Department of Physics and Astronomy

University of Heidelberg

Bachelor Thesis in Physics

submitted by

Markus Kreft

born in Berlin (Germany)

March 2019

Structural Plasticity for Feature Selection in Auditory Stimuli on Neuromorphic Hardware

This Bachelor Thesis has been carried out by Markus Kreft at the
Kirchhoff-Institute for Physics
Heidelberg University
under the supervision of
Dr. Schemmel and Prof. Meier

Abstract

This thesis presents an approach to processing auditory stimuli with recurrent neural networks implemented on the HICANN-DLS neuromorphic hardware prototype.
Auditory stimuli consist of many spike channels. For typical acoustic signals, however, only certain spike channels corresponding to particular frequency bands hold relevant information. While the employed HICANN-DLS chip has limited resources, especially in terms of possible synaptic connections, the implementation of structural plasticity, a constant rewiring of synapses, is possible.
In an attempt to select the relevant channels of auditory stimuli, a structural plasticity rule is realized that chooses one of multiple inputs provided to each synapse based on pre-postsynaptic spike-time correlation. In experiments performed with artificial stimuli constructed to mimic correlation patterns found in auditory stimuli, a preference of synapses to couple to strongly correlated inputs was observed. An increased coupling to channels with high correlation was also demonstrated for stimulation with actual auditory stimuli from a dataset of spoken digits. Correlation in these channels coincides with high information on the digits.

Zusammenfassung

Diese Arbeit präsentiert einen Ansatz zur Verarbeitung auditiver Stimuli mit Rekurrenten Neuronalen Netzwerken auf dem neuromorphen Hardware Prototypen HICANN-DLS.
Auditive Stimuli besitzen eine hohe Anzahl von Spike Kanälen. Bei typischen akustischen Signalen beinhalten allerdings nur einige zu bestimmten Frequenzbereichen gehörende Kanäle relevante Information. Die Ressourcen des verwendeten HICANN-DLS Chips sind, insbesondere im Hinblick auf die Anzahl realisierbarer synaptischer Verbindungen, begrenzt. Er ermöglicht jedoch die Implementierung von Struktureller Plastizität, einer ständigen Neuordnung der Synapsen.
In einem Versuch, bei auditiven Stimuli die relevanten Kanäle auszuwählen, wurde eine Strukturelle Plastizitätsregel eingesetzt, die in jeder Synapse einen von mehreren verfügbaren Eingängen basierend auf der Korrelation von prä- und postsynaptischen Spikezeiten auswählt. In Experimenten mit künstlichen Stimuli, die ähnliche Korrelationsmuster wie auditive Stimuli aufweisen, wurde eine Präferenz der Synapsen beobachtet, an stark korrelierte Kanäle zu koppeln. Bei Stimulation mit auditiven Stimuli aus einem Datensatz gesprochener Ziffern wurde ebenfalls eine erhöhte Koppelung an Kanäle mit hoher Korrelation demonstriert. In diesen Kanälen tritt starke Korrelation zusammen mit einem hohen Informationsgehalt über die gesprochenen Ziffern auf.

List of errata corrected in this version from March 15, 2019.

Location	**Error**	**Correction**
Table 5.1	$T_{\text{stim}} = 4.2\,\mu\text{s}$	$T_{\text{stim}} = 4.2\,\text{ms}$
Figure 3.3	Colorbar axis is labeled in percent	Should be fractions
Title	Auditoiry	Auditory
Acknowledgments	Vision(S)	Vision(s)
All occurrences	synaptic drivers	synapse drivers
All occurrences	initialisation	initialization
All occurrences	realisation	realization
All occurrences	hyperpolarisation	hyperpolarization
All occurrences	visualisation	visualization

Contents

Abbreviations

ADC	analog-to-digital converter
ANN	artificial neural network
BM	basilar membrane
CDF	cumulative distribution function
DAC	digital-to-analog converter
ERB	equivalent rectangular bandwidth
FPGA	field-programmable gate array
HC	hair cell
HICANN-DLS	High Input Count Analog Neural Network with Digital Learning System
LIF	leaky integrate-and-fire
LSB	least significant bit
LSM	liquid state machine
MI	mutual information
PPU	plasticity processing unit
SNN	spiking neural network
STDP	spike-timing dependent plasticity

List of Figures

1 Introduction

Computation in the human brain is achieved by a large network of interconnected units that can process, store, and transmit information. The dynamics of such systems are extremely complex and far from being fully understood [Yuste et al. 2014]. Even simple models designed to capture the basic properties of spiking neural networks (SNNs), most prominently the creation of action potentials that lead to neural spikes, show highly complex dynamics [Gerstner et al. 2014; Cessac 2008]. Such SNNs however have the potential to be computationally more powerful than more abstract artificial neural networks (ANNs) [Maass 1997]. Therefore they are an object of interest for future computing systems. A way to gain understanding of the network behavior is through experimental study of their dynamics. While simulation on conventional computers is possible, the parallel architecture of models leads to long simulation times and high power consumption. A scalable and low-power alternative is the emulation of the network dynamics on neuromorphic hardware.

The BrainScaleS hybrid platform [Schemmel et al. 2010] is a specialized hardware system for such emulation. It implements synapses and neurons in analog electrical circuitry while the network configuration is defined digitally. In addition to low power consumption, here the choice of analog hardware components results in a considerable shortening of the emulation run-time compared to its biological equivalent.

In this thesis a prototype hardware chip for the next generation platform with enhanced features is used. The High Input Count Analog Neural Network with Digital Learning System (HICANN-DLS) version 2 features an on-chip processor for implementation of synaptic plasticity [Friedmann et al. 2017]. This allows the network configuration to change over time based on plasticity rules. Such adaptation of the network is considered to play a major role in human learning [Hebb 1949; Hughes 1958].

This thesis deals with the processing of auditory stimuli by spiking neural networks implemented on the HICANN-DLS version 2. The hardware chip is configured in a way that allows later use in an liquid state machine (LSM) setup [Maass et al. 2002; Maass et al. 2004], where a recurrent network receives external input spikes from a preprocessing step. It processes the input in a non-linear way and provides output that can then be separated by a linear classifier for classification tasks. Such a setup with a simulated SNN containing hundreds of neurons has been used for isolated word recognition [Verstraeten et al. 2005]. In the network implemented in this thesis, the synaptic connections between neurons are subjected to a variant of spike-timing dependent plasticity (STDP). This has been shown to stabilize the network activity [Stöckel et al. in prep.]. The network recurrence allows to control the distance to the critical point between a phase where activity rapidly dies out and a phase of activity amplifying over time [Cramer et al. in prep.(a)]. This has the potential to enables high computational power [Beggs et al. 2012].

The generation of input spikes from sound signals is inspired by the human auditory system. Within the human ear in the cochlea sound waves are transformed into neural spike trains that serve as input channels to the human brain. Activity in these channels depends on the frequency bands present in the stimulus. Spike times show correlation patterns, both within a single channel and between different channels. The correlation is caused by resonance effects of the basilar membrane that is responsible for spike creation in the inner ear.

In the human auditory system there exist thousands of input channels to the brain. Typical auditory stimuli however do not exhaust the full available frequency range and therefore not all of these channels are relevant to the stimulus. The HICANN-DLS hardware prototype used in this thesis has only a low number of 32 neurons that can receive external spikes from 32 inputs. It however incorporates an address event coding system that allows to feed multiple spike trains on each input. The integrated on-chip processor can select which of the possible inputs are transmitted to the neurons. This can be used to implement structural plasticity.

Structural plasticity in general describes mechanisms that change a networks internal connections over time [Lamprecht et al. 2004; Butz et al. 2009]. When providing each input of the chip with multiple spike trains and subjecting the network to a structural plasticity rule that can change the event coding addresses, spike trains with certain features depending on the plasticity rule can be selected. Since this also results in less connections to spike trains not showing the desired features, potentially less noise is introduced and resources can be conserved.

This implementation of structural plasticity is used here to select highly correlated input channels in artificial stimuli that are constructed with different temporal and spacial correlation. In addition, a dataset of spoken digits is used as an example for an actual auditory stimulus. In chapter 2 the model implementation, the stimulation paradigm, and information theoretic methods for evaluation are presented. Results from the conducted experiments are given in chapter 3 and discussed in chapter 4.

2 Methods

The experiments presented in this thesis are conducted on the HICANN-DLS mixed-signal neuromorphic hardware chip. It implements biologically inspired neuron and synapse models in analog electrical circuitry. The chip is stimulated with signals of different correlation in time and between neurons. A structural plasticity rule acts on the external network connections to allow the selection of a preferred correlation pattern. Experiments are conducted both with artificial and biologically realistic stimuli.

Below a short biological background is provided, followed by a description of the hardware chip. Section 2.3 gives a detailed description of the implemented neurons, synapses and network model, as well as the plasticity rules. In section 2.4 the stimuli used to drive the network are specified. Information theoretic methods for evaluation are described in section 2.5.

2.1 Biological Principles

In the following a short overview of the most relevant aspects of spiking neural network models is given. A comprehensive introduction to the topic can be found in [Gerstner et al. 2014]. Computation in the human brain is achieved by a large network of cells, the *neurons*. They propagate pulses of electrical charges which allows for processing of information.

A neuron typically consists of three main parts, the dendrites, the soma, and the axon, as schematically drawn in figure 2.1. Dendrites act as inputs to the neuron by transmitting signals coming from other neurons. The soma, the central unit of the neuron, can process these signals in a non-linear way. Generated output is then taken through the axon to neighboring neurons. The axon can stretch over macroscopic distances in the brain enabling non-locally connected networks.

The signal processing in the soma is based on the collection of charges that build up a potential over the cellular membrane. Ion channels in the membrane allow certain charge carriers to cross. If no input is received by a neuron the membrane is in a resting state of constant polarization caused by the surrounding environment. The membrane dynamics can be modelled at different levels of abstraction by differential equations describing the

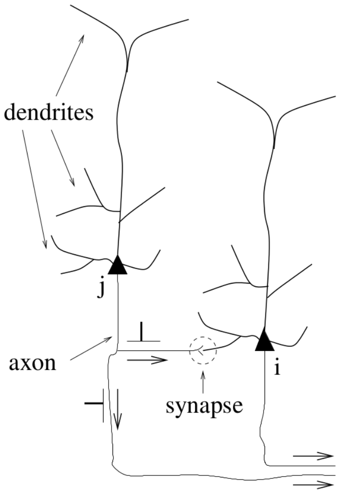


Figure 2.1: Schematic illustration of two neurons j and i connected by a synapse. The dendrites and axon of the presynaptic neuron are marked, the triangle highlights the position of the soma. A chemical synapse connects the pre- and postsynaptic neurons and shows a synaptic cleft. Figure taken from [Gerstner et al. 2014]

membrane potential. In most models, once external charges are built up and a threshold potential is reached, the neuron membrane shows a steep depolarization followed by a phase of hyperpolarization. If such an *action potential* is emitted, the neuron is said to *fire* a *spike*.

The connections between the axon of one neuron and the dendrites of other neurons are called *synapses*. They mostly occur in form of chemical synapses. The axon of the *presynaptic* neuron ends very close to the dendrite of the *postsynaptic* neuron leaving only a small gap in between, the synaptic cleft (figure 2.1). If an action potential arrives at the cleft, neurotransmitters are released. These allow the passage of certain ions through the channels in the membrane of the postsynaptic neuron, where a postsynaptic potential can build up. As a result the presynaptic electrical pulse is transmitted from one neuron to another.
The *synaptic weight* characterizes the strength of a synaptic connection. It depends on the amount of neurotransmitters released in the synapse and the type and number of ion channels activated in the postsynaptic neuron. If a presynaptic spike increases the postsynaptic membrane potential, a synapse is said to be *excitatory*, if it leads to a decrease of the potential it is called *inhibitory*.

In the human brain synaptic connections are not fixed, but change over time. This is an important part of learning and summarized by the term *plasticity*.
There are different processes that lead to a change in synaptic weights. The spike-timing dependent plasticity (STDP) mechanism [Bi et al. 1998], used here in a modified version,

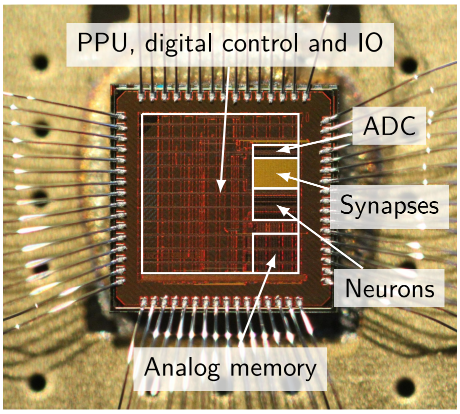


Figure 2.2: Photograph of a HICANN-DLS version 2 prototype chip. The digital components, neurons, and the synapse array are highlighted. Figure adapted from [Stöckel et al. in prep.].

is based on Hebb's Postulate [Hebb 1949]. There, a synaptic weight is adapted depending on the temporal correlation of pre- and postsynaptic neuron spikes. The general idea is that if a neuron takes part in invoking a postsynaptic spike of another, their connection is important for information propagation and should be developed in the future [Toyoizumi et al. 2007].
However, plasticity also incorporates the restructuring of a network by rewiring connections. This change in the network topology is termed *structural plasticity*. It can invoke reconnection of neurons based on different properties like neuron activity, current synaptic weight or temporal spike correlations. Structural plasticity usually acts on a longer time scale compared to typical STDP rules and is responsible for long term change of the network [Lamprecht et al. 2004; Butz et al. 2009; Deger et al. 2017].

2.2 Hardware

The HICANN-DLS version 2 prototype chip used in this thesis implements 32 LIF neurons (see 2.3.1) and a total of 1024 synapses. It is technically specified in [Friedmann et al. 2017] and [Aamir et al. 2018]. In the following a brief introduction to the basic components and working principles is given. A comprehensive overview can be found in [Stöckel 2017].

The chip has a mixed signal architecture with analog and digital parts, the most prominent components are highlighted in figure 2.2. Neuron membranes and synapses are emulated in analog electrical circuitry. The relevant capacitor time constants are decreased by a factor of approximately 1000 compared to the biological model parameters, which results in a corresponding time speed-up of the emulation compared to the biological time domain. The network configuration and spike distribution is implemented digitally. The chip can be configured from a host computer through an on-board field-programmable gate ar-

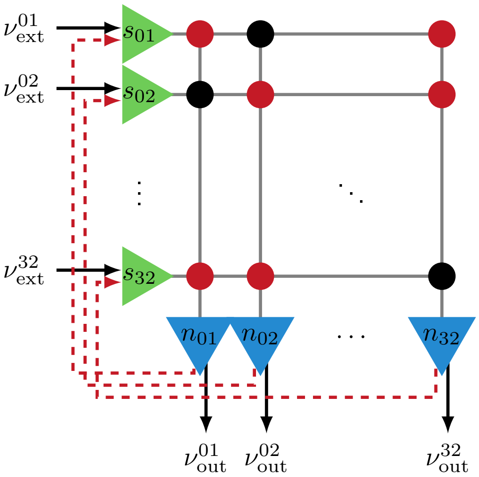

ν_{ext}^{01}
ν_{ext}^{02}
ν_{ext}^{32}
s_{01}
s_{02}
s_{32}
n_{01}
n_{02}
n_{32}
ν_{out}^{01}
ν_{out}^{02}
ν_{out}^{32}

Figure 2.3: Schematic overview of the synapse array: Green triangles on the left correspond to the synapse drivers transmitting the presynaptic input, blue triangles at the bottom represent the neurons. Synapses are indicated by circles at the intersection of vertical and horizontal lines. The black arrows from the left represent external input, the arrows at the bottom represent the spike readout. Recurrent spike routing is visualized by the red dashed lines. Figure taken from [Stöckel et al. in prep.].

ray (FPGA) clocked at 96 MHz. It has full read and write access to the neuron and synapse configuration and handles the recording and pass through of spikes. Analog values for model parameters are provided by a total of 17 digital-to-analog converters (DACs). Details on these parameters can be found in [Aamir et al. 2016].
Because of the implementation of neurons in independent analog circuits with individually configurable parameters the system is prone to manufacturing variations [Aamir et al. 2018]. Even if two parameters are set to the same digital value, their actual behavior may differ due to transistor mismatch. This fixed pattern noise can be reduced by calibration. Additional trial-to-trial variations arise from the configuration of parameter storage but have been found do be much smaller.

Synapses are implemented current based in a 32×32 array as schematically shown in figure 2.3. Input spikes are received by the 32 synapse drivers. These are provided in discretized time steps every 100 FPGA cycles. At each time step binary information, whether a spike is to be sent or not, is provided to all 32 inputs as a spike mask. Each driver routes the spikes to all synapses in a row. These convert the spikes to current pulses and send them along the column to the respective neuron at the bottom. Each synapse driver can be configured to be either excitatory or inhibitory. Thus all synapses in a row can either supply excitatory or inhibitory presynaptic input to the neurons.
The synaptic weights are configurable with 6 bit accuracy ranging from 0 LSB to 63 LSB and correspond to the electrical current of a transmitted pulse. A global scaling factor allows to adjust all weights simultaneously by changing the length of the current pulses.

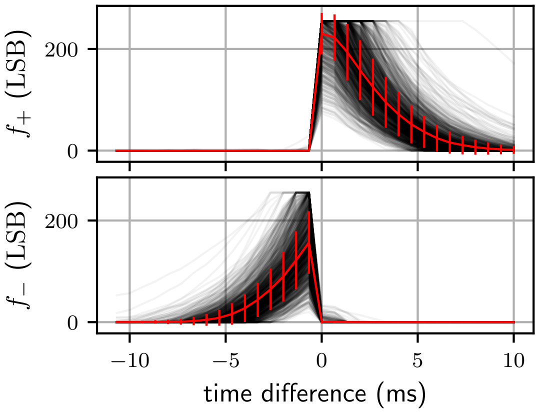


Figure 2.4: Causal and anti-causal correlation measurements provided by the dedicated analog circuitry of the HICANN-DLS chip that are accessible to the PPU. The red graph shows the average of 1024 single measurements, error bars indicate the standard deviation.

Each synapse is additionally configurable with a 6 bit decoder address. All spikes sent into the network are marked with an equivalent address. A synapse only transmits a spike if the spike address matches its decoder address. By sending spikes of different spike trains marked with different addresses one by one to a driver, the network can receive multiple inputs during a single experiment.

The FPGA is furthermore able to route spikes originating from the neurons back into the network as indicated by the red dashed lines in figure 2.3. This allows to configure recurrent networks by sending the internal spikes back into the network with a certain decoder address and configuring a fraction of synapses with the corresponding address.

The adaptation of synaptic weights, which corresponds to synaptic plasticity, is performed by the plasticity processing unit (PPU), an on-chip embedded processor. It not only has read and write access to the synaptic weights and decoder addresses but it can also read the pre-post spike correlation sensors of each synapse.

The PPU incorporates a vector extension allowing for parallel processing of 16 synapses at a time. It can act on 128 bit vectors with fractional saturation arithmetic. The PPU features the IBM Power instruction set architecture and is programmable with C code, while the vector unit is programmed with inline Assembler which makes machine code instructions directly accessible.

The correlation measurements are done by a dedicated analog circuit. For each synapse, analog accumulation rates, weighted with exponentially decaying time differences, are summed up and digitized by an 8 bit analog-to-digital converter (ADC). The correlation pairs are selected according to a reduced symmetric nearest neighbor pairing rule [Morrison et al. 2008]. The selection is done for causal and anti-causal correlations separately. Exemplary correlation measurements are depicted in figure 2.4.

2.3 Model

In this section the network model implemented on the HICANN-DLS is described. The neuron and synapse equations are given in section 2.3.1, followed by the plasticity rules in section 2.3.2. Section 2.3.3 provides an overview of the experimental procedure.

2.3.1 Neurons and Synapses

The leaky integrate-and-fire (LIF) model is an abstract model of neuron dynamics that captures the creation of a postsynaptic spike. A neuron's membrane is modeled by a capacitor circuit described by the time constant τ_{mem} and conductance $g_{\text{leak}} = C_{\text{m}}/\tau_{\text{mem}}$. A charge leaking term characterized by the leak potential u_{leak} represents ion channels in the membrane. The membrane potential u_i of neuron i obeys the differential equation

$$\tau_{\text{mem}} \frac{\mathrm{d}u_i}{\mathrm{d}t} = -\left[u_i(t) - u_{\text{leak}}\right] + \frac{I_i(t)}{g_{\text{leak}}}. \tag{2.1}$$

The current $I_i(t)$ serves as input and is attained from the sum of the currents of all presynaptic partners $I_{ij}(t)$ to neuron i. These currents are initiated by the arrival of the k-th spike, sent on input j at time t_j^k, at neuron i after the synaptic delay d_{syn} and follow the differential equation

$$\tau_{\text{syn}} \frac{\mathrm{d}I_{ij}}{\mathrm{d}t} = -I_{ij}(t) + \sum_k w_{ij} \cdot s_{\text{w}} \cdot \delta\left(t - t_j^k - d_{\text{syn}}\right), \tag{2.2}$$

which is characterized by the synaptic time constant τ_{syn} and a translation factor s_{w}. The synaptic weight w_{ij} specifies the strength of the connection between the presynaptic input j and neuron i. A value of $s_w > 0$ would correspond to an excitatory synapse, while $s_w < 0$ implies an inhibitory synapse.

Neuron i fires a spike when the membrane potential reaches the threshold value u_{thresh}:

$$u(t_i^k) \geq u_{\text{thresh}}, \tag{2.3}$$

where t_i^k denotes the k-th firing time. After firing, the membrane potential is fixed to the reset potential u_{reset} for a time duration of one refractory period τ_{ref}.

2.3.2 Plasticity

Two different plasticity rules operating on different time scales are implemented. Both alter synapses based on the correlation measurements of pre- and postsynaptic spikes. STDP adapts the synaptic weights, while structural plasticity can change the synapse decoder address which allows the network to choose spikes from different inputs.

Spike-Timing Dependent Plasticity

Spike-timing dependent plasticity (STDP) is a process that changes synaptic weights based on the timing of the spikes of pre- and postsynaptic neurons. A common kernel uses an exponential weighting of the time difference between these pre- and postsynaptic spikes [Gerstner et al. 2014] which is provided on hardware by the dedicated sensors described in section 2.2.

The update rule developed in [Stöckel et al. in prep. Cramer et al. in prep.(a)] contains three terms. Synaptic weights $w_{ij}(t)$ between neurons i and j are updated in discrete time steps T_{stdp} according to:

$$w_{ij}(t + T_{\mathrm{stdp}}) = w_{ij}(t) + \lambda_{\mathrm{stdp}} f_- \left(s_i^{[t,t+T_{\mathrm{stdp}})}, s_j^{[t,t+T_{\mathrm{stdp}})} \right) + \lambda_{\mathrm{w}} w_{ij}(t) + b_{ij}(t). \tag{2.4}$$

The term $s_i^{[t,t+T_{\mathrm{stdp}})}$ denotes the spikes of neuron i in the time interval $[t, t + T_{\mathrm{stdp}})$. The function f_- translates pre- and postsynaptic spike pairs to exponentially decaying weight terms

$$f_- \left(s_i^{[t,t+T_{\mathrm{stdp}})}, s_j^{[t,t+T_{\mathrm{stdp}})} \right) = \sum_{\Delta t \in \mathrm{NN}(t_i^k < t_j^k)} \eta_- \exp\left(-\frac{\Delta t}{\tau_-} \right), \tag{2.5}$$

with the time constant τ_-. The sum runs across all nearest-neighbor (NN) spike time differences $\Delta t = t_i^k - t_j^k$ for which the presynaptic spike time is larger than the postsynaptic. Sample measurements of f_- are plotted in figure 2.4. The factor λ_{stdp} characterizes the strength of STDP and is chosen to be negative. This way anti-causal spiking of pre- and postsynaptic neurons (the presynaptic neuron fires after the postsynaptic neuron) decreases the synaptic weights.

A constant drift term $\lambda_{\mathrm{w}} < 0$ pushes the synaptic weights further down.

The offset b_{ij} is drawn from a uniform random distribution

$$b_{ij} \sim \mathrm{unif}(-b_{\mathrm{amp}}, b_{\mathrm{amp}}) + \langle b \rangle, \tag{2.6}$$

with a positive bias $\langle b \rangle > 0$ and amplitude b_{amp}. The bias is chosen such that the network can develop a balanced state and over time reach constant weights. It also enables non-zero fixed points even for zero initialized weights.

Structural Plasticity

Structural plasticity describes the change of the internal network structure defined by the connections between neurons over time. The structural plasticity rule implemented in this thesis allows the network to choose from different inputs during an experiment by adapting the synapse decoder addresses based on causal spike-time correlation.
With a period of $T_{\text{struc}} = 500 \cdot T_{\text{stdp}}$ the PPU prunes the synapse addresses $a_{ij}(t)$:

$$a_{ij}(t + T_{\text{struc}}) = \begin{cases} a_{ij}(t), & \text{if } f_+ \left(s_i^{[t,t+T_{\text{stdp}})}, s_j^{[t,t+T_{\text{stdp}})} \right) \geq c_{ij} \\ d_{ij}, & \text{otherwise.} \end{cases} \tag{2.7}$$

The causal correlation f_+ is defined analogously to f_- (equation 2.5), but the sum runs over all spike pairs where the presynaptic spike occurs before the postsynaptic one. Only if f_+ is greater than or equal to a random value c_{ij} the respective address is kept. The reference values is drawn from an uniform distribution that is capped by the pruning threshold c_{th}:

$$c_{ij} \sim \text{unif}(0, c_{\text{th}}). \tag{2.8}$$

A value of $c_{\text{th}} = 7\,\text{bit}$ results in maximal pruning of synapses, since the 8 bit correlation measurements are shifted by one bit. If a correlation measurement is lower than c_{ij}, the address is reset to d_{ij}, a randomly selected addresses on which the network receives external input.
Random values are generated directly on the PPU using a Xorshift random number generator [Marsaglia 2003].

2.3.3 Network Initialization

The network model parameters used in the experiments in this thesis and their respective uncertainties are listed in table 5.1 in the appendix. Since STDP is used to stabilize the network no calibration is employed. In accordance with Dale's principle [Dale 1934], a number of N_{inh} synapses per neuron is randomly set to be inhibitory while the rest is excitatory.

The coding address system is used to configure a recurrent network of varying degree. All spikes originating from the neurons themselves are routed back to the synapse drivers encoded with a fixed address (see red dashed lines in figure 2.3). The number of synapses per neuron that transmit spikes with addresses other than this internal one is denoted by k_{in}. The choice of k_{in} decides the degree of network recurrence. A value of $k_{\text{in}} = 0$ results in a highly recurrent all-to-all connected network completely decoupled from external input, while $k_{\text{in}} = 32$ yields a pure feedforward network.

For an experiment, the network is initialized with a certain k_{in}. The recurrent synapses are not subjected to structural plasticity and therefore keep the network at a constant degree of recurrence. All weights are initiated with a value of 0 LSB and decoder addresses of the synapses sensitive to external input are chosen at random with a uniform distribution.
A full network stimulus contains 128 input channel, four spike trains for each of the 32 inputs. These are encoded with the addresses 0 to 3. Spikes of each address are sent into the network one after another in a rotating fashion.
While the network is stimulated with such external spikes, the synapses are subjected to STDP and structural plasticity allowing the network to adapt and select from the differently encoded spikes. A total of 100 structural plasticity updates are performed leading to an experiment duration of $100 \cdot T_{\text{struc}}$. After each pruning step the synapse array that holds the weights and decoder addresses of all synapses is saved.

Each experiment is conducted 10 times with different statistical seeds for the involved random processes. Results from the network analysis are averaged over these 10 independent trials, the standard deviation provides corresponding errors.

2.4 Stimuli

Two types of network stimuli are used. A dataset of spikes generated from recordings of spoken digits is used as an auditory stimulus. The creation of these spikes is based on biological models of the human auditory system. The dataset is briefly described in section 2.4.3.

The second type of stimuli are created with artificial correlation patterns that aim to reflect the correlations found in auditory stimuli. They are used to test the performance of the plasticity rules.
The spike generation is based on Poisson statistics. This is inspired by the spike creation in the auditory models [Boer 1980; Meddis 1986]. Two different correlation patterns between the auditory spikes can be distinguished. The spatially correlated stimulus captures the correlation between the spike times of different input channels and is described in section 2.4.1. The temporally correlated stimulus shows correlation between the spike times in a single spike train (section 2.4.2). The parameters characterizing each stimulus are listed in table 5.1 in the appendix.

2.4.1 Spatially Correlated Stimulus

The spatially correlated stimulus exhibits a correlation between the different input channels that is constant in time. The probability for spikes in all channels with the same

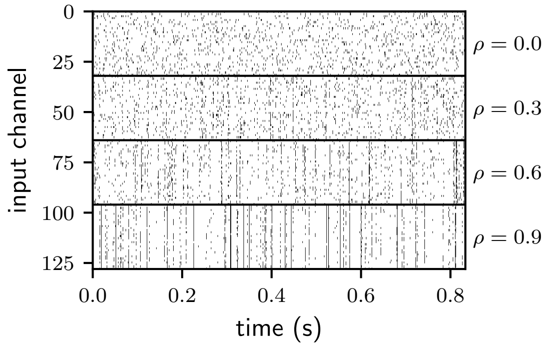


Figure 2.5: Spikes of a spatially correlated sample stimulus with mean firing rate $f = 13.4\,\mathrm{Hz}$. Black dots represent spikes in a given channel at a given time. The four stimuli of different correlation strength get sent into the network with different decoder addresses. With increasing ρ spikes tend to fire more synchronized.

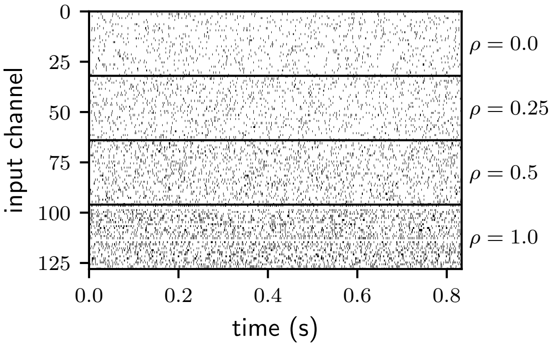


Figure 2.6: Spikes of a temporally correlated sample stimulus with an amplitude scaling of $A = 10\ \theta$. Black dots represent spikes in a given channel at a given time. The four stimuli with different amplitudes of the sine modulation get sent into the network with different decoder addresses.

correlation are calculated from the cumulative distribution function (CDF) of a normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ with mean $\mu = 0$ and standard deviation σ given by the correlation parameter ρ. This parameter determines the correlation strength of these channels. A spike is sent if the value is smaller than the acceptance probability:

$$\mathrm{CDF}(X) < p. \tag{2.9}$$

As a result the stimulus has a fixed correlation between spikes of all input channels at each point in time.
In different experiments the frequency f of the stimulus is varied. This is achieved by changing the probability p with which spikes are accepted. A higher probability leads to more spikes and therefore a higher frequency.

A full network stimulus consists of four spike masks with different correlation strengths, that are sent into the network with the four different encoder addresses 0 to 3. An equal spacing of $\rho \in \{0.0, 0.3, 0.6, 0.9\}$ is chosen. Figure 2.5 shows an excerpt of the spikes in a sample stimulus created with the actual model parameters used in the experiments.

2.4.2 Temporally Correlated Stimulus

The temporally correlated stimulus shows a temporal correlation separately for each input channel. The spike probability is modulated with a sine function with fixed offset θ. The frequency ν of the sine is chosen to be around one tenth of the inverse neuron refractory period and varied between input channels with a normally distributed jitter with amplitude $\Delta\nu = 0.1\ \nu$. In addition, a constant normally distributed phase shift φ is introduced

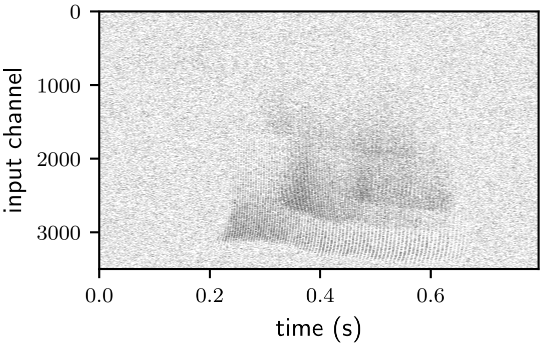


Figure 2.7: Full spike trains of a single spoken digit in the used dataset. Black dots represent spikes in a given channel at a given time. The different inputs are created by HCs located along the BM. High spike rates are a result of large membrane resonances.

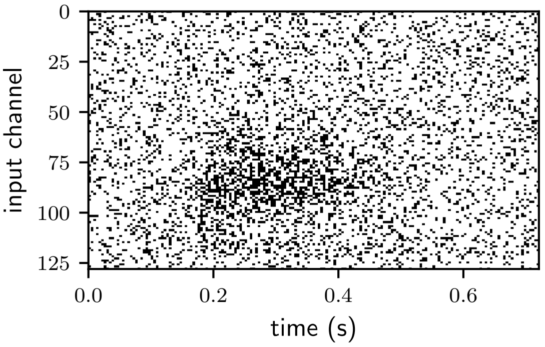


Figure 2.8: Down sampled spike trains of a spoken digit from the used dataset as they are sent into the hardware chip. Spikes are re-binned into spike routing time steps. Black dots represent spikes in a given channel at a given time.

for every input channel. Together, this avoids synchronous input on all channels and therefore reduces spatial correlation.

The strength of correlation is changed by scaling the amplitude of the sine with a factor ρ. When the amplitude of the sine becomes to high, probabilities could become negative. To avoid this, stimuli are cut off at zero. For high amplitudes the modulation pattern therefore is not actually a sine but shows sinusoidal peaks at constant intervals.

A spike is sent at time step i, if a random value $X \sim \text{unif}[0, 1)$ is smaller than the modulation probability:

$$X < \max \{0, A \cdot \rho \cdot \sin (2\pi \cdot (\nu + \Delta\nu) \cdot (i + \varphi)) + \theta\} . \tag{2.10}$$

Here, A is an additional scaling of the sine amplitude. It changes the number of spikes in the maxima of the sine to influence the firing rate of the stimulus.

A full stimulus is composed of four differently correlated spike masks with $\rho \in \{0.0, 0.25, 0.5, 1.0\}$. A section of a sample stimulus is shown in figure 2.6.

2.4.3 Auditory Stimulus

As an example for actual auditory stimuli a dataset of spikes generated from recordings of spoken digits is used [Cramer et al. in prep.(b)]. The dataset uses biologically inspired basilar membrane (BM) and hair cell (HC) models to generate neural spike data from audio files. A more detailed description of the models can be found in [Cramer 2016]. In the following a very brief overview of the involved steps is given.

Sound waves cause variations in pressure in the inner ear, leading to oscillations of the BM with perpendicular velocity $v_\perp$. The elasticity of the BM results in a spatial frequency separation [Boer 1980]. Different stimulus frequencies correspond to different resonance regions. The resulting translation of frequencies to regions of the membrane can be quantified by the equivalent rectangular bandwidth (ERB) concept [B. C. J. Moore 2003] which models each frequency selective segment of the BM by a rectangular filter.

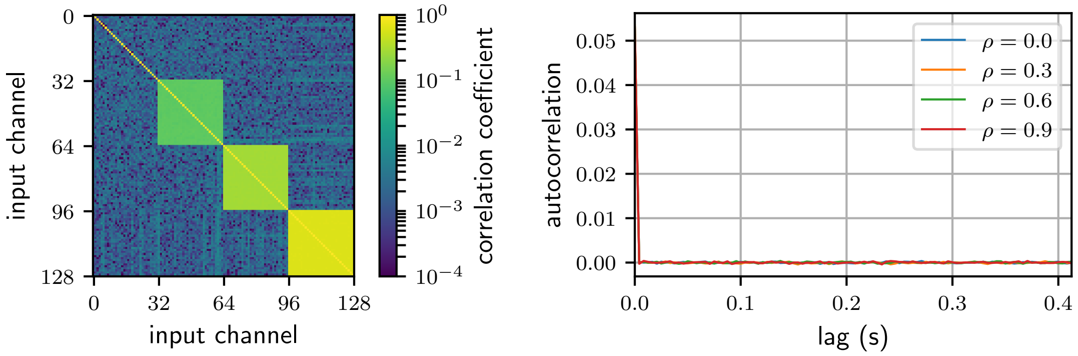
The HCs are placed along the BM. They translate the membrane oscillations into neural spikes with a transmitter based model. The propagation of transmitter molecules is modelled by differential equations allowing for a complete mathematical description [Meddis 1986; Meddis 1988]. Importantly, the transmitters are released through a permeable membrane into the cells synaptic cleft. While the membrane permeability depends on the momentary perpendicular velocity $v_\perp$ of the BM, the cells spiking probability is assumed to be linearly related to the amount of transmitters in the cleft. Thus, strong membrane oscillations generate high spike rates. Correlation between different spike channels results from the similar behavior of the BM at neighboring HCs.

The dataset is based on sound recordings of digits from “zero” to “nine” spoken by 12 different speakers. The model uses 3500 HCs linearly spaced along the BM in a region corresponding to frequencies from 1 Hz to 20 kHz.
The spikes of all 3500 channels for a single digits are shown in figure 2.7. It shows a region of high spike activity in certain spike channels. This is a result of the sensitivity of the BM to different frequencies of the sound waves. The uncorrelated background spiking in the beginning and end of the stimulus is due to thermal movement of the hairs in the HCs. This effect is also the reason for the offset θ in the temporally correlated stimulus. Because the effect is independent of the input sound waves, θ is also chosen to be constant for all correlation strengths.

For the purpose of this thesis the 3500 spike channels are downsampled to 128 inputs by simply choosing every 27th channel. These are split into four groups by the order of their position on the BM that are encoded with different addresses. The spikes are re-binned into spike routing time steps. Spike trains of multiple digits by different speakers are sent one after another in a random order for the duration of the experiment T_{exp}. The down sampled input spikes of a single digit are shown in figure 2.8.

2.4.4 Stimulus Characterization

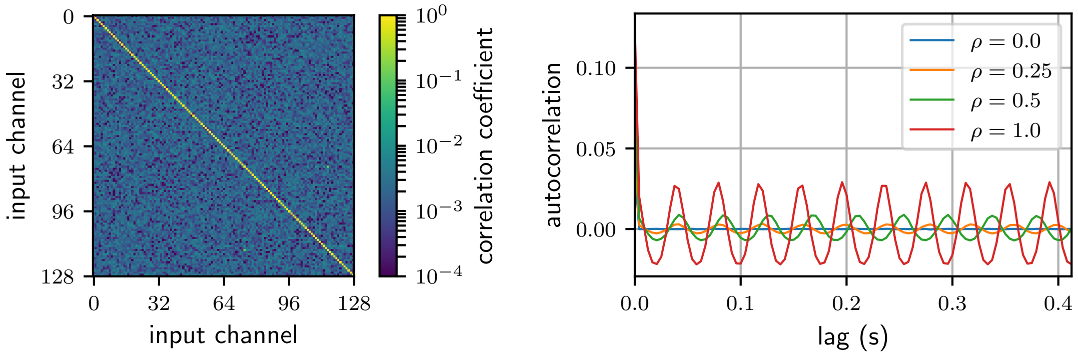
To demonstrate the correlation features of the stimuli the correlation coefficient matrices (Pearson product-moment correlation coefficients [Pearson et al. 1895]) between all input channels and the auto-correlation of individual channels are calculated.
The discrete correlation function $C_{xy}(k)$ of lag k between two spike trains x and y of length



(a) Cross-correlation matrix

(b) Autocorrelation

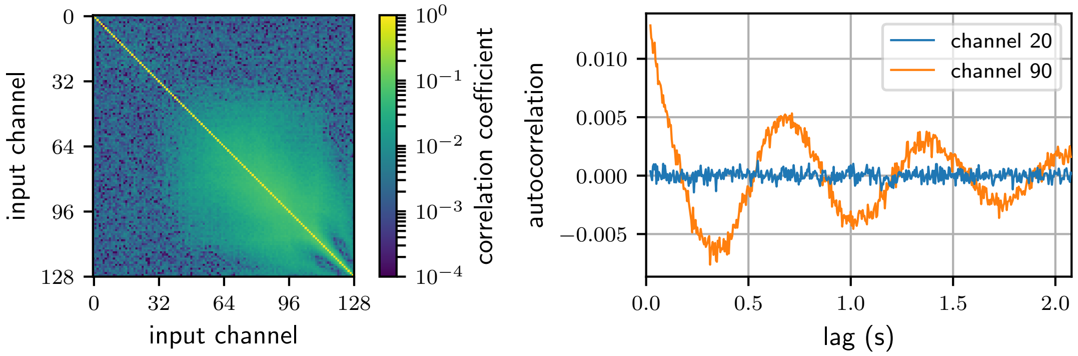
Figure 2.9: Correlation measures for the spatially correlated stimulus. The stimulus shows constant cross-correlation between neighboring channels of same correlation strength that increases with ρ but no autocorrelation within single channels.



(a) Cross-correlation matrix

(b) Autocorrelation

Figure 2.10: Correlation measures for the temporally correlated stimulus. The stimulus shows autocorrelation within each channels but no cross-correlation between channels.



(a) Cross-correlation matrix

(b) Autocorrelation

Figure 2.11: Correlation measures for the auditory stimulus. Both types of correlation are observed in certain channels.

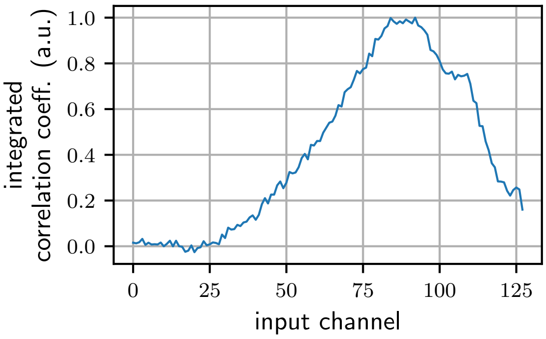


Figure 2.12: Integrated cross-correlation matrix for all channels of the auditory stimulus. The coefficients of each channel with all other channels are summed up.

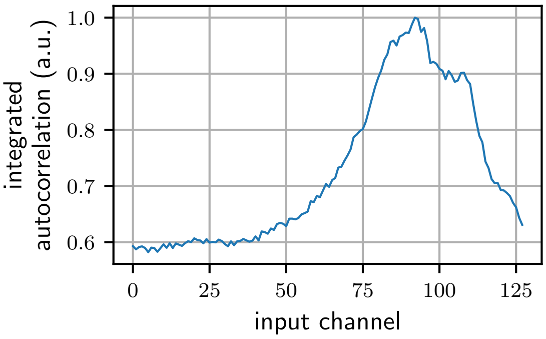


Figure 2.13: Integrated modulus of the autocorrelation function for each input channel of the auditory stimulus. The offset is due to the use of the modulus.

n with means μ_x and μ_y is defined by

$$C_{xy}(k) = \frac{1}{n}\sum_{i=1}^{n}(x_{i+k} - \mu_x) \cdot (y_i - \mu_y). \tag{2.11}$$

The spike trains are zero padded at the ends such that they always overlap. The autocorrelation function of x is defined as $C_{xx}(k)$.

Figures 2.9 through 2.11 visualize these measures for the three different stimuli.
The spatial stimulus shows cross-correlation between the different input channels dependent on the correlation parameter ρ. This becomes clear from the correlation coefficient matrix depicted in figure 2.9a. Groups of 32 neighboring channels exhibit roughly constant cross-correlation. The coefficients indicate a stronger correlation between channels with higher correlation parameter ρ. Figure 2.9b shows the autocorrelation of single spike trains with different spatial correlation. For any finite lag the value immediately drops to zero, the stimulus does not show any autocorrelation.
The temporally correlated stimulus in turn shows no cross-correlation. The off-diagonal elements of the correlation matrix are low for all temporal correlation parameters (figure 2.10a). However, the autocorrelation has periodic oscillations that increase in amplitude for higher ρ (figure 2.10b).
To calculate the correlation measures of the auditory stimulus a series of spike trains of multiple digits is used. Both types of correlation are observed. The off-diagonal coefficients in figure 2.11a show increased correlation for certain intermediate channels. In figure 2.11b the autocorrelations for two exemplary channels with low and high correlation are plotted.

To summarize the correlation per channel the integrated cross-correlation (sum of rows of the correlation matrix) is plotted in figure 2.12. The same is done for the autocorrelation by integrating the modulus of the autocorrelation over all lags (figure 2.13). Both distributions show a broad peak at higher input channels corresponding to lower frequencies in the stimulus.

2.5 Information Theoretic Measures

In the auditory stimulus the input channels exhibit differently pronounced correlation (section 2.4.4). However, for future processing of the digits it is also relevant how informative each channel is on the spoken digits. This is analyzed with measures from information theory. The mutual information (MI) between the spikes of each input channel and the labels of the spoken digits quantifies the amount of information each channel provides on the digits.

In the following a short introduction to the information theoretic measures used for evaluation is given. General definitions can be obtained from [Cover et al. 2006]. More details on the use of information theoretic measures for analyzing neural systems can be found in [Wibral et al. 2015].

Spikes of a single spike train are considered to be distributed according to a stationary random distribution. By sampling the spike patterns over time, their probability distribution can be constructed from the unique counts in the sampling. Such distributions can be analyzed with standard measures from information theory.

Subsequently, a stationary random process $\mathtt{X}$ of length n composed of random variables X_i, $i \in [1, n]$ is assumed. Realizations of such a process are denoted with lowercase letters x_i and the set of all possible outcomes is given by $\mathcal{X}_i$. The probability of a specific outcome is indicated by $p(x_i)$.

The information entropy H of a discrete random variable X quantifies the average amount of information contained in X. Its value can be calculated by

$$H(X) = \sum_{x \in \mathcal{X}} p(x) \log_2 \left(\frac{1}{p(x)} \right), \tag{2.12}$$

where the use of $\log_2$ generates results in binary digits (bits).

If the outcome of X is already known, the conditional entropy $H(Y|X)$ describes the additional information gained from observing a random variable Y. It is calculated by the weighted sum of the entropy $H(Y|X = x)$ of X taking a specific value x, over all possible values for x

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x)\, H(Y|X = x), \tag{2.13}$$

and can be rewritten to

$$H(Y|X) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log_2 \left(\frac{p(x)}{p(x,y)} \right), \tag{2.14}$$

with $p(x,y)$ being the joint probability distribution. It can also be understood as the amount of information contained in Y, that is not accessible from X directly.

These measures allow to define the mutual information (MI). The MI specifies the amount of information directly shared between two random variables

$$I(X;Y) = H(X) - H(X|Y). \tag{2.15}$$

It quantifies the information contained in X minus the additional information contained in X, provided knowledge of Y, leaving the information contained directly in both X and Y without knowledge of the respective other. Inserting the above yields

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in \mathcal{X}} p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)\, p(y)} \right). \tag{2.16}$$

This formula also shows that $I(X;Y)$ is symmetric in X and Y. The MI is calculated using the PyInform Python module, a wrapper for the Inform C library [D. G. Moore et al. 2017].

3 Results

The artificial stimuli are used to establish a baseline for the network behavior and the performance of the structural plasticity rule under differently correlated inputs. The results are given in section 3.1. The data obtained from stimulation with actual auditory stimuli is presented and analysed in section 3.2.

3.1 Artificial Stimuli

Experiments are conducted for different network recurrence degrees k_{in} and different spike rates. In the following, a pair of these parameters is referred to as a configuration. Configurations are swept to demonstrate different network characteristics.

3.1.1 Address Development

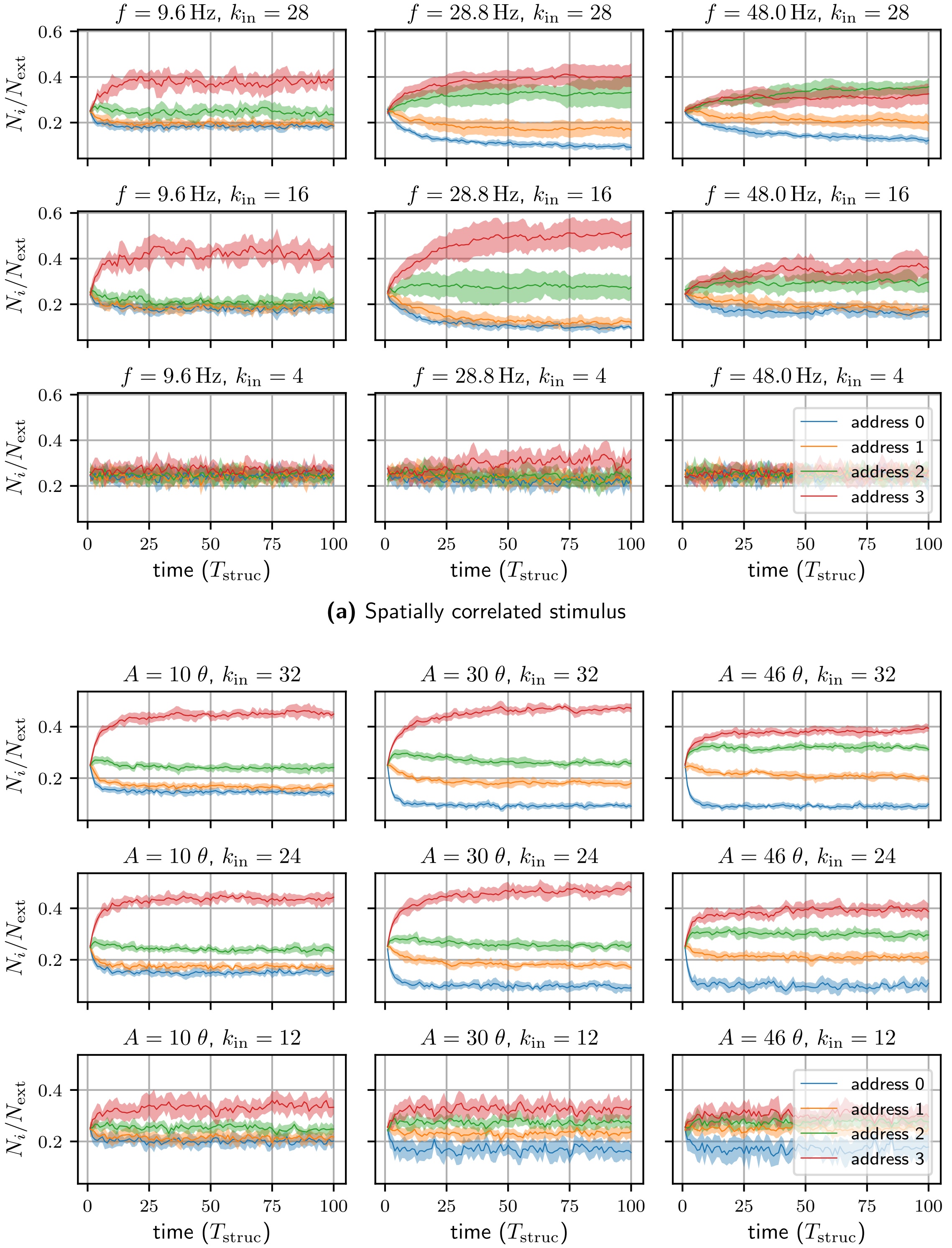
For the artificially correlated stimuli a simple evaluation of the network response can be performed by examining the addresses chosen by the synapses. If synapses show a preference for a certain decoder address after the network was stimulated and subjected to the plasticity rules, it can be interpreted as a preference of the network for the correlation strength sent with the corresponding address. This can be quantified by a preference measure for each address.

The number of synapses transmitting spikes with an external spike address i is denoted by N_i. The preference P_i for address i is defined as the share of N_i in all synapses that are sensitive to external input $N_{\mathrm{ext}} = k_{\mathrm{in}} \cdot N$:

$$P_i = \frac{N_i}{N_{\mathrm{ext}}}. \tag{3.1}$$

Since the addresses are pruned in every structural plasticity update, P_i changes over the course of an experiment.

The time development of this preference measure throughout the experiment is plotted in figure 3.1a for the spatially correlated stimulus. The figure shows sample configurations which highlight different characteristics observed in the whole sweep of configurations.



(a) Spatially correlated stimulus

(b) Temporally correlated stimulus

Figure 3.1: Time development of the address preference. The configurations shown capture the characteristic difference in observed network behavior. The presented data was recorded with a pruning threshold of 6 bit.

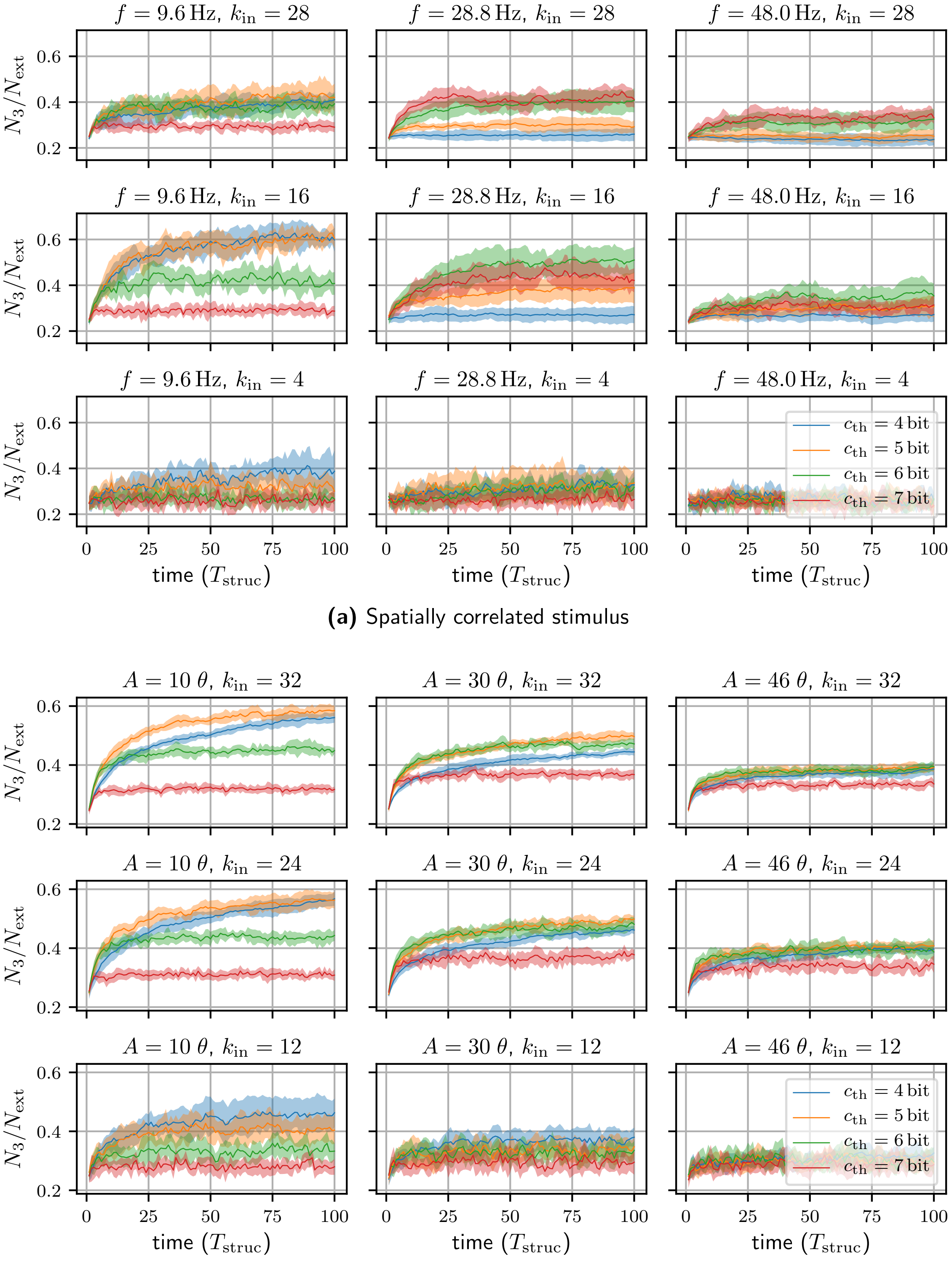
At very high recurrences the network does not show a preference for any of the stimuli. For all addresses, the share fluctuates around an equal, constant value.
Towards lower recurrences corresponding to overall higher coupling to the external input, the share of addresses splits up over the course of the experiment. The network starts to show a preference for address 3 which receives input with the highest correlation strength. This effect is particularly strong at intermediate frequencies and network recurrences. The split between addresses shows an asymptotic convergence to a maximal value.

The same graphs are shown for the temporally correlated stimulus in figure 3.1b. Again, multiple configurations are used to capture the qualitative behavior. The spike rate is controlled by scaling the amplitude A of the temporal modulation. Even though this also increases the correlation between the inputs of different addresses, the relative amplitude between channels ρ is kept constant for the whole sweep of A.
The overall development of the address share over time is comparable to the spatial stimulus. At high network recurrences only a small split between addresses is visible, which changes at lower recurrences with stronger coupling to the external input. Generally, the split seems to happen slightly faster for the temporal stimulus. The data is also less noisy as evidenced by the smaller errors.

3.1.2 Impact of Pruning Threshold

To analyse the impact of the pruning threshold c_{th} (section 2.3.2) on the address preference data is recorded for multiple thresholds in each configuration. For clearer presentation only the address with highest correlation is considered. It can be seen from figure 3.1 that this address alone is a good indicator of the networks ability to choose from the correlated input. Figure 3.2 therefore shows the development of the share of synapses with decoder address 3 for different pruning thresholds. Again, configurations are shown that represent the observed characteristics.
The pruning threshold shifts the frequency where the largest split happens. While for high spike frequencies around $f = 30\,\mathrm{Hz}$ the highest split is achieved with $c_{\mathrm{th}} = 7\,\mathrm{bit}$, for lower rates a threshold of $c_{\mathrm{th}} = 5\,\mathrm{bit}$ shows the best split performance. Values lower than $c_{\mathrm{th}} = 4\,\mathrm{bit}$ were not recorded because addresses would start to jump randomly in every plasticity update (also see section 3.1.4).
Similar behavior is observed for the temporally correlated stimulus (see the change of blue and green graphs in figure 3.2b). Here, a threshold of $c_{\mathrm{th}} = 7\,\mathrm{bit}$ results in a low ability to couple to the highest correlated input across all observed spike rates.

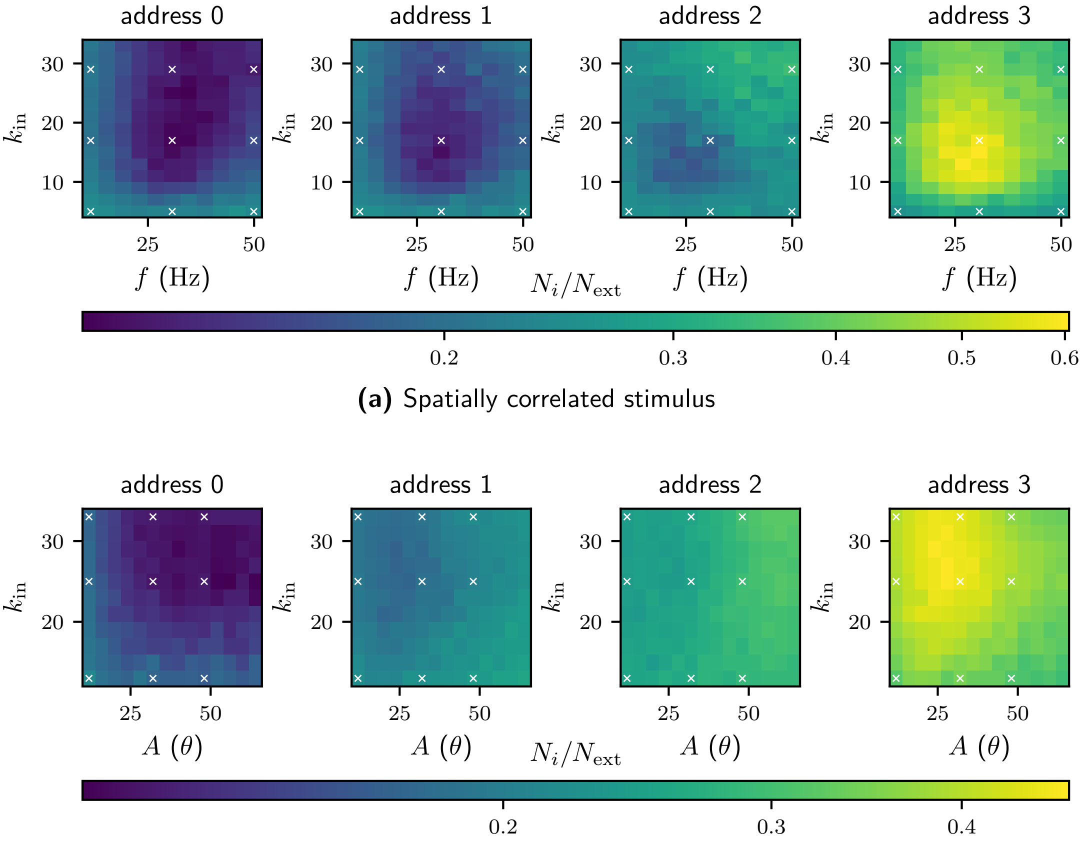
The connection between the pruning threshold and the frequency of the highest preference split can be explained by considering the correlation traces used in the pruning rule. At lower spike frequencies there is generally less correlation per time in the stimulus which



(a) Spatially correlated stimulus

(b) Temporally correlated stimulus

Figure 3.2: Time development of the preference for the highest correlated input compared between multiple pruning thresholds for multiple configurations.



(a) Spatially correlated stimulus

(b) Temporally correlated stimulus

Figure 3.3: Address preference averaged over the last 10 structural plasticity updates swept for different network configurations. The synapse pruning threshold is $c_{\mathrm{th}} = 6\,\mathrm{bit}$. The crosses mark the configurations for which the time development is shown in the previous figures.

results in lower correlation traces. Thus, when keeping the threshold fixed, more synapses are pruned than at higher frequencies. Reducing the pruning threshold counteracts this effect, the largest split happens at lower frequencies for lower thresholds.

3.1.3 Configuration Sweep

To systematically analyze the behavior for different configurations the networks preference at the end of an experiment is summarized by averaging the address share over the last 10 structural plasticity updates. This is calculated for each of the four addresses and plotted in a color mesh plot in figure 3.3. A pruning threshold of $c_{\mathrm{th}} = 6\,\mathrm{bit}$ was used, which generally provided good performance at intermediate frequencies. The configurations for which the time development was shown in the previous figures are marked with a cross. Network configurations are swept for the spatial stimulus from the lowest possible recurrence $k_{\mathrm{in}} = 32$ to a maximum recurrence of $k_{\mathrm{in}} = 6$ in steps of 2. For higher recurrences

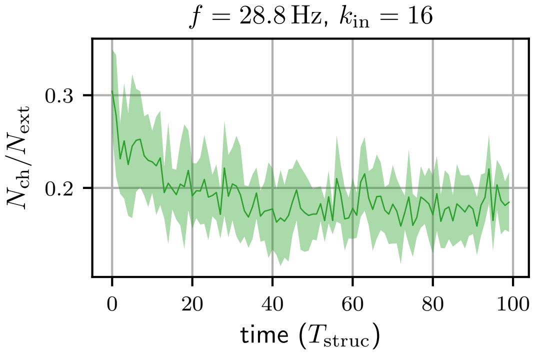


Figure 3.4: Time development of the number of changing decoder addresses N_{ch} in a structural plasticity update. A single configuration of the spatial stimulus with frequency f and recurrence k_{in} is shown. The pruning threshold is 6 bit.

the network did not become stable. The plot shows a broad maximum of preference for the highly correlated address at intermediate recurrences and spike rates (figure 3.3a).
For the temporally correlated stimulus such a sweep is shown in figure 3.3b. Here, network recurrences are swept from $k_{\mathrm{in}} = 14$, much lower values did not produce stable results. The plot also shows an area of high address share split, but it is generally located at higher recurrences towards the feedforward network.

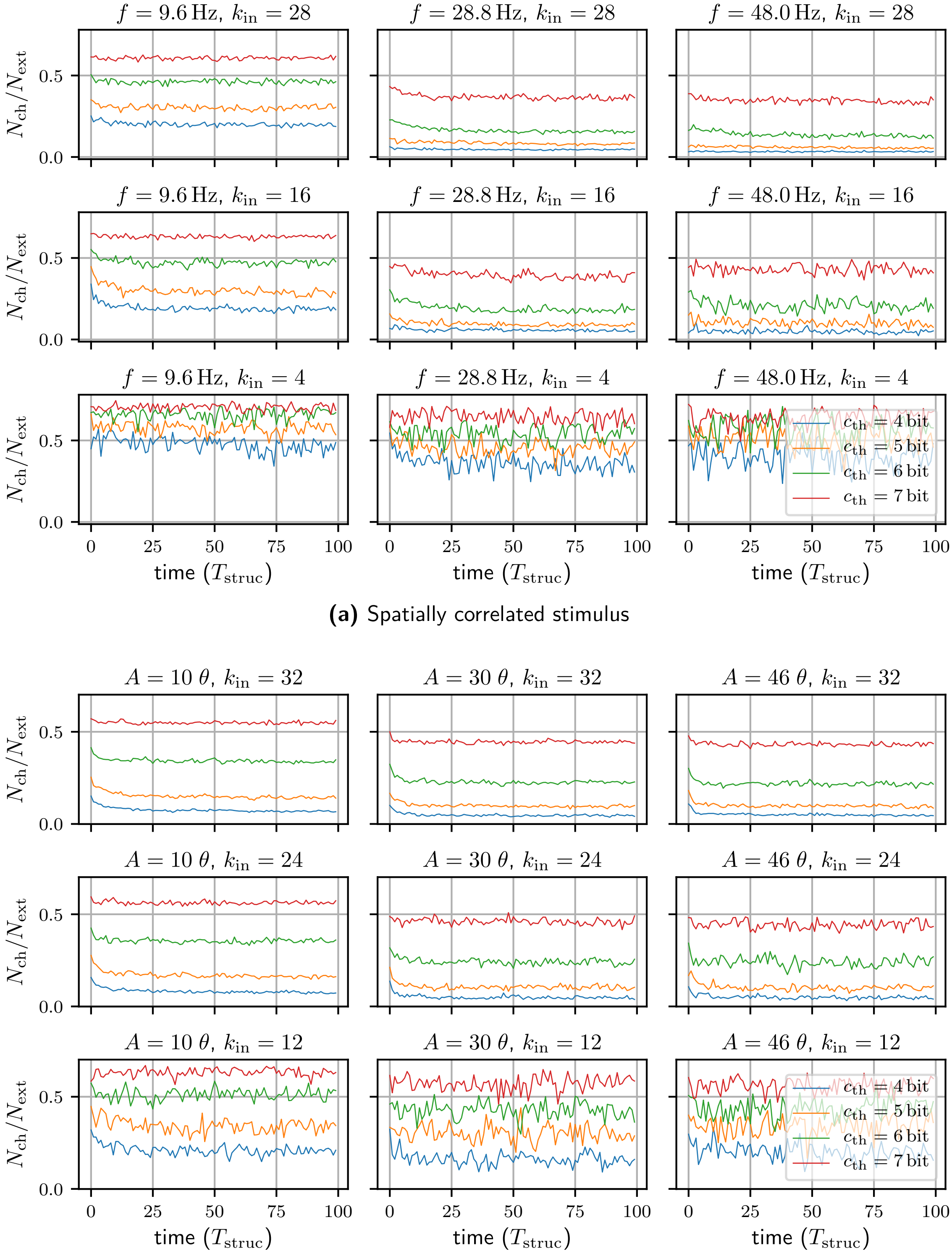
This shows that there are ranges of parameters for which the network is able to select spatially respectively temporally correlated input.

3.1.4 Address Stability

Another measure that characterizes the development of the network connectivity under a given plasticity rule is the overall stability of synaptic connections.

The fraction $N_{\mathrm{ch}}/N_{\mathrm{ext}}$ of external decoder addresses that change in a structural plasticity update is plotted in 3.4 for a single configuration of the spatially correlated stimulus. It demonstrates an example of how this ratio develops over time. A comparison between different configurations and pruning thresholds is depicted in figure 3.5. Error visualization is omitted for cleaner presentation. The standard deviation of all configurations is comparable to that shown in figure 3.4.

Configurations with low recurrences that show no split between addresses in figure 3.1 generally have unstable addresses that change frequently during the whole experiment. These are also more noisy, as visible from the jumpy lines.
Configurations that have a visible split exhibit significantly less address changes. Here, the number of jumps drops in the first few structural plasticity updates and then stays at a constant value (see 3.4). The pruning threshold impacts the timescale of this stabilization



(a) Spatially correlated stimulus

(b) Temporally correlated stimulus

Figure 3.5: Time development of the decoder address stability for different configurations. The number of addresses changed after each structural plasticity update is compared between multiple pruning thresholds. Data is averaged over all recorded trials, errors are omitted for cleaner visualization.

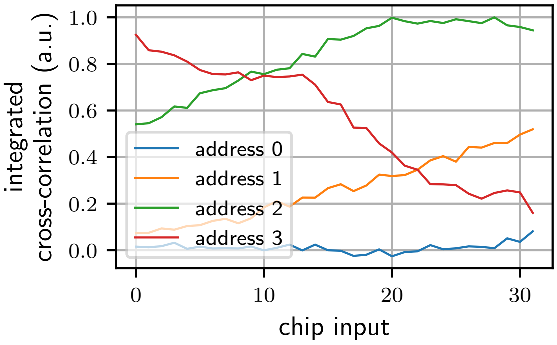


Figure 3.6: Correlation strength of the spike masks of the auditory stimulus. The integrated cross-correlation measure from section 2.4.4 is used. A similar assignment of correlation to addresses results from the integrated autocorrelation. The auditory channels are distributed across the inputs by sending 32 neighboring channels on a fixed address. With this distribution each input receives both strongly and weakly correlated spikes, from which the synapses can choose by selecting the corresponding addresses.

since it controls the number of synapses that can be pruned. For a high threshold, the final number of jumps is reached after very few updates, while more updates are needed for a lower threshold. This is especially pronounced in the temporally correlated stimulus (e.g. configuration with $A = 30\ \theta,\ k_{\text{in}} = 32$ in figure 3.5b).

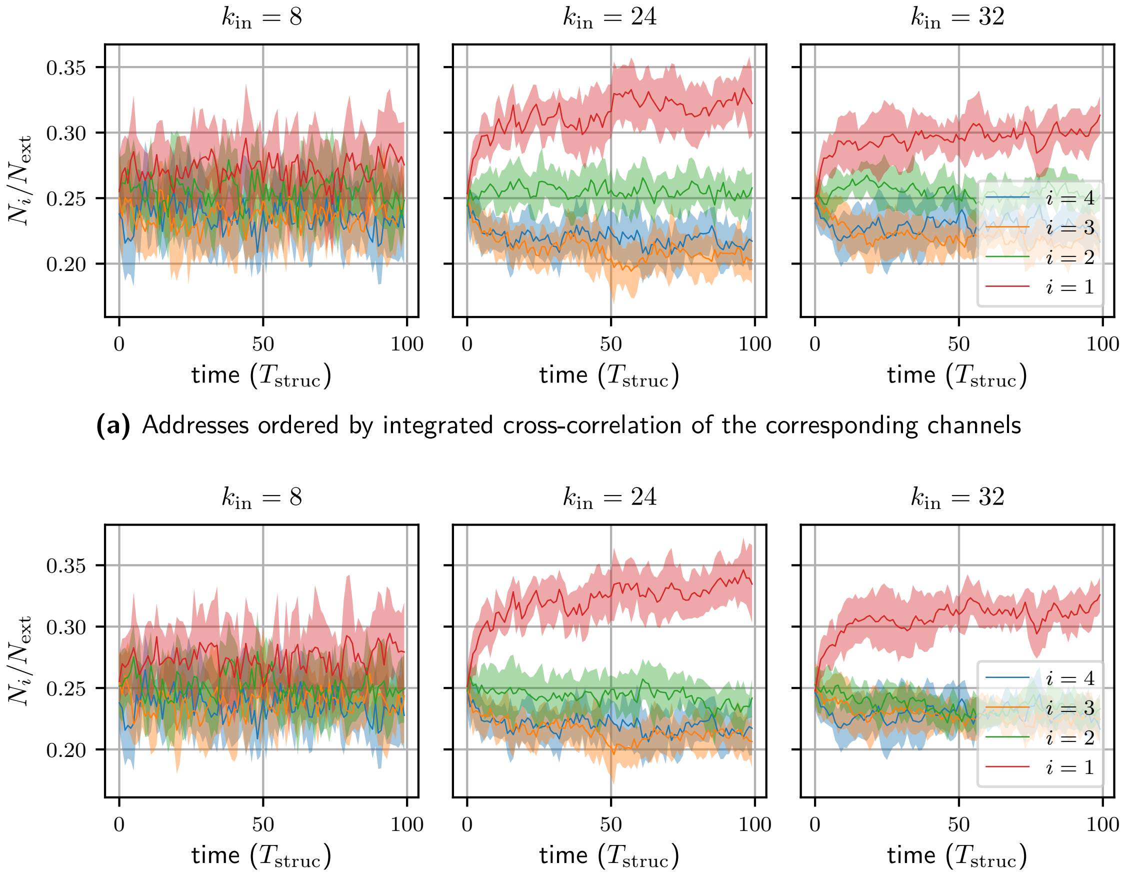
Naturally, the threshold also impacts the overall number of jumps across all configurations. A high threshold results in many address changes since synapses are pruned for a large range of correlation measurements. For $c_{\text{th}} = 7\,\text{bit}$, potentially all synapses could be pruned. Correspondingly, a low threshold results in less address jumps. This is however not necessarily wanted, because the network should maintain the ability to jump away from uncorrelated input.

3.2 Auditory Stimulus

The auditory stimulus has a fixed spike frequency. Because of the findings in section 3.1.2 the pruning threshold is not expected to have an impact on the network recurrence for which highest correlation preference is achieved. Therefore, a fixed value of $c_{\text{th}} = 6\,\text{bit}$ is used for all experiments with the auditory stimulus. This value generally provided good results for the artificial stimuli.

The 128 auditory channels of the stimulus have to be distributed across the chip's inputs and the available addresses. Because of the finite spike routing time, the number of addresses per input is limited (see discussion in chapter 4).

When sending 32 neighboring spike trains of the auditory stimulus on a fixed address at a time, the chip's 32 inputs each receive spikes of 4 different correlation strengths. As shown in figure 3.6, each input receives spikes with both high and low correlation. The addresses



(a) Addresses ordered by integrated cross-correlation of the corresponding channels

(b) Addresses ordered by integrated autocorrelation of the corresponding channels

Figure 3.7: Time development of the preference for correlated input of the auditory stimulus. The available addresses of each synapse are ordered by the correlation of the corresponding input channels. The graphs show the fraction of external synapses that are sensitive to the i-th highest correlation, i.e. $i = 1$ corresponds to synapses coupling to the strongest available correlation.

chosen by the synapses however loose their direct mapping to a fixed correlation strength. Yet, a correlation preference measure comparable to that of the artificial stimulus can be constructed.

For evaluation, the external addresses of every synapse are ordered according to the correlation strength of the corresponding spike trains they transmit. This allows to calculate the number N_i of synapses that chose the address with the i-th highest correlation. The fraction of these synapses in all external synapses is the preference for the i-th highest available correlation:

$$P_i = \frac{N_i}{N_{\text{ext}}}. \tag{3.2}$$

Figure 3.7a shows the development of this preference measure. Here, addresses are sorted with respect to the integrated cross-correlation of the corresponding input channels as determined in section 2.4.4. Figure 3.7b shows the same with addresses sorted by the integrated autocorrelation.

For both correlation patterns, configurations with different network recurrences are shown. There is no significant preference visible at high recurrence, especially, the distribution does not evolve over time.

Similarly to the artificial stimuli there is a large split between the correlation preference for intermediate recurrence. The network prefers such addresses that transmit input channels with high correlation. This is true for both temporal and spatial correlation. For the feedforward case, the split is slightly reduced, indicating less preference.

3.2.1 Input Channel Address Distribution

The performance of feature selection of relevant channels in the auditory stimulus is evaluated by analysing the total number of synapses realized for every input.

Each synapse driver receives input of four different channels that it transmits to 32 synapses. The external synapses can choose between these four input channels by selecting the respective addresses. The share of synapses with an address corresponding to each channel N_{ch} in all external synapses characterizes the total strength of coupling to the channels.

The time development of this distribution between channels after each structural plasticity update is plotted in figure 3.8. Channel numbers are mapped to corresponding position on the basilar membrane for a more general presentation.

The figure shows that the final distribution is reached relatively fast after the random initialization with an equal share in addresses. This allows to sample the distribution not only from the multiple trials, but also over time. Such a distribution of the average over the last 50 structural plasticity updates is plotted in figure 3.9 for multiple k_{in}.

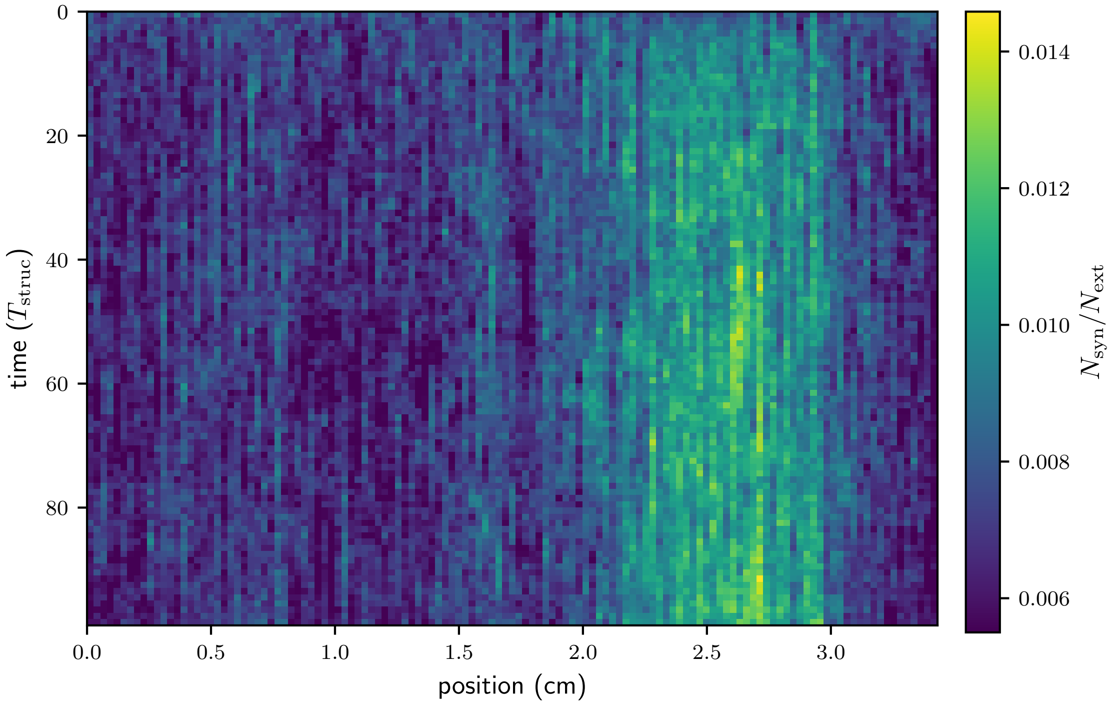


Figure 3.8: Time development of the share of synapses sensitive to each input channel of the auditory stimulus. Input channels are mapped to corresponding positions of the HCs on the BM. Data for a network recurrence of $k_{\mathrm{in}} = 24$ and a pruning threshold $c_{\mathrm{th}} = 6\,\mathrm{bit}$ is shown. The final distribution is reached relatively fast after a few structural plasticity updates.

The integrated correlation measures from section 2.4.4 are drawn for comparison. In addition, the MI between each input channel s_i (the stimulation spike trains) and the corresponding label l of the digits is plotted. This gives a measure for the amount of information on the digits that is contained in each channel.

All graphs show a broad peak at around 2.5 cm on the BM. There is an additional small peak at a lower BM position in the MI that is not present in the correlation measures. The address distribution is to noisy to make a clear statement about such a peak in the network's address choice. Thus, it can not be stated definitely whether the network only couples preferably to high correlation, or if other features of the stimulus, that are informative on the digits, are recognized.

The network recurrence degree k_{in} has a slight impact on the distinctiveness of the peak in the distribution. Towards high recurrence, it vanishes and overall noise increases as indicated by the larger errors.

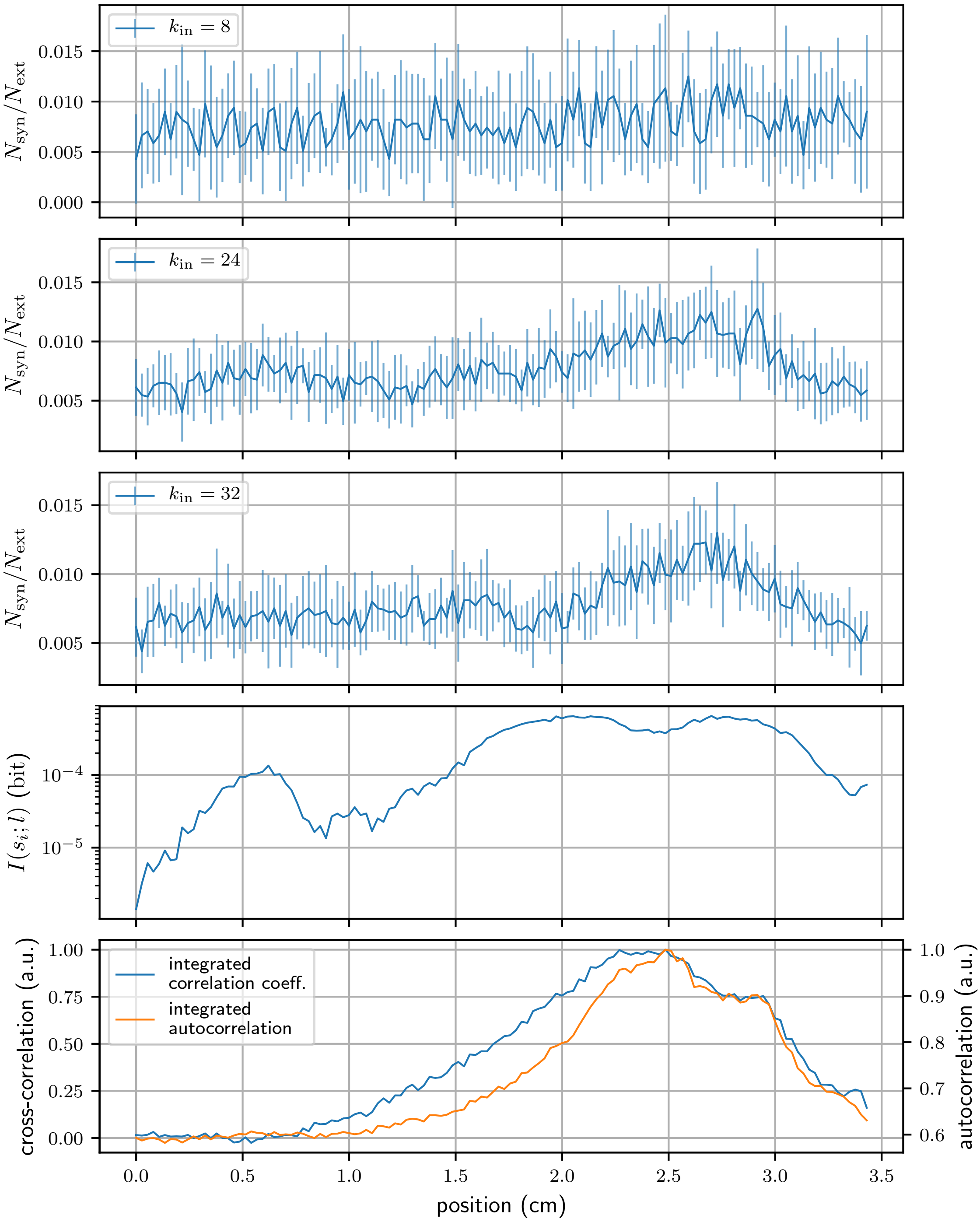


Figure 3.9: Distribution of synapses sensitive to each input channel of the auditory stimulus. Input channels are mapped to corresponding positions of the HCs on the BM. Data is sampled from the last 50 structural plasticity updates as well as being averaged over all recorded trials. For comparison the MI between input channels and digit label and the integrated correlation measures of the inputs are shown. All distributions show a peak at around $2.5\,\mathrm{cm}$.

4 Discussion and Outlook

In the experiments conducted as part of this thesis, feature selection with structural plasticity on the HICANN-DLS chip was successfully demonstrated. The chip received multiple spike trains on each of its inputs. The synaptic decoder address was used to select which of the spikes is transmitted by a synapse. The correlation based structural plasticity rule could select one of the available addresses in discrete pruning intervals. Performance was evaluated by analysing the fraction of realized decoder addresses that receive highly correlated input. The use of artificial stimuli allowed to establish a baseline for the networks capability to couple to correlated stimuli when subjecting synapse addresses to a correlation based pruning rule. The model's ability to couple stronger to input channels with higher correlation, both in time and between input channels was demonstrated.

The temporally correlated stimulus shows correlation of spike times in single input channels. This correlation can be measured directly with the correlation traces of the synaptic connections in the network when the postsynaptic neuron is activated to spike. Since these correlation measurements are used to adapt the decoder addresses of the synapses, the network quickly arrives at its final distribution of the share of addresses. In the spatially correlated stimulus the correlation is between different input channels. Spikes of multiple of these channels are routed into the same neuron. If these spikes are capable of invoking a postsynaptic spike, there will be correlation between pre- and postsynaptic spikes. Thus, high correlation of spike times of different input channels can lead to an increase in the correlation measurements of the synaptic connections involved. However, as multiple synapses are involved and neurons also receives spikes from uncorrelated channels, the capability of the network to select the correlated input is slower compared to the direct correlation in single channels of the temporally correlated stimulus (figure 3.1).

The threshold for pruning of synaptic decoder addresses allows to control the frequency for which optimal performance is achieved. Higher stimulation frequencies result in overall more activity per time and therefore higher correlation measurements for correlated input. To allow synapses to be pruned in such cases, a higher pruning threshold is required (figure 3.2a).
The pruning threshold also impacts the total amount of synapses that can be pruned in a plasticity update. Hence, it allows to tweak the responsiveness of the network to dynamic

input. Especially for the temporal stimulus the number of jumps stabilizes much faster for a higher pruning threshold, as the network is more responsive (figure 3.2a).
The actual value that is compared to a synapse's correlation measurement to decide if the synapse is pruned is chosen from a uniform random distribution. The pruning threshold just defines the maximum value of this distribution. Therefore, synapse pruning is to a large degree a random effect that introduces a form of noise in the address development. A clearer impact of the pruning threshold might be seen when comparing correlation measurements to a constant value globally selected for all synapses, which will have to be tested in future experiments. This would additionally reduce the number of random values that need to be calculated in each plasticity update and allow the pruning to happen more quickly.

By sweeping across a number of different network configurations it was shown that the optimal performance of correlation selection is achieved for a whole range of network recurrences and stimulation frequencies. The latter allows to use stimuli with different frequencies. Together with the knowledge gained above on the relation of the frequency to the pruning threshold, the setup can be tuned to stimuli with a range of different time scales.
The freedom to choose the network recurrence and the stimulation frequency while maintaining the capability of feature selection allows to control the computational power of the network. This is of interest for the second task analysed in this thesis.

As a step towards speech recognition with the HICANN-DLS chip the structural plasticity rule was used for feature selection from several provided input channels of auditory stimuli. Such stimuli were shown to exhibit both types of correlation patterns analysed above (figure 2.12 and 2.13). This is due to the spatial frequency separation of the BM involved in spike creation.
Furthermore, correlation coincides with high information on the label of the digit spoken, as measurable by the MI between spike activity and labeling vector. To understand this, the statistical spike probability of the HCs in the auditory model has to be considered. Without external stimulation the hairs of the HC are subject to thermal fluctuations that invoke random spikes following a Poisson distribution which is uncorrelated. An external stimulus however extrudes the BM at regions corresponding to frequencies relevant to the stimulus. This increases spiking rate of HCs in these regions which makes the corresponding channels informative on the input. Such channels now show correlation in spike times because the spike distribution is no longer just Poissonian.
With the results from the previous experiments in mind this explains the capability of the setup to preferably select auditory input channels with high information content on the digits (figure 3.9).

A relevant aspect to consider is the distribution of input channels across the synapse drivers. For an arbitrary stimulus not every synapse necessarily has the same chance to couple to strongly correlated input channels and some synapses would have to couple to inputs with low correlation. For the specific distribution of auditory channels used in this thesis this is not a problem. Figure 3.6 shows that every chip input receives high and low correlation. In general, the problem could be circumvented by distributing channels redundantly across the chip's inputs. If each spike train is provided to multiple synapse drivers on different addresses every synapse could potentially select the correlated input. This would however reduce the number of different input channels that can be used.
Cochlea implants with eight channels allow humans with an hearing impairment to understand speech with reasonable accuracy, with more channels seeming to increase performance [Croghan et al. 2017]. While the computational power of the employed network with only 32 neurons is significantly reduced compared to the human brain that processes inputs from the cochlea implants, for the rather simple task of classifying ten different digits, a low number of channels may still lead to good performance and allow for higher redundancy. This will have to be analysed in future experiments with an LSM setup (see below).

One way to increase the number of different inputs, either to improve redundancy or to allow for a more biologically realistic setup with more HCs, would be to raise the number of addresses from which the synapses can choose. However, more spikes would need to be send into the network on a single input. This would increase the time interval in which the network receives spikes belonging to the same input channel, leading to a decrease of the stimulation frequency. Therefore, the correlation per time drops which decreases temporal resolution of the correlation trace measurements and potentially effects synaptic pruning. The magnitude of this effect and the maximal number of addresses per input channel that still allows for correlation preference remain to be analysed.
A countermeasure would be to decrease the time between two spikes sent into the network. While this stimulation bin width (see table 5.1) would optimally be chosen infinitely small, the chip's spike input frequency is limited by the FPGA's clock rate. The hardware design is however expected to allow a reduction of the bin width by a factor of two in future experiments, doubling the effective number of input channels.

In an LSM setup for classification tasks the recurrent network is considered as a “liquid” capable of processing the inputs in a non-linear way by projecting them into the high-dimensional liquid space [Maass et al. 2002; Maass et al. 2004]. The outputs of a well designed liquid are expected to be linearly separable so information can be extracted with a simple linear classifier. Thus, the performance of an LSM setup can be evaluated by the classification accuracy of the linear classifier.

In the experiments conducted here, the network recurrence degree k_{in} remains as a control parameter to tweak performance since no significant impact on feature selection was found for a range of recurrences. It has been shown that k_{in} can influence the networks distance to the critical point between phases of amplifying spiking activity and rapid activity decline [Cramer et al. in prep.(a)]. This has the potential to impact the network's computational power and therefore the performance of the setup in a classification task, which should be analysed in future experiments.

The use of a higher number of addresses per synapse could also impact the classification and requires further investigation. To gain better understanding of the mechanisms governing the network, more extensive tools from information theory could be used to analyse the information flow between inputs, outputs, and neurons.

5 Appendix

Parameter	Symbol	Value
Membrane capacitance	C_m	$(2.38 \pm 0.02)\,\mathrm{nF}$
Time constant	τ_mem	$(1.6 \pm 1.0)\,\mathrm{ms}$
Threshold potential	u_thresh	$(551 \pm 22)\,\mathrm{mV}$
Leak potential	u_leak	$(350 \pm 100)\,\mathrm{mV}$
Reset potential	u_reset	$(318 \pm 18)\,\mathrm{mV}$
Refractory period	τ_ref	$(4.87 \pm 0.45)\,\mathrm{ms}$
Synaptic time constant	$\tau_\mathrm{syn}^\mathrm{exc}$	$(3.70 \pm 0.49)\,\mathrm{ms}$
	$\tau_\mathrm{syn}^\mathrm{inh}$	$(2.81 \pm 0.35)\,\mathrm{ms}$
Synaptic delay	d_syn	$(1.9 \pm 0.1)\,\mathrm{ms}$
Weight conversion	s_w	$(8.96 \pm 0.13)\,\mathrm{\mu A}$
Causal correlation time constant	τ_+	$(7.3 \pm 0.7)\,\mathrm{ms}$
Anticausal correlation time constant	τ_-	$(6.8 \pm 1.2)\,\mathrm{ms}$
Causal correlation amplitude	η_+	0.068 ± 0.022
Anticausal correlation amplitude	η_-	0.071 ± 0.023
STDP correlation scaling	λ_stdp	$-1/64$
Drift parameter	λ_w	$1/128$
Range of random variable	b_amp	$15/16$
Bias of random variable	$\langle b \rangle$	$3/16$
Number of neurons	N	32
Number of inhibitory neurons	N_inh	6
Initial weight	w_{ij}^init	$0\,\mathrm{LSB}$
STDP update period	T_stdp	$1.09\,\mathrm{s}$
Structural plasticity update period	T_struc	$9\,\mathrm{min}$
Experiment duration	T_exp	$15\,\mathrm{h}$
Stimulation bin width	T_stim	$4.2\,\mathrm{ms}$
Temporal modulation frequency	ν	$24\,\mathrm{Hz}$
Temporal jitter amplitude	$\Delta\nu$	$2.4\,\mathrm{Hz}$
Stimulus offset	θ	$9.6\,\mathrm{Hz}$

Table 5.1: List of model parameters. Times are given in the biological equivalents. Errors indicate the standard deviation. Hardware model parameters are taken from [Cramer et al. in prep.(a)]

6 References

Aamir, S. A., P. Müller, A. Hartel, J. Schemmel, and K. Meier (2016). „A highly tunable 65-nm CMOS LIF neuron for a large scale neuromorphic system“. *ESSCIRC Conference 2016: 42nd European Solid-State Circuits Conference*, pp. 71–74. DOI: 10.1109/ESSCIRC.2016.7598245.

Aamir, S. A., Y. Stradmann, P. Müller, C. Pehle, A. Hartel, A. Grübl, J. Schemmel, and K. Meier (2018). „An Accelerated LIF Neuronal Network Array for a Large-Scale Mixed-Signal Neuromorphic Architecture“. *IEEE Transactions on Circuits and Systems I: Regular Papers* (99), pp. 4299–4312. DOI: 10.1109/TCSI.2018.2840718.

Beggs, J. and N. Timme (2012). „Being Critical of Criticality in the Brain“. *Frontiers in Physiology* 3, p. 163. DOI: 10.3389/fphys.2012.00163.

Bi, G. and M. Poo (1998). „Synaptic Modifications in Cultured Hippocampal Neurons: Dependence on Spike Timing, Synaptic Strength, and Postsynaptic Cell Type“. *Journal of Neuroscience* 18 (24), pp. 10464–10472. DOI: 10.1523/JNEUROSCI.18-24-10464.1998.

Boer, E. de (1980). „Auditory physics. Physical principles in hearing theory. I“. *Physics Reports* 62 (2), pp. 87–174. DOI: 10.1016/0370-1573(80)90100-3.

Butz, M., F. Wörgötter, and A. van Ooyen (2009). „Activity-dependent structural plasticity“. *Brain Research Reviews* 60 (2), pp. 287–305. DOI: 10.1016/j.brainresrev.2008.12.023.

Cessac, B. J. (2008). „A discrete time neural network model with spiking neurons. Rigorous results on the spontaneous dynamics“. *Journal of Mathematical Biology* 56 (3), pp. 311–345. DOI: 10.1007/s00285-007-0117-3.

Cover, T. M. and J. A. Thomas (2006). *Elements of Information Theory.* 2nd ed. New York: Wiley & Sons. ISBN: 978-0-471-24195-9.

Cramer, B. (2016). „Modelling of the plastic binaural auditory System and age-related Alterations“. MA thesis. Kirchhoff Institute of Physics.

Cramer, B., D. Stöckel, J. Schemmel, K. Meier, and V. Priesemann (in prep.[a]). „Control of criticality in self-stabilizing neuromorphic spiking networks“.

Cramer, B., Y. Stradmann, and F. Zenke (in prep.[b]). „Spiking benchmarks - On the systematic evaluation of spiking neural networks“.

Croghan, N. B. H., S. I. Duran, and Z. M. Smith (2017). „Re-examining the relationship between number of cochlear implant channels and maximal speech intelligibility“. *The Journal of the Acoustical Society of America* 142 (6), pp. 537–543. DOI: `10.1121/1.5016044`.

Dale, H. (1934). „Pharmacology and Nerve Endings“. *British medical journal* 2 (3859), pp. 1161–1163. DOI: `10.1136/bmj.2.3859.1161`.

Deger, M., A. Seeholzer, and W. Gerstner (2017). „Multicontact Co-operativity in Spike-Timing–Dependent Structural Plasticity Stabilizes Networks“. *Cerebral Cortex* 28 (4), pp. 1396–1415. DOI: `10.1093/cercor/bhx339`.

Friedmann, S., J. Schemmel, A. Grübl, A. Hartel, M. Hock, and K. Meier (2017). „Demonstrating Hybrid Learning in a Flexible Neuromorphic Hardware System“. *IEEE Transactions on Biomedical Circuits and Systems* 11 (1), pp. 128–142. DOI: `10.1109/TBCAS.2016.2579164`.

Gerstner, W., W. M. Kistler, R. Naud, and L. Paninski (2014). *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition.* Cambridge University Press. ISBN: 978-1-107-63519-7.

Hebb, D. (1949). *The Organization of Behavior.* New York: Wiley & Sons. ISBN: 978-0-805-84300-2.

Hughes, J. R. (1958). „Post-Tetanic Potentiation“. *Physiological Reviews* 38 (1), pp. 91–113. DOI: `10.1152/physrev.1958.38.1.91`.

Lamprecht, R. and J. LeDoux (2004). „Structural plasticity and memory“. *Nature Reviews Neuroscience* 5, pp. 45–54. DOI: `10.1038/nrn1301`.

Maass, W. (1997). „Networks of spiking neurons: The third generation of neural network models“. *Neural Networks* 10 (9), pp. 1659–1671. DOI: `10.1016/S0893-6080(97)00011-7`.

Maass, W. and H. Markram (2004). „On the computational power of circuits of spiking neurons“. *Journal of Computer and System Sciences* 69 (4), pp. 593–616. DOI: 10.1016/j.jcss.2004.04.001.

Maass, W., T. Natschläger, and H. Markram (2002). „Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations“. *Neural Computation* 14 (11), pp. 2531–2560. DOI: 10.1162/089976602760407955.

Marsaglia, G. (2003). „Xorshift RNGs“. *Journal of Statistical Software, Articles* 8 (14), pp. 1–6. DOI: 10.18637/jss.v008.i14.

Meddis, R. (1986). „Simulation of mechanical to neural transduction in the auditory receptor“. *The Journal of the Acoustical Society of America* 79 (3), pp. 702–711. DOI: 10.1121/1.393460.

Meddis, R. (1988). „Simulation of auditory – neural transduction: Further studies“. *Journal of the Acoustical Society of America* 83 (3), pp. 1056–1063. DOI: 10.1121/1.396050.

Moore, B. C. J. (2003). *An Introduction to the Psychology of Hearing.* 5th ed. Boston: Academic Press. ISBN: 978-0-125-05628-1.

Moore, D. G., G. Valentini, S. I. Walker, and M. Levin (2017). „Inform: A toolkit for information-theoretic analysis of complex systems“. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–8. DOI: 10.1109/SSCI.2017.8285197.

Morrison, A., A. Diesmann, and W. Gerstner (2008). „Phenomenological models of synaptic plasticity based on spike timing“. *Biological cybernetics* 98 (6), pp. 459–478. DOI: 10.1007/s00422-008-0233-1.

Pearson, K. and F. Galton (1895). „VII. Note on regression and inheritance in the case of two parents“. *Proceedings of the Royal Society of London* 58 (347-352), pp. 240–242. DOI: 10.1098/rspl.1895.0041.

Schemmel, J., D. Brüderle, A. Grübl, M. Hock, K. Meier, and S. Millner (2010). „A wafer-scale neuromorphic hardware system for large-scale neural modeling“. *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pp. 1947–1950. DOI: 10.1109/ISCAS.2010.5536970.

Stöckel, D. (2017). „Exploring Collective Neural Dynamics under Synaptic Plasticity on Neuromorphic Hardware“. MA thesis. Kirchhoff Institute of Physics.

Stöckel, D., B. Cramer, A. Hartel, A. Heimbrecht, E. Müller, C. Pehle, Y. Stradmann, M. Petrovici, J. Schemmel, and K. Meier (in prep.). „Stabilizing Neural Activity with Long Term Synaptic Plasticity“.

Toyoizumi, T., JP. Pfister, K. Aihara, and W. Gerstner (2007). „Optimality Model of Unsupervised Spike-Timing-Dependent Plasticity: Synaptic Memory and Weight Distribution“. *Neural Computation* 19 (3), pp. 639–671. DOI: `10.1162/neco.2007.19.3.639`.

Verstraeten, D., B. Schrauwen, D. Stroobandt, and J. Van Campenhout (2005). „Isolated word recognition with the Liquid State Machine: a case study“. *Information Processing Letters* 95 (6), pp. 521–528. DOI: `10.1016/j.ipl.2005.05.019`.

Wibral, M., J. T. Lizier, and V. Priesemann (2015). „Bits from brains for biologically inspired computing“. *Frontiers in Robotics and AI* 2 (5). DOI: `10.3389/frobt.2015.00005`.

Yuste, R. and G. M. Church (2014). „The New Century of the Brain“. *Scientific American* 310 (3), pp. 38–45. DOI: `10.1038/scientificamerican0314-38`.

Acknowledgments

Firstly, I am grateful to Professor Karlheinz Meier for having allowed me to be part of the great research group he build up.

I would like to thank Dr. Johannes Schemmel for his purposeful leadership of the research group and all his dedication in developing the hardware that enabled my work.

I thank Professor Schäfer for his willingness to examine my thesis despite the aberration from his usual areas of research.

Very special thanks go to Benjamin Cramer for his careful and patient supervision of my work and his many insightful tips.

I thank Benjamin, Jakob and Christian for proofreading the thesis and discussing all the little details.

Thanks go to Oliver and Eric for all their help with technical problems during my experiments.

I would also like to thank the whole Electronic Vision(s) group for all the combined achievements and groundwork that enabled this thesis.

Finally, I want to thank my family and friends for supporting me throughout my thesis.

Statement of Authorship (Erklärung)

Ich versichere, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, 08.03.2019

Ort, Datum

Markus Kreft