

Computational Neuroscience of Speech Recognition



Irina Higgins
St Hilda's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Hilary Term 2015

Acknowledgements

I would like to start by thanking my supervisor, Dr Simon Stringer, for taking a chance on me when I first arrived at Oxford and for providing the most unprecedented level of support throughout my years at the Oxford Centre for Theoretical Neuroscience and Artificial Intelligence. This thesis would not be possible without your kindness, enthusiasm, encouragement and help.

I would also like to thank Prof Jan Schnupp for agreeing to become my co-supervisor and helping put my theory into context.

Special thank you is extended to the BBSRC, VoxGen and Oxford Foundation for Theoretical Neuroscience and Artificial Intelligence for financing my research throughout the years.

My somewhat infrequent and sporadic days at the lab would not have been nearly as nice if it were not for the wonderful company of my colleagues: James, Juan, Hector, Bedeho, Ben, Dan, Erica, Akihiro, Harry and Mihaela. Thank you for staying nice to me when my simulations broke the network over and over again, or when my bots accidentally took over your computers. I would also like to thank James for being the most inspirational mentor and for not only teaching me the ways of Computational Neuroscience, but also infecting me with your workaholism, entrepreneurial spirit and determination.

Finally I would like to thank my family and friends for keeping me sane when things got tough. In particular I would like to thank Vladimir and Ruzanna for always finding a way to cheer me up, Victor for showing me by example that we can achieve anything, Helen for adding cool weirdness into my life and sharing my misery at the low points, Peter for steering my career in the right direction and helping me make the most of my time at Oxford, Sarah for helping me take my mind off work, and Pomin for being the best.

Computational Neuroscience of Speech Recognition

Irina Higgins
St Hilda's College
University of Oxford

*A thesis submitted for the degree of
Doctor of Philosophy*

Hilary Term 2015

Abstract

Physical variability of speech combined with its perceptual constancy make speech recognition a challenging task. The human auditory brain, however, is able to perform speech recognition effortlessly. This thesis aims to understand the precise computational mechanisms that allow the auditory brain to do so. In particular, we look for the minimal subset of sub-cortical auditory brain areas that allow the primary auditory cortex to learn 'good representations' of speech-like auditory objects through spike-timing dependent plasticity (STDP) learning mechanisms as described by Bi & Poo (1998). A 'good representation' is defined as that which is informative of the stimulus class regardless of the variability in the raw input, while being less redundant and more compressed than the representations within the auditory nerve, which provides the firing inputs to the rest of the auditory brain hierarchy (Barlow 1961).

Neurophysiological studies have provided insights into the architecture and response properties of different areas within the auditory brain hierarchy. We use these insights to guide the development of an unsupervised spiking neural network grounded in the neurophysiology of the auditory brain and equipped with spike-time dependent plasticity (STDP) learning (Bi & Poo 1998). The model was exposed to simple controlled speech-like stimuli (artificially synthesised phonemes and naturally spoken words) to investigate how stable representations that are invariant to the within- and between-speaker differences can emerge in the output area of the model. The output of the model is roughly equivalent to the primary auditory cortex.

The aim of the first part of the thesis was to investigate what was the minimal taxonomy necessary for such representations to emerge through the interactions of spiking dynamics of the network neurons, their ability to learn through STDP learning and the statistics of the auditory input stimuli. It was found that sub-cortical pre-processing within the ventral cochlear nucleus and inferior colliculus was necessary to remove jitter inherent to the auditory nerve spike rasters, which would disrupt STDP learning in the primary auditory cortex otherwise.

The second half of the thesis investigated the nature of neural encoding used within the primary auditory cortex stage of the model to represent the learnt auditory object categories. It was found that single cell binary encoding (DeWeese & Zador 2003) was sufficient to represent two synthesised vowel classes, however more complex population encoding using precisely timed spikes within polychronous chains (Izhikevich 2006) represented more complex naturally spoken words in speaker-invariant manner.

Computational Neuroscience of Speech Recognition

Irina Higgins
St Hilda's College
University of Oxford

*A thesis submitted for the degree of
Doctor of Philosophy*

Hilary Term 2015

Long Abstract

The physical variability of speech combined with its perceptual constancy makes speech recognition a challenging computational problem. The human auditory brain, however, is able to recognise speech objects, such as phonemes or words, rapidly and accurately. Learning through exposure to the statistics of the auditory world is necessary to form invariant representations of auditory objects, such as speaker or noise independent word representations. This thesis tries to answer the following questions: 1) what are the computational mechanisms utilised by the auditory brain to learn stable invariant representations of auditory object classes, such as phonemes or words; 2) what is the minimal viable brain architecture necessary for the task; 3) and what is the potential encoding used by the brain to represent the learnt auditory objects.

In order to answer these questions, we develop a model grounded in the known neurophysiology of the auditory brain and expose it to speech-like stimuli. In this thesis we try to understand how stable invariant representations of auditory objects emerge through the interactions of spiking dynamics of the model neurons, their ability to learn through biologically inspired learning mechanisms, and the statistics of the auditory input stimuli. For the purposes of this thesis, the following assumptions, constraints, and definitions are used:

1) *Auditory objects* are simple isolated auditory stimuli derived from speech, such as synthesised vowels or naturally spoken single-syllable words. Simple controlled stimuli were chosen for this thesis in order to gain better understanding of the computational mechanisms that allow the model to learn good representations of the auditory inputs.

2) *Invariant representations* are representations that remain stable regardless of the within-class variations present in the raw input. For example, the same word spoken by the same speaker twice, spoken by a different speaker, or presented in

noise, will result in very different raw input audio waves, which in turn will result in very different patterns of firing within the auditory nerve - the first stage of the auditory brain hierarchy. We want to understand how this variability is reduced throughout the auditory brain hierarchy to produce stable informative compressed representations in the last stage of the proposed model in response to the different exemplars of the same auditory object class.

3) the *Input* to the model has to be as similar to the input to the auditory brain as possible, hence a good model of the auditory nerve firing is necessary to transform the raw audio inputs to neurophysiologically realistic firing patterns of the auditory nerve.

4) The model is built using *spiking neurons* with neurophysiologically realistic spiking dynamics, because we want to understand how the interaction between the statistics of the auditory inputs derived from speech and the spiking dynamics of the brain-inspired architecture equipped with long-term learning capabilities leads to the formation of invariant representations of the input auditory objects.

5) The model is equipped with neurophysiologically plausible long-term *learning* capabilities (spike-time dependent learning (Bi & Poo 1998)).

6) The broad *architecture and connectivity* within the model has to be grounded in the known neurophysiology of the auditory brain hierarchy.

The input to the auditory brain is provided by the mechanical cochlea followed by the firing within the auditory nerve (AN), which contains all information about sound identity and location. This information is passed through a large number of non-plastic sub-cortical layers before reaching the auditory cortex (A1), where learning takes place (Zhang et al. 2002). Long-term learning in the brain is driven by the presence of spike-time dependent plasticity (STDP) (Bi & Poo 1998), whereby the strength of connections between cortical neurons is modified according to the mechanism suggested by Hebb's postulate (Hebb 1949).

The auditory brain has to learn to associate together different versions of the same speech object in order to solve the invariant speech recognition problem. A similar problem has been solved in vision using both rate and spiking neural network models of the ventral visual stream using the biologically plausible, unsupervised, continuous transformation (CT) learning mechanism (Evans & Stringer 2012, Higgins & Stringer 2011, Stringer et al. 2006, Tromans et al. 2012). A good way to think about CT learning is essentially as a label describing the learning that emerges through the interactions of the Hebb postulate (Hebb 1949) with stimulus classes that contain high overlaps between their similar transforms. A version of CT learning adjusted to cope with the temporal dimension of auditory objects by introducing a distribution of axonal conduction delays to the afferent plastic connections is used in this thesis.

The minimal viable brain architecture necessary for learning invariant repre-

sentations of auditory object classes is currently unknown, and neither is the role played by the different brain areas (Bizley & Cohen 2014). Chapter 3 addressed these questions by first demonstrating that a simple AN-A1 spiking neural network model of the auditory brain that tries to learn class-selective invariant representations of two synthesised vowels, /i:/ and /a/, straight from the AN spike rasters performs poorly on the task. This is shown to be due to jitter present in the AN firing, which scrambles stable repeating firing patterns, and thus breaks CT STDP learning.

Based on neurophysiological evidence it is, therefore, hypothesised that jitter removal pre-processing within the non-plastic sub-cortical areas of the auditory brain may allow learning to take place. Two stages were found to be necessary. First, the chopper (CH), onset (ON) and primary-like (PL) sub-populations of the cochlear nucleus (CN) remove AN jitter. Then their inputs are converged onto the inferior colliculus (IC) layer, which serves as a whiteboard on which the de-jittered AN spike rasters are written and which serves as clean input to the A1 to allow CT STDP learning to develop class-selective invariant representations of the synthesised vowel stimuli (Delgutte 1982, Ferragamo et al. 1998, Godfrey et al. 1975, Liberman 1993, Rhode et al. 2010, Winslow et al. 1987).

The full AN-CN-IC-A1 model is, therefore, put forward, which incorporates the extra non-plastic pre-processing layers. It is demonstrated that this biologically realistic spiking neural network of the auditory brain then becomes capable of learning invariant representations of two synthesised vowels in an unsupervised manner purely through exposure to the different simulated pronunciations of the two vowels by the same synthetic speaker.

It was shown that all three CN sub-populations mentioned above (CH, ON and PL) are important for transforming the jittered AN input into the form suitable for CT STDP learning. When either CH, ON or PL sub-population is removed and the AN-CN-IC-A1 model parameters are re-optimised, the performance of the model on the two vowel recognition task drops.

Chapter 3 demonstrates that the reason why AN-CN-IC-A1 model performs better than the simple AN-A1 model when learning invariant representations of the two vowel stimuli is because of its precise architecture rather than the mere presence of extra layers of processing. A simple four layer model without the precise architecture of the AN-CN-IC-A1 model was shown to fail to learn transform invariant representations of the two synthesised vowel stimuli.

Chapter 4 analyses the performances of the simple AN-A1 and the full AN-CN-IC-A1 models further. First, AN jitter removal within the sub-cortical sub-populations within the CN is quantified. It is shown that the resulting IC firing rasters are essentially the same as AN rasters, albeit with less stochasticity. The jitter removal is achieved first through the high-pass filter response properties of

the CH cells, which amplify strong and dampen weak AN responses. This is then combined in the IC with the regularly timed priming received from the ON cells corresponding to the fundamental frequency of the stimulus, and the general activation from the PL sub-population, to result in the de-jittered IC spike rasters suitable for CT STDP learning.

Chapter 4 also analyses the response metrics within the hierarchy of the AN-CN-IC-A1 model. It is shown that synchronisation (inversely proportional to the amount of jitter) and the maximal amount of single cell information about the stimulus identity increase throughout the hierarchy of layers, while the average firing rate decreases. This is in line with neurophysiological findings, thus providing support for the validity of the AN-CN-IC-A1 model (Ito & Fukushima 1988, Kurogi 1991, Steinschneider et al. 1990).

Further analysis is performed in Chapter 4 to understand the changes within the simple AN-A1 and the full AN-CN-IC-A1 models due to training. It is shown that throughout training the rate of firing of the output A1 layer of the full AN-CN-IC-A1 model drops, the firing patterns of individual cells become more regular, and the cells become more stimulus selective, whereby they fire 1-2 spikes at similar times for most presentations of one vowel and to hardly any presentations of the other vowel. Essentially information encoding changes from being contained within a large population at the beginning of the auditory brain hierarchy, to being contained within a small population at the end of it, in line with Barlow (1961) and DeWeese & Zador (2003). This is not the case for the AN-A1 model, which only displays the reduction in firing rate but not the other two effects of training (regularity and stimulus selectivity) described for the full AN-CN-IC-A1 model.

Another difference between the full AN-CN-IC-A1 and the simple AN-A1 models of the auditory brain is that during training some IC→A1 afferent connections of the full model get strengthened while others are weakened, resulting in a bimodal distribution of afferent connection weights. The same is not the case for the AN-A1 model, whereby all of its AN→A1 afferents get weakened during training, resulting in a unimodal distribution. The presence of a bimodal distribution of afferent connections weights after training is a sign of learning, since the afferent connections can only strengthen if stable repeating patterns are present in the pre-synaptic layer during training (Masquelier et al. 2008). The global weakening of the AN→A1 afferents of the simple AN-A1 model, therefore, provides further evidence for AN jitter impeding CT STDP learning. Likewise, the strengthening of certain IC→A1 afferent connections of the full AN-CN-IC-A1 model provides further evidence for the role of the non-plastic CN and IC layers in removing AN jitter and enabling CT STDP learning.

The nature of stimulus selectivity of individual A1 cells of the full AN-CN-IC-A1 model after training is also analysed in Chapter 4. It is shown that even when

the distribution of axonal delays between the IC and A1 layers is kept constant, cell stimulus preference acquired during training changes depending on the different initialisation values of the IC→A1 afferent connections, whereby the same A1 cell can prefer vowel /i:/ after training when one set of initialisation values is used, but it can change its preference to vowel /a/ if a different set of initialisation values is used for the IC→A1 afferent connections. This does not affect the levels of information about the identity of the two synthesised vowel stimuli achieved by the network, however. The full AN-CN-IC-A1 still reaches high amounts of information about the identity of the two synthesised vowel stimuli even if it is trained from a very jumbled and distinct distribution of IC→A1 afferent connections.

In Chapter 4 the robustness of performance of the trained AN-CN-IC-A1 model is also tested. First, the model is presented with new never before seen exemplars of the two learnt vowels. It is demonstrated that the network is able to generalise its learning and recognise these new exemplars well. Second, the trained model is presented with exemplars of two new vowels in order to verify that the model has learnt the precise identities of the two original vowels rather than the ability to differentiate between any two different auditory objects. It is shown that the model struggles to tell apart the new vowels, thus passing the verification test.

Chapter 4 also starts the discussion of how the neural code changes throughout the auditory brain hierarchy. It is necessary that all information about the auditory world, such as auditory object identity, location, loudness, source and more, must be present in the AN. This is in line with the information inequality rule, whereby any data processing can result in at most as much information as was present in the original data (Nelken 2004). The role of the auditory brain hierarchy is, therefore, to transform the neural code into as simple a form as possible in the output layer. The search for the simplest sufficient encoding scheme for auditory object identification starts in Chapter 4 and continues throughout Chapters 5 and 6.

Chapter 4 demonstrates that the simple single cell binary on/off encoding scheme, which only takes into account whether a particular output A1 layer cell fires at least once during stimulus presentation or not, is sufficient to encode transform invariant representations of two synthesised vowel stimuli. Evidence for such an encoding scheme was found in the auditory cortex by DeWeese & Zador (2003). Chapter 5 describes whether the full AN-CN-IC-A1 model is able to cope with learning transform invariant representations of more than two synthesised vowels by presenting up to six vowel stimuli. It is shown that as the number of vowels is increased, the simple on/off encoding scheme used by the A1 cells of the model to represent two vowels becomes insufficient. Whereas Chapter 3 demonstrates that a single A1 cell is able to perfectly differentiate between two synthesised vowels using simple on/off single cell encoding scheme, such an encoding scheme provides

little information about the identity of three or more vowels. When population votes of A1 cells using on/off encoding are considered, the results are even worse, since the different A1 cells tend to disagree with each other on stimulus identity.

When temporal rather than rate information within the firing rasters of A1 cells is investigated, it is found that the precise firing of cells provides more differentiating responses to different vowels compared to the simple on/off encoding. Chapter 5, therefore, tests whether this additional stimulus discriminability helps separate out the representations of up to six different vowel classes. It is shown that single cell information using the precise firing times of A1 cells with respect to stimulus start provides more information about the vowel classes than the simpler rate binary (on/off) encoding scheme. This, however, is not enough to fully differentiate between up to six vowels. Multiple cell information using this new temporal encoding scheme is therefore considered, and it is shown to be enough to differentiate between up to six vowel stimuli. It is, therefore, suggested that when the network has to learn to categorise more than two synthesised vowel classes, it requires a temporal encoding scheme distributed within a population of cells in the primary auditory cortex. This is different to the simpler single cell binary (on/off) encoding scheme sufficient for representing the identities of just two vowel classes. The performance of the AN-CN-IC-A1 model is compared to that of the AN-A1 model and is found to be significantly better, thus further supporting the finding outlined in Chapters 3 and 4 that sub-cortical pre-processing is important for allowing CT STDP learning to form transform invariant representations of auditory objects.

The last chapter of this thesis, Chapter 6, investigates the performance of the full AN-CN-IC-A1 model on learning invariant representations of two naturally spoken digits 'one' and 'two'. The extra complexity of naturally spoken words makes transform invariance learning a more challenging task for the model compared to learning about synthesised vowels as described in the previous chapters. It is shown that the rate on/off encoding provides very poor information about the naturally spoken digit stimuli. A temporal population encoding scheme is, therefore, hypothesised to be necessary, however, the spike-timing encoding scheme used in Chapter 5 is shown to be inapplicable for the encoding of the naturally spoken digits 'one' and 'two' within the output layer A1 of the AN-CN-IC-A1 model. Instead a new temporal encoding scheme based on the presence of stable polychronous chains (Izhikevich 2006) is suggested.

Chapter 6 starts by providing evidence for the presence of stable polychronous chains in the full AN-CN-IC-A1 model after training. This includes the presence of stable patterns of firing within the IC that activates particular A1 cells; the strengthened IC→A1 afferent connections for certain A1 cells; and the general stability of firing patterns within the A1 layer. This evidence suggests that sta-

ble patterns of firing constituting polychronous chains (Izhikevich 2006) emerges within the A1 layer of the AN-CN-IC-A1 model after training. Such evidence is not present for the A1 layer of the simple AN-A1 model.

An information measure scheme is developed to analyse how much information is present about the identity of the two naturally spoken digits within the polychronous chains of firing within the A1. It is shown that this temporal polychronous chain based encoding scheme contains significantly more information about the two naturally spoken stimuli than the rate on/off encoding scheme. Furthermore, more information is present in the A1 layer of the AN-CN-IC-A1 model than the AN-A1 model.

The polychronous chains emerging in the A1 layer of the full AN-CN-IC-A1 model, however, do not reach the theoretically maximal amount of information about the two stimuli. Nevertheless, when an additional plastic cortical area (Belt) is attached on top of the A1 layer, it is shown that the A1→Belt afferent connections get strengthened and more information is present in the Belt than the A1 area of the full AN-CN-IC-A1-Belt model. This is not the case for the simple AN-A1-Belt model. While full information is not reached by the Belt layer of the full AN-CN-IC-A1-Belt model, the increase in the amount of information compared to the A1 layer suggests that extra cortical layers can eventually learn full transform invariant representations of the two naturally spoken digits. Simulations with more than two plastic cortical layers, however, are outside of the scope of this thesis due to computational restrictions.

This thesis, therefore, shows that invariance learning of auditory objects, such as synthesised vowels or naturally spoken words, requires sub-cortical pre-processing within the CN and IC in order to remove AN jitter and allow CT STDP learning to work. This results in the emergence of stimulus class selective invariant polychronous chains in the output layer of the model, which provides evidence for temporal population encoding of auditory objects in the auditory brain.

This thesis concludes with a discussion and evaluation of the research presented, followed by some themes for future research.

Abbreviations

AN: auditory nerve

CN: cochlear nucleus

CH: chopper cells within cochlear nucleus

PL: primary-like cells within cochlear nucleus

ON: onset cells within cochlear nucleus

IC: inferior colliculus

A1: primary auditory cortex

CT: continuous transformation learning

STDP: spike-timing dependent plasticity

LTP: long-term potentiation

LTD: long-term depression

CF: characteristic frequency

F_0 : vowel fundamental frequency

F_1 : vowel first formant

F_2 : vowel second formant

F_3 : vowel third formant

Contents

Abstract	i
Long Abstract	ii
List of Figures	xix
List of Tables	xxv
1 Literature Review	1
1.1 Introduction	1
1.2 Motivation	2
1.3 Definitions	5
1.4 Physics of Speech Sounds	6
1.4.1 Simple and Complex Sounds	6
1.4.2 Speech Sounds	8
1.4.2.1 Vowels	10
1.4.2.2 Consonants	12
1.5 Neurophysiology of the Auditory System	16
1.5.1 Rate or Temporal Coding	17
1.5.2 Plasticity in the Auditory Brain	22
1.5.3 Generalising from Animal Studies	24

1.5.4	Cochlea	27
1.5.5	Auditory Nerve	29
1.5.5.1	Tuning Curves	30
1.5.5.2	Amplitude Saturation	30
1.5.5.3	Adaptation	31
1.5.5.4	Two-Tone Suppression	33
1.5.5.5	Phase-Locking	35
1.5.5.6	Responses to Vowels	35
1.5.6	Cochlear Nucleus	40
1.5.6.1	Primary-Like Neurons	41
1.5.6.2	Primary-Like With Notch Neurons	42
1.5.6.3	Chopper Neurons	44
1.5.6.4	Onset Neurons	47
1.5.7	Superior Olivary Complex	50
1.5.7.1	Nuclei of the Trapezoid Body	50
1.5.7.2	Nuclei of the Lateral Lemniscus	51
1.5.8	Inferior Colliculus	51
1.5.9	Medial Geniculate Nucleus	54
1.5.10	Auditory Cortex	55
1.5.10.1	Vocalisation Response Properties: A1	59
1.5.10.2	Vocalisation Response Properties: pSTG and STS	61
1.5.10.3	Vocalisation Response Properties: vPFC	63
1.6	Past Models of Speech Recognition	65
1.6.1	Rate-Coded Models	66
1.6.1.1	Model by McClelland & Elman (1986)	66
1.6.1.2	Model by Kral & Majernik (1996)	67
1.6.1.3	Model by Kurogi (1991)	69

1.6.2	Spiking Models	71
1.6.2.1	Model by Meddis & O'Mard (2006)	71
1.6.2.2	Model by Masquelier et al. (2008)	73
1.6.2.3	Model by Nager et al. (2002)	74
1.6.2.4	Model by Coath et al. (2014)	76
1.6.2.5	Reservoir computing approaches	78
1.7	Models of Spiking Neurons	80
1.7.1	Hodgkin-Huxley	81
1.7.2	Integrate-and-Fire	83
1.7.3	Izhikevich	84
2	Methods	86
2.1	Introduction	86
2.2	Stimuli	86
2.2.1	Stimulus Presentation Paradigm	88
2.2.2	Extra vowel stimuli	90
2.2.3	Naturally spoken word stimuli	91
2.3	Learning in the Auditory Brain	93
2.3.1	Plasticity	93
2.3.2	Spike-Time Dependent Plasticity Learning	93
2.4	Delays	98
2.5	Learning Transform Invariant Representations of Auditory Objects	99
2.5.1	Continuous Transformation Learning in Visual Object Recognition	102
2.5.2	Continuous Transformation Learning in Auditory Object Recognition	106

2.5.3	Stimulus Presentation Paradigm Appropriate for Continuous Transformation Learning	110
2.6	Auditory Nerve Model	111
2.6.1	Generalising from Cat Neurophysiological Data	112
2.7	Auditory Nerve Input Layer	114
2.8	Jitter in the Auditory Nerve	115
2.8.1	Phase-Locking	116
2.9	Model of Spiking Neurons	118
2.9.1	Excitatory Cells	121
2.9.2	Inhibitory Cells	123
2.10	Quantitative and Plotting Measures	124
2.10.1	Neurogram	124
2.10.2	Synchronisation Index	125
2.10.3	Quantifying Spike Raster Similarity	125
2.10.3.1	Same Exemplar Index	126
2.10.3.2	Different Exemplars Index	127
2.10.3.3	Different Category Index	128
2.10.3.4	Polychronization index	129
2.10.4	Information Analysis	130
2.10.5	Information Encoding	132
2.10.5.1	Simple Binary On/Off Encoding	132
2.10.5.2	Bayesian Encoding	134
3	The Role of Subcortical Processing in Auditory Category Learning: Models	136
3.1	Introduction	136
3.2	Simple AN-A1 Model Architecture	138

3.3	Validating the Auditory Nerve Layer	141
3.4	Optimising the Simple AN-A1 Model	145
3.5	Quantifying Space and Time Jitter in the Auditory Nerve	155
3.6	Jitter Reduction in the Cochlear Nucleus and Inferior Colliculus . .	156
3.6.1	Space Jitter Removal	157
3.6.2	Time Jitter Removal	157
3.7	Full AN-CN-IC-A1 Model Architecture	161
3.8	Optimising the Full AN-CN-IC-A1 Model	165
3.8.1	Optimising Primary-Like With Notch/Chopper Cells	165
3.8.2	Optimising Onset Cells	169
3.8.3	Neurophysiological Validity of the Cochlear Nucleus Layer .	171
3.8.3.1	Primary-Like Cells	171
3.8.3.2	Primary-Like With Notch/Chopper Cells	172
3.8.3.3	Onset Cells	174
3.8.4	Cochlear Nucleus to Inferior Colliculus Connectivity	175
3.8.5	Optimising STDP and Finding Minimal Layer Set for Full AN-CN-IC-A1 Model	175
3.9	Simple Four-Layer Model	180
4	The Role of Subcortical Processing in Auditory Category Learn- ing: Results	184
4.1	Introduction	184
4.2	Analysing Space and Time Jitter Removal in the Cochlear Nucleus	185
4.3	Understanding the Role of Chopper Cells in Space Jitter Removal .	186
4.4	Comparing AN-A1 and AN-CN-IC-A1 Models	190
4.4.1	Change in Information Due to Training	190

4.4.2	Change in AN-CN-IC-A1 Model Response Properties Due to Training	190
4.4.3	Preferred Afferent Patterns for AN-A1 and AN-CN-IC-A1 Models	199
4.4.4	Changes in Afferent Connections Strength Due to Learning .	202
4.4.5	Distribution of Afferent Connection Weights for AN-A1 and AN-CN-IC-A1 Models	206
4.4.6	Cell Response Properties Within the AN-A1 and AN-CN-IC-A1 Models	208
4.5	Understanding AN-CN-IC-A1 Model	208
4.5.1	Change in Information Within Layers of the AN-CN-IC-A1 Model	208
4.5.2	Analysing AN-CN-IC-A1 Afferent Connections Strength Changes Due to Training	210
4.5.3	Role of Afferent Connections Strength Initialisation Values for Learning	210
4.5.4	Weight Convergence During Training	214
4.5.5	Testing AN-CN-IC-A1 Model on Novel Stimuli	214
4.5.6	Importance of Formant Peaks In Differentiating Between Vowel Stimuli	216
4.6	Change in Information Encoding Within the Auditory Brain	217
4.6.1	Evidence for Encoding Simplification Within the Auditory Brain	218
4.6.2	Evidence for Reduction in the Information Bearing Cell Subset Size Within the Auditory Brain	219

5	Learning Transform Invariant Representations of Six Vowels: the Importance of Temporal Coding	223
5.1	Introduction	223
5.2	Stimuli	224
5.3	Training With the New Vowel Sets	224
5.4	Changes in Vowel Representations With Increased Stimulus Set Sizes	225
5.5	Quantifying Distributed Encoding	228
5.5.1	Multiple Cell Information	229
5.6	Understanding Stimulus Preferences of the full AN-CN-IC-A1 Model	231
5.6.1	Stimulus Preference Plots	231
5.6.2	Changes in Stimulus Preference	232
5.6.3	Stimulus Preferences for Distributed Encoding	232
5.7	Understanding Cell Firing Patterns	235
5.8	Using Temporal Patterns of Firing of Single Cells	237
5.8.1	Bayesian Single Cell Information	237
5.8.2	Bayesian Multiple Cell Information	240
6	Towards Learning Transform Invariant Representations of Naturally Spoken Words	243
6.1	Introduction	243
6.1.1	Polychronous Chains	244
6.2	Stimuli	247
6.2.1	Analysing the Complexity of Naturally Spoken Stimuli	247
6.3	Analysing Network Responses to Naturally Spoken Stimuli	249
6.3.1	On/Off Encoding	249
6.3.1.1	Information About Naturally Spoken Digits Using On/Off Encoding	249

6.3.2	Single Cell Temporal Encoding	251
6.3.3	Choosing the Right Information Encoding Scheme	252
6.3.4	Analysing A1 Firing	253
6.3.4.1	Local Firing Stability Measure	254
6.3.4.2	A1 Local Firing Stability in Response to Naturally Spoken Digits	257
6.3.4.3	Preferred Firing Patterns of A1 Cells	257
6.3.5	Changes in Distributions of IC→A1 Afferent Connection Weights	260
6.3.6	Changes in IC→A1 Afferent Connection Weights Due to Learning	260
6.3.7	Measuring Polychronization Index in A1	263
6.3.8	Expanding the Models	264
6.3.8.1	Changes in A1→Belt Afferent Connection Weights Due to Learning	265
6.3.8.2	Changes in Distributions of A1→Belt Afferent Con- nection Weights	265
6.3.9	Analysing Population Encoding Information	268
6.3.9.1	Finding the Stable Repeating Discriminative A1 Firing Patterns	269
6.3.9.2	Temporal Encoding Results	270
6.3.10	Examining Polychronous Chains	273
7	Conclusion	276
7.1	Introduction	276
7.2	The Role of Sub-Cortical Pre-Processing	276
7.2.1	Future Work: Testing Neurophysiological Predictions	277

7.3	New View on the Volley Principle in the Auditory Nerve	278
7.4	Continuous Transformation Learning	278
7.5	Spiking Neuron Model by Izhikevich (2003)	279
7.6	Training Data	279
7.7	Increasing the Capacity of the AN-CN-IC-A1 Model	280
7.7.1	Future Work: Adding Extra Feed-Forward Cortical Layers .	280
7.7.2	Future Work: Increasing the Average Connection Weight Magnitudes	280
7.7.3	Future Work: Adding Recurrent Connections	281
7.7.3.1	Future Work: Testing Noise Robustness in the Re- current AN-CN-IC-A1 Model	282
7.8	Neural Code	282
7.9	Unique Attributes of the AN-CN-IC-A1 Model	284
References		286

List of Figures

1.1	Sound frequency and amplitude	7
1.2	Complex sound	8
1.3	Vowel production	10
1.4	Vowel formants	11
1.5	Common vowels	12
1.6	Consonants	13
1.7	Auditory brain overview	18
1.8	Auditory brain schematic	19
1.9	Rate and temporal information encoding in the brain	20
1.10	Different types of neural codes	21
1.11	Cochlea schematic	28
1.12	Auditory nerve tuning curve	31
1.13	AN amplitude saturation	32
1.14	Two-tone suppression in the AN	34
1.15	Phase-locking and the volley principle in the AN	36
1.16	AN amplitude modulation	37
1.17	AN response properties to /da/	38
1.18	AN phase-locking regions	39
1.19	CN neuron types	41

1.20	Primary-like neurons	43
1.21	Primary-like and primary-like with notch neurons	44
1.22	Phase-locking in auditory nerve and primary-like with notch neurons	45
1.23	Chopper cells	47
1.24	Onset cells	49
1.25	Lateral lemniscus pathways	52
1.26	Inhibition in the IC	53
1.27	Tuning curves of MGN neurons	56
1.28	'Where' and 'what' pathways in the auditory cortex	57
1.29	A1 response properties to vocalisations	60
1.30	Model by Kral & Majernik (1996)	68
1.31	Model by Kurogi (1991)	70
1.32	Model by Meddis & O'Mard (2006)	72
1.33	Spiking model by Masquelier et al. (2008)	75
1.34	Parameter effects on the model by Masquelier et al. (2008)	76
1.35	Spiking neural network model by Coath et al. (2014)	77
1.36	Spiking neuron models	82
2.1	Synthesising exemplars of two vowel stimuli	89
2.2	Synthesising exemplars of six vowel stimuli	92
2.3	Visualisation of two naturally spoked digits	94
2.4	Spike-timing dependent plasticity learning	95
2.5	Function of heterogeneous axonal conduction delays	100
2.6	Continuous transformation learning class boundaries	102
2.7	Continuous transformation learning	104
2.8	Effects of network parameters changes on continuous transformation learning	105

2.9	Auditory nerve model schematic	113
2.10	Phase-locking in the auditory nerve	117
2.11	Space and time jitter in the auditory nerve	119
2.12	Twenty response types of Izhikevich neuron model	122
3.1	Schematic of the simple AN-A1 model of the auditory brain	142
3.2	Comparison of auditory nerve response properties between neuro- physiological data and model fibers	143
3.3	Rasters and neurograms of auditory nerve responses to vowels /a/ and /i:/	144
3.4	Spectra of vowels /a/ and /i:/	145
3.5	Results of STDP parameter search for the simple AN-A1 model of the auditory brain	150
3.6	Single cell information plot	151
3.7	Results of inhibition and afferents weight initialisation range pa- rameter search for the simple AN-A1 model of the auditory brain .	154
3.8	Space and time jitter removal	159
3.9	Schematic of the full AN-CN-IC-A1 model of the auditory brain . .	166
3.10	Results of Gaussian connectivity parameter search between auditory nerve and primary-like with notch layers	168
3.11	Results of connectivity parameter search between auditory nerve and onset layers	170
3.12	Comparison of primary-like and chopper response properties be- tween neurophysiological data and model fibers	172
3.13	Comparison of primary-like with notch response properties between neurophysiological data and model fibers	173
3.14	Periodogram of onset fibers in response to vowel /a/	174

3.15	Results of parameter search for the full AN-CN-IC-A1 model of the auditory brain	177
3.16	Schematic of the simple four-layer model of the auditory brain . . .	182
4.1	Aggregate auditory nerve and inferior colliculus firing rasters	186
4.2	Changes throughout the layers of the AN-CN-IC-A1 model	187
4.3	Auditory nerve response properties to vowels /a/ and /i:/	188
4.4	Chopper cells' response properties to vowels /a/ and /i:/	189
4.5	Response properties of output cell 891 of the full AN-CN-IC-A1 model after training	192
4.6	Response properties of output cell 891 of the full AN-CN-IC-A1 model before training	193
4.7	Response properties of output cell 353 of the full AN-CN-IC-A1 model after training	194
4.8	Response properties of output cell 353 of the full AN-CN-IC-A1 model before training	195
4.9	Response properties of output cell 198 of the simple AN-A1 model after training	197
4.10	Response properties of output cell 547 of the simple AN-A1 model after training	198
4.11	Aggregate IC firing raster aligned on the spikes by cell 891 within the A1 layer of the full AN-CN-IC-A1 model	200
4.12	Aggregate IC firing raster aligned on the spikes by cell 353 within the A1 layer of the full AN-CN-IC-A1 model	201
4.13	Aggregate AN firing raster aligned on the spikes by cell 547 within the A1 layer of the simple AN-A1 model	201

4.14	Aggregate AN firing raster aligned on the spikes by cell 198 within the A1 layer of the simple AN-A1 model	202
4.15	Afferent connection weight changes after training	204
4.16	Distribution of strengthened afferents in the full AN-CN-IC-A1 model	205
4.17	Afferent connection strength histograms for AN-A1 and AN-CN-IC-A1 output cells	207
4.18	Response properties to vowels /i:/ and /a/	209
4.19	Output layer information retention after shuffling the trained afferent connection weights, then re-training the full AN-CN-IC-A1 model	212
4.20	Output layer information retention after re-initialising and re-training the full AN-CN-IC-A1 model	213
4.21	Response properties of the originally trained and re-trained after afferent weights shuffling AN-CN-IC-A1 model	215
4.22	Weight convergence during training	216
4.23	Schematic of back-propagating neural network	220
4.24	Decoder performance on the two-vowel recognition task with diminishing input cell populations	222
5.1	Single cell information in response to multiple vowels	227
5.2	Vowel recognition accuracy using population encoding	230
5.3	Distributed vowel representations in A1	233
5.4	Distributed vowel representations in A1	234
5.5	Firing rasters of A1 cells in response to four vowels	236
5.6	Probabilistic single cell vowel information	238
5.7	Firing rasters of A1 cells in response to four vowels	239
5.8	Probabilistic multiple cell vowel information	241

6.1	Toy illustration of two polychronous chains	245
6.2	A polychronous chain learnt by a recurrent spiking neural network .	248
6.3	Firing rasters of A1 cells in response to two naturally spoken digits	250
6.4	Rate information encoding of naturally spoken digits	251
6.5	Preferred firing patterns of A1 cells of trained AN-CN-IC-A1 model	255
6.6	Preferred firing patterns of A1 cells of untrained AN-CN-IC-A1 model and simple AN-A1 model	258
6.7	Afferent connection strength histograms for AN-A1 and AN-CN- IC-A1 output cells	261
6.8	Afferent connection weight changes after training	262
6.9	Polychronization index	264
6.10	Afferent connection weight changes after training	266
6.11	A1→Belt afferents histograms of the full AN-CN-IC-A1-Belt model after training	267
6.12	Multiple cell information for digits	272
6.13	Stimulus selective polychronous chains	274

List of Tables

1.1	Spontaneous activity of Type 1 auditory nerve fibers	29
2.1	Vowel formants	88
2.2	Parameters of excitatory and inhibitory cells	121
2.3	Joint probability table for on/off encoding	132
3.1	Parameters for the simple AN-A1 model of the auditory brain . . .	140
3.2	Optimal parameter sets for the two best performing simple AN-A1 models of the auditory brain	149
3.3	Space and time jitter measures in the auditory nerve	156
3.4	Parameters for the full AN-CN-IC-A1 model of the auditory brain .	164
3.5	Optimal parameter set for the Gaussian connectivity between audi- tory nerve and primary-like with notch/chopper layers	168
3.6	Optimal parameter set for the connectivity between auditory nerve and onset layers	170
3.7	Optimal parameter set for the best performing full AN-CN-IC-A1 models of the auditory brain	178
3.8	Effect of different sub-populations of cochlear nucleus on the per- formance of the full AN-CN-IC-A1 model of the auditory brain . . .	179
3.9	Parameters for the simple four-layer model of the auditory brain . .	181

3.10	STDP parameters for the simple four-layer model	183
4.1	Space and time jitter measures in the inferior colliculus	185
5.1	Selecting optimal subsets of multiple vowels	225
5.2	Optimal vowel sets	225
5.3	Joint probability table for on/off encoding for up to six vowels . . .	226
5.4	Joint probability table for multi-response information encoding . . .	229
6.1	Joint probability table for multiple cell encoding	271

1

Literature Review

1.1 Introduction

One of the fundamental characteristics of speech is its physical variability and perceptual constancy (Eggermont 2001). Speech signals have been found to vary greatly both between and within speakers depending on the speaker's emotional state, physiology, speaking style, rate of speech and accent among other parameters (Benzeghiba et al. 2007). This makes the problem of automatic speech recognition particularly hard. While the human auditory brain is able to recognise speech effortlessly, the mechanisms it employs to learn invariant representations of speech objects are currently unknown.

A lot of neurophysiological and psychophysical research has been conducted in order to understand how auditory objects are recognised by the auditory brain. The majority of the data collected, however, is of a descriptive nature, thus leaving explanatory gaps, whereby no insight is given into the mechanisms used by the

auditory brain to perform invariant auditory object recognition. Furthermore, no explanation is provided as to why the auditory brain is organised in its particular way. In order to go beyond purely descriptive accounts of the auditory brain, it is important to build a biologically realistic model of it that can be used to generate and test specific hypotheses about the structure of mechanisms that allow the auditory brain to perform invariant auditory object recognition, including speech perception (Arbib et al. 1995, Corchs & Deco 2002, Horwitz & Tagamets 1999, Husain et al. 2004, Salem et al. 2007). This thesis describes an attempt to understand some of the processes involved in learning stable invariant representations of speech-like auditory objects, such as phonemes and simple words, through biologically plausible spiking neural network modelling of the auditory brain.

This chapter presents a literature review providing the background information necessary for understanding the modelling work described in later chapters. First, the motivations for the newly proposed model are given. Then the nature of speech sounds is described, followed by an overview of the neurophysiology of the auditory brain involved in auditory object identification. Past relevant attempts at auditory brain modelling are then discussed. The chapter is concluded by an overview of spiking neuron models considered for the computational modelling work described in later chapters.

1.2 Motivation

It is important to understand how the early stages of the auditory brain hierarchy (up to approximately the primary auditory cortex) perform auditory object classification, because speech recognition is impossible if its basic acoustic components, e.g. phonemes or words, are not correctly classified first. In order to understand how the brain performs such acoustic pattern classification, it can be helpful to isolate the minimum subset of brain areas necessary for the task. One

way of selecting the minimal subset of brain areas necessary for auditory object recognition is to create biologically realistic computational models of them inspired by the available neurophysiological data, before testing different arrangements of these modules to find the set that transforms input AN spike trains in the most optimal way. An optimal transformation is defined as that which retains the most information about auditory object identity while compressing its representation encoding into as simple a code as possible to remove redundancy (Barlow 1961).

When constructing a model, one has to decide on the level of abstraction that is appropriate for the questions asked. In this thesis, we are interested in learning how the auditory brain transforms the noisy firing in the auditory nerve (AN) in response to the variable raw auditory input into stable compressed representations in the primary auditory cortex (A1) that are informative of the stimulus class membership. We put the following constraints on our model: 1) the input to the model has to be as similar to the input received by the real auditory brain (i.e. the model AN has to replicate the response properties of the real AN as closely as possible); 2) the building blocks of our model are spiking neurons; 3) the spiking dynamics of the neurons have to be neurophysiologically plausible; 4) the learning in the model has to be neurophysiologically plausible; 4) the broad architecture of the model has to be neurophysiologically plausible. Hence, in this thesis we investigate how the brain may learn robust representations of simple auditory objects using the unsupervised interactions of physiologically plausible circuits equipped with learning capabilities with the statistics of speech-like auditory inputs.

Neurophysiological studies (discussed later in this chapter) have provided evidence as to which auditory brain areas may be included or excluded from the minimal set necessary for auditory object identification that is our proposed model. The first area of the auditory brain hierarchy that is included within the minimal auditory object recognition subset of our model is the auditory nerve (AN), since

it is the first stage of auditory processing through which all auditory information is passed. As will be discussed later in this chapter, auditory processing is split between a number of different sub-cortical areas after leaving the AN. This thesis proposes a hypothesis that the chopper, onset and primary-like sub-populations of the cochlear nucleus (CN) are crucial for auditory object recognition and should also be included in the minimal subset of our model. On the other hand, it has been shown that the superior olivary complex (SOC) is primarily involved in sound localisation (Boatman 2006). The SOC is, therefore, not included into our proposed model. The first place where all processing streams converge again after being split into separate streams in the AN is the inferior colliculus (IC). The IC, therefore, is another likely candidate for the minimal subset necessary for auditory object recognition. Finally, it is also safe to include the primary auditory cortex (A1) into this minimal subset, since it is where most of learning happens, as suggested by neurophysiological evidence.

The proposed model of the auditory brain, therefore, consists of areas AN, CN, IC and A1. While the first two areas of the proposed model, AN and CN, are modelled to closely resemble the real AN and CN, the higher areas IC and A1 are meant to be more loose approximations of the higher parts of the auditory brain hierarchy. The exact locations of the homologues of the modelled IC and A1 in the brain are not stated. They could be a combination of neighbouring areas in the brain (for example, IC may be an approximation of a combination of the real SOC + IC + MGN; model A1 may be an approximation of the real MGN + A1). The inclusion of these areas in the model, however, is paramount in order to allow the model to learn robust representations of auditory objects, such as phonemes or words.

Learning is necessary to form invariant representations of auditory objects, such as speaker or noise independent word representations. Long-term learning

in the brain is driven by the presence of spike-time dependent plasticity (STDP) (Bi & Poo 1998), whereby the strength of connections between cortical neurons is modified according to the mechanism suggested by Hebb (1949). He proposed that when activity in one cell repeatedly drives activity in another cell, the synaptic connection between the two cells is strengthened, otherwise it is weakened. STDP learning is discussed in more details in Chapter 2 and it is the mechanisms that will allow our model to learn stable representations that are informative of the auditory stimulus class membership.

1.3 Definitions

This section provides operating definitions for some key concepts used in this thesis. This work attempts to understand how the auditory brain may learn stable transform invariant representations of auditory objects, defined as follows:

Auditory objects: since this thesis aims to make first steps in understanding how speech recognition may be performed within a neurophysiologically grounded spiking architecture using realistic auditory nerve firing inputs, we define auditory objects as any simple speech unit. An auditory object may, therefore, be a phoneme (a vowel or a consonant), a speech supra-segmental (either a larger acoustical pattern corresponding to a combination of phonemes or a smaller acoustical pattern corresponding to a part of a phoneme (Eggermont 2001)), or a word. In this thesis we do not move beyond the scope of simple single-syllable words. Using more complex auditory objects as stimuli is left for future work.

Transform invariant representations: this terminology comes from the work on vision, where researchers are concerned with understanding how the visual brain learns representations of visual objects that are invariant to transforms, such as translation or rotation. When applying this terminology to auditory objects, transform invariance means invariance to any variations in the raw auditory input

that does not affect the class membership of the auditory object. For example, transforms of a word would include different pronunciations of the same word by the same or a different speaker, the addition of background noise, etc. A transform invariant representation would remain unchanged for the different exemplars of the same auditory object class.

Speech recognition: the title of this thesis mentions speech recognition. The work described henceforth, however, is only the first step in understanding how a dynamic system consisting of a large number of neurophysiologically grounded neurons arranged in a neurophysiologically inspired architecture and equipped with a neurophysiologically realistic learning rule can start to learn transform invariant representations of auditory objects. Hence, for the purposes of running systematic tests and gaining good understanding of the computational principles underlying such learning, this thesis is limited in scope to learning about very simple auditory stimuli derived from speech. These include isolated synthesised vowels and naturally spoken single-syllable words. We hope that by expanding the model as described in Chapter 7, future work can address more interesting questions regarding speech recognition, such as segmentation of continuous speech or dealing with noise.

1.4 Physics of Speech Sounds

1.4.1 Simple and Complex Sounds

Sounds are one-dimensional, time-varying signals, that can be characterised using three parameters: frequency, intensity and duration (Fig. 1.1). All sounds can be split into simple and complex, whereby simple sounds are pure tones represented by a single sine wave with constant frequency and amplitude, while complex sounds are spectrally and/or temporally complex, comprising multiple frequencies that

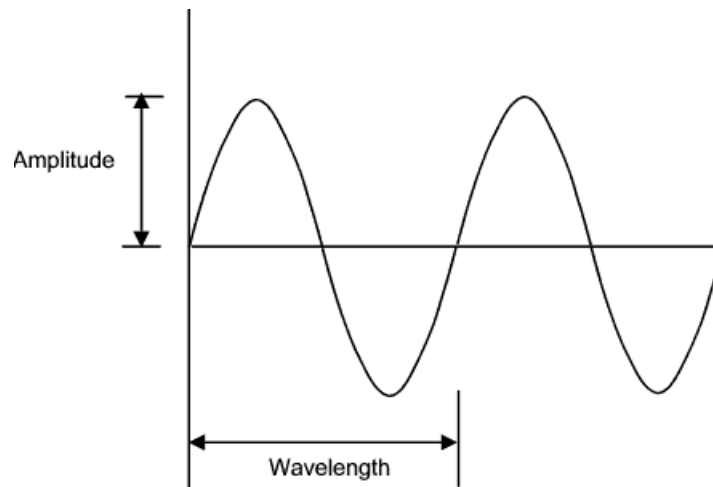


Figure 1.1: Illustration of sound wave properties of frequency and amplitude.

may or may not be harmonically related and which may vary in amplitude and change in content over time (Eggermont 2001) (Fig. 1.2). In pure tones, frequency (the rate of sound wave vibration measured in Hz) correlates with the perception of pitch, and intensity (the amplitude of the sound wave measured in dB) correlates with loudness. All natural sounds are complex, and simple sounds are not typically present in natural environments, however they are widely used as stimuli in neurophysiological studies of the auditory brain. Speech is a complex sound that is both temporally and spectrally modulated.

Complex sounds can be periodic or aperiodic. Periodic sounds consist of a number of frequencies (harmonics) that are related to each other by a common frequency known as the fundamental frequency (F_0). Examples of spectrally complex sounds are vowels and musical notes. Such sounds can be visually represented by a sum of pure tones, however, complex sounds normally produce more or less activity in the brain than the sum of its components (Eggermont 2001). Aperiodic spectrally complex sounds are noise bursts or white noise. These sounds consist of a number of uncorrelated frequencies.

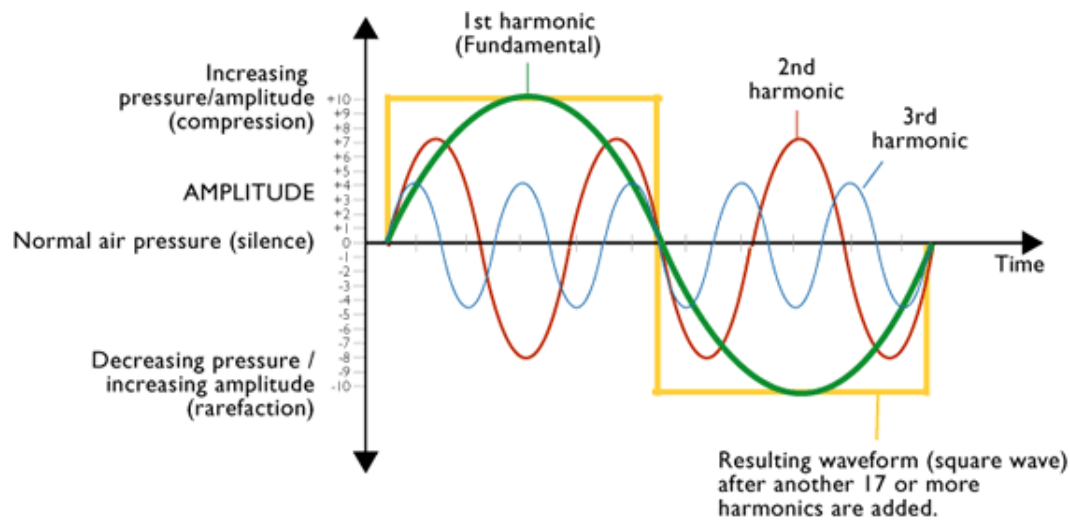


Figure 1.2: An example of a complex sound (square wave) made of a sum of its spectral components, which are simple sinusoids. Each component of the square wave is a harmonic of the fundamental frequency (F_0), whereby their frequency is an integer multiple of the F_0 frequency. Image adapted from www.planetoftunes.com (2015).

1.4.2 Speech Sounds

Human auditory system can detect sounds between 0.06-23 kHz, however, speech sounds take up only a small portion of the audible frequency range, with all speech sounds lying between 0.1-4 kHz (Yost 2005), and speech remaining intelligible even after being put through a bandpass filter of 0.5-1.5 kHz (Denes & Pinson 1993). Speech sounds have a source, which modulates the airflow to create an audible sound. Normally the vocal chords within the glottis are the source of speech sounds. Some speech sounds, however, have multiple sources. For example, /z/ is produced by both vibrations at the glottis and simultaneous frication sounds (Fant 1973). The source produces broadband noise with amplitude decreasing by around 6 dB per octave at frequencies above 300-500 Hz, called the 'radiation characteristic' (Stevens & House 1961). This sound is then filtered by the vocal articulators, such as the tongue and the jaw, to create the sound spectrum characterised by 'formants'. The jaw opening correlates with the frequency of the first

formant (F_1), the second formant (F_2) is most sensitive to the shape of the body of the tongue, and the third formant (F_3) is influenced by the position of the tip of the tongue. The frequencies of the lowest 3-4 formants normally lie in the range of 0.2-3 kHz and form a 'formant pattern'. It has been suggested that formant frequency patterns are the main discriminators of voiced speech sounds, with formant amplitudes and bandwidths playing a lesser role in the process (Klatt 1982).

Speech sounds can be either voiced or voiceless. Voiced sounds are quasi-periodic and are produced by vibrations of the vocal tract folds. Such sounds have a fundamental frequency (F_0) that corresponds to the repetition rate of vocal fold vibrations. Average F_0 of a male voice is around 120 Hz (80-160 Hz range) and of a female voice is around 220 Hz (140-280 Hz range) (Olive et al. 1993). Voiced sounds are vowels (/a/), nasal consonants (/m/), liquids (/r/) and glides (/w/). Unvoiced aperiodic sounds are produced by either: 1) turbulence noise as air flows through an open non-vibrating glottis (aspiration sounds, /h/); 2) turbulence noise as air flows rapidly through a narrow supralaryngeal constriction (frication sounds, /f/); or 3) brief pulse of excitation caused by a rapid change in oral air pressure (transient source sounds, /t/).

Speech sounds consist of two distinct classes of phonemes: vowels and consonants presented at the rate of around 10-12 per second on average in natural speech (Cooper 1980, Maclay & Osgood 1959). Although consonants and vowels have distinct acoustic properties, there is no simple one-to-one relationship between their acoustical features and their identity (Mullennix & Pisoni 1989). Acoustical representations of phonemes vary greatly depending on context. In different contexts identical phonemes can have very distinct acoustical features and different phonemes can have very similar acoustical features. Therefore, it seems unlikely that word identification is done by identifying its individual constituent phonemes first. Instead, speech suprasegmentals may be used, which can either be

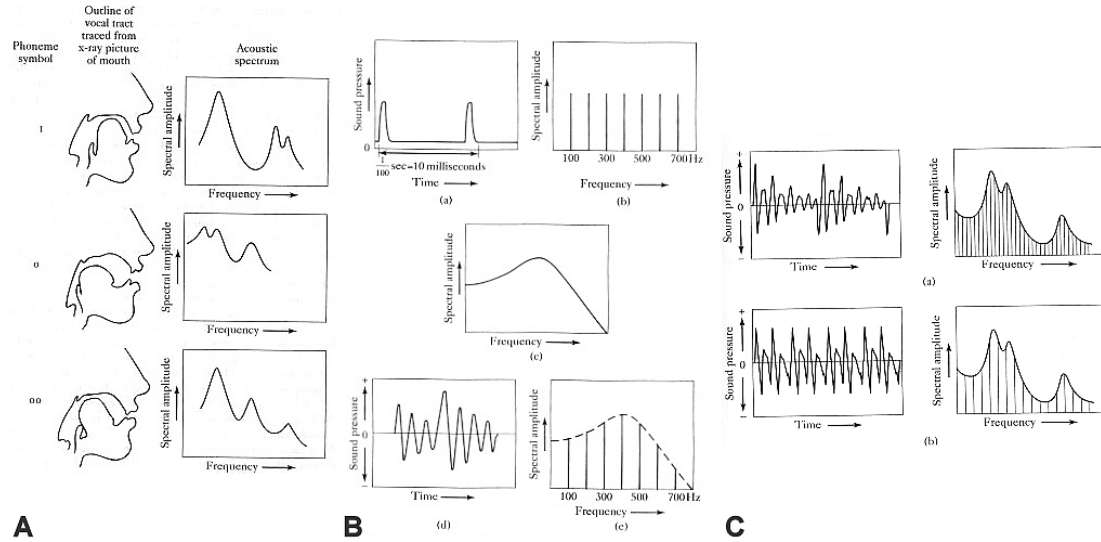


Figure 1.3: A: Vocal tract configurations for three different vowels. Peaks in the vowel spectra (formants) are products of resonances created by the vocal tract. B: a - pulse train as created by the vibrations of the vocal chords; b - pulse train spectrum. In speech this would be the fundamental frequency (F_0) of a vocalisation; c - frequency response of a resonator. In speech this would be created by the shape of a vocal tract configuration; d and e - sound waveform and spectrum created when the pulse train is applied to the resonator. In speech, this would be a vowel. C: waveforms and spectra of the same vowel produced by two different vocal tract frequencies of 90 Hz (a) and 150 Hz (b). Adapted from Denes & Pinson (1993).

larger acoustical patterns corresponding to combinations of phonemes or smaller acoustical patterns corresponding to parts of phonemes (Eggermont 2001).

1.4.2.1 Vowels

Vowels are produced by air passing through the vibrating vocal chord and an open vocal tract, with the shape of the vocal tract determining the identity of the vowel produced. All vowels are periodic sounds that last in the order of several hundred milliseconds. Vowels are characterized by formant patterns as described in the previous section (Fig. 1.3), however the exact formant frequencies can vary greatly for the same vowel depending on the context of the preceding and following phonemes (Fig. 1.4).

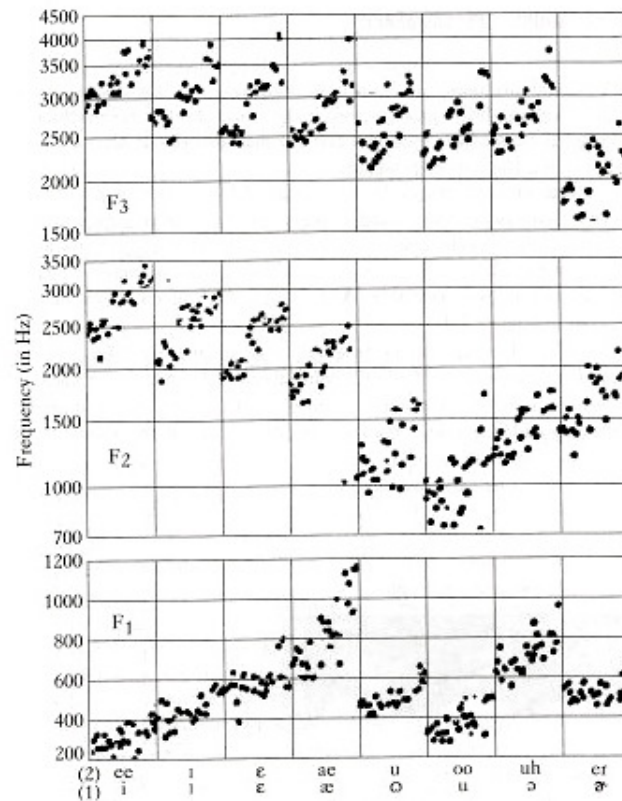


Figure 1.4: Individual formant frequency values of English vowels pronounced by different speakers. F_1 , F_2 and F_3 are the first three formants. Adapted from Denes & Pinson (1993).

Vowel identification is possible when at least two dominant formants (F_1 and F_2) are present (Eggermont 2001). For example Fig. 1.5 shows synthesised vowels /a/, /e/, /ε/, /i/, /o/ and /u/, which are among the most common across the world's languages (Lindblom 1986), and are used by phoneticians as reference sounds to identify vowels in new languages (Jones 1956). Even though only the first two formants were used to synthesise these vowels, participants were able to correctly identify them.

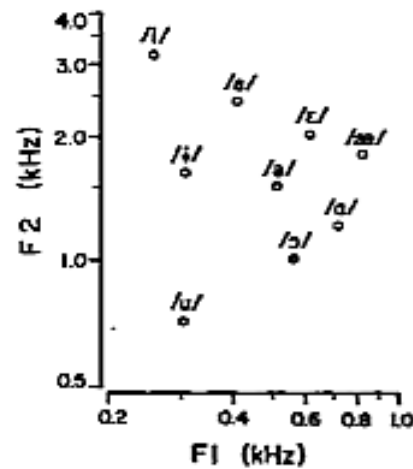


Figure 1.5: Second formant of nine vowels plotted against their corresponding first formants. Adapted from Delgutte & Kiang (1984).

1.4.2.2 Consonants

Consonants are shorter than vowels and can be voiced or voiceless depending on whether they are produced with vocal fold vibration. Voiced consonants are periodic and normally have their spectrum heavily weighted to low frequencies (< 1 kHz), while voiceless consonants are aperiodic and normally have their spectrum weighted to higher frequencies (Rosen 1992). Consonants can be discriminated from vowels by their fast changes in overall amplitude (Stevens 1980, 1981, Stevens & Blumstein 1982), whereby vowels fluctuate at rates of 0.05-0.5 kHz, while consonants fluctuate at rates of 2-10 kHz (Rosen 1992).

Consonant production depends on the degree of constriction in the vocal tract. This is reflected in the direction and rate of frequency change (formant transition) from the onset of the consonant to the onset of the following vowel. These normally occur in frequencies above 1000 Hz (Olive et al. 1993). Formant transitions are used as perceptual cues to consonant identification. Consonants can be plosive ([p], [t], [k]), fricatives ([f], [θ], [s], [ʃ]) or nasal ([m], [n], [ŋ]).

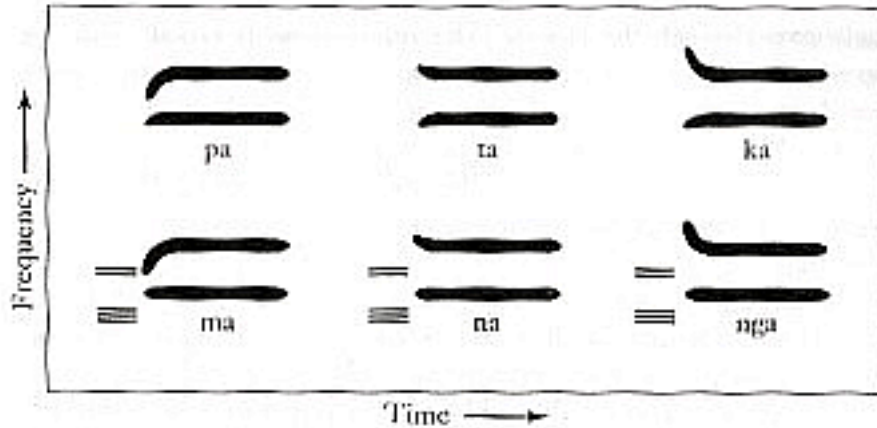


Figure 1.6: Schematic representations of a plosive consonant-vowel and nasal consonant-vowel syllables demonstrating the effect of formant transitions and the place of articulation of consonants on the identity of consonants. Adapted from Denes & Pinson (1993).

Plosive consonants Plosive consonants have a plosive burst centered at a certain frequency and sometimes followed by a second formant transition (FM sweep) of the following vowel (Fig. 1.6). The identification of plosive consonants is independent of the frequency of the plosive burst and depends more on the degree, steepness and direction of the FM sweep and of the following vowel (Denes & Pinson 1993). For example, words 'bait', 'date' and 'gate' are distinguished to a lesser degree by the frequency spectrum of the initial release burst and mostly by the dynamic formant transitions that follow (Hazan & Rosen 1991).

Formant transitions depend on the formant pattern of the vowel as well as the place of articulation of the consonant. For example, for a neutral vowel, where the cross-sectional area of the vocal tract is constant along its length, the formant frequencies are regular intervals of $\frac{c}{4l}$, $\frac{3c}{4l}$, $\frac{5c}{4l}$ and so on, where c is the speed of sound and l is the length of vocal tract. The amplitudes of the spectral peaks of the formants decrease with increased frequency according to the radiation characteristic. When a constriction is introduced to the vocal tract in order to

produce a consonant, according to acoustic theory F_1 is always lowered, while F_2 and F_3 can either be lowered or raised depending on the position and shape of the constriction. All formants then move to return to their original values for the vowel during the consonant duration.

Fricative consonants Fricative consonants are similar to plosives, however, they are longer in duration, resulting in relatively long fuzzy noise segments. If their duration is reduced to around 10 ms in length, they start to sound like plosives ([s] starts to sound like [t]). Fricative consonants have sound levels 10-20 dB lower than vowels (Delgutte 1980).

Although fricatives produce formant modulations when followed by a vowel (Fant 1960), these modulations are not as important for fricatives identification as they are for plosives. Instead, fricative identification depends on the general location of energy bands in stimulus spectra (Harris 1958, Heinz & Stevens 1961). For example, /s/ and /ʃ/ are perceived when the bandpass noise has most energy concentrated near 4-5 kHz with little energy in low frequencies. When the low-frequency cutoff is around 2-3 kHz, the noise is perceived as /ʃ/. When the low-frequency cutoff is around 3-4 kHz, the noise is perceived as /s/. Fricatives /f/ and /θ/ are perceived when weak flat noise is centered around low frequencies, while the perception of /χ/ requires a spectral peak in the 1.5-2.5 kHz region (Harris 1958, Heinz & Stevens 1961). The identification of fricatives is also facilitated by the fact that there are normally no more than five fricatives in a language, compared to the possibility of more than fifteen vowels in a language (Jakobson et al. 1963).

Nasal consonants Nasal consonants are characterised by a low intensity buzz followed by a second formant transition (Fig. 1.6). Lack of buzz makes nasal

consonants sound like plosives ([ma] sounds like [pa]). Nasals have low first formant frequency, broad resonances, and zeros in the spectrum (Fujimura 1962).

1.5 Neurophysiology of the Auditory System

The auditory pathway starts in the cochlea within the inner ear, followed by the auditory nerve and the cochlear nucleus. The information flow is then split into a series of pathways, some of which pass through the superior olivary complex (primarily used for sound localisation), before converging on the inferior colliculus. The auditory pathway then moves to the primary auditory cortex through the medial geniculate body of the thalamus (Boatman 2006) (Figs. 1.7 and 1.8). It is suggested that the auditory pathway performs three main operations: 1) it separates out the individual frequencies that make up a complex sound; 2) it enhances the signal to noise ratio by enhancing the spectral and temporal contrasts of the sound stimulus compared to background noise; and 3) it extracts behaviourally meaningful parameters from the results of the spectral analysis performed by the initial stages of sound processing (Evans 1992).

The early auditory brain areas up to the auditory cortex are tonotopically organised, forming a hierarchy where higher areas receive convergent tonotopic connections from lower areas. Early stages of the auditory brain hierarchy operate on small time scales virtually transcribing the spectral components of sound, while the later stages reflect changes in the stimulus on relatively slow time scales (Covey & Casseday 1999). Throughout the hierarchy of the auditory system responses of neurons change from predominantly sustained to predominantly transient, neurons responding to combinations of simple features appear, and frequency response selectivity of neurons declines (Hromádka et al. 2008, Ito & Fukushima 1988, Kurogi 1991, Wang et al. 2005). The main stages of the auditory brain hierarchy thought to be involved in auditory object recognition are discussed in more detail next, preceded by information on the nature of the neural code used in the brain to represent auditory objects; the role of synaptic plasticity in the auditory brain

hierarchy; and whether animals can be used as appropriate models to study speech processing in humans.

1.5.1 Rate or Temporal Coding

The nature of the neural code used by the auditory brain to represent auditory objects is unknown. The two potential types of encoding are rate and temporal (Fig. 1.9). Rate encoding presumes that auditory object identity information is contained within the overall firing rate of a single neuron or a unique population of neurons within a brain area over a set interval from tens to hundreds of milliseconds in response to an auditory stimulus (Cariani 1995). Temporal encoding involves identifying unique temporal spike patterns within the spike train of a single neuron or over a population of neurons within a brain area in response to an auditory stimulus. Temporal coding has been found in the auditory nerve (AN) (Carianni & Delgutte 1996), in the inferior colliculus (IC) (Delgutte et al. 1998) and in the auditory cortex (Eggermont 1998*b*). Rate and temporal information encoding can be seen as two limits of the same continuum, whereby temporal encoding involves spike timing information calculated over a very short temporal window, while rate encoding involves spike timing information integrated over a very long temporal window.

Rate and temporal codes can exist side by side within the firing of the same neurons, and so can the different types of temporal codes (Fig. 1.10) (Cariani 1995). Different types of information can then be extracted by different post-synaptic neurons, whereby each post-synaptic cell can extract a particular type of information from either one or a number of pre-synaptic cells. This is what happens in the AN, where speech is encoded within the firing rate of single cells (Sachs 1984), synchronised firing rate of populations of cells (Sachs 1984) and inter-spike intervals of either single or populations of cells (Cariani 1999, Carianni & Delgutte

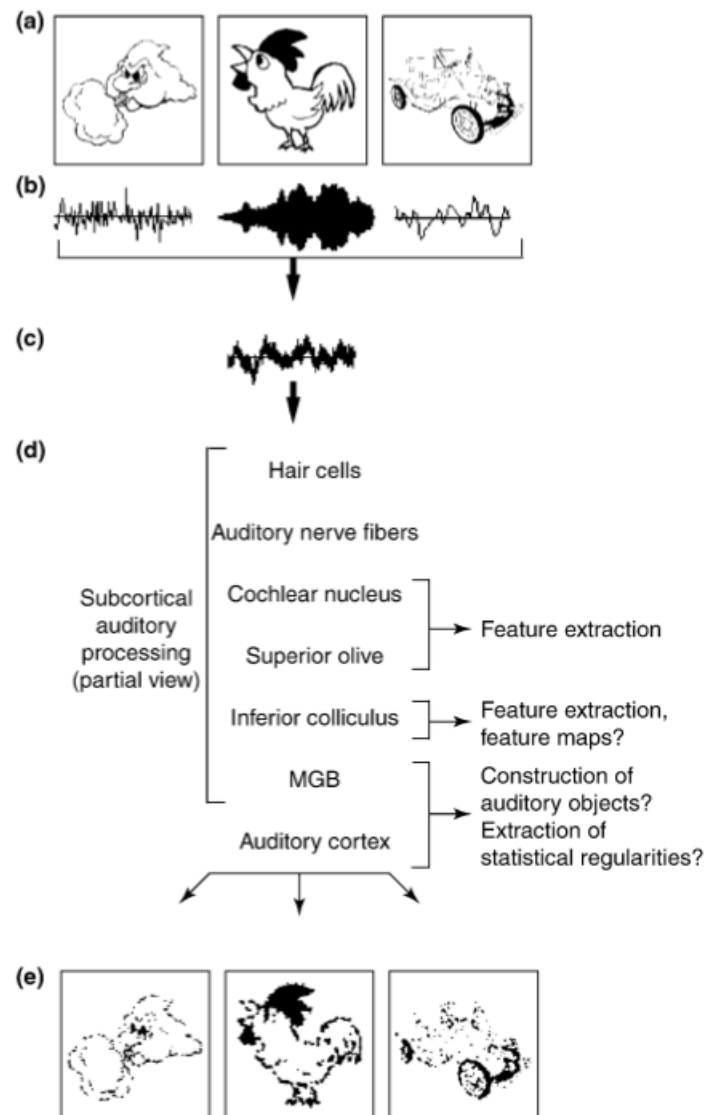


Figure 1.7: Schematic overview of auditory processing in the brain. **a:** multiple sound sources; **b:** sounds produced by the multiple sound sources, including animal vocalisations, man-made sounds and environmental sounds; **c:** input to the auditory system, which is a linear summation of all sounds in the environment; **d:** pre-cortical auditory brain extracts features from sound using a number of different pathways, all of which converge in the inferior colliculus (IC); **e:** auditory cortex transforms pre-processed sound features into auditory object representations (adapted from Nelken (2004)).

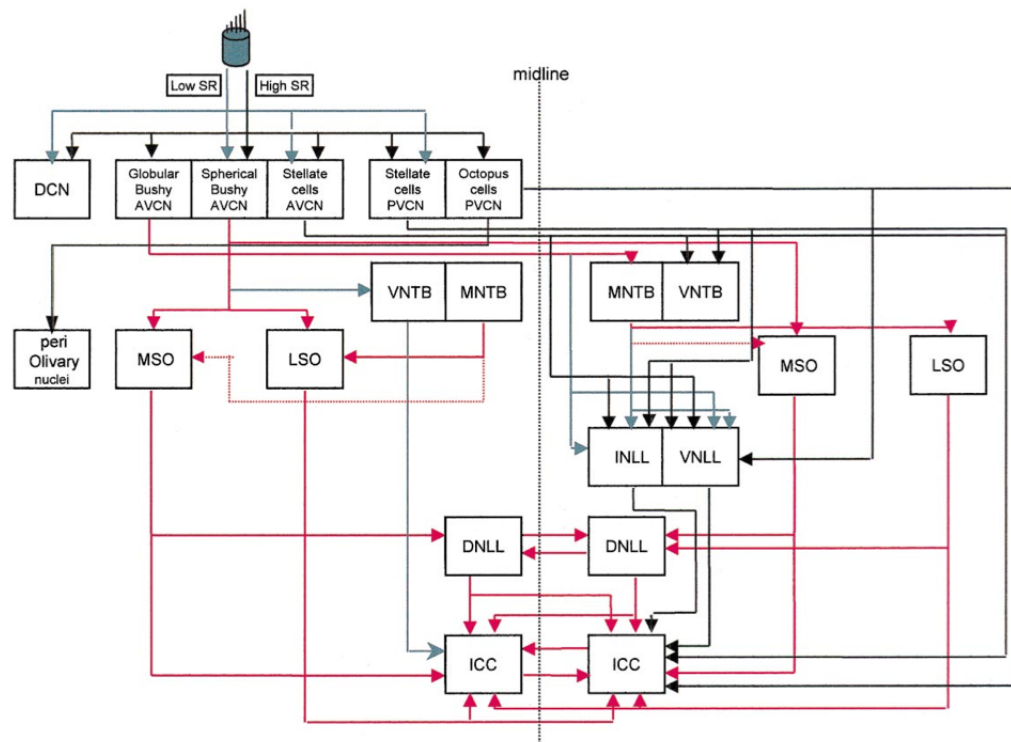


Figure 1.8: Schematic representation of the afferent auditory pathways up to the IC. Red pathways are involved in sound localisation, black pathways are involved in sound identification, grey pathways are involved in both (adapted from Eggermont (2001)).

1996). The different types of information are then read out by different types of neurons in the CN, with stellate cells extracting the firing rate information, and bushy and octopus cells extracting the firing synchrony information (Eggermont 2001).

Suga (1978) suggested two ways in which information about the identity of auditory objects can be encoded in the brain. The detector theory suggests that the identity of an auditory stimulus is encoded within the activity of a specialised set of neurons that fire selectively to that stimulus. The spatio-temporal pattern theory suggests that different auditory objects are represented by unique patterns of activity within populations of non-specialised neurons, whereby each neuron

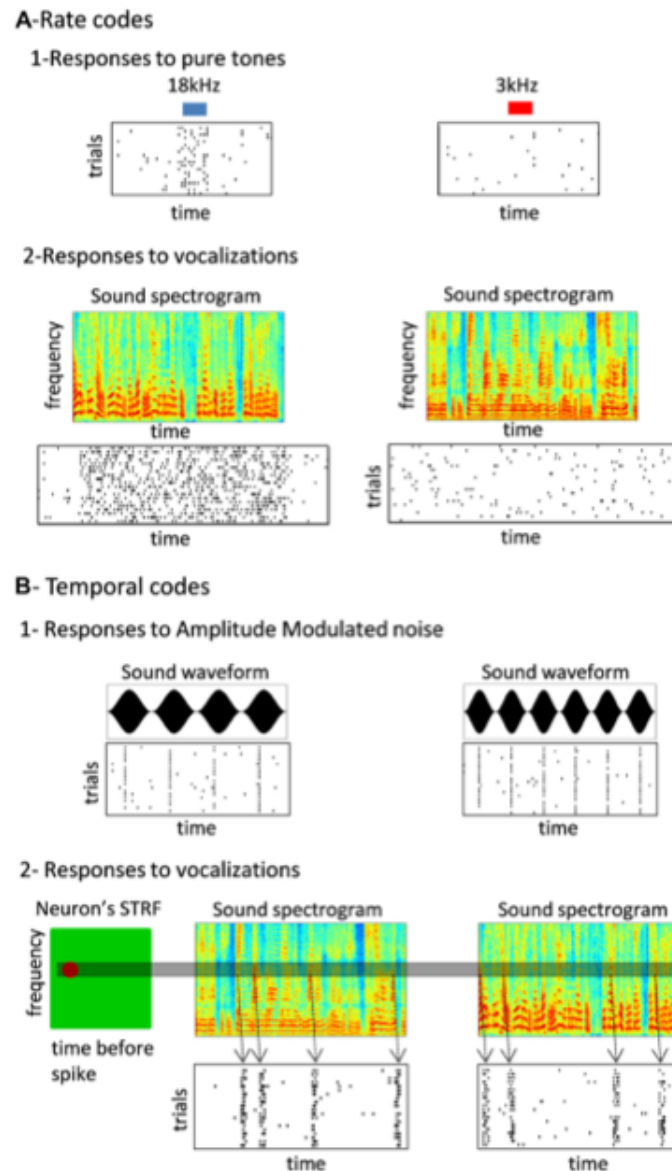


Figure 1.9: Schematic representations of rate (A) and temporal (B) encoding theories. A: Response properties of a hypothetical neuron according to the rate theory to a pure tone (1) and a natural vocalisation (2) stimuli. The neuron is expected to increase its firing rate in a Poisson-like way to some stimuli, but not others. The only information for stimulus discrimination is the average rate of firing of the neuron. B: Response properties of a hypothetical neuron according to the temporal theory to an amplitude-modulated (AM) (1) and a natural vocalisation (2) stimuli. The neuron phase-locks its firing to the phase of the AM sound or to the presence of a particular frequency in a natural vocalisation, thus providing information for stimulus discrimination within its firing pattern. Adapted from Huetz et al. (2011).

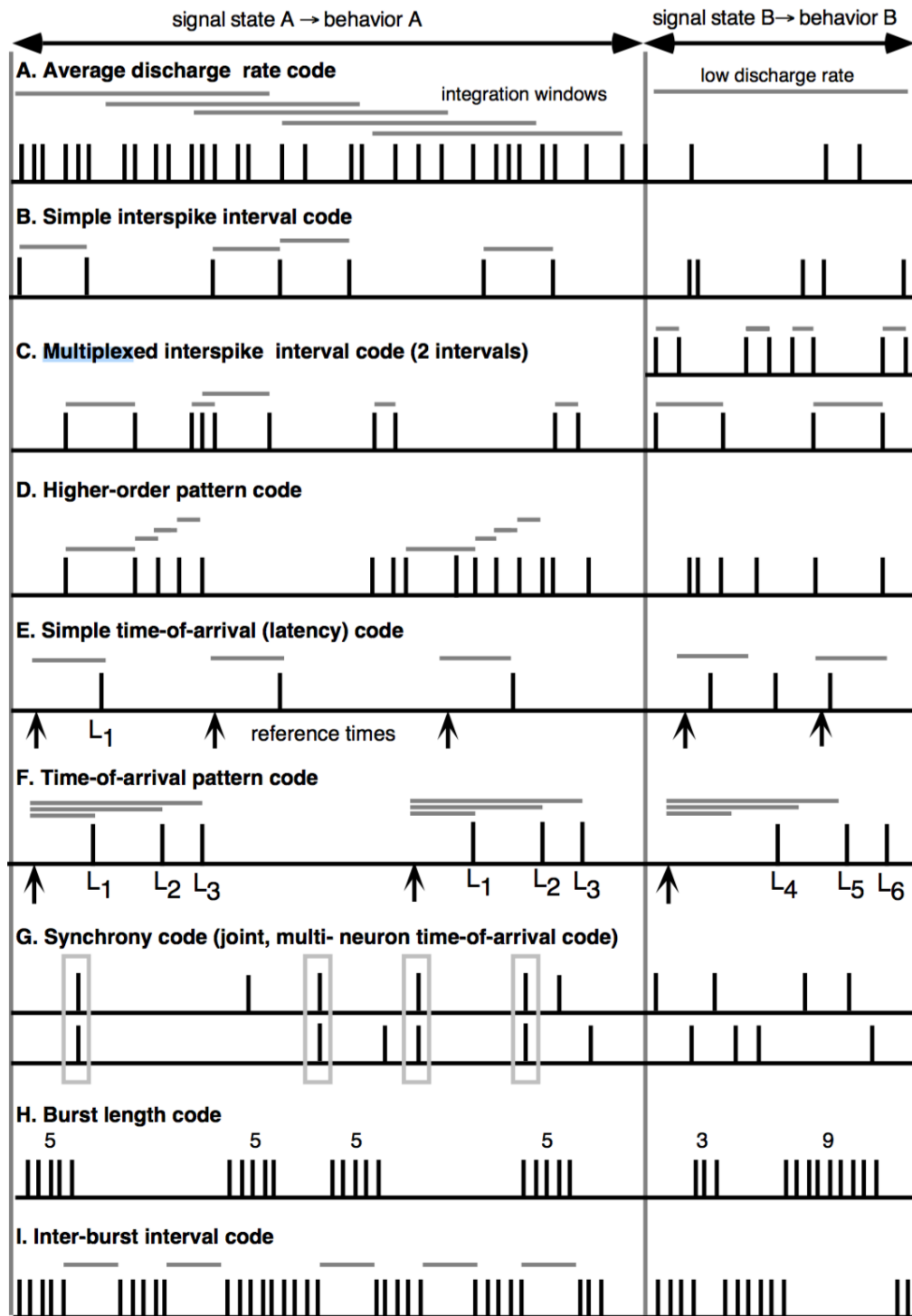


Figure 1.10: Schematic representation of some possible neural codes. Adapted from Cariani (1995).

might fire to a number of auditory objects. The detector theory assumes rate encoding, while the spatio-temporal pattern theory assumes temporal encoding to take place in the brain.

Another hypothesis is that rate and temporal codes dominate in different parts of the auditory brain hierarchy, whereby temporal encoding is used in the periphery of the auditory brain and rate encoding is used at the later stages, such as the auditory cortex (Eggermont 1998*a*). This idea is supported by the fact that the phase-locking property of neurons diminishes over the auditory brain hierarchy (Eggermont 1997, Frisina et al. 1985, Joris & Yin 1992, Rees & Møller 1983, Rouiller et al. 1981), with only very limited phase-locking remaining in the auditory cortex (Langner 1992). It is suggested that the transformation from temporal to rate encoding happens in the IC (Langner 1992). Binary encoding, whereby cells fire once at the onset of their preferred stimulus, has been described in the auditory cortex in response to simple tonal stimuli by DeWeese & Zador (2003).

1.5.2 Plasticity in the Auditory Brain

Learning in the brain is important to optimise the animal's behaviour within an environment. Learning is driven by the presence of activity dependent changes in the strength of synaptic connections between cells (Tzounopoulos & Kraus 2009). These changes can arise either through synaptic plasticity, whereby the connection strength between two neurons can be modified, if activity in one neuron drives activity in the other (Bi & Poo 1998, Hebb 1949), or through intrinsic plasticity, whereby long-lasting changes in the intrinsic excitability of neurons are induced by changing the function of the cells' voltage-gated ion channels (Trussell 1999, vonGersdorff & Borst 2002).

Traditionally it was believed that learning only happened in the auditory cortex, while the role of the subcortical brain areas was to pre-process and prepare

sensory information for the cortex (Atiani et al. 2009, Fritz et al. 2007, Schreiner & Winer 2007, Weinberger & Diamond 1987, Weinberger 2007, Weinberger et al. 1990). It was believed that the characteristics of auditory brain stem responses to auditory stimuli (they specialise in generating fast, reliable, and consistent signals) made the auditory brain stem areas ill suited for activity dependent learning. More recent evidence, however, suggests that experience-dependent learning is present in some auditory brain stem areas, such as the dorsal cochlear nucleus (synaptic plasticity) (Fujino & Oertel 2003, Tzounopoulos et al. 2004, 2007), and the superior olivary complex nuclei (intrinsic plasticity in MNTB, MSO and LSO) (Song et al. 2008, 2005, Steinert et al. Neuron). These areas of the brain stem, however, are primarily specialised in sound localisation rather than identification. Such brain stem learning is, therefore, not included in the models described in this thesis.

Neurophysiological evidence suggests that synaptic plasticity is present in the IC. It has been found that tuning curves and bandwidths of IC neurons can be affected by distorted acoustic environments during a critical period (Clopton & Winfield 1976, Poon & Chen 1992, Sanes & Constantine-Paton 1983, 1985). The existence of synaptic plasticity in the MGN has also been investigated, however, evidence has been inconclusive. It has been found that 15/57 recorded MGN cells showed a change in their response properties following classical conditioning in un-anaesthetised rats (Edeline 1990). These changes, however, may have been the result of cortical feedback connections, whereby the learning happened in the cortex and the effects in the thalamus were only modulatory. On the other hand, MGN lesions, but not cortical lesions, were found to prevent classical conditioning in frequency discrimination (Jarrell, Gentile, McCabe & Schneiderman 1986, Jarrell, Romanski, Gentile, McCabe & Schneiderman 1986, Ledoux et al. 1986, 1983).

Although relatively little data has been collected on neural plasticity in the subcortical auditory brain areas, evidence for neural plasticity in the auditory cortex is vast. Cortical sound representations have been reported to be shaped through early auditory experience (de Villers-Sidani E. et al. 2007, Zhang et al. 2002), whereby unpatterned acoustic experience during a critical period has been found to have similar developmental effects as no sensory input at all (Chang & Merzenich 2003). Furthermore, when visual input was re-wired to the auditory cortex through the auditory thalamus in infant ferrets and hamsters, it was found that the auditory cortex developed response properties appropriate for vision, such as feature-selective visual responses, retinotopic maps and orientation columns (Frost 1990, Sharma et al. 2000, von Melchner et al. 2000), suggesting that functional organisation of the auditory cortex depends on the spectro-temporal structure of the input stimulus. It was also reported that early speech experience in humans shapes the language-specific perception of phonemes, whereby contrasts between native speech sounds are enhanced, and contrasts between foreign speech sounds are reduced (Kuhl et al. 2006, 1992, Werker & Tees 1984).

Although most research on synaptic plasticity emphasises the importance of the early sensory experience, it has been found that cortical plasticity is present in adult animals too, both when attending behaviourally important sensory stimuli (Doupe & Kuhl 1999, Kilgard & Merzenich 1998, Pantev et al. 1998, Weinberger 1995) and through passive listening (Eggermont 2008).

1.5.3 Generalising from Animal Studies

Most of the auditory brain neurophysiological data comes from studies using anaesthetised animals. It is, therefore, important to know whether anaesthesia qualitatively affects neuronal response properties in the auditory brain; how well the anatomy of the auditory brain is matched between species; and whether data

collected from animals can be generalised to humans, especially with respect to speech recognition.

Studies investigating the effects of anaesthesia on the response properties of the auditory brain neurons have found that, although anaesthesia reduces inhibitory effects (Palmer et al. 1986) and stimulus level representation in the presence of background noise (May et al. 1991), it does not significantly alter neuronal response properties (Egorova et al. 2001, Peake et al. 1962, Willott et al. 1988).

While differences in the structure of the auditory brain exist between different species, successful work has been done to find and map auditory brain homologues (see Rutkowski et al. (2002), Stecker et al. (2003) for homologies between primates and non-primates, and Rauschecker (1998) for human homologues of animal auditory cortex areas). Furthermore, single-unit recording data from such areas as the AN, CN and IC have been found to be comparable across many species of mammals, including primates (Evans 1975, Portfors & Felix 2005, Rhode et al. 2010, Rose et al. 1967). Comparisons have also been made with data collected from humans, whereby AN compound action potentials in response to click and tone burst stimuli have been found to be similar in humans and other mammals (Eggermont 1976, Kiang & Peake 1960, Stephens et al. 1975); and primates and humans have been found to have the same anatomical CN cell types displaying the same physiological responses (Evans & Nelson 1973, Feng et al. 1993, Ostapoff et al. 1994, Palmer et al. 2003, Rhode & Greenberg 1992, Rhode et al. 1983, Rhode & Smith 1986, Young & Brownell 1976).

Different views exist on whether data collected from animals is appropriate to gain an understanding of how humans perform speech recognition. One view is that different auditory systems adapted to their species-specific auditory stimuli throughout evolution. Thus, in order to classify neurons correctly, it is important to use the appropriate auditory stimuli for the species studied, whereby it

does not make sense to study neurophysiological properties of auditory neurons in cats using human speech input (Ohlemiller et al. 1994). An alternative view is that human speech is not old enough in evolutionary terms to have affected the auditory systems of the vertebrates. It is, therefore, argued that the human auditory brain is not qualitatively different from the auditory systems of other vertebrates, and that speech developed according to the constraints of the fully developed auditory and vocalisation systems, rather than the other way round (Eggermont 2001). A theory of co-evolution of language and the physiology of the auditory brain hierarchy has also been put forward (Deacon 1997). It was suggested that the better articulation abilities in humans due to the dropped larynx put evolutionary pressure on the auditory brain to grow and develop in order to perform more complicated processing of the more complicated vocalisations. The relative enlargement of the auditory brain hierarchy in humans compared to other animals, however, is primarily restricted to the prefrontal cortex. This suggests that at least the early parts of the auditory brain hierarchy should be comparable between humans and animals.

In addition to the points discussed above, it has been suggested that all SSVs, including human speech, have similar acoustic properties (Patterson & Pepperberg 1994, Rauschecker et al. 1995, Suga 1992), which suggests that data obtained from different species might be broadly generalisable. It has been proposed that all animal vocalisations, including human speech, are made of three basic information bearing elements, such as steady-state harmonically related frequencies, frequency modulations and noise bursts (Fitch et al. 1997, Mullenix & Pisoni 1989, Suga 1988, 1989, 1992). While human speech is more complex and variable than animal vocalisations, with human vocabularies consisting of tens of thousands of words (de Saussure 1959, Hockett 1958, Tomasello 2005), it is suggested that the early

stages of SSV processing are shared, whereby humans and animals have similar lower levels of auditory processing (Fitch et al. 1997).

Evidence to support the idea that human and animal auditory systems respond to speech in a similar manner comes from neurophysiological and psychophysical studies comparing the performance of humans and animals in speech discrimination tasks. For example, human adults, infants, and babies, as well as monkeys, chinchillas and cats have all been found to have the same categorical perception of the voice onset time (VOT) characteristic of consonants (Eimas et al. 1971, Kuhl & Miller 1978, Kuhl & Padden 1982, Kuhl & Miller 1975, Sinnott & Adams 1987, Wong & Schreiner 2003). VOT is one of the main discriminating characteristics of consonants and corresponds to the gap between the consonant release time and the start of the following vowel. Apart from similarities in VOT perception, some other abilities believed to be important for speech recognition, such as perceptual constancy, category formation and non-monotonicities in acuity, have been found to be shared between humans and other animals (Dooling et al. 1989, Kluender et al. 1987, Kuhl 1981, 1986, 1987, Kuhl & Miller 1978, 1975, Miller 1977).

1.5.4 Cochlea

The cochlea is a part of the inner ear and is a fluid filled coiled structure about 35 mm long in humans, which receives sound waves caught by the pinna (outer ear) and funnelled down the external canal. The received sound wave first vibrates the tympanic membrane, which propagates the vibrations through the three ossicles in the middle ear to the oval window of the cochlea. The movement of the oval window moves the fluid inside the cochlea, which displaces the basilar membrane innervated by around 30,000 nerve fibers, thus transforming mechanical movements inside the cochlea into the cell firing pattern within the auditory nerve (Pickles 1988).

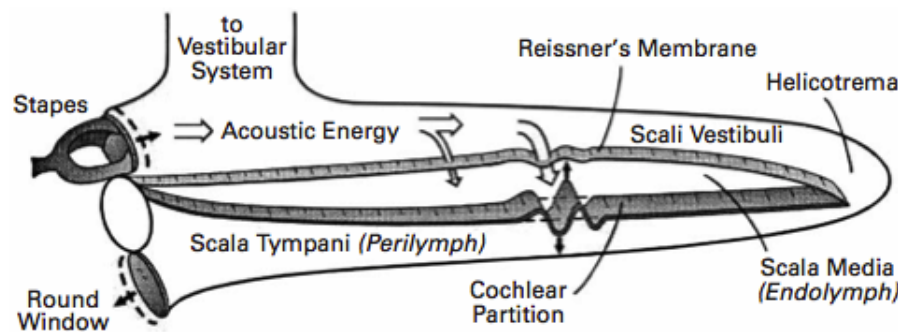


Figure 1.11: A schematic representation of an uncoiled cochlea showing basilar membrane displacement in response to a pure tone. The displacement results in movement of inner hair cells (IHC) attached to the area, which in turn results in depolarisation of auditory nerve fibers attached to the IHCs. Adapted from Geisler (1998).

Sound arrives to the cochlea as a complex, time-varying frequency signal, and the role of the cochlea is to act as a band-pass filter by breaking down the sound into narrow frequency bands (von Békésy 1960). Complex sounds cause a pattern of vibration in the basilar membrane equivalent to the superposition of the vibrations generated by the individual tones making up the complex sound. As the sound waves displace the basilar membrane in the cochlea, inner hair cells (IHC) attached to the basilar membrane become activated. These are tonotopically organised and with high frequencies represented at the basal and low frequencies represented at the apical end of the basilar membrane (Fig. 1.11). IHCs translate mechanical energy of the displacement of the basilar membrane into the electrical energy of the firing of the auditory nerve neurons through purely mechanical transduction channels (Hudspeth 1989). Hair cells are hyperpolarised and depolarised in synchrony with sound frequency and this is translated to the AN fibers with very little loss of time resolution (Hudspeth 1999, Trussell 1997).

1.5.5 Auditory Nerve

The auditory Nerve (AN) consists of Type 1 and Type 2 neurons. Type 1 neurons are myelinated, large and bipolar and constitute 90-95% of AN. Each Type 1 neuron contacts 1 IHC in the cochlea, while each IHC gets around 10-20 Type 1 neurons (Ehret & Romand 1996). Type 2 neurons are non-myelinated, small, pseudo-bipolar and constitute 5-10% of all AN fibers. Type 2 neurons contact outer hair cells in the cochlea which are involved in controlling the stiffness of the basilar membrane in order to enhance the dynamic range of perceived sound amplitudes and are not involved in the actual perception of sound (Ehret & Romand 1996). Type 2 neurons are therefore not discussed further.

Type 1 AN fibers are usually classified by their spontaneous firing rate (SR) into three groups: high-SR, medium-SR and low-SR (see Table 1.1). It has been suggested that the different SRs of AN fibers arise from morphological differences in AN neurons and their synapses from IHCs (Liberman 1980, 1982). Low- and medium-SR fibers have 15-30 dB higher thresholds and 30-40% narrower bandwidths than high-SR fibers (Liberman 1978, Temchin et al. 2008). They are also more robust in noisy conditions and contain more robust temporal and rate information about the auditory stimulus than high-SR fibers (Silkes & Geisler 1991).

Type 1 AN Fibers	Spontaneous Activity	Proportion	Threshold
High	≥ 18 spikes/s	66.1%	16.0 ± 11.4 dB
Medium	1-18 spikes/s	23%	21.4 ± 12.3 dB
Low	≤ 1 spikes/s	10.9%	28.2 ± 12.4 dB

Table 1.1: Properties of Type 1 Auditory Nerve Fibers (Temchin et al. 2008)

1.5.5.1 Tuning Curves

AN fibers are tonotopically organised with each fiber selectively responding to a particular range of frequencies. This frequency selectivity is defined through overlapping tuning curves, so that a pure tone of a particular frequency activates more than one AN neuron. The edge of the excitatory area of a tuning curve defines the amplitudes required at each frequency to produce equal responses in the neuron, which is 20% higher than the neuron's spontaneous rate (Sachs 1969) (Fig. 1.12). The peak of the inverted Gaussian of a tuning curve corresponds to the neuron's characteristic frequency (CF), which is the frequency of the pure tone that elicits the cell response at the lowest sound amplitude. Neurophysiological evidence suggests that neurons with a CF below 5 kHz have symmetrical excitatory tuning curves, while neurons with higher CFs have skewed tuning curves with long tails within the low frequency-high amplitude region (Ehret & Romand 1996, Evans 1975). The frequency tuning of AN fibers is level-dependent, whereby neurons respond to more frequencies as sound intensity increases (Covey 2000). The bandwidth of neurons' tuning curves has been reported to vary depending on the neurons' CF and threshold, whereby neurons with higher CF and lower threshold have slightly narrower tuning curves (Silkes & Geisler 1991).

1.5.5.2 Amplitude Saturation

Neurophysiological evidence suggests that the maximum firing rate of AN fibers is around 200 spikes/s (Eggermont 2001, Sachs & Abbas 1974, 1976). The AN fiber firing rate increases in a non-linear fashion until it reaches saturation over a dynamic range of 20-50 dB starting at the threshold defined by the excitatory tuning curve (Sachs & Abbas 1974, 1976, Viemeister 1984) (Fig. 1.13).

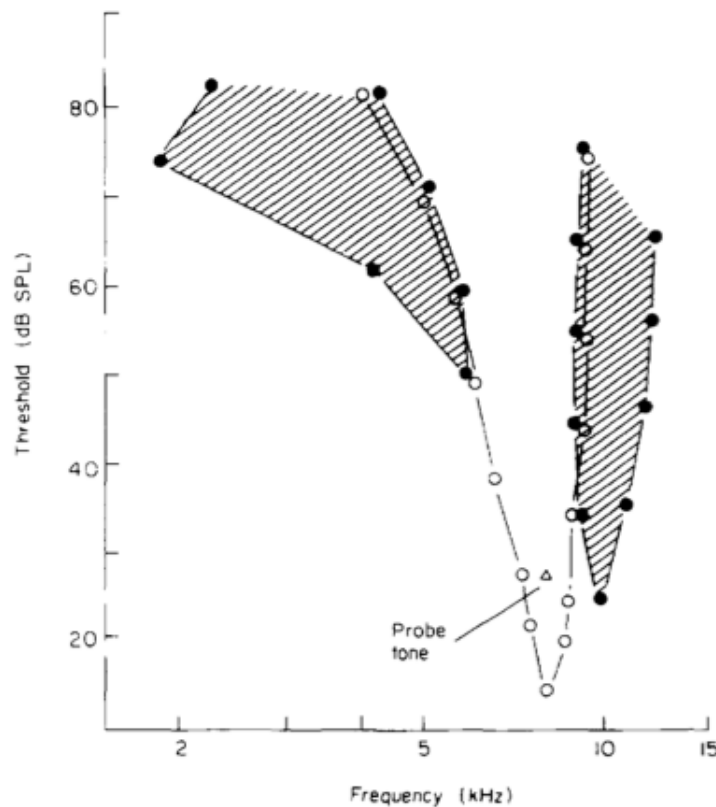


Figure 1.12: An example tuning curve (open circles) with suppressive areas (closed circles). Adapted from Kates (1995)

1.5.5.3 Adaptation

AN fibers have been shown to be prone to short-term adaptation, whereby a fiber's response to a stimulus is diminished if the fiber has been active shortly beforehand in response to a context stimulus. The effects of short-term adaptation last around 1-2 ms and are the strongest when the context stimulus has greater duration or intensity, when it is close to the AN fiber CF, and when the silent interval between the context stimulus and the test stimulus is shorter (Delgutte & Kiang 1984, Harris & Dallos 1979, Smith 1977).

Short-term adaptation means that AN response properties to speech sounds are

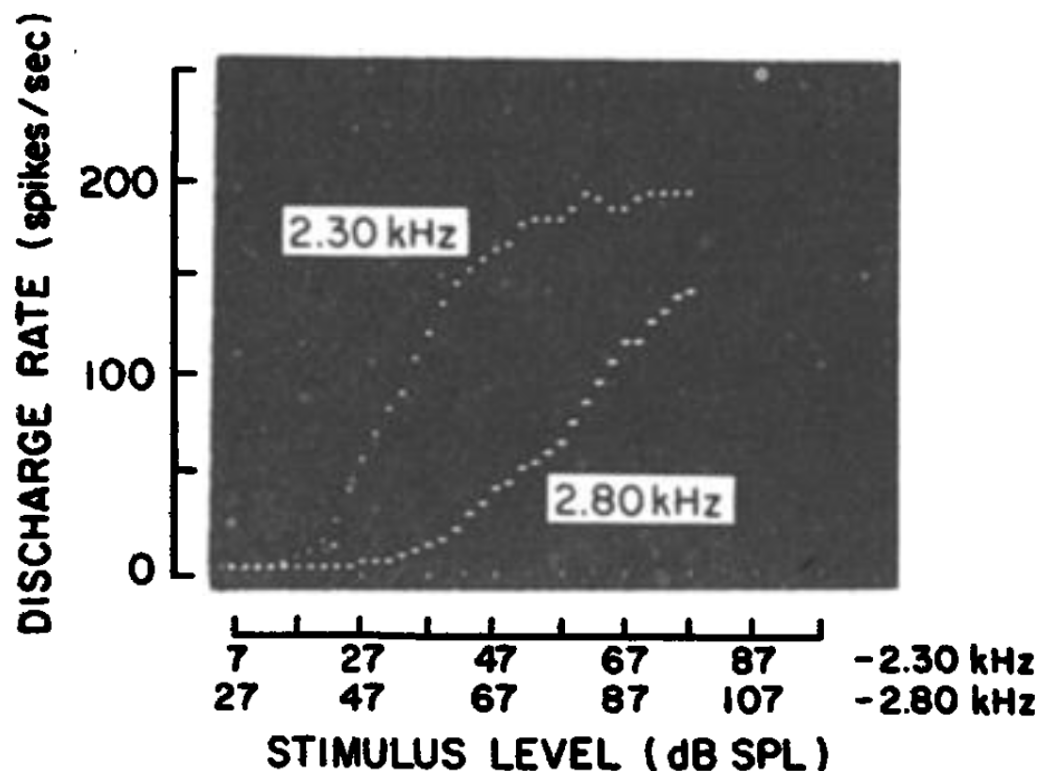


Figure 1.13: Neuronal discharge rate versus stimulus level functions of AN responses to 2.3 and 2.8 kHz tone bursts. Sound level was increased by 2 dB between each two points. Adapted from Sachs & Abbas (1974)

affected by the context. For example, the distinction between phonemes /f/ and /tʃ/ is based on the duration of the frication noise rise time or depending on the length of silence between the fricative and its preceding vowel (Dorman et al. 1979). A longer silence allows AN neurons to recover from adaptation and to respond to the fricative with more prominent peaks in their onset discharge rates. This is similar to a response to a decreased rise time of the frication noise. Adaptation is also present in response to steady-state stimuli, whereby each neuron's firing rate diminishes with time in response to a continuous simple tone stimulus (Delgutte & Kiang 1984, Sinex & McDonald 1988). Furthermore, it was found that background noise results in AN fiber adaptation, however, it was found that, while noise reduces

the AN fibers' discharge rate, it does not affect the temporal response patterns within the AN (Delgutte 1980).

1.5.5.4 Two-Tone Suppression

Each AN fiber has two inhibitory areas within its tuning curve, which slightly overlap the excitatory area of the fiber's tuning curve both above and below the fiber CF (Sachs & Kiang 1968) (Fig. 1.14). Sounds within the inhibitory areas suppress the neuron's firing rate, whereby a fiber's response to a pure tone from within the excitatory area is reduced, when another tone from within the inhibitory area is present. This phenomenon is called two-tone suppression, and it has been reported to produce inhibitory effects equivalent to a 40-50 dB reduction in the excitatory tone amplitude (Delgutte 1990, Duifhuis 1980). The level of suppression is higher if the suppressor frequency is below the fiber CF, and the fiber CF is above 2 kHz (Delgutte 1990). Furthermore, inhibition is stronger towards the middle of the inhibitory areas, where it reaches the maximum level, and is weaker at the edges, where it can be hardly noticeable (Arthur et al. 1971). It has also been found that, as the suppressor tone amplitude increases, the magnitude of two-tone inhibition also increases over a 15-20 dB range until it reaches its maximum value (Kates 1995). Two-tone suppression is caused by the mechanical properties of the basilar membrane in the cochlea (Cooper & Rhode 1992, Nuttall & Dolan 1993, Rhode 1978, Robles et al. 1991). Sounds from within both the excitatory and the inhibitory areas of the tuning curve excite the neuron when presented in isolation, however, they exhibit two-tone suppression when presented together (Fig. 1.14).

Although two-tone suppression is normally described in terms of two tones, a number of studies looked at the inhibitory interactions between a larger number of tones presented simultaneously (Stelmachowicz et al. 1982, Weber 1984). These studies found that the magnitude of inhibition is equivalent to the largest

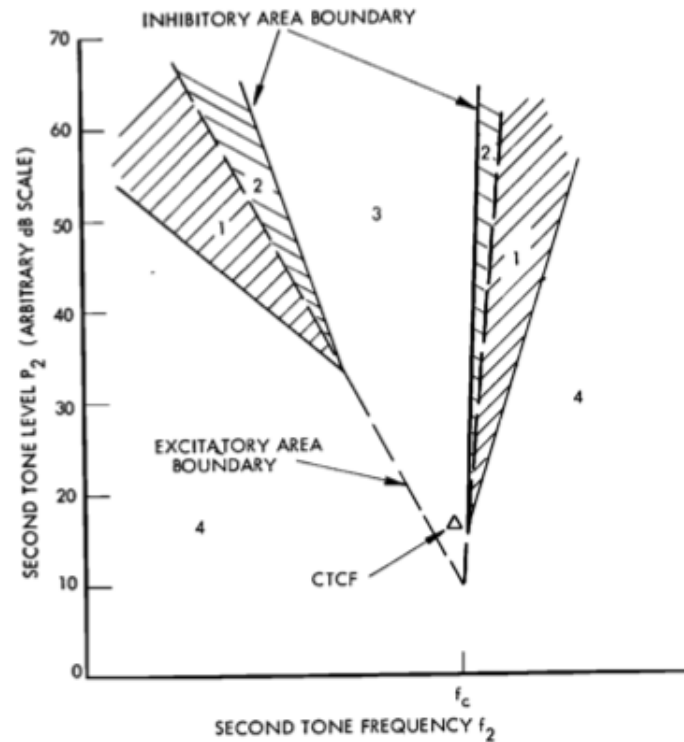


Figure 1.14: Schematic representation of the excitatory and inhibitory areas of an auditory nerve tuning curve. F_c is neuron's characteristic frequency with threshold being 10 dB. Single pure tones presented in areas 2 and 3 will elicit a response in the cell. Single pure tones presented in areas 1 and 4 will not elicit a response in the cell. A pure tone presented in area 4 simultaneously with another pure tone presented in areas 1 or 2 will elicit a suppressed response in the cell compared to the cell's response to the same tone presented in area 4 alone. This phenomenon is called two-tone suppression. Adapted from Sachs (1969).

magnitude of inhibition produced by an individual suppressor tone when multiple suppressor tones are present. Two-tone inhibition effects are relevant to speech processing, since they may result in the enhancing of formant peaks in a vowel (Houtgast 1974), and in potentially reducing AN fibers' response rates to vowel F_2 and F_3 due to the inhibition produced by F_1 (Delgutte 1990, Sachs & Young 1980).

1.5.5.5 Phase-Locking

Neurophysiological evidence suggests that the spike times of AN fibers in response to a pure tone are synchronised to the peaks of the sine wave corresponding to the pure tone, a property called phase-locking (Fig. 1.15 (a)) (Brugge et al. 1969, Carianni & Delgutte 1996). Since individual AN fibers are unable to fire more than 200 spikes/s (Eggermont 2001, Sachs & Abbas 1974, 1976), they can not fire to every peak of sound frequencies over 200 Hz. A population of AN fibers, however, can provide information about sound frequencies of up to 4-5 kHz (Brugge et al. 1969, Covey 2000) due to the Volley Principle suggested by Wever & Bray (1930), whereby each neuron in a population phase-locks to a different set of amplitude peaks of the sound wave (Fig. 1.15 (b)).

The phase-locking property of the AN fibers is important for the temporal coding of information in the auditory system. It involves identifying the stimulus periodicity by considering the inter-spike intervals of AN fibers without the need of knowing their CF (Carianni & Delgutte 1996). AN fibers preserve both the fine structure and the envelope of sounds (Joris et al. 2004) (Fig. 1.16). Furthermore, when presented with a complex tone, AN fibers phase-lock to the amplitude peaks of that complex tone, instead of phase-locking to the amplitude peaks of pure tones that make up the complex tone (Brugge et al. 1969)

1.5.5.6 Responses to Vowels

AN response properties to vowels have been described by Secker-Walker & Searle (1990), who found that AN fibers with higher CFs started to fire 3-4 ms earlier than those with lower CFs in response to a synthesised syllable /da/ (Fig. 1.17). Furthermore, different sub-populations of AN fibers emerged based on their phase-locking properties. They found that AN fibers with the lowest (< 350 Hz) and the highest (> 3.64 kHz) CFs synchronised to F_0 with distinct increases in activity

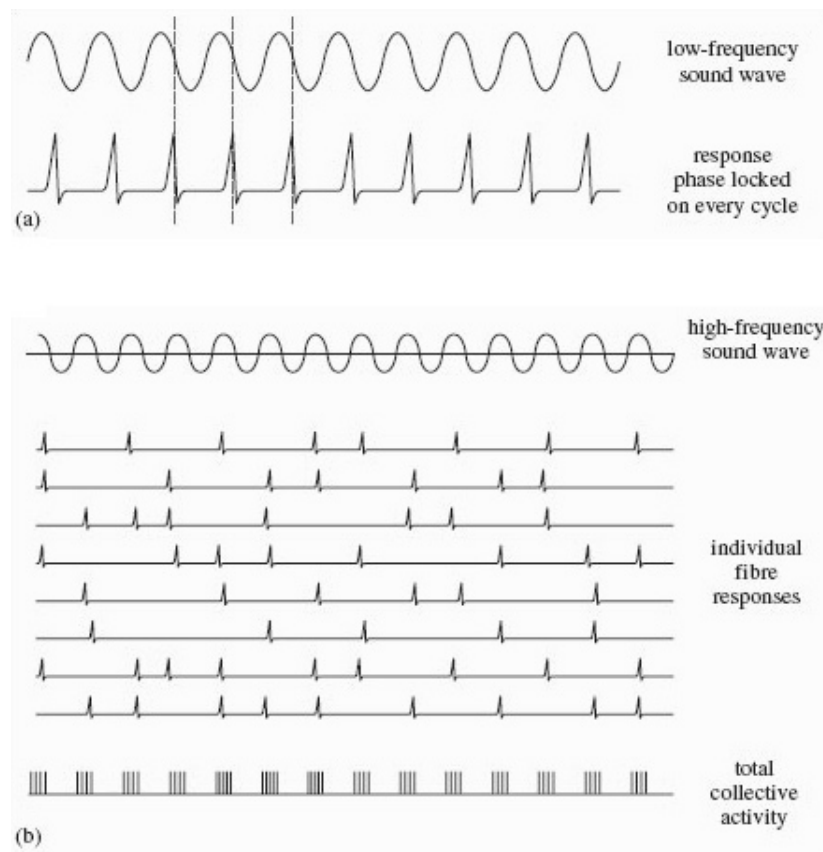


Figure 1.15: A: schematic representation of phase-locking in AN fibers, whereby spike timings of the cells correlate to the peaks in the sine wave corresponding to the frequency of the input sound. B: schematic representation of the Volley Principle in AN fibers, whereby each neuron in a population phase-locks to a different set of amplitude peaks of the sound wave. Therefore although there is not enough information about the input sound frequency in inter-spike intervals of each individual AN fiber, this information is found when the whole population of AN fibers responding to the particular frequency is considered.

every 8 ms, which corresponded to the glotal closures or the fundamental frequency (F_0) of the syllable. There were also three groups of neurons that phase-locked to the three formants of the vowel /a/. Fibers with CFs between 0.45-1.2 kHz synchronised to F_1 , between 1.2-2 kHz synchronised to F_2 and between 2-3.64 kHz synchronised to F_3 and F_0 .

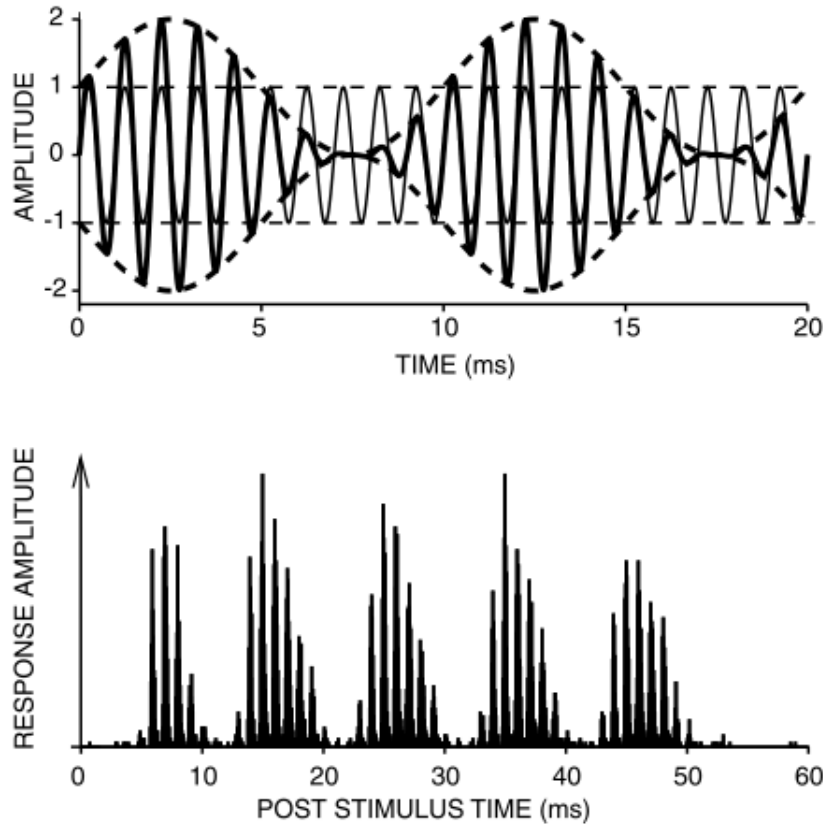


Figure 1.16: Top: superimposed waveforms of an unmodulated 1000 Hz wave (thin line) and the same wave amplitude modulated at 100% by a 100 Hz tone according to the formula $s(t) = [1 + m \sin(2\pi f_m t)] \sin(2\pi f_c t)$, where f_c is the carrier tone, f_m is the modulation tone and m is the modulation depth (thick line). Dashed line represented the envelope of the sound. Bottom: post-stimulus time histogram (PST) of an AN fiber to the amplitude-modulated tone above. It can be seen that both the fine structure and the envelope of sounds are preserved in AN response patterns. Adapted from Joris et al. (2004).

The quality of AN phase-locking was also estimated by measuring the interpeak intervals (IPIs) of AN fibers. IPIs estimated which frequency each AN fiber was phase-locked to at each point in time by measuring the time δs between the pairs of subsequent peaks within the instantaneous firing rate curve of each AN fiber

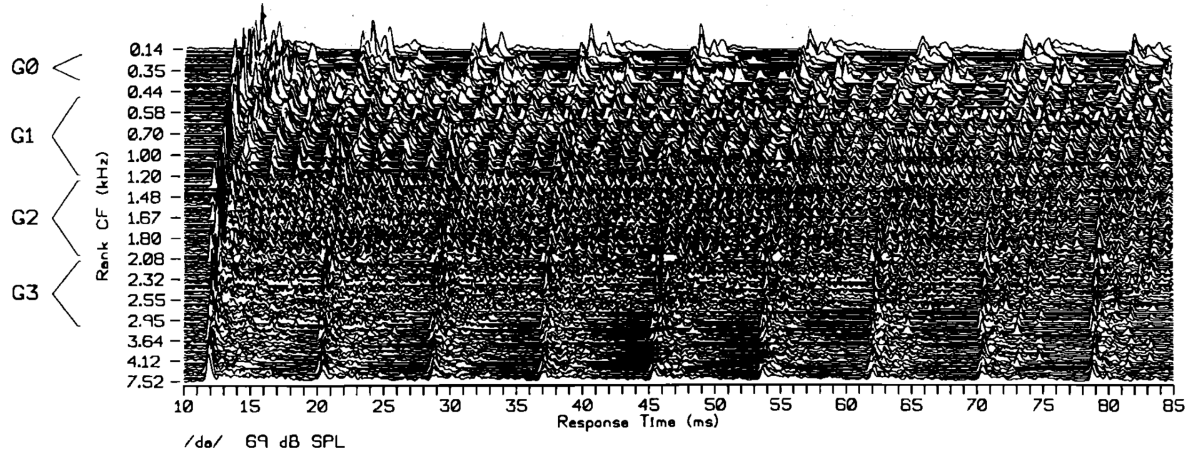


Figure 1.17: A neurogram of 178 auditory nerve (AN) fibers in response to synthetic syllable /da/ presented at 69 dB. Abscissa is response time (ms), the ordinate is the instantaneous firing rate (IFR) and the receding axis is fiber characteristic frequency. Scale bar represents 1000 spikes/s. IFRs were computed with a 0.1 ms bin width and smoothed using a 4-point Hamming window function. Groups of AN fibers that synchronise to the same syllable component are marked on the left. Adapted from Secker-Walker & Searle (1990).

as shown in Figure 1.17. The IPIs of the AN fibers described by Secker-Walker & Searle (1990) varied systematically both across CFs and across time as stimulus resonances changed. The temporal information present in the IPIs of AN fibers was enough to reproduce the formants of the synthesised syllable /da/. When formant frequencies were estimated purely from the IPIs recorded, they matched almost perfectly to the actual formants of the stimulus used in the study.

A study by Delgutte & Kiang (1984) used nine synthesised two-formant vowels and found similar results to Secker-Walker & Searle (1990), however they reported five rather than four different groups of AN fibers classified by their response properties (Fig. 1.18). Apart from the groups synchronised to F_0 , F_1 and F_2 matching the groups reported by Secker-Walker & Searle (1990), Delgutte & Kiang (1984) also reported a region where neurons synchronised to their CF. This region was found between F_1 and F_2 for spread vowels /i/, /e/, and /ɪ/, and in low-CF

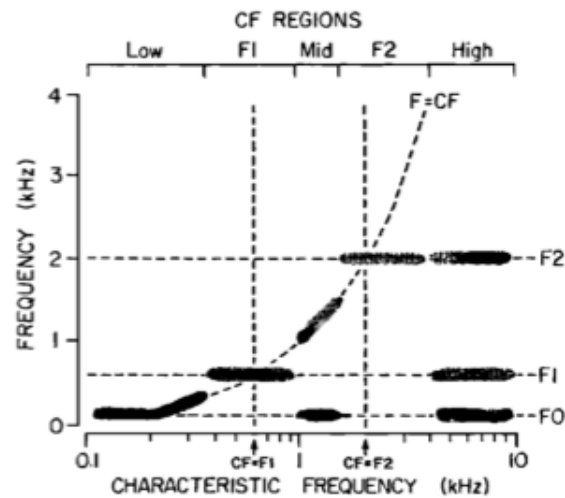


Figure 1.18: A schematic representation of 5 response regions of auditory nerve (AN) fibers in response to vowels. Abscissa is the AN fiber characteristic frequency and the ordinate is the frequency components of vowels. Dark areas represent the largest response components. The response regions in the schematic are exaggerated for display purposes. Adapted from Delgutte & Kiang (1984).

region for the open vowels /æ/, /ɔ/, /ɑ/ and /ɛ/. Furthermore, when examining AN fibers with high-CFs, Delgutte & Kiang (1984) reported synchronisation to F_1 in addition to the top formant and F_0 as reported by Secker-Walker & Searle (1990).

Overall AN fiber responses are generally dominated by formant components in response to periodic sounds, while responses to noise are normally dominated by components near fiber CFs (De Boer & De Jongh 1978, De Boer & Kuyper 1968, Delgutte & Kiang 1984, Moller 1977, Ruggero 1973). It was also found that neurons that synchronised to F_1 normally had CFs distributed symmetrically around F_1 , while neurons synchronised to F_2 normally had CFs equal to or greater than F_2 (Sinex & Geisler 1983).

1.5.6 Cochlear Nucleus

All AN fibers synapse at the cochlear nucleus (CN), the first central nucleus of the mammalian auditory system located in the medulla. CN is the place where physiological diversity arises from the information transmitted from the AN through convergence, inhibition and cell membrane specialisations (Clarey et al. 2004, Ehret & Romand 1996, Frisina 2001, Irvine 1986, Rhode & Greenberg 1992). Information flow is essentially divided into parallel streams, each of which performs specialised processing of the AN information, before converging at the inferior colliculus and then the auditory cortex (Lai et al. 1994).

CN is tonotopically organised and can be divided into ventral (VCN) and dorsal (DCN) parts. DCN is suggested to be primarily involved in sound localisation, therefore, it will not be discussed further. VCN extracts and enhances frequency and timing information received from the afferent connections from the AN. It then distributes the information through two pathways: sound localisation (primarily through the anterior VCN (AVCN) (Cant & Hyson 1992)), and sound identification (through both the AVCN and the posterior VCN (PVCN) (Eggermont 2001)).

VCN neurons have been categorised based on their morphology and temporal discharge characteristics in response to short supra-threshold tones into primary-like (including primary-like with notch) (47%), chopper (36%), onset (10%) and other (7%) (Evans 1992, Rhode et al. 2010) (Fig. 1.19). It has been suggested that in response to a vowel, chopper cells maintain rate-place representation of a vowel spectrum, whereby stimulus information is carried within the preferred frequency tuning and the rate of firing of chopper cells (Recio & Rhode 2000). It has also been suggested that primary-like cells maintain temporal representation of the vowel spectrum, whereby stimulus information is encoded within the precise temporal configuration of primary-like neurons' spikes (Recio & Rhode 2000).

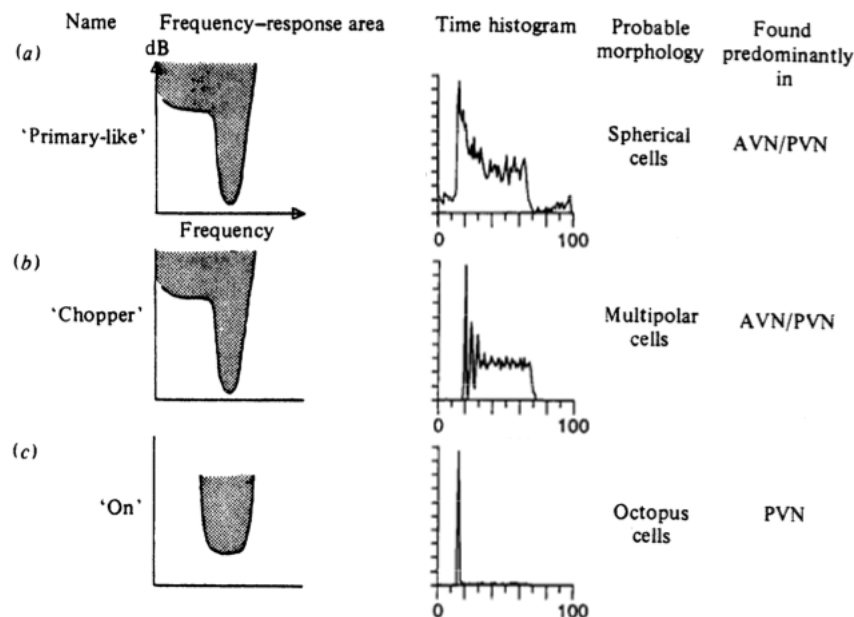


Figure 1.19: Schematic representation of the three main types of neurons in mammalian AVCN and PVCN based on their response properties. Left-hand column: neurons' typical frequency selectivity tuning curves. Middle column: post-stimulus time histograms normalised to maximum firing rate depicting neurons' typical response patterns to CF tones. Right-hand column: cell types and location. Adapted from Winter & Palmer (1990) and Evans (1992).

The three main sub-populations of the VCN (the primary-like, chopper and onset cells) are reviewed below. For the sake of brevity only the most important aspects of the diverse and complex response properties of these sub-populations are discussed.

1.5.6.1 Primary-Like Neurons

Primary-like (PL) response properties are normally recorded from small and large spherical bushy cells in the VCN (Joris, Carney, Smith & Yin 1994, Joris, Smith & Yin 1994, Palmer et al. 2003). These cells receive one or at most a few large calyx endings on their somata from AN fibers (Brawer et al. 1974, Cant & Morest 1984, Lorente de No 1981). Apart from the AN synapses, PL cells also receive

inhibitory connections. These may be intrinsic connections within the AVCN, projections from the DCN, contralateral AVCN or higher centres (Cant & Gaston 1982, Cant & Morest 1978, Comis 1970, Elverland 1977, Kane 1976, Lorente de No 1981, Wickesberg & Oertel 1988, Winter et al. 1989). PL cells synapse in the medial superior olive before reaching the inferior colliculus (IC) (Recio & Rhode 2000).

PL spherical bushy cells fire one or two small action potentials before remaining in a depolarised state for a few milliseconds (Oertel 1991). They re-polarise considerably faster after a depolarising current pulse compared to a hyper-polarising current pulse (Oertel 1991). These neurons have very short temporal integration windows, which means that they act as coincidence detectors, whereby sub-threshold inputs have to arrive simultaneously to produce a supra-threshold response in a spherical bushy cell (Smith et al. 1991) (Fig. 1.20)

PL neurons' response properties are characterised by a high probability of discharge straight after the stimulus onset, with a following exponential decline to a steady-state rate (Winter et al. 2003). PL neurons have response properties almost indistinguishable to AN fibers due to strong afferents from the AN synapsing straight onto the PL cell somata (Blackburn & Sachs 1989, 1990, Winter & Palmer 1990). Recio & Rhode (2000) found that in response to vowels PL neurons normally phase-lock to formant frequencies that are closest to their CF similar to AN fibers.

1.5.6.2 Primary-Like With Notch Neurons

Primary-like with notch (PLn) response properties within the CN are characterised as such due to a short (around 2ms long) drop in firing activity after the initial post-stimulus spikes in the cell PSTH plot (Fig. 1.21) (Winter et al. 1989). PLn response properties are normally recorded from globular bushy cells located in the caudal areas of the AVCN near the nerve root area (Joris, Carney, Smith & Yin

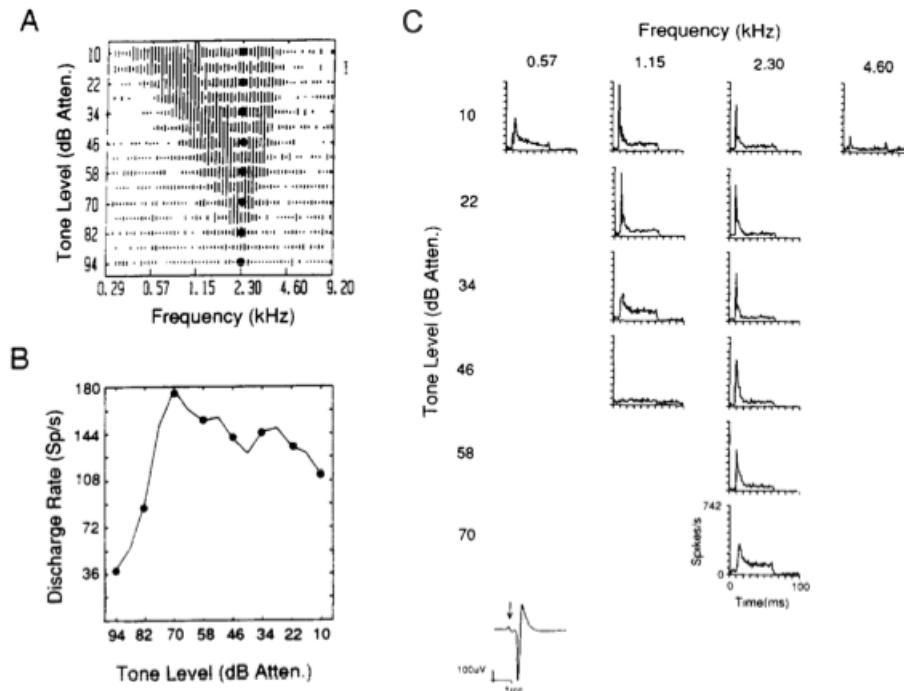


Figure 1.20: Response profile of a primary-like neuron in AVCN. A: response area; scale bar is 161 spikes/s. B: rate-level function at fiber CF. Filled circles correspond to the filled circles in the response profile. C: PSTH plots of variation in neuron's response properties depending on the frequency-amplitude combination of the stimulus. Bin width is 1 ms. Adapted from Winter & Palmer (1990).

1994, Joris, Smith & Yin 1994, Osen 1969, Palmer et al. 2003, Smith & Rhode 1987). These cells receive smaller end-bulb terminals from a few AN fibers (Osen 1969). Globular bushy cells project to the various nuclei within the trapezoid body, lateral lemniscus and superior olivary nucleus (Cant & Hyson 1992, Glendenning et al. 1981, Kuwabara & Zook 1991, Smith et al. 1993).

PLn neurons show enhanced synchronisation to the stimulus phase, whereby they fire regular spikes in response to the cycle peaks of a sinusoidal stimulus. This results in a stable pattern of coincident spike times across different presentations of the same pure tone. This is in contrast to the AN firing patterns, which are more stochastic, whereby AN fibers may skip firing in response to certain peaks

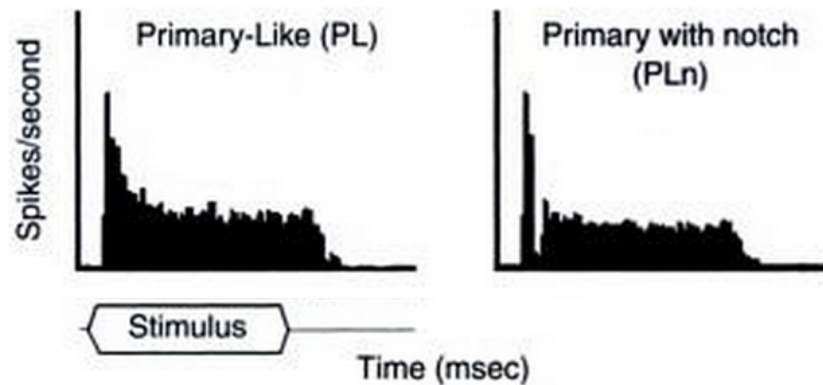


Figure 1.21: PSTH response profiles of primary-like and primary-like with notch neurons in the cochlear nucleus. Primary-like with notch neurons are characterised by the presence of a short (approximately 2ms long) period of reduced firing shortly after the onset of firing in response to an auditory stimulus. Adapted from Bronzino & Peterson (2006).

of a sinusoidal stimulus. This results in badly aligned spike trains in response to multiple presentations of the same sinusoidal stimulus. The differences in phase-locking between AN and PLn fibers are illustrated in Fig. 1.22 (Joris & Smith 2008). It can be seen that the spikes within the dot raster plot of a PLn cell (B) in response to 50 repetitions of a 670Hz sinusoidal tone are more synchronised than those within the equivalent plot of an AN fiber (A). The enhanced phase-locking of PLn fibers is further illustrated within the period-response histograms, polar phase-response plots (C, D) and inter-spike interval histograms (E, F) of the PLn and AN fibers in response to 100 presentations of the 670Hz pure tone stimulus.

1.5.6.3 Chopper Neurons

Chopper (CH) response properties are recorded from tonic T-stellate cells (or Type I multipolar cells) in PVCN (Rhode et al. 1983, Rouiller & Ryugo 1984, Smith & Rhode 1989). These cells receive afferents from 5 AN fibers on average (Ferragamo et al. 1998), as well as inhibitory backprojections from T- and D-

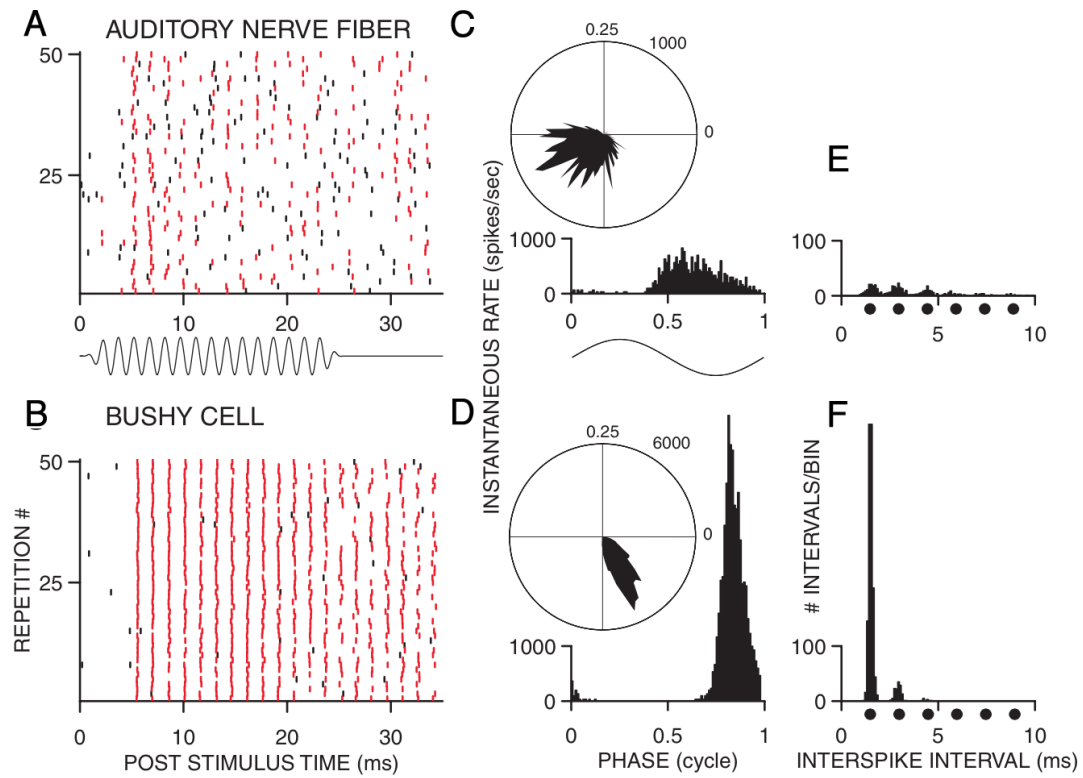


Figure 1.22: Enhanced phase-locking in primary-like with notch (bushy) neurons in the cochlear nucleus compared to auditory nerve fibers. A-B: Dot rasters of cell responses to 50 presentations of a 670Hz pure tone. Each tone was presented for 25 ms. Red indicates that at least two spikes occurred within 50 ms of each other. C-D: period response histograms and polar plots illustrating the distribution of spike times with respect to the phase of the 670Hz pure tone sinusoid. Data were collected in response to 100 stimulus presentations. The outer circle in the polar plots represents the instantaneous firing rate of the neuron (1000 and 6000 spikes/s for auditory nerve and bushy cells respectively). E-F: corresponding interspike-interval histograms. Adapted from Joris & Smith (2008).

stellate neurons from the olivary complex (OC) onto their dendrites and somata (Cant 1981, Smith & Rhode 1989). It has been suggested that CH T-stellate cells integrate information from high- and low-SR AN fibers with similar CFs (Delgutte 1982, Koch et al. 1982, Winslow et al. 1987). T-stellate cells have been found to project to the trapezoid body, and then to the contralateral superior olivary complex and to the ventral nucleus of lateral lemniscus (Oertel et al. 1990, Recio & Rhode 2000, Smith & Rhode 1989, Smith et al. 1993), and then the inferior colliculus, to which they form a major excitatory input (Adams 1979, Alibardi 1998, Beyerl 1978, Oertel et al. 2011, Oliver et al. 1997).

CH T-stellate cells have a linear voltage-current relationship and fire tonically when depolarised by a continuous current, or when presented with a tone at their CF (Smith & Rhode 1989, Wang & Sachs 1994). CH cells have a longer integration time window resulting in a first spike latency of around 1 ms after stimulus onset, which is longer than the first spike latency of PL or onset cells. Their first action potential is timed to reflect the onset of the tone burst, however, once a T-stellate cell starts firing, the frequency of firing is determined by the properties of the cell and by the stimulus intensity (Banks & Sachs 1991, Oertel et al. 2011). This enables these cells to function as rate processors of auditory information, whereby playing a role in intensity coding (Winter & Palmer 1990), and to follow amplitude modulation of sounds well (Frisina et al. 1990*a,b*).

CH cells have different response properties to the AN fibers, whereby they have low SRs (Recio & Rhode 2000), and they have regular spike trains, which are less prone to phase-locking (Winter & Palmer 1990), in particular to the higher frequency (> 1 kHz) and lower intensity stimuli (Oertel et al. 2000). Unlike the AN fibers, CH cells do not display adaptation (Oertel et al. 2011) (Fig. 1.23).

Clarey et al. (2004) reported the response properties of CH neurons to stop-consonant-vowel-consonant syllables. CH cells showed a clear response peak to

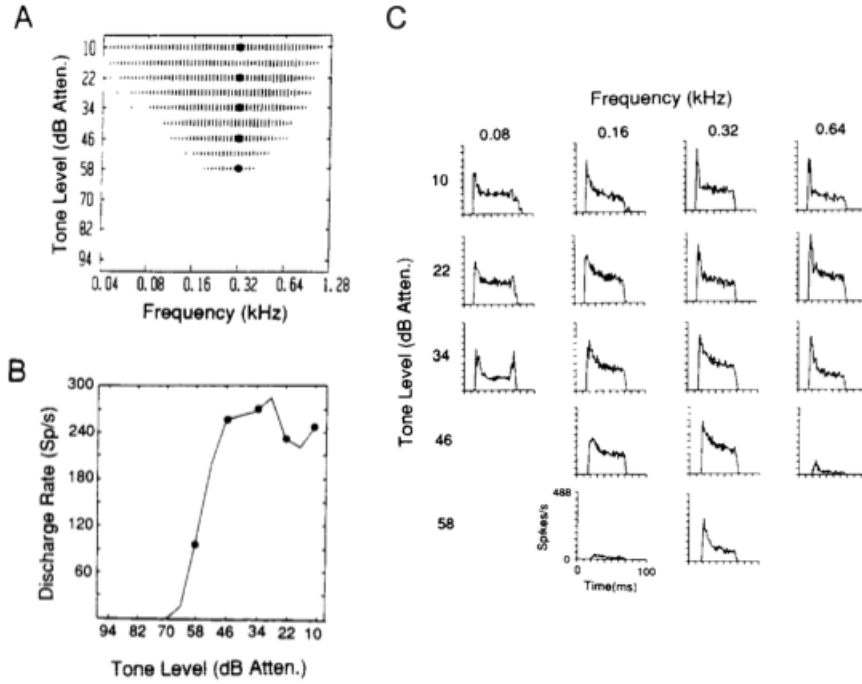


Figure 1.23: Response profile of a chopper neuron. A: response area; scale bar is 166 spikes/s. B: rate-level function at fiber CF. Filled circles correspond to the filled circles in the response profile. C: PSTH plots of variation in neuron's response properties depending on the frequency-amplitude combination of the stimulus. Bin width is 1 ms. Adapted from Winter & Palmer (1990).

stop-consonant release, however, the magnitude of the peak relative to the subsequent activity was smaller for CH cells than for PL cells. The stop-consonant release peak was followed by a narrow synchronised peak in activity at vowel onset. It was found that CH neurons with low CF (< 1 kHz) phase-locked to the vowel F_1 , whereas cells with higher CFs phase-locked to F_0 , however, the phase-locking was diminished in the presence of noise (Recio & Rhode 2000, Winter et al. 2003).

1.5.6.4 Onset Neurons

Onset (ON) responses are recorded from the octopus cells in PVCN (Joris, Carney, Smith & Yin 1994, Joris, Smith & Yin 1994, Palmer et al. 2003, Rhode et al. 1983,

Smith & Rhode 1989). Liberman (1993) has estimated that each ON cell receives afferent connections from up to 65 AN fibers with a large range of CFs, which results in very broadly tuned ON response properties (Godfrey et al. 1975, Rhode et al. 2010). Furthermore, octopus ON cells receive inhibitory connections from the D-stellate cells in the contralateral PVCN (Ferragamo et al. 1998, Needham & Paolini 2003, Wenthold 1987). Octopus ON cells are characterised by low voltage-activated potassium channels and hyper-polarization activated cation channels, which result in very low input resistance and fast membrane time constants. This makes these neurons very leaky with high spike thresholds. This means that octopus ON cells require strong synchronisation from many AN fibers in order to produce a discharge (Cant & Hyson 1992, Oertel et al. 2000, Osen 1969). Octopus ON cells project to the contralateral superior paraolivary nucleus and to the ventral nucleus of the lateral lemniscus, before converging in the inferior colliculus (Adams 1997, Schofield & Cant 1997, Smith et al. 1993, 2005).

ON responses are characterised by a prominent initial discharge peak indicating the start of a tone, with new peaks for each tone pip (Winter & Palmer 1990, Winter et al. 2003) (Fig. 1.24). Furthermore, ON cells can discharge at a precise phase to every cycle of a low-frequency tone up to 500-800 Hz (Godfrey et al. 1975). While ON cells lock to the waveform of periodic broadband signals precisely, high-frequency high intensity tones only generate one discharge independent of the tone duration (Oertel et al. 2000). ON cells also show strong adaptation, whereby their firing can be totally suppressed by an adaptation tone presented within 200 ms of the test tone (Bleeck et al. 2006).

ON cells have been found to have stronger phase-locking to vowel F_0 than chopper or AN fibers, which is not disrupted by the addition of noise (Winter et al. 2003). Furthermore, the phase-locking to F_0 in the ON cells is more extensive, whereby it is not restricted to ON regions with CFs between vowel formants, like

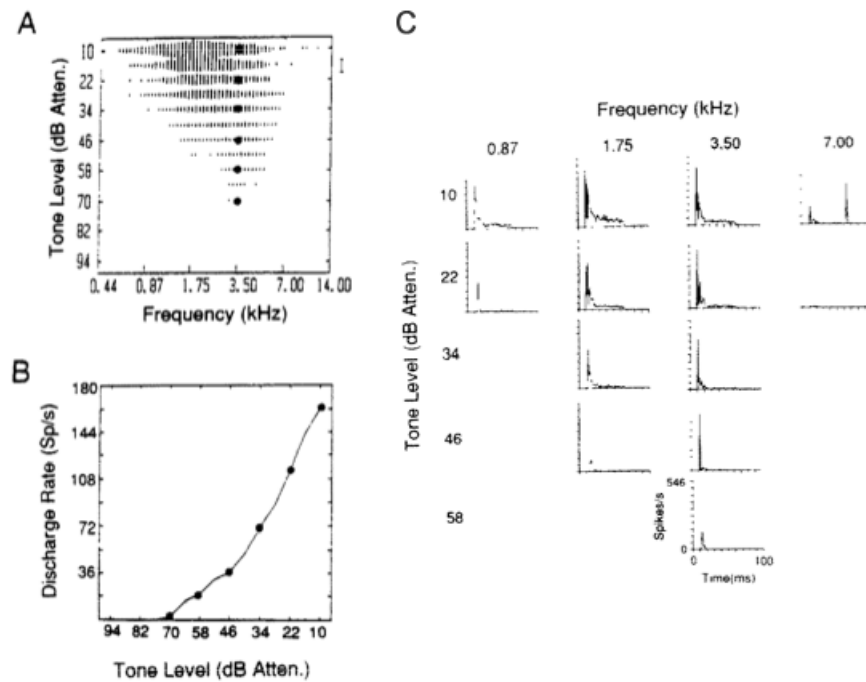


Figure 1.24: Response profile of an onset neuron. A: response area; scale bar is 166 spikes/s. B: rate-level function at fiber CF. Filled circles correspond to the filled circles in the response profile. C: PSTH plots of variation in neuron's response properties depending on the frequency-amplitude combination of the stimulus. Bin width is 1 ms. Adapted from Winter & Palmer (1990).

it is in the AN (Winter et al. 2003). This strong phase-locking property of the ON cells is suggested to be the result of across frequency coincidence detection (Kim et al. 1986, Palmer & Winter 1996, Rhode & Smith 1986), since ON neurons can detect coincident firing of many AN fibers due to its afferent connectivity (Bal & Oertel 2001, Ferragamo & Oertel 2002, Golding et al. 1999). In response to stop-consonant-vowel-consonant syllables, ON cells show a precise discharge peak at the stop consonant release, followed by another peak at the beginning of the vowel voice onset (Clarey et al. 2004).

1.5.7 Superior Olivary Complex

Superior Olivary Complex is a conglomeration of a number of nuclei including the trapezoid body and the lateral lemniscus discussed below.

1.5.7.1 Nuclei of the Trapezoid Body

Trapezoid body (TB) is a tonotopically organised (Warr & Beck 1996) relay fiber tract receiving afferent connections from the CN, and sending projections to the olivary nuclei, the lateral lemniscus nuclei, or the inferior colliculus (Stotler 1953, Van Noort 1969, Warr 1966, 1972). TB is normally divided into three nuclei : ventral (VNTB), lateral (LNTB) and medial (MNTB). Evidence suggests that VNTB is implicated in sound identification, while LNTB and MNTB are mostly involved in sound localisation (Irvine 1986). It has been found that in humans LNTB is the largest part, containing around 7,500 homogeneous stellate neurons; MNTB is the second largest part, containing around 3,600 neurons, which are mainly round (principal) but also some stellate neurons (about 11%); and VNTB is the smallest part of TB with approximately 1400 neurons, which are mostly round or stellate (Kulesza 2008). All TB neurons display primary-like responses with excellent phase-locking even for tones in excess of 1 kHz, and sigmoid rate-intensity functions, reaching maximum values around 30 dB above threshold (Brownell 1975). The similarity between the response properties of the neurons within the TB nuclei to those of the PL neurons in the CN and the AN fibers, coupled with the neurophysiological data suggesting that the largest nucleus of the TB (LNTB) is involved in sound localisation rather than auditory object identification, makes the role of TB nuclei in speech recognition unclear (Brownell 1975).

1.5.7.2 Nuclei of the Lateral Lemniscus

Nuclei of the Lateral Lemniscus (NLL) are normally subdivided into three parts: ventral (VNLL), intermediate (INLL), and dorsal (DNLL) which receive afferent inputs from the contralateral VCN (Browner & Webster 1975, Glendenning et al. 1981, Huffman & Covey 1995, Schofield & Cant 1997, Van Noort 1969, Vater & Feng 1990, Warr 1982, 1966, 1972). Neurophysiological evidence suggests that the DNLL, INLL and the medial part of the VNLL are primarily implicated in sound localisation and therefore they will not be considered further (Adams 1979, Batra & Fitzpatrick 2002, Xie et al. 2005). The caudal part of VNLL, however, receives afferent connections from the octopus ON cells in VCN, and this pathway has been found to be especially prominent in humans, thus suggesting that it may be important in speech processing (Adams 1997) (Fig. 1.25). Projections from the NLL to ipsilateral inferior colliculus (IC) form the largest input to the IC from any other brainstem area, with more than 50% of projections to combination-sensitive neurons in the IC coming from the LL (Wenstrup 1999). However, 40% of these afferents come from INLL and medial VNLL, which are specialised in sound localisation (Kudo 1981). The role of the NLL in speech recognition is, therefore, unclear.

1.5.8 Inferior Colliculus

Inferior colliculus (IC) is tonotopically organised (Ehret & Fischer 1991, Hage & Ehret 2003) and is located in the mid-brain. It is the largest subcortical structure of the auditory system where multiple auditory processing streams converge (Portfors & Felix 2005). IC receives around nineteen separate ascending and descending projections from different brain areas (Ehret 1997, Portfors & Felix 2005), however its main ascending afferent projections are from the CN (Frisina et al. 1998).

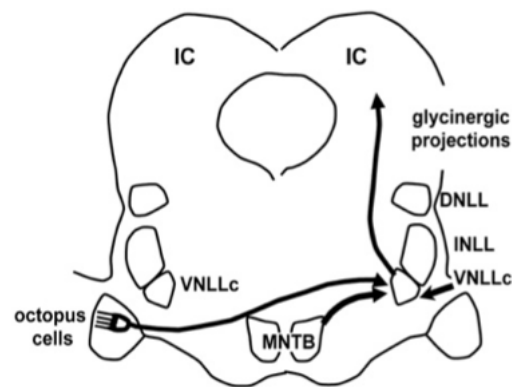


Figure 1.25: Schematic representation of afferent connections to the VNLL from the cochlear nucleus and the trapezoid body. Adapted from Pollak (2011).

IC projects ipsilaterally and contralaterally to the ventral division of the medial geniculate body (vMGB) of the thalamus (Ehret & Romand 1996).

It is suggested that the IC is similar to V1 in visual processing, whereby it plays the role of a 'whiteboard' on which acoustical information is written to be used by the higher brain areas (Nelken 2008). Furthermore, IC are considered to be the endpoint of time specialised part of the auditory pathway (Trussell 1997), as supported by electrophysiological studies in humans (Galbraith et al. 2004, Wible et al. 2005).

IC performs frequency analysis, frequency-modulated (FM) sound processing, sound filtering and acoustic cue integration by inheriting the response properties of the earlier auditory brain areas (Ramachandran et al. 1999), and refining them further by integrating excitatory and inhibitory inputs with different temporal properties (Covey & Casseday 1999, Davis 2005, Keller & Takahashi 2000). Within-layer inhibition in the IC has been found to be important, since response properties of the IC neurons change substantially when inhibition is blocked (Egorova et al. 2001, Gonzalez-Hernandez et al. 1996, Vater et al. 1992) (Fig. 1.26). This effect di-

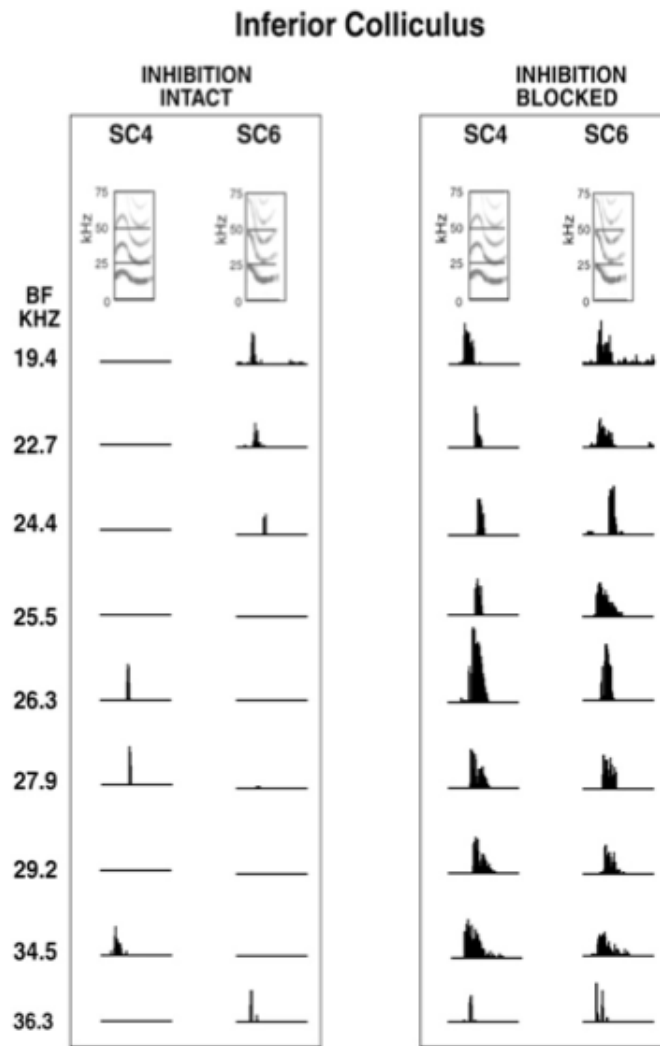


Figure 1.26: Responses of IC neurons to two vocalisations, SC4 and SC6, before and while blocking inhibition in the area. The calls have similar spectrotemporal features, however, different populations of IC neurons responded to each one. This selectivity disappeared once inhibition in the IC was blocked. Adapted from (Pollak 2011).

minishes, however, towards the edges of the IC (Egorova et al. 2001).

Neurons with a very diverse set of response properties have been reported in the IC (Egorova et al. 2001). These include Class 1 neurons (42%) with PL-like response properties, and Class 3 neurons (25%) with ON-like response properties,

which suggests direct connections from the CN (Egorova et al. 2001, Hage & Ehret 2003). The complexity of the afferent projections to the IC mean that it is still not fully understood how the complex response properties of the IC neurons emerge (Ehret et al. 2004).

1.5.9 Medial Geniculate Nucleus

Medial Geniculate Nucleus (MGN) of the thalamus is the last subcortical nucleus of the auditory pathway. MGN receives its main afferent projections from the IC. It also receives back-projections from the auditory cortex, which play a modulatory role (Andersen et al. 1972). It has been suggested that each IC neuron contacts a number of MGN cells, and each MGN cell receives afferent projections from a number of IC cells (Heierli et al. 1987). MGN neurons also form excitatory or inhibitory connections with each other, whereby firing of one neuron either increased or decreased the firing probability of the other neuron. MGN cells that either share a common afferent projection from the IC or a lateral excitatory/inhibitory connection are not necessarily spatially close to each other.

MGN is normally divided into three major parts the medial, ventral and lateral subdivisions (Webster & Aitkin 1975). Only the ventral part of MGN, however, is thought to be involved in auditory object processing. Ventral MGN sends afferent connections to the laminae III and IV of the auditory cortex (Burton & Jones 1976, Jones & Burton 1976, Kimura et al. 2005, Rouiller & de Ribaupierre 1985).

MGN response latencies varying between 3.5-27 ms from stimulus onset (Aitkin & Webster 1971). Most (69%) of the MGN cells show onset response properties (Aitkin & Webster 1971) and are broadly tuned, with 50% of the cells responding to almost the entire frequency range (Kiang 1965). The other MGN cells have complex response patterns, whereby their response varies dramatically depending on the frequency-intensity combination of the tone presented, with excitatory

regions becoming inhibitory and visa versa, as well as new response components appearing or disappearing at certain intensities (Allon et al. 1981) (Fig. 1.27). The majority of MGN neurons do not phase-lock (Rouiller et al. 1981), but show some adaptation (Barlett & Smith 2002) and FM sweep direction sensitivity (Allon et al. 1981). No selectivity for vocalisations have been found in MGN neurons (Allon & Yeshurun 1985).

1.5.10 Auditory Cortex

The auditory cortex receives excitatory afferent inputs into layers IIIb and IV from the IC through the ventral MGN in the thalamus, whereby each MGN cell contacts a number of cortical neurons (Eggermont 1996, Sugimoto et al. 1997). While the effective processing time constants in the IC are in the order of 10 ms, they are in the order of 100 ms in the auditory cortex (Nelken et al. 2003). Inhibitory signals in the auditory cortex arise from the within-cortex inhibitory recurrent collaterals (Creutzfeldt et al. 1980). Furthermore, it has been found that the pyramidal cells in the auditory cortex are interconnected with each other. Some researchers argue that these excitatory recurrent connections in the auditory cortex are very few and weak to make a difference (Abeles 1982*a*, Atzori et al. 2001), while others argue that this recurrency turns the auditory cortex into a stable attractor which can be switched from one state to another by sudden correlated changes in the firing rate of its units (Abeles et al. 1995).

The auditory cortex is normally subdivided into three major areas: primary auditory cortex (A1, although sometimes AAF (in rodents and carnivores) or R (in primates) is also included as a second primary field), belt and parabelt areas. In humans A1 is situated in the posterior two-thirds of the supra-temporal lobe in both hemispheres and includes Heschl's gyrus (Brodmann area 41) (Brugge & Reale 1985, Ehret & Romand 1996). The auditory belt and parabelt areas extend into

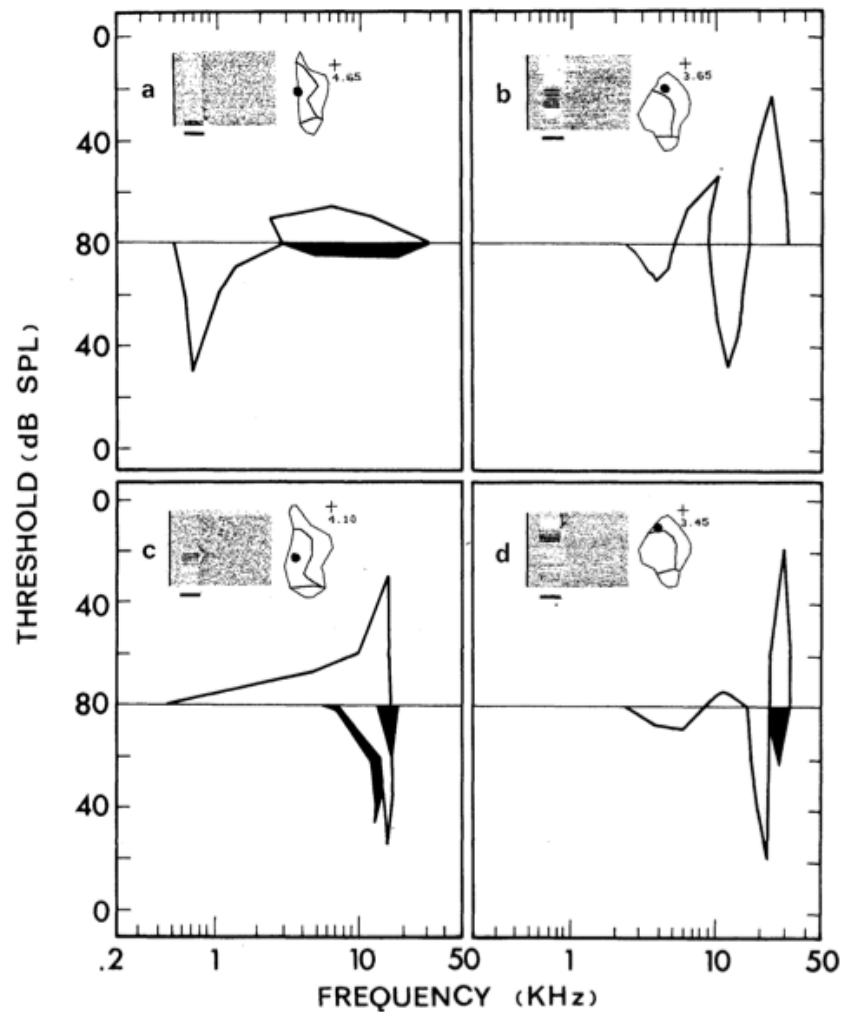


Figure 1.27: Schematic representation of complex tuning curves of four cells in the MGN. The areas above the midline represent the cells' inhibitory tuning curves, while the areas below the midline represent the excitatory tuning curves. Non-shaded tuning curves represent sustained excitation, whereby a cell fires continuously throughout the preferred stimulus presentation. Shaded areas represent excitatory off-components, whereby a cell does not respond throughout the stimulus duration, but instead starts firing once the stimulus presentation is over. The inset plots are a schematic representation of the cell's location within the MGN (*right*) and a dot display of the cell's response properties to a pure tone presented over a range of frequencies at 80 dB (*left*). Adapted from Allon et al. (1981).

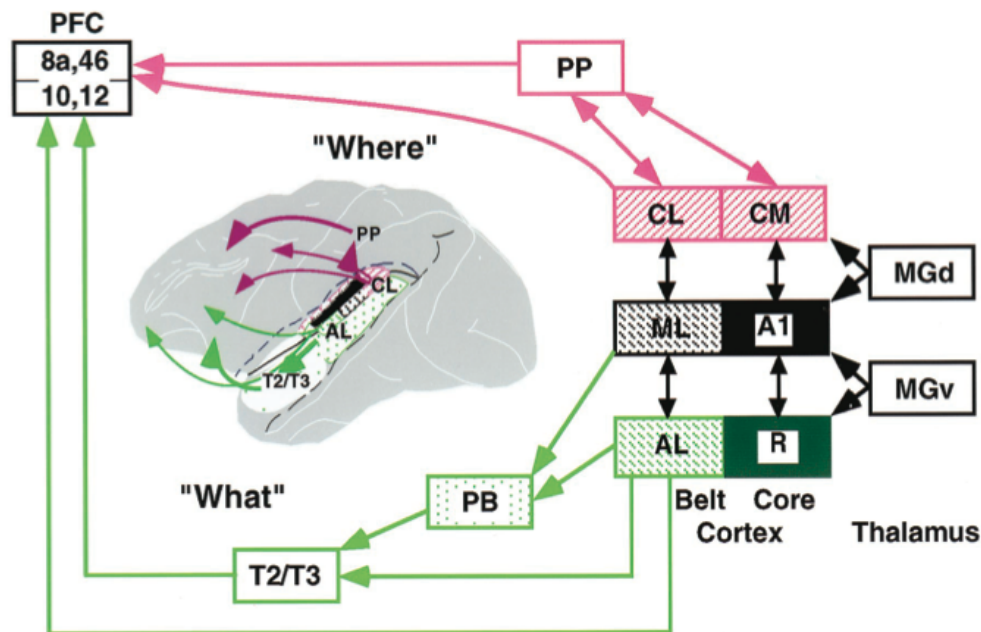


Figure 1.28: A schematic flow diagram of the 'where' (*red*) and 'what' (*green*) pathways in the auditory cortex. Adapted from Rauschecker & Tian (2000).

the posterior superior temporal gyrus (pSTG) (Brodmann area 22) (Galaburda & Pandya 1983, Hackett et al. 1998). Following the belt and parabelt areas of the auditory cortex, some neuroscientists believe that the auditory information is split into two different pathways (Romanski, Tian, Fritz, Mishkin, Goldman-Rakic & Rauschecker 1999). One pathway targets the caudal dorsolateral pre-frontal cortex, constituting the 'where' pathway of auditory processing; while the other pathway targets ventral pre-frontal cortex (vPFC) and orbitofrontal cortex, constituting the 'what' pathway of auditory processing (Kaas & Hackett 2000, Rauschecker 1998, Rauschecker & Tian 2000, Romanski & Goldman-Raciz 2002, Romanski, Tian, Fritz, Mishkin, Goldman-Rakic & Rauschecker 1999, Russ et al. 2007) (Fig. 1.28).

Human fMRI data suggest that the auditory cortex is hierarchically organised with areas close to A1 processing language-independent physical cues, areas further

up the hierarchy encoding acoustic features of speech, and eventually other areas encoding more abstract, non-acoustic representations of speech sounds (Davis & Johnsrude 2003, Uppenkamp et al. 2006). It has been found that categorical perception of phonemes is associated with activity in the belt and parabelt areas in the superior temporal region (Binder et al. 2004, Boatman et al. 1995, Chang et al. 2010, Husain et al. 2006, Lemus et al. 2009, Liebenthal et al. 2010, Selezneva et al. 2006, Steinschneider et al. 2003, 1994, 1995, 2005, 1999, Zatorre et al. 1992).

The majority (60%) of cortical neurons display onset response properties. The other prevailing response type is offset (30%), whereby a cell fires once after the auditory stimulus is finished. Only a small portion of cortical neurons (10%) show sustained firing in response to an auditory stimulus (Ehret & Romand 1996).

Neurons in the auditory cortex have been reported to fire to either a burst of 2-3 spikes received from a single neuron (Lisman 1997), or to a combined synchronised input from a population of neurons (Eggermont 2001). A number of studies thus argue that the functional properties of pyramidal cells in the auditory cortex resemble coincidence detectors more closely than rate integrators, having high response thresholds and short temporal integration windows (Abeles 1982*a,b*, Sadagopan & Wang 2009, Softky & Koch 1992, Windhorst 1988). It is suggested that cortical neurons are best driven by transient stimulus events (Phillips 1988, 1990).

Around half of the neurons in the auditory cortex respond to pure tones (Hromádka et al. 2008). Sadagopan & Wang (2009) studied neurons that do not respond to pure tones (the non-linear neurons) and found that they had more complex response properties than neurons that do (linear neurons). The researchers found that the non-linear neurons responded to two pips presented in a specific temporal order, to bandpassed noise with a particular bandwidth, and to natural, but not reversed, vocalisations. Non-linear neurons responded to fewer vocalisations

than linear neurons, which fired whenever a feature of a vocalisation entered their excitatory tuning curve. Furthermore, fewer non-linear neurons in a population responded to a particular vocalisation compared to linear neurons. All of the non-linear neurons described by Sadagopan & Wang (2009) had non-monotonic rate-level functions, whereby an increase in sound level resulted in the cessation of neuronal responses.

Hromádka et al. (2008) found that complex sounds are represented with sparse population encoding in the auditory cortex, whereby only around 5% of auditory cortex neurons respond to any particular stimulus. Each neuron took part in representations of numerous stimuli, and the same neuron could respond in a transient or a sustained manner to different stimuli. Furthermore, the authors reported that there was no topographic organisation of neurons, whereby neurons that fired strongly to a stimulus were not necessarily clustered together. A neuron that responded strongly to a stimulus could be surrounded by completely non-responding neurons.

1.5.10.1 Vocalisation Response Properties: A1

A study by Lu & Wang (2000) suggests that neurons in A1 can be divided into two sub-populations based on their response properties to vocalisations - the larger non-selective population responds to both vocalisation and non-vocalisation sounds, while the smaller selective population consists of neurons primarily responsive to natural vocalisation calls and to the identity of the callers (Fig. 1.29). It is suggested that the non-selective population is a general sound analyser, which provides a description of the auditory stimulus and is necessary for learning and analysing new sounds, whereas the selective population is necessary for identifying a known sound (Wang 2000). The non-selective population becomes smaller throughout the hierarchy of the 'what' pathway in the auditory cortex.

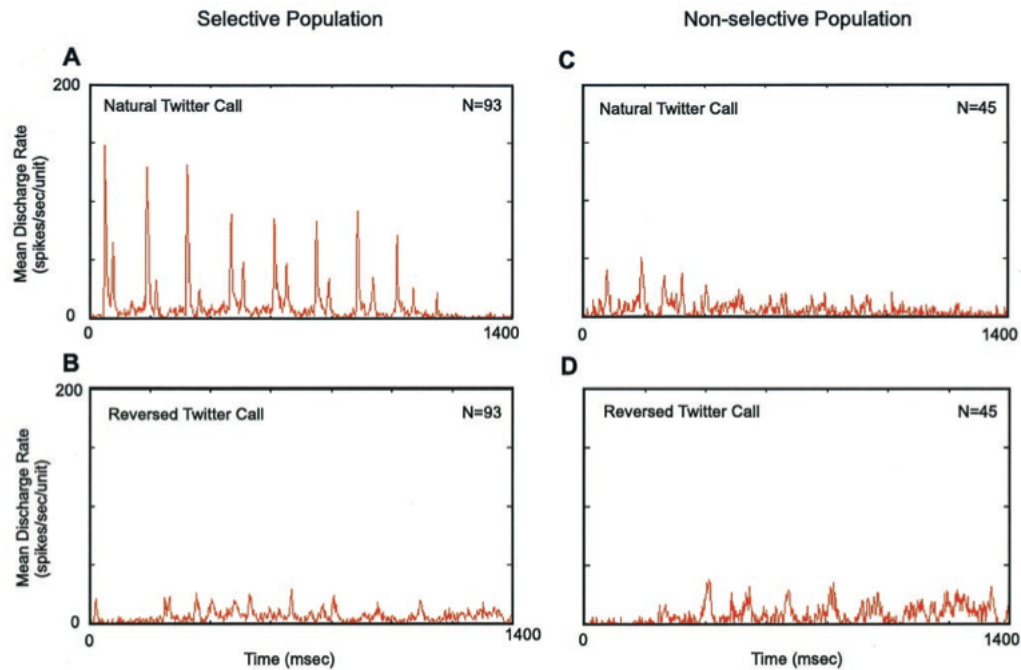


Figure 1.29: Averaged PSTHs of responses of A1 neurons to marmoset calls and time-reversed versions of the calls (bin width is 2 ms). Two populations of cells were found, selective and non-selective. A - B: selective population showing preferential response to a natural call compared to a time-reversed version of it. C - D: non-selective population showing no preferential response to a natural call compared to a time-reversed version of it. Adapted from Wang (2000).

It is suggested that behaviourally relevant sounds are represented differently in the cortex compared to behaviourally irrelevant sounds (Buonomano & Merzenich 1998, Schmidt et al. 1999). This was demonstrated using an experimental paradigm which compares response properties of neurons to natural and time-reversed versions of vocalisations, since time-reversed vocalisations have the same spectral content and acoustic complexity as the natural vocalisations, but without the behavioural relevancy (Wang et al. 1995). The results of such experiments have been inconclusive, however. Wang et al. (1995) found that A1 neurons had greater mean firing rates and displayed a clearer representation of the spectral shape of natural

vocalisations compared to reversed vocalisations. Furthermore, A1 cells that displayed a clearer representation of the spectral shape of the vocalisations contained less information about the behavioural meaning of the sounds than those that did not. Schnupp et al. (2006) on the other hand found no difference between A1 responses to natural and reversed vocalisations (these vocalisations were not species specific), however some cells contained more information about the vocalisation identity than others.

In order to test whether any differences in A1 response properties to behaviourally relevant compared to behaviourally irrelevant sounds were learnt through experience, Wang & Kadia (2001) presented natural and time-reversed marmoset vocalisations to cats, who would not have had any prior experience with the stimuli. No differences between the two classes of vocalisations were found in cat A1. Furthermore, when Schnupp et al. (2006) trained ferrets to differentiate between marmoset vocalisations, it was found that the amount of information in the temporal response patterns of non-selective neurons increased significantly compared to the naive animals.

1.5.10.2 Vocalisation Response Properties: pSTG and STS

Lateral belt neurons in macaque auditory cortex show greater vocalisation call selectivity than A1 neurons (Rauschecker 1998, Rauschecker & Tian 2000, Tian et al. 2001). It has been found that neurons in the lateral belt areas respond to 3-5 monkey calls out of 7, with a few neurons responding to only 1 of the calls (Rauschecker & Tian 2000). Vocalisation selectivity in the lateral belt can not be explained by their frequency tuning (Hauser 1996). Moreover, neurons in this area have minimal responses to each syllable of a vocalisation when presented alone, while responding maximally when all syllables are presented in the right temporal order to create a full vocalisation (Rauschecker & Tian 2000). Lateral belt neurons

have been found to respond similarly to calls that have similar acoustic features rather than similar behavioural and functional meaning (Rauschecker 1998, Tian et al. 2001), which suggests that semantic representations of vocalisations arise further in the auditory cortex.

Posterior superior temporal gyrus (pSTG) in humans (corresponding to the belt area in primates) may be an area of the auditory pathway where representations of complex auditory objects are formed. Studies reported that electrical stimulation of pSTG in humans during electro-cortical mapping resulted in auditory hallucinations including speech, melodies and human voices (Penfield & Jasper 1954). It was also reported that low-level auditory syllable discrimination maps to the pSTG (Ojemann & Mateer 1979), and that pSTG processes sounds that are both spectrally and temporally complex, but not only one or the other (Boatman & Miglioretti 2005). Furthermore, evidence from human imaging studies found stronger activation in the pSTG to phonemes compared to sounds of lesser complexity (Naatanen et al. 1997).

Imaging studies in humans showed acoustic invariance in response to speech stimuli within the superior temporal sulcus (STS) region of the auditory cortex, corresponding to the parabelt area, suggesting that this area is also implicated in invariant representations of auditory objects (Rauschecker & Scott 2009). Further analysis showed that there was increased invariance throughout the hierarchy of the auditory cortex, with the dorsal plane of the STG being sensitive to acoustic variance within both intelligible and unintelligible speech, and STS being less sensitive to the acoustic variance in intelligible speech while being more sensitive to the differences between intelligible and unintelligible speech (Okada et al. 2010).

1.5.10.3 Vocalisation Response Properties: vPFC

Ventral prefrontal cortex (vPFC), which receives direct connections from the parabelt area (Hackett et al. 1999, Romanski, Bates & Goldman-Rakic 1999, Romanski, Tian, Fritz, Mishkin, Goldman-Rakic & Rauschecker 1999), has been shown to respond to familiar vocalisations, however, it is unknown what exactly the neurons are responding to. It can be vocalisation category, caller identity, caller body size or their reproductive status (Bradbury & Vehrencamp 1998, Hauser 1996, Hauser & Marler 1993, Owings & Morton 1998). The human homologue of the vPFC is the inferior frontal gyrus, which plays a role in semantic processing (Demb et al. 1995, Poldrack et al. 1999).

While it was found that some vPFC neurons in awake behaving macaque monkeys showed vocalisation selectivity, whereby they responded to only 1/10 vocalisations, further analysis revealed that their responses were mostly based on the acoustic structure of the stimuli (Romanski et al. 2004), which is similar to the response properties of the belt and parabelt neurons (Tian et al. 2001). The authors thus concluded that vPFC may play a role in articulatory feedback control instead of vocalisation identification.

An alternative view of the vPFC was suggested by Gifford et al. (2005), who found that populations of vPFC neurons responded differentially to semantically different calls but not to spectrally different calls, or repeated presentations of the same call, or to artificial stimuli with different bandwidths. They, therefore, suggested that vPFC might have higher-order semantic representations of vocalisations. Similar conclusions were reached by other researchers, who found that the responses of vPFC neurons of rhesus monkeys could be reliably classified into two groups based on whether they were semantically associated with food or not (Cohen et al. 2006, Gifford et al. 2005). The response preferences of these neurons

could not be attributed purely to the acoustic properties of the sounds.

In humans vPFC overlaps with Broca's area, and imaging studies in humans showed that this area is activated during semantic processing, syntactic working memory, as well as maintenance and retrieval of abstract rules (Amunts et al. 2004, Bunge et al. 2003, Demb et al. 1995, Fiebach et al. 2005, Muller & Basho 2004, Price 1998, Reber et al. 2003, Strange et al. 2000).

1.6 Past Models of Speech Recognition

Section 1.5 has summarised some of the neurophysiological data that has been collected to describe the structure and response properties of several auditory brain areas believed to be potentially involved in auditory object recognition. This data by itself, however, provides little insight into *how* the auditory brain learns stable invariant representations of auditory objects, including speech. A number of researchers believe that the only way to understand how speech perception is learned and subsequently performed in the brain is through large-scale biologically plausible neural network model simulation work (Arbib et al. 1995, Corchs & Deco 2002, Horwitz & Tagamets 1999, Husain et al. 2004, Salem et al. 2007). Such models, however, are very rare, and no comprehensive biologically plausible large-scale neural network model of the auditory brain capable of speech recognition currently exists.

The general trends in auditory brain modelling and in understanding the mechanisms of speech recognition are the following: 1) Connectionist/phenomenological models trying to replicate psychophysical speech recognition results (McClelland & Elman 1986, Nelson & Carney 2004); 2) models trying to perform speech recognition in a more biologically plausible way (without backpropagation of error) (Maass et al. 2004); 3) models that try to replicate and understand the auditory brain, without necessarily achieving full speech recognition (Kral & Majernik 1996).

This section describes some example models for the trends mentioned above, concluding with the description of the relevant computational modelling research using spiking neural networks (Appeltant et al. 2011, Coath et al. 2014, Maass et al. 2004, Masquelier et al. 2008, Nager et al. 2002). This thesis attempts to understand how robust representations of auditory objects, such as phonemes or words, can be learnt within a spiking neural network architecture whose structure

is closely inspired by the relevant anatomy of the auditory brain. Little research has been done using spiking neural networks compared to the rate-coded networks, however, due to their additional computational cost. Nevertheless, the few spiking neural networks for speech processing that have been developed so far suggest that they are a powerful tool for the task (Appeltant et al. 2011, Coath et al. 2014, Liaw & Berger 1996, Maass et al. 2004, Masquelier et al. 2008, Nager et al. 2002).

1.6.1 Rate-Coded Models

1.6.1.1 Model by McClelland & Elman (1986)

An important model of speech recognition is the Trace model proposed by McClelland & Elman (1986). The connectionist network consists of simple units organised in hierarchical layers, whereby the first layer represents features of phonemes, the next layer represents full phonemes and the last layer represents words. The layers are interconnected with excitatory and inhibitory connections, whereby connections between layers are purely excitatory and connections within layers are purely inhibitory to create within-layer competition. The model uses a dynamic processing structure called 'the Trace', which acts as a perceptual processing mechanism and model's working memory.

The model can account for a number of psychophysical findings, such as the time course of word recognition, the role of lexical effects on phoneme and word perception, and categorical perception of phonemes. The model, therefore, provides insights into how speech recognition might be performed in the brain on the conceptual level. The network, however, is unable to recognise words longer than the length of its input layer, it cannot deal with time distorted speech, and it is unable to recognise two instances of the same word in a sentence. Furthermore, it lacks biological plausibility, because it is not based on the neurophysiology of the brain; the representations of phonemes in the input layer and the lexicon in the

output layer are symbolic and not realistic; the network relies on top-down information flow from the full word representations from within its vocabulary, which is hard-coded; the network is not capable of learning; and the network does not cope well with the temporal aspects of speech, since it copes with the time dimension of speech by concatenating a number of copies of the network together, whereby each copy of the network represents the activity of the network at a different moment in time.

1.6.1.2 Model by Kral & Majernik (1996)

Kral & Majernik (1996) suggested a simplified neural network architecture representing the auditory brain hierarchy. The network used neuron-like units, called processing elements (PEs), the activity of which was governed by a simple linear threshold mechanism. The input to the network was a simplified representation of auditory nerve spike rasters, whereby arbitrarily specified spike trains with particular regular inter-spike intervals were applied to the input layer. While the inputs to the model were spike rasters, the model itself cannot be considered a spiking model, since the response dynamics of its constituent PEs were those of rate-coded neurons.

The network consisted of three parallel pathways: 1) the lateral inhibition neural network (LINN) (Fig. 1.30) was supposed to preserve temporal information and was modelled using four feed-forward layers loosely corresponding to the cochlear nucleus, inferior colliculus, medial geniculate body and the cortex. The layers were interconnected with sparse excitatory and inhibitory connections, whereby each PE in the pre-synaptic layer sent one excitatory and a number of inhibitory connections to PEs in the post-synaptic layer. Such connectivity, however, is not biologically plausible due to Dale's principle (Dale 1934), which states that all afferent synapses belonging to one neuron have to be of the same type, whether that

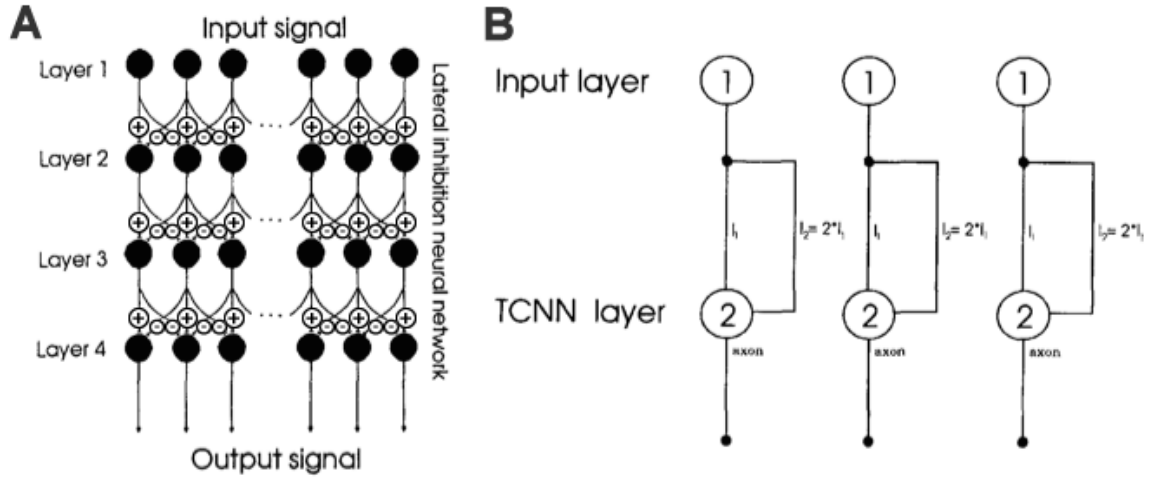


Figure 1.30: Schematic representation of the auditory brain model by Kral & Majernik (1996). A: Lateral inhibition neural network architecture B: Temporal coincidence neural network architecture. Thresholds of the neurons in the TCNN layer are twice as high as thresholds of the neurons in the input layer. Adapted from Kral & Majernik (1996).

is excitatory or inhibitory. 2) The second pathway was used to turn temporal code into place code and was modelled using a temporal coincidence neural network (TCNN) (Fig. 1.30), where each post-synaptic PE had a high firing threshold and a number dendrites of different lengths, where $l_1 = h * l_2$ (l_1 is the length of the dendrite with the longer delay and l_2 is the length of the dendrite with the shorter delay, h is a positive integer). Since each post-synaptic neuron needed at least two concurrent excitatory inputs in order to fire, these neurons were tuned to input spikes arriving at a particular interval. 3) The third pathway was used to detect frequency changes and was modelled by a TCNN, whereby each neuron had three dendrites of varying lengths set in a way that would make these neurons tuned to changes from one particular frequency to another. All delays and connectivity were hard-wired and no learning was used.

It was found that the network was able to simulate certain physiological properties of the auditory brain. These included frequency change detection and change

of information encoding from temporal to rate throughout the hierarchy of the network. While the authors claim that the layers of the network loosely correspond to the cochlear nucleus, inferior colliculus, medial geniculate body and the cortex, it is unlikely that the model would be able to deal with more complex tasks than the simple differentiation between various pure tone frequency sweeps, due to its very particular delay connectivity and lack of learning.

1.6.1.3 Model by Kurogi (1991)

A model by Kurogi (1991) was able to recognise isolated syllables and words, as well as words in continuous speech, from Fourier Transform representations of speech inputs (Fig. 1.31). The model was based loosely on the architecture of the auditory brain, whereby the different layers of the network were suggested to represent the cochlea and the cochlear nucleus (input Fourier Transform); superior olivary complex (Lateral Excitation and Lateral Inhibition Layers (LEL and LIL)); inferior colliculus, medial geniculate body and the cortex (Feature Detection Layers (FDLs) modelled by Kurogi (1987)). The FDLs were made of P-cells and B-cells, which are the equivalents of the pyramidal and basket cells in the cortex correspondingly. P-cells were set up to have a high threshold, which meant that they had to receive a particular pattern of inputs in order to fire. Once a particular P-cell fired, it was thought to have recognised a particular feature. It would then activate its corresponding B-cell, that would fire a burst of inhibitory spikes with connections to all of the P-cells in the layer, therefore resetting the activity in the P-cell layer (Fig. 1.31). This sort of architecture was found to be successful in recognising temporal patterns. The number of cells in the output layer matched the number of stimuli to be recognised. The weight vectors within the FDL layers were set up manually to achieve the best recognition performance on the training set.

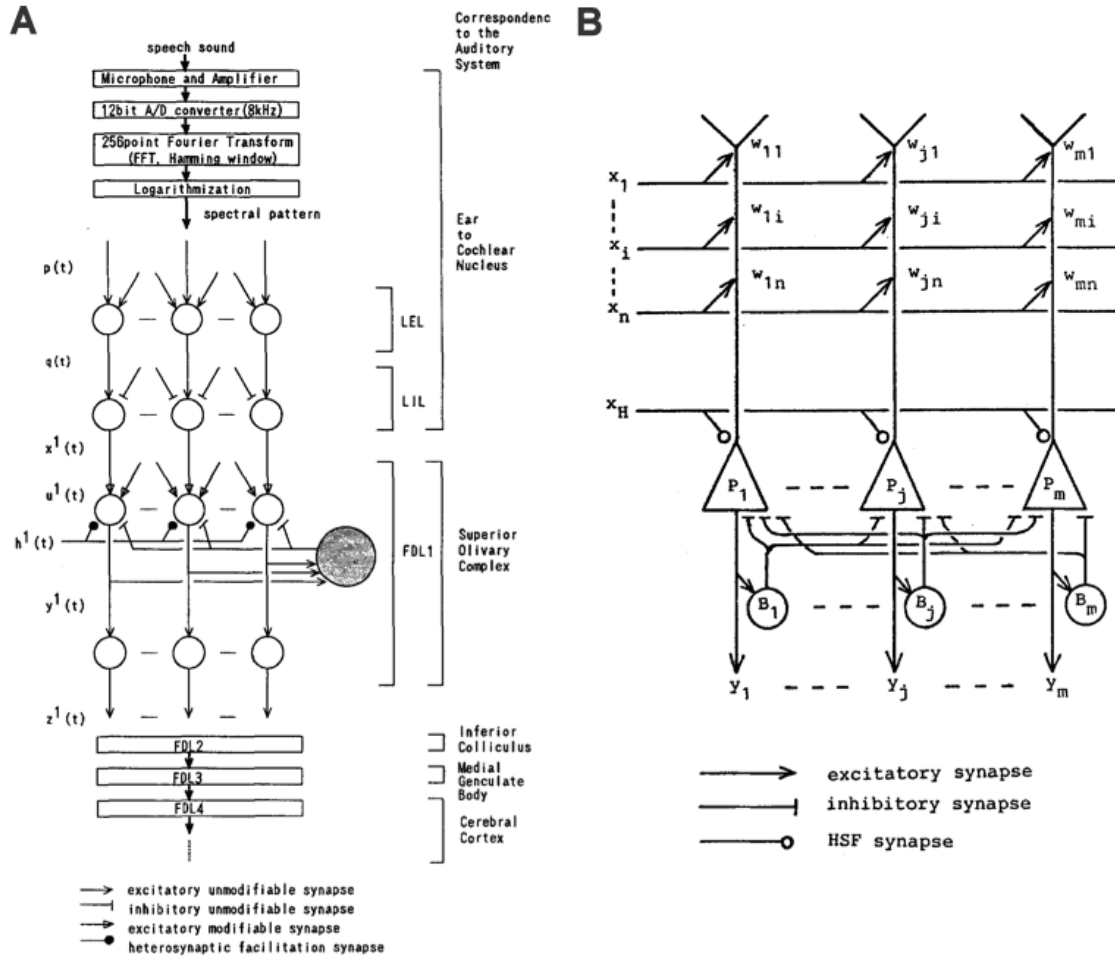


Figure 1.31: Schematic representation of the speech recognition model by Kurogi (1991). A: Full model. LEL is the lateral excitation layer, LIL is the lateral inhibition layer, FDL is the feature detection layer and its architecture is schematically depicted in B. B: Architecture of the feature detection layers based on the connectivity of the cortex used for higher auditory brain areas. Adapted from Kurogi (1991).

While the authors argue that their model is grounded in the neurophysiology of the auditory brain, this claim is rather loose, since the authors do not replicate the neurophysiology of the corresponding brain areas. Furthermore, their learning rule is not realistic, since no biologically plausible explanation of how their manual weight adjustment may take place in the brain is provided.

1.6.2 Spiking Models

1.6.2.1 Model by Meddis & O'Mard (2006)

A model relevant to the work described in this thesis was proposed by Meddis & O'Mard (2006). The authors modelled three stages of the sub-cortical auditory brain: the auditory nerve (AN), the chopper units of the ventral cochlear nucleus (CN), the inferior colliculus (IC). These were followed by the higher Level 4 neurons for which the corresponding brain area was not specified. All neurons (apart from the AN fibers) were modelled using MacGregor cells (MacGregor 1987).

It was shown that the model could accurately replicate many psychophysical results of pitch perception, thus demonstrating the potential computational mechanisms underlying pitch perception in the brain. The architecture of the model is shown in Figure 1.32. The AN fibers were organised into 40 characteristic frequency (CF) bands. Each CN chopper cell received inputs from 30 AN fibers from the same CF band. For each CF channel, there were 30 blocks of 10 CN cells each. Each CN block was characterised by a different intrinsic saturation chopping rate. Each IC unit received input from 10 CN cells with the same intrinsic chopping rate and acts as a coincidence detector, whereby they required coincident input from many CN neurons in order to fire. Each Level 4 cell received inputs from 40 IC cells across, one from each of the 40 CF bands, and with the same intrinsic chopping rate CN inputs. Their threshold was set low, so that Level 4 neurons would fire whenever any of their IC input neurons would fire, thus acting as spike counters.

As will be discussed later in this thesis, the architecture of the model suggested by Meddis & O'Mard (2006) is similar to the architecture we suggest for our model. Our model, however, relaxes the strong connectivity assumptions imposed on the model by Meddis & O'Mard (2006) and shown to be paramount for the model's

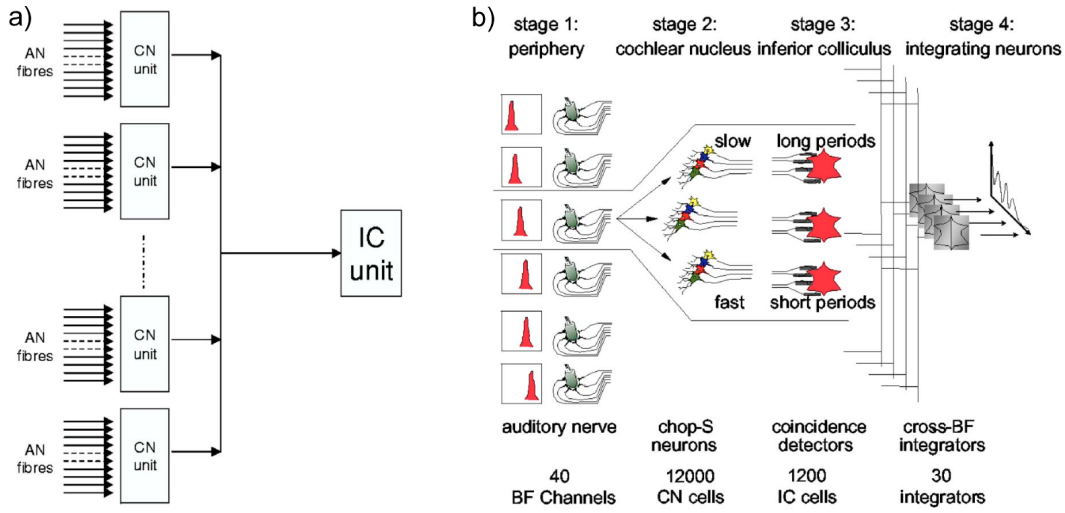


Figure 1.32: Schematic representation of the auditory periphery model by Meddis & O'Mard (2006). **a)** single module of the model. Each ventral cochlear nucleus (VCN) cell receives inputs from 30 auditory nerve (AN) fibers with the same characteristic frequency (CF). All VCN units within a module have the same saturated firing rate. Each inferior colliculus (IC) cell receives inputs from 10 VCN cells. **b)** arrangement of model modules. Each CF channel consists of 30 modules, each with a different saturating firing rate. There are 40 CF channels overall, with CFs ranging between 0.1 to 10 kHz. Each stage 4 unit receives one input from each CF channel from modules with the same firing rate.

operation. In our model the local connectivity for the AN→CN and CN→IC areas is drawn from a Gaussian distribution rather than hard-coded as in Meddis & O'Mard (2006). Furthermore, we extend the model by Meddis & O'Mard (2006) by adding onset and primary-like CN cells, as well as higher order A1 neurons with STDP-based learning. The addition of STDP-based learning in particular makes our model more powerful, whereby it is able to adapt to the underlying structure of the auditory stimuli through long-term STDP-based learning, while the model by Meddis & O'Mard (2006) lacks any learning capabilities.

1.6.2.2 Model by Masquelier et al. (2008)

A spiking neural network model capable of learning to pick out repeating spatio-temporal patterns from statistically identical noise using STDP-based learning was described by Masquelier et al. (2008). While the model is not aimed at replicating the auditory brain or performing speech recognition, it is included in this review because it demonstrates the power of spiking neural networks that can be applied in the context of speech recognition modelling.

The model described by Masquelier et al. (2008) is referred to as RPLM from here on (standing for Repeated Patterns Learning Model). RPLM consists of an input layer of 2000 spiking leaky integrate-and-fire neurons fully connected to one neuron in the output layer with equal weight. A spike pattern generated using Poisson process is applied to the input layer, with an arbitrarily chosen 50 ms long interval of the input raster acting as the repeating pattern, whereby it was presented a number of times during each simulation. It was found that the output cell learnt to only fire when the repeated pattern was presented and not during the statistically identical noise. The network worked as following: at the beginning of training the output cell fired at regular intervals during stimulus presentation. Due to the stronger LTD compared to LTP, random input firing resulted in the weakening of connections. Meanwhile, connections from the cells involved in the repeating pattern got strengthened every time the pattern was presented. Carefully chosen values for the relative strengths of LTD and LTP, as well as the initial strength of the afferents, resulted in the output cell pattern selectivity. The output cell learnt to fire whenever the pattern was present, and not to fire during the periods of noise (see Fig. 1.33). Masquelier et al. (2008) have shown that RPLM is robust, since its performance is preserved even when a combination of the following is present: the pattern makes up only a quarter of

the overall input duration and involves only half of all neurons in the input layer; 10 Hz Poissonian noise and 1 ms jitter is added to the pattern. Masquelier et al. (2008) have also shown that training conditions can be degraded even further while still preserving acceptable levels of performance (see Fig. 1.34 for the summary).

The principles described by Masquelier et al. (2008) can be used within the auditory cortex to learn statistical regularities underlying different auditory objects. While the model in its original form would only be able to learn short temporal patterns (on the order of single phonemes), the addition of temporal delays within the feedforward connections, recurrent connections, or the creation of a hierarchy of feedforward layers could allow neurons in the output layer to learn longer sequences spanning syllables and/or whole words.

1.6.2.3 Model by Nager et al. (2002)

A more advanced model of the auditory brain using a spiking neural network with leaky integrate-and-fire neurons was suggested by Nager et al. (2002), who used a multi-layer fully connected feed-forward spiking neural network with dynamic synapses to model the auditory cortex. The input to the model was pre-processed using a model of the cochlea and hair cells suggested by Lyon (1982), while the short-term dynamics and adaptation in the spiking neurons were modelled according to Reutimann et al. (1999).

It was found that when the network with four output neurons was trained on four different frequency modulation sweeps, it was able to learn to recognise them. The network was able to do that because short-term dynamics described by Reutimann et al. (1999) enabled each synapse to learn its preferred characteristic input pattern by encoding the spatio-temporal patterns into synaptic delays. After training, the input spikes corresponding to the presentation of a learnt stimulus arrived simultaneously to their corresponding post-synaptic neuron due to the

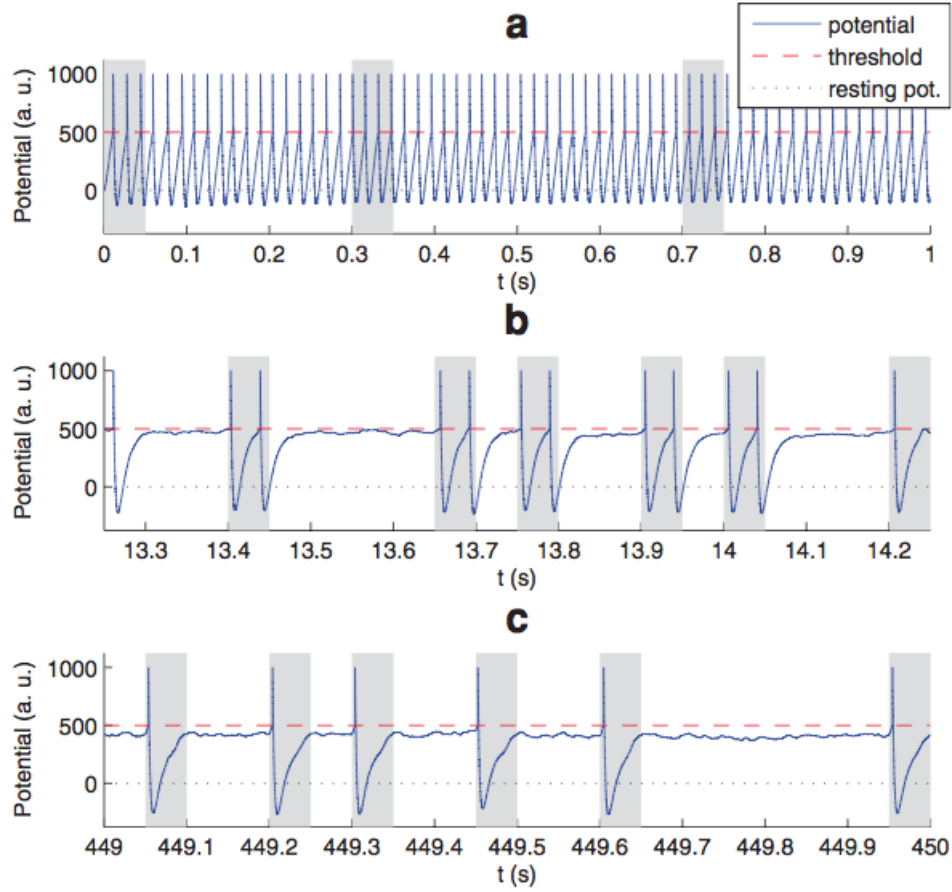


Figure 1.33: Schematic representation of how the repeated pattern learning model (RPLM) (Masquelier et al. 2008) functions. RPLM is a feed-forward one-layer spiking network with 2000 input cells and 1 output cell. **A:** at the beginning of training the output cell fires regularly to both Poisson noise (white regions) and the 50 ms windows where 50% of the input cells fire according to a pre-set pattern (grey regions); **B:** after about 70 pattern presentations the output cell mostly ceases to fire to noise; **C:** at the end of training the output cell fires to 99.1% of patterns and never to noise (adapted from Masquelier et al. (2008)).

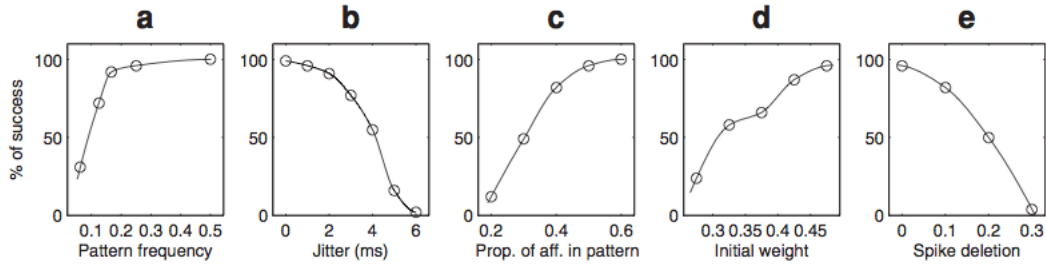


Figure 1.34: Plots illustrating the performance of repeated pattern learning model (Masquelier et al. 2008) in various conditions. The ordinate in all plots is the proportion of patterns recognised by the output neuron. This is measured after changing: **a** - pattern frequency (pattern duration divided by total input duration, where pattern duration is fixed at 50 ms); **b** - pattern jitter; **c** - proportion of input cells taking part in the pattern; **d** - initial connection strength; **e** - proportion of deleted pattern spikes (adapted from Masquelier et al. (2008)).

learnt synaptic delays. This network, however, was unable to perform long-term learning, since the delay learning described above operated on a short-term scale only. Furthermore, the model was unable to learn stationary signal and could only learn transitions in speech, since it failed at the single phoneme learning task where the phonemes were too spectro-temporally static for the network to learn.

1.6.2.4 Model by Coath et al. (2014)

Recently a spiking neural network model capable of recognising previously learnt frequency modulation (FM) sweeps implemented both in neuromorphic hardware and as a software simulation was suggested by Coath et al. (2014). Their model consisted of an array of 32 tonotopic channels interconnected through excitatory and inhibitory connections with axonal conduction delays and STDP-based learning as shown in Figure 1.35. Neurons A represent the tonotopic channels. Neurons B_2 are the outputs of the network, and their responses are modulated by neurons B_1 through direct inhibitory connections, as well as plastic excitatory connections with axonal conduction delays through neurons C .

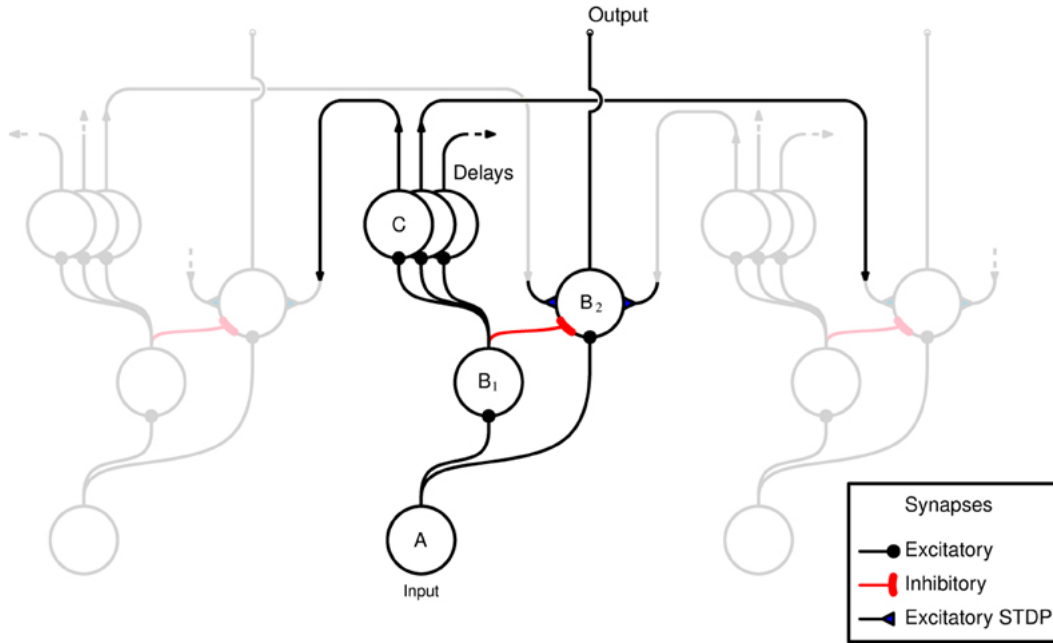


Figure 1.35: Schematic representation of the spiking neural network model as implemented in hardware and software by Coath et al. (2014). 32 A neurons form the tonotopic input layer. Only three A neurons are shown in the schematic. Neurons B_1 and B_2 receive excitatory connections from the corresponding A neurons in their tonotopic channel. Neurons B_2 serve as the output of the model. Their firing is modulated by an inhibitory connection from the B_1 neuron within their tonotopic channel and plastic excitatory connections with a distribution of axonal delays from a number of B_1 neurons from other tonotopic channels (adapted from Coath et al. (2014)).

The network was presented with FM sweeps and simplified representations of seven words: 'and', 'of', 'yes', 'one', 'two', 'three', and 'four'. The words were naturally spoken by a single male speaker and pre-processed, whereby approximate movements of the first two formants were extracted for each word, and the resulting FM sweeps were presented to the network by injecting stable currents at appropriate times into the corresponding tonotopic channels. Each time the network was trained on one stimulus, and tested on the full stimulus set. The authors were looking for higher levels of stimulus specific information using rate encoding of the output B_2 neurons in response to the previously trained stimulus compared

to the rest of the stimulus set. It was shown that the network was mostly able to achieve the expected levels of performance, even when the presented sweeps were embedded within some noise (same levels of noise were using for training and testing). The precise architecture of the model, however, was not grounded in the neurophysiology of the auditory brain, the stimuli were oversimplified, and the model was trained to do simple pattern matching rather than trying to achieve invariance learning.

1.6.2.5 Reservoir computing approaches

Another large class of models that use spiking neurons is that of liquid state machines (LSM), an instance of the more general reservoir computing approach. While these LSM models are not necessarily aimed at gaining insights into the operation of the brain, they nevertheless showcase the power of spiking neural networks as computational modules.

The LSM class of models takes a temporal stream $u(\cdot)$ as input, which is then fed into a randomly connected recurrent spiking neural network with fixed connectivity and weights (the 'liquid'). This 'liquid' acts as a non-linear kernel that performs an expansion and transformation of the input into a new state $x(\cdot)$, where $x(t)$ depends not only on $u(t)$, but also on previous inputs $u(t - \tau)$. This dependency is arbitrary and non-linear and it is governed by the random initialisation of the connectivity and connections weights of the 'liquid'. The 'liquid' is followed by a set of linear readout neurons that learn to map $x(t)$ to some target output $y(t)$. The linear readout neurons are the only parts of an LSM that are allowed to learn during training.

LSMs and reservoir computing approaches in general rely on three properties: separation, approximation and fading memory. Separation means that different inputs have to be mapped to different outputs by the 'liquid'. Approximation means

that inputs that are only slightly different from each other should be mapped to identical outputs. Otherwise, noisy instances of the same input would be classified as different classes. Fading memory property means that the 'liquid' should take into account the recent past when performing classification. For example, in speech the recent context of a phoneme can influence the phoneme's spectral form. The effect of context normally fades with time, so the 'liquid' should have fading memory. For a better overview of LSMs and reservoir computing in general, see Appeltant et al. (2011). The difficulty in designing LSMs comes from initialising the right 'liquid' in order for it to have the three properties described above for the types of stimuli used as input to the model.

Maass et al. (2004) demonstrated that an LSM where the 'liquid' was made up of randomly primarily locally connected integrate and fire neurons, 20% of which were chosen to be inhibitory, was capable of learning to correctly classify a subset of NIST TI-46 dataset, consisting of digits 'zero' to 'nine' spoken by 50 female speakers 10 times each. The raw auditory files were pre-processed before serving as input to the model. Each audio recording was transformed into 40 trains of transient spikes, whereby each spike generating input neuron was tuned to respond to the onset, offset or peaks of power modulated sine wave tones.

Maass et al. (2004) showed that their model was able to transform these arbitrarily pre-processed input spike rasters into 'liquid' representations suitable for linear readout neuron learning, hence demonstrating the power of the spiking 'liquid' in producing non-linear expansions of the input stream. The authors acknowledge, however, that the brain mechanisms to implement the linear supervised learning they suggest for readout neurons are yet unknown. Furthermore, the authors do not implement any learning in their 'liquid'. While they suggest that the addition of learning would only improve the performance of the model, they do not test this suggestion explicitly through simulations.

The work described in this thesis attempts to understand how the auditory brain transforms its input auditory nerve (AN) activity into robust invariant representations of auditory objects, such as words, in higher areas of the auditory brain hierarchy, such as the primary auditory cortex. While the work by Maass et al. (2004) demonstrates the power of spiking networks for implementing fading memory and non-linear kernel transformations, their work does not address the question above that this thesis tries to address. The input to the model by Maass et al. (2004) is a high level approximation of firing activity one might find in the higher parts of the auditory brain hierarchy (such as the primary auditory cortex) rather than the input auditory nerve (AN) stage. The model, therefore, provides no insights into how the auditory brain might transform the input AN activity into the robust encoding they suggest emerges in the 'liquid' of their model. Furthermore, the authors suggest that the encoding used by their proposed 'liquid' to represent auditory objects is different from rate, temporal and population codes commonly considered in the neuroscience literature. The authors, however, do not describe what this encoding implies.

1.7 Models of Spiking Neurons

One of the primary aims of this thesis is to understand the nature of computational mechanisms that allow the auditory brain to learn robust representations of speech-like auditory objects through the spiking dynamics of its constituent neurons, their ability to adapt through STDP-based learning mechanisms (Bi & Poo 1998), and their interactions with the statistics of the input stimuli. Furthermore, we are interested in investigating the nature of encoding used by the primary auditory cortex to represent the learnt stimulus class information. Temporal encoding presumes that the precise timing of individual spikes is informative, while rate encoding presumes that the information is carried by the average rate of firing of

neurons, and the precise spike timing does not carry information about the stimulus identity. Spiking neurons can implement both rate and temporal encoding of information, while rate neurons lose temporal information by definition. Hence, due to the nature of the questions we were trying to address in this thesis, our model had to be built using spiking rather than rate neurons. It was, therefore, important to select a good model of spiking neurons to be used as building blocks of our proposed model.

The main considerations when choosing a spiking neuron model were computational complexity and the range of neuron response properties. The model had to be computationally cheap to allow large scale simulations to be run in feasible amounts of time. Furthermore, the model had to be capable of different neuron response properties to model different types of neurons found in the auditory brain. A large number of spiking neural network models exist with varying degrees of trade-offs between computational complexity and richness of response properties (Fig. 1.36). Three spiking neuron models were primarily considered for the purposes of this thesis and are discussed next. A review of other spiking neural network models can be found in Izhikevich (2004) and Gerstner & Kistler (2004).

1.7.1 Hodgkin-Huxley

The Hodgkin-Huxley spiking neuron model is the oldest and one of the most important neuron models in computational neuroscience (Hodgkin & Huxley 1952). It is a conductance based model based on experiments with a giant squid axon that describes the evolution of cell membrane potential through time according to the following:

$$C_m \frac{dV}{dt} = - \sum_k I_k(t)$$

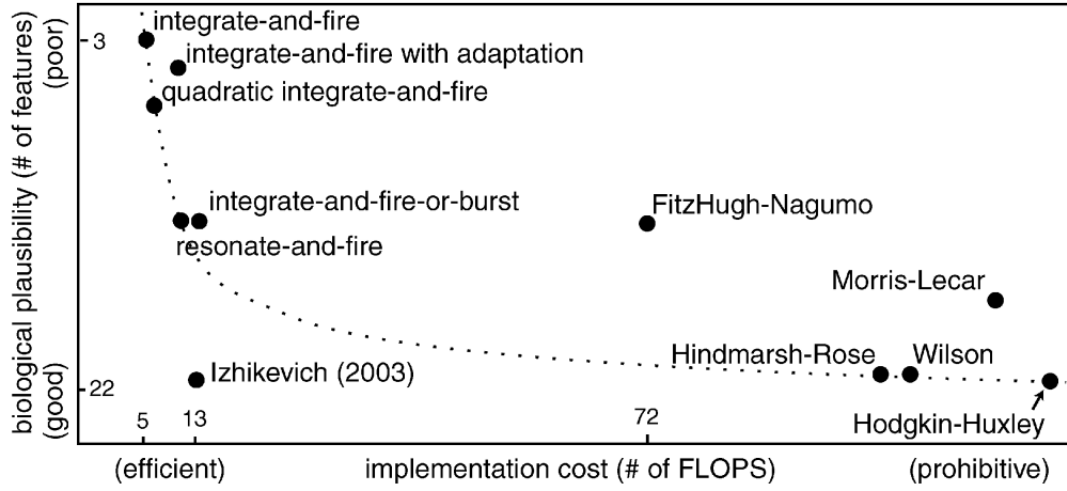


Figure 1.36: Comparison of spiking neuron models in terms of computational complexity (abscissa) and biological plausibility (ordinate). Computational complexity refers to the approximate number of floating point operations per simulation time step, and biological plausibility refers to the range of different neuron response properties available to the model. Adapted from Izhikevich (2004).

$$\sum_k I_k = g_{Na} m^3 h (V - E_{Na}) + g_K n^4 (V - E_K) + g_L (V - E_L)$$

where C_m is the cell membrane capacitance; V is the cell membrane potential; $\sum_k I_k(t)$ is the sum of ionic currents passing through the cell membrane (sodium (Na), potassium (K) and a leakage (L) ion thought to be chlorine); $g_{\{K,Na,L\}}$ is ion conductance; $E_{\{K,Na,L\}}$ is ion reversal potential; and m , n and h are gating variables corresponding to the number of channels open for each ion type. The gating variables evolve through time according to the following:

$$\begin{aligned} \frac{dm}{dt} &= \alpha_m(V)(1 - m) - \beta_m(V)m \\ \frac{dn}{dt} &= \alpha_n(V)(1 - n) - \beta_n(V)n \\ \frac{dh}{dt} &= \alpha_h(V)(1 - h) - \beta_h(V)h \end{aligned}$$

The measurements of ions' reversal potentials, membrane conductance, and functions $\alpha_x(V)$ and $\beta_x(V)$ were adjusted to fit the response properties of the giant squid axon.

The Hodgkin-Huxley spiking neuron model and its subsequent variants are considered to be some of the most detailed and accurate neuron models, whereby their parameters are biophysically meaningful and measurable, and they are capable of replicating many neuronal response properties. The Hodgkin-Huxley model, however, is prohibitively computationally expensive, requiring around 1200 floating point operations per cell per 1 ms of simulation time (Fig. 1.36) (Izhikevich 2004), which makes it impossible to model more than a very limited number of neurons at a time. The Hodgkin-Huxley model was, therefore, dismissed as a potential choice for the spiking neural network model of the auditory brain described in this thesis.

1.7.2 Integrate-and-Fire

The leaky integrate-and-fire (LIF) model is one of the most popular spiking neuron models in computational neuroscience due to its simplicity and computational cheapness (Fig. 1.36). The model makes a simplifying assumption by ignoring the shape of action potentials, whereby spikes are treated as uniform events defined by their time of occurrence, and the refractory period is explicitly implemented. The LIF neuron model is derived from a simple RC circuit driven by an input current ($I(t)$) according to the following:

$$\tau_m \frac{dV}{dt} = -V(t) + RI(t)$$

where τ_m is the membrane time constant, V is voltage, R is resistance and I is the input current. A LIF neuron model was considered suitable for the spiking neural network model of the auditory brain described in this thesis, since it is

computationally cheap, requiring only four floating point operations per neuron per 1 ms of simulation time (Fig. 1.36), and its simplifying assumptions were not expected to significantly affect the dynamics of the full auditory brain model. The model, however, lacks a full enough range of available neuron response properties, since in the original form it is only capable of one behaviour, whereby it acts as an integrator and fires tonic spikes with constant frequency. Since neurons in the auditory brain have a wide range of response properties (see Section 1.5), a spiking neuron model with well defined and easily interchangeable response properties is required to build a model of the auditory brain. While the LIF model parameters can be adjusted to achieve some extra response properties, this process is not straightforward. Furthermore, since the LIF model has only one variable, it is unable to achieve some response properties, including phasic spiking and bursting (Izhikevich 2004).

1.7.3 Izhikevich

The null-cline derived model of spiking neurons by Izhikevich (2004) achieves the same wide range of response properties as the Hodgkin-Huxley model, while being relatively computationally cheap (Fig. 1.36). Such behaviour is achieved by higher-order modelling, whereby the aim is to replicate the behaviour of neurons, rather than to faithfully replicate the low-level dynamics of the flow of ions in order to produce the neuronal behaviour. While the model parameters are not biophysically meaningful and measurable, this is not expected to affect the validity of the full spiking neural network model of the auditory brain described in this thesis. The computational cheapness of the Izhikevich neurons of thirteen floating point operations per neuron per 1 ms of simulation time (Fig. 1.36), combined with a wide range of well defined and parametrised response properties make the neuron

model by Izhikevich (2004) the model of choice for the purposes of this thesis. Full details of the model are provided in Chapter 2.

2

Methods

2.1 Introduction

This chapter discusses the methods used in the chapters to follow.

2.2 Stimuli

For Chapters 3 and 4 a stimulus set consisting of twelve exemplars of each of two vowels, /i:/ and /a/, was generated using the Klatt synthesiser (Klatt 1980) and sampled at 48,828 Hz. The stimuli were chosen to be as simple and as controlled as possible, while still being representative of natural speech. The two particular vowels, /i:/ and /a/, were chosen because they have one of the largest euclidean distances from each other based on their constituent formant frequencies as described by Peterson & Barney (1952) (see Fig. 1.5 and Table 2.1). These differences were expected to make the vowels easier to discriminate by the model. Synthesised vowels were chosen over naturally spoken equivalents in order to control for

pronunciation differences other than variations in formant frequencies.

Each 100 ms long artificial vowel sound was created by sampling each of the three vowel formants from a uniform 200 Hz distribution centered around the corresponding formant frequency as reported by Peterson & Barney (1952) for male speakers. The variability in formant frequencies among the twelve stimulus exemplars was consistent with the range of variation present in natural human speech as demonstrated in Figure 2.1 (Huckvale 2004). Informal tests showed that greater variation in vowel formant frequencies or sampling the formant frequencies from a different (Gaussian) distribution resulted in many vowel exemplars that sounded perceptually different to /i:/ or /a/. This is not surprising, since the range of within-speaker formant frequency variation described by Huckvale (2004) was estimated from natural speech, in which vowels could appear in different acoustic contexts. Here we define acoustic context as the set of phonemes immediately surrounding the chosen vowel. Such acoustic context can drastically affect the frequencies of formants making up the vowel while preserving the vowel’s perceptual identity. If such a vowel with distorted formant frequencies is presented without its original acoustic context, however, it may no longer be possible to identify the perceptual class of the vowel correctly. All sound files that were chosen as stimuli for the simulations described in this thesis could be correctly classified by human subjects when presented on their own without an acoustic context.

A fundamental frequency (F_0) of 100 Hz was used for all stimuli, which is within the range of fundamental frequencies reported in natural speech (male range is 80-160 Hz and female range is 140-280 Hz (Olive et al. 1993)). The 100 ms duration for the vowels was chosen, since it has been reported that in natural speech vowels are presented on average at the rate of around 10-12 per second (Maclay & Osgood 1959).

Phonetic Symbol	Example Word	F ₁ (Hz)	F ₂ (Hz)	F ₃ (Hz)
/ɔ:/	<i>bought</i>	570	840	2410
/u:/	<i>boot</i>	300	870	2240
/ʊ/	<i>foot</i>	440	1020	2240
/a/	<i>hot</i>	730	1090	2440
/ʌ/	<i>but</i>	640	1190	2390
/ɜ:/	<i>bird</i>	490	1350	1690
/æ/	<i>bat</i>	660	1720	2410
/e/	<i>bet</i>	530	1840	2480
/ɪ/	<i>bit</i>	390	1990	2550
/i:/	<i>beet</i>	270	2290	3010

Table 2.1: Formants of ten most common vowels averaged across 76 male speakers (Peterson & Barney 1952).

The stimuli were saved as .wav files and normalised for loudness using the root mean square measure of the average power of the signal. This was done to remove any potential effect of stimulus loudness on learning.

2.2.1 Stimulus Presentation Paradigm

The vowel stimuli belonging to the two classes, /i:/ and /a/, were presented in an interleaved fashion and separated by 100 ms of silence. The silence encouraged the models to learn separate representations of each individual vowel class and to avoid learning transitions between vowel classes. We introduced the silence, because in our initial simulations we were only interested in understanding the contributions of the different subcortical areas of the auditory brain in enabling the primary auditory cortex to learn the ability to classify two vowel classes invariantly to the perturbations of the vowels' constituent formants. Learning transitions between vowels was seen as a confounding factor for the initial simulations. The ability of the model to deal with phoneme transitions by learning to categorise simple naturally spoken words was addressed in Chapter 6.

This simple two vowel categorisation task was of an appropriate complexity for

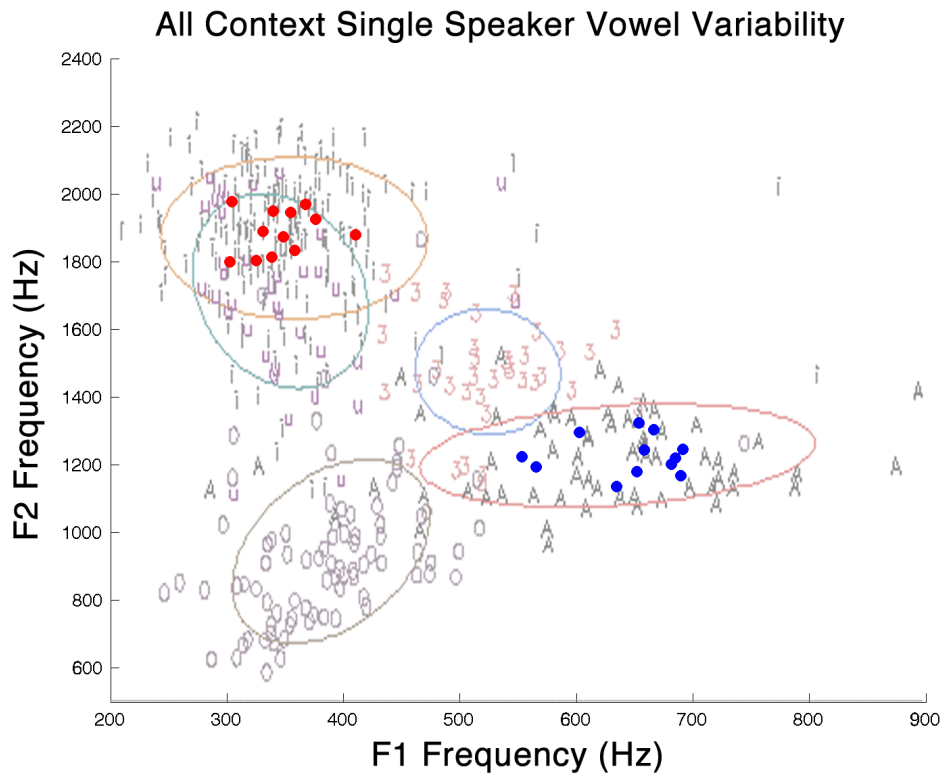


Figure 2.1: Schematic representation of twelve exemplars of two synthesised vowels (/a/ - blue, /i:/ - red) projected onto the two-dimensional plane defined by the first two formants of the vowels. Each exemplar was generated by randomly sampling a 200 Hz distribution centered around the corresponding formant frequency as reported by Peterson & Barney (1952) for each of the three vowel formants. It can be seen that the generated vowel exemplars are in line with the vowel distribution clouds produced from natural speech of a single speaker (Huckvale 2004). All exemplars have been manually checked by human subjects to ensure that they are recognisable as either an /a/ or an /i:/. The ellipses approximate the 70% within-speaker variability boundary for a particular phoneme class.

the initial simulations, because a simple model of the auditory brain lacking the subcortical pre-processing stages CN and IC was unable to solve the task using any of the extensive number of hyper-parameter value combinations sampled through a grid search. The task could only be solved by a model that incorporated the CN and the IC, and only when the hyper-parameter values were set within a particular area within the multi-dimensional hyper-parameter search space (see Chapter 3).

We used 200 training and twenty testing epochs, whereby each epoch consisted of the first exemplar of vowel /i:/ followed by the first exemplar of vowel /a/, followed by the second exemplar of vowel /i:/ up to the last twelfth exemplar of vowel /a/. Twenty (rather than one) test epochs were used because, due to the stochasticity of AN responses, input AN spike patterns in response to repeated presentations of the same sound were not identical. The fixed presentation schedule used for the simulations described in this paper was an arbitrary choice, however, informal tests demonstrated that the order in which the vowel exemplars were presented did not make a qualitative difference to the performance of the trained models, as long as any two consequent presentations of exemplars from one vowel class were not separated by too long a streak of exemplars from a different vowel class. This is in line with the work by Masquelier et al. (2008). See Section 2.5.3 for a further discussion of the chosen stimulus presentation paradigm.

2.2.2 Extra vowel stimuli

In Chapter 5 we investigated whether our proposed model is able to learn transform invariant representations of more than two vowels. Therefore, additional four vowel stimuli to make a set of six vowel classes in total were generated using the Klatt synthesiser (Klatt 1980) in the same way as described above for vowels /i:/ and /a/. The number of vowels was limited to six due to the computational restrictions that make simulations with more than six vowels prohibitively long. Each vowel had

twelve exemplars, whereby the frequency of each of the three vowel formants was picked from a uniform distribution spread over 200 Hz around the corresponding central vowel formant frequency as shown in Table 2.1.

In addition to the original two vowels /i:/ and /a/, the following vowels were generated: /ɔ:/, /u:/, /e/, and /æ/. These vowels were chosen because they occupy the least overlapping spaces in the two-dimensional plane defined by their first two formants (Fig. 2.2). Each of the twelve exemplars of the six vowels were manually checked by human subjects to ensure that they were perceptually recognisable. The stimuli were presented to the network according to the paradigm described in Section 2.2.1 over 100 training and 20 testing epochs.

2.2.3 Naturally spoken word stimuli

Digits 'one' and 'two' naturally pronounced twice each by 94 speakers were used as stimuli for Chapter 6. The first utterance by each speaker of each digit constituted the training set and the second utterance by the same speakers constituted the testing set. The utterances were on average 702 ms long (ranging between 234-1170 ms). Each utterance was normalised for loudness using the root mean square measure of the average power of the signal. This was done to remove any potential effect of stimulus loudness on learning. Ten training and four testing epochs of the 94 different pronunciations of each of the two digits ('one' and 'two') were presented to the network.

Example spectrograms of digit 'one' spoken by speaker 1 and of digit 'two' spoken by speaker 3, as well as their resulting AN firing rasters can be seen in Figure 2.3. It can be seen that the AN firing rasters roughly reflect their corresponding sound spectrograms, and that the vowel onset time, which results in the first major burst of activity within the AN fibers with lower characteristic frequencies, is different for the two vowels. It can also be seen that the fundamental frequencies

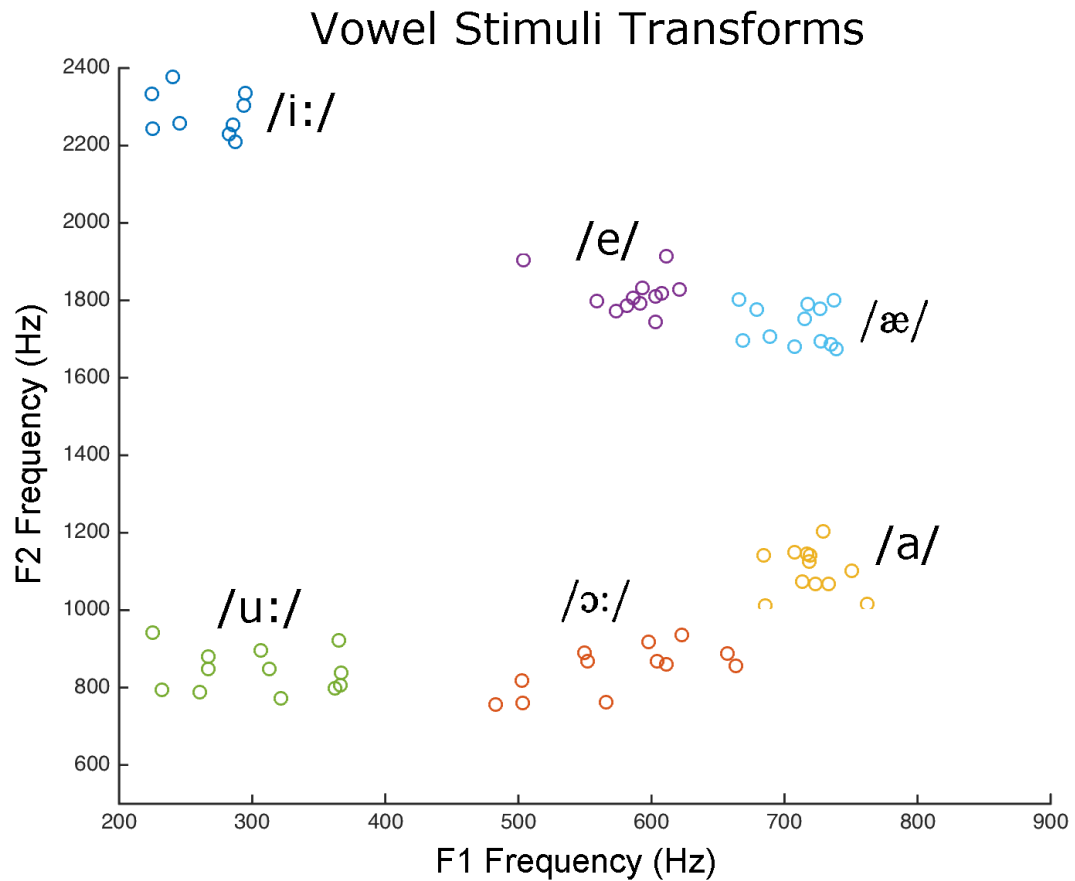


Figure 2.2: Schematic representation of twelve exemplars of six synthesised vowels (/a/, /i:/, /ɜ:/, /u:/, /e/, and /æ/) projected onto the two-dimensional plane defined by their first two formants. Each exemplar was generated by randomly sampling a 200 Hz distribution centered around the corresponding formant frequency as reported by Peterson & Barney (1952) for each of the three vowel formants. All exemplars have been manually checked by human subjects to ensure that they are recognisable as their corresponding vowels.

of the two digits are different for the two different speakers. This is evidenced by the larger gap between the columns of concurrent spikes in the AN firing raster for digit 'one' compared to that for digit 'two'.

2.3 Learning in the Auditory Brain

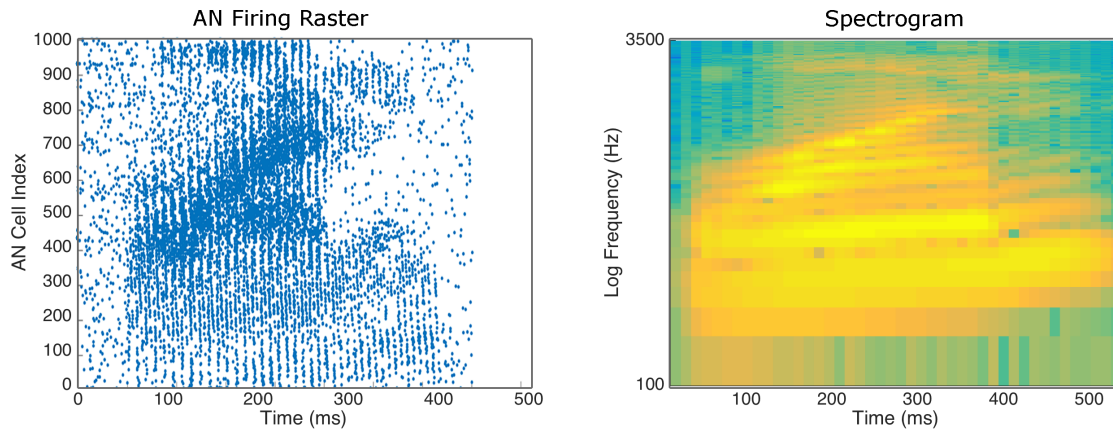
2.3.1 Plasticity

Neurophysiological studies have demonstrated that different parts of the cortex have similar anatomical and physiological structure even if they process different types of sensory inputs (Rolls & Treves 1998). For example, Frost (1990), Sharma et al. (2000), von Melchner et al. (2000) have demonstrated that if visual input is rewired to the auditory cortex in infant ferrets and hamsters, cells in the auditory cortex develop response properties appropriate for vision (such as feature-selective visual responses, retinotopic maps and orientation columns). This suggests that response properties of cortical cells are learnt through exposure to the spectrotemporal structure of sensory input using some form of synaptic plasticity. Neurophysiological evidence for spike-time dependent plasticity (STDP) learning in the auditory cortex has been reported by Dahmen et al. (2008) and Karmarkar et al. (2002).

2.3.2 Spike-Time Dependent Plasticity Learning

Spike-Time Dependent Plasticity (STDP) learning is a type of Hebbian learning (Hebb 1949), whereby when a pre-synaptic spike occurs just before a post-synaptic spike, the connection between the two cells is strengthened through Long Term Potentiation (LTP), whereas when a pre-synaptic spike occurs just after a post-synaptic spike, the connection is weakened through Long Term Depression (LTD) (Fig. 2.4). The magnitude of change depends on the time interval between the

ONE - speaker 1



TWO - speaker 3

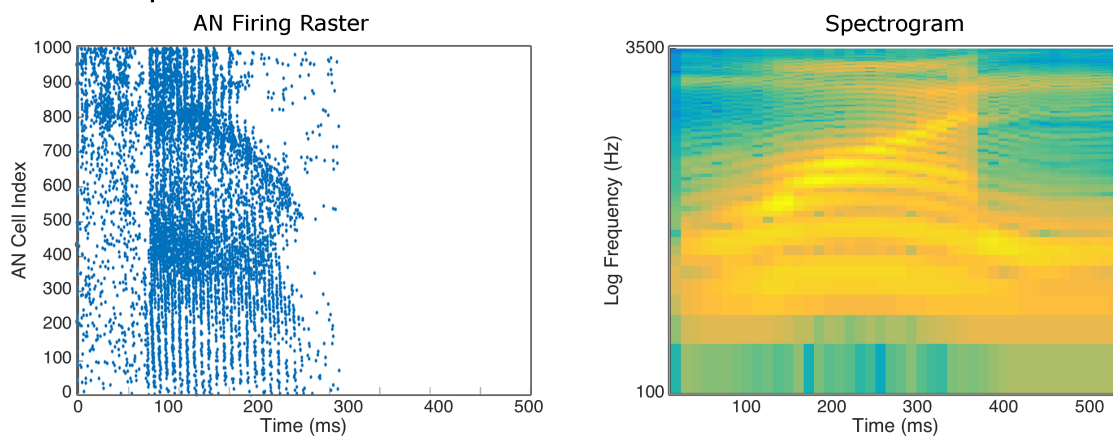


Figure 2.3: Spectrograms (right) and auditory nerve (AN) firing rasters (left) in response to naturally spoken digit 'one' pronounced by speaker 1 and naturally spoken digit 'two' pronounced by speaker 3. The abscissa of both types of graphs represents time (ms), while the ordinate represents the log frequency (Hz) in the spectrograms and the AN cell index for the firing rasters. The characteristic frequencies (CFs) of the AN fibers roughly correspond to the horizontally aligned frequencies in the spectrograms. It can be seen that the AN firing rasters roughly match their corresponding spectrograms. It can also be seen that the speaker fundamental frequencies are different for the two speakers, which is represented by the larger gaps between the vertical lines of concurrent AN cell firing in the 'one' firing raster compared to the 'two' firing raster. Furthermore, the vowel onset time happens later in response to digit 'two' compared to digit 'one', as evidenced by the later onset of firing within the AN fibers with lower CFs.

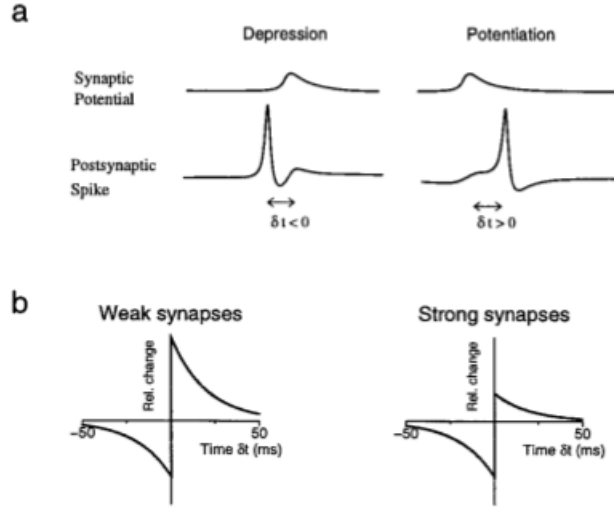


Figure 2.4: Illustration of spike-timing dependent plasticity learning. **a** - if a pre-synaptic spike occurs just before a post-synaptic spike, their connection is strengthened through Long Term Potentiation (LTP). If a pre-synaptic spike occurs just after a post-synaptic spike, their connection is weakened through Long Term Depression (LTD). **b** - the relative change in connection strength is exponential to the relative timing of the pre- and post-synaptic spikes. The magnitude of the change in connection strength depends on the current strength of the connection for LTP, but not for LTD. Adapted from van Rossum et al. (2000).

pre- and post-synaptic spikes, with smaller absolute differences resulting in exponentially stronger changes (Bell et al. 1997, Bi & Poo 1998, Markram et al. 1997).

The following equations describe the implementation of STDP-based learning within the proposed neural network model of the auditory brain. The weight update is scaled by:

$$f(s_{ij}) = \begin{cases} \alpha_p e^{-s_{ij}/\tau_p}, & \text{if } s_{ij} > 0 & LTP \\ \alpha_d e^{s_{ij}/\tau_d}, & \text{if } s_{ij} < 0 & LTD \end{cases} \quad (2.1)$$

where τ_p and τ_d are STDP time constants, α_p and α_d are constant coefficients, and s_{ij} is the time difference between a post- and a pre-synaptic spike calculated according to the following:

$$s_{ij} = t_i - (t_j + \Delta_{ij})$$

where t_i is the time of the post-synaptic spike, t_j is the time of the pre-synaptic spike, and Δ_{ij} is the magnitude of the axonal conduction delay between the pre- and post-synaptic cells. All delays were treated as axonal, whereby each pre-synaptic spike time was set to be the time of spike arrival to the post-synaptic cell, rather than the time of pre-synaptic spike discharge.

Neurophysiological evidence suggests that STDP time constants are asymmetric and are equal to 17 ± 9 ms for LTP and 34 ± 13 ms for LTD (Bi & Poo 1998). Therefore, the default values of τ_p and τ_d were set to 15 ms and 25 ms correspondingly as suggested by Perrinet et al. (2001).

The equations above calculate a scaling variable $f(s_{ij})$, which can then be used to update the synaptic weights according to one of three paradigms: additive, multiplicative, or mixed.

The additive paradigm produces synaptic weight changes independent of the instantaneous synaptic strength, and results in a bimodal distribution of afferent connection strength values, whereby all connection weights become either fully strengthened or fully weakened (Rubin et al. 2001, van Rossum et al. 2000). This requires artificial minimum and maximum connection strength bounds to be set to prevent the weights from running towards plus and minus infinity.

The multiplicative paradigm produces synaptic weight changes proportional to the current synaptic strength, and results in a stable unimodal distribution of afferent connection strength values. A unimodal distribution is more neurophysiologically plausible than a bimodal distribution (Rubin et al. 2001, van Rossum et al. 2000). When a multiplicative paradigm is used, an equilibrium state independent of the initial weight distribution is reached. It is defined as a state

whereby individual connection weights change, but the overall distribution of connection weights remains stable. Multiplicative learning results in an equilibrium because it creates two forces acting in different directions (van Rossum et al. 2000). For LTP, the first force pushes stronger connection strengths to strengthen even further, since their probability of firing and, therefore, their probability of potentiation increases. The second force keeps the connections from running to infinity due to the dependence of the weight change on the difference between the current synaptic strength and an upper bound w_{ij}^{max} , with stronger synapses resulting in less potentiation than weaker synapses (Bi & Poo 1998, Debanne et al. 1996, 1999).

Neurophysiological data suggests that the LTD magnitude is independent of the instantaneous synaptic strength (w_{ij}), while the magnitude of LTP changes inversely proportionally to connection strength, with stronger synapses resulting in less LTP than weaker synapses (Bi & Poo 1998, Debanne et al. 1996, 1999). These findings are modelled using a mixed STDP paradigm as shown in the equations below, whereby additive learning is used for LTD and multiplicative learning is used for LTP. The minimum bound of 0 nA was set to ensure that the additive LTD did not run away to $-\infty$. The differential equations shown below lead to the bounding of weights in the interval $[0, w_{ij}^{max}]$.

$$w_{ij}(t+1) = \begin{cases} w_{ij}(t) + (w_{ij}^{max} - w_{ij}(t))f(s_{ij}), & \text{if } s_{ij} > 0 \\ w_{ij}(t) + w_{ij}^{max}f(s_{ij}), & \text{if } s_{ij} < 0 \end{cases} \quad \begin{matrix} LTP \\ LTD \end{matrix} \quad (2.2)$$

For the purposes of the simulations described in this chapter, whenever the same pre-synaptic cell fired more than once within the STDP time window, only the first spike was used in STDP calculations (van Rossum et al. 2000). This 'nearest only' paradigm is in contrast with the alternative 'all pairings' paradigm, whereby all pre-synaptic spikes for each cell take part in STDP (Kempster et al. 1999, Song et al. 2000). It has been demonstrated that either paradigm results

in the same equilibrium state when used with mixed learning (van Rossum et al. 2000). The 'nearest only' paradigm was chosen for the simulations described in this chapter because it is less computationally expensive.

2.4 Delays

Neurophysiological data collected from different species suggests that cortical axonal connections, including those within the auditory brain, have conduction delays associated with them (Ferster & Lindstrom 1983, Liberman et al. 1967, Miller 1987, 1996, Salami et al. 2003, Swadlow 1985, 1988, 1992, 1994). These delays form a uniform distribution ranging approximately from 0.1 ms to 44 ms. The precise range depends on the type and location of the pre- and post-synaptic neurons. Each axonal delay between two neurons has been shown to be stable and precise with sub-millisecond precision (Swadlow 1985, 1994).

Axonal conduction delays are hypothesised to play an important role in auditory object representation learning through increasing the neurons' temporal integration time window and revealing otherwise hidden patterns within the afferent spike raster. Temporal integration time window is defined as the maximal length of input activity pattern that a particular neuron can learn to respond to.

The first role of axonal conduction delays is to increase temporal integration time window of neurons throughout the hierarchy of the auditory cortex, as suggested by neurophysiological evidence (Covey & Casseday 1999). Without conduction delays each coincidence detector cell has narrow temporal tuning ranging from around 3 ms to 12 ms as specified by the neuronal time constant (Gerstner et al. 1996). This means that neurons in the first cortical layer can learn input patterns that are no longer than 12 ms long at most. Since on average a word is around 500-600 ms long (Maclay & Osgood 1959), approximately 40-50 cortical layers would be necessary for an output layer cell to have temporal tuning spanning the

length of the full auditory object. However, when axonal conduction delays drawn from a 0-50 ms uniform distribution are introduced to afferent connections, each coincidence detector cell now has much wider temporal tuning, requiring around 10-12 cortical layers to enable output layer cells to learn representations of full auditory word objects.

Another mechanism through which axonal conduction delays are hypothesised to help with auditory object learning is by creating a large number of permutations of the original input pattern - one for each neuron in the post-synaptic layer of the network. Each post-synaptic cell has a different pattern of afferent axonal conduction delays, which forms a different transformation matrix for each post-synaptic cell. Input spike raster transformation through these different delay matrices can reveal otherwise hidden stable spike patterns within the original spike raster (see Fig. 2.5). Therefore, the more post-synaptic cells there are in a network, the more permutations of the original input raster are searched for patterns, and, hence, the more hidden patterns are revealed. The more patterns there are associated with a stimulus, the more likely it is to find a pattern that is transform invariant. Such stable patterns can be exploited by the network to learn transform invariant representations of auditory objects. The range of delay magnitudes determines the length of patterns that can potentially be discovered.

2.5 Learning Transform Invariant Representations of Auditory Objects

In order for a spiking neural network to learn robust auditory categories, it should use the statistical structure within its input to develop sparse, efficient and informative output representations that minimize the within-class output response differences, and maximize the between-class output response differences. In other words, a spiking neural network should effectively act as the encoder part of an

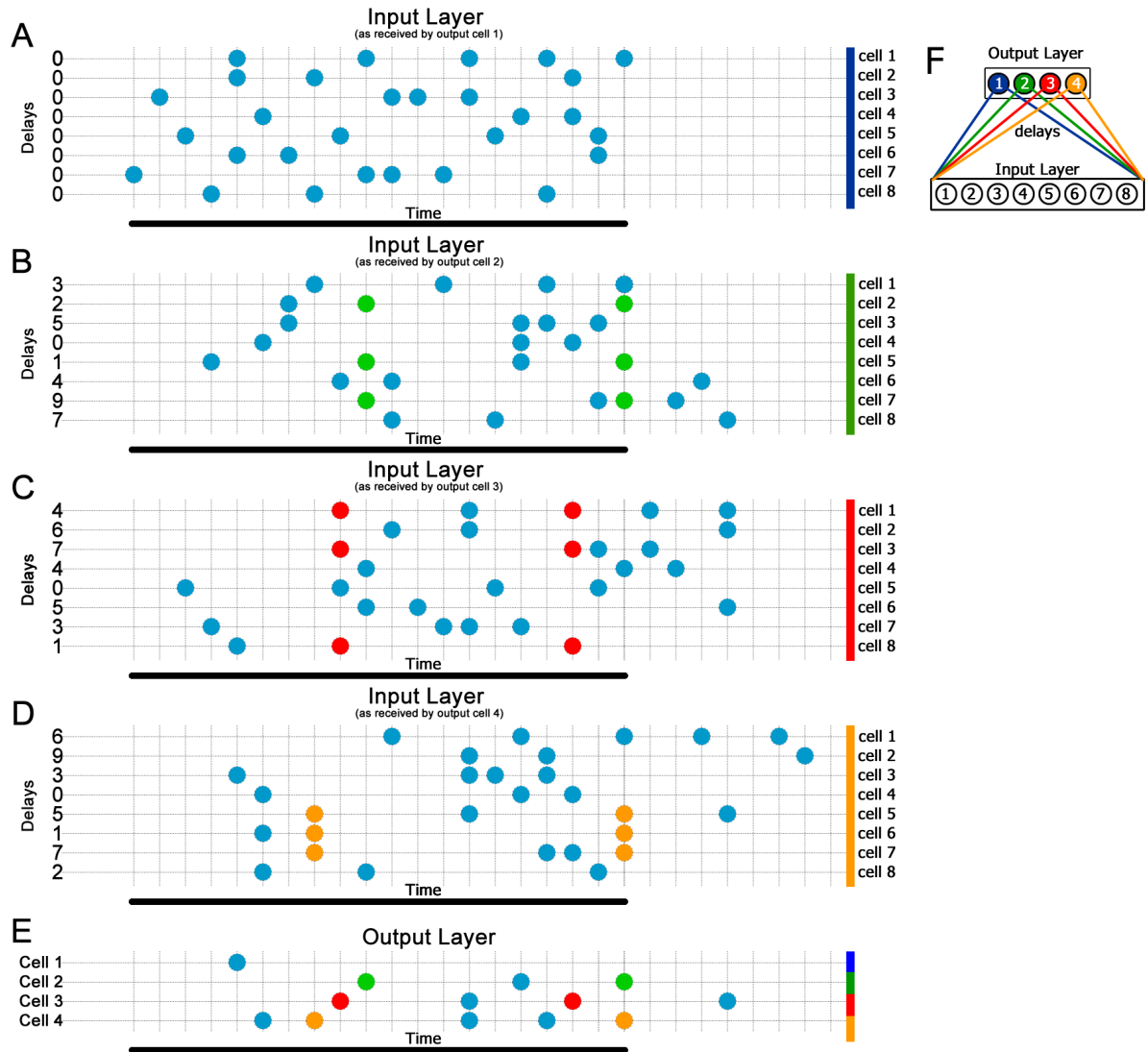


Figure 2.5: Schematic illustrating the role of heterogeneous axonal conduction delays in spike-timing dependent plasticity learning. In this simplified illustration a two layer model is used with 8 input cells and 4 output cells (F). For illustration purposes all output cells have a threshold of 3, and all connection strength magnitudes are set to 1. An output cell fires only if it receives at least 3 coincident spikes. Input layer spike raster in response to a stimulus (illustrated as a bold horizontal black line) is shown in A. Each of the 4 output layer cells has a distinct distribution of axonal conduction delays. For illustration purposes all connections between the input layer and cell 1 in the output layer have 0 ms delays. Therefore, output cell 1 sees the raster of spikes shown in A, and fires once at time 5 when it receives a coincident input from 3 input cells (E). Afferent connections for output cells 2-4 have randomly set delays (shown on the left of each corresponding sub-plot B-D). The spike rasters received by output layer cells 2-4 after the input spike raster shown in A is transformed by the corresponding delay vector for each of the output layer cells are shown in B-D correspondingly. The firing raster in A appears to have no repeating patterns of at least three simultaneously arriving spikes, and it produces one random spike in output cell 1. The repeating patterns (highlighted in colour) hidden within the firing raster in A emerge, when they get transformed using an appropriate distribution of delays into the spike rasters in sub-plots B-D. The resulting activation of output cells 2-4 are shown in plot E. Coloured nodes represent responses to repeated patterns, while blue nodes represent other responses.

autoencoder (Olshausen 1996), albeit using only physiologically realistic, local STDP-based learning (Bi & Poo 1998) without the need for feedback or backpropagation of error.

We hypothesised that such category learning is possible through the Continuous Transformation (CT) learning principle (Evans & Stringer 2012, Stringer et al. 2006). CT learning is a biologically plausible mechanism which assumes that highly similar, overlapping input patterns are more likely to be different exemplars of the same stimulus class. CT learning then binds these similar input patterns together onto the same subset of higher stage neurons, which, thereby, learn to be selective and informative about their learnt preferred stimulus class.

CT learning is essentially just a label given to Hebbian learning applied to inputs with highly overlapping neighbouring within-class exemplars. When Hebbian learning is applied to a dataset containing stimuli with continuous transformations (and the conditions for CT learning are therefore satisfied) a different class boundary is learnt compared to when Hebbian learning is applied to a dataset with discrete transformations (Fig. 2.6). In other words, a model with Hebbian learning trained on stimuli with high similarity between the neighbouring transforms of each stimulus class (thus operating using the CT principle) would learn a different type of class boundary compared to the same Hebbian model trained on stimuli without such statistics. In the latter case, the model will learn the average of each stimulus class and will classify the exemplars of the different stimulus classes in a similar way to a k-means clustering algorithm. A model operating under the CT principle, however, can learn class boundaries of arbitrary shapes (Fig. 2.6). Another way to interpret CT learning is as the biologically plausible mechanism for acquiring transform invariance orbits for visual objects in an unsupervised fashion as postulated by Liao et al. (2013).

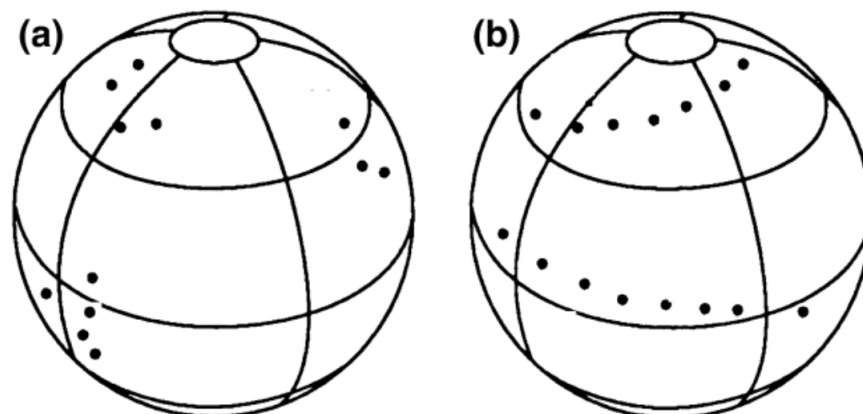


Figure 2.6: Schematic representation of how class boundaries learnt through continuous transformation (CT) Hebbian learning differ from those learnt through conventional competitive Hebbian learning. **a:** During conventional competitive learning, the model learns to point its weight vectors towards the centers of the object class clusters. Here three object classes are shown with 3-5 exemplars each. **b:** During CT learning, the model can learn arbitrarily shaped class boundaries. Here two object classes are shown. Their exemplars span an elongated region of space, and for each object class the exemplars at the opposite ends of the corresponding space have no overlap (correlation). CT learning allows the model to learn to categorise the exemplars of the two classes correctly.

The CT learning mechanism was originally described to account for geometric transform invariance learning in a rate-coded neural network model of visual object recognition in the ventral visual stream (Stringer et al. 2006), but has recently been shown to also work in a spiking neural network model of the ventral visual stream (Evans & Stringer 2012).

2.5.1 Continuous Transformation Learning in Visual Object Recognition

In vision, simple changes in the geometry of a scene, such as a shift in location or rotation (so-called 'affine transformations'), can generate a multitude of visual stimuli which are all different views, or 'transforms', of the same object. CT

learning was at its origin an attempt to understand how the brain can form representations of visual objects which are not confused by such transformations, i.e. they are 'transform invariant'.

CT learning requires local Hebbian learning (Hebb 1949) and lateral competition within layers of neurons. The mechanism utilised by the CT learning rule is the following (see Fig. 2.7): the first transform of a stimulus will activate a random cell in the output layer that wins the competition. The connections between the input layer cells representing the transform and the active output cell will become strengthened. If another transform of the same stimulus is sufficiently similar to the first one, it will share a large number of the input layer cells with the original transform. Due to the previously strengthened connections, the activation of these shared input cells is likely to result in the same output cell winning the competition. The concurrent firing of the cells activated by the second stimulus transform and the original output cell strengthens the connections between them, thus teaching the output cell to respond to both transforms of the same stimulus. This process is repeated for a large number of overlapping stimulus transforms in order to achieve transform invariance learning. CT learning, therefore, can be seen as a mechanism for adding new cells to the set of input cells that are known to represent a particular stimulus.

In the spiking neural network model of visual object recognition developed by Evans & Stringer (2012), the individual stimulus transforms were static, whereby each transform of an object was represented by a continuous injection of a steady current into a particular set of neurons in the input layer of the model. Therefore, there was no need to incorporate distributions of heterogeneous axonal delays in order for the network to learn transform invariance. These authors found three conditions were necessary in producing transform invariant object representations. These included having an appropriate level of lateral inhibition within the pre-

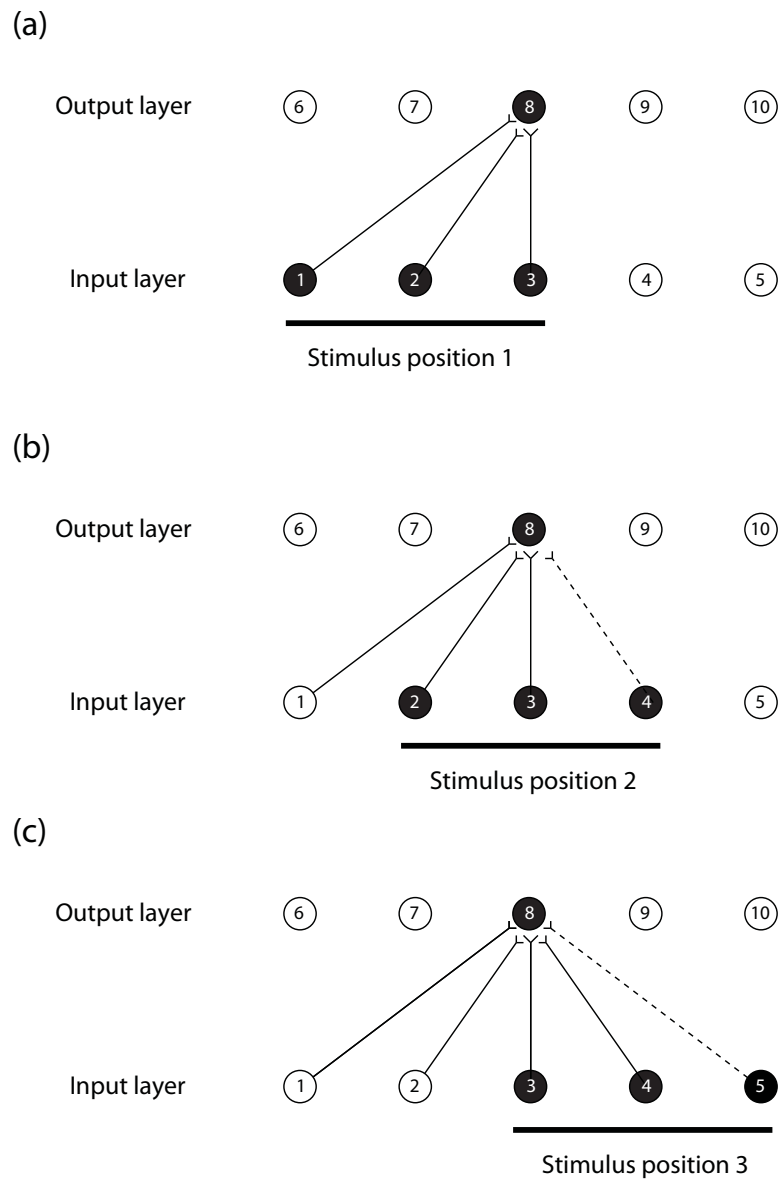


Figure 2.7: Schematic representation of how continuous transformation learning functions in a feed-forward one-layer network. Activation of overlapping neurons during the transformation of the object from position to position leads to the activation of the same neuron in the output layer. After each stimulus presentation afferent connections are strengthened according to a Hebbian learning rule.

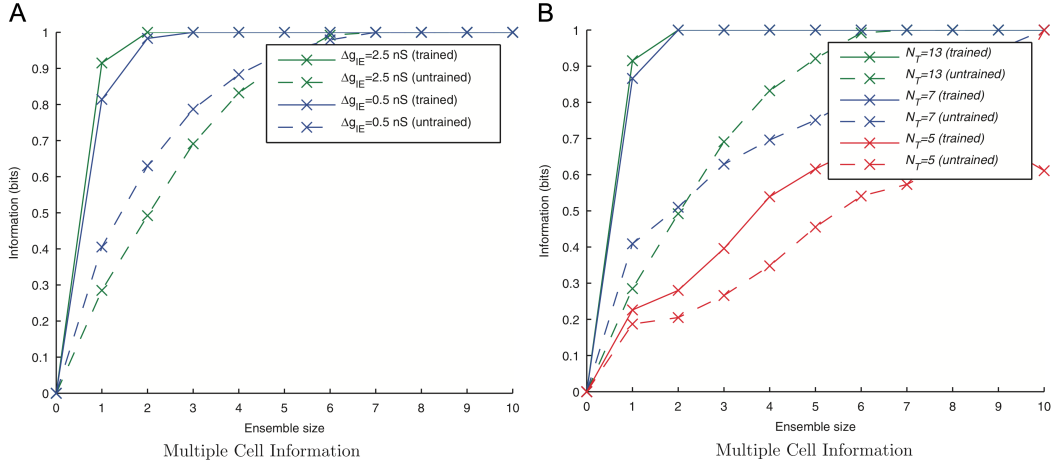


Figure 2.8: Performance of continuous transformation learning in a spiking neural network (Evans & Stringer 2012) measured with multiple cell information analysis, whereby populations of output layer cells are analysed to check whether there is enough information present within their firing to differentiate between two simple stimuli. Each stimulus was represented by a steady continuous injection of current into a number of neighbouring input layer neurons, whereby the two stimuli were represented by completely non-overlapping populations of input layer neurons. Different neighbouring transforms of a stimulus were represented by the activation of highly overlapping sets of input neurons, however the representations of the first and the last transforms were completely disjoint. If the maximum information of 1 bit is reached, then learning is considered to have been successful, whereby the model has developed populations of output cells that respond to all transforms of one stimulus and to no transforms of the other. **A:** as the level of lateral inhibition, mediated by the strength of synaptic conductance Δg_{ie} , is decreased and input layer cells desynchronise, learning is reduced. **B:** as stimulus overlap is decreased (as the number of transforms N_T decreases), learning also decreases. Adapted from Evans & Stringer (2012).

synaptic layer, having significant overlap between different stimulus transforms, and having homogenous axonal delays between layers (see Fig. 2.8).

An appropriate level of lateral inhibition was necessary to synchronise the firing of the pre-synaptic layer cells in response to a pattern to ensure that all important pattern driven spikes would have arrived at the post-synaptic cell during the time window conducive of significant LTP. Without synchronisation some important spikes will arrive in the LTP time window, while others may have arrived during

the LTD time window, thus resulting in no significant strengthening of synapses over a large number of stimulus presentations.

The second requirement was that there needed to be significant overlap between different transforms of the stimulus in order for CT to work. Evans & Stringer (2012) showed that the reduction in overlap between stimulus transforms from 44/56 (78.6%) neurons to 32/56 (57.1%) or 20/56 (35.7%) resulted in significantly diminished invariance learning, with no single cell firing to all five transforms of one stimulus and to no transforms of the other stimulus in the last condition. It was also shown that the order in which the different transforms are presented to the network does not matter, whereby the highly overlapping transforms of the stimulus do not need to be presented to the network in a particular order. As long as over the whole set of stimulus transforms the overlap is present, the network would be able to learn transform invariant representations of the object. Therefore, a training paradigm where transforms of one stimulus randomly sampled from the set of highly overlapping consequent transforms of that stimulus are interleaved with the similarly randomly sampled transforms of a different stimulus would be successful in enabling the network to learn transform invariant representations of the two stimuli.

The last requirement was that only homogenous axonal delays are present between the pre- and post-synaptic neurons. Having a distribution of heterogeneous axonal delays resulted in the destruction of CT learning in the vision model by Evans (2012).

2.5.2 Continuous Transformation Learning in Auditory Object Recognition

At first glance it may seem that there is no obvious analogue of 'affine transformations' of visual objects in the auditory world. For many classes of natural auditory

stimuli, however, their location in 'frequency space' depends on the physical characteristics of the sound source. For example, the changes in physical dimensions of the resonators of the vocal tract would create 'affine transformations' of vocalisation sounds. Such changes would happen due to variations in the placement of the tongue or the jaw when the same or different speakers produce the same speech sound. Thus, many natural auditory objects are prone to shifts in frequency space that are not too unlike the shifts in retinotopic space observed when visual objects undergo affine transformations. We, therefore, propose that CT learning may play a crucial role in auditory category learning.

The original CT learning mechanism relies on the presence of a significant overlap between input representations of temporally static stimulus transforms; in other words, neural representations of 'snapshots' of the same object taken from somewhat different points of view often exhibit areas of high correlation which can be discovered and exploited by an associative learning mechanism (Evans & Stringer 2012, Stringer et al. 2006). Unlike snapshots of visual objects, auditory stimuli have an essential temporal structure. In order for CT learning to associate similar temporal pre-synaptic patterns of firing onto the same output neuron by STDP, it is important that the volley of spikes from the pre-synaptic neurons arrive at the post-synaptic neuron almost simultaneously (Evans & Stringer 2012). If this is not the case, connections corresponding to the pre-synaptic spikes that arrive after the post-synaptic neuron fires will get weakened due to the nature of STDP, thus preventing effective CT learning of the input patterns.

In order to allow CT learning to work for the temporal auditory stimuli, therefore, a distribution of heterogeneous axonal conduction delays needs to be added to the plastic afferent connections. These axonal delays would transform temporal input sequences into patterns of spikes arriving simultaneously at individual post-synaptic cells. The patterns of coincident spikes received by each post-synaptic

cell would depend on the cell's transform matrix of axonal delays. If an appropriate delay transformation matrix is applied to the input spike pattern, a subset of post-synaptic neurons will receive synchronized spikes from the subset of input neurons encoding similar exemplars of a particular stimulus class, such as a vowel, thus enabling CT learning. Neurophysiological data collected from different species suggests that cortical axonal connections, including those within the auditory brain, may have conduction delays associated with them of the order of milliseconds to tens of milliseconds (Miller 1996, Salami et al. 2003).

It is important that the volley of spikes from the pre-synaptic neurons arrive at the post-synaptic neuron almost simultaneously (Evans & Stringer 2012). This is because learning by STDP strengthens a synaptic connection if and only if the spike from the pre-synaptic neuron arrives a few milliseconds before the spike emitted by the post-synaptic neuron. On the other hand, STDP weakens a synaptic connection if and only if the spike from the pre-synaptic neuron arrives a few milliseconds after the spike emitted by the post-synaptic neuron. So, to associate a subset of pre-synaptic neurons onto the same output neuron by STDP, the spikes from the pre-synaptic neurons need to arrive at the post-synaptic neuron at approximately the same time. Otherwise, while some connections will be strengthened, other connections will be weakened. Consider a particular subset of pre-synaptic neurons in the preceding layer with a regularly repeating temporal pattern of spikes representing an auditory object such as a vowel. In a large network with many post-synaptic neurons in the next layer, it is likely that by chance some of these post-synaptic neurons will have an appropriate set of axonal delays on their input connections to allow them to receive coincident spikes from the aforementioned subset of pre-synaptic neurons in the preceding layer. These post-synaptic neurons can then learn to respond to the temporal firing pattern carried by the pre-synaptic neurons. CT learning can then enable post-synaptic neurons

to bind together similar temporal firing patterns carried by the pre-synaptic neurons representing other transforms of the same vowel, as long as the spikes from each firing pattern arrive nearly simultaneously at the post-synaptic neuron after passing through the axonal delays.

CT with axonal delays would work as follows. An AN firing pattern in response to the first transform of an auditory stimulus will result in a number of different sets of coincident spikes arriving at and activating different post-synaptic cells after being passed through the randomly initialised axonal conduction delays (Fig. 2.5). Afferent connections between the coincident input pattern cells and their corresponding post-synaptic cells will become strengthened. When a different transform of the same auditory object is presented, it will result in a similar AN firing pattern, however some of the spikes will be different. The similar parts of the AN firing pattern will turn into the same coincident spikes arriving at the post-synaptic neurons as those for the previous stimulus transform. Parts of the AN firing raster that differ from the first stimulus transform may or may not result in spikes coincident with the unchanged part of the AN firing raster when they arrive at the post-synaptic neuron. This depends on the particular distribution of delays for each post-synaptic cell. The new spikes need to arrive coincidently with the old spikes in order for the CT mechanism to learn that these new spikes are part of the same stimulus representation. Therefore, the more output cells there are, the more likely it is that enough AN firing raster spikes arrive coincidently with the previously learnt spikes to a post-synaptic cell for all stimulus transforms, thus allowing CT learning to work.

Previous research in modelling visual object recognition suggested that the addition of heterogeneous axonal conduction delays breaks CT learning when the input neurons are synchronised and the STDP time constants are relatively short (Evans & Stringer 2012). In this case, delay transformation matrices can scramble

the synchronisation of spikes reaching the post-synaptic cells. However, in the case when input neurons are not synchronised, such as in response to auditory stimuli, the addition of heterogeneous axonal delays should enable CT learning. If a correct delay transformation matrix is applied to the input spike pattern, a subset of post-synaptic neurons will receive synchronised spikes from the subset of input neurons encoding all transforms of a particular stimulus, such as a vowel, thus enabling CT learning.

It is, therefore, suggested that the CT mechanism can enable a spiking neural network to learn class identities of temporal auditory stimuli if over the whole space of different stimulus exemplars belonging to one class, stimuli that are similar to each other physically also evoke similar spatio-temporal firing patterns (i.e. have 'sufficient overlap'). Stochasticity present in the AN adds noise to the spatio-temporal firing patterns, and therefore make responses to similar stimuli more dissimilar, hence preventing effective CT learning.

This thesis starts by using simple vowel stimuli, which are static in terms of their constituent fundamental and formant frequencies. These simple vowels, however, still have a temporal dimension, since they are complex periodic sounds with a regularly repeating temporal wave pattern. Such stimuli were chosen as one of the simplest examples of speech-like auditory inputs that can be used to investigate how CT learning might work to learn temporal auditory stimuli.

2.5.3 Stimulus Presentation Paradigm Appropriate for Continuous Transformation Learning

Vowel presentation order does not affect learning in the model, since the CT learning mechanism does not rely on any particular order of presentation of the transforms of the different object classes. As long as there is high pairwise similarity for the consequent members of the full set of transforms for each object class, the

exact order in which each individual transform is presented does not matter. For example, if the stimulus set consists of two object classes, A and B , with five transforms each, $\{a_1, a_2, a_3, a_4, a_5\}$ and $\{b_1, b_2, b_3, b_4, b_5\}$; and each pairwise combination of transforms for each class are highly overlapping (ie a_i and a_{i+1} ; b_i and b_{i+1} are highly overlapping for all $i \in \{1, \dots, 4\}$), CT learning will operate when presented with any permutation of the set $C = A \cup B$.

The learning in the model, however, can be affected by the presence of long gaps between the presentations of the different exemplars of one vowel due to the stronger LTD vs LTP magnitudes as described by Masquelier et al. (2008). In such a scenario connections strengthened in response to a repeating pattern within one stimulus class would get weakened during the long gap of being presented with the exemplars of the other stimulus class, during which these connections may transmit noise.

In order to compare the performance of the different models presented in this thesis, we have decided to remove the confounding variable of stimulus presentation order. Hence, the transforms of the different stimulus classes are presented in an ordered interleaved fashion (see Section 2.2.1) during training and testing for all simulations described in this thesis.

2.6 Auditory Nerve Model

In order for the auditory brain to process sounds, it needs to transform the varying air pressure levels of the auditory wave into a spike raster. The transformation is done within the cochlea, which depolarises auditory nerve fibers through afferent synapses from inner hair cells. A number of computational models of this process have been put forward, replicating the transformation of the varying air pressure levels into the firing of auditory nerve (AN) fibers in the brain (Bruce et al. 2003, Carney 1993, Delgutte et al. 1998, Deng & Geisler 1987, Sumner et al.

2003, Zilany et al. 2009). Such models have been used as inputs for traditional speech recognition systems based on Hidden Markov Models, as well as to investigate speech coding in the peripheral auditory system without the need for animal experimentation (Sumner et al. 2003).

For the purpose of simulations described in this chapter the model by Zilany et al. (2009) was chosen to transform the instantaneous pressure waveforms of the digital audio recordings of the stimuli into AN spike rasters. These rasters were then used as the input to the rest of the spiking neural network model of the auditory brain. The AN model by Zilany et al. (2009) was chosen because it had been shown to accurately replicate single-unit temporal responses of AN fibers to a large number of stimuli, such as pure tones, broadband noise, amplitude-modulated sounds, forward-masking paradigm stimuli and speech sounds. Furthermore, this model is largely accepted within the field of auditory neuroscience, having been extensively cited since its initial publication (Henry et al. 2011, Jennings et al. 2011). The main advantage of the Zilany et al. (2009) model is that it accounts for most of the non-linearities present in the AN responses, including those in the frequency domain (such as two-tone suppression) and time domain (such as adaptation). A schematic representation of the model is shown in Figure 2.9. The AN model has two sources of stochasticity: 1) the spike generator is implemented as a heterogeneous Poisson process; 2) the distribution of spontaneous responses of AN fibers is modelled by inserting Gaussian noise within the slow power-law adaptation path within the model (Zilany et al. 2009).

2.6.1 Generalising from Cat Neurophysiological Data

Zilany et al. (2009) fitted the response properties of their AN model to neurophysiological data obtained from cats, which raises the question of how generalisable such a model is to humans. Research suggests that it is, since cats and humans

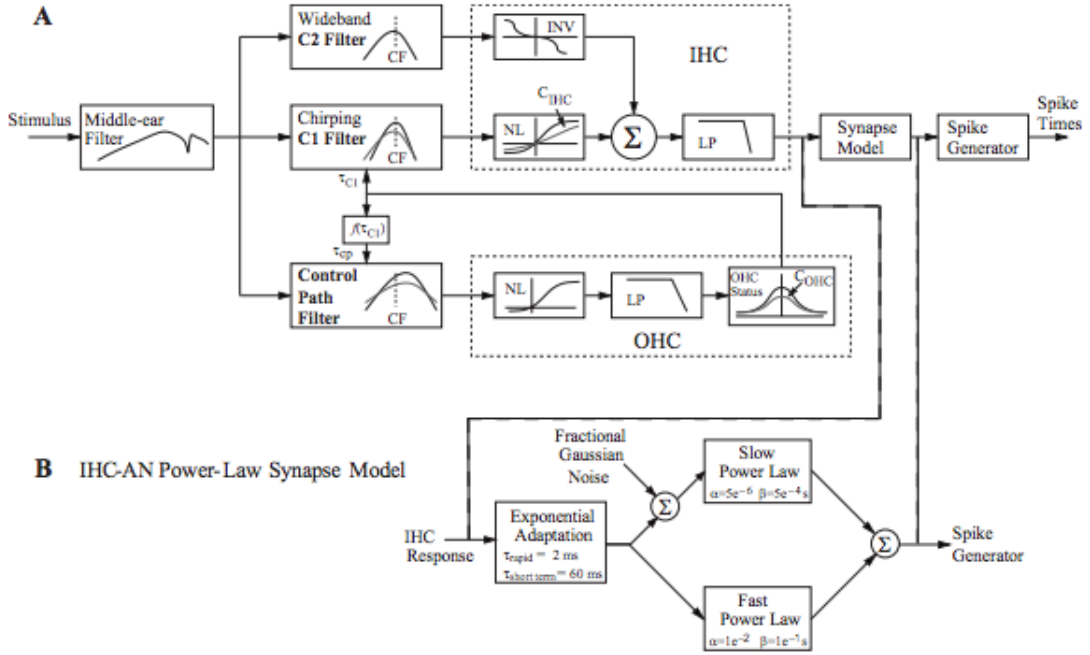


Figure 2.9: Schematic representation of the auditory nerve model by Zilany et al. (2009). The input to the model is the instantaneous pressure waveform of a sound and the output is the firing times of each modelled auditory nerve fiber. The model consists of a middle-ear filter, followed by a feed-forward control path, a signal-path (C1) filter and a parallel-path (C2) filter, an outer-hair cell (OHC) and an inner-hair cell (IHC) sections, converging on the synapse model and the spike generator. The synapse model is expanded in section B and includes power-law adaptation models that result in better recreation of many non-linearities present in the auditory nerve. Fractional Gaussian noise is added to recreate the desired spontaneous activity profiles. This Gaussian noise and the Poissonian nature of the spike generator result in stochastic output spike trains generated by the model. Adapted from Zilany et al. (2009).

have been found to share many similarities in their auditory systems, including the anatomy of the cochlea and the AN (Schuknecht 1974); and the psychophysical tuning curves from masking and two-tone suppression experiments (Bauer 1978, Houtgast 1971, Zwicker 1974). While the two species have different auditory ranges, the differences are not detrimental to the generalisation of the model, since the human auditory range (0.06-23 kHz) is a subset of the cat auditory range (45-

64,000 Hz) (Delgutte & Kiang 1984). Furthermore, research has shown that cats are able to discriminate between vowels, such as /i/ and /u/ (Dewson 1964), which suggests that the AN model by Zilany et al. (2009) is appropriate for pre-processing the stimuli used in this thesis.

2.7 Auditory Nerve Input Layer

The input layer of the spiking neural network model of the auditory brain described consists of 1,000 AN fibers with preferred frequencies logarithmically spaced between 100-3,500 Hz. A neuron's preferred frequency is defined as the frequency of a pure tone that discharges the cell above 10% of its spontaneous firing rate at the lowest intensity. The particular range of preferred frequencies within the AN layer was chosen because the simulations described later in this thesis investigate auditory speech processing, and speech sounds take up only a small portion of the audible frequency range in humans, with all speech sounds lying between 100-4,000 Hz (Yost 2005), and psychophysical studies showing that speech remains intelligible even after being put through a bandpass filter of 500-1,500 Hz (Denes & Pinson 1993). The 100-3,500 Hz range of the model AN fibers was specifically chosen to include the F_0 of the stimuli (100 Hz), and to cover the third formant of the vowels, which goes up to 3,060 Hz. The fiber preferred frequencies were logarithmically spaced to account for the log distribution of preferred frequencies in the real AN (Gelfand 1998). One of the parameters of the AN model is the threshold of AN fibers. All AN fiber thresholds were set to 60 dB SPL, to ensure that the stimuli presented at the conversational speech level of 60-65 dB SPL did not saturate the firing of AN fibers. Each fiber was set to have medium spontaneous firing rate. The AN was set up to operate at 0.1 ms time step resolution.

The size of the AN layer was set to 1,000 fibers to ensure that the cell preferred frequencies are closely spaced with sufficient overlap in the fibers' tuning curves.

Such overlap exists in the real AN, whereby approximately 30,000 fibers cover the frequency range of 64-23,000 Hz (Pickles 1988). Additionally, the response properties of chopper and primary-like with notch cells in the cochlear nucleus depend on receiving afferent connections from a number of AN fibers with similar preferred frequencies (Ferragamo et al. 1998).

2.8 Jitter in the Auditory Nerve

Since CT learning is essentially a label for the interaction between Hebbian STDP-based learning (Bi & Poo 1998) and input stimuli with strongly overlapping transforms, CT learning should operate even in the presence of some stochasticity in the spike rasters representing the input stimuli, as long as the level of stochasticity is not disruptive enough to break STDP-based learning. Past research suggested that the intrinsic properties of spiking neural networks and STDP-based learning can cope with noise on the scale of a few milliseconds (Bohte & Mozer 2004, Diesmann et al. 1999). However, AN spike rasters in response to the different versions of the same auditory object class change on the order of tens rather than single milliseconds, thus breaking STDP-based learning. This means that a certain level of pre-processing of the input AN spike rasters is required to enable STDP-based learning to develop invariant representations of auditory object classes. One of the main hypotheses put forward within this thesis, therefore, is the role of CN and IC in de-noising AN input and preparing it for STDP-based learning in the plastic cortical areas of the auditory brain.

The stochasticity within the AN spike rasters is described in terms of time and space jitter within this thesis. Such a framework has been chosen in order to make it easier to explain the role of particular brain areas in de-noising the AN spike raster inputs. The framework does not imply strict separation of AN stochasticity

into time and space jitter, neither does it imply the existence of the two types of jitter.

There are two reasons why AN spike rasters produced in response to the same auditory stimulus differ: AN stochasticity and the interplay of refractoriness and phase-locking. The inherent stochasticity present in the AN and modelled using fractional Gaussian noise and Poissonian spike generator in the model by Zilany et al. (2009) results in time jitter within AN spikes, whereby the precise timing of each information bearing spike changes from one stimulus presentation to another. The second reason for differences in AN firing is the phase-locking property of AN neurons, which is discussed below.

2.8.1 Phase-Locking

The phase-locking property of AN fibers postulates that the timing of AN spikes correlates with the peaks of the pure tone driving cell activity (Fig. 1.15 (a)). The fact that AN fibers fire in synchrony with the sound wave peaks means that a 100 Hz tone would be expected to produce on average 100 spikes/s in an AN neuron if the tone is presented at high enough sound level. The relationship between sound frequency, loudness and phase-locking is summarised in Figure 2.10.

Since AN cells are unable to fire more than 200 spikes/s on average (Eggermont 2001, Sachs & Abbas 1974, 1976), they cannot fire in response to every cycle of tones with frequencies above 200 Hz. Such 'place' information encoding, therefore, fails to represent tones with frequencies above 200 Hz. Nevertheless, it was shown that AN firing can carry information about sound frequencies up to 4-5 kHz (Brugge et al. 1969, Covey 2000). This discrepancy is solved using the Volley Principle suggested by Wever & Bray (1930), whereby each neuron in a population phase-locks to a different subset of amplitude peaks of the sound wave. Therefore, although there is not enough information about the input sound frequency in

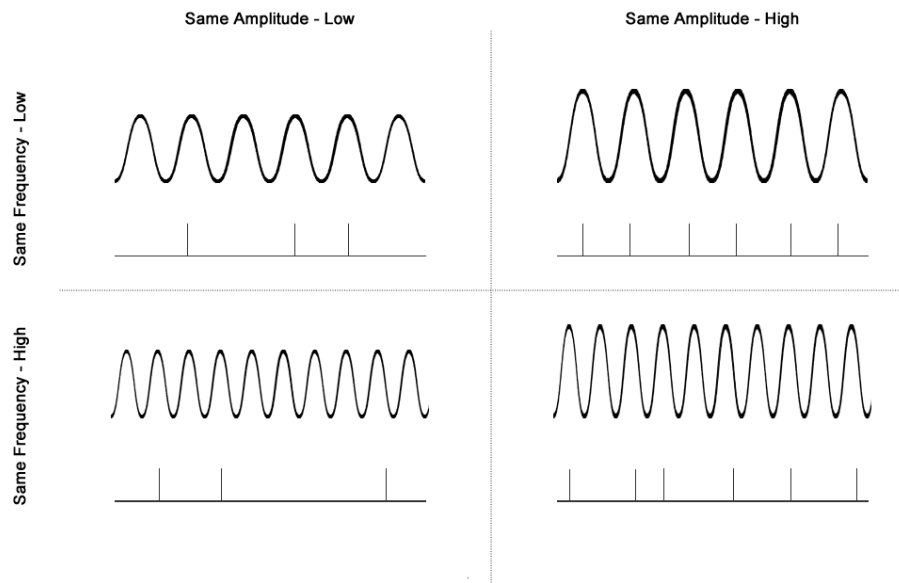


Figure 2.10: Schematic representation of phase-locking to pure tones of varying frequency and amplitude of an individual neuron in the auditory nerve. While the rate of firing changes with sound intensity (amplitude), the precise timing of spikes remains phase-locked to the sinusoid peaks.

inter-spike intervals of each individual AN fiber, this information is found when the whole population of AN fibers responding to the particular frequency is considered (Fig. 1.15 (b)).

The interplay of stochasticity, refractoriness and phase-locking produces time and space jitter in the AN. Time jitter means that the precise timing of spikes in the AN can vary when presented with the same auditory stimulus. Although the spikes are phase-locked to the peaks of the sound wave, unless the sound is presented at saturating intensity, and the frequency of the sound is less or equal to the maximum frequency of firing for the neuron, AN fibers will fire to a random subset of sound wave peaks. Furthermore, phase-locking creates space jitter, whereby spikes phase-locking to a particular pure tone peak may belong to a random subset of AN cells with that characteristic frequency, as postulated by

the Volley Principle (Wever & Bray 1930).

This means that an information bearing spike can move in two dimensions between different presentations of the same stimulus. It can either occur for the same AN cell as during the original stimulus presentation, but at a different point in time (time jitter); or it may appear at the same time as during the original stimulus presentation, but for a different cell (space jitter) (Fig. 2.11). There can also be a combination of both space and time jitter. This makes AN spike rasters very variable, and it is argued in the rest of this thesis that this variability makes it extremely hard for the Hebbian learning mechanisms within the auditory brain (such as STDP-based learning) to learn invariant representations of auditory objects. This task, however, becomes much simpler after de-noising pre-processing performed within subcortical areas CN and IC.

2.9 Model of Spiking Neurons

Physiological properties of neurons were modelled according to the spiking neuron model suggested by Izhikevich (2003). This was chosen over the Hodgkin-Huxley or integrate-and-fire models because it combines the best features of both model classes, having the biological accuracy of Hodgkin-Huxley-type dynamics and the computational efficiency of integrate-and-fire neurons. The spiking behaviour of Izhikevich's neurons is governed by the following equations:

$$\frac{dv_i}{dt} = 0.04(v_i(t))^2 + 5v_i(t) + 140 - u_i(t) + I_i(t) \quad (2.3)$$

$$\frac{du_i}{dt} = a(b \cdot v_i(t) - u_i(t)) \quad (2.4)$$

where $v_i(t)$ is the membrane potential of the post-synaptic neuron i measured on a mV scale, $u_i(t)$ is the membrane recovery variable of the post-synaptic neuron i ,

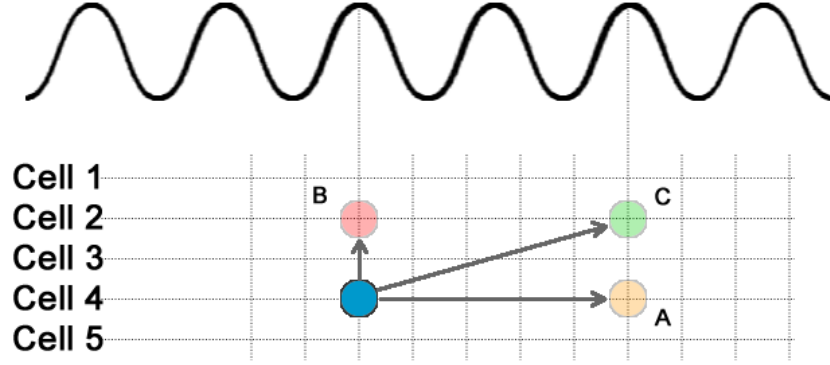


Figure 2.11: Illustration of space and time jitter in the auditory nerve. Blue circle represents the original spike phase-locked to a sound wave peak. **A-C** show possible locations of the original blue spike during a subsequent presentation of the same sound wave. **A** - time jitter. The same neuron spikes, however the spike phase-locks to a different sound wave peak. **B** - space jitter. A different cell spikes to the original sound wave peak in accordance to the volley principle (Wever & Bray 1930). **C** - time and space jitter. The spike both phase-locks to a different sound wave peak and is produced by a different cell.

which accounts for the activation of K^+ and closure of Na^+ ionic currents, and t is time measured on a ms scale. $I_i(t)$ represents a combination of incoming externally injected ($I_i^{ext}(t) = 0$ nA for the simulations described in this thesis) and synaptic currents for the post-synaptic neuron i according to the following formula:

$$I_i(t) = \sum_j \sum_l w_{ij} \delta(t - \Delta t_{ij} - t_j^l) + I_i^{ext}(t)$$

where w_{ij} is the magnitude of the synaptic connection strength between the pre-synaptic neuron j and the post-synaptic neuron i , and $\delta(t - \Delta t_{ij} - t_j^l)$ is the Dirac delta function evaluated at time t minus the magnitude of the axonal conduction delay Δt_{ij} between the pre- and with the post-synaptic cells j and i

and minus the times of the pre-synaptic spike train t_j^l indexed by l . Dirac delta function is defined as the following:

$$\delta(x) = \begin{cases} \infty & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{where, } \int_{-\infty}^{+\infty} \delta(x) dx = 1$$

Equations 2.3-2.4 aim to approximate the nullclines of conductance-based models, such as the Hodgkin-Huxley model, in the local region where the decision is made whether the neuron is to fire or not. This is done by simplifying the fast V -nullcline to the locally quadratic function, and the slow u -nullcline to the locally linear function. The constant factors in equations 2.3-2.4 were arbitrarily chosen to fit the spike initiation dynamics of a real cortical neuron while achieving the required mV/ms time scale. The model threshold potential, the cell potential at which a spike is emitted, lies between -55 mV and -40 mV. The exact value of the threshold potential is either fixed or varies depending on prior membrane potential depending on the type of bifurcation of the neuron model. In general cells that act as integrators have a well defined threshold manifold, whereas cells that act as resonators do not. The precise threshold dynamics for different types of neurons are beyond the scope of this review but are described in Izhikevich (2000) and Izhikevich (2007). The after-spike resetting is governed by the following equations:

$$\text{if } v_j(t) \geq 30 \text{ mV, then } \begin{cases} v_j(t) \leftarrow c \\ u_j(t) \leftarrow u_j(t) + d \end{cases} \quad (2.5)$$

where c accounts for the after-spike reset of cell membrane potential caused by the fast high-threshold K^+ conductances, and d accounts for the after-spike reset of the membrane recovery variable caused by the slow high-threshold K^+ and Na^+ conductances. Parameters $a - d$ are dimensionless and can be adjusted to achieve the required spiking behaviour of neurons. Figure 2.12 depicts the spiking behaviours that can be achieved by Izhikevich's neurons.

Parameter	Excitatory Cortical (Spike Adaptation - F)	Excitatory Subcortical (Class 1 - G)	Inhibitory (Phasic bursting - D)
a	0.01	0.02	0.02
b	0.2	-0.1	0.25
c	-65	-55	-55
d	8	6	0.05
Threshold (mV)	30	30	30

Table 2.2: Values of the $a - d$ constants used in equations 2.3-2.5 for the excitatory and inhibitory neurons used in the model of the auditory brain. *Neuron Type* refers to model neurons reported by Izhikevich (2003) and seen in Fig. 2.12. **a**: time scale of $u(t)$. Smaller values mean slower recovery. **b**: sensitivity of $u(t)$ to subthreshold fluctuations in $v(t)$. Larger values result in subthreshold oscillations and low-threshold spiking. **c**: after-spike reset value of $v(t)$. **d**: after-spike reset value of $u(t)$.

The differential equations described above were replaced by finite difference equations and solved using a Forward-Euler numerical scheme with a 0.1 ms time step. This was chosen to match the value used by Izhikevich (2003). Furthermore, it has been shown experimentally that varying the time step resulted in no significant changes in network performance. The Dirac delta function was replaced by an indicator function $1(t - \Delta t_{ij} - t_j^l)$ in the finite difference equations returning 1 if when $t - \Delta t_{ij} - t_j^l = 0$.

2.9.1 Excitatory Cells

Table 2.2 summarises the parameters chosen for the excitatory neurons within the auditory brain models described in chapters 3-4. All subcortical excitatory cells were set to be Class 1 (Fig. 2.12 (G)), while cortical excitatory cells were set to have Spike Frequency Adaptation (Fig. 2.12 (F)) as defined by Izhikevich (2003).

Neurophysiological evidence suggests that many neurons in the subcortical auditory brain have high spiking thresholds and short temporal integration windows, thus acting more like coincidence detectors than rate integrators (Frisina 2001). This is similar to the behaviour of the Izhikevich's Class 1 neurons, which can be considered to be coincidence detectors, since they return back to resting state

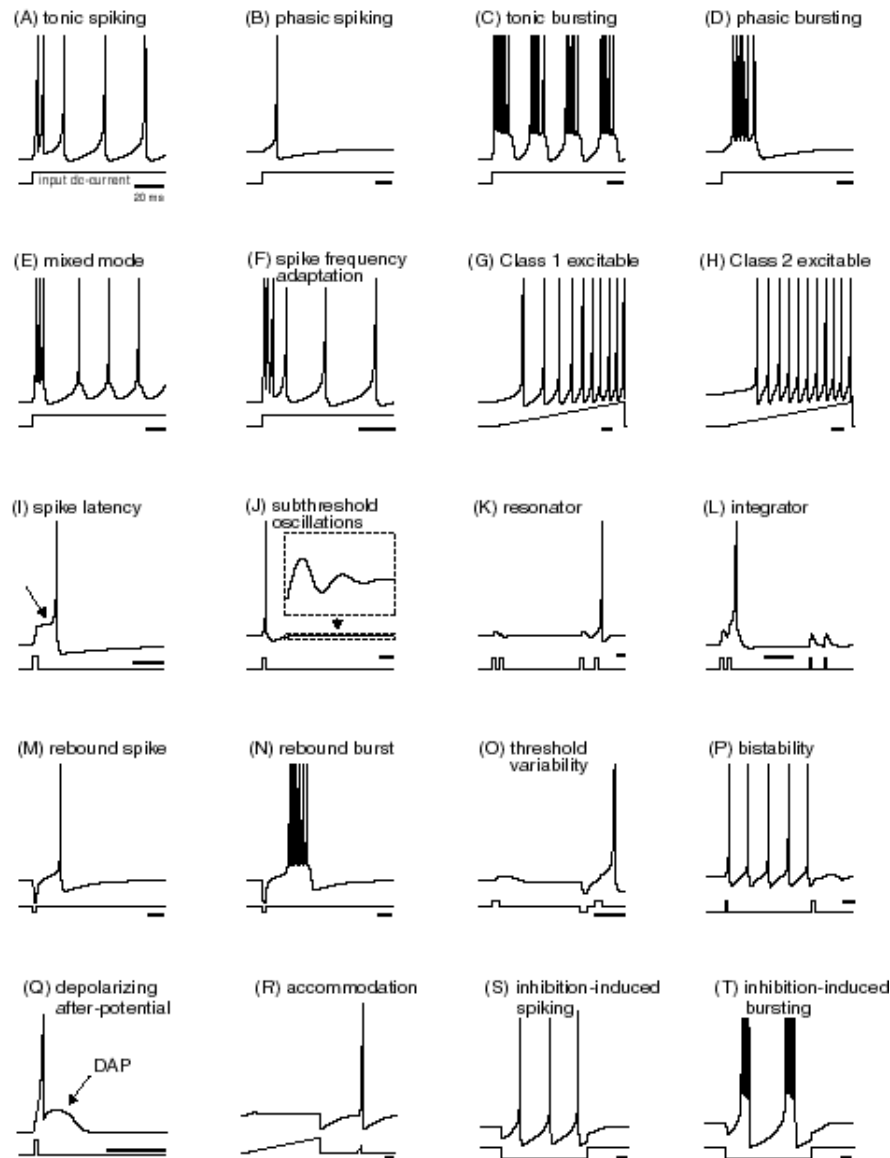


Figure 2.12: Schematic representation of twenty different response types of the spiking neural model by Izhikevich (2003). Each response type corresponds to a different set of values chosen for the $a - d$ constants used in equations 2.3-2.5. Adapted from Izhikevich (2003).

quickly and require a large concurrent input in order to fire.

Izhikevich's Spike Frequency Adaptation neurons were chosen to model the excitatory cells in the auditory cortex, since neurons in the auditory cortex show adaptation to a continuous stimulus (Ulanovsky et al. 2004). Furthermore, preliminary simulation results showed that the adaptation implemented within the Izhikevich's Spike Frequency Adaptation neurons is important in aiding inhibition-induced within-layer competition.

2.9.2 Inhibitory Cells

Each layer of the auditory brain model (apart from the input AN layer) contains both excitatory and inhibitory neurons. Inhibitory cells are implemented as Izhikevich's Phasic Bursting neurons (Fig. 2.12 (D)). The choice of neuron type for inhibitory neurons was somewhat arbitrary, with preliminary simulations showing that other neuron types can work similarly to the Phasic Bursting type. The main two criteria for choosing the within-layer inhibition set up was that it achieved a dual purpose - competition and within layer firing rate regulation.

For the simulations described in this thesis, competition within different layers of the neural network model of the auditory brain is implemented in the following way. Sparse connectivity between excitatory and inhibitory cells within a layer is implemented using one-to-one connections to limit the computational cost of simulations. Each excitatory cell within a layer is connected to a corresponding inhibitory cell through a one-to-one connection with a strong synapse to ensure that an inhibitory cell always fires when its corresponding excitatory cell fires. Each inhibitory cell is fully connected to all excitatory cells within the same layer, thus ensuring global suppression of firing within the layer once one excitatory cell within the layer fires. Therefore, excitatory cells with stronger afferent connections quickly activate their corresponding inhibitory cells, which in turn suppress the

firing of other excitatory cells within the same layer that have weaker afferent connections, thus enabling within-layer competition.

The other role of inhibitory cells is to regulate the total level of firing within a layer, ensuring that it does not explode. The more excitatory cells fire within a layer, the more inhibitory cells get activated. This means that within-layer inhibition level increases with increased excitatory activity, eventually preventing further increase in excitatory activity.

The bursting behaviour of inhibitory cells is necessary to ensure sufficiently long inhibitory activity to affect excitatory cells. Strong rate adaptation is necessary to ensure that the two inhibitory effects described above are stronger immediately after an excitatory cell fires and weaken with time.

Informal tests demonstrated that the exact implementation of within-layer inhibition does not have a significant impact on the results presented in this thesis, as long as the implementation still achieves an appropriate level of within-layer competition and activity modulation. The behaviour of the proposed inhibitory circuit is in line with the dynamic and tightly balanced inhibition as described in Deneve & Machens (2016).

2.10 Quantitative and Plotting Measures

2.10.1 Neurogram

Neurograms of firing within a population of neurons are plotted by first creating period histograms (PH) (Shimazaki & Shinomoto 2007) for each neuron, whereby firing rasters are split into period slices of length corresponding to the stimulus fundamental frequency. These slices are overlaid on top of each other and the number of spikes within each corresponding 0.1 ms bin contained is calculated. Stimulus period histograms are used to compute each neuron's 'synchronized rate' (Sinex &

Geisler 1983), defined at each frequency as the magnitude of the PH's FFT component at that frequency, whereby higher magnitudes signify better phase-locking to that particular frequency. A neuron is considered to be well phase-locked to a frequency if the timing of its action potentials is consistent with the phase of the sound at that frequency.

2.10.2 Synchronisation Index

Synchronisation index is used to quantify how synchronised the firing of neurons within a network layer is. A higher synchronisation index indicates the presence of neural oscillations, whereby there is a high correlation between spike trains of a large number of neurons within a layer.

The single cell synchronisation index reported by Winter et al. (2003) was adapted to calculate the population synchronisation index within the ON CN layer. First the firing rate histogram of ON cells in response to a chosen stimulus was calculated, resulting in matrix X of size $M \times N$, where $M = 100$ is the number of 1 ms bins making up the stimulus, and N is the number of ON cells. The matrix was then summed row-wise, resulting in vector $x \in R^M$, which represented the rate of firing within the ON layer in response to the stimulus discretised within 1 ms bins. Fast Fourier Transform (FFT) was applied to the resulting data vector x , and the normalised magnitude of the FFT component corresponding to the F_0 of the vowel stimuli was used as the synchronization index. The mean of vector x was used as the normalisation factor.

2.10.3 Quantifying Spike Raster Similarity

In this thesis we hypothesise that a spiking neural network can learn auditory categories through the continuous transformation (CT) learning mechanism. CT

learning relies on a high degree of similarity/overlap between spike rasters in response to different exemplars of one stimulus class. Here we describe three indices which we have devised to quantify the degree of similarity/dissimilarity between spike rasters in response to different presentations of the same stimulus (Same Exemplar Index), different exemplars of the same stimulus class (Different Exemplars Index) or different stimulus classes (Different Category Index). The network responds differently to different presentations of the same stimulus, due to the stochasticity present in the AN. Each index varies between 0 and 1, with higher scores indicating a higher degree of similarity between the corresponding firing rasters. Lower scores suggest that the firing rasters being compared are dissimilar due to either the inherent differences between the input stimuli, or due to the presence of jitter that diminishes the otherwise high similarity between the firing rasters being compared.

2.10.3.1 Same Exemplar Index

The Same Exemplar (SE) index quantifies the degree of similarity between the firing rasters within a particular area (such as AN or IC) in response to different presentations of the same exemplar of a stimulus. Broadly, it calculates the average number of identical spikes across the different presentations of each exemplar $e_{k(s)} \in \{e_{1(s)}, \dots, e_{12(s)}\}$ of a stimulus $s \in \{s_1, s_2\}$ in proportion to the total number of stimulus exemplar presentations ($n \in N$, where $N = 20$ testing epochs). For each presentation of each stimulus we, therefore, constructed a $T \times J$ matrix $M_{e_{k(s)}}^n$ (where $T = 100$ ms is the number of 1 ms time bins spanned by the auditory input, and $J \in \{100, 1000\}$ neurons is the size of the chosen area of the model). Each element m_{tj} of matrix $M_{e_{k(s)}}^n$ contained the number of spikes produced by the particular neuron $j \in J$ within the time bin $t \in T$ in response to the stimulus exemplar $e_{k(s)}$.

If the firing rasters of the chosen area of the model in response to the different presentations $n \in N$ of the same stimulus exemplar $e_{k(s)}$ are similar to each other, then the same slots of the firing pattern matrices $M_{e_{k(s)}}^n$ should be non-zero for different $n \in N$. Consequently, the following becomes more likely when the proportion of stimulus presentation epochs n for which elements of $M_{e_{k(s)}}^n$ are non-zero across the different presentations of the same stimulus exemplar becomes large: 1) the firing responses within the model area are more likely to be similar; 2) it is likely that less jitter is present in the chosen area of the model; 3) CT learning is more likely to enable postsynaptic cells to learn that the similar, stable, jitter-less responses within the model area belong to the same stimulus class.

We, therefore, computed the matrix $Me_{k(s)} = \langle M_{e_{k(s)}}^n \rangle$, where $\langle \cdot \rangle$ signifies the mean over all the presentation epochs $n \in N$, and then identified the mean $\mu_{e_{k(s)}}$ of the hundred largest elements of $Me_{k(s)}$. These were used to compute the final SE_s score for each stimulus $s \in \{s_1, s_2\}$ as $SE_s = \langle \mu_{e_{k(s)}} \rangle$, where $\langle \cdot \rangle$ signifies the mean over all exemplars $e_k(s)$ of stimulus s . Higher SE_s index points to more similarity between the chosen firing rasters in response to the different presentations of the same exemplar of stimulus s . Consequently, this also signifies lower levels of jitter present within the layer, since high levels of jitter would disrupt the similarity in firing patterns and result in lower SE_s index.

2.10.3.2 Different Exemplars Index

The Different Exemplars (DE) index quantifies the similarity of the firing rasters within a chosen area of the model in response to the different exemplars of the same stimulus class. It is somewhat similar to the SE_s index described above, however, instead of comparing the firing matrices across the different presentations n of the same stimulus exemplar $e_k(s)$, the firing matrices are compared across the different exemplars $e_k(s)$ of each stimulus class $s \in \{s_1, s_2\}$. Consequently, firing raster

matrices $M_{e_{k(s)}}^n$ were calculated once again, but this time the average was taken over all the different exemplars $e_{k(s)} \in \{e_{1(s)}, \dots, e_{12(s)}\}$ of stimulus s ($M_s^n = \langle M_{e_{k(s)}}^n \rangle$, where $\langle \cdot \rangle$ signifies the mean over all the stimulus exemplars). We then identified the mean μ_s^n of the hundred largest elements of M_s^n and used them to compute the final DE_s score for each stimulus $s \in \{s_1, s_2\}$ as $DE_s = \langle \mu_s^n \rangle$, where $\langle \cdot \rangle$ signifies the mean over all $n \in \{n_1, \dots, n_{20}\}$ presentation epochs of each exemplar $e_{k(s)}$ of stimulus s . Higher DE_s index points to more similarity between the firing rasters within the chosen model area in response to the different exemplars $e_{k(s)}$ of stimulus s . Consequently, this also signifies lower levels of jitter present within the layer, since high levels of jitter would disrupt the similarity in firing patterns and result in lower DE_s index.

2.10.3.3 Different Category Index

The Different Category (DC) index quantifies the similarity of the different firing rasters within a chosen area of the model in response to different stimulus classes. This score is somewhat similar to the SE_s and DE_s scores described above, however, here the rasters are compared across the different stimulus categories $s \in \{s_1, s_2\}$. To this accord, firing raster matrices $M_{e_{k(s)}}^n$ were calculated once again, but this time the average $M^n = \langle M_{e_{k(s)}}^n \rangle$ was taken over all the different exemplars $e_k(s) \in \{e_{1(s)}, \dots, e_{12(s)}\}$ and over all the stimuli $s \in \{s_1, s_2\}$. We then identified the mean μ^n of the hundred largest elements of each matrix M^n and used them to compute the final DC score as $DC = \langle \mu^n \rangle$, where $\langle \cdot \rangle$ signifies the mean over all $n \in \{n_1, \dots, n_{20}\}$ presentation epochs. Lower DC index points to more differences between the chosen firing rasters in response to the different stimulus categories s .

2.10.3.4 Polychronization index

The continuous transformation (CT) learning mechanism based on STDP described earlier in this chapter relies on stable temporal patterns of firing within the pre-synaptic layer. In order to understand whether such stable firing patterns, or polychronous groups (PGs) (Izhikevich 2006), emerge within the A1 layer of our models in response to naturally spoken words, such as digits 'one' and 'two', a Polychronization Index (PI) measure was devised. It is similar to the SE, DE and DC indices described above, however it can be applied to analysing the response properties of the network when presented with more complex naturally spoken word stimuli as used in Chapter 6. Unlike the SE, DE and DC measures, PI does not rely on the identical duration, fundamental frequency and rate of speech of the stimuli, which is true for the synthesised vowels /i:/ and /a/ used in Chapters 3-5, but false for the naturally spoken digits 'one' and 'two' used in Chapter 6. The 'polychronization index' (PI), therefore, quantifies the prevalence of reproducible patterns across the population of neurons in one layer.

The PI was calculated for each cell j in turn, by searching for spikes fired by other cells at fixed intervals relative to spikes fired by cell j . PGs which are responsive to a stimulus class should produce such stable and reproducible firing patterns repeatedly across the presentation of stimulus exemplars ($e_k(s) = [e_k(s), \dots, e_{376}(s)]$) of each class $s \in [s_1, s_2]$. For each presentation of stimulus $e_k(s)$ we randomly selected one spike by cell j and noted its time t post stimulus onset. We then constructed a 1000 neuron by 101 ms matrix $M_{e_k(s)}^j$ representing the firing of the other neurons within $t \pm 50$ ms at 1 ms resolution. The ± 50 ms time window reflects the maximum conduction delay in our model, which puts an upper limit on the temporal extent over which PGs can produce synchronized input to neurons in the next layer. If neuron j is part of a PG which produces reproducible

firing patterns, then this reproducible pattern should fill the same the slots in the firing pattern matrices $M_{e_k(s)}^j$ for different $e_k(s)$. Consequently, elements of $M_{e_k(s)}^j$ which are non-zero across the stimulus exemplars more frequently than would be expected by chance (where chance levels can be estimated from the average firing rate f) are diagnostic of the a PG firing pattern, and the larger the proportion of stimulus exemplars for which these elements are non-zero, the more established and reproducible the PG firing pattern is across the responses to the stimuli from the given category.

We therefore computed $M_s^j = \langle M_{e_k(s)}^j \rangle$, where $\langle \cdot \rangle$ signifies the mean over all the exemplars $e_k(s)$ of stimulus class s , and then identified the ten elements of M_s^j with the largest mean spike counts $(\mathbf{m}_s^j)_n$, where $n \in \{1, \dots, 10\}$ (the element corresponding to the randomly sampled spike by cell j occurring at time t was ignored). These were used to compute $a_s^j = \langle (\mathbf{m}_s^j)_n \rangle / f$, where f is the average firing rate within the layer and $\langle \cdot \rangle$ signifies the mean over the ten largest elements indexed by n .

Thus, a_s^j quantifies the evidence that cell j participates in polychronous firing in its responses to stimulus class s . To calculate an overall polychronization index which is not stimulus specific we simply compute $PI^j = \max_s(a_s^j)$. The larger the PI^j , the stronger the evidence that cell j takes part in a PG.

2.10.4 Information Analysis

Information analysis measures the stimulus-response relationship by quantifying the amount of information present in the firing of a cell about the identity of the input stimulus. If neuronal activity is different in response to different stimulus classes, and is the same for different exemplars of the same stimulus class, then full information about the stimulus class is present. The maximum amount of information, or full entropy, is calculated as follows (Cover & Thomas 1991):

$$H(p) = \sum_s p(s) \log_2 \frac{1}{p(s)}$$

where $p(s)$ is the a priori probability of stimulus class s being present. Therefore, the entropy for two stimuli with equal presentation probability is 1 bit.

Single cell information analysis quantifies how much information about which stimulus identity is present within the responses of a single cell and is calculated as following:

$$I(S; R) = \sum_{s \in S, r \in R} p(s, r) \log_2 \frac{p(s, r)}{p(s)p(r)}$$

where S is the set of all stimuli and R is the set of all possible cell responses, $p(S, R)$ is the joint distribution of stimuli and responses and $p(s) = \sum_{r \in R} p(s, r)$ and $p(r) = \sum_{s \in S} p(s, r)$ are the marginal distributions of stimuli and responses (Nelken & Chechik 2007).

When calculating information measures, it is important to account for bias that occurs when the number of recording trials is not significantly larger than the number of cells in the stimulus/response joint distribution table. In this case the calculated single cell information may be an over-estimate, and the amount of bias needs to be approximated and subtracted from the single cell information value according to the following formula (Panzeri & Treves 1996, Treves & Panzeri 1995):

$$Bias = \frac{\#bins}{2N \log_2 2}$$

where $\#bins$ is the number of potentially non-zero cells in the joint probability table, and N is the number of recording trials. The amount of bias, however, was found to be negligible for the simulations described in this thesis (approximately 0.008 bits) and it was, therefore, ignored.

	Stimulus 1	Stimulus 2
On	$\frac{\#CellOn2Stim1}{\#Stim1*N}$	$\frac{\#CellOn2Stim2}{\#Stim2*N}$
Off	$\frac{\#CellOff2Stim1}{\#Stim1*N}$	$\frac{\#CellOff2Stim2}{\#Stim2*N}$

Table 2.3: Joint probability table used to calculate single cell information for each cell using on/off encoding. $\#CellOn2Stim1$ ($\#CellOn2Stim2$) stands for the number of presentations of stimulus 1 (2) during which the cell fired at least once. Similarly $\#CellOff2Stim1$ ($\#CellOff2Stim2$) stands for the number of presentations of stimulus 1 (2) during which the cell did not fire. $\#Stim1$ ($\#Stim2$) stands for the total number of presentations of stimulus 1 (2). N stands for the total number of stimuli, which is two for the two vowel discrimination paradigm.

2.10.5 Information Encoding

2.10.5.1 Simple Binary On/Off Encoding

In order to measure single cell information, it is important to have a good hypothesis about the way stimulus identity is encoded in the responses of auditory cells. A bad encoding hypothesis will result in poor information scores by being either too simple, thus ignoring some important information about the stimulus identity, or by being overly complicated, thus creating an infeasibly large space of responses R , a large and sparse joint probability table and, therefore, a large bias. For the purposes of chapters 3-4 a very simple 'on/off' encoding scheme was chosen, whereby the response of each neuron unit in the output layer of the network to each stimulus exemplar was considered to be binary - a neuron unit could either be 'on' (if it fired at least once during stimulus presentation), or 'off' (if it never fired during stimulus presentation), resulting in the joint distribution table shown in Table 2.3. Binary encoding has been described in the auditory cortex by DeWeese & Zador (2003).

Such simplistic on/off encoding was chosen because neurophysiological evi-

dence suggests that, throughout the hierarchy of the auditory brain, neurons' response properties change from predominantly sustained to predominantly transient, whereby spectro-temporal sound information found in the early auditory brain, such as the auditory nerve, is transformed into spatio-temporal patterns of neural activity as found in the auditory cortex (Bizley et al. 2010, Huetz et al. 2011). Ehret & Romand (1996) have reported that most neurons in the auditory cortex are onset type, whereby they show little phase locking, and they fire only once to tones of higher frequencies, which is in line with the binary encoding described in the auditory cortex by DeWeese & Zador (2003). Evidence suggests that similar transient response properties exist in the human auditory cortex too (Seifritz et al. 2002).

Since neurons in the auditory cortex have been reported to fire a very limited number of times throughout the duration of tone stimuli used for the neurophysiological studies cited above, it is hypothesised that, at least for the short and simple vowel stimuli used in this study, stimulus information is not encoded within the precise temporal patterns of spike trains of each individual cell, but instead within temporal or spatial patterns over a population of cells. In other words, it is hypothesised that, in the early auditory cortex, each cell might fire once to its preferred phoneme, thus creating a spatio-temporal pattern of firing over the whole population of such early cells in response to a word. Neurons in the later layers of the auditory cortex use these early stable spatio-temporal patterns to learn about the identity of the whole word through spike-time dependent plasticity (STDP) learning (as discussed in Section 2.3.2). STDP-based learning relies on the detection of stable spatio-temporal patterns within the firing of a population of neurons. If these patterns are unique to different stimuli, then cells should emerge at the end of the auditory brain pathway hierarchy that would preferentially fire to one stimulus over the other. Each one of these cells might not fire to all exemplars of the

stimulus, and they may not fire to every presentation of the stimulus, however over the whole population of stimulus selective cells, if most of them are on, it should provide the brain with enough evidence to make a decision that the particular stimulus is being heard. This is in line with the detector theory of auditory object encoding, which suggests that perception of a specific vocalisation is directly related to the activity of a specialised set of neurons that fire selectively to that vocalisation (Suga 1978). Thus, neurons will exist at the end of the auditory brain hierarchy that fire transiently whenever their preferred auditory object is being heard. Such neurons, however, will be discussed in later chapters, while the early chapters 3-4 of this thesis look at the response properties of cells in the primary auditory cortex in response to phonemes, which are sub-units of the full auditory word objects, and are appropriately simple stimuli to be encoded using the simple on/off grandmother-like encoding scheme proposed here.

2.10.5.2 Bayesian Encoding

In order to test whether more information was present about the identity of the Klatt generated vowels within the precise timing of spikes of A1 cells, rather than through the binary on/off encoding described above, a procedure described next was devised for Chapter 5. Intuitively this procedure helps quantify the probability of a sound belonging to a particular stimulus class by using the timing of spikes produced by neurons with respect to the stimulus onset time. If the temporal spike pattern produced by a neuron is more similar to a pattern produced by that neuron when presented with a particular stimulus class in the past, then it is more likely that a stimulus belonging to that particular stimulus class is being presented to the model currently. The quantitative details of the procedure are presented next.

A Bayesian approach was taken to calculate the posterior probability of a particular stimulus being presented to a cell based on the cell's response ($P(s|R)$, where $s \in S$ represents the stimulus, and R represents the set of cell's spikes during stimulus presentation). Based on the Bayesian rule:

$$P(s|R) = \sum_r \frac{P(r|s)P(s)}{P(r)}$$

where $r \in R$ is the timing of each individual spike emitted by a cell during stimulus presentation, and $P(s) = \frac{1}{S}$ with S denoting the stimulus set size.

$$P(r|s) = \frac{count_{rs}}{exemplars_s * epochs_s}$$

where $count_{rs}$ is the number of times the chosen cell had a particular response r (a spike at a particular 1 ms time bin t) during all epochs ($epochs_s$) of all exemplars ($exemplars_s$) of a particular stimulus s .

$$P(r) = \frac{count_r}{\sum_s exemplars_s * epochs_s}$$

where $count_r$ is the number of times the chosen cell had a particular response r (a spike at a particular 1 ms time bin t) during all epochs ($epochs_s$) of all exemplars ($exemplars_s$) of all stimuli $s \in S$.

Stimulus identity vote assigned to a particular cell in response to each stimulus presentation was chosen as $argmax_s P(s|R)$. If two or more stimuli were equally probable, then the response of the cell was classified as 'undecided' and quantified as $S + 1$.

3

The Role of Subcortical Processing in Auditory Category Learning: Models

3.1 Introduction

The hypothesis being tested is that primary auditory cortex (A1) would be unlikely to learn invariant representations of auditory objects such as vowels if it were directly provided with auditory nerve (AN) spike rasters in response to those stimuli as input, as opposed to being provided inputs that have received essential pre-processing from earlier stages in the pathway. This is because spike-time dependent plasticity (STDP) learning will be unable to cope with time and space jitter present in the auditory nerve. Therefore, extra subcortical pre-processing is hypothesised to be necessary between the AN and A1 in order to remove time and space jitter. This pre-processing is hypothesised to happen in the cochlear nucleus (CN) and the inferior colliculus (IC).

In order to test the hypothesis stated above, two models were constructed -

a simple two-layer model consisting of the input AN and the output A1 (AN-A1 model); and a more complex four-layer model, consisting of the input AN, an intermediate CN with distinct sub-populations of neurons, an inferior colliculus (IC) where the different sub-populations of the CN converge, and finally the A1 (AN-CN-IC-A1 model). The IC is expected to have similar spiking behaviour to the AN, albeit with time and space jitter removed through the physiologically plausible pre-processing starting in the CN, thus enabling the A1 to learn invariant auditory object representations of the two input stimuli. It is, therefore, expected that the full AN-CN-IC-A1 model will be able learn transform invariant representations of the two vowel stimuli, whereas the simple AN-A1 model will struggle to do so.

As mentioned in Chapter 1, one of the primary goals of this thesis is to understand which parts of the auditory brain hierarchy between the AN and A1 are important for learning categorical representations of speech-like auditory objects. In this chapter we try to achieve this goal by looking at how spiking neural networks with or without modules representing certain particular auditory brain areas learn to perform classification of simple speech-like auditory stimuli. We take inspiration from the known neurophysiology of the auditory brain in order to construct these spiking neural network models. However, as with any model, a number of simplifying assumptions are made with regards to certain aspects that we believe are not crucial for testing our hypothesis. These simplifications include the lack of superior olivary complex or thalamus in our full AN-CN-IC-A1 model, the nature of implementation of within-layer inhibition in both the AN-A1 and AN-CN-IC-A1 models, and lack of top-down or recurrent connectivity in either model. While we believe that all of these aspects do affect the learning of auditory object categories to some extent, we believe that their role is not crucial for the task. Therefore, we leave the investigation of these effects for future work.

3.2 Simple AN-A1 Model Architecture

The simple AN-A1 spiking neural network model of the auditory brain consists of two fully connected layers of spiking neurons. The input layer of the model represents the auditory nerve (AN). The output layer of the model is a loose approximation of the MGN and A1 in the real brain. The AN cells are based on the auditory nerve model by Zilany et al. (2009) (see Section 2.6), while the output A1 cells are based on the Spike Frequency Adaptation neuronal model by Izhikevich (2003) (see Section 2.9). While there is no neurophysiological evidence of inhibitory inter-neurons in the auditory nerve, such evidence exists for the A1 (Razak & Fuzessery 2007, Zhang et al. 2003), therefore the output layer contains both excitatory (1000) and inhibitory (1000) neurons, thus creating competition within the layer (see Table 3.1 and Section 2.9). In order to manage the computational requirements of the model, the sparse connectivity between the excitatory to inhibitory cells within the A1 is modelled using strong one-to-one connections from each excitatory cell to an inhibitory cell. Each inhibitory cell is fully connected to all excitatory cells in A1. While excitatory cortical cells are more numerous than inhibitory cortical cells in the real brain, the two populations were modelled in the way described above in order to implement and investigate within-layer competition and activity modulation in a controlled way. Informal tests demonstrated that the exact implementation of within-layer inhibition does not have a significant impact on the results presented in this thesis, as long as the implementation still achieves an appropriate level of within-layer competition and activity modulation.

The AN and A1 layers are fully connected using feed-forward connections modifiable through spike-time dependent plasticity (STDP) learning (see Section 2.3.2). The connections are initialised with a uniform distribution of axonal delays (Δ_{ij}) between 0 and 50 ms (see Section 2.4). A range of conduction delays between

layers is a key feature of our models. In real brains, these delays might be axonal, dendritic, synaptic or due to indirect connections, but in the model, for simplicity, all delays were implemented as axonal. The $[0, 50]$ ms range was chosen to approximately match the range reported by Izhikevich (2006). The randomly chosen axonal delay matrix was fixed for all simulations described in this thesis (unless stated otherwise) to remove the confounding effect of different delay initialisation values on learning. Informal testing has demonstrated that the different delay initialisation values do not result in qualitatively different network performance. The delays were fixed, however, to allow for an un-confounded evaluation of the effect of different model parameters on the model performance. The initial afferent connection strength magnitudes (w_{ij}^{BL}) were randomly initialised using values drawn from a uniform distribution. Model parameters can be seen in Table 3.1 and model architecture is depicted in Figure 3.1.

Within Layer Parameters

Corresponding Brain Area	Layer 1 Auditory Nerve	Layer 2 Primary Auditory Cortex
Excitatory Cell Number	1000	1000
Inhibitory Cell Number	0	1000
E→I Connection Strength, w_{ij}^{EI}	N/A	1000 nA
I→E Connection Strength, w_{ij}^{IE}	N/A	{ -6 , -5, -10, ..., -95} nA

Between Layer Parameters

	AN-A1
Axonal Delays, Δ_{ij}	[0, 50] ms
Connectivity	full
Initial Connection Strength, w_{ij}^{BL}	{[30 , 35], [15, 20], [45, 50]} nA
Spike Rule	nearest
STDP Rule	mixed
LTP Time Constant, τ_p	{ 15 , 3, 75} ms
LTD Time Constant, τ_d	{ 25 , 5, 125} ms
LTP Magnitude, α_p	{ 0.005 , 0.05, 0.0005}
LTD Magnitude, α_d	[-1, ..., -20 , ..., -54]*{0.001, 0.01, 0.0001}
Maximum Synaptic Strength, w_{max}	2*Initial Connection Strength nA

Table 3.1: Parameters used for the simple AN-A1 model of the auditory brain. Curly brackets represent sets of parameters explored. Square brackets represent ranges of parameters explored. The first value presented within sets defined by curly brackets indicates the parameter value that was explored as standard, having been found to result in good network performance in informal tests. Parameters highlighted in bold were found to be optimal through grid parameter search. Parameter symbols printed in italics refer to the corresponding parameters in equations 2.1-2.5 in Chapter 2

3.3 Validating the Auditory Nerve Layer

The response properties of the model AN were compared to neurophysiological data to ensure the validity of the model input. First, AN model fibers were individually compared to real AN fibers in response to a synthetic vowel /æ/ (Delgutte & Kiang 1984). Harmonic spectra of period histograms were plotted for the real and model AN fibers (Fig. 3.2), illustrating the variation in neural responses as a function of both the fiber characteristic frequency (CF) and the vowel formant frequencies. Similarity between the real and model AN fibers' response properties suggests that the model performance is comparable to the neurophysiological data.

Population encoding within the model AN was validated by examining layer raster plots and neurograms in response to synthetic vowels /a/ and /i:/ (Fig. 3.3). While vowel spectra are constant in time when measured over the duration of the vowel (see Fig. 3.4), since each vowel is a composition of three formant frequencies which do not change throughout the stimulus duration, it can be seen that AN spike rasters are periodic. This is due to phase-locking and other non-linearities inherent to AN fiber response properties (see Chapter 1 Section 1.5.5 for details).

Overall the rasters and the neurograms of the AN (Fig. 3.3) are in accordance with neurophysiological findings. AN spike rasters show that higher CF neurons start firing before neurons with lower CFs, which is due to the transmission properties of the cochlea and is in line with what has been reported by Secker-Walker & Searle (1990) (Fig. 1.17). It can also be observed that the AN spike rasters contain 100 Hz periodic cycles of activity that reflect the periodicity of the stimuli, since the fundamental frequency (F_0) of the vowels is 100 Hz. The same has been reported in neurophysiological literature (Brugge et al. 1969, Joris et al. 2004).

Looking at the AN neurograms, five different groups of cells can be observed differentiated by whether they synchronise to one of the three vowel formants, to

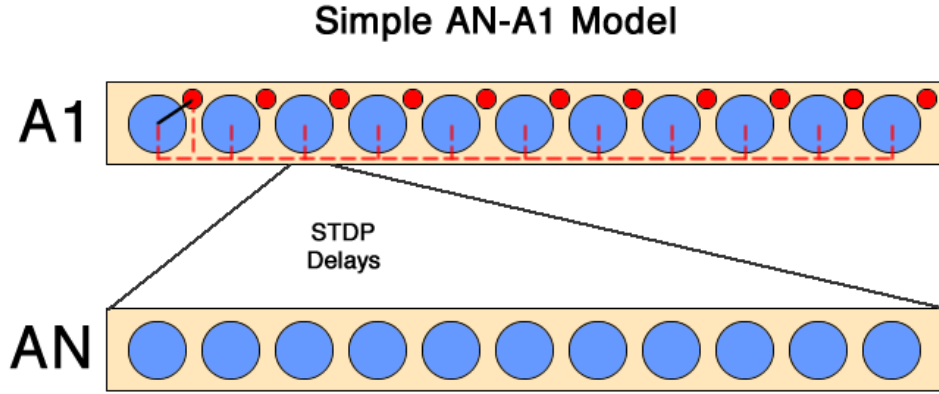


Figure 3.1: Schematic representing the simple AN-A1 model of the auditory brain. Blue circles represent excitatory and red circles represent inhibitory neurons. Auditory nerve (AN) and primary auditory cortex (A1) layers are fully connected using feed-forward connections with randomly initialised weights (w_{ij}^{BL}) and a distribution of axonal conduction delays (Δ_{ij}) drawn from uniform distributions shown in Table 3.1. Feed-forward connections (w_{ij}^{BL}) are modifiable through spike-timing dependent plasticity learning. The sparse connectivity between the excitatory and inhibitory cells within A1 (w_{ij}^{EI}) is modelled as one-to-one feed-forward connections from each excitatory cell to a corresponding inhibitory cell (shown in black), and each inhibitory cell is fully connected to all of the excitatory A1 cells (w_{ij}^{IE}) (red connections).

the fundamental frequency, or to their own CF (Fig. 1.18). This is in line with the neurophysiological data reported by Delgutte & Kiang (1984) and Secker-Walker & Searle (1990). AN neurograms also show that cells synchronised to the first formant (F_1) have their CFs distributed symmetrically around F_1 , while cells synchronised to higher formants have CFs equal to or greater than their corresponding formant. This is in line with the neurophysiological data reported by Sinex & Geisler (1983). The neurograms also show that cells with high and low CFs tend to synchronise to the vowel fundamental frequency (Secker-Walker & Searle 1990), and high CF fibers also tend to synchronise to F_1 , as described by Delgutte & Kiang (1984). Finally, Delgutte & Kiang (1984) also reported variability in the localisation of the AN regions where neurons synchronise to their CFs. This region was found

Response Properties of Auditory Nerve Neurons

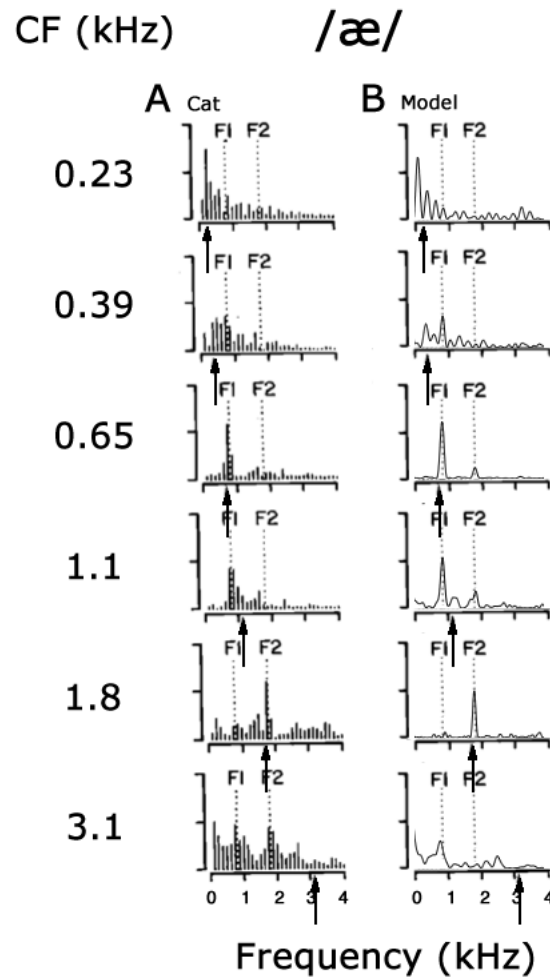


Figure 3.2: Spectra (computed as FFTs of period histograms) of auditory nerve fiber responses to a synthetic vowel /æ/ generated using Klatt synthesiser (Klatt 1980) (see Table 2.1 for formant frequencies). The abscissa represents sound frequencies that the fibers phase-lock to. The ordinate represents the level of phase-locking, calculated as described in Section 2.10.1. High values indicate good phase locking to the corresponding frequency. Dotted lines show the positions of the vowel formant frequencies F_1 and F_2 . Arrows below the graphs show the neurons' characteristic frequencies. **A:** Data collected from cat auditory nerve fibers (Delgutte & Kiang 1984); **B:** Data collected from the corresponding model auditory nerve fibers.

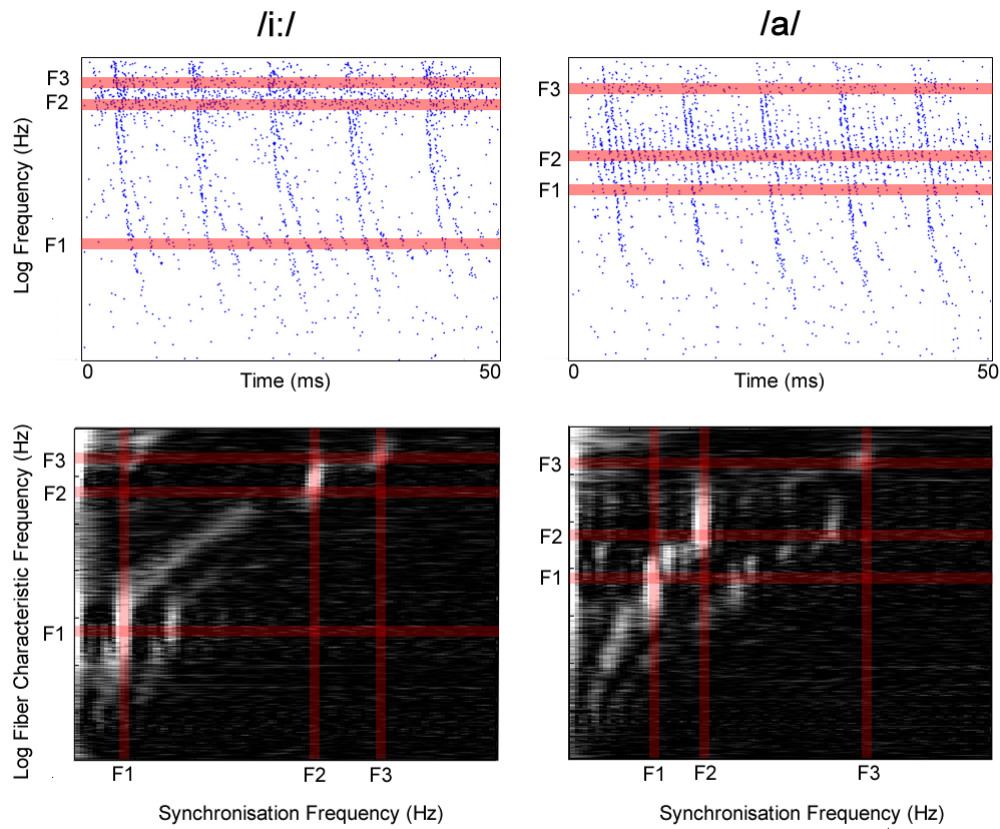


Figure 3.3: Rasters (top row) and neurograms (bottom row) of the first 50 ms of two vowel stimuli, /a/ and /i:/, presented to the model auditory nerve fibers. Within the raster plots the abscissa represents time in ms and the ordinate represents the fibers' characteristic frequency (CF) in log scale. Each blue point represents a spike and red horizontal bands roughly correspond to the frequencies of vowel formants. Within the neurograms the abscissa represents sound frequencies in linear scale, while the ordinate represents the fibers' CF in log scale. Dark areas represent bad phase-locking of fibers with particular CF to particular frequencies, while light areas represent good phase-locking, calculated as described in Section 2.10.1. The horizontal and vertical red bands highlight the location of vowel formant frequencies, and light areas underneath their intersection points represent good phase-locking of auditory nerve fibers to formant frequencies that match the fibers' CF.

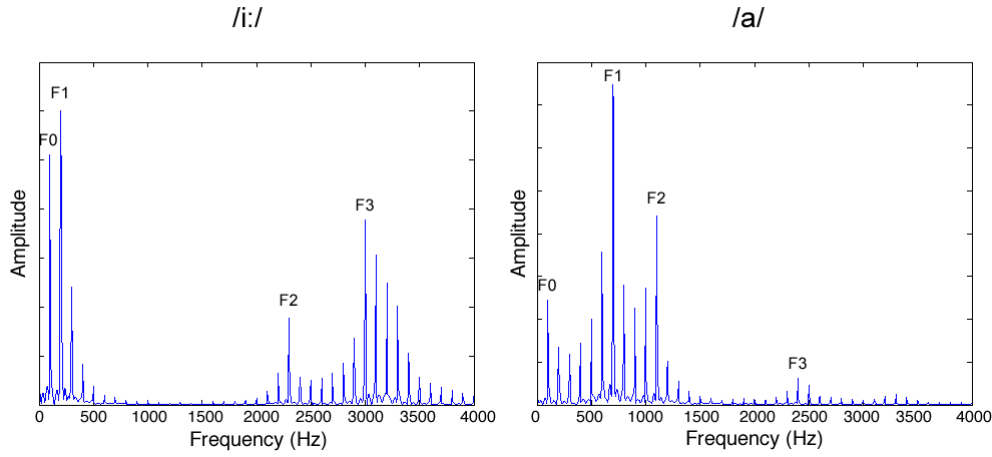


Figure 3.4: Spectra of vowels /a/ and /i:/. The vowels were generated using Klatt synthesiser (Klatt 1980) (see Table 2.1 for formant frequencies).

between F_1 and F_2 for spread vowels /i:/, /ɜ:/ and /ɪ/ and in low-CF region for the open vowels /æ/, /ɔ/, /ɑ/ and /ɛ/. This is exactly what can be observed in the neurograms of the model AN fiber responses to vowels /i:/ and /a/ shown in Figure. 3.3.

The analysis of the AN layer outlined above suggests that the model is representative of the real auditory nerve, thus validating it as the input layer to the spiking neural network model of the auditory brain described in this chapter.

3.4 Optimising the Simple AN-A1 Model

A parameter search of STDP parameter values for the simple AN-A1 model was performed in order to test the hypothesis that a simple spiking neural network model of the auditory brain without extra sub-cortical pre-processing between the input AN and the output A1 is unable to learn invariant representations of auditory objects, and, therefore, will underperform a more complex AN-CN-IC-A1 model with extra sub-cortical pre-processing. While it is easy to construct a model that

does not learn, the architecture of the AN-A1 model is very basic and does not include anything to prevent STDP-based learning from operating, as long as the input contains regular stable repeating patterns that are not distorted with too much noise (Liaw & Berger 1996, Masquelier et al. 2008). Furthermore, the last stage of the AN-CN-IC-A1 model (the IC-A1 stage) is architecturally an exact copy of the AN-A1 model, with the one difference that the input to the A1 of the AN-CN-IC-A1 model has gone through the jitter reduction pre-processing in CN and IC, while the input to the A1 of the AN-A1 model has not. Hence, the only reason why the AN-CN-IC-A1 model should learn good vowel categorisation, while the AN-A1 model should not, is the effect of CN and IC pre-processing.

A grid heuristic was used to sample a representative number of points in the parameter space to get an estimate of the AN-A1 network's optimal performance on the two vowel recognition task. The performance was measured using single cell information analysis as described in Chapter 2 Section 2.10.4. Single cell information measures how much information about the identity of a particular stimulus is present in the responses of one particular cell within a network layer. For two stimulus classes, the theoretical maximum of 1 bit of information is achieved when the cell fires at least once to all presentations of all exemplars of its preferred stimulus class, and it does not fire to any presentations of any transforms of the other stimulus class.

While it is recognised that the best performance may be at a point not sampled during the parameter search, it is assumed that the network performance changes gradually and continuously as a function of parameters, since any learning mechanism employed by the brain has to be robust to mild variations in biophysical constants. Therefore, any missed global maximum for the network performance is expected to be not significantly dissimilar to the best network performance found through the grid parameter search.

The parameter that most affects the network learning performance is the relative magnitude of LTD (α_d) compared to LTP (α_p). Therefore, a number of LTD values (α_d) were sampled while LTP (α_p) was kept fixed. This was done for three different values of LTP changed by a factor of ten, resulting in the following set of values for LTP magnitude (α_p): $\{0.005, 0.05, 0.0005\}$ (the best performing parameter is placed first based on the informal simulations results). Fifty different equally spaced magnitudes of LTD (α_d) were sampled for each LTP value, whereby each LTD (α_d) value was set as a multiple of the LTP (α_p) magnitude, with the multiplication constant chosen from the closed interval $[-0.2, -10.8]$. Furthermore, the LTP/LTD time constants (τ_p/τ_d) were varied by shortening or lengthening the original values of 15/25 ms by a factor of five while maintaining the 3/5 ratio described in neurophysiological literature (Bi & Poo 1998), resulting in the set $\{15/25, 3/5, 75/125\}$ ms (see Table 3.1).

Informal tests demonstrated that the AN-A1 afferent connection weights (w_{ij}^{BL}) initialised uniformly between 30 and 35 nA produced best results. Furthermore, exploratory tests demonstrated that the optimal within-layer inhibition value (w_{ij}^{IE}) was -6 nA. These values were used for the main STDP parameter search simulations.

The results of the parameter search can be seen in Figure 3.5. The color of each cell of the figure corresponds to the maximal single cell information reached within the output layer A1 of the AN-A1 model after the model was initialised with the different combinations of parameters sampled during the grid search as described above. It is evident that the network reaches its maximum performance when run with the parameters found to be optimal using informal testing (LTP magnitude (α_p) of 0.005, LTP/LTD time constants (τ_p/τ_d) of 15/25 ms and within A1 inhibition (w_{ij}^{IE}) set to -6 nA). The maximum performance of 0.57 bits of information is reached when LTD magnitude (α_d) is set to -0.02, however this is

not enough for an individual cell to differentiate between the two vowel stimuli across all exemplars. Furthermore, only one cell reaches this level of information and the next most informative cell contains only 0.255 bits. The second best performing network after training reached the maximum of 0.51 bits of information with a smaller drop in information content within the top 10 information bearing cells (see Fig. 3.6, which plots the descending order of single cell information achieved by all 1000 cells within the different layers of the AN-A1 and AN-CN-IC-A1 models). The parameter sets for the two best performing networks can be seen in Table 3.2. Since the second model shows more stable information retention, from this point onwards it will be the model considered when describing the best performing simple AN-A1 model. The amount of information for the longer LTP/LTD (τ_p/τ_d) time constants of 75/125 ms is mostly 0, because these values result in full strengthening of all connections and an explosion in network activity. Therefore, this suggests that STDP time constant lengthening does not improve the network performance on the transform invariant vowel recognition task, even though it changes the network learning rule from relying on the precise temporal spike information, which is contaminated with time jitter, to relying on rate information, which is mostly unaffected by time jitter.

	Model 1	Model 2
Single Cell Information	0.57	0.51
<i>LTP Time Constant, τ_p (ms)</i>	15	15
<i>LTD Time Constant, τ_d (ms)</i>	25	25
<i>LTP Magnitude, α_p</i>	0.005	0.005
<i>LTD Magnitude, α_d</i>	-0.02	-0.033
<i>I→E Connection Strength, w_{ij}^{IE} (nA)</i>	-6	-6
<i>AN→A1 Connection Strength, w_{ij}^{BL} (nA)</i>	[30, 35]	[30, 35]

Table 3.2: Parameters found to maximise the amount of single cell information within a simple AN-A1 neural network model of the auditory brain on the two vowel discrimination task. Model 1 reaches higher maximal single cell information of 0.57 bits (out of maximal 1 bit), however this number drops fast for the second most informative cell (down to 0.255 bits). Model 2 reaches second highest maximal single cell information of 0.51 bits and retains it better, whereby the second most informative cell contains 0.39 bits of information (see Fig. 3.6).

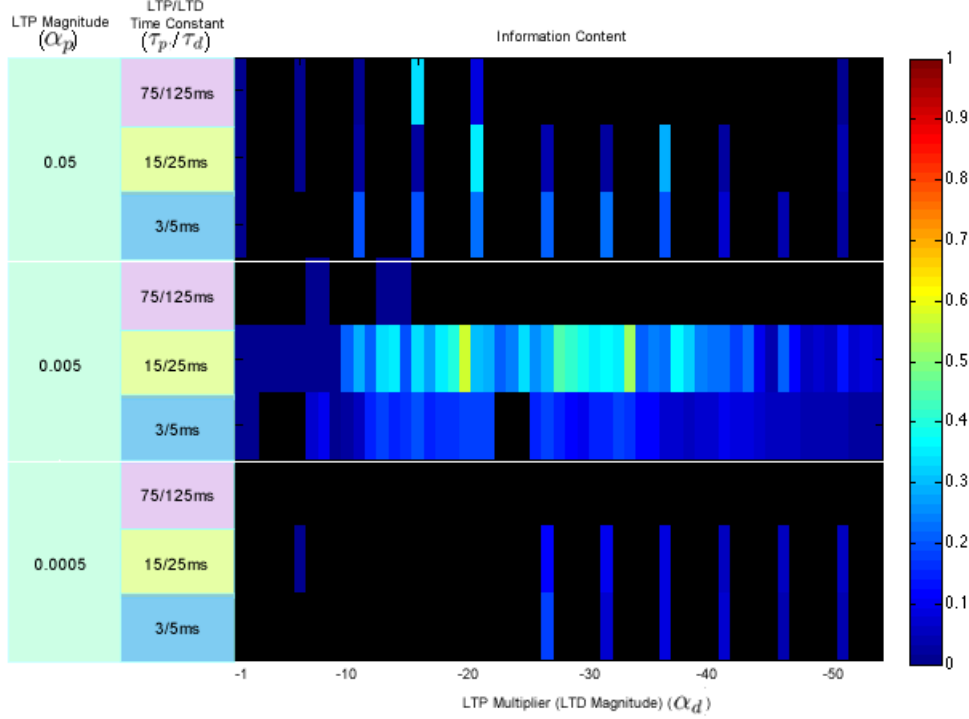


Figure 3.5: Results of the spike-timing dependent plasticity parameter search for a simple AN-A1 neural network model of the auditory brain on the two vowel discrimination task. The rows represent maximum (out of 1000 excitatory A1 cells) single cell information content within the output A1 layer when the relative LTD magnitude (α_d) was varied while keeping the LTP magnitude (α_p) and LTP/LTD time constants (τ_p/τ_d) fixed. In order to calculate the actual LTD values used in different simulations, it is necessary to multiply the abscissa value by the corresponding LTP value (α_p) divided by 5. The maximum amount of information of 0.57 bits was reached for the -20 LTD multiplication constant when LTP magnitude (α_p) was set to 0.005, resulting in the -0.02 LTD magnitude (α_d). The theoretical maximum amount of information is 1 bit, therefore the best performing network did relatively poorly in the two vowel identification task. Black indicates no data due to either no sampling of the particular point in the parameter space, or due to inability of the network to complete a simulation because of firing explosion in the output layer. Such a firing explosion happens when too many AN→A1 afferents (w_{ij}^{BL}) strengthen too much.

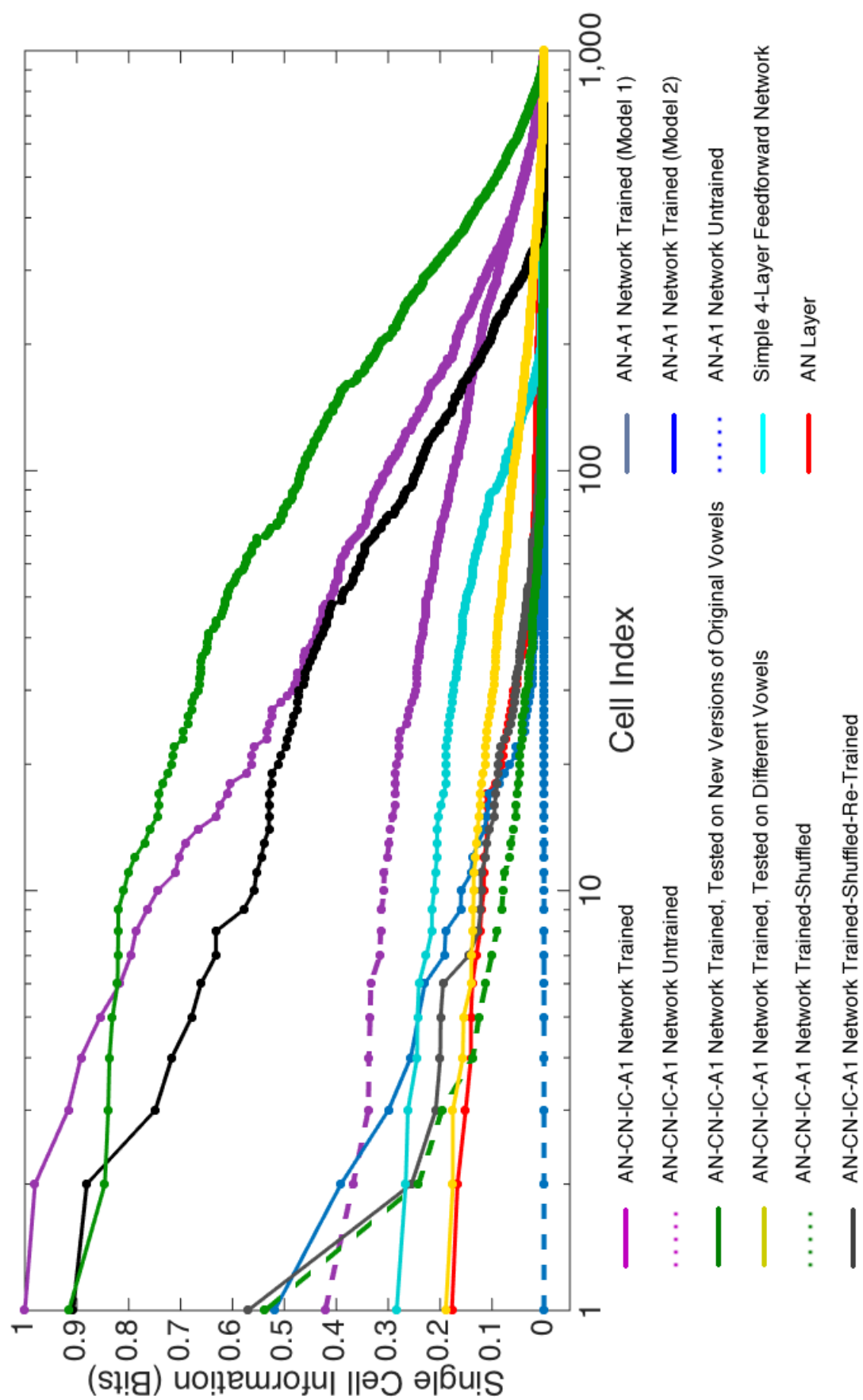


Figure 3.6: (Figure legend on next page)

(Image on previous page) Single cell information in response to the two vowel discrimination task carried by single cells within a layer arranged in descending order. Maximum theoretical entropy for the task is 1 bit. It can be seen that both the simple AN-A1 and the full AN-CN-IC-A1 spiking neural network models of the auditory brain carry more information about the two stimuli than the input auditory nerve (AN) layer or the untrained networks. Only cells within the output A1 layer of the full AN-CN-IC-A1 model reached the maximum 1 bit of information. The performance of the AN-CN-IC-A1 model is significantly above that of any other network pictured.

AN-CN-IC-A1 Network Trained: the full AN-CN-IC-A1 network tested on twelve exemplars of each of the two vowels, /i:/ and /a/, after training.

AN-CN-IC-A1 Network Untrained: full AN-CN-IC-A1 network tested on twelve exemplars of each of the two vowels, /i:/ and /a/, before training.

AN-CN-IC-A1 Network Trained, Tested on New Versions of Original Vowels: full AN-CN-IC-A1 network tested on twelve exemplars of each of the new versions of the original two vowels, /i:/ and /a/, after training. The new vowels were synthesised using different formants than the original vowels used during training (see Section 4.5.5).

AN-CN-IC-A1 Network Trained, Tested on Different Vowels: full AN-CN-IC-A1 network tested on twelve exemplars of each of the two new vowels, /ɜ:/ and /æ/, after having been trained on twelve exemplars of each of the original two vowels, /i:/ and /a/ (see Section 4.5.5).

AN-CN-IC-A1 Network Trained-Shuffled: full trained AN-CN-IC-A1 network tested on twelve exemplars of each of the two vowels, /i:/ and /a/, after the magnitudes of the afferent connections for each A1 neuron were shuffled.

AN-CN-IC-A1 Network Trained-Shuffled-Re-Trained: full trained AN-CN-IC-A1 network tested on twelve exemplars of each of the two vowels, /i:/ and /a/, after the magnitudes of the afferent connections for each A1 neuron were shuffled, and then the shuffled network was re-trained using the original training paradigm.

AN-A1 Network Trained (Model 1): simple AN-A1 network tested on twelve exemplars of each of the two vowels, /i:/ and /a/, after training. This network reached the best maximum single cell information of 0.57 bits, however information drops off rapidly for subsequent cells.

AN-A1 Network Trained (Model 2): simple AN-A1 network tested on twelve exemplars of each of the two vowels, /i:/ and /a/, after training. This network reached the second best maximum single cell information of 0.51 bits, and information drops off less rapidly for subsequent cells of Model 2 compared to Model 1. Model 2, therefore, is used as the AN-A1 network for comparison with the full AN-CN-IC-A1 network.

AN-A1 Network Untrained: simple AN-A1 network tested on twelve exemplars of each of the two vowels, /i:/ and /a/, before training.

Simple 4-Layer Feedforward Network: simple feed-forward four-layer network tested on twelve exemplars of each of the two vowels, /i:/ and /a/, after training. The network was used to test whether the mere presence of extra layers could account for the better performance of the full AN-CN-IC-A1 network (see Section 3.9).

AN layer: input AN layer tested on twelve exemplars of each of the two vowels, /i:/ and /a/.

An additional parameter search was run to test the effect of different levels of within-layer inhibition (w_{ij}^{IE}) in the output A1 layer of the network, as well as the different ranges of afferent weight initialisation (w_{ij}^{BL}) on the performance of the simple AN-A1 model on the two vowel discrimination task (see Fig. 3.7). No improvement on the maximal information magnitude of the main parameter search was found, thus validating the earlier result that the best performing AN-A1 network achieves no more than 0.6 bits of information on the two vowel discrimination task. This is significantly below the theoretical maximum of 1 bit of information on a two stimulus discrimination task. While one might think that the information deficit could be the result of the simplicity of the chosen encoding scheme (binary 'on/off'), this is in fact not the case, due to the following reasons: 1) the full AN-CN-IC-A1 model is able to reach the theoretical maximum of 1 bit of information using the same simple 'on/off' encoding scheme; 2) in later chapters we will show that when the AN-A1 and the AN-CN-IC-A1 models are presented with more complex stimuli and more complex information encoding schemes are tested, the AN-CN-IC-A1 model still significantly outperforms the simple AN-A1 model, thus supporting the hypothesis that the simple AN-A1 auditory brain model is unable to learn invariant representations of auditory stimuli due to the presence of AN jitter, which impedes STDP-based learning, since the only difference between the AN-A1 and the full AN-CN-IC-A1 models is the presence of jitter removing stages CN and IC.

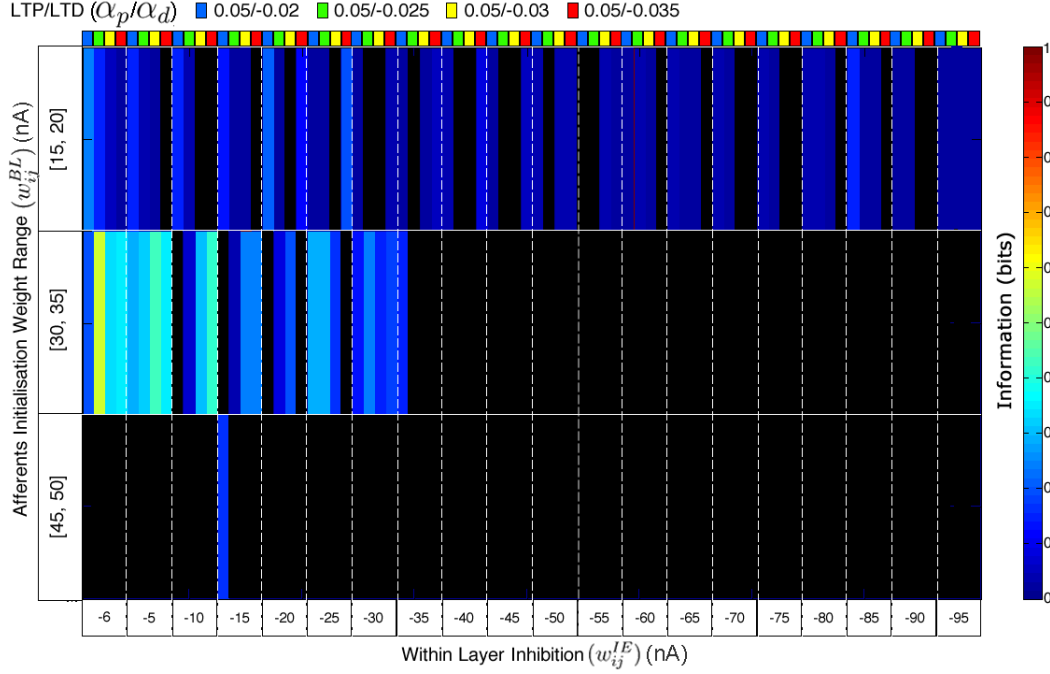


Figure 3.7: Results of an additional parameter search for a simple AN-A1 neural network model of the auditory brain. This parameter search was designed to test the role of within-layer A1 inhibition (w_{ij}^{IE}) and different uniform distribution ranges used for the afferent AN→A1 connection weight (w_{ij}^{BL}) initialisation on the network performance on the two vowel discrimination task. The rows represent maximum single cell information content within the output A1 layer when LTD magnitude (α_d) was varied for different levels of inhibition (w_{ij}^{IE}) and afferents weight initialisation (w_{ij}^{BL}). The theoretical maximum amount of information is 1 bit. Single cell information data is compared to the best performing AN-A1 networks, which had inhibition (w_{ij}^{IE}) set to -6 nA and weight initialisation range (w_{ij}^{BL}) set to [30, 35] nA. It can be seen that any other parameter combinations result in worse network performance, thus validating the finding that the best performing AN-A1 network achieves approximately 0.5 bits of information. Black indicates no data due to inability of the network to complete a simulation because of firing explosion in the output layer. Such a firing explosion happens when too many AN→A1 afferents (w_{ij}^{BL}) strengthen too much, and is the reason why most of the data corresponding to the [45, 50] nA afferents weight initialisation range is black.

3.5 Quantifying Space and Time Jitter in the Auditory Nerve

In order to quantify space and time jitter in the model AN, each of the twelve exemplars of the two vowels /a/ and /i:/ were presented to the layer over 20 epochs as described in Chapter 2 Section 2.2.1. AN rasters in response to the two vowel stimuli were binned within 1 ms bins. Whenever each of the 1 ms bins contained one or more spikes, its spike count was set to one. The rasters were used to calculate three scores to measure AN jitter: Same Exemplar, Different Exemplars and Different Category Index (see Chapter 2 Section 2.10.3).

The three scores indicate how many spikes occur on average within 100 bins with the most number of spikes in proportion to the maximum number of spikes possible within a single bin. The higher the number, the more stable AN rasters are, whereby they have spikes that appear at the same time and at the same place for many stimulus presentations. Low numbers mean that space and time jitter move spikes between stimulus presentations, resulting in unstable firing patterns in the AN. Without jitter/stochasticity present in the AN, Same Exemplar Index is expected to be the highest score, as AN rasters in response to the same auditory input are expected to be very similar. The Different Exemplars Index is supposed to be the second highest score, since AN responses to different exemplars of the same stimulus are still expected to be quite similar to each other, but not as much as the responses to exactly the same auditory input. Finally, the Different Categories Index is supposed to be the lowest score, since AN rasters in response to the exemplars of different stimuli are expected to be very different.

It was found that all three index scores were similar to each other (Table 3.3), which suggests that jitter is present in the model AN. This AN jitter is strong, since AN rasters in response to the same exemplar are almost as dissimilar to each other as AN rasters in response to different exemplars of the same vowel. Furthermore, AN rasters in response to the two different vowel stimuli are almost as dissimilar to each other as those in response to different exemplars of the same vowel.

	/i:/	/a/
Same Exemplar Index	0.45	0.57
Different Exemplars Index	0.52	0.63
Different Categories Index	0.42	0.42

Table 3.3: Same Exemplar, Different Exemplars and Different Category Indices measuring space and time jitter within the auditory nerve.

3.6 Jitter Reduction in the Cochlear Nucleus and Inferior Colliculus

As discussed above and in Chapter 2 Section 2.8 AN time and space jitter are significant enough to potentially impair spike-timing dependent learning (STDP) in the auditory cortex. It is hypothesised, therefore, that there needs to be an intermediate stage before the auditory cortex where AN spike rasters are cleared of jitter. This can happen through the following mechanisms: 1) information from a number of AN fibers with similar characteristic frequencies is integrated in order to remove space jitter; and 2) AN spike trains for each cell are synchronised, whereby spikes are re-aligned to occur at set points in time rather than in continuous time, thus removing time jitter.

It is hypothesised that time and space jitter removal occurs in the cochlear nucleus (CN) and the inferior colliculus (IC). CN is well suited for the task, because it is the first central nucleus of the mammalian auditory system where all AN neurons synapse, and it is the place where physiological diversity arises from essentially homogeneous information from the AN through convergence, inhibition and cell membrane specialisations (Ehret & Romand 1996, Frisina 2001, Irvine 1986, Rhode & Greenberg 1992). The AN information is essentially divided into parallel streams, each of which undergoes further processing before converging at the inferior colliculus and then the auditory cortex (Lai et al. 1994). The CN has three anatomical sub-divisions of cells with afferent connectivity and response properties that suggest that they are well suited for the jitter removal task: the primary-like with notch, chopper and onset neurons. While their general neurophysiology was discussed in Chapter 1 Section 1.5.6, their properties in relation to time and space jitter removal are discussed below. The initial jitter reduction produced within the CN is finished within the IC, where the different sub-populations of the CN are converged and integrated in such a way as to maximise the performance of the CT learning mechanism for stimulus selective transform invariant auditory object recognition. It is hypothesised that the following processes take place within the

CN and IC: primary-like with notch or chopper cells (PLn/CH) within the CN remove space jitter, onset (ON) cells within the CN remove time jitter, IC produces spike rasters with reduced space and time jitter by combining the afferent activity from the PLn/CH and ON cells in the CN. The neurophysiological background to support this hypothesis is discussed next.

3.6.1 Space Jitter Removal

Either chopper (CH) or primary-like with notch (PLn) neurons in the CN are suitable for the space jitter removal task due to their similar afferent connectivity patterns from the AN. Each PLn and CH cell receives a small number of afferent connections from AN neurons with similar characteristic frequencies (CFs) (Delgutte 1982, Ferragamo et al. 1998, Winslow et al. 1987). The incoming signals are integrated to produce regular spike trains (Fig. 3.8). The two types of cells, however, have different intrinsic properties, with PLn cells showing strong phase-locking to the stimulus (Joris & Smith 2008), while CH cells showing firing patterns that reflect the amplitude of the stimulus without good phase-locking (Koch et al. 1982, Winslow et al. 1987, Winter & Palmer 1990).

The results of a parameter search aimed at finding the optimal settings for space jitter removal can guide the selection of either PLn or CH sub-populations for the space jitter removal task, since the different response properties of the two types of cells correspond to different regions of the same parameter space.

3.6.2 Time Jitter Removal

If strong inhibitory feedback is present in the PLn/CH layer, it may help remove some of the time jitter by synchronising the firing of PLn/CH neurons. This effect, however, is not expected to be enough for complete time jitter removal, which is hypothesised to be done by onset neurons in the CN. Onset cells are relatively rare, constituting only 10% of the ventral CN (Rhode et al. 2010). They have been estimated to each receive connections from up to 65 AN fibers across a large area of the cochlea, which results in very broadly frequency tuned response properties (Godfrey et al. 1975, Liberman 1993, Rhode et al. 2010). These cells are characterised by fast membrane time constants, which makes them very leaky with high spike thresholds. This means that onset cells require strong synchronisation from many AN fibers with a wide range of CFs in order to produce a discharge (Cant & Hyson 1992, Oertel et al. 2000, Osen 1969). In response to high-frequency tones onset cells discharge once at the beginning of the stimulus independently of the stimulus duration (Oertel et al. 2000). In response to a low-frequency tone up to 500-800 Hz, onset cells generate a single spike at the beginning of the stimulus, followed by a discharge to every cycle of the tone (Godfrey et al. 1975). Therefore,

onset cells are expected to fire to every cycle of vowel fundamental frequency, since the range of F_0 is 80-280 Hz (Denes & Pinson 1993, Olive et al. 1993). This is supported by neurophysiological evidence. For example, Winter et al. (2003) reported that onset units with CF near F_1 of vowel /a/ phase-locked to the vowel fundamental frequency. This was not affected by the addition of noise and was not restricted to fibers with CFs between formants, like in the AN. This ability to represent sound fundamental frequency comes from across-frequency coincidence detection inherent to the onset cells (Bal & Oertel 2001, Ferragamo & Oertel 2002, Golding et al. 1999, Kim et al. 1986, Palmer & Winter 1996, Rhode & Smith 1986).

It is hypothesised that the interplay between the onset cells in the CN and the IC can help reduce time jitter. Since onset cells synchronise to the vowel fundamental frequency, they can introduce regularly spaced strong afferent input to the IC. If these afferent currents are sub-threshold, then their role within the IC would be to prime the post-synaptic cells to discharge at times corresponding to the cycles of vowel fundamental frequency. If the IC cells also receive afferent connections from PLn/CH cells, then onset afferents will help synchronise PLn/CH spike trains within the IC by increasing the likelihood of the IC cells firing at the beginning of each fundamental frequency cycle. This would be the same as snapping incoming PLn/CH spikes that occur during the time bins defined by the vowel F_0 to the edges of the bins (Fig. 3.8). The theoretical principles of such an encoding scheme have been described by Hopfield (1995) in the context of olfactory perception. Hopfield (1995) demonstrated that since stable spatio-temporal patterns arise locally around each peak of the subthreshold oscillation in the system, such a system is robust to the different oscillation frequencies and the multiplicative scaling of the input currents. In the context of auditory object recognition of the current thesis, such an encoding scheme means that the system is robust to the different F_0 frequencies and to the multiplicative scaling of the PLn/CH→IC afferent connections (w_{ij}^{BL}). The robustness of the system to different stimulus F_0 will be demonstrated empirically in Chapter 6, where our model will be shown to be capable of learning informative speaker-independent representations of two naturally spoken words, "one" and "two", pronounced by 94 different speakers with different F_0 .

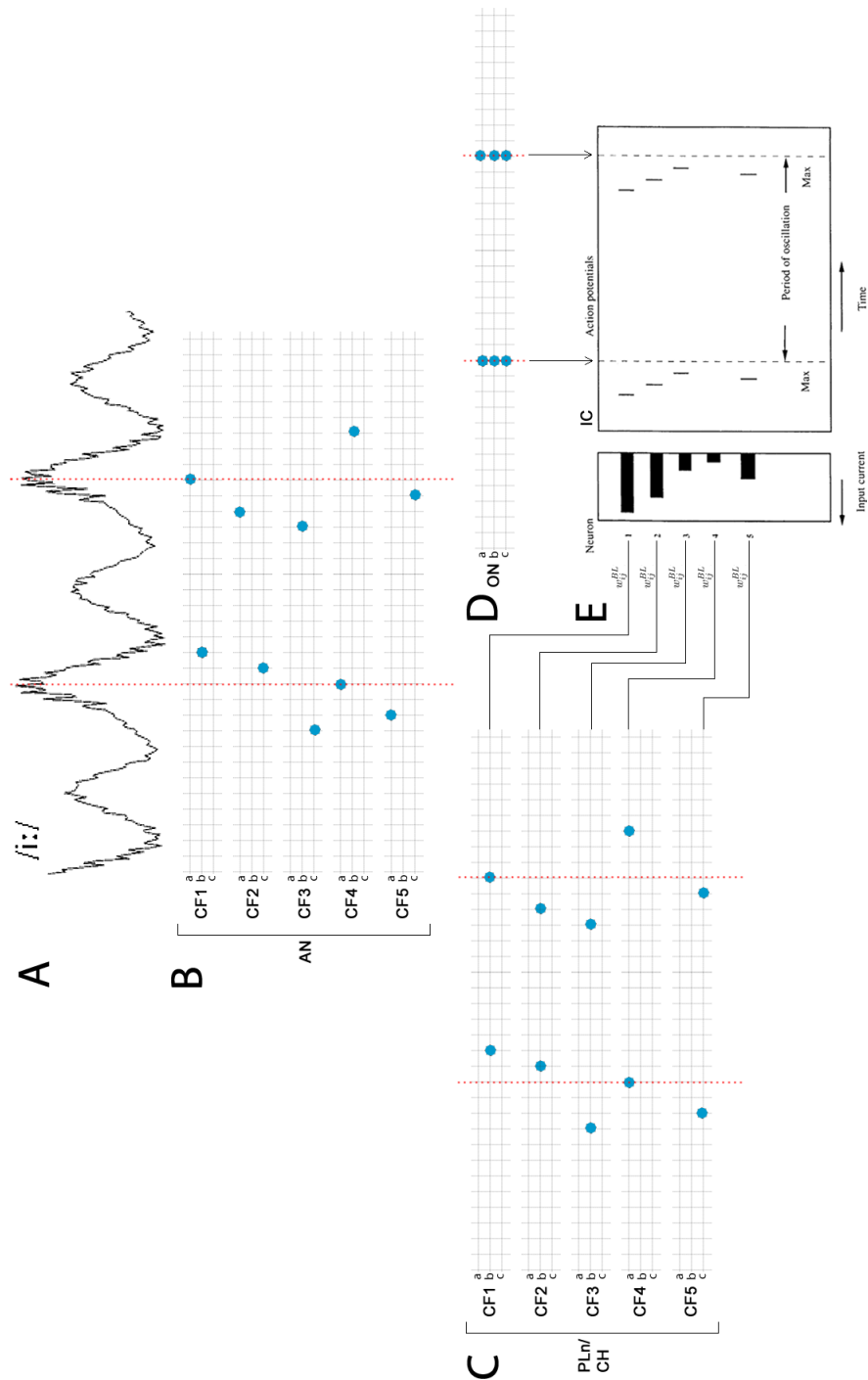


Figure 3.8: (Figure legend on next page)

(Image on previous page)

Schematic representation of space and time jitter removal by primary-like with notch (PLn)/chopper (CH) and onset (ON) cells in the cochlear nucleus (CN) and inferior colliculus (IC) cells. **A** - complex sound wave stimulus corresponding to vowel /i:/. Red vertical dashed lines indicate the fundamental frequency (F_0) of the vowel. **B** - schematic representation of the spikes produced by five sets of auditory nerve (AN) fibers with shared respective characteristic frequencies (CF1, CF2, ..., CF5) in response to the auditory stimulus shown in A. Each group of AN fibers with the same CF consists of three neurons (a, b, and c). The spikes of AN neurons in each group are corrupted with space and time jitter. "Space jitter" refers to the fact that neighbouring AN fibers have almost identical tuning properties so that an action potential that might be expected at a particular fiber at a particular time may be observed in one of the neighbouring fibers. When the amplitude of the sine wave reaches its peak around F_0 , any of the three depicted AN fibers (a, b, or c) with similar characteristic frequency (CF) may fire. "Time jitter" arises when the AN fiber propensity to phase lock to temporal features of the stimulus is degraded to a greater or lesser extent by poisson-like noise in the nerve fibers, refractoriness and the neuron's intrinsic firing properties. This stochastic delay in spike time is illustrated by the difference in spike times with respect to F_0 within each CF group for the two different cycles of the vowel. **C** - responses of five PLn/CH neurons receiving afferent connections from the AN fibers within the different CF groups shown in B. Each PLn/CH fiber integrates the firing of all the AN fibers within a group that shares a similar CF. It can be seen that space jitter is removed, since the same AN fiber within each CF group reliably fires to every F_0 peak of the auditory stimulus. **D** - response of ON neurons receiving afferent connections from the AN fibers shown in B. The ON neurons fire around the stimulus F_0 peaks indicated by the red vertical lines. **E** - adapted from Hopfield (1995). Responses of IC neurons. These neurons integrate the activity received from the PLn/CH neurons (shown in C) and the ON neurons (shown in D). Input received from the ON neurons creates subthreshold oscillation in the IC fibers. The input currents of varying strength (w_{ij}^{BL}) from the PLn/CH neurons result in a reproducible spatio-temporal pattern of firing with respect to the stimulus F_0 oscillations due to the interactions of PLn/ON inputs with the subthreshold depolarisation created by the ON inputs. Such an encoding scheme is robust to the exact frequency of stimulus F_0 (Hopfield 1995) .

3.7 Full AN-CN-IC-A1 Model Architecture

The full AN-CN-IC-A1 spiking neural network model of the auditory brain consists of four layers of spiking neurons as shown in Figure 3.9. In contrast to the simple AN-A1 network described in Section 3.2, the full AN-CN-IC-A1 model includes intermediate cochlear nucleus (CN) and inferior colliculus (IC) between the input AN and output A1. In the brain sub-populations of the CN do not necessarily synapse on the IC directly. Instead, they pass through a number of nuclei within the superior olivary complex (SOC). The nature of processing done within the SOC in terms of auditory object recognition (rather than sound localisation), however, is unclear (see Chapter 1). The information from the different CN sub-populations does converge in the IC eventually, and for the purposes of the current model we model this convergence as direct. A similar simplified connectivity pattern (direct CN-IC projections) was implemented by Meddis & O’Mard (2006) for their model of the sub-cortical auditory brain (see Chapter 1 Section 1.6.2.1 for differences from our model). Similar reasoning applies when justifying our decision to skip the thalamus and model the projections from the IC to A1 as direct.

The addition of the CN and IC is hypothesised to remove time and space jitter present in the AN. Ventral cochlear nucleus consists of distinct sub-populations of cells: the primary-like (PL), primary-like with notch (PLn), chopper (CH) and onset neurons (ON), discussed in more detail in Chapter 1 Section 1.5.6. As mentioned above in Section 3.6, it is hypothesised that ON neurons help remove time jitter, while either CH or PLn neurons primarily remove space jitter (see Section 3.6). PL neurons make up approximately 47% of the ventral CN in the brain (Winter & Palmer 1990), suggesting that they might play a significant role in auditory processing. Although their contribution to the preprocessing of AN discharge patterns is perhaps less clear than that of the PLn/CH and ON subpopulations, PL cells were included in the model architecture to investigate their effect on auditory class learning.

While the newly introduced CN is hypothesised to start the removal of time and space jitter present in the AN, the IC is hypothesised to be the place where the jitter removal is completed through the convergence of the different sub-populations of CN neurons. The jitter free IC spike rasters can then serve as inputs to the A1. Such architecture with the intermediate IC serving as the layer on which the different CN sub-populations converge is supported by neurophysiological evidence. It suggests that the IC is the area where multiple auditory processing streams in the brainstem converge, before moving up to the cortex (Portfors & Felix 2005). The main projections to the IC have been reported to come from the cochlear nucleus (Frisina et al. 1998), and it is suggested that the IC plays the role of a ‘whiteboard’ on which acoustical information is written for use by the higher brain areas, which is similar to the role of V1 in visual processing (Nelken 2008).

The input AN cells of the full AN-CN-IC-A1 model are based on the auditory nerve model by Zilany et al. (2009) (see Section 2.6), while the cells within the remaining CN, IC and A1 layers are based on the model by Izhikevich (2003) (see Section 2.9). The excitatory cells in layers CN and IC are modelled using type G Izhikevich neurons, while the excitatory cells in A1 are modelled as type F Izhikevich neurons to introduce adaptation. While there is no neurophysiological evidence of inhibitory inter-neurons in the auditory nerve, such evidence exists for the other model layers (Davis 2005, Needham & Paolini 2003, Razak & Fuzessery 2007, Winter & Palmer 1990, Zhang et al. 2003), which, therefore, contain both excitatory and inhibitory neurons. The excitatory to inhibitory cell connectivity within the CN, IC and A1 layers is the same as described for the A1 layer within the simple AN-A1 model (see Section 3.2). Each of the three sub-populations of cells within the CN layer implemented in this network has its own independent pool of inhibitory cells.

As discussed in Chapter 1 Section 1.5.7, neurophysiological evidence suggests that the nuclei of the superior olivary complex are primarily specialised for sound localisation rather than identification. We, therefore, simplify these areas away by modelling direct connections from the CN to the IC, resulting in the following CN connectivity: each ON cell receives afferents from a number of randomly distributed AN cells, and each ON cell synapses to all IC cells; and PLn/CH cells have Gaussian topological afferent and efferent connectivity, whereby each PLn/CH cell receives afferents from a larger topological region of the AN, and each cell in the IC receives afferents from a smaller topological region of the PLn/CH sub-population of the CN layer. Neurophysiological evidence suggests that PL neurons receive strong one-to-one AN afferents, which synapse straight onto the cell somata, thereby producing a high degree of fidelity in information transmission from the AN fibers to the primary-like fibers (Blackburn & Sachs 1989, 1990, Palmer et al. 1986, Young & Sachs 1988). PL cells essentially transcribe AN firing in the brain (Winter & Palmer 1990) and were, therefore, modelled using strong ($w_{ij}^{BL} = 1000$ nA) one-to-one AN→PL afferent connections and no inhibition ($w_{ij}^{IE} = 0$ nA) within the PL area.

The IC and A1 layers are fully connected using feed-forward connections (w_{ij}^{BL}) modifiable through spike-timing dependent plasticity (STDP) learning (see Section 2.3.2). The connections are initialised with a uniform distribution of axonal delays (Δ_{ij}) between 0 and 50 ms (see Section 2.4). Similarly to the reduced AN-A1 model, these delays might be axonal, dendritic, synaptic or due to indirect connections, but in the AN-CN-IC-A1 model, for simplicity, all delays were implemented as axonal. The randomly chosen axonal delay matrix was fixed for all simulations described in this chapter to remove the confounding effect of different delay initialisation values on learning. The initial afferent connection strength magnitudes

(w_{ij}^{BL}) were randomly initialised using values drawn from a uniform distribution. Model parameters can be seen in Table 3.4.

The proposed model of the auditory brain, therefore, consists of areas AN, CN, IC and A1. While the first two areas of the proposed model, AN and CN, are modelled to closely resemble the real AN and CN, the higher areas IC and A1 are meant to be more loose approximations of the higher parts of the auditory brain hierarchy. The exact locations of the homologues of the modelled IC and A1 in the brain are not stated. They could be a combination of neighbouring areas in the brain (for example, IC may be an approximation of a combination of the real SOC + IC + MGN; model A1 may be an approximation of the real MGN + A1). The inclusion of these areas in the model, however, is paramount in order to allow the model to learn robust representations of auditory objects, such as phonemes or words. While this full AN-CN-IC-A1 model is necessarily a simplification of the real auditory brain hierarchy up to the primary auditory cortex, we believe that the model captures some of the core principles present in the brain and is more biologically plausible than many previous attempts at modelling the auditory brain using spiking neural networks (Appeltant et al. 2011, Coath et al. 2014, Cohen et al. 2006, Maass et al. 2004, Nager et al. 2002). We, therefore, believe that the model can be used to produce neurophysiologically testable hypotheses about the computational mechanisms taking part in the auditory brain in order to learn auditory categories.

AN-CN-IC-A1 Model Parameters

Within Layer Parameters						
	Layer 1	Layer 2a	Layer 2b	Layer 2c	Layer 3	Layer 4
Corresponding Brain Area	AN	PLn/CH (CN)	PL (CN)	ON (CN)	IC	A1
Excitatory Cell Number	1000	1000	1000	100	1000	1000
Inhibitory Cell Number	0	1000	1000	100	1000	1000
E→I Connection Strength, w_{ij}^{EI} (nA)	N/A	0	0	1000	1000	1000
I→E Connection Strength, w_{ij}^{IE} (nA)	N/A	0	0	-75	{-6, 0, -15}	-6
Between Layer Parameters						
	AN-PLn/CH	AN-PL	AN-ON	PLn/CH-IC	PL-IC	ON-IC
Axonal Delays, Δ_{ij} (ms)	N/A	N/A	N/A	N/A	N/A	N/A
Connectivity	Gaussian ($\sigma = 26$)	1-to-1	sparse (0.46)	Gaussian ($\sigma = 2$)	1-to-1	full
Initial Connection Strength, w_{ij}^{BL} (nA)	[25, 30]	1000	26	{0, 200, 300, 400}	{0, 400, 500, 600}	{0, 1, 3, 5}
Spike Rule	N/A	N/A	N/A	N/A	N/A	N/A
STDP Rule	N/A	N/A	N/A	N/A	N/A	N/A
LTP Time Constant, τ_p (ms)	N/A	N/A	N/A	N/A	N/A	N/A
LTD Time Constant, τ_d (ms)	N/A	N/A	N/A	N/A	N/A	N/A
LTP Magnitude, α_p	N/A	N/A	N/A	N/A	N/A	N/A
LTD Magnitude, α_d	N/A	N/A	N/A	N/A	N/A	N/A
Max Synaptic Strength, w_{max} (nA)	N/A	N/A	N/A	N/A	N/A	N/A

Table 3.4: Parameters used for the full AN-CN-IC-A1 model of the auditory brain trained to discriminate between two vowels /a/ and /i:/. The four layers are: AN - auditory nerve; CN - cochlear nucleus with three sub-populations of cells discriminated based on their connectivity: primary-like with notch (PLn)/chopper (CH), primary-like (PL) and onset (ON); IC - inferior colliculus; A1 - primary auditory cortex. Curly brackets represent sets of parameters explored. Square brackets represent ranges of parameters explored. Parameters highlighted in bold were found to be optimal through grid parameter search. Parameter symbols printed in italics refer to the corresponding parameters in equations 2.1-2.5 in Chapter 2

3.8 Optimising the Full AN-CN-IC-A1 Model

Before a parameter search could be run to optimise the performance of the full AN-CN-IC-A1 model of the auditory brain for the two vowel discrimination task, two additional parameter searches were conducted in order to find the settings that achieve the desired jitter removal properties of the CN sub-populations.

3.8.1 Optimising Primary-Like With Notch/Chopper Cells

As mentioned in Section 3.6, both primary-like with notch (PLn) and chopper (CH) cells receive afferent connections from a tonotopic region of the auditory nerve, and are themselves topologically organised according to their preferred frequency (PLn cells, however receive fewer afferents than CH cells). In order to model this connectivity, each AN cell was assigned a PLn/CH cell as the centre of a Gaussian distribution. The Gaussian was used to determine the probability of that AN cell synapsing onto each of the PLn/CH cells. Each AN cell was assigned a PLn/CH cell in an equidistant manner to uniformly cover the whole PLn/CH layer. The variance of the Gaussian, determined by parameter σ , affects the number of AN cells that each PLn/CH cell receives afferents from, with smaller values of σ resulting in fewer and more localised afferents and, therefore, more sharply frequency tuned input.

The PLn/CH layer was made of 1000 Class 1 neurons as shown in Figure 2.12 (G) (Izhikevich 2003), and was initialised with Gaussian connectivity from the AN layer. A parameter search was performed, whereby the values of σ , the magnitude of afferent connections from AN to PLn/CH cells (w_{ij}^{BL}), and the level of inhibition (w_{ij}^{IE}) in the PLn/CH layer was varied. The parameters were evaluated based on the PLn/CH layer performance on the space jitter removal task. Time jitter removal could also be done by the onset sub-population of the CN, whereas space jitter removal is hypothesised to only be done in the PLn/CH sub-population, hence the performance of the layer on the space jitter removal task was considered to be of the most importance for choosing the parameter settings (see Chapter 2 Section 2.8 for a reminder of our definition of space and time jitter). The ON layer is not capable of space jitter removal, since all ON cells synchronise to the stimulus fundamental frequency. The performance of the PLn/CH layer on the space jitter reduction task was evaluated by calculating the amount of maximal single cell information present in the firing of PLn/CH neurons. Information analysis was chosen as an appropriate measure of PLn/CH layer performance, because, due to the characteristics of PLn/CH cells described in Section 3.6, the more space jitter they remove from the AN input, the more reliably and more selectively they would respond to vowel formants, resulting in more single cell information about the stimuli. Information analysis is not sensitive to time jitter present in the

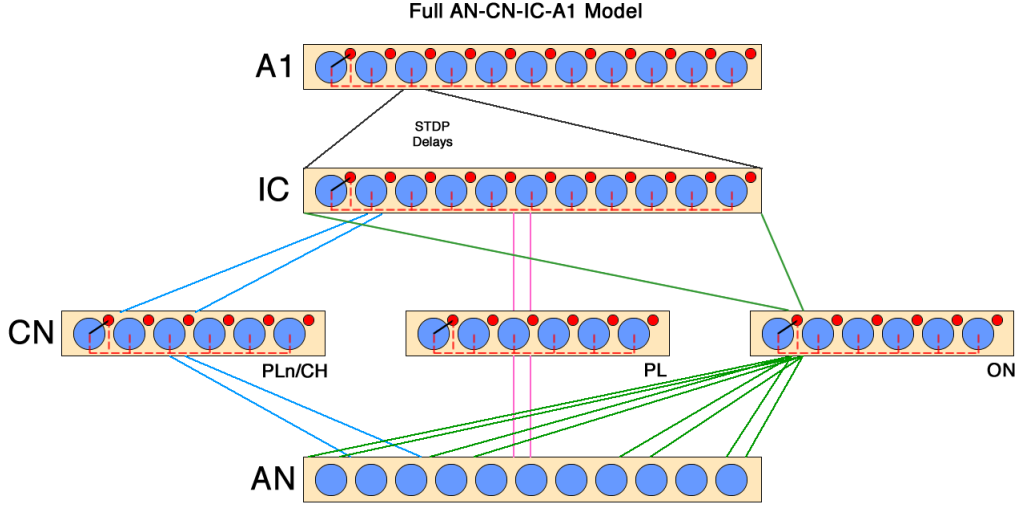


Figure 3.9: Schematic representation of the full AN-CN-IC-A1 model of the auditory brain trained to discriminate between two vowels /a/ and /i:/. Blue circles represent excitatory and red circles represent inhibitory neurons. The sparse connectivity between the excitatory and inhibitory cells (w_{ij}^{EI}) within all layers is modelled as one-to-one feed-forward connections from each excitatory cell to a corresponding inhibitory cell (shown in black), and each inhibitory cell is fully connected to all of the excitatory cells (w_{ij}^{IE}) within its layer (red connections). AN - auditory nerve; CN - cochlear nucleus with three sub-populations of cells discriminated based on their connectivity: primary-like with notch (PLn)/chopper (CH), primary-like (PL) and onset (ON); IC - inferior colliculus; A1 - primary auditory cortex. The connections involving PLn/CH cells are shown in blue (w_{ij}^{BL}). PLn/CH layer has Gaussian topological connectivity for both its afferent and efferent connections, whereby each PLn/CH cell receives afferents from a larger topological region of the AN ($\sigma = 26$), and each cell in the IC receives afferents from a smaller topological region of the PLn/CH CN ($\sigma = 2$). The one-to-one afferent and efferent connections involving PL cells are shown in pink (w_{ij}^{BL}). The connections involving ON cells are shown in green (w_{ij}^{BL}), whereby each ON cell receives afferent connections from a number of randomly distributed AN cells (sparsity=0.46), and each ON cell synapses to all IC cells. The IC and the A1 layers are fully connected using feed-forward connections with randomly initialised weights (w_{ij}^{BL}) and a distribution of axonal delays (Δ_{ij}). These feed-forward connections are modifiable through spike-timing dependent plasticity learning.

spiking of an individual neuron, because small differences in the precise timing of a spike are ignored when considering whether a neuron fires one or more spike throughout the duration of a stimulus. Information analysis, however, is sensitive to space jitter across neurons, since space jitter can affect on/off encoding of an individual cell.

The performance of the PLn/CH cells in space jitter removal was tested using a ten vowel recognition paradigm. The ten vowel paradigm was used instead of the two vowel paradigm, because it was found that the two vowel task described in Chapter 2 Section 2.2 was too easy to assess the space jitter removal performance of the PLn/CH layer, since 1 bit of information was reached for many parameter choices. Therefore, a different stimulus set was constructed using the Klatt synthesiser, whereby ten 100 ms long vowels with three exemplars each were synthesised. The formant frequencies were taken from Peterson & Barney (1952) and are shown in Table 2.1.

During testing the stimuli were presented in a similar fashion to that described in Chapter 2 Section 2.2.1 over 80 epochs. The number of testing epochs was increased from 20 for the two vowel paradigm to 80 for the ten vowel paradigm, to allow enough sampling to reduce information bias. The maximum amount of information necessary to perfectly differentiate between ten vowels is $\log_2 10 \approx 3.32$ bits. Figure 3.10 illustrates the outcome of the parameter search. The colour of each cell of the figure corresponds to the maximal single cell information reached within the CH/PLn area when the AN-CH/PLn connections were initialised using the different combinations of parameters sampled during the grid search as described above. It was found that the best parameter set reached 0.74 bits of information (Table 3.5). It is interesting to note that no within-layer inhibition ($w_{ij}^{IE} = 0$ nA) was necessary for the CH/PLn neurons to reach their best space jitter removal performance, which suggests that the CH/PLn layer of the model is not a competitive layer and instead it acts as an integrator of firing from topological regions of the AN.

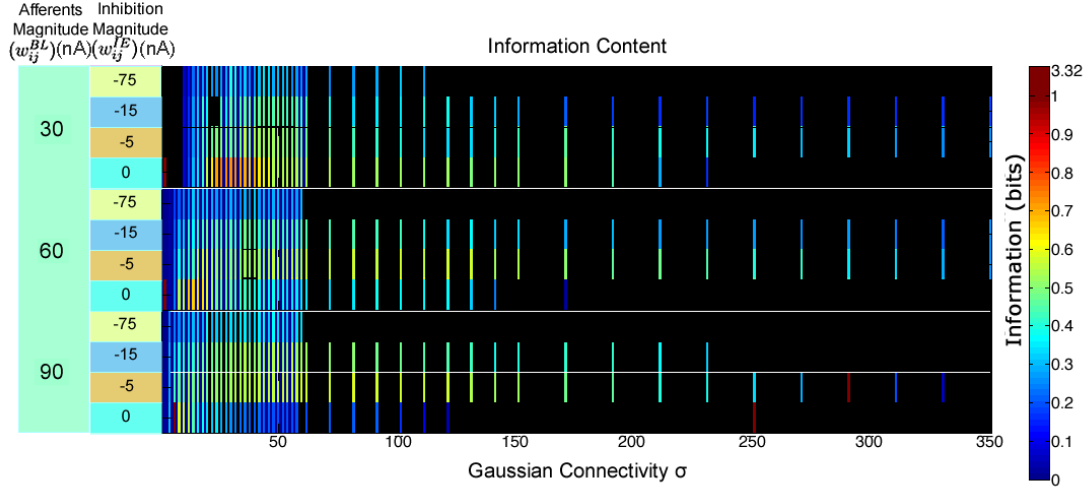


Figure 3.10: Parameter search results to determine the optimal topological Gaussian connectivity (σ) between the auditory nerve layer and the primary-like with notch (PLn)/chopper (CH) sub-population of the cochlear nucleus using ten vowel discrimination task. The maximum theoretically possible single cell information for discriminating between ten stimuli is 3.32 bits. The rows represent maximum single cell information content within the PLn/CH layer when the σ parameter of the Gaussian connectivity was varied, while keeping the strength of the afferents (w_{ij}^{BL}) and the magnitude of inhibition (w_{ij}^{IE}) within the PLn/CH layer fixed. The maximum amount of information of 0.74 bits was reached for the parameters shown in Table 3.5. Black indicates no data.

I→E Connection Strength, w_{ij}^{IE} (nA)	0
AN→PLn/CH Gaussian Connectivity σ	26
AN→PLn/CH Connection Strength, w_{ij}^{BL} (nA)	[30, 35]

Table 3.5: Primary-like with notch/chopper cell parameters found to maximise the amount of information about the identity of ten vowel stimuli.

3.8.2 Optimising Onset Cells

As mentioned in Section 3.6, onset cells receive afferent connections from a large number of AN fibers with different characteristic frequencies. Therefore, sparse connectivity was used between the AN layer and the onset layer, and the sparseness (the proportion of active synapses) was a parameter to be optimised. The onset layer was made of 100 Class 1 neurons as shown in Figure 2.12 (G) (Izhikevich 2003). The lower number of cells within the onset layer was chosen to match the neurophysiological findings that onset cells account for only 10% of all cells in the cochlear nucleus. Apart from the sparseness of afferent connections, other parameters searched were the strength of inhibition within the onset layer (w_{ij}^{IE}), and the strength of AN-to-onset connections (w_{ij}^{BL}). The performance of onset cells on the time jitter removal task was evaluated using the synchronisation index described in Chapter 2 Section 2.10.2. The layer was tested in response to the two vowel testing paradigm as described in Chapter 2 Section 2.2. Simulations where onset cells fired less than once per stimulus presentation were excluded from analysis. The results of the parameter search are illustrated in Figure 3.11, which reveals that the best time jitter removal performance as quantified by the synchronisation index of 0.79 was achieved with the parameters shown in Table 3.6.

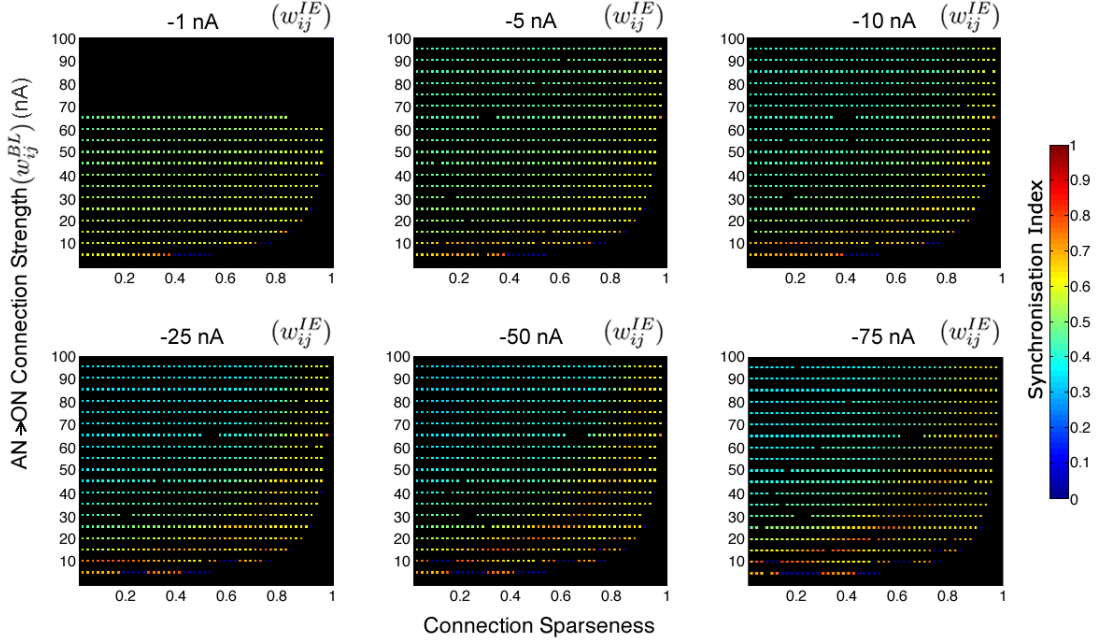


Figure 3.11: Parameter search results to determine the best connectivity between the auditory nerve layer and the onset sub-population of the cochlear nucleus. The synchronisation index ranging from 0 to 1 was calculated to measure the layer's performance on the two vowel discrimination task. Each sub-plot represents one part of the parameter search, whereby the magnitude of inhibition within the onset layer (w_{ij}^{IE}) was fixed to the value stated above the plot, while the sparseness of AN→ON connections (abscissa) and the magnitude of AN→ON connections (w_{ij}^{BL}) (ordinate) varied. The sparseness value indicates the proportion of AN→ON synapses set to 0. The maximum synchronisation index of 0.79 was reached for the parameters shown in Table 3.6. Black indicates no data.

I→E Connection Strength, w_{ij}^{IE} (nA)	-75
AN→ON Connection Sparseness	0.46
AN→ON Connection Strength, w_{ij}^{BL} (nA)	21

Table 3.6: Onset cell parameters found to maximise the synchrony index in the onset layer in response to the two vowel discrimination task. Afferent connection sparseness indicates the proportion of all synapses between the auditory nerve and the onset layers set to 0.

3.8.3 Neurophysiological Validity of the Cochlear Nucleus Layer

In order to verify that the CN layer of the full AN-CN-IC-A1 model resembled a real CN, the response properties of the three sub-populations of the model CN were compared to the corresponding neurophysiological data.

3.8.3.1 Primary-Like Cells

Harmonic spectra of period histograms were plotted for the real and model PL fibers (Fig. 3.12), illustrating the variation in neural responses as a function of both the fiber characteristic frequency (CF) and the vowel formant frequencies. Similarity between the real and model PL fibers' response properties suggests that the model's performance is comparable to the neurophysiological data.

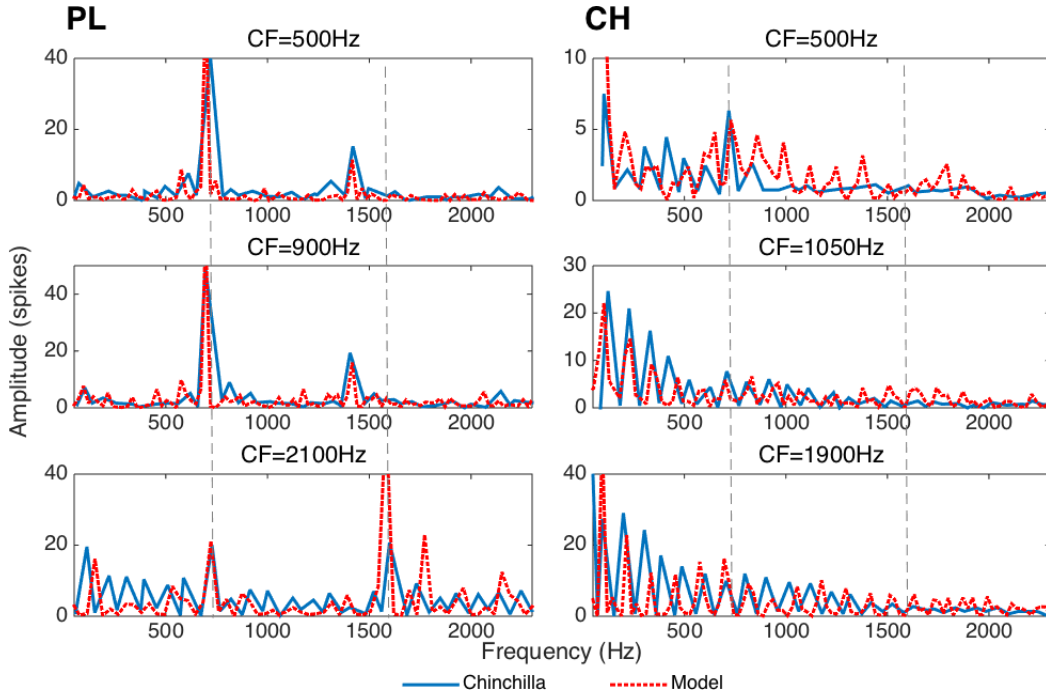


Figure 3.12: Spectra (computed as Fast Fourier Transforms of period histograms) of primary-like (PL) (left column) and chopper (CH) (right column) cochlear nucleus neuron responses to a synthetic vowel /a/ generated using the Klatt synthesiser (Klatt 1980). The ordinate represents the level of phase-locking to the stimulus at frequencies shown along the abscissa. Dotted lines show the positions of the vowel formant frequencies F_1 and F_2 . Data from chinchilla CN fibers reproduced from Recio & Rhode (2000) is shown in solid blue. Data collected from the corresponding model CN fibers is shown in dashed red. Similarity between the real and model fibers response properties suggests that the models performance is comparable to the neurophysiological data.

3.8.3.2 Primary-Like With Notch/Chopper Cells

After fine tuning the parameters of the primary-like with notch (PLn)/chopper (CH) sub-population within the CN layer of the full AN-CN-IC-A1 model for the optimal performance on the space jitter removal task, the response properties of the model PLn/CH neurons were compared to the response properties of the real CH and PLn cells in order to determine whether they resembled real PLn or CH cells more.

First, PLn/CH neurons' response properties to a synthetic vowel /a/ were compared to the real CH fibers' response properties as reported by Recio & Rhode (2000) (Fig. 3.12). Second, the responses of a PLn/CH layer cell were compared to its neurophysiological PLn equivalent recorded from the cat CN in response to

a 670 Hz pure tone as reported by Smith et al. (1991) (Fig. 3.13). Similarities between the real CH and model PLn/CH fibers' response properties, and dissimilarity between the real PLn and model PLn/CH fibers' response properties suggest that the parameter search has resulted in more CH-like response properties within the PLn/CH layer. This layer will, therefore, be referred to as the CH layer from now on.

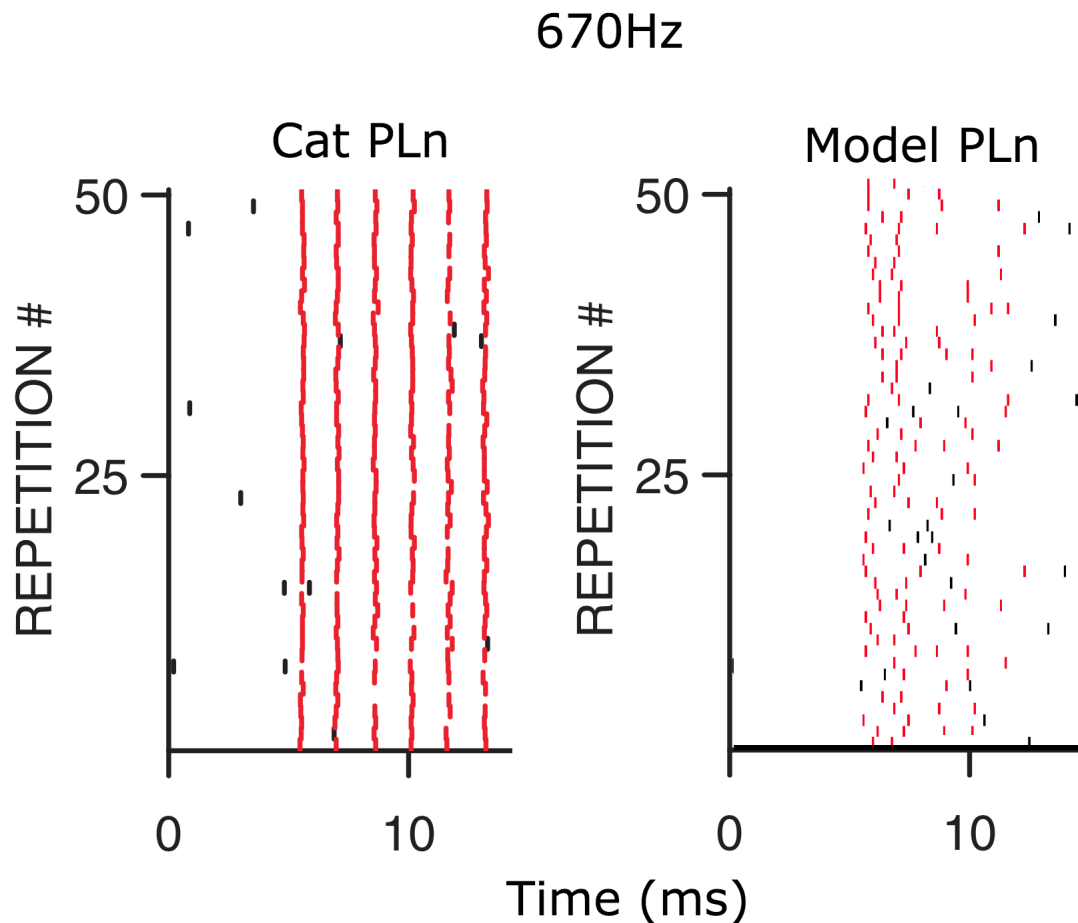


Figure 3.13: Cell firing rasters for a cat PLn cell (Joris & Smith 2008) and a model PLn/CH cell in response to 50 presentations of a 670 Hz pure tone. Red spikes indicate that at least two spikes occurred within $50\mu\text{s}$ of each other in the raster. It can be seen that the response properties of the real and model cells seem to differ - the real cell shows strong phase-locking to the stimulus, which is not present for the model cell.

3.8.3.3 Onset Cells

Since onset (ON) CN cells are relatively rare (about 10% of all CN neurons), there is little neurophysiological data available for direct comparison with the data collected from the model. Nevertheless, neurophysiological data suggests that ON neurons phase-lock to the vowel fundamental frequency (F_0) (Winter et al. 2003). A periodogram of model ON neurons' response properties to vowel /a/ is shown in Figure 3.14. It can be seen that most neurons strongly synchronise to the vowel's F_0 component of 100 Hz, therefore suggesting that the response properties of the model ON cells are comparable to the neurophysiological data.

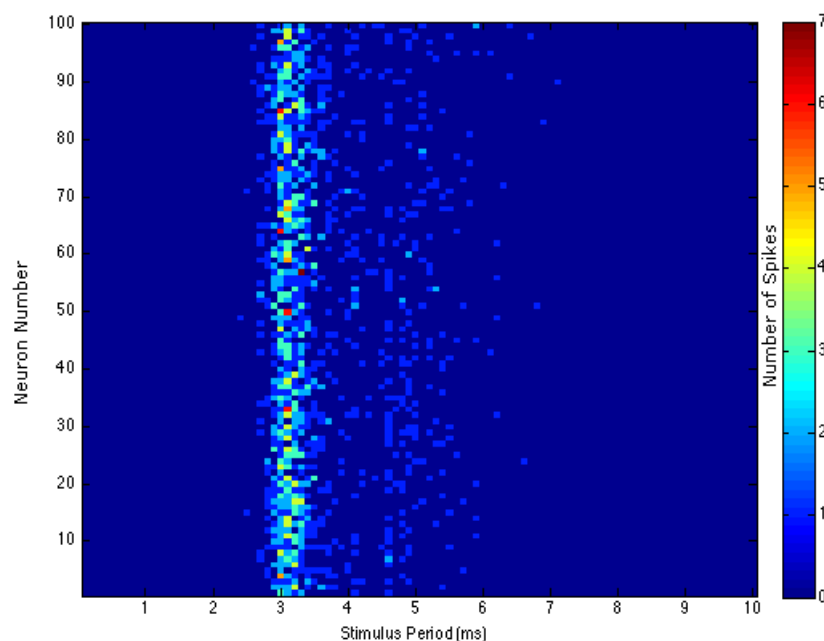


Figure 3.14: Periodogram of model onset fibers in response to a synthetic vowel /a/ (see Table 2.1). The abscissa represents time in ms of a stimulus period. The ordinate represents the population of onset (ON) cells within the cochlear nucleus. Since the fundamental frequency (F_0) of the vowel is 100 Hz, each period is 10 ms long. The periodogram shown is generated using a 400 ms long presentation of vowel /a/ to the network. It can be seen that ON cells phase-lock to the F_0 of the vowel and fire at a regular interval of approximately 10 ms long.

3.8.4 Cochlear Nucleus to Inferior Colliculus Connectivity

The excitatory feed-forward connections between the three different sub-populations of the CN (PL, CH and ON) and the IC were chosen in a heuristic manner, since a parameter search would have been too computationally expensive. The connections from the primary-like neurons up to the IC cells were modelled to be one-to-one, similar to the primary-like afferent connections from the AN.

The role of onset cells is hypothesised to be time jitter removal through sub-threshold activation of all IC cells at times phase-locked to the fundamental frequency of vowels. Therefore, full connectivity between each onset cell and all IC cells was chosen as the most appropriate connectivity scheme.

The connectivity between chopper (CH) and the IC layers was chosen to be Gaussian topological (same as the AN to CH connectivity), however σ was set to 2. A narrow Gaussian was used to maintain the space jitter removal properties of CH cells in the IC. While one-to-one connectivity would have preserved CH outputs perfectly, it was not used, because there is no neurophysiological evidence for one-to-one connectivity between CH cells and IC cells.

3.8.5 Optimising STDP and Finding Minimal Layer Set for Full AN-CN-IC-A1 Model

Once the optimal parameters for the CH and ON layers were found, the performance of the full AN-CN-IC-A1 network was evaluated on the same two vowel recognition paradigm as the simple two-layer AN-A1 model (Chapter 2 Section 2.2). Certain model parameters were varied to find the architecture with the best performance on the task as quantified by single cell information analysis of neurons in the output A1 layer. The parameters modified were the afferent connection weights to the IC layer from the CH, ON and PL sub-populations within the CN layer (w_{ij}^{BL}), the strength of inhibition within the IC layer (w_{ij}^{IE}), the strength of afferent connections from the IC to the A1 layer (w_{ij}^{BL}), and the STDP parameters for the modifiable connections between IC and A1 (τ_p , τ_d , α_p and α_d) (see Table 3.4 and Figure 3.9).

During the parameter search the relative strength of the afferent connections from the three CN sub-populations (PL, CH and ON) to the IC (w_{ij}^{BL}) were varied from 0 up to a level found to be adequate through exploratory testing, whereby the maximum level of ON afferents was set to the maximum sub-threshold level, while the maximum levels of PL and CH afferents were set to the level that transcribed their firing onto the IC layer in the presence of no other input. The parameter search, therefore, investigated both the optimal STDP parameters, and the optimal levels of convergence of the three sub-populations of the CN on the IC.

The parameter search revealed that the full AN-CN-IC-A1 model performed

significantly better on the two vowel recognition task than the best simple AN-A1 model described in Section 3.2, whereby a full 1 bit of information was reached (see Fig. 3.15).

The optimal parameters for the full AN-CN-IC-A1 model are shown in Table 3.7. It can be seen that the network reaches its best performance when no within layer inhibition is present in the IC layer ($w_{ij}^{IE} = 0$ nA), which suggests that the IC layer is not a competitive layer and instead it acts as an integrator of firing from the CN sub-populations. Neurophysiological evidence suggests that within-layer inhibition exists in the IC, and that different neuron types emerge based on the shape of the inhibitory sidebands of their tuning curves (Egorova et al. 2001, Gonzalez-Hernandez et al. 1996, Vater et al. 1992). This literature also suggests that some IC neuron types in the brain may be less affected by inhibition than others, and it may be that those neurons are the ones being modelled the closest by the IC neurons of the full AN-CN-IC-A1 model.

It can be seen in Table 3.7 that all three CN sub-populations - the primary-like (PL→IC), the chopper (CH→IC) and the onset neurons (ON→IC) - have non-zero afferent connection strength magnitudes (w_{ij}^{BL}). This supports the hypothesis that all three sub-populations of the CN are necessary to remove space and time jitter and enable STDP-based learning. In order to verify that all of these sub-populations are necessary to enable the network to reach full information about the two vowel stimuli, each of the three CN sub-populations were knocked out one by one, the network was re-optimised and the change in the maximum single cell information within the best performing network output A1 layer was measured (see Table 3.8). It can be seen that knocking any part of the CN out results in a significant drop in maximum single cell information within the A1 layer, hence supporting the hypothesis that all three sub-populations are important and necessary for invariant auditory object learning.

It is interesting to note that information reaches the theoretical maximum of 1 bit in the IC layer but drops in the A1 layer of the full AN-CN-IC-A1 model when either PL or PL and ON sub-populations of the CN are knocked out (rows 3-4 in Table 3.8). As mentioned in Section 3.8.1 the two vowel recognition task was too easy for the CH layer, hence the full 1 bit of information was reached by CH neurons. Since CH→IC afferent connections were strong enough to transfer the firing pattern to the IC, a full 1 bit of information was achieved by the IC when PL and ON sub-populations of the CN were knocked out. The reason why information content dropped in the A1 level is because without the ON sub-population time jitter was not removed, and therefore A1 learning from the IC input was impaired. In the condition when only the PL sub-population was knocked out, both space and time jitter was removed in the IC layer due to the presence of afferents from both the CH and ON sub-populations of the CN. A full 1 bit of information

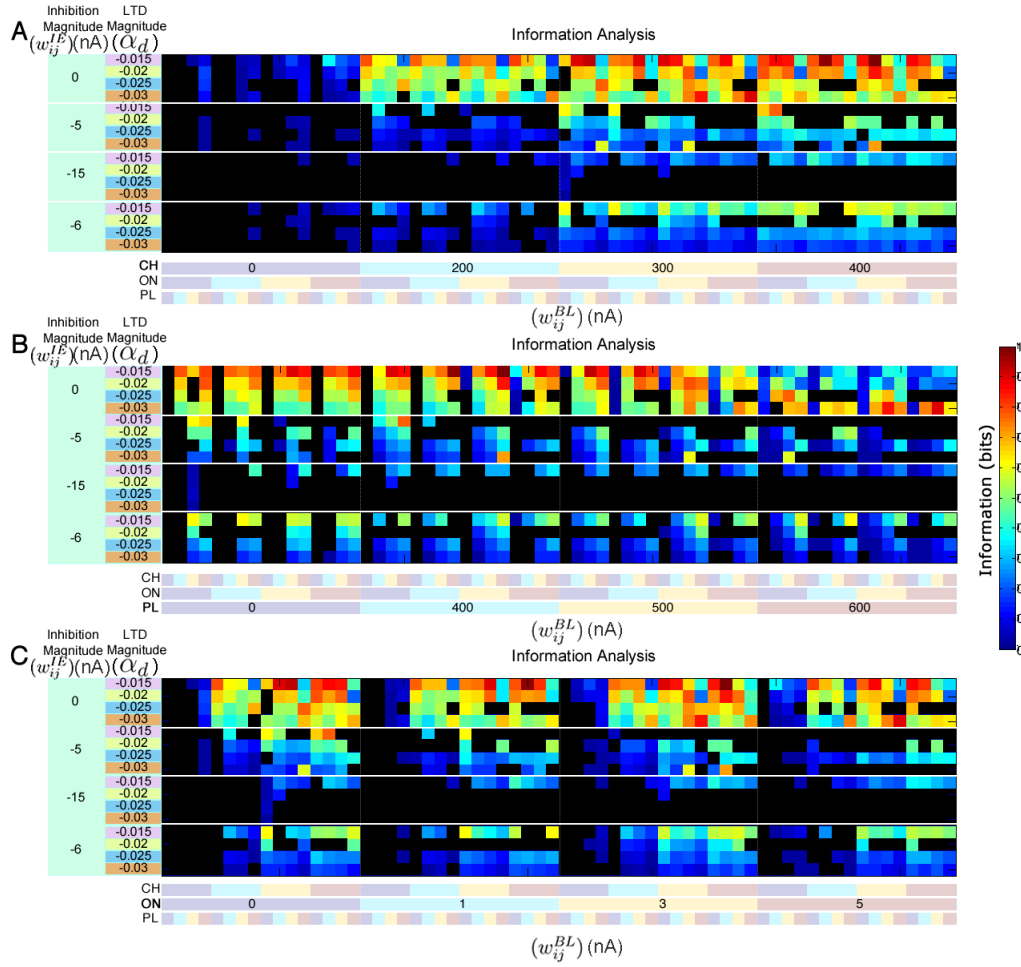


Figure 3.15: Parameter search results to determine the best connectivity between the three sub-population of the cochlear nucleus (CN) - the chopper (CH), onset (ON) and primary-like (PL) neurons - and the inferior colliculus (IC) (w_{ij}^{BL}). The parameter search also investigated the best level of inhibition within the IC (w_{ij}^{IE}), and the appropriate level of LTD (α_d). The two vowel discrimination task was used, with maximum theoretically possible single cell information of 1 bit. The performance of the model was assessed by the maximal amount of single cell information reached within the output A1 layer. Sub-plots **A-C** show different sorted orders of the same data to illustrate the effect of varying the afferents' strength from the CH, PL and ON layers. The rows represent maximum single cell information content within the A1 layer when the relative strength of afferents (w_{ij}^{BL}) of the three CN sub-populations were varied while keeping the magnitude of inhibition within the IC (w_{ij}^{IE}) and the LTD magnitude between the IC and A1 layers (α_d) fixed. **A**: data sorted by the magnitudes of CH-IC afferents (w_{ij}^{BL}); **B**: data sorted by the magnitudes of PL-IC afferents (w_{ij}^{BL}); **C**: data sorted by the magnitudes of ON-IC afferents (w_{ij}^{BL}). Transparent plots at the bottom of each heat plot indicate the strength of afferents (w_{ij}^{BL}) from the CH, ON and PL layers for each simulation. The maximum amount of information of 1 bits was reached for the parameters shown in Table 3.7. Black indicates no data due to restricted sampling.

IC-A1 LTD Magnitude, α_d	-0.015
IC I→E Connection Strength, w_{ij}^{IE} (nA)	0
CH→IC Connection Strength, w_{ij}^{BL} (nA)	400
ON→IC Connection Strength, w_{ij}^{BL} (nA)	3
PL→IC Connection Strength, w_{ij}^{BL} (nA)	400

Table 3.7: Parameters found to maximise the amount of information about the identity of two vowel stimuli within the full AN-CN-IC-A1 model of the auditory brain shown in Figure 3.9. The four layers are: AN - auditory nerve; CN - cochlear nucleus with three sub-populations of cells discriminated based on their connectivity: chopper (CH), primary-like (PL) and onset (ON); IC - inferior colliculus; A1 - primary auditory cortex.

was reached within the IC due to strong afferent connections from the CH layer, which was not disturbed by noise from the PL layer. The reason why there was a reduction in information content within the A1 layer is because without the additional input from the PL sub-population there was nothing to provide the A1 layer with a background level of activation, and therefore the A1 neurons simply failed to fire to certain stimulus presentations, hence resulting in reduced learning. Stronger initialised IC→A1 afferents (w_{ij}^{BL}) could potentially remove the need for the PL input. Informal tests and simple AN-A1 model parameter search (Fig. 3.7), however, showed that IC→A1 afferents initialised from a [30, 35] nA uniform distribution (w_{ij}^{BL}) produced the best network performance.

It is interesting to note that the amount of information retained by the model depended significantly on CH cells, with the average amount of information increasing with the increasing CH afferent connection strength (CH→IC w_{ij}^{BL}) (see Fig. 3.15 (A)). This is hypothesised to be due to the fact that CH pre-processing results in the highest increase in information content compared to that of the AN fibers out of all three sub-populations of the CN. Stronger CH→IC afferent connections, therefore, transfer more of this information to the IC layer. Figure 3.15 (B) suggests that the opposite is true for the primary-like population, whereby the stronger their afferents are (PL→IC w_{ij}^{BL}), the less information is retained by the network. This could be because stronger PL→IC afferents re-introduce too much noise to the IC firing. Some input from the primary-like population is nevertheless necessary for learning. There seems to be no linear correlation between the onset afferents' strength (ON→IC w_{ij}^{BL}) and the amount of information retained by the network (Fig. 3.15 (C)). Instead it seems like the network performance increases as the sub-threshold level of activation from the ON sub-population grows up to a particular level (3 nA for the ON→IC afferent connections (w_{ij}^{BL})), with further increase in the strength of the ON→IC afferent connections resulting in reduced

<i>Chopper</i>	<i>Onset</i>	<i>Primary-Like</i>	<i>IC Information (bit)</i>	A1 Information (bit)
✓	✓	✓	<i>0.87</i>	1
✓	-	✓	<i>0.92</i>	0.93
✓	✓	-	<i>1</i>	0.89
✓	-	-	<i>1</i>	0.81
-	✓	✓	<i>0.64</i>	0.36
-	-	✓	<i>0.18</i>	0.18
-	✓	-	<i>0</i>	0

Table 3.8: Maximum single cell information within the output A1 layer of the best performing re-optimised full AN-CN-IC-A1 model when different CN sub-populations of neurons (chopper, onset or primary-like) are selectively knocked out. Theoretical maximum for the single cell information measure for two stimuli is 1 bit.

information within the A1 level.

3.9 Simple Four-Layer Model

In order to confirm that jitter removal done by chopper (CH) and onset layers followed by the integration within the IC, rather than the pure addition of extra layers, was the reason why the four layer AN-CN-IC-A1 network performed better than the two layer AN-A1 network, an extra simulation study was performed. A new simple four-layer fully connected model without the detailed CH, ON, PL micro-structure converging on the IC was tested.

The new simple four-layer model was initialised with randomly initiated weights (w_{ij}^{BL}) and delays (δ_{ij}) between each of the layers, and with STDP-based learning for the Layer 3→A1 connections (see Fig. 3.16 and Table 3.9). Since parameter searches for both the simple two-layer model and the complex four-layer model revealed that similar STDP parameters result in best performance of both networks, an average of these parameters was used for the simple four-layer model (see Table 3.10). The reason why STDP-based learning was used only between the last two layers of the simple four-layer model is because learning was only implemented for the IC→A1 connections (w_{ij}^{BL}) in the full AN-CN-IC-A1 network described in Section 3.7. The magnitudes of between layer connections (w_{ij}^{BL}) were chosen to ensure that the rate of firing in the third layer is similar to that of the IC layer of the full AN-CN-IC-A1 model (approximately 9 Hz). Such weight adjustment to ensure the correct firing rate within a layer was implemented in previous modelling work, and was named activity dependent scaling (Evans & Stringer 2012, van Rossum et al. 2000).

Simple Four-Layer Model Parameters

<i>Within Layer Parameters</i>				
Corresponding Brain Area	Layer 1	Layer 2	Layer 3	Layer 4
	Auditory Nerve	N/A	N/A	A1
Excitatory Cell Number	1000	1000	1000	1000
Inhibitory Cell Number	0	1000	1000	1000
E→I Connection Strength, w_{ij}^{EI} (nA)	N/A	1000	1000	1000
I→E Connection Strength, w_{ij}^{IE} (nA)	N/A	-6	-6	-6
<i>Between Layer Parameters</i>				
	Layer 1-2	Layer 2-3	Layer 3-4	
Axonal Delays, Δ_{ij} (ms)	[0, 50]	[0, 50]	[0, 50]	
Connectivity	full	full	full	
Connection Strength, w_{ij}^{BL} (nA)	[20, 25]	[12, 17]	[30, 35]	
Spike Rule	N/A	N/A	nearest	
STDP Rule	N/A	N/A	mixed	
LTP Time Constant, τ_p (ms)	N/A	N/A	15	
LTD Time Constant, τ_d (ms)	N/A	N/A	25	
LTP Magnitude, α_p	N/A	N/A	0.005	
LTD Magnitude, α_d	N/A	N/A	-0.017	
Maximum Synaptic Strength, w_{max} (nA)	N/A	N/A	60	

Table 3.9: Parameters used for the simple four-layer model of the auditory brain used to test whether the mere presence of extra pre-cortical layers helped the four-layer AN-CN-IC-A1 model described in Section 3.7 learn to discriminate between two vowels /a/ and /i:/.

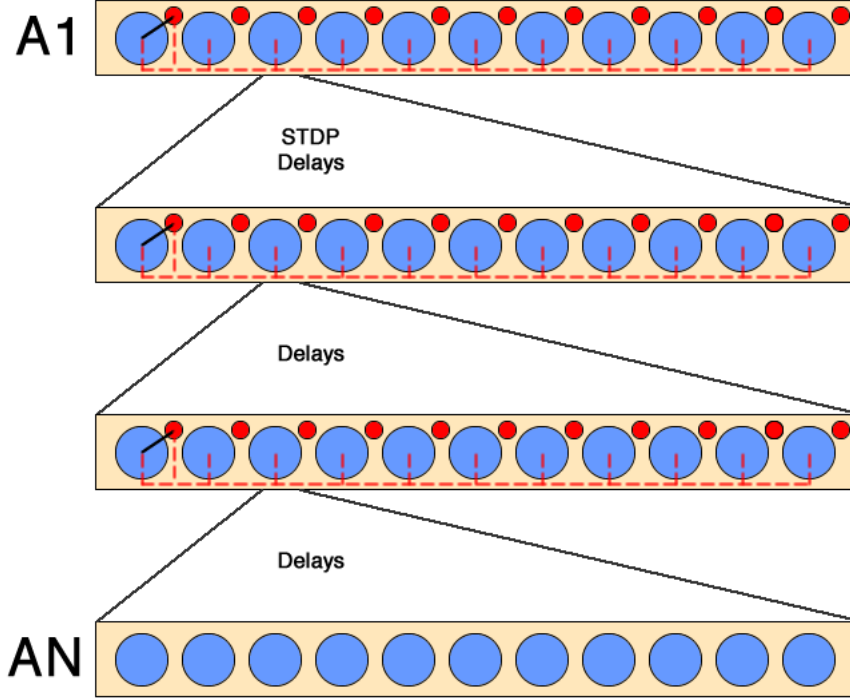


Figure 3.16: Schematic representation of the simple four layer model of the auditory brain. This architecture was used to test whether the mere presence of extra pre-cortical layers, that is without necessarily having the complex CH, ON, PL micro-structure converging into the IC, helped the four-layer AN-CN-IC-A1 model described in Section 3.7 learn to discriminate between two vowels /a/ and /i:/. Blue circles represent excitatory and red circles represent inhibitory neurons. The sparse connectivity between the excitatory and inhibitory cells (w_{ij}^{EI}) within all layers is modelled as one-to-one feed-forward connections from each excitatory cell to a corresponding inhibitory cell (shown in black), and each inhibitory cell is fully connected to all of the excitatory cells within its layer (w_{ij}^{IE}) (red connections). Each of the four layers are fully connected using feed-forward connections with randomly initialised weights (w_{ij}^{BL}) and a distribution of axonal conduction delays (Δ_{ij}). Feed-forward connections between the last two layers are modifiable through spike-timing dependent plasticity learning.

	AN-A1 Model	AN-CN-IC-A1 Model	Simple 4-Layer Model
Spike Rule	nearest	nearest	<i>nearest</i>
STDP Rule	mixed	mixed	<i>mixed</i>
LTP Time Constant, τ_p (ms)	15	15	<i>15</i>
LTD Time Constant, τ_d (ms)	25	25	<i>25</i>
LTP Magnitude, α_p	0.005	0.005	<i>0.005</i>
LTD Magnitude, α_d	-0.02	-0.015	<i>-0.017</i>

Table 3.10: Optimal STDP parameters for the simple AN-A1 model, the full AN-CN-IC-A1 model, and the simple 4-layer model of the auditory brain. Since the parameters for the AN-A1 and AN-CN-IC-A1 models are very similar, the same STDP parameters were assumed to be optimal for a simple four-layer model. The small discrepancy in the LTD magnitude (α_d) between the AN-A1 and the AN-CN-IC-A1 models was solved by taking the average of the two values as LTD magnitude for the simple four-layer model.

It can be seen that a simple four-layer network achieves very little information, i.e. 0.28 bits maximum (see Fig. 3.6). This suggests that purely adding extra layers does not help with invariant auditory object representation learning. This further supports the hypothesis that space and time jitter reduction, as performed by the CH and ON neurons in the cochlear nucleus, and finished through the integration within the IC, is necessary for such learning to occur.

4

The Role of Subcortical Processing in Auditory Category Learning: Results

4.1 Introduction

Two models of the auditory brain were described in Chapter 3: the simple AN-A1 model and the full AN-CN-IC-A1 model with extra sub-cortical pre-processing. Both models were evaluated on the same two vowel discrimination task, and after extensive parameter searches it was shown that the simple AN-A1 model does not achieve more than 0.6/1 bits of information, while the full AN-CN-IC-A1 model reaches the full 1/1 bits. It was hypothesised that this was due to the disruptive nature of jitter present within the AN, which impedes continuous transformation (CT) STDP-based learning in building transform invariant representations of the two vowel stimuli in the simple AN-A1 model. It was hypothesised that the two sub-populations of the cochlear nucleus (CN) present in the full AN-CN-IC-A1 model, the chopper (CH) and onset (ON) neurons, and the specialised integration of their afferent firing input within the inferior colliculus (IC) help remove AN jitter and therefore enable CT STDP-based learning .

This chapter investigates the role of the CN and IC in enabling transform invariant learning of auditory objects. First, the firing patterns within the IC are compared to those within the AN to confirm jitter reduction produced by the sub-cortical pre-processing within the full AN-CN-IC-A1 model. Then the response properties of the CH and ON sub-populations of the CN layer of the AN-CN-IC-

	/i:/		/a/	
	AN	IC	AN	IC
Same Exemplar Index	0.45	0.90	0.57	1
Different Exemplars Index	0.52	0.91	0.63	1
Different Categories Index	0.42	0.67	0.42	0.67

Table 4.1: Same Exemplar, Different Exemplars and Different Category Indices measuring space and time jitter within the auditory nerve (AN) and inferior colliculus (IC) layers of the full AN-CN-IC-A1 model of the auditory brain.

A1 model of the auditory brain are analysed in more detail. This is followed by the analysis of the performance of the simple AN-A1 and the full AN-CN-IC-A1 models to understand how information encoding changes throughout the different layers of the models and how the different response properties in their output layers emerge. The effect of variations in certain parameters on the network performance is also discussed.

4.2 Analysing Space and Time Jitter Removal in the Cochlear Nucleus

In order to test whether the optimised full AN-CN-IC-A1 spiking neural network model of the auditory brain described in Chapter 3 Section 3.7 was able to learn transform invariant representations of two vowels, /i:/ and /a/, due to space and time jitter removal by the chopper and onset sub-populations within the cochlear nucleus followed by further pre-processing through the integration of the afferent signals from the CN within the inferior colliculus (IC), jitter scores were compared between the input auditory nerve (AN) and the IC layers (see Chapter 2 Section 2.10.3 for the description of jitter measurement). Table 4.1 demonstrates that the amount of jitter is almost completely reduced in the IC layer, whereby IC rasters in response to the different presentations of the same vowel exemplar, or in response to different exemplars of the same vowel, contain spikes that almost always appear in the same place and at the same time. Furthermore, it can be seen in Figure 4.1 that the overlaid spike rasters of all presentations of all exemplars for each of the two vowel stimuli are qualitatively similar for the AN and the IC layers, however the AN versions are more blurry than the IC versions, indicating the presence of substantial jitter in the AN.

The synchrony index was also calculated for the different layers within the network (see Fig. 4.2). It can be seen that synchrony is increased throughout the

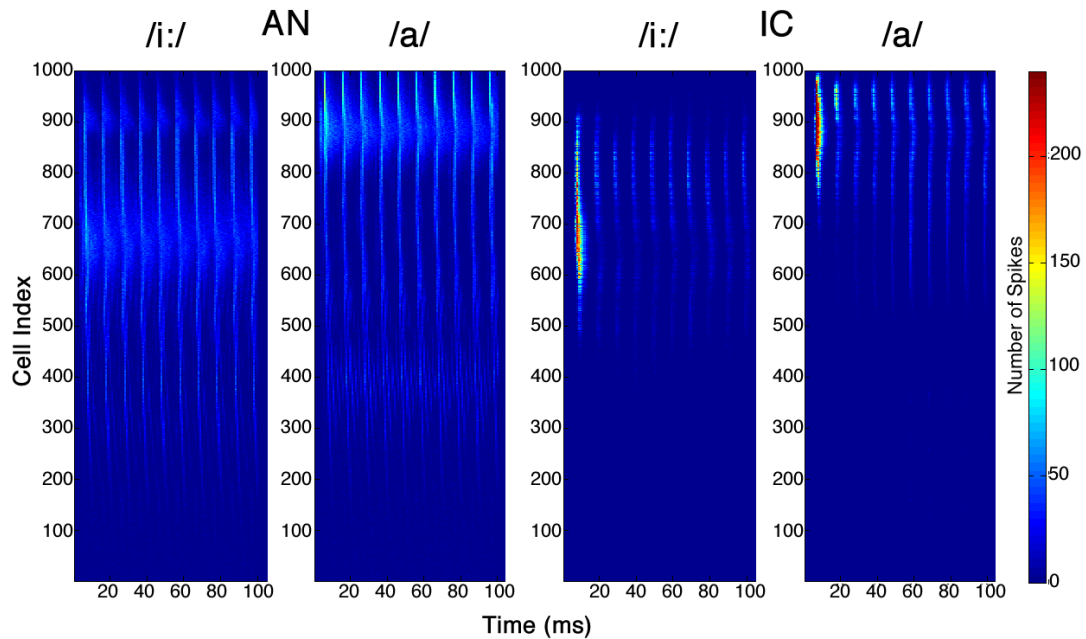


Figure 4.1: Spike rasters of AN and IC layers of the full AN-CN-IC-A1 model of the auditory brain in response to the two vowel discrimination task. Each raster shown is the sum of rasters in response to 20 presentations of all 12 exemplars for each vowel. The maximum number of spikes within each 1 ms time bin is therefore 240. It can be seen that the IC rasters are less noisy compared to the AN rasters. Furthermore, some bins in the IC rasters contain up to 240 spikes, thus providing evidence for the presence of stable jitter free spikes within the IC layer.

hierarchy of the auditory model, suggesting time jitter removal.

4.3 Understanding the Role of Chopper Cells in Space Jitter Removal

In order to understand the role of chopper cells better, the following analysis was performed. Bar plots were drawn for each cell in the AN and in the CH layers, whereby each bar indicated the cell's rate of firing in response to each of the 12 exemplars of each of the two vowel stimuli averaged over 20 presentation epochs. Figures 4.3 and 4.4 show that while cells in the AN layer tend to fire to both stimuli, which is also reflected in poor information analysis scores, the behaviour of chopper neurons is different. The responses of CH cells change from not firing to either vowel, to selectively firing to one or the other vowel. It appears that chopper

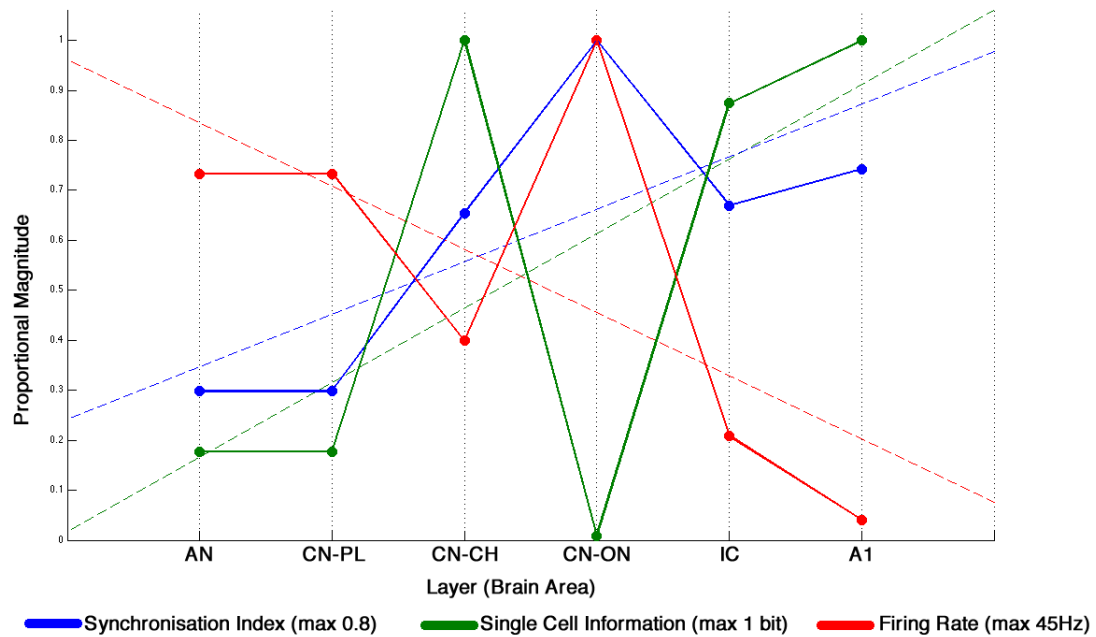


Figure 4.2: Graph depicting changes in three metrics throughout the hierarchy of the full four-layer AN-CN-IC-A1 spiking neural network model of the auditory brain. The three metrics are: the amount of maximum single cell information (green), synchronisation index (blue) and average firing rate (red) in response to the twelve exemplars of each of the two vowel stimuli. The regression lines (dashed) show that synchronisation and the maximal amount of single cell information increase throughout layers, while the average firing rate decreases. This is in line with neurophysiological findings (Ito & Fukushima 1988, Kurogi 1991, Steinschneider et al. 1990). CH, ON and PL are parallel sub-populations of the CN. ON information is 0 bits, since it does not carry any information about the vowel class identity. Instead ON cells purely phase-lock to the vowel F_0 . PL sub-population essentially transcribes the AN, hence its information content is identical to AN.

cells amplify strong responses and dampen weak responses within the AN, thus acting like highpass filters. This happens because chopper cells receive tonotopic AN afferents, have high firing thresholds, and act like coincidence detectors. These properties ensure that chopper cells only fire when a large number of neighbouring AN cells discharge synchronously, which happens when AN fibers respond to vowel formants, which have relatively high amplitudes, and thus result in high rates of firing in the corresponding parts of the AN. Therefore, the chopper layer transforms areas of high firing within the AN into regular spike trains of a small subset of chopper cells.

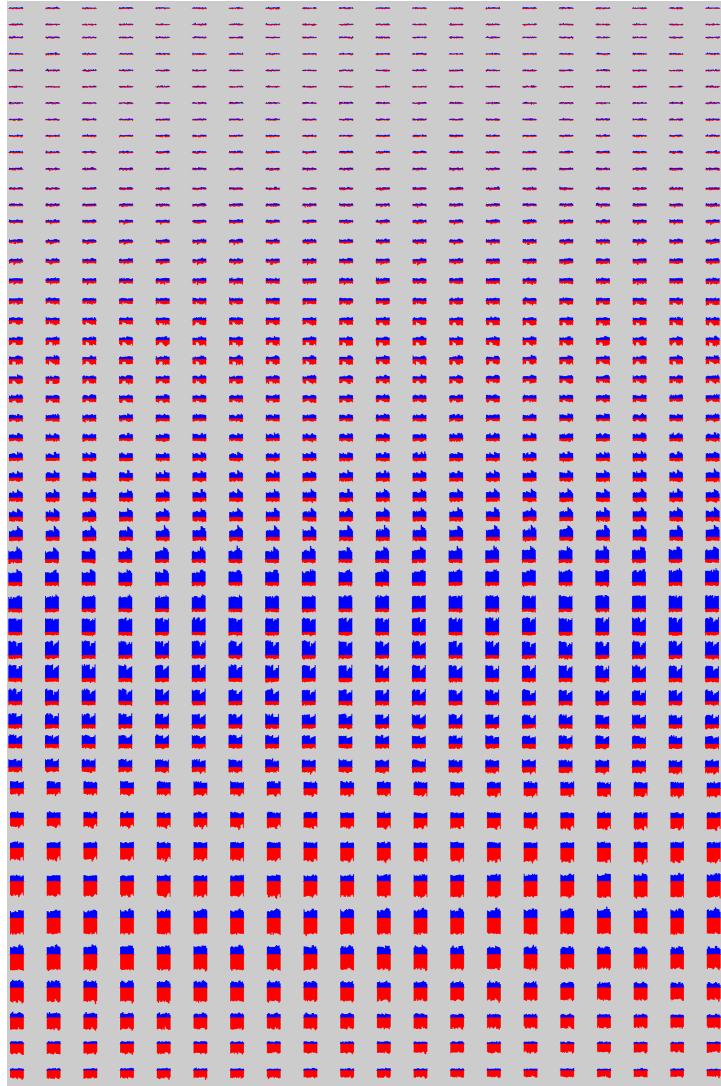


Figure 4.3: Average firing responses of 1000 AN fibers of the full AN-CN-IC-A1 model of the auditory brain to twelve exemplars of each of the two vowel stimuli, /a/ and /i:/. Each bar plot represents the response properties of one AN cell. Each bar plot contains twelve bars above the abscissa and twelve bars below the abscissa. The magnitude of each bar represents the average number of spikes elicited by the cell when presented with each of the twelve corresponding exemplars of the two vowels. Red bars represent exemplars of /a/ and blue bars represent exemplars of /i:/. It can be seen that while some cells prefer one stimulus over the other by responding on average more to its exemplars, no cell appears to respond to only one stimulus preferentially, i.e. by not responding to any of the exemplars of the other stimulus.

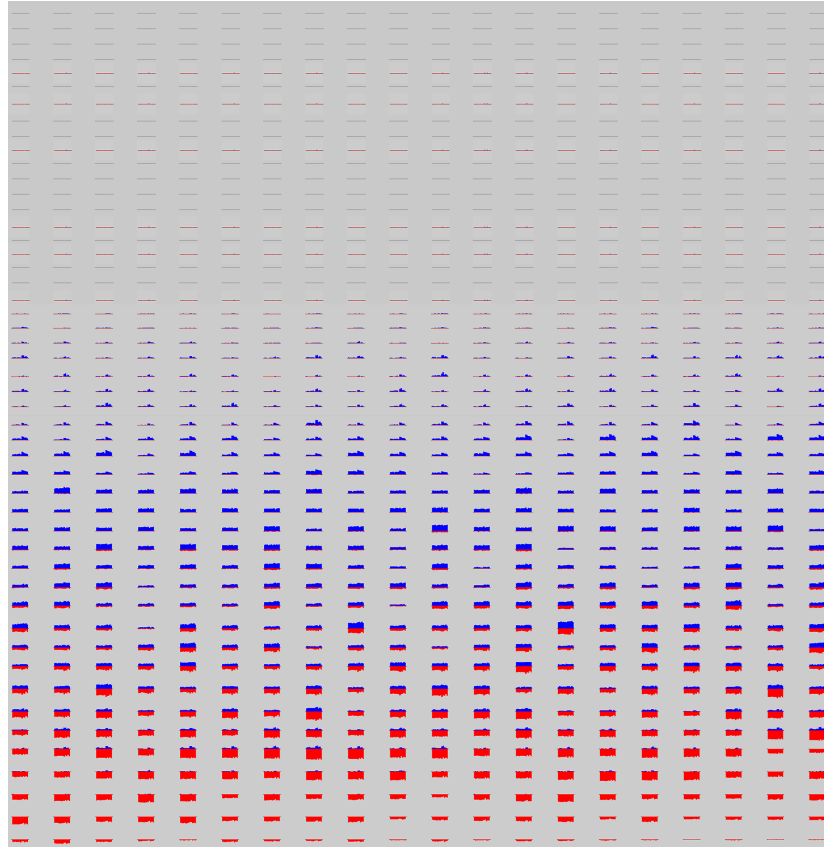


Figure 4.4: Average firing responses of 1000 chopper cells within the cochlear nucleus layer of the full AN-CN-IC-A1 model of the auditory brain to twelve exemplars of each of the two vowel stimuli, /a/ and /i:/. Each bar plot represents the response properties of one chopper cell. Each bar plot contains twelve bars above the abscissa and twelve bars below the abscissa. The magnitude of each bar represents the average number of spikes elicited by the cell when presented with each of the twelve corresponding exemplars of the two vowels. Red bars represent exemplars of /a/ and blue bars represent exemplars of /i:/. It can be seen that while some cells do not fire at all (dotted lines), and some fire to both stimuli, there are a number of cells that fire to only one stimulus preferentially, whereby not responding to any of the exemplars of the other stimulus. Such cells exist for both of the two vowel stimuli.

4.4 Comparing AN-A1 and AN-CN-IC-A1 Models

4.4.1 Change in Information Due to Training

It has been shown that the full AN-CN-IC-A1 auditory brain model with CN and IC layers between the input AN and the output A1 layers is able to reach the maximum of 1 bit single cell information when presented with twelve exemplars of each of the two vowel stimuli, /a/ and /i:/. The simple AN-A1 network was not able to match this performance, reaching approximately 0.57 bits of information at most. In order to verify that the performance of either of these two models was the result of training, rather than inherent network properties, single cell information was calculated for the un-trained models in response to the same stimuli. Figure 3.6 illustrates that training significantly increases the amount of information within both of the networks. Furthermore, the amount of information in the A1 layer of either network is more than that of the input AN layer, which suggests that some useful learning is happening in the simple AN-A1 model, even if this learning is not enough to fully differentiate between the two vowels.

It is important to note the slope of the lines within Figure 3.6, since it indicates how much information is retained within the top fifty most informative cells. The more gradual the slope, the more informative cells the layer contains. If one cell within a layer reaches maximal information, while the second most informative cell has very little information (resulting in a very steep slope), such a representation would not be considered as good as a layer where no cell reaches maximal information, but a large number of cells still have significantly large information (resulting in a very gradual slope). This is because information encoding in the brain needs to be very robust, and that can only be achieved if a large number of cells learn distributed representations of the stimuli, compared to only one cell learning everything, and other cells learning little. In the first scenario damage to a few cells does not dramatically damage the organism's ability to perform recognition tasks, while in the second scenario damage to one cell can completely disable this process.

4.4.2 Change in AN-CN-IC-A1 Model Response Properties Due to Training

Raster plots of the firing of neurons from the output layer of the full AN-CN-IC-A1 model to the 20 exemplars of each stimulus before and after training were plotted and are depicted in Figures 4.5-4.8. It can be seen that after training the output layer cells from the full AN-CN-IC-A1 model that learnt transform invariant

representations of the two vowels /i:/ and /a/ (cells 353 and 891 correspondingly) fire to all presentations of all exemplars of their preferred stimulus and to very few of the other stimulus exemplars (see Figs. 4.5 and 4.7). These cells also fire primarily once during stimulus presentation and their spike times are precisely aligned.

In contrast to the trained behaviour, before training cells 353 and 891 have similar response properties to both vowel stimuli (Figs. 4.6 and 4.8). It can be seen that before training both cells fire to both stimuli indiscriminately, thus providing no information about the identity of the stimulus, their firing rate is much higher than after training and they discharge at random times throughout the presentation of the stimuli. By contrasting the spiking behaviour of the two cells before and after training, it can be seen that network training had a significant effect on the emergence of auditory object selectivity and invariance within the network output layer.

The raster plots shown in Figures 4.5-4.8 demonstrate that in general before training A1 cells fire more, they fire at random times and they fire to different exemplars of both stimuli. After training, however, the cells fire typically only once during each stimulus presentation, the time of their spike is more constant and they are selective to one vowel, whereby they fire to most presentations of most exemplars of one vowel, and to very few of the other. Further evidence for the change in firing comes from calculating the average (median) rate of firing during the different stages within the model hierarchy (see Fig. 4.2). It can be seen that the rate of firing goes down throughout the hierarchy, supporting the neurophysiological findings that neurons' firing pattern changes from predominantly sustained to predominantly transient in the auditory brain (Ito & Fukushima 1988, Kurogi 1991, Steinschneider et al. 1990). It is important to note here that on average the rate of firing is reduced for *all* cells within each layer of the hierarchy, rather than just the stimulus discriminating cells. This is important, because this means that the hypothetical next readout stage of the auditory brain hierarchy will not be overwhelmed by the high rate of firing of uninformative cells. On the contrary, due to the nature of STDP, the rate of firing of uninformative cells falls more than that of the informative cells due to the strong LTD (Masquelier et al. 2008). Since uninformative cells do not take part in regularly repeating patterns of firing by definition, their firing will essentially constitute background noise, which will then be dampened through the strong LTD, thus enabling making the informative spikes appear more prominent and hence easier to readout by the hypothetical next area of the auditory brain.

In order to better understand the limitations of object selectivity and invariance learning within the output layer of the simple AN-A1 model, raster plots in response to the two vowel discrimination paradigm similar to the ones described

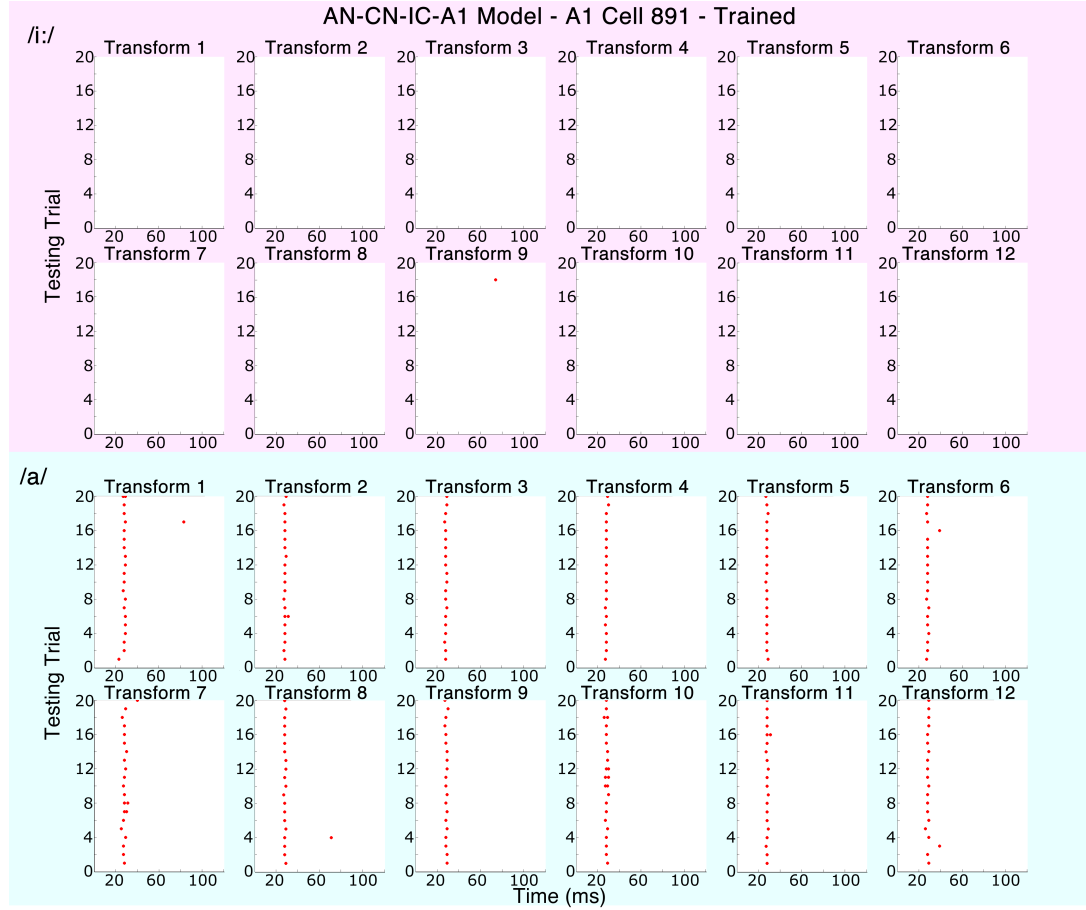


Figure 4.5: Response properties of cell 891 within the output A1 layer of the full AN-CN-IC-A1 spiking neural network model of the auditory brain after training. Each plot represents a raster of firing of the output layer cell 891 in response to twenty testing trials of one of the twelve exemplars of each of the two vowels, /i:/ and /a/. It can be seen that the cell responds to all presentations of all exemplars of /a/, while firing only once in response to /i:/ (during the 18th epoch of the 9th exemplar of the vowel). Therefore, this cell has learnt a transform invariant representation of the vowel /a/, which is reflected in the 0.981/1 bits of information achieved by this cell in the two vowel discrimination task. It can also be seen that the timing of firing of the cell is precise and exemplar independent.

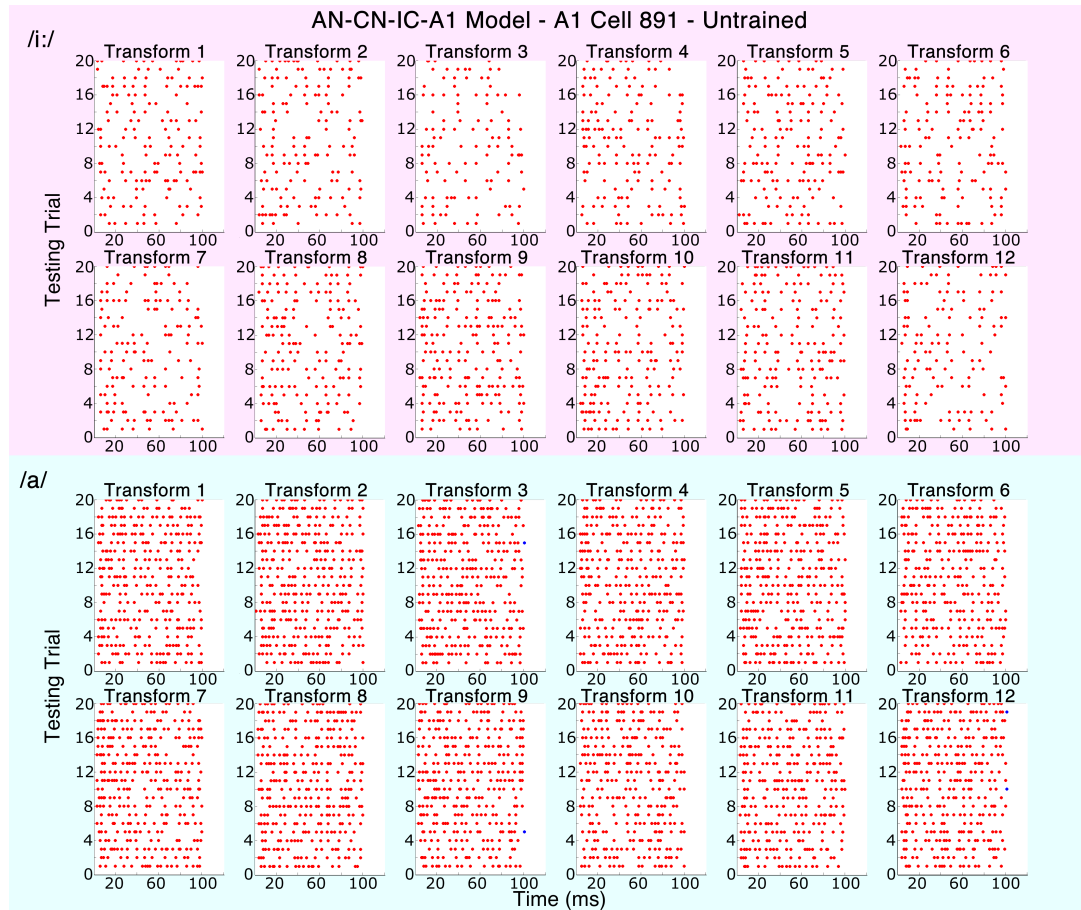


Figure 4.6: Response properties of cell 891 within the output A1 layer of the full AN-CN-IC-A1 spiking neural network model of the auditory brain before training. Each plot represents a raster of firing of the output layer cell 891 in response to twenty testing trials of one of the twelve exemplars of each of the two vowels, /i:/ and /a/. It can be seen that the cell responds strongly to all exemplars of both vowels, thus providing no information about the vowel identity. It can also be seen that the timing of firing of the cell is random and the firing rate is high.

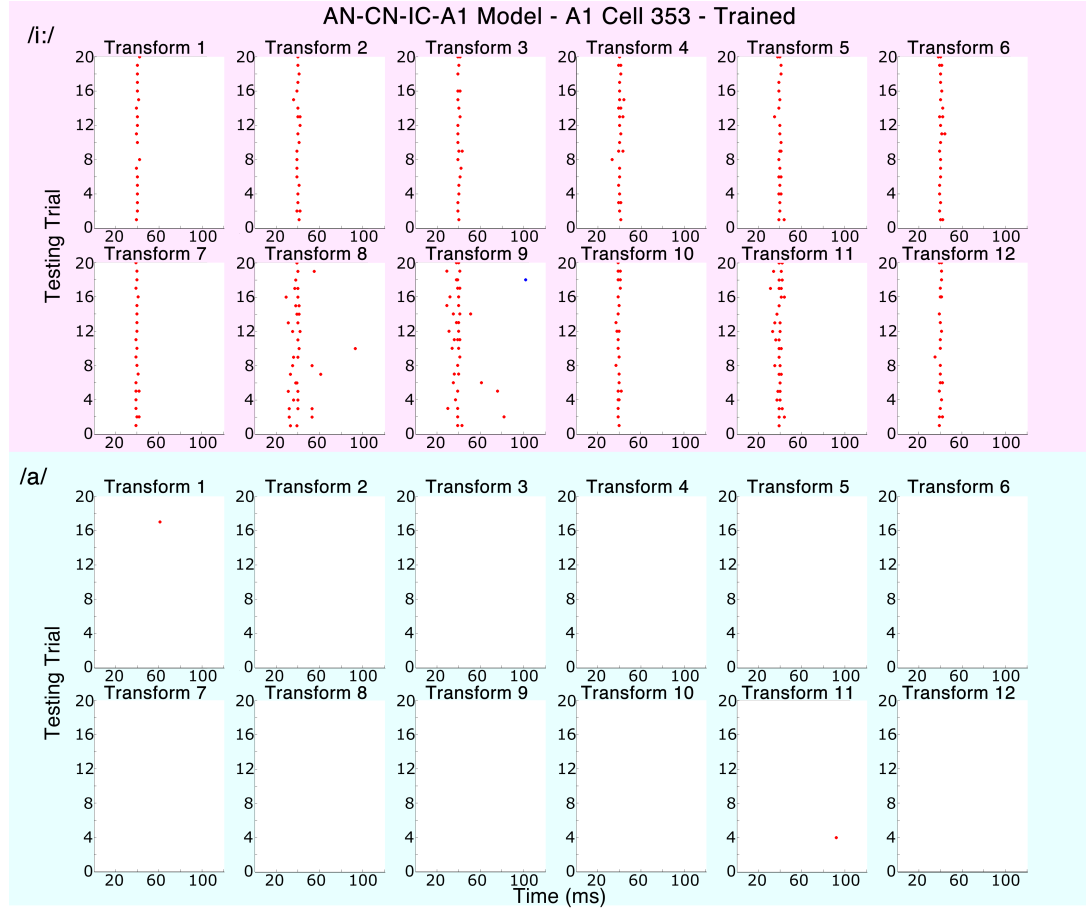


Figure 4.7: Response properties of cell 353 within the output A1 layer of the full AN-CN-IC-A1 spiking neural network model of the auditory brain after training. Each plot represents a raster of firing of the output layer cell 353 in response to twenty testing trials of one of the twelve exemplars of each of the two vowels, /i:/ and /a/. It can be seen that the cell responds to all presentations of all exemplars of /i:/, while firing only twice in response to /a/ (during the 17th testing trial of the first exemplar and the 4th epoch of the 11th exemplar of the vowel). Therefore, this cell has learnt a transform invariant representation of the vowel /i:/, which is reflected in the 0.917/1 bits of information achieved by this cell in the two vowel discrimination task. It can also be seen that the timing of firing of the cell is precise and exemplar independent.

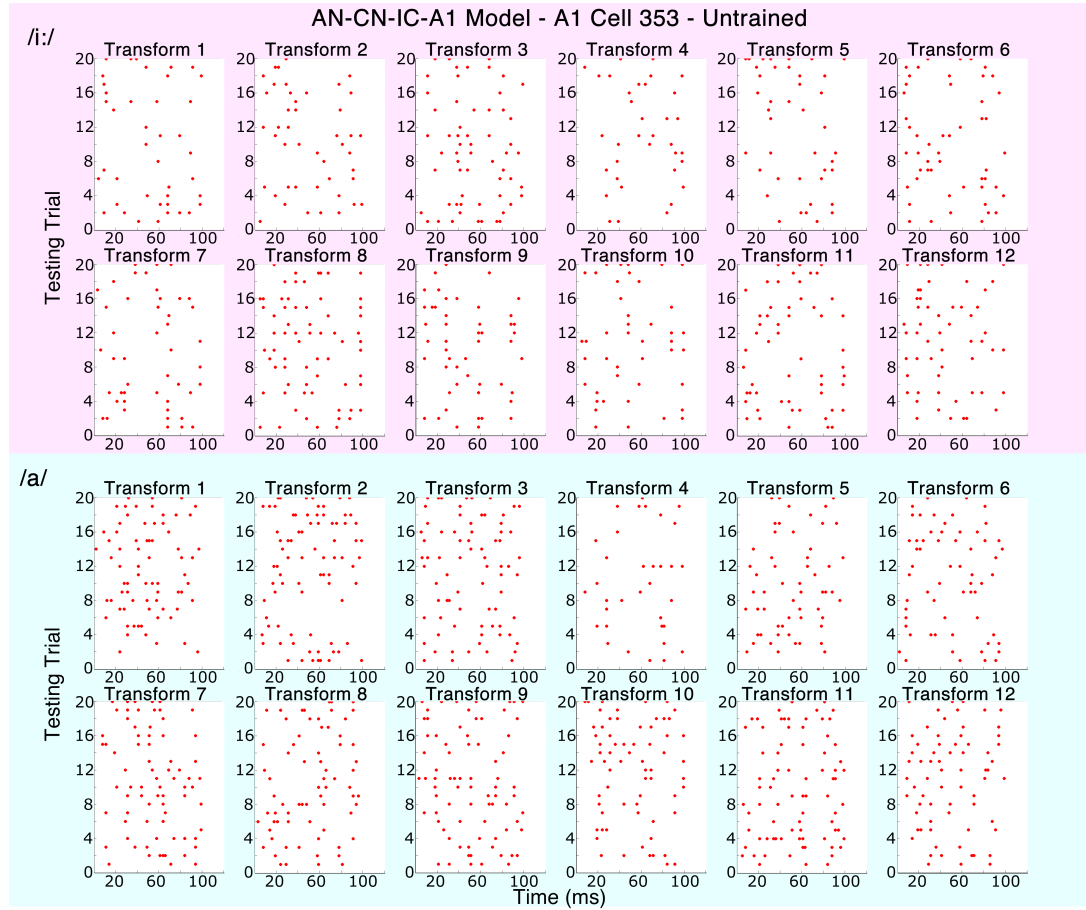


Figure 4.8: Response properties of cell 353 within the output A1 layer of the full AN-CN-IC-A1 spiking neural network model of the auditory brain before training. Each plot represents a raster of firing of the output layer cell 353 in response to twenty testing trials of one of the twelve exemplars of each of the two vowels, /i:/ and /a/. It can be seen that the cell responds strongly to all exemplars of both vowels, thus providing no information about the vowel identity. It can also be seen that the timing of firing of the cell is random and the firing rate is high.

in Section 4.4.2 were plotted. There were no cells that preferred vowel /a/ among the top hundred output cells with most single cell information about the two vowel stimuli. Any output cell in the simple AN-A1 model which fired to most exemplars of vowel /a/ also fired strongly to vowel /i:/, resulting in very little information content (see Fig. 4.9). Furthermore, output cells from the simple AN-A1 model that had preference for vowel /i:/ also fired to many exemplars of the non-preferred vowel /a/ (see Fig. 4.10). It can be seen that the output cells within the simple AN-A1 model did not learn stimulus selective, transform invariant representations of the two vowels, and their firing is randomly timed for different presentations of the stimuli.

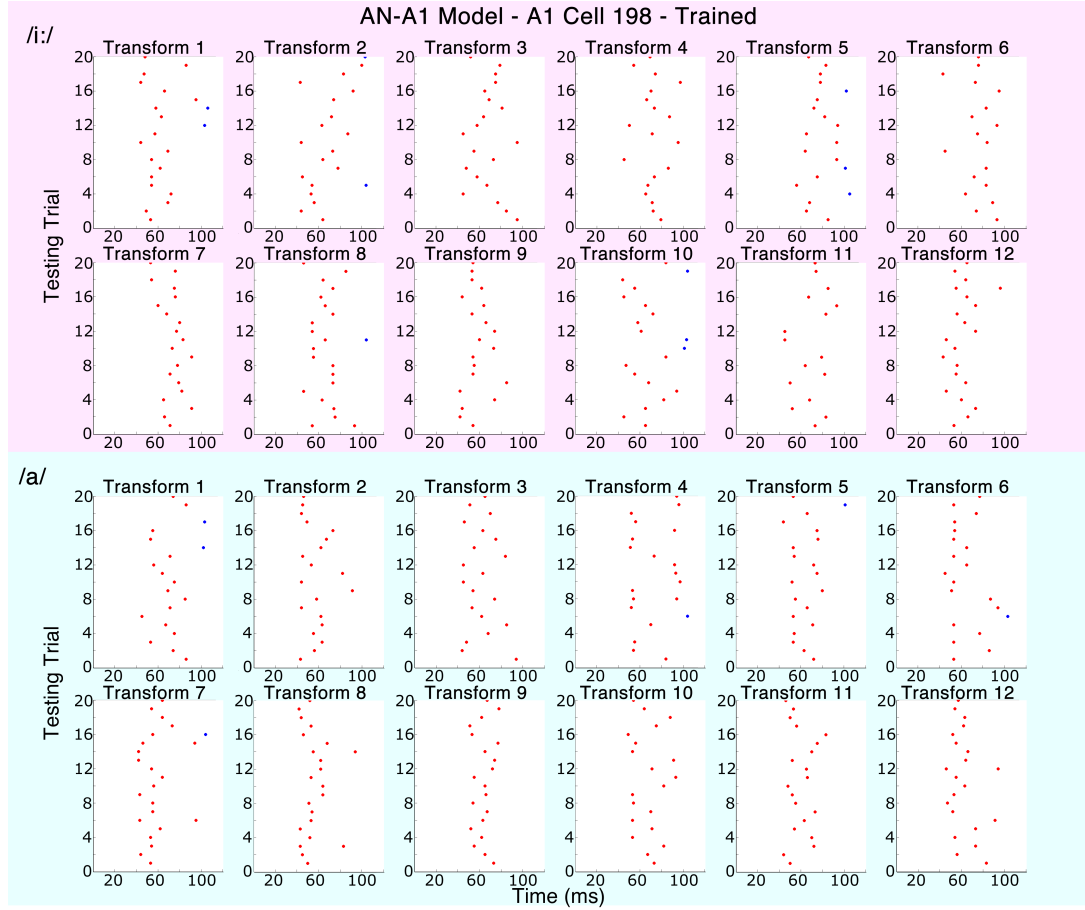


Figure 4.9: Response properties of cell 198 within the output A1 layer of the simple AN-A1 spiking neural network model of the auditory brain after training. Each plot represents a raster of firing of the output layer cell 198 in response to twenty testing trials of one of the twelve exemplars of each of the two vowels, /i:/ and /a/. It can be seen that the cell responds similarly to both stimuli, whereby it fires to most presentations of most exemplars of both /i:/ and /a/. This is reflected in the low 0.011/1 bits of information achieved by this cell in the two vowel discrimination task, which is not enough to discriminate between the two vowels. It can also be seen that the timing of firing of the cell is not precise, with the cell firing at different times during different presentations of the same stimulus.

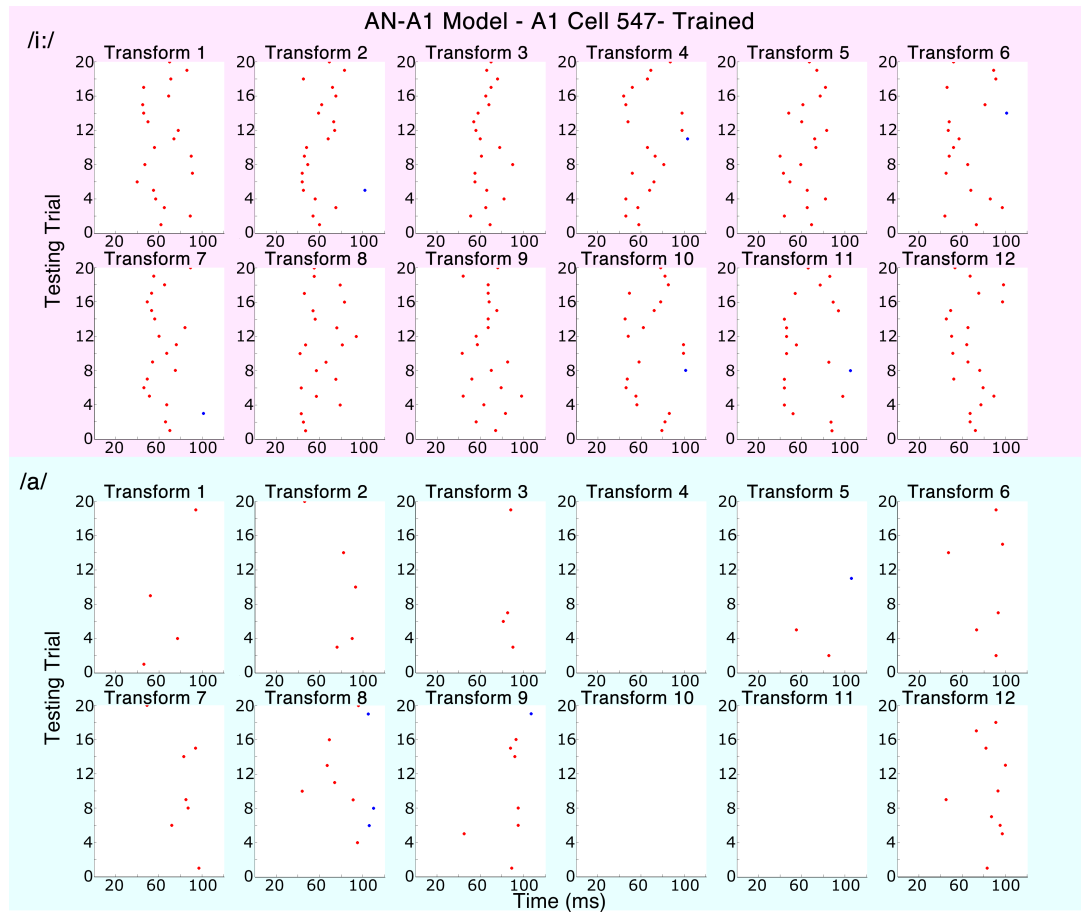


Figure 4.10: Response properties of cell 547 within the output A1 layer of the simple AN-A1 spiking neural network model of the auditory brain after training. Each plot represents a raster of firing of the output layer cell 547 in response to twenty testing trials of one of the twelve exemplars of each of the two vowels, /i:/ and /a/. It can be seen that the cell responds to all presentations of all exemplars of /i:/, however it also fires in response to /a/. This is reflected in the 0.519/1 bits of information achieved by this cell in the two vowel discrimination task, which is not enough to discriminate between the two vowels. It can also be seen that the timing of firing of the cell is not precise, with the cell firing at different times during different presentations of the same stimulus.

4.4.3 Preferred Afferent Patterns for AN-A1 and AN-CN-IC-A1 Models

It is of interest to visualise what particular auditory patterns the maximally informative cells within the output A1 layers of the simple AN-A1 and full AN-CN-IC-A1 models learnt to respond to. One instructive way in which this can be done is by taking 50 ms worth of 1 ms binned AN or IC firing rasters (for the simple and full model correspondingly) in response to the different stimulus transform presentations that immediately precede the firing time bin of the chosen A1 cell (this is similar to spike-triggered average calculations). The AN or IC firing rasters are transformed into a binary matrix form (M_{ic} , where i is the index of each spike emitted by each A1 or AN cell c in response to all exemplars of all stimuli), whereby the dimensionality of the matrices is the same (1000 cells by 50 ms), and the matrix components corresponding to the time bins when particular AN or IC cells fired are set to 1. These matrices are then summed ($M_c = \sum_i M_{ic}$) and plotted to help visualise whether the maximally informative A1 cells within the output layer of the simple AN-A1 and full AN-CN-IC-A1 models have learnt to respond to stable afferent firing patterns (Figs. 4.11, 4.12, 4.13, and 4.14).

In order to understand which stable patterns of firing in the IC were learnt by the A1 cells 353 (selective to vowel /i:/) and 891 (selective to vowel /a/), the following calculations were done. First, the total number of times each of the two output cells fired during testing was calculated (in response to the presentation of the 20 testing trials of 12 exemplars of each of the two vowels). Then the summed IC raster matrices, as described above, were analysed to determine how often spikes appeared in different bins during the testing paradigm. If certain bins contained as many spikes as the number of times their corresponding output cell fired, then it is likely that the spikes in that bin caused the output cell to spike. By examining IC bins with the most number of spikes, it is possible to determine a potential stable temporal pattern of activity in the IC that a particular A1 cell has learnt to respond to.

It was found that the A1 cell 353 fired 385 times and the A1 cell 891 fired 555 times during the presentation of the 20 testing trials of 12 exemplars of each of the two vowels to the full AN-CN-IC-A1 model. When looking at the summed IC raster matrix, the ten highest valued matrix components had between 84% (324) and 100% (385) of all spikes emitted by the A1 cell 353, and between 78% (438) and 90% (497) of all spikes emitted by the A1 cell 891, indicating that the IC layer of the full AN-CN-IC-A1 model contains stable firing patterns during the presentations of different exemplars of the stimuli over different epochs (see Figs. 4.11 and 4.12). This stability allows the A1 layer to learn stimulus selective and transform invariant representations of the two vowels.

In contrast to the above, patterns of firing that cause activity within the most

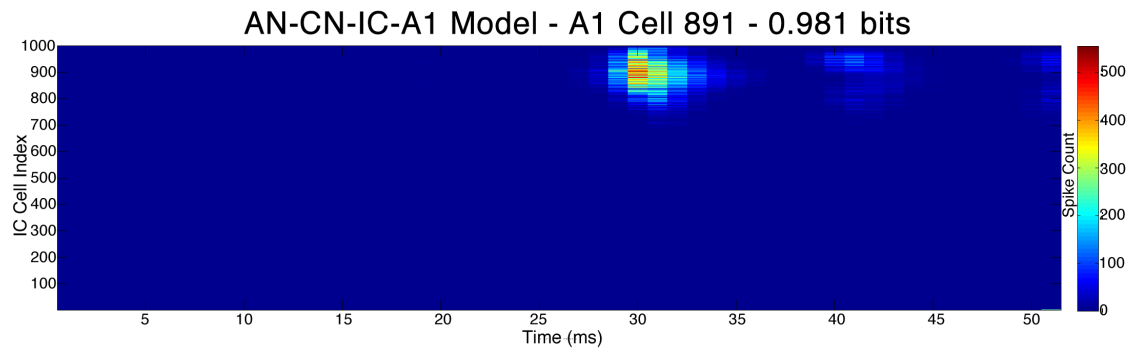


Figure 4.11: An aggregate representation of afferent firing received by cell 891 within the output layer A1 of the full AN-CN-IC-A1 neural network model of the auditory brain in response to the presentation of twelve exemplars of each of the two input vowels /i:/ and /a/. The afferent IC spikes were aligned on the spike time of neuron 891 in A1, whereby neuron 891 would fire at the 51 ms time bin in the current plot. It can be seen that while cell 891 fired 555 times during testing, the afferent cells fired up to 497 times, indicating stable spiking patterns within the IC, which allow neurons in the A1 to learn stimulus selective and transform invariant representations of vowels.

informative output layer cells from the simple AN-A1 model are less well defined (see Figs. 4.13, and 4.14). A1 cell 547 fired 297 times and A1 cell 198 fired 482 times during the presentation of the 20 epochs of 12 exemplars of each of the two vowels to the simple AN-A1 model. The highest values of the AN firing raster matrix components that cause cell 547 in A1 to discharge are between 51 (17%) and 54 (18%), and between 87 (18%) and 104 (22%) for the A1 cell 198, indicating that the AN patterns are not stable during the presentations of different exemplars of the stimuli over different epochs (see Figs. 4.13 and 4.14). This instability prevented the A1 layer of the simple AN-A1 model from learning stimulus selective and transform invariant representations of the two vowels.

Cells 891 and 353 were chosen arbitrarily as examples of informative A1 cells within the full AN-CN-IC-A1 models, and their receptive fields (as characterised by the matrices visualised in Figures 4.11 and 4.12) are not representative of *all* informative A1 cells. Instead, a large variety of different receptive fields were learnt in the A1.

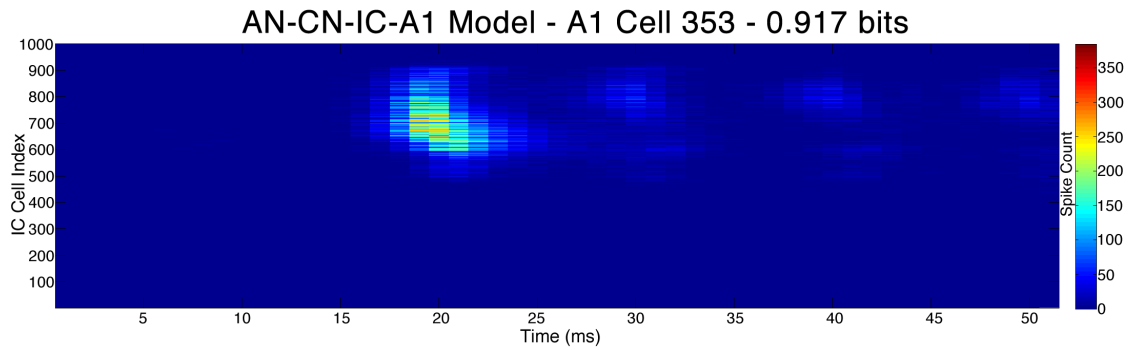


Figure 4.12: An aggregate representation of afferent firing received by cell 353 within the output layer A1 of the full AN-CN-IC-A1 neural network model of the auditory brain in response to the presentation of twelve exemplars of each of the two input vowels /i:/ and /a/. The afferent IC spikes were aligned on the spike time of neuron 353 in A1, whereby neuron 353 would fire at the 51 ms time bin in the current plot. It can be seen that while cell 353 fired 385 times during testing, the afferent cells fired up to 385 times as well, indicating stable spiking patterns within the IC, which allow neurons in the A1 to learn stimulus selective and transform invariant representations of vowels.

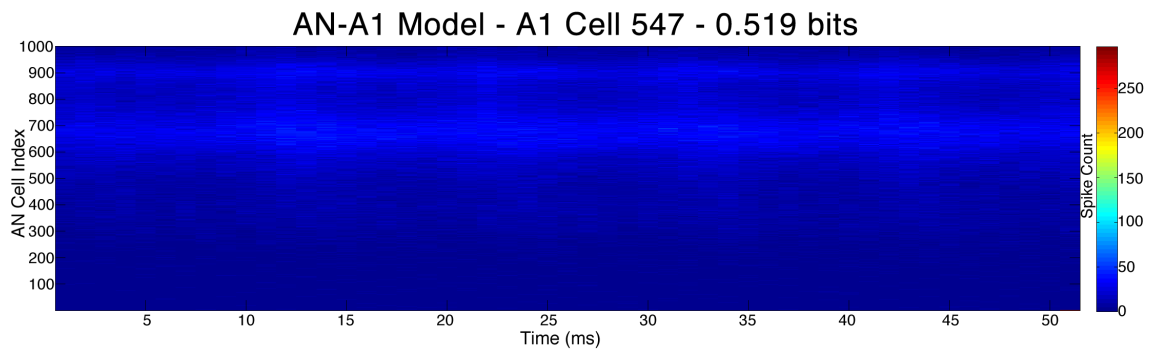


Figure 4.13: An aggregate representation of afferent firing received by cell 547 within the output layer A1 of the simple AN-A1 neural network model of the auditory brain in response to the presentation of twelve exemplars of each of the two input vowels /i:/ and /a/. The afferent AN spikes were aligned on the spike time of neuron 547 in A1, whereby neuron 547 would fire at the 51 ms time bin in the current plot. It can be seen that while cell 547 fired 297 times during testing, the afferent cells fired only up to 54 times, indicating unstable spiking patterns within the AN, which prevented neurons in the A1 from learning stimulus selective and transform invariant representations of vowels.

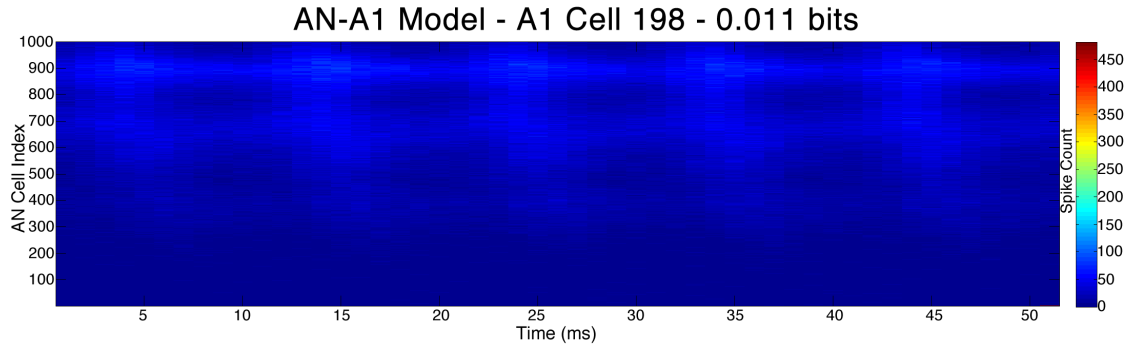


Figure 4.14: An aggregate representation of afferent firing received by cell 198 within the output layer A1 of the simple AN-A1 neural network model of the auditory brain in response to the presentation of twelve exemplars of each of the two input vowels /i:/ and /a/. The afferent AN spikes were aligned on the spike time of neuron 198 in A1, whereby neuron 198 would fire at the 51 ms time bin in the current plot. It can be seen that while cell 198 fired 482 times during testing, the afferent cells fired only up to 104 times, indicating unstable spiking patterns within the AN, which prevented neurons in the A1 from learning stimulus selective and transform invariant representations of vowels.

4.4.4 Changes in Afferent Connections Strength Due to Learning

When examining histograms and heat plots of all afferent connection weights (w_{ij}^{BL}) between the IC and A1 of the full AN-CN-IC-A1 model of the auditory brain before and after training (see Fig. 4.15 A-B), it can be seen that little learning happened in the low characteristic frequency part of the layer, since most connections remained within the [30, 35] nA range used for the afferent connection strength initialisation. The part of the layer that corresponded to the vowel formants (see Figure 3.3), however, changed significantly compared to the untrained network. While most connections weakened due to random firing within the IC, a large number of afferents strengthened up to the maximum of 60 nA (see Fig. 4.16, as these connections are hard to see in the histogram shown in Fig. 4.15 due to the scale). These afferents connect the IC cells that took part in the regular patterns of firing in response to the vowels to the A1 cells that have the right distribution of delays to extract stable firing patterns from the afferent IC firing raster (see Section 2.4). When compared to the best performing simple AN-A1 models (see Fig. 4.15 C-D), it can be seen that a different pattern of afferent connection weights (w_{ij}^{BL}) emerges. One of the simple AN-A1 models did not get any strengthening at all during training, while the other got too much, which resulted in all afferents to certain A1 cells getting maximally strengthened. This makes

such A1 cells non-discriminative, since any pattern of firing within the AN would discharge these cells. In comparison, the maximally strengthened afferents within the full AN-CN-IC-A1 model are randomly scattered within the upper half of the IC-A1 connection weight matrix.

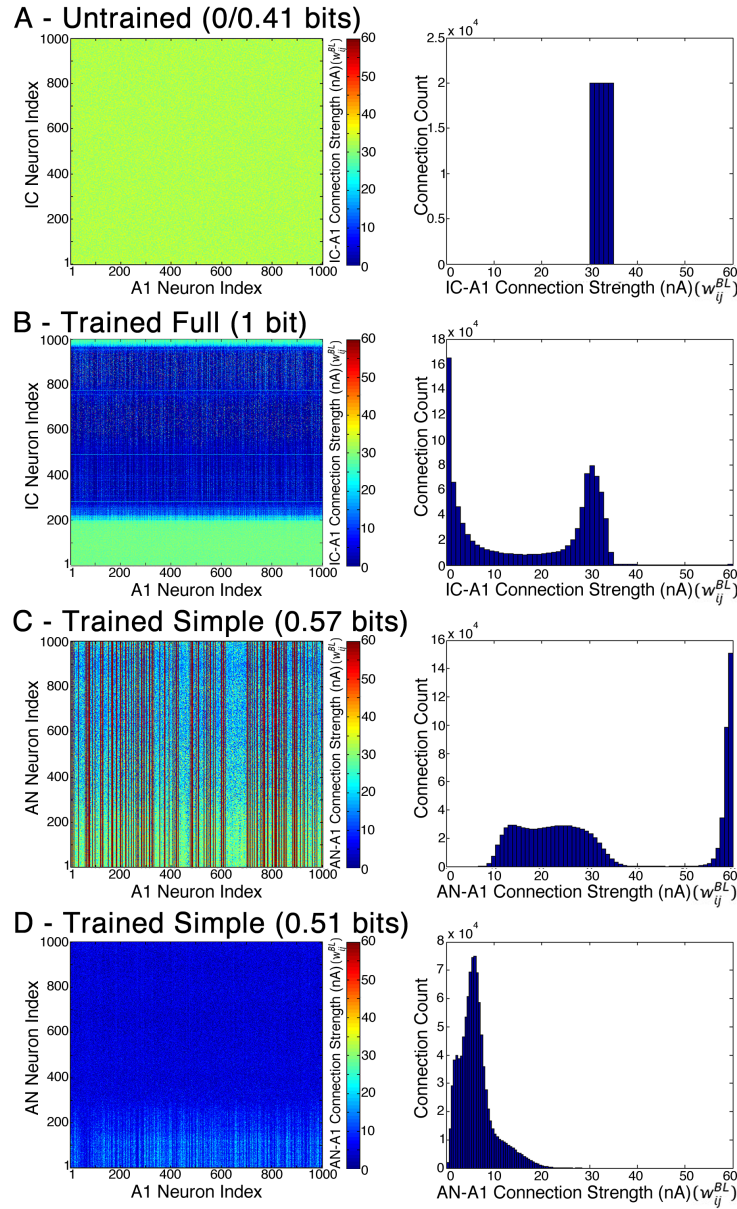


Figure 4.15: Afferent connection weight matrices (w_{ij}^{BL}) (left column) and histograms (right column) for the output A1 layer. **A:** illustration of newly initialised afferent connections for the output layer A1 of the simple AN-A1 and the full AN-CN-IC-A1 network models. The numbers in parentheses indicate the maximal single cell information present within the output layer of the two models respectively. **B:** illustration of the IC-A1 afferent connections of the full AN-CN-IC-A1 model after training. **C-D:** illustration of the AN-A1 afferent connections of the two best performing simple AN-A1 models of the auditory brain after training. The number in parentheses indicates the maximal single cell information present within the output layer of each corresponding model after training. The maximal theoretical single cell information for a two vowel discrimination task is 1 bit. The plots illustrate how the distributions of the connection weights changed after training. The extensive strengthening (C) or unimodal weakening (D) of the A1 connection weights results in uninformative cells within the A1 (maximum of 0.57 or 0.51 bits respectively). A multimodal distribution of A1 connections, whereby most connections weaken, while some get strengthened (B), results in maximally informative cells in the A1 (full 1 bit of information).

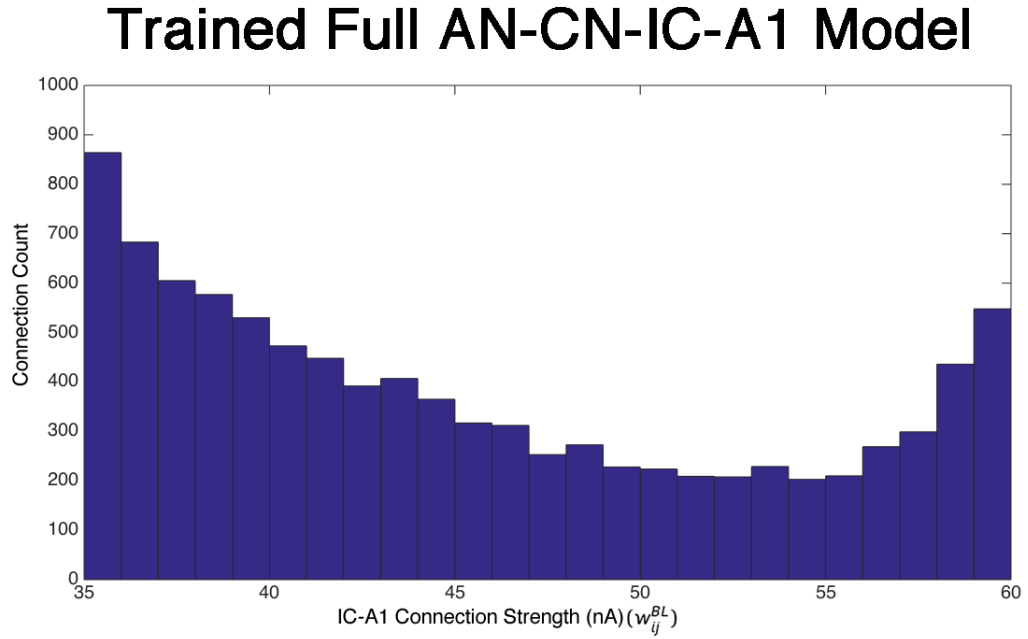


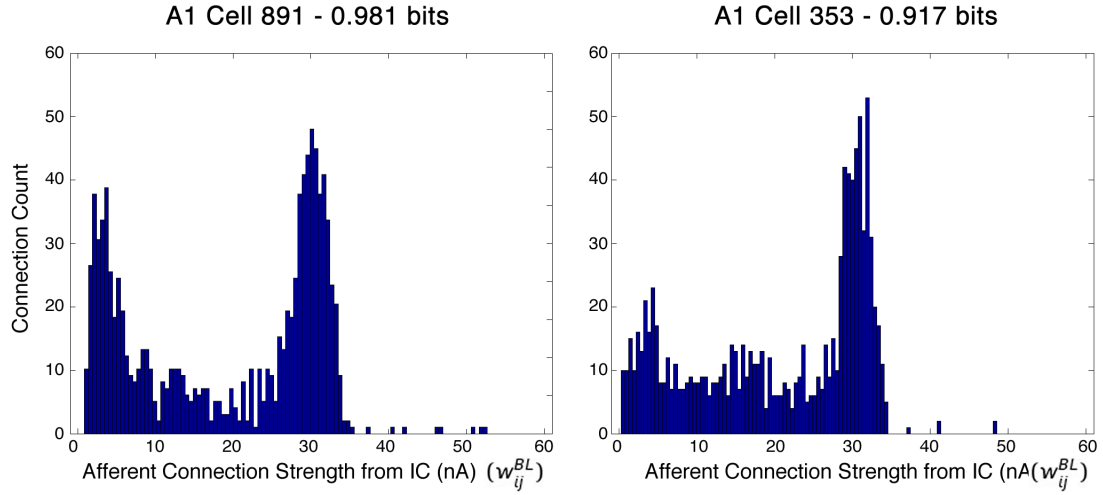
Figure 4.16: Histogram summarising the distribution of afferent connections (w_{ij}^{BL}) that were strengthened between IC and A1 layers of the full AN-CN-IC-A1 model of the auditory brain when trained on the synthesised vowels /i:/ and /a/. Connections are defined as strengthened if their magnitude is in the range $(35, 60]$ nA, which is above the upper bound of the uniform distribution used for the weight initialization (35 nA). It can be seen that 548 afferents were strengthened maximally ($[59, 60]$ nA). They correspond to 6% of all strengthened afferents.

4.4.5 Distribution of Afferent Connection Weights for AN-A1 and AN-CN-IC-A1 Models

When plotting the distribution of afferent connection weights (w_{ij}^{BL}) for the maximally informative cells, it can be seen that the output layer cells from the full AN-CN-IC-A1 model that have learnt transform invariant representations of the two vowels, /i:/ and /a/ (cells 353 and 891 correspondingly) have bimodal distributions of afferent connection magnitudes. Such bimodal distributions are indicative of learning, because they could only arise if the output A1 layer was exposed to stable repeating patterns of IC firing embedded within noise during training. While no noise was explicitly added to the auditory stimuli, not all spikes received by the A1 were information bearing. Those other non-information bearing spikes then essentially acted as statistically equivalent 'noise' in which the stable repeating informative patterns of firing were embedded. The interplay between such IC input and the right STDP parameters would result in the weakening of the majority of IC→A1 connections due to their exposure to this 'noise' of non-informative spikes, and the strengthening of some IC→A1 connections due to their exposure to the information-bearing spikes making up the stable repeating patterns as described by Masquelier et al. (2008).

It can be seen that a large number of connections from the IC layer to A1 cells 353 and 891 were weakened ($w_{ij}^{BL} < 30$ nA), many connections remained within the initialisation range ($30 \leq w_{ij}^{BL} \leq 35$ nA), and a number of connections from the IC got strengthened ($35 < w_{ij}^{BL}$ nA) (see Fig. 4.17 top row). At the same time the maximally informative output cells from the simple AN-A1 model have retained their initial untrained unimodal shape of distribution of afferent connection weights (w_{ij}^{BL}) from the AN, albeit with weakened majority of afferent connections from the AN layer (see Fig. 4.17 bottom row). This is to be expected, since time and space jitter make AN firing more random, which weakens the afferent connections due to the stronger LTD compared to LTP, as necessary for CT STDP-based learning

AN-CN-IC-A1 Network



AN-A1 Network

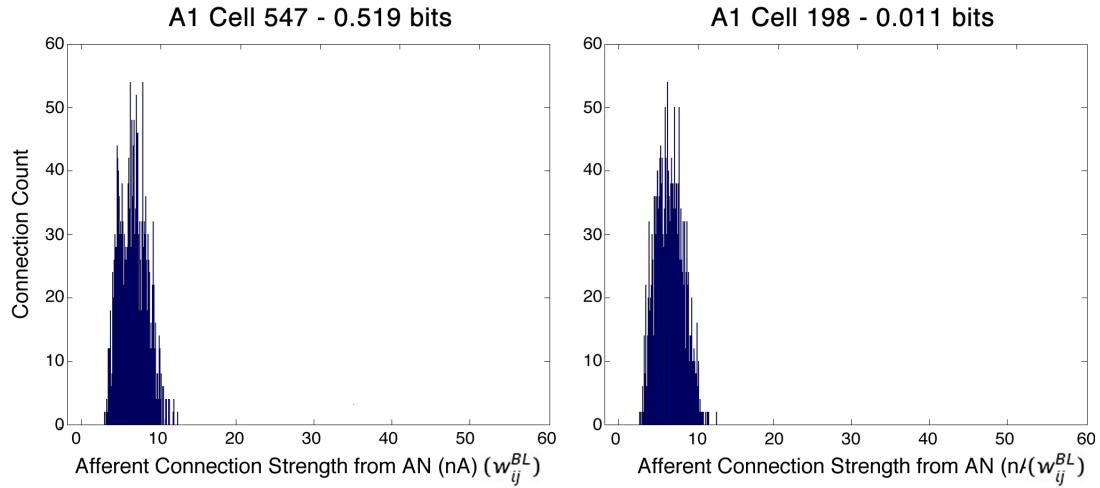


Figure 4.17: Histograms of afferent connection magnitudes (w_{ij}^{BL}) of four cells, 891 and 353 from the output A1 layer of the full AN-CN-IC-A1 network, and 547 and 198 from the output A1 layer of the simple AN-A1 network. Cells from the full AN-CN-IC-A1 network have achieved almost perfect single cell information (0.981/1 bit and 0.917/1 bit correspondingly) and have a bimodal distribution of afferent connection weights from the IC, which is indicative of learning. Cells from the simple AN-A1 network have achieved poor single cell information (0.519/1 bit and 0.011/1 bit correspondingly) and have retained their initial untrained unimodal shape of distribution of afferent connection weights from the AN, albeit most of the connections got weakened from their initial magnitudes between 30 and 35 nA.

4.4.6 Cell Response Properties Within the AN-A1 and AN-CN-IC-A1 Models

Figure 4.18 illustrates how different cells within the output layers of the simple AN-A1 model and the full AN-CN-IC-A1 model, as well as the input AN layer respond to different exemplars of the two vowel stimuli. Each of the two vowels, /a/ and /i:/, has twelve exemplars, and each exemplar is presented over twenty epochs, resulting in the total of 240 presentations of each vowel. In order to visualise how individual cells respond to these presentations, nested pie charts were created, whereby each circle was divided into 240 segments, with each segment representing a stimulus presentation. Filled in segments indicate that the cell fired at least once to the particular stimulus presentation, and empty segments indicate that the cell did not fire at all to the particular stimulus presentations. A cell with maximum information about the two stimuli would have either the inner or the outer circle fully filled, with the remaining circle completely empty, depending on the cell's stimulus preference. It can be seen that 20 most informative A1 cells of the full AN-CN-IC-A1 network almost reach that level of performance (Fig. 4.18, C). The network learnt selectivity for both input vowels, since cells are present that have primarily green and primarily blue pie charts, representing invariant responses to vowels /a/ and /i:/ correspondingly.

It can also be seen that cells from the output layer of the simple AN-A1 model, as well as from the input AN layer, do not reach the maximum of 1 bit of information. They have nested pie charts with both green and blue circles partially filled, which indicates that those cells fire to some presentations of both stimuli, and they are not able to provide enough information to detect which stimulus is being presented (Fig. 4.18, A-B).

4.5 Understanding AN-CN-IC-A1 Model

4.5.1 Change in Information Within Layers of the AN-CN-IC-A1 Model

The change in maximal single cell information throughout the different layers of the full AN-CN-IC-A1 model can be seen in Figure 4.2. Overall information content increases throughout the hierarchy of the model's four layers, with the exception of chopper cells, which reach full 1 bit of information. This, however, is an artefact of the simplicity of the task and of the encoding scheme used (see Figs. 5.8 and 6.12 for a demonstration of the increasing information throughout the hierarchy of the full AN-CN-IC-A1 model when more complex stimuli and encoding schemes are used). Differentiating between two vowels specifically picked because of their dissimilarity

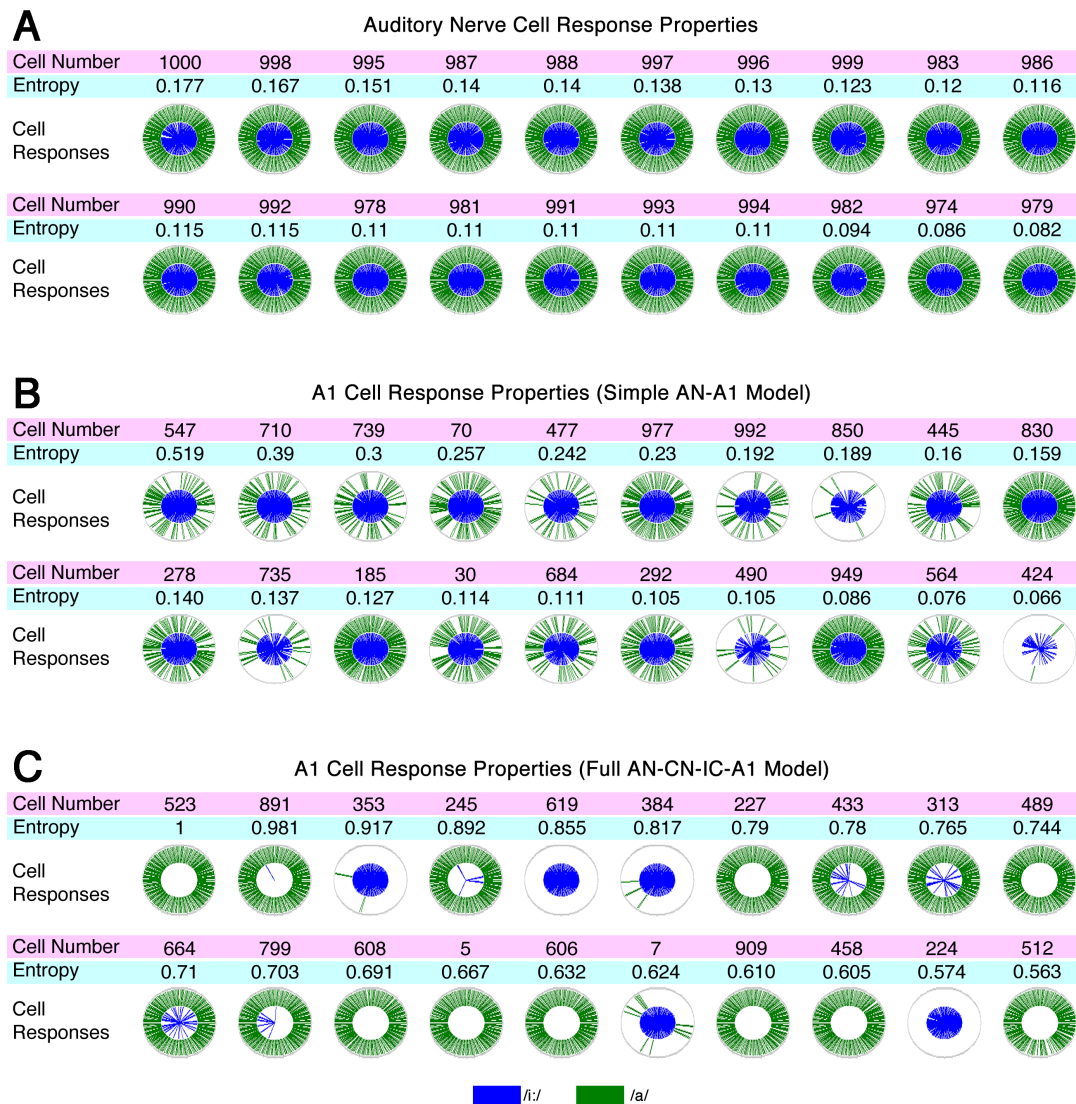


Figure 4.18: Schematic representation of response properties of 20 AN cells (A), A1 cells of the reduced AN-A1 network (B) and A1 cells of the full AN-CN-IC-A1 network (C) containing the most single cell information about the identity of the two input vowels /i:/ and /a/. The nested pie charts illustrate whether each cell fired to each of the 20 presentations of the 12 exemplars of each vowel, resulting in 240 segments. Coloured segments represent at least one discharge in response to that stimulus presentation. Blue pie charts represent responses to /i:/ and green pie charts represent responses to /a/. The magnitude of single cell information for each cell is positively correlated with the number of examples of the preferred stimulus class that the cell responds to, and negatively correlated with the number of examples of the other (non-preferred) stimulus class that the cell responds to. This is why the information of cell 224 (subplot C) is lower than the information of cells 7 or 353. Even though cell 224 doesn't respond to any examples of the non-preferred vowel /a/, it responds to much fewer examples of vowel /i:/ than cells 7 or 353, which results in less single cell information overall than cells 7 or 353.

is a very simple task for the chopper layer to solve. This task, however, was appropriate for testing the hypothesis that the sub-cortical pre-processing present within the full AN-CN-IC-A1 model enables it to outperform the reduced AN-A1 model without such pre-processing on the auditory object transform invariance learning task. Extensive parameter searches had to be run in order to test the hypothesis, and the simple two vowel discrimination task chosen demonstrated the difference in the performances of the two models using the simplest (on/off) encoding scheme and in the computationally cheapest way. As was mentioned in Chapter 3 Section 3.8.1, the same chopper cells perform much worse on the harder ten vowel discrimination task, reaching only 0.74 bits out of the maximal 3.32 bits of information. Therefore, it is necessary to have subsequent layers learning from chopper cell inputs in order for the network to solve tasks harder than the two vowel test paradigm described in Chapter 3, which was chosen specifically for its simplicity.

4.5.2 Analysing AN-CN-IC-A1 Afferent Connections Strength Changes Due to Training

In order to test whether the connections strengthened through learning within the full AN-CN-IC-A1 model are indeed important for the transform invariant recognition of the two vowel stimuli, the weight matrix of afferent connections (w_{ij}^{BL}) between the IC and A1 was shuffled and the performance of the network on the two vowel recognition task was measured. It was found that the performance dropped significantly (see Fig. 3.6). This suggests that the strengthening of connections was indeed due to the network learning about the repeating patterns within the IC layer in response to the presentation of the different stimuli. When the shuffled network was allowed to learn again, its performance returned back to almost perfect information (0.91 bits) (see Fig. 3.6).

4.5.3 Role of Afferent Connections Strength Initialisation Values for Learning

In order to check whether during training the network adapted its afferent weight matrix (w_{ij}^{BL}) to represent the underlying structure of the input stimuli transformed through the particular delay matrix (Δ_{ij}), the IC-A1 weight matrix of the originally trained full AN-CN-IC-A1 model was compared to that of the model re-trained after IC-A1 weight shuffling. It was found that after weight shuffling 3.1% of originally strengthened connections remained strengthened. This number rose to 21.25% after re-training the network. Strengthened weights are defined as having magnitude above 35 nA, which is the upper bound of the uniform distribution

used for weight initialisation. Furthermore, among the top 100 most informative cells within the originally trained A1 layer of the full AN-CN-IC-A1 model, 19% remained within the top 100 most informative cells of the model re-trained after weight shuffling (see Fig. 4.19). The reason why many of the same cells within the output layer learn to represent the two vowel stimuli with high information after the model is re-trained following weight shuffling is that the distribution of heterogeneous axonal delays (Δ_{ij}) between the IC and A1 was kept constant, thus encouraging the same A1 cells to learn about the vowel stimuli despite the shuffled afferent weight matrix (w_{ij}^{BL}). When examining the information content within these cells, it can be seen that the re-trained model has more information on average, which could be due to the fact that the afferent weight shuffling preserves the magnitude of the weights while changing their location. This could potentially facilitate learning, meaning that the same 200 epochs of training as used during the original training results in extra learning. It can also be seen that hardly any information is preserved after shuffling the weights. Furthermore, the information content within the two versions of the network is strongly correlated. Pearson correlation was calculated for the data collected, whereby the two vectors used in the calculation were the magnitudes of information content for the A1 cells in the original AN-CN-IC-A1 network, and the AN-CN-IC-A1 network re-trained after weight shuffling, thus resulting in a repeated measures design. The Pearson's correlation was found to be 0.7346 ($p < 0.001$).

In another simulation, the original AN-CN-IC-A1 network was trained from a naive randomly initialized state, whereby all the parameters, including the distribution of heterogeneous axonal delays (Δ_{ij}) between the IC and A1 layers, were kept the same as in the original simulations. The only change was a different random initialization of the afferent connection weights (w_{ij}^{BL}), albeit still drawn from the same uniform distribution between 30 and 35 nA. It was found that 62% of the top 100 cells with the most information content were the same in the output A1 layer of the newly trained AN-CN-IC-A1 model and the original network model, and the Pearson correlation between the information content vectors of their corresponding A1 layers was 0.8385 ($p < 0.001$) (see Fig 4.20).

These results of the two simulations described above indicate that the initial distribution of connection weights (w_{ij}^{BL}) has an effect on the final network organisation, since other parameters between the two training conditions were kept the same. When the two initialised weight vectors are similar (both drawn from a narrow uniform distribution between 30 and 35 nA), the differences in the final network organisation and neurons' stimulus selectivity profiles are little. Weight initialisation drawn from very different distributions (uniform distribution between 30 and 35 nA, and complex distribution arising after training as seen in Fig. 4.16), results in more profound changes in the stimulus selectivity profiles of the 100 most

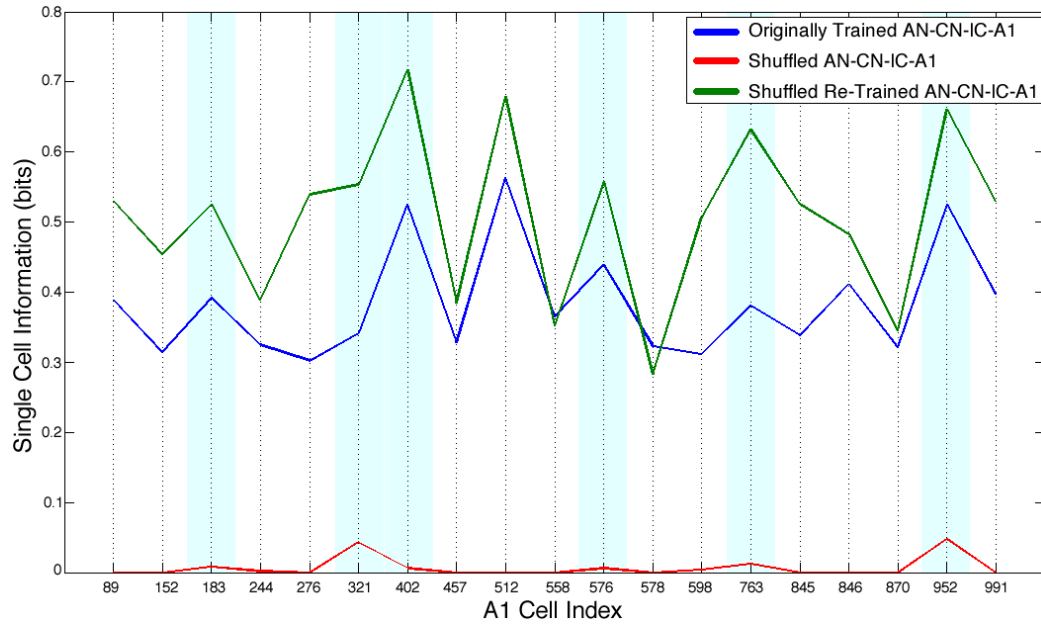


Figure 4.19: Amount of single cell information in the output layer of the full AN-CN-IC-A1 model after original training, once its trained afferent connection weights (w_{ij}^{BL}) were shuffled, and once the shuffled network was re-trained. Cells plotted are the intersection of the top 100 most informative cells of the original and the shuffled re-trained networks. Light blue shaded cells are also within the top 100 most informative cells of the shuffled network before re-training. It can be seen that on average there is more information present in the re-trained condition. Furthermore, it can be seen that the correlation between the information magnitudes between the two networks is strong (0.7346, $p < 0.001$).

informative output layer neurons (see Fig. 4.21). For example, it can be seen that output neurons 152, 457, 558 and 578 are all among the 100 cells with the highest amount of information within the output layer of both the original AN-CN-IC-A1 network and the AN-CN-IC-A1 network re-trained after weight shuffling. The stimulus preference of these cells, however, differs dramatically between the two networks. This change in stimulus selectivity arose due to the different distributions of initial afferent connection weights (w_{ij}^{BL}) between the two networks. It has to be noted though, that the maximal single cell information magnitude in the output layer of the AN-CN-IC-A1 network remains similar whether the network was trained originally, re-trained after weight shuffling or trained with the re-initialized afferent connection weights. The only thing that differs is which cells are among the top 100 most informative, and which stimulus these cells learn to prefer.

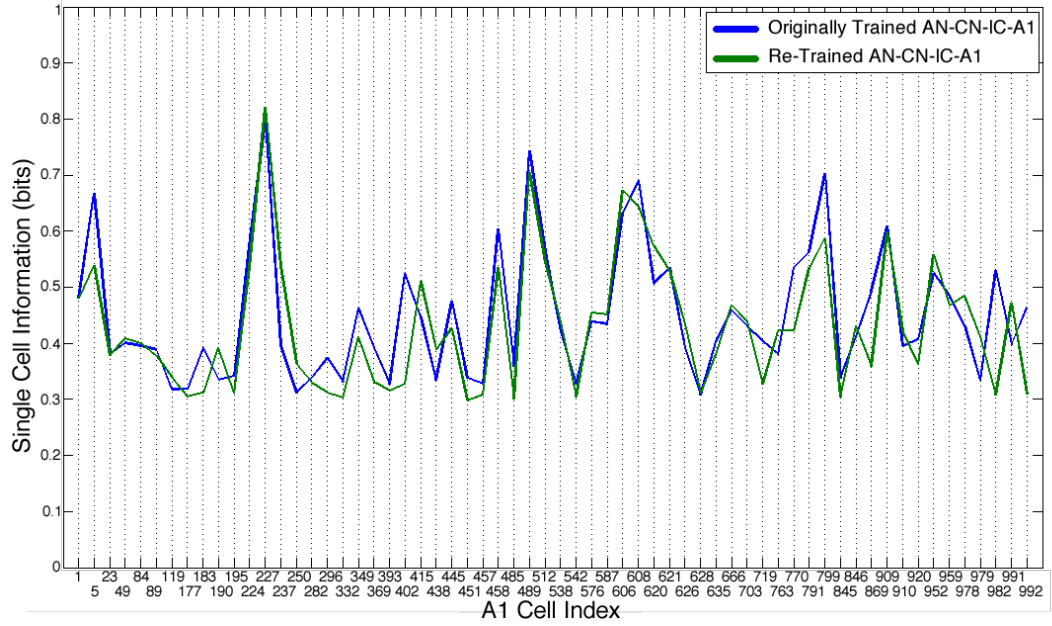


Figure 4.20: Amount of single cell information in the output layer of the full AN-CN-IC-A1 model after original training, and after re-initialising the afferent connection weights (w_{ij}^{BL}) and re-training the network. Cells plotted are the intersection of the top 100 most informative cells of the original and the re-trained networks. It can be seen that on average information present in the two conditions is very similar. Furthermore, it can be seen that the correlation between the information magnitudes between the two networks is very strong (0.8385, $p < 0.001$).

When the response properties of the 19 common most informative cells before shuffling and re-training and after are considered, it can be seen that while some cells remain selective to the same vowel, others change (see Fig. 4.21). This could be due to a number of reasons. First, it could be that weight shuffling changes output cell afferents significantly, which means that the input pattern that used to activate that cell does not do so anymore at the start of re-training. Hence, the cell is not able to learn to respond to that pattern during re-training.

Furthermore, it may be that a cell has a combination of afferent weights and delays which means that it can be selective to stable patterns within both vowel stimuli. If nothing intervenes, such a cell would become one of the less informative ones, since it would fire to all exemplars of both stimuli. However, it may instead learn to be selective to only one of the stimuli if an appropriate number of other output cells fire just before it, thus activating strong within-layer inhibition, which in turn would prevent the original cell from firing in response to one of the two stimuli. After such training, the cell in question would appear very informative,

being selective to the exemplars of one vowel and not the other.

However, weight shuffling may disturb the configuration of afferent weights (w_{ij}^{BL}) and delays (Δ_{ij}) that activated the right number of other output cells just before the cell under question would fire to a particular stimulus. During the original training condition such activation of other output cells was preventing the cell under question from firing to exemplars of one of the two stimuli through strong within-layer inhibition (w_{ij}^{IE}). If this mechanism gets broken through weight shuffling, the cell under question would fire to both stimuli, thus changing from a very informative cell after original training, to an uninformative cell after weight shuffling. The cell under question may also change its stimulus preference after weight shuffling through the same process if the appropriate within-layer inhibition for the originally preferred stimulus gets switched on, while within-layer inhibition for the other stimulus gets switched off.

4.5.4 Weight Convergence During Training

When IC-A1 weight matrices (w_{ij}^{BL}) are compared at different snapshots spaced 20 s apart during training (lasting 95.9 seconds in total), it was found that the matrix was becoming more stable throughout training, whereby, while the exact weight magnitudes continued to fluctuate, the overall magnitude of weight changes declined with time, thus suggesting weight convergence (Fig. 4.22).

4.5.5 Testing AN-CN-IC-A1 Model on Novel Stimuli

In order to test whether the full AN-CN-IC-A1 network learnt to recognise the general class of vowels /a/ and /i:/, or whether it learnt the particular training examples only, the network was presented with 20 epochs of 12 exemplars of each of the new synthetic vowels /a/ and /i:/, whereby a new stimulus set was generated, which was perceptually still recognisable as vowels /a/ and /i:/, however its constituent formants had never been heard by the network in the past. It can be seen in Figure 3.6 (green) that when the trained AN-CN-IC-A1 network was tested on this new set of vowels, it still achieved good information content in its output layer (0.92 bits maximum). This suggests that the network learnt transform invariant representations of vowels /a/ and /i:/ that were also generalisable to new examples of these vowels that had never been heard by the network before.

In order to test whether the network was simply differentiating between two different classes of stimuli, rather than responding selectively to vowels /a/ and /i:/ in particular, another set of testing stimuli was constructed. This time the new testing set consisted of 20 epochs of 12 exemplars each of two new vowels, /ɜ:/ and /æ/. It can be seen in Figure 3.6 (yellow) that the network failed to differentiate

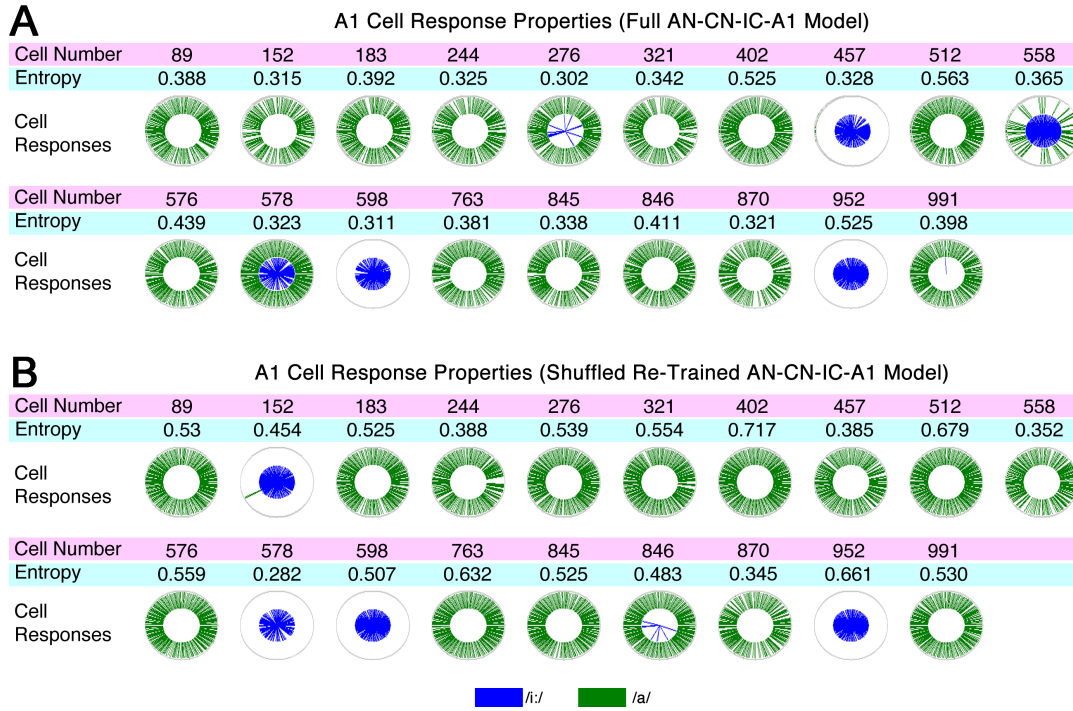


Figure 4.21: Schematic representation of response properties of 19 A1 cells common between the top 100 most informative cells of the originally trained full AN-CN-IC-A1 network, and the re-trained model once its IC-A1 afferents (w_{ij}^{BL}) were shuffled after original training. **A:** the amount of single cell information contained within A1 cells of the originally trained full AN-CN-IC-A1 model about the identity of the two input vowels /i:/ and /a/. **B:** the amount of single cell information contained within A1 cells of the shuffled re-trained AN-CN-IC-A1 model about the identity of the two input vowels /i:/ and /a/. The nested pie charts illustrate whether each cell fired to each of the 20 presentations of the 12 exemplars of each vowel, resulting in 240 segments. Coloured segments indicate that the cell discharged at least once in response to the corresponding stimulus presentation. Blue pie charts represent responses to /i:/ and green pie charts represent responses to /a/.

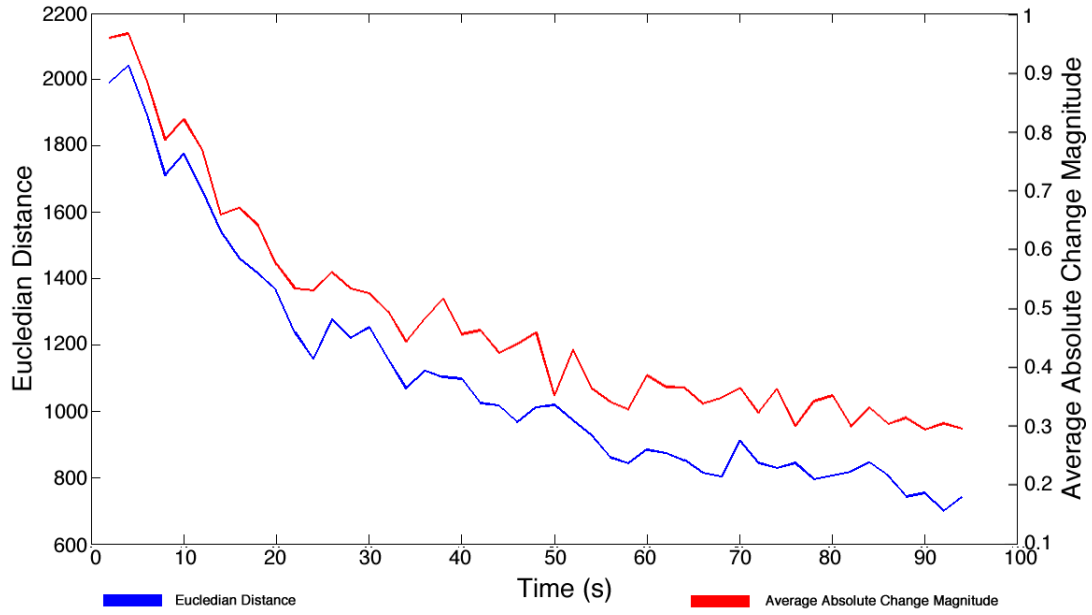


Figure 4.22: Measure of changes in afferent weight (w_{ij}^{BL}) magnitudes through STDP-based learning between the IC and A1 layers of the full AN-CN-IC-A1 spiking neural network model of the auditory brain during training. Twelve exemplars of each of the two 100 ms vowel stimuli separated by 100 ms gaps were presented over 200 epochs, taking 95.9 seconds. Euclidean distance (blue line, left scale) and average absolute weight change magnitude (red line, right scale) were calculated between afferent weight matrices taken every 20 s during training. It can be seen that both measures decrease throughout training, suggesting weight convergence.

between these two vowels, reaching a maximum of 0.19 bits of information out of 1 bits.

4.5.6 Importance of Formant Peaks In Differentiating Between Vowel Stimuli

Vowel stimuli /i:/ and /a/ were chosen as stimuli for the initial simulations due to their dissimilarity in the space characterised by the first three formant frequencies. The vowel also have stable spectra throughout their duration. Hence, the easiest way for a model to learn to discriminate between the two vowel classes is to learn to cue off just one of the formants. Figures 4.1 and 4.11 -4.12 suggest that this is indeed what has happened for the most informative A1 cells of the full AN-CN-IC-A1 model. This does not, however, mean that the model is only capable of learning to cue off single spectral features without learning any feature

combinations. This will be shown in the next two chapters, when the network will be exposed to up to six vowel stimuli with overlapping spectral forms, and to naturally spoken words. We will show that the model is able to learn complex spectro-temporal patterns of firing in response to the stimuli, however the encoding of the more complex information also becomes more complex, whereby precise timing of spikes within a population of neurons becomes important.

4.6 Change in Information Encoding Within the Auditory Brain

All information about the identity of the input stimuli is present within the AN layer, even if its single cell information scores are low. This has to be true, because otherwise the later brain areas would never be able to reach full single cell information, since throughout the hierarchy of the auditory brain no new information is introduced, and information can only be lost. This is in line with the information inequality rule, whereby any data processing can result in at most as much information as was present in the original data (Nelken 2004). If full information is present about the stimuli in the input AN layer, then the question arises as to the role of the rest of the auditory brain. It is hypothesised that the auditory hierarchy performs the following functions: 1) it links the separate AN representations for each stimulus exemplar/identity together into stimulus class representations, thus achieving better transform invariant representations of the stimuli than the AN; 2) it performs information compression by representing the stimulus class membership using a simpler encoding scheme than the AN.

An example of the first function would be the following. In the AN the twelve exemplars of vowel /i:/ would be encoded as twelve separate stimuli, while the same exemplars in the A1 would be linked together into a single stimulus class representation of being vowel /i:/. Evidence for the presence of this first function in the full AN-CN-IC-A1 model was presented in Chapters 3 and 4, whereby it was demonstrated that the output A1 layer of the AN-CN-IC-A1 model learns invariant representations of the two vowel stimuli. Such representations, however, are not present in the input AN layer of the model.

The second function is similar to the principle exploited by auto-encoders in the Machine Learning literature (Bengio 2009). Stimulus representations get compressed into a more sparse and simpler code in the A1 compared to the AN, which can only happen if the A1 learns some statistical regularities underlying the class membership of the stimuli and uses that to perform the lossy compression. The partial loss of information about the identity of each stimulus exemplar (vowel /i:/ produced with a particular combination of formants) is compensated by the

enhanced information about the stimulus class identity (the sound heard is vowel /i:/) in the A1 compared to the AN. Evidence for this second function of the auditory brain hierarchy is discussed in more detail below.

The AN contains more information than is necessary for auditory object recognition. Some of this information is necessary for the auditory object localisation task, and some of the information is redundant. Information within the AN is obscured by noise and spread over a very large population of cells. Furthermore, the information is potentially encoded simultaneously spatially, temporally and through rate encoding. Therefore, in order for the brain to act on this data, it needs to integrate and process different types of information from a large number of cells. It is hypothesised that the role of the auditory hierarchy is to split the information extraction task into smaller manageable sub-tasks, which are performed one after another. These smaller information extraction tasks would essentially transform the complex encoding system spread over a large population of cells in the AN, to a simpler encoding spread over a small subset of information bearing cells in the output layer. It is more neurophysiologically plausible for a cell in the readout brain area to receive afferents from a small information bearing subset of cells at the end of auditory brain hierarchy, rather than having to receive information from the thousands of AN fibers. Furthermore, a cell in the readout brain area receiving afferents from the smaller information bearing subset of output neurons would have to do less processing if the encoding it receives is a simple one. We demonstrate how such a readout might happen in Chapter 6, where an additional Belt layer is added on top of the A1.

4.6.1 Evidence for Encoding Simplification Within the Auditory Brain

Changes in average firing rate from approximately 37 Hz in the input AN layer to approximately 9 Hz in the IC layer and 2 Hz in the A1 layer of the full AN-CN-IC-A1 model of the auditory brain in response to the two vowel discrimination paradigm (see Fig. 4.2), provides evidence that information encoding changes from the more complex distributed rate or temporal code in the AN to the more simple localist on/off or binary code in the A1 (as described in Section 2.10.5). Further evidence comes from looking at rasters of firing of the most informative cells in the output A1 layer of the full AN-CN-IC-A1 model in response to the two vowel stimuli after training (see Figs. 4.5 and 4.7). These figures demonstrate that the cells fire mostly once for each stimulus presentation, suggesting on/off encoding. Such binary encoding has been described in the auditory cortex by DeWeese & Zador (2003), and is in line with the redundancy reduction hypothesis first proposed by Barlow (1961).

4.6.2 Evidence for Reduction in the Information Bearing Cell Subset Size Within the Auditory Brain

In order to test whether information encoding changes from being spread among a large population of AN layer cells to being spread among a small population of A1 layer cells, a multi-layer perceptron (MLP) neural network was used. The backpropagation network was trained on input vectors that represented activity patterns in the AN and A1 layers of the AN-A1 and AN-CN-IC-A1 models as they responded to the two vowels /i:/ and /a/. The network included a varied number of input cells (N_{in}) (depending on the size of the cell population used to create the input vectors), $N_{hid} = N_{in} * 0.1$ hidden layer cells and 2 output layer cells (one for each of the two stimulus classes) (see Fig. 4.23). The connections between the layers of the MLP were initialised randomly to small initial values. The network was tested on its ability to classify input vectors into those arising in response to one of the two stimuli vowels /a/ and /i:/. The same stimulus presentation paradigm was used as described in Section 2.2.1. The network was trained on 50% of the data, 15% of data was used for validation to avoid over-fitting, and the performance of the network was assessed on the 35% of testing data. The data were normalised before being presented to the network. Training stopped when the network started over-fitting, whereby its mean squared error rate on the validation set started to rise while the error rate on the training data continued to drop.

The MLP was used to explicitly demonstrate the reduction in the minimal cell subset size between the AN and the A1 of the AN-CN-IC-A1 model necessary to correctly categorise the input vowel stimuli. It was also used to show how the inability of STDP-based learning to produce good representations of vowel categories in the reduced AN-A1 model also translated into the larger minimal A1 cell subset sizes necessary for vowel categorisation compared to the A1 of the full AN-CN-IC-A1 model.

In order to test the hypothesis that the number of cells in the minimal subset of cells necessary to fully discriminate between the two vowels, /a/ and /i:/, becomes smaller throughout the auditory brain hierarchy, the backpropagation neural network was trained on data collected from the AN layer, the A1 layer of the simple AN-A1 model or the A1 layer of the full AN-CN-IC-A1 model in response to the two vowel discrimination paradigm. The network input vectors were binary, containing 1 for each cell that fired at least once during stimulus presentation, and 0 for each cell that never fired during stimulus presentation. The input cell population sizes (N_p) were varied from 120 cells down to 2. The upper bound (120) was picked as the total number of unique neurons within the union of the 100 most informative cells for each of the two stimuli (N_t , where $N_p \subset N_t$). Each population N_p was chosen by randomly sampling N_t . Such population sampling

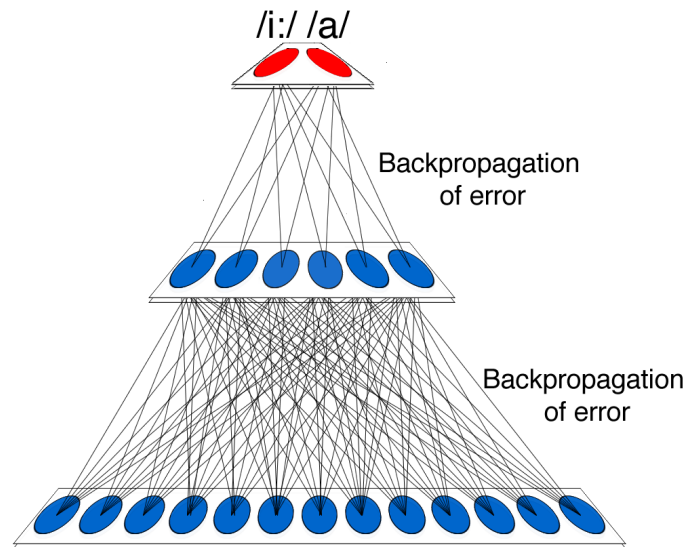


Figure 4.23: Schematic representation of a backpropagation neural network used to test the encoding scheme within the simple two-layer AN-A1 and full four-layer AN-CN-IC-A1 spiking neural network models of the auditory brain discussed in this chapter. The input vectors to the backpropagation network were patterns of activity taken from the AN and A1 layers of the AN-A1 and AN-CN-IC-A1 models as they responded to either of the two vowels, /i:/ or /a/, where an on/off encoding scheme was used to represent these activity patterns. The output to be learnt was the identity of the vowel, either /i:/ or /a/. The number of hidden layer nodes was always set to be 10% of the number of input nodes. There were two output nodes representing the two vowels.

was repeated 50 times, and for each sample the network was trained 20 times, each starting with different random weight initialisation, and the maximal performance from the 100 resulting training operations was recorded. As the input layer size changed to account for the cell population size, the number of hidden neurons also changed. It was set to be 10% of the input layer size for all simulations.

It was found that when tested on the AN layer, the network's performance accuracy dropped from about 99% for a population of 120 most informative cells to about 72% for a population of 2 most informative cells. For the A1 layer of the simple AN-A1 model, the network's performance accuracy dropped from about 97% for a population of 120 most informative cells to about 90% for a population of 2 most informative cells. For the A1 layer of the full AN-CN-IC-A1 model, the network's performance accuracy remained at 100% for any number of input cells between 120 and 2 (see Fig. 4.24). It can be seen that the performance of

the backpropagation network on the smallest population of 2 cells depends on the complexity of the spiking neural network model of the auditory brain considered, moving from 72% for the simplest model containing just the AN layer, to 90% for the simple AN-A1 model, and 100% for the full AN-CN-IC-A1 model. This supports the hypothesis that information encoding changes from being contained within a large population at the beginning of the auditory brain hierarchy, to being contained within a small population at the end of it. Essentially, in the full AN-CN-IC-A1 model, the encoding of the two vowels was transformed from a distributed representation in the AN layer to a more localist representation in the A1 layer, in which individual neurons responded selectively to only one of the two vowels.

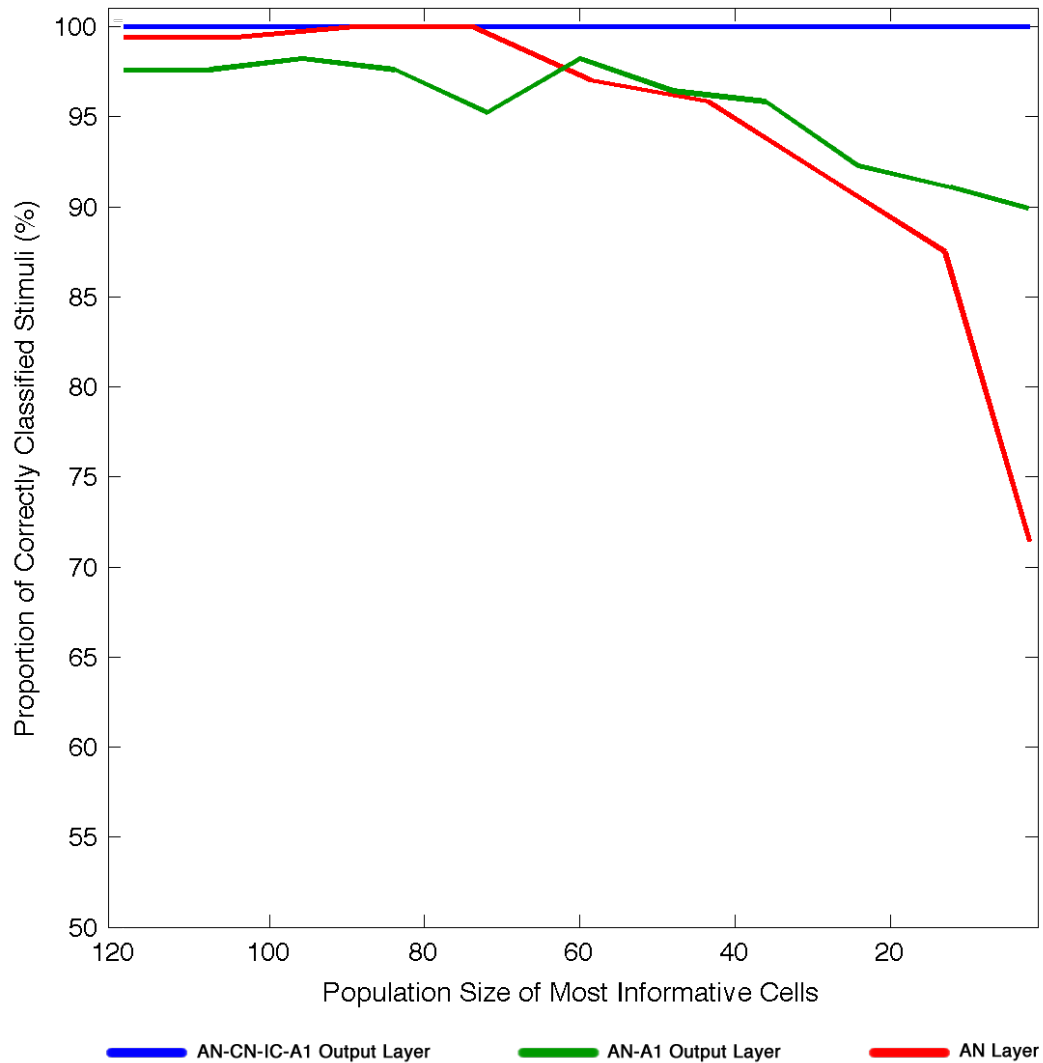


Figure 4.24: Performance of a backpropagation decoder neural network on the two vowel recognition task using data from the output A1 layer of the simple AN-A1 and full AN-CN-IC-A1 spiking neural network models of the auditory brain described in this chapter, as well as the input AN layer. The red line represents data from the input AN layer, green line represents data from the output A1 layer of the simple AN-A1 model, and blue line represents data from the output A1 layer of the full AN-CN-IC-A1 model. For each presentation of each exemplar of each vowel on/off encoding was used to transform spiking neural network firing rasters into backpropagation neural network input vectors. The size of these input vectors was diminished from the initial 120 cells to 2 cells. A number of different combinations of cells were picked and tested for each population size. Each combination was chosen from the hundred most informative cells about each of the two vowel stimuli. The best vowel classification performance for each population size was recorded. It can be seen that the decoder performance decreases with diminishing input size for the AN and AN-A1 conditions, but not for the AN-CN-IC-A1 network.

5

Learning Transform Invariant Representations of Six Vowels: the Importance of Temporal Coding

5.1 Introduction

Chapter 3 outlined two models of the auditory brain: a simple AN-A1 model, in which the primary auditory cortex (A1) receives afferent connections directly from the auditory nerve (AN); and a full AN-CN-IC-A1 model, where AN activity is pre-processed within the non-plastic sub-cortical cochlear nucleus (CN) and inferior colliculus (IC) layers, before being sent up to the output layer A1. It was shown that the simple AN-A1 model was unable to learn transform invariant representations of auditory objects, such as vowels, because of the jitter present in the AN spike rasters. This jitter was, however, removed through a two-stage process beginning within the sub-populations of the CN and finishing within the IC, which allowed the full AN-CN-IC-A1 model to learn stimulus selective transform invariant representations of two synthesised vowels, /i:/ and /a/.

Chapter 4 looked in more detail into the mechanisms that enable the full AN-CN-IC-A1 model to learn transform invariant representations of auditory objects, such as vowels. The chapter discussed the changes happening throughout the hierarchy of the model and characterised the response properties of the most informative cells within the model's output layer A1.

The simulations so far, however, have looked into learning transform invariant representations of only two synthesised vowels. Such simple stimuli were well suited for investigating whether extra sub-cortical pre-processing is necessary to remove AN jitter and enable transform invariance learning to take place within the output layer A1. It is important, however, to understand whether the full AN-CN-IC-A1 model can cope with more realistic situations where the number of stimuli is increased. This chapter addresses this question by presenting up to six synthesised vowels to the network.

5.2 Stimuli

In order to understand whether the full AN-CN-IC-A1 model as parametrised in Table 3.4 is able to learn transform invariant representations of more than two vowels, six vowel stimuli were generated using the Klatt synthesiser (Klatt 1980) as described in Chapter 2 Section 2.2.2.

5.3 Training With the New Vowel Sets

In order to avoid the extra computations of having to test the performance of the network with all combinations of vowels in the 3-5 set size range, we ran a number of simulations in the fashion described next (the two- and six-vowel sets were pre-determined, whereby the original two vowels /i:/ and /a/ were kept in the two-vowel set, and the six-vowel set contained all synthesised vowels). For each simulation the number of vowels presented to the full AN-CN-IC-A1 network was increased from the initial two, /i:/ and /a/, up to the maximum of all six vowels from the newly generated set. Each time the set size of the vowels presented to the network was incremented using the vowel with the best network performance. For example, four different simulations were run in order to find the optimal set of three vowels (see Table 5.1). Each simulation contained the original two vowels, /i:/ and /a/, combined with each of the four potential options for the third vowel: /ɔ:/, /u:/, /e/, or /æ/. Maximal single cell information (see Chapter 2 Section 2.10.4) within the output layer (A1) of the AN-CN-IC-A1 model was used as the measure of network performance. Simulation results can be seen in Table 5.1. It can be seen that the performance of the network is not significantly dissimilar for the different sets of vowels, however we always chose the set with the best network performance for further analysis in this chapter. Table 5.2 shows the chosen sets of vowels for each stimulus set size. As the number of vowels in the training set grows, the maximal theoretical amount of information carried by the cells in the output layer of the full AN-CN-IC-A1 model also increases (see Table 5.2). The

		Single Cell Information (Bits)			
		/ɔ:/	/u:/	/e/	/æ/
3 vowels	/i:/ + /a/ +	0.876	0.802	0.892	0.89
4 vowels	/i:/ + /a/ + /e/ +	0.762	0.891	NA	0.752
5 vowels	/i:/ + /a/ + /e/ + /u:/ +	0.72	NA	NA	0.828

Table 5.1: The full AN-CN-IC-A1 model was trained on different sets of vowels starting from a set of size two that included the original vowels /i:/ and /a/ used in the simulations described in the preceding chapters, and finishing with the set of six vowels, whereby additional vowels /ɔ:/, /u:/, /e/, and /æ/ were included. As the size of the vowel set was incremented, the optimal vowel to add to the set was found through measuring the maximal single cell information within the output layer (A1) of the full AN-CN-IC-A1 model. This table summarises the A1 maximal single cell information after the model was trained on different vowel combinations for each set size. Values shown in bold are maximal, and thus the corresponding vowel was found to be optimal for the particular set size. Table 5.2 shows the finalised optimal sets of vowels for each stimulus set size.

Set Size	Set Vowels	Maximum Information (Bits)
Two vowels	/i:/, /a/	1
Three vowels	/i:/, /a/, /e/	1.6
Four vowels	/i:/, /a/, /e/, /u:/	2
Five vowels	/i:/, /a/, /e/, /u:/, /æ/	2.3
Six vowels	/i:/, /a/, /e/, /u:/, /æ/, /ɔ:/	2.6

Table 5.2: Vowels used to train the full AN-CN-IC-A1 and the simple AN-A1 models of the auditory brain as the size of the training set increased from two to six vowels. The last column contains the maximum theoretical information for each stimulus set size, which is computed as $\log_2 S$ bits, where S is the stimulus set size (number of vowels).

maximal theoretical amount of information was calculated as $\log_2 S$, where S is the size of the stimulus set with the equiprobable stimuli.

5.4 Changes in Vowel Representations With Increased Stimulus Set Sizes

Figure 5.1 shows the maximal single cell information within the output (A1) layer of the full AN-CN-IC-A1 network in proportion to the theoretical maximum single

	Stimulus 1	Stimulus 2...	...Stimulus S
On	$\frac{\#CellOn2Stim1}{\#Stim1*S}$	$\frac{\#CellOn2Stim2}{\#Stim2*S}$	$\frac{\#CellOn2StimS}{\#StimS*S}$
Off	$\frac{\#CellOff2Stim1}{\#Stim1*S}$	$\frac{\#CellOff2Stim2}{\#Stim2*S}$	$\frac{\#CellOff2StimS}{\#StimS*S}$

Table 5.3: Joint probability table used to calculate single cell information for each cell using on/off encoding for more than two vowels. $\#CellOn2Stim1$ ($\#CellOn2StimS$) stands for the number of presentations of stimulus 1 (S) during which the cell fired at least once. Similarly $\#CellOff2Stim1$ ($\#CellOff2StimS$) stands for the number of presentations of stimulus 1 (S) during which the cell did not fire. $\#Stim1$ ($\#StimS$) stands for the total number of presentations of stimulus 1 (S). S stands for the total number of stimuli in the training/testing set.

cell information when the network was presented with the transform invariant vowel recognition problem, where the number of vowels presented to the network increased from two to six. Single cell information was calculated using the simple on/off encoding similarly to the method described in Chapter 2 Section 2.10.4, however the joint probability table was modified to accommodate the additional vowel stimuli (Table 5.3).

It can be seen that as the number of vowels in the training set increases, single cells in the output layer (A1) of the full AN-CN-IC-A1 model become less informative. They are still, however, more informative than the A1 cells within the full AN-CN-IC-A1 model before training or the output A1 cells of the simple AN-A1 model before or after training (see Tables 3.1 and 3.2 (Model 2) for the parametrisation of the simple AN-A1 model). This provides further support for the main hypothesis of this thesis that extra sub-cortical pre-processing within the CN and IC is necessary to enable transform invariant learning of auditory objects, including vowels. It can also be seen that the intermediate CH layer reaches slightly more information than the IC or the output A1 layers of the full AN-CN-IC-A1 model. This is an artefact of the interaction between the relatively simplistic vowel stimuli and the on/off encoding scheme. The reason why information drops between the CH and the IC is because the inputs from the ON and PL sub-populations of the CN reduce the information content of the CH inputs. The reason why information content drops between the IC and the A1 is because the simple on/off encoding scheme might not be appropriate for representing the extra vowel stimuli, as discussed next.

Figure 5.1 shows that the amount of maximal single cell information within the A1 cells of the full AN-CN-IC-A1 model decreases rapidly from 100% to around

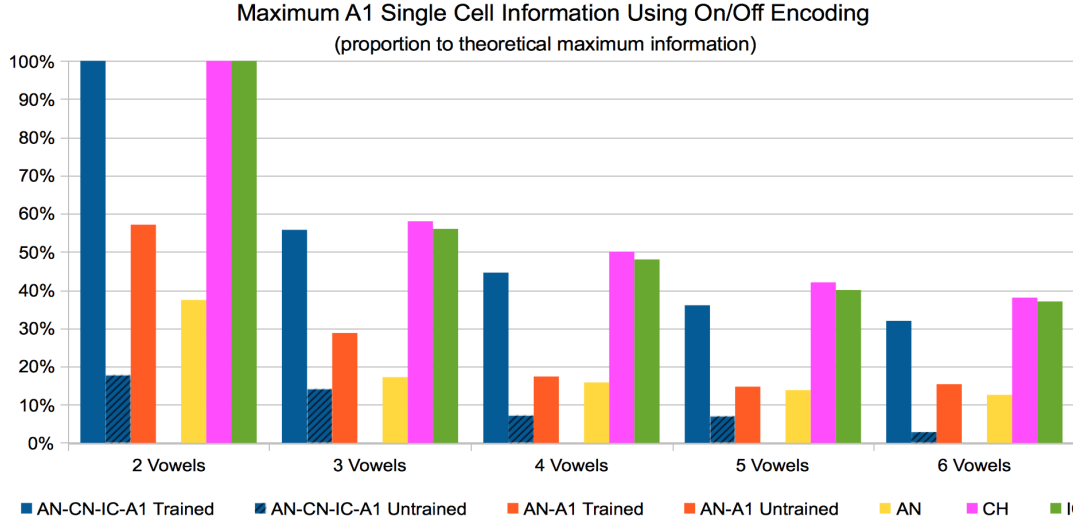


Figure 5.1: Maximal single cell information in the output (A1) layer of the full AN-CN-IC-A1 (blue) and the simple AN-A1 (red) models of the auditory brain in proportion to the theoretical maximum information for the different stimulus set sizes (as shown in Table 5.2). Data are shown for both models before (hatched bars) and after (solid bars) training and in response to two to six vowel stimuli. Each data point represents the best performing cell in the output A1 layer. Similarly data for the input AN (yellow), and intermediate CH (CN) (magenta) and IC (green) layers are presented. Data for ON and PL sub-populations of the CN are not presented, since ON does not carry any information about the vowel class identity, while PL is essentially identical to AN. It can be seen that the full AN-CN-IC-A1 model reaches more maximal single cell information in its output layer (A1) than the simple AN-A1 model, thus supporting the hypothesis that sub-cortical pre-processing is necessary for transform invariant auditory object recognition, including vowels. The output A1 layer of either model results in more information compared to the input AN layer. It can also be seen that the intermediate CH layer reaches slightly more information than the IC or the output A1 layers of the full AN-CN-IC-A1 model. This is an artefact of the interaction between the relatively simplistic vowel stimuli and the on/off encoding scheme.

30% of the theoretical maximum ($\log_2 S$, where S is the stimulus set size) as the number of vowels in the training set is increased from two to six. This suggests either the loss of transform invariance or that the type of stimulus encoding used by the network becomes more distributed with the increased stimulus number. Distributed encoding presumes that stimulus information is encoded within the firing of a population of cells rather than one cell, whereby each cell in the population responds to more than one auditory object, but enough information is present within the population of cells to perfectly identify each individual vowel.

5.5 Quantifying Distributed Encoding

In order to test whether the distributed representation in the output layer (A1) of the full AN-CN-IC-A1 model of the auditory brain contains enough information to differentiate between the different vowels within the training set, the following procedure was performed. Response properties of 125 A1 cells with the most single cell information were considered. During each stimulus presentation the most likely stimulus identity vote was assigned to each cell based on on/off encoding. The vote was decided by randomly sampling from a multinomial distribution with s categories, where $s \in S$ is the number of vowels presented to the network. The probabilities assigned to each of the s categories were calculated by counting the number of times the cell fired in response to the presentation of a particular stimulus s and normalising by the total number of spikes the cell emitted during testing:

$$p(s) = \frac{spikes_s}{\sum_s spikes_s}$$

Once the votes for all of the 125 A1 cells were assigned, the population decision was considered. The population vote for each stimulus presentation was chosen to be the majority vote of the 125 chosen A1 cells. This was supposed to emulate the information received by the hypothetical next layer of the auditory brain, which would have to decide on the vowel identity based on the majority votes received from the A1. The stimulus identity decoding in this hypothetical next layer would be performed in the following manner. Each neuron in this next layer would essentially act as the output neuron in the model by Masquelier et al. (2008). They will receive A1 spikes, some of which will be information bearing (i.e. the ones arriving from the 125 A1 cells chosen above), and others not. The information bearing spikes will essentially be equivalent to the spikes making up the repeating pattern in Masquelier et al. (2008), while the others will essentially constitute the statistically equivalent noise described in Masquelier et al. (2008). If the majority of the 125 chosen A1 cells vote for the same stimulus identity,

		Actual	
		Stimulus 1...	...Stimulus S
Predicted	Stimulus 1...	$\frac{\#Response1}{\#Stim1*S}$	$\frac{\#ResponseS}{\#StimS*S}$
	...Stimulus S	$\frac{\#ResponseS}{\#Stim1*S}$	$\frac{\#ResponseS}{\#StimS*S}$

Table 5.4: Joint probability table used to calculate cell information when more than two cell response types were considered. $\#Response(s)$ stands for the number of presentations of stimulus $s \in S$ during which cell response was classified as voting for stimulus s (see Section 5.5). $\#Stim(s)$ stands for the total number of presentations of stimulus s . S stands for the total number of stimuli.

then it will be equivalent to the hypothetical next layer cell receiving more input belonging to one pattern, and hence the next layer cell should learn through the same mechanisms as described in Masquelier et al. (2008) to become selective for the particular stimulus identity. We will show in the next chapter that this is in fact what happens in the network once we attach a Belt layer on top of the A1 layer.

Information analysis was performed on the resulting population votes as described in Chapter 2 Section 2.10.4. The confusion matrix, however, was calculated differently due to the increased number of possible cell responses (S) compared to the two possible cell responses for the simple on/off encoding scheme described in Chapter 2 Section 2.10.4. Table 5.4 gives an overview of the way the new confusion matrix was calculated.

5.5.1 Multiple Cell Information

Figure 5.2 demonstrates multiple cell information for the population of 125 most informative A1 cells of the trained and untrained versions of the AN-CN-IC-A1 and the AN-A1 models of the auditory brain in response to 2-6 synthesised vowels. Each value is presented as a proportion of the maximal theoretical information for the particular stimulus set size ($\log_2 S$, where S is the stimulus set size).

It can be seen that the population of the 125 most informative A1 cells of the full AN-CN-IC-A1 model after training reaches 100% accuracy when differentiating between two vowels. Its accuracy then drops dramatically to around 20%. This is still, however, significantly more than the 1% achieved by the trained AN-A1 model and 0% achieved by the untrained models, which suggests that the original

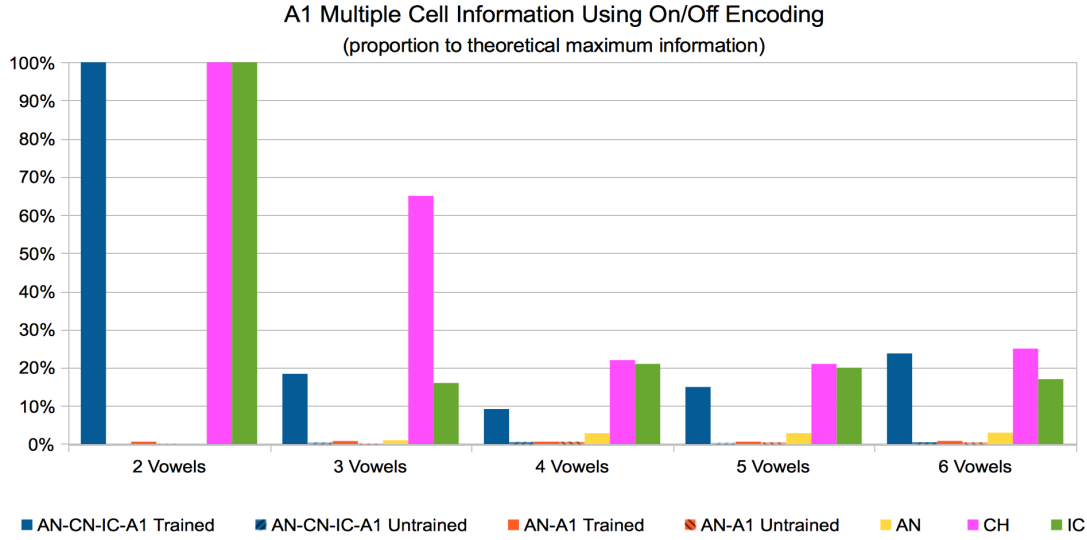


Figure 5.2: Multiple cell information for the population of 125 most informative cells in the output (A1) layer of the full AN-CN-IC-A1 (blue) and the simple AN-A1 (red) models of the auditory brain. Simple on/off encoding is assumed (see Section 5.4). Data are shown for both models before (hatched bars) and after (solid bars) training and in response to two to six vowel stimuli. Similar data are presented for the input AN (yellow), and intermediate CH (CN) (magenta) and IC (green) layers. Data for ON and PL sub-populations of the CN are not presented, since ON does not carry any information about the vowel class identity, while PL is essentially identical to AN. It can be seen that the full AN-CN-IC-A1 model reaches higher accuracy than the simple AN-A1 model, thus supporting the hypothesis that sub-cortical pre-processing is necessary for transform invariant auditory object recognition, including vowels. Furthermore, the output A1 layer of either model results in more information compared to the input AN layer. It can also be seen that the intermediate CH layer reaches significantly more information than the IC or the output A1 layers of the full AN-CN-IC-A1 model. This is an artefact of the interaction between the relatively simplistic vowel stimuli and the on/off encoding scheme, whereby the population of the 125 most informative CH neurons is unambiguous in its stimulus classification voting, while the similar IC and A1 populations are more unsure about the stimulus classification.

hypothesis for the importance of sub-cortical pre-processing still holds.

The reason why multiple cell information is lower than single cell information shown in Figure 5.1 for more than two vowels is that the cells within the population provide conflicting votes about the identity of the stimuli, whereby some correct single cell votes get outnumbered by incorrect votes by other cells in the population.

The low levels of information reached by the on/off population encoding sug-

gests that a different encoding scheme might be used by the network to encode more than two synthesised vowels. The apparent inadequacy of the on/off encoding scheme for representing multiple vowel classes is also demonstrated by the significantly higher amounts of multiple cell information within the CH layer of the full AN-CN-IC-A1 model. The nature of this layer makes it highly suitable for representing stimulus classification information using on/off type encoding (see Chapter 4 Section 4.3 for details). The convergence of the CH inputs with the ON and PL inputs within the IC, however, results in a more complex encoding scheme. It is important to understand which encoding scheme is likely to be used within the output layer of the AN-CN-IC-A1 model to differentiate between more than two vowel stimuli. The network performance on the vowel recognition task was therefore analysed in more detail.

5.6 Understanding Stimulus Preferences of the full AN-CN-IC-A1 Model

5.6.1 Stimulus Preference Plots

Stimulus preference plots were created in order to understand how the nature of stimulus selectivity changes within the most informative A1 cells of the full AN-CN-IC-A1 model before and after training, and with the increase in the vowel training set size (Figs. 5.3). Stimulus preference plots help visualise the proportion of presentations of all exemplars of each stimulus that produced one or more spike in a cell. Such plots help de-tangle the cell's responsiveness to preferred stimuli (RPS) from its non-responsiveness to non-preferred stimuli (NRNS), which are both encoded within the single cell information measure.

RPS measures how many times a cell fires at least one spike during all the different presentations of a particular stimulus. For example, an A1 cell would have 75% RPS for vowel /a/ if it fires at least one spike when presented with 180/240 versions of vowel /a/ (240 corresponds to 20 epochs of each of the 12 exemplars of the vowel).

NRNS measures how well a cell is tuned to respond to its preferred vowel only. For example, an A1 cell would have 75% NRNS for its preferred vowel /a/ if it fires at least one spike in response to 120/480 presentations of vowels /i:/ and /e/ in a three vowel training paradigm (480 corresponds to 20 epochs of each of the 12 exemplars of each of the vowels /i:/ and /e/).

Maximum single cell information is achieved if both RPS is at 100% (the cell responds to all presentations of all exemplars of its preferred stimulus) and NRNS is at 100% (the cell responds to no presentations of any exemplars of the non-preferred stimuli). Stimulus preference plots de-tangle and highlight the cell RPS

and NRNS, thus helping visualise how much a cell has learnt about each of the individual stimuli in the training set.

5.6.2 Changes in Stimulus Preference

It can be seen in Figure 5.3 that the nature of stimulus preferences changes in the output layer (A1) of the full AN-CN-IC-A1 model after training. Whereas before training cells have similar RPS for all stimuli, after training each cell becomes tuned to a number of preferred stimuli which keeps the same high RPS for a subset of preferred stimuli, but also increases the cell NRNS. The role of training, therefore, is for the cells to become un-responsive to a subset of stimuli while remaining responsive to others.

Figure 5.4 illustrates stimulus preference of the most informative cells of the AN layer. It can be seen that due to the formant frequency compositions of the six vowels, certain AN cells tend to fire to most presentations of vowels /a/, /u:/ and /ɔ:/. These cells, however, also fire to a large number of presentations of the other three vowels. Throughout the hierarchy of the full AN-CN-IC-A1 model the network increases the cell NRNS. Cells in A1 emerge that respond to different vowels preferentially, whereby they respond to many exemplars of their preferred vowel(s) while remaining primarily unresponsive to the other non-preferred vowels (Fig. 5.3).

5.6.3 Stimulus Preferences for Distributed Encoding

Figure 5.3 shows that the output layer (A1) cells of the full AN-CN-IC-A1 model have indeed learnt distributed representations of vowels, whereby each vowel is represented by a distinct population of A1 cells, and each A1 cell responds to a number of vowels. In order for the next hypothetical cortical layer to accurately discriminate between the different vowel stimuli using the distributed encoding within the A1 layer of the full AN-CN-IC-A1 model, it is important that the different A1 cells have diverse response properties. For example, if all A1 cells have the same response properties, whereby they all respond to all presentations of vowels /a/ and /u:/ and to no presentations of vowel /i:/ in a three vowel recognition paradigm, then the next cortical layer will not be able to differentiate between vowels /a/ and /u:/. However, if one A1 cell (C1) responds to presentations 1-120 of vowel /a/ and presentations 121-240 of vowel /u:/, and another A1 cell (C2) responds to presentations 1-120 of vowel /a/ and presentations 1-120 of vowel /u:/, then a cell in the next cortical layer can learn that if both A1 cells C1 and C2 are firing, then vowel /a/ is being presented, while the firing of only one A1 cell, either cell C1 or cell C2, suggests that vowel /u:/ is being presented.

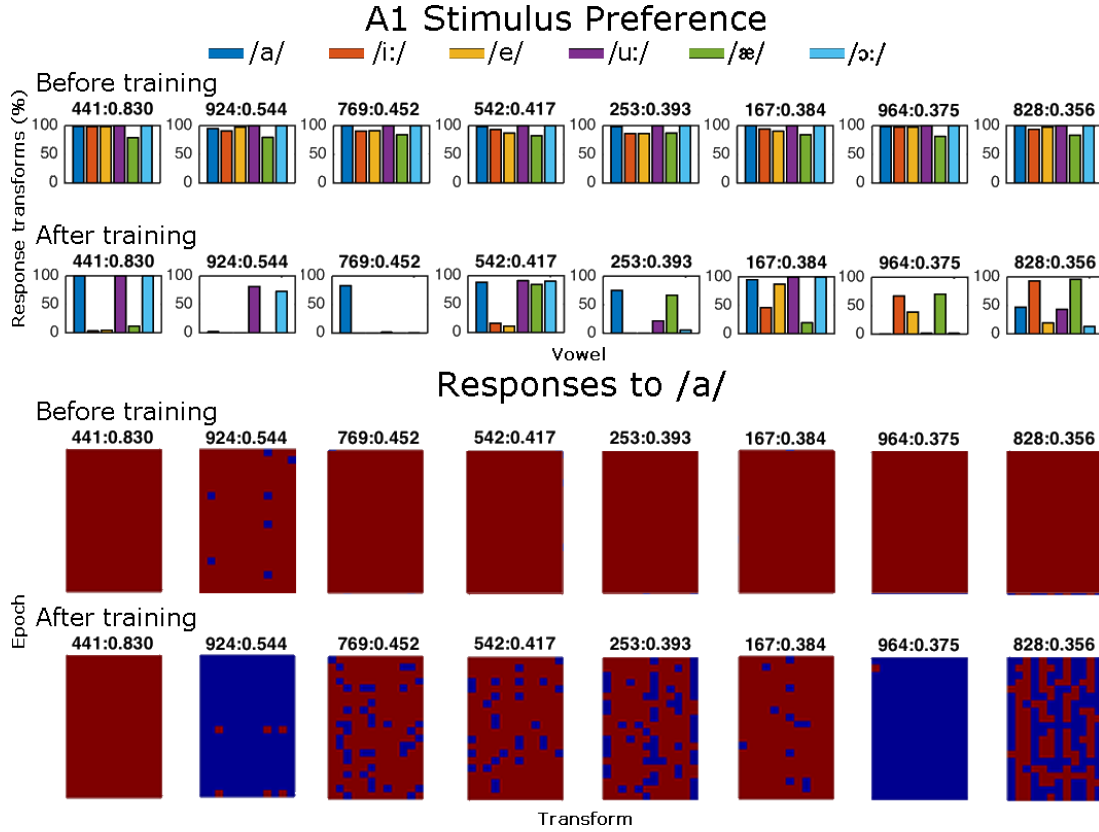


Figure 5.3: Stimulus preference plots for eight A1 cells of the full AN-CN-IC-A1 model of the auditory brain trained to differentiate between six vowels. The **top row** represents cell response properties before training, while the **second row** represents response properties of the same cells after training. Each bar plot shows the proportion of all stimulus presentations (20 epochs of 12 exemplars resulting in 240 presentations of each vowel during testing) that the cell responded to with one or more spikes. At the top of each graph is the cell number and single cell information carried by the cell calculated as per Chapter 2 Section 2.10.4. It can be seen that cell response properties change after training, whereby cells become more selective as to which vowels they respond to, thus increasing the NRNS while keeping the RPS for their preferred stimuli high. It can also be seen that distributed representations are used in the output layer of the network, whereby each cell responds to a number of vowels and different cells have varied stimulus preference profiles (**middle row**). For example, cells 441, 769, 542, 253, 167 and 828 are all responsive to /a/, albeit with a varying degree of selectivity and invariance. The differences in the preference profiles for other stimuli of these cells allow the hypothetical next layer of the brain as modelled by the backpropagating neural network to learn a complete transform invariant and stimulus selective representation of vowel /a/. A population of 300 most informative A1 cells provides enough information to achieve 98% vowel recognition accuracy by the backpropagating neural network. **Last two rows** demonstrate which of the 240 presentations of vowel /a/ each cell responds to before (row three) and after (bottom row) training. Each subplot is a matrix of 12 vowel exemplars (abscissa) presented over 20 epochs (ordinate). Red signifies that the cell fired one or more times during the stimulus presentation, while blue signifies that the cell did not fire. It can be seen that before training all cells have high levels of transform invariance for vowel /a/ (high RPS). This changes for cells 924 and 964 after training, as these cells learn to be selective to other stimuli instead. Training results in cells like 253 and 828 forgetting about certain exemplars of vowel /a/ during training, as evidenced by vertical blue lines in their sub-plots.

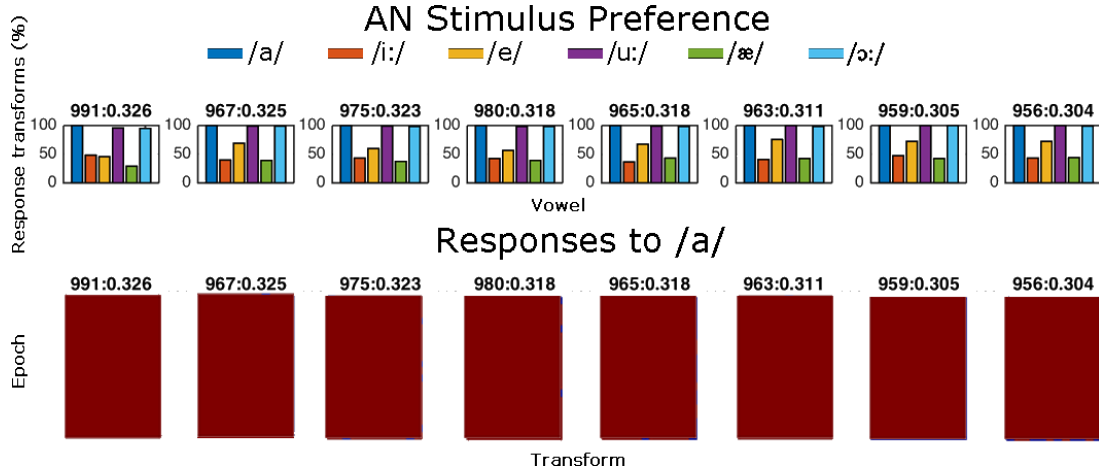


Figure 5.4: Stimulus preference plots for eight most informative (calculated as per Chapter 2 Section 2.10.4) AN cells when presented with six vowels. Each bar in the **top row** shows the proportion of all stimulus presentations (20 epochs of 12 exemplars resulting in 240 presentations of each vowel during testing) that the cell responded to with one or more spikes. At the top of each graph is the cell number and single cell information carried by the cell. The **bottom row** demonstrates which of the 240 presentations of vowel /a/ each cell responds to. Each subplot is a matrix of 12 vowel exemplars (abscissa) presented over 20 epochs (ordinate). Red signifies that the cell fired one or more times during the stimulus presentation, while blue signifies that the cell did not fire. It can be seen that the AN cells have a natural preference for some vowels (/a/, /u:/ and /ɔ:/) than others (/i:/, /e/ and /æ/) due to the formant distributions of the vowels. Each of the presented AN cells naturally fires to all presentations of vowel /a/. The role of the hierarchy of the auditory brain as simulated by the full AN-CN-IC-A1 model is to reduce the cells' responses to the non-preferred stimuli (/i:/, /e/ and /æ/), while maintaining their stable response properties to the preferred stimuli (/a/, /u:/ and /ɔ:/). Furthermore, some diversity in the cells' response properties has to be introduced to ensure that a population of output A1 cells can differentiate between all six vowels.

The diversity in the A1 cells' response properties of the full AN-CN-IC-A1 model trained on six vowels can be seen in Fig. 5.3 (second row). The figure shows some of the A1 layer cells that have high selectivity for vowel /a/ (cells 441, 769, 542, 253, 167 and 828). These cells also respond to a combination of other vowels, whereby each cell has a different stimulus preference profile. For example, cell 441 responds to vowels /a/, /u:/ and /ɔ:/, while cell 253 mostly responds to vowels /a/ and /æ/.

Each of the plotted A1 cells also has a different response profile for the exemplars of its preferred vowel /a/. The bottom row of Figure 5.3 shows that during

training of the full AN-CN-IC-A1 model as the NRNS is improved, the RNS of some A1 cells that are selective to /a/ drops (cells 769, 542, 253 and 167). These cells develop a number of exemplars of /a/ that they do not respond to during many presentation epochs of that exemplar, thus forming vertical blue lines. The A1 cells have therefore lost some transform invariance in order to gain better stimulus selectivity. The non-preferred exemplars of /a/, however, are different for the different A1 cells, thus allowing the whole population of A1 cells to respond to all exemplars of vowel /a/.

5.7 Understanding Cell Firing Patterns

This section examines whether extra information is present within the precise firing times of A1 cells in response to the different vowels to introduce enough diversity in cell responses to de-tangle linked vowel representations and enable the hypothetical next layer of the auditory brain to differentiate between the synthesised vowel stimuli. Figure 5.5 shows firing rasters of five A1 cells of the full AN-CN-IC-A1 model in response to four vowels, /a/, /i:/, /e/ and /u:/. Stimulus preference plots for each cell are also shown at the top of each firing raster.

Stimulus preference plots for cells 255, 422 and 340 indicate that they fire to a large number of exemplars of around three vowels, which makes them have low single cell information based on on/off encoding. When the cells' firing rasters are examined, however, it can be seen that the cells fire at different times for different stimuli. For example, cell 255 has a lot more information about the identity of the four vowels if the cell spike time information is taken into account than if on/off encoding is used. If the cell fires a single spike around 20 ms after the beginning of the stimulus, then it is likely that vowel /u:/ is being presented; if the cell fires a single spike around 60 ms after the beginning of the stimulus, then it is likely that vowel /e/ is being presented; if the cell fires both at 20 ms and at 60 ms after the beginning of the stimulus, then it is likely that vowel /i:/ is being presented; finally if the cell does not fire, then it is likely that vowel /a/ is being presented.

It is also interesting to note the response properties of cells 150 and 991. It can be seen that both cells have a strong preference for vowels /i:/ and /e/, however the two cells fire at different times with respect to stimulus start (around 40 ms and 20 ms correspondingly). The cells tend to fire at the same time for most of the presentations of their preferred two stimuli, thus forming a stable temporal firing pattern constituting a part of a polychronous chain (Izhikevich 2006). Such a stable temporal firing pattern can then be learnt through STDP-based learning by the cells in the hypothetical next layer of the auditory cortex (Masquelier et al. 2008).

Figure 5.5 suggests that as the full AN-CN-IC-A1 model of the auditory brain

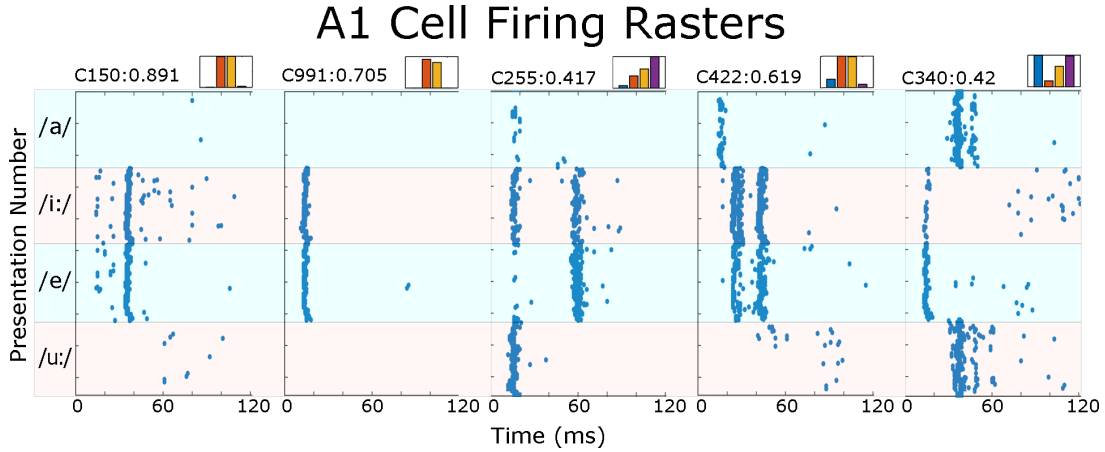


Figure 5.5: Firing rasters of five cells within the A1 layer of the full AN-CN-IC-A1 model of the auditory brain in response to four vowels: /a/, /i:/, /e/ and /u:/. The abscissa is the time (ms) from stimulus start. The ordinate represents stimulus presentation number, whereby each vowel was presented 240 times (20 epochs of each of the 12 exemplars for each vowel). Each dot within a raster represents a spike produced by the cell. Cell index as well as its single cell information using on/off information encoding is provided at the top of each firing raster. Cell stimulus preference plots (calculated as per Chapter 2 Section 2.10.4) are also shown at the top of each firing raster. It can be seen that while all of the cells respond to many of the presentations of vowels /i:/ and /e/, the cells fire at different times after the stimulus presentation start. The precise stable firing patterns produced by populations of such cells in response to different vowels serve as part of polychronous chains (Izhikevich 2006), which increase the network learning capacity and provide more information about the identity of the learnt vowels compared to the on/off encoding.

is exposed to more vowels, its transform invariant representations become more distributed and the nature of vowel representations learnt by output layer (A1) of the model are no longer sufficiently represented by the simple grandmother-like on/off encoding (see Chapter 2 Section 2.10.5). Instead a temporal encoding scheme looks more likely, whereby the exact timing of spikes produced by each cell within a population starts to matter. When temporal information about the precise timing of a cell's spikes are taken into account, a lot more information becomes available to the next layer of the cortex about the identity of the stimulus being presented. All of the five cells shown in Figure 5.5 respond to vowels /i:/ and /e/, however they fire with different stable temporal patterns to the two stimuli.

5.8 Using Temporal Patterns of Firing of Single Cells

5.8.1 Bayesian Single Cell Information

In order to test whether more information was present about the identity of the Klatt generated vowels within the precise timing of spikes of A1 cells, maximal single cell information was calculated using the approach described in Chapter 2 Section 2.10.5.2 and shown in proportion to the theoretical maximum information ($\log_2 S$) for each stimulus set size S in Figure 5.6.

It can be seen that more information about the synthesised vowel stimuli is present in the A1 cells of the models when the precise firing times of individual cell spikes are considered compared to on/off encoding (Fig. 5.1). On average there is around 9% more single cell information about the identity of the vowel stimuli in the spiking responses of A1 cells of the AN-CN-IC-A1 and AN-A1 models compared to on/off encoding. This number, however, is significantly higher (48%) for the untrained AN-CN-IC-A1 model. It can be seen that single cell information of the untrained AN-CN-IC-A1 model reaches the level of the trained model. This suggests that if precise spike times of individual cells are taken into account, then IC→A1 plasticity is not as important for the particular simple task of developing transform invariant representations of synthesised vowel stimuli as the sub-cortical pre-processing within the CN and IC. It will be shown in the next chapter, however, that learning does play a significant role in the development of representations of naturally spoken words.

The examples of firing rasters of typical A1 cells in response to four vowels of the untrained AN-CN-IC-A1 model and both trained and untrained AN-A1 model can be seen in Figure 5.7. This figure shows that a typical A1 cell of the untrained AN-CN-IC-A1 model does develop similar diverse firing response properties in response to four synthesised vowels as does a trained AN-CN-IC-A1 model. This is not the case for the typical A1 cells of the trained or untrained AN-A1 model, which respond in a similar way to all four vowels.

Figure 5.6 demonstrates that the amount of information increases throughout the hierarchy of the trained AN-CN-IC-A1 model apart from the slight drop in information between the CH and IC layers. This drop is due to the convergence of the less informative ON and PL inputs on the IC in addition to the CH inputs. The overall information increase trend throughout the hierarchy of the AN-CN-IC-A1 model, however, suggests that the Bayesian temporal information encoding scheme described in Chapter 2 Section 2.10.5.2 is more appropriate for differentiating between the vowel stimuli than the simple on/off encoding scheme.

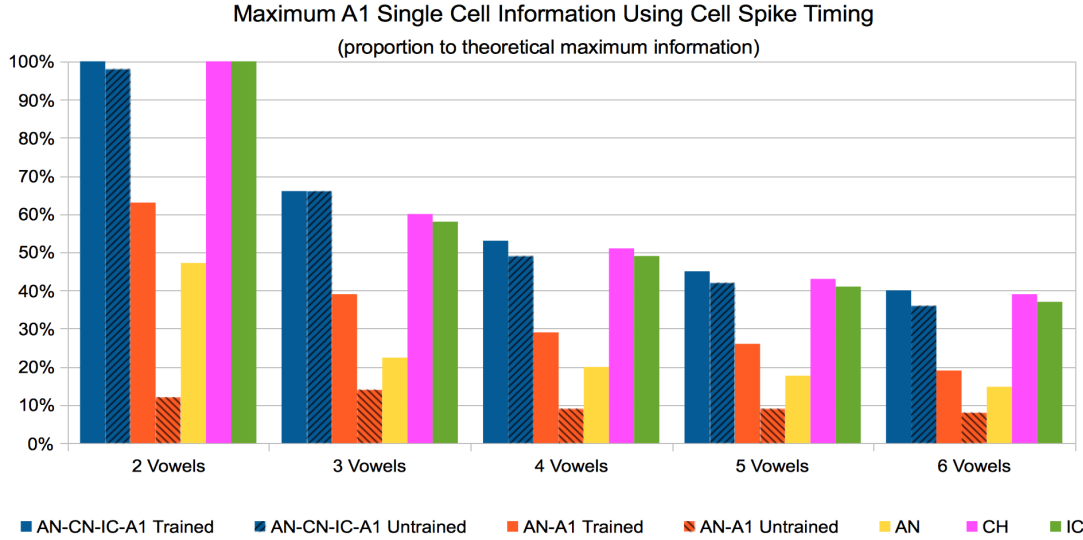


Figure 5.6: Maximal single cell information in the output (A1) layer of the full AN-CN-IC-A1 (blue) and the simple AN-A1 (red) models of the auditory brain in proportion to the theoretical maximum information for the different stimulus set sizes (as shown in Table 5.2). Data are shown for both models before (hatched bars) and after (solid bars) training and in response to two to six vowel stimuli. It can be seen that the full AN-CN-IC-A1 model reaches more maximal single cell information in its output layer (A1) than the simple AN-A1 model, thus supporting the hypothesis that sub-cortical pre-processing is necessary for transform invariant auditory object recognition, including vowels. The proportion of single cell information, however, drops as the number of vowels within the training set is increased from two to six, suggesting the emergence of distributed encoding. This figure demonstrates that single cell information encoding using the precise spike timing of A1 cells with respect to stimulus onset carries more information than the simple on/off encoding as plotted in Figure 5.1. Similar data are presented for the input AN (yellow), and intermediate CH (CN) (magenta) and IC (green) layers. Data for ON and PL sub-populations of the CN are not presented, since ON does not carry any information about the vowel class identity, while PL is essentially identical to AN. The output A1 layer of the trained full AN-CN-IC-A1 model results in more information compared to the input AN, or the intermediate CH or IC layers (apart from the two vowel recognition task due to its simplicity).

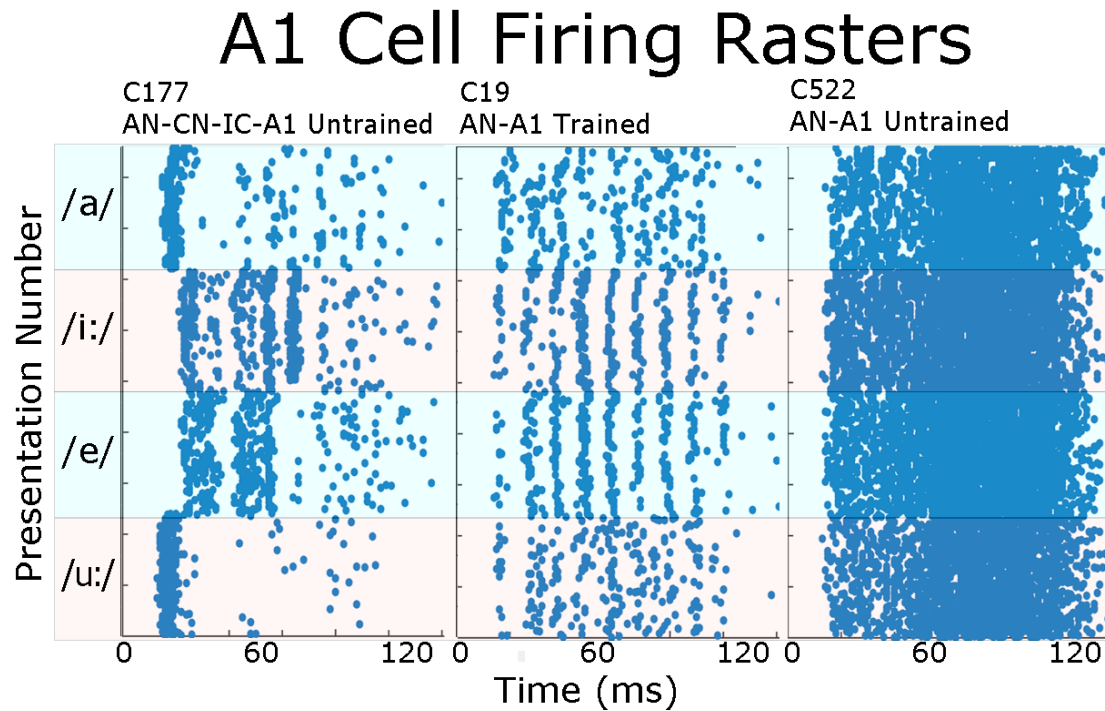


Figure 5.7: Firing rasters of three typical cells of the untrained AN-CN-IC-A1 model (cell 177) and the trained and untrained AN-A1 model (cells 19 and 522 correspondingly) in response to four vowels: /a/, /i:/, /e/ and /u:/. The abscissa is the time (ms) from stimulus start. The ordinate represents stimulus presentation number, whereby each vowel was presented 240 times (20 epochs of each of the 12 exemplars for each vowel). Each dot within a raster represents a spike produced by the cell. Cell index is provided at the top of each firing raster. It can be seen that cell 177 of the untrained AN-CN-IC-A1 model displays similar response properties to the trained AN-CN-IC-A1 model (Fig. 5.5), whereby different stable patterns of firing are produced in response to the different vowels. The same is not true for the AN-A1 model before or after training (cells 522 and 19 correspondingly), which have similar patterns of firing in response to the four vowels.

5.8.2 Bayesian Multiple Cell Information

In order to test whether more information is present about the identity of the synthesised vowel stimuli within the firing of a population of cells in the output A1 layer of the AN-CN-IC-A1 and the AN-A1 models of the auditory brain, a similar procedure was performed to the one described in Section 5.5. Information analysis was performed on the majority votes of a population of 125 most informative A1 cells as determined by the calculations described in Section 5.8.1. Figure 5.8 shows the resulting figures in proportion to the maximal theoretical information for the given stimulus set size ($\log_2 S$, where S denotes stimulus set size).

It can be seen that population information encoding based on the spiking patterns of A1 cells of the trained AN-CN-IC-A1 model results in high amounts of information about the identity of up to six synthesised vowel stimuli. The same is not true for the trained and untrained versions of the simple AN-A1 model, which reach significantly lower levels of multiple cell information. The difference in multiple cell information between the AN-CN-IC-A1 and the AN-A1 models provides further support for the importance of sub-cortical pre-processing within the CN and IC layers for auditory stimuli transform invariance learning.

The amount of information carried by the populations of A1 cells of the trained AN-CN-IC-A1 model based on their precise spiking times is significantly more than the amount of information carried by the populations of A1 cells using simple on/off encoding as described in Section 5.5. This suggests that the AN-CN-IC-A1 model is capable of learning about the identity of up to six simple synthesised vowel stimuli as long as the precise firing patterns of individual A1 cells are taken into account. The amount of information is increased throughout the hierarchy of the AN-CN-IC-A1 model, which further suggests that the temporal probabilistic information encoding is more appropriate than the simple rate on/off encoding scheme for representing multiple vowel classes.

Similar amounts of information are reached by the populations of A1 cells of the trained and untrained AN-CN-IC-A1 model. This was also the case for single A1 cells using the precise firing timing information encoding as discussed in Section 5.8.1, but not when simple on/off encoding was considered (Figs. 5.1 and 5.2). Such results are to be expected, however. Cell spike rasters show that A1 cells of the full AN-CN-IC-A1 model before (Fig. 5.7) and after (Fig 5.5) training show qualitatively similar patterns of firing in response to four vowels, whereby additional clues about stimulus identity are present within the precise firing timing of these cells. Simple on/off encoding only considers whether a cell fired at all during stimulus presentation or not. Figure 5.3 demonstrates that before training A1 cells fire to most presentations of most stimuli, which results in low single and multiple cell on/off information. The same is not the case after training using on/off encoding, since the firing rate of A1 cells drops due to the weakening of

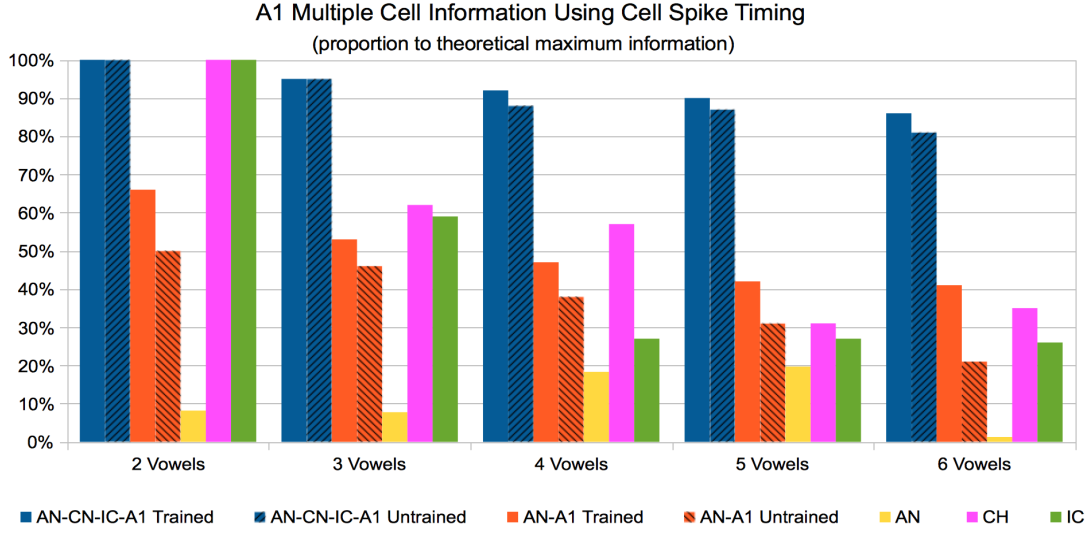


Figure 5.8: Vowel recognition accuracy for population responses of the 125 most informative cells in the output (A1) layer of the full AN-CN-IC-A1 (blue) and the simple AN-A1 (red) models of the auditory brain. It is assumed that information is encoded within the unique precise firing patterns of A1 cells in response to the different vowel stimuli. Data are shown for both models before (hatched bars) and after (solid bars) training and in response to two to six vowel stimuli. It can be seen that the full AN-CN-IC-A1 model reaches higher accuracy than the simple AN-A1 model, thus supporting the hypothesis that sub-cortical pre-processing is necessary for transform invariant auditory object recognition, including vowels. The high vowel recognition accuracy of the untrained AN-CN-IC-A1 model (hatched blue bars) suggests that plasticity within the IC→A1 connections might not be necessary for differentiating between the simple synthesised vowel stimuli. This figure demonstrates that multiple cell information encoding using the precise spike timing of A1 cells with respect to stimulus onset carries significantly more information than the simple on/off encoding as plotted in Figure 5.2. Similar data are presented for the input AN (yellow), and intermediate CH (CN) (magenta) and IC (green) layers. Data for ON and PL sub-populations of the CN are not presented, since ON does not carry any information about the vowel class identity, while PL is essentially identical to AN. The output A1 layer of the trained full AN-CN-IC-A1 model results in more information compared to the input AN, or the intermediate CH or IC layers (apart from the two vowel recognition task due to its simplicity).

many IC→A1 synapses. This results in less responsiveness of A1 cells to their un-preferred stimuli after training, and therefore higher on/off information.

The similar levels of multiple cell information within the trained and untrained versions of the AN-CN-IC-A1 model suggest that plasticity within the IC→A1 connections is not important for the development of transform invariant representations of simple synthesised vowel stimuli. IC→A1 plasticity, however, might be necessary for learning more complex naturally spoken word stimuli, which are discussed in the next chapter.

6

Towards Learning Transform Invariant Representations of Naturally Spoken Words

6.1 Introduction

This thesis has so far compared the performance of two models of the auditory brain, the full AN-CN-IC-A1 and the simple AN-A1 models, using synthesised vowels. It was shown that the full AN-CN-IC-A1 model is capable of developing transform invariant representations of up to six vowel stimuli. The encoding scheme used by the model A1 cells for representing the vowel identities, however, changed as the number of stimuli was increased from two to six. It was shown that simple on/off encoding was sufficient to provide full information about the identity of two vowels, whereby the only thing considered was whether an A1 cell fired at least once during a stimulus presentation or not. Such a simple encoding scheme was shown to fail to provide enough information about the stimuli as the number of vowels was increased. A more complex encoding scheme which took into account the precise firing times of individual cells was shown to be a more adequate way to differentiate between more vowels. Furthermore, while single cell information was shown to be enough to differentiate between two vowels, multiple cell information was necessary to differentiate between three and more vowels. This suggested that information encoding changed from 'grandmother-like' to distributed as the

number of synthesised vowel stimuli was increased from two to six.

This chapter explores whether the full AN-CN-IC-A1 model is capable of learning transform invariant representations of naturally spoken words rather than synthesised vowels. The complexity of naturally spoken words is much higher than that of synthesised vowels, which makes the transform invariance learning task a lot more challenging for the network. This chapter discusses how the full AN-CN-IC-A1 and the simple AN-A1 models of the auditory brain cope with learning transform invariant representations of naturally spoken digits 'one' and 'two' taken from the TIDIGITS database (Leonard & Doddington 1993). Furthermore, this chapter shows that due to the complexity of naturally spoken word stimuli a more elaborate encoding scheme based on the idea of polychronous chains as described by Izhikevich (2006) is required to encode transform invariant representations emerging in the output layer of the network.

6.1.1 Polychronous Chains

In order to form speaker independent representations of different words, the auditory brain has to be able to respond in a manner that differentiates between different words but not between different exemplars of the same word. This is a challenging task, given the great variability in the auditory inputs due to differences in pronunciation both within and between speakers. This input variability is further compounded by the stochasticity present in the firing patterns generated at the first neural stage of auditory processing, the AN. How can the brain discover the statistical regularities differentiating various words in such noisy inputs? We believe an answer to this question can be partially provided by the independent yet overlapping theories described in the previous chapters (Evans & Stringer 2012, Masquelier et al. 2008). However, in order to account for the complexity of naturally spoken words, a new concept of 'polychronous groups' (PGs) also has to be introduced.

A PG is a group of neurons that 'exhibit reproducible time-locked but not synchronous firing patterns with millisecond precision' (Izhikevich 2006). When a cell is a part of a polychronous group/chain, it can fire to a number of stimuli, with its different spikes helping encode different stimuli as part of different polychronous chains. Polychronous temporal representations of objects have been shown to dramatically increase the capacity of the learning system, since each cell can take part in a number of polychronous chains by producing a number of spikes, where each spike belongs to a different polychronous chain without disturbing any of the other chains.

The idea of a polychronous chain is similar to that of a synfire chain (Abeles 1982a, 1991, Diesmann et al. 1999, Herrmann et al. 1995). Synfire chains are response patterns within a feedforward neural network that are made of a number

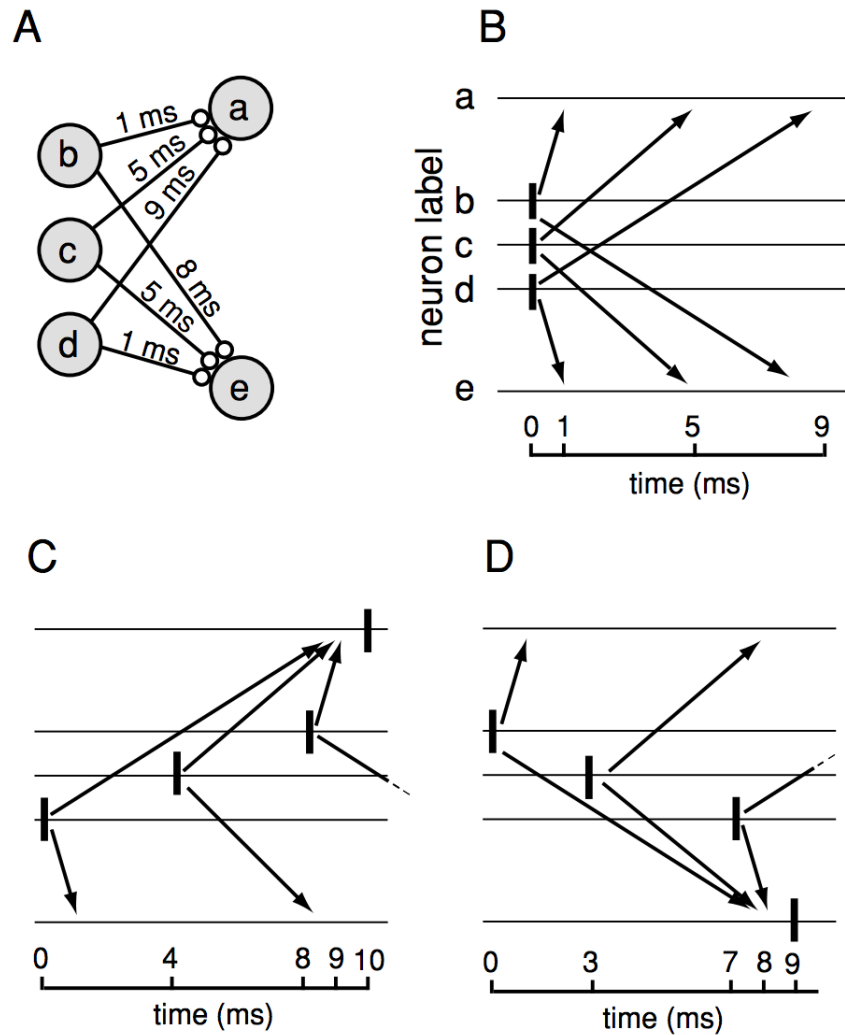


Figure 6.1: **A:** an illustration of two polychronous chains: *b, c, d, a* and *b, c, d, e*. The axonal delays from neurons *b*, *c* and *d* to neurons *a* and *e* are different, thus enabling the existence of the two polychronous chains. **B:** synchronous firing of neurons *b*, *c* and *d* is not a part of a polychronous chain, since it does not activate neurons *a* or *e*. Each spike is denoted by a vertical black line, and the arrows point at the spike arrival times to the corresponding cells after the appropriate axonal delays. **C:** due to the particular distribution of axonal delays, the depicted pattern of firing of neurons *b*, *c* and *d* results in the activation of neuron *a*. This, therefore, constitutes a polychronous chain *b, c, d, a*. **D:** due to the particular distribution of axonal delays, the depicted pattern of firing of neurons *b*, *c* and *d* results in the activation of neuron *e*. This, therefore, constitutes a polychronous chain *b, c, d, e*. Adapted from Izhikevich (2006).

of pools of neurons, whereby each neuron in pool i contacts a number of neurons in pool $i + 1$, and each neuron in pool $i + 1$ receives excitatory connections from a number of neurons in pool i . Such an architecture results in two patterns of activity, whereby activity either quickly dies down, or the initial activity is built up into a wave that propagates along a chain of neurons in a particular repeatable way, whereby the same initial activity will result in the same propagation path of the wave (Abeles 1982*a*, 1991, Diesmann et al. 1999, Herrmann et al. 1995). A synfire chain assumes synchronous firing of its constituent neurons, and the axonal delays between the different pools of neurons are assumed to be equal. This means that a particular population of cells can represent only one stimulus, since they can either fire synchronously or not. Polychronous chains are essentially synfire chains that allow for asynchronous firing of neurons. Polychronous groups have higher representational capacity, because different precise patterns of firing of one population of cells can encode different stimuli. This becomes possible because of the heterogeneous distribution of axonal delays in the network, which allows different temporal spike patterns to arrive simultaneously to different receiving neurons, thus providing them with super-threshold activation, if the appropriate matching distributions of axonal delays is in place (Fig. 6.1). An example of a polychronous chain arising within a recurrent network of spiking neurons as described by Izhikevich (2006) is shown in Figure 6.2.

PGs are shaped by the interplay between delay lines, STDP-based learning and stimulus structure. A random distribution of conduction delays sets up a repertoire of PGs in a network. When a stimulus is presented to the network, it may activate a subset of PGs, whose output propagates through a further set of connections with random delays. If that set of connections is large then it may contain subsets of delays which happen to match the characteristic spatio-temporal firing pattern of a PG in a manner that allows the PG output to converge synchronously on a receiving neuron. This receiving neuron thereby receives super-threshold activation, and its connections to the PG are strengthened by STDP-based learning .

In this manner, different output layer neurons become sensitized to the characteristic activity of different input layer PGs. The temporal structure of the input stimuli may cause the output layer neurons themselves to generate reproducible spatio-temporal firing patterns, giving rise to 'higher order' PGs, which may in turn be learned by the next layer in the network. Such a feedforward hierarchy could take advantage of cumulative delays over several layers of connectivity, enabling PGs to discover regularities in the temporal structure of input stimuli over an ever wider temporal scale. In addition, such networks should be able to cope with variable or degraded inputs while maintaining stable, informative output representations. For example, it may learn to cope with the presence of temporal jitter

in the order of a few milliseconds, spontaneous firing at up to 10 Hz, or loss of up to half of the neurons participating in an input PG (Masquelier et al. 2008).

In this chapter we will show that stable spatio-temporal patterns of firing (PGs) spontaneously emerge within the plastic cortical stage of the full AN-CN-IC-A1 model due to CT learning mediated by STDP (Evans & Stringer 2012). It is useful to think of CT learning as a mechanism which ‘expands PGs’ to respond to increasing numbers of similar exemplars from a given stimulus class. Networks using CT learning should thus be able to discover similarities in similar pronunciations of the same word and map these onto responses of PGs. In essence the network should operate much like an autoencoder network, and be able to exploit reproducible spatio-temporal input patterns and overlap among inputs representing members of each stimulus class, to develop an efficient, robust and compact temporal encoding of the stimulus set.

6.2 Stimuli

Digits ‘one’ and ‘two’ naturally pronounced twice each by 94 speakers were used as stimuli in this chapter. A more thorough description of the stimuli can be found in Chapter 2 Section 2.2.3.

6.2.1 Analysing the Complexity of Naturally Spoken Stimuli

The task of learning transform invariant representations of naturally spoken digits ‘one’ and ‘two’ is in a way similar to the task of learning transform invariant representations of synthesised vowels /i:/ and /a/. In the case of synthesised vowels, the network has to learn to respond similarly to 240 (20 epochs of each of the 12 exemplars) presentations of each vowel, while keeping the responses to the two vowels different. In the case of naturally spoken digits, the network has to learn to respond similarly to 376 (4 epochs of 94 pronunciations by different speakers) presentations of each digit, while keeping the responses to the two digits different. The two tasks, however, are very different in their complexity due to the nature of transform invariance that the network has to learn in the two conditions.

The only changes between the different exemplars of the synthesised vowels are the exact frequencies of their three constituent formants. This roughly corresponds to the different pronunciations of the vowels by the same speaker. The different exemplars of the naturally spoken digits ‘one’ and ‘two’ taken from the TIDIGITS database (Leonard & Doddington 1993) also vary in their constituent formants (both within and between speakers), however they also differ in their fundamental frequency (for different speakers), duration, rate of speech and envelope (due to

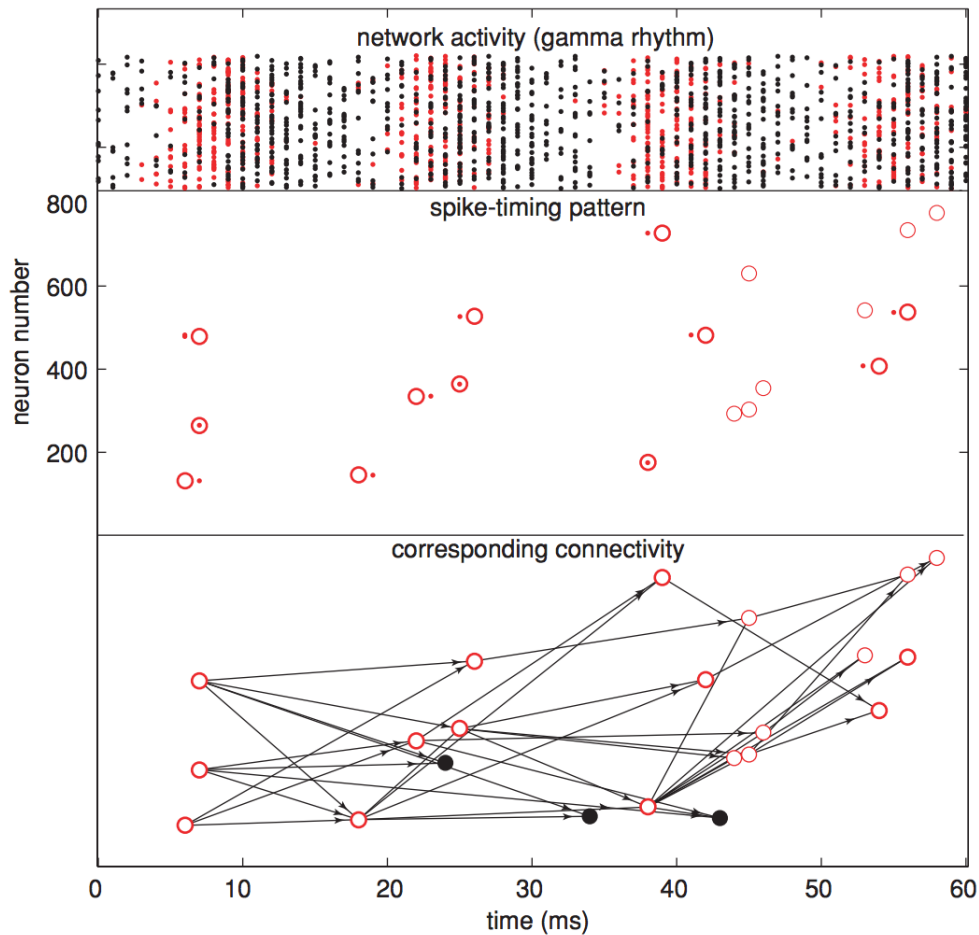


Figure 6.2: An example of a polychronous chain learnt by a recurrent spiking neural network with STDP-based learning and a distribution of axonal delays between 1000 randomly connected excitatory (80%, red) and inhibitory (20%, black) cells as described by Izhikevich (2006). The activity of the entire network is shown in the top plot. While the presence of any polychronous chains is not obvious from the raster of firing shown in the top plot, a polychronous chain shown in the middle plot appears more often than would be expected by chance. The chain is learnt because of the matching distribution of axonal delays shown in the bottom plot. Dots denote the actual firing of cells in the model, while circles represent the potential firing times based on the connectivity shown in the bottom plot. Adapted from Izhikevich (2006).

the different accents). Furthermore, while individual vowels are periodic yet temporally static stimuli, since their constituent formants do not change over time, words (such as 'one' and 'two') contain a number of phonemes, and are therefore temporally dynamic and contain additional data for the network to learn, such as phoneme co-articulation.

6.3 Analysing Network Responses to Naturally Spoken Stimuli

6.3.1 On/Off Encoding

Different types of information encoding have been assumed throughout this thesis to analyse the response properties of the output layer A1 of the full AN-CN-IC-A1 and the simple AN-A1 auditory brain models. For the two vowel recognition task described in Chapters 3-4 a simple on/off encoding was assumed, whereby each A1 cell was hypothesised to encode the vowel /a/ or /i:/ by firing at least once or not firing at all while being presented with the stimuli (see Chapter 2 Section 2.10.5 for details). Full 1 bit of information for the two synthesised vowel stimuli was reached if a single A1 cell fired at least once during every presentation of all exemplars of one vowel, but never fired to any presentations of any exemplars of the other vowel.

6.3.1.1 Information About Naturally Spoken Digits Using On/Off Encoding

Spike rasters of A1 cells of the full AN-CN-IC-A1 model after training (Fig. 6.3) suggest that simple on/off encoding used in previous chapters is not sufficient to provide enough information to differentiate between two naturally spoken digits 'one' and 'two'. On/off encoding assumes that stimulus information is represented according to whether a particular A1 cell fires at least one spike during stimulus presentation or not. Figure 6.3 suggests no differences in such binary cell responses between the depicted typical A1 cell responses to the two naturally spoken digits after training.

Figure 6.4 verifies that simple on/off encoding used in previous chapters is not sufficient to represent the two naturally spoken stimuli. It can be seen that the maximum multiple cell information reached by the 201 A1 layer cells of the full AN-CN-IC-A1 model after training is very low (0.012/1 bits) (see Chapter 5 Section 5.5 for methods).

It can be seen, however, that the A1 layer of the full AN-CN-IC-A1 model after training still contains more single and multiple cell information than the A1

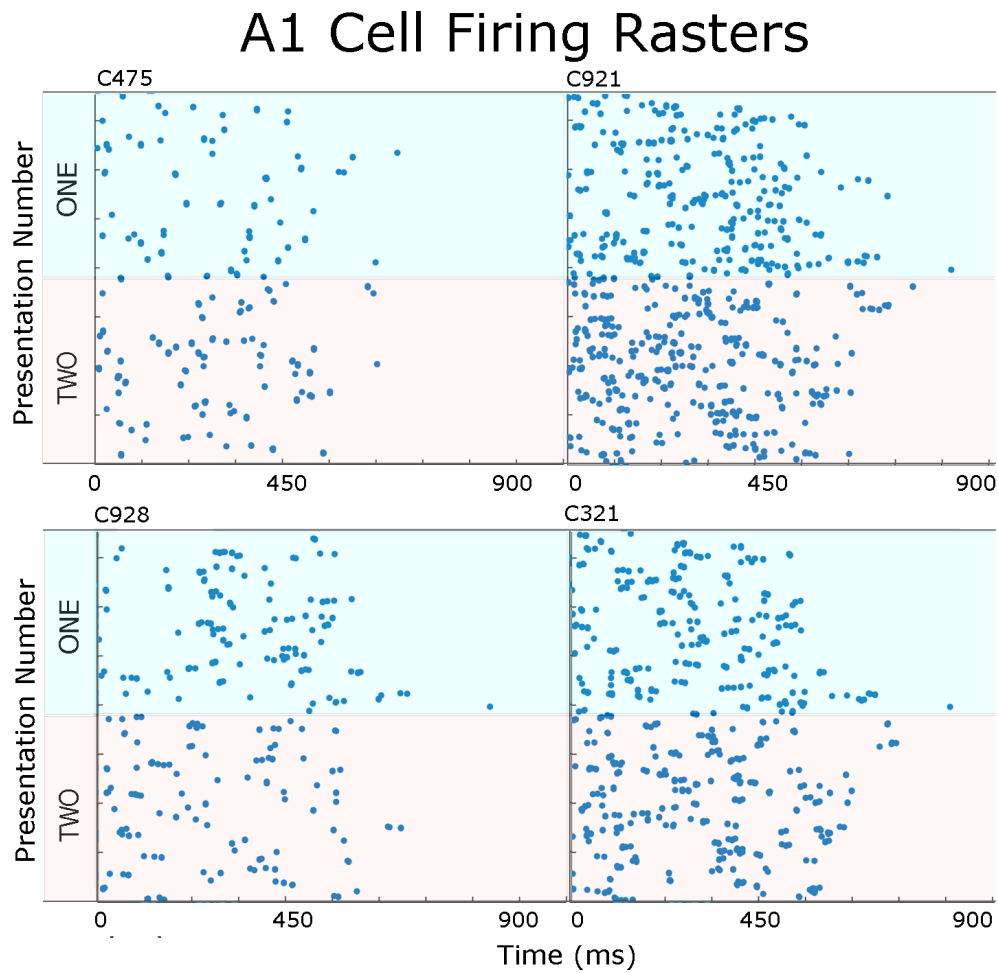


Figure 6.3: Firing rasters of four cells within the A1 layer of the full AN-CN-IC-A1 model of the auditory brain in response to two naturally spoken digits 'one' and 'two' after training. The abscissa is the time (ms) from stimulus start. The ordinate represents stimulus presentation number, whereby each digit was presented 188 times (4 epochs of each of the 94 exemplars for each digits). Each dot within a raster represents a spike produced by the cell. Cell index is provided at the top of each firing raster.

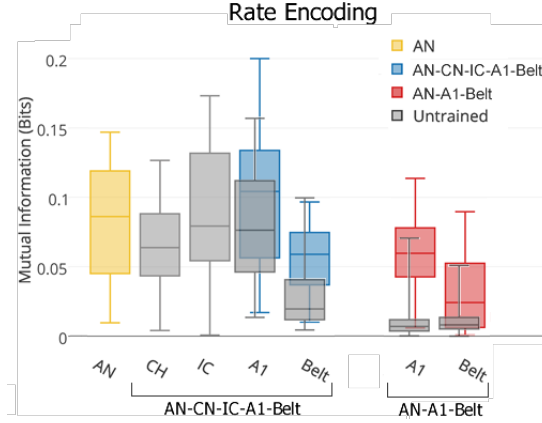


Figure 6.4: Box-and-whisker plots showing the distribution of information over twenty different subsamples of cells within various layers of the full AN-CN-IC-A1-Belt and reduced AN-A1-Belt models based on on/off rate encoding scheme in response to two digits 'one' and 'two' naturally spoken by 94 different speakers and presented over four epochs. Maximum theoretical entropy for the task is 1 bit. It can be seen that simple on/off encoding scheme is not sufficient to represent the two naturally spoken stimuli and a more elaborate temporal encoding scheme using precise spike times might be necessary. Nevertheless it can be seen that the full AN-CN-IC-A1 model reaches higher information than the simple AN-A1 model, thus confirming the hypothesis that sub-cortical pre-processing is helpful for transform invariant auditory object recognition, including naturally spoken words. PL and ON layers are not shown. The ON layer conveys 0 bits of information, and PL is equivalent to AN.

layer of the trained AN-A1 model, or the A1 layers of either of the two models before training. This provides further evidence for the importance of sub-cortical pre-processing for transform invariance learning of auditory objects.

6.3.2 Single Cell Temporal Encoding

Chapter 5 demonstrated that as the number of vowels presented to the network increases from two to six, the simple on/off encoding becomes insufficient to represent enough information to differentiate between the vowels in a transform invariant manner. It was shown that the precise timing of spikes emitted by a cell is useful to consider, as it carries more information about the stimulus identity. Evidence for such encoding was present in the single cell spike rasters (Fig. 5.5), which highlighted the different stable temporal patterns of firing of individual A1 cells in response to the different vowels.

Figure 6.3 plots spike rasters of a number of A1 cells within the AN-CN-IC-A1

model of the auditory brain in response to the naturally spoken digits 'one' and 'two' after training. It can be seen that no obvious structure of firing is present within the firing rasters of these cells to differentiate between the two digits. This is due to the increased stimulus complexity of naturally spoken digits compared to synthesised vowels as discussed in Section 6.2.1.

The single cell temporal information encoding scheme based on the precise spike times of individual A1 cells with respect to the stimulus start described in Chapter 5 and used to differentiate between three to six synthesised vowels relies on the regularity of precise spike times of individual cells with respect to the stimulus onset, which is achievable in response to the highly controlled synthesised vowel stimuli. This regularity breaks when the extra complexity and transform variance of the naturally spoken digits as discussed in Section 6.2.1 is added. Such an encoding scheme, therefore, is not appropriate any more to represent information about the identity of naturally spoken digits.

It is therefore hypothesised that the identity of the naturally spoken digit stimuli is encoded within the precise temporal patterns of firing of a number of A1 cells within polychronous chains. Such patterns of firing are expected to appear regularly, whereby the spikes that make up a pattern happen at the same time Δs with respect to each other rather than the stimulus onset.

6.3.3 Choosing the Right Information Encoding Scheme

The temporal information encoding scheme based on the precise spike times of individual A1 cells with respect to the stimulus start described in Chapter 5 is a simpler scheme than the one based on polychronous chains suggested above. The former is based on the spiking information of single A1 cells, while the latter requires spiking information of a population of A1 cells. Furthermore, the former is easier to intuitively interpret, since the differences in A1 cell firing patterns in response to different vowel stimuli are obvious in the cell raster plots (Figs. 5.5 and 5.7), while polychronous chains are virtually impossible to spot with the naked eye (Fig. 6.3).

While polychronous chains were likely to be present in the A1 layer of the full AN-CN-IC-A1 model when trained on two to six synthesised vowels, they have not been discussed so far, because each chapter assumed the simplest sufficient encoding scheme for the task. This was done because one of the roles of the auditory brain hierarchy is assumed to be the transformation of stimulus information encoding from the more complex one within the input AN layer, to the simplest possible one within the output A1 layer (see Chapter 4 Section 4.6 for further discussion of this). This is why simple on/off encoding was used to analyse the model performance when learning transform invariant representations of two vowels, but the more complex single cell temporal encoding scheme was used when three to

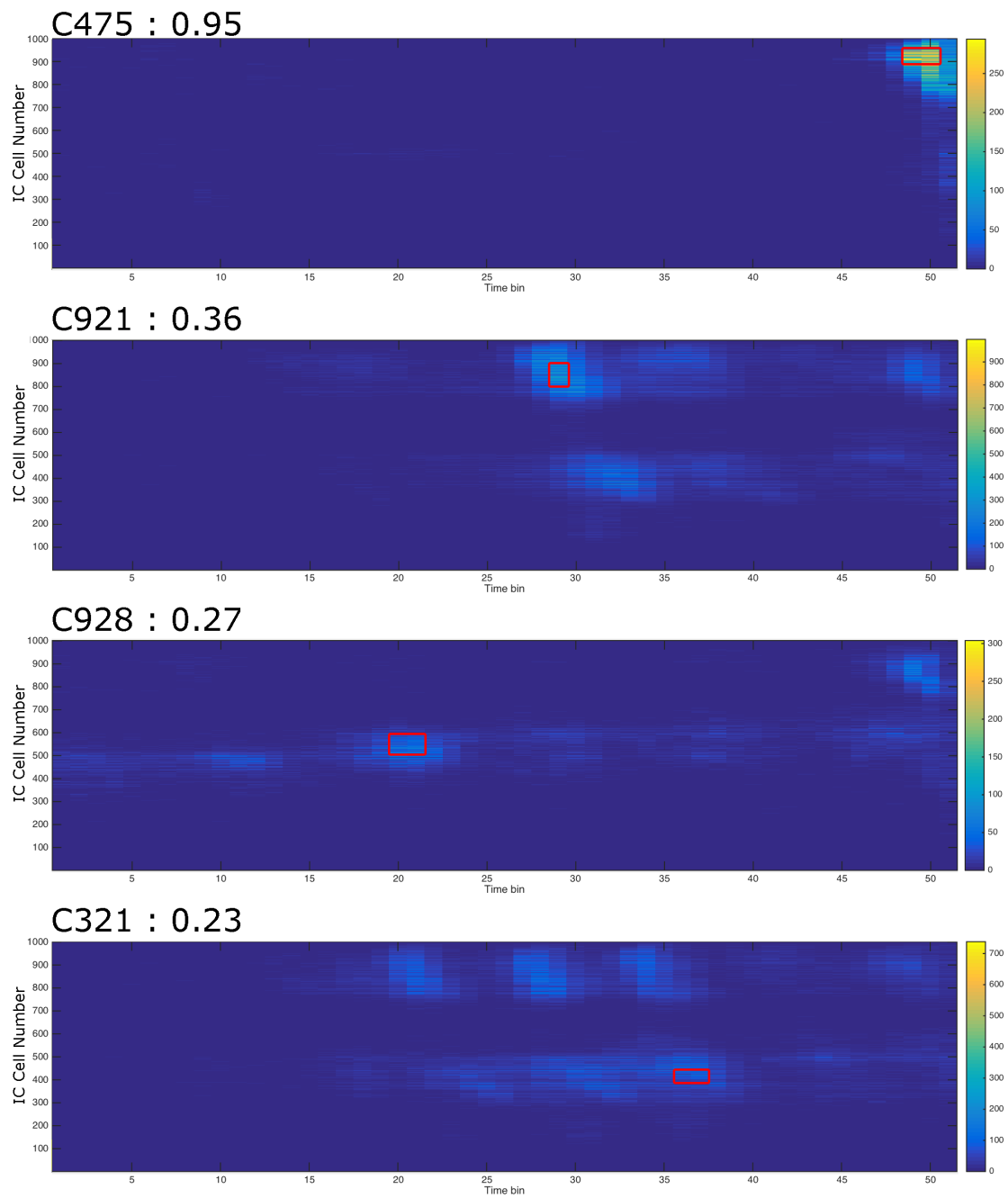
six vowels were presented to the network. On/off encoding was shown to be not sufficient any more to fully differentiate between three to six vowels, and therefore the next simplest encoding scheme was used. The single cell temporal information encoding scheme based on the precise spike times of individual A1 cells with respect to the stimulus start used to represent three to six synthesised vowels is not sufficient any more to represent naturally spoken digits 'one' and 'two' used in this chapter. This is why the next simplest encoding scheme based on the precise firing times of A1 cells within polychronous chains is discussed in this chapter.

6.3.4 Analysing A1 Firing

The continuous transformation (CT) learning mechanism based on STDP (described in Chapter 2 Section 2.5) relies on stable temporal patterns of firing within the pre-synaptic layer. In order to understand whether such stable firing patterns emerge within the A1 layer of the full AN-CN-IC-A1 model in response to naturally spoken words, such as digits 'one' and 'two', a Local Firing Stability (LFS) measure was used. The LFS is similar to the Same Exemplar (SE), Different Exemplars (DE) and Different Stimuli (DS) indices (Chapter 2 Section 2.10.3) used to measure the firing pattern stability within the AN and IC layers of the full AN-CN-IC-A1 model in response to synthesised vowels /i:/ and /a/. The SE, DE and DS indices, however, are not suitable to analyse the firing pattern stability within the A1 layer of the full AN-CN-IC-A1 model in response to the naturally spoken digits 'one' and 'two'. This is due to the fact that the indices described above measure global pattern stability within the whole layer. These measures look for repeating patterns of firing that happen for exactly the same A1 cells at exactly the same time Δ with respect to the stimulus start (± 0.5 ms due to the 1 ms binning of the firing rasters). These measures, therefore, rely on the identical duration, fundamental frequency and rate of speech of the stimuli, which is true for the synthesised vowels /i:/ and /a/, but false for the naturally spoken digits 'one' and 'two'. This is why a different approach based on the local firing stability of individual A1 cells was used to assess the firing stability in the output layer of the full AN-CN-IC-A1 model in response to the naturally spoken digits 'one' and 'two'. LFS index is somewhat similar to the spike-triggered average technique (Simioncelli et al. 2004) - a standard tool used in neuroscience to characterise the response properties of neurons. While spike-triggered average calculates the average stimulus preceding the chosen neuron's spikes, LFS index calculates average spiking activity of neurons within the same neural area as the chosen neuron and within a time delta both before and after the chosen neuron's spikes. The mathematical formulation of the LFS index is described in more detail next.

6.3.4.1 Local Firing Stability Measure

The new LFS measure for the A1 layer of the full AN-CN-IC-A1 model relies on the following principle. It is assumed that if an individual A1 cell learns to respond to a particular pattern of activity in the preceding layer (IC), and it responds to that pattern in a persistent manner at a fixed time Δ after the end of the pattern (± 0.5 ms), then the cell is likely to be a part of a stable firing pattern within a polychronous chain (Izhikevich 2006). This is because it is assumed that the different exemplars of the naturally spoken stimuli will consist of particular combinations of these micro-patterns of activity in the IC, which will drive the stable patterns of activity within the polychronous chains in the A1, which can then be learnt by the next cortical layers. In other words, if an A1 cell consistently fires to a particular pattern of activity in the preceding layer IC, then this cell carries information about the presence of that particular micropattern in the stimulus, and therefore is useful for the next cortical layer in learning transform invariant representations of natural auditory objects, such as naturally spoken digits 'one' and 'two'. On the other hand, if A1 cells fire to undistinguishable random patterns in the preceding IC layer, then the relative timing of such cells would be random, and thus they would not form stable patterns of firing constituting a polychronous chain within the A1, which in turn means that they would not provide the stable firing patterns necessary for the next cortical layers to learn anything.

**Figure 6.5:** (Figure legend on next page)

(Figure 6.5 legend) Visualisation of the IC firing patterns that A1 cells 475, 921, 928 and 321 of the full AN-CN-IC-A1 model trained on two naturally spoken digits ('one' and 'two') learnt to respond to. Each plot represents the sum of the binary matrices representing the IC firing over 50 ms preceding each of the spikes of the corresponding A1 cell (see Chapter 4 Section 4.4.3). The A1 cells fire at the 51 ms time bin. 1 ms bins are used for the plots. The colour of each plot component represents the number of spikes occurring for that IC cell (ordinate) at that time bin (abscissa) preceding the A1 cell spikes. The total number of spikes emitted by each of the four A1 cells during testing corresponds to the maximum value of the colour bar shown to the right of each plot. Higher values of plot components indicate more stable patterns of afferent IC firing that the A1 cells have learnt to respond to. The red rectangles indicate the areas of the plot that contain the 10 components with the highest magnitudes which serve as the firing patterns that drive the corresponding A1 cells. The title of each plot shows the cell number followed by the local firing stability (LFS) measure score - the maximum magnitude of the plot components in proportion to the total number of spikes emitted by the A1 cell during testing. The higher the LFS measure score, the more stable is the coupling between the particular patterns of firing within the IC and the firing of the particular A1 cell. It is important to remember that the delays in the model might be axonal, dendritic, synaptic or due to indirect connections. Furthermore, the model A1 is an approximation of MGN and A1 in the real brain.

In order to calculate this new LFS measure score the following calculations were made. First, preferred pattern plots (Chapter 4 Section 4.4.3) were constructed for the 100 most informative cells of the full AN-CN-IC-A1 model in response to the naturally spoken digits 'one' and 'two' (Fig. 6.5). Each plot is a visualisation of a matrix $M_c = \sum_i M_{ic}$, where i is the index of each spike emitted by each A1 cell c in response to all exemplars of all stimuli, and M_{ic} is a binary matrix representing an IC firing raster for the 50 ms preceding the i th spike of an A1 cell c . Each component of the matrix M_c is therefore the number of spikes that were emitted by a particular IC cell at a particular time Δ (± 0.5 ms) before the A1 cell c emitted a spike in response to any of the stimuli. The top ten components (m_j , where $j \in \{1, \dots, 10\}$) with the highest magnitude for each matrix M_c were taken, and their magnitude in proportion to the number of times A1 cell c fired in total in response to any stimulus presentation was calculated ($p_c = \frac{[m_1, m_2, \dots, m_{10}]}{n_c}$, where n_c is the number of times A1 cell c fired in total in response to any stimulus presentation). The components of vectors p_c for each A1 cell c indicate how likely it is that the A1 cell c was driven by the presence of a spike corresponding to the component m_j of matrix M_c , and therefore whether the A1 cell c has learnt to respond to a particular stable firing pattern within the IC which includes m_j .

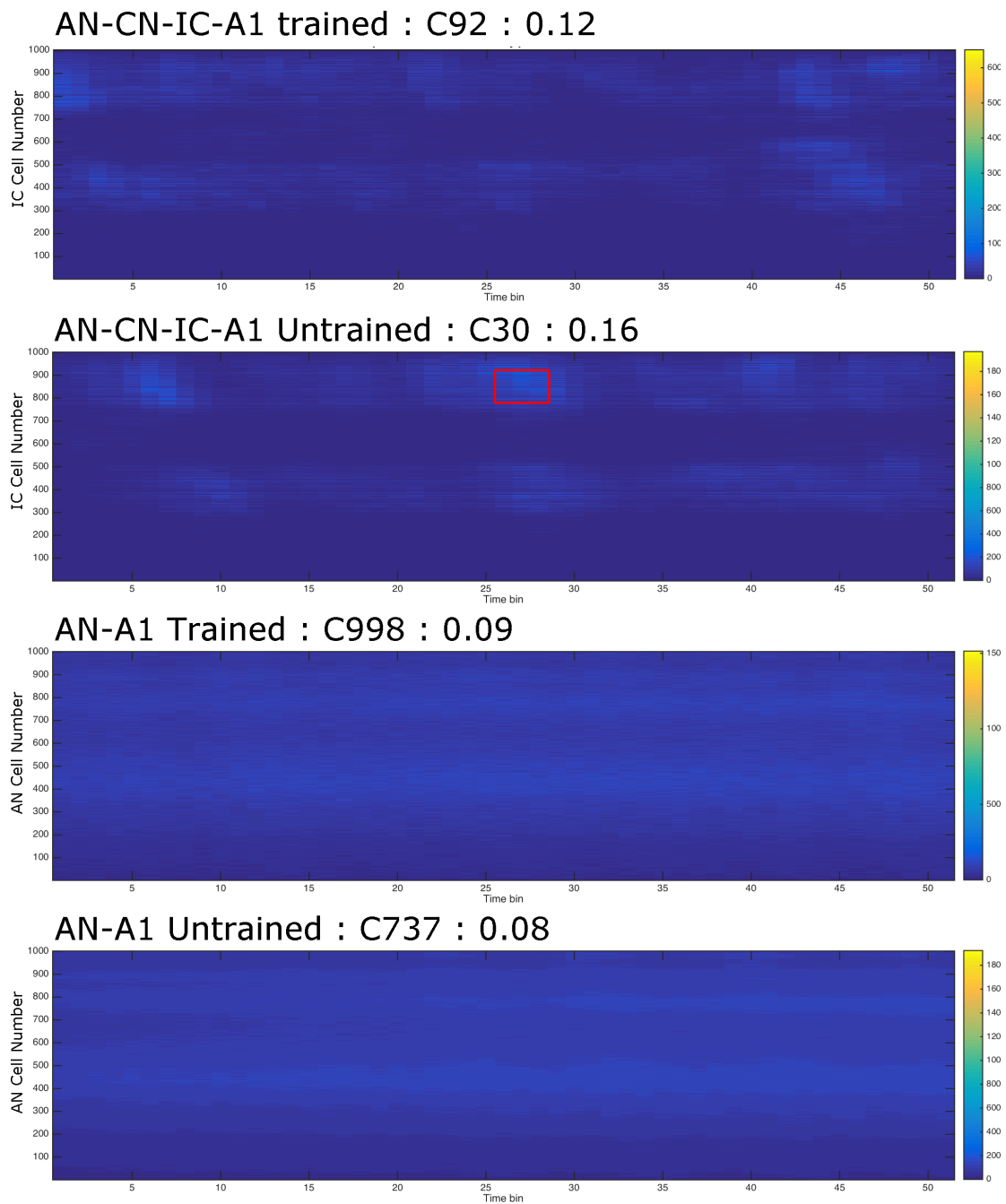
6.3.4.2 A1 Local Firing Stability in Response to Naturally Spoken Digits

The average of all components within vectors p_c for the 100 most informative A1 cells of the full AN-CN-IC-A1 and the simple AN-A1 models in response to the naturally spoken digits 'one' and 'two' before and after training was calculated. It was found that the LFS measure score was the highest for the full AN-CN-IC-A1 model after training at 0.44, compared to 0.17 before training. These were both higher than the LFS measure scores for the simple AN-A1 model before (0.08) and after (0.09) training. This suggests that the A1 cells of the trained full AN-CN-IC-A1 model of the auditory brain learn to respond to particular stable afferent patterns of activity within the IC and therefore are likely to form stable patterns of activity themselves as part of polychronous chains (Izhikevich 2006). This should enable the next cortical layers to continue building stimulus specific transform invariant representations of naturally spoken auditory objects, such as naturally spoken digits 'one' and 'two', using the firing with the A1 layer as input.

6.3.4.3 Preferred Firing Patterns of A1 Cells

When examining the preferred pattern plots of A1 cells of the full AN-CN-IC-A1 model in response to the naturally spoken digits 'one' and 'two' it can be seen that some A1 cells learnt to respond to particular IC firing patterns which only happens after a period of little activity within the IC (either at stimulus onset, since it is preceded by 100 ms of silence, or at a particular part of the stimulus which is preceded by at least around 50 ms of very low IC firing rate) (Fig. 6.5, cell 475). These cells have very high LFS measure scores (0.95 for cell 475), which indicate good learning.

Other A1 cells have lower LFS measure scores (which are still higher than those of the untrained AN-CN-IC-A1 or either trained or untrained AN-A1 models) (0.36, 0.27 and 0.23 for cells 921, 928 and 321 respectively in Fig. 6.5). This indicates that these A1 cells learnt to respond to particular afferent IC firing patterns that can appear at different points within the stimulus and can thus drive the A1 cells to fire quicker or slower depending on the A1 cells' state of activation at the time. The state of activation of each of the A1 cells depends on the particular IC firing preceding the onset of the preferred pattern. Due to the variability in the A1 cells' state of activation, the same preferred firing pattern can result in different time delta before the particular A1 cell fires. This in turn misaligns the IC firing binary matrices (M_{ic}), which in turn results in lower LFS measure scores.

**Figure 6.6:** (Figure legend on next page)

(Figure 6.6 legend) Visualisation of the preferred IC firing pattern for the A1 cell 92 of the trained full AN-CN-IC-A1 model with low local firing stability (LFS) measure score; the A1 cell 30 of the untrained full AN-CN-IC-A1 model of the auditory brain; and the preferred AN firing patterns for the A1 cells 737 and 998 of the simple AN-A1 model of the auditory brain before and after training, respectively. The firing data were collected by presenting the models with four epochs of two digits 'one' and 'two' naturally spoken by 94 speakers. Each plot represents 50 ms of IC/AN firing binned within 1 ms bins preceding each of the spikes of the corresponding A1 cell. The A1 cells fire at the 51 ms time bin. The colour of each plot component represents the number of spikes occurring for that IC/AN cell (ordinate) at that time bin (abscissa) preceding the A1 cell spikes. The total number of spikes emitted by each of the three A1 cells during testing corresponds to the maximum value of the colour bar shown to the right of each plot. Higher values of plot components indicate more stable patterns of afferent IC/AN firing that the A1 cells have learnt to respond to. The red rectangle indicates the area of the plot that contains the 10 components with the highest magnitudes which serve as the firing pattern that drives the A1 cell 30. Such a rectangle is not present for the A1 cells 92, 998 and 737, since there was no single area of the plot that contained the components with the higher magnitude, thus suggesting that the A1 cells of the simple AN-A1 model did not respond to particular stable afferent firing patterns either before or after training. The title of each plot shows the cell number, followed by the LFS measure score - the maximum magnitude of the plot components in proportion to the total number of spikes emitted by the A1 cell during testing. The higher the LFS measure score, the more stable is the coupling between the particular patterns of firing within the IC/AN and the firing of the particular A1 cell. The lack of yellow areas within the plots and the low LFS scores indicate that none of the cells presented in this plot learnt to respond to any particular afferent firing patterns from their respective input layers.

Figure 6.6 shows preferred pattern plots for the most informative A1 cell of the full AN-CN-IC-A1 model before training (cell 30) and for the most informative cells of the simple AN-A1 model before and after training (cells 737 and 998 respectively). It can be seen that these cells do not learn to respond to any particular afferent firing patterns from the IC or AN respectively, as indicated by the low LFS measure scores and the blurred preferred pattern plots. Figure 6.6 also shows the preferred pattern plot of A1 cell 92 within the full AN-CN-IC-A1 model after training, which has a low LFS measure score. Not all A1 cells learn to become a part of polychronous chains after training, and cell 92 is an example of such cells. It can be seen that its preferred pattern plot looks similar to that of cell 30 of the untrained AN-CN-IC-A1 model. This is different to the plots indicative of cells that do become a part of polychronous chains after training (Fig. 6.5).

6.3.5 Changes in Distributions of IC→A1 Afferent Connection Weights

When the distributions of afferent connection weights (w_{ij}^{BL}) for the cells described in Section 6.3.4.3 are examined (Fig. 6.7), it can be seen that cells 475, 921, 928 and 321 of the AN-CN-IC-A1 model (Fig. 6.5) and cell 998 of the AN-A1 model (Fig. 6.6) after training have significantly different distributions of afferent connection weights compared to cells 30 and 998 of the corresponding models before training.

It can be seen that cells 475, 921, 928 and 321 of the trained AN-CN-IC-A1 model have both weakened (< 30 nA) and strengthened (> 35 nA) connections (the afferent IC→A1 connections (w_{ij}^{BL}) were initialised randomly from a uniform [30, 35] nA distribution, see Table 3.4). Cells 475 and 921 developed the strongest afferent connections. These strong afferent connections correspond to the higher LFS measure scores of 0.95 and 0.36 correspondingly (Fig. 6.5). This suggests that these cells have indeed learnt to respond to a particular stable repeating pattern within the IC, since such a stable repeating afferent firing pattern is necessary to enable LTP to strengthen the cell's afferent connection weights even when the model LTP magnitude (α_p) is lower than LTD magnitude (α_d) (Table 3.4) (Masquelier et al. 2008). The less stable IC firing patterns for cells 928 and 321 as evidenced by their lower LFS measure scores (0.27 and 0.23 correspondingly) (Fig. 6.5) resulted in the lower maximal afferent connection strength magnitude developed through STDP-based learning for these cells.

Figure 6.7 shows that cell 998 of the trained AN-A1 model has only weakened afferent connection weights (< 30 nA). This cell also has a very low LFS measure score of 0.09 (Figs. 6.6 and 6.7). This further suggests that this cell did not receive stable afferent firing patterns from the AN layer, which resulted in the global depression of its afferent connection weights (w_{ij}^{BL}) due to the lower LTP magnitude (α_p) compared to the LTD magnitude (α_d) within the AN-A1 model (Table 3.2, Model 2) (Masquelier et al. 2008).

6.3.6 Changes in IC→A1 Afferent Connection Weights Due to Learning

Histograms and heat plots of all afferent connection weights (w_{ij}^{BL}) between the IC/AN and A1 layers of the full AN-CN-IC-A1 and simple AN-A1 models of the auditory brain were plotted before and after training (Fig. 6.8). When examining the changes in afferent connection weights (w_{ij}^{BL}) between the IC and A1 layers of the full AN-CN-IC-A1 model of the auditory brain before and after training, it can be seen that little learning happened in the low characteristic frequency part of the layer, since most connections remained within the [30, 35] nA range used for the afferent connection strength initialisation (see Fig. 6.8 A). The part

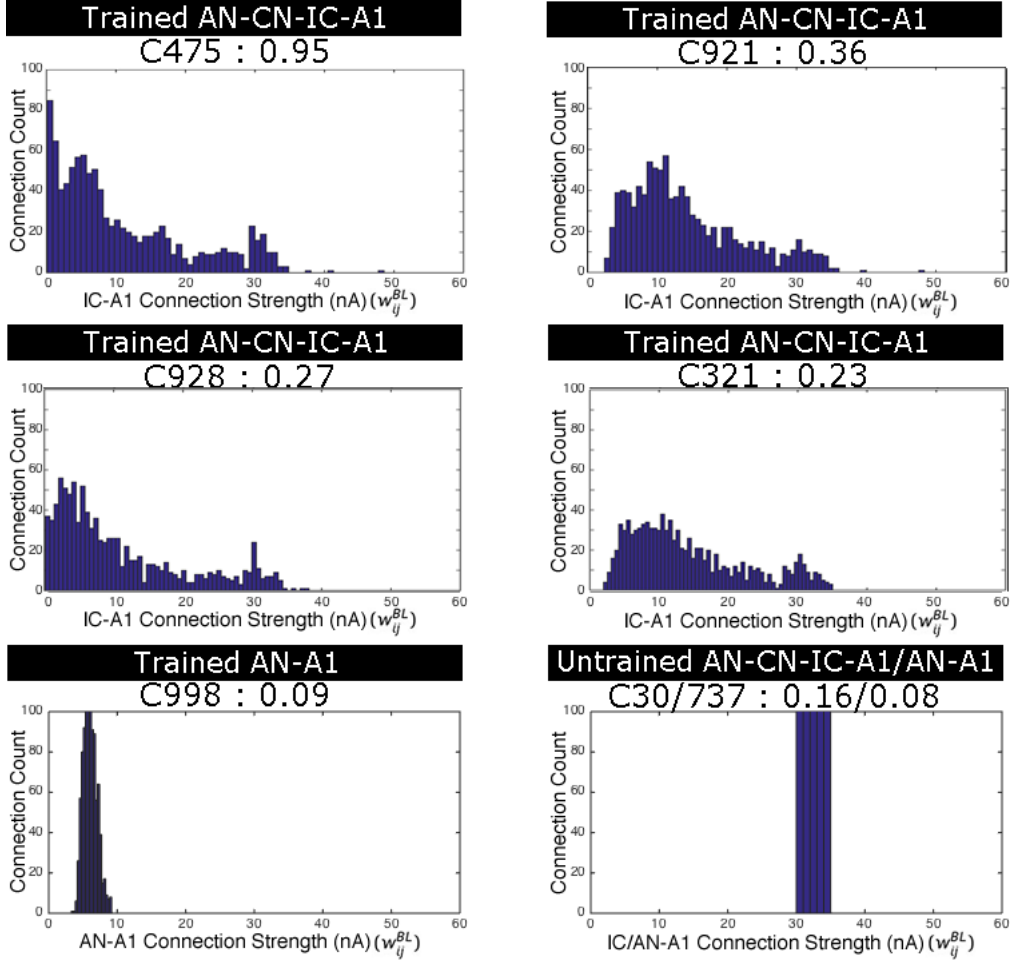


Figure 6.7: Histograms of afferent IC→A1 connection magnitudes (w_{ij}^{BL}) of seven cells: cells 475, 921, 928 and 321 of the trained AN-CN-IC-A1 model; cell 998 of the trained AN-A1 model; and cells 30 and 737 of the untrained AN-CN-IC-A1 and AN-A1 models respectively. The title of each plot shows the cell number, followed by the local firing stability (LFS) measure score (see Section 6.3.4.1). The higher the LFS measure score, the more stable is the coupling between the particular patterns of firing within the IC/AN and the firing of the particular A1 cell. Cells from the trained AN-CN-IC-A1 network have achieved higher single cell information and LFS measure scores, and have strengthened afferent connection weights from the IC, which is indicative of learning. Cell 998 from the trained simple AN-A1 network has achieved poor single cell information and a low LFS measure score, and has only weakened afferent connection weights from the AN, which indicates that no stable firing patterns were learnt. Cells 30/737 are example A1 cells within the untrained AN-CN-IC-A1 and AN-A1 models correspondingly. Since the afferent connections received by A1 were randomly initialised from the same [30, 35] nA uniform distribution for both the AN-CN-IC-A1 and the AN-A1 models, the histograms of afferent IC→A1 connection magnitudes for every untrained A1 cell of either of the two models look identical. An example of such a histogram is depicted in the bottom right sub-plot.

of the layer that received more input from the naturally spoken digits 'one' and 'two', however, changed significantly compared to the untrained network. While most connections weakened due to noisy firing within the IC and the lower LTP magnitude (α_p) compared to the LTD magnitude (α_d) (Table 3.4) (Masquelier et al. 2008), a large number of afferents strengthened up to the maximum of 60 nA. These afferents connect the IC cells that took part in the regular patterns of firing in response to the naturally spoken digits to the A1 cells that have the right distribution of delays to extract stable firing patterns from the afferent IC firing raster (see Section 2.4).

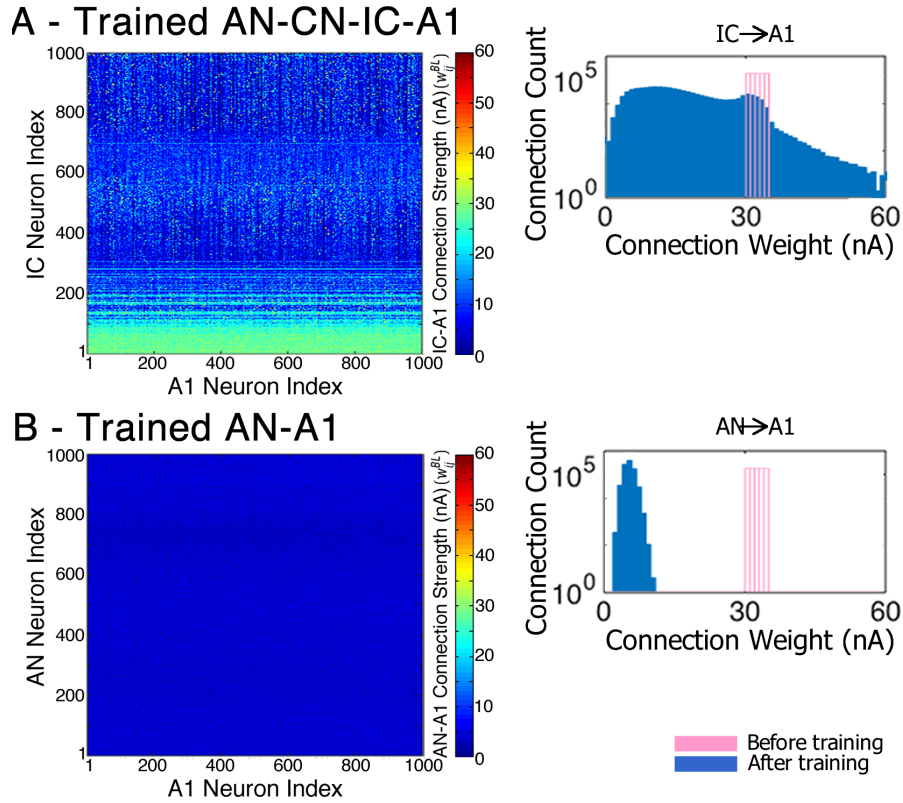


Figure 6.8: (Figure legend on next page)

(Figure 6.8 legend) Afferent connection weight matrices (w_{ij}^{BL}) (left column) and distributions of synaptic connection weights before (pink) and after (blue) training (right column) for the full AN-CN-IC-A1 model (**A**) and reduced AN-A1 model (**B**). The ordinate of the subplots in the right column is log scaled. It can be seen that STDP-based learning resulted in a multi-modal distribution of afferent $IC \rightarrow A1$ connections of the full AN-CN-IC-A1 model, whereby some connections got strengthened, some remained unchanged from the initialisation values in the [30, 35] nA range (these connections correspond primarily to the afferents received from the low frequency tuned region of the IC), and some got weakened. This is different from the unimodal distribution of $AN \rightarrow A1$ connections of the simple AN-A1 model after training, whereby *all* connections got weakened.

It can be seen that all of the afferent connection weights (w_{ij}^{BL}) of the simple AN-A1 model got weakened during training (see Fig. 6.8 B). This suggests that none of the A1 cells of the AN-A1 model received stable patterns of firing from the AN during training (Masquelier et al. 2008). This further supports the hypothesis that sub-cortical pre-processing within the CN and IC is necessary to remove AN noise and enable afferent connection strengthening through STDP-based learning.

6.3.7 Measuring Polychronization Index in A1

As outlined earlier in this chapter, we envisage that the full AN-CN-IC-A1 network will exhibit unsupervised learning by forming a hierarchy of PGs, where each PG is a group of neurons which 'exhibit reproducible time-locked but not synchronous firing patterns with millisecond precision' (Izhikevich 2006). An exhaustive search for PGs through the network would be prohibitive (Paugam-Moisy et al. 2008), so we developed a numerical score, the 'polychronization index' (PI) to quantify the prevalence of reproducible patterns across the population of neurons in one layer (see Chapter 2 Section 2.10.3.4).

Figure 6.9 shows the calculated PI scores. It can be seen that subcortical preprocessing in the CN and IC led to more stable spatio-temporal discharge patterns, as evidenced by higher PI scores in the IC compared to AN. These more reproducible firing patterns also carried through to A1 and Belt. The AN-A1-Belt model, which lacked subcortical preprocessing layers CN and IC, did not achieve the same stability of firing patterns seen in IC, A1 and Belt of the full AN-CN-IC-A1-Belt model even after training.

6.3.8 Expanding the Models

Sections 6.3.4-6.3.7 have examined evidence that the output A1 layer cells of the full AN-CN-IC-A1 model have learnt about the naturally spoken digits 'one' and 'two' and thus have become parts of polychronous chains of stable repeating patterns of activity in response to the different exemplars of the two stimuli. This evidence included the following differences between the AN-CN-IC-A1 and AN-A1 models: higher LFS measure scores (see Section 6.3.4.1); strengthening of the afferent IC→A1 connections during training; and higher within layer synchrony scores for the A1 cells of the trained AN-CN-IC-A1 model.

Higher local LFS measure scores found for the full AN-CN-IC-A1 model compared to the AN-A1 model suggest that the activity of A1 cells of the AN-CN-IC-A1 model was driven by stable patterns of firing within the IC. These stable patterns resulted in the strengthening of the afferent connections between the IC and A1 cells of the full AN-CN-IC-A1 model during training. The two factors described

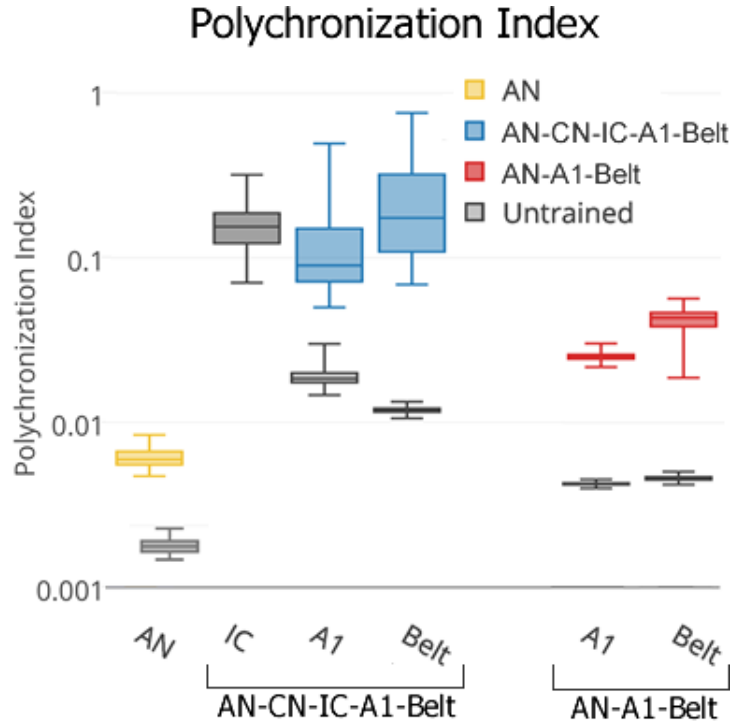


Figure 6.9: Box-and-whisker plot showing the distribution of polychronization indices (PI) over all the cells within the relevant layers of the full AN-CN-IC-A1-Belt and reduced AN-A1-Belt models. The ordinate is log scaled. A higher score indicates that the cells within the layer are more likely to form stable information bearing polychronous chains.

above would also give rise to stable firing within the A1 through the strengthened IC→A1 afferents, as evidenced by the corresponding higher LFS measure score for the A1 layer of the AN-CN-IC-A1 model. This, therefore, suggests that informative polychronous chains emerge in the A1 layer which could encode transform invariant representations of the naturally spoken digits 'one' and 'two'.

In order to provide additional evidence as to whether stable patterns of activity emerge in the output layer A1 of the full AN-CN-IC-A1 model of the auditory brain, an extra cortical layer loosely corresponding to the Belt area of the auditory cortex was added after the A1. The Belt layer was initialised in the same way as the A1 layer (see Table 3.4). It comprised 1000 excitatory neurons with plastic feed-forward A1→Belt afferent connections initialised with random connection weights (w_{ij}^{BL}) drawn from a uniform distribution of [30, 35] nA and with a random distribution of axonal delays (Δ_{ij}) drawn from a uniform distribution of [0, 50] ms. The layer included inhibitory inter-neurons identical to the A1 layer. An additional Belt area was also attached to the end of the simple AN-A1 model in a similar manner. The two models were trained on the exemplars of the naturally spoken digits 'one' and 'two' according to the same paradigm as was described in Section 6.2.1. The LTP/LTD time constants (τ_p/τ_d) and magnitudes (α_p/α_d) were kept the same as in the original AN-CN-IC-A1 model simulations at 15/25 ms and 0.005/-0.015 correspondingly.

6.3.8.1 Changes in A1→Belt Afferent Connection Weights Due to Learning

It was hypothesised that the A1→Belt afferent connections (w_{ij}^{BL}) would get strengthened if stable repeating firing patterns corresponding to polychronous chains were present within the A1 layer. If stable firing patterns of polychronous chains were not present, then the A1→Belt afferent connections were expected to weaken during training. The former was found to be the case for the AN-CN-IC-A1-Belt model (Fig. 6.10, A), while the latter was found to be the case for the A1→Belt afferents of the simple AN-A1 model (Fig. 6.10, B).

6.3.8.2 Changes in Distributions of A1→Belt Afferent Connection Weights

Section 6.3.4.3 suggested that A1 cells 475, 921, 928 and 321 of the full AN-CN-IC-A1 model took part in stable polychronous chains of firing in response to the naturally spoken digits 'one' and 'two' based on the evidence of these cells having higher LFS scores. If these A1 cells were indeed a part of stable polychronous chains, then some of their connections (w_{ij}^{BL}) to certain Belt neurons with the right axonal delays (Δ_{ij}) to capture the repeating polychronous patterns would get strengthened during training. This was indeed found to be the case (see Fig.

6.11, cells 475, 921, 928 and 321).

The A1 cells of the full AN-CN-IC-A1 model after training which had low LFS measure scores did not exhibit any strengthening of their afferent connections (w_{ij}^{BL}) to the Belt area (Fig. 6.11, cells 92 and 1000), thus supporting the claim made in Section 6.3.4.3 that lower A1 LFS measure scores indicated that these A1 cells did not learn much information about the stimuli and thus were not a part of any informative polychronous chains.

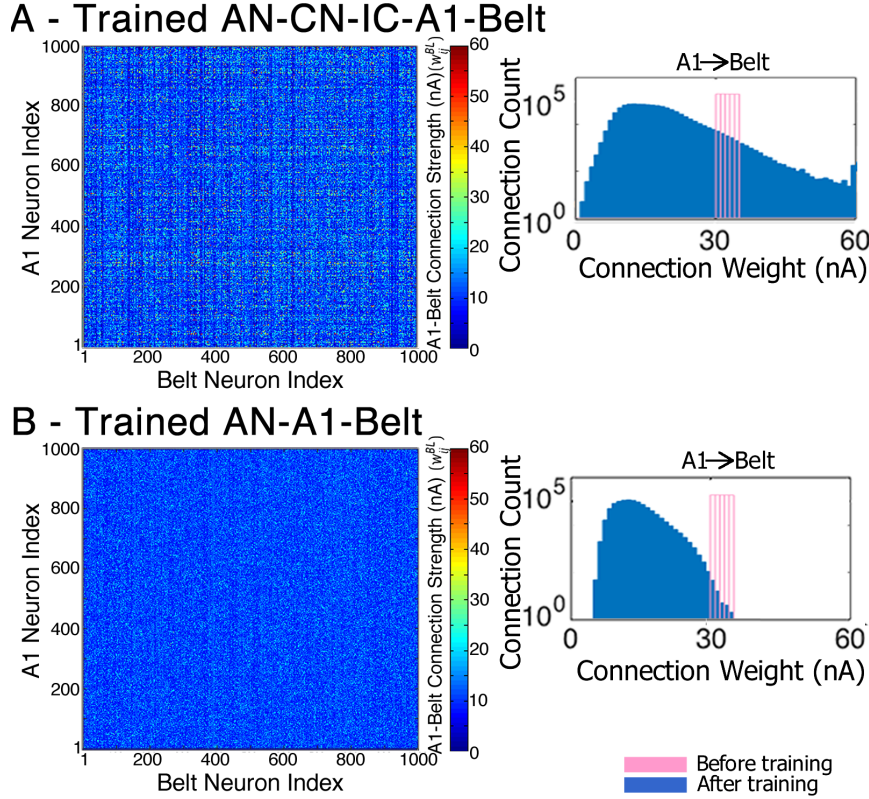


Figure 6.10: Afferent connection weight matrices (w_{ij}^{BL}) (left column) and distributions of synaptic connection weights before (pink) and after (blue) training (right column) for the full AN-CN-IC-A1-Belt model (A) and reduced AN-A1-Belt model (B). The ordinate of the subplots in the right column is log scaled. It can be seen that STDP-based learning resulted in strengthening of some afferent A1→Belt connections of the full AN-CN-IC-A1-Belt model, while all A1→Belt connections of the simple AN-A1-Belt model got weakened after training. This suggests that the A1 layer developed stable firing patterns in response to the naturally spoken digits 'one' and 'two' in the first instance but not in the second one.

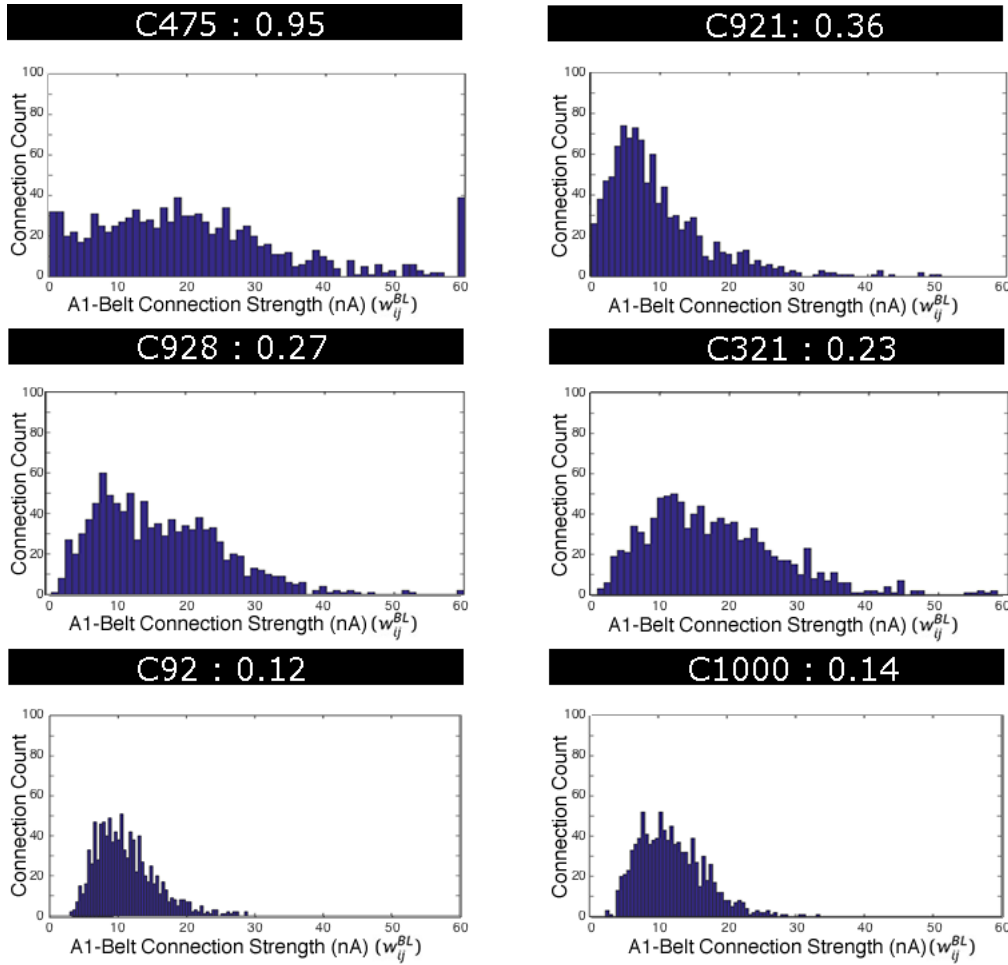


Figure 6.11: Histograms of A1→Belt connection magnitudes (w_{ij}^{BL}) of six A1 cells of the trained AN-CN-IC-A1-Belt model. The title of each plot shows the cell number, followed by the local firing stability (LFS) measure score (see Section 6.3.4.1). The higher the LFS measure score, the more stable is the coupling between the particular patterns of firing within the IC/AN and the firing of the particular A1 cell. Cells 475, 921, 928 and 321 have developed some strengthened A1→Belt afferents during training (the connections were randomly initialised from a [30, 35] nA uniform distribution). These cells also have higher LFS measure scores. Both of these pieces of evidence are consistent and suggest that A1 cells 475, 921, 928 and 321 might be parts of information carrying polychronous chains. A1 cells 92 and 1000 do not have any strengthened A1→Belt connections after training, which suggests that they are not part of information carrying polychronous chains. These cells also have lower LFS measure scores.

6.3.9 Analysing Population Encoding Information

The strengthening of the A1→Belt afferent connections (w_{ij}^{BL}) within the full AN-CN-IC-A1-Belt model of the auditory brain discussed in Section 6.3.8.1 and the higher A1 synchrony scores discussed in Section 6.3.7 indicate that stable patterns of firing emerge in the A1 layer of the model in response to the naturally spoken digits 'one' and 'two' during training. It does not, however, provide any information whether these stable firing patterns are discriminative between the two stimuli. If the same firing pattern is produced within the A1 in response to every presentation of every exemplar of both stimuli, then such a pattern would carry no information about the identity of the stimuli while resulting in maximal strengthening of the afferent A1→Belt connections (w_{ij}^{BL}) with the appropriate axonal delays (Δ_{ij}).

In order to test whether discriminative patterns of firing were present in the A1, the following calculations were made. First, it was assumed that A1 cells which take part in non-discriminative but still *stable* firing patterns (ie stable repeating patterns of firing that are common to the both stimulus classes) would have maximally strengthened afferent connections to Belt neurons. This was hypothesised to happen because such cells would not have a preference between the naturally spoken digits 'one' and 'two', and would therefore fire to most presentations of either stimulus. Since Belt neurons would be presented with these repeated firing patterns often during training, the corresponding afferent A1 connection weights (w_{ij}^{BL}) would be strengthened much more often than they would be weakened through STDP-based learning (Masquelier et al. 2008), and thus the maximum (w_{max}) connection strengthening would be achieved.

On the other hand, A1 cells that take part in discriminative stimulus selective stable firing patterns (patterns that are more likely to happen in response to exemplars of one stimulus rather than the other) are hypothesised to have medium rather than maximally (w_{max}) strengthened A1→Belt connections (w_{ij}^{BL}). This is because the Belt cells with the right distribution of axonal delays (Δ_{ij}) to be activated by the discriminative stimulus selective repeating patterns in the A1 would be exposed to these repeating patterns during half of all stimuli presentations at most during training. Their afferent connections from A1 would therefore be pulled into an equilibrium around a certain non-maximal connection strength weight value by STDP-based learning (Masquelier et al. 2008). The range of such non-maximal connection strength equilibrium was hypothesised to be between 45-55 nA for afferent connections initialised between 30-35 nA with the maximal strength (w_{max}) of 60 nA. This was found to be the case for the maximally informative cells within the A1 layer of the full AN-CN-IC-A1 model trained on the synthesised vowels /i:/ and /a/ (see Fig. 4.17).

For the reasons described above, a set of 109 A1 cells with afferent connections

to the Belt layer in the 45-55 nA range were selected, whereby the strengthened connections synapsed on Belt cells that had at least four A1 afferents with magnitude within the 45-55 nA range and no A1 afferents with magnitude above 55 nA. These cells were hypothesised to take part in discriminative stable firing patterns within the A1 in response to the two naturally spoken digit stimuli, because they synapsed on Belt cells which received enough strengthened afferents to have learnt about the stimuli, while not being affected by the activity of maximally strengthened non-discriminative afferents from the A1.

It was hypothesised that each of the 109 A1 cells could take part in a number of stable firing patterns comprised by a number of A1 cells. An individual A1 neuron was assumed to fire to exemplars of both digits, however the exact patterns of firing over a population of A1 cells were assumed to be different for digit 'one' and digit 'two'.

6.3.9.1 Finding the Stable Repeating Discriminative A1 Firing Patterns

Using MLP decoders to evaluate network performance: Decoder multi-layer perceptrons (MLPs) were used to evaluate the ability of the spiking neural network (SNN) autoencoders to represent stimulus identity using either rate or temporal encoding schemes. The MLPs were presented with 201×752 input data matrices $X = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{752}]$, consisting of 201 element column vectors $\mathbf{x}_i \in \mathbf{R}^{201}$ each representing features of the neural population response to one of the 752 sound stimuli in the stimulus set, and each column vector needed to be classified as belonging to one of the two stimulus classes, 'one' or 'two'. The response feature vectors were constructed as follows:

Temporal code: Temporal encoding assumes that information about stimulus class is provided by stable spatio-temporal firing patterns. In order to be informative, particular firing patterns must appear more frequently in response to the exemplars of one stimulus class than the other. The column vectors $\mathbf{x}_i$ of the MLP input data matrix X should, therefore, capture these spatio-temporal firing patterns within an SNN layer. We quantified reproducible spatio-temporal patterns using methods which were entirely analogous to those used to calculate the polychronization index (see above) except that, to reduce computational load, we based calculations on a random sample of $J = 100$ cells from a given layer rather than all cells in the layer. Separate MLP _{j} decoders were trained for each cell $j \in J$. For each MLP _{j} a separate input data matrix $(X_j)_{201 \times 752}$ was created. First, we computed a 100 neuron by 101 ms matrix $M_{e_k(s)}^j$ for the given cell j and for each stimulus $e_k(s)$. Then we computed a 100 element vector of row sums and a 101 element vector of column sums of matrix $M_{e_k(s)}^j$ and concatenated the two to form the 201 element column vectors $\mathbf{x}_i \in \mathbf{R}^{201}$ of the MLP input data matrix X_j .

Summing and concatenating reduces the dimension of the input feature vectors from 10,100 to 201, while still capturing key features of the pattern of firing across the sample population of 100 cells in response to a single presentation of stimulus $e_k(s)$. Each of the decoders MLP_j was then trained as described above and then used to classify response vectors $\mathbf{x}_i$ as responses to an exemplar of word one or word two. The majority classification votes among the 100 trained MLP_j in response to the column vectors of their corresponding matrices X^j made up the final vector of stimulus classifications $\mathbf{c} = [\mathbf{c}_1, \dots, \mathbf{c}_{752}]$ for the layer. The confusion matrix used in the information analysis calculations described above was constructed based on these final layer classifications contained in $\mathbf{c}$.

Estimating information content using MLP decoders: Our MLP decoders had a small, single hidden layer of 20 units (10% of the input layer size) to ensure that their performance would depend heavily on good preprocessing by the SNN autoencoder, making the relative merits of the different SNN models and encoding schemes under investigation readily apparent in the MLP's classification performance (Bengio 2009). The MLPs had hidden layer neurons with tan-sigmoid transfer functions. They were trained to categorize the columns of their input matrices as belonging to one of the two stimulus classes using scaled conjugate gradient descent (Moller 1993) and a cross-entropy loss function. Once the MLPs were trained, their classification outputs were used to fill the confusion matrices (see Table 6.1) used in the information analysis calculations described in Chapter 2 Section 2.10.4. The process was repeated 20 times, each time with a different random subsample of J cells, to obtain a distribution of stimulus information estimates for each cell population.

6.3.9.2 Temporal Encoding Results

Figure 6.12 shows that the A1 cells with A1→Belt afferent connections (w_{ij}^{BL}) strengthened up to within the 45-55 nA range have the most multiple cell information compared to the other three ranges tested, including the population of A1 cells with the maximally strengthened A1→Belt connections. This supports the hypothesis outlined above that stable patterns of firing within the A1 that are discriminative of the two naturally spoken digit stimuli result in moderate, rather than maximal, strengthening of the A1→Belt afferent connections. Furthermore, it can be seen that there is significantly more information about the naturally spoken digits when the precise spike times of A1 cells are considered compared to the simple on/off encoding shown in Figure 6.4. This provides evidence for temporal rather than rate encoding of complex auditory objects, such as naturally spoken words, in the auditory cortex.

Due to the complexity of the naturally spoken digits, the stimulus selective transform invariant learning is not finished within the A1 layer of the AN-CN-IC-

		Actual	
		Stimulus 1	Stimulus 2
Predicted	Stimulus 1	$\frac{\#PredictStim1}{\#Stim1*N}$	$\frac{\#PredictStim1}{\#Stim2*N}$
	Stimulus 2	$\frac{\#PredictStim2}{\#Stim1*N}$	$\frac{\#PredictStim2}{\#Stim2*N}$

Table 6.1: Joint probability table used to calculate multiple cell information using classifier generated encoding. $\#PredictStim1$ ($\#PredictStim2$) stands for the number of presentations of stimulus 1 (2) during which a pre-trained backpropagation network classified the input matrix as Stimulus 1 (2). $\#Stim1$ ($\#Stim2$) stands for the total number of presentations of stimulus 1 (2). N stands for the total number of stimuli, which is two for the two naturally spoken digit discrimination paradigm.

A1 model. This could be due to the fact that the duration of naturally spoken digits (702 ms on average) is significantly longer than the maximum axonal conduction delay present within a single cortical layer (50 ms). The addition of more hierarchical feed-forward cortical layers with [0, 50] ms distributions of axonal delays would result in output layer cells that are exposed to and are capable of learning longer stimulus segments. Alternatively, the addition of plastic recurrent connections within the cortical layers can help, however this is beyond the scope of this thesis. The A1 layer, therefore, serves as the initial stage for the development of stimulus selective transform invariant representations of the auditory objects, whereby extra plastic cortical layers following the A1 layer may continue developing stimulus specific transform invariant representations of the complex naturally spoken words, such as digits 'one' and 'two'. This is supported by the increased multiple cell information within the Belt area of the full AN-CN-IC-A1-Belt model. Multiple cell information for the Belt layer (as well as all other layers which do not have strengthened afferent connections to the next layer in the hierarchy) was calculated using random samples of 100 cells. When the trained model was tested on additional data of the same words 'one' and 'two' being spoken by twenty novel speakers (ten male and ten female speakers pronouncing each word twice), the network reached a similar level of performance as described above (0.27 vs 0.28 bits in the A1, and 0.48 vs 0.43 bits in Belt).

It was found that less information was present within the A1 of the trained or untrained simple AN-A1-Belt model, or the corresponding Belt layer. This suggests that the simple AN-A1 model without sub-cortical pre-processing is un-

able to learn transform invariant representations of naturally spoken digits, even if extra cortical layers are added to the hierarchy.

The lower level of multiple cell information in the A1 layer of the untrained AN-

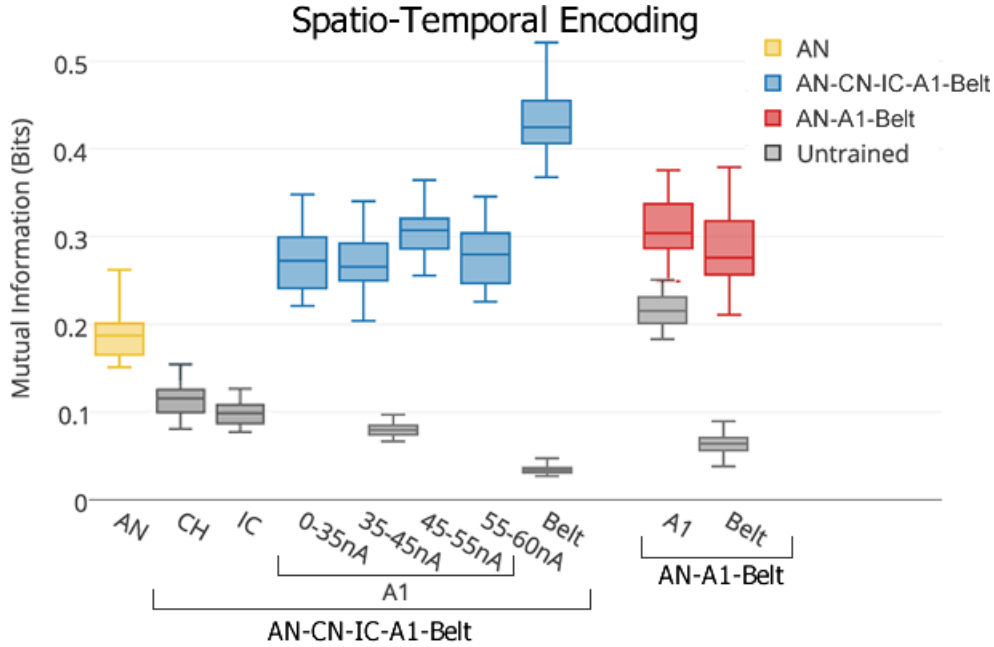


Figure 6.12: Box-and-whisker plots showing the distribution of multiple cell information over twenty different subsamples of cells within various layers of the full AN-CN-IC-A1-Belt and reduced AN-A1-Belt models based on an encoding scheme using polychronous chain firing patterns (see text for the description of the method). For the A1 layer of the trained AN-CN-IC-A1 model different sub-populations were used for the calculations depending on the strength of their connections to the Belt layer. It was hypothesised that the most stimulus discriminative stable polychronous chain firing patterns would come from A1 cells which have bounded strengthened A1→Belt connections within the [45, 55] nA range. Indeed, it can be seen that these cells result in the higher amount of multiple cell information among the four different sub-groups (first four left-most bars of the plot). Due to the complexity of the naturally spoken stimuli, extra cortical layers are necessary following the A1 in order to continue learning transform invariant representations of auditory objects. Indeed, it can be seen that the Belt layer of the full AN-CN-IC-A1-Belt model has more information than the A1 layer. It can also be seen that the amount of information rises from the CH layer within the CN, to the IC, to A1 and then Belt within the AN-CN-IC-A1-Belt model. No layer within the reduced AN-A1-Belt model is able to reach the same amount of multiple cell information as the A1 or Belt layers of the full AN-CN-IC-A1-Belt model. PL and ON layers are not shown. The ON layer conveys 0 bits of information, and PL is equivalent to AN.

CN-IC-A1-Belt model compared to the trained AN-CN-IC-A1-Belt model suggests that plasticity within the IC→A1 afferent connections is necessary to allow the model to start learning transform invariant representations of naturally spoken digits, even though this was not shown to be the case when learning transform invariant representations of synthesised vowels (see Chapter 5 Section 5.8.2).

An additional control simulation was run to test the ability of an untrained AN-CN-IC-A1-Belt model with the same distribution of IC-A1 and A1-Belt afferent connection weights as the trained model to discriminate between the two stimulus classes. This was done by randomly shuffling the corresponding connection weights of the trained AN-CN-IC-A1-Belt model before presenting it with the two naturally spoken word stimuli. It was found that such a model performed at the same level as the untrained model with afferent connection weights initialised from the [30, 35] nA uniform distribution (0.05 vs 0.08 bits in the A1, and 0.04 vs 0.03 bits in Belt).

6.3.10 Examining Polychronous Chains

An example of polychronous chains emerging in the A1 layer of the trained AN-CN-IC-A1 model when presented with naturally spoken digits 'one' and 'two' can be seen in Figure 6.13. The figure consists of two sub-plots, each one representing a potential polychronous chain that has learnt to be stimulus selective with a preference either for the exemplars of digit 'one' or 'two'. The red dots represent spike time locations for A1 cells that are more likely to have a spike in response to the chain's preferred stimulus. They were found when calculating the within layer firing synchrony as described in Section 6.3.7, whereby they correspond to the unique matrix components for each stimulus that have the highest scores. These red dots correspond to the cell firing times that are likely to be a part of a polychronous chain if the right distribution of axonal delays between these A1 cells and a Belt cell exist, which would result in the A1 spikes arriving simultaneously to a Belt cell, thus exciting it.

Randomly initialised combinations of axonal delays to all Belt cells from the A1 cells pinpointed by the red dots within Figure 6.13 were examined to find whether a combination exists that results in the spikes from at least four A1 cells within the potential polychronous chains to arrive to a single Belt cell simultaneously. If such a combination of axonal delays exists, then the A1 cells would provide super-threshold activation to the corresponding Belt cell. The co-activity of the A1 cells within the polychronous chain and their corresponding Belt cell would allow STDP-based learning to reinforce the polychronous chain and make it more informative about its preferred stimulus.

When the randomly initialised combinations of axonal delays to all Belt cells from the A1 cells within the potential polychronous chains for the two naturally

spoken digits were considered, it was found that two good combinations exist for stimulus 'one', and one good combination exists for stimulus 'two'. The red dots of the potential polychronous chains are projected through these matching axonal delays and the resulting rasters are plotted in green and yellow in Figure 6.13.

It can be seen that the red dots within the left subplot of Figure 6.13 which were assumed to be potentially a part of a single polychronous chain selective for 'one' are in fact two polychronous chains. A1 cells 958, 921, 573, 540, 187 and 65 have the right combination of axonal delays to arrive to the Belt cell 971 simultaneously. Whereas A1 cells 573, 374, 42 and 2 have the right combinations of axonal delays to arrive to the Belt cell 412 simultaneously. It can also be seen that A1 cell 573 takes part in both polychronous chains, whereas A1 cell 318 appears to take part in none. Similarly, all A1 cells highlighted to be a likely part of the polychronous chain selective for stimulus 'two' (Fig. 6.13, right sub-plot) apart from cells 772 and 811 were found to have the right combination of axonal delays to arrive to the Belt cell 642 simultaneously.

This chapter has demonstrated for the first time how a biologically plausible

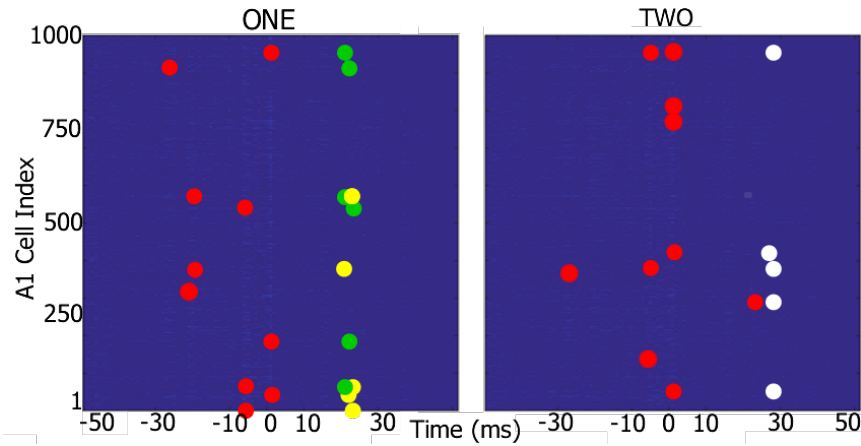


Figure 6.13: Evidence for PGs responding selectively to word 'one' or 'two' in the A1 layer of the trained AN-CN-IC-A1-Belt model. Red dots show the ten elements of the firing pattern matrix M_s^j with the largest mean spike counts $(\mathbf{m}_s^j)_n$ (see Chapter 2 Section 2.10.3.4). They can be thought of as representative subsets of distributed firing patterns often observed in the neural responses to the respective stimuli. When projected through the A1→Belt connections, these patterns produce near-synchronous input from several A1 neurons onto a subset of Belt neurons. The green and yellow dots show such inputs for two Belt neurons which in this manner respond selectively for a number of different pronunciations of the word 'one', the white dots show inputs for a neuron that responded selectively to exemplars of the word 'two'.

spiking neural network of the auditory brain is able to start learning transform invariant representations of naturally spoken words. It further strengthens the claim of this thesis that two stage sub-cortical pre-processing within the sub-populations of the CN followed by convergence within the IC is important for auditory object learning and recognition, since otherwise CT STDP-based learning is unable to cope with the high degree of jitter present in the AN. This chapter has also demonstrated the importance of learning within a number of cortical layers using the reduced jitter inputs from the IC. Such learning results in increased amounts of multiple cell information throughout the hierarchy of plastic cortical layers with axonal delays. The information about the naturally spoken digits is encoded within the precise spike times of populations of cortical cells forming polychronous chains (Izhikevich 2006), since rate on/off encoding is shown to carry significantly less information about the naturally spoken digits. Examples of such stimulus selective polychronous chains are shown in Figure 6.13. This chapter, therefore, provides evidence for the necessity of temporal rather than rate population encoding in the auditory brain when learning transform invariant representations of naturally spoken words.

7

Conclusion

7.1 Introduction

How does the auditory brain learn to robustly recognise auditory objects, such as naturally spoken words, in a transform invariant manner even despite the huge variability within the raw auditory wave inputs? What are the areas within the extensive auditory brain hierarchy that are important for this task? What is the simplest neural code sufficient to represent the learnt auditory objects within the output layers of the auditory brain hierarchy? Does the brain use rate or temporal encoding to represent auditory objects? These are some of the questions that have been addressed within this thesis. This chapter discusses the strengths and limitations of the work described in this thesis, as well as suggesting some avenues for future research.

7.2 The Role of Sub-Cortical Pre-Processing

This first major result of this thesis is that a simple AN-A1 biologically plausible spiking neural network model of the auditory brain is unable to learn transform invariant representations of auditory objects, such as synthesised vowels or naturally spoken words, through unsupervised spike-time dependent plasticity (STDP) learning (Bi & Poo 1998) if it has to learn straight from the input auditory nerve (AN) firing rasters. This is due to space and time jitter present in the AN primarily due to its phase-locking property and the volley principle (Wever & Bray 1930).

This jitter breaks stable AN firing patterns that are necessary for STDP-based learning to operate.

This thesis has shown that non-plastic sub-cortical pre-processing is necessary to remove AN jitter and thus enable transform invariance STDP-based learning to take place. This pre-processing was shown to happen in the sub-populations of the ventral cochlear nucleus (CN) followed by the inferior colliculus (IC), which serves as a whiteboard for the CN sub-populations to converge on. In particular, it has been suggested that chopper and onset CN sub-populations have the right connectivity and response properties to remove space and time jitter respectively. It has also been demonstrated that primary-like CN cells are important for auditory object learning by providing background activation to the IC, even though they do not play a direct role in jitter removal.

The role of sub-cortical pre-processing for transform invariance learning of auditory objects is in line with the neurophysiological data, which suggests that early sub-cortical auditory areas (AN, CN and TB) prepare sensory input for processing in the higher order, highly plastic auditory brain areas (IC, MGN and the cortex) (Covey & Casseday 1999, Nelken et al. 2003, Sanes & Constantine-Paton 1985, Weinberger & Diamond 1987, Weinberger et al. 1990). While the neurophysiologists have been able to speculate about the role of sub-cortical auditory brain areas, no precise hypothesis for their exact role in learning auditory object representations has been agreed on (Bizley & Cohen 2014). This thesis hopes to provide a neurophysiologically testable hypothesis for the role of the chopper, onset and primary-like sub-populations of the CN and their convergence on the IC for learning transform invariant representations of auditory objects.

7.2.1 **Future Work: Testing Neurophysiological Predictions**

The first major result of this thesis outlined above provides a neurophysiologically testable hypothesis for the role of the CN and IC layers in learning transform invariant representations of auditory objects. It is suggested that these areas act as the jitter removal stage of the auditory brain. While neurophysiological studies have provided good characterisations of response properties of neurons in the sub-cortical areas of the auditory brain (Delgutte 1982, Ferragamo et al. 1998, Godfrey et al. 1975, Liberman 1993, Rhode et al. 2010, Winslow et al. 1987), their exact computational role in auditory object learning and recognition tasks has so far been a mystery. This thesis suggests the mechanisms through which the chopper and onset sub-populations of the CN remove jitter from the AN spike rasters, before converging together with the primary-like CN sub-population on the IC to create stable spike rasters suitable for STDP transform invariance learning.

One avenue of future research, therefore, would be to obtain neurophysiological evidence to verify the jitter-removal role of the CN. This could be done by

measuring the levels of jitter in the AN, CN and IC responses to auditory stimuli.

Furthermore, this thesis suggests the potential minimal subset of auditory brain areas sufficient for learning transform invariant representations of auditory objects. It is suggested that such a minimal subset consists of the auditory nerve, followed by the chopper, onset and primary-like sub-populations of the cochlear nucleus, which in turn converge on the inferior colliculus. The inferior colliculus then provides input to the auditory cortex. This prediction could be tested using neurophysiological techniques by selectively knocking out the auditory brain areas within the proposed minimal subset and investigating whether it significantly disrupts transform invariant learning of auditory objects.

7.3 New View on the Volley Principle in the Auditory Nerve

This thesis proposes a novel view on stochasticity and noise within the AN firing by categorising it into space and time jitter and highlighting the role of this jitter in obstructing transform invariance learning of auditory objects, including speech. While the volley principle is a well studied phenomenon within the AN, its role in auditory processing has primarily been seen as allowing AN fibers to represent high frequency sounds through temporal encoding despite their 200 Hz maximum firing rate threshold.

This thesis suggests that there is a flip side to the beneficial role of the volley principle, whereby it introduces additional jitter to the AN firing, which breaks STDP-based learning and creates the necessity for the sub-cortical pre-processing within the CN and IC. This also explains why the sub-cortical architecture of the auditory brain is more complex than that of the visual brain, since jitter is not present in the inputs to the visual cortex to the same extent as it is present in the AN. Thus no jitter reducing equivalent of the sub-cortical pre-processing suggested for the auditory brain is necessary for the visual cortex to learn transform invariant representations of visual objects (Higgins & Stringer 2011, Stringer et al. 2006, Tromans et al. 2012).

7.4 Continuous Transformation Learning

The auditory brain has to learn to associate together different versions of the same speech object in order to solve the invariant speech recognition problem. A similar problem has been solved in vision using both rate and spiking neural network models of the ventral visual stream using the biologically plausible, unsupervised, continuous transformation (CT) learning mechanism (Evans & Stringer 2012, Hig-

gins & Stringer 2011, Stringer et al. 2006, Tromans et al. 2012). CT learning is essentially an emergent property of Hebbian learning (Hebb 1949) that requires high overlap between examples of the same stimulus class.

This thesis has demonstrated that a version of CT learning adjusted to cope with the temporal dimension of auditory objects by introducing a distribution of axonal conduction delays to the afferent plastic connections can enable the auditory brain to solve the auditory object invariance problem using the pre-processed jitter-free IC input. This, therefore, suggests that CT learning is a general mechanism utilised by the cortex to learn statistical regularities of stimuli regardless of the sensory domain.

7.5 Spiking Neuron Model by Izhikevich (2003)

One of the limitations of the work discussed in this thesis is the choice of the spiking neuron model by Izhikevich (2003). While this choice has allowed the computationally expensive simulations described in this thesis to run by limiting their computational cost, the parameters of the model have no direct biophysical meaning. This makes it hard to make certain neurophysiological predictions and to tune the model parameters to represent certain exact values found in the auditory brain.

7.6 Training Data

This thesis was a first step towards finding the minimal taxonomy of brain areas and computations important for speech recognition. For this reason, the stimuli selected for training the models in this thesis, as well as the training paradigm are rather simplistic. This was necessary to allow for controlled and systematic investigation of how the auditory brain may learn to extract important information about auditory class identity from the stochastic auditory nerve input, whereby the information encoding is simplified and redundancy is reduced (Barlow 1961). The stimuli and the training paradigm, however, were a simplification of the richness of the natural auditory world, where sounds of interest almost never occur in isolation and have to be disentangled from other auditory signals, they are presented in random order and as part of continuous speech rather than in isolation. Hence, future work should investigate the nature of representations learnt by the model suggested in this thesis when exposed to the more realistic training paradigms.

7.7 Increasing the Capacity of the AN-CN-IC-A1 Model

This thesis has demonstrated that the full AN-CN-IC-A1 model of the auditory brain is able to learn transform invariant representations of up to six synthesised vowels. It is also capable of learning significantly better transform invariant representations of two naturally spoken digits 'one' and 'two' than the simple AN-A1 model. The output A1 layer of the full AN-CN-IC-A1 model, however, is unable to reach the theoretical maximum of 1 bit of information for the two naturally spoken digit recognition task. This is due to the increased complexity of the naturally spoken stimuli. Future work, therefore, could investigate how the learning capacity of the network can be increased through either changes to its architecture, or through the changes in the network parameters.

7.7.1 Future Work: Adding Extra Feed-Forward Cortical Layers

It has been shown that the addition of an extra cortical layer following the A1 results in a significant increase in the amount of information learnt about the two naturally spoken words. Future work, therefore, could investigate how the learning capacity of the network can be increased when more plastic cortical layers with randomly initialised axonal delays are added. Such an extension to the network architecture would allow the new output layer neurons to receive convergent information from longer temporal patterns of the input stimuli. If each additional cortical layer is initialised with axonal delays drawn from the uniform distribution of $[0, 50]$ ms, then neurons in the fourth cortical layer would be able to learn temporal patterns of up to 200 ms long. The addition of too many extra layers, however, can quickly become infeasible due to the computational constraints.

7.7.2 Future Work: Increasing the Average Connection Weight Magnitudes

Another approach to increasing the network capacity would be to increase the average connection weight magnitudes of the plastic cortical afferent connections. This would affect the average number of pre-synaptic neurons necessary to activate a post-synaptic neuron. Higher average connection weight magnitudes would result in fewer concurrently arriving inputs from pre-synaptic cells being necessary to activate a post-synaptic cell. There would, therefore, be a higher probability that the right distribution of axonal conduction delays is present within the post-synaptic layer to receive concurrent inputs from the smaller sets of pre-synaptic

neurons needed to activate the post-synaptic cells. This, in turn, would mean that there will be a larger vocabulary of temporal spike patterns that the network could potentially learn. Such a larger vocabulary would make it more likely that the input stimuli would match some patterns from this vocabulary, which would then activate the post-synaptic cells corresponding to these matching patterns, and thus enable STDP-based learning. Increasing the average connection weight magnitudes within a cortical layer would, therefore, be analogous to increasing the number of neurons within the layer.

7.7.3 Future Work: Adding Recurrent Connections

Another approach to increasing the network capacity would be to add recurrent lateral plastic connections with a distribution of axonal delays within the A1. While networks with recurrent plastic connections are notoriously hard to optimise to ensure that the activity within the recurrent layer does not explode (Izhikevich 2004), such a change would serve three important purposes: 1) it would increase the temporal tuning of A1 neurons, whereby each cell would be able to learn longer segments of the input stimuli; 2) it would increase the capacity of the network, since the learnt polychronous chains would not be strictly feed-forward any more, and therefore more polychronous chains would be able to co-exist within the same population of neurons (Izhikevich 2004); 3) it would make the model more robust when being presented with degraded or noisy inputs. Furthermore, the addition of plastic recurrent connections to the A1 stage of the model will essentially transform A1 into a 'liquid' of a Liquid State Machine (LSM) (see Chapter 1 Section 1.6.2.5). It will, therefore, be interesting to compare the performance of the AN-CN-IC-A1 model with such recurrent connections to the performance of LSMs on similar auditory tasks (Maass et al. 2004) in order to investigate the effects of STDP-based learning and the CN-IC sub-cortical preprocessing present in our model but lacking in traditional LSMs. Additionally, it would be interesting to investigate the differences in learnt polychronous chains in the recurrent A1 layer when the STDP mode is switched from the 'nearest only' paradigm, whereby only the closest pre-synaptic spikes participate in STDP, to the 'all pairings' paradigm, whereby all pre-synaptic spikes for each cell take part in STDP (Kempster et al. 1999, Song et al. 2000). Humble et al. (2012) have demonstrated that the two different paradigms result in qualitatively different polychronous chain profiles in a recurrent spiking neural network.

7.7.3.1 Future Work: Testing Noise Robustness in the Recurrent AN-CN-IC-A1 Model

The noise robustness property of the recurrent A1 layer of the AN-CN-IC-A1 model as suggested for future research could be tested in a simulation described next. The proposed recurrent architecture is expected to preserve higher levels of information about the identity of speech stimuli in lower signal to noise ratio (SNR) conditions. This would happen because at each point in time the A1 neurons representing the part of the auditory signal that is most likely to come next would be primed through the strengthened recurrent connections. In high levels of noise these primed connections can help the network activate the right stimulus selective neurons in the A1 layer, and hence the auditory object will be more likely to be recognised. This priming is expected to happen because the particular pattern of statistical regularity within the stimulus set will be represented in the strengthened network connections acquired through STDP-based learning during training. The network is expected to fail to recognise input stimuli that do not follow the learnt statistical regularity rules in low SNR levels.

Future research may test the hypothesis outlined above by using an artificial simplified speech-like stimulus set created using statistical transitional probability rules. A control stimulus set can also be created using the same sounds as the test set, but with reversed transitional probability distributions. The network responses can then be compared to the neurophysiological data that can be obtained by raising and testing animals using the same experimental paradigm as the one used in the proposed simulation study.

7.8 Neural Code

The hierarchy of the auditory brain is very complex, with numerous inter-connected predominantly non-plastic sub-cortical and plastic cortical areas. It is unknown how sensory features of auditory objects, such as naturally spoken words, are transformed into higher-order representations of object identities in the auditory brain (Russ et al. 2008). Even the nature of auditory object representations themselves is yet to be confirmed (Eggermont 2001, Mesgarani & Chang 2012, Russ et al. 2008). Apart from the question of whether auditory objects are encoded by single cells or populations of neurons, another ongoing debate is whether they are encoded within the temporal structure in the timings of spikes, or are effectively rate-coded, whereby the precise timing of spikes emitted by the cells would not convey additional information about the auditory stimulus (Cariani 1999, Eggermont 2001, Huetz et al. 2011).

Due to the information inequality rule (Nelken 2004), AN rasters have to con-

tain all available information about auditory objects. This has to be true, because throughout the hierarchy of the auditory brain no new information is introduced, and information can only be lost. If full information is present about the auditory objects in the input AN layer, then the question arises as to the role of the rest of the auditory brain. This thesis argues that one of the roles of the auditory brain hierarchy is to transform the information about the identity of auditory objects from the more complex form in the AN, to the simplest sufficient form at the end of the hierarchy. The nature of this simplest sufficient neural code is discussed throughout the thesis. The simplification of the neural code should happen consequently with the acquisition of better stimulus selectivity and transform invariance.

One of the main restrictions for understanding the nature of the auditory neural code in the real brain is lack of simultaneous firing information from a sufficiently large number of individual neurons necessary to detect potentially temporal codes in a distributed population of neurons (Schnupp et al. 2011). Numerous researchers have suggested that this problem can be circumvented by analysing the response properties to natural auditory stimuli of a sufficiently realistic model of the auditory brain (Arbib et al. 1995, Corchs & Deco 2002, Horwitz & Tagamets 1999, Husain et al. 2004, Salem et al. 2007). This is exactly what has been done using the full AN-CN-IC-A1 model trained on two to six synthesised vowels or two naturally spoken words.

It was found that as the complexity of the task was increased, the simplest type of encoding in the output layer of the full AN-CN-IC-A1 model sufficient for representing the full stimulus set in a transform invariant manner changed. For the simplest two synthesised vowel recognition task a simple 'grandmother'-like on/off rate information encoding scheme was found to be sufficient. This scheme presumed that the information about the vowel identity was encoded in whether a single output layer neuron fired at least once during stimulus presentation or not. Once the number of vowels in the training set was increased from two up to six, the 'grandmother'-like rate on/off encoding was not sufficient any more. Instead a simple temporal encoding scheme which took into account the precise timing of spikes of individual output layer cells with respect to stimulus onset was considered. It was found that population encoding like this was able to provide enough information to differentiate between up to six synthesised vowels.

When the full AN-CN-IC-A1 model was presented with the most complex set of stimuli, the two naturally spoken digits 'one' and 'two', the nature of neural code for natural speech encoding could be investigated. It was found that precise temporal patterns of firing of polychronous chains (Izhikevich 2004) spread over populations of output layer cells could encode the stimuli well. This is in line with the neurophysiological evidence that found that individual neurons at the end of the auditory processing hierarchy (areas pSTG and vPFC) respond to a

large number of species specific vocalisations (SSVs), suggesting that SSVs are encoded by overlapping populations of neurons instead of single cell call detectors (Russ et al. 2008, Schnupp et al. 2006).

7.9 Unique Attributes of the AN-CN-IC-A1 Model

The full AN-CN-IC-A1 model described in this thesis possesses a unique combination of components necessary to simulate the emergent neurodynamics of natural speech processing and recognition in the brain, such as biologically accurate spiking dynamics of individual neurons, STDP-based learning, neurophysiologically guided architecture and exposure to natural speech input. Due to its biological plausibility, the model can be used to make neurophysiologically testable predictions, and its responses can be used to provide insights into the nature of the neural code used by the auditory brain to represent auditory objects.

References

- Abeles, M. (1982*a*), *Local Cortical Circuits: An Electrophysiological Study*, Springer Berlin.
- Abeles, M. (1982*b*), 'Role of the cortical neuron: integrator or coincidence detector', *Israel Journal of Medical Sciences* **18**, 83–92.
- Abeles, M. (1991), *Corticonics, Neural Circuits of the Cerebral Cortex*, Cambridge University Press.
- Abeles, M., Bergman, H., Gat, I., Meilijson, I., Seidemann, E., Tishby, N. & Vaadia, E. (1995), 'Cortical activity flips among quasi-stationary states', *Proc. Natl. Acad. Sci. U. S. A.* **92**, 8616–8620.
- Adams, J. (1979), 'Ascending projections to the inferior colliculus', *J Comp Neurol* **183**, 519–538.
- Adams, J. (1997), 'Projections from octopus cells of the posteroventral cochlear nucleus to the ventral nucleus of the lateral lemniscus in cat and human', *Aud. Neurosci.* **3**, 335–350.
- Aitkin, L. & Webster, W. (1971), 'Tonotopic organization in the medial geniculate body of the cats', *Brain Research* **26**, 402–405.
- Alibardi, L. (1998), 'Ultrastructural and immunocytochemical characterization of neurons in the rat ventral cochlear nucleus projecting to the inferior colliculus', *Ann Anat* **180**, 415–426.
- Allon, N. & Yeshurun, Y. (1985), 'Functional organization of the medial geniculate body's subdivisions of the awake squirrel monkey', *Brain Research* **360**, 75–82.
- Allon, N., Yeshurun, Y. & Wollberg, Z. (1981), 'Responses of single cells in the medial geniculate body of awake squirrel monkey', *Exp Brain Res* **41**, 222–232.

- Amunts, K., Weiss, P. H., Mohlberg, H., Pieperhoff, P., Eickhoff, S., Gurd, J. M., Marshall, J. C., Shah, N. J., Fink, G. R. & Zilles, K. (2004), 'Analysis of neural mechanisms underlying verbal fluency in cytoarchitectonically defined stereotaxic space: The roles of brodmann areas 44 and 45', *NeuroImage* **22**(1), 42–56.
URL: <http://www.sciencedirect.com/science/article/pii/S1053811904000059>
- Andersen, P., Junge, K. & Sveen, O. (1972), 'Cortico-fugal facilitation of thalamic transmission', *Brain Behav. Evol.* **6**, 170–184.
- Appeltant, L., Soriano, M. C., Van der Sande, G., Danckaert, J., Massar, S., Dambre, J., Schrauwen, B., Mirasso, C. R. & Fischer, I. (2011), 'Information processing using a single dynamical node as complex system', *Nature Communications* **2**(468).
- Arbib, M., Bischoff, A., Fagg, A. & Grafton, S. (1995), 'Synthetic pet: analyzing large-scale properties of neural networks', *Hum. Brain Mapp.* **2**, 225–233.
- Arthur, R. M., Pfeiffer, R. R. & Suga, N. (1971), 'Properties of 'two-tone inhibition' in primary auditory neurones', *J. Physiol.* **212**, 593–609.
- Atiani, S., Elhilali, M., David, S. V., Fritz, J. B. & Shamma, S. A. (2009), 'Task difficulty and performance induce diverse adaptive patterns in gain and shape of primary auditory cortical receptive fields', *Neuron* **61**, 469–480.
- Atzori, M., Lei, S., Evans, D., Kanold, P., Phillips-Tansey, E., McIntyre, O. & McBain, C. (2001), 'Differential synaptic processing separates stationary from transient inputs to the auditory cortex', *Nat. Neurosci.* **4**, 1230–1237.
- Bal, R. & Oertel, D. (2001), 'Potassium currents in octopus cells of the mammalian cochlear nucleus', *J. Neurophysiol.* **86**, 2299–2311.
- Banks, M. I. & Sachs, M. B. (1991), 'Regularity analysis in a compartmental model of chopper units in the anteroventral cochlear nucleus', *J Neurophysiol* **65**, 606–629.
- Barlett, E. L. & Smith, P. H. (2002), 'Effects of paired-pulse and repetitive stimulation on neurons in the rat medial geniculate body', *Neuroscience* **113**(4), 957–974.
- Barlow, H. B. (1961), 'Possible principles underlying the transformations of sensory messages', *Sensory Communication* **1**.

- Batra, R. & Fitzpatrick, D. C. (2002), ‘Monaural and binaural processing in the ventral nucleus of the lateral lemniscus: a major source of inhibition to the inferior colliculus’, *Hearing Research* **168**, 90–97.
- Bauer, J. W. (1978), ‘Tuning curves and masking functions of auditory-nerve fibers in cat’, *Sens Proc* **2**, 156–172.
- Bell, C. C., Han, V. Z., Sugawara, Y. & Grant, K. (1997), ‘Synaptic plasticity in cerebellum-like structure depends on temporal order’, *Nature* **387**, 278–281.
- Bengio, Y. (2009), ‘Learning deep architectures for ai’, *Foundations and Trends in Machine Learning* **2**(1), 1–127.
- Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V. & Wellekens, C. (2007), ‘Automatic speech recognition and speech variability: A review’, *Speech Communication* **49**, 763–768.
- Beyerl, B. D. (1978), ‘Afferent projections to the central nucleus of the inferior colliculus in the rat’, *Brain Research* **145**, 209–223.
- Bi, G.-Q. & Poo, M.-M. (1998), ‘Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type’, *J Neurosci* **18**, 10464–10472.
- Binder, J. R., Liebenthal, E., Possing, E. T., Medler, D. A. & Ward, B. D. (2004), ‘Neural correlates of sensory and decision processes in auditory object identification’, *Nat Neurosci* **7**(3), 295–301.
URL: <http://dx.doi.org/10.1038/nn1198>
- Bizley, J. K. & Cohen, Y. E. (2014), ‘The what, where and how of auditory-object perception’, *Nat Rev Neurosci* **14**(10), 693–707.
- Bizley, J. K., Walker, K. M., King, A. J. & Schnupp, J. W. (2010), ‘Neural ensemble codes for stimulus periodicity in auditory cortex’, *J Neurosci* **30**(4), 5078–91.
- Blackburn, C. C. & Sachs, M. B. (1989), ‘Classification of unit types in the anteroventral cochlear nucleus: Pst histograms and regularity analysis’, *J Neurophysiol* **62**, 1303–1329.
- Blackburn, C. C. & Sachs, M. B. (1990), ‘The representations of the steady-state vowel sound /ε/ in the discharge patterns of cat anteroventral cochlear nucleus neurons’, *Journal of Neurophysiology* **63**(5), 1191–1212.

- Bleeck, S., Sayles, M., Ingham, N. J. & Winter, I. M. (2006), 'The time course of recovery from suppression and facilitation from single units in the mammalian cochlear nucleus', *Hearing Research* **212**, 176–184.
- Boatman, D. F. (2006), 'Cortical auditory systems: Speech and other complex sounds', *Epilepsy & Behavior* **8**, 494–503.
- Boatman, D., Lesser, R. P. & Gordon, B. (1995), 'Functional organization of auditory speech processing in the left temporal cortex: evidence from functional lesions', *Brain Lang* **51**, 269–290.
- Boatman, D. & Miglioretti, D. (2005), 'Cortical sites critical for speech discrimination in normal and impaired listeners. j neurosci 2005;25:547580.', *J Neurosci* **25**, 5475–5480.
- Bohte, S. M. & Mozer, M. C. (2004), 'Reducing spike train variability: A computational theory of spike-timing dependent plasticity', *NIPS* **14**.
- Bradbury, J. W. & Vehrencamp, S. L. (1998), *Principles of Animal Communication*, Oxford: Blackwell.
- Brawer, J. R., Morest, D. K. & Kane, E. C. (1974), 'The neuronal architecture of the cochlear nucleus of the cats', *J. Comp. Neur.* **155**, 251–300.
- Bronzino, J. & Peterson, D. (2006), *Biomedical Engineering Fundamentals*, CRC Press Inc.
- Brownell, W. E. (1975), 'Organization of the cat trapezoid body and the discharge characteristics of its fibers', *Brain Research* **94**, 413–433.
- Browner, R. & Webster, D. (1975), 'Projections of the trapezoid body and the superior olivary complex of the kangaroo rat (*dipodomys merriami*)', *Brain Behav. Evol.* **11**, 322–354.
- Bruce, I. C., Sachs, M. B. & Young, E. D. (2003), 'An auditory-periphery model of the effects of acoustic trauma on auditory nerve responses', *J Acoust Soc Am* **113**(1), 369–388.
- Brugge, J. F., Anderson, D. J., Hind, J. E. & Rose, J. E. (1969), 'Time structure of discharge in single auditory nerve fibers of the squirrel monkey in response to complex periodic sounds', *Neurophysiology* **32**, 386–401.
- Brugge, J. F. & Reale, R. A. (1985), *Cerebral cortex*, New York: Plenum, chapter Auditory cortex, pp. 229–271.

- Bunge, S. A., Kahn, I., Wallis, J. D., Miller, E. K. & Wagner, A. D. (2003), 'Neural circuits subserving the retrieval and maintenance of abstract rules', *J Neurophysiol* **90**, 3419–3428.
- Buonomano, D. V. & Merzenich, M. M. (1998), 'Net interaction between different forms of short-term synaptic plasticity and slow-ipsps in the hippocampus and auditory cortex', *J Neurophysiol* **80**, 1765–1774.
- Burton, H. & Jones, E. (1976), 'The posterior thalamic region and its cortical projection in new world monkeys', *J. Comp. Neur.* **168**, 249–302.
- Cant, N. B. (1981), 'The fine structure of two types of stellate cells in the anterior division of the anteroventral cochlear nucleus of the cat', *Neuroscience* **6**(12), 2643–2655.
URL: <http://www.sciencedirect.com/science/article/pii/0306452281901093>
- Cant, N. B. & Gaston, K. C. (1982), 'Pathways connecting the right and left cochlear nuclei.', *J. Comp. Neur.* **212**, 313–326.
- Cant, N. B. & Hyson, R. L. (1992), 'Projections from the lateral nucleus of the trapezoid body to the medial superior olivary nucleus in the gerbil', *Hearing Research* **58**, 26–34.
- Cant, N. B. & Morest, D. K. (1978), 'Axons from non-cochlear sources in the anterior division of the anteroventral cochlear nucleus of the cat. a study with the rapid golgi method.', *Neuroscience* **3**, 1003–1029.
- Cant, N. B. & Morest, D. K. (1984), *Hearing Science Recent Advances*, College Hill: San-Diego, California, chapter Response characteristics of neurons in the cochlear nuclei, pp. 371–422.
- Cariani, P. A. (1995), 'As if time really mattered: Temporal strategies for neural coding of sensory information', *Communication and Cognition-Artificial Intelligence* **12**(1-2), 157–219.
- Cariani, P. A. (1999), 'Temporal coding of periodicity pitch in the auditory system: an overview', *Neural Plast* **6**, 147–172.
- Carianni, P. & Delgutte, B. (1996), 'Neural correlates of the pitch of complex tones. ii. pitch shift, pitch ambiguity, phase-invariance, pitch circularity, rate pitch, and the dominance region of pitch', *J Neurophysiol* **76**, 1717–1734.
- Carney, L. H. (1993), 'A model for the responses of low-frequency auditory-nerve fibers in cat', *J Acoust Soc Am* **93**(1), 410–417.

- Chang, E. F. & Merzenich, M. M. (2003), 'Environmental noise retards auditory cortical development', *Science* **300**, 498–502.
- Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M. & Knight, R. T. (2010), 'Categorical speech representation in human superior temporal gyrus', *Nat Neurosci* **13**(11), 1428–1432.
URL: <http://dx.doi.org/10.1038/nn.2641>
- Clarey, J. C., Paolini, A. G., Grayden, D. B., Burkitt, A. N. & Clark, G. M. (2004), 'Ventral cochlear nucleus coding of voice onset time in naturally spoken syllables', *Hearing Research* **190**, 37–59.
- Clopton, B. M. & Winfield, J. A. (1976), 'Effect of early exposure to patterned sound on unit activity in rat inferior colliculus', *J Neurophysiol* **39**, 1081–1089.
- Coath, M., Sheik, S., Chicca, E., Indiveri, G., Denham, S. L. & Wennekers, T. (2014), 'A robust sound perception model suitable for neuromorphic implementation', *Frontiers in Neuroscience* .
- Cohen, Y. E., Hauser, M. D. & Russ, B. E. (2006), 'Spontaneous processing of abstract categorical information in the ventrolateral prefrontal cortex', *Biology Letters* **2**, 261–265.
- Comis, S. D. (1970), 'Centrifugal inhibitory processes affecting neurones in the cat cochlear nucleus.', *J. Physiol. London* **210**, 751–760.
- Cooper, F. (1980), 'Acoustics in human communication: evolving ideas about the nature of speech', *J. Acoust. Soc. Am.* **68**, 18–21.
- Cooper, N. P. & Rhode, W. S. (1992), 'Basilar membrane mechanics in the hook region of the cat and guinea-pig cochleae: Sharp tuning and nonlinearity in the absence of baseline position shifts', *Hearing Research* **63**(1/2), 163–190.
- Corchs, S. & Deco, G. (2002), 'Large-scale neural model for visual attention: integration of experimental single-cell and fmri data', *Cereb. Cortex* **12**, 339–348.
- Cover, T. & Thomas, J. (1991), *Elements of Information Theory*, Wiley and Sons, NY.
- Covey, E. (2000), 'Neural population coding and auditory temporal pattern analysis.', *Physiol. Behav.* **69**, 211–220.
- Covey, E. & Casseday, J. H. (1999), 'Timing in the auditory system of the bat', *Annu. Rev. Physiol.* **61**, 457–476.

- Creutzfeldt, O., Hellweg, F. C. & Schreiner, C. (1980), 'Thalamocortical transformation of responses to complex auditory stimuli', *Exp Brain Res* **39**, 87–104.
- Dahmen, J. C., Hartley, D. E. H. & King, A. J. (2008), 'Stimulus-timing-dependent plasticity of cortical frequency representation', *Journal of Neuroscience* **28**(50), 13629–13639.
- Dale, H. H. (1934), 'Pharmacology and nerve-endings', *Proceedings of the Royal Society of Medicine* **28**(3), 319–330.
- Davis, K. A. (2005), Spectral processing in the inferior colliculus, in M. S. Malmierca & D. R. F. Irvine, eds, 'Auditory Spectral Processing', Vol. Volume 70, Academic Press, pp. 169–205.
URL: <http://www.sciencedirect.com/science/article/pii/S0074774205700064>
- Davis, M. H. & Johnsrude, I. S. (2003), 'Hierarchical processing in spoken language comprehension', *J. Neurosci.* **23**, 34233431.
- De Boer, E. & De Jongh, H. R. (1978), 'On cochlear encoding: Potentialities and limitations of the reverse-correlation technique', *J. Acoust. Soc. Am.* **63**, 115–135.
- De Boer, E. & Kuyper, P. (1968), Triggered correlation, in 'IEEE Trans. Biomed. Eng.'.
- de Saussure, F. (1959), *Course in General Linguistics*, McGraw-Hill, New York.
- de Villers-Sidani E., Chang, E. F., Bao, S. & Merzenich, M. M. (2007), 'Critical period window for spectral tuning defined in the primary auditory cortex (a1) in the rats', *J Neurosci* **27**, 180–189.
- Deacon, T. W. (1997), *The Symbolic Species: The Co-evolution of Language and the Brain*, W. W. Norton & Company.
- Debanne, D., Gahwiler, B. H. & Thompson, S. M. (1996), 'Cooperative interactions in the induction of long-term potentiation and depression of synaptic excitation between hippocampal ca3-ca1 cell pairs in vitro.', *Proc Natl Acad Sci USA* **93**, 1122511230.
- Debanne, D., Gahwiler, B. H. & Thompson, S. M. (1999), 'Heterogeneity of synaptic plasticity at unitary ca1-ca3 and ca3-ca3 connections in rat hippocampal slice cultures', *J Neurosci* **19**, 1066410671.

- Delgutte, B. (1980), 'Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers', *J. Acoust. Soc. Am.* **68**(3), 843–857.
- Delgutte, B. (1982), *Representations of Speech in the Peripheral Auditory System*, Elsevier: Amsterdam, chapter Some correlates of phonetic distinction at the level of the auditory nerve, pp. 131–149.
- Delgutte, B. (1990), 'Two-tone rate suppression in auditory-nerve fibers: Dependence on suppressor frequency and level', *Hearing Research* **49**, 225–246.
- Delgutte, B., Hammond, B. M. & Cariani, P. A. (1998), *Psychophysical and Physiological Advances in Hearing*, Whurr Publishing, London, chapter Neural coding of the temporal envelope of speech: relation to modulation transfer functions, pp. 595–603.
- Delgutte, B. & Kiang, N. Y.-S. (1984), 'Speech coding in the auditory nerve: I. vowel-like sounds', *J. Acoust. Soc. Am.* **75**(3), 866–878.
- Demb, J. B., Desmond, J. E., Wagner, A. D., Vaidya, C. J., Glover, G. H. & Gabrieli, J. D. (1995), 'Semantic encoding and retrieval in the left inferior prefrontal cortex: a functional mri study of task difficulty and process specificity', *J Neurosci.* **15**(9), 5870–5878.
- Denes, P. B. & Pinson, E. N. (1993), *The Speech Chain: The Physics and Biology of Spoken Speech*, W. H. Freeman.
- Deneve, S. & Machens, C. K. (2016), 'Efficient codes and balanced networks', *Nature Neuroscience* **19**(3), 375–382.
- Deng, L. & Geisler, C. (1987), 'A composite auditory model for processing speech sounds', *J Acoust Soc Am* **82**(6), 2001–2012.
- DeWeese, M. R. & Zador, A. M. (2003), 'Binary coding in auditory cortex', *NIPS* .
- Dewson, J. H. (1964), 'Speech sound discrimination by cats', *Science* **144**, 555–556.
- Diesmann, M., Gewaltig, M. O. & Aertsen, A. (1999), 'Stable propagation of synchronous spiking in cortical neural networks', *Nature* **402**, 529–533.
- Dooling, R. J., Okanoya, K. & Brown, S. D. (1989), 'Speech perception by budgerigars (*melopsitacus undulatus*): the voiced-voiceless distinction', *Percept. Psychophys.* **46**, 65–71.

- Dorman, M. F., Raphael, L. J. & Liberman, A. M. (1979), 'Some experiments on the sound of silence in phonetic perception', *J. Acoust. Soc. Am.* **65**, 1518–1532.
- Doupe, A. J. & Kuhl, P. K. (1999), 'Birdsong and human speech: common themes and mechanisms', *Annu. Rev. Neurosci* **22**, 567–631.
- Duifhuis, H. (1980), 'Level effects in psychophysical two-tone suppression', *J. Acoust. Soc. Am.* **67**(3), 914–927.
- Edeline, J. M. (1990), 'Frequency-specific plasticity of single unit discharges in the rat medial geniculate body', *Brain Research* **529**, 109–119.
- Eggermont, J. J. (1976), 'Analysis of compound action potential responses to tone bursts in the human and guinea pig cochlea', *J Acoust Soc Am* **60**, 1132–1139.
- Eggermont, J. J. (1996), 'How homogeneous is cat primary auditory cortex? evidence from simultaneous single-unit recordings', *Audiol. Neurosci.* **2**, 79–96.
- Eggermont, J. J. (1997), *Acoustical Signal Processing in the Central Auditory System*, New York: Plenum Press, chapter Representation of amplitude modulated sounds in two fields in auditory cortex of the cat, pp. 303–319.
- Eggermont, J. J. (1998a), 'Is there a neural code?', *Neurosci. Biobehav. Rev.* **22**, 355–370.
- Eggermont, J. J. (1998b), 'Representation of spectral and temporal sound features in three cortical fields of the cat. similarities outweigh differences', *J Neurophysiol* **80**, 2743–2764.
- Eggermont, J. J. (2001), 'Between sound and perception: reviewing the search for a neural code', *Hearing Research* **157**, 1–42.
- Eggermont, J. J. (2008), 'The role of sound in adult and developmental auditory cortical plasticity', *Ear Hear* **29**(6), 819–29.
- Egorova, M., Ehret, G., Vartanian, I. & Esser, K. H. (2001), 'Frequency response areas of neurons in the mouse inferior colliculus. i. threshold and tuning characteristics', *Exp Brain Res* **140**, 145–161.
- Ehret, G. (1997), *The Central Auditory System*, Oxford University Press, New York, chapter The auditory midbrain, a "shunting yard" of acoustical information processing, pp. 259–316.

- Ehret, G. & Fischer, R. (1991), 'Neuronal activity and tonotopy in the auditory system visualized by c-fos gene expression', *Brain Research* **567**, 350–354.
- Ehret, G., Hage, S. R., Egorova, M. & Muller, B. A. (2004), *Auditory Signal Processing: Physiology, Psychoacoustics, and Models*, Springer Verlag, chapter Auditory maps in the midbrain: The inferior colliculus.
- Ehret, G. & Romand, R. (1996), *The Central Auditory System*, Oxford University Press, USA.
- Eimas, P. D., Siqueland, E. R., Jusczyk, P. & Vigorito, J. (1971), 'Speech perception in infants', *Science* **171**, 303–306.
- Elverland, H. H. (1977), 'Descending connections between the superior olivary and cochlear nuclear complexes in the cat studied by autoradiographic and horseradish peroxidase methods.', *Exp Brain Res* **27**, 397–412.
- Evans, B. (2012), PhD thesis, University of Oxford.
- Evans, B. D. & Stringer, S. M. (2012), 'Transformation-invariant visual representations in self-organizing spiking neural networks', *Front. Comput. Neurosci.* **6**(46), 1–19.
- Evans, E. F. (1975), *Handbook of sensory physiology*, Vol. 5, Springer-Verlag, Berlin, chapter Cochlear nerve and cochlear nucleus.
- Evans, E. F. (1992), 'Auditory processing of complex sounds: An overview reviewed', *Philosophical Transactions: Biological Sciences* **336**(1278), 295–306. Unread.
- Evans, E. & Nelson, P. (1973), 'The responses of single neurons in the cochlear nucleus of the cat as a function of their location and anesthetic states', *Exp Brain Res* **117**, 402–427.
- Fant, C. G. M. (1960), *Acoustic Theory of Speech Production*, Mouton, The Hague.
- Fant, G. (1973), *Speech sounds and features*, Cambridge, MA: MIT Press.
- Feng, J. J., Kuwada, S., Ostapoff, E. M., Batra, R. & Morest, D. K. (1993), 'A physiological and structural study of neuron types in the cochlear nucleus. i. intracellular responses to acoustic stimulation and current injection', *J Comp Neurol* **346**(1), 1–18.

- Ferragamo, M., Golding, N. & Oertel, D. (1998), 'Synaptic input to stellate cells in the ventral cochlear nucleus', *J Neurophysiol* **79**, 51–63.
- Ferragamo, M. J. & Oertel, D. (2002), 'Octopus cells of the mammalian ventral cochlear nucleus sense the rate of depolarization', *J. Neurophysiol.* **87**, 2262–2270.
- Ferster, D. & Lindstrom, S. (1983), 'An intracellular analysis of geniculocortical connectivity in area 17 of the cat.', *Journal of Physiology* **342**, 181–215.
- Fiebach, C., Schlesewsky, M., Lohmann, G., von Cramon, D. & Friederici, A. (2005), 'Revisiting the role of broca's area in sentence processing: Syntactic integration versus syntactic working memory', *Hum. Brain Mapp.* **24**(2), 79–91.
URL: <http://dx.doi.org/10.1002/hbm.20070>
- Fitch, R. H., Miller, S. & Tallal, P. (1997), 'Neurobiology of speech perception', *Annu. Rev. Physiol.* **20**, 331–353.
- Frisina, R. D. (2001), 'Subcortical neural coding mechanisms for auditory temporal processing', *Hearing Research* **158**, 1–27.
- Frisina, R. D., Smith, R. L. & Chamberlain, S. C. (1985), 'Differential encoding of rapid changes in sound amplitude by second-order auditory neurons', *Exp Brain Res* **60**, 417–422.
- Frisina, R. D., Smith, R. L. & Chamberlain, S. C. (1990a), 'Encoding of amplitude modulation in the gerbil cochlear nucleus: I. a hierarchy of enhancement', *Hearing Research* **44**, 99–122.
- Frisina, R. D., Smith, R. L. & Chamberlain, S. C. (1990b), 'Encoding of amplitude modulation in the gerbil cochlear nucleus: Ii. possible neural mechanisms', *Hearing Research* **44**, 123–142.
- Frisina, R. D., Walton, J. P., Lynch-Armour, M. A. & Byrd, J. D. (1998), 'Inputs to a physiologically characterized region of the inferior colliculus of the young adult cba mouse', *Hearing Research* **115**(12), 61–81.
URL: <http://www.sciencedirect.com/science/article/pii/S0378595597001767>
- Fritz, J. B., Elhilali, M., David, S. V. & Shamma, S. A. (2007), 'Auditory attention focusing the searchlight on sounds', *Curr. Opin. Neurobiol.* **17**, 437–455.
- Frost, D. O. (1990), 'Sensory processing by novel, experimentally induced crossmodal circuits', *Ann. NY Acad. Sci.* **608**, 92–112.

- Fujimura, O. (1962), *Analysis of nasal consonants*, MIT Press: Cambridge, Mass.
- Fujino, K. & Oertel, D. (2003), 'Bidirectional synaptic plasticity in the cerebellum-like mammalian dorsal cochlear nucleus', *Proc Natl Acad Sci USA* **100**, 265–270.
- Galaburda, A. M. & Pandya, D. N. (1983), 'The intrinsic architectonic and connectional organization of the superior temporal region of the rhesus monkeys', *J Comp Neurol* **221**, 169–184.
- Galbraith, G. C., Gutterson, R. P., Levy, D. S., Mussey, J. L., Sabatasso, F. A. & Wasserman, R. I. (2004), 'Correlated brain stem and cortical evoked responses to auditory tone change', *Neuroreport* **15**, 2613–2616.
- Geisler, C. (1998), *From Sound to Synapse: The Physiology of the Mammalian Ear*, Oxford University Press, New York.
- Gelfand, S. A. (1998), *Hearing: An Introduction to Psychological and Physiological Acoustics.*, New York: Marcel Decker.
- Gerstner, W., Kempter, R., van Hemmen, J. L. & Wagner, H. (1996), 'A neuronal learning rule for sub-millisecond temporal coding', *Nature* **383**, 76–81.
- Gerstner, W. & Kistler, W. M. (2004), *Spiking Neuron Models: Single Neurons, Populations, Plasticity*, Cambridge: Cambridge University Press.
- Gifford, G. W., MacLean, K. A., Hauser, M. D. & Cohen, Y. E. (2005), 'The neurophysiology of functionally meaningful categories: Macaque ventrolateral prefrontal cortex plays a critical role in spontaneous categorization of species-specific vocalizations', *J Cogn Neurosci* **17**(9), 1471–1482.
- Glendenning, K. K., Brunso-Bechtold, J. K., Thompson, G. C. & Masterton, R. B. (1981), 'Ascending auditory afferents to the nuclei of the lateral lemniscus', *J. Comp. Neur.* **197**, 673–703.
- Godfrey, D., Kiang, N. & Norris, B. (1975), 'Single unit activity in the posteroventral cochlear nucleus of the cat', *J. Comp. Neur.* **162**, 247–268.
- Golding, N., Ferragamo, M. & Oertel, D. (1999), 'Role of intrinsic conductances underlying responses to transients in octopus cells of the cochlear nucleus', *J. Neurophysiol.* **19**, 2897–2905.
- Gonzalez-Hernandez, T., Mantoln-Sarmiento, B., Gonzalez-Gonzalez, B. & Prez-Gonzalez, H. (1996), 'Sources of gabaergic input to the inferior colliculus of the rat', *J. Comp. Neur.* **372**, 309–326.

- Hackett, T. A., Stepniewska, I. & Kaas, J. H. (1998), 'Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys', *J Comp Neurol* **394**, 475–495.
- Hackett, T. A., Stepniewska, I. & Kaas, J. H. (1999), 'Prefrontal connections of the parabelt auditory cortex in macaque monkeys', *Brain Research* **817**(12), 45–58.
URL: <http://www.sciencedirect.com/science/article/pii/S0006899398011822>
- Hage, S. R. & Ehret, G. (2003), 'Mapping responses to frequency sweeps and tones in the inferior colliculus of house mice', *European Journal of Neuroscience* **18**(8), 2301–2312.
URL: <http://dx.doi.org/10.1046/j.1460-9568.2003.02945.x>
- Harris, D. M. & Dallos, P. (1979), 'Forward masking of auditory nerve fiber responses', *Journal of Neurophysiology* **42**(4), 1083–1107.
- Harris, K. S. (1958), 'Cues for the discrimination of american english fricatives in spoken syllables', *Language and Speech* **1**, 1–7.
- Hauser, M. D. (1996), *The Evolutin of Communication*, MIT Press: Cambridge, Mass.
- Hauser, M. D. & Marler, P. (1993), 'Food-associated calls in rhesus macaques (*macaca mulatta*): I. socioecological factors', *Behavioral Ecology* **4**(3), 194–205.
URL: <http://beheco.oxfordjournals.org/content/4/3/194.abstract>
- Hazan, V. & Rosen, S. (1991), 'Individual variability in the perception of cues to place contrasts in initial stops', *Percept. Psychophys.* **49**, 187–200.
- Hebb, D. O. (1949), *The Organization of Behavior*, Wiley.
- Heierli, P., de Ribaupierre, F. & de Ribaupierre, Y. (1987), 'Functional properties and interactions of neuron pairs simultaneously recorded in the medial geniculate body of the cat', *Hearing Research* **25**, 209–225.
- Heinz, J. M. & Stevens, H. N. (1961), 'On the properties of voiceless fricative consonants', *J. Acoust. Soc. Am.* **34**, 179–188.
- Henry, K. S., Gall, M. D., Bidelman, G. M. & Lucas, J. R. , . S. (2011), 'Songbirds tradeoff auditory frequency resolution and temporal resolution', *Journal of Comparative Physiology* **197**(4), 351–359.
- Herrmann, M., Hertz, J. & Prugel-Bennett, A. (1995), 'Analysis of synfire chains', *Network: Computation in Neural Systems* **6**, 403–414.

- Higgins, I. & Stringer, S. (2011), ‘The role of independent motion in object segmentation in the ventral visual stream: Learning to recognise the separate parts of the body’, *Vision Research* **51**, 533–562.
- Hockett, C. (1958), *A Course in Modern Linguistics*, Macmillan.
- Hodgkin, A. L. & Huxley, A. F. (1952), ‘A quantitative description of membrane current and its application to conduction and excitation in nerve.’, *J. Physiol.* **117**(4), 500–544.
- Hopfield, J. J. (1995), ‘Pattern recognition computation using action potential timing for stimulus presentation’, *Nature* .
- Horwitz, B. & Tagamets, M. (1999), ‘Predicting human functional maps with neural net modeling’, *Hum. Brain Mapp.* **8**, 137–142.
- Houtgast, T. (1971), ‘Psychophysical evidence for lateral inhibition in hearing’, *J Acoust Soc Am* **51**, 1885–1894.
- Houtgast, T. (1974), ‘Auditory analysis of vowel-like sound’, *Acustica* **31**, 320–324.
- Hromadka, T., DeWeese, M. R. & Zador, A. M. (2008), ‘Sparse representation of sounds in the unanesthetized auditory cortex’, *PLoS Biol* **6**(1), 124–137.
- Huckvale, M. (2004), ‘How to: Phonetic analysis using formant measurements’.
URL: <http://www.phon.ucl.ac.uk/resource/sfs/howto/formant.htm>
- Hudspeth, A. J. (1989), ‘How the ear’s works work’, *Nature* **341**, 397–404.
- Hudspeth, A. J. (1999), ‘How hearing happens’, *Neuron* **19**, 947–950.
- Huetz, C., Gourevitch, B. & Edeline, J. M. (2011), ‘Neural codes in the thalamocortical auditory system: From artificial stimuli to communication sounds’, *Hearing Research* **271**, 147–158.
- Huffman, R. F. & Covey, E. (1995), ‘Origin of ascending projections to the nuclei of the lateral lemniscus in the big brown bat, *ptesicus fuscus*’, *J. Comp. Neur.* **357**, 532–545.
- Humble, J., Denham, S. & Wennekers, T. (2012), ‘Spatio-temporal pattern recognizers using spiking neurons and spike-timing-dependent plasticity’, *Front Comput Neurosci* .

- Husain, F. T., Fromm, S. J., Pursley, R. H., Hosey, L. A., Braun, A. R. & Horwitz, B. (2006), 'Neural bases of categorization of simple speech and nonspeech sounds', *Hum. Brain Mapp.* **27**(8), 636–651.
URL: <http://dx.doi.org/10.1002/hbm.20207>
- Husain, F., Tagamets, M.-A., Fromm, S., Braun, A. & Horwitz, B. (2004), 'Relating neuronal dynamics for auditory object processing to neuroimaging activity: a computational modeling and an fmri study', *NeuroImage* **21**(4), 1701–1720.
URL: <http://www.sciencedirect.com/science/article/pii/S105381190300733X>
- Irvine, D. (1986), *Progress in Sensory Physiology*, Vol. 7, Springer, Berlin, chapter The auditory brainstem, pp. 1–279.
- Ito, T. & Fukushima, K. (1988), 'A neural network model extracting features from speech signals', *Systems and Communications in Japan* **19**(3), 32–45.
- Izhikevich, E. M. (2000), 'Neural excitability, spiking and bursting', *International Journal of Bifurcation and Chaos* **10**(6), 1171–1266.
- Izhikevich, E. M. (2003), Simple model of spiking neurons, in 'IEEE Transactions on Neural Networks', Vol. 14, pp. 1569–1572.
- Izhikevich, E. M. (2004), 'Which model to use for cortical spiking neurons?', *IEEE Transactions on Neural Networks* **15**(5), 1063–1070.
- Izhikevich, E. M. (2006), 'Polychronization: Computation with spikes', *Neural Computation* **18**, 245–282.
- Izhikevich, E. M. (2007), *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*, MIT Press.
- Jakobson, R., Fant, C. G. M. & Halle, M. (1963), *Preliminaries to Speech Analysis: the Distinctive Features and their Correlates*, MIT Press: Cambridge, Mass.
- Jarrell, T. W., Gentile, C. G., McCabe, P. M. & Schneiderman, N. (1986), 'The role of the medial geniculate region in differential pavlovian conditioning of bradycardia in rabbits', *Brain Research* **374**, 126–136.
- Jarrell, T. W., Romanski, L. M., Gentile, C. G., McCabe, P. M. & Schneiderman, N. (1986), 'Ibotenic acid lesion of the medial geniculate region prevent the acquisition of differential pavlovian conditioning to acoustic stimuli in rabbits', *Brain Research* **412**, 285–294.

- Jennings, S. G., Heinz, M. G. & Strickland, E. A. (2011), 'Evaluating adaptation and olivocochlear efferent feedback as potential explanations of psychophysical overshoot', *Journal of the Association for Research in Otolaryngology* **12**, 345–360.
- Jones, D. (1956), *Outline of English Phonetics*, Heifer, Cambridge.
- Jones, E. & Burton, H. (1976), 'Areal differences in the laminar distribution of thalamic afferents in cortical fields of the insular parietal and temporal regions of primates', *J. Comp. Neur.* **168**, 197–248.
- Joris, P. & Smith, P. (2008), 'The volley theory and the spherical cell puzzle', *Neuroscience* **154**, 65–76.
- Joris, P. X., Carney, L. H., Smith, P. H. & Yin, T. C. T. (1994), 'Enhancement of neural synchronization in the anteroventral cochlear nucleus. i. responses to tones at the characteristic frequency', *Journal of Neurophysiology* **71**(3), 1022–1036.
- Joris, P. X., Schreiner, C. E. & Rees, A. (2004), 'Neural processing of amplitude-modulated sounds', *Physiol Rev* **84**, 541577.
- Joris, P. X., Smith, P. H. & Yin, T. C. T. (1994), 'Enhancement of neural synchronization in the anteroventral cochlear nucleus. ii. responses in the tuning curve tail', *Journal of Neurophysiology* **71**(3), 1037–1051.
- Joris, P. X. & Yin, T. C. T. (1992), 'Responses to amplitude-modulated tones in the auditory nerve of the cats', *J Acoust Soc Am* **91**, 215–232.
- Kaas, J. & Hackett, T. (2000), 'Subdivisions of auditory cortex and processing streams in primates', *Proc. Natl. Acad. Sci. U. S. A.* **97**, 11793–11799.
- Kane, E. S. (1976), 'Descending projections of specific regions of cat cochlear nucleus: A light microscopic study.', *Exp. Neurol.* **52**, 372–388.
- Karmarkar, U. R., Najarian, M. T. & Buonomano, D. V. (2002), 'Mechanisms and significance of spike-timing dependent plasticity', *Biological Cybernetics* **87**, 373–382.
- Kates, J. M. (1995), 'Two-tone suppression in a cochlear model', *IEEE Transactions on Speech and Audio Processing* **3**(5), 396–406.
- Keller, C. H. & Takahashi, T. T. (2000), 'Representation of temporal features of complex sounds by the discharge patterns of neurons in the owl's inferior colliculus', *J Neurophysiol* **84**, 2638–2650.

- Kempter, R., Gerstner, W. & van Hemmen, J. L. (1999), 'Hebbian learning and spiking neurons', *Phys Rev E* **59**, 44984514.
- Kiang, N. & Peake, W. T. (1960), 'Components of electric responses recorded from the cochlea', *Ann Otol Rhinol Laryngol* **69**, 448–458.
- Kiang, N. Y. S. (1965), *Discharge patterns of single fibers in the cat's auditory nerve*, MIT Press: Cambridge, Mass.
- Kilgard, M. P. & Merzenich, M. M. (1998), 'Cortical map reorganization enabled by nucleus basalis activity', *Science* **279**, 1714–1718.
- Kim, D., Rhode, W. & Greenberg, S. (1986), *Auditory Frequency Selectivity: A NATO Advanced Research Workshop*, Plenum Press, New York, chapter Responses of cochlear nucleus neurones to speech signals: neural encoding of pitch, intensity and other parameters, pp. 281–288.
- Kimura, A., Donishi, T., Okamoto, K. & Tamai, Y. (2005), 'Topography of projections from the primary and non-primary auditory cortical areas to the medial geniculate body and thalamic reticular nucleus in the rat', *Neuroscience* **135**, 13251342.
- Klatt, D. H. (1980), *Perception and production of fluent speech*, Hillsdale, NJ: Lawrence Erlbaum Associates., chapter Speech perception: a model of acoustic-phonetic analysis and lexical access.
- Klatt, D. H. (1982), Prediction of perceived phonetic distance from critical-band spectra: a first step, in 'Proc. IEEE Int. Conf. Speech, Acoustic Signal Process', Vol. 82, p. 12781281.
- Kluender, K. R., Diehi, R. L. & Killeen, P. R. (1987), 'Japanes quail can learn phonetic categories', *Science* **237**, 1195–1197.
- Koch, C., Poggio, T. & Torres, V. (1982), 'Retinal ganglion cells: A functional interpretation of dendritic morphology', *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **298**(1090), 227–263.
URL: <http://www.jstor.org/stable/2395919>
- Kral, A. & Majernik, V. (1996), 'Neural networks simulating the frequency discrimination of hearing for non-stationary short tone stimuli', *Biol. Cybern.* **74**, 359–366.
- Kudo, M. (1981), 'Projections of the nuclei of the lateral lemniscus in the cat: An autoradiographic study', *Brain Research* **221**, 57–69.

- Kuhl, P. (1981), 'Discrimination of speech by non human animals: Basic auditory sensitivities conducive to the perception of speech-sound categories', *J Acoust Soc Am* **70**, 340–349.
- Kuhl, P. K. (1986), 'Theoretical contributions of tests on animals to the special-mechanisms debates in speech', *Exp Biol* **45**, 233–265.
- Kuhl, P. K. (1987), *Categorical perception*, Cambridge University Press, Cambridge, chapter The special-mechanism debate in speech research: facts and hypotheses from animal studies, pp. 355–386.
- Kuhl, P. K. & Miller, J. D. (1978), 'Speech perception by the chinchilla: Identification functions for synthetic vot stimuli', *J. Acoust. Soc. Am.* **63**, 905–917.
- Kuhl, P. K. & Padden, D. M. (1982), 'Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques', *Percept. Psychophys.* **32**, 542–550.
- Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S. & Iverson, P. (2006), 'Infants show a facilitation effect for native language phonetic perception between 6 and 12 months', *Dev Sci* **9**, 13–21.
- Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N. & Lindblom, B. (1992), 'Linguistic experience alters phonetic perception in infants by 6 months of age', *Science* **255**, 606–608.
- Kuhl, P. & Miller, J. (1975), 'Speech perception by the chinchilla: voicedvoiceless distinction in alveolar plosive consonants', *Science* **190**, 69–72.
- Kulesza, R. J. (2008), 'Cytoarchitecture of the human superior olivary complex: Nuclei of the trapezoid body and posterior tier', *Hearing Research* **241**, 52–63.
- Kurogi, S. (1987), 'A model of neural network for spatiotemporal pattern recognition', *Biological Cybernetics* **57**(1), 103–114.
URL: <http://dx.doi.org/10.1007/BF00318720>
- Kurogi, S. (1991), 'Speech recognition by an artificial neural network using findings on the afferent auditory system', *Biol. Cybern.* **64**, 243–249.
- Kuwabara, N. & Zook, J. M. (1991), 'Classification of the principal cells of the medial nucleus of the trapezoid body', *J. Comp. Neurol.* **314**(4), 707–720.
URL: <http://dx.doi.org/10.1002/cne.903140406>

- Lai, Y.-C., Winslow, R. L. & Sachs, M. B. (1994), 'The functional role of excitatory and inhibitory interactions in chopper cells of the anteroventral cochlear nucleus', *Neural Computation* **6**, 1127–1140.
- Langner, G. (1992), 'Periodicity coding in the auditory systems', *Hearing Research* **60**, 115–142.
- Ledoux, J. E., Sakaguchi, A., Iwata, J. & Reis, D. (1986), 'Interruption of projections from the medial geniculate body to an archi-neostriatal field disrupts the classical conditioning of emotional responses to acoustic stimuli', *Neuroscience* **17**, 615–627.
- Ledoux, J. E., Sakaguchi, A. & Reis, D. J. (1983), 'Subcortical efferent projections of the medial geniculate mediate emotional responses conditioned to acoustic stimuli', *J Neurosci* **4**, 683–398.
- Lemus, L., Hernandez, A. & Romo, R. (2009), 'Neural codes for perceptual discrimination of acoustic flutter in the primate auditory cortex', *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9471–9476.
- Leonard, R. G. & Doddington, G. (1993), 'Tidigits speech corpus', *Texas Instruments, Inc* .
- Liao, Q., Leibo, J. Z. & Poggio, T. (2013), 'Learning invariant representations and applications to face verification', *NIPS* .
- Liaw, J.-S. & Berger, T. W. (1996), 'Dynamic synapse: A new concept of neural representation and computation', *Hippocampus* **6**(6), 591–600.
URL: [http://dx.doi.org/10.1002/\(SICI\)1098-1063\(1996\)6:6;591::AID-HIPO4;3.0.CO;2-K](http://dx.doi.org/10.1002/(SICI)1098-1063(1996)6:6;591::AID-HIPO4;3.0.CO;2-K)
- Liberman, A., Cooper, F., Shankweiler, D. & Studdert-Kennedy, M. (1967), 'Perception of the speech code', *Psychol. Rev.* **74**, 431–461.
- Liberman, M. C. (1978), 'Auditory-nerve response from cats raised in a low-noise chamber', *J. Acoust. Soc. Am.* **63**(2), 442–455.
- Liberman, M. C. (1980), 'Morphological differences among radial afferent fibers in the cat cochlea: an electron-microscopic study of serial sections.', *Hearing Research* **3**, 45–63.
- Liberman, M. C. (1982), 'Single-neuron labeling in the cat auditory nerve.', *Science* **216**, 12391241.

- Liberman, M. C. (1993), 'Central projections of auditory nerve fibers of differing spontaneous rate. ii. posteroventral and dorsal cochlear nuclei', *J. Comp. Neur.* **327**, 17–36.
- Liebenthal, E., Desai, R., Ellingson, M. M., Ramachandran, B., Desai, A. & Binder, J. R. (2010), 'Specialization along the left superior temporal sulcus for auditory categorization', *Cereb. Cortex* **2**, 2958–2970.
- Lindblom, B. (1986), *Experimental Phonology*, Academic Press, Orlando, FL, chapter Phonetic universals in vowel systems, pp. 13–44.
- Lisman, J. (1997), 'Bursts as a unit of neural information: making unreliable synapses reliable', *Trends Neurosci* **20**, 38–43.
- Lorente de No, R. (1981), *The Primary Acoustic Nuclei*, Raven: New York.
- Lu, T. & Wang, X. (2000), 'Temporal discharge patterns evoked by rapid sequences of wide- and narrowband clicks in the primary auditory cortex of cats', *J Neurophysiol* **84**, 236–246.
- Lyon, R. (1982), A computational model of filtering, detection, and compression in the cochlea, in 'Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82. DOI - 10.1109/ICASSP.1982.1171644', Vol. 7, pp. 1282–1285.
- Maass, W., Natschlager, T. & Markram, H. (2004), 'Fading memory and kernel properties of generic cortical microcircuit models', *Journal of Physiology* **98**, 315–330.
- MacGregor, R. J. (1987), *Neural and Brain Modeling*, Academic, San Diego.
- Maclay, H. & Osgood, C. (1959), 'Hesitation phenomena in spontaneous english speech', *Word* **15**, 19–44.
- Markram, H., Lubke, J., Frotscher, M. & Sakmann, B. (1997), 'Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps', *Science* **275**, 213215.
- Masquelier, T., Guyonneau, R. & Thorpe, S. J. (2008), 'Spike timing dependent plasticity finds the start of repeating patterns in continuous spike trains', *PLoS ONE* **3**(1).
- May, B. J., Aleszczyk, C. M. & Sachs, M. B. (1991), 'Single-unit recording in the ventral cochlear nucleus of behaving cats', *J Neurosci Methods* **40**, 155–169.

- McClelland, J. L. & Elman, J. L. (1986), 'The TRACE model of speech perception', *Cognitive Psychology* **18**, 1–86.
- Meddis, R. & O'Mard, L. P. (2006), 'Virtual pitch in a computational physiological model', *J Acoust Soc Am* **120**(6), 3861–3869.
- Mesgarani, N. & Chang, E. F. (2012), 'Selective cortical representation of attended speaker in multi-speaker speech perception', *Nature* **485**, 233–237.
- Miller, J. D. (1977), *Recognition of complex acoustic signals*, Abakon Verlagsgesellschaft, Berlin, chapter Perception of speech sounds in animals: evidence for speech processing by mammalian auditory mechanisms.
- Miller, R. (1987), 'Representation of brief temporal patterns, hebbian synapses, and the left-hemisphere dominance for phoneme recognition', *Psychobiology* **15**, 241–247.
- Miller, R. (1996), *Axonal conduction times and human cerebral laterality. A psychobiological theory.*, Harwood Academic.
- Moller, A. R. (1977), 'Frequency selectivity of single auditory-nerve fibers in response to broadband noise stimuli', *J. Acoust. Soc. Am.* **62**, 135–142.
- Moller, M. F. (1993), 'A scaled conjugate gradient algorithm for fast supervised learning', *Neural Networks* **6**(4), 525–533.
- Mullennix, J. & Pisoni, D. (1989), *The Comparative Psychology of Audition: Perceiving Complex Sounds*, Lawrence Erlbaum, Hillsdale, NJ, chapter Speech perception: analysis of biologically significant signals, pp. 97–128.
- Muller, R.-A. & Basho, S. (2004), 'Are nonlinguistic functions in "broca's area" prerequisites for language acquisition? fmri findings from an ontogenetic viewpoint', *Brain and Language* **89**(2), 329–336.
URL: <http://www.sciencedirect.com/science/article/pii/S0093934X03003468>
- Naatanen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huottilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmonemi, R. J. & Luuk, A. e. a. (1997), 'Language-specific phoneme representations revealed by electric and magnetic brain responses', *Nature* **385**, 432–434.
- Nager, C., Storck, J. & Deco, G. (2002), 'Speech recognition with spiking neurons and dynamic synapses: a model motivated by the human auditory pathway', *Neurocomputing* **44**(46), 937–942.

- Needham, K. & Paolini, A. (2003), 'Fast inhibition underlies the transmission of auditory information between cochlear nuclei', *J. Neurosci.* **23**, 6357–6361.
- Nelken, I. (2004), 'Processing of complex stimuli and natural scenes in the auditory cortex', *Curr. Opin. Neurobiol.* **14**, 474–480.
- Nelken, I. (2008), 'Processing of complex sounds in the auditory system', *Curr. Opin. Neurobiol.* **18**, 413–417.
- Nelken, I. & Chechik, G. (2007), 'Information theory in auditory research', *Hearing Research* **229**, 94–105.
- Nelken, I., Fishbach, A., Las, L., Ulanovsky, N. & Farkas, D. (2003), 'Primary auditory cortex of cats: feature detection or something else?', *Biol. Cybern.* **89**, 397–406.
- Nelson, P. C. & Carney, L. H. (2004), 'A phenomenological model of peripheral and central neural responses to amplitude-modulated tones', *J Acoust Soc Am* **116**, 2173–2186.
- Nuttall, A. L. & Dolan, D. F. (1993), 'Two-tone suppression of inner hair cell and basilar membrane responses in the guinea pig', *J. Acousr. Soc. Amer.* **93**(1), 390–400.
- Oertel, D. (1991), 'The role of intrinsic neuronal properties in the encoding of auditory information in the cochlear nuclei', *Current Opinion in Neurobiology* **1**, 221–228.
- Oertel, D., Bal, R., Gardner, S. M., Smith, P. H. & Joris, P. X. (2000), 'Detection of synchrony in the activity of auditory nerve fibers by octopus cells of the mammalian cochlear nucleus', *Proceedings of the National Academy of Sciences* **97**(22), 11773–11779.
URL: <http://www.pnas.org/content/97/22/11773.abstract>
- Oertel, D., Wright, S., Cao, X.-J., Ferragamo, M. & Bal, R. (2011), 'The multiple functions of t stellate/multipolar/chopper cells in the ventral cochlear nucleus', *Hearing Research* **276**, 61–69.
- Oertel, D., Wu, S. H., Garb, M. W. & Dizack, C. (1990), 'Morphology and physiology of cells in slice preparations of the posteroventral cochlear nucleus of mice', *J. Comp. Neurol.* **295**(1), 136–154.
URL: <http://dx.doi.org/10.1002/cne.902950112>

- Ohlemiller, K., Kanwal, J., Butman, J. & Suga, N. (1994), 'Stimulus design for auditory neuroethology: synthesis and manipulation of complex communication sounds', *Aud. Neurosci.* **1**, 19–37.
- Ojemann, G. & Mateer, C. (1979), 'Human language cortex: localization of memory, syntax and sequential motor-phoneme identification systems', *Science* **205**, 1401–1403.
- Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I.-H., Saberi, K., Serences, J. T. & Hickok, G. (2010), 'Hierarchical organization of human auditory cortex: Evidence from acoustic invariance in the response to intelligible speech', *Cereb. Cortex* **20**, 2486–2495.
- Olive, J., Greenwood, A. & Coleman, J. (1993), *Acoustics of American English speech*, New York: Springer-Verlag.
- Oliver, D. L., Beckius, G. E., Bishop, D. C. & Kuwada, S. (1997), 'Simultaneous anterograde labeling of axonal layers from lateral superior olive and dorsal cochlear nucleus in the inferior colliculus of cat', *J. Comp. Neur.* **382**, 215–229.
- Olshausen, B. A. (1996), 'Emergence of simple-cell receptive field properties by learning a sparse code for natural images', *Nature* **381**(6583), 607–609.
- Osen, K. (1969), 'Cytoarchitecture of the cochlear nuclei in the cat', *J. Comp. Neur.* **136**, 453–484.
- Ostapoff, E. M., Feng, J. J. & Morest, D. K. (1994), 'A physiological and structural study of neuron types in the cochlear nucleus. ii. neuron types and their structural correlation with response properties', *J Comp Neurol* **346**(1), 19–42.
- Owings, D. H. & Morton, E. S. (1998), *Animal Vocal Communication: A New Approach*, Cambridge, MA: Cambridge Univ. Press.
- Palmer, A. R., Wallace, M. N., Arnott, R. H. & Shackleton, T. M. (2003), 'Morphology of physiologically characterised ventral cochlear nucleus stellate cells', *Exp Brain Res* **153**, 418426.
- Palmer, A. R., Winter, I. M. & Darwin, C. J. (1986), 'The representation of steady-state vowel sounds in the temporal discharge patterns of the guinea pig cochlear nerve and primarylike cochlear nucleus neurons', *J. Acoust. Soc. Amer.* **79**, 100–113.
- Palmer, A. & Winter, I. (1996), 'The temporal window of two-tone facilitation in onset units of the ventral cochlear nucleus', *Audiol. Neurotol.* **1**, 12–30.

- Pantev, C., Oostenveld, R., Engelien, A., Ross, B., Roberts, L. E. & Hoke, M. (1998), 'Increased auditory cortical representation in musicians', *Nature* **392**, 811–814.
- Panzeri, S. & Treves, A. (1996), 'Analytical estimates of limited sampling biases in different information measures', *Network* **7**, 87–101.
- Patterson, D. K. & Pepperberg, I. M. (1994), 'A comparative study of human and parrot phonation: acoustic and articulatory correlates of vowels', *J Acoust Soc Am* **96**, 634–648.
- Paugam-Moisy, H., Martinez, R. & Bengio, S. (2008), 'Delay learning and polychronization for reservoir computing', *Neurocomputing* **71**(7-9), 1143–1158.
- Peake, W. T., Goldstein, M. H. & Kiang, N. Y. S. (1962), 'Response of the auditory nerve to repetitive auditory stimuli', *J Acoust Soc Am* **34**, 562–570.
- Penfield, W. & Jasper, H. (1954), *Epilepsy and the functional anatomy of the human brains*, Boston: Little, Brown.
- Perrinet, L., Delorme, A., Samuelides, M. & Thorpe, S. J. (2001), 'Networks of integrate-and-fire neuron using rank order coding a: how to implement spike time dependent hebbian plasticity', *Neurocomputing* **38-40**, 817–822.
- Peterson, G. E. & Barney, H. L. (1952), 'Control methods used in a study of the vowels', *Journal of the Acoustical Society of America* **24**.
- Phillips, D. P. (1988), 'Effect of tone-pulse rise time on rate-level functions of cat auditory cortex neurons: Excitatory and inhibitory processes shaping responses to tone onset', *J Neurosci* **59**, 1524–1539.
- Phillips, D. P. (1990), 'Neural representation of sound amplitude in the auditory cortex: effects of noise masking', *Behavioural Brain Research* **37**(3), 197–214.
URL: <http://www.sciencedirect.com/science/article/pii/016643289090132X>
- Pickles, J. O. (1988), *An Introduction to the Physiology of Hearing*, Academic Press, London.
- Poldrack, R. A., Wagner, A. D., Prull, M. W., Desmond, J. E., Glover, G. H. & Gabrieli, J. D. (1999), 'Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex', *NeuroImage* **10**(1), 15–35.
URL: <http://www.sciencedirect.com/science/article/pii/S105381199990441X>

- Pollak, G. D. (2011), 'Inhibitory projections from the ventral nucleus of the lateral lemniscus and superior paraolivary nucleus create directional selectivity of frequency modulations in the inferior colliculus: A comparison of bats with other mammals', *Hearing Research* **273**, 134–144.
- Poon, P. W. & Chen, X. (1992), 'Postnatal exposure to tones alters the tuning characteristics of inferior collicular neurons in the rats', *Brain Research* **585**, 391–394.
- Portfors, C. V. & Felix, R. A. (2005), 'Spectral integration in the inferior colliculus of the CBA/CaJ mouse', *Neuroscience* **136**, 1159–1170.
- Price, C. J. (1998), 'The functional anatomy of word comprehension and production', *Trends in Cognitive Sciences* **2**(8), 281–288.
URL: <http://www.sciencedirect.com/science/article/pii/S1364661398012017>
- Ramachandran, R., Davis, K. A. & May, B. J. (1999), 'Single-unit responses in the inferior colliculus of decerebrate cats. i. classification based on frequency response maps', *J Neurophysiol* **82**, 152–163.
- Rauschecker, J. P. (1998), 'Parallel processing in the auditory cortex of primates', *Audiol. Neurotol.* **3**, 86–103.
- Rauschecker, J. P. & Scott, S. K. (2009), 'Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing', *Nat Neurosci* **12**, 718–724.
- Rauschecker, J. P. & Tian, B. (2000), 'Mechanisms and streams for processing of "what" and "where" in auditory cortex', *Proc. Natl. Acad. Sci. U. S. A.* **97**(22), 11800–11806.
- Rauschecker, J. P., Tian, B. & Hauser, M. (1995), 'Processing of complex sounds in the macaque nonprimary auditory cortex', *Science* **268**, 111–114.
- Razak, K. A. & Fuzessery, Z. M. (2007), 'Development of inhibitory mechanisms underlying selectivity for the rate and direction of frequency-modulated sweeps in the auditory cortex', *J Neurosci* **27**, 1769–1781.
- Reber, P. J., Gitelman, D. R., Parrish, T. B. & Mesulam, M. M. (2003), 'Dissociating explicit and implicit category knowledge with fmri. journal of cognitive neuroscience', *J Cogn Neurosci* **15**, 574–583.
- Recio, A. & Rhode, W. S. (2000), 'Representation of vowel stimuli in the ventral cochlear nucleus of the chinchilla', *Hearing Research* **146**, 167–184.

- Rees, A. & Moller, A. R. (1983), 'Responses of neurons in the inferior colliculus of the rat to am and fm tones', *Hearing Research* **10**, 301–330.
- Reutimann, J., Dr, P. & Mey, H. (1999), 'A model of synaptic short-term plasticity and its adaptation'.
- Rhode, W. & Greenberg, S. (1992), *The Mammalian Auditory Pathway: Neurophysiology*, Springer, New York, chapter Physiology of the cochlear nuclei, pp. 94–152.
- Rhode, W. S. (1978), 'Some observations on cochlear mechanics', *J. Acoust. Soc. Amer.* **64**(1), 158–176.
- Rhode, W. S., Oertel, D. & Smith, P. H. (1983), 'Physiological response properties of cells labeled intracellularly with horseradish peroxidase in the cat ventral cochlear nucleus', *J. Comp. Neur.* **213**, 448–463.
- Rhode, W. S., Roth, G. L. & Recio-Spinoso, A. (2010), 'Response properties of cochlear nucleus neurons in monkeys', *Hearing Research* **259**, 1–15.
- Rhode, W. & Smith, P. (1986), 'Encoding timing and intensity in the ventral cochlear nucleus of the cat', *J. Neurophysiol.* **56**, 261–286.
- Robles, L., Ruggero, M. A. & Rich, N. C. (1991), 'Two-tone distortion in the basilar membrane of the cochlea', *Nature* **349**, 413–414.
- Rolls, E. T. & Treves, A. (1998), *Neural Networks and Brain Function*, Oxford: Oxford University Press.
- Romanski, L., Bates, J. & Goldman-Rakic, P. (1999), 'Auditory belt and parabelt projections to the prefrontal cortex in the rhesus monkey', *J. Comp. Neurol.* **403**(2), 141–157.
URL: [http://dx.doi.org/10.1002/\(SICI\)1096-9861\(19990111\)403:2<141::AID-CNE1>3.0.CO;2-V](http://dx.doi.org/10.1002/(SICI)1096-9861(19990111)403:2<141::AID-CNE1>3.0.CO;2-V)
- Romanski, L. M., Averbeck, B. B. & Diltz, M. (2004), 'Neural representation of vocalizations in the primate ventrolateral prefrontal cortex', *J Neurophysiol* **93**, 734–747.
- Romanski, L. M. & Goldman-Raciz, P. S. (2002), 'An auditory domain in primate prefrontal cortex', *Nat. Neurosci.* **5**, 15–16.
- Romanski, L. M., Tian, B., Fritz, J. B., Mishkin, M., Goldman-Rakic, P. S. & Rauschecker, J. P. (1999), 'Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex', *Nat. Neurosci.* **2**(12), 1131–1136.

- Rose, J. E., Brugge, J. F., Anderson, D. J. & Hind, J. E. (1967), 'Phase-locked responses to low frequency tones in single auditory nerve fibers of the squirrel monkey', *J Neurophysiol* **30**, 769–793.
- Rosen, S. (1992), 'Temporal information in speech: Acoustic, auditory and linguistic aspects', *Phil. Trans. R. Soc. Lond.* **336**, 367–373.
- Rouiller, E., De Ribaupierre, Y., Toros-Morel, A. & De Ribaupierre, F. (1981), 'Neural coding of repetitive clicks in the medial geniculate body of cat', *Hearing Research* **5**, 81–100.
- Rouiller, E. M. & de Ribaupierre, F. (1985), 'Origin of afferents to physiologically defined regions of the medial geniculate body of the cat: ventral and dorsal divisions', *Hearing Research* **19**, 97–114.
- Rouiller, E. M. & Ryugo, D. K. (1984), 'Intracellular marking of physiologically characterised cells in the ventral cochlear nucleus of the cat', *J. Comp. Neur.* **225**, 167–186.
- Rubin, J., Lee, D. D. & Sompolinsky, H. (2001), 'Equilibrium properties of temporally asymmetric hebbian plasticity', *Phys Rev Lett* **86**, 364367.
- Ruggero, M. A. (1973), 'Response to noise of auditory nerve fibers in the squirrel monkey', *J Neurophysiol* **36**, 569–587.
- Russ, B. E., Ackelson, A. L., Baker, A. E. & Cohen, Y. E. (2008), 'Coding of auditory-stimulus identity in the auditory non-spatial processing stream', *J Neurophysiol* **99**, 87–95.
- Russ, M., Lee, Y. S. & Cohen, Y. E. (2007), 'Neural and behavioral correlates of auditory categorization', *Hearing Research* **229**, 204–212.
- Rutkowski, R., Shackleton, T. M., Schnupp, J. W., Wallace, M. N. & Palmer, A. R. (2002), 'Spectrotemporal receptive field properties of single units in the primary, dorsocaudal and ventrostral auditory cortex of the guinea pigs', *Audiol. Neurotol.* **7**, 214–227.
- Sachs, M. B. (1969), 'Stimulus-response relation for auditory-noise fibers: Two-tone stimuli', *The Journal of the Acoustical Society of America* **45**(4), 1025–1036.
- Sachs, M. B. (1984), *Hearing Science*, College Hill: San-Diego, California, chapter Speech encoding in the auditory nerve, pp. 263–307.

- Sachs, M. B. & Abbas, P. J. (1974), 'Rate versus level functions for auditory-nerve fibers in cats: tone-burst stimuli', *J. Acoust. Soc. Am.* **56**(6), 1835–1847.
- Sachs, M. B. & Abbas, P. J. (1976), 'Phenomenological model for two-tone suppression', *J. Acoust. Soc. Am.* **60**(5), 1157–1163.
- Sachs, M. B. & Kiang, N. Y. S. (1968), 'Two-tone inhibition in auditory-nerve fibers', *J. Acoust. Soc. Am.* **43**(5), 1120–1128.
- Sachs, M. B. & Young, E. D. (1980), 'Effects of nonlinearities on speech encoding in the auditory nerve', *J. Acoust. Soc. Amer.* **68**, 858–875.
- Sadagopan, S. & Wang, X. (2009), 'Nonlinear spectrotemporal interactions underlying selectivity for complex sounds in auditory cortex', *J Neurosci* **29**(36), 11192–11202.
- Salami, M., Itami, C., Tsumoto, T. & Kimura, F. (2003), 'Change of conduction velocity by regional myelination yields constant latency irrespective of distance between thalamus and cortex', *PNAS* **100**, 6174–6179.
- Salem, Z. N. B., Boougrain, L. & Alexnandre, F. (2007), 'Spatio-temporal biologically inspired models for clean and noisy speech recognition', *Neurocomputing* **71**, 131–136.
- Sanes, D. H. & Constantine-Paton, M. (1983), 'Altered activity patterns during development reduce neural tuning', *Science* **16**, 1183–1185.
- Sanes, D. H. & Constantine-Paton, M. (1985), 'The sharpening of frequency tuning curves requires patterned activity during development in the mouse, *mus musculus*', *J Neurosci* **5**, 1152–1166.
- Schmidt, K. E., Galuske, R. A. & Singer, W. (1999), 'Matching the modules: cortical maps and long-range intrinsic connections in visual cortex during development', *J Neurophysiol* **41**, 10–17.
- Schnupp, J. W. H., Hall, T. M., Kokelaar, R. F. & Ahmed, B. (2006), 'Plasticity of temporal pattern codes for vocalization stimuli in primary auditory cortex', *J Neurosci* **26**(18), 4785–4795.
- Schnupp, J. W., Nelken, I. & King, A. (2011), *Auditory Neuroscience. Making Sense of Sound*, MIT Press.

- Schofield, B. & Cant, N. (1997), 'Ventral nucleus of the lateral lemniscus in guinea pigs: cytoarchitecture and inputs from the cochlear nucleus', *J. Comp. Neur.* **379**, 363–385.
- Schreiner, C. E. & Winer, J. A. (2007), 'Auditory cortex mapmaking: principles, projections, and plasticity.', *Neuron* **56**, 356–365.
- Schuknecht, F. (1974), *Pathology of the Ear*, Harvard U. P., Cambridge, MA.
- Secker-Walker, H. E. & Searle, C. L. (1990), 'Time-domain analysis of auditory-nerve-fiber firing rates', *J. Acoust. Soc. Am.* **88**(3), 1427–1436.
- Seifritz, E., Esposito, F., Hennel, F., Mustovic, H., Neuhoff, J. G., Bilecen, D., Tedeschi, G., Schefer, K. & Di Salle, F. (2002), 'Spatiotemporal pattern of neural processing in the human auditory cortex', *Science* **297**, 1706–1708.
- Selezneva, E., Scheich, H. & Brosch, M. (2006), 'Dual time scales for categorical decision making in auditory cortex', *Current Biology* **16**(24), 2428–2433.
URL: <http://www.sciencedirect.com/science/article/pii/S0960982206023645>
- Sharma, J., Angelucci, A. & Sur, M. (2000), 'Induction of visual orientation modules in auditory cortex', *Nature* **404**, 841–847.
- Shimazaki, H. & Shinomoto, S. (2007), 'A method for selecting the bin size of a time histogram', *Neural Computation* **19**, 1503–1527.
- Silkes, S. M. & Geisler, C. D. (1991), 'Responses of "lower-spontaneous-rate" auditory-nerve fibers to speech syllables presented in noise. i: General characteristics', *J. Acoust. Soc. Am.* **90**(6), 3122–3139.
- Simoncelli, E. P., Paninski, L., Pillow, J. & Swartz, O. (2004), *The Cognitive Neurosciences*, MIT Press, chapter Characterization of neural responses with stochastic stimuli, pp. 327–338.
- Sinex, D. G. & McDonald, L. P. (1988), 'Average discharge rate representation of voice onset time in the chinchilla auditory nerve', *J. Acoust. Soc. Am.* **83**(5), 1817–1827.
- Sinex, D. & Geisler, C. (1983), 'Responses of auditory-nerve fibers to consonant-vowel syllables', *J. Acoust. Soc. Am.* **73**(2), 602–615.
- Sinnott, J. M. & Adams, F. S. (1987), 'Differences in human and monkey sensitivity to acoustic cues underlying voicing contrasts', *J Acoust Soc Am* **82**, 1539–1547.

- Smith, P. H., Joris, P. X., Carney, L. H. & Yin, T. C. T. (1991), 'Projections of physiologically characterized globular bushy cell axons from the cochlear nucleus of the cat', *The Journal of Comparative Neurology* **304**, 387–407.
- Smith, P. H. & Rhode, W. S. (1987), 'Characterization of hrp-labeled globular bushy cells in the cat anteroventral cochlear nucleus', *The Journal of Comparative Neurology* **266**, 360–376.
- Smith, P. H. & Rhode, W. S. (1989), 'Structural and functional properties distinguish two types of multipolar cells in the ventral cochlear nucleus', *J. Comp. Neur.* **282**, 595–616.
- Smith, P., Joris, P., Banks, M. & Yin, T. (1993), *The Mammalian Cochlear Nuclei. Organization and Function*, Plenum Press, New York and London, chapter Responses of cochlear nucleus cells and projections of their axons, p. 349360.
- Smith, P., Massie, A. & Joris, P. (2005), 'Acoustic stria: anatomy of physiologically characterized cells and their axonal projection patterns', *J. Comp. Neur.* **482**, 349–371.
- Smith, R. L. (1977), 'Short-term adaptation in single auditory nerve fibers: Some poststimulatory effects', *Journal of Neurophysiology* **40**(5), 1098–1112.
- Softky, W. & Koch, C. (1992), 'Cortical cells do not perform temporal integration of small epsp's', *Soc. Neurosci. Abstr.* **18**, 740.
- Song, J. H., Skoe, E., Wong, P. C. & Kraus, N. (2008), 'Plasticity in the adult human auditory brainstem following short-term linguistic training', *J Cogn Neurosci* **20**(10), 1892–1902.
- Song, P., Yang, Y., Barnes-Davies, M., Bhattacharjee, A., Hamann, M., Forsythe, I. D., Oliver, D. L. & Kaczmarek, L. K. (2005), 'Acoustic environment determines phosphorylation state of the kv3.1 potassium channel in auditory neurons.', *Nat Neurosci* **8**, 1335–1342.
- Song, S., Miller, K. D. & Abbott, L. F. (2000), 'Competitive hebbian learning through spike-timing-dependent synaptic plasticity', *Nat Neurosci* **3**, 919926.
- Stecker, G., Mickey, B. J., Macpherson, E. A. & Middlebrooks, J. C. (2003), 'Spatial sensitivity in field paf of cat auditory cortex', *J Neurophysiol* **89**, 2889–2903.

- Steinert, J. R., Kopp-Scheinflug, C., Baker, C., Challiss, R. A., Mistry, R., Haustein, M. D., Griffin, S. J., Tong, H., Graham, B. P. & Forsythe, I. D. (Neuron), 'Nitric oxide is a volume transmitter regulating postsynaptic excitability at a glutamatergic synapse', *2008* **60**(4), 642–656.
- Steinschneider, M., Fishman, Y. I. & Arezzo, J. C. (2003), 'Representation of the voice onset time (vot) speech parameter in population responses within primary auditory cortex of the awake monkey', *J Acoust Soc Am* **114**(1), 307–321.
- Steinschneider, M., Schroeder, C., Arezzo, J. & Vaughan, H. (1990), 'Tonotopic features of speech evoked activity in primate auditory cortex', *Brain Research* **519**, 158–168.
- Steinschneider, M., Schroeder, C., Arezzo, J. & Vaughan, H. (1994), 'Speech-evoked activity in primary auditory cortex: effects of voice-onset time', *Electroencephalogr. Clin. Neurophysiol.* **92**, 30–43.
- Steinschneider, M., Schroeder, C. E., Arezzo, J. C. & Vaughan, H. G. (1995), 'Physiologic correlates of the voice onset time boundary in primary auditory cortex (a1) of the awake monkey: temporal response patterns', *Brain and Language* **48**, 326–340.
- Steinschneider, M., Volkov, I. O., Fishman, Y. I., Oya, H., Arezzo, J. C. & Howard, M. A. (2005), 'Intracortical responses in human and monkey primary auditory cortex support a temporal processing mechanism for encoding of the voice onset time phonetic parameter', *Cereb. Cortex* **15**, 170–186.
- Steinschneider, M., Volkov, I. O., Noh, D. M., Garell, C. P. & Howard, M. A. (1999), 'Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex', *J Neurophysiol* **82**, 2346–2357.
- Stelmachowicz, P. G., Small, A. M. & Abbas, P. J. (1982), 'Suppression effects for complex stimuli', *J. Acoust. Soc. Am.* **71**(2), 410–420.
- Stephens, S. D. G., Charlet de Sauvage, R. & Aran, J. M. (1975), 'Gross responses from the cochlear nerve from the man and in the guinea pig', *Symp Zool Soc London* **37**, 167–186.
- Stevens, K. N. (1980), 'Acoustic correlates of some phonetic categories', *J. Acoust. Soc. Am.* **68**, 836–842.

- Stevens, K. N. (1981), *The cognitive representation of speech*, Amsterdam: North Holland, chapter Constraints imposed by the auditory system on the properties used to clarify speech sounds: data from phonology, acoustics and psycho-acoustics, pp. 61–74.
- Stevens, K. N. & Blumstein, S. E. (1982), *Perspectives on the study of speech*, Hillsdale, New Jersey: Lawrence Erlbaum, chapter The search for invariant acoustic correlates of phonemic features, pp. 1–38.
- Stevens, K. N. & House, A. S. (1961), ‘An acoustical theory of vowel production and some of its implications’, *J. Speech Hear. Res.* **4**, 303–320.
- Stotler, W. A. (1953), ‘An experimental study of the cells and connections of the superior olivary complex of the cat’, *J. Comp. Neur.* **98**, 401–432.
- Strange, B., Henson, R., Friston, K. & Dolan, R. (2000), ‘Brain mechanisms for detecting perceptual, semantic, and emotional deviance’, *NeuroImage* **12**(4), 425–433.
URL: <http://www.sciencedirect.com/science/article/pii/S1053811900906372>
- Stringer, S., Perry, G., Rolls, E. & Proske, J. (2006), ‘Learning invariant object recognition in the visual system with continuous transformations’, *Biological Cybernetics* **94**, 128–142.
- Suga, N. (1978), ‘Specialization of the auditory system for reception and processing of species-specific sounds’, *Fed. Proc.* **37**, 2342–2354.
- Suga, N. (1988), *Auditory Function. Neurobiological Bases of Hearing*, J. Wiley and Sons, New York, chapter Auditory neuroethology and speech processing: complex-sound processing by combination-sensitive neurons, pp. 679–720.
- Suga, N. (1989), ‘Principles of auditory information-processing derived from neuroethology’, *J Exp Biol* **146**, 277–286.
- Suga, N. (1992), ‘Philosophy and stimulus design for neuroethology of complex-sound processing’, *Phil. Trans. R. Soc. Lond.* **336**, 423–428.
- Sugimoto, S., Sakurada, M., Horikawa, J. & I., T. (1997), ‘The columnar and layer-specific response properties of neurons in the primary auditory cortex of mongolian gerbils’, *Hearing Research* **112**, 175–185.
- Sumner, C., O’Mard, L. P., Lopez-Poveda, E. A. & Meddis, R. (2003), ‘A non-linear filter-bank model of the guinea pig cochlear nerve: Rate responses’, *J Acoust Soc Am* **113**, 3264–3274.

- Swadlow, H. A. (1985), 'Physiological properties of individual cerebral axons studied in vivo for as long as one years', *J Neurophysiol* **54**, 1346–1362.
- Swadlow, H. A. (1988), 'Efferent neurons and suspected interneurons in binocular visual cortex of the awake rabbit: Receptive fields and binocular properties', *J Neurophysiol* **88**, 1162–1187.
- Swadlow, H. A. (1992), 'Monitoring the excitability of neocortical efferent neurons to direct activation by extracellular current pulses', *J Neurophysiol* **68**, 605–619.
- Swadlow, H. A. (1994), 'Efferent neurons and suspected interneurons in motor cortex of the awake rabbit: Axonal properties, sensory receptive fields, and subthreshold synaptic inputs', *J Neurophysiol* **71**, 437–453.
- Temchin, A. N., Rich, N. C. & Ruggero, M. A. (2008), 'Threshold tuning curves of chinchilla auditory-nerve fibers. i. dependence on characteristic frequency and relation to the magnitudes of cochlear vibrations', *J Neurophysiol* **100**(5), 2889–2898.
- Tian, B., Reser, D., Durham, A., Kustov, A. & Rauschecker, J. P. (2001), 'Functional specialization in rhesus monkey auditory cortex', *Science* **292**, 290–293.
- Tomasello, M. (2005), *Constructing a language: A usage-based theory of language acquisition.*, Boston, MA: Harvard University Press.
- Treves, A. & Panzeri, S. (1995), 'The upward bias in measures of information derived from limited data samples.', *Neural Computation* **7**, 399–407.
- Tromans, J. M., Higgins, I. V. & Stringer, S. M. (2012), 'Learning view invariant recognition with partially occluded objects', *Frontiers in Computational Neuroscience* **6**(48).
- Trussell, L. (1997), 'Cellular mechanisms for preservation of timing in central auditory pathways', *Curr. Opin. Neurobiol.* **7**, 487–492.
- Trussell, L. O. (1999), 'Synaptic mechanisms for coding timing in auditory neurons', *Annu. Rev. Physiol.* **61**, 477–496.
- Tzounopoulos, T., Kim, Y., Oertel, D. & Trussell, L. O. (2004), 'Cell-specific, spike timing-dependent plasticities in the dorsal cochlear nucleus', *Nat Neurosci* **7**(7), 719–725.

- Tzounopoulos, T. & Kraus, N. (2009), 'Learning to encode timing: Mechanisms of plasticity in the auditory brainstem', *Neuron* **62**(4), 463–469.
- Tzounopoulos, T., Rubio, M. E., Keen, J. E. & Trussell, L. O. (2007), 'Coactivation of pre- and postsynaptic signaling mechanisms determines cell-specific spike-timing-dependent plasticity', *Neuron* **54**, 291–301.
- Ulanovsky, N., Las, L., Farkas, D. & Nelken, I. (2004), 'Multiple time scales of adaptation in auditory cortex neurons', *J Neurosci* **24**, 10440–10453.
- Uppenkamp, S., Johnsrude, I. S., Norris, D., Marslen-Wilson, W. & Patterson, R. D. (2006), 'Locating the initial stages of speech-sound processing in human temporal cortex', *Neuroimage* **31**, 1284–1296.
- Van Noort, J. (1969), *The Structure and Connections of the Inferior Colliculus. An Investigation of the Lower Auditory System*, Van Gorcum, Assert.
- van Rossum, M. C. W., Bi, G. Q. & Turrigiano, G. G. (2000), 'Stable hebbian learning from spike timing-dependent plasticity', *Journal of Neuroscience* **20**(23), 8812–8821.
- Vater, M. & Feng, A. S. (1990), 'Functional organization of ascending and descending connections of the cochlear nucleus of horseshoe bats', *J. Comp. Neurol.* **292**(3), 373–395.
URL: <http://dx.doi.org/10.1002/cne.902920305>
- Vater, M., Kssl, M. & Horn, A. K. E. (1992), 'Gad- and gaba-immunoreactivity in the ascending auditory pathway of horseshoe and mustached bats', *J. Comp. Neurol.* **325**(2), 183–206.
URL: <http://dx.doi.org/10.1002/cne.903250205>
- Viemeister, N. F. (1984), 'Intensity coding and the dynamic range problem', *Hearing Research* **34**, 267–274.
- von Békésy, G. (1960), *Experiments in hearing*, McGraw-Hill, New York.
- von Melchner, L., Pallas, S. L. & Sur, M. (2000), 'Visual behaviour mediated by retinal projections directed to the auditory pathways', *Nature* **404**, 871–876.
- vonGersdorff, H. & Borst, J. G. (2002), 'Short-term plasticity at the calyx of held', *Nat. Rev. Neurosci* **3**, 53–64.
- Wang, X. (2000), 'On cortical coding of vocal communication sounds in primates', *PNAS* **97**(22), 11843–11849.

- Wang, X. & Kadia, S. C. (2001), 'Differential representation of species-specific primate vocalizations in the auditory cortices of marmoset and cat', *Journal of Neurophysiology* **86**, 2616–2620.
- Wang, X., Lu, T., Snider, R. K. & Liang, L. (2005), 'Sustained firing in auditory cortex evoked by preferred stimuli', *Nature* **435**, 341–346.
- Wang, X., Merzenich, M. M., Beitel, R. & Schreiner, C. E. (1995), 'Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: temporal and spectral characteristics', *J Neurophysiol* **74**, 2685–2706.
- Wang, X. & Sachs, M. B. (1994), 'Neural encoding of single-formant stimuli in the cat. ii. responses of anteroventral cochlear nucleus units', *Journal of Neurophysiology* **71**(1), 59–78.
- Warr, W. (1982), *Contributions to Sensory Physiology*, Vol. 7, Academic Press, New York, chapter Parallel ascending pathways from the cochlear nucleus: neuroanatomical evidence of functional specialization, pp. 1–38.
- Warr, W. B. (1966), 'Fiber degeneration following lesions in the anterior ventral cochlear nucleus of the cat', *Exp Neurol* **14**, 453–474.
- Warr, W. B. (1972), 'Fiber degeneration following lesions in the multipolar and globular cell areas in the ventral cochlear nucleus of the cat', *Brain Research* **40**, 247–270.
- Warr, W. B. & Beck, J. E. (1996), 'Multiple projections from the ventral nucleus of the trapezoid body in the rat', *Hearing Research* **93**, 83–101.
- Weber, D. L. (1984), 'Combined effect of two suppressors', *J. Acoust. Soc. Am.* **75**(5), 1563–1569.
- Webster, W. & Aitkin, L. (1975), *Handbook of Psychobiology*, Academic Press, New York, chapter Central auditory processing, pp. 325–364.
- Weinberger, N. & Diamond, D. (1987), 'Physiological plasticity in auditory cortex: rapid induction by learning', *Prog. Neurobiol.* **29**, 1–55.
- Weinberger, N. M. (1995), 'Dynamic regulation of receptive fields and maps in the adult sensory cortex', *Annu. Rev. Neurosci* **18**, 129.
- Weinberger, N. M. (2007), 'Associative representational plasticity in the auditory cortex: a synthesis of two disciplines', *Learn. Mem* **14**, 1–16.

- Weinberger, N. M., Ashe, J., Metherrate, R., McKenna, T. M., Diamond, D. M. & Bakin, J. (1990), 'Returning auditory cortex by learning: a preliminary model', *Concepts Neurosci.* **1**, 91–132.
- Wenstrup, J. J. (1999), 'Frequency organization and responses to complex sounds in the medial geniculate body of the mustached bat', *J. Neurophysiol.* **82**, 2528–2544.
- Wenthold, R. (1987), 'Evidence for a glycinergic pathway connecting the two cochlear nuclei: an immunocytochemical and retrograde transport study', *Brain Research* **415**, 183–187.
- Werker, J. F. & Tees, R. C. (1984), 'Cross-language speech perception: evidence for perceptual reorganization during the first year of life', *Infant Behav Dev* **7**, 49–63.
- Wever, E. G. & Bray, C. W. (1930), 'The nature of acoustical response: the relation between sound frequency and frequency of impulses in the auditory nerve', *J. Exper. Psychol.* **13**, 373–387.
- Wible, B., Nicol, T. & Kraus, N. (2005), 'Correlation between brainstem and cortical auditory processes in normal and language-impaired children', *Brain* **128**(2), 417–423.
URL: <http://brain.oxfordjournals.org/content/128/2/417.abstract>
- Wickesberg, R. E. & Oertel, D. (1988), 'Tonotopic projection from the dorsal to the anteroventral cochlear nucleus of mice.', *J. Comp. Neur.* **268**, 389–399.
- Willott, J. F., Parham, K. & Hunter, K. P. (1988), 'Response properties of inferior colliculus neurons in young and very old cba/j mice', *Hearing Research* **37**, 1–14.
- Windhorst, U. (1988), *How Brain-Like is the Spinal Cord? Interacting Cell Assemblies in the Nervous Systems*, Berlin: Springer-Verlag.
- Winslow, R. L., Barta, P. E. & Sachs, M. B. (1987), *Auditory Processing of Complex Sounds*, Lawrence Erlbaum, Hillsdale, NJ, chapter Rate coding in the auditory nerve, pp. 212–224.
- Winter, I. M. & Palmer, A. R. (1990), 'Responses of single units in the anteroventral cochlear nucleus of the guinea pig', *Hearing Research* **44**, 161–178.

- Winter, I. M., Palmer, A. R., Wiegrebe, L. & Patterson, R. D. (2003), 'Temporal coding of the pitch of complex sounds by presumed multipolar cells in the ventral cochlear nucleus', *Speech Communication* **41**, 135–149.
- Winter, I. M., Robertson, D. & Cole, K. S. (1989), 'Origin of descending projections to cochlea and cochlear nucleus in the guinea pig.', *J. Comp. Neur.* **280**, 143–157.
- Wong, S. W. & Schreiner, C. E. (2003), 'Representation of cv-sounds in cat primary auditory cortex: intensity dependence', *Speech Communication* **41**, 93–106.
- www.planetoftunes.com (2015).
URL:
<http://www.planetoftunes.com/sound-audio-theory/complex-soundwaves.html>
- Xie, R., Meitzen, J. & Pollack, G. D. (2005), 'Differing roles of inhibition in hierarchical processing of species-specific calls in auditory brainstem nuclei', *J Neurophysiol* **94**, 4019–4037.
- Yost, W. (2005), *Fundamentals of hearing: an introduction*, Academic Press, San-Diego.
- Young, E. D. & Brownell, W. E. (1976), 'Responses to tones and noise of single cells in dorsal cochlear nucleus of unanesthetized cats', *Journal of Neurophysiology* **39**(2), 282–300.
- Young, E. D. & Sachs, M. B. (1988), 'Interactions of auditory nerve fibers and ventral cochlear nucleus cells studied with cross-correlation', *Soc. Neurosci. Abstr.* **14**, 648.
- Zatorre, R. J., Evans, A. C., Meyer, E. & Gjedde, A. (1992), 'Lateralization of phonetic and pitch discrimination in speech processing', *Science* **256**, 946–949.
- Zhang, L. I., Bao, S. & Merzenich, M. M. (2002), 'Disruption of primary auditory cortex by synchronous auditory inputs during a critical period', *Proc. Natl. Acad. Sci. U. S. A.* **99**, 2309–2314.
- Zhang, L. I., Tan, A. Y. Y., Schreiner, C. E. & Merzenich, M. M. (2003), 'Topography and synaptic shaping of direction selectivity in primary auditory cortex', *Nature* **424**(6945), 201–205.
URL: <http://dx.doi.org/10.1038/nature01796>

- Zilany, M. S., Bruce, I. C., Nelson, P. C. & Carney, L. H. (2009), ‘A phenomenological model of the synapse between the inner hair cell and auditory nerve: long-term adaptation with power-law dynamics’, *J Acoust Soc Am* **126**(5), 2390–2412.
- Zwicker, E. (1974), *Facts and models of hearing*, Springer Verlag, Berlin, chapter On a psychoacoustical equivalent of tuning curves, pp. 132–141.