

Monostatic Acoustic Localization Using Convolutional Neural Networks

Stef Brits¹, Robin Kerstens^{1,2} and Jan Steckel^{1,2}

¹CoSys-Lab, Faculty of Applied Engineering, University of Antwerp, Antwerp, Belgium

²Flanders Make Strategic Research Centre, Lommel, Belgium

Abstract

Many applications require knowledge about the position of objects in a room. Popular ways to tackle this issue is to use either vision based sensors, or several communicating beacons placed at known positions which allow beamforming or triangulation methods. However, in some cases, vision is limited due to a lack of light or the presence of airborne obscurants and also the placement of several beacons can be seen as impractical. This paper suggests a method using a monostatic setup where a sensor uses a limited set of known Room Impulse Responses to then accurately estimate its position in that environment using a Regression Convolutional Neural Network. The research is performed using a Finite-difference Time-domain simulation method to generate realistic data and achieves results with an average estimation error of 14,7 cm.

1. Introduction

With an ever increasing demand for automation applications and technologies, localization is an issue that often needs to be handled. For outdoor situations this problem can easily be solved using (D)GPS which is known to achieve accurate results [1]. However, there are many cases in which a GPS system can not be used for accurate measurements since there is a lacking line of sight (LOS) between the satellites and the object. The obstacles between the object and the satellites cause disturbances in the communication, which prohibit reliable use. These drawbacks are further described by Gonzalo Seco-Granados et al. [2]. For cases like this, that are either indoors or located in heavily obscured places (e.g. mining shafts, greenhouses, ...) other solutions are required that rely on more robust techniques utilizing little infrastructure.


Employing the information found in sound waves as a base of a localization technique possesses some advantages over its alternatives. The most important properties being low cost and highly accurate indoor localization. These advantages are further established by Ureña et al. [3]. Applying (ultra)sound as a localization medium can make it possible to attain an accuracy close to one centimeter, as also stated by Ureña. There have been a great number of studies that research acoustic localization and auralization. Dokmanic et al. [4] research how acoustics can be used to estimate the shapes of rooms, which can be practical when researching indoor localization.

IPIN 2021 WiP Proceedings, November 29 – December 2, 2021, Lloret de Mar, Spain

✉ jan.steckel@uantwerpen.be (J. Steckel)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

It is important to note that in many cases which implement sound for localization purposes, multiple microphones are employed. Such an infrastructure is called a microphone array. These arrays collect data by measuring incoming sound waves in a synchronous manner which can then be compared to each other. Using the concepts of the speed of sound and the way sound waves propagate, the differences between the times of arrival in the microphones make it possible to calculate the angle between the microphone array and the sound source [1]. Another localization technique utilizing microphones uses synchronized static beacons. These beacons can send out sound waves which are received by the object to localize. This object then triangulates its position relative to the static beacons. This localization can be accomplished in a multitude of ways, one of which is the time-of-flight (ToF) method. When ToF is used, the time that the sound wave took to travel between the beacon and the object is used to find the location of the object. The concept of static beacons and the conventional localization methods is explained more in depth by [3].

Research in the field of sonar technology is of great interest when other forms of localization are not applicable to a situation. For example, when visual localization is employed and there is not enough light for normal cameras or there is a substantial amount of dust in the environment, as explored by Shehryar Khattak et al. [5]. The dissimilarity in the approaches executed by this paper is that most sonar systems use static beacons or microphone arrays. In our research, no such infrastructure is provided and a monostatic setup is used without the use of supplementary beacons. In this paper, we will propose a machine learning approach using data obtained from finite-difference time-domain (FDTD) [6] simulations. We will debate the design and research choices for the data, simulations, and lastly, the convolutional neural network (CNN) constructed for localizing with a single transceiver. To the author's knowledge, at the time of writing, the exact approach taken in paper has not yet been published in literature. The proposed method finds inspiration in popular Wi-Fi fingerprinting methods [7] where pre-calculated radio maps are used to determine the location of a user. In [8] a similar approach is used, but with a passive measuring scenario where environmental ultrasound is being analyzed. Vera-Diaz [9] does another passive approach, tracking human speech using CNNs. The question this paper proposes to answer is: "Is it possible to localize an object inside a known room using only one sound transceiver? If it is possible, how accurate can the measurement be without the help of this additional infrastructure and which additional intelligent algorithms will be needed?"

In section 2, we will discuss the importance of a room impulse response for localization purposes in this paper. Thereafter, Section 3 contains information on the data generation employed for this research. We explain the methods of data generation and further discuss room modeling techniques. Section 4 explains the design choices of three different neural networks used to localize an object based on the simulated data. Additionally, section 5 shows the results of the networks, localizing an object in a simulated room, with a greater focus on the convolutional neural network for regression. Lastly, we conclude this paper in 6, discussing results, a mean localization error of 0,14 m, and implications of the executed research.

2. Defining the Room Impulse Response

For an object to localize itself inside a room without external help using sound, prior knowledge about the room can be used to help the process. To obtain this knowledge, this research aims to employ a set of Room Impulse Responses (RIR). The RIR can be described as the transfer function of a room between a transmitting sound source and a receiving microphone. The object can send out a broadband signal (e.g. a sine sweep, or an Additive White Gaussian Noise (AWGN) sequence) and record, for a specified amount of time, all reflections that originate from the available surfaces in the room. This can be done by placing the transmitter and receiver on different sides of a room (bistatic) but also when they are located at the same exact position. When using this monostatic approach, the measurement forms a location-specific RIR of which the content changes as the transceiver moves through the room. As every independent position has a unique set of distances towards the reflecting surfaces, the RIR can also be expected to be unique. In our research, we will create such RIRs in using a FDTD simulation in MATLAB [6], using the same exact location for both the transmitter and the receiver. For this work, an omnidirectional transceiver is assumed.

In the current research, we always use the same room in a single dataset. With this knowledge we can state that spatial properties of the room and the relative position between sender and receiver stay the same. A lot of research has already gone into accurately detecting acoustic reflections [10, 11], and it can be concluded that as the bandwidth of the measuring sequence increases, localization will become more robust to noise and will allow for a more accurate result if the frequency increases. However, because of computational needs of the FDTD simulations, this first attempt uses a sequence that limits the time required for running the simulations. A pseudo-random AWGN sequence that lasts 6 ms and that has a bandwidth between 2 kHz and 4 kHz, sampled at 10 kHz.

The main challenge proposed in this paper consists of finding the connection between the location-specific RIR and that same exact location in that room using a limited set of prior info. Antonello et al. [12] describe the importance of measuring and using the RIR where an infrastructure with multiple microphones is used. This was a recurring problem when studying research because conventional approaches do not include single transceiver localization or monostatic localization, as stated by El Badawy et al [13].

One of the RIRs used for this research is depicted in Fig. 1. A typical RIR consists of the measured pressure over time at a certain place in a room. In some cases, the pressure is expressed as energy in the air. Mathematically, the RIR represents the transfer function of the room between the sound source and the microphone. And because a monostatic setup is used, the RIR shows information about the location in the room. According to Cecchi et al. [14], it is possible to split the information provided by a RIR in three sections:

1. Direct sound: this is the sound measured by the first LOS transmission. This can be helpful for evaluating the transmitted sound, as it is the direct, unreflected signal. In this research, the source and receiver would be placed on the same object, a couple of centimeters apart, this means that the direct sound is measured almost instantaneously. In simulation it is possible to measure and transmit at the exact same position. Thus, the

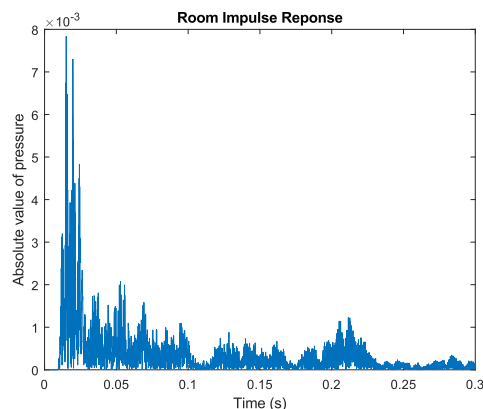


Figure 1: Graph of a typical room impulse response, discrete time on the x-axis, absolute pressure on the y-axis. Measuring for a period of 0,3 seconds allows for the capture of large amount of the reverberation that occurs after the direct reflections have been recorded.

direct sound is not relevant for our simulated results, as the transmitter and receiver are positioned exactly on the same pixel.

2. Early reflections: the early reflections are the first reflected waves, created by first order reflections. In a rectangular room, this part is likely to consist of six reflections originating from the walls, floor and ceiling. Furthermore, these reflections are influenced by the directivity of the used microphone and sound source. Also the distances of the walls in relation to the object impact these early reflections. A larger distance will result in the sound waves taking longer to reach a wall and reflect back.
3. Late reflections: the late reflections consist of all measured information after direct sound and early reflections. This part of the RIR contains a large amount of noise relative to the signal, since the measured signals have dissipated over time. Nevertheless, the late reflections contain a lot of information on the rooms' spatial properties. Through multi-path propagation, late reflections could show the dimensions of the room and can aid in the aural localization in the room.

3. Generating Acoustic Data

Firstly, we employed the intuitive approach to measure in real life. A very basic procedure by measuring in a dorm room with a laptop (Lenovo Y520-15IKBN) was used. That laptop simultaneously sent and received a sub 20kHz sine-sweep signal sampled at 44100 Hz, based on methods used by Stan et al. [15]. This method proved to be inconsistent. When conducting multiple equal tests at different days, different measurements were found. This was thought to be caused by noise drowning out the information gathered from measurements. Also, obtaining the large amount of training data that is required to train the network in this manner can be seen as cumbersome. For this reason the research would first be validated using simulated measurements, that can be run in parallel to limit the required time.

3.1. Simulation Methods

Simulation is a strong, well known tool for creating substitute data. For example Vargas et al. [16] showed that using simulated data could be used to train machine learning algorithms for sound recognition. Vargas et al. also noted that transfer learning could be used to expand neural networks trained on simulated data to be tested on real, measured data. Such methods of simulating are explored by Markovic et al. [17] and Deines et al. [18] more thoroughly. Multiple ways of simulating the acoustic properties of a room exist. These can be split in two categories.

1. Solving wave equations: This approach considers numerically solving the wave equations to find the physical properties of a room. This method is more accurate than geometrical acoustics. The drawback of this method consists of the large computational cost of solving the wave equations.
2. Geometrical acoustics (GA): The geometrical acoustics approach simplifies the acoustics modeling problem by assuming sound waves to be rays. This simplification creates the advantage of a favorably lower computational cost at the price of accuracy. Savioja et al. [19] further expand upon these concepts in practice. Maa et al. [20] also provide useful insights, favouring geometrical acoustics for practitioners, while favouring wave equations for theoretical studies.

Both solving wave equations as GA contain useful modeling techniques for localization and auralization purposes. The choice was made to use a wave equation solving technique due to its accurate nature.

3.2. Finite-Difference Time-Domain Simulation

We used the finite-difference time-domain method for modeling the room and more precisely its reflective properties. This choice was based on multiple successful researches comparing modeling methods and preferring FDTD over multiple GA methods like ray-tracing. De Sena et

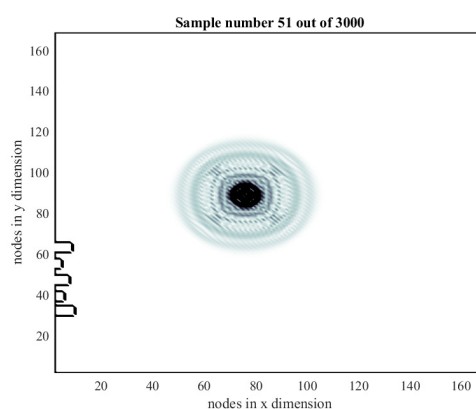


Figure 2: A 2-dimensional projection of an example 3-D FDTD simulation. The wavefront originates from the location of the acoustic transceiver, where also the RIR will be recorded after the signal has passed through the environment.

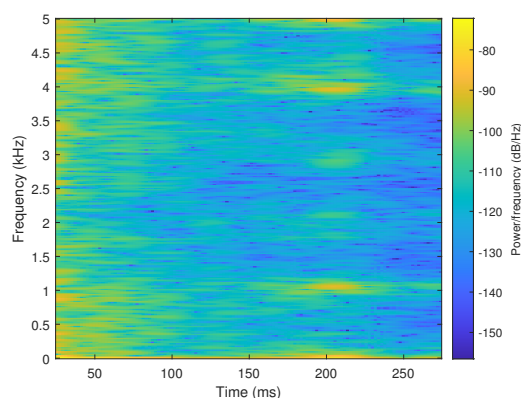


Figure 3: An example spectrogram of a RIR, used by the CNN to estimate the location of the transceiver.

al. [21] and Yokota et al. [22] used FDTD in comparative studies, where they show the relevance of the numerical approach to localization problems in acoustics. It could be possible to execute our research with other methods of simulation. For example, when using ray-tracing methods, a higher frequency signal could have been simulated, as explored by Vasiou et al. [23].

For this research, it is more important to know what FDTD does than knowing how the calculations are performed. For a detailed mathematical definition of FDTD, we refer to [24] by Schneider, in chapter twelve, the topic of acoustics is discussed separately. FDTD calculates the ‘next’ state of the pressure field based on all previous states. It calculates every next state of the field, given the previous states. Which in turn gives a result that is close to what is expected to be measured in practice. Using this algorithm we could define a room in MATLAB and calculate the pressure (sound) fields over a time of 0,3 seconds. 0,3 seconds is sufficiently long, as we can deduce that 0,3 seconds of measurements simulate paths over 100 meters long, inside a room of 10 by 10 meters. A snapshot of the FDTD simulation is depicted in Fig. 2, where the transceiver is located at the center of the visible wavefront, and the room is shown as the borders of the plot, with a small amount of reflectors located at the left portion of the room, breaking the symmetry of the room. The figure shows 170 by 170 nodes, as opposed to the expected 10 meters by 10 meters. This effect appeared due to the FDTD simulations requiring space to be discretized for calculating the sound waves in the room. The coordinates in the room that contained the location of the object were chosen to be random values in the xy-plane. Every simulation consisted of 0,3 seconds of sampling at a random location, with the signal source at the same location, simulating the transceiver.

4. Using Neural Networks for Localization

4.1. Fully Connected Neural Networks

The first simulation results consisted of 3000 samples, measuring pressure at the location of the object, which will be seen as the room impulse response (RIR). Along with the RIR, the ground

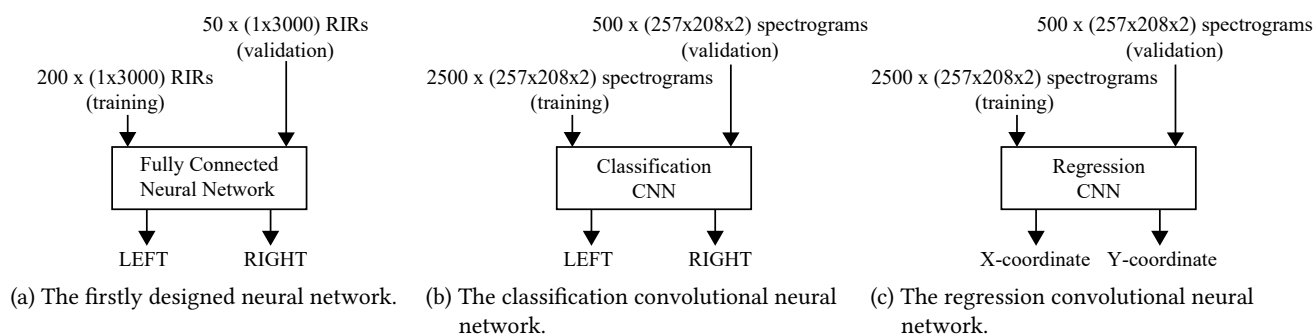


Figure 4: Illustrations of the three different neural networks produced for this paper. Notice that the (a) utilizes time-domain data, while the other networks use both the time and frequency information present in the spectrogram images.

truth location was added for every simulation to later use as a label in the neural networks. The location contains the x-, and y-coordinate in the simulated room. 250 simulations were constructed based on the same room with the same additive white Gaussian noise pulse. These simulations were the first real dataset that could be used as input for the fully connected neural network. It was important to use a simple network at first, for clarity reasons and to know that there is indeed the possibility to extract locations from the simulated data.

The main goal in this step was not to produce a precise network, but to produce an accurate network. This meant that the importance lies in the consistence of localization guesses, not the precision of those guesses. For this we designed a simple, fully connected classification network that had a high accuracy in contrast to its small dataset of 500 simulations, as seen in the confusion matrix in Fig. 6.

4.2. Classification Convolutional Neural Networks

After designing the first fully connected network, it could be remarked that training a fully connected network on the time series data would be less efficient than employing the frequency domain counterpart of the data. To achieve this, the fast Fourier transform (FFT) helped in making images (spectrograms) out of the time series RIR, which was also done in [8]. The dimensions of those spectrogram images change as the bin size used to store them changes, where a higher bin size or resolution contains more information. The downside of using a higher resolution is higher computational load. In the end, the original matrix was a complex 257 by 208 matrix. That matrix was split up into two channels. With the first channel being the amplitude and the second channel containing the phase. By using the spectrogram, the amplitude and phase could be used as input features. A downside of using this type of network is that the localization output is limited to a fixed set of outcomes, which drastically limits the potential accuracy obtained by the system.

4.3. Regression Convolutional Neural Networks

The final iteration of the algorithm needed to perform monostatic localization consisted of a convolutional network with three convolution layers, each followed by a ReLU and normalization layer. The structure of this network can be seen in Fig. 5. The input consists of the same type of spectrogram used in the classification type of network. The benefit of using a regression network is that the output is not limited to a fixed set of predetermined outcomes, but returns a set of coordinate estimations.

4.4. Overfitting

Overfitting was a recurring obstacle during the execution of this research. It is well known that overfitting is bound to be a problem in every research involving (convolutional) neural networks. The frequent occurrence of over-fitting makes it so that large quantities of different methods exist to counteract the over-fitting problems. As an example, Srivastava et al. [25] use dropout layers in deep neural networks. For this research, a multitude of methods were used to counteract overfitting. Firstly, we used lower initial learning rates to stop the weights from reaching their end values too fast. If the learning rate is too high, the network will learn too many features from the training data and will overfit. Additionally, we used larger datasets, containing 3000 simulations, helping restrain the overfitting. If a network uses more (diverse) data during training, it is intuitive that the network will learn less trivial, wrong features. Also the switch from classification to regression CNNs made overfitting occur after less epochs, so in a later stage of the training stage. Since using a binary output ‘left’ and ‘right’ generates no difference in interpreting ‘far left’ or ‘close left’, while regression generates an exact coordinate.

4.5. Designing the Layers and Hyperparameters

When designing layers for a neural network, two general starting points can be chosen. The first option is to make a minimalist network which has as little complexity as possible, slowly adding complexity until the desired specifications are reached. The other starting option is the

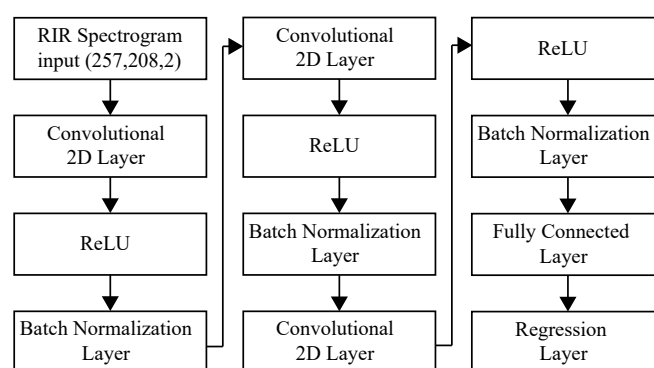


Figure 5: The structure of the regression Convolutional Neural Network used to obtain the final results in this paper.

exact opposite, starting with an especially complex network and whittling down the complexity until reaching the desired results.

In this research, we employed a combination of these two approaches. The model on which the first CNN was based, was a classification network made for lung sound analysis available within our lab. This was a more complex network than what was needed for this application but it laid the ground work for spectrogram images as input data. This meant using the second starting point, a complex neural network that can be whittled down into a usable network. We lowered the number of weights and biases by changing the kernel and stride sizes. The art of designing layers for neural networks relies on trial-and-error, as explored by Suganuma et al. [26].

5. Results and Discussion

5.1. Fully Connected Neural Network Results

After making a small dataset of 100 RIRs consisting of 3000 samples. The first, fully connected network using time-domain data as input, was able to be trained and tested. The earliest results were then reached by testing the network and plotting the confusion matrix on the limited dataset. This is illustrated in Fig. 6.

As seen in Fig. 4a, the output consists of guessing whether the object is LEFT or RIGHT in the room. Which corresponds to the left hand side and the right hand side of the simulated room. The room is split up in left and right by dividing the x coordinate in two and deciding the border at that x value. Fig. 6 shows the confusion matrix of that network, with zero corresponding to LEFT and one corresponding to right. The network is capable of estimating the rough location of the transceiver with an accuracy of 81%.

5.2. Classification With Convolutional Neural Network

The results from the first versions of the convolutional network were not optimized to a useful degree. The added complexity and small dataset made classification harder. The used network was too complex for the amount of data available. Overfitting was substantially big that learning would quickly halt. This did not mean that making this network was in vain, the goal of the classification network was so that the regression network had firm fundamentals. We learned the importance of the relation between complexity of the network and the amount of available data. Also ways of lowering learning rates were learned while minimizing overfitting. This network served as a stepping stone to the next results, as this network validated the possibility to use spectrogram images of the RIR data to perform rough localization. This paved the way for the third version of the network that adds regression to perform more accurate estimations.

Confusion Matrix

Output Class	0	30 30.0%	3 3.0%	90.9% 9.1%
	1	16 16.0%	51 51.0%	76.1% 23.9%
		65.2% 34.8%	94.4% 5.6%	81.0% 19.0%
	\	Target Class		

Figure 6: Illustration of the confusion matrix of the first fully connected network which made rough estimations on the transceiver being on either the left side, or the right side of the room. with zero corresponding to ‘left’ and one corresponding to ‘right’. The obtained accuracy is 81%.

5.3. Regression With Convolutional Neural Network

The benefit of using regression instead of classification is that regression trains continuous variables instead of specified labels. This makes it more suitable for applications such as this one, where two parameters need to be estimated accurately. In Fig. 8 an error histogram is shown that shows the estimation error distribution between the true locations and location estimates by the most accurate network we could design and train during this research. Note that the error function is the euclidean distance between the two points. The simulated room wherein these predictions took place is visible in Fig. 2. The data is extracted out of 300 position estimates and shows that the majority of estimations have an estimation error below 20 cm, with a total average of 14,7 cm.

In Fig. 7 a single example of a set of estimated coordinates, generated using the network shown in Fig. 5, is depicted. The room is displayed within the boundaries of the plot. The results are promising and encourage future research on this topic. The height dimension could be added in a later iteration of this research, but was not deemed relevant for the current application.

To contribute to the robustness of the research, tests were performed on the trained networks used for the last results. The goal was to show that the localization is not random and is far more precise and accurate than random guessing. When using the same 300 simulations used

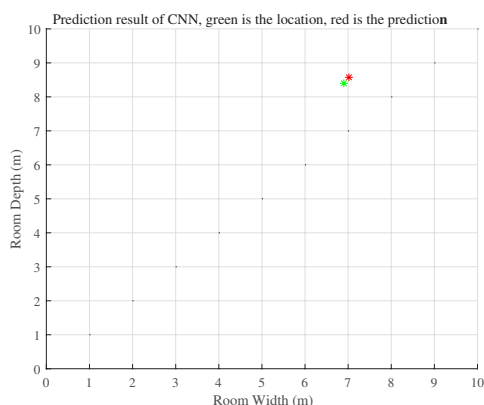


Figure 7: Example of a set of coordinates, estimated by the regression CNN when using the spectrogram or the measured RIR as input. Ground truth coordinate: [6,9 m; 8,4 m], estimated coordinate: [7,0183 m; 85,744 m]

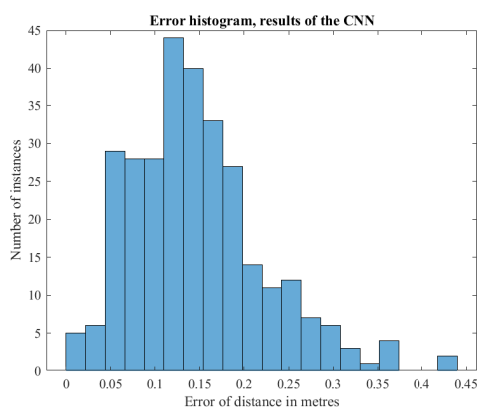


Figure 8: Histogram that plots the errors of validation data passed through the final regression convolutional neural network. The error is the distance between guess and ground truth in a simulated 10x10x4 (LxWxH in m) room.

in Fig. 8, random guesses resulted in a mean error distance of 4,0961 meter. This proves that the results discussed, with a mean error of 0,1473 meters in the same environment, are more accurate.

6. Conclusion and Discussion

The research question that was handled in this paper was the following: “Is it possible to localize an object inside a known room using only one sound transceiver? If it is possible, how accurate can the measurement be without the help of this additional infrastructure and which additional intelligent algorithms will be needed?” This paper came to the conclusion that it

is indeed possible to accurately localize an object in a room, simulated using finite-difference time-domain numerical techniques. Different types of networks were tested, starting with a classification approach where rough estimates about the position of the transceiver were made based on time-domain recordings. To improve accuracy and extract more information out of the time-domain data, the research switched to working with spectrogram images of the recorded data, which made it possible to use convolutional neural networks. To reach a mean accuracy of under 15 cm, a regression convolutional neural network was needed. This network was trained on more than 2000 different spectrograms of room impulse responses.

7. Future Work

This work was constrained to use only one microphone/transceiver. This made the research interesting, differing from conventional auralization and localization research. Future research may consist of transferring the simulated networks to real life scenarios. Bianco et al. [27] suggest the use of transfer learning and make a summary of multiple successful studies reaching accurate localization in real world scenarios by employing transfer learning. Due to limitations in time and computational power, the research in this paper was forced to use a sub-optimal measuring sequence for this type of application. For future research, it may also be useful have a measuring sequence with a larger bandwidth and explore the use of coded emissions which would allow multiple objects to be tracked simultaneously. Also the influence of the transceiver beam pattern should be examined. We hope that future research may build upon the concepts and approaches formed in this study.

References

- [1] B. Siciliano, O. Khatib, Springer Handbook of Robotics, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. URL: <http://link.springer.com/10.1007/978-3-540-30301-5>. doi:10.1007/978-3-540-30301-5.
- [2] G. Seco-Granados, J. López-Salcedo, D. Jiménez-Baños, G. López-Risueño, Challenges in indoor global navigation satellite systems: Unveiling its core features in signal processing, *IEEE Signal Processing Magazine* 29 (2012) 108–131. doi:10.1109/MSP.2011.943410.
- [3] J. Urena, A. Hernandez, J. J. García, J. M. Villadangos, M. Carmen Perez, D. Gualda, F. J. Álvarez, T. Aguilera, Acoustic local positioning with encoded emission beacons, *Proceedings of the IEEE* 106 (2018) 1042–1062.
- [4] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, M. Vetterli, Acoustic echoes reveal room shape, *Proceedings of the National Academy of Sciences* 110 (2013) 12186 – 12191.
- [5] S. Khattak, C. Papachristos, K. Alexis, Visual-thermal landmarks and inertial fusion for navigation in degraded visual environments, *CoRR* abs/1903.01656 (2019). URL: <http://arxiv.org/abs/1903.01656>. arXiv:1903.01656.
- [6] T. J. Cox, P. D’Antonio, Acoustic Absorbers and Diffusers, volume 4, 2009. URL: <http://arxiv.org/abs/1011.1669><http://dx.doi.org/10.1088/1751-8113/44/8/085201>. doi:10.4324/9781482266412.
- [7] N. Le Dortz, F. Gain, P. Zetterberg, WiFi fingerprint indoor positioning system using

- probability distribution comparison, *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings (2012)* 2301–2304. doi:10.1109/ICASSP.2012.6288374.
- [8] Y. Nagama, T. Umezawa, N. Osawa, Indoor localization based on analysis of environmental ultrasound, in: *IPIN (Short Papers/Work-in-Progress Papers)*, 2019, pp. 423–430.
- [9] J. M. Vera-Diaz, D. Pizarro, J. Macias-Guarasa, Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates, *Sensors (Switzerland)* 18 (2018). URL: <https://pubmed.ncbi.nlm.nih.gov/30322007/>. doi:10.3390/s18103418. arXiv:1807.11094.
- [10] P. Stoica, H. He, J. Li, Optimization of the receive filter and transmit sequence for active sensing, *IEEE Transactions on Signal Processing* 60 (2012) 1730–1740. doi:10.1109/TSP.2011.2179652.
- [11] H. He, J. Li, P. Stoica, *Waveform design for active sensing systems: a computational approach*, Cambridge University Press, 2012.
- [12] N. Antonello, E. De Sena, M. Moonen, P. A. Naylor, T. van Waterschoot, Room impulse response interpolation using a sparse spatio-temporal representation of the sound field, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25 (2017) 1929–1941. doi:10.1109/TASLP.2017.2730284.
- [13] D. El Badawy, I. Dokmanić, Direction of arrival with one microphone, a few legos, and non-negative matrix factorization, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26 (2018) 2436–2446. doi:10.1109/TASLP.2018.2867081.
- [14] S. Cecchi, A. Carini, S. Spors, Room response equalization—a review, *Applied Sciences* 8 (2017) 16. URL: <https://doi.org/10.3390/app8010016>. doi:10.3390/app8010016.
- [15] A. D. Stan G, Embrechts J, Comparison of different impulse response measurement techniques, *AES: Journal of the Audio Engineering Society* 50 (2002) 249–262.
- [16] E. Vargas, J. R. Hopgood, K. Brown, K. Subr, On improved training of cnn for acoustic source localisation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021) 720–732. doi:10.1109/TASLP.2021.3049337.
- [17] M. Markovic, S. K. Olesen, D. Hammersho/i, Three-dimensional point-cloud room model in room acoustics simulations 133 (2013) 3532–3532. URL: <https://sfx.aub.aau.dk/sfxaub?sid=pureportal&doi=10.1121/1.4806371>. doi:10.1121/1.4806371.
- [18] E. Deines, M. Hering-Bertram, J. Mohring, J. Jedorovs, F. Oberste-Dommes, G. Nielson, Comparative visualization for wave-based and geometric acoustics, *Visualization and Computer Graphics, IEEE Transactions on* 12 (2006) 1173–1180.
- [19] L. Savioja, U. P. Svensson, Overview of geometrical room acoustic modeling techniques, *The Journal of the Acoustical Society of America* 138 (2015) 708–730. URL: <https://doi.org/10.1121/1.4926438>. doi:10.1121/1.4926438.
- [20] D. Y. Maa, The flutter echoes, *The Journal of the Acoustical Society of America* 13 (1941) 170–178. URL: <https://doi.org/10.1121/1.1916161>. doi:10.1121/1.1916161.
- [21] E. De Sena, N. Antonello, M. Moonen, T. van Waterschoot, On the modeling of rectangular geometries in room acoustic simulations, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23 (2015) 774–786. doi:10.1109/TASLP.2015.2405476.
- [22] T. H. Yokota T, Sakamoto S, Comparison of room impulse response calculated by the simulation methods based on geometrical acoustics and wave acoustics, *Institute of*

- Industrial and Science, University of Tokyo (2002) 2715–2716.
- [23] E. Vasiou, K. Shkurko, I. Mallett, E. Brunvand, C. Yuksel, A detailed study of ray tracing performance: render time and energy cost, *The Visual Computer* 34 (2018). doi:10.1007/s00371-018-1532-8.
- [24] J. B. Schneider, *Understanding the Finite-Difference Time-Domain Method*, Washington, 2020.
- [25] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, *Journal of Machine Learning Research* 15 (2014) 1929–1958. URL: <http://jmlr.org/papers/v15/srivastava14a.html>.
- [26] M. Suganuma, S. Shirakawa, T. Nagao, A genetic programming approach to designing convolutional neural network architectures, in: *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO '17*, Association for Computing Machinery, New York, NY, USA, 2017, p. 497–504. URL: <https://doi.org/10.1145/3071178.3071229>. doi:10.1145/3071178.3071229.
- [27] M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, C.-A. Deledalle, Machine learning in acoustics: Theory and applications, *The Journal of the Acoustical Society of America* 146 (2019) 3590–3628. URL: <https://doi.org/10.1121/1.5133944>. doi:10.1121/1.5133944.