Improving accuracy of sound reflection estimation using neural networks

Ekko Scholtens[1]

Supervisor(s): Elmar Eisemann[1]**, Jorge Martinez Castaneda**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 29, 2023

Name of the student: Ekko Scholtens
Final project course: CSE3000 Research Project
Thesis committee: Elmar Eisemann, Jorge Martinez Castaneda, Marcus Specht

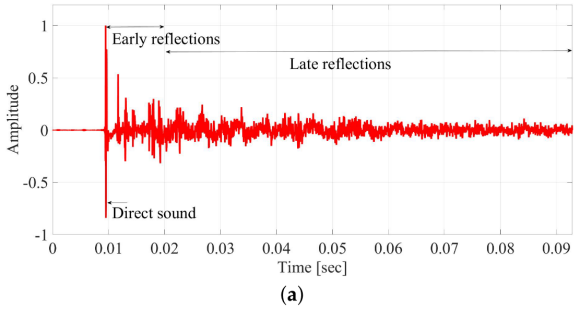
Abstract

In this paper, we present a study to improve using neural networks for acoustic reflection localization. Our study focuses on the reimplementation of the proposed neural network model by Bologni et al. and investigates the effect of adding a third microphone to the microphone array. We reimplemented and trained the neural network using the same framework and hyperparameters as the original model, and then evaluated it using the same metrics. Our results show that the addition of a third microphone improves the amount of detected sources from 43% to 58%, it also improved the front-back ambiguity from 25% to 18%. Conclusively, have demonstrated the potential benefits of adding a third microphone to the neural network approach for acoustic reflection localization.

1 Introduction

When it comes to sound, the acoustics of a room can make all the difference. Imagine being fully immersed in a concert, with each note resonating perfectly, making the sound come to life. But unfortunately, bad room acoustics can render audio inaudible and murky[1], robbing us of that immersive experience.

The acoustics of a room can be divided into the following three main parts, depicted in Figure 1, the direct sound, early reflections, and reverberation (late reflections). The direct sound is the sound that comes directly from the source, used by humans to localize sound[2]. Next the early reflections, which are the first reflected sounds by, for example, walls as depicted in Figure 2. This can be useful for speaker manufacturers in order to create phantom channels with a limited set of real channels[3, 85–86]. However, this become an issue when the phantom sources appear in places we do not want them.



(a)

Figure 1: Acoustic parts: direct sound, early reflections, and reverberation (late reflections); from [1]

Filtering out all the reflections is not always a good idea, because we want some reverberation to create the feeling of spaciousness in the room[2]. However too much reverberation can cause the sound to become murky and speech to become inaudible[4]. This leads to the problem of filtering some

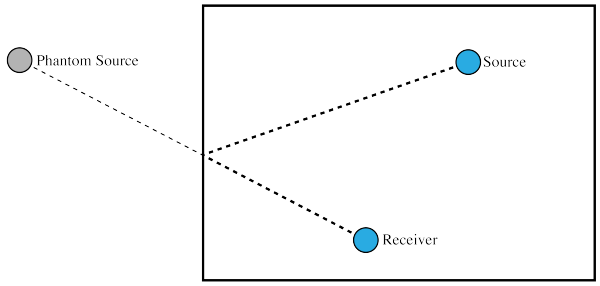


Figure 2: Reflected phantom source in a room

but not all reflections. One solution is to install diffusion or absorption panels[5] to filter some of the most problematic, but the question then becomes where to place these panels. The first step is to locate the most prevalent reflections in the room, and this is where the current research comes in.

1.1 Current solutions

One important question when trying to improve the acoustics of a room is where to place sound-diffusing or absorbing panels using an inexpensive way to find locations. A key step in this process is identifying the areas where reflections are most prevalent. One approach to find the location of reflections would be to match reflections received by a receiver, but unfortunately this can be reduced to a known NP-Hard problem known as the “maximum independent set” problem[6].

Currently, most of the research has focused on large rooms like concert halls or idealized rooms. While other research has been conduct for smaller rooms these are not without issue, most of these approaches uses custom or expensive microphones/arrays. For example Tatsuya et al. use a custom 6-channel microphone to measure the angle of the incoming sound[7], or Farina and Tronchin used an expensive 32-channel Eigenmike[8]. While these studies have yielded positive results, they do not meet the criterion of being cost-effective for small scale use.

Other research groups use arrays of many separate microphones like Sasaki et al.[9] and Tamai et al.[10] who used 64 and 32 microphones respectively in a spherical array on a mobile robot. These have the problem that they are impractical to uses in an ordinary room due to the physical size of the array. Another approach by Riemens et al. is to use multiple kinds of sensors, they added a system called LIDAR (Laser Imaging, detection, and ranging[11]) where they use lasers to detect walls[12].

Another approach for finding reflections use is the utilization of neural networks with only two inexpensive subcardioid microphones, like Bologni et al. describes in their paper[13]. In their paper, they describe a way for someone to use two inexpensive subcardioid microphones to measure angles and distance of reflections in a room, and subsequently estimates the room size. However, there are some limitations that need to be addressed.

One limitation is the capability of the network to detect reflections in real room where noise can interfere. Additionally, there is also the issue of front-back ambiguity which

can cause uncertainty in the output, making it difficult to determine the true direction of the reflections. These issues are particularly important when trying to locate reflections in a room, as it can significantly impact the accuracy of the network. Overall, while the neural network approach is a promising solution for measuring reflections in a room, there are still some limitations that need to be addressed.

In this paper, we aim to address these limitations of the neural network-based approach by answering the following questions: Can the addition of a third microphone reduce the front-back ambiguity, and improve accuracy in detecting reflections in a room?

2 Methodology

In order to answer the questions asked in the previous section, we first describe how we generated the dataset we used for training and validation. We then outline the steps taken to reimplement the network, and train and evaluate the model. The reimplementation serves as a case study for understanding the capabilities and limitations of using the neural networks approach for real-world applications, and if a third microphone can reduce the error.

We begin by describing how we generated the dataset we used for training, and validation. We generate the data using the same constraints mentioned in the paper by Bologni et al.[13], namely:

- The microphone array and source can not be with in 50 cm of each other and must be 50 cm from the walls.
- The microphone array cannot be directly in front or behind the source with a tolerance of ± 3 degrees.
- Only simulate three image orders.

To generate the dataset we used the python package pyroomacoustics[14] to generate the synthetic dataset, using the image source model as the simulation.

The generated dataset consists of 9 rooms with different sizes ranging from 2 by 2 meter to just below 8 by 8 meter. For each room we generate a number of location pairs, these consist of a source and receiver location, using the constraints mentioned previously. For each receiver position we generate 150 random rotations of the microphone array and save each as a 16kHz wav file. We simulate the microphone array as an array of three subcardioid microphones with an inter microphone distance of 10 cm, as depicted in Figure 3. We chose these parameters based on experimentation, these gave a good balance between training time and results.

We use subcardioid microphones in this study because they have some directional sensitivity, where they are most sensitive to sounds coming from the front, and least sensitive to sounds coming from the rear. This makes them well suited for capturing sounds coming from a specific direction, such as a sound source in a room. And as a result reduce front-back ambiguity[13].

Next, we reimplemented the neural network using the same open-source neural network framework NNabla[15] as the original paper[13]. We implemented the network as described in the paper. After validating the two channel version against the results Bologni et al. found in their paper, we altered the network by adding a third channel to the feature extractor block. We will be using the same hyperparameters as given in the paper.

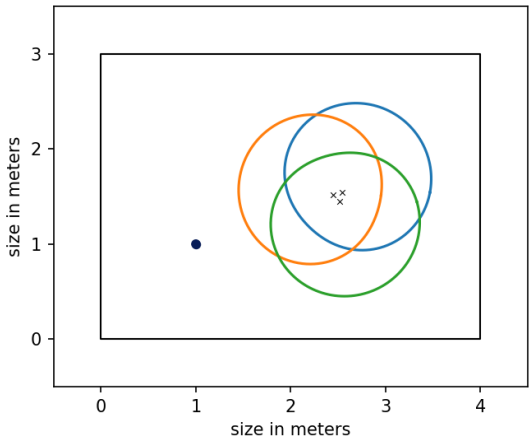


Figure 3: Subcardioid microphone array (depicted as $\times$) with source (blue circle) in a room.

3 Experimental Setup and Results

In this section we present the results of our reimplementation and alterations. We will compare our results against the original results precented in the paper by Bologni et al.[13]. We trained two networks: a version with two channels to compare our implementation against the original, and one with the proposed changes to make it three channels. We have trained all these versions on the Delft HPC[16] for around 800 epochs, after which there was no more notable improvement.

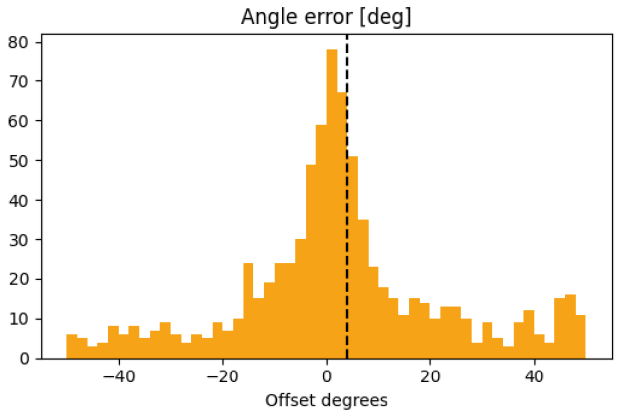
For training, we used the implementation and dataset described in the previous section. In Table 1 are the results of the first two trained networks and the results presented by Bologni et al.[13]. Our implementation of the two channel version detected 43% of the sources in the test dataset, this is comparable to the original version. Our version with three channels did however outperform both with 58% of the sources detected.

	Bologni et al.[13]	two-channel	three-channel
Detected sources	49%	43%	58%
Front-back	–	25%	18%

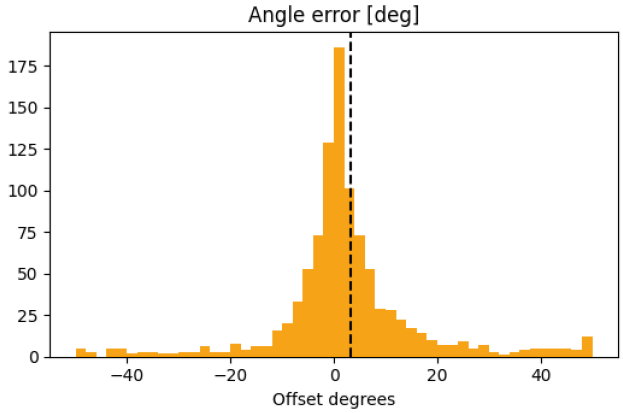
Table 1: Results comparison between Bologni[13] and our proposed implementation.

In the original paper there was no mention of the front-back ambiguity metric for rooms with multiple reflected surfaces, however adding the third channel does improve this metric too.

In Figure 4 we show a comparison of the angular error of the two versus three microphone trained networks. The three microphone version is more concentrated around the centre,



(a) Two microphone network predicted angle error distubution



(b) Three microphone network predicted angle error distubution

Figure 4: Angular error comparison between two(a) microphones and three(b) microphones

with a smaller standard deviation of 44.3 degrees compared against 51.5 degrees for the two microphone network.

4 Responsible Research

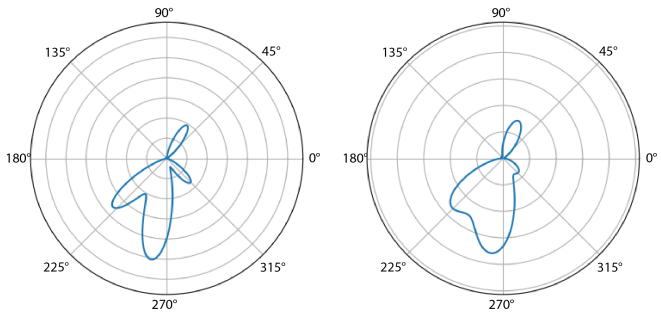
In this paper we have shown that the original paper is reproducible by reimplementing the network and training it. We have followed the methods and procedures outlined in the original paper, also used the same performance metrics to evaluate the performance of our reimplementation, and have compared our results to the results reported in the original paper. This allows us to demonstrate that the original results are reproducible, and that our reimplementation replicates the performance of the original network.

In addition to demonstrating reproducibility, we have also made clear what our extension is and what we have changed for our results. Specifically, we have added another channel of data to the network, which has improved its performance. By providing this information, we are transparent in our methodology and show how it differs from the original paper, and we can provide a clear and detailed explanation of how our results differ from the original results.

5 Discussion

The main objective of this study is to evaluate the effect of adding a third microphone channel to the neural network architecture by Bologni et al.[13]. We want to show that adding a third microphone can reduce the front-back ambiguity and that it can improve the accuracy. Our results show that our implementation of the two-channel version performed comparably to the original version, with 43% of the sources detected. When we added a third microphone channel, the performance of the network improved significantly, with 58% of the sources detected.

These results demonstrate that adding a third microphone can help improve the accuracy of the localization by increasing the number of detected source by 15% and decreasing the standard deviation by 6 degrees, which can be beneficial in real-world scenarios. Furthermore, the improvement in the front-back ambiguity decreased from 25% to 18% by adding the third microphone, which shows that it also helps to reduce the ambiguity.



(a) Expected angle (b) predicted angle

Figure 5: Expected(a) and predicted(b) angles of image sources in a single simulated sample

In Figure 5 we show a single sample comparison between the expected and the predicted output of the neural network. It shows that the predicted result could be used to find reflections in a room, and in turn find place where to place sound absorption panels. However, it also shows that the error in the predicted angle quadrant 1 is the wrong angle, it should be more to the 45 degrees grid line. In a small room this should not mater as much, but when the size of the room increases the error does too, leading to that the reflected image being in a wrong place.

In our study we find that with the current network architecture the network has to train for variables we do not use such as the distance of the reflection. Since we are only interested in the angle to determine where the reflections are, we could redesign the network to only output the angle. We conducted some initial experiments that showed promising results. However, this requires further research due to the nature of the network to overfit and the need for additional adjustments to the network architecture.

6 Conclusions and Future Work

In conclusion, the main goal of paper is to evaluate if adding a third microphone to the neural network approach for detecting sound reflections by Bologni et al.[13] could reduce the front-back ambiguity and could it improve accuracy. Our results show that adding a third microphone can significantly improve the accuracy by detecting more sources and reduces the front-back ambiguity. These findings provide insights for future research and can guide the design of real-world applications that utilize this method.

One limitation of this study is that we trained and validated the network using a simulated dataset, which does not capture the complexity of real-world rooms. In future work, it would be interesting to evaluate the proposed method on real-world rooms. Another thing to consider in future work is to test the network with more complex rooms shapes and acoustic properties, in order to get a better understanding of its capabilities and limitations.

Another avenue for future research is to change the neural network to only output the angles of the reflections, we did conduct some initial experiments that showed promising results. However, due to the tendency of the network to overfit and the need for additional adjustments to the network architecture this required future research.

References

[1] Stefania Cecchi, Alberto Carini, and Sascha Spors. Room Response Equalization—A Review. *Applied Sciences*, 8(1):16, January 2018. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/app8010016.

[2] Andrew D. Brown, G. Christopher Stecker, and Daniel J. Tollin. The Precedence Effect in Sound Localization. *JARO: Journal of the Association for Research in Otolaryngology*, 16(1):1–28, February 2015. doi:10.1007/s10162-014-0496-2.

[3] Floyd E. Toole. *Sound Reproduction: Loudspeakers and Rooms*. Taylor & Francis, 2008. ISBN: 9780240520094.

[4] Murray Hodgson and Eva-Marie Nosal. Effect of noise and occupancy on optimal reverberation times for speech intelligibility in classrooms. *The Journal of the Acoustical Society of America*, 111(2):931–939, February 2002. Publisher: Acoustical Society of America. doi:10.1121/1.1428264.

[5] Peter D'Antonio and Trevor J Cox. Diffusor application in rooms. *Applied Acoustics*, 60(2):113–142, June 2000. doi:10.1016/S0003-682X(99)00054-7.

[6] Ingmar Jager, Richard Heusdens, and Nikolay D. Gaubitch. Room geometry estimation from acoustic echoes using graph-based echo labeling. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, March 2016. ISSN: 2379-190X. doi:10.1109/ICASSP.2016.7471625.

[7] Ohta Tatsuya, Yano Hiroo, Yokoyama Sakae, and Tachibana Hideki. Sound Source Localization by 3-D Sound Intensity Measurement using a 6-channel microphone system Part 2 : Application in room acoustics. In *INTERNOISE 08*, 2008.

[8] Angelo Farina and Lamberto Tronchin. 3D Sound Characterisation in Theatres Employing Microphone Arrays. *Acta Acustica united with Acustica*, 99(1):118–125, January 2013. doi:10.3813/AAA.918595.

[9] Yoko Sasaki, Mitsutaka Kabasawa, Simon Thompson, Satoshi Kagami, and Kyoichi Oro. Spherical microphone array for spatial sound localization for a mobile robot. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 713–718, October 2012. ISSN: 2153-0866. doi:10.1109/IROS.2012.6385877.

[10] Y. Tamai, Y. Sasaki, S. Kagami, and H. Mizoguchi. Three ring microphone array for 3D sound localization and separation for mobile robot audition. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4172–4177, August 2005. ISSN: 2153-0866. doi:10.1109/IROS.2005.1545095.

[11] Travis S. Taylor. *Introduction to Laser Science and Engineering*. CRC Press, August 2019. ISBN: 9781138036390.

[12] Ellen Riemens, Pablo Martínez-Nuevo, Jorge Martinez, Martin Møller, and Richard C. Hendriks. On the Integration of Acoustics and LiDAR: a Multi-Modal Approach to Acoustic Reflector Estimation, June 2022. arXiv:2206.03885 [eess]. doi:10.48550/arXiv.2206.03885.

[13] Giovanni Bologni, Richard Heusdens, and Jorge Martinez. Acoustic Reflectors Localization from Stereo Recordings Using Neural Networks: ICASSP 2021. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 461–465, 2021. Place: Piscataway Publisher: IEEE. doi:10.1109/ICASSP39728.2021.9414473.

[14] Pyroomacoustics. URL: https://pyroomacoustics.readthedocs.io/.

[15] Neural Network Libraries by Sony. URL: https://nnabla.org/.

[16] Delft High Performance Computing Centre (DHPC). DelftBlue Supercomputer (Phase 1), 2022. URL: https://www.tudelft.nl/dhpc/.