Neurocomputing 470 (2022) 432–442

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Goal-driven, neurobiological-inspired convolutional neural network models of human spatial hearing

Kiki van der Heijden [a,b,c,*], Siamak Mehrkanoon [d]

[a] Donders Institute for Brain Cognition and Behavior, Radboud University, The Netherlands
[b] Zuckerman Mind Brain Behavior Institute, Columbia University, New York, United States
[c] Maastricht Center for Systems Biology, Maastricht University, Maastricht, Netherlands
[d] Department of Knowledge Engineering, Maastricht University, Maastricht, Netherlands

ARTICLE INFO

Article history:
Received 8 January 2021
Revised 2 April 2021
Accepted 3 May 2021
Available online 23 July 2021

Keywords:
Convolutional neural network
Human sound localization
Binaural integration
Deep learning

ABSTRACT

The human brain effortlessly solves the complex computational task of sound localization using a mixture of spatial cues. How the brain performs this task in naturalistic listening environments (e.g. with reverberation) is not well understood. In the present paper, we build on the success of deep neural networks at solving complex and high-dimensional problems [1] to develop goal-driven, neurobiological-inspired convolutional neural network (CNN) models of human spatial hearing. After training, we visualize and quantify feature representations in intermediate layers to gain insights into the representational mechanisms underlying sound location encoding in CNNs. Our results show that neurobiological-inspired CNN models trained on real-life sounds spatialized with human binaural hearing characteristics can accurately predict sound location in the horizontal plane. CNN localization acuity across the azimuth resembles human sound localization acuity, but CNN models outperform human sound localization in the back. Training models with different objective functions - that is, minimizing either Euclidean or angular distance - modulates localization acuity in particular ways. Moreover, different implementations of binaural integration result in unique patterns of localization errors that resemble behavioral observations in humans. Finally, feature representations reveal a gradient of spatial selectivity across network layers, starting with broad spatial representations in early layers and progressing to sparse, highly selective spatial representations in deeper layers. In sum, our results show that neurobiological-inspired CNNs are a valid approach to modeling human spatial hearing. This work paves the way for future studies combining neural network models with empirical measurements of neural activity to unravel the complex computational mechanisms underlying neural sound location encoding in the human auditory pathway.

1. Introduction

Humans use spatial hearing to rapidly localize events in the environment and to separate sound sources into coherent auditory objects in multi-source listening environments (e.g. to focus on the voice of a friend from background sounds 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. That is, prior studies of neural sound location processing mainly focus on simple sounds (i.e. tones, clicks, noise bursts) in controlled listening environments (i.e. without reverberation) that have low ecological validity [2] Additionally, computational studies targeting the representational and computational mechanisms underlying the transformation from binaural sound wave to neural location representation are rare [2,3].

In the present paper, we develop goal-directed, neurobiological-inspired convolutional neural network (CNN) models of human spatial hearing. CNNs have proven very successful as computational models of neural sensory encoding [4], for example to unravel processing in visual cortex [5,6]. However, neural network models of sensory processing in the auditory system are still in its infancy (but see for example [7]). Here, we propose CNN architectures that loosely resemble the anatomy of the early stages of the subcortical human auditory pathway. Crucially, these neurobiological-inspired CNNs operate on real-life sounds as they are perceived by humans in a real-world listening environment (including reverberation) and in an anechoic listening environment (without reverberation). We originally introduced this approach in

* Corresponding author.
E-mail address: kiki.vanderheijden@donders.ru.nl (K. van der Heijden).

https://doi.org/10.1016/j.neucom.2021.05.104
0925-2312/

our previous study [8]. Here, we extend our approach by testing the effect of three different implementations of binaural integration corresponding to the various forms of neuronal encoding of binaural disparity cues that take place in the human auditory pathway [9]. Additionally, we examine the effect of different objective functions on localization acuity by training CNNs to minimize either the Euclidean or the angular distance. Finally, in order to obtain an understanding of how the networks form spatial representations, we investigate the computational and representational mechanisms emerging after training by analyzing the feature representations of intermediate layers. This provides crucial insights for future development of deep neural network models of human neural sound location encoding.

1.1. Deep neural network models of sound localization

Most prior neural network models of sound localization were developed in the context of computational environmental analysis, i.e. focusing on advanced signal processing methods to retrieve information from everyday listening scenes (for recent overviews, see [10,11]). Typically, these network architectures contain a number of convolutional layers followed by one or multiple recurrent layers. Neural networks are trained to localize sound sources using either a classification task or a multi-output regression task to estimate sound location on a continuous scale in Cartesian coordinates. Often, networks are trained to perform dual tasks such as sound localization and event detection. Using such set-ups, neural network models have become very successful at predicting sound location accurately, with average location prediction errors as small as 3°(e.g. [10,11]).

While these models provide important insights into the utility of neural networks for signal processing and acoustic scene analysis, they provide little insight into human spatial hearing. That is, input to these networks is typically derived from microphone arrays that consist of four or more channels, while humans only have two channels at their disposal. Moreover, networks typically operate on pre-processed sounds or a priori extracted features such as phase, time, and/or spectral inter-channel differences [10,11]. Neural network models of human sound location encoding – that is, with binaural, unprocessed sounds – are rare. Recently, Francl and McDermott (2020)[25] succeeded at reproducing human sound localization behavior by training deep neural network models on real-life spatialized sounds. Yet, this work did not investigate sound location processing in the human auditory pathway. Hence, the study does not explore neurobiological-inspired network architectures, or the representational mechanisms employed by the neural networks.

1.2. Human spatial hearing

Humans localize sounds in the horizontal plane utilizing binaural spatial cues: interaural level and time differences (ILD and ITD, respectively). These binaural disparity cues arise from the position of the head in between the two ears (Fig. 1 A). Monaural, spectral cues provide an additional source of information to disambiguate sound locations in the front from locations in the back [2]. The extraction and computation of these cues takes place in the subcortical auditory pathway, with binaural integration starting at the level of the superior olivary complex ([12], Fig. 1 C). That is, ILDs are processed in the lateral superior olive (LSO) based on ipsilateral excitatory and contralateral inhibitory (EI) input from the cochlear nuclei. In contrast, ITDs are computed in the medial superior olive (MSO) based on excitatory-excitatory input (EE) from the cohlear nuclei. At the level of the inferior colliculus, processing of spatial cues is mostly completed [9]. This information is then propagated via the thalamus to the auditory cortex, which is implicated in goal-oriented sound localization, spatial processing of complex sounds, and spatial hearing in complex listening scenes [2].

Human spatial hearing acuity is highest around the interaural midline and deteriorates towards the periphery and back ([13,14], Fig. 1 B). Further, humans are prone to making front-back reversals: ILDs and ITDs are identical for source locations in the front and back that are at equal angular distance from the interaural axis, making it difficult to resolve the front from the back. Humans make use of very small head movements or, in the absence thereof, monaural, spectral cues to resolve these front-back ambiguities [15]. Finally, localization acuity is affected by both low-level and high-level sound properties. For example, mammals localize broadband sounds more accurately than narrow-band sounds [16,17], and humans localize behaviorally relevant sounds more accurately than less relevant sounds [18].

2. Methods

2.1. Data generation and pre-processing

We created a database of binaural nerve representations of spatialized, real-life sounds in different acoustic environments (500 ms duration). Real-life sounds included speech, music, animal sounds, nature, tools, and urban environments. First, sound clips were spatialized to 36 locations covering the entire azimuth (elevation = 0°, distance = 1.5 m) at an angular resolution of 10° (starting from 0°, Fig. 1 D) using human binaural hearing characteristics. In total, we spatialized 2,087 monaural sound clips to 36 locations in two different acoustic environments, resulting in a total of 150,263 training sounds. We also created a separate data set of 13,608 spatialized sounds to evaluate model performance (i.e. 169 monaural sound clips spatialized to 36 locations in two different listening environments); this data was not used for training.

Sound clips were spatialized to the relevant azimuth location in two different acoustic environments in order to encourage the network to predict sound location irrespective of environment-specific acoustic properties (e.g. differences in reverberation). To

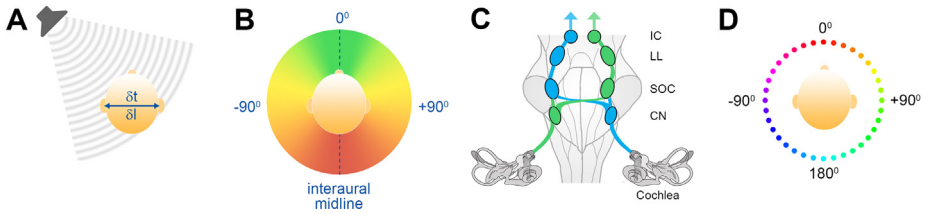

Fig. 1. Overview of human spatial hearing. (A) Humans utilize binaural disparity cues – interaural time and level differences (ITD and ILD) – to localize sounds in the horizontal plane. (B) Human localization acuity is highest around the interaural midline and deteriorates towards the periphery and especially the back, as indicated here by the green to red color scale. (C) Schematic overview of the first stages of the human subcortical auditory pathway. CN = cochlear nucleus. SOC = superior olivary complex. LL = lateral lemniscus. IC = inferior colliculus. (D) Colored circles indicate the 36 target locations used for training and testing the neural networks.

spatialize the monaural sound clips to a given azimuth location and acoustic environment, we used a head related transfer function (HRTF) describing human binaural hearing characteristics, and a 3D sound rendering technique capturing room-specific acoustic properties (e.g. reflections). That is, the HRTF describes listener-specific properties and simulates listening to a loudspeaker at a given location in an anechoic environment. In addition, we simulated listening to a sound source in a real-life acoustic environment - i.e. a lecture hall (10 x 14 m) with reflections - we modelled a binaural room impulse response (BRIR [19]). Specifically, we modelled the BRIR by convolving the HRTF with the direct sound part of the b-format recorded room impulse response and by linearly combining the later reflections part of the b-format signal channels. This results in a BRIR with the same spectral and spatial cues as would be expected for a BRIR recorded in the same circumstances (for full details, see [19]). In this way, the BRIR reflects the combination of listener and room specific acoustic properties. Thus, in total, we spatialized sounds to two acoustic environments using a HRTF and BRIR: an anechoic environment, and a large lecture hall with early and late reflections. Listening environments did not include background noise as the aim of the present paper was to model single-source sound localization. This procedure resulted in spatialized, stereo sound clips with clear and realistic binaural disparity cues as they are typically present in human hearing (Fig. 2).

Finally, we use a model of cochlear sound processing to convert each channel of the stereo sound clip into the expected activation pattern at the level of the left and right auditory nerve, respectively. Specifically, we modelled a bank of 100 gammatone filters to simulate movement of the basilar membrane in the cochlea resulting in a multi-channel spectral analysis over time [20]. Center frequencies of the filters spanned 50–8,000 Hz and were equidistantly spaced in terms of auditory filter bandwidth [21]. Here, the filter gain reflects the transfer function of the outer and middle ear (utilizing the implementation of [22,23]). In this way, we simulate the output of the cochlea by generating auditory nerve (AN) representations: Spectrogram representations of the sound wave at the spectral and temporal resolution (i.e. 1,500 Hz) of the human auditory nerve fibers (Fig. 2).

2.2. Neural network architecture

The neural network architectures evaluated here are inspired by the basic feedforward, hierarchical architecture of the first stages of the human subcortical auditory pathway (Fig. 1 C). Specifically, the input to the network comprises the simulated bilateral AN representations described in the aforementioned section which correspond to the input to the human auditory pathway. These bilateral AN representations are passed to the first layer, which consists of two uncoupled branches resembling the initial stage of the left and right auditory pathway, i.e. the cochlear nucleus (CN). Here, spectral and temporal features in the AN representations are learned in a convolution-pooling block. A block comprises a 2D convolutional layer (CNN) with 32 kernels and a rectified linear unit (ReLu) activation function [24], a drop-out layer (drop-out rate = 20%) and a max pooling layer reducing the dimensionality along the frequency axis (pool size = 1 x 2).

The next layer simulates binaural integration in the human superior olivary nucleus by merging the feature maps of the two branches. We created three different architectures corresponding to three different implementations of binaural integration: subtraction, addition, or concatenation (Fig. 3). These architectures were selected for their correspondence to human binaural integration for ILD encoding (excitatory-inhibitory integration in the LSO modeled here as subtraction [see Human sound localization]) and for ITD encoding (excitatory-excitatory integration in the MSO modeled here as addition [see Human sound localization]), and to explore the effects of an unconstrained architecture in which features from the left and right stream remain available throughout the network (concatenation).

In order to test specific effects of the binaural integration mechanisms, we model binaural integration solely in one of the bilateral streams the human auditory pathway (Fig. 1 C). tThat is, it is still unclear how the left and right auditory pathway interact during sound localization [2]. We therefore aimed to investigate the location representations and localization accuracy that emerge based on binaural integration on one side of the auditory pathway. Here, the implementation of subtraction models binaural integration in the left superior olivary complex: Subtraction is modeled as left minus right, and concatenation is modeled by concatenating the feature maps of the right stream to the feature maps of the left stream. Because the order of the feature maps is not relevant for addition and concatenation, the model does not map a specific (i.e. left or right) site for these methods of binaural integration.

The resulting merged feature maps are used as input to a series of three convolution-pooling blocks, each consisting of a 2D CNN layer with ReLu activation function, a drop-out layer (drop-out rate = 20%) and a max pooling layer (pool size = 2 x 2 for first

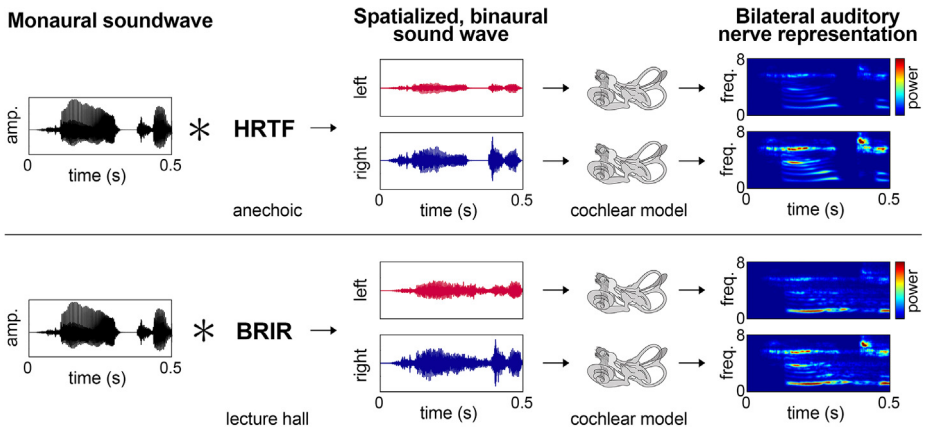

Fig. 2. Schematic illustration of sound pre-processing procedure in two listening environments. Top pane: A monaural sound clip (left) is spatialized to a position in the right hemifield (+40 degrees) in an anechoic environment by convolving the sound clip with an HRTF. This results in spatialized, binaural sound waves corresponding to the sound wave arriving at the left and right ear respectively (middle). The spatialized, binaural sound waves are converted into a simulation of bilateral auditory nerve representation using a model of cochlear processing (right, see main text). Bottom: A monaural sound clip (left) is spatialized to the same position in a reverberant listening environment (lecture hall) by convolving the sound clip with a BRIR (see main text). Middle and right panels similar to top.

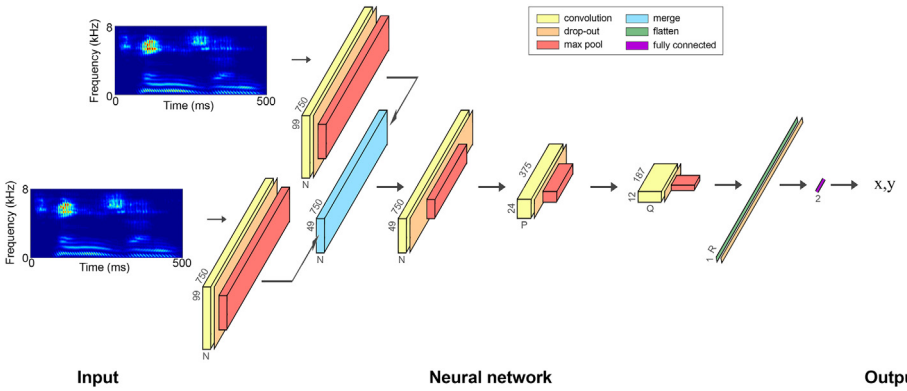

Fig. 3. Schematic of neurobiological-inspired neural network architectures. Bilateral auditory nerve representations are fed into two uncoupled convolution-pooling blocks. The outputs of this stage are merged (layer in blue) using one of three implementations of binaural integration (concatenation, subtraction, addition). The merged outputs are fed into a series of three convolution-pooling blocks, followed by a flattening layer. The final step consists of a fully connected layer with two nodes corresponding to the x and y-coordinate of the location predictions. The number and size of convolution kernels evaluated here vary, see Table 1 for an overview.

two blocks, and 3 x 2 for last block). After flattening and a final drop-out layer (drop-out rate = 20%), the output activation of the final convolution-pooling block is fed into a fully connected (FC) layer with two nodes corresponding to the two outputs (x and y coordinates) and a tanh activation function (Fig. 3). The tanh activation function ensures that location estimates are restricted to the unit circle with axes [−1,1] [10].

Note that we selected a CNN architecture for the present study for the following reasons. First, as the network learns to predict sound location of stationary sounds, a one-time output of sound location fits the task of the model. Further, sound statistics of stationary sounds are translation invariant in the temporal dimension and temporal convolution therefore plausible. Additionally, in the frequency dimension, local approximate translation invariance is plausible [25]. Finally, using a CNN architecture enables us to investigate the internal representations of the network in a straightforward manner.

Finally, to find the optimal network parameters, we tested various numbers of convolution kernels per layer and two different kernel heights. Further, as mentioned above, we varied the model of binaural integration as well as the use of two different loss functions (Table 1).

2.3. Training procedure

We trained CNN architectures to predict sound location on a unit circle around the head (axes [−1,1]), using two regressors that correspond to the x and y coordinates. Networks were trained twice, utilizing a different loss function each time: mean square error (MSE) and angular distance (AD). These loss functions are of interest here because they quantify different aspects of the localization task. First, MSE quantifies the Euclidean distance between two points in 2D Cartesian coordinates (i.e. x,y-coordinates) and is commonly used in DNN approaches to sound localization [10,11].

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

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, the MSE is independent of the direction of the error and is therefore influenced not only by the azimuth position (i.e. the angle), but also by the distance with respect to the 'listener' (i.e. the microphones). For example, the MSE will be the same for a prediction that is at the correct azimuth position but at an incorrect distance, and a prediction that has a deviant azimuth position but that is at the correct distance. Because we aim to develop CNNs estimating sound azimuth position, we therefore also trained the CNNs with an angular distance (AD) loss.

$$AD = \frac{\cos^{-1}\left(\frac{\sum_{i=1}^{n}(\hat{x}_i,\hat{y}_i)(x_i,y_i)}{\sqrt{\sum_{i=1}^{n}(\hat{x}_i,\hat{y}_i)^2}\sqrt{\sum_{i=1}^{n}(x_i,y_i)^2}}\right)}{\pi} \quad (2)$$

Similar to Eq. (1), the predicted x,y-coordinates are indicated as $\hat{x}_i, \hat{y}_i$, and the actual x,y-coordinates (i.e. the labels) are indicated as x_i, y_i.

Training sounds were divided into a train and test set (75% and 25%, respectively). We trained the networks using Adam optimizer (default parameters) and early stopping to prevent overfitting to the training data (training was stopped if performance did not improve for 10 epochs). Training occurred in mini batches. We conducted a search for optimal batch size per loss function, resulting in batch size = 64 for models trained with MSE loss and batch size = 128 for models trained with AD loss. Networks were implemented with Keras library [26] with a Tensorflow backend [27].

The proposed architectures were evaluated on a held-out dataset consisting of 13,608 spatialized sounds, using the MSE and AD as metrics. Given that our aim is to develop neural networks that accurately predict sound location in the horizontal plane (that is, the angle with respect to the listener), the most relevant metric is the AD.

Table 1
Tested network parameters. All models had the same number of layers, and convolution kernels always had the same width. We varied number of kernels and kernel height, as well as the model of binaural integration (merging).

Layer	Layer type	Nr. of kernels	Kernel height	Merging
1	Convolution	32	[5, 7]	–
2	Merge	–	–	[subtraction, addition, summation]
3	Convolution	[32,64]	5,7]	–
4	Convolution	[64,128]	[5,7]	–
5	Convolution	128	[5,7]	–
6	Flatten	–	–	–

2.4. Extracting internal feature representations

To better understand the representational mechanisms utilized by the neural networks for encoding sound location - and to compare feature encoding strategies across loss functions and implementations of binaural integration – we visualize and quantify the internal feature representations of the trained networks. That is, the intermediate layers of neural networks extract feature representations that contribute to the task at hand, in this case: sound localization. Thus, the internal feature representations are expected to differentiate between the target sound locations. To reveal these feature representations, we construct a feature selectivity vector for all nodes within a feature map of a given layer. Following the approach of Nagamine and Mesgarani [28], the feature selectivity vector of node m is computed as the average activation h_m to a target location k across all samples in the evaluation dataset, for all target locations $k \in \{1, \ldots, K\}$. In this way, the feature selectivity of a given feature map in a layer is summarized by $[K \times M \times N]$ matrix consisting of $[M \times N]$ feature selectivity vectors. Here, K corresponds to the number of target locations, M to the number of nodes in the time dimension and N to the number of nodes in the frequency dimension.

To characterize the complexity of the feature representations within each layer, we perform unsupervised hierarchical clustering based on the similarity of the feature maps. Here, we focus on spatiospectral feature representations, collapsing feature maps over the time dimension (i.e. averaging over time). To perform unsupervised hierarchical clustering, we first compute the pairwise Euclidean distance between all feature maps within a layer. Next, we perform unsupervised hierarchical clustering using the 'ward' method [29]. Using a distance cut-off criterion of 0.4 of the maximum distance, we extract the number of distinct clusters within each layer as a measure of the complexity of feature representations. That is, if feature representations reveal broad feature tuning, feature representations are not strongly differentiated from each other. This will result in a relatively small number of different clusters at the level of the distance cut-off criterion. In contrast, if feature representations show sharp feature tuning, feature representations are highly differentiable. This will result in a relatively large number of different clusters.

3. Results

3.1. Overall model performance

We evaluated the trained networks on an unseen data set of 13,608 sounds that were spatialized in the same manner as the training sounds. We compared performance across the different implementations of binaural integration (concatenation, addition and subtraction) and the type of loss function used during training (MSE loss or AD loss). Table 2 shows the average model performance for each combination of binaural integration implementation and training loss (raw data without correction for front-back reversals). For the angular distance metric, models trained with AD loss perform better than models trained with MSE loss (lowest error = 3.7° versus 4.8°). For the MSE metric, models trained with MSE loss outperform models trained with AD loss (lowest error = 0.011 verus 0.107).

Table 2
Model performance on a held-out evaluation data set. Depicted are the evaluation metrics for the best performing model within the tested parameters per combination of binaural integration and type of training loss. Best scores on evaluation metrics are in bold and marked by an asterisk.

Training loss	Binaural integration	Angular distance (AD)	Mean square error (MSE)
MSE	concatenation	4.8°	**0.011***
	addition	4.8°	0.013
	subtraction	5.3°	0.014
AD	concatenation	**3.7°***	0.139
	addition	3.9°	0.107
	subtraction	5.2°	0.160

Comparing the performance of the proposed networks to the performance of other neural network models of sound localization is not straightforward. As mentioned above, existing models are mostly developed in the context of computational environmental analysis and are often trained to localize sounds in the horizontal and vertical plane simultaneously, or to localize multiple sound sources at the same time (i.e. operating on more complex acoustic environments). Further, these models typically operate on multichannel input at a higher temporal resolution than the input to the present networks, or on a priori extracted features (see Introduction). Nevertheless, the best performing model for single-source sound localization in the horizontal plane proposed here, localizes sound sources with a precision that is comparable to or higher than the precision of such neural networks. For example, the best performing models at the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events report localization errors of 3.2° or larger [11]. Adavanne et al. (2018) [10] report error ranges as small as 3.4° for their proposed network architecture, although errors are higher for more challenging environments. Thus, although our models received different inputs (two channels versus multi-channel, lower temporal resolution, no a priori feature extraction) and were tested in different acoustic environments, the present results provide a good indication that our models adequately predict sound location in the horizontal plane and may compete with other models of sound localization.

The CNN models of Francl and McDermott (2020) [25], which aim to reproduce human sound localization behavior, are closest to the networks considered here. These models predict single-source sound location with an error range of 5–13% across azimuth positions. This is higher than the error of our best performing models, but this may be due to the inclusion of the horizontal as well as the vertical plane, and the addition of noise to the acoustic environments tested [25].

Furthermore, we observe that the selected implementation of binaural integration affects model performance. Specifically, comparing localization performance across models of binaural integration shows that concatenation and addition models perform better than subtraction models. This applies both to CNNs trained with MSE loss and CNNs trained with AD loss. Interestingly, performance of addition and concatenation models is nearly equal, especially in terms of angular distance. This shows that constrained models mimicking human spatial hearing by implementing binaural integration similar to binaural integration in the auditory pathway (i.e. the addition models) perform equal to unconstrained models that have all data at their disposal (i.e. the concatenation models).

3.2. Localization acuity as a function of target location

To obtain a more detailed understanding of model performance, we examine location prediction error as a function of target location. That is, human sound localization errors are not uniform across the azimuth (Fig. 1 B). Therefore, we expect CNNs trained on sounds spatialized with human binaural hearing characteristics to depict similar non-uniform error patterns. Fig. 4 A shows that *concatenation* and *addition* models indeed exhibit a specific pattern of prediction errors (i.e. angular distance) that resembles approximately localization errors made by humans. Specifically, localization acuity is highest for frontal locations close to the interaural midline, and decreases towards more peripheral locations (com-

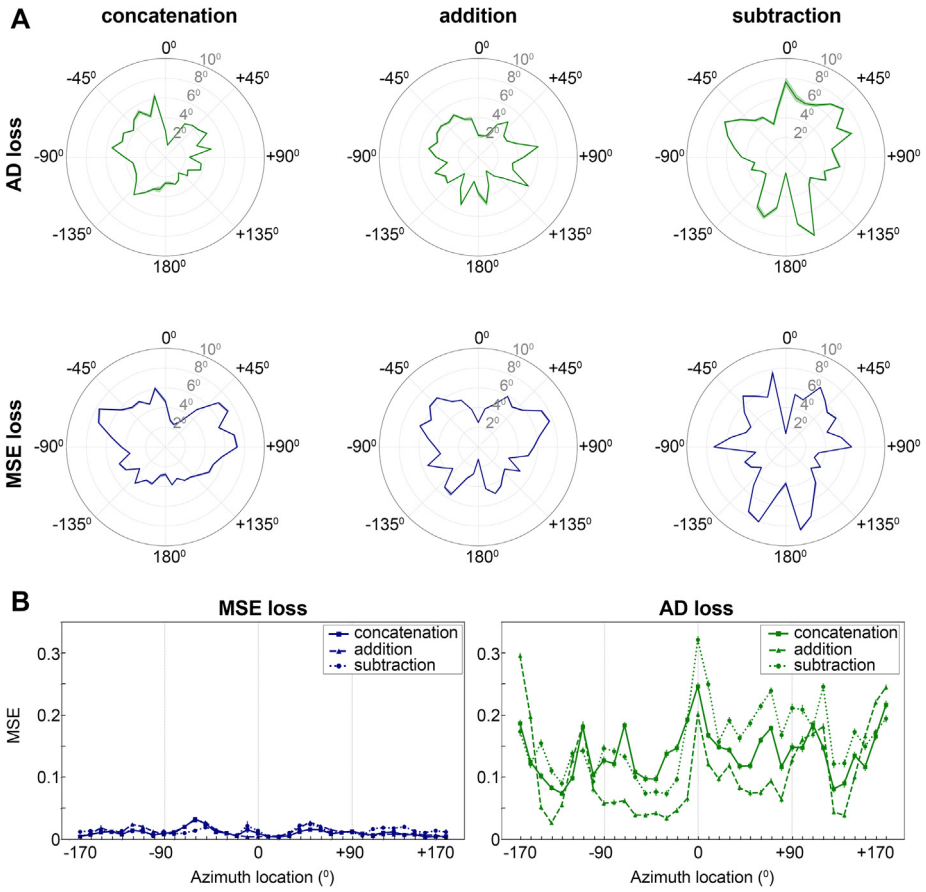

Fig. 4. Location prediction error per target location. (A) Polar plots display the average angular distance (AD) for each target location. Shaded area reflects standard deviation. Blue colors refer to CNNs trained with MSE loss, green colors to CNNs trained with AD loss. (B) Mean square error per azimuth location for CNNs trained with MSE loss (blue, left panel) and models trained with AD loss (green, right panel). Error bars reflect standard error of the mean.

pare to Fig. 1 B). This is similar to human sound localization acuity [13,14]. However, unlike humans, CNNs also predict locations in the back close to the interaural midline relatively accurately (Fig. 4 B). Further, the *subtraction* CNN trained with AD loss does not predict sound location around the interaural midline accurately, while it predicts locations around the interaural axis (−90°and +90°) more accurately than would be expected for humans.

Fig. 4 also shows that prediction errors made by the models trained with MSE loss are relatively similar for target locations on the left and on the right, while models trained with AD loss produce more asymmetric error patterns. A statistical comparison between the errors in the left and in the right hemifield confirms that neural networks trained with MSE loss are unbiased and make similar errors for locations in the left and in the right hemifield (paired samples t-test, $p > 0.05$, Fig. 5 A and Fig. 6). In contrast, several CNNs trained with AD loss indeed exhibit biased localization acuity. Specifically, the *subtraction* CNN trained with AD loss had lower MSE for locations in the left hemifield than in the right hemifield (paired samples t-test, p = 7.28E-6 corrected for multiple comparisons using the False Discovery Rate [FDR][30] at $q < 0.05$). The AD also appeared smaller for locations in the left hemifield than locations in the right hemifield for this network (Fig. 5 B), but this difference was not statistically significant ($p > 0.05$).

Strikingly, this pattern of localization errors (smallest errors in ipsilateral hemifield) is in close agreement with observations of neural location encoding in the mammalian left LSO. That is, ILD encoding in the LSO takes place through the integration of ipsilateral excitatory and contralateral inhibitory input, and most neurons represent ipsilateral sound locations (Pickles, 2015). Additionally, lesion studies of mammalian LSO report mainly ipsilateral localization errors (Celesia, 2015), indicating that the LSO contributes mainly to ipsilateral sound localization. Note that our implementation of binaural integration used here – i.e. left–right – corresponds to the excitatory-inhibitory binaural integration taking place in the left human LSO. As the output of the LSO mostly crosses to the contralateral auditory stream [12] (see also Fig. 1 C), the stages of our CNN after binaural integration conceivably reflect encoding in the right human auditory pathway. Hence, the biased localization acuity favoring locations in the left hemifield by *subtraction* CNNs trained with AD loss, indicates that our implementation of excitatory-inhibitory integration in CNN models as a simple subtraction may indeed resemble human neural location processing in the LSO.

Further, the *concatenation* CNN trained with AD loss also exhibits biased location predictions with lower AD scores in the right hemifield (paired samples t-test, p = 0.023, FDR corrected for multiple comparisons, $q < 0.05$, Fig. 5 A and Fig. 6). Thus, the direction of the localization acuity bias for the *concatenation* CNN is the opposite as for the *subtraction* CNN, favoring the right hemifield rather than the left hemifield. Interestingly, this is in agreement with expectations based on ITD encoding in the left human MSO based on excitatory-excitatory binaural integration, which is tuned mostly to contralateral locations [9,12]. Our results indicate that the *concatenation* CNNs may learn a similar excitatory-excitatory mechanism for binaural integration. As the output of the MSO is uncrossed, the stages after binaural integration for the *concatenation* CNNs conceivably reflect location processing in the left auditory pathway, predicting optimal localization in the contralat-

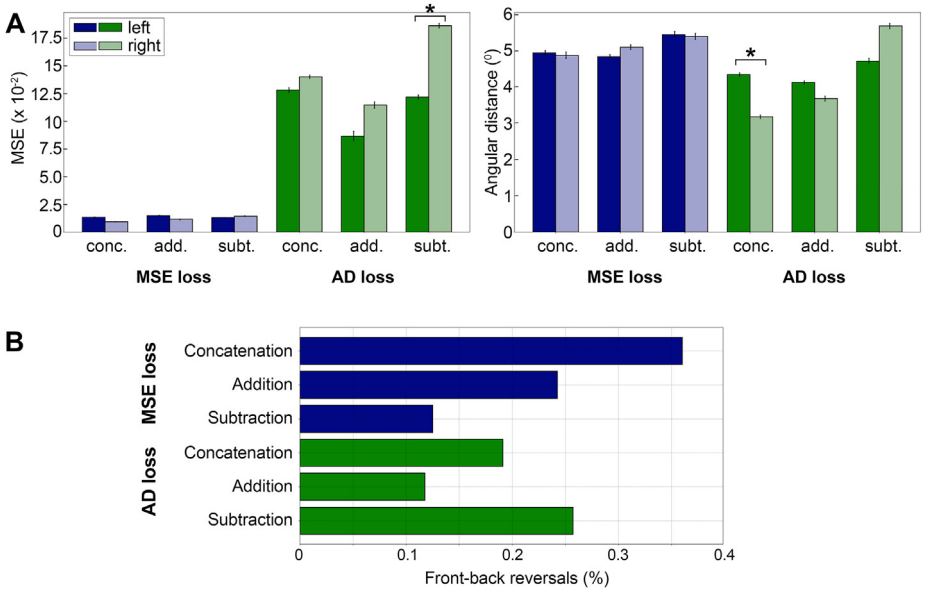

Fig. 5. Prediction errors specified per hemifield, and front-back reversals. (A) Bars depict the average error across locations in the left hemifield and locations in the right hemifield (left pane = MSE, right pane = AD). A horizontal line with asterisk indicates a significant difference between the left and right hemifield (paired samples t-test, $p < 0.05$, FDR corrected for multiple comparisons ($q < 0.05$). Error bars indicate standard error of the mean. (B) Plotted is the percentage of front-back reversals in location predictions for the evaluation data set. Each bar depicts the result of the best performing model for each type of binaural integration architecture (i.e. concatenation, addition or subtraction). Blue bars depict scores for models trained with MSE loss, green bars depict scores for models trained with AD loss.

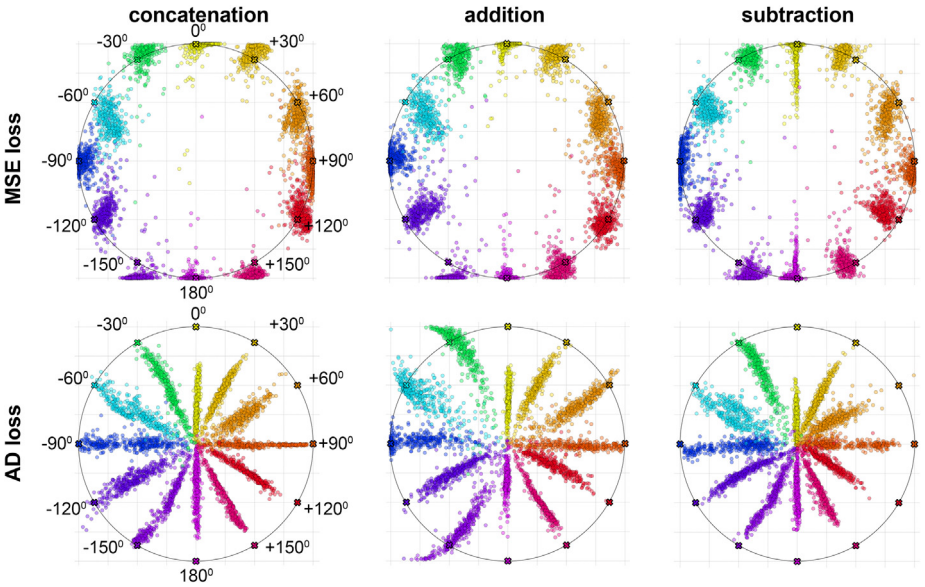

Fig. 6. Location predictions for a series of target locations per implementation of binaural integration and type of loss function. Colored circles indicate location predictions for all sounds at a given target location in the evaluation test set. Colored crosses indicate target locations. The large circle depicts the unit circle on which all target locations lie.

eral - i.e. right - hemifield. Finally, the *addition* CNNs trained with AD and MSE loss do not exhibit significant differences in error scores between the left and right hemifield ($p > 0.05$). This was expected given that the addition CNN does not model a specific auditory stream (i.e. left or right) because the order of feature maps for integration through addition is irrelevant (see Section 2.2 Neural network architecture).

Taken together, our results provide a first indication that implementing binaural integration in a neurobiological-inspired CNN as subtraction or concatenation, gives rise to location encoding mechanisms that correspond to different types of binaural integration in the human auditory pathway that underlie ITD and ILD encoding.

Remarkably, CNNs made very few front-back reversals in location predictions (less than 1% of the predicted locations for all models). In contrast, humans make front-back reversals up to 10% of localization judgements at zero elevation, especially for locations at the back of the head [14]. Other neural network architectures report front-back reversal rates of around 35–55% [25]. However, the latter study included localization judgements at varying elevations, which is expected to lead to higher front-back reversal rates [14].

Finally, visualizing the actual location predictions highlights the different location encoding mechanisms employed by the CNNs (Fig. 6). That is, training models to minimize the Euclidean distance (i.e. using MSE loss) results in location predictions that minimize not only the Euclidean distance between the target and prediction in a given direction, but also between the prediction and the unit circle. That is, these CNNs simultaneously optimize the distance in a given direction and the distance with respect to the origin (i.e. the 'listener') . As a consequence, the angular distance for some of target locations is relatively high (see for example the AD for target location +90°in Fig. 6). Training models with AD loss results in a very different pattern of location predictions. Specifically, Fig. 6 shows that – in contrast to the predictions made by the MSE loss models – a large proportion of the location predictions lies close to the origin (i.e. the listener). However, because the models trained with AD loss only aim to minimize the angular distance, the AD is relatively small at many locations. Note that these location predictions show that the relatively high MSE score at frontal locations that we observed for CNNs trained with AD loss (Fig. 5 B, right panel) is a consequence of predictions lying very close to the origin, rather than to large angular distances. Additionally, these plots also highlight the relative absence of front-back reversals.

Thus, while MSE loss is utilized most often to train deep neural network models of sound localization (for example [10,11]), our findings emphasize that the optimal loss function is dependent on the specific goal: If the distance to the target is not relevant for the task at hand, training neural networks to minimize angular distance may give better results than training models to minimize the Euclidean distance.

3.3. Representational mechanisms underlying sound location encoding in CNNs

We computed feature selectivity vectors for all nodes in the feature maps of intermediate layers of trained networks to reveal the representational mechanisms underlying sound location encoding. Fig. 7 shows representative examples of the resulting feature selectivity maps of each intermediate layer. Maps are collapsed (i.e. averaged) over the time dimension of the input, resulting in spatiospectral representations. Visual inspection of the feature maps highlights several observations. First, similar to feedforward CNN models of visual processing [1,6], the present CNN models of auditory processing exhibit a hierarchical gradient of spatial encoding. This is visible in the relatively broad feature (i.e. spatial) selectivity in early layers, and increasingly sparse representations of features (i.e. locations) in deeper layers. This resembles findings in visual deep neural network (DNN) models. For example, in DNN models trained to perform visual object recognition, feature representations in deep layers are specific and sparse, comprising entire objects [5,6]. In the auditory CNN model of sound localization developed here, feature representations in deep layers contain sparse representations of a specific target location.

A further confirmation of the existence of a gradient of increasing spatial selectivity in the auditory CNN models is provided by the results of a hierarchical clustering analysis. Specifically, Table 3 shows that for all proposed models, the number of clusters increases in deeper layers as compared to early layers. The high number of clusters in the final layer indicates that feature maps contain sparse and highly differentiable feature representations. In contrast, the low number of clusters in the first layer point at broad and relatively similar feature representations.

A second observation is that the feature representations at the first CNN layer are very similar across CNN models, irrespective of the implemented binaural integration and type of training loss (Fig. 7, Conv. Layer 1). Specifically, at the first layer, there appear to be three distinct feature representations in CNNs trained with

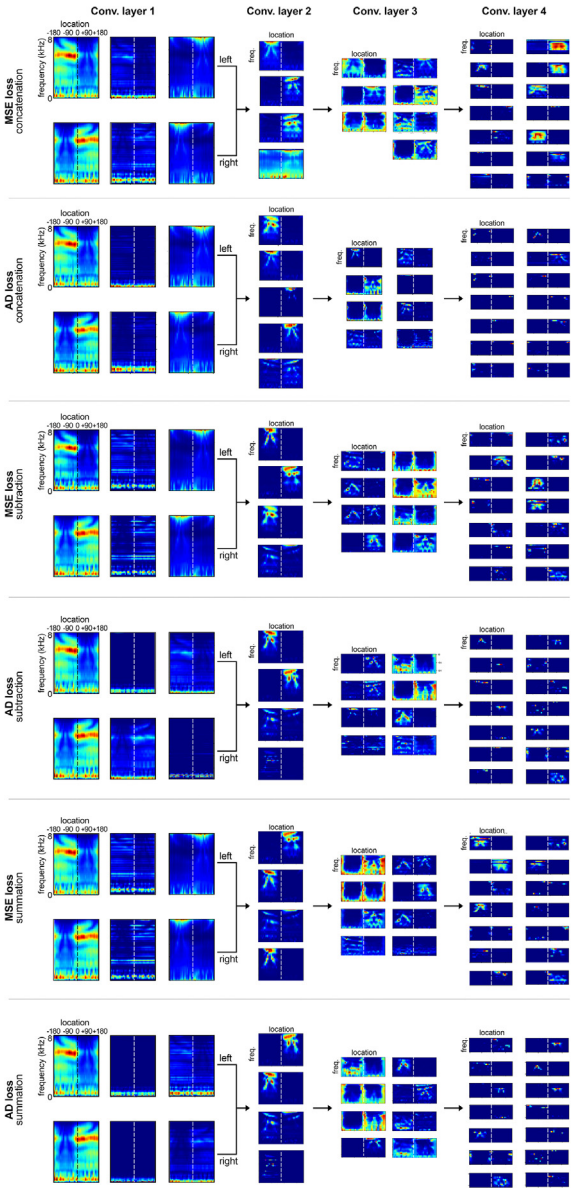

Fig. 7. Visualization of features in intermediate layers of trained CNN models. For each layer, we depict some typical feature maps. Each map belongs to an individual cluster resulting from the hierarchical clustering analysis. Note that the first layer ('Conv. layer 1') consists of two uncoupled streams 'left' and 'right'.

MSE loss and two distinct feature representations in CNNs trained with AD loss. That is, the most frequent feature representation contains high frequency nodes exhibiting a strong response to sound locations in the ipsilateral hemifield, in combination with low frequency nodes with a uniform response to all sound location. The second largest cluster of distinct feature representations contains low-frequency nodes exhibiting strong activation to all sound locations, without a pronounced activation of high frequency nodes. The third recurring feature representation concerns a group of high frequency nodes with strong activations to sound locations around in the interaural axis in the contralateral hemifield (only for CNNs trained with AD loss and concatenation CNNs; examples of these three feature representations can be found in every row in Fig. 7, panel 'Conv. layer 1').

The consistent presence of these three feature representations suggests that the first layer of the models - prior to binaural integration - learns a spatial representation that is invariant to model of binaural integration and type of training loss. Interest-

Table 3
Number of clusters in feature representations defined with unsupervised hierarchical clustering. We used a distance cut-off criterion of 0.4. Prior to clustering, feature maps without activation were removed.

Training loss	Binaural integration	Layer 2 (left)	Layer 3 (right)	Layer 9	Layer 12	Layer 15
MSE	concatenation	7	6	7	15	15
	addition	6	5	7	18	18
	subtraction	6	5	11	7	25
AD	concatenation	6	6	8	31	43
	addition	5	5	8	16	33
	subtraction	6	5	9	12	71

ingly, these representations are in close agreement with neuronal responses to sound location in the cochlear nucleus (CN), the first stage of the human auditory pathway after the auditory nerve. That is, neurons in the CN tend to exhibit strongest responses to sounds at ipsilateral locations, with little further differentiation [12].

Even in the first layer after binaural integration (Fig. 7, Conv. layer 2), feature representations appear relatively similar across model classes despite the different implementations of binaural integration (concatenation, addition, subtraction). Feature representations start to diverge between models from the third convolutional layer onwards. Note that the final layer of CNNs trained with MSE loss has some feature maps with broad spatial representations, while CNNs trained with AD loss only have highly selective spatial representations in feature maps in the final layer (Fig. 7, panels on the right). This observation is supported by the results of the unsupervised hierarchical clustering, which demonstrates that the final layer of models trained with AD loss contains sparser representations than the final layer of models trained with MSE loss (Table 3, Fig. 8).

Ultimately, the proposed CNNs and their internal feature representations may be used as encoding models in empirical studies investigating neural sound location encoding at various stages of the human auditory pathway. Here, we can already begin to make qualitative comparisons between feature representations and our current knowledge of neuronal spatial tuning. For example, the internal representations of the first layer that we described previously are in close agreement with neuronal responses to sound location in the cochlear nucleus (CN). The CN is the first stage of the human auditory pathway after the auditory nerve. Neurons in the CN tend to exhibit strongest responses to sounds at ipsilateral locations with little further differentiation [12], similar to the majority of the observed feature representations resulting after the first convolutional layer (Fig. 7). Interestingly, several of the feature representations in the second convolutional layer exhibit clusters of nodes at high frequencies that exhibit strong responses to a range of peripheral locations centered around the interaural axes (i.e. −90°or +90°). Most neurons throughout the spatial auditory pathway display similar spatial tuning, responding broadly to a range of locations – most often in the contralateral hemifield [31–34].

However, there are also discrepancies between the observed feature representations and our current understanding of neuronal spatial tuning. That is, many of the feature representations in the third convolutional layer exhibit specific activations to the interaural midline (0°and 180°, Fig. 7). While the auditory pathway also contains neurons responding specifically to these locations, these are less prevalent. Further, the systematic relationship between location and frequency tuning that appears to be present in several feature representations in the third convolutional layer (see inverted V-shapes) is not known for neuronal spatial tuning [2]. Finally, the feature representations that we observe in the final layer of the proposed CNNs – especially for CNNs trained with AD loss – are more selective for sound location than neuronal tuning at higher stages of the human auditory pathway. Specifically, even though neuronal spatial tuning sharpens during active, goal-oriented sound localization [35,36], the feature representations at the final stages appear more selective than neuronal spatial tuning.

Taken together, the feature representations emerging from the first layers of the proposed goal-oriented, neurobiological-

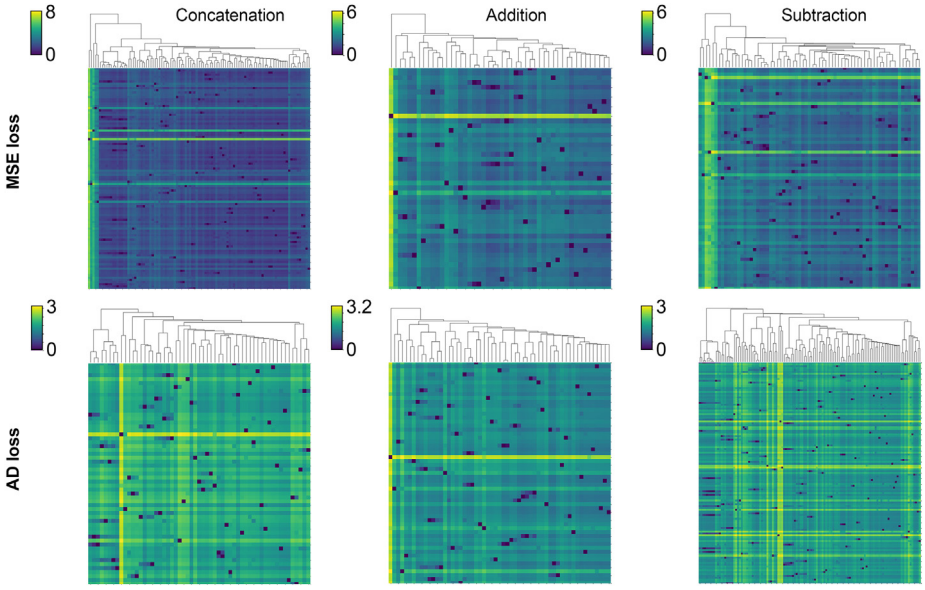

Fig. 8. Unsupervised hierarchical clustering of feature maps in the final convolutional layer. Columns represent feature maps.

inspired CNNs appear to bear some resemblance to current knowledge of neuronal spatial tuning in the human brain. At the same time, feature representations emerging in later layers are less in agreement with brain responses to sound location. Future studies may explore other network architectures such as recurrent neural networks or spiking neural networks [37,38] that may resemble brain processing of sound location more closely. Importantly, our work shows that deep neural network models generate testable hypotheses about neuronal spatial tuning whose validity can be assessed in with empirical measurements of neuronal responses to sound location, an approach that has been led to many valuable insights into neuronal sensory encoding in the visual domain (see for example [5]).

4. Conclusions and future work

In this paper we present goal-driven, neurobiological-inspired CNN models of human sound localization. We showed that such CNNs make accurate location predictions, that training on minimizing the Euclidean distance versus the angular distance results in different location encoding strategies, and that there is a hierarchical gradient of increasing spatial selectivity in the feature representations from shallow to deeper layers. The proposed models make a crucial contribution to the development of biologically as well as ecologically valid computational models of naturalistic spatial hearing, and opens up avenues for empirical work testing neural sound location encoding in the human auditory pathway. Our future work focuses on developing auditory neural network models that mimic more closely sound (location) processing in the brain, for example through the use of spiking neural networks [37,38] and biologically plausible back propagation algorithms (e.g. [39]).

5. Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 898134.

CRediT authorship contribution statement

Kiki van der Heijden: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration, Funding acquisition. **Siamak Mehrkanoon:** Methodology, Software, Writing - review & editing, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[2] K. van der Heijden, J.P. Rauschecker, B. de Gelder, E. Formisano, Cortical mechanisms of spatial hearing, Nature Reviews Neuroscience 20 (10) (2019) 609–623.
[3] W. Młynarski, The opponent channel population code of sound location is an efficient representation of natural binaural sounds, PLoS Comput Biol 11 (5) (2015) e1004294.
[4] N. Kriegeskorte, P.K. Douglas, Cognitive computational neuroscience, Nature Neuroscience 21 (9) (2018) 1148–1160.
[5] U. Güçlü, M.A. van Gerven, Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream, Journal of Neuroscience 35 (27) (2015) 10005–10014.
[6] D.L. Yamins, J.J. DiCarlo, Using goal-driven deep learning models to understand sensory cortex, Nature Neuroscience 19 (3) (2016) 356–365.
[7] M. Keshishian, H. Akbari, B. Khalighinejad, J.L. Herrero, A.D. Mehta, N. Mesgarani, Estimating and interpreting nonlinear receptive field of sensory neural responses with deep neural network models, Elife 9 (2020) e53445.
[8] K. van der Heijden, S. Mehrkanoon, Modelling human sound localization with deep neural networks, Proceedings of the European Symposium on Artificial Neural Networks (ESANN)..
[9] B. Grothe, M. Pecka, D. McAlpine, Mechanisms of sound localization in mammals, Physiological Reviews 90 (3) (2010) 983–1012.
[10] S. Adavanne, A. Politis, J. Nikunen, T. Virtanen, Sound event localization and detection of overlapping sources using convolutional recurrent neural networks, IEEE Journal of Selected Topics in Signal Processing 13 (1) (2018) 34–48.
[11] IEEE, Ieee aasp challenge on detection and classification of acoustic scenes and events (November 2019). URL:http://dcase.community/challenge2019/task-sound-event-localization-and-detection-results..
[12] J.O. Pickles, Auditory pathways: anatomy and physiology, in: Handbook of Clinical Neurology, vol. 129, Elsevier, 2015, pp. 3–25..
[13] J.C. Makous, J.C. Middlebrooks, Two-dimensional sound localization by human listeners, The Journal of the Acoustical Society of America 87 (5) (1990) 2188–2200.
[14] S.R. Oldfield, S.P. Parker, Acuity of sound localisation: a topography of auditory space. i. normal hearing conditions, Perception 13 (5) (1984) 581–600.
[15] J. Blauert, Spatial hearing: the psychophysics of human sound localization, MIT Press (1997).
[16] R.A. Butler, The bandwidth effect on monaural and binaural localization, Hearing Research 21 (1) (1986) 67–73.
[17] D.J. Tollin, J.L. Ruhland, T.C. Yin, The role of spectral composition of sounds on the localization of sound sources by cats, Journal of Neurophysiology 109 (6) (2013) 1658–1668.
[18] K. Derey, J.P. Rauschecker, E. Formisano, G. Valente, B. de Gelder, Localization of complex sounds is modulated by behavioral relevance and sound category, The Journal of the Acoustical Society of America 142 (4) (2017) 1757–1773.
[19] F. Menzer, C. Faller, H. 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) (2010) 396–405.
[20] R.D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, M. Allerhand, Complex sounds and auditory images, in: Auditory Physiology and Perception, Elsevier, 1992, pp. 429–446.
[21] B.R. Glasberg, B.C. Moore, Derivation of auditory filter shapes from notched-noise data, Hearing Research 47 (1–2) (1990) 103–138.
[22] N. Ma, P. Green, J. Barker, A. Coy, Exploiting correlogram structure for robust speech recognition with multiple speech sources, Speech Communication 49 (12) (2007) 874–891.
[23] I. Zulfiqar, M. Moerel, E. Formisano, Spectro-temporal processing in a two-stream computational model of auditory cortex, Frontiers in Computational Neuroscience 13 (2020) 95.
[24] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, in: ICML, 2010..
[25] A.F. Francl, J.H. McDermott, Deep neural network models of sound localization reveal how perception is adapted to real-world environments, bioRxiv..
[26] F. Chollet, et al., keras (2015)..
[27] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: A system for large-scale machine learning, in: 12th {*USENIX*} Symposium on Operating Systems Design and Implementation ({*OSDI*} 16), 2016, pp. 265–283..
[28] T. Nagamine, M.L. Seltzer, N. Mesgarani, Exploring how deep neural networks form phonemic categories, in: Sixteenth Annual Conference of the International Speech Communication Association, 2015.
[29] J.H. Ward Jr, Hierarchical grouping to optimize an objective function, Journal of the American Statistical Association 58 (301) (1963) 236–244.
[30] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the Royal statistical Society: Series B (Methodological) 57 (1) (1995) 289–300.
[31] K. Derey, G. Valente, B. De Gelder, E. Formisano, Opponent coding of sound location (azimuth) in planum temporale is robust to sound-level variations, Cerebral Cortex 26 (1) (2016) 450–464.
[32] M. Ortiz-Rios, F.A. Azevedo, P. Kuśmierek, D.Z. Balla, M.H. Munk, G.A. Keliris, N. K. Logothetis, J.P. Rauschecker, Widespread and opponent fmri signals represent sound location in macaque auditory cortex, Neuron 93 (4) (2017) 971–983.
[33] G.H. Recanzone, D.C. Guard, M.L. Phan, T.-I.K. Su, Correlation between the activity of single auditory cortical neurons and sound-localization behavior in the macaque monkey, Journal of Neurophysiology 83 (5) (2000) 2723–2739.
[34] B. Tian, D. Reser, A. Durham, A. Kustov, J.P. Rauschecker, Functional specialization in rhesus monkey auditory cortex, Science 292 (5515) (2001) 290–293.
[35] C.-C. Lee, J.C. Middlebrooks, Auditory cortex spatial sensitivity sharpens during task performance, Nature Neuroscience 14 (1) (2011) 108–114.

[36] K. van der Heijden, J.P. Rauschecker, E. Formisano, G. Valente, B. de Gelder, Active sound localization sharpens spatial tuning in human primary auditory cortex, Journal of Neuroscience 38 (40) (2018) 8574–8587.
[37] J.H. Lee, T. Delbruck, M. Pfeiffer, Training deep spiking neural networks using backpropagation, Frontiers in Neuroscience 10 (2016) 508.
[38] A. Tavanaei, M. Ghodrati, S.R. Kheradpisheh, T. Masquelier, A. Maida, Deep learning in spiking neural networks, Neural Networks 111 (2019) 47–63.
[39] N. Ahmad, M.A. van Gerven, L. Ambrogioni, Gait-prop: A biologically plausible learning rule derived from backpropagation of error, arXiv preprint arXiv:2006.06438..

Kiki van der Heijden received a B.A. in Cultural Sciences from Maastricht University (The Netherlands) in 2006, a M.A. in Media and Communications Management from Middlesex University (London, United Kingdom) in 2007, and a M.Sc. in Cognitive Neuroscience from Maastricht University (The Netherlands) in 2012. She conducted her Ph.D. research at Maastricht University and Georgetown University (United States) and was awarded the Ph.D. degree in 2017. After completing her Ph.D., she worked as a Post-Doctoral Research Fellow at the Cognitive Neuroscience Department at Maastricht University, and the Ear-, Nose and Throat (ENT) Department of the Maastricht University Medical Center. She is currently a Research Fellow at the Donders Institute at Radboud University (Nijmegen, Netherlands) and a Visiting Research Fellow at Columbia University (New York, United States). In her research, she uses an interdisciplinary approach combining cognitive neuroscience, computational modelling (focusing on deep neural network models) and clinical audiology to unravel the computational mechanisms underlying neural sound encoding in normal and hearing-impaired listeners. She has been awarded several research grants, including a Marie-Curie Individual Global Fellowship by the European Commission in 2019, and a NWO Veni by the Dutch Scientific Council (NWO) in 2020.

Siamak Mehrkanoon received the B.Sc. degree in pure mathematics and the M.Sc. degree in applied mathematics from the Iran University of Science and Technology, Tehran, Iran, in 2005 and 2007, respectively, and the Ph.D. degrees in numerical analysis and machine learning from Universiti Putra Malaysia, Seri Kembangan, Malaysia, and Katholieke Universiteit Leuven (KU Leuven), Leuven, Belgium, in 2011 and 2015, respectively. He was a Visiting Researcher with the Department of Automation, Tsinghua University, Beijing, China, in 2014, a Post-Doctoral Research Fellow with the University of Waterloo, Waterloo, ON, Canada, from 2015 to 2016, and a Visiting Post-Doctoral Researcher with the Cognitive Systems Laboratory, University of T?bingen, T?bingen, Germany, in 2016. He was an FWO Post-Doctoral Research Fellow with the Stadius Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven from 2016 to 2018. He is currently an Assistant Professor at the Department of Data Science and Knowledge Engineering (DKE), Maastricht University, the Netherlands. His current research interests include deep learning, neural networks, kernel-based models, numerical algorithms, optimization and computational science. He has been awarded several Grants including PDM from KU Leuven and prestigious Fund for Scientific Research from FWO Flanders.