# Modelling human sound localization with deep neural networks

Kiki van der Heijden<sup>1</sup> and Siamak Mehrkanoon<sup>2</sup>

1- Maastricht University - School for Mental Health and Neuroscience (MHeNs), Maastricht Centre for Systems Biology (MaCSBio). Universiteitssingel 40-60, 6229 ER Maastricht - Netherlands

2- Maastricht University - Department of Knowledge Engineering. Paul-Henri Spaaklaan 1, 6229 EN Maastricht - Netherlands

**Abstract**. How the brain transforms binaural, real-life sounds into a neural representation of sound location is unclear. This paper introduces a deep learning approach to address these neurocomputational mechanisms: We develop a biological-inspired deep neural network model of sound azimuth encoding operating on auditory nerve representations of real-life sounds. We explore two types of loss functions: Euclidean distance and angular distance. Our results show that a network resembling the early stages of the human auditory pathway can predict sound azimuth location. The type of loss function modulates spatial acuity in different ways. Finally, learning is independent of environment-specific acoustic properties.

### 1 Introduction

Humans use spatial hearing to rapidly localize events in the environment, and to separate sound sources in multi-source listening environments (e.g. to filter out the voice of an acquaintance in a noisy bar). Despite extensive research into human sound localization, it remains unclear how the brain computes the location of real-life sounds in real-world listening environments. Specifically, empirical studies of neural sound location processing mainly focus on simple sounds (e.g. noise bursts) in controlled listening environments (e.g. without reverberation) that have little ecological validity [1]. In addition, computational studies investigating the representational mechanisms of neural sound location encoding are sparse [1][2]. Here, we explore a biological-inspired deep neural network (DNN) model that estimates the location of real-life sounds in real-world listening situations. Such a model can be used in future empirical studies to investigate the complex computational and representational mechanisms underlying neural location encoding of real-life sounds in humans.

#### 1.1 Existing DNNs and present approach

Deep neural networks developed in the context of computational environmental audio analysis are highly successful, with error scores on combined azimuthelevation estimation as small as  $3^{\circ}$  (even in reverberant listening conditions [3]). Although the specific architectures vary, most models consist of a number of convolutional layers followed by one or multiple recurrent layers, and use a multi-output regression task to estimate sound location on a continuous scale in Cartesian coordinates [3][4]. However, these models differ from neurobiological sound location processing in humans: Input is typically derived from microphone arrays (consisting of four or more channels), while human hearing is based on binaural (i.e. two-channel) input. Moreover, most models do not receive sound waves as input but a priori extracted features such as phase, time, or spectral inter-channel differences [3][4].

Here, we aim to develop a DNN that resembles the functioning of the early stages of the human auditory pathway. DNN models are ideally suited for studying neurobiological systems because they operate in a hierarchical manner abstracting from simple to more complex representations akin to neural encoding of sensory stimuli. Further, DNNs use an input-driven learning procedure to extract relevant features and thus avoid making a priori assumptions relevant features [5].



Fig. 1: Spatial hearing. (A) Binaural disparity cues in humans. (B) Human spatial hearing acuity indicated by the green to red color scale. (C) Schematic overview of human subcortical auditory pathway. CN = cochlear nucleus; SOC = superior olivary complex; LL = lateral lemniscus; IC = inferior colliculus. (D) Azimuth locations included in the present study.

#### 1.2 Human sound location processing

Human spatial hearing acuity is highest around the interaural midline and deteriorates towards the sides and back [6][7]. Humans localize sound sources in the horizontal plane using binaural spatial cues: Interaural time and level differences (ITDs and ILDs; Fig. 1A). Processing and computing of these cues occurs in the subcortical auditory pathway, with binaural integration starting at the level of the superior olivary complex (Fig. 1 B). At the level of the inferior colliculus (IC), extraction and computation of spatial cues is mostly completed [8].

# 2 Methods

## 2.1 Data generation and pre-processing

We created a database of spatialized real-life sounds in different listening scenes. Sound clips were spatialized to 36 locations covering the entire azimuth (elevation =  $0^{\circ}$ ) at an angular resolution of  $10^{\circ}$  (starting from  $0^{\circ}$ , Fig. 1 D). At every azimuth location, we randomly sampled 500 sound clips out of a database of 6,500 mono audio clips (1 s) of real-life sounds such as speech, music, animal sounds, nature, tools, and urban environments. Sound clips were spatialized to the relevant azimuth location in two acoustic environments in order to encourage the network to learn sound location irrespective of environment-specific acoustic properties (e.g. differences in reverberation). To this end, we simulated an ane-choic environment without any reverberation, and a large lecture hall containing early and late reflections. We used a head related transfer function (HRTF) describing human binaural hearing, and a binaural room impulse response (BRIR) capturing the combination of listener and room specific acoustic properties [9]. We randomly divided the sounds at each azimuth location into two sets corresponding to the two acoustic environments (N = 250 each). Using a model of cochlear sound processing [10], we then converted each stereo sound clip into the output of the left and right cochlea. This corresponds to a spectrogram representation of the sound wave at the temporal and spectral resolution of the auditory nerve (AN) fibers.

#### 2.2 Neural network architecture

The DNN architecture developed here is a simplified representation of the first stages of the subcortical auditory pathway (Fig. 1 C). Bilateral AN representations are fed to the neural network and the first layer consists of two uncoupled branches that operate on the left and right AN representation, respectively (Fig. 2). Here, frequency invariant features in the AN representations are learned using a 2D convolutional layer (CNN) with 16 kernels (size  $1 \times 3$ ) with a rectified linear unit (ReLu) activation function. The dimensionality of the output of the CNN layer is reduced along the frequency axis using max pooling (pool size =1 x 2). Next, the two branches are merged by concatenating the feature maps of the left and right branch along the channel axis. This resembles binaural integration in the right olivary nuclei (Fig. 1 C). The concatenated feature maps are then used as input to the next 2D CNN layer. This layer learns time and frequency invariant features from the merged representation using 32 kernels (size 3 x 3) with a ReLu activation function. We reduced dimensionality further along both the time and frequency axis using max pooling (pool size  $= 2 \ge 2$ ). After flattening, the output activation of the CNN layer is fed into a fully connected (FC) layer with 2 nodes corresponding to the two outputs (x and y coordinates) and a tanh activation function.



Fig. 2: Proposed DNN architecture.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

#### 2.3 Training procedure

We trained the DNN using four different loss functions: mean square error (MSE), angular distance (AD), a combination with equal weight for MSE and AD, and a combination in which MSE was given twice the weight of AD. The mean square error (MSE) is commonly used in DNN approaches to sound localization [3][4] and minimizes the Euclidean distance between two points in 2D Cartesian x,y-coordinates:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2$$
(1)

Here,  $\hat{x}_i, \hat{y}_i$  refers to the predicted x,y-coordinates, and  $x_i, y_i$  to the actual x,y-coordinates (i.e. the label). However, as MSE is not dependent on the direction of the distance we also implemented a novel approach by training the model using AD as a loss function, or a combination of MSE and AD.

$$AD = \frac{\cos^{-1}\left(\frac{\sum_{i=1}^{n}(\hat{x}_{i},\hat{y}_{i})(x_{i},y_{i})}{\sqrt{\sum_{i=1}^{n}(\hat{x}_{i},\hat{y}_{i})^{2}}\sqrt{\sum_{i=1}^{n}(x_{i},y_{i})^{2}}}\right)}{\pi}$$
(2)

Sounds were divided into a train (80%) and test (20%) set. We trained the network using Adam optimizer (default parameters) and early stopping (10 epochs). The network was implemented with Keras with a Tensorflow backend.

#### 3 Results

#### 3.1 DNN predictions of azimuth location

Model performance was evaluated on an unseen data set of 1,800 sounds (50 per location) using MSE and AD as evaluation metrics. As the proposed DNN is – to the best of our knowledge – the first biological-inspired DNN operating on binaural auditory nerve representations, there is no clear baseline to evaluate the model against. Non-biological-inspired DNNs operating on four channel-input currently achieve direction of arrival (DOA) error scores smaller than  $10^{o}[3]$ .

Table 1 shows that in terms of average Euclidean distance, the MSE loss function performs best, the AD loss function worst, and the combinations in between. In terms of angular distance, the difference between the models is smaller. Moreover, azimuth predictions for the evaluation data set reveal that the angular distance varies per target azimuth (Fig. 3). At a number of azimuth locations, the AD loss function or the combination of AD and MSE have a comparable or lower angular distance than the model trained with the MSE loss function. Further, Figure 4 shows that prediction accuracy is not dependent on acoustic environment. This indicates that learning is robust to room-specific acoustic properties. ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

Loss: $\gamma_1 MSE + \gamma_2 AD$	Euclidean distance (MSE)	Angular distance (AD)
$\gamma_1 = 1, \gamma_2 = 0$	0.13	18.0°
$\gamma_1 = 0, \gamma_2 = 1$	0.27	$21.4^{o}$
$\gamma_1 = 1, \gamma_2 = 1$	0.18	$17.7^{o}$
$\gamma_1 = 2, \gamma_2 = 1$	0.17	$19.9^{o}$

Table 1: Evaluation metrics per loss function



Fig. 3: Azimuth location predictions for evaluation set. (A) Predicted azimuth location by the DNN trained with an MSE loss function, averaged across the evaluation data set (circles) for a given target azimuth (triangles). Colors indicate target azimuth for each predicted azimuth. (B) Same as (A) but for the DNN trained with an AD loss function. (C) Mean angular distance as a function of target azimuth. Error bars represent standard error of the mean.

# 4 Conclusion and future work

Here we explored a biological-inspired DNN-model trained with different loss functions. Our results show that this approach is successful and that the type of loss function affects acuity of location estimates. Although prediction errors are somewhat larger overall than human spatial hearing acuity, the pattern of prediction errors resembles human localization behavior [6][7], with the exception of the relatively high error at frontal locations. In future work, we will expand our current approach by exploring different DNN architectures and loss functions. Further, we will test how learning transfers to unseen acoustic environments with different reverberation characteristics. Ultimately, we aim to arrive at a DNN that can be used as a neurobiological model for investigating the transformation from binaural sound to neural representation of sound location in humans.



Fig. 4: Angular distance as a function of acoustic environment. Boxes show the results between the 25th and 75th percentile. The line inside the box reflects the median, and the lower and upper error bars the 10th and 90th percentiles. Crosses reflect data falling outside the 90th percentile.

# References

- Kiki van der Heijden, Josef P Rauschecker, Beatrice de Gelder, and Elia Formisano. Cortical mechanisms of spatial hearing. *Nature Reviews Neuroscience*, 20(10):609–623, 2019.
- [2] Wiktor Młynarski. The opponent channel population code of sound location is an efficient representation of natural binaural sounds. *PLoS computational biology*, 11(5):e1004294, 2015.
- [3] IEEE. Ieee aasp challenge on detection and classification of acoustic scenes and events, November 2019.
- [4] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):34–48, 2018.
- [5] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. nature, 521(7553):436– 444, 2015.
- [6] James C Makous and John C Middlebrooks. Two-dimensional sound localization by human listeners. The journal of the Acoustical Society of America, 87(5):2188–2200, 1990.
- [7] Simon R Oldfield and Simon PA Parker. Acuity of sound localisation: a topography of auditory space. i. normal hearing conditions. *Perception*, 13(5):581–600, 1984.
- [8] Benedikt Grothe, Michael Pecka, and David McAlpine. Mechanisms of sound localization in mammals. *Physiological reviews*, 90(3):983–1012, 2010.
- [9] Fritz Menzer, Christof Faller, and Hervé Lissek. Obtaining binaural room impulse responses from b-format impulse responses using frequency-dependent coherence matching. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(2):396–405, 2010.
- [10] Roy D Patterson, KEN Robinson, John Holdsworth, Denis McKeown, C Zhang, and Michael Allerhand. Complex sounds and auditory images. In Auditory physiology and perception, pages 429–446. Elsevier, 1992.