Radboud Universiteit Nijmegen

Faculteit Sociale Wetenschappen

Azimuth Sound Localization of Binaural Neural Networks

Evaluating the biological plausibility and the exploitation of ILDs

Masterthesis Artificial Intelligence

Supervisors:
Dr. Yagmur Güçlütürk

Dr. Marc van Wanrooij

Jan-Willem Wasmann

Author:
Alex Tichter

April 2021

Contents

Abstract

Normal-hearing humans can accurately localize sounds and use this ability to focus on a specific sound source in noisy environments. This accuracy can only be achieved through the processing of acoustic cues in our brain. Recently, it has been shown that also artificial neural networks are able to mimic human sound localization performance under various listening conditions. However, it is still unknown whether the artificial networks process the same cues in a biologically plausible manner.

In addition to the commonly used broadband, high-pass and low-pass stimulus-response plots for sound localization, this master thesis analyzed the spatial, frequency and ILD tuning of artificial binaural neural networks with varying complexity. The results suggest that ILDs contribute to, but can not fully explain the localization strategy of the networks. Especially narrow frequency tuning and an insensitivity to low-pass sounds was absent in all tested models. Based on these results, we hypothesize that a more complex input, such as complex sounds, and the addition of a new learning goal (i.e. pitch detection or musical categorization) is needed to facilitate the development of biologically-plausible artificial neural networks.

1 Introduction

Over the last decades, artificial neural networks (ANNs) have been used for solving many tasks involving the mapping of a set of inputs to a set of corresponding outputs. Current ANNs are able to reach, or even exceed, human level performance in classical tasks such as object detection [1, 2], speech processing [3] or playing strategic board games such as Go [4] or chess [5]. ANNs have also been used to reveal representational gradients in biological neural networks (BNNs), when trained on similar tasks. For example, an ANN trained on music tagging was able to predict the fMRI activity of a human brain when presented with the same stimuli [6].

Recently, ANNs have also been used in the field of binaural sound localization, with the goal to understand real-world localization better [7] or to test for the presence of spatial cues under perturbed listening conditions [8]. In contrast to earlier proposed binaural sound localization projects, these two employed ANNs to predict the location of a sound source directly from cochleograms. In the past, binaural localization with neural networks often depended on extensive preprocessing to extract binaural information before using a ANN for predicting the location of the sound [9]. Other approaches modelled the whole auditory pathway mathematically [10] or analyzed the directional reverberations to derive simple inverse functions [11].

Moving from mathematically well-explained approaches [9, 10, 11] to ANN models [7, 8] automatically leads to the question of which of the various strategies these ANNs use to predict the location of a sound source when listening with two ears. Especially, when the goal of using ANNs is to better understand human localization or testing for the presence of spatial cues in hearing aids that might be beneficial for the user. In other words, does the ANN learn inverse functions such as [11], does it extract binaural cues [9] in a biologically plausible manner or is it using a completely different localization strategy?

Before answering this question, we first need to establish how normal hearing humans localize sound an why sound localization is vital in our everyday lives. Binaural hearing is essential to distinguish between background noise and a target sound source, especially in noisy environments such as a cocktail-party [12]. Depending on the location of the sound source relative to the head of the listener, interaural level differences (ILDs) and interaural time differences (ITDs) provide the brain with the necessary information to localize sounds [13]. So-called head related transfer function (HRTFs) capture the location dependent spectral reverberations that originate from the head, shoulder, body, pinna and ear-canal. These HRTFs are monaural impulse responses measured at the eardrums that vary with the location of the sound. For each location relative to the head of the listener HRTFs can be measured for the two ears. Differences in the HRTFs of the left an right ear for the same location yield the ILDs and ITDs. ILDs originate from the head shadow (Fig. 1A), where ipsilaterally presented high frequency sounds (>1.5 kHz) are attenuated on the contralateral side. ITDs describe binaural phase differences resulting from differences in travelling distance between the two ears for low frequency sounds (<1.5 kHz). ILDs and ITDs are informative of the horizontal or azimuth location of the sound. The vertical location of elevation of the sound can only be inferred from monaural spectral pinna cues that are present for high frequency sounds (>3-4 kHz). The sound that arrives at the eardrum is the result of a convolution of the sound emitted from a sound source and the pair of HRTFs of the respective location. Inferring the location of the sound is thereby an ill-posed problem for the brain, as both, the original sound spectra and the HRTF are unknown [14].

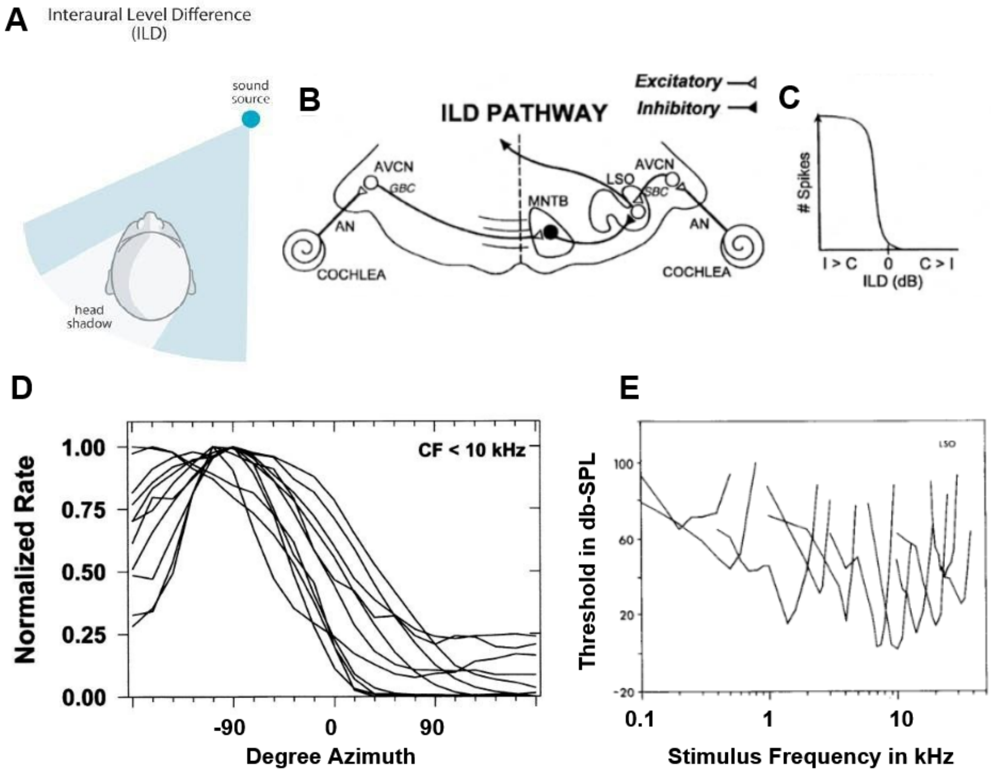


Figure 1: A: Origin of the head shadow, extracted from [8]. B: Schematic overview of the ILD pathway, extracted from [15]. C: Characteristic sigmoidal spatial tuning curve of an ILD sensitive neuron in the brain, the number of spikes are plotted against varying ILD. extracted from [15] D: Spatial tuning curves of multiple ILD-sensitive neurons in the left LSO of a cat [16]. Similar to the schematic depiction of the ILD tuning curve in C, you can see that the ILD sensitive neurons increase their activity the further a sound is coming from the left side (0° to -90°). E: Frequency tuning of multiple neurons in the LSO [17]. You can see that each neuron is sharply tuned to a specific center frequency.

This project was concerned with predicting the sound source azimuth of sounds in the frontal half-circle based on the spectral information from both ears. This meant that if the ANN processes acoustic space in a similar manner as the BNN, ILDs should be encoded by the neurons of the ANN. To elaborate, we continue with a neurological explanation of ILD processing in the brain. Looking at the schematic overview of the head shadow (Fig. 1A), an emitted sound is travelling to each of the two ears. Depending on the location relative to the head, the acoustic shadow results in the earlier mentioned ILDs. Then, per side, the cochlea extracts the frequency components of the sound and relays them via the auditory nerve to the anteroventral cochlear nucleus (AVCN) (Fig 1B). Finally, the information gets binaurally integrated at the LSO. Important to note is that while the ipsilateral neural processing chain is purely excitatory, the arriving contralateral signal contains an inhibitory synapse at the medial nucleus of the trapezoid body (MNTB). This ipsilateral excitation and contralateral inhibition is the driving mechanism behind the mammalian ILD calculation and leads to the characteristic sigmoidal neural activation of LSO neurons (Fig 1C schematic, Fig 1D neuronal measurements) [18]. Note that the LSO on the other side has the same spatial tuning curves, but mirrored at zero degree azimuth, the so-called opponent coding. Next to the hemispherical spatial tuning of LSO neurons, it is also known that LSO neurons have a sharp frequency tuning (Fig. 1E), meaning that each neuron is only sensitive to a narrow frequency band, maximally sensitive to a center frequency and rapidly decreasing its sensitivity for neighbouring frequency bands. Summarizing this means that the ILD processing of ANNs only resembles the ILD processing of BNNs if binaural amplitude differences are compared in narrow frequency bands, resulting in a sigmoidal shaped and hemispherically tuned neuron activations.

Normal-hearing humans are experts in exploiting binaural cues like the ILD to localize sounds with high accuracy [8]. However, if hearing is impaired and partially restored with a cochlear implant (CI), localization accuracy drops drastically [8]. This leads to the problem that speech perception in noisy conditions is poor for all CI recipients [19]. To date, the exact reason for the poor localization ability of CI recipients is unknown. The neural interface between the CI and the auditory nerve, the neural survival rate and insertion depth differences (in case of bilateral implantation), are considered patient-related limiting factors. But also device related issues, such as the coding strategy or the inter-implant asynchrony (for bilateral CI recipients) can distort available binaural cues.

Moreover, hearing loss is connected to a high psychological burden for both, adults and children [20]. Hearing impairment in children negatively affects language development, education and increases the risk of abuse. In adults, hearing impairment can lead to a wide range of psychological issues, including depression, social isolation, and restricted career choices. While current CIs are able to partially restore the ability to hear, it is questionable whether the lack of speech reception in noisy conditions contributes to a reduction of social issues in hearing impaired people. After all, humans are usually socializing in noisy environments such as (cocktail) parties, conferences or class rooms. Therefore, increasing our understanding of the mechanisms behind binaural hearing and further developing CIs to improve speech reception in noisy conditions is an important step to further reduce the global burden of hearing impairment.

Here, we want to establish a broader set of analysis methods for binaural ANNs to explain the localization strategy that allows ANNs to accurately localize sounds. We propose that, next to the usual stimulus-response plots that are used to assess the localization accuracy for broadband, high-pass and low-pass sounds, also the spatial, frequency and ILD tuning of the artificial neurons should be analyzed.

In the future, ANN architectures that use biologically valid processing of ILDs can be used to evaluate various CI coding strategies on the preservation of binaural cues. In contrast to current manual CI fitting procedures, which are time-consuming for both the patient and clinic, and depend on the local expertise and equipment, such a simulation could be used to make an educated guess on a set of parameters that leads to biologically valid binaural localization in the simulation. Such simulations are only possible if ANNs are able to mimic the overall sound localization behaviour of stimulus-response plots and simultaneously exploit biologically plausible ILDs in a frequency specific manner. In the long run, this is a key first step in broadening our knowledge on binaural hearing and helping people with hearing impairments to participate in an increasing number of everyday activities.

2 Methods

2.1 Acoustic Material

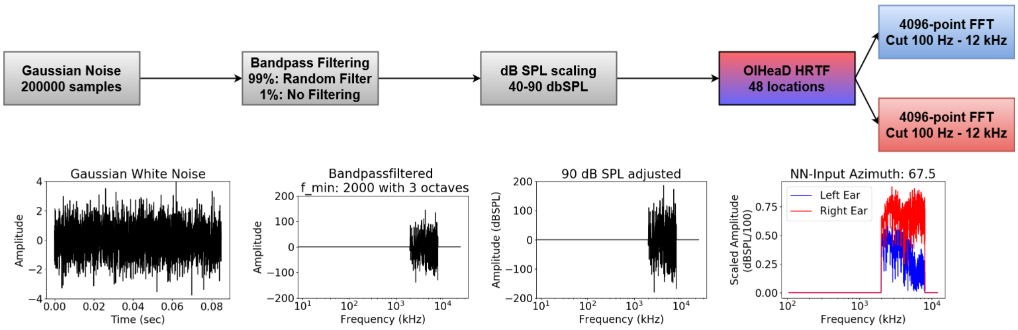


Figure 2: Overview of the acoustic material processing pipeline. Each of the 200000 randomly generated gaussian noise sound is transformed to the frequency domain, and is band-pass filtered between a randomly drawn start frequency f_{min} and an end frequency f_{max} which is at least 100 Hz above the start frequency in 99% of all cases. For 1% of the sounds, no band-pass filtering is applied. Afterwards the amplitudes are scaled to a random value between 40 and 90 dB SPL. Lastly, the dB scaled sound is artificially placed to each of the 48 locations available in the OlHeaD HRTF database. The sound after applying the HRTF of the left ear is depicted in blue, and the resulting sound after applying the right HRTF is depicted in red.

All data was synthetically generated, similar to the sounds that are used to assess the human sound localization ability. In Fig. 2 you can see an overview of the processing pipeline. First, 200000 gaussian white noise sounds (duration = 0.085 sec, sampling frequency = 48 kHz) are generated and transformed to the frequency domain via a 4096-point FFT (fast fourier transform). Each generated sound has a probability of 0.99 to be band-pass filtered with a randomly selected $f_{min} < 6kHz$ and an randomly selected f_{max} that is at least 100 Hz above f_{min}, whereas the other sounds keep their broadband spectrum. Afterwards, the intensity of the (filtered) sounds is adjusted to a random level between 40 and 90 dB SPL (1).
Afterwards, each sound is convolved with one of the 48 HRTF pairs of the left and right ear drum, corresponding to the 48 locations recorded by the OlHeaD HRTF database [21]. From these recordings, 25 were evenly spaced in the azimuth direction between -90° and 90° with 0° elevation. The azimuth angle of the other 23 HRTFs (with non-zero elevation) was transformed from the original navigational (spherical) coordinate system,

via the cartesian coordinates (3), to the double polar coordinate system (4). Similar to the experiments by Knudsen and Konishi [22], utilizing double polar coordinates is more suitable for biological sound localization [23, 24], as the azimuthal angle is measured relative to the vertical plane and is thereby independent of the elevation. The HRIR with 90° elevation was excluded, as it is not possible to define a meaningful azimuth label for a sound that is coming from directly above the head.
In Fig. 3 you can see the ILDs of the 25 azimuth directions with 0° elevation, which have been calculated by transforming each HRTF pair to the frequency domain with a 4096-point FFT and subtracting the absolute value of the left HRTF from the right. Note that the unit of the HRTFs is in dB and describes a relative change; +6 dB in the ILD curve refers to the sound pressure of the right ear being twice as loud compared to the left ear [25] 2. Hence, we transformed the dB values of the HRTF to a relative sound pressure change and multiplied them with the sound in the frequency domain.

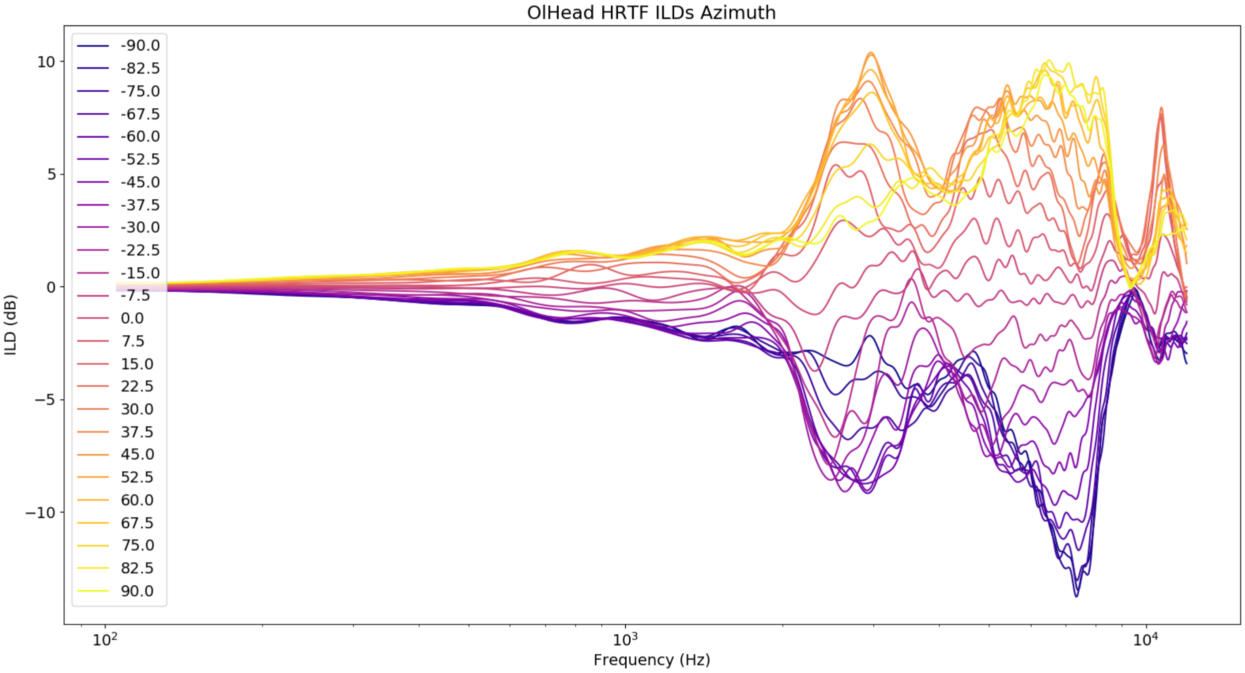


Figure 3: An overview of ILDs of the zero elevation HRTFs in the OlHeaD HRTF database. You can see how sounds on the right side (yellowish) have positive ILDs, while sounds on the left side (blueish) have negative ILDs. Around 1 kHz, the ILDs start to diverge. Around 2 and 8 kHz the biggest ILDs can be observed, but the 2 kHz divergence is only large for locations between ±67.5°. Note that around 10 kHz the ILD for all locations is almost zero.

In the end, the magnitudes of the 200000 frequency spectra are calculated, transformed to dB SPL values and normalized by dividing each value by 100. The normalization of the dB magnitude values facilitated the training of the neural network because all magnitude values are between 0 and 1. Additionally, all frequency bins outside of the 100 to 12 kHz frequency range were discarded, resulting in 1015 frequency bins per

ear, which were fed into the neural network. The double polar azimuth angles were used as targets during the training of the binaural ANN.
Fig. 4 depicts two exemplary inputs to the neural network and their respective ILD curves. You can see that the largest ILD around 2 kHz for the sound coming from 15° to the left (Fig. 4 C). For a sound presented from 90° to the right, the largest ILD (+10 dB) is observed around 8 kHz. This is in line with what we would expect from the ILD curves 3 that were used to artificially create these stimuli.

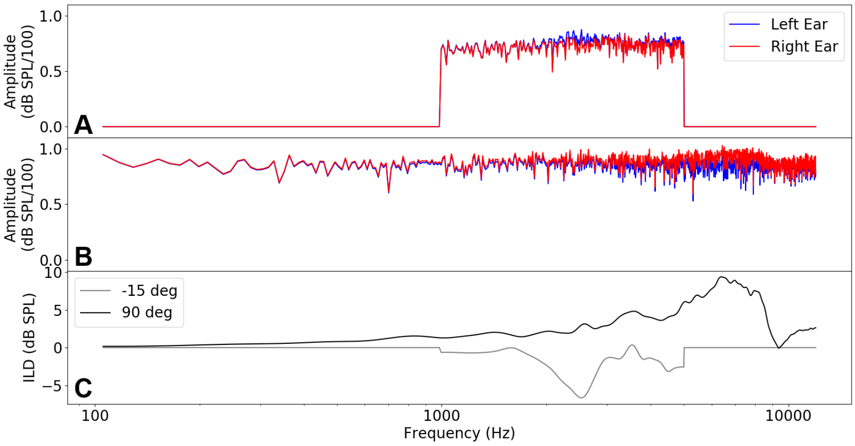


Figure 4: A: Example input with Azimuth = -15°, Gaussian Noise Filtered between 1000-5000 Hz and an intensity of 60 dB SPL. B: Example input with Azimuth = 90°, Gaussian Noise Filtered between 100-12000 Hz and an intensity of 80 dB SPL. C: ILD curves of the two sounds. The ILD curve of the stimulus depicted in A is light grey and the ILD curve of the stimulus in B is black.
In C, the head shadow effect can be seen for frequencies 4-9 kHz, where the ILD of the sound depicted in B is most pronounced. Additionally, for the sound in B we can see that the ILD up to 2 kHz is mostly constant and close to 0.

2.1.1 Formulas

Calculate the dB SPL of a sound and set it to a *targetDB*.

$$\begin{aligned} P_0 &= 20 \cdot 10^{-6} \\ pressure &= \sqrt{\frac{\sum abs(sound)^2}{len(sound)}} \\ dBSPL &= 20 * \log(\frac{pressure}{P_0}) \\ newpressure &= 10^{\frac{targetDB}{20}} \cdot P_0 \\ sound &= sound \cdot \frac{newpressure}{pressure} \end{aligned} \tag{1}$$

Calculating the sound pressure change y from a dB value and applying this to a HRTF to convolve it with a *sound* in the frequency domain.

$$
\begin{aligned}
y &= 10^{\frac{db}{20}} \\
convolved &= sound_f \cdot 10^{\frac{hrtf}{20}}
\end{aligned}
\tag{2}
$$

Converting Spherical (azimuth,elevation,r) coordinates to Cartesian (x,y,z) coordinates.

$$
\begin{aligned}
rcosele &= r. * \cos(elevation) \\
z &= r. * \sin(elevation) \\
x &= rcosele. * \cos(azimuth) \\
y &= rcosele. * \sin(azimuth)
\end{aligned}
\tag{3}
$$

Converting Cartesian (x,y,z) coordinates to double polar (x,y,z,r) coordinates. RTD: Radians To Degree conversion-

$$
\begin{aligned}
RTD &= \frac{180}{\pi} \\
Azimuth &= RTD \cdot arctan2(y, \sqrt{x^2 + z^2}) \\
Elevation &= RTD \cdot arctan2(x, \sqrt{y^2 + z^2})
\end{aligned}
\tag{4}
$$

2.2 Neural Network Architectures

Overall, we tested four architectures with increasing complexity in the number of nodes and layers. All architectures had an input layer consisting of 2030 nodes corresponding to the concatenated 1015-sample frequency arrays of the left and right ear and a single node output layer which directly predicted the sound-source azimuth. The first two architectures had a single fully connected hidden layer with 2 or 20 hidden units, whereof the former is the simplest model we could think of and the latter being the architecture inspired by Ausili [8]. The other two networks had two fully connected hidden layers with 40 and 4 hidden nodes. For one of these architectures, the Late Fusion model, we introduced a layer which only receives monaural input by splitting the first hidden layer in two parts of 20 hidden nodes whereof each only is connected to one ear. Thereby, only the second hidden layer receives binaural input. In the other architecture, the Early Fusion model, all the neurons in the first layer received binaural inputs. In Table 1 you can find an overview of the configurations of the four models. Figure 5 depicts each of the networks together with the respective formulas.

Table 1: Overview of the four network configurations. (Rectified Linear Unit (ReLU))

Name	Hidden Layer 1	Hidden Layer 2	Activation Function
Simple	2 hidden nodes	-	Sigmoid
Ausili	20 hidden nodes	-	Sigmoid
Early Fusion	40 hidden nodes	4 hidden nodes	ReLU/Sigmoid
Late Fusion	20/20 hidden nodes	4 hidden nodes	ReLU/Sigmoid

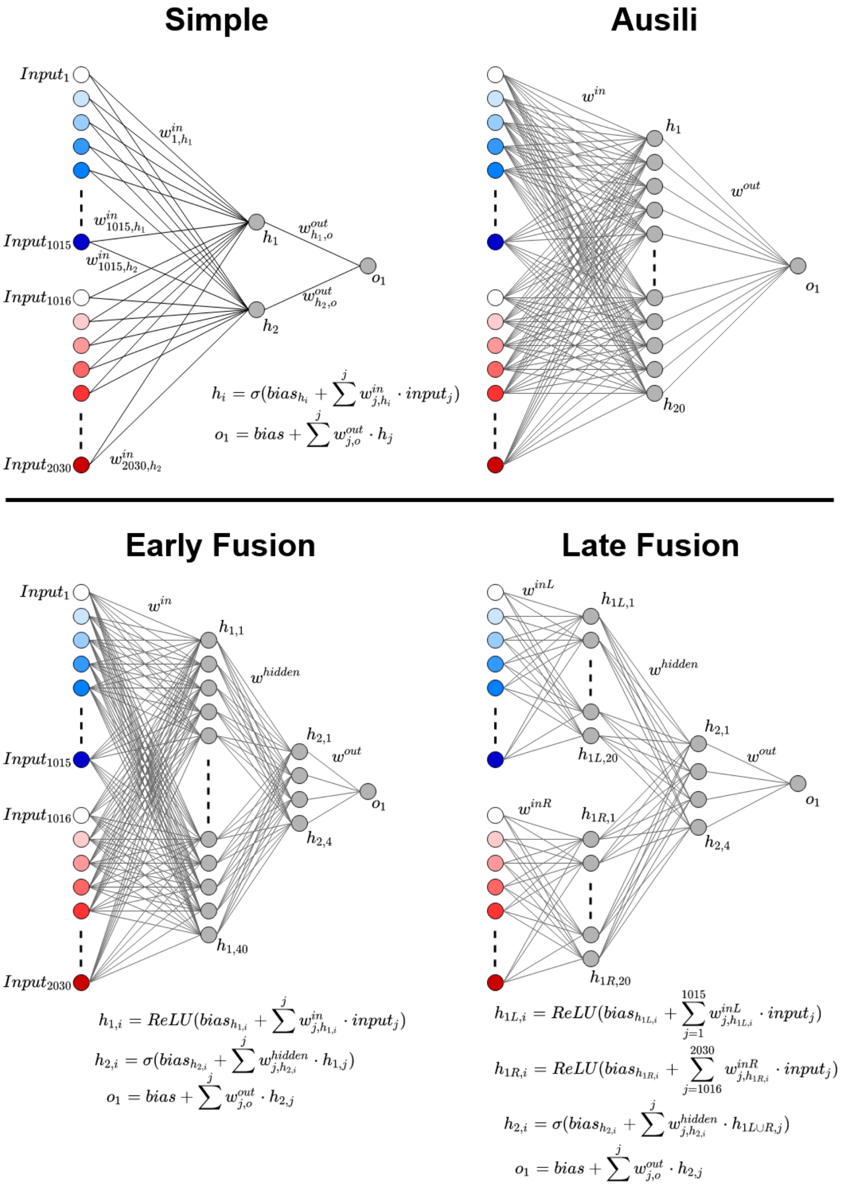

Simple
$Input_1$
w^{in}_{1,h_1}
w^{in}_{1015,h_1}
$Input_{1015}$
w^{in}_{1015,h_2}
$Input_{1016}$
h_1
$w^{out}_{h_1,o}$
o_1
$w^{out}_{h_2,o}$
h_2
$h_i = \sigma(bias_{h_i} + \sum^{j} w^{in}_{j,h_i} \cdot input_j)$
$o_1 = bias + \sum^{j} w^{out}_{j,o} \cdot h_j$
$Input_{2030}$
w^{in}_{2030,h_2}
Ausili
w^{in}
h_1
w^{out}
o_1
h_{20}
Early Fusion
$Input_1$
w^{in}
$h_{1,1}$
w^{hidden}
$Input_{1015}$
$h_{2,1}$
w^{out}
o_1
$Input_{1016}$
$h_{2,4}$
$h_{1,40}$
$Input_{2030}$
$h_{1,i} = ReLU(bias_{h_{1,i}} + \sum^{j} w^{in}_{j,h_{1,i}} \cdot input_j)$
$h_{2,i} = \sigma(bias_{h_{2,i}} + \sum^{j} w^{hidden}_{j,h_{2,i}} \cdot h_{1,j})$
$o_1 = bias + \sum^{j} w^{out}_{j,o} \cdot h_{2,j}$
Late Fusion
w^{inL}
$h_{1L,1}$
w^{hidden}
$h_{2,1}$
w^{out}
$h_{1L,20}$
o_1
w^{inR}
$h_{1R,1}$
$h_{2,4}$
$h_{1R,20}$
$h_{1L,i} = ReLU(bias_{h_{1L,i}} + \sum_{j=1}^{1015} w^{inL}_{j,h_{1L,i}} \cdot input_j)$
$h_{1R,i} = ReLU(bias_{h_{1R,i}} + \sum_{j=1016}^{2030} w^{inR}_{j,h_{1R,i}} \cdot input_j)$
$h_{2,i} = \sigma(bias_{h_{2,i}} + \sum^{j} w^{hidden}_{j,h_{2,i}} \cdot h_{1L\cup R,j})$
$o_1 = bias + \sum^{j} w^{out}_{j,o} \cdot h_{2,j}$

Figure 5: For each of the network architectures you can see how the left ear (blue) and right ear (red) frequency bins are mapped onto the respective hidden layers. The Simple and Ausili architectures only involve a single hidden layer with 2 or 20 hidden units. The Early and Late Fusion architectures have an additional hidden layer. The Early Fusion model has an additional binaural hidden layer compared to the other networks, whereas the Late Fusion model has two hidden layers that only receives monaural input.

2.3 Neural Network Training

All networks have been developed in PyTorch [26] and have been trained for 400 epochs with a batch size of 100, using the Adam [27] optimizer with a learning rate of $1e-4$. In addition to the earlier described training set, we generated a validation set with 10000 sounds according to the same procedure. The loss was defined by the mean squared error (MSE) of the predicted azimuth angle of the network and the actual horizontal orientation of the sample. After each epoch, we calculated the error of the validation set and only kept the network with the lowest validation error for further analysis.

2.4 Analysis

We have employed four different analysis methods to analyse the neural networks. First of all, the overall sound localization performance was evaluated to grasp whether the neural networks are able to perform sound localization in the azimuth direction. Secondly, we analyzed the spatial tuning of the neurons in the hidden layer(s) by plotting the activation of the neurons against the location of the sound. Thirdly, we plotted the weight matrices of the hidden units to analyze the frequency tuning. Lastly, we analyzed the ILD tuning of the hidden neurons by plotting the activation of the neurons against sounds with different ILDs.

2.4.1 Sound Localization

The overall sound localization performance of the model is analyzed with the same techniques as sound localization performance is assessed in humans [8]. We presented the network with broadband (BB) 100-12000 Hz, high-pass (HP) 3000-12000 Hz and low-pass (LP) 100-1500 Hz band-pass filtered noise. To that end, we generated new noise samples, and applied the same processing pipeline as for the training data, but used the aforementioned band-pass filters and the five dB SPL levels (45,60,75,90,100 dB SPL). We plotted the localization results together with the ideal localization diagonal. To quantify the sound localization results, we performed linear regression between the true angle and the predicted angle. Then, we extracted the slope of the fit (dimensionless) and the MSE between the estimated angle by the polynomial fit and the predicted angle (in deg). A slope of 1 and a MSE of 0° indicates perfect sound localization; while a slope of near 0 and a large MSE indicate inaccurate and imprecise sound localization, respectively. For comparison, normal-hearing humans have a slope near 1 and a MSE of 5° [8].

2.4.2 Spatial Tuning

Next to the sound localization of the whole network, we can also assess the spatial tuning of each hidden neuron. When we present the network with a sound in one of the three band-pass filters, we store the activity of each hidden unit and plot it against the location of the presented sound. This allows us to plot the spatial tuning of each hidden neuron to the three bands.
If the neural network is inferring the location of a sound in a biologically plausible manner, we expect to find sigmoidal shaped spatial tuning curves (Fig. 1D) for BB and HP sounds. LP filtered sounds however, should have a flat curve, as there is no ILD information for low frequencies.

2.4.3 Frequency Tuning

LSO recordings revealed that neurons are sharply tuned to frequency [16]. In order to assess the frequency tuning in the ANN, we plot the connection strength of the weight vector against the frequency each weight is referring to. For the neurons in the first hidden layer, the frequency tuning of hidden neuron X h_x for the left and right ear is acquired by plotting w^{in}_{1-1015,h_x} and $w^{in}_{1016-2030,h_x}$ (Fig. 5). For the second hidden layer, we calculate the sum of the frequency tuning of the hidden units and weigh each contribution according to the weight vector of w^{hidden}. For a hidden neuron $h_{2,1}$ the frequency tuning is thereby defined by

$$\sum_{i=1}^{40} w_i^{hidden} \cdot w_{1-1015,h_i}^{in}$$

for the left ear and

$$\sum_{i=1}^{40} w_i^{hidden} \cdot w_{1016-2030,h_i}^{in}$$

for the right ear.

This approach reveals which frequencies in the input affect each hidden neuron. If the hidden neurons infer the location of a sound in a biologically plausible manner, then the neurons should be sharply tuned to a set of neighbouring frequencies (Fig. 1E). Additionally, we should be able to see that a neuron having positive connections to certain frequencies in the one ear, should also have negative connections to the same frequencies in the other ear. If we can find this, we can conclude that the neuron is sensitive two binaural frequency specific level differences. If however, the neuron is only sensitive to frequencies of one ear or is showing other combinations , we must conclude that this neuron is exploiting monaural cues to infer the location of a sound.

2.4.4 ILD Tuning

Neurons processing ILDs are excited by ipsilateral presented sounds and inhibited by contralateral presented sounds. The louder the sound is on the ipsilateral side compared to the contralateral side, the higher the response of the neuron (Fig. 1C). To test for ILD tuning, we artificially create base broadband noise sounds with 40,60,80 dB SPL. We then present the network with the base sound at the one ear and with sounds that are up to 10 dB SPL (in steps of 1 dB SPL) louder at the other ear and store the normalized activity for each neuron. In the end, we can plot the ILD tuning curve for the three different base levels for an ILD of -10 to 10 dB SPL.
If we can find sigmoidal shaped ILD tuning curves (Fig. 1C) we can assume that the ANN bases its predictions on the binaural level differences.

3 Results

In the following sections we will present the results of the four analysis methods. We will start with the assessing the overall sound localization of the four models in the typical stimulus-response plots. Afterwards, we will discuss the spatial, frequency and ILD tuning per model in order to relate per neuron the outcomes of these three analyses. For the simple model, we can show all tuning curves on one page. For the other models, we made a selection of up to three neurons that were indicative of all the neurons. Please refer to the GitHub repository if you would like to see all the tuning curves.

In Appendix A you can see the train and validation loss of each model. For all four models, no signs of overfitting have been observed and the validation error always followed the training error closely.

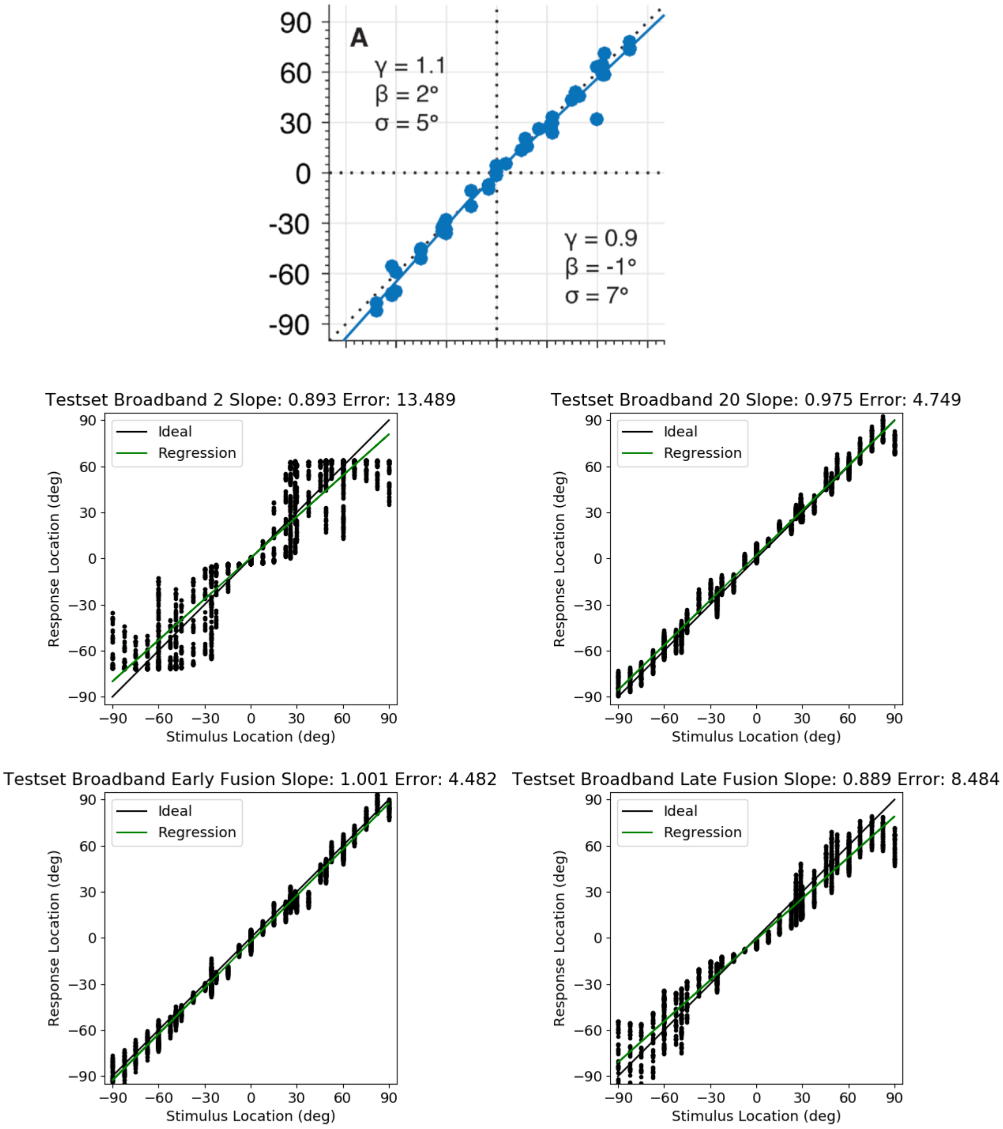
3.1 Overall Performance

Classically, the sound localization performance is assessed in a stimulus-response plot. On the x-axis of every plot you can find the location of the presented stimulus. The predicted location of every model can be found on the y-axis. Additionally, you can see the ideal localization (black) and the linear regression results (green). The average human can localize sounds with an average error of 5° and an accuracy (slope) of 1 (Fig. 6A). Looking at the four models, we can see that the only the Ausili architecture with 20 hidden nodes (error 4.75°, slope 0.98) and the Early Fusion model (error 4.48°, slope 1.00) are able to mimic the human sound localization ability well. The Late Fusion model (error 8.48°, slope 0.89) is also able to localize sounds fairly well, but struggles to localize sounds at the far left or right accurately, leading to an increase in the error while the slope decreases. The simple model (error 13.49°, slope 0.89) has the worst performance. Despite all sounds being localized at the correct hemisphere, there are no predictions $> 60°$ or $< -70°$.

In table 2 the results of the high-pass and low-pass localization are depicted. Overall, the high-pass results are similar to the broadband results for all models with only slight shifts in the localization of the Simple and Late Fusion models (Appendix B). Looking at the low-pass localization results, we can see that all models perform far worse compared to the human baseline. Especially the simple model is performing quite poorly with a slope of 0.39 and an error of 22.16. However, we need to take into account that humans exploit ITDs to localize low-frequency sounds. ITDs are not available to the model because we remove the phase information by taking the absolute value of the complex HRTF to generate the sounds.

Table 2: Overview of the localization results for the three bands. The human baselines are extracted from [8].

Name	Broadband	High-pass	Low-pass
Human	1 / 5	1 / 5	1 / 6
Simple	0.89 / 13.49	0.85 / 12.97	0.39 / 22.16
Ausili	0.98 / 4.75	0.95 / 6.71	0.82 / 12.88
Early Fusion	1.00 / 4.48	0.97 / 4.19	0.83 / 10.97
Late Fusion	0.89 / 8.48	0.86 / 6.62	0.65 / 17.68



Testset Broadband 2 Slope: 0.893 Error: 13.489

Testset Broadband 20 Slope: 0.975 Error: 4.749

Testset Broadband Early Fusion Slope: 1.001 Error: 4.482

Testset Broadband Late Fusion Slope: 0.889 Error: 8.484

Figure 6: Broadband localization plots the four architectures (Top left to bottom right: Simple, Ausili, Early Fusion, Late Fusion) and the human reference from [8] (Top)

3.2 Tuning of the Simple model

The simple model performed worst in the conventional localization task. In this section, we will investigate the different tuning curves of the two hidden neurons and relate them to the neural tuning properties of biological neural networks.
Looking at the spatial tuning (Fig. 7 Top), we can see that the two hidden neurons code for the two hemispheres, but only show an increase in activation ±50° and a slight decrease for more laterally presented sounds. This is the opponent coding we expected from the spatial tuning curves of animals (Fig. 1D). The opponent hemispheric coding of the two neurons can also be observed in the frequency tuning. The first observation here is that Neuron 0, which is coding for the left hemisphere, has mostly positive weights to all frequencies from the left ear, while the frequencies arriving at the right ear are mostly negatively weighted. The opposite is true for the other neuron, coding for the right hemisphere. Additionally, both neurons are responsive to almost all frequencies, which is contrary to our expectation of sharply tuned neurons (Fig. 1E).
Moreover, looking at the four bottom plots of Fig. 7, we can see that the ILD tuning and spatial tuning of each neuron are closely related. For high-pass and broadband sounds, the activity of the respective neuron is rapidly saturating. The same can be observed in the ILD tuning curves, where already a slight increase or decrease leads to almost full saturation. Still, the spatial tuning curves for low-pass filtered sounds shows us that the two neurons are also tuned to frequencies below 1.5 kHz.

From the spatial tuning curves, it becomes apparent that the simple neuron model is not able to localize far laterally presented sounds because the activity of the neurons quickly saturates. Despite the fact that the ILD tuning curves are closely related to the BB and HP spatial tuning, we can see large contributions of low frequencies to the frequency tuning and that the neurons are spatially responsive to low frequencies.
From these results we can conclude that the simple model is not a good model for sound localization. It seems that two neurons are not enough to cover the whole frontal semicircle.

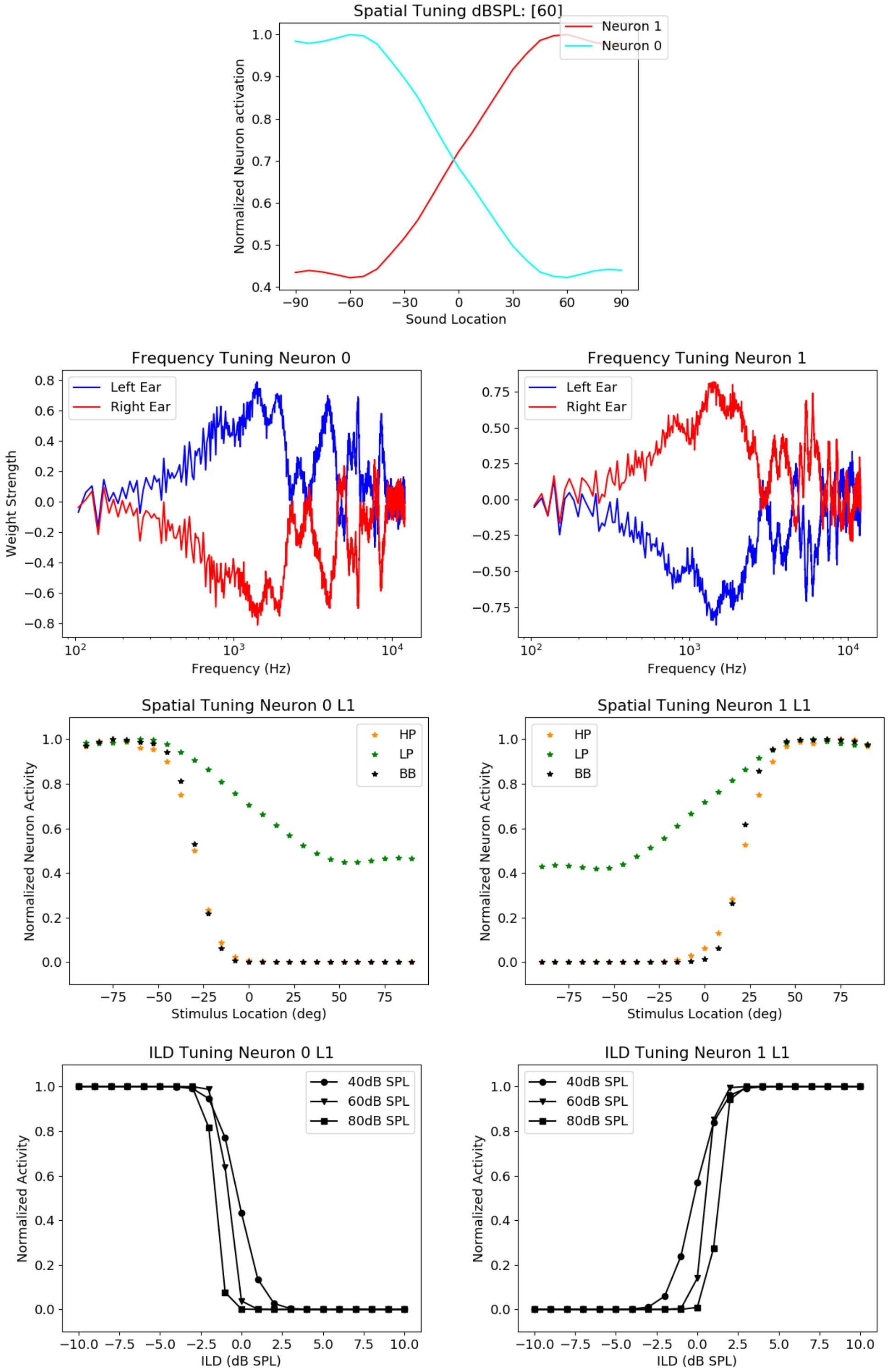


Figure 7: The different tuning curves of the simple model. From top to bottom: Spatial Tuning of the hidden neurons, Frequency Tuning of the hidden neurons, Spatial Tuning of the hidden neurons per frequency band, ILD Tuning of the hidden neurons for three levels.

3.3 Tuning of the Ausili model

The Ausili architecture had the second best score in the conventional measures and outperformed the simple model by a large margin. Looking at the overall spatial tuning in Fig. 8 Top, we can see that multiple neurons cover each hemisphere, with slightly different spatial tuning. Overall the slightly varying sigmoidal spatial tuning curves seem to be in line with what we would expect from animal measurements. Similar to the two hidden neuron model, we can find neurons 13 and 15 which are hemispherical tuned and have inverted frequency tuning. Still, the frequency tuning is broad, covering all frequencies. Looking at the ILD tuning we can see that the activity doesn't saturate as quickly, allowing for a smoother spatial tuning. Nevertheless, these two neurons are still spatially tuned to low frequencies.
In addition to the hemispherical tuned neurons 13 and 15, a new type of neuron (neuron 2) emerged that specifically is interested in the far left spatial area. This neuron is also broadly tuned to frequency but has a single large spike between 9 and 10 kHz. With respect to the ILD tuning, we can see that this neuron is only sensitive to quiet sounds around 40 dB SPL.

In conclusion, we can see that the neural network uses the additional number of hidden units to further specialize for different locations. While the simple architecture struggled to localize far laterally presented sounds, this architecture employed neurons that specifically code for the far lateral locations (Neuron 2). With regard to the sharp frequency tuning, there was no improvement found. Despite Neuron 2 having a distinct spike in the high frequencies, all frequencies seem to contribute to the overall spatial tuning of the neuron. Additionally, the spatial tuning cannot be explained by the ILD-tuning in many cases. Thereby, also this model cannot be seen to be biologically plausible despite the human like localization plots.

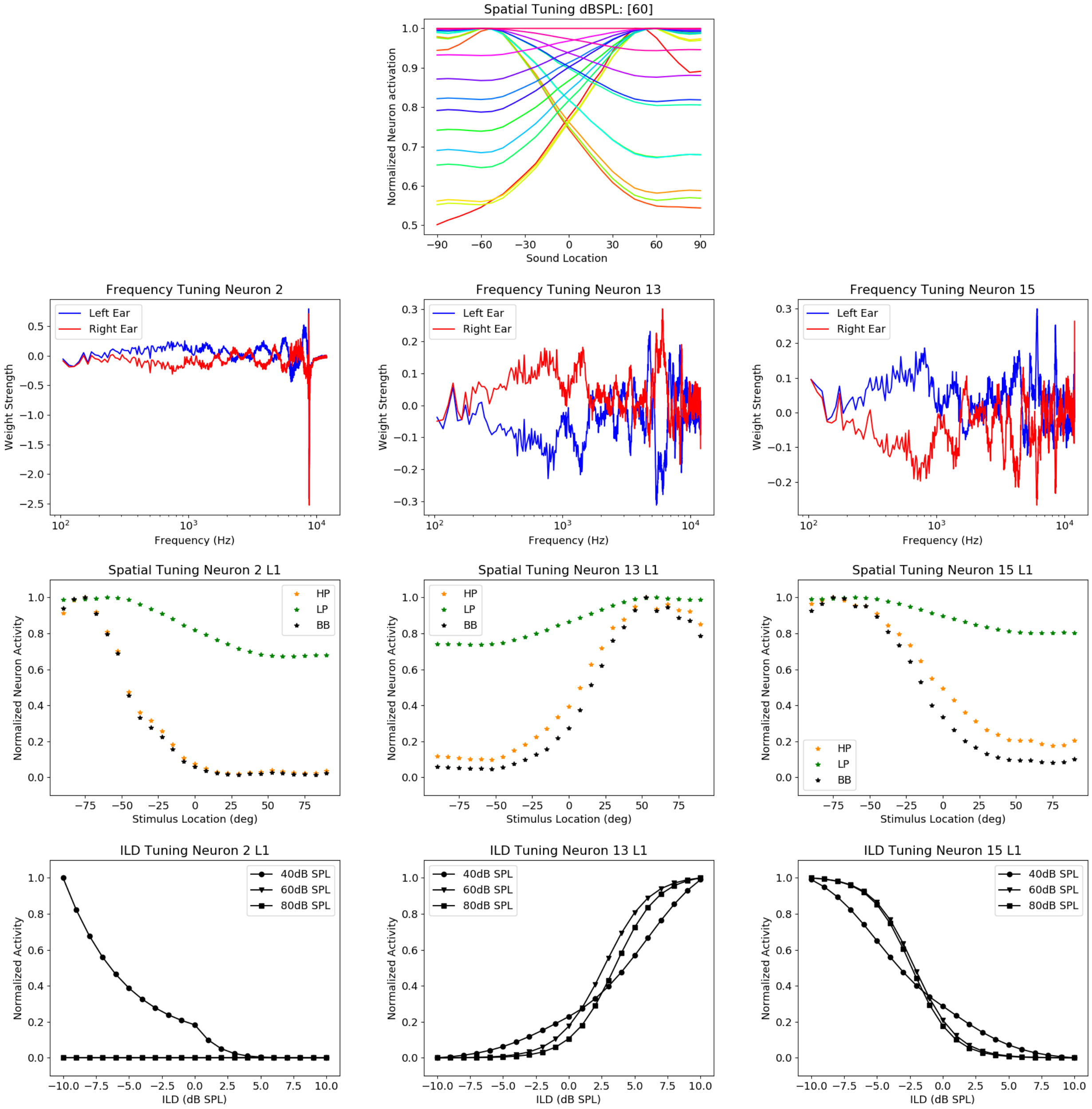


Figure 8: The different tuning curves of the Ausili model. Top to bottom row: Spatial Tuning of the hidden neurons, Frequency Tuning of the hidden neurons, Spatial Tuning of the hidden neurons per frequency band, ILD Tuning of the hidden neurons for three levels.

3.4 Tuning of the Early Fusion model

The results of the Early Fusion model are broken up in two parts referring to the two hidden layers. We will first discuss the tuning of the first hidden layer (Fig. 9), followed by the analysis of the second hidden layer (Fig. 10).
In the spatial tuning of the first hidden layer (Fig. 9) we can find two different groups of neurons. On the one hand, we have neurons that show a sigmoidal tuning to for either hemisphere with no activation on the other hemisphere (blue-green-yellow). On the other hand, we have neurons that have maximal activation on the one hemisphere, while the activation only slightly decreases for the other hemisphere (purple and red). While the former group follows our sigmoidal expectation of the spatial tuning, the latter shows a different type of spatial tuning.
Looking at the frequency tuning, we can find neurons (Neuron 13 and 22) that only have weak connections in the low frequencies. Still most neurons, like neuron 14, show large weights in the low frequency bands. Moreover, all neurons are broadly tuned to all frequencies.
If we take a closer look on the spatial tuning of the above mentioned neurons we can make three observations. First, broadly tuned neurons in frequency, like neuron 14, are show a similar spatial tuning in all frequency bands. Neuron 22, which shows little frequency tuning to low frequencies, is also spatially not tuned to low-pass sounds, while showing a similar tuning for high-pass and broadband sounds. However, neuron 13, which also shows little frequency tuning to low frequency sounds, is spatially tuned to the right hemisphere for low frequency sounds but tuned to the left hemisphere for high and broadband frequencies. This might be explainable by the slight inversion of frequency tuning between 1 and 2 kHz, as the frequencies of the right ear are positively weighted compared to the frequencies of the left ear, which is the other way round for all other frequency bands.
The ILD tuning in the first hidden layer shows hemispheric linear, rather than sigmoidal, tuning for all neurons.

In the second hidden layer, three of the four neurons are spatially tuned to the right hemisphere while only one neuron is weakly tuned to the left hemisphere. Looking at the frequency tuning, we can find that the two opposing neurons (Neuron 2 and Neuron 3) also show an inverted frequency tuning with each other, similar to the 2 hidden neuron model. Again, these neurons are well tuned to low frequency sounds, but show broad tuning to all frequencies. For neuron 3 this leads to the expected sigmoidal shaped spatial tuning for all frequency bands. But the spatial tuning of neuron 2 reveals that it is only weakly tuned to low-pass sounds. In other words, despite the broad frequency tuning of neuron 2, only low frequency sounds affect the spatial prediction of the neuron. The ILD tuning of the neurons in the second hidden layer show the expected sigmoidal shapes for the respective hemisphere.

In conclusion we can see that especially the frequency tuning is contrary to our expectations. But also the spatial tuning shows a group of neurons that are tuned to low-pass sounds and continue to have large activations for all spatial locations. This means that despite mimicking the human localization well, the Early Fusion model does not reach these results in a biological plausible manner.

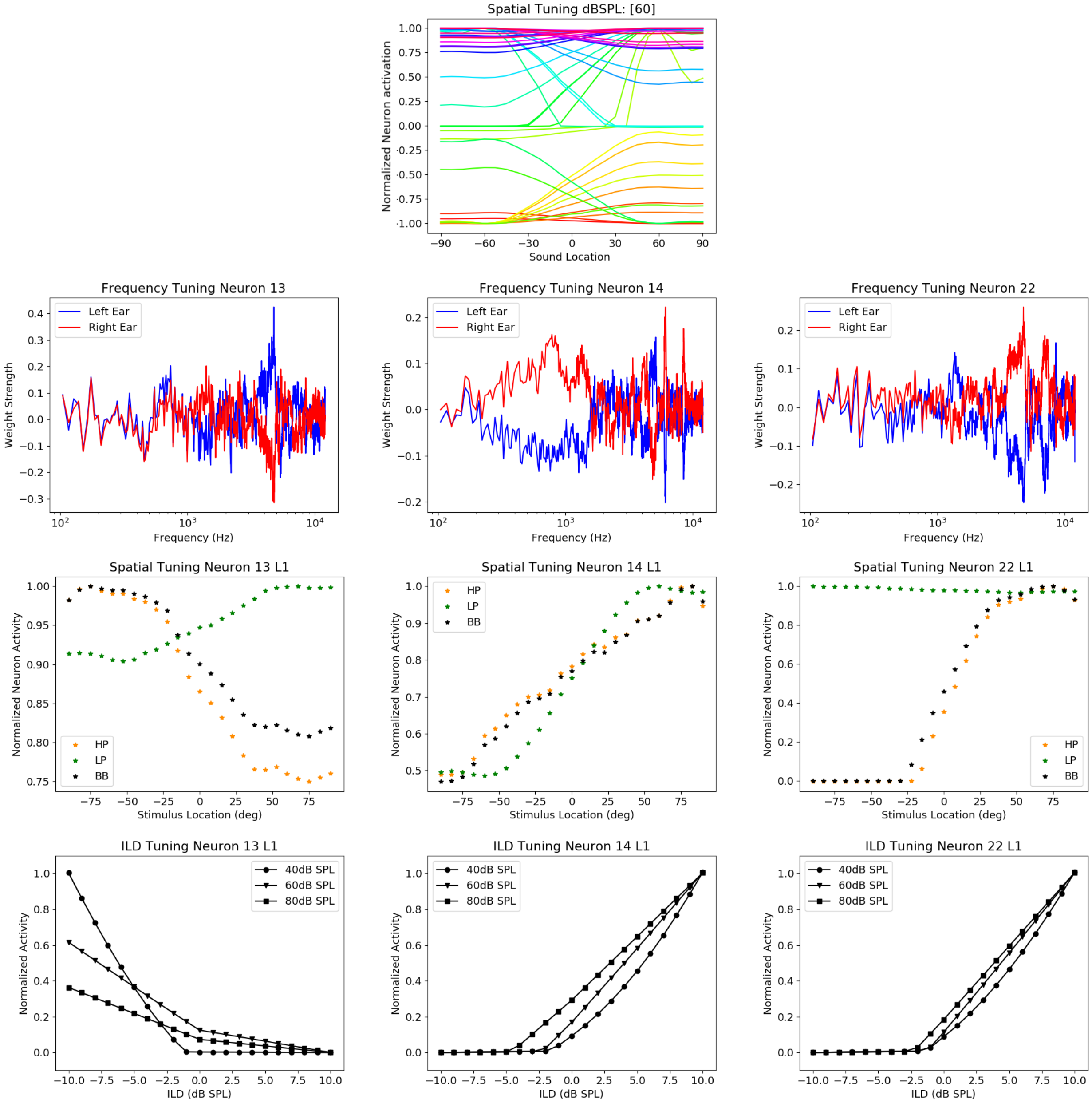


Figure 9: The different tuning curves of the first hidden layer of the Early Fusion model. Top to bottom row: Spatial Tuning of the hidden neurons, Frequency Tuning of the hidden neurons, Spatial Tuning of the hidden neurons per frequency band, ILD Tuning of the hidden neurons for three levels.

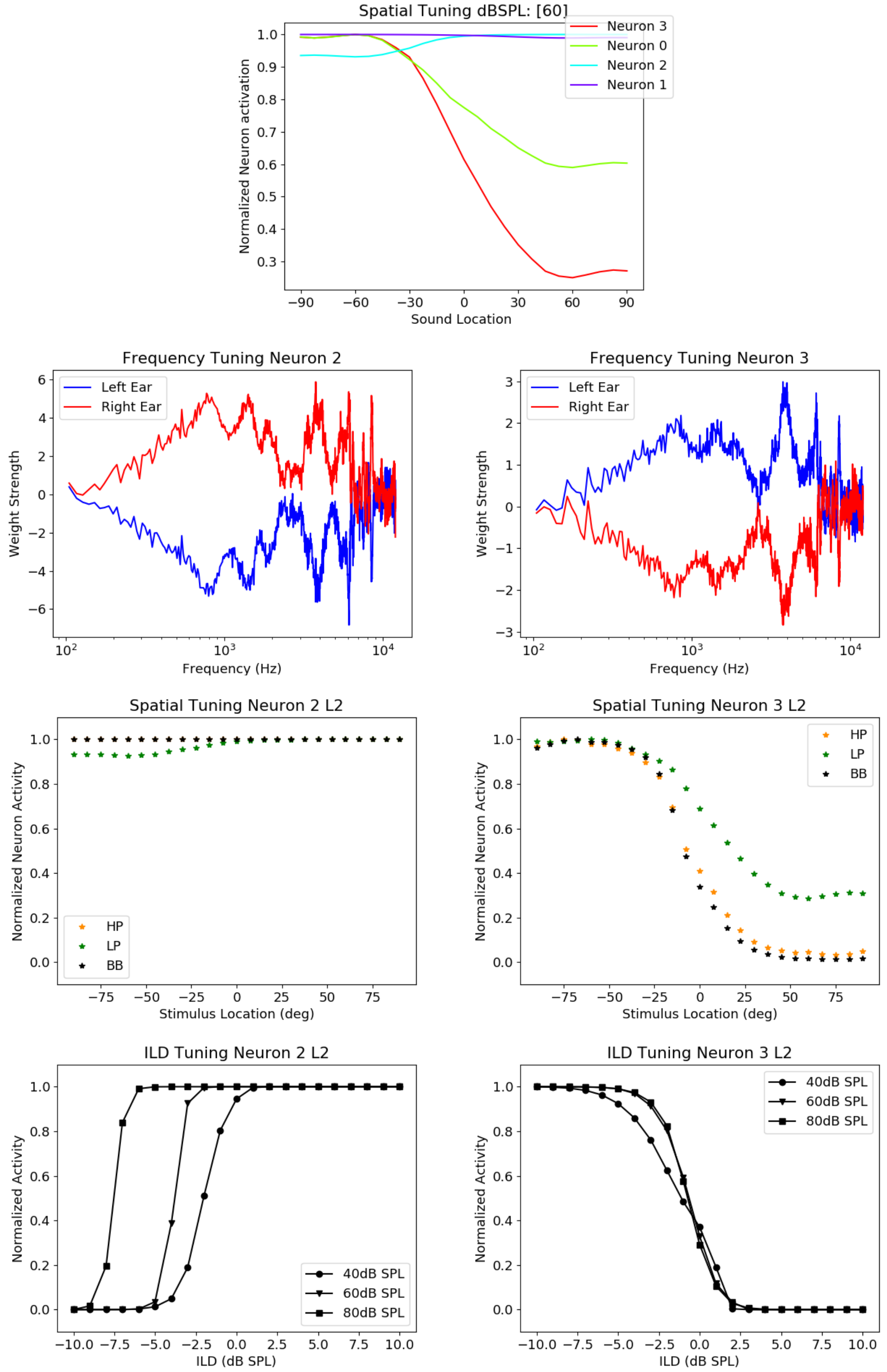


Figure 10: The different tuning curves of the second hidden layer of the Early Fusion model. Top to bottom row: Spatial Tuning of the hidden neurons, Frequency Tuning of the hidden neurons, Spatial Tuning of the hidden neurons per frequency band, ILD Tuning of the hidden neurons for three levels.

3.5 Tuning of the Late Fusion model

The tuning of the Late Fusion model is discussed in two parts. First we will look at the tuning of the first two hidden layers that only receive monaural input from one ear (Fig. 11). Afterwards, we will analyse the tuning of the third hidden layer where the two monaural hidden layers are fused (Fig. 12).

Looking at the spatial tuning of the two monaural hidden layers (Fig. 11), we can see that almost all neurons show no spatial tuning. This is expected as monaural input should not allow for spatial tuning. Still, a single neuron receiving only input from the right hemisphere, shows an increase in activity around -50°. This is unexpected as the neuron does not receive any input from the left ear.

Most neurons show broad frequency tuning, similar to Neuron 8 from the left side. Additionally, upon closer investigation of the spatial tuning of the single neurons, we can see in the spatial tuning that the activity of the neurons that receive monaural input still changes with the location of the sound source across both hemispheres. Also the ILD tuning reveals that the monaural neurons are sensitive to variations in ILDs of their respective hemisphere, but are keep their activation for the opposing hemisphere. The ILD tuning thereby cannot explain the contralateral spatial tuning of the neurons.

Moreover, the special neuron of the right hemisphere, which was coding for the left hemisphere seems to be Neuron 12R. From its spatial tuning we can see that it has a large increase for low-pass filtered sound presented in the left hemisphere, with maximal activity around -50°. However, this neuron is mostly tuned to frequencies above 1 kHz.

The spatial tuning of the second hidden layer (Fig. 11) shows a similar distribution of spatial tuning as we observed in the Early Fusion model, although for the opposite hemispheres. But the frequency tuning reveals that the spatial tuning is not achieved by the earlier observed frequency tuning which shows inverted frequency tuning for the left and right ear. Rather, the neurons are dominated by frequencies of the left ear with weak connections for frequencies of the left ear. This effect is visible for neurons coding for either hemisphere. Again, we can see in the spatial tuning curves that the only neuron coding for the left hemisphere is only spatially tuned for low frequencies, while the neuron coding for the right hemisphere is spatially tuned to all frequency bands.

In conclusion this means that also the Late Fusion model is not biologically plausible. First and foremost because the neurons of the first hidden layer seem to be spatially tuned to both hemispheres, which should not be possible from monaural input only. Also in the second hidden layer, we can see a larger contribution of the left frequencies compared to the right frequencies. Moreover, we can see that the only neuron coding for the left hemisphere in the second hidden layer is solely tuned to low frequencies. This is also contrary to our expectations.

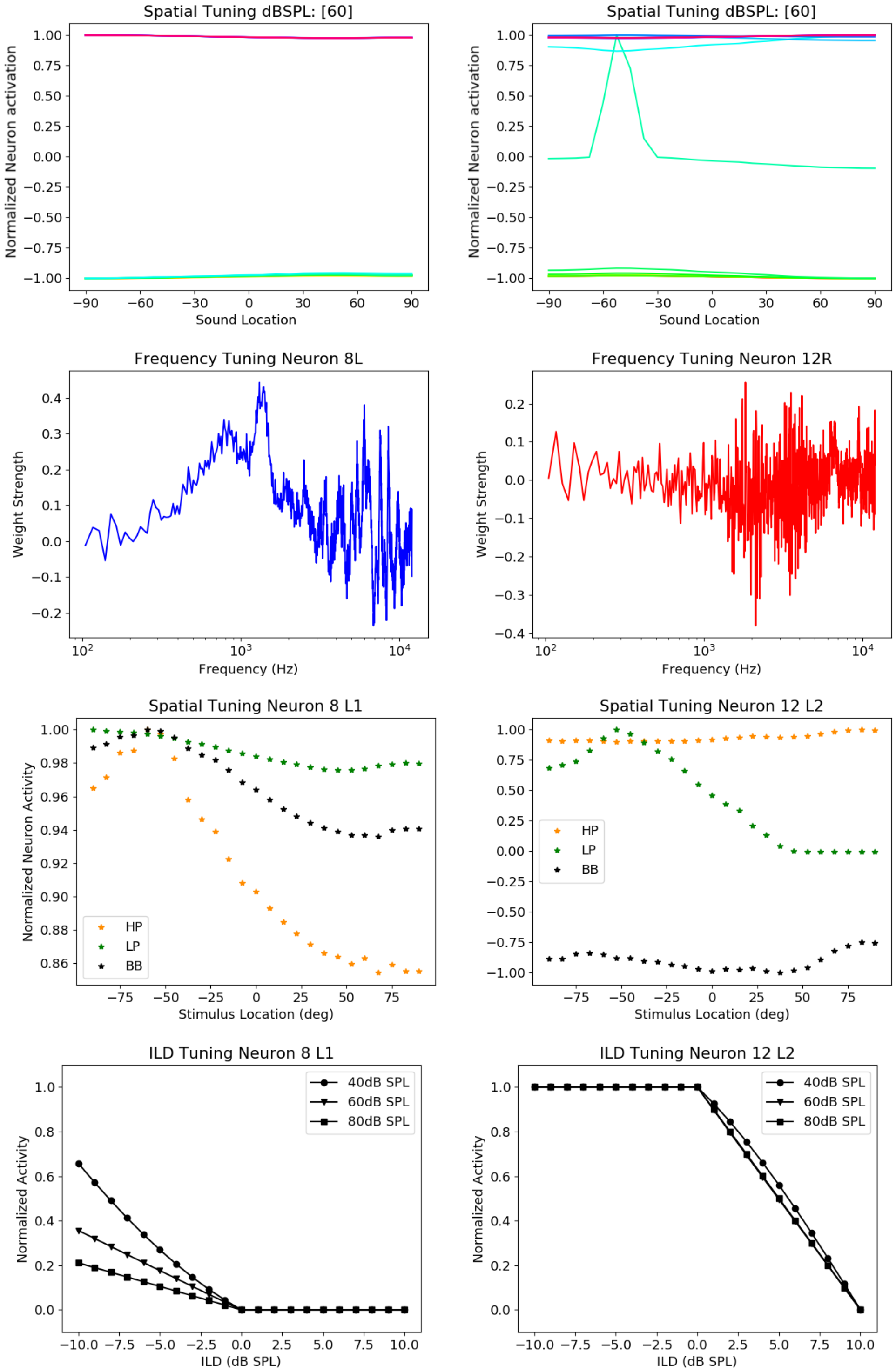


Figure 11: The different tuning curves of the Late Fusion model. Top to bottom row:: Spatial Tuning of the hidden neurons, Frequency Tuning of the hidden neurons, Spatial Tuning of the hidden neurons per frequency band, ILD Tuning of the hidden neurons for three levels.

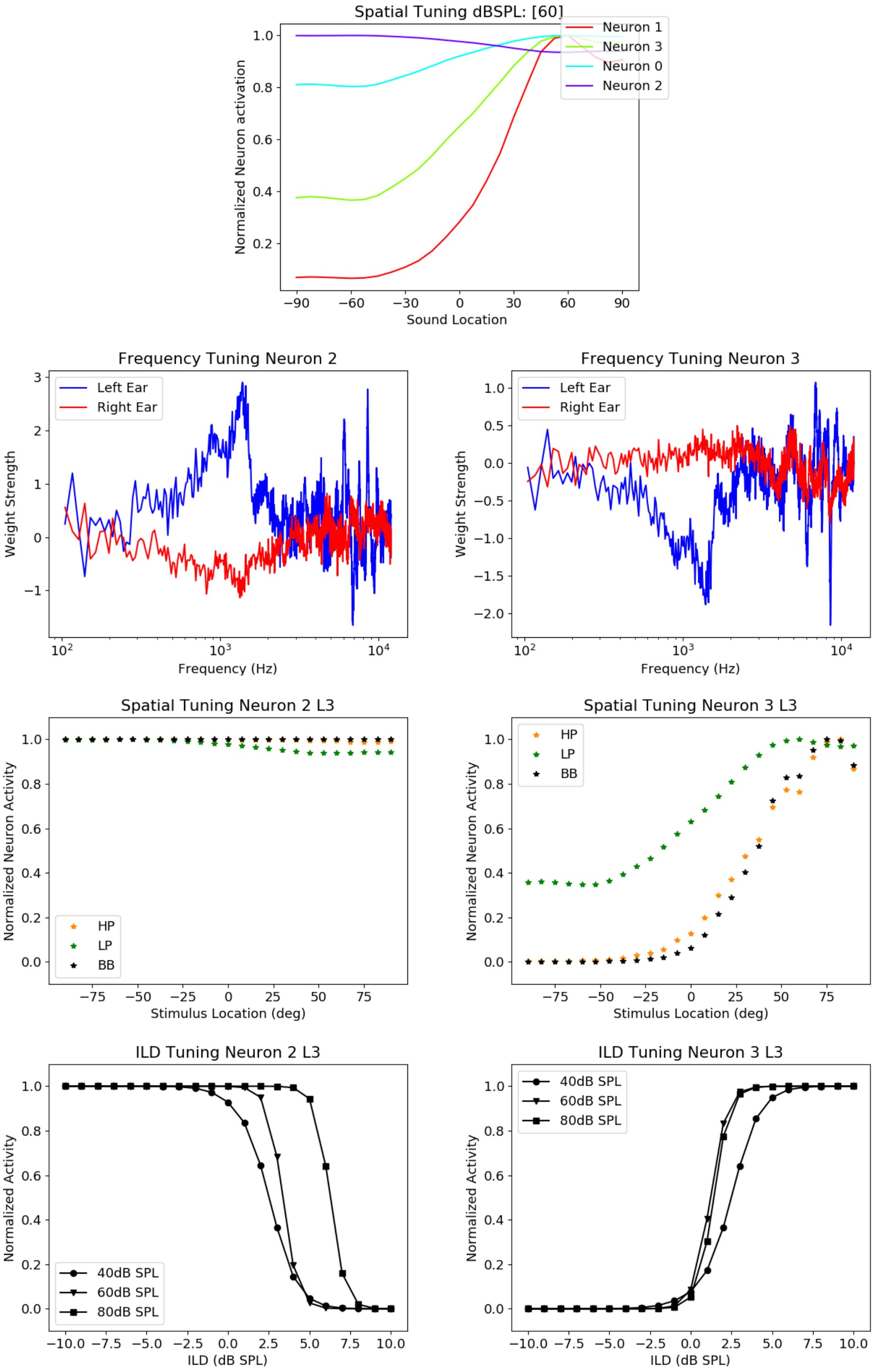


Figure 12: The different tuning curves of the Late Fusion model. From top to bottom: Spatial Tuning of the hidden neurons, Frequency Tuning of the hidden neurons, Spatial Tuning of the hidden neurons per frequency band, ILD Tuning of the hidden neurons for three levels.

4 Discussion

During this master thesis we assessed the biological plausibility of binaural neural networks with different complexity by introducing three new evaluation criteria, next to the stimulus-response localization plots. During the introduction, we established that the individual artificial neurons need to have hemispherical sigmoidal tuning curves in space and for ILDs, while also being sharply tuned to narrow frequency bands. To rule out architecture specific effects, we tested four different architectures. The Ausili model is a recent model that reportedly mimicked the human localization behavior accurately. In order to establish a baseline for our additional evaluations, we also trained and evaluated a simple version of the Ausili architecture with only 2 hidden nodes.
During the development of the project, we quickly realised that neither of these two architectures showed signs of sharp frequency tuning. Therefore, we introduced architectures with an additional hidden layer. We hypothesized that the additional layer might give the neural network the opportunity to extract narrow frequency information before calculating binaural interactions. The idea of the two hidden layer models was further supported by the schematic overview of the ILD pathway (Fig. 1B), where we saw that an additional nucleus (AVCN) processes the information from the respective ear before the ILDs are calculated at the LSO. To that end, we designed the Early- and Late-Fusion models. The rational behind the two monaural layers of the Late Fusion model was to force the network to extract narrow frequency bands per ear that are then binaurally integrated in the second hidden layer. The Early Fusion model was designed to control for the effects of the additional hidden layer and the higher number of artificial neurons.

The localization results of the four models revealed that only the Ausili and Early Fusion architectures mimicked the human localization results accurately for Broadband and High-pass sounds. The late-fusion model was close to the human performance, but had difficulties to accurately predict far laterally presented sounds. The simple model was only able to accurately predict the hemisphere of the sound, with no predictions outside ±60°. However, all models performed worse than the human baseline for low-pass sounds. This perfectly suits our assumptions because during the stimulus creation we focused on ILDs and did not include any phase information. Thereby, the ITD cue, which allows humans to localize low frequency sounds, was absent in the data.
Overall these results show that the Ausili and Early Fusion architectures are able to model the human sound localization behaviour under the conventional evaluation criterion's for artificial binaural neural networks.

However, upon closer investigation of the architectures with our three proposed evaluation criteria, we realised that the conventional measures are not related to the biological plausibility of the neural networks.
The most convincing argument for this conclusion is that narrow frequency tuning was never observed. Almost all neurons are broadly tuned to all frequencies. Across all models, we only found two neurons in the Early Fusion architecture that were not tuned to frequencies below 1 kHz, but even these neurons showed broad tuning for higher frequencies.

Still, looking only at the frequency tuning, we concluded that the neurons are sensitive to ILDs, due to the symmetry between the frequency tuning of the left and right ear. If the neuron is positively weighing the a frequency at the one ear, it is negatively weighing the same frequency from the other ear. This was further supported by the ILD tuning curves, that showed that neurons coding for a specific hemisphere are also

sensitive to ILD changes of that hemisphere. Together, the sigmoidal shaped ILD tuning curves and the ipsilateral excitation and contralateral inhibition of the frequency, form a strong argument that the ANNs utilize binaural ILDs to predict the location of the sound.
Also the spatial tuning curves of all models have a sigmoidal shape and the neurons are distributed in two opponent neural populations that increase their activity the further a sound is laterally presented. From the simple model we learned that even two neurons follow the opponent coding strategy to localize sounds. But it also seemed like two neurons are not enough to cover the whole hemisphere.

Despite of all of this evidence, we still have the suspicion that ILDs are not the driving mechanism behind the localization behavior of the binaural networks because of the neurons being spatially tuned to low frequency sounds. If the neurons would use only ILDs to predict the location of a sound, we would expect no spatial tuning to low frequency sounds. Additionally, the spatial tuning curves of all models show a relation between the spatial tuning and the ILD tuning. But especially because this relation is also present in the monaural layers of the Late Fusion model, we think that the ILD tuning curves can be an artifact of the overall input level change. The ILDs are constructed by changing the overall level of a sound with respect to a certain base frequency. Hence, with increasing negative ILD the sound at the left ear is increasingly getting louder and thereby the activity of the neuron increases. Moreover, the monaural neurons are also, although weakly, spatially tuned. This spatial tuning even continues in the hemisphere where they receive no input.

5 Conclusion

Putting all of this together, we conclude that the tested binaural neural networks are not biologically plausible, yet alone due to their broad frequency tuning. Additionally, the ILD and spatial tuning curves let us believe that the networks exploit a different kind of information than relying purely on ILDs.
Still, the presented work is an important step towards biologically plausible neural networks, because we showed that any single evaluation criterion is not enough to validate binaural models. In addition to that, we also think that the training of the models shouldn't purely rely on sound localization of band-pass filtered noise. Despite our utmost efforts in creating a wide variety of sounds, band-pass filters and sound levels, non of the models developed narrow frequency tuning. We belief that a more complex task and input might force the artificial neurons to narrow frequency bands, which will be discussed further in the future work section.
Lastly, we conclude that the tested artificial neural networks are not yet ready to simulate human binaural hearing for the development of hearing aids like cochlear implants. This thesis showed that the tested models do not fully rely on ILDs to predict the location of a sound. This means that some other feature must also influence to localization behaviour. If we would train these models on CI input with different settings, it might be that the ANNs reach good performance based on an input feature that is not usable for the human auditory pathway. We can only utilize ANNs for the development of better CI processors or support audiologists in their fitting, after we can create a biologically plausible model of binaural hearing with artificial neural networks.

6 Future Work

Based on this research we can foresee three possible directions how future research could further pursue the goal of developing better CIs with the help of ANNs.

First of all, future research could revolve around the exact explanation of the localization strategy of the analyzed networks. While ILDs might play some role in the prediction of the sound location, there is reason to belief that ILDs are not the only cue the ANN exploits. To that end, a future project could try to explain the localization behaviour with spectral template matching [11] or come up with other hypothesis on how the binaural neural networks could localize sounds.

A second research trajectory could try to use the presented evaluation measures to test other network architectures or training paradigms in order to develop biologically plausible neural networks. While the architectures seem to be well suited for the task of sound localization, we think that the introduction of a different task and complex sounds could force the networks to develop sharp frequency tuning. We hypothesize that complex sounds and the introduction of background noise might facilitate the development of sharper frequency tuning because being sensitive to all frequencies might not be suitable for this type of data. Currently, all present frequencies contain the location information because they have been convolved with the HRTF, while all other frequencies are 0. Thereby, broad tuning gives the ANN no penalty, as the input already contains the information of the important frequencies. However, if background noise would be introduced, the neural network would need to focus on the frequencies containing spatial information while ignoring all other frequencies. We also expect that a task that enforces the need to distinguish between frequencies will also help the frequency tuning. Training the model on pitch detection, musical genre discrimination or speech recognition, next to sound localization, is also biologically closer to how humans learn sound localization. Humans do not learn to localize sounds with a bunch of noise sounds from different locations; only the sound localization ability is assessed in this way. Rather, humans want to listen to a certain conversation in noisy environments, for example, and thereby we concentrate on a certain pitch in a certain direction.

Lastly, despite these networks not being biologically plausible, a future line of research could still analyse networks trained under perturbed listening conditions. We think that the differences in spatial and ILD tuning, e.g. for normal hearing input compared to CI input, could lead to interesting observations that might already form the basis for CI settings or processing strategies. However, it is questionable whether these findings can be translated to advances for humans, as we can not explain how these networks localize.

7 Funding

This project was funded by AI for Health (`https://www.ai-for-health.nl/about/`), which is a collaboration between the Radboud University and Radboudumc, aimed at developing innovations based on artificial intelligence that solves clinical problems.

8 Data availability

The data and program code used for this project is publicly available at `lhttps://github.com/tichter/simulatedbinauralhearing`. If you are interested, you can also find all the tuning curves in separate folder.

9 Acknowledgements

I want to thank my supervisors for their critical questions and constructive feedback. I also want to thank them for the opportunity to participate in their daily activities, which facilitated the shaping of my future career prospects.

I want to thank my family and friends for their continued support during my whole university studies and this thesis.

A special thank you to Ignacio Calderon De Palma for always being open to discuss this thesis and sound-processing in general.

References

[1] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[2] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[3] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[4] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[5] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

[6] Umut Güçlü, Jordy Thielen, Michael Hanke, and Marcel AJ van Gerven. Brains on beats. *arXiv preprint arXiv:1606.02627*, 2016.

[7] Andrew F Francl and Josh H McDermott. Deep neural network models of sound localization reveal how perception is adapted to real-world environments. *bioRxiv*, 2020.

[8] S. A. Ausili. Spatial hearing with electrical stimulation listening with cochlear implants, doctoral thesis, 2019. `https://repository.ubn.ru.nl/handle/2066/203054` accessed: 1.04.2021.

[9] Ning Ma, Tobias May, and Guy J Brown. Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2444–2453, 2017.

[10] Daryl Kelvasa and Mathias Dietz. Auditory model-based sound direction estimation with bilateral cochlear implants. *Trends in Hearing*, 19:2331216515616378, 2015.

[11] Martin Rothbucher, David Kronmüller, Marko Durkovic, Tim Habigt, and Klaus Diepold. Hrtf sound localization. *Advances in Sound Localization. Rijeka, Croatia: InTech*, pages 79–94, 2011.

[12] E Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979, 1953.

[13] Jens Blauert. *Spatial hearing: the psychophysics of human sound localization.* MIT press, 1997.

[14] Bahram Zonooz, Elahe Arani, Konrad P Körding, PAT Remco Aalbers, Tansu Celikel, and A John Van Opstal. Spectral weighting underlies perceived sound elevation. *Scientific reports*, 9(1):1–12, 2019.

[15] Tom CT Yin. Neural mechanisms of encoding binaural localization cues in the auditory brainstem. In *Integrative functions in the mammalian auditory pathway*, pages 99–159. Springer, 2002.

[16] Daniel J Tollin and Tom CT Yin. The coding of spatial location by single units in the lateral superior olive of the cat. i. spatial receptive fields in azimuth. *Journal of Neuroscience*, 22(4):1454–1467, 2002.

[17] CHIYEKO Tsuchitani. Functional organization of lateral cell groups of cat superior olivary complex. *Journal of neurophysiology*, 40(2):296–318, 1977.

[18] Benedikt Grothe, Michael Pecka, and David McAlpine. Mechanisms of sound localization in mammals. *Physiological reviews*, 90(3):983–1012, 2010.

[19] Blake S Wilson and Michael F Dorman. Cochlear implants: a remarkable past and a brilliant future. *Hearing research*, 242(1-2):3–21, 2008.

[20] Bolajoko O Olusanya, Katrin J Neumann, and James E Saunders. The global burden of disabling hearing impairment: a call to action. *Bulletin of the World Health Organization*, 92:367–373, 2014.

[21] Florian Denk, Stephan MA Ernst, Jan Heeren, Stephan D Ewert, and Birger Kollmeier. The oldenburg hearing device (olhead) hrtf database. Technical report, Technical report, 2018.

[22] Eric I Knudsen and Masakazu Konishi. Mechanisms of sound localization in the barn owl (tyto alba). *Journal of Comparative Physiology*, 133(1):13–21, 1979.

[23] Marc M Van Wanrooij and A John Van Opstal. Sound localization under perturbed binaural hearing. *Journal of neurophysiology*, 97(1):715–726, 2007.

[24] Paul M Hofman and A John Van Opstal. Spectro-temporal factors in two-dimensional human sound localization. *The Journal of the Acoustical Society of America*, 103(5):2634–2648, 1998.

[25] sengpielaudio. Poll: Is 3 db, 6 db or 10 db spl double the sound pressure?, -. `http://www.sengpielaudio.com/calculator-levelchange.htm` accessed: 1.04.2021.

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[27] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

A Losses

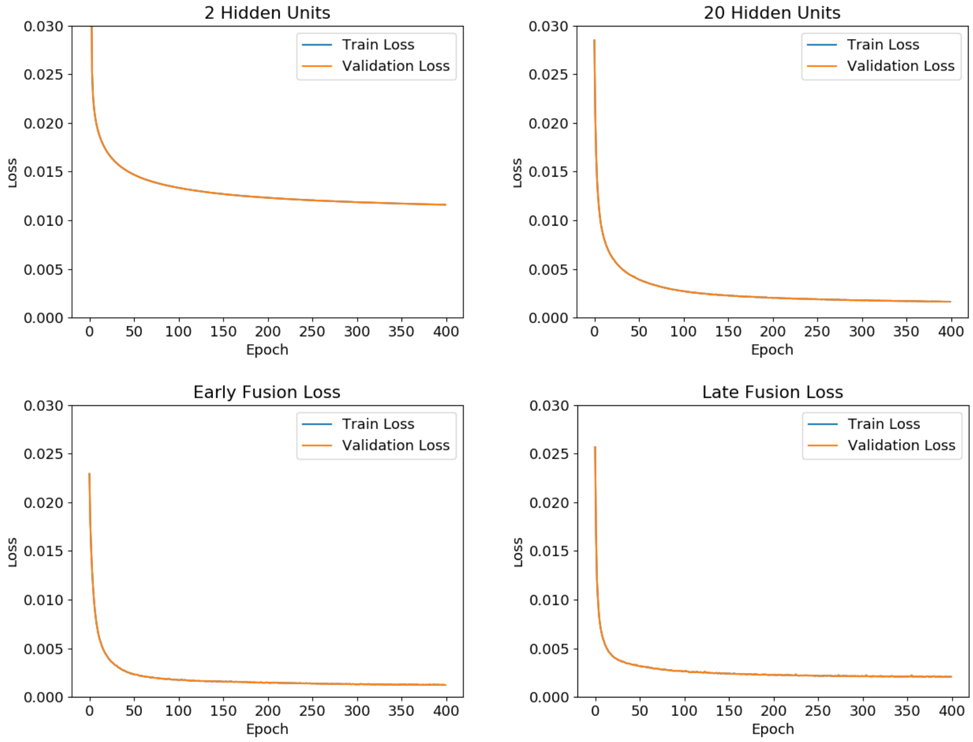


Figure 13: Training and Validation Loss of the four architectures (Top left to bottom right: Simple, Ausili, Early Fusion, Late Fusion). For all models, the validation Loss decreased at roughly the same rate as the training loss.

B Overall Performance

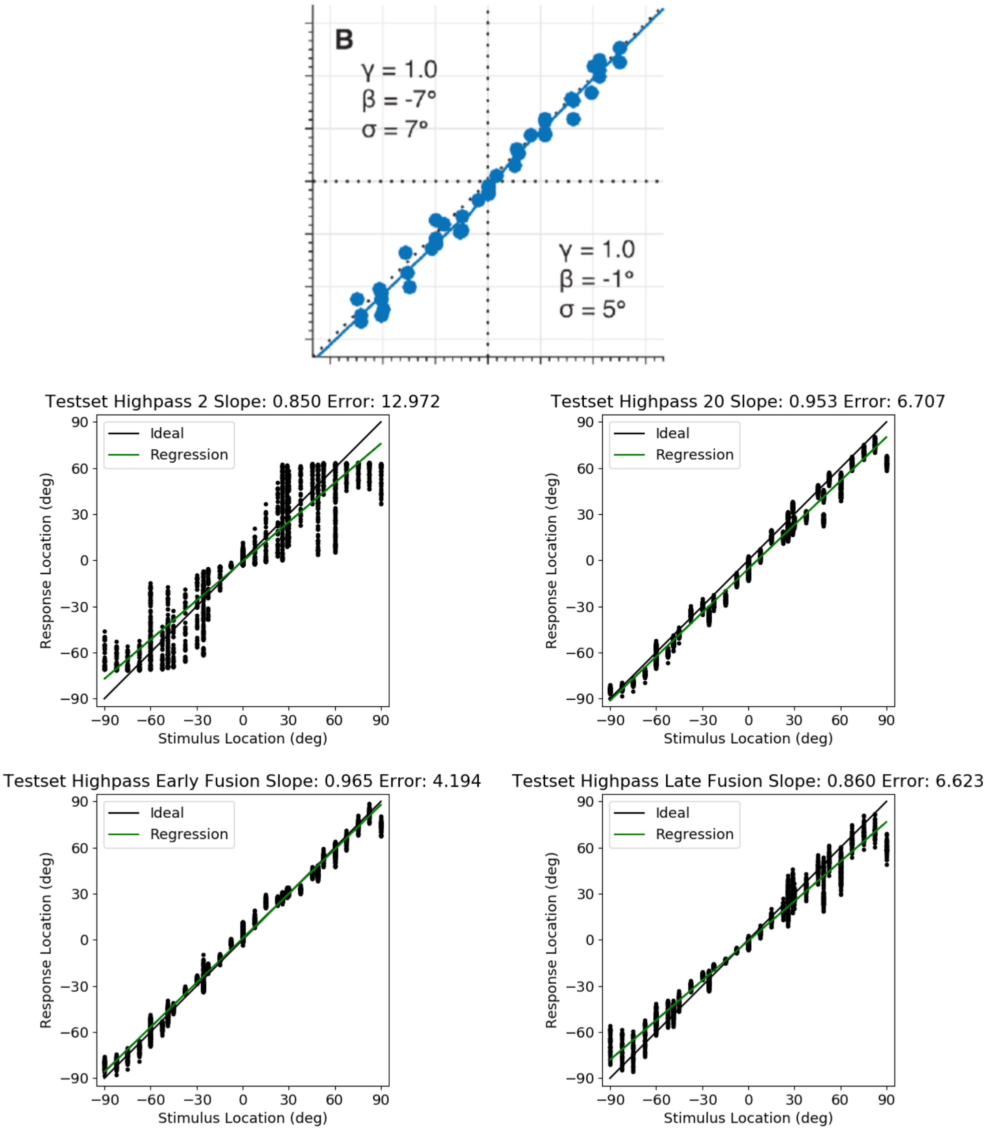


Figure 14: High-pass localization plots the four architectures (Top left to bottom right: Simple, Ausili, Early Fusion, Late Fusion) and the human reference from [8] (Top)

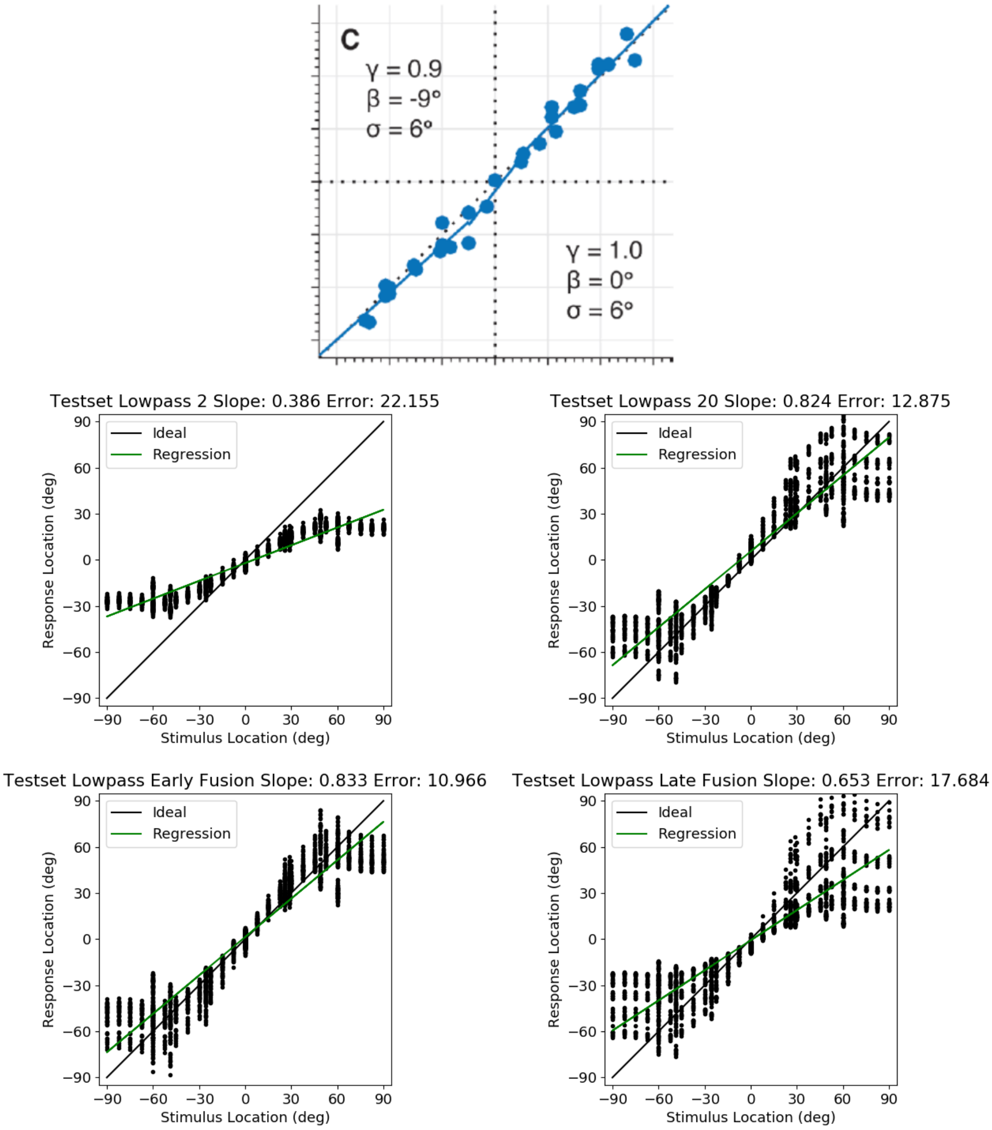


Figure 15: Low-pass localization plots the four architectures (Top left to bottom right: Simple, Ausili, Early Fusion, Late Fusion) and the human reference from [8] (Top)