Radboud University Nijmegen

Faculty of Social Sciences

Sound Source Localisation in a Simulated Environment

using Convolutional Neural Networks

Thesis BSc Artificial Intelligence

Author:
Freek van den Bergh

Supervisors:
Dr. Umut Güçlü
Prof. Marcel van Gerven

28th June 2019

Contents

1 Introduction 2

2 Methods 3

2.1 Data Generation . . . 3

2.2 Simulated Environment . . . 4

2.3 Neural Network Architecture . . . 5

2.4 Penalised Cosine Loss . . . 6

3 Experiments 8

3.1 Experiment 1: accuracy of DOA estimation in terms of x- and y-coordinates and in terms of angles . . . 8

3.1.1 Experimental setup . . . 8

3.1.2 Evaluation . . . 8

3.1.3 Results and Discussion . . . 8

3.2 Experiment 2: the effect of distance between sound sources and microphones on the accuracy of SSL . . . 9

3.2.1 Experimental setup . . . 9

3.2.2 Evaluation . . . 9

3.2.3 Results and Discussion . . . 10

3.3 Experiment 3: interaural components learned by the network . . . 11

3.3.1 Experimental setup . . . 11

3.3.2 Results and Discussion . . . 12

4 Future Research 14

5 Conclusion 15

1 Introduction

In the field of sound source localisation (SSL), the goal is to compute the position of one or multiple sound sources with the signals received at multiple microphones. This position is often referred to as the Direction Of Arrival (DOA). SSL is a multifaceted problem with numerous different solutions. Variables that make SSL such a multifaceted problem include the possible locations/distances of sound sources relative to the microphones; the number of microphones in the environment; the number of sound sources in the environment; different kinds of real-world environments, for example empty rooms, noisy public spaces, anechoic chambers, controlled experimental rooms, et cetera. Due to the large amount of variability in the aforementioned variables, the number of possible waveforms, that signals in their environments can take on, is infinite. Fortunately, recent approaches with neural networks have proven to be an effective and robust solution to this issue [1, 10, 12, 13, 21–24, 26].

In state-of-the-art research, numerous different methods have been proposed to solve the problem of SSL with neural networks. Regarding these methods, there are three main categories: time delay based, beamforming based and high-resolution spectral-estimation based approaches [4, 7, 23]. In the time delay based approach, the time difference of arrival (TDOA) between every pair of microphones is computed and is passed as input to the neural network, along with the signals. The TDOA is usually computed with generalised cross-correlation (GCC) and is sometimes combined with the phase transform filtering (PHAT) to improve robustness [3, 4, 11, 20]. In the beamforming approach, the position of the sound source(s) is estimated by optimising a spatial statistic based on the TDOAs of the different pairs of microphones. This spatial statistic, as well as the signals, are then fed into the neural network. An algorithm widely used for optimisation of these spatial statistics is the steered response power algorithm. This algorithm, like GCC, is often combined with PHAT [5–8, 14]. Lastly, an example of a high-resolution spectral-estimation based approach, is the multiple signal classification algorithm. Again, the computed features are given as input to the neural network [18]. Moreover, across these three main categories, the position of a sound source is represented in several different ways. For instance, the position can be represented in terms of the x-, y- and z-coordinates of the signal [10, 23]; by specifying the direction of arrival (DOA) in degrees or radians and, optionally, the distance to the sound source(s) [12, 21]; by estimating a probability distribution over a pre-determined amount of coordinates/angles [21, 22, 26]; and even by estimating the clean source signal [13].

All three aforementioned approaches require pre-processing between the time point of arrival at the microphones and the moment the extracted features become input of the neural network. In the case of human SSL, however, none of the auditory signals that are received at the ears are pre-processed with these algorithms. Instead, humans rely on binaural cues, namely the interaural intensity and phase/time differences (IID, IPD/ITD respectively), to perform SSL [2, 15]. These cues come to exist due to the shape of our ears and the head forming an obstruction between both ears. Specifically, human SSL, in the horizontal plane, relies more on one interaural component depending on the frequency of the signal. In humans, for signals with a frequency below 1500Hz, the ITD is the strongest feature, as the waveform of these signals is larger than the average human head, making the signals distinguishable by time difference. For signals with frequencies above 3000Hz, however, the waveform becomes too small to distinguish them by time difference. Hence, in this case, humans are more dependent on the IID. For frequencies between 1500 and 3000Hz, both cues are equally important for performing SSL [2, 15].

In this thesis, the main focus is to address SSL with neural networks by only providing the networks with the raw audio signals received at the microphones in a simulated environment with two microphones. The approach of feeding the raw propagated signals to the network has been shown equally as effective as pre-processing methods and is still very new in the field of SSL. To our knowledge, the paper of [23] is the only paper that performed SSL without pre-processing their signals. Moreover, when the signals are not pre-processed, the network has to learn to distinguish the relevant features itself. Whereas, in the aforementioned conventional pre-processing based approaches, these features are already available to the network. The ITD and IID are among these important features and, therefore, one of the questions that will be answered in this paper is: does the neural network learn to use the ITD and IID? We hypothesise that the network will learn to use the interaural components, since these components are very simple in nature and extracted easily from the raw signals. Furthermore, the effect of different location representations of the sound sources on the accuracy of SSL is researched. Specifically, a coordinate representation, as compared to an angular representation. Because the angular representation suffers from the circular property, we believe that the coordinate representation will prove to be more accurate. Lastly, the effect of different distances between the sound sources and microphones on the accuracy of SSL is researched. The distance of a sound source should have no effect on the accuracy of SSL, since the interaural components will always stay equal relative to the distance between the sound source and microphones. Therefore, we hypothesise that there will be no significant effect of distance on the accuracy of SSL.

This thesis is organised as follows. The data generation, neural network topology and the environment simulation are explained in Section 2. All three experiments and their discussed results are presented in separate paragraphs in Section 3. The ideas for future research and limitations of our research are discussed in Section 4. Finally, the concluding remarks are provided in Section 5.

2 Methods

This section describes how an infinite amount of data was generated, how the environments were simulated, what the architecture of the neural networks was and a custom loss function used during training and testing of networks with an angular location representation.

2.1 Data Generation

In this paper, the only audio signal was a sine wave. When using real-world data, such as a dataset of voice excerpts, one is limited in the amount of data we can use. In contrast, sine waves can be generated in an infinite amounts of forms by altering their frequency and amplitude. The only restriction imposed by our neural network structure, explained in Section 2.3, on the data generation was that the input signals of a network need to be of the same length. For example, if the network was modeled such that it expected input signals of 100 total samples, then any signal of more than 100 samples was unusable for the network. On the contrary, signals smaller than 100 samples could be made suitable by padding the signals with zeros until they were 100 samples long.

Regarding the actual data generation, each individual wave was given a length of one second and a random frequency and amplitude. The average range of audible frequencies for humans is 20 to 20000Hz, hence, to further emulate human SSL, the chosen frequency range for the sine waves was [20, 20000]. Moreover, the signal was amplified by an integer in the range [1, 500]. These ranges allow for $500 * 19881 = 9990500$ unique sine waves. If the infinite number of angles and locations, where the signal can be propagated from, are taken into account, the number of unique sine waves that can be simulated is infinite.

2.2 Simulated Environment

The environment is a vital component of SSL. A vast number of factors can be changed in the environment that will change the outcome and difficulty of the problem. These factors include the number of microphones; the kind of environment that is simulated (e.g. noisy, anechoic, small, big); the number of sound sources; the distance between a sound source; the type of microphones; et cetera. Therefore, to simplify this issue, the Python library `pyroomacoustics` was used to simulate sound propagation [17]. `pyroomacoustics` simulates real-world sound propagation according to the variables and environment that were chosen.

The environment that was chosen was a reverberant square room of size x cm where sound sources were placed only on the circumference of a circle with diameter x cm and center ($x/2$, $x/2$), where the (0, 0) point of the room was placed at the top-left. In this thesis, this circle will be referred to as the *sound source circle*. The first microphone was placed on coordinates ($x/2 - 11$ cm, $x/2 - 10$ cm) and the second microphone on coordinates ($x/2 + 11$ cm, $x/2 - 10$ cm). An example environment with an exemplary sound source is visualised in Figure 1. The distance between the microphones on the x-axis was set equal to 22cm to simulate the average human head, which is approximately 22cm wide. Furthermore, due to the *cone of confusion* problem which occurs when the center of the microphones is equal to the center of the sound sources [25], the two microphones were placed 10cm above the center of the sound source circle. Lastly, in this environment, the sampling rate was set to $2.2 * 20000 = 44000$Hz, as, according to the Nyquist-Shannon sampling theorem [16,19], the sampling rate should exceed twice the maximal frequency found in the sine wave data to be able to properly distinguish all signals.

With these parameters, one second of a randomly generated sine wave had 44000 samples. However, due to the nature of sound propagation in `pyroomacoustics` and the fact that some signals arrived later at one microphone than the other, some propagated signals were slightly longer than others. Hence, all signals were padded with a number of zeros at their tail such that the length of each generated signal was equal to a number, M, greater than the maximum original signal length, such that

$$M > max(|s| : s \in S),$$

where S is the set of all signals for the current environment. M is relative to the radius of the circle of sound sources. A larger radius induces a longer signal propagation time and, hence, a longer signal length.

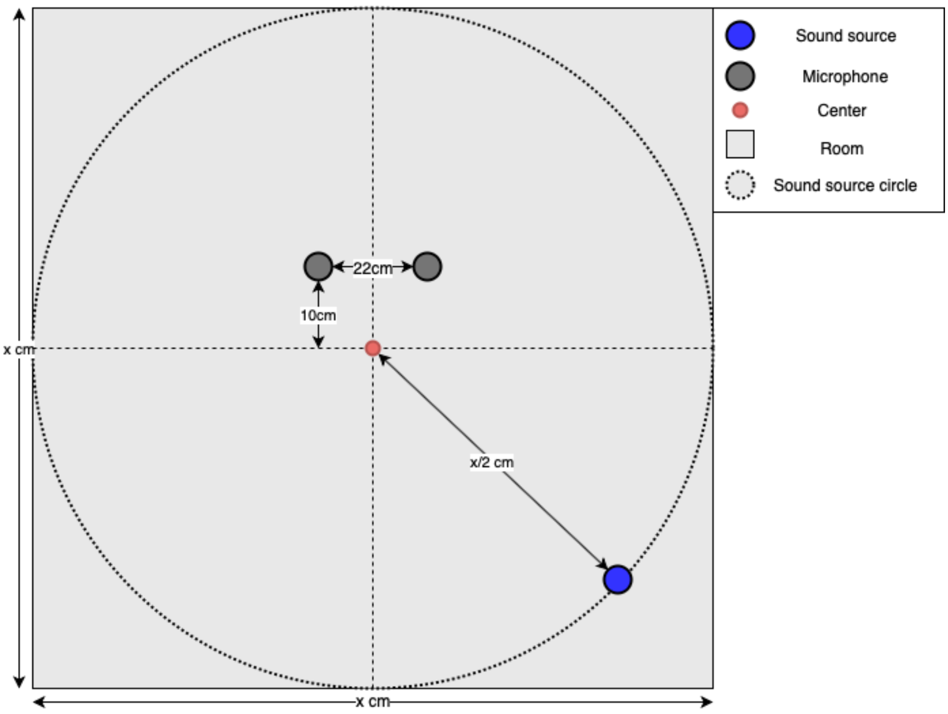

Figure 1: An exemplary environment with sound source. The legend is given in the top-right and the (0, 0) coordinate is in the top-left.

2.3 Neural Network Architecture

The neural network architecture is shown in Figure 2. This architecture was based on [23] and is a convolutional neural network (CNN). One difference between the CNN in [23] and our CNN is that no dropout was used in the fully connected layers. Since the problems posed in this thesis are less complex than in [23], our network did not converge when dropout was applied, likely due to underfitting. Another difference is that uniform Xavier weight initialisation was used in this thesis to overcome vanishing gradients [9]. Lastly, in this thesis, coordinate estimation was only concerned with the x- and y-coordinates, whereas the network in [23] was also concerned with the distance. Hence, the *FC Output* layer of the network in Figure 2 consisted of two blocks, or one block in the case of angle estimation.

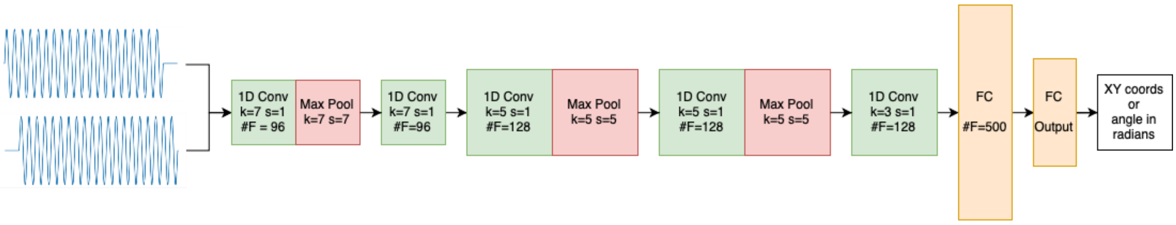

Figure 2: The used neural network architecture. k stands for kernel size, s stands for stride and $\#F$ stands for number of filters/nodes. *1D Conv* is an abbreviation for a one-dimensional convolutional layer and displayed in green, *Max Pool* for a max pooling layer and displayed in red, and *FC* for a fully connected layer and displayed in orange.

2.4 Penalised Cosine Loss

To solve the problem of SSL, where the DOA is estimated in terms of the angle of the sound sources, a *circular* loss function was required. A linear loss function, such as the Mean Squared Error (MSE) loss, did not suffice because it did not adhere to the circular property, which was defined such that

$$\theta \bmod 2\pi = \theta. \tag{1}$$

For example, the angle of 0 is equal to the angle 2π and nearer to the angle of $\frac{3\pi}{2}$ than π.

Thus, due to the circular property in Equation 1, the number of correct predictions for any angle was virtually infinite. To ensure that a neural network trained to estimate an angle converged, the number of possible correct predictions needed to be limited. More specifically, to restrict the problem, a penalty P, with a penalty value p, needed to be introduced for all estimated angles θ that are in a multiple k of the range $[0, 2\pi)$, such that

$$P = |kp|, \text{ for all } k \text{ where } \theta \in [0 + 2\pi k, 2\pi + 2\pi k) \text{ and } k \in \mathbb{Z}. \tag{2}$$

Because there was no loss function implemented in PyTorch that satisfied the properties in Equations 1 and 2, a custom loss function was implemented, called the penalised cosine loss.

The penalised cosine loss function consisted of two parts where each part satisfied one property respectively. The circular property in Equation 1 was satisfied by the first part of the loss function, which is visualised in Figure 3,

$$|\cos(y - \hat{y}) - \cos(0)| \tag{3}$$

where y is the target angle and $\hat{y}$ is the predicted angle, both in radians. The subtraction of $\cos(0)$ was necessary to center the zero point of the loss function at equal target and predicted values and to reduce the number of extrema to one, instead of two. Moreover, because the absolute value was taken, the cosine showed two maxima in the range $[0,\ 2\pi]$, instead of one maximum and one minimum.

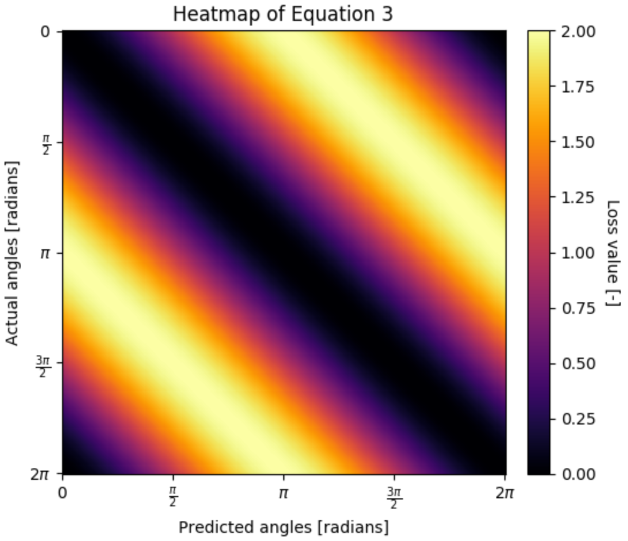

Figure 3: Heatmap of Equation 3 for angles in the range of $[0,\ 2\pi]$. The bar on the right indicates the values corresponding to the colours.

The property in Equation 2 was satisfied by taking the predictions and putting them in the following equation, which is visualised in Figure 4,

$$\max(0, \hat{y} - 2\pi) + \max(0, -\hat{y}) \tag{4}$$

where $\hat{y}$ is the predicted angle in radians. With this equation, the penalty value p is equal to θ such that

$$P = |k\theta|, \text{ for all } k \text{ where } \theta \in [0 + 2\pi k, 2\pi + 2\pi k] \text{ and } k \in \mathbb{Z}.$$

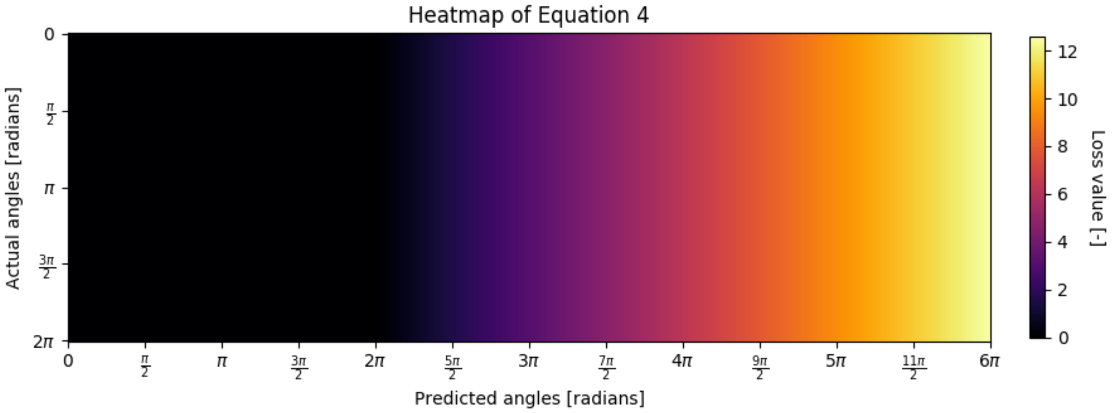

Figure 4: Heatmap of Equation 4 for angles (in radians) in the range of [0, 2π] on the y-axis and angles (in radians) in the range [0, 6π]. The bar on the right indicates the values corresponding to the colours.

Thus, the combination of Equations 3 and 4 produced the following loss function that satisfied all necessary properties, which is visualised in Figure 5,

$$L = |\cos(y - \hat{y}) - \cos(0)| + \max(0, \hat{y} - 2\pi) + \max(0, -\hat{y}) \tag{5}$$

where L is the final loss, y is the target angle and $\hat{y}$ is the predicted angle.

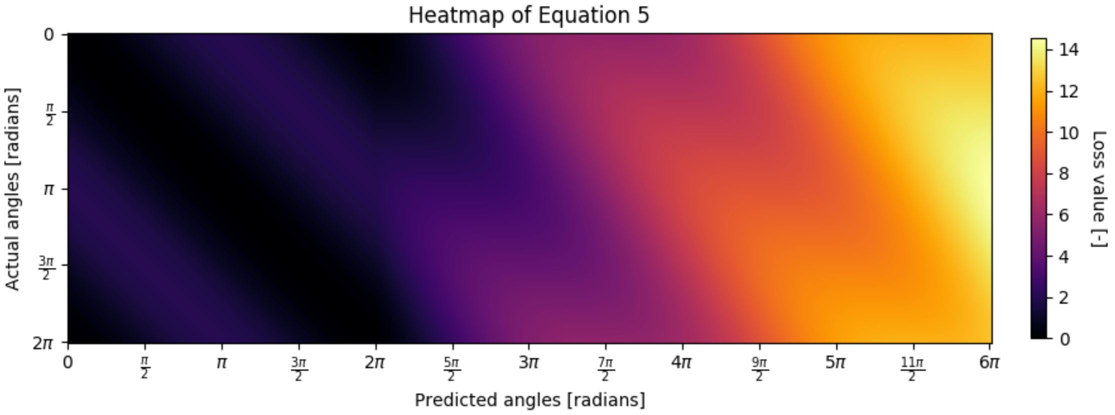

Figure 5: Heatmap of Equation 5 for angles (in radians) in the range of [0, 2π] on the y-axis and angles (in radians) in the range [0, 6π]. The bar on the right indicates the values corresponding to the colours.

3 Experiments

In this paper, three experiments were conducted: firstly, the accuracy of SSL via coordinate and angle estimation were compared; secondly, accuracy of SSL was measured and compared between different distances between the microphones and sound source; lastly, the networks in experiment two were analysed on whether they used the interaural components to perform SSL.

3.1 Experiment 1: accuracy of DOA estimation in terms of x- and y-coordinates and in terms of angles

3.1.1 Experimental setup

In this experiment, two networks were trained and tested and, finally, their results were compared to each other and a baseline model. One network was trained to estimate the x- and y-coordinates of the sound sources placed on the sound source circle. The other network was trained to estimate the angle of the sound sources placed on the sound source circle. Both networks adhere to the topology described in Section 2.3. The baseline model was a simple fully connected neural network with 6 layers, which was also trained on the x- and y-coordinate representation.

Regarding the training parameters, all networks were trained on a batch size of 64 signals, M, the length of the input signals, of 48000 for 7000 epochs. The MSE loss function was used for the estimation in terms of x- and y-coordinates, such that the total loss was the sum of the MSE loss on the x-coordinates and the MSE loss on the y-coordinates. The penalised cosine loss function, described in Section 2.4, was used for the estimation in terms of the angle in radians. The width and height of the room, and therefore also the diameter of the sound source circle, were set to 100cm. The signals of the sound sources were generated as described in Section 2.1 and placed on random angles generated in the range $[0, 2\pi)$ along the sound source circle. Lastly, both networks were tested on 10000 total sine waves.

3.1.2 Evaluation

The networks were trained with different loss functions and the loss values could, therefore, not be compared directly. Nevertheless, due to the constraint that sound sources were only placed on the circumference of the sound source circle, the angles could be converted to coordinates and vice versa. Thus, the penalised cosine loss on the predictions in terms of angles was compared to the cosine loss on the predictions in terms of x- and y-coordinates *converted* to angles and vice versa, such that both networks could be objectively compared to each other.

To determine whether the results were significantly different a dependent t-test was used. The dependent t-test was used because the data for the network trained with x- and y-coordinate representation and for the network trained with the angular representation were sampled from the same dataset.

3.1.3 Results and Discussion

In Table 1 the results for this experiment are shown. From the table, it follows that the location representation in terms of x- and y-coordinates performed significantly better than the location representation in terms of angles. Moreover, both representations clearly performed significantly better than the baseline. For the MSE loss, $t(9998) = -20.08$ at $p < 0.001$ and for the penalised cosine loss, $t(9998) = -20.26$ at $p < 0.001$. The standard deviations of the losses of the coordinate representation imply that the loss values could be negative. However, it is important to note that both loss function are non-negative and that, therefore, their loss values can only reach a minimum of zero.

The x- and y-coordinate representation was likely more accurate due to the fact that the x- and y-coordinate representation is easier to predict than the angular representation. This is because circular data, such as angles, suffer from the circular property from Equation 1 which results in an infinite number of possible correct predictions. Whereas coordinate systems do not suffer from the circular property.

Table 1: Cross-evaluated results for the coordinate and angular representation. * indicates significance at $p < 0.001$. The standard deviation has been given after the ±-sign.

	Coordinate representation	Angular representation	Baseline
MSE Loss	4.835 ± 6.702	$86.253 \pm 83.226^*$	3884.853 ± 1985.833
Penalised Cosine Loss	0.019 ± 0.026	$0.433 \pm 0.416^*$	4.682 ± 1.543

3.2 Experiment 2: the effect of distance between sound sources and microphones on the accuracy of SSL

3.2.1 Experimental setup

To measure the effect of distance between sound sources and microphones on the accuracy of SSL, three neural networks, following the architecture in Section 2.3, were trained in three different simulated environment where the room width/height and diameter of the sound source circle was 100cm, 1000cm or 10000cm, respectively. Subsequently, each networks was tested on signals from every simulated environment.

To train the networks, the batch size was set equal to 32 and the number of epochs to 12000. The batch size was kept low due to computational constraints. However, the epochs were increased to make up for the lower batch size. M was set to 65000, since a value of 48000 could not capture the whole signal due to propagation delays in the room of 100m long/wide. The signals of the sound sources were generated identically to Experiment 1. Finally, both networks were tested on 10000 total sine waves.

3.2.2 Evaluation

Similar to Experiment 1, the location of the sound source was estimated in terms of the x- and y-coordinates and, hence, the loss was set to the sum of the MSE on the x- and y-coordinates. Furthermore, the relative MSE (rMSE) loss was employed to compare across tested distances. The absolute MSE loss showed the differences within tested distances better and the relative MSE loss enabled comparison between tested distances. The relative MSE loss was obtained by taking the square root of the MSE loss and dividing the outcome by the radius the network was *tested* on as follows:

$$rMSE = \frac{\sqrt{MSE}}{r_{test}}.$$

Furthermore, an independent Welch's t-test was employed to determine whether there were significant differences between the losses on the diagonal in Table 5. Welch's t-test was chosen due to the violation of the homoscedasticity assumption of the standard t-test. The assumption was tested with a Levene's test on every pairwise combination of the diagonals of Table 5, the results of which can be found in Table 2.

Table 2: The F-statistics of the Levene's tests for every pairwise combination of the diagonals of Table 5. * indicates significance at $p < 0.001$.

Radius pairs	F-statistic
$r = 50$cm versus $r = 500$cm	5.63*
$r = 50$cm versus $r = 5000$cm	51.03*
$r = 500$cm versus $r = 5000$cm	21.37*

3.2.3 Results and Discussion

The absolute MSE loss values are presented in Table 4 and the rMSE loss values in Table 5. In both tables, the radius that the networks were trained on is indicated in the columns and the radius that the network was tested on is indicated in the rows. From the data presented in Table 4 it is apparent that all values not on the main diagonal are disproportionate to the value on the diagonal in their respective row, indicating that networks trained on one radius did not generalise to another radius. Furthermore, the relative MSE losses in Table 5 indicate that networks trained on a smaller radius performed significantly more accurately, when given test data from their own radius, than networks trained on a larger radius. The differences between all pairwise combinations of the diagonals were significant (see Table 3).

Table 3: Degrees of freedom and t-statistics for all pairwise combinations of the diagonals of Table 5. * indicates significance at $p < 0.001$.

Radius pairs	Degrees of freedom	t-statistic
$r = 50$cm vs $r = 500$cm	9998	-16.10^*
$r = 50$cm vs $r = 5000$cm	9998	51.04*
$r = 500$cm vs $r = 5000$cm	9998	-5.31^*

Contrary to our hypothesis, the network trained on a smaller radius performed relatively more accurate on its own data, than networks trained on a larger radius. This might be due to the fact that the room was reverberant and that this reverberation interfered more with the original signal for larger radii, than for smaller radii. Because it took longer for the signal to arrive at the microphones for larger radii, the signal had more time to reverberate and interfere with the original signal. The time window of reverberation for smaller radii was narrower, resulting in less interference with the original signal.

Table 4: Cross-evaluated absolute MSE losses (± standard deviation) for the networks trained and tested on different radii (r_{train} and r_{test} respectively). * indicates significance at $p < 0.001$.

	$r_{train} = 50$cm	$r_{train} = 500$cm	$r_{train} = 5000$cm
$r_{test} = 50$cm	$4.834 \pm 12.974^*$	$6.257 * 10^5 \pm 6.507 * 10^4$	$6.841 * 10^7 \pm 1.424 * 10^7$
$r_{test} = 500$cm	$6.643 * 10^5 \pm 2.093 * 10^5$	$1.068 * 10^3 \pm 3.804 * 10^{3*}$	$4.321 * 10^7 \pm 4.850 * 10^6$
$r_{test} = 5000$cm	$7.466 * 10^7 \pm 2.183 * 10^7$	$6.821 * 10^7 \pm 2.075 * 10^7$	$1.573 * 10^5 \pm 5.572 * 10^{5*}$

Table 5: Cross-evaluated rMSE losses (± standard deviation) for the networks trained and tested on different radii (r_{train} and r_{test} respectively). * indicates significance at $p < 0.001$.

	$r_{train} = 50$cm	$r_{train} = 500$cm	$r_{train} = 5000$cm
$r_{test} = 50$cm	$0.044 \pm 0.072^*$	15.821 ± 5.102	165.423 ± 75.484
$r_{test} = 500$cm	1.630 ± 0.915	$0.065 \pm 0.123^*$	13.146 ± 4.405
$r_{test} = 5000$cm	1.728 ± 0.934	1.652 ± 0.911	$0.079 \pm 0.149^*$

3.3 Experiment 3: interaural components learned by the network

3.3.1 Experimental setup

For the final experiment, the activations of the linear layer of the x- and y-coordinate network in Experiment 1 were recorded for 4500 sine waves, 1500 for each frequency category. These three categories included a low frequency (LowFreq) group, where signals were generated with a frequency (in Hz) in the range $[20, 1500]$; a medium frequency (MedFreq) group, where signals were generated with a frequency (in Hz) in the range $[1500, 3000]$; a high frequency (HigFreq) group, where signals were generated with a frequency (in Hz) in the range $[3000, 20000]$. Subsequently, the means of these activations were correlated with the interaural components of the received signals. These components included the ITD, the TDOA between one microphone and the other, and the IID, the difference in amplitude between one microphone and the other. In the case of the ITD, the time of arrival at the right microphone, TOA_R, was subtracted from the time of arrival at the left microphone, TOA_L, as follows

$$ITD = TOA_R - TOA_L.$$

For the IID, the amplitudes of all peaks of the received signal at the left, $peaks_L$ and right microphones, $peaks_R$ were computed. Subsequently, the (mean) interaural intensity difference was obtained by

$$IID = \frac{1}{N} \sum_{i=0}^{N} peaks_L^i - peaks_R^i,$$

where $N = |peaks_L|$. If $|peaks_L| \neq |peaks_R|$, then the first $||peaks_L| - |peaks_R||$ elements were removed from whichever vector was longer before the IID was computed.

To correlate the activations with the interaural differences, Pearson's r correlation coefficient was used. A two-tailed p-value was computed to test whether the correlations were significant.

3.3.2 Results and Discussion

The plot of the ITDs correlated with the activations of the neural network are shown in Figure 6 and the plot of the IIDs correlated with the activations of the neural network are presented in Figure 7. For the ITD-activations significant correlations were found for the low, medium and high frequency categories (r=0.75, 0.74 and 0.76, respectively, all significant at $p < 0.001$). Furthermore, for the IID-activations significant correlations were observed for the low, medium and high frequency categories (r=-0.34, -0.23 and -0.26, respectively, all significant at $p < 0.001$).

These results indicate that the neural network did learn to use the interaural differences, which are also employed by humans when performing SSL. However, the results were not completely consistent with human SSL. For frequencies below 1500 Hz, i.e. the low frequency category, human SSL is more dependent on the ITD than the IID. For frequencies above 3000, i.e. the high frequency category, this effect is reversed; human SSL is, in the high frequency case, more dependent on the IID than the ITD. When frequencies between 1500 Hz and 3000 Hz are presented, human SSL is equally dependent on both interaural differences [2,15]. This effect was not observed in the correlation coefficients. In fact, the ITD stayed equally relevant throughout all frequency categories and the IID showed the opposite effect, where the correlation coefficient for the low frequency group was higher than for the high frequency group. We presume that this might be caused by the fact that, in human SSL, the head forms an obstruction between the two ears, whereas, in the simulated environment, there was no obstruction between the two microphones.

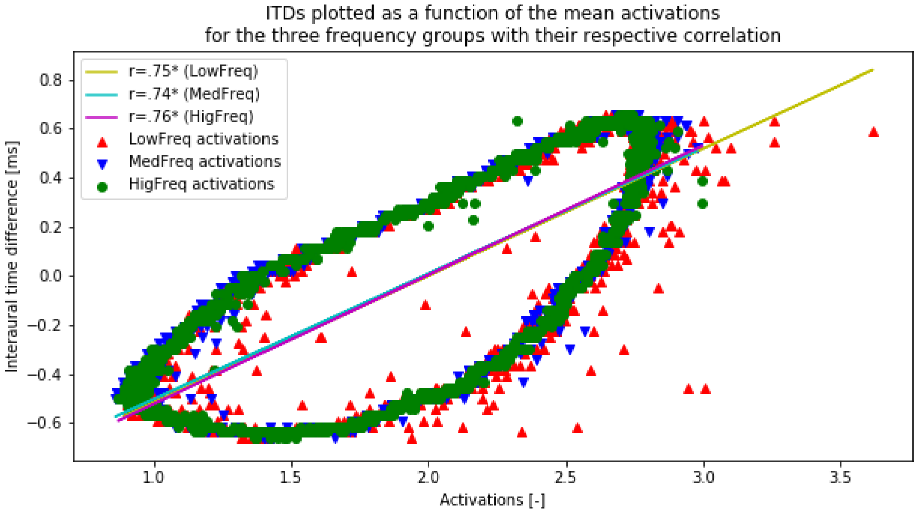

Figure 6: The ITDs (y-axis) correlated with the activations of the neural network (x-axis) for the three frequency categories. * indicates significance at $p < 0.001$.

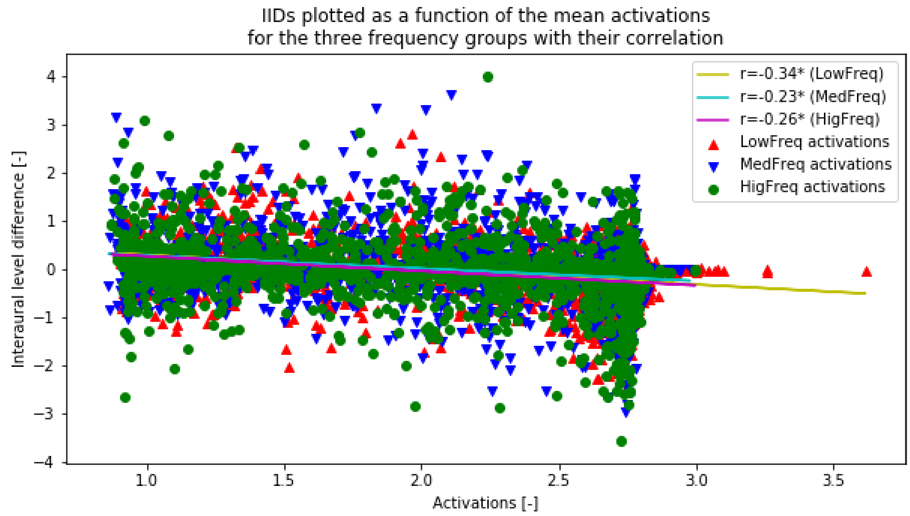

Figure 7: The IIDs (y-axis) correlated with the activations of the neural network (x-axis) for the three frequency categories. * indicates significance at $p < 0.001$.

Nevertheless, Figure 6 showed an interesting pattern; the representation of all datapoints in Figure 6 shows a circular pattern, which implies that there is no simple linear correlation between the activations (of all categories) and the ITDs. When a line was drawn across the zero point on the y-axis and another line was drawn from the smallest ITD to the largest ITD, as is shown in Figure 8, the ITDs plotted against the activations showed interesting correspondence to the simulated environment, shown in Figure 9. The area marked in cyan represents negative ITDs, indicating that the signal arrived at the left microphone earlier than at the right microphone. Similarly, the area marked in yellow indicates that the signal arrived at the right microphone earlier than at the left microphone.

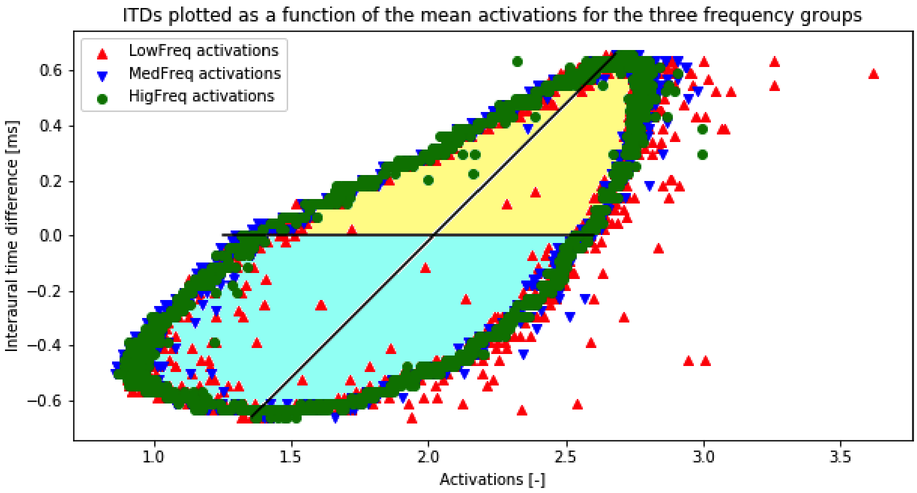

Figure 8: The activations plotted as a function of the ITDs corresponded with the simulated environment. All negative ITDs are marked in cyan and all positive ITDs in yellow. The markings show correspondence to Figure 9.

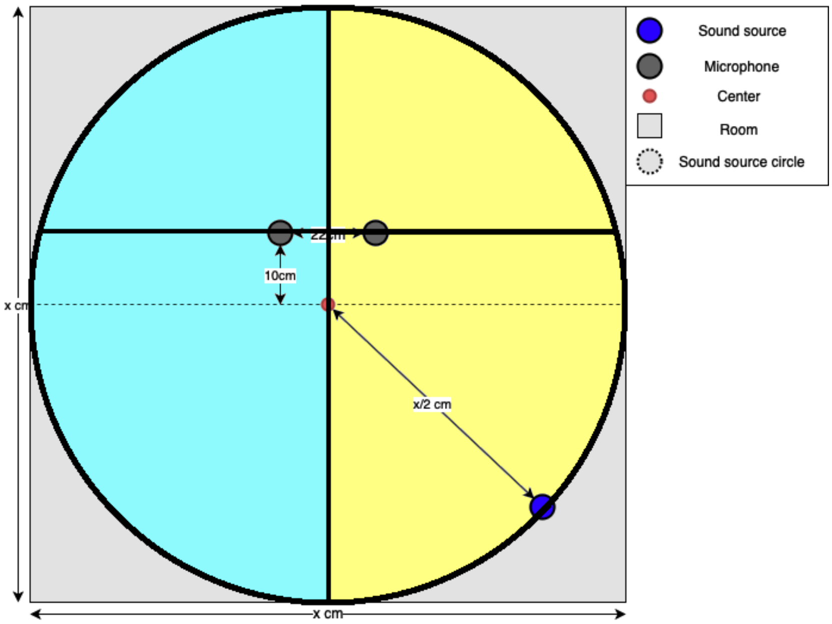

Figure 9: The simulation environment where the side of the sound source circle that produces negative ITDs is marked in cyan and the side that produces positive ITDs is marked in yellow. The markings show correspondence to the correlation in Figure 8.

Hence, we observe from Figures 8 and 9 that, when the simulated environment was rotated by $-\frac{\pi}{2}$ radians, the simulated environment could be mapped to the pattern of the activations plotted against the ITDs. However, there was not a one-on-one relationship between the simulated environment and the activations. The sound source circle in the simulated environment was a perfect circle, whereas the correlation is more oval-like. Firstly, this discrepancy might be due to the fact that the neural network does did solely rely on ITDs to perform SSL. The IID also correlated with the activations of the neural network. Hence, the correlation might not show a perfect circle shape because the activations also represented other features than the ITDs. Secondly, the microphones in the simulated environment were placed off-center which resulted in asymmetric ITDs above and below the horizontal midline of the sound source circle. The asymmetric ITDs might be another explanation for the oval-like, i.e. non-symmetric, shape of the correlation.

4 Future Research

Future research could aim to study the effect of distance or different location representations on accuracy, or the use of interaural components in more ecologically valid situations, given the lack of ecological validity of the approach used in this thesis. For example, by using noisier and differently shaped environments or real-world data such as voices. Furthermore, networks in Experiment 2 did not generalise outside of their own trained distance. Hence, it could prove interesting to train the networks on signals from different distances to obtain a network that generalises across different distances. Lastly, the effect of different variables on the correlations in Experiment 3 could be researched to gain more insight into the representations of the neural network. These variables might include the addition of some form of obstruction in between the two microphones to simulate a human head or different microphone positions within or even outside the sound source circle.

5 Conclusion

In this paper, we have investigated SSL by giving the raw signals to a CNN. The results of the first experiment showed that an x- and y-coordinate representation of the location of a sound source achieved higher SSL accuracy than an angular representation of the location of a sound source. The x- and y-coordinate representation outperformed the angular representation because coordinates are not influenced by the circular property, whereas angles are. In the second experiment, the results showed that networks trained on one specific distance between the microphones and the sound source did not generalise to other distances than the one they were trained on. Furthermore, it was shown that SSL was more accurate for smaller distances than for larger distances. Finally, the results of the third experiment showed that the approach of the neural network to SSL was partially consistent with human SSL. It was consistent with human SSL in the sense that the activations correlated with the ITD and IID, which are both fundamental to human SSL. However, it was inconsistent in the sense that the correlation did not differ between signals of different frequencies for both the ITD and IID, whereas this is the case for humans. Furthermore, the circular pattern of the datapoints showed interesting similarity to the simulated circular environment.

References

[1] Bialer, O., Garnett, N., & Tirer, T. (2019). Performance Advantages of Deep Neural Networks for Angle of Arrival Estimation. *arXiv preprint arXiv:1902.03569.*

[2] Blauert, J. (1997). Spatial hearing: the psychophysics of human sound localization. MIT press.

[3] Brandstein, M. S. (1999). Time-delay estimation of reverberated speech exploiting harmonic structure. *The Journal of the Acoustical Society of America, 105*, 2914-2919.

[4] Brandstein, M. S., & Silverman, H. F. (1997). A practical methodology for speech source localization with microphone arrays. *Computer Speech & Language, 11*(2), 91-126.

[5] Cobos, M., García-Pineda, M., & Arevalillo-Herráez, M. (2017). Steered response power localization of acoustic passband signals. *IEEE Signal Processing Letters, 24*(5), 717-721.

[6] DiBiase, J. H. (2000). *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays.* Providence, RI: Brown University.

[7] DiBiase, J. H., Silverman, H. F., & Brandstein, M. S. (2001). Robust localization in reverberant rooms. In *Microphone Arrays* (pp. 157-180). Springer, Berlin, Heidelberg.

[8] Dmochowski, J. P., & Benesty, J. (2010). Steered beamforming approaches for acoustic source localization. In *Speech processing in modern communication* (pp. 307-337). Springer, Berlin, Heidelberg.

[9] Glorot, X., & Bengio, Y. (2010, March). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249-256).

[10] He, W., Motlicek, P., & Odobez, J. M. (2018, May). Deep neural networks for multiple speaker detection and localization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 74-79). IEEE.

[11] Knapp, C., & Carter, G. (1976). The generalized correlation method for estimation of time delay. *IEEE transactions on acoustics, speech, and signal processing, 24*(4), 320-327.

[12] Ma, N., Brown, G. J., & May, T. (2015). Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions. In *Sixteenth Annual Conference of the International Speech Communication Association.*

[13] Ma, W., & Liu, X. (2018). Phased microphone array for sound source localization with deep learning. *arXiv preprint arXiv:1802.04479.*

[14] Marti, A., Cobos, M., Lopez, J. J., & Escolano, J. (2013). A steered response power iterative method for high-accuracy acoustic source localization. *The Journal of the Acoustical society of America, 134*(4), 2627-2630.

[15] Middlebrooks, J. C., & Green, D. M. (1991). Sound localization by human listeners. Annual review of psychology, 42(1), 135-159.

[16] Nyquist, H. (1928). Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers, 47*(2), 617-644.

[17] Scheibler, R., Bezzam, E., & Dokmanić, I. (2018, April). Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 351-355). IEEE.

[18] Schmidt, R. (1986). Multiple emitter location and signal parameter estimation. *IEEE transactions on antennas and propagation, 34*(3), 276-280.

[19] Shannon, C. E. (1998). Communication in the presence of noise. *Proceedings of the IEEE, 86*(2), 447-457.

[20] Stoica, P., & Li, J. (2006). Lecture notes-source localization from range-difference measurements. *IEEE Signal Processing Magazine, 23*(6), 63-66.

[21] Takeda, R., & Komatani, K. (2016, March). Sound source localization based on deep neural networks with directional activate function exploiting phase information. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 405-409). IEEE.

[22] Takeda, R., & Komatani, K. (2016, December). Discriminative multiple sound source localization based on deep neural networks using independent location model. In *2016 IEEE Spoken Language Technology Workshop (SLT)* (pp. 603-609). IEEE.

[23] Vera-Diaz, J., Pizarro, D., & Macias-Guarasa, J. (2018). Towards End-to-End Acoustic Localization Using Deep Learning: From Audio Signals to Source Position Coordinates. *Sensors, 18*(10), 3418.

[24] Wang, Z. Q., Zhang, X., & Wang, D. (2019). Robust Speaker Localization Guided by Deep Learning-Based Time-Frequency Masking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27*(1), 178-188.

[25] Woodworth, R. S., & Schlosberg, H. (1938). Experimental psychology. New York: Henry Holt and Company.

[26] Yalta, N., Nakadai, K., & Ogata, T. (2017). Sound source localization using deep learning models. *Journal of Robotics and Mechatronics, 29*(1), 37-48.