BeBeC-2020-D21

DEEP NEURAL NETWORK MODELS FOR ACOUSTIC SOURCE LOCALIZATION

Pengwei Xu, Elias J. G. Arcondoulis and Yu Liu
Department of Mechanics and Aerospace Engineering, Southern University of Science and Technology
1088 Xueyuan Blvd, Nanshan, 518055, Shenzhen, P. R. China

Abstract

The localization of acoustic sources using acoustic imaging methods such as acoustic beamforming can be limited by the Rayleigh resolution limit at relatively low frequencies and the output of these methods may also produce spatial aliasing images referred to as sidelobes, particularly at high frequencies. To date, there are very few Deep Neural Network (DNN) applications for acoustic imaging to help alleviate some of these issues. In this study, we developed several DNN models using DenseNet-201 with a training strategy specifically designed for an acoustic imaging task. The DNN method is able to recognize the pattern behind microphone array signals for different source positions, by using the real-component of the cross-spectral matrix of the received pressure vectors at the microphone positions. DNN models with both a fixed and random number of input sources are simulated for a range of specific frequencies. The DNN model with a random number of input sources is tested against conventional beamforming, CLEAN-SC and DAMAS, revealing a far improved source localization and source strength estimation. These DNN models represent a very promising proof-of-concept for the use of DNN models in the field of acoustic imaging.

1 INTRODUCTION

Aeroacoustic sources, such as aircraft landing gear [6, 34], small-scale drones [38] and airfoil trailing edge noise [3], have gained significant attention in recent years leading to deeper research into the fundamental mechanisms responsible for their noise generation. Quantification and localization of the acoustic source regions are important stages in understanding the noise generation and thus the accurate mapping of sound sources in wind tunnel environments attracts considerable interest. Conventional beamforming (CB) [8, 17] is a popular and robust method, yet the acoustic maps it generates potentially suffer from spatial aliasing images (sidelobes) and poor resolution between acoustic sources defined by the Rayleigh resolution limit [26, 30]. Furthermore, the array design can be responsible for the quality of the acoustic source map, due to various acoustic frequencies and locations of the source relative to the source plane of investigation [4]. Various post-processing techniques have been developed to localize all possible sound sources and to significantly clear the acoustic map of sidelobes, such as CLEAN-SC [33], DAMAS [5, 27] and several other variations [15, 23]. One of the challenges of the existing acoustic imaging methods is the inability to resolve complicated sound source distributions, especially at low frequencies where the main lobe of an acoustic source can dominate the acoustic source map as a result of the Rayleigh resolution limit. Using an adaptation of CLEAN-SC, namely adaptive High-Resolution CLEAN-SC (HR CLEAN-SC), Luesutthiviboon et al. [24] were able to resolve acoustic sources that would be otherwise unable to be resolved using CB (beyond the Rayleigh resolution limit) yet this was achieved using an in-house optimized array design.

Due to the rapid development of computational power in recent years, Deep Learning (DL) [21] has become an emerging field of research in a wide range of applications. Specifically to this study, it potentially provides a fresh perspective to tackle existing problems in acoustic imaging, such as CB. The concept of DL was originated by the progress of artificial neural networks, where Hinton [13] proposed the concept of a Deep Neural Networks (DNN), attracting a tremendous amount of attention due to its remarkable performance for pattern recognition even with very limited labeled data [21]. The process of DL involves observing the characteristic or representation of data by each layer, among which a higher layer continues to learn from a lower layer. By learning or training DNNs, they are able to approximate such complex relationships within the data that could not be achieved otherwise.

The remarkable capability of DNNs has led to their use in a wide range of applications in mechanical systems and acoustic signal processing, such as machine health monitoring [10, 19, 32, 39], speech enhancement [1, 7], acoustic source localization under water [16, 18, 28, 36], bioacoustics [2, 11, 37] and many others. In regards to acoustic source localization, Vera-Diaz et al. [35] employed a convolution neural network for indoor acoustic source localization of a single source using a microphone array and Niu et al. [28] applied residual neural networks [20] using extremely large data sets (often referred to as big data) to locate broadband acoustic sources using a single hydrophone in an ocean environment. Despite these DNN applications, these studies do not reveal details of source distribution, including position and strength of multiple sources, which is the aim of this study. Based on an extensive search of recent literature as of 2019, Ma et al. [25] has applied a DNN method for acoustic imaging of multiple acoustic sources using an array of microphones, which is analogous to CB. They were able to achieve good performance in predicting three dispersed point sources of frequencies 3000 Hz and 8000 Hz by training convolutional neural networks yet due to the spacing of these sources and their frequency, they were not able to reveal any improvements in acoustic source resolution relative to the Rayleigh limit that restricts CB [26] but showed in most cases identical results to DAMAS.

In this paper, we propose a novel acoustic imaging method based on a densely connected convolutional network (namely DenseNet) [14] with a specifically designed training strategy. DenseNet is one of the most sophisticated DNN architectures with many compelling advantages. For example, in machine learning the vanishing gradient problem is a difficulty found in training neural networks with gradient-based learning methods and back propagation [31].

In some cases, the gradient, which is used to update the weights of the neural network, will be vanishingly small, and effectively prevent the weight from updating. DenseNet is able to alleviate this vanishing-gradient problem by constructing tight ensembles of many short networks together. In a standard convolutional network [22], input features pass through multiple convolution stages and obtain higher-level features. But in DenseNet, each layer obtains additional inputs from all preceding layers and passes on its own feature maps to all subsequent layers. The network becomes thinner and compact due to each layer receiving feature maps from all preceding layers and thus the number of parameters is substantially reduced.

A proposed DenseNet-based acoustic imaging method is presented. By training a DNN model with significant amounts of simulated acoustic source data that radiates toward a 64-channel logarithmic spiral microphone array that spans over a 1 m area, the DNN model is developed such that it is capable of locating single and multiple sources over a wide frequency range, commonly observed in aeroacoustic applications (200 Hz to 20,000 Hz) located 1.2 m from the array plane. The model is tested with a challenging acoustic source distributions, ranging from one to twenty-five acoustic sources over a 1 m × 1 m area. The DNN model locates these sources with minimal sidelobe contributions and acoustic source resolution that well exceeds the restrictions of the Rayleigh limit. A thorough investigation of the parameters that comprise the DNN model are also presented, such as the introduction of a Mean Absolute Error (MAE) to help quantify the errors in source localization quantification in acoustic source imaging. The MAE is seen to converge with increasing DNN training time for each acoustic source frequency investigated and its variation with number of acoustic sources within the map is also discussed. Overall, the DNN model presented here possesses far greater source resolution, accuracy and sidelobe characteristics than CB, CLEAN-SC and DAMAS.

2 Densely Connected Convolutional Networks

Generally speaking DNNs contain input layers, hidden layers and output layers where each layer is built with multiple neurons (units). In order to train a DNN model, a loss function is computed that measures the error (or distance) between the output and the true values. The machine then modifies its internal adjustable parameters to reduce this error by using a weighting function, **w**, for each connection between the units. The entries of **w** are real numbers that define the input-output function of the machine. For each unit k the output a_k [14] is defined as

$$a_k = F\left\{\sum_i a_i w_{ki} + \varepsilon\right\}, \qquad k = g, g+1, \ldots, K \tag{1}$$

where F is an activation function (see Eq. (9) for an example), i is the upstream unit index, w_{ki} is the weight of each path connecting unit k per i-th connection and a_i is the output of the i-th unit from upstream. The term ε is a small threshold constant, g is the number of input units and K is the total number of units in the model.

To help explain Eq. (1), an illustrative example of a multilayer neural network [21] is presented in Fig. 1. The open circles represent the units, the lines connecting the units represent the information transfer between units and the vertical column of units represents a layer. Every unit in the input layer (denoted as k = 1 to 8) sends information to the every unit in the first hidden layer (denoted as k = 9 to 14) via a complex web of connections. In Fig. 1 the unit k =

15 is arbitrarily highlighted to reveal its upstream connections, illustrated as blue lines. It can be seen that for $k = 15$ that the $I(k)$ has six elements, as there exist six upstream connections for $k = 15$. Each of those connections have weights $w_{15,i}$, where $i = 9$ to 14.

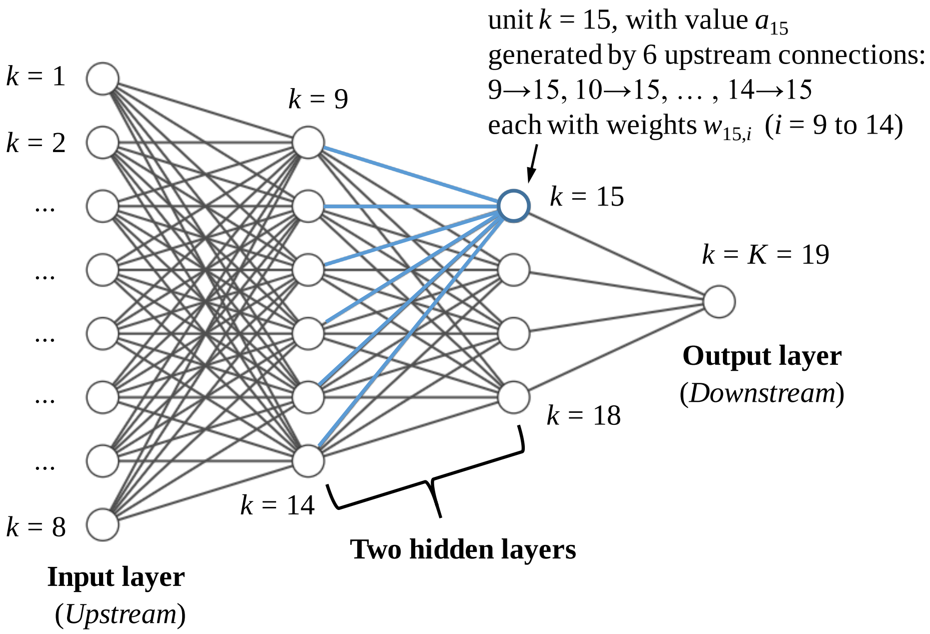

Figure 1: Illustrative example of a multilayer neural network, showing the various layers within the DNN required to achieve the output layer. In this basic example, K = 19 and the unit $k = 15$ *is highlighted in blue to help illustrate the concept of Eq. (1). Only two hidden layers are depicted here for illustrative purposes; DNNs typically can use hundreds of hidden layers (refer to Table 1 for the hidden layer architecture used in this study).*

From the basic illustrative example in Fig. 1, it can be seen that a DNN model with hundreds of hidden layers and input layers consisting of hundreds of units that the number of computations between each unit can be immense and that the selection of an appropriate DNN model for the behavioral dynamics to be investigated (in this case acoustic imaging) must be selected. Recently Huang et al. [14] introduced a unique type of DNN model, namely the Dense Convolutional Network (DenseNet). This unique DNN model ensures that the maximum possible information flow between the layers in the network is achieved, and that every layer is connected directly with each other. This leads to the advantageous consequence that DenseNet has fewer network parameters during computation relative to other network architectures. Therefore, based on the clear advantages of its network structure and its well-recognized performance the DenseNet-201 architecture (where 201 denotes the number of layers within the DNN model) is employed in this study for acoustic imaging.

3 Simulation Methodology

A comparison between acoustic imaging using a DNN and existing acoustic beamforming techniques is presented in this study and this section details the simulation methodology. Firstly the principles of acoustic beamforming are introduced, as the synthetic data generation for the DNN model follows the spherical wave propagation equation needed for acoustic beamforming sound source generation and steering vector formulation. The DNN model architecture is then explained based on the array sizes and dimensions of the acoustic data produced from the synthetic noise generation required for acoustic beamforming.

3.1 Acoustic Beamforming

Multiple spherical (monopole) wave sources, s, of unit source strength in still conditions are simulated. The propagation of a wave from each acoustic source s to a microphone m on an array plane (where $m = 1$ to M) can be represented as $p_s(m)$, which is a complex pressure (Pa) in the frequency domain defined as

$$p_s(m) = \frac{\mathrm{e}^{-\mathrm{j}2\pi f \mathbf{r}_s / c_0}}{4\pi \left|\mathbf{r}_s\right|} \tag{2}$$

where c_0 is the speed of sound in air (343 m/s). The vector $\mathbf{r}_s$ links the simulated source s location to each of microphone m in the array plane. The total pressure contribution at microphone m due to s sources is calculated via the sum of the sound propagation of each source (Pa)

$$\mathbf{p} = \left\{ \sum_1^s p_s(1), \sum_1^s p_s(2), \ldots, \sum_1^s p_s(M) \right\} \tag{3}$$

where $\mathbf{p}$ is an $M \times 1$ vector that is used to produce a Cross-Spectral Matrix (CSM), C [17, 26], which is an $M \times M$ matrix (Pa^2) defined as

$$C = \mathbf{p}\mathbf{p}^H \tag{4}$$

where H represents the complex transpose and conjugate. The beamforming output is calculated over a square-planar discretised grid of N data points (herein referred to as the scanning grid) at a known distance from the array, z, positioned in line with the center of the microphone array. Steering vectors, $\widehat{\mathbf{v}}$, relate the propagated pressure from an assumed scanning grid point source to each microphone, m. One possible and commonly used steering vector formulations for the m^{th} microphone is an $N \times 1$ array defined as

$$\widehat{\mathbf{v}} = \frac{\mathrm{e}^{-\mathrm{j}2\pi f \mathbf{r}_m / c_0}}{4\pi \left|\mathbf{r}_m\right|} \tag{5}$$

where $\mathbf{r}_m$ is the vector between the scanning grid point to the microphone m. The cross-spectral beamforming output (i.e., the output of CB, Y) [5] is computed using

$$Y(\widehat{\mathbf{v}}) = \frac{\widehat{\mathbf{v}}^H C \widehat{\mathbf{v}}}{M^2 - M} \tag{6}$$

The scanning grid in this study is a square grid of N points on a plane, located z = 1.2 m from the array plane. The distance between two points in the x- and y-directions are the same and are referred to as Δx. To remove any concerns regarding the quality of an array design, when the DNN performance is of chief importance in this study, a typical logarithmic spiral array possessing M = 64 channels is used here [4, 9, 29]. The array spans an approximate 1 m × 1 m area with an aperture of D = 1.03 m. To ensure that the array performs well over a range of frequencies (from approximately frequencies f = 800 Hz to 20,000 Hz), a densely spaced region of microphones is required near the array geometric center (for high frequency performance) and some microphones are required near the array outer perimeter (a maximized aperture assists low frequency performance) [4, 29]. The array design and a schematic diagram representing the array location relative to the source plane are presented in Figs. 2(a) and (b), respectively.

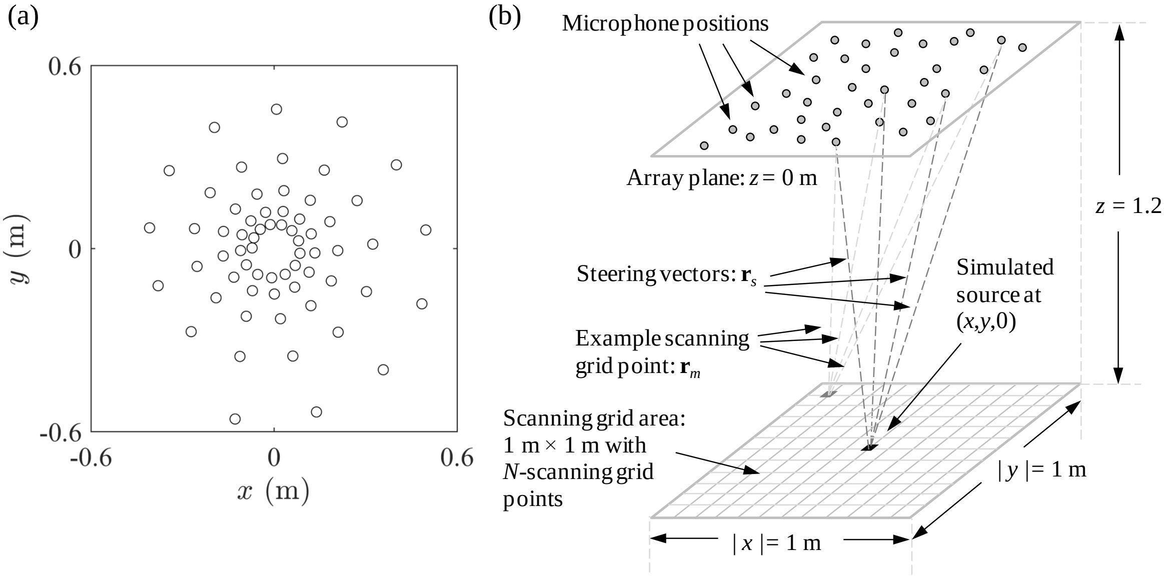
(a)
0.6
0
-0.6
-0.6
0
0.6
y (m)
x (m)
(b)
Microphone positions
Array plane: z = 0 m
Steering vectors: $\mathbf{r}_s$
Example scanning grid point: $\mathbf{r}_m$
Simulated source at (x,y,0)
z = 1.2 m
Scanning grid area: 1 m × 1 m with N-scanning grid points
$|y|$ = 1 m
$|x|$ = 1 m

Figure 2: (a) Logarithmic-spiral array pattern containing M = 64-channels used for the acoustic beamforming and DNN synthetic data generation (b) Schematic diagram representing the beamforming array plane and scanning grid dimensions.

The source resolution performance of an acoustic beamformer output can be estimated by comparing its resolution against the Rayleigh Resolution Limit (RRL) which represents the angular resolution for an optical system [30]. For the case of acoustic imaging based on Fig. 2(b) it can be represented as a length term (m)

$$\mathrm{RRL} \approx 1.22 z \frac{c_0}{fD} \tag{7}$$

where it is assumed that $z >>$ RRL to satisfy a small-angle approximation ($\sin\theta \approx \theta$). It can be seen that the spatial resolution of CB will therefore be proportional to z and inversely proportional to f. Consequently we define a dimensionless parameter directly related to the CB output, namely the Rayleigh Resolution Ratio (RRR) as

$$\mathrm{RRR} = \frac{\Delta x}{\mathrm{RRL}} \tag{8}$$

To provide the best comparison between acoustic beamforming and DNN acoustic imaging, we apply the well-known deconvolution methods CLEAN-SC and DAMAS to Y to minimize sidelobe contributions and reduce the main lobe width of the acoustic source. These results are compared against CB and the output of the DNN acoustic imaging model.

3.2 DNN Model

In this study, we employ the typical architecture of a DenseNet-201 DNN model as developed by Huang et al. [14]. The input layer, classification layer, and output size of DenseNet are adjusted according to the input feature of microphone array signal and resolution of scanning grid. Due to recent advances of the DenseNet method, it has been implemented successfully in discovering intricate structures in high-dimensional data and thus can be applied to obtain microphone array data and relate the information into sound source locations.

Input Feature

The same procedure to simulate acoustic sources for typical acoustic beamforming array design [4] is used here. The pressure contribution at each scanning grid point is calculated as the sum of each simulated pressure vector $\mathbf{p}$ defined in Eq. (3). For the purpose of comparison between the acoustic imaging methods, the microphone array is also identical to the array used for acoustic beamforming calculations and so is the relative position of the scanning grid plane to the array plane, z = 1.2 m.

Due to the convenience of synthetic data-set generation, we apply a dynamic training data strategy. The ground truth of the DNN model (i.e., the value which the DNN model is trained upon) is defined as $\mathbf{q}_0$, which is the randomly generated source strength (each source possesses equal source strength) over the scanning grid ($N \times 1$ vector). The acoustic propagation of the $\mathbf{q}_0$-sources towards the array plane produce pressures $\mathbf{p}$ detected by each microphone m (which allows us to predict the received microphone signal of a random sound source distribution). The sparseness of $\mathbf{q}_0$ (i.e., how many sources exist in the scanning grid) varies per DNN model, as discussed in Section 4.

The CSM is chosen to be the input feature of the proposed DNN model due to its ability to reveal the intrinsic structure of the microphone array signals. From extensive preliminary testing, it was revealed that the real-component of the CSM was the simplest term that contained sufficient information required to conduct DNN training. Thus the DNN model input feature is $\Re\{C\}$.

The training data is input in batches, referred to as an epoch. In each epoch the DNN uses 512-unique randomly generated source patterns (i.e., 512-unique $\mathbf{q}_0$ vectors) to optimize the weighting functions w within the hidden layers based on gradient [14]. Once this training is complete, another randomly generated set of 512-unique $\mathbf{q}_0$ vectors are produced and are used as the input layer. This allows the training data-set to dynamically grow as the model continues to fit data and produce a massive number of acoustic source patterns to be learned by the DNN model. Typically several thousand epochs are run to train the DNN model, which is discussed in detail in Section 4.1.

Hidden Layers

There are 199-hidden layers (= 201−Input layer−Output layer) between the input and output layers and they are presented in Table 1. The pooling layers represent a sample-based discretisation process [12], the objective is to down sample feature maps by summarizing the presence of features in patches of the feature map and the Dense Block is a set of convolutional layers. In Table 1, a 1 × 1 conv means a convolutional layer with a 1 × 1 kernel size [22] and subsequently a 3 × 3 conv means a convolutional layer with a 3 × 3 kernel size. The term 3 × 3 max pool, stride 2 refers to a max pooling layer with 3 × 3 kernel size and striding 2 steps each time. These parameters are consistent with Huang et al. [14]. The activation function F in Eq. (1) used here is simply a Rectified Linear Unit (ReLU) activation function, such that

$$F(a_k) = \begin{cases} a_k & \text{if } a_k \geq 0 \\ 0 & \text{if } a_k < 0 \end{cases} \tag{9}$$

Table 1: Typical architecture of DenseNet-201 (using a growth rate of 32) [14] showing each of the hidden layers between the input and output layers. Each conv layer shown in the table corresponds to the sequence Batch Normalization (BN) - ReLU - Convolutional layer.

Hidden Layers	Output Size	DenseNet-201
Convolution	112 × 112	7 × 7 conv, st ride 2
Pooling	56 × 56	3 × 3 max pool, stride 2
Dense Block (1)	56 × 56	$\begin{bmatrix} 1\times1 \text{ conv} \\ 3\times3 \text{ conv} \end{bmatrix} \times 6$
Transition Layer (1)	56 × 56 28 × 28	1 × 1 conv 2 × 2 average pool, stride 2
Dense Block (2)	28 × 28	$\begin{bmatrix} 1\times1 \text{ conv} \\ 3\times3 \text{ conv} \end{bmatrix} \times 12$
Transition Layer (2)	28 × 28 14 × 14	1 × 1 conv 2 × 2 average pool, stride 2
Dense Block (3)	14 × 14	$\begin{bmatrix} 1\times1 \text{ conv} \\ 3\times3 \text{ conv} \end{bmatrix} \times 64$
Transition Layer (3)	14 × 14 7 × 7	1 × 1 conv 2 × 2 average pool, stride 2
Dense Block (4)	7 × 7	$\begin{bmatrix} 1\times1 \text{ conv} \\ 3\times3 \text{ conv} \end{bmatrix} \times 48$
Classification Layer	1 × 1	7 × 7 global average pool 1000D fully-connected, softmax

Output and Loss Function

The output layer is defined as $\mathbf{q}$ which is the estimated source strength (Pa) over the scanning grid ($N \times 1$ vector). During post-processing (outside of the DNN model) to assist visualizing the source map over a square area, this vector is reshaped into a square matrix of size $\sqrt{N} \times \sqrt{N}$.

In neural networks, a loss function is used to examine the model prediction errors. During the training of neural networks, the loss function will be minimized constantly by adjusting the weights w. In this study, the term MAE is chosen to be the loss function, which is the difference between the ground truth and the source map predicted by the DenseNet model, divided by the maximum input strength of the ground truth (1 Pa). Therefore MAE is a singular non-dimensional value defined as

$$\mathrm{MAE} = \frac{|\overline{\mathbf{q}} - \overline{\mathbf{q}}_0|}{\mathbf{q}_{0,\max}} \tag{10}$$

where $\overline{\mathbf{q}}$ represents the mean source strength over the scanning grid predicted by DenseNet model (Pa), $\overline{\mathbf{q}}_0$ is the mean ground truth source strength (Pa) over the scanning grid and $\mathbf{q}_{0,\max}$ is the maximum input strength of the ground truth (1 Pa). Note that the elements of $\mathbf{q}$ correspond to the a_K values depicted in Fig. 1. The output layer consists of N units (note that for simplicity purposes the illustration in Fig. 1 shows only one unit in the output layer) and effectively $\mathbf{q} = \{q(1), q(2), \ldots, q(N)\} = \left\{a_{K,1}, a_{K,2}, \ldots, a_{K,N}\right\}$. Note that the singular MAE value in Eq. (10) can then be averaged over a range of test cases. In this case, the averaged MAE is referred to as $\mathrm{MAE}_{\mathrm{av}}$.

Model Summary

To summarize the main characteristics of the DNN model: the feature input is the real-part of the CSM ($\Re\{C\}$, Pa^2) of size $M \times M$, the output is the acoustic source strength $\mathbf{q}$ (Pa) of size $N \times 1$ and the loss function is the difference between the estimated source strength and the ground truth, being the singular value of MAE as defined in Eq. 10. The input consists of an epoch which contains 512-$\mathbf{q}_0$ vectors run as a batch. Once the MAE is calculated, it is then passed back into the hidden layers to adapt the weights w, which is referred to as a training cycle (ten training cycles are applied per epoch). Each of these terms and processes can be visualized in Fig. 3. In this paper, a number of DenseNet models were built for different purposes and compared with each other, all of which were trained within a few hours on two NVIDIA Tesla v100 GPU PCs.

4 Results and Discussion

4.1 DNNs Training Convergence and Prediction Accuracy

A training epoch can be regarded as a unit of time, assuming that each epoch is of the same size and array sparseness. It is important to determine how many epochs are required to achieve converge of MAE, as training a DNN model with many hidden layers and large input structures can be very computationally expensive. At the end of each epoch, one-thousand (1000) $\mathbf{q}_0$ vectors possessing six randomised sources (i.e., six random non-zero entries equal to one) would be generated to validate the DNN model at this point in its training life. The values of

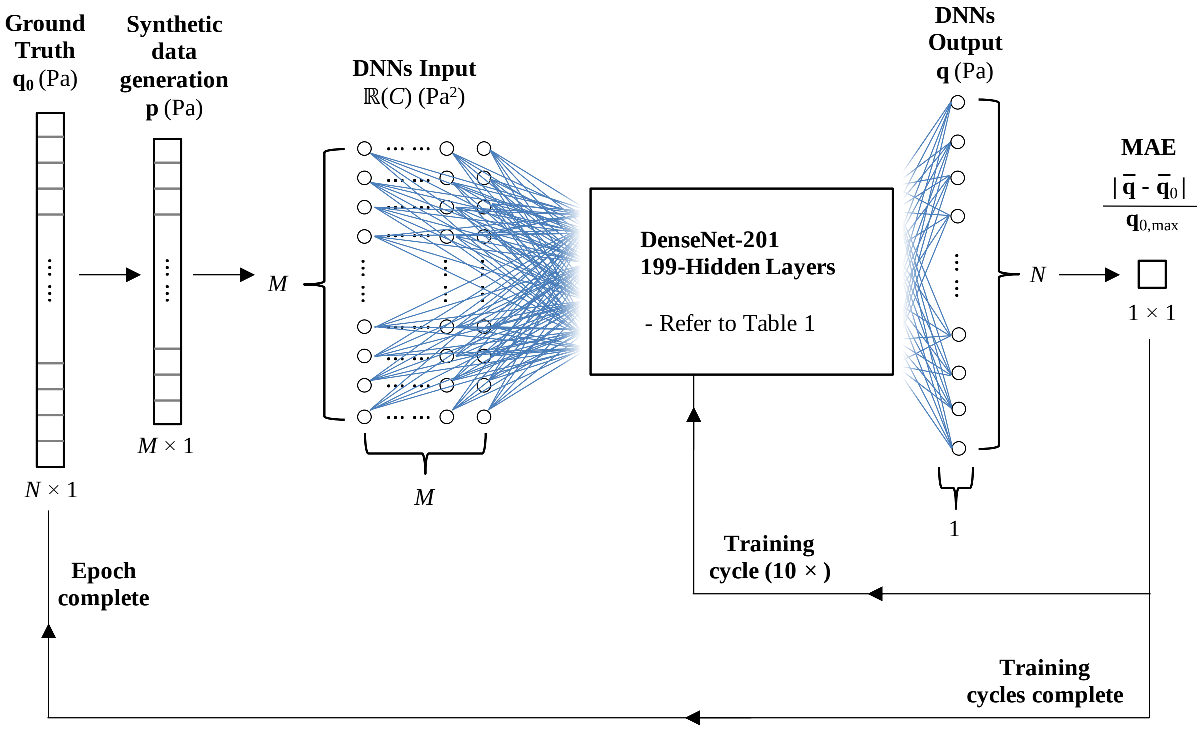

Figure 3: A schematic diagram of the DNN model used in this study.

$\mathrm{MAE}_{\mathrm{av}}$, averaged with respect to these 1000 validation data, are used to present the accuracy of prediction of the DNN model.

Figure 4 presents the progress of $\mathrm{MAE}_{\mathrm{av}}$ with respect to the number of training epochs for various acoustic frequencies. Each curve represents a different DNN model, trained specifically for that acoustic frequency. It can be observed in Fig. 4 that the DNN models trained for higher frequencies appear to converge using fewer epochs and possess lower $\mathrm{MAE}_{\mathrm{av}}$ values. This indicates that it is easier for a DNN model to recognize and be trained using a CSM comprised of higher frequency pressure vectors. Note this study was conducted for every DNN model over a wider range of frequencies than presented in Fig. 4, being f = 100, 200, 300, 400, 500, 600, 800, 1000, 2000, 5000, 8000 and 20,000 Hz.

To formally quantify whether a DNN model has converged for a specific frequency, a convergence criteria is introduced for $\mathrm{MAE}_{\mathrm{av}}$ defined as

$$\mathrm{MAE}_{\mathrm{av},n} - \mathrm{MAE}_{\mathrm{av},n-100} < 0.001 \tag{11}$$

where n represents the number of epochs simulated during training and $n > 100$. Using this criterion, the number of required epochs to achieve convergence of $\mathrm{MAE}_{\mathrm{av}}$ for each frequency were recorded and used as an end-point for DNN training.

Acoustic source maps using the DNN models are presented in Fig. 5. These maps are also generated using 1000 randomly-generated validation data, to ensure that the source maps (i.e., $\mathbf{q}$) are averaged and are representative of the DNN performance. The ground truth values, $\mathbf{q}_0$, are varied for each frequency-specific DNN model, to showcase a wider range of the DNN

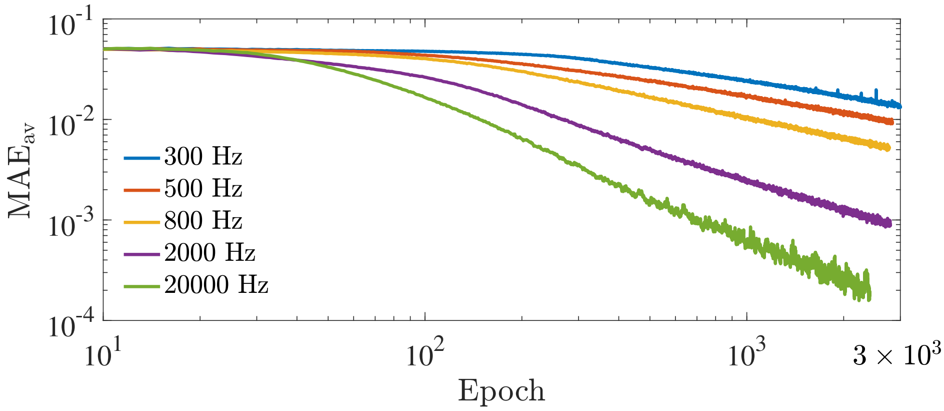

Figure 4: $\mathrm{MAE}_{\mathrm{av}}$ *recorded using 1000 randomly-generated validation data versus training epoch. Each curve corresponds to the* $\mathrm{MAE}_{\mathrm{av}}$ *values of the DNN model that is trained specifically for the frequency shown and for six random sources in* $\mathbf{q_0}$.

source prediction capability. Note that all of the entries within $\mathbf{q}_0$ are either 0 or 1, so that the ground truth source locations in Fig. 5 can simply be identified as a cyan dot. The predicted source strengths are shown in color and their source strengths are presented in dB, normalized to the true source strength, being unity. Recall that each of the DNN models investigated here are only capable of localizing exactly six monopole sources with unit source strength at one specific frequency.

In Figs. 5(f) through 5(i) both the source location and strength are perfectly estimated, corresponding to $f = 2000$ Hz to 20,000 Hz, respectively. Excellent acoustic imaging performance at higher frequencies is relatively common, especially with the use of CLEAN-SC and DAMAS upon post-processing CB data, Y. These methods can typically resolve the acoustic sources and significantly reduce any sidelobes [5, 33]. The key observation from Fig. 5 is the impressive source localization capability at lower frequencies. Acoustic frequencies such as $f = 200$ Hz to 800 Hz are relatively low frequencies for acoustic beamforming, assuming the parameters D, z and $|x| \times |y|$ as defined in Figs. 2(a) and 2(b). Nonetheless, using the frequency-specific DNN models within this frequency range, reasonable source location and strength estimation can be observed in Figs. 5(a) through 5(d). At $f = 200$ Hz to 600 Hz, some of the acoustic sources are not perfectly estimated yet the estimated source locations still indicate a source in the true region and the map is still a reasonable representation of the ground truth. Technically the misplaced sources are not sidelobes; they can be regarded as a slightly smeared main lobe area, rather than an exact single source location. Note that even at $f = 400$ Hz every source is located. At slightly higher frequencies of $f = 800$ Hz to 1000 Hz, as shown in Figs. 5(d) and 5(e) respectively, every source is accurately located while in some cases the acoustic source strengths are not equal to unity (and thus not equal to 0 dB). The greatest source strength error in Fig. 5 is approximately -8 dB, as observed in Fig. 5(d).

By observation of each source map in Fig. 5 we can observe that the source map at $f = 800$ Hz in Fig. 5(d) presents the lowest frequency acceptable source map. This is based on the lack of sidelobes observed and overall good quality sound source strength quantification. At $f = 600$ Hz as shown in Fig. 6(c), the source map is still reasonable yet the smearing of the main

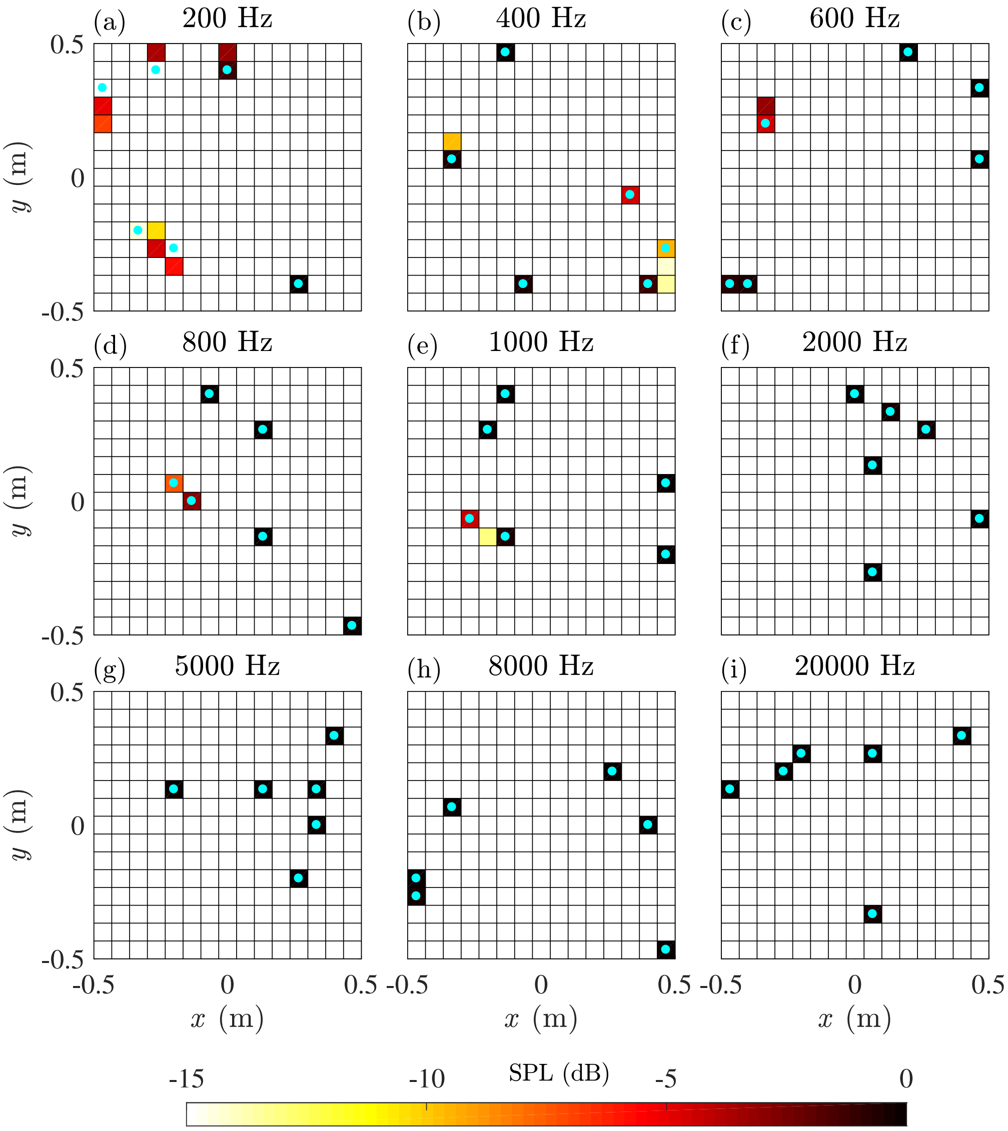

Figure 5: Application of the DNN model to locate various acoustic point sources for a range of different frequencies, with results presented at (a) 200 Hz to (i) 20,000 Hz. Each map possesses six acoustic sources at unique positions and the dB scale is defined relative to a unit strength source. The scanning grid consists of $N = 15 \times 15 = 225$ points.

lobe at $x = -0.3$ m and $y = 0.3$ m makes it unsuitable, which is a conservative estimate. By assuming that $f = 800$ Hz is the threshold for high quality DNN model performance, we can then determine an $\mathrm{MAE}_{\mathrm{av}}$ allowable upper limit.

An investigation of MAE with respect to each DNN model trained for a specific frequency was conducted where the frequency was converted into a RRR value using Eq. (8), as presented in Fig. 6. It can be observed that the $\mathrm{MAE}_{\mathrm{av}}$ of each DNN model decreases with respect to RRR. Similarly, each $\mathrm{MAE}_{\mathrm{av}}$ value was calculated using 1000 randomly-generated validation data. At $0.012 < \mathrm{RRR} < 0.212$, the $\mathrm{MAE_{av}}$ values are seen to slightly decrease with frequency yet there are significant changes of $\mathrm{MAE_{av}}$ beyond this RRR range. It appears that RRR $\approx$ 0.12 is a RRR cut-off value for the DNN models investigated here in terms of $\mathrm{MAE_{av}}$. A fitted curve of $\mathrm{MAE_{av}}$ and RRR for $0.012 < \mathrm{RRR} < 0.212$ is included, showing that there exists an inverse relationship between these parameters. Note that the constant of this empirically fitted relationship would specific to the conditions, such as $\Delta x, D$ and z.

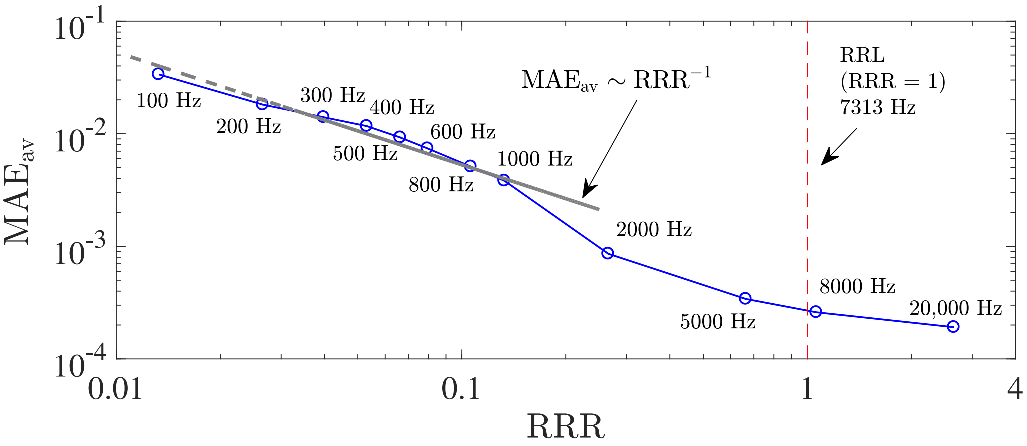

Figure 6: A curve of converged $\mathrm{MAE_{av}}$ *values with respect to RRR. Each data point represents a DNN model that is specifically trained for corresponding RRR frequency. Each* $\mathrm{MAE_{av}}$ *value is calculated using 1000 randomly-generated validation data with six sources in* $\mathbf{q}_0$. *The data are calculated based on* Δx *= 1/15 m. A fitted curve between* $\mathrm{MAE_{av}}$ *and RRR is shown as a dashed line.*

Values of RRR ≥ 1 imply that the acoustic sources can be well resolved via acoustic beamforming techniques. If RRR < 1, it would be difficult to accurately calculate the acoustic source distribution by conventional beamforming methods. As $\Delta x = 1/15$ m and $z = 1.2$ m are fixed in Eqs. (7) and (8), the resolution between acoustic sources can only be improved at higher frequencies via conventional methods. By basing the lower frequency threshold at f = 800 Hz (RRR = 0.11), we see that the corresponding $\mathrm{MAE_{av}}$ value is 5×10^{-3} which represents the upper limit of acceptable DNN model performance. This also implies that high quality DNN model performance occurs at a frequency that is approximately one order of magnitude smaller than RRL. This shows significant promise in terms of source resolution capability relative to existing acoustic beamforming techniques.

4.2 Influence of Source Number

All of the presented DNN results have been based on a trained model using a fixed number of sources (six sources) in $\mathbf{q}_0$. It is important to determine the influence of the source number (i.e., the number of non-zero entries in $\mathbf{q}_0$) on the performance of the DNN output, $\mathbf{q}$ and thus $\mathrm{MAE_{av}}$. To conduct an analysis of $\mathrm{MAE_{av}}$ with varying source number, N_s, a DNN model was made for a specific number of sources N_s. These models, namely fixed input DNN models, were trained for a single specific frequency to reduce overall computation time. In addition, a more complex DNN model was created that can localize a random number of sources, namely a random input DNN model. The upper limit of sources that it is trained to detect is twenty-five (25). This model, while much more complex to simulate, provides a much more realistic and practical DNN model, as the number of sources to be investigated can rarely be known prior to conducting experiments. It is important to determine a specific frequency that represents the challenge of accurately estimating $\mathbf{q}_0$ for both of the models. By observation of Fig. 5, f = 800 Hz was selected, as it revealed primarily accurate source localization and yet not a perfect estimation of the source strengths.

Figure 7 reveals both DNN model performances in terms of MAE_av with respect to N_s. Note that the data point at $N_s = 6$ for the fixed input DNN corresponds to the same data presented in the source map in Fig. 5(d). For both models, each MAE_av value was calculated using 1000 randomly-generated validation data per N_s-DNN model.

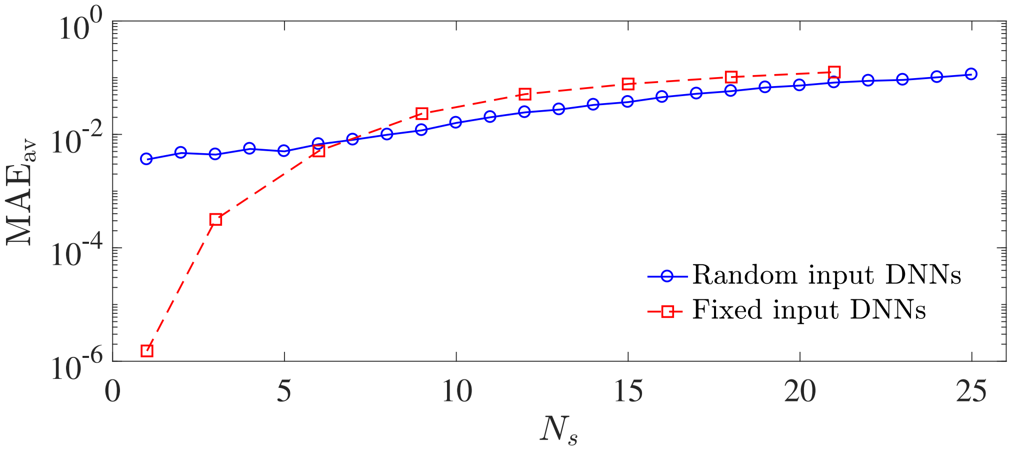

Figure 7: MAE_av *curves with increasing source number N_s, using both a fixed and random input f = 800 Hz DNN model. Each fixed input DNN data point represents a model that is specifically trained for the corresponding N_s value whereas only one random input DNN model is used for all N_s values presented. For both DNN models, the* MAE_av *values were calculated using 1000 randomly-generated validation data.*

From Fig. 7 it can be seen that for both DNN models MAE_av increases with N_s, which is expected due to the increased complexity of the ground truth (fewer non-zero values in $\mathbf{q}_0$) and thus the increased complexity of relaying the CSM input layer through the hidden layers of each DNN model. Note that from MAE_av alone, it cannot be determined whether the increased difference in prediction error between $\mathbf{q}$ and $\mathbf{q}_0$ is due to errors in the main lobe width and/or sidelobe contributions. Despite the increase of MAE_av with respect to N_s, it can be seen that the average error of each scanning grid point never exceeds 10^{-1} (i.e., 10%, by using acoustic sources of unity strength). For fewer sources, the fixed input DNN yields a lower MAE_av, yet for N_s = 6 to 7, the random input DNN possesses lower MAE_av values. Therefore the additional computational requirements to produce a random input DNN model appear to be justified. Note that this comparison is only conducted at a single frequency of f = 800 Hz. The ultimate goal, and will be future work, is to develop a random input DNN model for a frequency range. This model will require significant computational power to train and create yet it will be a far more practical and robust alternative to the models presented in this study thus far.

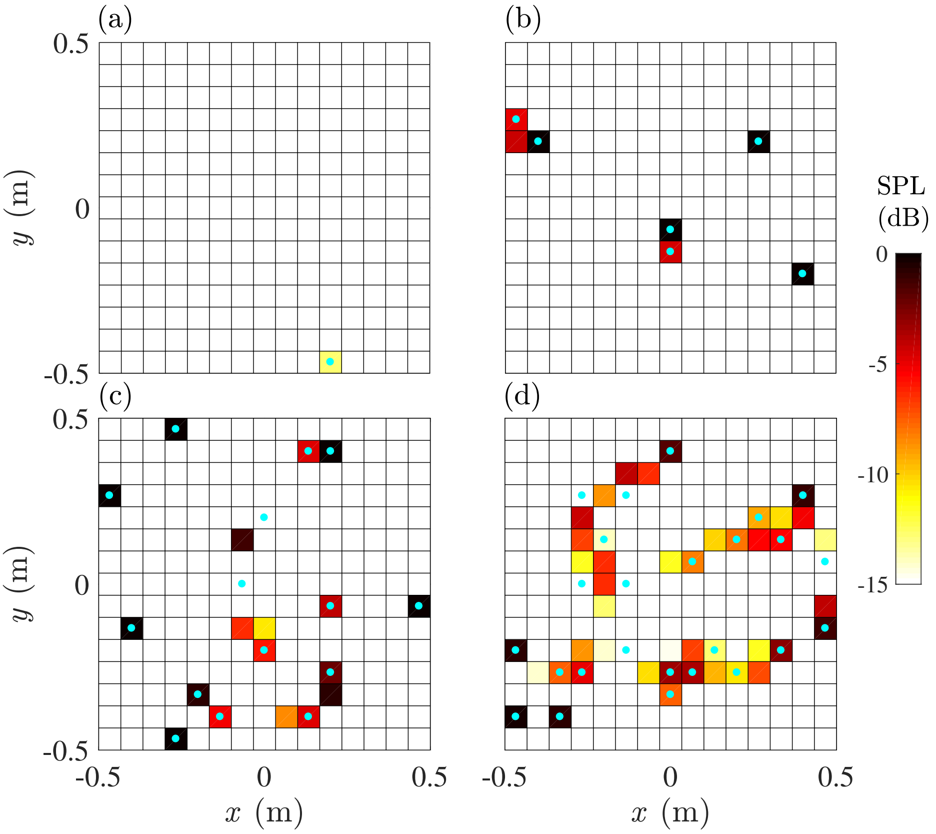

Figure 8: Application of the random input DNN model to locate various acoustic source distributions at f = 2000 Hz for different source numbers, N_s (a) 1, (b) 6, (c) 15 and (d) 25. The dB scale is defined relative to a unit strength source. The scanning grid consists of $N = 15 \times 15 = 225$ points.

The random input DNN model is now tested over a range of source locations and source numbers at f = 2000 Hz, as presented in Fig. 8. This model differs from the fixed input DNN model in that a randomized number of sources are simulated such that the model has no prior knowledge of the number of sources to detect, thus depicting a more realistic acoustic testing scenario. Four cases are considered, where one (1), six (6), fifteen (15) and twenty-five (25) sources are simulated and tested, as presented in Figs. 8(a), (b), (c) and (d), respectively. Figure 8(a) reveals that for a simple single source case, the source is accurately located without any sidelobes. Despite the accurate source localization, the DNN model has not accurately estimated its true source strength (the true source strength is 0 dB whereas it is estimated as -13 dB). Such misrepresentation of the acoustic source strength is less noticeable for the more realistic, multiple source cases, such as in Fig. 8(b) where all six sources are accurately located. The source strength values vary from 0 dB to -6 dB and the only observable error is some main lobe smearing at $x = -0.5$ m and $y = 0.2$ m. This is due to the proximity of the two sources that are located within 45° to each other, leading to some spreading of the source location over an adjacent scanning grid point. Increasing the number of simulated sources to fifteen, as revealed in Fig. 8(c), shows on average excellent source location capability of the randomized DNN model. Two sources exist near the center of the scanning grid, that are estimated by a single source that is approximately equispaced between them. All of the other sources are accurately estimated yet some of the sources near the lower part of the scanning grid show some distortion. The acoustic source strengths are well captured and range from 0 dB to -6 dB. The most complex source distribution and thus the greatest challenge for the randomized DNN model is shown in Fig. 8(d) where twenty-five sources are simulated. It can be casually observed that the majority of the acoustic sources are accurately located by the randomized DNN model. In cases where the sources are not accurately located, such as $x \approx -0.25$ m and $y \approx 0$ m to 0.3 m, the estimated source locations lay between the true sources. It is postulated that if a more refined scanning grid were to be used, such as a $\sqrt{N} \times \sqrt{N} = 50 \times 50$ grid, then some of these sources may be more accurately estimated. To summarize, the randomized DNN model is a more practical model for acoustic testing, in that the number of true acoustic sources will be unknown, yet is possess a small increase in sidelobe and source estimation errors relative to the DNN models that specifically seek a number of sources.

4.3 Acoustic Beamforming Comparison

The results using the random input DNN model are compared against the acoustic beamforming techniques discussed in Section 3.1 to provide some benchmark against the performance of the DNN models. Two frequencies are considered, f = 800 Hz and 2000 Hz. In Fig. 9(a), at f = 800 Hz CB is unable to accurately locate the sources as the RRR value is much less than 1 at this frequency. By passing the CB data through CLEAN-SC, it can be observed that much better source localization can be obtained, as shown in Fig. 9(b). Similarly, DAMAS attempts to locate the sources as shown in Fig. 9(c). However, neither of these deconvolution methods are capable of accurately locating the sources, with CLEAN-SC yielding the better result. The random input DNN model however locates all of the sources accurately and only shows some small main lobe smearing for the two adjacent sources near the center of the scanning grid, as observed in Fig. 9(d). Clearly at f = 800 Hz, the DNN model far outperforms any of the beamforming techniques considered here.

At f = 2000 Hz, the CB map provides a better representation of the acoustic sources as compared to f = 800 Hz. By observation of Fig. 9(e), the CB is even able to separate the source located near the upper-left corner of the map from the other sources. Using CLEAN-SC at f = 2000 Hz reveals a far improved result relative to the CB map, as shown in Fig. 9(f). Interestingly, DAMAS fails to recognize the source locations consistently as can be seen in Fig. 9(g) yet it shows some improvement relative to the result at f = 800 Hz. It is suspected that this may be due to the relationship between the parameters D, z, f and the size of the scanning grid $|x| \times |y|$, that inhibits the performance of DAMAS under these conditions. The random input DNN model again shows excellent performance, where the MAE = 0, as presented in Fig. 9(h). All of the sources are accurately located and their strengths are all 0 dB (i.e., the unit strength has been recovered in $\mathbf{q}$). While these tests reveal that the random input DNN model can locate sources well for two different frequencies with sources scattered over a scanning grid, the comparison between this DNN model and the beamforming techniques is only an indication of the expected performance over a wider range of source location conditions. It is future work to present a more comprehensive set of acoustic beamforming data over which the DNN model

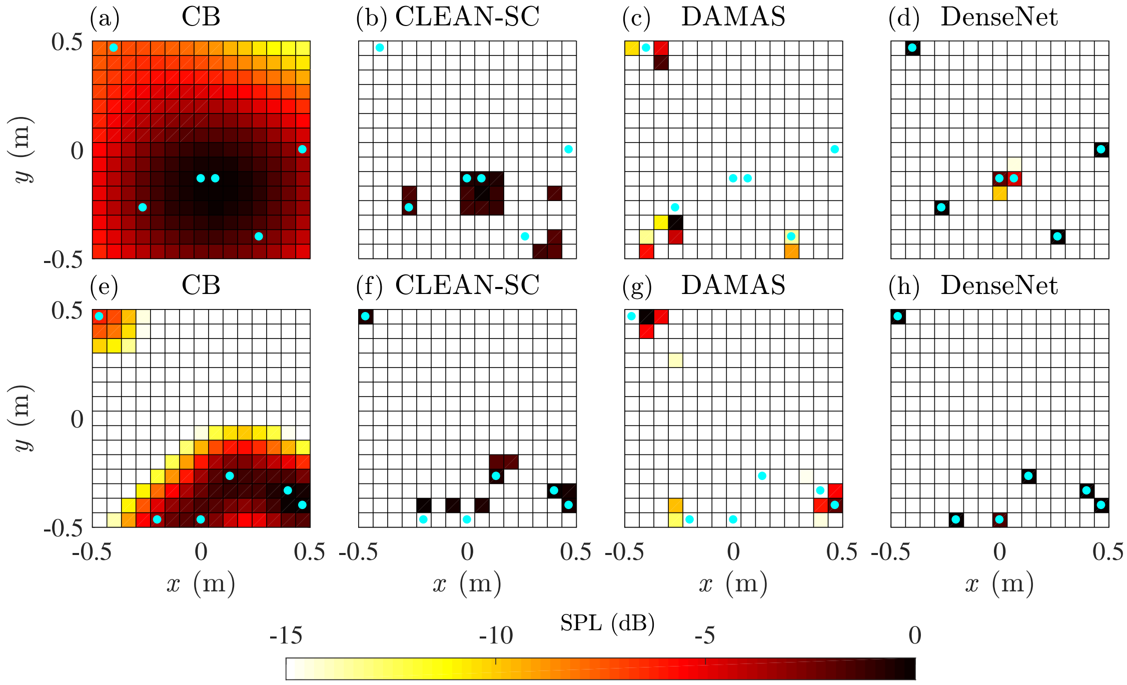

Figure 9: Comparison between (a) (e) CB, (b) (f) CLEAN-SC, (c) (g) DAMAS and (d) (h) the random input DNN model. Images (a) through (d) are conducted at f = 800 Hz and images (e) through (h) are conducted at f = 2000 Hz. The dB scale is defined relative to the maximum beamformer output for the beamforming techniques, and relative to a unit strength source for the DNN model. The scanning grid consists of $N = 15 \times 15 = 225$ points.

can be comparatively tested.

5 Conclusions

One of the challenges of conventional acoustic imaging techniques such as acoustic beamforming is the difficulty of resolving complex acoustic source distributions especially at relatively low frequencies. Deconvolution methods such as CLEAN-SC and DAMAS are capable of significantly improving the conventional beamforming source map yet sometimes fail to recognize every acoustic source and accurately quantify them. As an alternative to these methods, a deep learning approach was presented in this paper. The DNN is developed from DenseNet-201, a recently developed DNN model that has been previously implemented for large scale deep learning models. To successfully develop a DNN model, the input feature must be clearly defined and contain necessary information for the model to be trained. From preliminary investigations, it was found that the real-component of the CSM sufficed. This was a convenient input feature that corresponds directly with acoustic beamforming processes. The hidden layers within the DNN model were identical to those of the previously published DenseNet-201 model. The output layer was a vector of source strengths that map to a scanning grid, presented in a similar manner to an acoustic beamforming result. To determine the validity of a DNN model for acoustic imaging, a range of DNN models were created for specific frequencies ranging from 100 Hz to 20,000 Hz. Using the DNN models trained for a specific frequency and for six acoustic sources, a resolution of approximately an order of magnitude below the Rayleigh resolution limit can be obtained which is a significant improvement from existing acoustic imaging methods. Furthermore, the DNN models show superior source location capability relative to existing acoustic beamforming methods over a range of frequencies and challenging source distributions. A random input DNN model was also created that was capable of locating a random number of sources between one and twenty-five with no prior inputs defining the expected source number, which acted as a significant challenge for the DNN model.

The presented results are very promising but are not a completed work. Future work will include developing a more generalized DNN model that is capable of detecting a random number of sources for a range of acoustic frequencies. The current model can detect a random number of sources but must be only trained for a single frequency. As there are several key parameters that define the performance of acoustic beamforming (such as the array aperture, distance of the source to the array plane, number of microphones, etc) much more work is required to develop a complete DNN model. Nonetheless, the results presented in this study show a significant improvement in this current field of study and serve as a promising proof-of-concept.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant No. 11772146) and the Science, Technology and Innovation Commission of Shenzhen Municipality (Grant Nos. JCYJ20170817110605193 and ZDSYS201802081843517).

References

[1] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen. "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks." *IEEE Journal of Selected Topics in Signal Processing*, 13(1), 34–48, 2019.

[2] O. M. Aodha, R. Gibb, K. E. Barlow, E. Browning, M. Firman, R. Freeman, B. Harder, L. Kinsey, G. R. Mead, S. E. Newson, I. Pandourski, S. Parsons, J. Russ, A. Szodoray-Paradi, F. Szodoray-Paradi, E. Tilova, M. A. Girolami, G. J. Brostow, and K. E. Jones. "Bat detective—deep learning tools for bat acoustic signal detection." *PLOS Computational Biology*, 14(3), 2018.

[3] E. Arcondoulis, C. Doolan, A. Zander, L. Brooks, and Y. Liu. "An investigation of airfoil dual acoustic feedback mechanisms at low-to-moderate Reynolds number." *Journal of Sound and Vibration*, 460, 114887, 2019. doi:10.1016/j.jsv.2019.114887.

[4] E. Arcondoulis and Y. Liu. "An iterative microphone removal method for acoustic beamforming array design." *Journal of Sound and Vibration*, 442, 552–571, 2019.

[5] T. F. Brooks and W. M. Humphreys. “A deconvolution approach for the mapping of acoustic sources (DAMAS) determined from phased microphone arrays.” *J. Sound Vib.*, 294(4), 856–879, 2006.

[6] L. Brusniak. “DAMAS2 validation for flight test airframe noise measurements.” In *Proceedings of the 2nd Berlin Beamforming Conference*, page 12. 2008.

[7] S. Chakrabarty and E. A. P. Habets. “Multi-speaker doa estimation using deep convolutional networks trained with noise signals.” *IEEE Journal of Selected Topics in Signal Processing*, 13(1), 8–21, 2019.

[8] P. Chiariotti, M. Martarelli, and P. Castellini. “Acoustic beamforming for noise source localization – reviews, methodology and applications.” *Mechanical Systems and Signal Processing*, 120, 422–448, 2019.

[9] R. P. Dougherty. “Spiral-shaped array for broadband imaging.” *US Patent, No. 5838284, 1998*, 1998.

[10] M. Gan, C. Wang, and C. Zhu. “Construction of hierarchical diagnosis network based on deep learning and its application in the fault pattern recognition of rolling element bearings.” *Mechanical Systems and Signal Processing*, 72, 92–104, 2016.

[11] H. Goëau, H. Glotin, W.-P. Vellinga, R. Planqué, and A. Joly. “Lifeclef bird identification task 2016: The arrival of deep learning.” In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.*, pages 440–449. 2016.

[12] B. Graham. “Fractional max-pooling.” *arXiv preprint arXiv:1412.6071*, 2014.

[13] G. E. Hinton and R. R. Salakhutdinov. “Reducing the dimensionality of data with neural networks.” *Science*, 313(5786), 504–507, 2006.

[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. “Densely connected convolutional networks.” In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708. 2017.

[15] X. Huang, L. Bai, I. Vinogradov, and E. Peers. “Adaptive beamforming for array signal processing in aeroacoustic measurements.” *Journal of the Acoustical Society of America*, 131(3), 2152–61, 2012.

[16] Z. Huang, J. Xu, Z. Gong, H. Wang, and Y. Yan. “Source localization using deep neural networks in a shallow water environment.” *Journal of the Acoustical Society of America*, 143(5), 2922–2932, 2018.

[17] W. Humphreys, Jr., T. Brooks, W. Hunter, Jr., and K. Meadows. “Design and use of microphone directional arrays for aeroacoustic measurements.” *36th AIAA Aerospace Sciences Meeting and Exhibit*, AIAA Paper 1998-471 (1998).

[18] Z. Y. M. J. Y. Yangzhou and X. Huang. “A deep neural network approach to acoustic source localization in a shallow water tank experiment.” *Journal of the Acoustical Society of America*, 146(6), 4802–4811, 2019.

[19] F. Jia, Y. Lei, J. Lin, X. Zhou, and N. Lu. “Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data.” *Mechanical Systems and Signal Processing*, 72, 303–315, 2016.

[20] H. Kaiming, Z. Xiangyu, R. Shaoqing, and J. Sun. “Deep residual learning for image recognition.” In *IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[21] Y. Lecun, Y. Bengio, and G. Hinton. “Deep learning.” *Nature*, 521(7553), 436, 2015.

[22] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. “Gradient-based learning applied to document recognition.” *Proceedings of the IEEE*, 86(11), 2278–2324, 1998.

[23] Y. Liu, A. R. Quayle, A. P. Dowling, and P. Sijtsma. “Beamforming correction for dipole measurement using two-dimensional microphone arrays.” *Journal of the Acoustical Society of America*, 124(1), 182–191, 2008.

[24] S. Luesutthiviboon, A. M. Malgoezar, R. Merino-Martinez, M. Snellen, P. Sijtsma, and D. G. Simons. “Enhanced HR-CLEAN-SC for resolving multiple closely spaced sound sources.” *International Journal of Aeroacoustics*, page 1475472X19852938, 2019.

[25] W. Ma and X. Liu. “Phased microphone array for sound source localization with deep learning.” *Aerospace Systems*, pages 1–11, 2019.

[26] R. Merino-Martínez, P. Sijtsma, M. Snellen, T. Ahlefeldt, J. Antoni, C. J. Bahr, D. Blacodon, D. Ernst, A. Finez, S. Funke, T. F. Geyer, S. Haxter, G. Herold, X. Huang, W. M. Humphreys, Q. Leclère, A. Malgoezar, U. Michel, T. Padois, A. Pereira, C. Picard, E. Sarradj, H. Siller, D. G. Simons, and C. Spehr. “A review of acoustic imaging methods using phased microphone arrays.” *CEAS Aeronautical Journal*, 10(1), 197–230, 2019.

[27] F. Ning, F. Pan, C. Zhang, Y. Liu, X. Li, and J. Wei. “A highly efficient compressed sensing algorithm for acoustic imaging in low signal-to-noise ratio environments.” *Mechanical Systems and Signal Processing*, 112, 113–128, 2018.

[28] H. Niu, Z. Gong, E. Ozanich, P. Gerstoft, H. Wang, and Z. Li. “Deep-learning source localization using multi-frequency magnitude-only data.” *Journal of the Acoustical Society of America*, 146(1), 211–211, 2019.

[29] Z. Prime and C. Doolan. “A comparison of popular beamforming arrays.” In *Proceedings of the Australian Acoustical Society AAS2013 Victor Harbor*, volume 1, page 5. 2013.

[30] L. Rayleigh. “Xxxi. investigations in optics, with special reference to the spectroscope.” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 8(49), 261–274, 1879.

[31] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. “Learning internal representations by error propagation.” *Neurocomputing: foundations of research*, pages 673–695, 1988.

[32] H. Shao, H. Jiang, H. Zhao, and F. Wang. "A novel deep autoencoder feature learning method for rotating machinery fault diagnosis." *Mechanical Systems and Signal Processing*, 95(95), 187–204, 2017.

[33] P. Sijtsma. "Clean based on spatial source coherence." *International Journal of Aeroacoustics*, 6, 357–374, 2007.

[34] T. Takaishi, T. Inoue, H.-H. Lee, M. Murayama, Y. Yokokawa, Y. Ito, T. Kumada, and K. Yamamoto. "Noise Reduction Design for Landing Gear toward FQUROH Flight Demonstration." In *Proceedings of the 23rd AIAA/CEAS Aeroacoustics Conference*. AIAA Paper 2017-4033 (2017).

[35] J. M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa. "Towards end-to-end acoustic localization using deep learning: From audio signal to source position coordinates." *Sensors*, 18(10), 3418, 2018.

[36] Y. Wang and H. Peng. "Underwater acoustic source localization using generalized regression neural network." *Journal of the Acoustical Society of America*, 143(4), 2321–2331, 2018.

[37] K. Xu, H. Cai, X. Liu, Z. Gao, and B. Zhang. "North atlantic right whale call detection with very deep convolutional neural networks." *Journal of the Acoustical Society of America*, 141(5), 3944–3945, 2017.

[38] Y. Yang, Y. Liu, Y. Li, E. Arcondoulis, Y. Wang, W. Li, and B. Huang. "Aerodynamic and aeroacoustic characteristics of a multicopter propeller during forward flight." In *2018 Joint Propulsion Conference*, page 4892. 2018.

[39] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao. "Deep learning and its applications to machine health monitoring." *Mechanical Systems and Signal Processing*, 115, 213–237, 2019.