Citation: Hiroki Terashima (2014) *Computational Model for Auditory Cortex: An Analogy to Visual Cortex*. PhD thesis, The University of Tokyo.

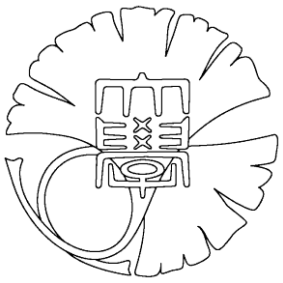
Computational Model for Auditory Cortex: An Analogy to Visual Cortex

（大脳皮質聴覚野の計算論的モデル：視覚野との類比）

TERASHIMA, Hiroki

寺島裕貴

Submitted for the degree of
Doctor of Philosophy



The University of Tokyo, Japan
December 2013

ABSTRACT

The neocortex, the locus of intelligence, is functionally diverse, but structurally uniform, which has drawn researchers into an odyssey to identify computational principles. However, computational cross-areal comparison has been lacking, primarily due to the incomparability of function. We propose a model-driven strategy using an analogy of the primary visual cortex (V1), the only area that is well understood. Through a top-down approach, this thesis demonstrates a computational model of the primary auditory cortex (A1), which has been difficult to interpret owing to its apparently disorganised characteristics. The analogous modelling enables us not only to model A1 but also to further understand the neocortex as a whole.

To model A1 in an analogy with V1, we hypothesise that A1 and V1 share learning strategies, while adapting to different input statistics. Thus, before modelling A1, we simply compare statistics of natural images and natural sounds. In accordance with our hypothesis, three specific concepts of V1 are analogously discussed for A1: (1) the receptive field of simple cells, (2) topography, and (3) complex cells. We discuss effective use of the model-driven approach in neuroscience in an analogy to linguistics. Learning of sparse representations, which has come to the forefront of the machine learning literature, may prove to be a general strategy for understanding high-dimensional data in a world of amazing complexity and harmony.

論文要旨

知能の座である大脳皮質は領野毎に多様で高度な機能を呈す一方，解剖学的には一様であり何らかの計算原理の存在が示唆される．原理解明には計算理論のレベルにおける領野間比較が必要だが，異なる領野の計算や機能の直接比較が困難なため行われていない．この「計算論的領野比較」の実現方策として，本論文は，唯一理解が進んでいる一次視覚野（V1）を起点とする類比（アナロジー）を提案する．具体的には，このモデル駆動戦略を用いて，これまで V1 との相違が大きく解釈困難だと考えられてきた一次聴覚野（A1）を計算論的にモデル化できることを示す．このモデル化は単に A1 の理解を深めるだけでなく，皮質計算原理解明の一助となる．

V1 との類比を用いた A1 の計算論的モデル化において我々は，両領野の差異は自然刺激統計性の違いに由来し，学習戦略自体は共通であるという仮説を立てた．よって本論文ではまず自然音と自然画像を比較し，その統計的な差異を示す．そして仮説に従い，(1) 単純細胞受容野，(2) 地図構造，(3) 複雑細胞という三つの V1 概念を類比によって A1 に移し，モデル駆動で議論できることを示す．また，多様性の中に普遍性を見出すという意味において，計算論的領野比較と言語学の類似を指摘する．スパースな表現の獲得は機械学習においても重要なアプローチであり，複雑な—しかし根底には調和が織り込まれているだろう—多次元入力を理解するための一般戦略だと考えられる．

Acknowledgements

I would like to express my deepest appreciation to my supervisor, Professor OKADA Masato, who has supported and encouraged me throughout my graduate studies. I was influenced by his broad, interdisciplinary perspective that typically tries to bridge different scales. The most significant contribution came from clear "pictures" that he often drew as a physicist, particularly those about what it means to find the universality or to establish a theory or a model. I am also in debt to the members of my thesis committee, too: Dr ICHINOHE Noritaka, MD, PhD, Lecturer TAKAHASHI Hirokazu, Professor NOSE Akinao, and Associate Professor KUNIHIRO Noboru. In particular, this thesis would be lacking important results if Dr Ichinohe did not provide his recording of marmoset vocalizations.

I also thank those who gave me opportunities to dive into this exciting research field. Dr NISHIO Hirokazu, whom I met in Nada High School, helped develop most of my intellectual background: programming languages, artificial intelligence, human intelligence, and natural languages. Moreover, he is an outstanding man who made me ponder the nature and potential of human beings regarding intelligence and creativity. Dr HOSOYA Haruo enthusiastically argued that now is the time for computer scientists to address the mystery of the neocortex, which is in fact the most important topic of this thesis. He first introduced me the joy of research, at the same time showing his own bravery when he shifted his research field.

During my graduate studies, I sometimes encountered difficult situations, but each time, I was supported by my colleagues. In particular, I thank those who began the graduate program with me and have always provided fun and laughter: IGARASHI Yasuhiko, OTSUBO Yosuke, and SAITO Hiroshi, without whom it would have been difficut to complete this thesis. I also thank other lab members including Assistant Professor NAGATA

Kenji and the alumni who welcomed me five years ago. My graduate life has also been supported by public organisations: to name few, the Japanese Neural Network Society, the Japanese Society for Artificial Intelligence, and the Japan Society for the Promotion of Science. I am proud of having been a research fellow of the JSPS.

Kazuo, Keiko, and Yumi have always been supportive, giving selfless love as my family even though my perspective was often obscure. Kazuo also used his experience in academia to advise me. Lastly, I sadly have to mention the biggest change that occured since the beginning of my graduate studies: I unfortunately lost three of my four grandparents; I thank TERASHIMA Yasuharu and Toshie and FUKUDA Kazuo and Tsuyako, who recognised the importance of higher education when it was not necessarily common and kept an unbroken line under the overwhelming power of nature in harmony.

Contents

List of Figures

List of Tables

Chapter 1

Why computationally model the auditory cortex?

Out of clutter, find simplicity;
From discord make harmony; and finally
In the middle of difficulty lies opportunity.

ALBERT EINSTEIN[1]

1.1 Neocortex: diversity and pursuit of computational principles

When you read this thesis, you perceive vibrations of the air at the same time. Our recognition of the world crucially depends on the perpetual use of the two major sensory modalities: vision and audition. Continually exposed to massive data streams in the natural environments, both systems have adapted and their functions are surprisingly specialised for survival. For example, if you enter a room where there are delicious cookies on the centre table, you cannot help quickly glancing at them; even in the middle of a crowd, you can identify the voice of an individual you love. The methods of computation underlying these processes, which have developed through both evolution and development, may seem very different. Nonetheless, in the central nervous system, the two ways of computation are processed by a single organ, the neocortex.

Similar to the two distinct sensory modalities, the neocortex is functionally heterogeneous. Neuroscience has revealed that it can be spatially divided into many functional areas. Even within the sensory cortex, in addition to the visual and auditory cortices, it

[1]The three rules were attributed to him by Wheeler, J. A. [135].

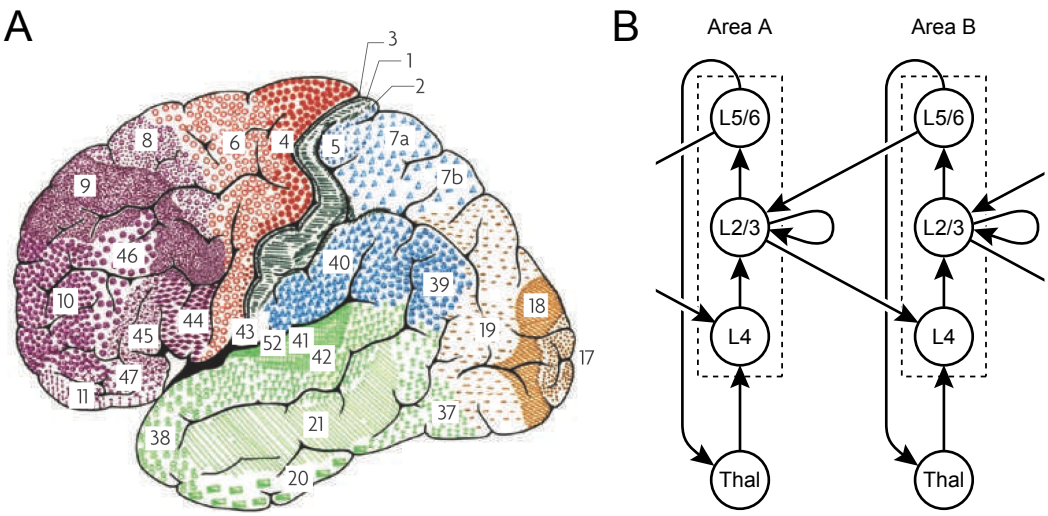


Figure 1.1: **Introduction to the "computational cross-areal comparison" for the neocortex.** (*A*) The neocortex consists of functionally diverse areas. (They roughly correspond to the coloured topography that are anatomically defined.) (*B*) The functional diversity seems underpinned by the "canonical circuits", which has suggested the existence of some computational principles. To reveal them, we need to compare multiple areas at the computational level. Specifically, the thesis focuses on the primary visual cortex (V1; area 17 of (A)) and the primary auditory cortex (A1; area 41 and 42 of (A)). The arrows in the image indicate the direction of primary information flow between layers, indicating that the canonical circuits are mainly composed of three types of loops. A dotted rectangle shows a single functional area of the neocortex. L5/6: layers 5 and 6; L2/3: layers 2 and 3; L4: layer 4; Thal: thalamus. (The image (A) was adapted from [140]; the idea of the component arrangement of (B) is by courtesy of Dr Jun-nosuke Teramae at Osaka University.)

also encompasses the somatosensory cortex, the olfactory cortex, and the gustatory cortex. While the sensory cortices are considered the gates of perception, it also commands actions of the animal from the motor cortex, the centre of motor control. More generally, it is considered the locus of intelligence because the frontal and association cortices play central roles in integrating information and making inferences and decisions. While it consists of nearly everything associated with intelligence, the repertoire of its functions is surprisingly diverse (Figure 1.1A).

On the other hand, the mystery of the neocortex lies in the observation that distinct functional areas seem underpinned by almost common anatomical features [19, 39]. Somas of the neocortex neurons are distributed only in the grey matter, a thin six-layered

structure on the surface. Its spatial structure is essentially two-dimensional because the responses of radially aligned neurons are typically similar. The direction of information flow is primarily radial, and basic relationships between the layers are seemingly universal across functional areas, which have been called the 'canonical circuit' [80, 29, 30] (Figure 1.1B). The uniformity over the entire neocortex suggests that its diverse functions are realised through common "intelligent" processes or computational principles.

How can we understand the apparent discrepancy between diversity and uniformity? Marr [71] argued that if we desire to "understand" complex computing machinery, we need to investigate it at multiple levels in parallel. In addition to the implementation level and the representation and algorithm level, Marr proposed to study a complex system (the visual system in his case) at the level of computational theory. For our issue, it has been difficult to achieve the goal by working only at the implementation level or the level of neurophysiology. Thus, in order to understand an ultimately complex system, we need to challenge the issue of diversity and uniformity at the level of computational theory. However, Marr's theory itself does not provide a practical strategy to identify computational principles within this diversity.

1.2 Lack and difficulty of computational area-to-area comparison

What does it mean to identify universal features and how are these features typically found? The word "universal" inherently assumes diversity, at least on a superficial level. To discuss diversity, we first need to gather multiple instances of data systematically. Based on analysis of the observation, diverse features are established. Then, we can seek universality underlying them, i.e., a set of concise, human-understandable rules that can explain most of the diversity. This procedure, that is, *comparison* of multiple data, will also be needed for our ultimate goal of identifying computational principles, where the comparison must be done at the level of computational theory.

The necessity of computational area-to-area comparison is, in a sense, trivial. However, in fact, there has essentially been no attempt in the field of computational neuroscience. Why? To understand the reason, we provide two examples of successful cross-areal comparisons for the neocortex.

The first example is anatomy. The historical study by Brodmann [19, 39] established

the cross-areal comparison for the cortex at the level of histology. He first systematically collected histological data, which following a detailed analysis showed distinct characteristics, and indicated that areal demarcation forms a topography on the neocortex. However, he argued that based on the analysis all diverse anatomical features of distinct areas may stem from a single prototypical six-layered structure, that is, he emphasised the universality.

The second, relatively new example is genetics. Rapid advancements in genetic tools now allow us to systematically investigate gene expression patterns in the rodent brain [82, 43, 83], which can be one of the foundations of cross-areal comparison for the neocortex. The expression atlas has begun to be statistically analysed, revealing many area-specific distinct patterns. This illustrates that cross-areal comparison can naturally follow observations, once the necessary techniques are established, because the results of the observations are easily comparable.

What of the comparison of neural activities? Most neuroscientists agree that the minimal unit of neural representations is a spike. In fact, statistical analysis of spike patterns can be conducted across neocortical areas without any area-specific manipulations [113, 77]. However, in reard to "computation" or "function" represented by the spikes, the situation dramatically changes, and it is difficult to compare different functional areas. First, how can we collect functions discussed for multiple areas? We face an issue that the functions of only a few areas (e.g., the primary visual cortex and primary motor cortex) have been discussed and established, whereas most areas have no explicit functional description. Second, even if we select two areas whose functions are known, how can we compare the two functions (e.g., edge detection in the eye field and motor command)? A function of a specific area is arbitrarily defined in order to interpret the way information is processed in the area, which means that there are generally no guarantees of *comparability*. Thus, the typical procedure of comparison cannot be applied for function or computation (Table 1.1).

Does the difficulty prevent us from comparing neocortical areas in any other way? In this thesis, we propose to tackle the issue using a strategy different from examples of anatomy and genetics. In these two successful examples, the strategy used was induction, which attempts to extract rules in a data-driven way. By contrast, for the issue of what we call "computational cross-areal comparison", the data-driven approach is unpromising; our

Table 1.1: **Three levels of neocortical area-to-area comparison.**

	Genetics	**Anatomy**	**Computation**
Global data availability	High	High	Low
Comparability	High	High	Low
Strategy of inference	Induction	Induction	Analogy

functional understanding is limited to a few areas, in particular, computational modelling has flourished only in the primary visual cortex (V1). The strategy we propose is to use V1 as a source of *analogy* in order to understand other areas whose functions are less known (Table 1.1). Analogy is a cognitive process for transferring concepts from a source to a target, and thus, it can be applied even when equality of understanding and natural comparability are not satisfied.

1.3 Why analogise the auditory cortex to the visual cortex?

We suggested an analogy with V1 as a strategy to understand other areas by overcoming the function incomparability. The richness of V1 neurophysiological literatures enables researchers to quantitatively discuss the results of computational studies. As a result, V1 is currently the area that is considered to be the most understood among the sensory cortices and that has the richest repertoire of computational models, making it the most qualified as the source of analogy. Which functional area should be analogised with V1? Our target in this thesis is another primary sensory cortex, the primary auditory cortex (A1).

The analogy is primarily supported by the similarity of the structures, which is the base of analogy according to structure mapping theory [40]. Here, it is useful to examine the results of systematic comparisons at the level of anatomy and genetics. From the very beginning of neuroanatomy, V1 and A1 have been similarly characterised as the neocortical first gates in each sensory pathway; these regions have the thickest layer 4, which receives primary inputs from the modality-specific thalamic nuclei, showing (macroscopic) topographies on their surface. Recent findings also show their outstanding proximity in terms of gene expressions [82, 43, 83]. Even when compared with the primary somatosensory cortex (S1), V1 appears more similar to A1. These similarities justify our attempt of analogy between the two areas.

Furthermore, more direct support for the analogy between V1 and A1 can be found in neurophysiology. Sur et al. [117, 2, 111] performed surgeries on immature ferrets to divert their peripheral visual inputs into the auditory subcortical pathway. After development, A1 neurons exhibited responses to visual stimuli similarly to normal V1 neurons. The results suggest that even though the functions of V1 and A1 are distinct and seem incomparable, both are underpinned by a single adaptive computational strategy.

The structural similarity and the rewiring experiment validate the purpose of this thesis: a computational modelling of A1 in an analogy with V1. Computational models for A1, "the last frontier" of the auditory pathway [73], have been by far less fruitful than models for V1, mainly because the neurophysiological features of A1 neurons seem so disorganised that we have difficulties in interpreting them intuitively. This discrepancy might suggest that V1 and A1 adopt different computational strategies; even so, they remain very similar to each other in terms of neuroanatomy. What is common between V1 and A1? And what other factors are not? Answering this pair of questions will lead to the realisation of computational models for A1, which would further our understanding not only of A1 but also the entire neocortex.

1.4 Thesis themes

1.4.1 Computational cross-areal comparison for the neocortex

To identify universality in the functional diversity of the neocortex, we must simultaneously consider multiple areas at the level of computational theory, what we have named a "computational cross-areal comparison". However, computational modelling has typically been tried for specific single areas with no comparison. This contrasts with the cases of anatomy and genetics, mainly due to the incomparability of computation, and the limited number of areas whose functions have been discussed. How can we approach computational cross-areal comparisons?

We discuss the potential of this strategy using an analogy to V1. This can be viewed as model-driven, whereas the successful achievements in anatomy and genetics were data-driven or inductive. Universal features at the level of computational theory will be found through this approach, and some of its first examples will be discussed in this thesis.

Finally, we will discuss another, more abstract analogy between computational cross-areal comparison and linguistics, which involves another pursuit to derive universals out of diversity.

1.4.2 The core hypothesis for modelling of A1

The purpose of this thesis is to computationally model A1 in comparison to V1, considering the question 'what is similar and what is not?'. To answer this question, we focus on the highly adaptive nature of the neocortex. Although the spatial distribution of functional areas is basically innate in normal adults, the areas are flexible and can compensate for each other in special cases. For example, when a man loses his sight even after birth, his visual cortex adapts to the new situation and begins to process auditory information. Particularly for V1 and A1, more direct evidence has been demonstrated by physiological studies. Sur et al. [117, 2, 111] performed surgeries on immature ferrets to divert their peripheral visual inputs into the auditory subcortical pathway. After development, A1 neurons exhibited responses to visual stimuli that were similar to normal V1 neurons. The result suggests that the physiological characteristics of V1 and A1 neurons are mainly shaped by the statistics of their inputs based on a single adaptive strategy.

Throughout this thesis, we consistently discuss our view that what differentiates V1 and A1 is not adaptive strategies, but the statistics of stimuli in natural environments. We propose to model A1 using the same learning rules with V1, in particular, ones that learn sparse representations from natural image statistics [87, 88, 89]. From this viewpoint, the difficulty in interpreting A1 neurons results from the statistics of natural sounds. The simple view, "learning is same, but input is different", enables us to model A1 in a manner integrated with V1.

1.4.3 Simple cells, topography, and complex cells

The thesis attempts to validate the hypothesis by modelling A1 based on this approach. First, we will show how natural stimulus statistics differ in vision and audition. Then, we discuss how this difference is reflected in three neurophysiological aspects, which are seemingly contrstive between V1 and A1, through a single adaptive strategy.

The first contrast is the form of receptive fields of neurons that basically respond lin-

early. Receptive fields of V1 simple cells are strictly limited within a very localised area in the eye field, while those of A1 neurons can be non-localised and tend to be harmonically structured. What is the source of this contrast? Computational models for V1 have explained the V1 receptive fields as a result of adaptation to natural image statistics. Based on the hypothesis, we applied the learning rule to natural sounds, instead of natural images. The approach reproduced receptive fields that are non-localised and harmonically structured, which are similar to A1 neurons.

The second contrast is the smoothness of topographic maps. V1 and A1 have been known for retinotopic and tonotopic maps, respectively, which have been considered similar structures. However, recent advances in optical imaging have revealed that the A1 map is much more disordered at a microscopic scale. Why is the A1 map more scattered? V1 models have shown that the smooth V1 map can emerge from adaptation to natural image statistics. Based on the hypothesis, we applied the learning rule for V1 to natural sounds. The obtained tonotopic map was more disorganised than the V1 map, which suggests that the scattered frequency map can be efficient at integrating distant frequencies.

Following the two known contrasts above, we considered the third contrast, which has not yet been noted: complex cells. The complex cells in V1 nonlinearly respond to visual stimuli and are considered to hold representations at a higher level than simple cells. The concept has been firmly established in the visual cortex, but there has been relatively no discussion about their counterparts in other modalities. Can we even consider "complex cells" of A1? If we can, what is their function? V1 models have suggested that complex cells can also emerge from natural image statistics, which are closely related to the smooth retinotopic map. We applied the learning rule of V1 complex cells to natural sounds. The learned "complex cells" of A1 showed nonlinearity similar to a psychoacoustic phenomenon called the "missing fundamental", which resemble the pitch cells recently identified in the core field of the monkey auditory cortex, including A1. This suggests that A1 pitch cells are computationally analogous to V1 complex cells.

The three cases discussed consistently support our core hypothesis. The results provide computational models for A1 and further the understanding of A1. Moreover, these models would also contribute to our understanding of the neocortex.

1.5 Outline of the thesis

The thesis is constructed as follows (Figure 1.2). Chapter 2 introduces the common background of the two functional areas we will focus on, V1 and A1. In Chapter 3, following the discussion of their similarities and differences, the core hypothesis is presented. Chapter 4 shows the representative contrast between natural image statistics and natural sound statistics. Based on the hypothesis, Chapters 5 and 6 model A1 neurons as the result of adaptation to natural sound stimuli using V1 learning models. Extrapolating these results, Chapter 7 deductively discusses "complex cells" of the auditory cortex. Following the discussion in Chapter 8, the last chapter concludes the thesis.

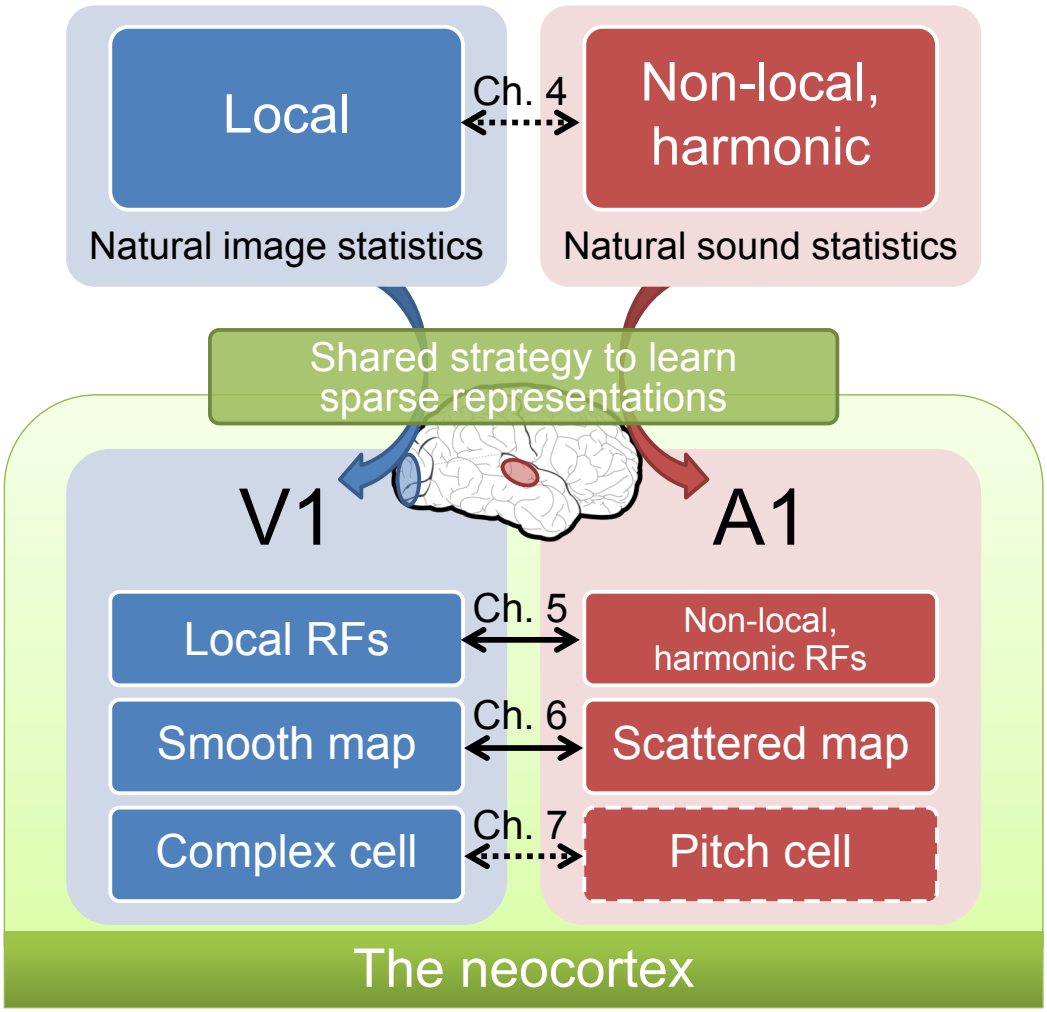


Figure 1.2: **Schematic summary of the thesis.** The purpose of the thesis is to computationally model A1 in analogy with V1. Our core hypothesis was "learning is same, but input is different". Prior to modelling, a clear contrast was shown between natural image statistics and natural sound statistics. This contrast was shown to be reflected, through common learning rules, in the features of A1 neurons that contrast with three specific features of V1 neurons: the receptive field of simple cells, topographic maps, and complex cells. RF: receptive field; Ch.: chapter; dotted lines: concepts that previous studies have not discussed in a contrastive way.

Chapter 2

The visual and auditory cortices

> For men of science nothing is so important as the clear definition of differences.
>
> HERMANN HESSE

This chapter briefly introduces the neurophysiological characteristics of the primary visual cortex (V1) and the primary auditory cortex (A1) as the common background for this thesis. Their similarities and differences will be discussed, and a unified interpretation will be presented using our hypothesis.

2.1 The primary visual cortex (V1)

2.1.1 Subcortical pathway

Where does the input to V1 originate from? The subcortical visual pathway starts at the retinas of the eyes. Onve the incoming light is received, photoreceptors on the retina generate spikes. The spikes are processed in the layered structure of the retina and eventually pass to the retinal ganglion cells, which are the projection neurons from the retina. The firing of the ganglion cells is characterised by either an on-centre or off-centre pattern of visual stimuli [65]. Spatial organisation on the retina is maintained in the topography of the output neurons, i.e., the adjacent points of the visual image are mapped onto the adjacent ganglion cells.

The ganglion cells mainly project to a specific part of the thalamus known as the lateral

geniculate nucleus (LGN). Projections from an eye on one side are split to the LGNs in both hemispheres, according to the specific side of the visual fields (i.e., right or left) in which the ganglion cell's receptive field lies in. As a result, LGN neurons have receptive fields in the contralateral side of the visual field. The receptive fields of LGN neurons have typically been described as the on- or off-centre type, whose function is a type of whitening or decorrelation. LGN neurons then project to the layer 4 neurons of V1 in the ipsilateral hemisphere.

2.1.2 Receptive field of simple cell

V1 neurons require more complex structures of stimuli to induce firing than subcortical neurons. They are selective to edges of a specific orientation shown at a specific position in the visual field [47]. Since the very early stage of studies on V1, its neurons have been classified into two classes: simple cells and complex cells [47]. Responses of V1 simple cells are characterised by their linearity, which enables us to describe their receptive fields as linear filters. Shapes of the filters are usually identified using a reverse correlation method [55, 28, 100].

The class of forms of their receptive fields are well described as two-dimensional Gabor filters [55], i.e., two-dimensional sinusoidal waves that are spatially restricted by Gaussians. They show selectivity to a specific orientation at a specific retinal position, which typically occurs with inhibitory fields on its sides. The typical size of receptive fields are approximately $\sim 1\,\mathrm{deg}$ for monkeys and $\sim 35\,\mathrm{deg}$ for mice [127].

2.1.3 Topographic maps

As the basic structure of the neocortex is two-dimensional, we can discuss the topography of a specific area or a spatial distribution of a receptive field property of its neurons. A topography with regards to the retinotopic position of receptive fields is called retinotopy. The retinotopic map of V1 is the best preserved retinotopy in cortical visual areas, which can be clearly visualised using functional magnetic resonance imaging (fMRI) even in humans [130, 15] (Figure 2.1A). Although V1 also has other types of topographies (e.g., the orientation map and the ocular dominance map), these topographies will not be discussed in this thesis.

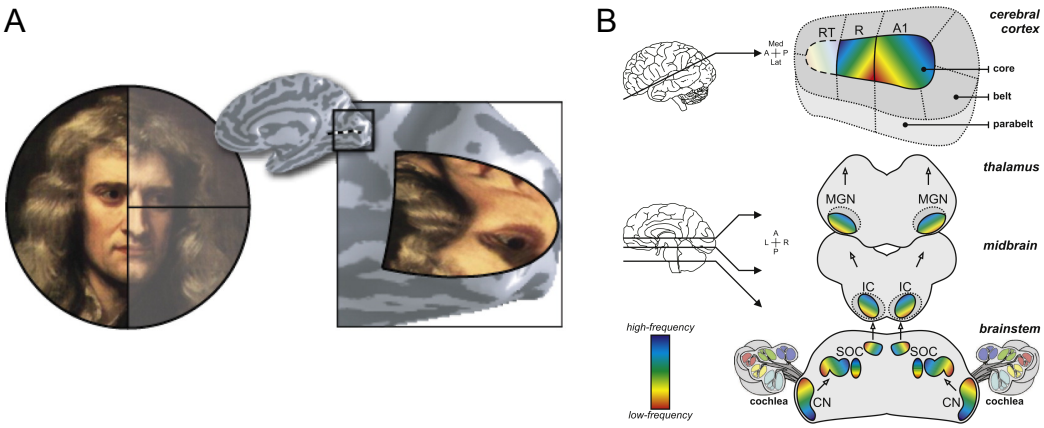


Figure 2.1: **Retinotopic map and tonotopic map**. (*A*) Activity patterns on the surface of V1 exhibit a correspondence with those on the retina, accompanying the cortical magnification to some extent. (*B*) Frequency representation is preserved across the auditory pathway leading to A1. (The images ware adapted from [130, 104].)

Classic discussions on the retinotopic map were focused on a macroscopic scale in the sense that there was no way to systematically map response properties of each neuron while excluding sampling bias. Recent advances in recording techniques, however, have enabled us to directly map neural activities at the single-cell scale in vivo. Use of these techniques has revealed that the retinotopic map is still relatively smooth even at a microscopic scale [17]. Nearby neurons are selective to similar retinotopic positions, whose scatter is approximately the size of a neuron's receptive field [17].

2.1.4 Complex cell

Similarly to a simple cell, a complex cell has a spatially restricted receptive field and is selective to a specific orientation. However, its response cannot be characterised by a single linear filter. Whereas a simple cell is selective to both the orientation and the phase of sinusoids, a complex cell is selective to the orientation, but not to the phase. This property is called phase invariance.

The nonlinear response has been modelled as the pooling of activities of simple cells, which are selective to the same orientation and different phases. This leads to a view that representations of complex cells can be found at a higher level than simple cells. In fact, in monkey V1, simple and complex cells are typically found in layers 4 and 2/3, respectively,

which means there are major projections from simple cells to complex cells according to the canonical circuit model.

2.2 The primary auditory cortex (A1)

2.2.1 Subcortical pathway

The starting point of the auditory pathway is the ear. Amplified through the ear canal, vibrations of the air are transformed into spiking activities by the hair cells at the cochlea. A hair cell is selective to a specific frequency, which reminds us of Fourier transform or wavelet transform. The topography of the receptors is in accordance with the logarithm of frequency, which is called tonotopy.

On the way to the neocortex, the auditory pathway has several relay nuclei. These nuclei primarily integrate information from both ears for the purpose of sound localisation [57]. Less is done for the integration of different, particularly distant, frequency channels, which basically preserves the tonotopic map throughout the subcortical nuclei [76, 74, 75, 104] (Figure 2.1B). Auditory neurons in the medial geniculate body (MGB) of the thalamus, the last relay nucleus, project to the layer 4 neurons of A1.

2.2.2 Receptive field: frequency integration

It is difficult to simply describe the repertoire of the receptive fields of A1 neurons, even on the frequency domain alone [107]. The biggest population shows sharp selectivity to a single frequency, while some have a broad frequency tuning curve and others are selective to multiple distant frequencies. The neurophysiological diversity has not provided a simple description of its function. The frequency tuning properties will be a focus of this thesis, as the time domain is less evident in A1 receptive fields [33].

Can we find any structures in the non-localised receptive fields? Kadia and Wang reported that multiple peaks of the frequency tuning curve tend to be harmonically related in marmoset monkey A1 using single cell recording [56]. A local field potential study on marmoset monkey A1 [34] also suggests a harmonic structure. For cats, Qin et al. [95] showed that the excitatory peaks are harmonically related and, in addition, showed a form of sensitivity to fundamental frequency, although little is yet known regarding nonlinear

computations in A1.

2.2.3 Topographic maps

Following the subcortical structures, the tonotopy is the most evident topographic structure on A1 [108]. As the frequency topography requires only one dimension, the two-dimensional cortical map has a redundant topographic structure. A discussion of the other "axis" on the map has been controversial, and there is no clear accepted notion [108].

Recent advances of recording techniques have revealed that the tonotopic map is substantially disorganised at the single-cell scale: nearby A1 neurons in layers 2/3 can be selective to frequencies distant by up to four octaves [8, 103]. This scatter is quite large, as the hearing range is ~ 10 octaves at most [91]. On the other hand, at a macroscopic level, the global tonotopic map still exists based on the averaged activity. Thus, we can discuss the global tonotopy of A1, similar to the retinotopy of V1, while it is much more locally scattered than the V1 map.

Chapter 3

The hypothesis

孟方水方，孟圜水圜。
If the basin is square, the water is square;
if the basin is round, the water is round.

孔子
CONFUCIUS

3.1 Similarities between V1 and A1

A classic view for V1 and A1 is that they are basically similar. V1 and A1 are the first gates to the neocortex in the visual and auditory pathways, respectively, which start to form complex receptive fields. The primary topographies of both areas, i.e., the retinotopy and the tonotopy, reflect topographies of the peripheral receptors in each pathway (Figure 3.1).

A similarity can also be found in the neuroanatomical features. The layer 4 of V1 and A1 are two of the thickest in the neocortex, both of which primarily receive projections from the thalamus. Analyses of their radial projections have supported the view of canonical circuits [29]. For V1 and A1, we can identify the best similarity of two different functional areas in the neocortex.

3.2 Dissimilarities between V1 and A1

Despite of the similarities between V1 and A1, they also have some dissimilarities or contrasts that have made our understanding of A1 more difficult than V1. We discuss

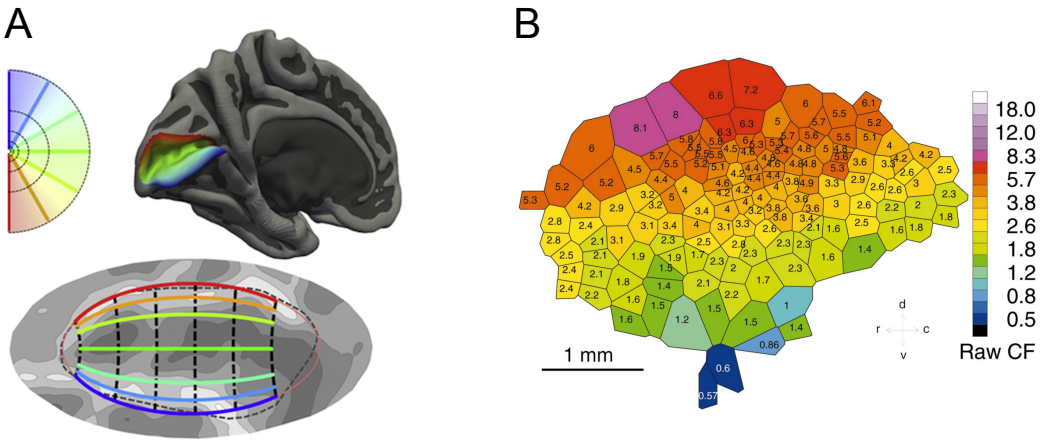


Figure 3.1: **Similar topographic structures in V1 and A1 at a macroscopic scale**. (*A*) Schematic of the retinotopic map of V1. (*B*) The tonotopic map (frequency map) of A1 of a squirrel monkey. (The images were adapted from [15] and [21], respectively.)

three of these dissimilarities in this thesis.

The first dissimilarity is related to the localisation of the receptive fields. The receptive field of a V1 neuron is restricted within a very small area in the eye field, whereas the receptive fields of A1 can be non-localised in the frequency domain. The contrast seems to suggest that we should consider that V1 and A1 adopt different computational strategies.

The second dissimilarity is the degree of scatter seen in the topography. In a classic view, V1 and A1 were analogous in terms of the topographies that reflect those of peripheral receptors. However, recent findings at a microscopic scale have showen that the A1 map is much more scattered than the V1 map, which seemingly suggests distinct strategies (Figure 3.2).

The third viewpoint is somewhat different: for A1, there has been no discussion on complex cells. The concept of complex cells in V1 has been established for more than 50 years, since the very early stage of neurophysiological studies of V1. Interestingly, there has been almost no neurophysiological discussion about complex cells in other modalities, including audition. Are there any analogues of V1 complex cells? If there are, what is their function? It is hard to answer these questions in a neurophysiological manner, regardless of their existence.

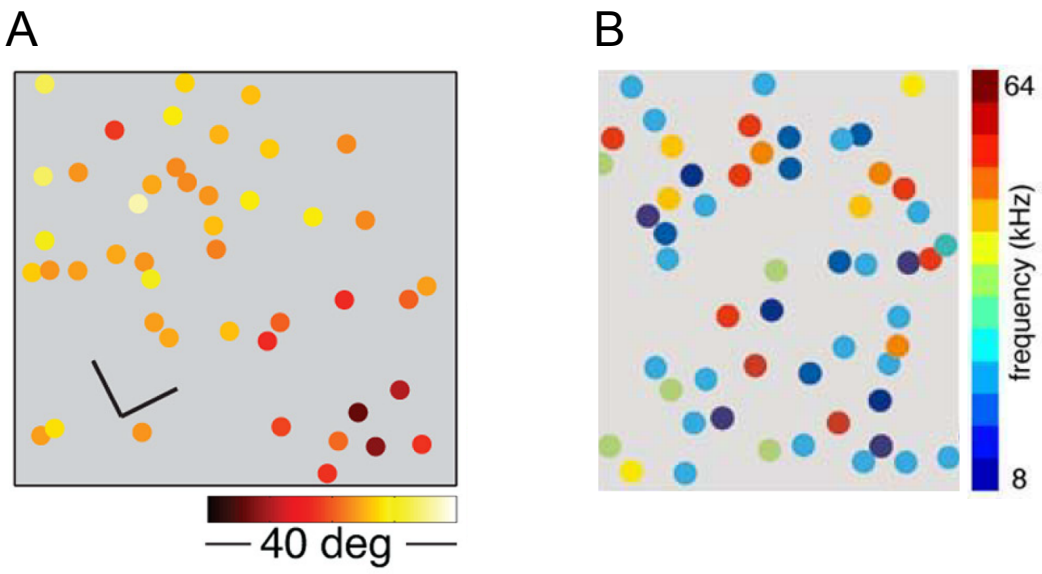


Figure 3.2: **Dissimilar topographic structures in V1 and A1 at a microscopic scale**. (*A*) A dot represents a single cell of mouse V1, showing its stimulus selectivity (the vertical retinotopic position of its receptive field) by its colour. Nearby neurons have similar selectivity, forming a relatively smooth map even at the single-cell scale. (*B*) A dot represents a single cell of mouse A1, showing its frequency selectivity by its colour. In contrast to V1, the map is much more scattered; neighbouring neurons typically have different frequency tunings. (The images were adapted and modified from [17] and [8], respectively.)

3.3 A unified interpretation: shared learning and different inputs

How can we interpret the similarities and the dissimilarities in an integrated manner? We focus on the neocortex's highly flexible adaptability, which for V1 and A1, was directly shown by a line of studies on ferrets. For immature ferrets, Sur et al. [117, 2, 111] diverted retinal inputs to a relay nucleus of the auditory pathway (Figure 3.3). After the surgery, they were reared in a normal conditions, and the responses of A1 were investigated. Surprisingly, neurons in the area usually called A1 showed visual responses similar to normal V1 neurons. The results suggest that the major contributions to the development of neurophysiological features of the areas do not originate from genetic factors, but from the statistics of inputs they receive from the natural environment.

In order to model A1 in comparison with V1, we propose an integrated computational

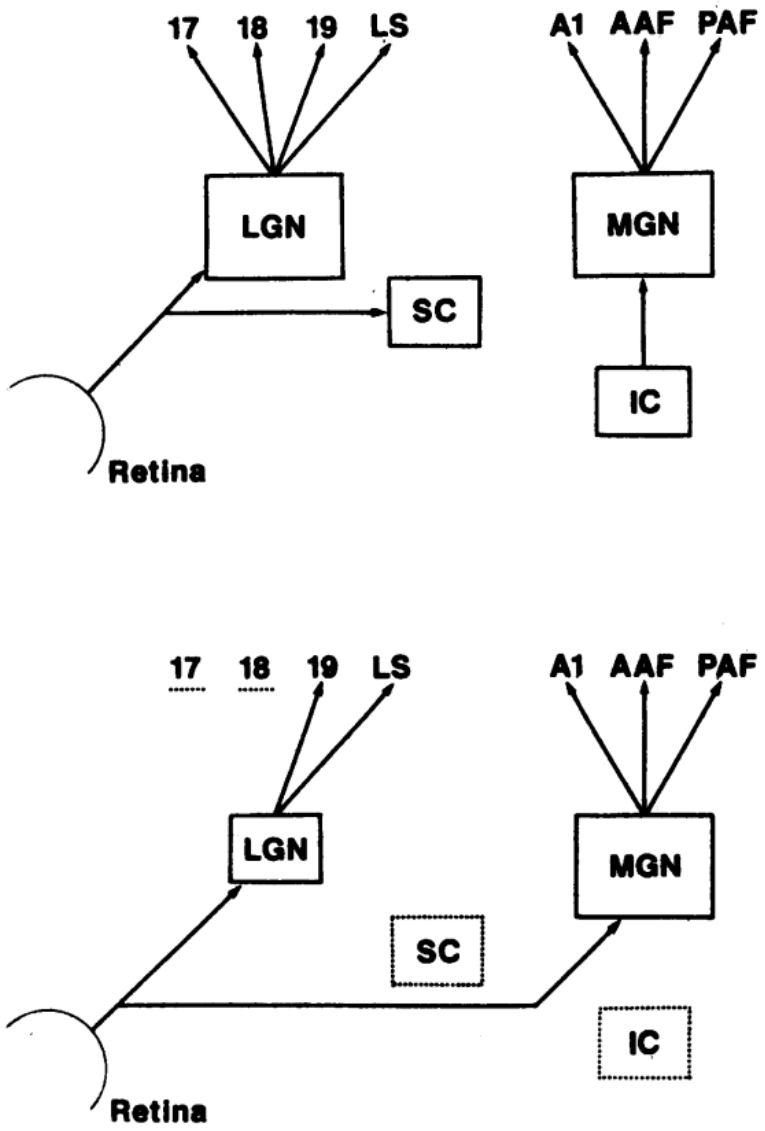


Figure 3.3: **Schematic of the cortical rewiring experiment**. The normal visual and auditory pathways (top) were modified by diverting the visual input to a nucleus in the auditory pathway (bottom). After normal rearing, neurons in the area that is usually called the auditory cortex behaved similarly to V1 neurons. (The image was adapted from [117].)

view for their similarities and dissimilarities: the disorganised properties of A1 neurons result from distinct statistics of natural sounds that contrast to those of natural images, while A1 still shares the learning rule with V1. Even with a single learning rule, the learning result will be dissimilar when the inputs are different, which we think is what happens in V1 and A1. Based on the hypothesis "learning is same, but input is different", we modelled A1 on the three aspects discussed above, which seemingly contrast with V1.

Chapter 4

Natural image statistics vs natural sound statistics

Nature, to be commanded, must be obeyed.

Novum Organum
FRANCIS BACON

4.1 Natural image statistics are localised

Given that V1 is supposed to adapt to natural images and that A1 is supposed to adapt to natural sounds, the first analysis in this thesis simply compared the statistics for natural images and natural sounds[1].

The natural images were taken from the van Hateren database [128] and were reduced four times from their original size. Vertical arrays of 120 pixels each were extracted from the reduced images (Figure 4.1A (left)), each of which covered approximately 8° ($\frac{1}{15}$ of the vertical range of the human field of view). Figure 4.1A (right) illustrates the correlation matrix for these images, which is a simple structure that contains local correlations that span approximately 6°. Because the human field of view spans over $\sim 8^\circ$, this indicates that the statistical dependency in the visual field is localised. This result was not surprising, as distant pixels typically depict different objects.

[1] Preliminary results of this chapter were published in [122].

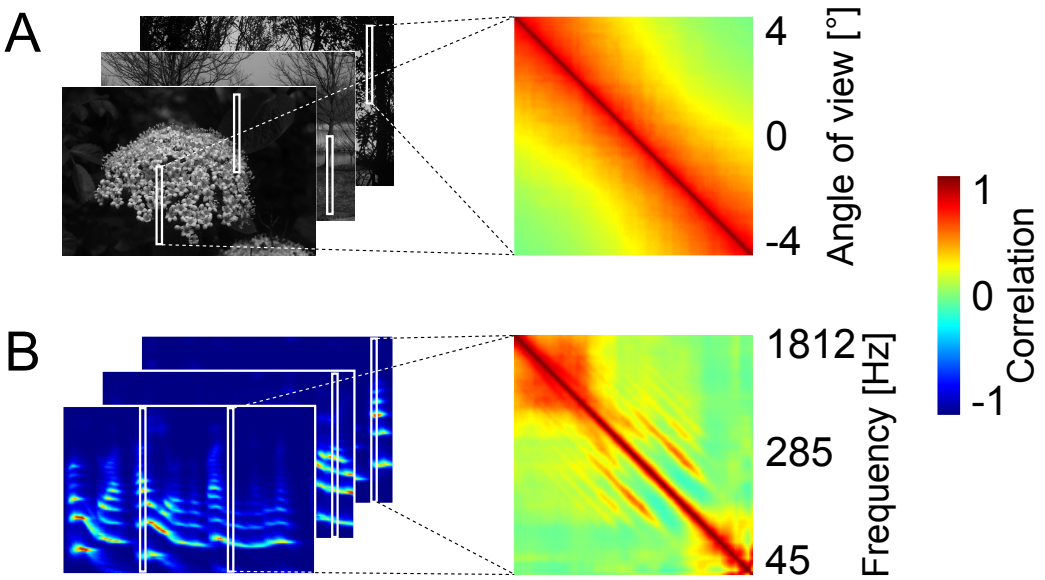


Figure 4.1: **Local statistics of natural images and non-local statistics of natural sounds.** (*A*) The correlation matrix of image strips (right) demonstrated only local correlations ($\sim 6°$) in the field of view ($\sim 120°$). (*B*) The correlation matrix of the human voice spectra (right) demonstrated not only local correlations but also off-diagonal distant correlations, which were produced by harmonics.

4.2 Natural sound statistics are not localised but are structured harmonically

For representative natural sounds, we used human narratives from the Handbook of the International Phonetic Association [53]; efficient representations of human voices have been successful in facilitating studies of various components of the auditory system [68, 114], including A1 [62, 106]. The narrative collection consists of recordings in 27 languages of both male and female voices. After these sounds were downsampled to 4 kHz, their spectrograms were generated using the NSL toolbox [22] to approximate peripheral auditory processing, as shown in Figure 4.1B (left). Short-time spectra were extracted from the spectrograms, each of which were 128 pixels wide on a logarithmic scale (24 pixels = 1 octave). Note that the frequency range (> 5 octaves) spans approximately half of a typical mammalian hearing range (~ 10 octaves [91]), whereas the image pixel array spans only $\frac{1}{15}$ of the field of view.

Figure 4.1B (right) illustrates the correlation matrix for these sounds, which is a complex structure that incorporates distant, off-diagonal correlations. The most prominent off-diagonal correlation, which was 1 octave away from the main diagonal, corresponded to the second harmonic of a sound, i.e., the frequencies at a ratio of 1:2. Similarly, other off-diagonal peaks indicated correlations due to higher harmonics, i.e., frequencies that were related to each other by simple integral ratios. These distant correlations represent relatively typical results for natural sounds, and they differ greatly from the strictly local correlations observed for natural images.

We showed a clear contrast between natural image statistics and natural sound statistics. In addition to the linear correlation shown in our example, the nature of the natural sounds is much more complex [126]. If the different statistics are reflected by the neurophysiological features of V1 and A1 neurons through a common adaptive strategy, the result of adaptation that we can observe may well be contrastive.

Chapter 5

Receptive field of simple cell

> Nature will tell you what to do on the spot,
> fully and adequately.
>
> MICHEL DE MONTAIGNE

5.1 Localised V1 receptive fields vs harmonic A1 receptive fields

In Chapter 4, we showed a contrast between natural image statistics and natural sound statistics. From here on, based on our hypothesis, we will model how the natural sound statistics are reflected in neurophysiological features of A1 neurons, focusing on three specific aspects.

The first contrast discussed is the locality of receptive fields[1]. In contrast to the receptive fields of V1 neurons that are localised in the visual field, those of A1 neurons are not necessarily localised on the frequency domain. Additionally, recent physiological studies have revealed the presence of "harmonic neurons" in A1 of primates, i.e., neurons that specifically respond to multiple frequencies that are related by simple integer ratios, or harmonically-related [56, 12, 13]. According to the detailed description of such neurons of marmoset monkeys [56], their tuning curves to pure- and two-tone stimuli tended to be modulated (facilitated or inhibited) at harmonically-related frequencies. Those features seem totally different from V1 neurons. Does this suggest that A1 does not share computational principles with V1?

[1] This chapter is based on a published article [121].

We hypothesise that such harmony-related responses are the result of sparse coding of harmonic vocalisations. Sparse coding is a computational model that relates the organisation of the neural response patterns, or the receptive fields, to statistics of natural input signals [87, 88, 89]. Several simulation studies showed that the features optimally learned from natural images had patterns similar to the receptive fields of neurons in V1, suggesting that V1 adapts to natural scene statistics. Since V1 shares basic anatomical features with A1 [101] and the activities of A1 neurons are sparse [46], we hypothesise that the model is also applicable to A1. Previously, we have preliminarily shown that a set of harmony-related responses of A1 neurons reported in the marmoset experiment [56] can be reproduced using a sparse coding algorithm applied to highly harmonic sounds [119, 120]. However, the results were limited because the specific data set we used in that simulation was the recording of a piano performance, which is unlikely to be present in either the evolution or development of monkeys. In the present study, we recorded a set of sounds behaviourally important for marmosets, namely voices of marmosets [133, 132]. We show that the same model applied to such conspecific vocalisations explains the harmony-related responses. Thus, these results provide more direct support for the hypothesis that A1 adapts to natural harmonic sounds, in particular harmonic vocalisations, on the principle of sparse coding. This view might also shed light on the long-held puzzle in auditory research on the qualitative contrast between V1 and A1 regarding receptive field localisation.

5.1.1 Related work

Some studies have discussed the efficient coding strategy in peripheral and cortical auditory systems. Lewicki [68] as well as Smith and Lewicki [114] show that sparse representations of natural sounds result in a set of filters similar to those of cochlea and auditory nerve fibres, suggesting that the peripheral auditory system may adapt to natural sounds in a quite efficient manner. Klein et al. [62] showed that A1 may also encode auditory information using an efficient, sparse coding strategy. They applied independent component analysis to small patches extracted from spectrograms of human speech, and obtained a set of filters whose receptive fields are spectro-temporally localised, similarly to those of A1 neurons. In comparison to our work, the time domain is additionally considered,

though they only replicated basic linear responses to pure-tones, without any reference to the harmony-related responses that we dealt with here in detail.

5.2 Material and methods

This section reviews the sparse coding model [87] and describes the data set used in our simulation.

5.2.1 Sparse coding model

We assume an amplitude spectrum $A(x)$ as an input instance, where x is the frequency. It can be generated by Fourier transformation of a short-time sound waveform, which approximates peripheral auditory processing. This input spectrum corresponds to the original model's image $I(x, y)$.

The model minimises the cost function E:

$$E = \frac{1}{2} \sum_x \left[A(x) - \sum_i a_i \phi_i(x) \right]^2 + \lambda \sum_i |a_i| \tag{5.1}$$

where $\phi_i(x)$ are the real-valued bases or the model neurons indexed by i, and a_i are their activity levels. The first term means that the input can be reconstructed from the activities of the model neurons. Specifically, the input $A(x)$ is reconstructed from a linear superposition of the bases $\phi_i(x)$.

$$A(x) \simeq \sum_i a_i \phi_i(x) \tag{5.2}$$

Figure 5.1A shows a schematic diagram of the reconstruction. Each basis $\phi_i(x)$ corresponds to a model neuron and its coefficient a_i to its degree of activity when presented the input $A(x)$. In what follows, a basis will also be referred to as a unit.

The second term imposes a penalty on the model neuron activity so that its distribution is sparse, as shown in Figure 5.1B. As the first term does not determine the combination of a_i and $\phi_i(x)$ uniquely, the minimisation problem is solved by adding a constraint that

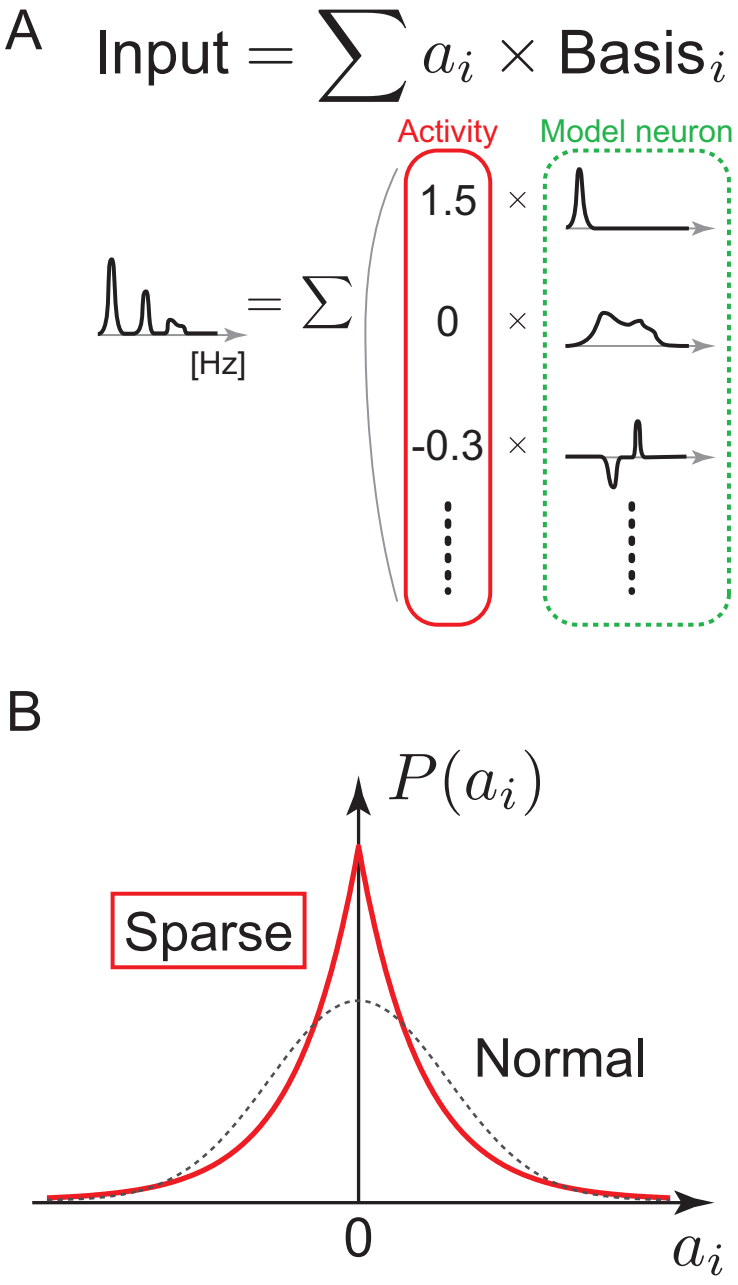


Figure 5.1: **A schematic of sparse coding by A1 neurons.** (*A*) Each input spectrum is reconstructed as a linear superposition of basis vectors. A basis vector and its coefficient represent a model neuron and its activity level, respectively. (*B*) An example of sparse distribution of the coefficient a_i. Compared with a normal distribution with the same variance (dotted line), it has a sharper peak around 0 and longer tails.

the activity levels a_i are mostly zero. λ is the relative weight of the sparseness constraint against the first term.

Specifically, the minimisation problem can be decomposed into two nested iterations. The inner iteration reconstructs the input: given a set of $\phi_i(x)$ and an input $A(x)$, the coefficients a_i are determined under condition (5.2). In particular, the coefficients are determined so that their distribution is sparse as shown in Figure 5.1B, by using gradient descent, where α is the learning rate.

$$\Delta a_i = \alpha \left[\sum_x \phi_i(x) A(x) - \sum_j a_j \sum_x \phi_i(x) \phi_j(x) - \lambda \cdot \mathrm{sign}(a_i) \right]$$

For the rest of the study, the coefficients determined this way is referred to as the activities of the model neurons when a stimulus $A(x)$ is presented. The outer iteration updates the basis set. The solutions obtained in the inner iteration for several inputs are pooled, and the bases $\phi_i(x)$ are updated so that these solutions can represent the inputs $A(x)$ more sparsely. The bases $\phi_i(x)$ are also updated by using gradient descent with the pooled results a_i in the inner iteration.

$$\Delta \phi_i(x) = \eta \left\langle a_i \left[A(x) - \sum_j a_j \phi_j(x) \right] \right\rangle$$

Here, η is the learning rate, and $\langle \cdots \rangle$ is the average over the results of the inner iteration. After each update, the bases are normalised to avoid divergences; specifically, they are scaled so that the average of their squares is a constant value v_{const} (this specific process slightly differs from the original one [87]).

5.2.2 Inputs

The vocalisations of marmoset monkeys (*Callithrix jacchus*) were recorded as biologically plausible sounds. Conspecific vocalisations are behaviourally important, and the same species was used in the neurophysiological experiment [56]. The vocalisations were from ten adults (six males, four females; eighteen to forty months old) in five cages in the animal room and were recorded about two metres away from the cages (44.1 kHz stereo; Olympus

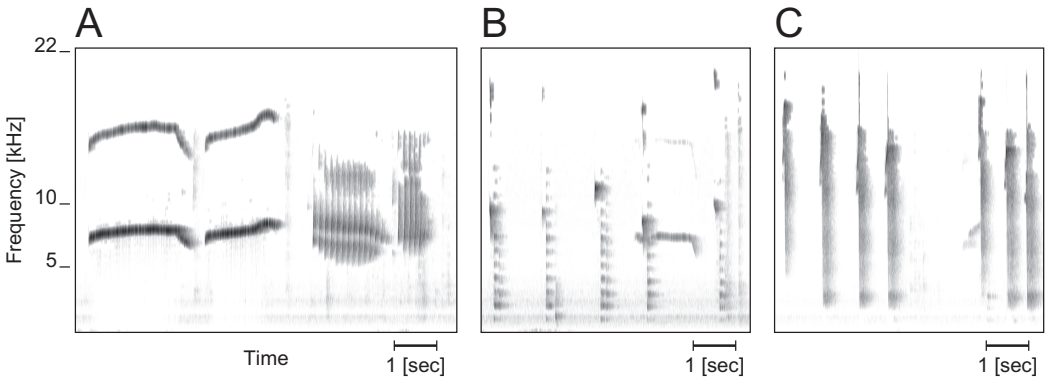


Figure 5.2: **Representative spectrograms of the recorded marmoset vocalisations.** (*A*) Typical vocalisation. The left part consists of two clear peaks at about 7 and 14 kHz, which are related harmonically. (*B*) A low fundamental frequency and its overtones. (*C*) There is a wide range of frequency components present at an instant.

Voice-Trek V-82). The recording included noise originating from the cage contacts, air conditioner, and humidifier. All experimental procedures were approved by the Animal Research Committee, Hirosaki University.

Figure 5.2 shows representative examples of spectrograms of the recorded vocalisations. Figure 5.2A shows typical voices, called phee [54] and twitter [131]. In particular, phee consists of two clear peaks at about 7 and 14 kHz, which are related harmonically. Figure 5.2B and C are vocalisations called egg and chatter, respectively [16]. Note that they each consist of various frequency components at any instant; that is, the vocalisations show correlations between distant frequencies.

Table 5.1: **Sound sources and corresponding parameters**

Sound source	**Sampling frequency**	**# of bases**	**Sample length**
Marmoset vocalisation	44.1 kHz	80	320
Piano performance [2]	44.1 kHz $\rightarrow$ 4 kHz	80	320
Human voice [3]	22 kHz $\rightarrow$ 4 kHz	128	256
Non-harmonic sound [4]	16 kHz $\rightarrow$ 4 kHz	128	256

[2]Recording of a Mozart work: Sonata for Two Pianos in D major, K. 448.

[3]American-English narratives from the Linguistics Handbook of IPA [53]

[4]Non-speech sound dry source (except for musical instruments and electronic sounds) from the RWCP

For comparison, the present study refers to three sound sources that our previous research [119] used as the input. Table 5.1 shows all sound sources including the three previous ones, and their corresponding parameters. The piano performance recording consisted of highly harmonic tones. The human voice consists of intermediately harmonically-related components. The non-harmonic sound was a "non-speech sound dry source" (collision, friction, crush, etc.) from the RWCP Sound Scene Database in Real Acoustical Environments [98], excluding harmonic sounds such as musical instruments and electronic sounds.

For each sound source, the amplitude spectra $A(x)$ ($N = 100,000$) were generated as follows. Sound sources were first bandpassed, monoralised, and downsampled if needed, and then, N segments of the length shown in Table 5.1 were randomly extracted. Next, the segments were converted by fast Fourier transformation into spectra. Finally, the spectra were scaled so that the average of squares of every element value is 1.0.

Note that the marmoset vocalisation was not downsampled, whereas all sound sources were downsampled to 4 kHz in our previous research [119]. Our choice not to downsample in this study was due to a difference in the frequency ranges where harmonic overtones appear [131]. Piano or human voice has rich harmonic components under 2 kHz, while the marmoset voice has ones at $\sim$ 10 kHz, as shown in Figure 5.2. In accordance with this difference, the marmoset's hearing range is shifted to higher frequencies [91].

5.3 Results

The electrophysiological experiment using pure- and two-tone stimuli on marmoset A1 [56] revealed that A1 neurons exhibit the following complex response properties. (1) The tuning curves to the pure-tone stimuli can classify the A1 neurons into single- or multi-peaked units (the frequency of the largest peak is called the characteristic frequency (CF)). (2) The peak frequencies of the multi-peaked units tend to be related harmonically. (3) When presenting two-tone stimuli (CF plus varying second tone), some of the single-peaked units are modulated (facilitated or inhibited) at frequencies that are distant from and harmonically-related to their CFs. Our purpose here is to investigate whether the sparse coding model can reproduce such complex, harmony-related response characteris-

Sound Scene Database in Real Acoustical Environments [98]

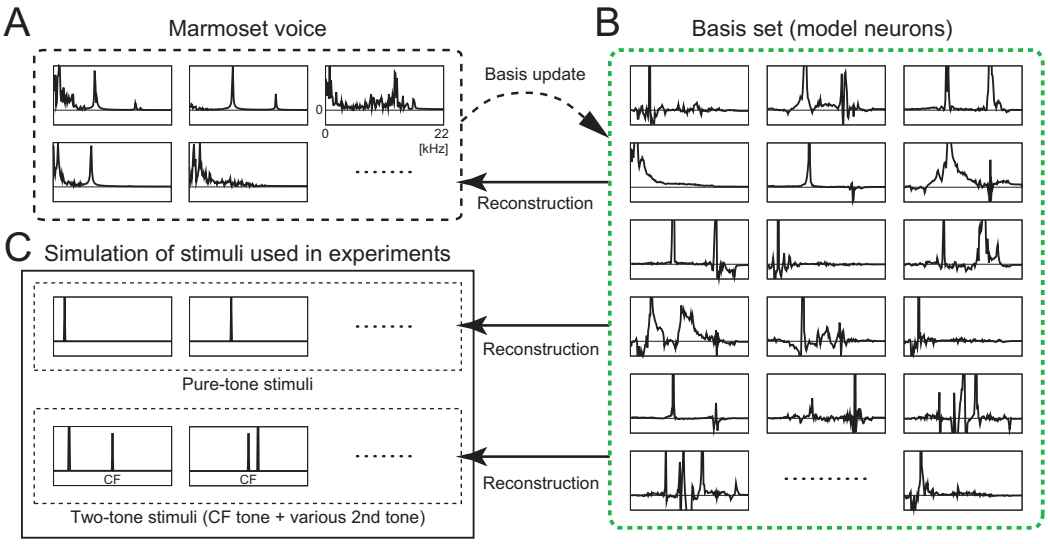


Figure 5.3: **Schematic diagram of learning and analysing model neurons.** (*A*) Input spectra of marmoset voices. (*B*) Bases that adapted to the marmoset voice. Basis learning was done by iteratively performing reconstructions and basis updates (see Section 5.2.1). (*C*) Input spectra of simulated pure- and two-tone (CF plus varying second tone) stimuli. The physiological experiment [56] was simulated by reconstructing those spectra from the bases.

tics, using marmoset vocalisations as input.

5.3.1 Basis learning by adapting to monkey voice

Basis learning was conducted on the model formulated in Section 5.2.1 using the marmoset voice as the input. Figure 5.3A illustrates some input spectra of the conspecific vocalisation. The bases that can efficiently represent the inputs were learned as follows. First, eighty bases were randomly initialised. Next, the input reconstruction (the "inner iteration" described in Section 5.2.1) was iteratively solved using randomly selected inputs, and the unit activity levels (coefficients) were determined. The gradient-descent optimisation was performed; the learning rate was $\alpha = 0.005$, and the termination rule was that the change rate of the cost function E (Equation 5.1) be less than 1%. The basis was updated after every 100 iterations of this input reconstruction; the normalisation constant was $v_{\mathrm{const}} = 0.1$. The basis was updated $\sim 100,000$ times. Here, $\lambda = 0.14$, and η was manually decreased from 1.0 to 0.1.

Figure 5.3B illustrates examples of the obtained bases. We can see they tend to have multiple peaks at harmonically-related frequencies similarly to the inputs. Note that, since the basis forms differ from the inputs, the learning was not just copying of the inputs, but adaptation to them.

5.3.2 Responses to pure-tone stimuli

The marmoset experiment recording neural responses to pure-tone stimuli revealed that (1) A1 neurons can be classified into single- or multi-peaked units, and (2) the peak frequencies of the multi-peaked units tend to be related harmonically. Following the experimental schema, the present study simulated the responses to pure-tones of the model neurons that had adapted to the marmoset voice, and compared them with the experimental results.

First, the responses to various pure-tone stimuli were simulated. Figure 5.3C (top) shows examples of the input spectra $A(x)$ that have a peak at a single frequency and 0 at others. The responses to the pure-tones were determined by the input reconstruction (the "inner iteration" described in Section 5.2.1). Specifically, the results of ten simulations that randomly initialised a_i were averaged, and the CF of each unit [107] was defined as the frequency with the maximum absolute coefficient. Since the activity level a_i should be positive, the coefficient a_i and the basis $\phi_i(x)$ were sign-inverted, if necessary.

Next, based on the obtained tuning curves, the units were classified into the two groups, single- or multi-peaked units. The peak frequencies were defined as the frequencies whose activity was more than $\rho_{\mathrm{peak}}\times$ (response to a pure-tone at CF) and more than the activities at the four nearest frequencies on each side. The multi-peaked units were defined as those units with more than one peak, and the single-peaked units were the others. The peak frequencies of the multi-peaked units were consecutively labelled CF_n ($n = 1, 2, \ldots, N$) from low to high frequency, and their peak frequency ratios were defined as $\mathrm{CF}_n/\mathrm{CF}_1$ ($n = 2, 3, \ldots, N$).

As a result, the single- and multi-peaked units appeared as in the marmoset experiment [56] and our previous research [119]. The proportions of the units were 57% (46/80) single-peaked and 43% (34/80) multi-peaked. The tendency that the single-peaked units were more numerous than the multi-peaked ones was in common with the marmoset experiment (single-peaked: $\sim$80%) and our previous research that used a recorded piano

performance as input (single-peaked: 77%).

Figure 5.4A shows a tuning curve of a single-peaked unit that saliently responded to just one frequency. Its CF was 16.3 kHz, and there were typical lateral inhibitions on both sides of the peak. Even though the tuning curve can be roughly estimated from its basis form, the curve is not identical to the basis form because the activity is determined by sparsification or divisive interaction with other units.

Figure 5.4B shows a tuning curve of a multi-peaked unit that had four peaks. Peak frequencies were 6.6 kHz (CF_1), 9.2 kHz (CF_2), 9.9 kHz (CF_3), and 13.4 kHz (CF_4). Note that $\mathrm{CF}_3/\mathrm{CF}_1 = 1.5$ and $\mathrm{CF}_4/\mathrm{CF}_1 = 2.0$; that is, the peak frequency ratios were harmonic.

As in the experiment paper [56], Figure 5.4C shows the distribution of peak frequency ratios of all multi-peaked units. The ratios tended to concentrate at ratios of simple integers such as 1.5 or 2.0, which shows that the peak frequencies in Figure 5.4B were not harmonically-related by chance. This harmonic bias was similar to the marmoset experiment study, as well as our previous model study using piano performance.

5.3.3 Responses to two-tone stimuli

The marmoset experiment also studied how A1 neurons respond to two-tone stimuli, namely simultaneous presentations of a CF tone and a second tone of varying frequency. Some single-peaked units were modulated (facilitated or inhibited) at frequencies distant from their CFs, and moreover, the modulation frequency and CF tended to be related harmonically.

The present study examined whether the model can reproduce such modulatory responses to simulated two-tone stimuli, that is, the input spectra with two peaks at CF and another frequency (its intensity is 20 dB above that of CF) such as shown in Figure 5.3C (bottom). In the same way as was done in the marmoset experiment, the change rate in activity δ is defined as $(R_2 - R_1)/R_1$, where R_1 and R_2 are the activities when a CF pure-tone or two-tone stimuli are presented, respectively. The frequencies of the CF-adjacent typical inhibitory peaks ($\delta < 0$) are defined as CF_-, CF_+. A distant inhibitory peak is defined as the off-CF ($< \mathrm{CF}_-$ or $> \mathrm{CF}_+$) negative peak whose δ is less than a constant δ_-. Likewise, a distant facilitatory peak is defined as the off-CF positive peak, where $\delta > \delta_+$.

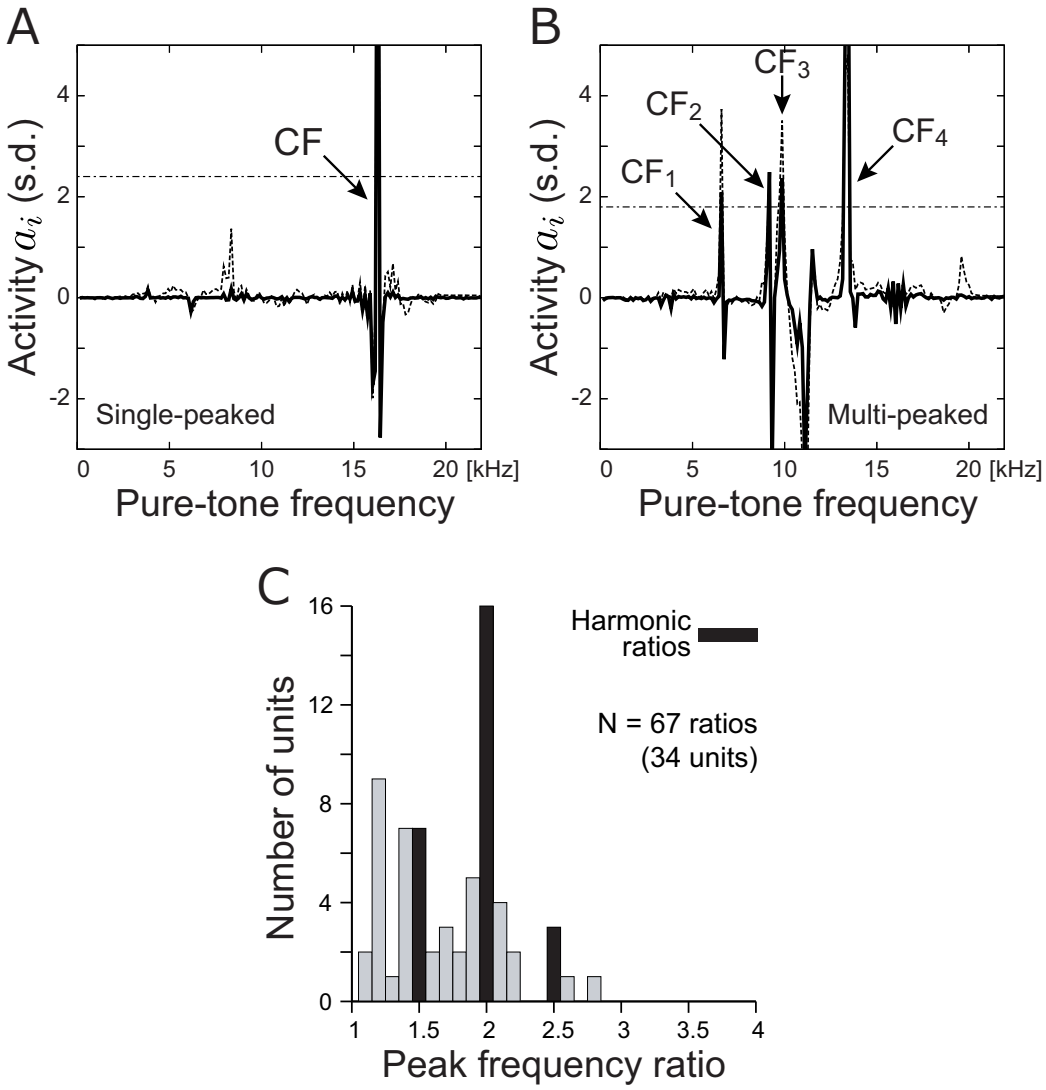


Figure 5.4: **Responses to pure-tone stimuli.** (*A*) Tuning curve of a single-peaked unit (CF = 16.3 kHz). Horizontal dashed line: the peak threshold; Dotted line: the basis form. (*B*) Tuning curve of a multi-peaked unit (CF = 13.4 kHz). Note that the peak frequencies $\mathrm{CF}_1, \mathrm{CF}_3, \mathrm{CF}_4$ are harmonically-related. (*C*) Distribution of peak frequency ratios of multi-peaked units. The ratios tend to concentrate on integer multiples of 0.5 (filled black). Ratios more than 4.0 are excluded. $\rho_{\mathrm{peak}} = 0.2$.

The best inhibitory or facilitatory peak of each unit is the inhibitory or facilitatory peak with the largest modulation.

As a result, in the current simulation, some of the single-peaked units exhibited off-CF response modulations, that is, distant facilitation or inhibition. Facilitation and inhibition were seen in 37% (17/46) and 26% (12/46) of the single-peaked units, respectively; 17% (8/46) had both of facilitation and inhibition. These results were in accordance with our previous results using piano performance recordings (47%, 27%, 11%) and the marmoset experiment (45%, 45%, 24%).

Figure 5.5A shows a tuning curve of a single-peaked unit that had an off-CF facilitatory peak. Its CF was 15.6 kHz, accompanying the CF-adjacent inhibitory peaks. In addition, the unit activity was facilitated by a second tone at 19.3 kHz, which is distant from the CF.

Figure 5.5B shows a tuning curve of another single-peaked unit that had an off-CF inhibitory peak. The inhibitory modulation appeared at 15 kHz, which is double the frequency of its CF, 7.5 kHz. Note that the frequency of the inhibitory modulation is harmonically-related to the CF.

The relation between CF and the modulatory second frequency is illustrated by Figure 5.5C, D, E, which show the distributions of peak frequency ratios of best facilitation, best inhibition, and all modulations (any facilitation or inhibition). All distributions tended to concentrate at harmonic ratios such as 0.5 or 2.0. This was also the case in the marmoset experiment, although the experiment reported the tendency was more salient in inhibition than in facilitation. Our previous study also yielded a similar result using piano performance.

Finally, the proportions of harmonic frequency ratios were calculated and compared with the marmoset experiment [56], as well as our previous results [119] (Figure 5.6). Specifically, proportions of integer multiples of 0.5 were calculated within two categories, all peaks (Figure 5.4C and Figure 5.5E) and best modulation of single-peaked units (Figure 5.5C and Figure 5.5D). The proportion of all peaks was 46% (46/99), and the best modulation was 52% (15/29). The ratios in this study tended to be harmonic, as was the case with the marmoset experiment (50%, 40%) and our previous study using a piano performance (39%, 46%), and their proportions were quantitatively comparable.

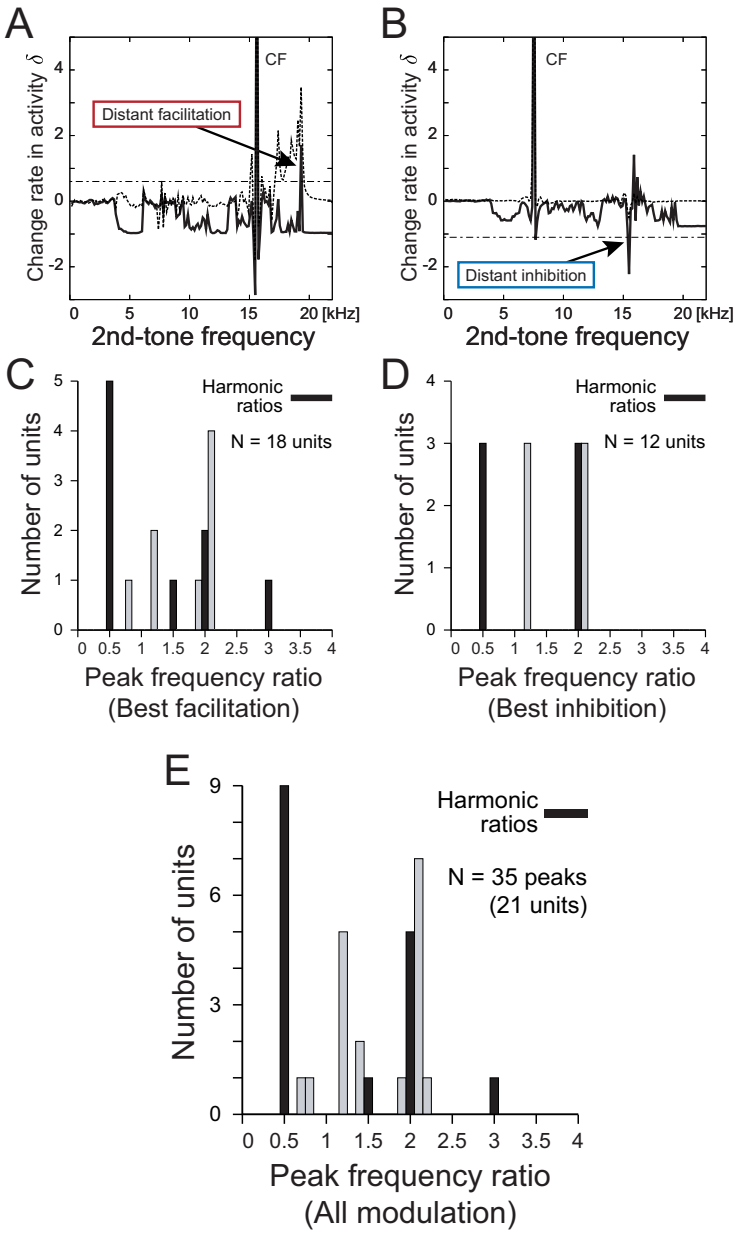


Figure 5.5: **Responses of single-peaked units to two-tone stimuli.** (*A*) Tuning curve of a single-peaked unit with distant facilitation. Horizontal dashed line: $\delta_+ = 0.6$. (*B*) Tuning curve with distant inhibition. Horizontal dashed line: $\delta_- = -1.1$. (*C*) Distribution of ratios of best facilitatory peak frequencies to CFs. (*D*) Distribution of best inhibitory peak frequency ratios. (*E*) Distribution of all modulatory (facilitatory and inhibitory) peak frequency ratios. Black bars: harmonic peak frequency ratios (integer multiples of 0.5).

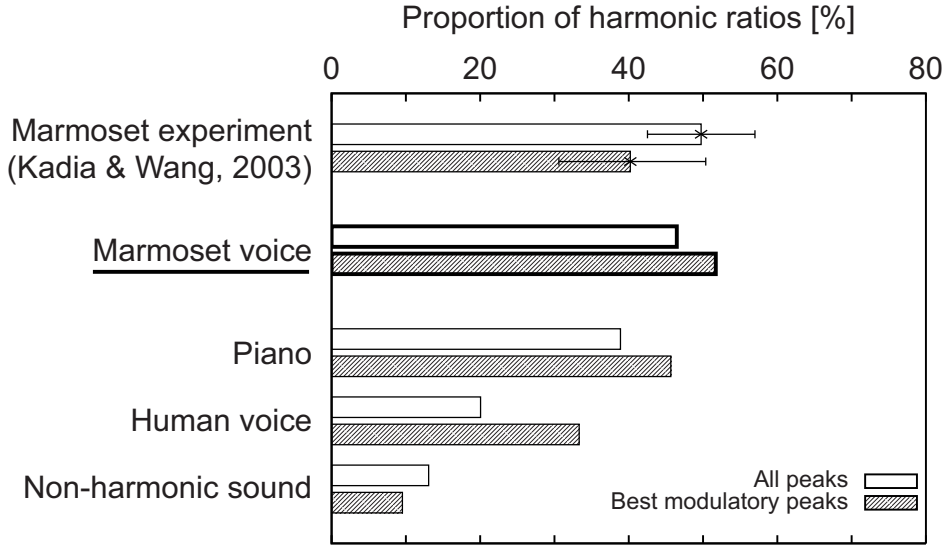


Figure 5.6: **Proportions of harmonic frequency ratios** (integer multiples of 0.5). The proportion in this study with the marmoset voice was comparable to that of the marmoset experiment or our previous study with the piano performance. For the comparison, the results for the piano performance, human voice, and non-harmonic sound were taken from [119], and the result of marmoset experiment was calculated from [56]. The "all peaks" result was calculated from Figure 5.4C and 5.5E; "Best modulatory peaks": Figure 5.5C and 5.5D; Horizontal bars: 95% confidence intervals.

5.4 Discussion

In this chapter, we hypothesise that the harmonically-related receptive fields of A1 neurons [56] may emerge from sparse coding of harmonic sound. Previously, we provided a simulation result supporting this hypothesis [119, 120]; however, the specific input sound we used was not biologically plausible. In the present study, we recorded marmoset vocalisations and used them as the input to the same model. The recorded conspecific voice comprised rich harmonic overtones (Figure 5.2, 5.3A). Learning an efficient representation of the voice resulted in model neurons, or bases, with multiple peaks whose frequencies were harmonically-related (Figure 5.3B). Next, their responses to pure- and two-tone stimuli were simulated following the experimental schema [56] (Figure 5.3C). As a result, the model reproduced the repertoire of harmony-related responses: (1) the units were classified into two groups, namely single- or multi-peaked (Figure 5.4A, B), (2) peak frequencies of the multi-peaked units tended to be related harmonically (Figure 5.4C),

(3) single-peaked units were modulated by off-CF two-tone stimuli whose second frequency tended to be harmonically-related to CF (Figure 5.5). Moreover, the proportion of harmonic peak frequency ratios was high and quantitatively comparable to those of the marmoset experiment (Figure 5.6).

Overall, the results were qualitatively similar to the marmoset experiment [56]. In a sense, it may be obvious that supplying stimuli which have harmonic relations will result in coding systems which are sensitive to these; however, it is not obvious that not only piano sounds but also the marmoset voices including background noise will have such property. The use of marmoset vocalisations would give a stronger and more direct support for the hypothesis that the repertoire of harmony-related A1 receptive fields may result from adapting to natural harmonic sounds such as conspecific vocalisations, through an efficient coding strategy [10] similarly to V1.

It is important for social animals to capture the harmonic structure in natural sounds with their neural activities, since behaviourally important sounds often consist of rich harmonic overtones due to their production mechanism. Such harmonic overtones are well known in vocalisations of mammals such as mice [31], cats [92], marmoset monkeys [16], and humans [53]. In association with the harmonics, recent studies have shown that the auditory cortices specifically respond to conspecific vocalisations [133, 132, 93] and that the acquisition of harmony-related responses requires exposure to conspecific vocalisations during infancy [125]. Our study has modelled how the sound statistics in the environment affect neural responses, and thus, our model would be a useful tool to explore how tuning curves change in subsequent experiments on environmental switches. Specifically, this study suggests that the degree of such modulations at harmonically-related frequencies in other species may correlate with degree of harmonic relations in their vocalisations. For instance, Figure 5.6 suggests that the degree of harmonic modulation of human A1 neurons may be intermediate or lower than that of marmoset neurons.

One of the characteristics of A1 neurons is that their receptive fields are not necessarily localised in the frequency domain or the hearing range [107, 56]: A1 neurons' activity can be modulated by frequencies that are distant from their CFs, as seen in the multi-peaked units (Figure 5.4B). This non-localised receptive field of A1 contrasts with that of V1, which is strictly spatially localised in the eye field [48]; however, the contrastive feature can appear through a common adaptive algorithm, because they can be explained by an

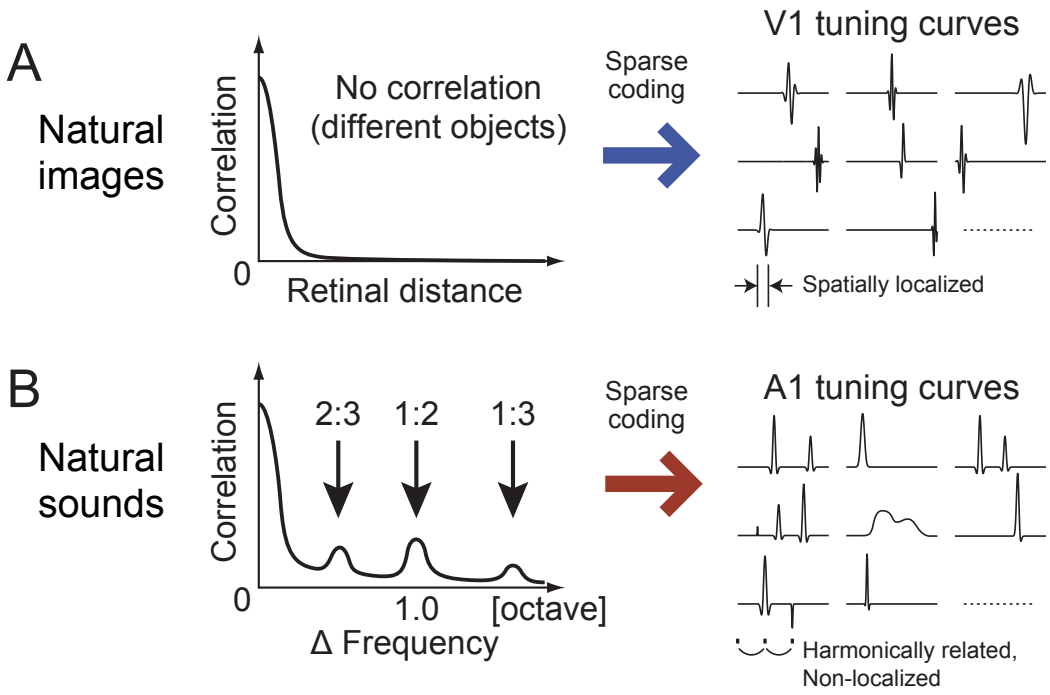


Figure 5.7: **Suggested relationship between natural stimulus statistics and the locality of receptive fields.** (*A*) In the visual environment, since distant eye fields usually show different objects, visual stimuli correlate only locally. Adapting to such image statistics results in the localised receptive fields of V1 neurons. (*B*) In the auditory environment, sound statistics shows correlation not only locally, but with distant frequencies. In particular, those distant frequencies are often harmonically-related in behaviourally important sounds such as animal vocalisations. The A1 receptive fields are not localised, because the A1 neurons need to integrate information embedded in the distant frequencies.

identical model, sparse coding. In other words, the apparent dissimilarity may simply reflect different statistics of natural images and sounds. As shown in Figure 5.7A, natural images of the visual environment strongly correlate locally, and the correlations rapidly decrease with increasing retinal distance. In contrast, Figure 5.7B illustrates that sound statistics from the auditory environment exhibit correlation not only locally, but also between distant frequencies. Natural sound tends to contain various frequency components at any instant, and moreover, due to the physical production mechanism, those frequencies are often harmonically-related in behaviourally important sounds such as animal vocalisations (Figure 5.2). We suggest that this statistical discrepancy is responsible, despite the

common adaptive algorithm, for the dissimilar receptive fields of V1 and A1 (Figure 5.7).

In order to capture the statistics of natural sound, auditory neurons need to integrate information conveyed by distant frequencies, which may be initially done at A1 in the hierarchical auditory system. The integration process might be underpinned by the anatomical specificity of A1, namely a large number of long-range connections [97] that in the process of plasticity play a different role from those in V1 [64]. Our results provide an interpretation for the dissimilarity between V1 and A1: the difference may be due to the different stimulus statistics. This idea dovetails with previous experiments that exchanged peripheral inputs of the visual and auditory system [117, 2, 111], suggesting that at least the two cortex areas share a fundamental function [101] and that the function and the inputs define the direction of development. Moreover, although this study has only discussed the receptive fields of individual neurons, its conclusions might be also applicable to the neural population. Previous theoretical studies on V1 suggested that the formation of the receptive field of each neuron is related to their spatial arrangement, that is, the map formation [49, 52, 90, 59, 60]. On the other hand, recent physiological studies have revealed that V1 maps are very smooth at a single-cell precision [84, 85, 115], whereas the A1 map is disordered [103, 8]. This discrepancy might also be accounted for by our view that V1 and A1 similarly adapt to natural stimulus statistics, which will be discussed in the next chapter.

5.5 Conclusion

We previously showed that the harmonically-related responses of marmoset A1 neurons can emerge by sparse coding of harmonic sound, but the specific sound we used was biologically implausible. In the present study, we recorded voices of marmosets, which consists of rich harmonic overtones, and used the conspecific vocalisation as the input to the model. Overall, the model reproduced results qualitatively similar to those of the marmoset experiment. Its results therefore more directly support the hypothesis that A1 adapts to the natural stimulus statistics in the same way as V1. In addition, this view provides an interpretation for the qualitative dissimilarity of V1 and A1 regarding receptive field localisation: it may emerge from discrepancies between the natural stimulus statistics of vision and audition.

Chapter 6

Topographic map

> Jamais la nature ne nous trompe;
> c'est toujours nous qui nous trompons.
>
> JEAN-JACQUES ROUSSEAU

The previous chapter modelled the receptive field of individual cells, whose arrangement in the physical space was not considered. In this chapter, we will discuss spatial arrangement of those cells known as the topograhic map[1].

6.1 Smooth V1 map vs scattered A1 map

Despite of the anatomical and functional similarities between A1 and V1, the computational modelling of A1 has proven to be less fruitful than V1, primarily because the responses of A1 cells are more disorganized. For instance, the receptive fields of V1 cells are localized within a small portion of the field of view [47], whereas certain A1 cells have receptive fields that are not localized, as these A1 cells demonstrate significant responses to multiple distant frequencies [107, 56]. An additional discrepancy that has recently been discovered between these two regions relates to their topographic structures, i.e., the retinotopy of V1 and the tonotopy of A1; these structures had long been considered to be quite similar, but studies on a microscopic scale have demonstrated that in mice, the tonotopy of A1 is much more disordered [8, 103] than the retinotopy of V1 [115, 17]. This

[1] Preliminary results of this chapter were published in [122].

result is consistent with previous investigations involving other species [32, 41], suggesting that the discrepancy in question constitutes a general tendency among mammals. This disorderliness appears to pose significant difficulties for the development of computational models of A1.

A number of computational modelling studies have emphasized the close associations between V1 cells and natural image statistics, which suggests that the V1 adopts an unsupervised, efficient coding strategy [51]. For instance, the receptive fields of V1 simple cells were reproduced by either sparse coding [87] or the independent component analysis [11] of natural images. This line of research yields explanations for the two-dimensional topography, the orientation and retinotopic maps of V1 [49, 50, 70]. Similar efforts to address A1 have been attempted by only a few studies, which demonstrated that the efficient coding of natural, harmonic sounds, such as human voices or piano recordings, can explain the basic receptive fields of A1 cells [62, 106] and their harmony-related responses [119, 121]. However, these studies have not yet addressed the topography of A1.

In an integrated and computational manner, the present study attempts to explain why the tonotopy of A1 is more disordered than the retinotopy of V1. We hypothesized that V1 and A1 still share an efficient coding strategy, and we therefore proposed that the distant correlations in natural sounds would be responsible for the relative disorder in A1. To test this hypothesis, we first demonstrated the significant differences between natural images and natural sounds. Natural images and natural sounds were then each used as inputs for topographic independent component analysis, a model that had previously been proposed for the smooth topography of V1, and maps were generated for these images and sounds. Due to the distant correlations of natural sounds, greater disorder was observed in the learned map that had been adapted to natural sounds than in the analogous map that had been adapted to images. For natural sounds, this model predicted harmonic relationships between neighbouring cells. These results suggest that the apparently dissimilar topographies of V1 and A1 may reflect statistical differences between natural images and natural sounds; however, these two regions may employ a common adaptive strategy.

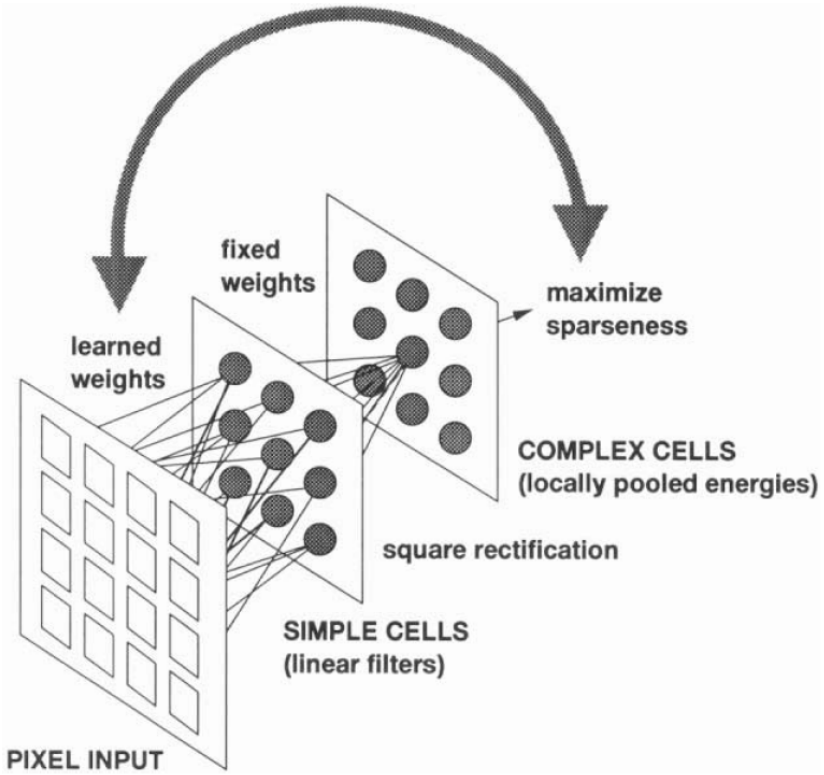


Figure 6.1: **Schematic of the model architecture of TICA**. The TICA model consists of the input layer and two layers of neurons, each of which models simple cells and complex cells of V1, respectively. The filters that characterises responses of simple cells are learned by maximising the sparseness of the complex cells' activities. (The image was adapted from [49].)

6.2 Methods

6.2.1 Topographic independent component analysis (TICA)

Herein, we discuss an unsupervised learning model termed topographic independent component analysis (TICA), which was originally proposed for the study of V1 topography [49, 50]. This model comprises two layers: the first layer of N units models the linear responses of V1 simple cells, whereas the second layer of N units models the nonlinear responses of V1 complex cells, and the connections between the layers define a topography (Figure 6.1). Given a whitened input vector $\boldsymbol{I}(x) \in \mathbb{R}^d$ (here, $d = N$), the input is reconstructed by the linear superposition of a basis $\boldsymbol{a}_i \in \mathbb{R}^d$, each of which corresponds to the first-layer units

$$\boldsymbol{I} = \sum_i s_i \boldsymbol{a}_i \tag{6.1}$$

where $s_i \in \mathbb{R}$ are activity levels of the units or model neurons. Inverse filters $\boldsymbol{w}_i$ to determine s_i can typically be obtained, and thus $s_i = \boldsymbol{I}^T \boldsymbol{w}_i$ (inner product). Using the activities of the first layer, the activities of the second-layer units $c_i \in \mathbb{R}$ can be defined as follows:

$$c_i = \sum_j h(i,j) s_j^2 \tag{6.2}$$

where $h(i,j)$ is the neighbourhood function that takes the value of 1 if i and j are neighbours and is 0 otherwise. The neighbourhood is defined by a square window (e.g., 5×5) in cases of two-dimensional topography. The learning of $\boldsymbol{w}_i$ is accomplished through the minimisation of the energy function E or the negative log likelihood:

$$E = -\log L(\boldsymbol{I}; \{\boldsymbol{w}_i\}) = -\sum_i G(c_i) \tag{6.3}$$

$$\Delta \boldsymbol{w}_i \propto \left\langle \boldsymbol{I} s_i \left(\sum_j h(i,j) g(c_j) \right) \right\rangle \tag{6.4}$$

where $G(c_i) = -\sqrt{\epsilon + c_i}$ imposes sparseness on the second-layer activities ($\epsilon = 0.005$ for the stability), and $g(c_i)$ is the derivative of $G(c_i)$. The operator $\langle \cdots \rangle$ is the mean over the iterations.

6.2.2 An extension for overcomplete representation

Ma and Zhang [70] extended the TICA model to account for overcomplete representations ($d < N$), which are observed in the V1 of primates. In this extension, inverse filters cannot be uniquely defined; therefore, a set of first-layer responses s_i to an input is computed by minimizing the following extended energy function:

$$E = -\log L(\boldsymbol{I}; \{\boldsymbol{a}_i\}, \{s_i\}) = \left| \boldsymbol{I} - \sum_i s_i \boldsymbol{a}_i \right|^2 - \lambda \sum_i G(c_i) \tag{6.5}$$

$$\Delta s_i \propto \boldsymbol{a}_i^T \left(\boldsymbol{I} - \sum_j s_j \boldsymbol{a}_j \right) - \lambda s_i \left(\sum_j h(i,j) g(c_j) \right) \tag{6.6}$$

where λ is the relative weight of the activity sparseness, in accordance with sparse coding [87]. The initial value of s_i is set equal to the inner product of $\boldsymbol{I}$ and $\boldsymbol{a}_i$. Every 256 inputs, the basis is updated using the following gradient. In this study, we used the learning rate $\eta = 0.08$.

$$\Delta \boldsymbol{a}_i = \eta \left\langle s_i \left(\boldsymbol{I} - \sum_j s_j \boldsymbol{a}_j \right) \right\rangle \tag{6.7}$$

6.2.3 The discontinuity index for topographic representation

To compare the degrees of disorder in topographies of different modalities, we defined a discontinuity index (DI) for each point i of the maps. Features defining a topography $f(i)$ (e.g., a retinotopic position or a frequency) were normalized to the range of $[0, 1]$. Features $f(j)$ within the neighbourhood of the ith unit defined by $h(i,j)$ were linearly fitted using the least squares method, and the DI value at i was then determined using the following equation:

$$\mathrm{DI}(i) = \sqrt{\frac{\sum_j h(i,j) r(j)^2}{N_{\mathrm{NB}}}} \tag{6.8}$$

where $r(j)$ is the residual error of linear regression at j and N_{NB} is the number of units within a neighbourhood window. If the input space is a torus (see Section 6.3.2), another DI value is computed using modified f values that are increased by 1 if they were initially within $[0, \frac{1}{2})$, and the smaller of the calculated DI values is used.

6.3 Results

6.3.1 Greater disorder for the tonotopy than the retinotopy

To test the hypothesis that V1 and A1 share a learning strategy, the TICA model was applied to natural images and natural sounds, which exhibit different statistical profiles, as discussed above. Learning with natural images was accomplished in accordance with the original TICA study [49, 50]. Images from the van Hateren database were reduced four times from their original size, and 25×25 pixel image patches were randomly extracted (n = 50,000). The patches were whitened and bandpassed by applying principal component analysis, whereby we selected 400 components and rejected certain components with low variances and the three components with the largest variances [49]. The topography was a 20×20 torus, and the neighbourhood window was 5×5.

Figure 6.2 illustrates the visual topographic map obtained from this analysis, a small square of which constitutes a basis vector $\boldsymbol{a}_i$. As previously observed in the original TICA study [49, 50], each unit was localized, oriented, and bandpassed; thus, these units appeared to be organized similarly to the receptive fields of V1 simple cells. The orientation and position of the units changed smoothly with the coordinates that were examined, which suggested that this map evinces an ordered topography. To quantify the retinotopic discontinuity, each unit was fitted using a two-dimensional Gabor function, and DI was calculated using the y values of the centre coordinates of the resulting Gabor functions as the features. Figure 6.4 graphically indicates that the obtained DI values were quite low, which is consistent with the smooth retinotopy illustrated in Figure 6.5A.

Next, another TICA model was applied to natural sounds to create an auditory topographic map that could be compared to the visual topography. As detailed in the previous section, spectrograms of human voices (sampled at 8 kHz) were generated using the NSL toolbox to approximate peripheral auditory processing. Spectrogram patches of 200 ms (25 pixels) in width were randomly extracted (n = 50,000) and vertically reduced from 128 to 25 pixels, which enabled these spectrogram patches to be directly compared with the image patches. The sound patches were whitened, bandpassed, and adapted using the model in the same manner as was described for the image patches.

Figure 6.3 shows the resulting auditory topographic map, which consists of spectro-

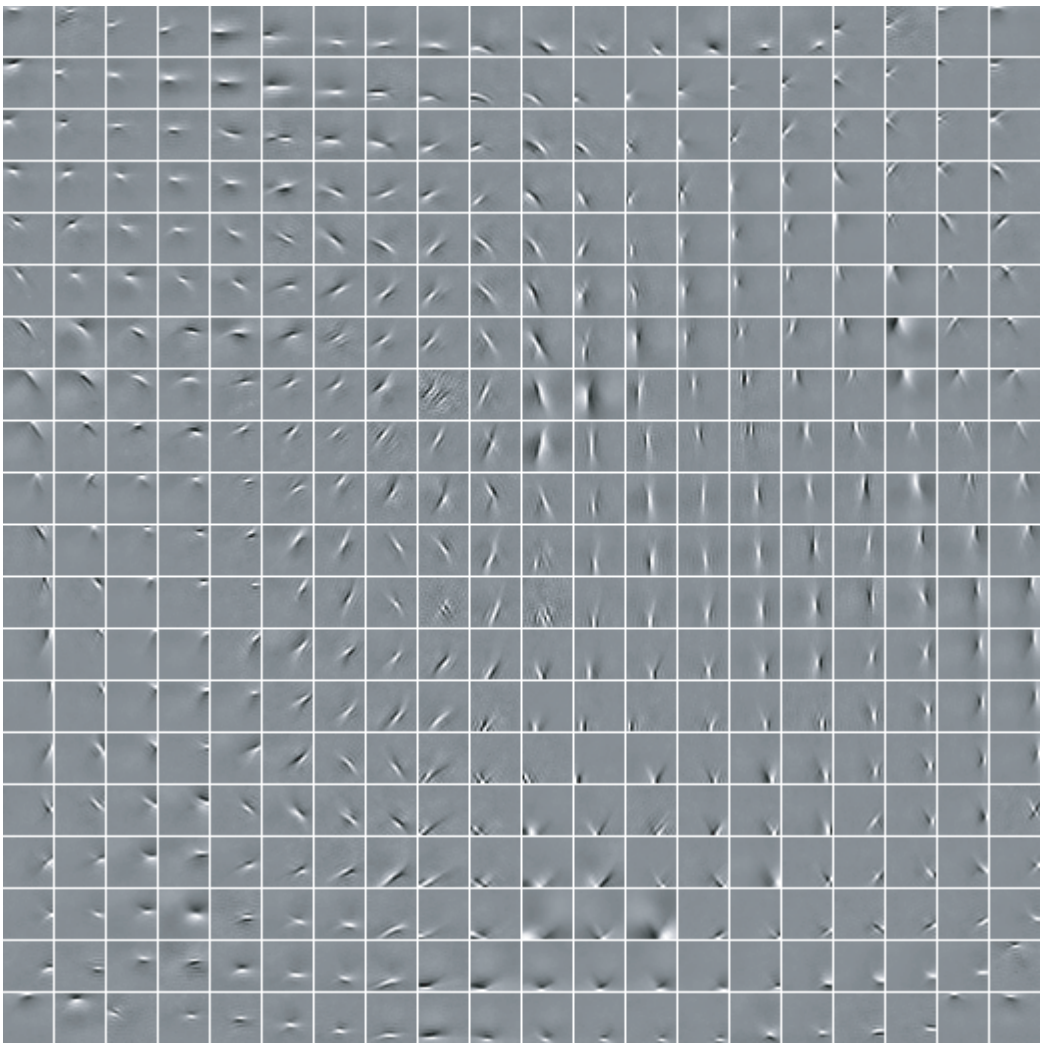


Figure 6.2: **The smooth topographic map adapted to natural images.** The topography of units adapted to natural images. A small square indicates a unit $\boldsymbol{a}_i$ (i.e., the receptive field of unit i) (grey: 0; white: max value).

temporal units of $\boldsymbol{a}_i$ that are represented by small squares. The units were localized temporally and spectrally, and some units demonstrated multiple, harmonic peaks; thus, these units appeared to reasonably represent the typical spectro-temporal receptive fields of A1 cells [62, 56]. The frequency to which an auditory neuron responds most significantly is called its characteristic frequency (CF) [107]. In this analysis, the CF of a unit was defined as the frequency that demonstrated the largest absolute value for the unit in question.

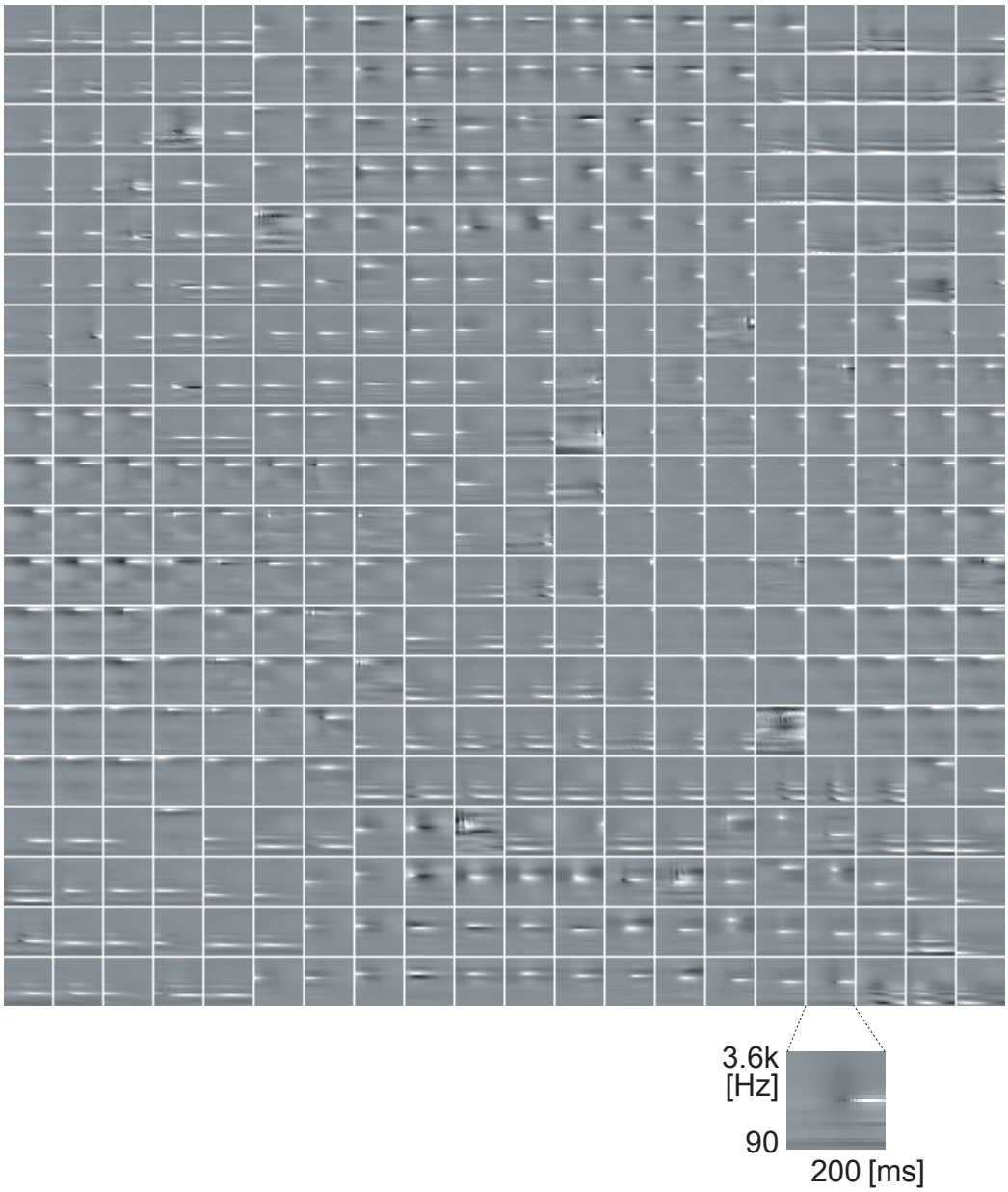


Figure 6.3: **The disorganised topographic map adapted to natural sounds.** The topography of spectro-temporal units that have been adapted to natural sounds (grey: 0; white: max value). It has some local scatter, i.e., nearby units can be selective to distant frequencies.

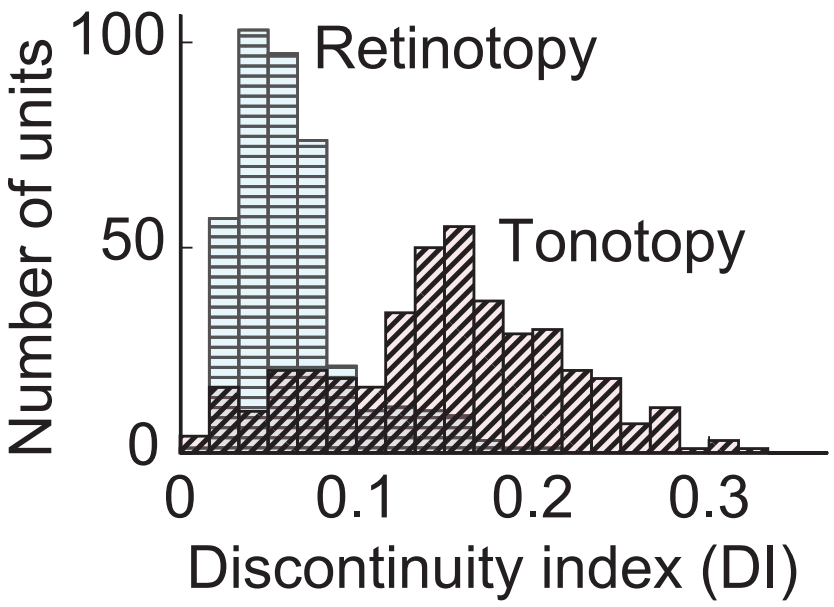


Figure 6.4: **Distributions of DI for the visual and auditory topographies.** The tonotopy is more disordered than the retinotopy. They were calculated using Figure 6.3 and Figure 6.2, respectively.

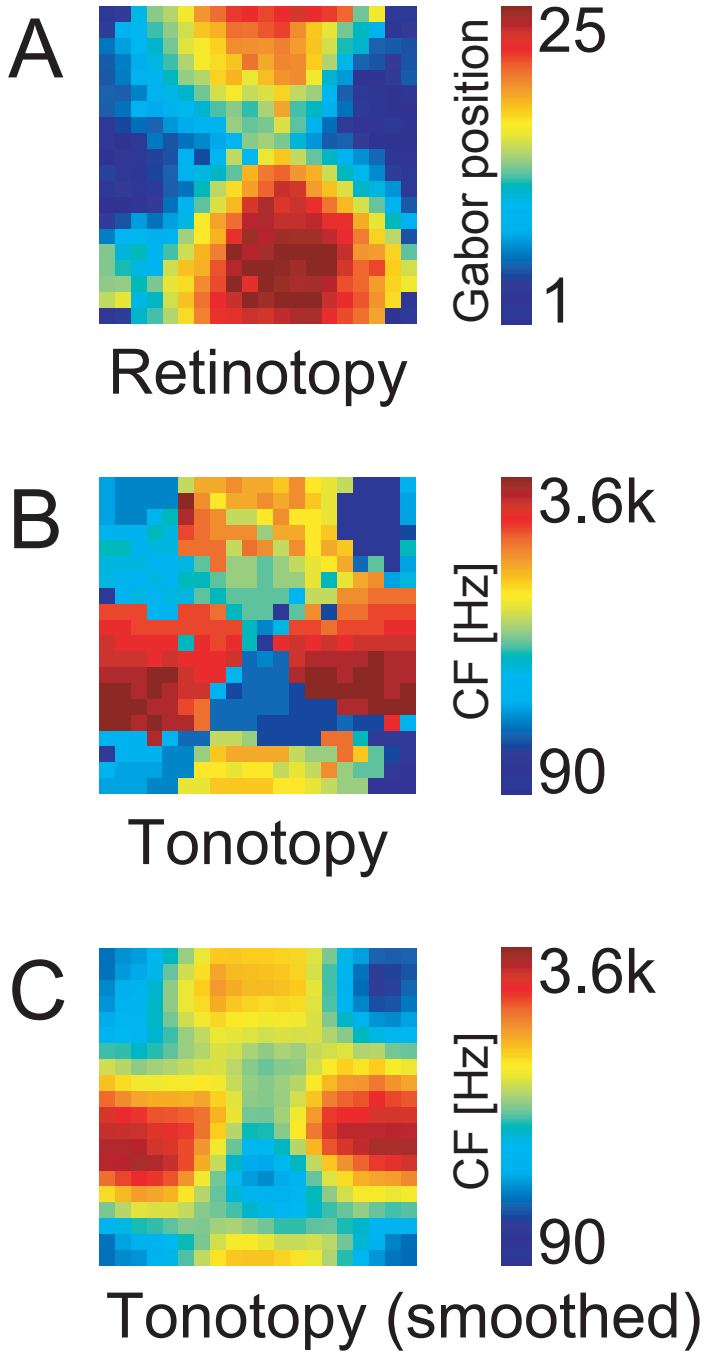


Figure 6.5: **The ordered retinotopy and disordered tonotopy.** (*A*-*C*) The retinotopy of the visual map (A) is smooth, whereas the tonotopy of the auditory map (B) is more disordered, although global tonotopy still exists (C).

Figure 6.5B illustrates the spatial distribution of CFs, i.e., the tonotopic map. Within local regions, the tonotopy was not necessarily smooth, i.e., neighbouring units displayed distant CFs. However, at a global level, a smooth tonotopy was observed (Figure 6.5C). Both of these findings are consistent with established experimental results [8, 103]. The distribution of tonotopic DI values is shown in Figure 6.4, which clearly demonstrates that the tonotopy was more disordered than the retinotopy ($p < 0.0001$; Wilcoxon rank test).

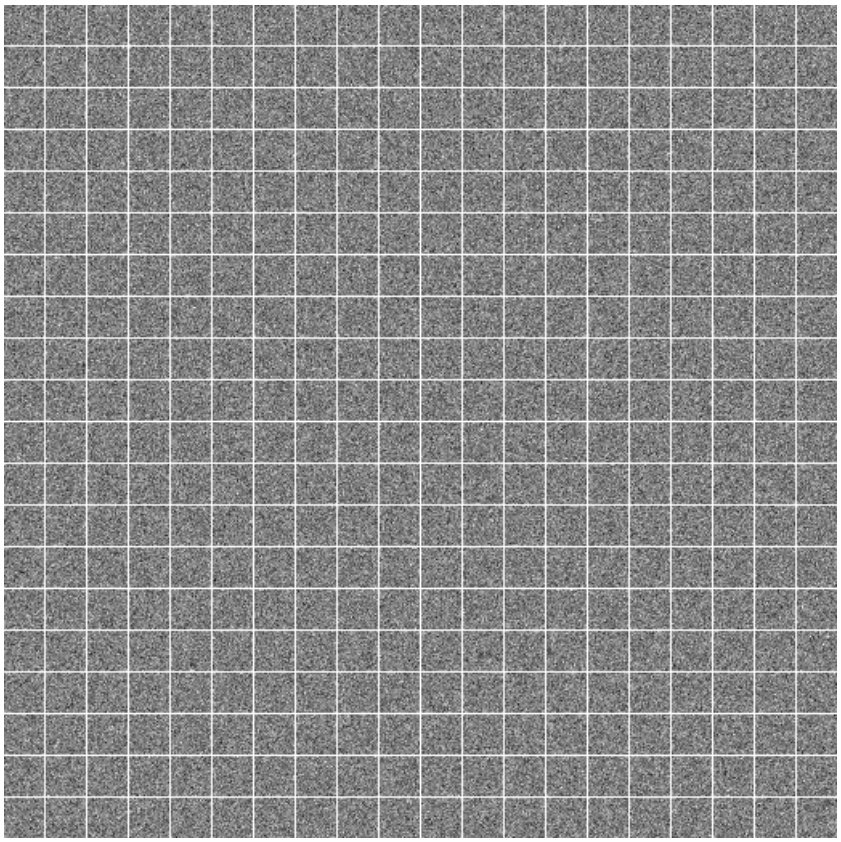


Figure 6.6: **The control topographic map adapted to random stimuli.** The topography learned with random white noise stimuli shows no clear structures.

Figure 6.6 shows the control result for random stimuli. Pixel values of the two-dimensional stimuli were independent white noise, which were used as the input to the same model with the same parameters. Learning of the input resulted in a topography that has no clear structures.

6.3.2 The topographic disorder due to distant input correlations

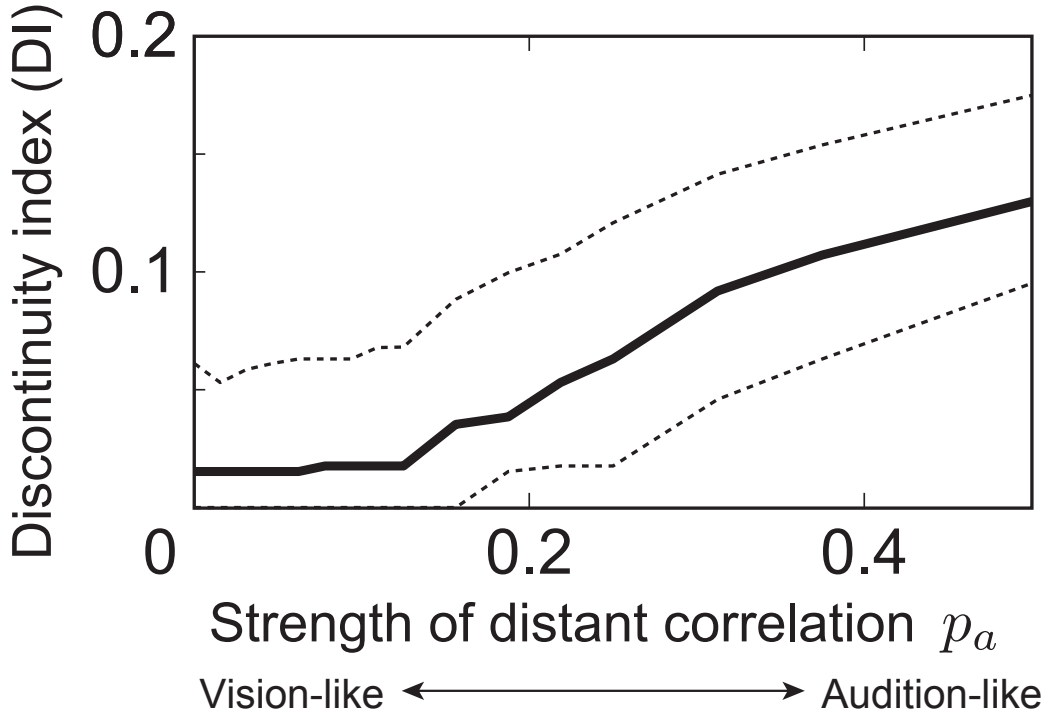


Figure 6.7: **The correlation between discontinuity and input "auditoriness".** When inputs only correlated locally ($p_a \sim 0$: vision-like inputs), DI was low, and DI increased with the input "auditoriness" p_a. Three lines: the quartiles (25, 50 (bold), and 75%) obtained from 100 iterations.

The previous section demonstrated that natural sounds could induce greater topographic disorder than natural images, and this section discusses the attempts to elucidate the disorder resulting from a specific characteristic of natural sounds, namely, distant correlations. For this purpose, we generated artificial inputs ($d = 16$) with a parameter $p_a \in [0, 1]$ that regulates the degree of distant correlations. After the inputs were initially generated from a standard normal distribution, a constant value of 4 was added at k points of each input, where k was from a uniform distribution over $\{3, 4, 5, 6\}$ and the points' coordinates x were from a normal distribution with a random centre and $\sigma = 2$. After adding this constant value at x, we also added another at $x_{dist} = x + 5$ with a probability p_a that defines its "auditoriness", i.e., its degree of distant correlations. For greater simplicity and to avoid border effects, the input space was defined to be a one-dimensional torus.

The topography was also set as a one-dimensional torus of 16 units with a neighbourhood window size of 5.

Figure 6.7 shows the positive correlation between the input "auditoriness" p_a and the DI of the learned topographies. In computations of DI, the feature f of a unit was considered to be its peak coordinate with the largest absolute value, and a toric input space was used (Section 6.2.3). If the input only demonstrated local correlations like visual stimuli ($p_a \sim 0$), then its learned topography was smooth (i.e., its DI was low). The DI values generally increased as distant correlations appeared more frequently, i.e., more "auditoriness" of the inputs grew. Thus, the topographic disorder of auditory maps results from distant correlations presented by natural auditory signals.

6.3.3 The harmonic relationship among neighbouring units

Several experiments [8, 103] have reported that the CFs of neighbouring cells can differ by up to 4 octaves, although these studies have failed to provide additional detail regarding the local spatial patterns of the CF distributions. However, if the auditory topography is representative of natural stimulus statistics, the topographic map is likely to possess certain additional spatial features that reflect the statistical characteristics of natural sounds.

To enable a detailed investigation of the CF distribution, we employed a model that had been adapted to finer frequency spectra of natural sounds, and this model was then used throughout the remainder of the study. As the temporal structure of the auditory receptive fields was less dominant than their spectral structure (Figure 6.5C), we focused solely on the spectral domain and did not attempt to address temporal information. Therefore, the inputs for the new model (n = 100,000) were short-time frequency spectra of 128 pixels each (24 pixels = 1 octave). The data for these spectra were first obtained from the spectrograms of human voices (8 kHz) using the method detailed in Chapter 2, and these data were then whitened, bandpassed, and reduced to 100 dimensions prior to input into the model. To illustrate patterns more clearly, the results shown below were obtained using the overcomplete extension of TICA described in Section 6.2.2, which included a 14×14 torus (approximately $2\times$ overcomplete) and 3×3 windows. The CF of a unit was determined using pure-tone inputs of 128 frequencies.

Figure 6.8A illustrates the full distribution of the distance and CF difference between

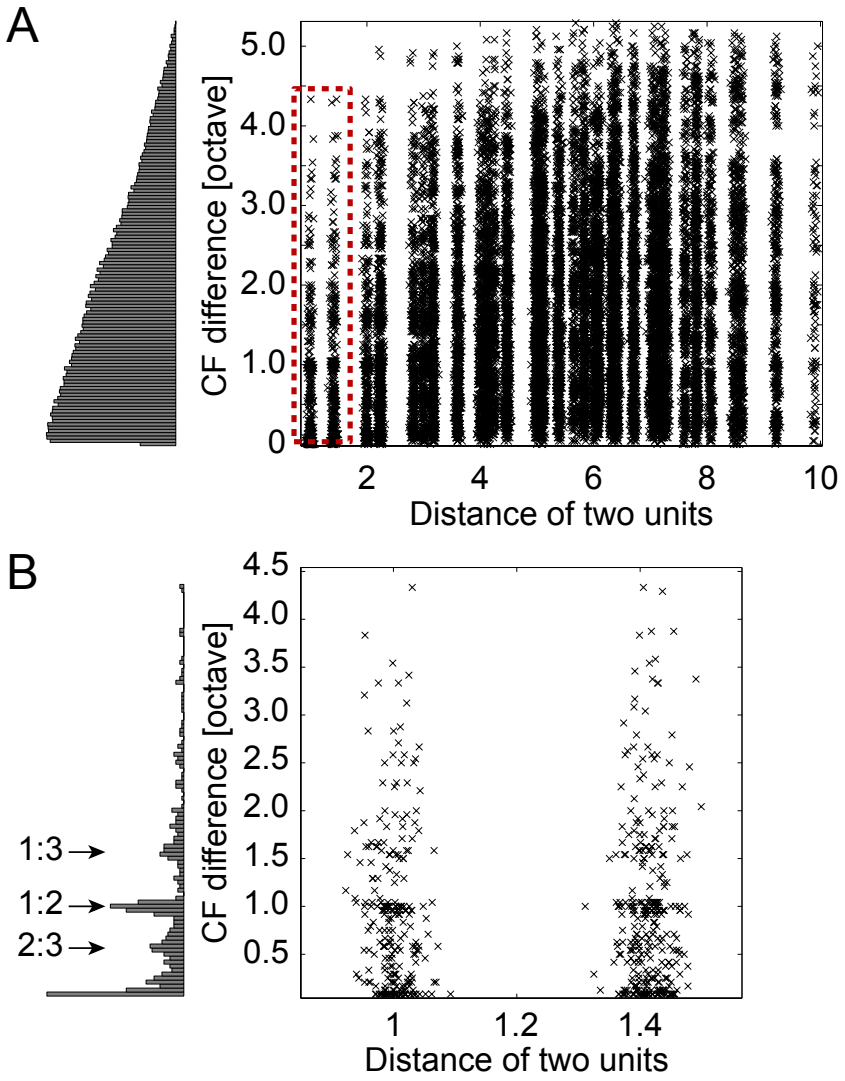


Figure 6.8: **Harmonic relationships between CFs of neighbouring units.** (*A*) The full distribution of distance and CF difference between two units. (*B*) The distribution of CF differences within neighbourhoods (the red-dotted rectangle in (A)). There were three peaks that indicate harmonic relationships between neighbouring units. The distances were jittered to obtain the visualisation.

two units in a learned topography. The CFs of even neighbouring units differed by up to ~ 4 octaves, which is consistent with recent experimental findings [8, 103]. A closer inspection of the red-dotted rectangular region of Figure 6.8A is shown in Figure 6.8B. The histogram in Figure 6.8B demonstrates several peaks at harmony-related CF differences, such as 0.59 ($= \log_2 1.5$), 1.0 ($= \log_2 2$; the largest peak), and 1.59 ($= \log_2 3$). These examples indicate that CFs of neighbouring units did not differ randomly, but tended to be harmonically related. A careful inspection of published data (Figure 5d from [103]) suggests that this relationship may be discernible in those published results; however, the magnitude of non-harmonic relationships cannot be clearly established from the inspection of this previously published study, as the stimuli used by the relevant experiment [103] were separated by an interval of 0.25 octaves and were therefore biased towards being harmonic. Thus, this prediction of a harmonic relationship in neighbouring CFs will need to be examined in more detailed investigations.

6.4 Discussion

Using a single model, we have provided a computational account explaining why the tonotopy of A1 is more disordered than the retinotopy of V1. First, we demonstrated that there are significant differences between natural images and natural sounds; in particular, the latter evince distant correlations, whereas the former do not. The topographic independent component analysis therefore generated a disordered tonotopy for these sounds, whereas the retinotopy adapted to natural images was locally organized throughout. Detailed analyses of the TICA model predicted harmonic relationships among neighbouring neurons. The results suggest that A1 and V1 may share an adaptive strategy, and the dissimilar topographies of visual and auditory maps may therefore reflect significant differences in the natural stimuli.

Figure 6.9 summarizes the ways in which the organisations of V1 and A1 reflect these input differences. Natural images correlate only locally, which produces a smooth retinotopy through an efficient coding strategy (Figure 6.9A). By contrast, natural sounds exhibit additional distant correlations (primarily correlations among harmonics), which produce the topographic disorganisation observed for A1 (Figure 6.9B). To extract the features of natural sounds in the auditory pathway, A1 must integrate multiple channels of distant

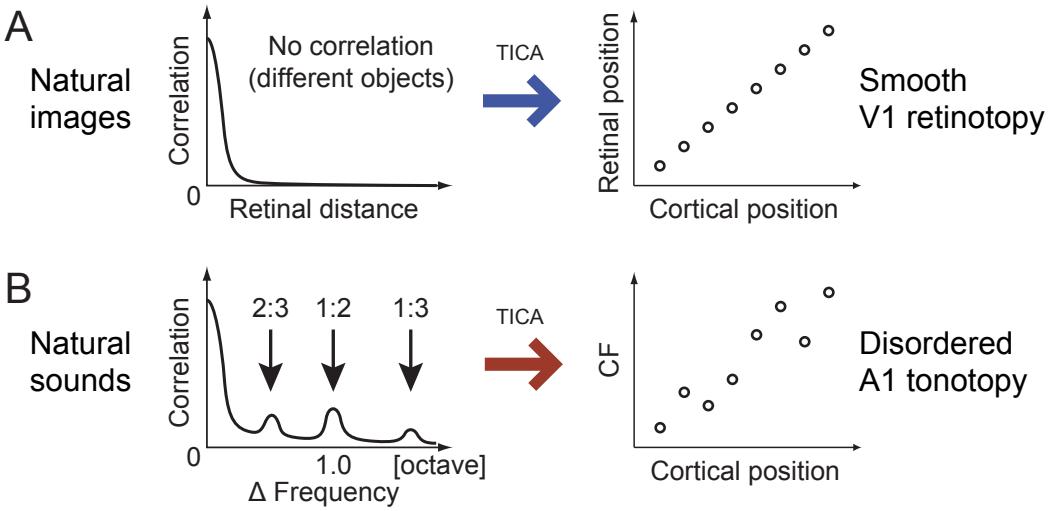


Figure 6.9: **Suggested relationships between natural stimulus statistics and topography.** While the local statistics of natural images are reflected in the smooth retinotopy of V1, the non-local harmonic statistics of natural sounds are reflected in the scattered tonotopy of A1.

frequencies [107]; for this purpose, the disordered tonotopy can be beneficial because a neuron can easily collect information regarding distant (and often harmonically related) frequencies from other cells within its neighbourhood. Our result suggests the existence of a common adaptive strategy underlying V1 and A1, which would be consistent with experimental studies that exchanged the peripheral inputs of the visual and auditory systems and suggested the sensory experiences had a dominant effect on cortical organisation [117, 2, 111].

This chapter begins with a neurophysiological report of scattered A1 topography, which were recorded in layer 2/3 of mice [8, 103]. By contrast, a more recent study reported the tonotopy of layer 4 is basically smooth [136] even in mice. How can we interpret it? Our model suggests that A1 (i) actively disorganise its tonotopy to integrate distant frequency channels. Although it is likely, the degree of scatter seen in the model result seems smaller than that first seen in the original observation, which suggests the existence of additional factor of disorganisation. This reminds us of the absence of columnar structure in layer 2/3 of mice V1 [84, 85], both of which may toghther suggest that the mice neocortex has (ii) an intrinsic spatial scattering mechanism through the laminar transformation from layer 4 to 2/3. This view seems to be supported by recent reports on

biased connections between sister cells, or ontogenetic columns [138, 139, 69, 86, 38]. Future models specifically for mice A1 would need to incorporate these observations.

6.5 Conclusion

We used a V1 model that was selected for its smooth retinotopy to provide a computational model of the tonotopic disorder in A1. In contrast to natural images, natural sounds exhibit distant (and often harmonic) correlations, which are learned and reflected in the apparent tonotopic disorder. The auditory model that had been adapted to human voices predicted harmonic relationships among neighbouring A1 cells. The result contributes to the understanding of the sensory cortices of different modalities in a novel and integrated manner.

Chapter 7

Complex cell

> Different roads sometimes lead to
> the same castle.
>
> GEORGE R. R. MARTIN

7.1 What are "complex cells" of A1?

Chapter 5 and 6 discussed the two contrasts known for V1 and A1 (i.e., the locality of the receptive fields and the smoothness of maps). The disorganised A1 properties, which do not seem easy to interpret intuitively, were reproduced using the learning models for V1 adapted to natural sounds. As the results support our hypothesis, in this chapter, we deductively discuss "complex cells" of A1, trying to propose a computational analogy yet unknown[1].

The V1 complex cells were found at the very early stage of V1 studies [47]. While they are selective to a specific orientation as well as simple cells, their responses are nonlinearly invariant to shift of position or phase. This nonlinear invariance has been modelled as pooling of activities of simple cells that are selective to a specific orientation and various phases [1] (Figure), which can be learned by computational models that adapt to natural image statistics [49]. Even though the concept of complex cells have been established for more than 50 years [47], interestingly, there have been few discussion on their counterparts in different modalities. Can we somehow discuss them? If we can, what is their function?

[1] Preliminary results of this chapter were published in [122].

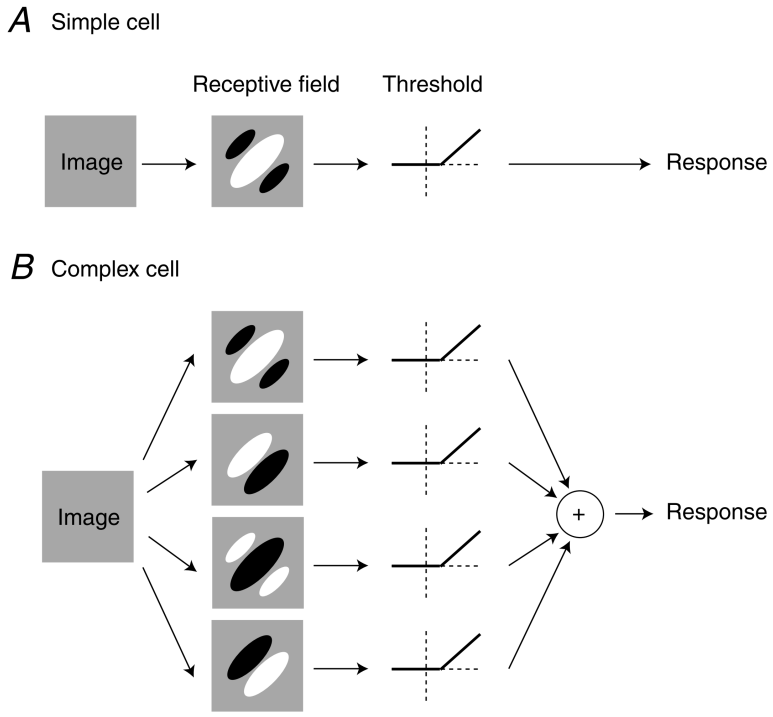


Figure 7.1: **Simple cells and complex cells of V1**. (*A*) Responses of a simple cell can be basically described by a single Gabor linear filter. (*B*) The energy model of a complex cell. The complex cell is selective to the same orientation as the simple cell described above; but its response is insensitive to the phase. This nonlinearity has been modelled as the sum of responses of simple cells that are selective to a single orientation but to different phases. (The image was adapted from [20].)

In fact, it is hard to show the existence of their analogues electrophysiologically; but we can approach them using computational methods based on our hypothesis. Assuming that the learning rule is shared by V1 and A1, we apply a V1 complex cell model to natural sounds, in lieu of natural images. We found that some of the learned "complex cells" showed a nonlinearity similar to the pitch cells recently found in the core field of marmoset monkey auditory cortex, which includes A1 [12]. The result suggests that the pitch cells of the auditory cortex are computationally analogous to V1 complex cells.

7.2 Method

7.2.1 The overcomplete TICA

The model used in this chapter is the extension of the TICA model [70] to account for overcomplete representations ($d < N$), which are observed in the V1 of primates. This was same with the one described in Section 6.2.2 and used here to show clearer localisation in the topography (Figure 7.5), even though the basically similar result can be reproduced with the original TICA.

The model is a two-layer neural network, as well as the original TICA, for learning simple cells, topographies, and complex cells in V1, using natural image statistics. An input to the model is an image patch, which induces a response pattern in the first layer for simple cells. The topography is defined by a fixed connection between the layers. The second layer models complex cells, whose responses are defined as pooling of activities of neighbouring simple cells.

We consider to reconstruct a whitened input $\boldsymbol{I}(x) \in \mathbb{R}^d$ ($d < N$) using a basis comprising N vectors $\boldsymbol{a}_i \in \mathbb{R}^d$,

$$\boldsymbol{I} = \sum_i s_i \boldsymbol{a}_i \tag{7.1}$$

where $s_i \in \mathbb{R}$ are the activities of the first layer units or simple cells. The activities of the simple cells are used to define those of complex cells $c_i \in \mathbb{R}$ as

$$c_i = \sum_j h(i,j) s_j^2 \tag{7.2}$$

where $h(i,j)$ is the neighbourhood function that defines the topography, which takes a value of either 1 or 0 depending on whether the indices i and j are neighbours or not. A 3×3 square windows was used in the two-dimensional topography of the present study.

Contrary to the original TICA, inverse filters cannot be uniquely defined because of the overcompleteness. A set of first-layer responses s_i to an input is computed by minimising the following extended energy function:

$$E = -\log L(\boldsymbol{I}; \{\boldsymbol{a}_i\}, \{s_i\}) \tag{7.3}$$

$$= \left| \boldsymbol{I} - \sum_i s_i \boldsymbol{a}_i \right|^2 - \lambda \sum_i G(c_i). \tag{7.4}$$

The first and second terms request preservation of the information and sparseness of activities of the complex cell layer, respectively. The parameter λ is the relative weight of the activity sparseness, in accordance with sparse coding [87]. We used $G(c_i) = -\sqrt{\epsilon + c_i}$. Minimisation of the function used its gradient

$$\Delta s_i \propto \boldsymbol{a}_i^T \left(\boldsymbol{I} - \sum_j s_j \boldsymbol{a}_j \right) - \lambda s_i \left(\sum_j h(i,j) g(c_j) \right) \tag{7.5}$$

where $g(c_i)$ is the derivative of $G(c_i)$. The initial value of s_i is set equal to the inner product of $\boldsymbol{I}$ and $\boldsymbol{a}_i$. Every 256 inputs, the basis is updated using the following gradient.

$$\Delta \boldsymbol{a}_i = \eta \left\langle s_i \left(\boldsymbol{I} - \sum_j s_j \boldsymbol{a}_j \right) \right\rangle \tag{7.6}$$

The operator $\langle \cdots \rangle$ represents the average over iterations. The parameter values used in the study was $\epsilon = 0.005$ (for numerical stability), $\lambda = 0.91$, and $\eta = 0.08$.

7.3 Results

7.3.1 Nonlinear responses similar to pitch-selectivity

In Chapter 5 and 6, neurophysiological features of A1 neurons were associated with the harmonic statistics of natural sounds. Psychoacoustics have long demonstrated interesting phenomena related to harmony, namely, the perception of pitch, which represents a subjective attribute of perceived sounds. Forming a rigid definition for the notion of pitch is difficult; however, if a tone consists of a stack of harmonics $(f_0, 2f_0, 3f_0, \ldots)$, then its pitch is the frequency of the lowest harmonic, which is called the fundamental frequency f_0. The perception of pitch is known to remain constant even if the sound lacks power at

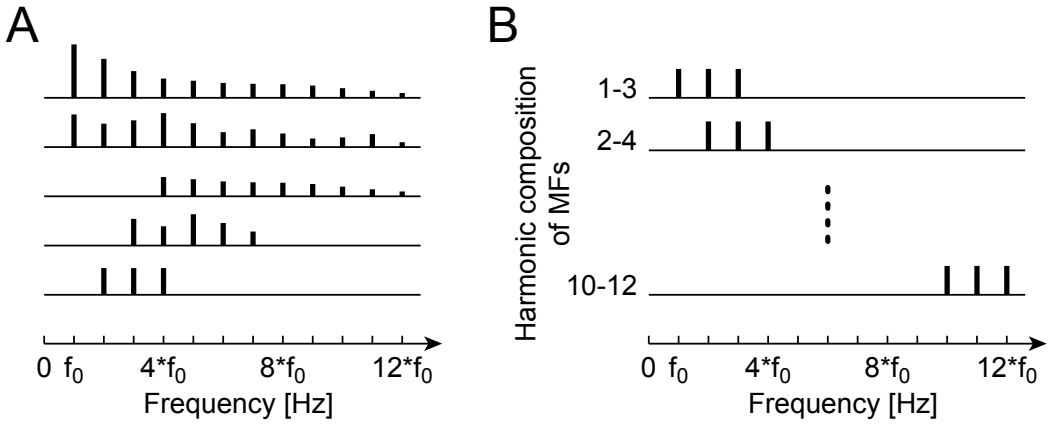


Figure 7.2: **Spectra of the missing fundamental sounds** (*A*) The spectra that share a single f_0, all of which are perceived with the same pitch. (*B*) The spectra of MFs used in the present study.

lower harmonics; in fact, pitch at f_0 can be perceived from a sound that lacks f_0, a phenomenon known as "missing fundamental" (Figure 7.2) [79]. Nonlinear pitch-selective responses similar to this perception have recently been demonstrated in certain neurons of the auditory cortex [12] that localise in the low-frequency area of the global tonotopy where A1 and area R share a border (Figure 7.3A).

To investigate pitch-related responses, previously described complex tones [12] that consisted of harmonics were selected as inputs for the model described in Section 6.3.3. For each unit, responses were calculated to complex tones termed missing fundamental complex tones (MFs) [12]. The MFs were composed of three consecutive harmonics sharing a single f_0; the lowest frequency for these consecutive harmonics varied from the fundamental frequency (f_0) to the tenth harmonic ($10f_0$), as shown in Figure 7.2B. For each unit, five patterns of f_0 around its CF (~ 0.2 octave) were tested, resulting in a total of $10 \times 5 = 50$ variations of MFs. The activity of a unit was normalised to its maximum response to the MFs. Pitch-selective units were defined as those that significantly responded (normalised activity > 0.4) to all of the MFs sharing a single f_0 with a lowest harmonic from 1 to 4.

We found certain pitch-selective units in the second layer (n = 66; 6 simulations), whereas none were found in the first layer. Figure 7.4 illustrates the response profiles of the pitch-selective units, which demonstrated sustained activity for MFs with a lowest harmonic below the sixth harmonic ($6f_0$), and this result is similar to previously published

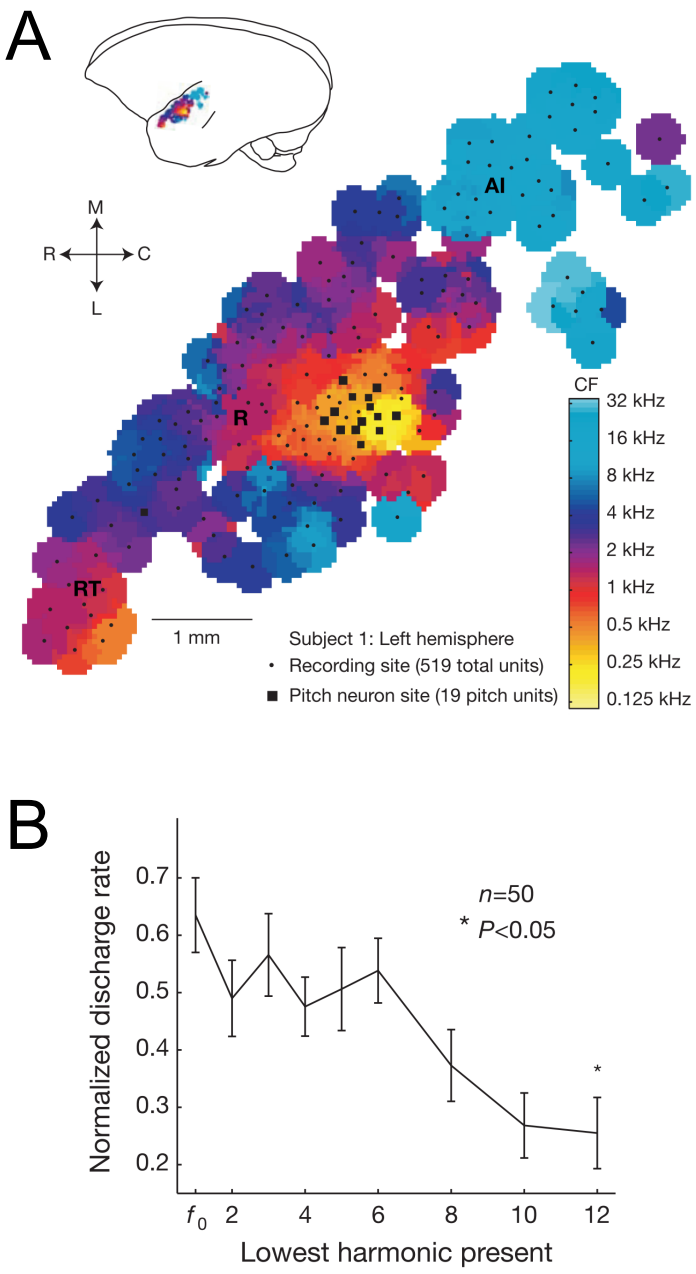


Figure 7.3: **Pitch neurons reported in the monkey auditory cortex**. (*A*) In the core field of marmoset monkeys, the pitch cells were found only in the restricted area where A1 and area R share a border as the low-frequency area of the global tonotopy. (*B*) The pitch cells showed nonlinear responses like missing fundamental, keeping to discharge for complex missing fundamental sounds composed of three consecutive harmonics. (Both images were adapted from [12].)

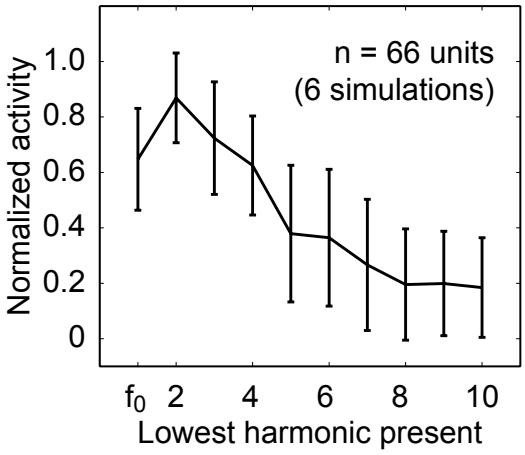


Figure 7.4: **Nonlinear responses similar to the pitch selectivity.** Those units continue to fire even when the simuli lack of power at lower harmonics, similarly to Figure 4C of [12] (Figure 7.3B). Since f_0 is the CF of these units, if their responses were linear, the activity should have quickluy dropped.

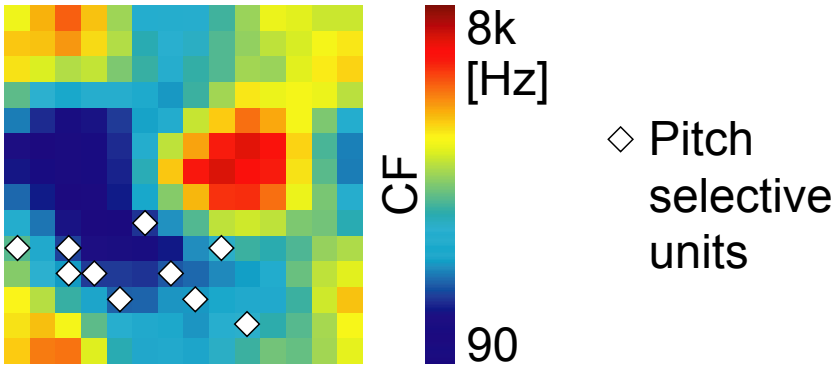


Figure 7.5: **The spatial distribution of pitch-selective units.** The positions are superimposed on the smoothed tonotopy. (An example of a single session)

data [12]. Other units did not show strong responses to MFs comprising of only high harmonics, as the first harmonic or the fundamental frequency f_0 is their CF.

Additionally, these units were located in a low-frequency region of the global tonotopy, as shown in Figure 7.5, and this feature of pitch-selective units is also consistent with previous findings [12] (Figure 7.3A). The second layer of the TICA model, which contained the pitch-selective units, was originally designed to represent the layer of V1 complex cells, which have nonlinear responses that can be modelled by a summation of

"energies" of neighbouring simple cells [49, 50, 70]. Our result suggests that the mechanism underlying V1 complex cells may be similar to the organisational mechanism for the pitch-selective cells of the auditory cortex, i.e., their computational analogy.

7.3.2 Biased pooling mechanism underlying the pitch-selectivity

What kinds of mechanisms underlie the nonlinear response like missing fundamental? In the model, activities of the complex layer are defined by pooling those of neighbouring simple cells. To characterise the profile of pooling that underpins the pitch selectivity, distributions of CF difference between a pair of simple cells in the same neighbourhood were investigated.

Figure 7.6A shows the distribution of CF difference between a pair of simple cells that share a pitch-selective unit as a parent. It has peaks at 0.59 ($= \log_2 1.5$), 1.0 ($= \log_2 2$; the biggest peak), 1.59 ($= \log_2 3$), and 2.0 ($= \log_2 4$) octave, which correspond to harmonic ratios 2:3, 1:2, 1:3, and 1:4, respectively. This means there are harmonic relationships among the simple cells that are pooled by the pitch-selective cells. The pitch selectivity like missing fundamental can emerge from integrating multiple frequency channels that are related harmonically.

Figure 7.6A shows another distribution obtained from other pairs that share a parent unit that show no pitch-selectivity. In contrast to the previous one, this distribution is flat and long-tailed with no clear structure. This difference of pooling profiles between pitch-selective cells and other cells is a testable prediction of the computational model.

7.4 Discussion

We found that the pitch cells of the auditory cortex can be modelled using a V1 complex cell model. The model demonstrated that the pitch cells achieve the robust selectivity to the shared fundamental frequency of various harmonic spectra (Figure 7.2) by integration of distant frequency channels that are related harmonically.

Our result suggested that a common mechanism may underlie the complex cells of V1 and the pitch-selective cells of the auditory cortex including A1. Additional support for this notion was provided by recent evidence indicating that the pitch-selective cells are most commonly found in the supragranular layer [12], and V1 complex cells display

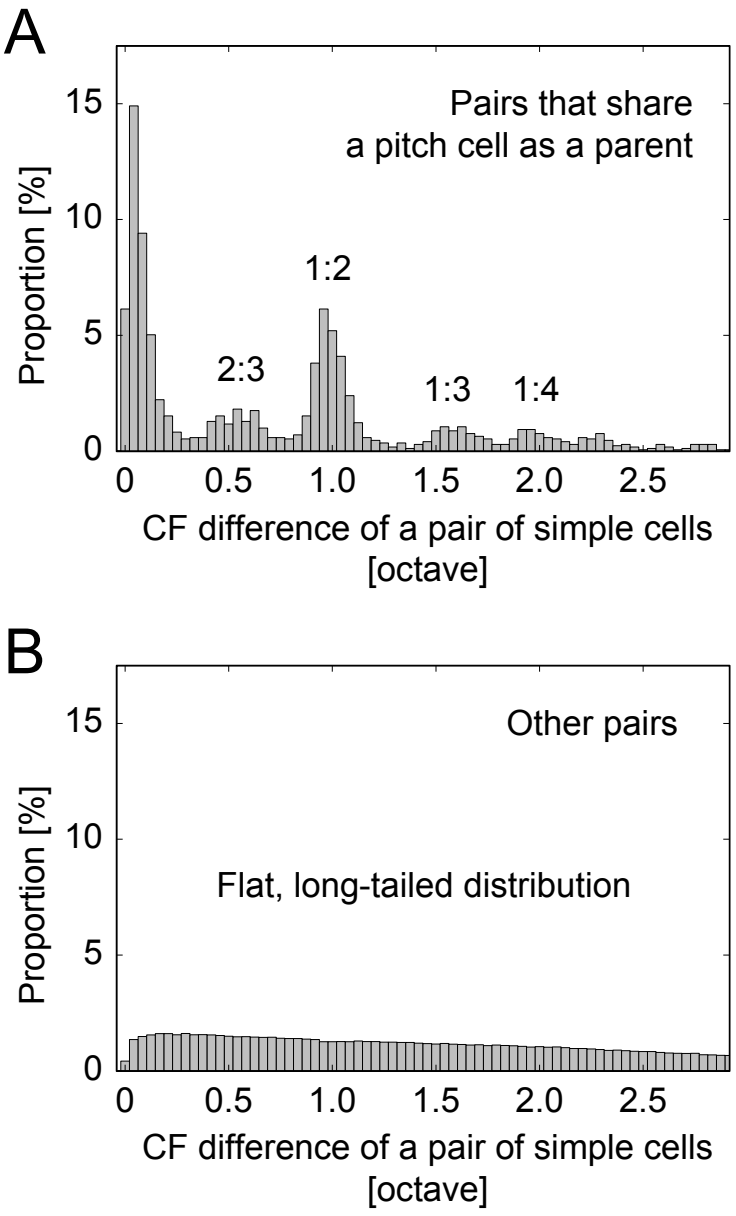


Figure 7.6: **Harmonically biased pooling by the pitch-selective units.** (*A*) The distribution of CF difference of simple cell pairs that share a pitch selective unit as a parent in the complex cell layer. (*B*) The distribution of CF difference taken from other simple cell pairs that share a parent that is not selective to the pitch.

Table 7.1: **Suggested analogy between complex cells and pitch cells**

	Response	**Selectivity**	**Invariance**	**Model**
Complex cells	Nonlinear	Light bars of a specific orientation	Phase	TICA + natural images
Pitch cells	Nonlinear	Harmonic sounds of a specific pitch	Spectrum	TICA + natural sounds

a similar tendency. It has been hypothesised that V1 complex cells collect information from neighbouring cells that are selective to different phases of similar orientations; in an analogous way, the pitch-selective cells could collect information from the activities of neighbouring cells, which in this case could be selective to different frequencies sharing a single f_0. To the best of our knowledge, no previous studies in the literature have attempted to use this analogy of V1 complex cells to explain the pitch-selective cells (however, other potential analogues have been mentioned [112, 3]). Our results and further investigations should help us to understand these pitch-selective cells from an integrated, computational viewpoint.

Another issue that must be addressed is what functional roles the other units in the second layer play. One possible answer to this question may be multipeaked responses related to harmony [56], which have been explained in part by sparse coding [119, 121] in Chapter 5; however, this answer has not yet been confirmed by existing evidence and must therefore be assessed in detail by further investigations.

What kind of computational analogy can we find in the comparison of V1 complex cell and the pitch cell? Table 7.1 summarises comparable viewpoints. Both nonlinearly respond to stimuli by extracting hidden essential features, i.e., the orientation and the pitch (or the fundamental frequency). In terms of the invariance, V1 complex cells are invariant to phase shifts as they pool simple cells with different phases of a single orientation; as an analogy, the pitch cells of the auditory cortex can be said to be invariant to spectral transformations that keeps a single fundamental frequency f_0 or a single pitch. Finally, what the model demonstrated is that both can emerge from different natural stimulus statistics, through a single learning model, strongly suggesting their computational analogy.

Complex cells hold representations that can be seen one step higher than those of simple cells. While we dealt with a two-layer hierarchy within a single functional area, we

can also discuss deeper architectures over multiple areas. In the visual system, whose hierarchy is clearer, there have been a lot of attempts to model the deep architecture such as Pandemonium [110], Neocognitron [35, 36, 37], Convolutional neural networks [67], and recent rapid advances that are called Deep Learning [45, 63, 66]. Most of these architectures stack simple and complex cell layers in an alternative manner, being inspired by the architecture of the visual cortical system. In contrast, though some deep models achieve good performance on auditory tasks [78, 109, 27], there have been no discussion on how we can interpret them in terms of comparison with the auditory cortical system. The analogy we showed would be the first step to further our neural understanding of the auditory deep learning.

7.5 Conclusion

The complex cells are a concept established in the primary visual cortex (V1), whereas their counterparts in other sensory modalities have been unidentified despite of the structural uniformity of the neocortex. Computational studies have shown that learning rules for some V1 properties can be applied to the primary auditory cortex (A1), although the discussion is limited to those for simple cells. The present study discusses "complex cells" of A1, using a V1 complex cell model adapted to natural sounds instead of natural images. We found that some of the auditory "complex cells" resemble the pitch cells recently found in the core field of auditory cortex including A1, showing nonlinear invariance under a spectral transformation that keeps a constant pitch perception. The result suggests that the pitch cells of the auditory cortex may be computationally analogous to V1 complex cells.

Chapter 8

Discussion

> We are all the same and we are all different.
> What great friends we will be.
>
> KELLY MORAN

8.1 A1 models based on sparse representation

8.1.1 Coverage of the models

We have shown successful examples of A1 modelling, but these examples are not panaceas for all aspects of the functional area. Here, we will mention three specific issues; the first one is the treatment of time. Except for some results that dealt with time, our models have primarily focused only on the frequency domain. Although neurophysiological A1 properties showed the relative importance of the frequency domain [99], the temporal structures of sound surely play crucial roles in accordance with our intuition. Lewicki [68, 114] and Karklin [58] and colleagues have proposed statistical models that can capture both domains simultaneously using sparse codes; we can benefit from these models, although they are not clearly specified as models for A1. Note that any discussion on the detailed temporal structure of the A1 model will require precise input patterns provided by subcortical computational models [73].

Second, the interpretation of the results is influenced by species-specific differences. The models have been validated by matching neurophysiological knowledge from the neocortex of mammals, mostly monkeys. The columnar structures found in monkeys were

considered units in the TICA model; however, recent imaging techniques have revealed that rodents do not have a columnar structure in the neocortex [84, 85]. Because the implicit assumption of the model has been basically falsified for rodents, we must carefully interpret the model behaviour in comparison with rodent data. While we currently cannot state that the model result is quantitatively consistent with mouse A1 data, the model may explain the qualitative difference between V1 and A1.

The last point is that our models are non-exhaustive and unreliable in terms of quantification. The works presented in this thesis have primarily focused on specific aspects of the auditory cortex, such as those related to harmonics. Quantitative results were partly given only in the first part of the results. It is unlikely that the models cover all important features of A1; thus, we need to be careful regarding what is lacking, which necessitates additional systematic physiological studies on A1. While our knowledge of A1 is limited, our models have succeeded in explaining as many things as possible, although never everything.

8.1.2 What are "natural" stimuli?

One remaining issue is that the present results necessitate the need to reconsider the type of natural signals used as the input in the models that learn sparse representations. In the present study, the result similar to the monkey experiment was reproduced using only a set of behaviourally important sounds or the mammal vocalisations, albeit in the environment the actual periods of hearing such highly harmonic sounds are quite limited. The original models for V1 [87, 88, 89, 49, 70] accepted a broad class of natural images as the input, which robustly produced V1-like properties because natural scenes generally share their statistical features regardless of behavioural importance. However, this is not the case of audition because the statistical features of behaviourally relevant sounds differ from those of irrelevant ones [72]. The behavioural importance of sensory stimuli affects the degree of neural plasticity within the sensory cortices including A1 and V1 [105], through neuromodulatory signals such as acetylcholine [7, 61] or dopamine [9], disinhibition by interneurons [94], and the selective attention [137, 102]. Thus, the input data set used in the sparse coding model might need to be built considering its behavioural relevance. Further investigation on this topic should be conducted in the future.

8.1.3 Interpretation of the sparsity

How can we interpret the objective functions of the computational models? All models discussed in this thesis aimed to realise representations of a higher level according to a criterion called sparseness [14] or independence, the latter of which might be more important for our understanding of the neocortex. Another fundamental feature of the neocortex as a computing machinery that was not mentioned in this thesis is hyper-parallelism. Computer science literatures has noted that efficient realisation of parallel computing requires the consideration of data parallelism [44]. Prior to distribution, data must be decomposed to small chunks that have as little interdependency as possible because the cost of inter-node communication is the largest overhead. Interestingly, this view (i.e., independent factorisation) can also be found in our own scheme of recognising the complex world (e.g., the concept of pseudo-articles in physics), which suggests that this type of decomposition might be the general strategy used when trying to understand massive data with complex structures.

The intelligence of mammals has rapidly grown along with the size explosion of the neocortex, which coincides with the presence of cortical functional columns. The emergence of the columnar structure might be fundamental in realising the independent factorisation, which could have rendered the neocortex a scalable system [118] that can flexibly develop to understand the amazingly complex world. For example, in our nocortex, our experience of reading this thesis could have led to the formation of a "pseudo-article" representation composed of keywords, such as V1 and A1.

8.2 Computational cross-areal comparison

8.2.1 Beyond A1: S1, M1, and more

In this thesis, we introduced three attempts to model A1 in an analogy with V1. This was a first step to computational cross-areal comparison, which is a more general approach that can be essentially applied to other neocortical areas. We argued that because V1 has the richest repertoire of computational models, the most probable approach would be an analogy that uses V1 as its source.

Which functional areas other than A1 might be hopeful as a target of the analogy?

Theoretical frameworks of analogy have suggested that the most reliable criterion is structure [40], which would correspond here to anatomy and genetics. Anatomical [19, 39, 18] and genetic [82, 43, 83] evidence suggests that the second closest target of the analogy is the primary somatosensory cortex (S1).

When applying models with sparseness to an area of a specific modality, we need a huge dataset of "natural stimuli", which is often a major issue. However, this issue has been partly solved by two studies. The first one is an application of the sparse coding model [106]. The authors established a new "natural somatosensory stimulus" dataset; they systematically collected hand-touch patterns when subjects grasp objects in various ways using white powder and a digital camera. Applying the model to the dataset, they successfully obtained S1-like multi-digit receptive fields.

The other study [81] took a different approach for data acquisition. They previously developed a simulator for the general movements of a foetus in the uterus that systematically simulates the "natural" sensory pre-born experience. The TICA model applied to the data produced S1-like receptive fields; in addition, the obtained topographic map was more similar to the S1 map compared with what was learned with an "unnatural" simulated sensory experience outside of the uterus. Thus, the authors suggested that the in utero sensory experience is working as a guide to the normal development of S1. The success of these two studies also illustrates the dawn of the paradigm.

Are there any other areas that are potentially suitable for the computational analogy with V1? Contrary to the structure-based bottom-up inference, if seen from the model-based top-down viewpoint, a potential region resides in the primary motor cortex (M1). Sparse representations of motor commands have long been investigated as "muscle synergies" or "motor primitives", and the mathematical formulation for learning is surprisingly similar to the sparse coding theory [124, 123].

However, we should not jump to a hasty conclusion; the anatomical structure and gene expression patterns of M1 have been known contrastive to primary sensory areas. In addition, the sparse coding of outputs (motor commands) and that of inputs (sensory stimuli) cannot be easily integrated without an appropriate extension of the sparse coding theory. Although we need to be careful, bridging the two contrastive ends of the functional areas would be an important challenge for computational cross-areal comparison.

8.2.2 General workflow of cross-areal analogous modelling

The analogous modelling from V1 to A1 can be seen as the first systematic application of a more general modelling approach for multiple functional areas. The research procedure in analogous modelling as a computational cross-areal comparison may be generalised as described below.

```
while (true) {
  Pick out two areas A and B to be compared
                                    based on anatomy/genetics.
  Pick out a 'nice' (transferrable) model X of Area A.
  Apply Model X to Area B.
  foreach (behaviour Q of Model X applied to B) {
    if (exist(comparable physiological knowledge)) {
      if (consistent) {
        Model X generalises to Area B!
      } else {
        Back to anatomy/genetics.
        Modify Model X and retry.
      }
    } else {
      Predict behaviour Q of Area B.
    }
  }
}
```

Table 8.1: **Contents of variables in our analogous modelling**

	A	**B**	**X**	**Q of A**	**Q of B**
Chap. 5	V1	A1	Sparse coding	Local RFs	Non-local RFs
Chap. 6	V1	A1	TICA	Smooth retinotopy	Scattered tonotopy
Chap. 7	V1	A1	Overcomplete TICA	Complex cells	Pitch cells

Table 8.1 summarises contents of variables A, B, X, and Q in our attempts of A1 modelling described in Chapter 5, 6, and 7.

This description of the procedure suggests how we should build a model that is suitable for modelling the computational principles. Model X needs to be "transferrable" to other areas, i.e., the model must exclude any overfit to a specific area. This condition seems to be satisfied by both of the two models used in this thesis (sparse coding and TICA), which underlies our successful analogous modelling.

8.2.3 An analogy to linguistics as another pursuit of universals

From a more general point of view, the quest for computational principles can be stated as a pursuit of universality in diversity. While a pursuit at the level of anatomy and genetics results in a data-driven approach, this thesis argued that at the level of computation, we need a model-driven approach. How can we further understand this picture? Here, we discuss, at a different level, another analogy between neocortical neuroscience and linguistics. Although human languages are very diverse, linguists have long been interested in finding their common features or "language universals". To tackle the issue in its long history, the field of linguistics has developed two approaches that are contrastive and complementary. Analogising this with our issue, we can identify where we are, what we should do, and where we are going.

8.2.3.1 Linguistic typology: data-driven, inductive

The pursuit of linguistic universals originated as linguistic typology, a data-driven, inductive approach [129, 96]. This type of research typically (1) creates an appropriate database across a wide range of languages, (2) compares elements in the database, and (3) tries to depict underlying language universals [42, 134]. For example, a universal is described as "All languages have nouns and verbs". This widely accepted procedure assumes that we can access a wide range of language resources, that we can compare cross-linguistic data, and that we know what we are going to compare.

We can see good agreement with some cross-areal comparisons of the neocortex stated in Section 1.2, that is, cross-areal comparisons at the levels of anatomy, genetics, and spike statistics. These comparisons share the data-driven research framework with linguistic

typology, which is underpinned by the non-biased access to data and the comparability (Table 1.1). However, these conditions cannot be satisfied by all comparative approaches; in fact, linguistic typology seems less effective in the search for universals at a higher level, where the comparability becomes more unclear. How can we address such situations with non-trivial comparability?

8.2.3.2 Theoretical linguistics: theory-driven, deductive

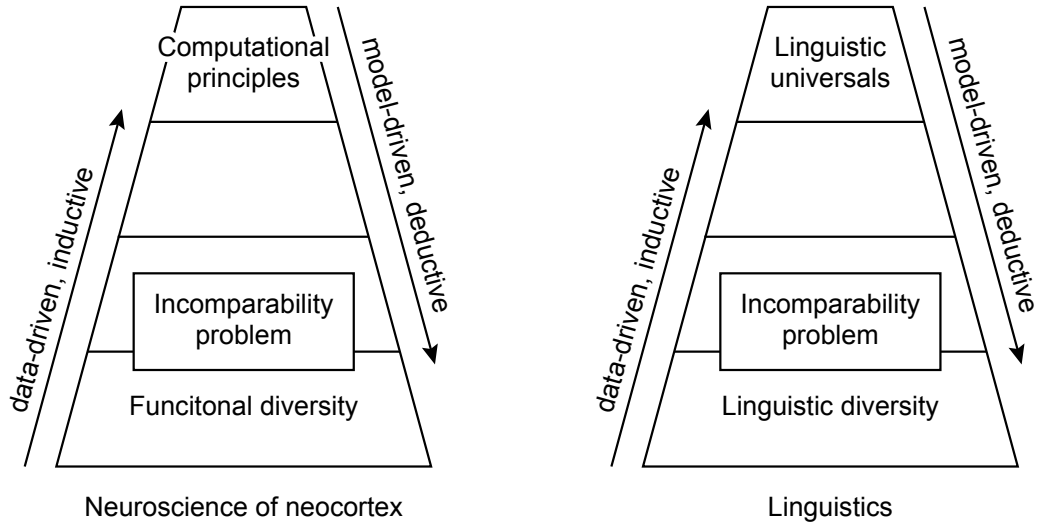


Figure 8.1: **Analogy between neocortical cross-areal comparison and linguistics**. Neocortical cross-areal comparison is analogous to linguistics in the sense that both aim to find universals out of diversity. An approach is a data-driven, inductive strategy like anatomy or typology, which faces the problem of incomparability at a relatively abstract level. By contrast, the model-driven approach like computataional cross-areal comparison or generative grammar, can deductively propose non-trivial comparisons. The two approaches are contrastive and complementary to each other.

Theoretical linguistics developed relatively later than typology and has exhibited astonishing advancements after Chomsky presented the theory of generative grammar [23, 24, 25]. In this approach, the existence of universality is "given" because all humans can learn a language without any special efforts unless deprived of a natural linguistic experience. Thus, researchers first assume a specific generative model. Then, they continue to make an endless loop of fitting to the linguistic data and refining the model. The model-

driven approach is effective, particularly at higher levels where comparability is non-trivial (e.g., syntaxes of English and Japanese, which apparently have distinct structures, can be described just by difference of a few parameters [4]).

This is in good agreement with what we have proposed in this thesis. As theoretical linguistics argues the need to compare multiple languages at the level of generative models, not of the linguistic features, we argued the need to compare multiple neocortical areas at the level of computational models, not of the neurophysiological functions themselves. This has enabled us to discuss computational homologies that are not easily comparable; the most illustrative example of this is the finding of the new analogy between complex cells and pitch cells. To date, this computational cross-areal comparison has not been substantially developed, which is similar to the situation of generative grammar in the past that has not yet emphasised the importance of cross-linguistic comparison. We might regard our analogous modelling from V1 to A1 as the cross-linguistic modelling in the first stage, e.g., from English to Dutch.

8.2.3.3 Future perspectives implied by the analogy

What can we learn from the analogy with linguistics? The contrast between the two approaches in linguistics can be restated as a contrast between the number of languages considered and the depth of linguistic features considered. Ideally, the two approaches should learn from and complement each other, but in fact, they have unfortunately not had a strong relationship [116, 6, 26], similar to the two approaches in neuroscience.

However, to date, there has been a move to resolve the conflict and identify "the middle way" [5, 116]. Typology began to consider more knowledge of languages, and formal theory started to be applied to more languages. A similar move also seems to be apparent in neuroscience (Figure 8.2). Anatomy began to consider more complex structures like microcircuits; computational modelling has begun to compare multiple functional areas [119, 106, 122, 121]. It is difficult to prove the "true universality" of computational principles as well as of generative grammar, which has tended to consider only Indo-European languages. However, further investigations are required in order to reveal the universality.

Before closing this discussion, we would like to mention two differences between neuroscience and linguistics. First, computational cross-areal comparison has fewer clues in

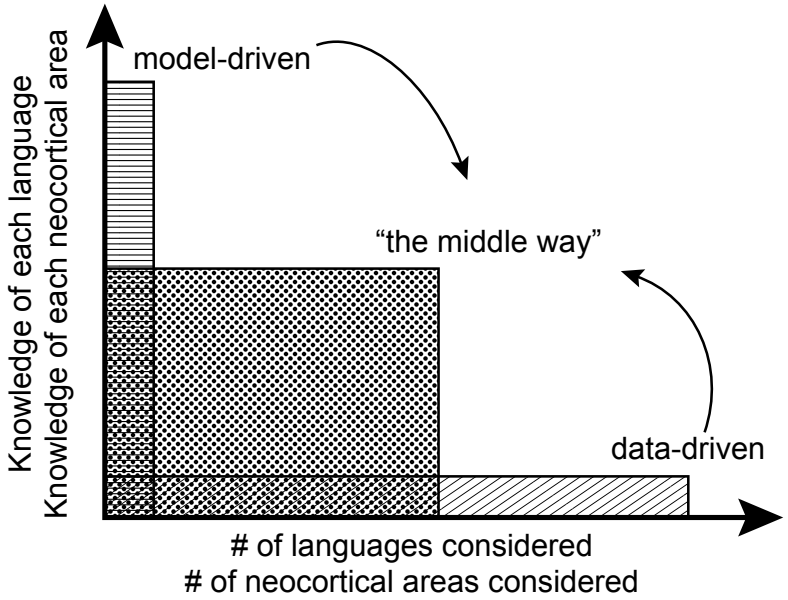


Figure 8.2: **Future of cross-areal comparison implied by the analogy with linguisitcs**. In search for universals, linguists have begun to seek for "the middle way" of the two contrastive approach. A similar move will be achieved by neuroscientists, too, in the future. This thesis provided its first step from the model-driven side: computational modelling has started to increase the number of cortical areas considered, from one to two. (The image is based on [5].)

the form of models of each functional areas. As for linguistics, major languages have been systematically modelled for each, which has laid the foundation to consider what to compare and how to build a specific formal theory. These are unavailable for computational cross-areal comparison, which allows us only to make an analogy. For a more flexible approach to the issue, we also need to develop models for each functional area.

The other issue is the cost of data acquisition. Language data can be obtained as needed, often through the Internet recently; but neurophysiology requires costly measures and an appropriate design of experiments in order to discuss functions. The cost difference seems to be one of the factors that cause a higher barrier between experimentalists and theoreticians in neuroscience compared to that in linguistics. Their reconciliation will face serious challenges, but we must solve them by moving closer together.

Chapter 9

Conclusion

True happiness resides in things unseen.

Night Thoughts
EDWARD YOUNG

9.1 V1 and A1: learning is same, but input is different

The purpose of this thesis was to computationally model A1 in comparison to V1. The key issue, which is also related to the mystery of the neocortex, was "what is common in V1 and A1, and what is not?". Computational models of A1 were proposed based on our core hypothesis: learning is same, but input is different. The models used the same learning rules as the original V1 models; however, they were adapted to natural sounds, instead of natural images.

Prior to modelling, in Chapter 4, we showed a representative clear contrast between the statistics of natural sounds and images. In the eye field, the statistical dependency of natural images is limited within a very small, localised region. That of natural sounds, however, is non-localised; there are strong dependencies between distant frequencies, which in particular are structured harmonically. If V1 and A1 are adapted to the distinct statistics, then the results of adaptation, which are what we can neurophysiologically observe, will be dissimilar even though they share a single learning strategy. We argued this hypothesis for V1 and A1, and provided explanations for the three neurophysiological contrasts in accordance with the hypothesis.

The first contrast discussed in Chapter 5 was related to the receptive fields of what we call simple cells in V1. Receptive fields of V1 simple cells are spatially restricted within a very small region of the eye field; however, those of A1 cells can have multiple distant peaks in their frequency tuning curve, which tend to be harmonically related. We modelled the A1 receptive fields, applying a V1 model called sparse coding, to natural sounds in lieu of natural images. The model adapted to harmonic sounds, such as voices, and successfully reproduced non-local and harmonic receptive fields similar to A1 cells.

The second contrast was topographic maps (Chapter 6). While the topographic structures are similar in V1 and A1, recent advances in recording techniques have revealed that the tonotopic map of A1 is much more scattered than the retinotopic map of V1 at a microscopic scale. To answer why this discrepancy emerges, we applied a V1 map computational model, TICA, to natural sounds, instead of natural images. The model adapted to sounds showed more scattered topography and predicted that the disorganisation may be partially organised in terms of harmonics. The results suggest that the A1 disorder resulted from the adaptation to natural sound statistics, through the same strategy as V1, in order to integrate distant, but statistically dependent, frequencies.

Following the two known contrasts, in Chapter 7, we attempted to discuss the last unknown contrast: complex cells. Whereas the concept of complex cells has been well discussed and established in the visual cortex, there was almost no discussion about their counterparts in other modalities. We applied a variant of the TICA model, which can explain not only maps but also complex cells, to natural sounds. We found that some of the learned "complex cells" showed nonlinear responses similar to the psychoacoustic phenomenon known as a missing fundamental, which resemble the pitch cells recently found in the auditory cortex including A1. We proposed an analogous view of V1 complex cells and A1 pitch cells in terms of invariance, suggesting that pitch cells are "complex cells" of the auditory cortex.

We discussed three computational models for A1 and showed that the neurophysiological characteristics of A1, which seemingly contrast with those of V1, can still be reproduced through the same learning strategy with V1. The results consistently support our core hypothesis, suggesting the computational analogy between V1 and A1. The distinct characteristics of A1, which have been harder to interpret than V1, would not result from its own specialised computational strategy, but instead from the distinct harmonic

statistics of natural sounds.

9.2 Analogous modelling advances our understanding of the neocortex

The attempts to model A1 in the analogy contribute not only to our understanding of the specific functional area but also to an understanding of the entire neocortex; they can be seen as the first set of instances of the area-to-area comparative approach, which we named the "computational cross-areal comparison". Because we dream of revealing the computational principles of the neocortex, we must compare its multiple functional areas at the level of computational theory; however, this comparison has been lacking because we can only discuss the functions of a few cortical areas, and even well-known functions can be difficult to compare.

This thesis proposed an approach that can solve the issue by virtue of an analogy. Using V1 as the source of the analogy, we may be able to target other functional areas in a model-driven way. In other words, we can compare multiple areas in the space of the models (they could be called "proto-functions"), even though we cannot accomplish this in the space of the area-specific, learned functions. It would be illustrative that, based on the approach, we found an unknown computational mapping from the complex cells to the pitch cells, which can never be directly compared in the space of their apparent functions. The approach is becoming more useful because modern measuring techniques are accumulating cross-areal findings in anatomy and genetics, on which the analogous modelling relies.

The model-driven approach contrasts with the classical data-driven approaches (e.g., anatomy and genetics), which reminds us of the discussion on linguistics, another pursuit of universals in diversity through two contrastive and complementary approaches. One of them is linguistic typology, a data-driven and inductive approach, which has revealed some linguistic universals but is less effective for non-trivial comparisons. These difficult comparisons can be discussed through the other contrastive approach known as theoretical linguistics (e.g., generative grammar), which is, in turn, model-driven and deductive. The theoretical approach has succeeded at more abstract levels, which paves the way for the analogous approach in computational neuroscience that aims to find, in a different domain, universals out of diversity.

In summary, this thesis emphasised the importance of computational cross-areal comparison and proposed a specific strategy employing analogy at the level of computational theory. Even though the cross-areal comparison was emphasised, it is never meant to debauch any attempts to model specific functional areas; rather, area-specific models will be more in demand because they are the foundations of the cross-areal comparison. Successful models for a wide repertoire of areas are lacking, which makes computational cross-areal comparisons more difficult than the case of linguistics. The analogy with linguistics implies the fusion of the two approaches in the future, which suggests what should be done by experimentalists and theoreticians. In linguistics, it is difficult to prove the "true" universality of a specific version of universal grammar. In computational neuroscience of the neocortex, it would also be difficult, but we hope that more and more researchers will tackle this big mystery and contribute to the understanding of intelligence (including language) and underlying "universal computation".

Appendix A

List of Acronyms and Abbreviations

A1 The primary auditory cortex

CF Characteristic Frequency

DI Discontinuity Index

fMRI functional Magnetic Resonance Imaging

LGN Lateral Geniculate Nucleus

MGB Medial Geniculate Body

M1 The primary motor cortex

RF Receptive Field

S1 The primary somatosensory cortex

TICA Topographic Independent Component Analysis

V1 The primary visual cortex

References

[1] Adelson EH and Bergen JR. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, **2**(2):284–299, 1985.

[2] Angelucci A, Clascá F, and Sur M. Brainstem inputs to the ferret medial geniculate nucleus and the effect of early deafferentation on novel retinal projections to the auditory thalamus. *The Journal of Comparative Neurology*, **400**(3):417–439, 1998.

[3] Atencio CA, Sharpee TO, and Schreiner CE. Hierarchical computation in the canonical auditory cortical circuit. *Proceedings of the National Academy of Sciences*, **106**(51):21894–21899, 2009.

[4] Baker MC. *The atoms of language: The mind's hidden rules of grammar*. Basic Books, 2001.

[5] Baker MC. Formal generative typology. In Heine B and Narrog H, editors, *The Oxford Handbook of Linguistic Analysis*, pages 285–312. Oxford University Press, 2009.

[6] Baker MC and McCloskey J. On the relationship of typology to theoretical syntax. *Linguistic Typology*, **11**(1):285–296, 2007.

[7] Bakin JS and Weinberger NM. Induction of a physiological memory in the cerebral cortex by stimulation of the nucleus basalis. *Proceedings of the National Academy of Sciences*, **93**(20):11219–11224, 1996.

[8] Bandyopadhyay S, Shamma SA, and Kanold PO. Dichotomy of functional organization in the mouse auditory cortex. *Nature Neuroscience*, **13**(3):361–368, 2010.

[9] Bao S, Chan VT, and Merzenich MM. Cortical remodelling induced by activity of ventral tegmental dopamine neurons. *Nature*, **412**(6842):79–83, 2001.

[10] Barlow HB. Possible principles underlying the transformations of sensory messages. In Rosenblith W, editor, *Sensory Communication*, pages 217–234. MIT Press, Cambridge, MA, 1961.

[11] Bell AJ and Sejnowski TJ. The "independent components" of natural scenes are edge filters. *Vision Research*, **37**(23):3327–3338, 1997.

[12] Bendor D and Wang X. The neuronal representation of pitch in primate auditory cortex. *Nature*, **436**(7054):1161–1165, 2005.

[13] Bendor D and Wang X. Cortical representations of pitch in monkeys and humans. *Current Opinion in Neurobiology*, **16**(4):391–399, 2006.

[14] Bengio Y, Courville A, and Vincent P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**(8):1798–1828, 2013.

[15] Benson NC, Butt OH, Datta R, Radoeva PD, Brainard DH, and Aguirre GK. The retinotopic organization of striate cortex is well predicted by surface topology. *Current Biology*, **22**(21):2081–2085, 2012.

[16] Bezerra BM and Souto A. Structure and usage of the vocal repertoire of callithrix jacchus. *International Journal of Primatology*, **29**(3):671–701, 2008.

[17] Bonin V, Histed MH, Yurgenson S, and Reid RC. Local diversity and fine-scale organization of receptive fields in mouse visual cortex. *The Journal of Neuroscience*, **31**(50):18506–18521, 2011.

[18] Boyle MP, Bernard A, Thompson CL, Ng L, Boe A, Mortrud M, Hawrylycz MJ, Jones AR, Hevner RF, and Lein ES. Cell-type-specific consequences of reelin deficiency in the mouse neocortex, hippocampus, and amygdala. *Journal of Comparative Neurology*, **519**(11):2061–2089, 2011.

[19] Brodmann K. *Vergleichende Lokalisationslehre der Grosshirnrinde: in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Ja Barth, 1909.

[20] Carandini M. What simple and complex cells compute. *The Journal of physiology*, **577**(2):463–466, 2006.

[21] Cheung SW, Bedenbaugh PH, Nagarajan SS, and Schreiner CE. Functional organization of squirrel monkey primary auditory cortex: responses to pure tones. *Journal of Neurophysiology*, **85**(4):1732–1749, 2001.

[22] Chi T and Shamma S. NSL Matlab Toolbox. `http://www.isr.umd.edu/Labs/NSL/Software.htm`, 2003.

[23] Chomsky N. *Syntactic Structures*. Mouton, 1957.

[24] Chomsky N. *Aspects of the Theory of Syntax*. MIT press, 1965.

[25] Chomsky N. *The Minimalist Program*, volume 28 of *Current Studies in Linguistics Series*. MIT Press, 1995.

[26] Cinque G. A note on linguistic theory and typology. *Linguistic Typology*, **11**(1):93–106, 2007.

[27] Dahl GE, Yu D, Deng L, and Acero A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, **20**(1):30–42, 2012.

[28] DeAngelis GC, Ohzawa I, and Freeman RD. Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. ii. linearity of temporal and spatial summation. *Journal of Neurophysiology*, **69**(4):1118–1135, 1993.

[29] Douglas RJ and Martin KAC. Neuronal circuits of the neocortex. *Annual Review of Neuroscience*, **27**:419–451, 2004.

[30] Douglas RJ, Martin KAC, and Whitteridge D. A canonical microcircuit for neocortex. *Neural Computation*, **1**(4):480–488, 1989.

[31] Ehret G and Riecke S. Mice and humans perceive multiharmonic communication sounds in the same way. *Proceedings of the National Academy of Sciences*, **99**(1):479–482, 2002.

[32] Evans EF, Ross HF, and Whitfield IC. The spatial distribution of unit characteristic frequency in the primary auditory cortex of the cat. *The Journal of Physiology*, **179**(2):238–247, 1965.

[33] Fastl H and Stoll G. Scaling of pitch strength. *Hearing Research*, **1**(4):293–301, 1979.

[34] Fishman YI, Micheyl C, and Steinschneider M. Neural representation of harmonic complex tones in primary auditory cortex of the awake monkey. *The Journal of Neuroscience*, **33**(25):10312–10323, 2013.

[35] Fukushima K. Neural network model for a mechanism of pattern recognition unaffected by shift in position —neocognitron—. *The Transactions of the Institute of Electronics and Communication Engineers of Japan. A*, **62**(10):658–665, 1979.

[36] Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, **36**(4):193–202, 1980.

[37] Fukushima K. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, **1**(2):119–130, 1988.

[38] Gao P, Sultan KT, Zhang XJ, and Shi SH. Lineage-dependent circuit assembly in the neocortex. *Development*, **140**(13):2645–2655, 2013.

[39] Garey LJ. *Brodmann' s localisation in the cerebral cortex*. London: Smith-Gordon, 1994.

[40] Gentner D. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, **7**(2):155–170, 1983.

[41] Goldstein Jr MH, Abeles M, Daly RL, and McIntosh J. Functional architecture in cat primary auditory cortex: tonotopic organization. *Journal of Neurophysiology*, **33**(1):188–197, 1970.

[42] Greenberg JH. A quantitative approach to the morphological typology of language. *International Journal of American Linguistics*, pages 178–194, 1960.

[43] Hawrylycz M, Bernard A, Lau C, Sunkin SM, Chakravarty MM, Lein ES, Jones AR, and Ng L. Areal and laminar differentiation in the mouse neocortex using large scale gene expression data. *Methods*, **50**(2):113–121, 2010.

[44] Hillis WD and Steele Jr GL. Data parallel algorithms. *Communications of the ACM*, **29**(12):1170–1183, 1986.

[45] Hinton GE and Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science*, **313**(5786):504–507, 2006.

[46] Hromádka T, DeWeese MR, and Zador AM. Sparse representation of sounds in the unanesthetized auditory cortex. *PLoS Biology*, **6**(1):e16, 2008.

[47] Hubel DH and Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, **160**(1):106–154, 1962.

[48] Hubel DH and Wiesel TN. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, **195**(1):215–243, 1968.

[49] Hyvärinen A and Hoyer PO. A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research*, **41**(18):2413–2423, 2001.

[50] Hyvärinen A, Hoyer PO, and Inki M. Topographic independent component analysis. *Neural Computation*, **13**(7):1527–1558, 2001.

[51] Hyvärinen A, Hurri J, and Hoyer PO. *Natural Image Statistics: A probabilistic approach to early computational vision*. Springer-Verlag London Ltd., 2009.

[52] Hyvärinen A, Hurri J, and Väyrynen J. Bubbles: a unifying framework for low-level statistical properties of natural image sequences. *Journal of the Optical Society of America A*, **20**(7):1237–1252, 2003.

[53] International Phonetic Association. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press, 1999.

[54] Jones BS, Harris DHR, and Catchpole CK. The stability of the vocal signature in phee calls of the common marmoset, callithrix jacchus. *American Journal of Primatology*, **31**(1):67–75, 1993.

[55] Jones JP and Palmer LA. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, **58**(6):1233–1258, 1987.

[56] Kadia SC and Wang X. Spectral integration in A1 of awake primates: Neurons with single- and multipeaked tuning characteristics. *Journal of Neurophysiology*, **89**(3):1603–1622, 2003.

[57] Kandel E, Schwartz J, and Jessell T. Principles of neural science. 2000.

[58] Karklin Y, Ekanadham C, and Simoncelli EP. Hierarchical spike coding of sound. In *Advances in Neural Information Processing Systems*, volume 25, pages 3041–3049, 2012.

[59] Karklin Y and Lewicki MS. A hierarchical Bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. *Neural Computation*, **17**(2):397–423, 2005.

[60] Karklin Y and Lewicki MS. Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, **457**(7225):83–86, 2009.

[61] Kilgard MP and Merzenich MM. Cortical map reorganization enabled by nucleus basalis activity. *Science*, **279**(5357):1714–1718, 1998.

[62] Klein DJ, Konig P, and Kording KP. Sparse spectrotemporal coding of sounds. *EURASIP Journal on Applied Signal Processing*, **2003**(7):659–667, 2003.

[63] Krizhevsky A, Sutskever I, and Hinton G. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1106–1114, 2012.

[64] Kudoh M and Shibuki K. Importance of polysynaptic inputs and horizontal connectivity in the generation of tetanus-induced long-term potentiation in the rat auditory cortex. *The Journal of Neuroscience*, **17**(24):9458–9465, 1997.

[65] Kuffler SW. Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, **16**(1):37–68, 1953.

[66] Le QV, Ranzato M, Monga R, Devin M, Chen K, Corrado GS, Dean J, and Ng AY. Building high-level features using large scale unsupervised learning. In Langford J and Pineau J, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML 2012, pages 81–88, New York, NY, USA, July 2012. Omnipress.

[67] LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, and Jackel LD. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, **1**(4):541–551, 1989.

[68] Lewicki MS. Efficient coding of natural sounds. *Nature Neuroscience*, **5**(4):356–363, 2002.

[69] Li Y, Lu H, Cheng PL, Ge S, Xu H, Shi SH, and Dan Y. Clonally related visual cortical neurons show similar stimulus feature selectivity. *Nature*, **486**(7401):118–121, 2012.

[70] Ma L and Zhang L. A hierarchical generative model for overcomplete topographic representations in natural images. In *International Joint Conference on Neural Networks (IJCNN2007)*, pages 1198–1203. IEEE, 2007.

[71] Marr D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc. New York, NY, USA, 1982.

[72] McDermott JH and Simoncelli EP. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron*, **71**(5):926–940, 2011.

[73] Meddis R, Popper AN, Fay RR, and Lopez-Poveda EA. *Computational Models of the Auditory System*, volume 35 of *Springer Handbook of Auditory Research*. Springer, 2010.

[74] Miller LM, Escabı MA, Read HL, and Schreiner CE. Functional convergence of response properties in the auditory thalamocortical system. *Neuron*, **32**(1):151–160, 2001.

[75] Miller LM, Escabı MA, and Schreiner CE. Feature selectivity and interneuronal cooperation in the thalamocortical system. *The Journal of Neuroscience*, **21**(20):8136–8144, 2001.

[76] Miller LM and Schreiner CE. Stimulus-based state control in the thalamocortical system. *The Journal of Neuroscience*, **20**(18):7011–7016, 2000.

[77] Mochizuki Y and Shinomoto S. Analog and digital codes in the brain. *Physical Review E*, **89**:022705, 2014.

[78] Mohamed A, Dahl G, and Hinton G. Deep belief networks for phone recognition. In *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, 2009.

[79] Moore BCJ. *An introduction to the psychology of hearing*. London: Emerald Group Publishing Ltd., 5th edition, 2003.

[80] Mountcastle VB. Introduction. computation in cortical columns. *Cerebral Cortex*, **13**(1):2–4, 2003.

[81] Nakashima A, Yamada Y, and Kuniyoshi Y. Uterine environment guides organization of somatosensory area: a computational approach. In *Proceedings of Humanoids 2012 Workshop on Developmental Robotics: Can developmental robotics yield human-like cognitive abilities?*, pages 44–46, 2012.

[82] Ng L, Bernard A, Lau C, Overly CC, Dong HW, Kuan C, Pathak S, Sunkin SM, Dang C, Bohland JW, et al. An anatomic gene expression atlas of the adult mouse brain. *Nature Neuroscience*, **12**(3):356–362, 2009.

[83] Ng L, Lau C, Sunkin SM, Bernard A, Chakravarty MM, Lein ES, Jones AR, and Hawrylycz M. Surface-based mapping of gene expression and probabilistic expression maps in the mouse cortex. *Methods*, **50**(2):55–62, 2010.

[84] Ohki K, Chung S, Ch'ng YH, Kara P, and Reid RC. Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. *Nature*, **433**(7026):597–603, 2005.

[85] Ohki K, Chung S, Kara P, Hübener M, Bonhoeffer T, and Reid RC. Highly ordered arrangement of single neurons in orientation pinwheels. *Nature*, **442**(7105):925–928, 2006.

[86] Ohtsuki G, Nishiyama M, Yoshida T, Murakami T, Histed M, Lois C, and Ohki K. Similarity of visual selectivity among clonally related neurons in visual cortex. *Neuron*, **75**(1):65–72, 2012.

[87] Olshausen BA and Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, **381**(6583):607–609, 1996.

[88] Olshausen BA and Field DJ. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, **37**(23):3311–3325, 1997.

[89] Olshausen BA and Field DJ. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, **14**(4):481–487, 2004.

[90] Osindero S, Welling M, and Hinton GE. Topographic product models applied to natural scene statistics. *Neural Computation*, **18**(2):381–414, 2006.

[91] Osmanski MS and Wang X. Measurement of absolute auditory thresholds in the common marmoset (callithrix jacchus). *Hearing Research*, **277**(1–2):127–133, 2011.

[92] Peters G, Baum L, Peters MK, and Tonkin-Leyhausen B. Spectral characteristics of intense mew calls in cat species of the genus felis (mammalia: Carnivora: Felidae). *Journal of Ethology*, **27**(2):221–237, 2009.

[93] Petkov CI, Kayser C, Steudel T, Whittingstall K, Augath M, and Logothetis NK. A voice region in the monkey brain. *Nature Neuroscience*, **11**(3):367–374, 2008.

[94] Pi HJ, Hangya B, Kvitsiani D, Sanders JI, Huang ZJ, and Kepecs A. Cortical interneurons that specialize in disinhibitory control. *Nature*, **503**(7477):521–524, 2013.

[95] Qin L, Sakai M, Chimoto S, and Sato Y. Interaction of excitatory and inhibitory frequency-receptive fields in determining fundamental frequency sensitivity of primary auditory cortex neurons in awake cats. *Cerebral Cortex*, **15**(9):1371–1383, 2005.

[96] Ramat P. Typological comparison: towards a historical perspective. *Approaches to Language Typology*, pages 27–48, 1995.

[97] Read HL, Winer JA, and Schreiner CE. Modular organization of intrinsic connections associated with spectral tuning in cat auditory cortex. *Proceedings of the National Academy of Sciences*, **98**(14):8042–8047, 2001.

[98] Real Acoustic Environments Working Group, Real World Computing Partnership. RWCP Sound Scene Database in Real Acoustic Environment. `http://tosa.mri.co.jp/sounddb/indexe.htm`, 2000.

[99] Reale RA, Brugge JF, and Feng JZ. Geometry and orientation of neuronal processes in cat primary auditory cortex (AI) related to characteristic-frequency maps. *Proceedings of the National Academy of Sciences*, **80**(17):5449–5453, 1983.

[100] Ringach D and Shapley R. Reverse correlation in neurophysiology. *Cognitive Science*, **28**(2):147–166, 2004.

[101] Rockel AJ, Hiorns RW, and Powell TPS. The basic uniformity in structure of the neocortex. *Brain*, **103**(2):221–244, 1980.

[102] Roelfsema PR, van Ooyen A, and Watanabe T. Perceptual learning rules based on reinforcers and attention. *Trends in Cognitive Sciences*, **14**(2):64–71, 2010.

[103] Rothschild G, Nelken I, and Mizrahi A. Functional organization and population dynamics in the mouse primary auditory cortex. *Nature Neuroscience*, **13**(3):353–360, 2010.

[104] Saenz M and Langers DRM. Tonotopic mapping of human auditory cortex. *Hearing Research*, **307**:42–52, 2014.

[105] Sasaki Y, Nanez JE, and Watanabe T. Advances in visual perceptual learning and plasticity. *Nature Reviews Neuroscience*, **11**(1):53–60, 2009.

[106] Saxe AM, Bhand M, Mudur R, Suresh B, and Ng AY. Unsupervised learning models of primary cortical receptive fields and receptive field plasticity. In Shawe-Taylor J, Zemel RS, Bartlett P, Pereira FCN, and Weinberger KQ, editors, *Advances in Neural Information Processing Systems 24 (NIPS2011)*, pages 1971–1979, 2011.

[107] Schreiner CE, Read HL, and Sutter ML. Modular organization of frequency integration in primary auditory cortex. *Annual Review of Neuroscience*, **23**(1):501–529, 2000.

[108] Schreiner CE and Winer JA. Auditory cortex mapmaking: principles, projections, and plasticity. *Neuron*, **56**(2):356–365, 2007.

[109] Seide F, Li G, and Yu D. Conversational speech transcription using context-dependent deep neural networks. In *Interspeech 2011*, pages 437–440, 2011.

[110] Selfridge OG. Pandemonium: a paradigm for learning. In Blake DV and Uttley AM, editors, *Proceedings of the Symposium on Mechanisation of Thought Processes*, pages 511–529, 1959.

[111] Sharma J, Angelucci A, and Sur M. Induction of visual orientation modules in auditory cortex. *Nature*, **404**(6780):841–847, 2000.

[112] Shechter B and Depireux DA. Nonlinearity of coding in primary auditory cortex of the awake ferret. *Neuroscience*, **165**(2):612–620, 2010.

[113] Shinomoto S, Kim H, Shimokawa T, Matsuno N, Funahashi S, Shima K, Fujita I, Tamura H, Doi T, Kawano K, et al. Relating neuronal firing patterns to functional differentiation of cerebral cortex. *PLoS Computational Biology*, **5**(7):e1000433, 2009.

[114] Smith EC and Lewicki MS. Efficient auditory coding. *Nature*, **439**(7079):978–982, 2006.

[115] Smith SL and Häusser M. Parallel processing of visual space by neighboring neurons in mouse visual cortex. *Nature Neuroscience*, **13**(9):1144–1149, 2010.

[116] Song JJ, editor. Oxford University Press, 2010.

[117] Sur M, Garraghty P, and Roe AW. Experimentally induced visual projections into auditory thalamus and cortex. *Science*, **242**(4884):1437–1441, 1988.

[118] Takahashi H, Yokota R, and Kanzaki R. Response variance in functional maps: Neural darwinism revisited. *PloS ONE*, **8**(7):e68705, 2013.

[119] Terashima H and Hosoya H. Sparse codes of harmonic natural sounds and their modulatory interactions. *Network: Computation in Neural Systems*, **20**(4):253–267, 2009.

[120] Terashima H and Hosoya H. Sparse codes of harmonic sound and their interaction explain harmony-related response of auditory cortex. *BMC Neuroscience*, **11**(Suppl 1):O19, 2010.

[121] Terashima H, Hosoya H, Tani T, Ichinohe N, and Okada M. Sparse coding of harmonic vocalization in monkey auditory cortex. *Neurocomputing*, **103**:14–21, 2013.

[122] Terashima H and Okada M. The topographic unsupervised learning of natural sounds in the auditory cortex. In *Advances in Neural Information Processing Systems 25*, pages 2321–2329, 2012.

[123] Ting LH. Dimensional reduction in sensorimotor systems: a framework for understanding muscle coordination of posture. *Progress in Brain Research*, **165**:299–321, 2007.

[124] Ting LH and McKay JL. Neuromechanics of muscle synergies for posture and movement. *Current Opinion in Neurobiology*, **17**(6):622–628, 2007.

[125] Tsukano H, Kubota Y, Komagata S, Hishida R, Kudoh M, and Shibuki K. Experience-dependent formation of cortical responses to "missing" fundamentals during presentation of harmonic sounds in the mouse auditory cortex. program no. 851.2. In *2008 Neuroscience Meeting Planner*, Washington, DC, USA, 2008. Society for Neuroscience.

[126] Turner RE. *Statistical models for natural sounds*. PhD thesis, UCL (University College London), 2010.

[127] Van den Bergh G, Zhang B, Arckens L, and Chino YM. Receptive-field properties of V1 and V2 neurons in mice and macaque monkeys. *The Journal of Comparative Neurology*, **518**(11):2051–2070, 2010.

[128] van Hateren JH and van der Schaaf A. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **265**(1394):359–366, 1998.

[129] von Humboldt CWF. *Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluss auf die geistige Entwicklung des Menschengeschlechts*. Dr. d. Kgl. Akad. d. Wiss., 1836.

[130] Wandell BA, Dumoulin SO, and Brewer AA. Visual field maps in human cortex. *Neuron*, **56**(2):366–383, 2007.

[131] Wang X. On cortical coding of vocal communication sounds in primates. *Proceedings of the National Academy of Sciences*, **97**(22):11843–11849, 2000.

[132] Wang X and Kadia SC. Differential representation of species-specific primate vocalizations in the auditory cortices of marmoset and cat. *Journal of Neurophysiology*, **86**(5):2616–2620, 2001.

[133] Wang X, Merzenich MM, Beitel R, and Schreiner CE. Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: temporal and spectral characteristics. *Journal of Neurophysiology*, **74**(6):2685–2706, 1995.

[134] Whaley LJ. *Introduction to typology: the unity and diversity of language*. Sage, 1997.

[135] Wheeler JA and Gearhart M. FORUM: From the big bang to the big crunch. *Cosmic Search*, **1**(4):2–8, 1979.

[136] Winkowski DE and Kanold PO. Laminar transformation of frequency organization in auditory cortex. *The Journal of Neuroscience*, **33**(4):1498–1508, 2013.

[137] Yotsumoto Y and Watanabe T. Defining a link between perceptual learning and attention. *PLoS Biology*, **6**(8):e221, 2008.

[138] Yu YC, Bultje RS, Wang X, and Shi SH. Specific synapses develop preferentially among sister excitatory neurons in the neocortex. *Nature*, **458**(7237):501–504, 2009.

[139] Yu YC, He S, Chen Y, S.and Fu, Brown KN, Yao XH, Ma J, Gao KP, Sosinsky GE, Huang K, et al. Preferential electrical coupling regulates neocortical lineage-dependent microcircuit assembly. *Nature*, **486**(7401):113–117, 2012.

[140] Zilles K and Amunts K. Centenary of Brodmann's map – conception and fate. *Nature Reviews Neuroscience*, **11**(2):139–145, 2010.

List of Publications

Peer reviewed papers

Terashima H, Hosoya H (2009) **Sparse codes of harmonic natural sounds and their modulatory interactions**. *Network: Computation in Neural Systems*, **20**(4):253–267.

Terashima H, Okada M (2012) **The topographic unsupervised learning of natural sounds in the auditory cortex**. *Advances in Neural Information Processing Systems **25** (NIPS2012)*, 2321–2329.

Terashima H, Hosoya H, Tani T, Ichinohe N, Okada M (2013) **Sparse coding of harmonic vocalization in monkey auditory cortex**. *Neurocomputing*, **103**:14–21.

Selected conference presentations

Terashima H, Hosoya H (2010) Sparse codes of harmonic sound and their interaction explain harmony-related response of auditory cortex. *The 19th Annual Computational Neuroscience Meeting (CNS*2010)*, O19, San Antonio, Texas, USA, 24–30 July 2010.

Terashima H, Okada M (2011) An integrated interpretation of adaptive maps in visual and auditory cortices. *The 25th Annual Conference of the Japanese Society for Artificial Intelligence (JSAI2011)*, 2C1-OS2a-8in, Morioka, Iwate, Japan, 1–3 June, 2011.

Terashima H, Okada M (2011) Modelling disordered A1 map and ordered V1 map. *The 21st Annual Conference of the Japanese Neural Network Society (JNNS2011)*, P2-32, Okinawa, Japan, 15–17 December 2011.

Terashima H, Okada M (2012) V1 and A1 maps: different topographies, a common organizing principle. *The 9th Computational and Systems Neuroscience meeting (Cosyne 2012)*, II-73, Salt Lake City, Utah, USA, 23–26 February 2012.

Terashima H, Okada M (2012) The topographic unsupervised learning of natural sounds in the auditory cortex. *Neural Information Processing System 2012 (NIPS2012)*, Th88, Lake Tahoe, Nevada, USA, 3–6 December 2012.

Terashima H, Okada M (2013) A computational discussion on "complex cells" of the auditory cortex. *The 27th Annual Conference of the Japanese Society for Artificial Intelligence (JSAI2013)*, 3H3-OS-05b-7in, Toyama, Japan, 4–7 June, 2013.

List of Awards

1. **Organization for Computational Neuroscience Student Travel Award**
   The 19th Annual Computational Neuroscience Meeting (CNS*2010), San Antonio, Texas, USA, July, 2010.

2. **SNSS2011 Data Analysis Challenge Award**
   Systems Neurobiology Spring School 2011 (SNSS2011), Kyoto, Japan, March, 2011.

3. **JSAI Annual Conference Award**
   The 25th Annual Conference of the Japanese Society for Artificial Intelligence (JSAI2011), Morioka, Japan, June, 2011.

4. **JNNS Young Presenter Award and Student Travel Support**
   The 21st Annual Conference of the Japanese Neural Network Society (JNNS2011), Okinawa, Japan, December, 2011.

5. **Outstanding Poster Award**
   Tohoku Winter School for Neuroscience, Miyagi, Japan, February, 2012.

6. **JSAI Annual Conference Student Incentive Award**
   The 27th Annual Conference of the Japanese Society for Artificial Intelligence (JSAI2013), Toyama, Japan, June, 2013.