Audio Engineering Society

Late Breaking Demo Paper

Presented at the AES International Conference on
Artificial Intelligence and Machine Learning for Audio
2025 September 8–10, London, UK

Investigating the use of Deep Convolutional Neural Networks for Direction-of-Arrival Estimation on Raw Stereo Audio

Samuel Hobern[1], Alan Archer-Boyd[2], and Damian Murphy[1]

[1]*AudioLab, School of Physics, Engineering and Technology, University of York, York, UK*
[2]*BBC Research and Development, The Lighthouse, White City Place, 201 Wood Lane, London, W12 7TQ*

Correspondence should be addressed to Samuel Hobern (sam.hobern@york.ac.uk)

ABSTRACT

This paper outlines the use of raw stereo audio as the input to a Deep Convolutional Neural Network for the purpose of Direction-of-Arrival (DOA) estimation based on source localisation over the frontal hemisphere of a 50-point Lebedev loudspeaker array using a pair of spaced stereo microphones. Results show the model is capable of a high classification accuracy with evidence of some ability to generalise to unseen data highlighting the benefits of raw audio as the input feature to models.

1 Introduction

Sound Source Localisation (SSL) is a challenging task with numerous applications including robotics, fault detection in industrial machinery, and audio upmixing [1]. This challenge can be solved by estimating the Direction-of-Arrival (DOA). SSL is typically tackled using Machine Learning (ML) based methods due to their high accuracy and robustness to noise [1].

2 Methods

In this work, ML based DOA estimation employed a classification approach using a modified version of the *SampleCNN* architecture [2]. In order to avoid confusion, the model created in this paper shall be referred to as *SampleDOA*.

Many models for this task use empirically-derived features such as the Generalised Cross Correlation Phase Transform (GCC-PHAT) or Mel-Spectrograms. However, these can inadvertently bias the model by limiting the information the model has to train on. Instead, the models were trained on raw, stereo audio. Impulse responses were recorded within the University of York AudioLab's 50 loudspeaker Lebedev sampled sphere for numerous stereo microphone configurations including a 40cm spaced pair of AKG C414s set to the omnidirectional mode and denoted as 'AB_Omni_40' [3]. The impulses responses were convolved with the NIGENS database [4]. The 8 splits allocated by NIGENS were used with 6 for training, 1 for validation, and 1 for testing.

A model was created, named *SampleDOA_5_6* with

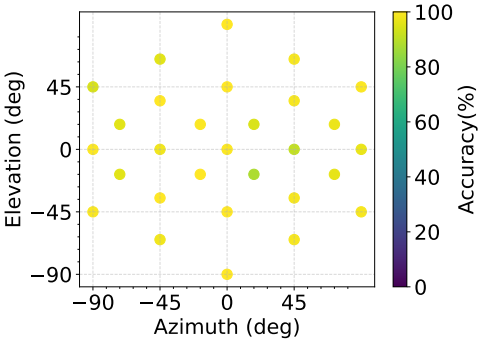

Fig. 1: *SampleDOA*_5_6 Azimuth Accuracy on the AB_Omni_40 Dataset

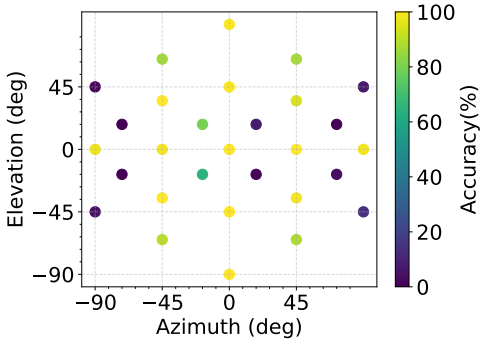

Fig. 2: *SampleDOA*_5_6 Azimuth Accuracy on the AB_Omni_30 Dataset

6 convolutional layers each with a convolutional filter size of 5. With an input of 78125 samples (9.77s at an 8kHz sampling rate), the model is capable of extracting sample level features. Each layer consists of a 1D Convolution, followed by Batch Normalisation, ReLU activation, and a Max Pooling operation. The *SampleDOA* model was tasked with estimating the azimuth of sound sources in the frontal hemisphere only; the azimuth (θ) and elevation (ϕ) were both bounded between $[-90°, 90°]$. The Adam optimiser was used with a learning rate of $1e-5$, and a batch size of 8.

3 Results

After 1000 epochs, the SampleDOA_5_6 model achieved an accuracy of 97.6% as shown for each loudspeaker position in Figure 1. The azimuth estimation accuracy is largely consistent across loudspeaker positions with slight deviations potentially caused by having fewer samples of those loudspeaker positions being included in the training data. To evaluate the performance of the model, it was tested against the 'AB_Omni_30' dataset. This dataset follows the same microphone configuration as 'AB_Omni_40' except the spacing between the microphones is 30cm instead of 40cm. As seen in Figure 2, the classification accuracy of the model decreases to approximately 62.5%. However, the model begins to show evidence of generalisation. At an elevation of 0° the model maintains the same level of accuracy along all loudspeakers as it did for the AB_Omni_40 dataset. This is also true for the loudspeakers positioned along an azimuth of 0°. Performance suffered most significantly at loudspeaker positions at the boundaries of the hemisphere as denoted by the darker spots in Figure 2.

4 Discussion

The results indicate the *SampleDOA* architecture is capable of performing Direction-of-Arrival estimation on raw stereo audio at a high degree of accuracy. The performance on the AB_Omni_30 dataset suggests evidence of the model beginning to generalise. However, further investigation is required. Training the model on multiple datasets of differing microphone configurations may improve its ability to generalise and will be investigated in future work.

References

[1] Khan, A., Waqar, A., Kim, B., and Park, D., "A review on recent advances in sound source localization techniques, challenges, and applications," *Sensors and Actuators Reports*, 9, p. 100313, 2025, ISSN 2666-0539, doi:10.1016/j.snr.2025.100313.

[2] Lee, J., Park, J., Kim, K. L., and Nam, J., "SampleCNN: End-to-End Deep Convolutional Neural Networks Using Very Small Filters for Music Classification," *Applied Sciences*, 8(1), p. 150, 2018, ISSN 2076-3417, doi:10.3390/app8010150, number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[3] Turner, D. and Murphy, D., "Dataset of stereo and multi-channel IRs for a 50-point Lebedev quadrature." 2023, doi:10.5281/zenodo.7990195.

[4] Trowitzsch, I., Taghia, J., Kashef, Y., and Obermayer, K., "The NIGENS General Sound Events Database," 2020, doi:10.48550/arXiv.1902.08314, arXiv:1902.08314 [cs].