Functional Sensory Representations of Natural Stimuli: The Case of Spatial Hearing

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr.rer.nat.)

im Fachgebiet

Informatik

vorgelegt

von M.Sc. Wiktor Młynarski
geboren am 01.10.1986 in Kraków (Polen)

Die Annahme der Dissertation wurde empfohlen von:

1. Professor Dr. Jürgen Jost (MPIMN, Leipzig)
2. Professor Dr. Joshua McDermott (MIT, Cambridge, USA)

Die Verleihung des akademischen Grades erfolgt mit Bestehen der Verteidigung am 21.01.2015 mit dem Gesamtprädikat summa cum laude.

W nieustannym poszukiwaniu Białego Patyka,
Rodzicom dedykuję

Contents

It is the brain, the little gray cells on which one must rely.
One must seek the truth within - not without.

Hercules Poirot

Acknowledgements

I read once that science is only rarely a story of tremendously brilliant individuals who "got it right" in instant sparks of genius. It is rather a story of error, a crucial part of the creative human activity, and it involves a "broad social context". Being much more a fan of various procrastination activities than a "tremenduously brilliant individual", I was incredibly lucky to find myself in an "appropriate social context" at an appropriate time, which resulted in this thesis. Here, I would like to thank all those who contributed to it.

Firstly, many thanks to my supervisor Jürgen Jost. Thank you for giving me the credit of the doubt, and accepting me as a member of your group - a very close approximation of an ideal academic environment, full of intellectual freedom. I am also very grateful to Rudolf Rübsamen - my co-supervisor. Thank you for showing me the world of experimental neurobiology from a perspective of a true naturalist. It is an experience, every theoretician should have at least once. My studies in Leipzig would not be possible without Andreas Reichenbach - the founder and co-director of the InterNeuro graduate college, of which I was the part for three years. Your passion and devotion for science (Müller cells!), as well as humour are contagious - thank you for spreading them so actively!

Ideas described in this thesis are also a product of interactions with many colleagues, both in Leipzig and in other places. Timm Lochmann, a great discussion partner, who has selflessly proofread drafts of my papers and early versions of the thesis. Thank you for long hours of discussions (in different parts of the world), which resulted in foundation of (not yet well known) Reudnitz Institute for Computational Philosophy. Many thanks to Nils Bertschinger (also a careful proofreader), Philipp Benner and Lilya Avdiyenko for the time spent listening to, and commenting on my rambles at one of the Institute's tea tables or in front of one of the blackboards. I also want to mention Bernhard Englitz, my InterNeuro predecessor - thanks for sharing with me your deep insights into the inner workings of the world of science, inviting me to Paris and time spent to talk about the auditory system. A collective expression of grattitude goes to everyone, who shared their time with me and were kind enough to provide me with interesting

comments and suggestions: Josh McDermott, Charles Cadieu, Yan Karklin, Urs Köster, Bruno Olshausen, Michael Lewicki, Elias Issa, Guido Montúfar, Eckehard Olbrich and many others, to whom I apologize if I failed to put their names here. Thank you!

During my time in Leipzig I was lucky to have an opportunity to work on experimental projects, which are not mentioned in this thesis. This was done in collaboration with members of Rudolf Rübsamen's group: Claudia Freigang, Jan Bennemann, Marc Stöhr, Mikaella Sarrou. For me - it was a great science lesson, eye opening experience and (crucially) lots of fun! Thank you all for many hours spent in the acoustic free field, discussions and draft corrections!

It is not an exaggeration, if I say that I would not manage to do anything constructive in Leipzig, if I was not taken care of in almost all not-directly-scientific aspects of life by the Institute's staff. A thousand thanks to Antje Vandenberg for not only managing the administration of the group in a way that makes the bureaucracy invisible to us, but (perhaps most importantly) for creating a warm and joyful atmosphere. I will miss our mensa lunches greatly! Thanks to Heike Rackwitz for her help in finding a roof over my head and keeping it in one piece. Many thanks to Ingo Brüggemann, Katarzyna Bajer, Christine Breitschopf and the rest of the library staff for creating the most advanced and user-friendly library I have ever seen.

During the last four years, my social life has been also greatly boosted by presence of all the colleagues and friends from the Institute: Yangjing, Leonhard, Tobias, Ilona, Frank, Gerardo, Enno, Pierre-Yves, Stephan, Martin, Özge, Christian, Camilo, Vasilis, Guido, Yuri, Felix, Anna, Georg, Jörg, Sylvia, Marius, Sophia, Damián. Also my science-related flatmates Lorenzo and Alexander made this time (positively) memorable. Thank you all!

Somewhat paradoxically, I want to devote the last paragrapgh of this acknowledgements to thank those with whom everything began - my Family. My Parents - Kasia and Kajetan for triggering and nursing the curiosity. For showing me how interesting the natural world is. For endless discussions and explanations, which began when I learned to speak, which continue till now, and which had a profound scientific impact on the shape of this thesis. Thank you! To my sister - Maria - for your (often sarcastic) words of wisdom, which keep me at the right track, and make me question my principles. And finally, a thank-you kiss to Sandra - the happiest Monkey I have found in the Jungle of Life.

Chapter 1

Introduction

The brain does not "see" the world the way its owner does. It exists concealed in the silent and dark space confined by the skull (or other anatomical structure) leading the animal through its surrounding. Every movement is planned and exectuted basing on the information present in physical stimuli - light, sound etc, but represented by series of electric potentials generated by sensory neurons.

Understanding the way in which the environment reflects itself in the neuronal activity is perhaps one of the primary goals of neurobiology. Over time sensory neuroscience - the subfield specifically devoted to this problem has developed. Using a broad repertoire of experimental and theoretical approaches it attempts to find answers to the great question - why and how do animals perceive the world in a way they do?

1.1 The notion of neuronal function

Following the great research tradition, which can be dated back to Gustav Fechner, Hermann von Helmholtz and Edgar Adrian, sensory neuroscience attempts to quantitatively characterize the relationship between properties of the stimulus, neuronal activity and perception. This at first very courageous and unusual thought that the way one perceives the world emerges from the electric activity of "little gray cells" has lead to a largely successful research program. Still however, many fundamental questions remain unanswered.

At the end of the XIXth and at the beginning of the XXth century it has been observed that the electric activity of nerve fibers can be triggered by specific properties of the environment. Adrian has accidentally discovered that when he walked in the toad's field of vision the optic nerve was eliciting electric pulses [2]. Basing on physical considerations Helmholtz proposed that nerve cells located at

different positions along the cochlea decompose complex sounds into pure tones [54]. Activity of each cell would therefore underlie the perception of pitch.

Research into the function of the nervous system started gaining pace in the second half of the XXth century. Results, which are now considered the foundation of visual neuroscience, have been obtained by Hubel and Wiesel in 1959 [59]. After arduous testing the visual cortex of the cat by exposing the animal to numerous stimuli types they observed a previously unknown effect. Neurons located in early visual areas seemed to be responsive to light bars of a particular orientation. Depending on the degree of invariance to the spatial shift of the stimulus, those cells have been named simple and complex. Since in laboratory conditions they increased firing rates in the presence of edge-like structures their function has been decided to be "edge detection".

Almost in parallel interesting discoveries were made in the frog's visual system. Horace Barlow in Cambridge demonstrated that neurons in the retina of the frog respond to presence or lack of black dots [10]. In the frog's world fast moving round black objects often corespond to the presence of food - a fly. Considering this behavioural importance Barlow called discovered cells "fly detectors" automatically ascribing them a particular function.

The observation that neuronal populations seem to extract information about different aspects of the environment from the raw stimulus input provoked theoretical considerations. Polish neurophysiologist Jerzy Konorski working in Łódź combined many results arising at the time into a coherent conceptual framework [70]. He suggested that neurons hierarchically extract and represent more and more abstract properties of the stimulus in a processing cascade. According to the Konorski's hypothesis the highest level of the processing hierarchy would consist of "gnostic units" - cells which represent abstract concepts. Not much later Jeremy Lettvin in Boston developed a similar concept. He named it, however with a dose of flamboyant humour - "the grandmother cells" (Lettvin introduced the term in 1969 during the course "*Biological Foundations for Perception and Knowledge*" taught at MIT [48]). This name precisely defines the function such neurons are expected to implement - they would elicit electric pulses only in the presence of the individual's grandmother. Years later experimental results supporting Konorski's and Lettvins predictions have been delivered. Neurons modulated by the identity of a person depicted on an image have been found both in humans [113] and monkeys [32]. One mismatch with the Lettvin's theory has been that it was not the grandmother, who influenced the neuronal activity of recorded units. It was the actress - Jennifer Aniston [113].

From edges to faces. From pure tones to the musical rhythm. Neurons in different parts of the nervous system seem to extract and represent very different, but often complex and subtle properties of the stimulus, while being non-responsive to changes in other parameters. The aspect of the physical world,

which is being made explicit by the activity of the neuron is often associated with this neuron's *function*. Edge detectors, face cells, pitch-sensitive neurons, sound-source localizers. In an emerging view sensory systems consist of basic units of clearly segregated and well defined functions. In some cases existence of such segregation is firstly being hypothesized, with no precise definition of how the separation should be performed. A prominent example comes from visual neuroscience, where dorsal and ventral streams are supposed to independently process spatial ("where") and identity-related ("what") information.

The postulated concept of sensory segregation is not free from theoretical limitations. Firstly, computational mechanisms which underlie neuronal functions (either hypothesized or experimentally observed ones) are often hard to understand or imagine. How is the grandmother's identity extracted from an image? How is a position of a sound source separated from its timbre? It is not clear whether those questions reflect just our limited algorithmic knowledge required to perform described computations, or are of a more fundamental nature. The second (perhaps largest) conceptual drawback of the notion of functional sensory segregation is the well-known "binding problem". Let's imagine one listens to a jazz quartet. The ears perceive overlapping waves of air pressure generated by four instruments. Auditory neurons process the music of each instrument in separated channels devoted to pitch, timbre, spatial location, and other perhaps unknown, or unnamed features of sound. If such a strict segregation happens, how is the information fused together correctly to form a percept? Why does one not perceive a piano playing a melody of a cello, located where the drums stand, rather than the real quartet?

Taking above mentioned issues into consideration, one can ask broadly - does the nervous system consist of a loose collection of functional subsystems, each devoted and "pre-designed" to have a separate function and process a different stimulus aspect? From a theoretical point of view this may lead to a more practical problem - which conceptual frameworks and theories can be useful in functional characterization of sensory neurons? And the big question - what is *the function* implemented by sensory neurons - what do they actually do?

1.2 Neuronal function in the laboratory - the experimenter's point of view

Perhaps the earliest and most broadly applied approach to characterization of neuronal function, can be exemplified by already mentioned work of Adrian, Hubel and Wiesel among many others. The experimenter pre-assumes that the function of a neuron is to represent a particular physical parameter ϕ say the angular position of a sound source in the head entered coordinate system. The animal is exposed to a range of parameter values, while the neuronal activity is

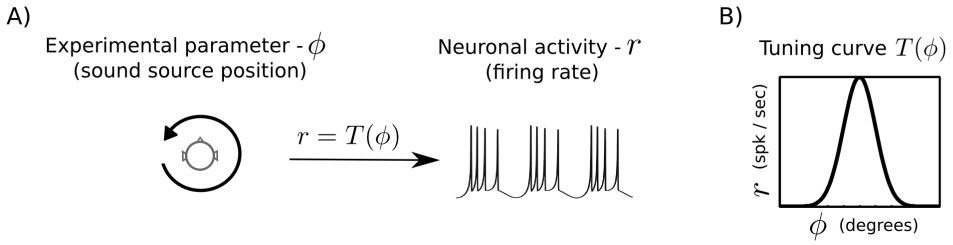

Figure 1.1: A classical approach to the neuronal function characterization. A) Variation of an experimental parameter ϕ leads to a modification of neuronal activity r (e.g. firing rate). B) A plot of a function T, which maps the parameter in question to activity value is often referred to as "the tuning curve"

being recorded. In the following step the experimenter computes a feature r of the electrophysiological signal (very often it is the firing rate - an average amount of action potentials elicited by the neuron in a time interval). Collected data is used to estimate the mapping from the stimulus parameter to the neurophysiological activity, as implemented by the sensory neuron[1], which is known as the *tuning curve*. The tuning curve T is therefore a function:

$$T(\phi) = r \tag{1.1}$$

A schema of this procedure is depicted on figure 1.1. If a modification of the parameter ϕ systematically triggers a change of the neuronal response r, one may draw a conclusion that the neuron *represents* this particular parameter. The form of the representation is defined by the shape of the tuning curve.

This research philosophy has tremendously advanced our understanding of the nervous system. However, as with any approach, among numerous advantages it has certain drawbacks. At the most fundamental level stimuli live in the high-dimensional "natural space". Its dimensionality is defined by the number of sensory receptors (cells transforming physical stimuli into the electric activity) in the sensory epithelium - in the human auditory system this corresponds to roughly 20000 hair cells per cochlea. The number of dimensions additionally expands, when one considers temporal change of the stimulus. A sensory neuron implements a mapping from the space of the time-varying stimulus s into the (typically much lower dimensional) space of neuronal activity[2]r:.

[1]One should note that in a more general (and also broadly used) setting the mapping of the stimulus on spiking activity is not deterministic. Due to different sources of noise and uncertainty the relationship is defined by a conditional probability distribution $p(r|\phi)$. For simplicity however, here I use the deterministic notation.

[2]It is important to stress that at most stages of the processing hierarchy down-stream neurons operate on inputs from up-stream neurons not on the actual stimulus. They are however characterized by the function they play in the stimulus processing, or a feature they repre-

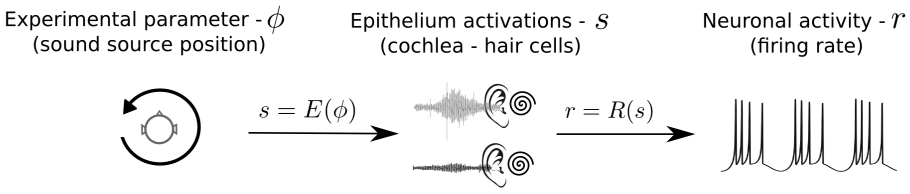

Figure 1.2: An actual mapping from the parameter space to the neuronal activity. The preselected experimental parameter ϕ is mapped to the space of epithelium activations s by the function $E(\phi)$. In a most fundamental sense neuronal responses r are defined in the epithelium space by the function $R(s) = r$. The mapping between the experimental parameter and the neuronal activity is therefore defined by a composition of two functions $r = R(E(\phi))$.

$$R(s) = r \tag{1.2}$$

The function R can be thought of as a conceptualization of the neuron's *true* receptive field i.e. the computation that phylogeny, ontogeny and learning have led it to perform. It of course remains unknown to the experimenter, and full understanding of the mapping described by equation 1.2 remains one of the "holy-grails" of sensory neuroscience.

When analysing the sensitivity of a neuron to the preselected parameter ϕ, one gradually changes its value, and observes a systematic change in the neural response. The important, yet subtle caveat is that observed modulation is defined by the receptive field R not a tuning curve[3] T. Modulation of the parameter ϕ (for instance the angular position of a sound source) yields a physical stimulus (a waveform), which excites sensory receptors generating the sensory signal s (spatiotemporal activation of left and right cochlea). The mapping between ϕ and s is defined by another function E. Modulation of neuronal firing in a response to the change of ϕ is therefore described by a composition of two functions:

$$r = R(E(\phi)) \tag{1.3}$$

This situation is depicted on figure 1.2. The experimental approach described at the beginning of this section, and depicted on figure 1.1 can easily overlook the

sent. The function of previously mentioned face-specific cells is defined by the properties of the represented stimulus not spike trains received from the lower parts of the visual system.

[3]The terms "receptive field" and "tuning curve" are not strictly defined and are often used as equivalent in the literature. In this discussion I use them separately. The receptive field R is the real mapping from the stimulus space into neuronal activity that a neuron implements. The tuning curve T is an experimentally obtained estimate of a mapping of a pre-defined parameter ϕ to neuronal activity.

existence of the intermediate function E. What it does instead of approximating R with T, is an approximation of $R(E)$. Function R implemented by a neuron can be very different from the experimentally observed tuning curve T. Yet, if different responses to a change of the parameter ϕ are experimentally obtained, it may be erroneously concluded that the sole function of the observed neuron is to represent the value of ϕ. Briefly speaking - characterizing neuronal response to the change of a certain parameter is not necessarily equivalent to determining this neuron's function.

The conceptual problems discussed above are closely related to specific issues encountered in auditory neurophysiology. A particularly important one is the sensitivity-selectivity dichotomy. Neurons which respond exclusively to one parameter and do not carry any information about any other stimulus features are called *stimulus-specific*. Specificity of a unit allows to draw a strong conclusion about its supposed function. On the other hand it is possible that the neuron is merely *sensitive* to a particular parameter i.e. it non-exclusively responds to a variation of multiple parameters, including the one in question. Observed change in the neuronal firing may yield a conclusion that the neuron's function is to represent a single feature only, while in reality it represents numerous other aspects. In the auditory cortex the majority of neurons seem to be sensitive to timbre, pitch and sound position [15], while units selective exclusively to one of those parameters are hardly found [14].

Experimental methodology described in this section does not include more explorative approaches based on the analysis of response conditional ensembles (RCEs). The RCE is a set of stimuli (typically generated by a random process), which preceded the spike elicitation. RCE based methods can be classified depending on the statistic they analyze. Prominent examples are the spike triggered average (STA), spike triggered covariance (STC) [126] and information theoretic methods such as maximally-informative dimensions [128]. These methods do not presuppose sensitivity to any high-level parameter. Instead, they sample the stimulus space and attempt to infer the neuronal mapping R from the observed RCE. For this reason they suffer much less from conceptual drawbacks described here. Their results, however maybe hard to interpret, since the RCE lies in the high-dimensional "epithelium space". I do not discuss them in detail, since they are not broadly used in the analysis of spatial hearing mechanisms.

1.3 Neuronal function in the natural environment - the organism's point of view

The perspective of the organism (or rather its nervous system) differs in fundamental ways from that of the experimenter. It is exposed to a constant stream of sensory information, consisting not of well defined and interpretable parame-

ters but of the raw, high-dimensional stimulus signal exciting sensory receptors. Stimuli are generated by numerous simultaneously active objects (for instance sound sources), which overlap, interact with each other and are affected by the environmental background (acoustics in case of sound). In such a setting it is extremely hard to define clearly separated aspects of the environment. If, for instance, one wants to refer to the sound source position, one has to decide first, which out of many sound sources is of interest. Additional questions are: is it separable from the background? Do other sources of similar quality overlap with it? Is the identity of the sound also important (perhaps it can carry additional information about the sound location)? Real world situations rarely reflect reductionistic and well-controlled experimental settings, where only a single, well interpretable aspect of the experimental setup (not necessarily of the stimulus) is being modulated. Interested reader should refer to an opinion article, which addresses conceptual problems of natural scene analysis [80].

A crucial property of stimuli generated by natural scenes is their *ambiguity*. If an unknown number of sources generates signals of an also unknown structure a situation can arise, where many different scene configurations correspond to the same stimulus value. It is also possible that sensory data do not suffice to find a clear solution. Such problems are known as *ill-posed*. In auditory neuroscience a classical example of an ill-posed problem is known as "the cocktail party problem" [90], where multiple overlapping sound sources collapse on the single stimulus waveform.

Due to ambiguity and presence of multiple noise sources extraction of stimulus features useful in accomplishing meaningful behavioural tasks is an inherently statistical problem. Properties of the environment, which are modelled by simple experimental parameters in the laboratory have to be *inferred* from the sensory stream [109, 110, 83]. To successfully interact with the environment the internal states of the organism (specifically its nervous system) have to be correlated with aspects of a scene relevant for survival, which happens by transforming the raw stimulus stream.

The above considerations raise important questions. In order to remain informed about the environment do neurons need to have clearly defined functions, which are sharply segregated and easily interpretable as experimental parameters? Does the separation into high-level functions defined by the natural language of human observers ("timbre encoding", "sound position encoding") happen at all?

1.4 Outline and scope of this thesis

Prior to answering questions about sensory representations employed by a particular system the level of abstraction at which the analysis will be performed

should be chosen. Very different questions can be asked about the same system - their choice depends in most cases on individual preferences rooted in the background and the research culture of the particular field. Questions naturally asked by a neurophysiologist could be whether the spike timing or rates carry sensory information, or what is the role of inhibitory neurons in sound localization. A statistician would rather ask what are the quantities that the nervous system attempts to estimate and how can this computation be carried out given sensory data.

A conceptual framework, which relates these perspectives to each other was provided by David Marr in his seminal book [87]. Marr has proposed that any information processing system (and the nervous system in particular) should be analyzed at three levels of abstraction. He ordered them from most to least abstract (definitions presented below are direct quotations from [87]):

1. Computational theory - *What is the goal of the computation, why is it appropriate, and what is the logic of strategy it can be carried out?*

2. Representation and algorithm - *How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?*

3. Hardware implementation - *How can the representation and algorithm be realized physically?*

Marr's levels are not fully independent, for instance the choice of an algorithm will be very often constrained by the available hardware. Despite that the hierarchy has proven to be useful to study neural systems. In this thesis which considers the function sensory neurons play in representing sensory information as exemplified by the binaural hearing system I focus on the two highest levels - computational and algorithmic. Even though auditory neurophysiology provides numerous fascinating examples of physical mechanisms (for instance the submilisecond spike coding of interaural phase differences), I will not discuss them. Instead I will consider tasks of spatial hearing from an information-processing perspective.

Following the existing research field, which attempts to connect neuronal function with statistical properties of the natural sensory environment, I will argue for two general, closely related tenets:

1. The function of sensory neurons can not be fully elucidated without understanding statistics of natural stimuli they process. While reductionist experimental designs using artificial stimuli may raise easily interpretable results and provide intuitions they can not suffice to fully elucidate the function of a sensory neuron. Artificial stimuli hardly reflect the complexity of naturally

encountered information which the system evolved to deal with. Simplistic stimuli considered in isolation can lead to too fast conclusions and misconceptions about the function of the system. Finally, one can not predict the richness of the natural environment - artificial stimuli will not include all possible cases faced by the organism. In order to discover them one has to explore natural data.

2. Function of sensory representations is determined by redundancies present in the natural sensory environment. As discussed in the first section of this chapter there are many possible experimental parameters, modulation of which correlates with a change of neuronal activity. Taking the interpretation of such results to the extreme one may conclude that for each ad-hoc defined experimental parameter there should be a subsystem within the brain which represents it. One may, however, counter-argue taking the evolutionary perspective, which stresses the necessity for adaptation to the environment. According to it the nervous system must not consist of a loose collection of "problem-solving circuits". It may be rather encoding correlated structures i.e. *redundancies* present in its natural input. Such structures in turn correspond to interpretable and potentially behaviourally relevant environmental states. This statement is a consequence of a hypothesis known as the *efficient coding hypothesis* or *redundancy reduction* [10, 7].

The argumentation is based on a concrete example. I study statistics of natural stereo sounds and compare them with known properties of the binaural hearing system. This part of the nervous system has been a subject of extensive physiological and psychophysical research, yet not much work has been done from the theoretical perspective presented here.

As mentioned in the second tenet the general theoretical framework I use in this thesis is provided by the efficient coding hypothesis (described in detail in chapter 2). Briefly stated, the hypothesis says that sensory neurons encode redundant stimulus patterns while minimizing mutual dependencies. The hypothesis has been succesfully applied in many domains of sensory neuroscience. Here I refer to it in an attempt to explain certain mechanisms of spatial hearing.

The thesis is structured as follows:

Chapter 2 presents the theoretical toolbox provided by the efficient coding hypothesis. It discusses formal tools derived from information theory and statistics to identify optimal representations of sensory data. Statistical algorithms (sparse coding and independent component analysis) inspired by the efficient coding hypothesis are discussed. The chapter concludes with speculations about the role of redundancy reduction in formation of neuronal functions.

Chapter 3 provides a crude overview of the exemplary sensory task studied here - spatial hearing. It presents anatomy and physiology of the binaural auditory system, and discusses the current knowledge of neuronal representations of the auditory space. In the final section it describes what is known about connections between natural statistics of auditory stimuli and hearing mechanisms.

Chapter 4 is the first out of three chapters, which describe original contributions of this thesis. It describes analysis of marginal statistics of natural binaural sounds. It compares observed cue distributions with knowledge from reductionist experiments. Such comparison allows to argue that the complexity of the spatial hearing task in the natural environment is much higher than analytical, physics-based predictions. It is discussed that early brain stem circuits such as LSO and MSO do not "compute sound localization" as is often being claimed in the experimental literature. Instead it is proposed that they perform a signal transformation, which constitutes a first step of a complex inference process. Results of this chapter have been published in [99].

Chapter 5 develops a hierarchical statistical model, which learns a joint sparse representation of the amplitude and phase information from natural stereo sounds. It is demonstrated that learned higher order features reproduce properties of auditory cortical neurons when probed with spatial sounds. Reproduced aspects were hypothesized to be a manifestation of a fine-tuned computation specific to the sound-localization task. Here it is demonstrated that they rather reflect redundancies present in the natural stimulus. Moreover, the learned representation couples "what" and "where" information, and does not separate them into distinct streams which also matches experimental observations. The article resulting from this chapter is currently under review [98].

Chapter 6 demonstrates that, in principle, learning a sparse factorial code of natural spectrograms can lead to the extraction and separation of spatial / identity relevant information. The results of this chapter suggest that efficient coding is a strategy useful for discovering structures (redundancies) in the input data. Their meaning has to be determined by the organism via environmental feedback. Results of this chapter have been published in [97].

Chapter 7 concludes this work by summarizing results presented in chapters $4-6$, and discussing them in the light of the initial tenets. It discusses strengths as well as drawbacks and limitations of the proposed approach.

Chapter 2

Efficient Coding in Sensory Systems

"The wing would be a most mystifying structure if one did not know that birds flew" wrote Horace Barlow in the opening of his famous paper [10]. Indeed, without theories, which are able to account for large sets of empirical regularities, our scientific efforts would be limited to collecting detailed observations with no connection between any two. In neurobiology theoretical approaches do not yet have the same status as in physics - they are rarely able to form quantitative predictions basing on solely analytical considerations. However candidate principles, which can potentially provide useful theoretical frameworks exist. Perhaps the most prominent one stems from the work of Barlow [10] and Attneave [7] done in the fifties and sixties. It is known as the *efficient coding hypothesis*.

In this chapter I begin by introducing information-theoretic concepts on which the efficient coding hypothesis builds. I proceed by describing the hypothesis itself and discussing statistical models which emerged from the considerations based on it: sparse coding and independent component analysis. The chapter concludes with a discussion of a potential role that efficient coding can play in the formation of functional sensory representations.

2.1 Information Theory

Information theory has been developed by Claude Shannon [127], an electrical engineer working at Bell Laboratories. Even though originally it was supposed to be applied to electric communication channels, such as telephones, it has quickly been picked up by researchers in multiple different areas, including neurobiology.

In this section I introduce selected information-theoretic concepts required to

define the notion of efficient coding, as used in neurobiology. For an in-depth overview of information theory, the interested reader may refer to a classical textbook [29], on which most of this section is based.

2.1.1 Entropy

Let X and Y be discrete random variables with alphabets $\mathcal{X}, \mathcal{Y}$ and the probability mass functions $p(x) = Pr\{X = x\}, x \in \mathcal{X}$ and $p(y) = Pr\{Y = y\}, y \in \mathcal{Y}$ [1].

The entropy H(X) of the variable X is defined by:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{2.1}$$

$0 \log 0$ is assumed here to be equal to 0. If the logarithm is of base 2, the unit of entropy is a *bit*. Entropy can be interpreted as a measure of uncertainty, which an observer has about the outcome of a random trial. In other words, observing a draw from the distribution $p(x)$ carries $H(X)$ bits of information on average.

The joint entropy of two random variables X, Y is simply a function of their joint probability mass function:

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \tag{2.2}$$

Joint entropy can be considered as entropy of a vector-valued random variable. The entropy is associated with the uncertainty, which is reduced by the information gain. Here the notion of conditional entropy becomes useful:

$$H(X|Y) = -\sum_{y \in \mathcal{Y}} p(y) \sum_{x \in \mathcal{X}} p(x|y) \log p(x|y) \tag{2.3}$$

Conditional entropy defines the average amount of uncertainty that remains about X after Y has been observed. The relationship between conditional and joint entropy of two variables is known as the *chain rule for entropies* which is defined as follows:

$$H(X, Y) = H(X) + H(Y|X) \tag{2.4}$$

$$= H(Y) + H(X|Y) \tag{2.5}$$

The chain rule can be extended for a vector of random variables $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. It takes the following form:

[1] In principle variables X and Y can have different probability mass functions $p_x(x)$ and $p_y(y)$. For simplicity, I use the same notation $p(x), p(y)$.

$$H(\mathbf{X}) = H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) \tag{2.6}$$

2.1.2 Mutual Information

Entropy of a variable quantifies its uncertainty - it defines the amount of information (the number of bits) required on average to obtain that variable's description. The related concept is mutual information $I(X;Y)$ - a measure of information that variable X carries about a different variable Y. It is defined in the following way:

$$I(X;Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{2.7}$$

.

Mutual information is a non-negative quantity ($I(X;Y) \geq 0$). It describes the reduction of uncertainty of one variable after observation of the other. It can be also understood as a measure of statistical dependence between two variables. If X and Y are independent their mutual information is equal to 0 and vice versa.

The entropy of a variable X can be then decomposed into the conditional entropy given variable Y and their mutual information:

$$H(X) = H(X|Y) + I(X;Y) \tag{2.8}$$

An important property of mutual information is known as the data processing inequality. Let three random variables X, Y, Z form a Markov chain (denoted as $X \to Y \to Z$) i.e. :

$$p(x,y,z) = p(x)p(y|x)p(z|y) \tag{2.9}$$

The data processing inequality states that there exist no transformation of Y (either deterministic or random), which could increase information about X:

$$I(X;Z) \leq I(X;Y) \tag{2.10}$$

see [29] for a proof. It is a property of a particular importance from the point of view of neurobiology. It states that without any additional sensory data no downstream sensory neuron can have more information about the input than its predecessors in the processing stream.

2.2 Efficient Coding Hypothesis

The development of information theory in the late forties and early fifties almost immediately drew the attention of researchers in disciplines outside of telecommunications. It has been quickly noted that information theoretic concepts should

be of special interest to brain sciences - psychology and neurobiology. After all, those fields attempt to understand how the nervous system processes information, and Shannon's formal tools address precisely this problem.

Fred Attneave [7] and Horace Barlow [10] suggested a hypothesis rooted in information theory which exerted profound influence on the study of perception and neuronal processing. They had observed that natural stimuli are redundant both in space and time. These homogeneities provide structure, which is determined by the state of the environment, hence carries relevant behavioural information. Sensory neurons which are supposed to encode stimulus patterns informative about the organism's surrounding should therefore transmit stimulus redundancies. Moreover, in order to successfully interact with the environment the organism may need all the information it can get. This means that information flow from the environment to the nervous system should be maximized.

If one neuron transmits information about some stimulus aspect there is no need to transmit it for a second time - it would be an uninformative and non-economical use of processing resources. Neurons should therefore represent mutually non-redundant features of the stimulus. This postulated maximization of *intra* (adaptation of receptive fields to correlated stimulus structures) - and minimization of *inter* (encoding of non-overlapping patterns) - neuronal redundance is known as the efficient coding hypothesis.

More formally let $\mathbf{X} = (X_1, \ldots, X_n)$ be the sensory signal and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ the output of a neuron which encodes it. The postulated goal of sensory coding is to maximize the information flow[1] from the environment into the nervous system i.e. the mutual information $I(\mathbf{X}; \mathbf{Y})$. We know that mutual information can be decomposed as:

$$I(\mathbf{X}; \mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}) \tag{2.11}$$

Assuming that noise which determines the encoding distribution $p(\mathbf{y}|\mathbf{x})$ is stationary and of constant variance, the conditional entropy $H(\mathbf{Y}|\mathbf{X})$ is constant as well, and does not depend on the input value $\mathbf{x}$. One can therefore see that the information maximization is equivalent to the maximization of the entropy of the neuronal code $H(\mathbf{Y})$.

Based on equations 2.6, 2.8, the joint entropy of the code can be decomposed into the difference of a sum of single-neuron entropies and mutual information that any of them carries about all others:

$$H(\mathbf{Y}) = H(Y_1, \ldots, Y_n) = \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} I(Y_i; Y_{i-1}, \ldots, Y_1) \tag{2.12}$$

[1] It should be noted that the term "information flow" has often a separate technical meaning (see [8]) for instance. In this context, by information flow I mean mutual information.

What follows from equations 2.11, 2.12 is that maximization of the information flow is achieved when mutual information between single neurons is minimized ($\sum_{i=1}^{n} I(Y_i; Y_{i-1}, \ldots, Y_1 \rightarrow 0)$). When it goes to zero the total entropy of the code is equal to the sum of single-neuron entropies:

$$H(\mathbf{Y}) = \sum_{i=1}^{n} H(Y_i) \tag{2.13}$$

This is equivalent to stating that the activity of single neurons is independent. In this case their joint probability $p(y)$ is equal to the product of marginal probabilities:

$$p(y) = p(y_1, \ldots, y_n) = \prod_{i=1}^{n} p(y_i) \tag{2.14}$$

An illustration of non-efficient and efficient codes is depicted on figure 2.1 A) and B) respectively. Stimulus information (equivalent to it's entropy $H(\mathbf{X})$) is represented by the area of the large circle. Entropies of single neurons are depicted by small circles. Their overlap with the stimulus corresponds to the mutual information $I(\mathbf{X}; \mathbf{Y})$, which is marked by shaded gray. The goal of the efficient code is to maximize the total gray area constrained by the number of neurons and their information coding capabilities. It becomes clear that coding efficiency is maximized when all neurons are coding stimulus-related information (small circles overlap with the large one) and encode non-redundant stimulus aspects (the total dark-gray area, marking the neuronal overlap is minimized).

The impact of the efficient coding hypothesis on neuroscience can be ascribed to the fact that it is able to form experimental predictions applicable to a broad range of sensory systems. The first prediction says that activity of sensory neurons at consecutive stages of processing should be progressively more independent when the system is exposed to a natural stimulus. The second (and perhaps most important one) is that neurons should encode correlated structures of the sensory data they typically encounter. This means that their tuning properties i.e. stimulus features, which modulate their activity, should be predictable from statistics of the natural sensory input. Over the years numerous experiments delivered results supporting those predictions.

The importance of the hypothesis for brain sciences can be also explained by its intellectual descendence and scientific zeitgeist. These can be well illustrated by the person of Horace Barlow - one of its proponents. Being himself a neurobiologist and a great-grandson of Charles Darwin he realized how important it is to consider anatomy and physiology of an organism not in an isolation, but in the context of its natural environment. At the same time being a contemporary of Claude Shannon through lecture of his work he became aware that in order to function properly any information-processing system should be "aware" of its

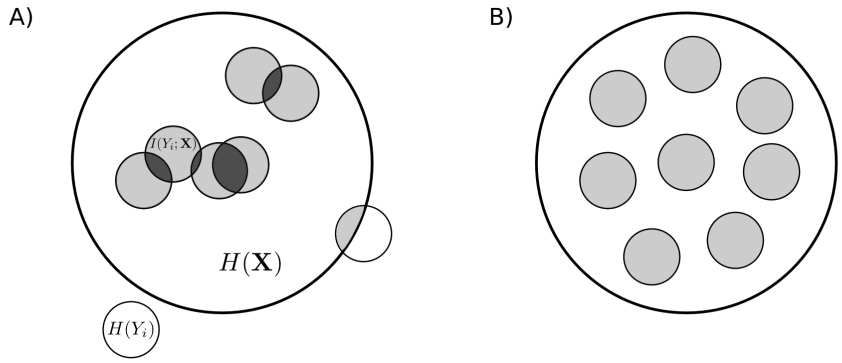

Figure 2.1: A graphical representation of non-efficient and efficient codes. The large circle corresponds to the stimulus entropy $H(\mathbf{X})$. Small circles represent entropies of individual neurons $H(Y_i)$. The total gray area (both light and dark) represents the mutual information between single units and the stimulus. Dark gray marks redundancies between single neurons. A) A non-efficient code. Redundancies are present, and one unit is not adapted to the stimulus. B) An efficient code. Neurons overlap with the stimulus and form a non-redundant representation.

input's statistics. Those two concepts when combined together by Barlow must have led to the insight that the nervous system is (at least to a certain extent) a product of data it processes - its sensory environment. "*Nothing in Biology Makes Sense Except in the Light of Evolution*" says the famous title of Theodosius Dobzhansky's essay [34] - the efficient coding hypothesis embeds theories of neuronal coding in the broad framework of evolutionary theory. It does so, by stressing the importance of the adaptation to the sensory niche.

One should note, however, that the idea of maximizing coding efficiency by progressive redundancy reduction is not free of theoretical limitations. For instance, when transmitting information over a noisy channel one may on purpose introduce redundancies in a controlled manner to reassure the quality of the transmission [29]. The nervous system may be doing that as well, in fact codes responsible for motion generation are known to be redundant [42]. An other important critique is that the organism may not need to encode the entire stimulus stream. Perhaps only certain bits carry an important value and should be encoded by sensory neurons. I discuss this concern further in the following subsection 2.4.

Even if sensory neurons do not form an optimal, exactly efficient code of the natural stimulus the hypothesis discussed in this section provides a *normative* account of the nervous system. By suggesting a theoretically optimal solution to the task of information transmission it suggests a research direction and provides

a benchmark with which biological systems can be compared.

2.3 Sparse Coding

The idea that sensory systems are adapted to natural stimuli has started a separate branch of research in theoretical neurobiology - natural scene statistics. The goal of this field is to explore natural sensory data (sounds, images, etc.) and find statistical regularities i.e. redundancies, which can be exploited by the brain. A particularly successful statistical model which directly builds on the efficient coding hypothesis is known as *sparse coding* [104]. Learning sparse codes of natural stimuli has led to substantial developments both in neuroscience and machine learning. Sparse coding is also one of the fundamental concepts of the present thesis. In this section I discuss its basic version.

Let $x_t \in \mathbb{R}^N$ be the $t-th$ sample of a $N-$dimensional stimulus. The sparse coding model assumes that each stimulus vector x_t can be represented as a linear superposition of basis vectors $b_n \in \mathbb{R}^N$ in the following way:

$$\hat{x}_{t,i} = \sum_{n=1}^{M} s_{t,n} b_{n,i} \tag{2.15}$$

where t is an index over data samples, and i over data dimensions. Linear coefficients $s_t = (s_{t,1}, \ldots, s_{t,M})$ form a representation of the data vector x_t in the space spanned by basis vectors [2] $b = (b_1, \ldots, b_M)$.

In order to handle noisy data an additive, stationary Gaussian noise term can be explicitly incorporated into the model:

$$x_{t,i} = \hat{x}_{t,i} + \eta \tag{2.16}$$

where $\eta \sim \mathcal{N}(0, \sigma^2)$. Equation 2.16 defines a likelihood function of the data $p(x_t|s_t, b)$. Assuming the conditional independence between data dimensions given basis b and coefficients s it is equivalent to:

$$p(x_t|s_t, b) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^N} \prod_{i=1}^{N} \exp\left[-\frac{(x_{t,i} - \hat{x}_{t,i})^2}{2\sigma^2}\right] \tag{2.17}$$

As mentioned above, the model forms a new representation (a code) of sensory data with basis vectors b. Since the sparse coding model is an implementation of the efficient coding hypothesis the new representation is supposed to be maximally efficient in an information theoretic sense. According to equations 2.13, 2.14, this can be achieved when coefficients s are independent i.e. their joint probability is equal to the product of marginals. This constraint allows to formulate a

[2] Traditionally basis vectors b_n are often referred to as "basis functions". I use those terms interchangeably even though they are discrete and do not span a function space.

prior distribution over coefficients s. Assuming that marginal distributions $p(s_n)$ are the same for all coefficients, and that they can be written in the exponential form $p(s_n) = \frac{1}{Z} \exp(-\lambda S(s_n))$ the prior over coefficients becomes:

$$p(s_t) = \prod_{n=1}^{M} p(s_{t,n}) = \frac{1}{Z} \exp\left[-\lambda \sum_{n=1}^{M} S(s_{t,n})\right] \tag{2.18}$$

Function $S(s)$ determines the shape of the coefficient distribution. Crucially, the coefficient distribution is typically assumed to be *sparse* i.e. the majority of the probability mass is densely allocated around 0. This implies that for a typical sample the most coefficients have very small absolute values, and only few largely deviate from 0.

The notion of sparsity and the motivation for use of sparse priors is discussed in detail in section 2.3.2.

Dimensionality of the sparse representation

An important property of a sparse code is the number M of basis functions b used to represent $N-$dimensional data vectors. It is possible that the number of relevant directions in the data space is different from its dimensionality. In such cases the number of basis vectors used to encode the data should be therefore appropriately selected.

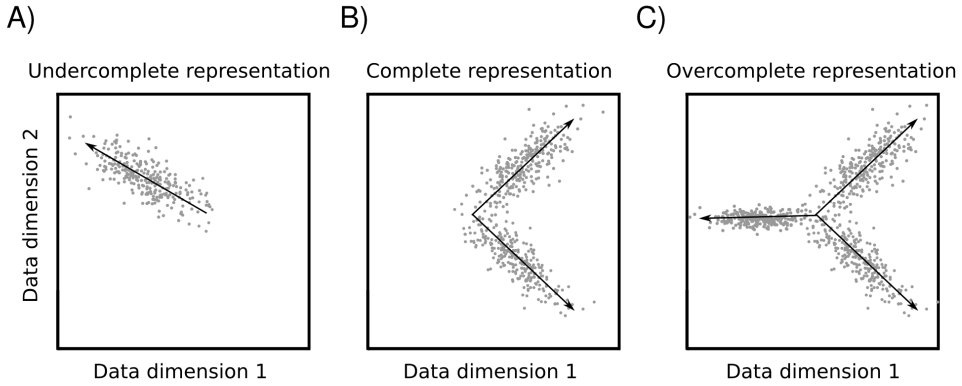

Figure 2.2: Dimensionality of the representation. Axes of each plot correspond to data dimensions. Data points are clustered along directions in the data space. The number M of those directions determines completeness of representation, which shall be used. A) undercomplete representation ($M < N$) B) complete representation ($M = N$) C) overcomplete representation ($M > N$)

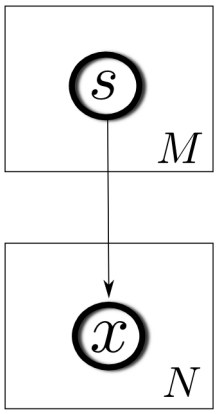

Figure 2.3: A graphical model representing variable dependencies

These situations are depicted on figure 2.2. Depending on the number of basis vectors the representation can be either undercomplete ($M < N$), complete ($N = M$), or overcomplete ($M > N$).

It has been argued that natural signals are well matched by overcomplete representations [105, 81]. Codes with a larger number of dimensions have a greater robustness in the presence of noise and can be sparser [81] than complete representations.

The notion of overcompletness has an important biological meaning. It has been early observed that the dimensionality of sensory representations largely expands in the nervous system [105]. Number of cortical neurons exceeds the number of sensory receptors by orders of magnitude (e.g. in the cat the auditory nerve consists of $\sim 10^4$ fibres, while in the auditory cortex there are $\sim 10^8$ neurons). It has been therefore postulated that overcomplete sparse codes approximate the representational strategy employed by the nervous system [105].

Sparse coding as a generative model

Sparse coding specifies a joint probability distribution over data vectors x and latent coefficients s. For this reason it can be understood as a *generative model* of the sensory input. The dependence between latent coefficients and data is depicted in figure 2.3. Each $N-$dimensional data vector depends on M sparse coefficients. Their joint probability factorizes in the following way[3]:

$$p(x, s) = p(s)p(x|s) \tag{2.19}$$

[3]One should note that this specific factorization applies for any two random variables - it is a straightforward consequence of the *chain rule* for probabilities.

The joint probability distribution $p(s, x)$ provides a holistic description of a relationship between data and model coefficients. It allows to generate data samples which match previously learned structure from the *generating distribution* $p(x|s)$. Simultaneously, it enables the inference of the underlying structure given noisy data samples. This can be done via the *recognition distribution* $p(s|x)$.

One should note that while coefficients s_n are assumed to be independent, a dependence can be introduced by conditioning on a particular data vector x_t. This property of the model is known as *explaining away*, and becomes useful in tasks such as classification or memory retrieval with sparse representations.

A sparse code represents sensory input as a composition of sparse independent *causes*. In this perspective, finding a representation of the sensory input is an active inference process. This corresponds well with the ideas that perception [103] as well as neuronal coding [83] are of inferential nature - they try to estimate the environmental causes which gave rise to the stimulus.

Learning and inference

Parameter estimation of a sparse coding model can be separated into two subtasks. The first one is finding an encoding of a data vector x_t given basis vectors b. This process is known as *inference*. In neural systems modelling it is thought to correspond to the neuronal encoding of a stimulus which happens over a short time-scale. The second task is finding a set of basis vectors b given the training dataset $x = (x_1, \ldots, x_T)$. It is typically referred to as *learning*, and is thought to model the process of receptive-field formation which may happen over longer time-scales (developmental or evolutionary). Learning and inference technics described in this subsection are known together as the sparsenet algorithm and have been introduced by Olshausen and Field [104, 105].

Inference amounts to estimating a coefficient vector s_t given data x_t and a vector basis b. According to the Bayes rule the posterior over latent coefficients is equal to:

$$p(s_t|x_t, b) \propto p(x_t|s_t, b)p(s_t) \tag{2.20}$$

The likelihood term is defined by equation 2.17 and the coefficient prior by 2.18. The negative log-posterior becomes:

$$-\log p(s_t|x_t, b) \propto \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_{t,i} - \hat{x}_{t,i})^2 + \lambda\sum_{n=1}^{M} S(s_{t,n}, \theta) \tag{2.21}$$

A common approach to finding an appropriate encoding, which I also use in this thesis, is the maximum a-posteriori (MAP) estimation. It corresponds to approximating the optimal value of s_t with the peak of the posterior. This can be done by performing a gradient decent on the negative log-posterior function 2.21. The gradient over the sparse coefficients is given as:

$$\frac{\partial}{\partial s_{t,n}}\Big[-\log p(s_t|x_t,b)\Big] \propto -\frac{1}{\sigma^2}\sum_{i=1}^{N} b_{n,i}(x_{t,i}-\hat{x}_{t,i}) + \frac{\partial}{\partial s_{t,n}}\lambda S(s_{t,n}) \tag{2.22}$$

Even though the generative process defined by equation 2.15 is linear in nature the inference is a non-linear task. Basis vectors b_n "compete" among themselves during the inference. Those which match the data vector well have high values of associated coefficients $s_{t,n}$ and suppress activations of other basis functions. The encoding of the observed stimulus emerges from such competitive interactions.

Learning of basis functions b can be achieved by finding a maximum-likelihood (ML) estimate for each vector b_n. This corresponds to minimizing the negative log-likelihood function, which is defined as:

$$-\log p(x_t|s_t,b) \propto \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_{t,i}-\hat{x}_{t,i})^2 \tag{2.23}$$

It can be performed by an iterative stochastic-gradient procedure. Basis vectors are initialized with white noise. Then, for every data vector x_t optimal coefficient values are inferred using gradient descent defined by equation 2.22. Given a MAP coefficient estimate s_t^{MAP} a gradient step is performed on basis functions' elements $b_{n,i}$ according to the following equation:

$$\frac{\partial}{\partial b_{n,i}}\Big[-\log p(x_t|s_t^{MAP},b)\Big] \propto -\frac{1}{\sigma^2}s_{t,n}^{MAP}(x_{t,i}-\hat{x}_{t,i}) \tag{2.24}$$

Those two steps are iterated until convergence. During learning the norm of basis functions has to be monitored and normalized in order to avoid singular solutions.

Numerous other learning algorithms for sparse representations have been proposed [81, 71, 75]. Since in the present work I rely mostly on Sparsenet and similar approaches I do not discuss other algorithms in detail.

2.3.1 Sparse coding and independent component analysis

Another algorithm inspired by the efficient coding hypothesis is known as independent component analysis (ICA) [11, 62]. ICA has evolved in parallel to sparse coding, and can be considered as its special case [102].

In the ICA model data vectors x_t are also assumed to be a linear combination of basis vectors b as defined by equation 2.15. The number of basis functions is equal to the number of data dimensions - the representation is complete (or *quadratic*). Moreover, the noise variance σ^2 is assumed to be 0. The data likelihood becomes then a Dirac delta function:

$$p(x_t|s_t,b) = \delta(x_t-\hat{x}_t) \tag{2.25}$$

Linear coefficients s are also assumed to be sparse and independent - see equation 2.18. Since coefficient values become equivalent to linear projections of data vectors x_t on basis functions b_n a matrix notation is being often used:

$$X = BS \tag{2.26}$$

where $X \in \mathbb{R}^{N \times T}$ is the data matrix $B \in \mathbb{R}^{N \times N}$ is the matrix of basis functions (each column corresponds to a separate basis function) and $S \in \mathbb{R}^{N \times T}$ is the coefficient matrix (each row is a single coefficient). The inference process is fully linear and can be performed using a filter matrix $W = B^{-1}$:

$$WX = S \tag{2.27}$$

ICA can be therefore understood as a rotation of the data vectors x_t into a new set of coordinates where coefficients s are maximally independent.

Depending on the definition numerous objective functions for learning of the basis matrix B can be determined. The famous ICA algorithm of Bell and Sejnowski [11] basing on ideas of the InfoMax transform introduced by Linsker [82] uses a gradient over basis vectors b to explicitly maximize the coefficient entropy. Other approaches attempt to maximize kurtosis or negentropy of coefficient distributions [62]. In this thesis a maximum-likelihood approach to basis function learning is used. Given the data matrix $X \in \mathbb{R}^{N \times T}$ the likelihood function of the model can be defined as:

$$p(X|W) = \prod_{t=1}^{T} \prod_{i=1}^{N} p(w_i^\top x_t) = \prod_{t=1}^{T} \prod_{i=1}^{N} p(s_i) \tag{2.28}$$

where w are filter vectors corresponding to rows of the matrix W. Assuming the exponential form of the marginals $p(s_{t,i}) = \frac{1}{Z} \exp\left(-\lambda S(s_{t,i})\right)$ the negative log-likelihood becomes:

$$-\log p(X|W) \propto \lambda \sum_{t=1}^{T} \sum_{i=1}^{N} S(s_i) = \lambda \sum_{t=1}^{T} \sum_{i=1}^{N} S(w_i^\top x_t) \tag{2.29}$$

Filter vectors w, which determine basis functions b, can be learned by gradient descent in a manner similar to sparse coding.

2.3.2 The notion of sparsity

Information theoretic considerations presented at the beginning of this chapter allow to conclude that the information transmission is maximized when elements of the code represent mutually non-redundant features of the stimulus. Formally this is defined by the product-of-marginals form of the joint distribution 2.18. An important question, which has not yet been addressed in this thesis is - what

shall be the functional form of these marginals i.e. their shape? I alluded before that they are often assumed to be *sparse*. Here, I discuss this notion in more detail.

What is sparsity?

A sparse code is a data representation where the majority of coefficients remain close to 0 when encoding a typical stimulus sample. The sparsity[4] of a population encoding $s_t \in \mathbb{R}^M$ of a data vector $x_t \in \mathbb{R}^N$ can be therefore understood as its L^p norm (with $p \in \{0, 1\}$) defined as:

$$\|s_t\|_p = \sqrt[p]{\sum_{n=1}^{M} |x_{t,n}|^p} \tag{2.30}$$

For $p = 0$ this amounts to the total number of non-zero elements (active neurons) of a vector s_t. If $p = 1$ the norm measures total activation of units. Increasing the sparsity of a representation amounts to finding a code, which can represent each training data sample with a vector of the smallest norm.

The notion of sparsity can be also considered for individual units s_n. It translates then to rare activations of the $n-$th code element when encoding all samples in the training ensemble. This is known as the *lifetime sparseness* [152]. It means that the distribution $p(s_n)$ should be highly concentrated around 0. The concentration in turn, can be measured by the fourth standardized moment i.e. *kurtosis* $\kappa(s_n)$:

$$\kappa(s_n) = \frac{\int_{-\infty}^{\infty} p(s_n)(s_n - \bar{s_n})^4 ds_n}{\left(\int_{-\infty}^{\infty} p(s_n)(s_n - \bar{s_n})^2 ds_n\right)^2} \tag{2.31}$$

where $\bar{s_n}$ denotes the mean which for sparse distributions should be 0. Highly kurtotic, zero-centred distributions are more sparse - they rarely generate samples of large absoulte values. Kurtosis has been proposed as a measure of sparsity by [38]. Notions of population sparseness and lifetime sparseness can be shown to be equivalent under certain conditions [63].

If, as discussed in the previous section, coefficient marginals are assumed to have a general form:

$$p(s_n) = \frac{1}{Z} \exp\left(-\lambda S(s_n)\right) \tag{2.32}$$

then the shape of the distribution is defined by the function $S(s_n)$. In order to induce a sparse coefficient distribution, different sparsity-promoting functions can be used. For instance $S(s_n) = \frac{|s_n|}{d}$ induces a zero-centered Laplace distribution of scale defined by the parameter d. Another often used function is

[4] In the literature one can encounter terms "sparsity" or "sparsness". Here, I use them interchangeably.

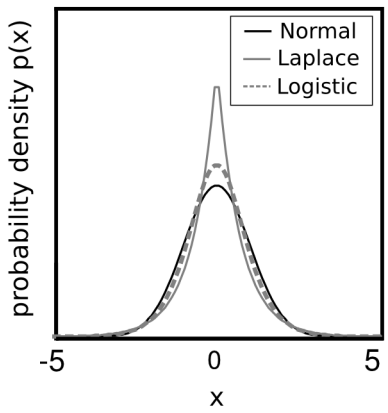

Figure 2.4: Two sparse distributions - Laplace (gray solid line) and logistic (gray dotted line) are contrasted with the Gaussian distribution of equal variance.

$S(s_n) = \log(1 + s_n^2)$ as proposed by [105]. The choice of the sparsity-promoting function determines the norm of the representation vector s_t which is going to be minimized.

Examplary sparse i.e. kurtotic distributions are depicted on figure 2.4 together with a normal distribution of the same variance. Sparse distributions are visibly more "peaked" i.e. concentrated around 0.

Why sparsity ?

A natural question to ask is - what is the advantage of sparse representations? Why should a typical stimulus sample be represented by only a few active neurons? From the point of view of statistics the primary reason for utilizing sparse codes is the structure of natural sensory signals. Natural images, videos and sounds share a curious property - at a small spatiotemporal scale they seem to be well described as a combination of only a few discrete sensory events [38]. Sparse coding forms a representation which makes this underlying structure explicit [106, 31].

Let us look more carefully at that notion using natural sounds as an example. Figure 2.5 A) depicts two log-histograms. The broader one plotted with a black line is an empirical distribution of air pressure values constituting a 5 second long recording of a forest environment. The estimated entropy of this distribution is 4.7 bits. When short, 23 millisecond long epochs (513 samples at 22050 sampling rate) of this recording were projected onto a set of 513 gammatone filters (four exemplary ones are visible on figure 2.5 B), the distribution of resulting coefficients was much sparser - it is plotted on panel A with a gray line. Even though the dimensionality of those representations (raw waveform chunks and

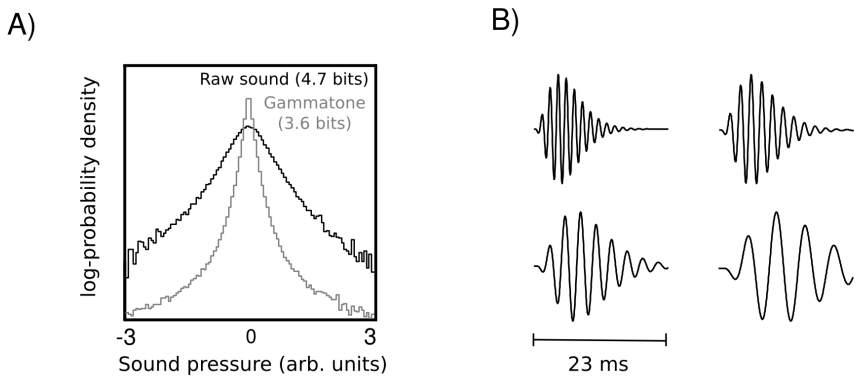

Figure 2.5: Sparse underlying structure of a natural sound. A) A log-histogram of a raw sound waveform (black line) plotted with a log-histogram of projections of 23 ms waveform chunks on gammatone filters (gray line). Projections are visibly more sparse and have more than 1 bit lower entropy. B) Four exemplary gammatone filters which likely constitute sparse dimensions underlying natural sounds.

projections) is the same, and they encode the identical information, the sparse encoding yields lower entropy - 3.6 instead of 4.7 bits. Since the entropy of the data distribution provides the lower bound on the code length, representations yielding lower entropies should better approximate the ground-truth data distribution [79][5]. For a more detailed discussion comparing codes based on coefficient entropies please refer to [79, 81].

The above example shows that indeed - natural sounds have a sparse underlying structure which can be approximated by gammatone filters. Natural images in turn, yield sparse, low-entropic coefficient distributions when encoded with 2−dimensional Gabor filters [31]. Interestingly, features very similar to Gabor and gammatone filters can be learned from statistics of natural stimuli in an unsupervised way by finding a maximally sparse representation [104, 133]. Existence of a sparse structure in sensory data has been therefore confirmed in top-down and bottom-up manners, by observing sparse responses of designed filters and recovering similar filter shapes when maximizing the sparsity of a representation.

In addition to making the statistical structure of natural signals explicit, sparse representations seem to have numerous other advantages. It has been suggested that a representation of a high-dimensional stimulus, which uses only a few active dimensions may be tracing out a smooth and low-dimensional manifold on which the sensory data live [106, 74]. Due to a small overlap between coef-

[5]In this case coefficients s are intepreted as codewords

ficients sparse codes provide a good addressing scheme for associative memories [153]. Finally, they are energy efficient since (what may seem at first like a trivial statement) sparsely spiking neurons consume less energy. This argument is, however, quite strong since it has been shown that due to metabolic constraints sparse neuronal activations in the cortex are not an option - they are a necessity [76, 77].

2.4 Emergence of Function via Efficient Coding

If, as ideas initiated by Attneave and Barlow suggest, sensory neurons form an efficient and sparse representation of natural stimuli, one may ask whether their presumed *function* can be predicted from the input statistics. This question (perhaps worded somewhat differently) has been notably asked in the title of Joseph Atick's paper "*Could information theory provide an ecological theory of sensory processing?*" [5].

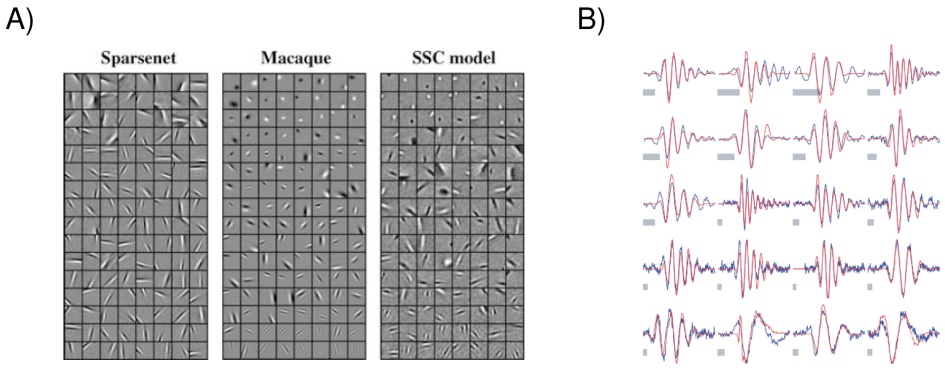

Figure 2.6: Sparse coding of natural stimuli reproduces receptive fields. A) Receptive fields in the macaque V1 (middle pannel) compared with basis functions of two sparse coding algorithms (left and right panels). Figure reproduced from [114]. B) Revcor filters of the auditory nerve of a cat (red lines) are reproduced by a sparse representation of natural sounds (blue lines). Figure reproduced from [133].

A milestone step on the way to providing an answer has been made in the years 1996/1997. Olshausen and Field [104] almost in parallel with Bell and Sejnowski [12] demonstrated that sparse codes and independent components learned from small patches of natural images yield features strongly resembling receptive fields of simple cells in the visual cortex (see figure 2.8 A). Over the following decade Lewicki [78] and Smith and Lewicki [133] provided results of similar importance for the study of the auditory system. They demonstrated that frequency-localized

cochlear filters of a cat can be predicted by a sparse code of a natural sound ensemble (see figure 2.8 B). These observations shed a new light on the notion of function implemented by sensory systems. One does not have to ask any more why do auditory nerve fibers "implement Fourier transform" (as described by Helmholtz) and why do simple cells perform "edge detection". It turns out that both computations can be unified by a single abstract principle - efficient coding of naturally encountered stimuli. Perhaps the very same principle can also explain higher-level computations performed by the brain.

2.4.1 Efficient codes and inference

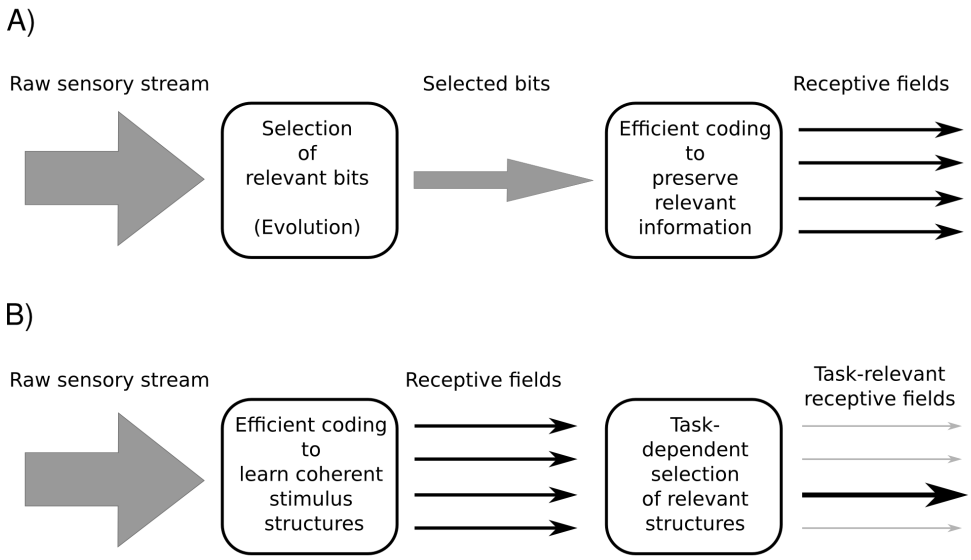

Figure 2.7: Possible roles of efficient coding. A) It serves to preserve maximal amount of pre-selected, narrow information stream. B) Redundancy reduction serves the purpose of finding correlated structures present in the data which may be useful in inferences about the environment (redundancy exploitation). Relevant data dimensions are then selected according to the task at hand.

Sensory signals are a reflection of the environment. They are generated or otherwise influenced by physical objects, and propagate through their medium (let it be the air or the electromagnetic spectrum), until intercepted by sensory receptors of an organism. Correlations present in receptor activations are therefore informative about the state of the environment. It means that encoding redundant activity patterns of the sensory epithelium is vital not only from an abstract, information-theoretic point of view. It may lead to extraction of coherent, interpretable features, which in turn inform the organism about its surrounding.

Following the naming convention of Barlow redundancy reduction becomes then redundancy exploitation.

The above mentioned ideas can be related to a common criticism of the efficient coding hypothesis. One may argue that the goal of the organism is not to reconstruct the stimulus as faithfully as possible. It is rather the extraction of behaviourally relevant information which should guide the design of neuronal codes. A possible answer is that efficient coding can serve two purposes (both are illustrated on figure 2.7).

Firstly it can preserve as many bits as possible from a substream of sensory data pre-selected according to its behavioural relevance (figure 2.7 A). A good example of such a process comes from the auditory system of a grasshopper. It has been observed that auditory receptors transmit a higher amount of information about conspecific calls than about different types of sound [85]. In this system the function of auditory neurons has been determined over the evolutionary time scale, and neurons maximize information transmission about a narrow, but relevant aspect of the sensory niche.

The second possible role of efficient coding is more general (figure 2.7 B). It is possible that redundancy reduction may serve the purpose of discovering coherent stimulus structures during learning. Their behavioural relevance i.e. *meaning* has to be determined by the environmental feedback. Mammals and animals more developed than a grasshopper perform numerous tasks and need different sorts of sensory information to achieve them. In such cases redundancy exploitation becomes a relevant concept. The organism has to use various features of a stimulus to achieve different goals. Those features can be discovered by recoding redundant data structures in the process of unsupervised learning. The function of sensory neurons is then fully determined by the stimulus structure. The *meaning* of the information they represent may vary from task to task.

2.4.2 Efficient codes and experimental design

Let us assume that according to the second strategy suggested above neuronal receptive fields form an efficient representation of ecological stimuli which in turn determines their function. One may then consider the relationship between stimulus statistics and experimental parameters, which has been introduced in the section 1.2 of the introduction.

Figure 2.8 illustrates a schematic "experiment". Gray dots represent stimulus samples observed by the organism, and generated by the environment. Two vectors R_1 and R_2 form an efficient representation. One may consider them as receptive fields of two neurons which are adapted to stimulus redundancies. The mini-population of two units encodes sensory data coordinates along two relevant dimensions. The physical parameter of interest to the experimenter (for instance the sound source position) is denoted by ϕ. Each ϕ value can be mapped into the

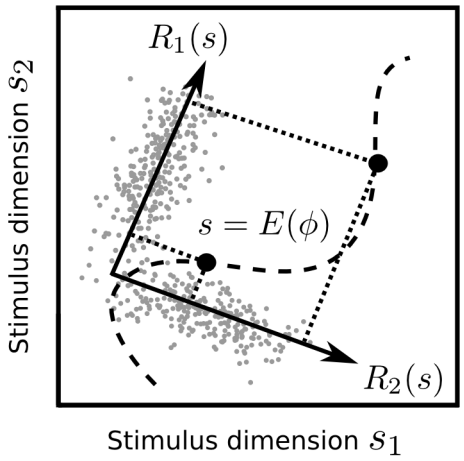

Figure 2.8: Relationship between redundancies in sensory data, neuronal representation and experimental parameters. Gray points represent samples of sensory data encountered by the organism. Two neurons encode position of stimuli along directions marked by black vectors $R_1(s), R_2(s)$. The range of experimental parameter values ϕ, traces a curve in the stimulus space $s = E(\phi)$. Black circles mark experimental measurements. Their projections on neuronal representation give differentiated values.

stimulus space by the function $E(\phi)$. Variation of ϕ traces out a complex, and perhaps not well understood trajectory in the stimulus space. When the experimenter presents the organism with two different ϕ values (black circles), both neurons give a differentiated response defined by projections of two parameter values onto relevant dimensions R_1 and R_2 (black dotted lines). Differentiated response to different stimuli may give an impression that the function of neurons is to represent ϕ. In general this is not true - they are adapted to stimulus statistics - not to the experimental parameter.

The use of natural stimuli in understanding neuronal representations is a subject of debate [122]. In the study of the spatial hearing system for instance, one can not deny that estimation of "spatial tuning curves" of auditory neurons has brought tremendous increase of knowledge. It reduces the complexity of the study and allows to obtain interpretable results. Variation of a point stimulus position in a head-centred, polar coordinate system can not, however, reveal the entire mapping from the stimulus space to the neuronal activity.

In this thesis I propose that mechanisms of binaural sound coding in mammals can be understood as a manifestation of efficient coding as a structure-learning strategy (illustrated on figure 2.7 B). I also suggest that experimental observations

about this system can be explained by a process depicted on figure 2.8. In chapter 5 I demonstrate that sparse codes of natural stereo sounds reproduce important properties of auditory cortical neurons, which were thought to implement a very task-specific computation. The following chapter suggests that sparse codes are also capable of learning auditory invariances from natural sounds. These observations allow to argue that the function of sensory neurons located away from the sensory periphery can be explained by the efficient coding hypothesis. This perspective can clarify a number of experimental observations, and embed them in a broad theoretical framework.

Chapter 3

Spatial Hearing

Among many, the world we evolved in has one particular physical property - it vibrates. Vibrating and otherwise moving objects generate waves of air pressure. They in turn, propagate through the environment, overlapping, interfering and distorting each other additionally being affected by acoustic reflections and attenuated by the surrounding matter.

Despite the presence of noise and ambiguity waves of air pressure, or as we call them *sounds* carry large amounts of information about vibrating objects. The nervous system has developed capabilities to infer abstract and complex properties of the environment solely from two one-dimensional, highly correlated time series - sounds entering the left and the right ear.

To realize how daunting a task that is one can look at figure 3.1. Temporal data presented there are meaningless to the visual system and it is almost impossible to infer the underlying source just by looking. However, after transforming numerical values plotted as lines into displacements of two headphone loudspeakers one would easily recognize a female voice speaking at the left side of the listener. One may even recognize Spanish words said with a Mexican accent and a forest environment.

In this chapter I provide a general overview of the physiology and anatomy of the mammalian hearing system with a special focus on the main subject of this thesis - spatial hearing. I review experimental and theoretical studies which relate the auditory system to the structure of natural stimuli. In the last section I briefly discuss the notion of functional separation in the auditory system into spatial and non-spatial channels.

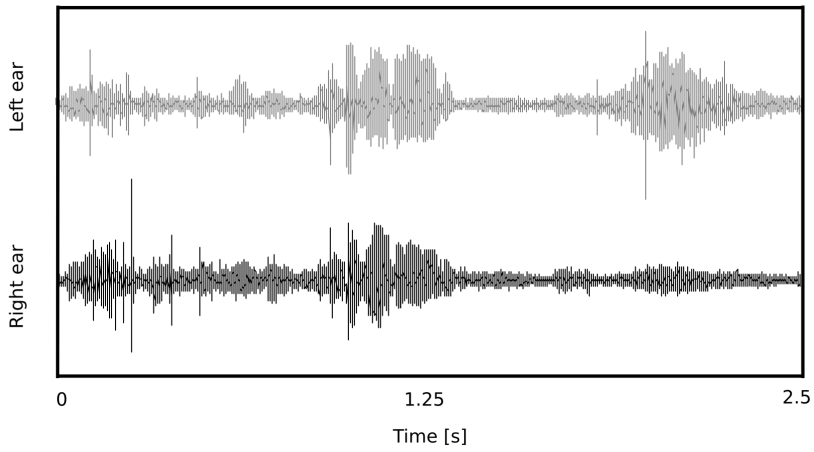

Figure 3.1: An exemplary stereo sound depicted graphically. Gray color corresponds to the left and black to the right ear. Even though numerical data presented here are the same as in the chunk of the sound wave, the visual system is not able to infer the exact content of the underlying physical scene.

3.1 Foundations of spatial hearing

The sound wave as such does not carry any information about the position of its generating source[1]. The spatial frame of reference used by organisms is therefore predominantly relative - sounds are localized in an organism-centred system of coordinates. This is possible due to the sound filtering and differentiation which occurs after the air-pressure waveform interacts with the head and outer ears of the listener.

In mammals (on which this review focuses) outer ears (or *pinnae*) are located at opposite poles of the head separated by the skull. A sound first reaches the ear ipsilateral to the generating source, and then after a very short time delay the contralateral one. This results in a temporal difference between the time of arrival to each ear called the *interaural time difference* (ITD). If the generating sound consists of a single, pure frequency component, or is decomposed into narrowly tuned frequency channels ITDs correspond to *interaural phase differences* (IPDs). ITDs depend on the position of the sound source relative to the organism's head, and constitute one of the major sources of spatial information in hearing (spatial

[1]One should note though that in the natural environment the *quality* of the sound source can be very strongly correlated with its position relative to the organism, and in this way carry indirect spatial information useful to an experienced listener. Hearing an elephant trumpet would very rarely require the listener to raise her head. After all flying elephants are quite a rare breed. This observation is closely related to the *Pratt effect* - see subsection 3.3.2

cues).

Another source of spatial information originates in the fact that the head acts as an acoustic filter. Bones, skin and the brain attenuate sounds contralateral to the listening ear. This attenuation gives rise to a variation in the sound level - *interaural level differences* (ILDs). The general relationship between the ILD and the position of a generating source is quite intuitive - when a sound source is located on the left hand side of the listener, relative sound intensity in the left ear is large. It becomes much smaller in the opposite case, when the source is located at the right hand side.

The role played by interaural level and time differences in the computation of a sound position by humans has first been discussed by John Strutt - Third Baron Rayleigh [143] in the beginning of the XX century. He performed calculations demonstrating that IPDs become a highly ambiguous localization cue when the sound wavelength is much shorter than the diameter of the listener's head. For pure tones of high frequency the IPD value stops corresponding to a single position on a circle surrounding the listener. Such sounds must be therefore localized using another cue - the ILD.

By performing psychoacoustical experiments with tuning forks Lord Rayleigh verified his theoretical predictions. He concluded that human listeners use IPDs to localize sounds of low frequency (lower than 1.5 kHz) and ILDs to identify the position of high frequency tones. Due to the dual nature of spatial information in binaural sound this concept is known as the *duplex theory* of sound localization. It constitutes a fundamental scientific theory explaining how animals can identify sound position.

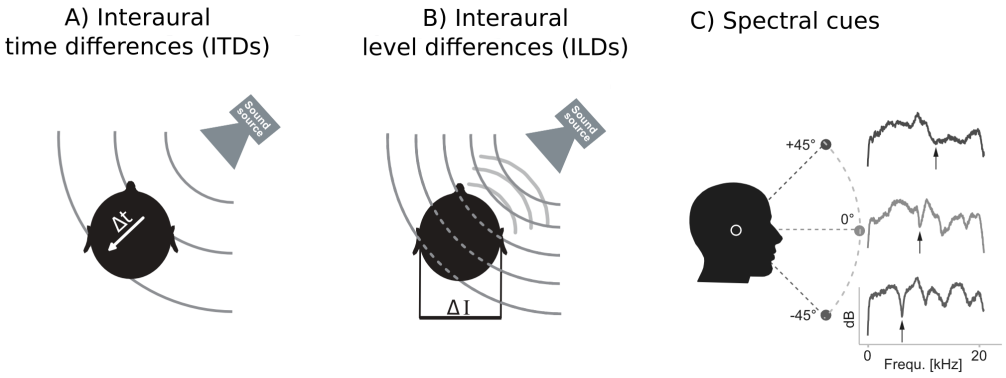

Figure 3.2: A sketch of binaural cues i.e. features of sound which result from interaction of the stimulus with the organism. A) Interaural time differences B) Interaural level differences C) Spectral cues imposed by the pinnae. Figure modified from [49]

In addition to two classes of binaural cues considered by Rayleigh (illustrated on figure 3.2 A and B) there is a third source of spatial information accessible to the listener - spectral cues (panel C of the same figure). Mammalian pinnae are typically of a complex shape. When entering each pinna sound waves are reflected multiple times - the form of this reflection strongly depends of the direction in which the sound propagates. Ear induced sound distortion is well characterized by linear filters known as *head related transfer functions* (HRTFs). HRTFs are typically measured in an anechoic environment where a click sound is played from a carefully controlled position. Microphones located in the ears of the listening subject register the sound waveform which is interpreted as a finite impulse response specific for a particular position. HRTF filtering provides a position-specific information, which can be recovered monaurally i.e. from a sound in a single ear channel. While binaural cues are hypothesized to play the most prominent role in localization of sounds on the azimuthal plane, monaural spectral cues are considered to be vital for estimation of sound elevation (as depicted on figure 3.2 C) [49]. Measurements of human HRTFs strongly support the duplex theory. They show that low frequency sounds are very weakly attenuated by the head which can be considered as a low-pass filter. Low attenuation results in hardly measurable ILDs. Sounds of higher frequencies such as 10 kHz can in turn generate pronounced ILDs as high as 40 dB [68]. Such pronounced cues can be easily detected and utilized in spatial hearing tasks.

According to the recently emerging view the separation of low and high frequency sounds into two nonoverlapping classes (localized with ITDs and ILDs) may not be describing functioning of the nervous system very well [49]. Useful temporal localization cues are generated also by high frequency sounds and carried in their temporal envelopes (*envelope ITDs*). Low-frequency sounds very close to the listener can also generate pronounced ILDs [129] informative about the source position. In summary, according to the modern understanding duplex theory describes frequency dependence of time and level cues rather than their absolute segregation.

3.2 Gross anatomy and physiology of the binaural auditory system

When considering spatial hearing within the framework provided by David Marr, the duplex theory lies somewhere between computational and algorithmic levels of analysis. In this section, I provide a crude overview of the implementation level i.e. of known physiology and anatomy of the binaural hearing system. For a more detailed information one can refer to the recent review article [49] or the book [124] on which this section is based.

A cartoon sketch of the ascending auditory pathway is presented on figure

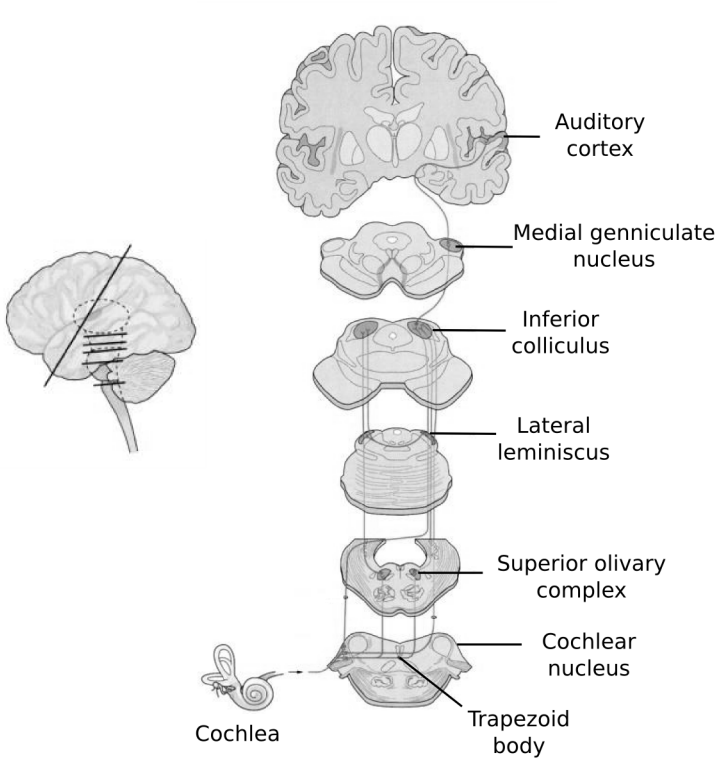

Figure 3.3: A schematic view of the ascending auditory pathway. Input from only single cochlea is depicted.

3.3. The inner ear transforms air-borne sounds into the motion of the cochlear fluid. By vibrating together with the fluid waves the cochlear membrane excites hair cells which convert mechanical energy into electric signal - action potentials. Since parts of the basilar membrane have different resonance frequencies this organ performs "spectral decomposition" of the sound. The underlying hair cells are aligned in a *tonotopic* map following a precise frequency ordering. Separate auditory nerve fibres projecting from the cochlea to the cochlear nucleus encode sound energy in different frequency channels. Importantly for binaural hearing mammalian auditory fibers are capable of representing the fine structure of sounds lower than 4000 Hz by phase-locking to the stimulus (i.e. eliciting spikes precisely aligned to waveform peaks). This provides a physiological constraint on the representation of fine-structure phase and IPDs.

Spatial information is first time processed in the dorsal cochlear nucleus (DCN). The principal neurons of the DCN are capable of determining notch frequencies with a high accuracy. It has been concluded that it makes them well suited for processing monaural, spectral localization cues.

Monaural input converges in the superior olivary complex where time and level

cues are extracted. ILDs are predominantly computed in the lateral superior olive (LSO). Neurons in this structure are excited by projections from the ipsilateral ear and inhibited from the contralateral side. For this reason they are typically referred to as "IE" neurons. It seems that their computational task can be understood as a subtraction of signal power in a narrow frequency channel at one side of the head from the power perceived by the opposite ear. An interesting fact about ILD sensitive neurons in the LSO is that in contrast to the majority of sensory neurons they prefer ipsilateral stimuli not contralateral ones.

Neuronal mechanisms of ILD computation seem to be well understood. In contrast means by which brainstem neurons extract ITDs are still a matter of debate. Smallest detectable ITDs are temporal intervals at the order of microseconds - almost three orders of magnitude shorter than the duration of action potentials. Despite that the mammalian nervous system is capable of extracting and representing them. Even though a solution to this paradox has been proposed [43], mechanisms of submilisecond coding are still a subject of an ongoing research. According to the traditional view the majority of ITD sensitive neurons is located in the medial superior olive (MSO) (current evidence points to the existence of ITD detectors also in the LSO [49]). Those cells receive a converging excitatory binaural input in corresponding frequency channels (EE neurons). Due to the narrow frequency selectivity they can be characterized in terms of IPD tuning. A prominent physiological model of ITD computation has been proposed by Jeffress [65]. He suggested that monaural neurons converge in arrays of delay lines - each corresponding to a particular ITD value. Such array would form a *labelled line code* or a *place code* where high activity of a single unit represents a specific ITD and effectively a location of a sound source. Neurons would therefore be arranged along a spatial gradient into a *spatiotopic map*. The Jeffreys model has dominated thinking about sound localization in mammals for a long time. Recent evidence however, points to the fact that ITDs in the mammalian auditory system are encoded in a different way - by the joint activity of two broadly tuned channels [88].

Outputs of many brainstem nuclei - LSO, MSO and DCN converge in the inferior colliculus (IC). Neurons in the IC are sensitive to multiple binaural cues. Interestingly many IC cells can be characterized with binaural, spectro-temporal receptive fields [112]. Identified sensitivity to the spectral-temporal composition of the binaural signal suggests that binaural hearing mechanisms expand beyond the cue extraction already in the brainstem.

Processed further by the auditory thalamus - medial genniculate nucleus (MGN), auditory information reaches the auditory cortex - the primary auditory field (A1). The functional role played by this structure in audition, and in spatial hearing in particular, remains a mystery. Stimulus transformations performed by subcortical structures can apparently account for a *localization* of a

point source of sound. However, lesions or silencing of neuronal activity in the auditory cortex lead to decreased sound localization performance in human and animal subjects. This apparent contradiction constitutes one of major challenges in understanding the function of this region.

In a manner similar to SOC spatial tuning of mammalian cortical neurons does not match the spatiotopic model. Tuning curves are very broad, and single neuron activity is modulated by sounds located at numerous positions surrounding the animal. They are characterized by steep slopes close to the midline area [137, 138]. Within each hemisphere, neurons seem to prefer positions close to the contralateral ear. These observations suggest that the position of a sound source could be encoded by the joint activity of two "opponent channels". Steep slopes at the midline would according to the theory serve the purpose of precisely encoding the position of the sound in this behaviorally important region. Another puzzling finding was that a linear function of the binaural spectrum suffices to predict with a high accuracy spatial selectivity of auditory cortical neurons [125], even though sound localization is a nonlinear operation. Taken together, the role of A1 in (spatial) audition is far from being understood [101].

The monaural ascending auditory pathway has been investigated in experiments guided by theoretical principles of efficient coding. It has been demonstrated that redundancy between neuronal responses to natural sounds (bird chirps) decreases between the auditory cortex and IC [26] in the cat. In this way a direct experimental evidence for the efficient coding hypothesis has been provided. Studies of auditory cortical responses have shown that cortical neurons are very sparsely active - i.e. firing rates remain below 1 Hz and less than 5% of neurons in a population are activated by a typical stimulus [57, 33]. Based on these results one may risk the statement that notions of sparse and efficient coding provide an appropriate theoretical framework to understand the functioning of the auditory system.

3.3 Processing of natural sounds in the auditory system

Historically, auditory neuroscience has been divided into two camps [145]. The first may be associated with Hermann von Helmholtz - the XIXth century German polymath. Followers of his tradition use simple and well controlled artificial stimuli such as pure sinusoids to characterize the neuronal processing of sound. The origins of the second approach can be traced back to the Austrian ethologist - Konrad Lorenz. He stressed the importance of behaviourally relevant sounds and the use of stimuli such as conspecific calls.

Nowadays mostly due to the rapid development of mathematical and computational tools both trends can merge. It becomes possible to understand the statistical structure of natural sound and test the nervous system using artificial

stimuli which preserve desired aspects of the natural environment while being well controllable. In this way hypotheses about adaptation of sensory systems to the natural environment can be directly tested.

This section provides a brief and concise review of theoretical and experimental investigations that used natural stimuli to study the auditory system. Firstly studies of non-spatial hearing are discussed followed by a discussion of research using spatial sound.

3.3.1 Non-spatial sound

Starting with the most simple characteristics of natural sounds, Rieke et al demonstrated that auditory neurons in the frog increase information transmission when the spectrum of the white-noise stimulus is shaped to match the spectrum of a frog call [116]. In a more recent experiment Hsu and colleagues [58] have shown similar facilitation effects in the zebra finch auditory system using stimuli with the power and phase modulation spectrum of a conspecific song. Modulation spectra of natural sounds were shown to display a characteristic statistical signature. This observation allowed to form quantitative predictions about neural representations and coding of sounds [131].

Simple statistical models of natural auditory scenes have led to interesting theoretical predictions and observations. Low-order, marginal statistics of amplitude envelopes, for instance, seem to be preserved across frequency channels as shown by Attias and Schreiner [6]. This means that all locations along the cochlea may be exposed to (on average) similar stimulation patterns in the natural environment. Strong evidence for adaptation of the early auditory system to natural sounds was provided by two complementary studies by Lewicki [78] and Smith and Lewicki [133]. The authors modelled high order statistics of natural stimuli by learning sparse representations of short sound chunks. In such a way they reproduced filter shapes of the cat's cochlear nerve. This result implies that the function of the cochlea should not be understood as a frequency decomposition per-se. It has rather evolved to maximize coding efficiency in the natural auditory environment. Results of Smith and Lewicki were recently extended by Carlson et al [25] who obtained features resembling spectro-temporal receptive fields in the cat's IC by learning sparse codes of speech spectrograms. This constitutes a strong suggestion that neural representations of acoustic stimuli reflect structures present in the natural environment.

Human perceptual capabilities have also been related to natural sound statistics in a recent study by McDermott and Simoncelli [91]. In a series of psychophysical experiments the authors have shown that the perceived realism and recognizability of sound "textures" by human subjects depends on how well the time-averaged statistics of a stimulus correspond to those of natural sounds.

3.3.2 Spatial sound

In line with the efficient coding hypothesis, binaural hearing mechanisms have also been studied in terms of adaptation to natural stimulus statistics. Even though research in this area has not been very extensive interesting results have been delivered.

Spitzer and Semple have demonstrated that already in the early stages of the binaural processing (in the IC) higher firing rates are elicited by neurons stimulated with dynamic IPD sequences rather than static ones [135, 136]. Moreover, neural responses were more informative about the relative change of the stimulus than its absolute value. Used stimuli were not fully natural, however the study provided an important step by showing that more ecologically valid IPD sequences are preferred by the brainstem circuits. Those studies raise questions whether the function of early binaural neurons in the natural environment is to only extract instantenous cues.

In an attempt to predict IPD coding strategies from theoretical principles, Harper and McAlpine [52] have shown that tuning properties of IPD sensitive neurons in a number of species can be predicted from distributions of this cue naturally encountered by the organism. This was done by forming a model neuronal representation of maximal sensitivity to the stimulus change as quantified by Fisher information. Obtained results stand against predictions of the Jeffress model, and provide one of the key theoretical arguments against its implementation in the mammalian brainstem.

Two recent experimental studies revealed a rapid adaptation of binaural neurons and perceptual mechanisms to changing cue statistics. These research did not utilize natural stimuli as such, however they provided evidence supporting adaptation of neuronal and perceptual mechanisms to the stimulus statistics. Dahmen and colleagues [30] stimulated human and animal subjects with non-stationary ILD sequences. They collected electrophysiological and psychophysical evidence in favor of an adaptation to the stimulus distribution. After a brief exposition to an adapting stimulus shapes of tuning curves of ILD coding neurons as well as human psychophysical curves were shifting towards the side of the adapter. Maier et al [86] in turn, have shown that neural tuning curves in the guinea pig and human performance in a localization task can be adapted to varying ITD distributions. Both - neural representation and human performance were, however, constrained to represent midline locations with the highest accuracy. One has to note that Maier et al. provide an alternative interpretation of the results obtained by Dahmen et al. suggesting that they may be explained by an adaptation to the monaural sound level and not ILDs per se.

A stunning relationship between the frequency of a sound and its position in the natural environment has been recently shown by Parise et al [108]. By analysing a dataset of recordings performed by a freely moving human subject

the authors have demonstrated that sounds originating from high locations systematically have higher frequencies. This stimulus property of not fully understood origins has been further shown to be incorporated as a perceptual prior and speculated to be reflected in the HRTF structure. It has been previously known that the perceived elevation of pure tones is almost entirely determined by their frequency, not actual position (the phenomenon known as the Pratt effect [111, 118]). These results and observations are of a particular importance for the subject of the present thesis. They show that the spatial position and the sound quality are not independent in the natural environment.

3.4 Functional sensory representations of spatial sound - separation of "what" and "where"?

An ongoing debate in the field is whether the auditory system processes sensory information in two functionally separate channels - spatially invariant "what" and identity invariant "where" [100, 67]. This would be a mechanism analogous to the long postulated separation between ventral and dorsal streams in the visual system [148].

Anatomical evidence in favour of a clear dissociation between spatial and non-spatial representations has been delivered by Romanski et al [120]. In their study two separate pathways have been traced from rostral and caudal regions of the auditory cortex. The authors concluded that identified pathways constitute the anatomical basis of "what" and "where" streams. These results have been further supported by a physiological study of Lomber and Malhotra [84]. By cooling down the posterior auditory field in the cat's brain they observed behavioural deficits in a sound localization task. Cooling of anterior areas lead to impaired discrimination ability. These observations led the authors to conclude that those regions are functionally segregated, and that spatial information is processed exclusively by the posterior auditory cortex.

There exists plenty of experimental data which complicate this interpretation. Ventral prefrontal cortex (vPFC) has been identified by Romanski and Goldman-Rakic as the final stage of the identity processing stream ("what") [119]. Cohen et al, however, have shown that vPFC neurons in a monkey are more selective to the sound location than identity (the type of a monkey call) [28]. Moreover, Bizley et al [15] demonstrated that neurons in multiple regions of the ferret auditory cortex are sensitive to the location of a sound source as well as its identity-specific features (pitch and timbre). Observation that the majority of cortical neurons seems to be *sensitive* rather than *selective* to the sound position has triggered a discussion, whether clearly separated "what" and "where" streams are a useful concept in understanding the function of the auditory system [14]. These doubts may be strenghtened by considering perceptual biases such as the Pratt effect

described in the previous section.

In the remaining parts of this thesis I will sketch a theoretical perspective on the functional separation of spatial and non-spatial auditory information. I will also demonstrate, how ideas of adaptation to natural stimulus statistics and efficient coding can provide a computational account of the sensitivity patterns found in physiological measurements.

Chapter 4

Statistical Characterization of Natural Binaural Sounds

4.1 Overview

Prior to understanding higher-level representations employed by the auditory system in spatial hearing tasks it is crucial to know simple characteristics of the sensory input. Low-order statistics of natural stimuli are relatively easy to describe and analyse. Despite this simplicity, they may provide important information, which elucidates functioning of the early sensory systems [130]. In spatial hearing, the low-order stimulus analysis amounts to describing distributions of binaural cues marginalized over relatively long time periods. In result, one should be able to (at least partially) predict input to early cue coding neurons processed when the organism explores real acoustic environments.

Binaural sound statistics determine also the complexity of the sound localization task. Natural sounds are typically generated by multiple independent sources, scattered in different configurations at both sides of the head. In such cases, binaural cues do not correspond to a position of a single object - its identification has to rely on algorithms more complicated than those useful in a simple, laboratory setting. One could assess to which extent this is the case in real auditory scenes, by quantifying the degree of dependence of sounds in each ear.

This chapter addresses the points raised above. Firstly it characterizes marginal statistics of binaural cues encountered in natural hearing conditions, which to my best knowledge, has not been done previously. Secondly, it analyses the redundance of monaural waveforms and in this way estimates the difficulty of a sound localization task in real environments. To achieve those goals three real-world auditory scenes of different acoustic and spatial characteristics were recorded. In the next step binaural cues - IPDs and ILDs were extracted and their marginal

distributions were analysed. Using Independent Component Analysis, it has been demonstrated that in real-world auditory scenes, monaural waveforms are mutually much less interdependent than in a simple, point-source case. Overall, this chapter demonstrates that understanding the function of early binaural neurons in the brainstem can not fully rely on simple, artificial stimuli. It provides a first step towards understanding functioning of the auditory system during spatial hearing in ecological conditions.

4.2 Methods

4.2.1 Recorded scenes

The main goal of research described in this chapter was to analyse cue distributions in different auditory environments. To this end, three auditory scenes of different spatial configuration and acoustic properties were recorded. Each of the recordings lasted 12 minutes.

1. **Nocturnal nature** - the recording subject sat in a randomly selected position in the garden during summer evening. During the recording the subject was keeping his head still, looking ahead, with his chin parallel to the ground. The dominating background sound were grasshopper calls. Other acoustic events included sounds of a distant storm and a few cars passing by on a near-by road. The spatial configuration of this scene did not change much in time - the scene was almost static.

2. **Forest walk** - this recording was performed by a subject freely moving in the wooded area. The second speaker was present, engaged in a free conversation with the recording subject. In addition to speech, this scene included environmental sounds such as flowing water, cracks of broken sticks, leave crunching, wind etc. Binaural signal was affected not only by the spatial scene configuration, but also by head and body motion patterns of the recording subject.

3. **City center** - the recording subject sat in a tourist area of an old part of town, fixating the head as in the previous case. During the recording many moving and static human speakers were present. Contrasted with the previous example, the spatial configuration of the scene varied continuously.

Two of the analysed auditory scenes (nocturnal nature and city center) were recorded by a non-moving subject, therefore sound statistics were unaffected by listener's motion patterns and self generated sounds. In the third scene (forest walk) the subject was moving freely and speaking sparsely. Scene recordings are publicly available at the following URL: http://figshare.com/articles/Statistics_of_Natural_Binaural_Sounds_Supplementary_Material/1157161

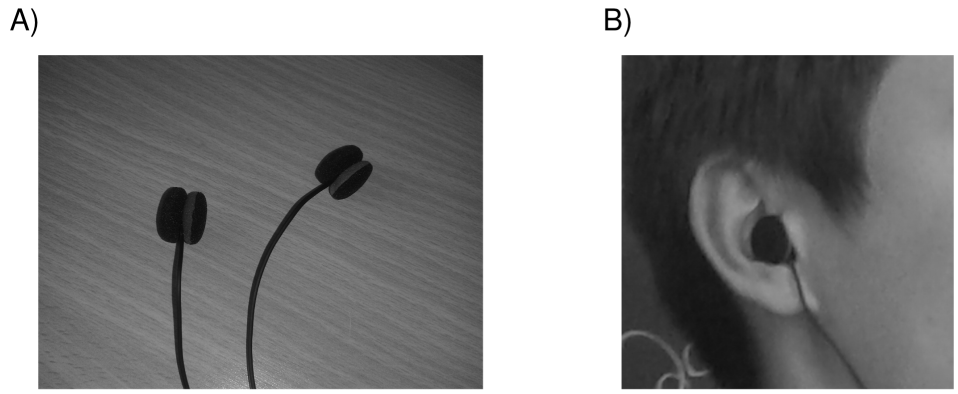

Figure 4.1: Binaural microphones Soundman OKM-II. A) Both microphones B) One of the microphones placed at the entrance to the ear channel of the recording subject.

4.2.2 Binaural recordings

Recordings were performed using the Soundman OKM-II binaural microphones which were placed in the left and the right ear channels of the recording subject (see figure 4.1). Soudman DR2 recorder was used to simultaneously record sound in both channels in an uncompressed wave format at 44100 Hz sampling rate. The circumference of the recording subject's head was equal to 60 cm.

4.2.3 Frequency filtering and cue extraction

Prior to analysis, raw recordings were down-sampled to 22050 Hz sampling rate. The filtering and cue extraction pipeline is schematically depicted in figure 4.2

To obtain a spectral decomposition of the signal, sound waveforms from each ear were transformed using a filterbank of 64 linear gammatone filters. Filter center frequencies were linearly spaced between 200 and 3000 Hz for IPD analysis and 200 and 10000 Hz for ILD analysis. Biological cochlear filters are spaced logarithimcally. Here however, a linear spacing was utilized. This resulted in a more uniform coverage of the frequency range than in the case of a biologically plausible filterbank. Within the limits of the analysis performed here, results should not be significantly different when using different filterbanks for preprocessing.

A Hilbert transform of each frequency channel was performed. In result, instantaneous phase $\phi_{L,R}(\omega, t)$ and amplitude $A_{L,R}(\omega, t)$ were extracted, separating level and time information. Instantaneous binaural cue values were computed in corresponding frequency channels ω from both ears according to the following

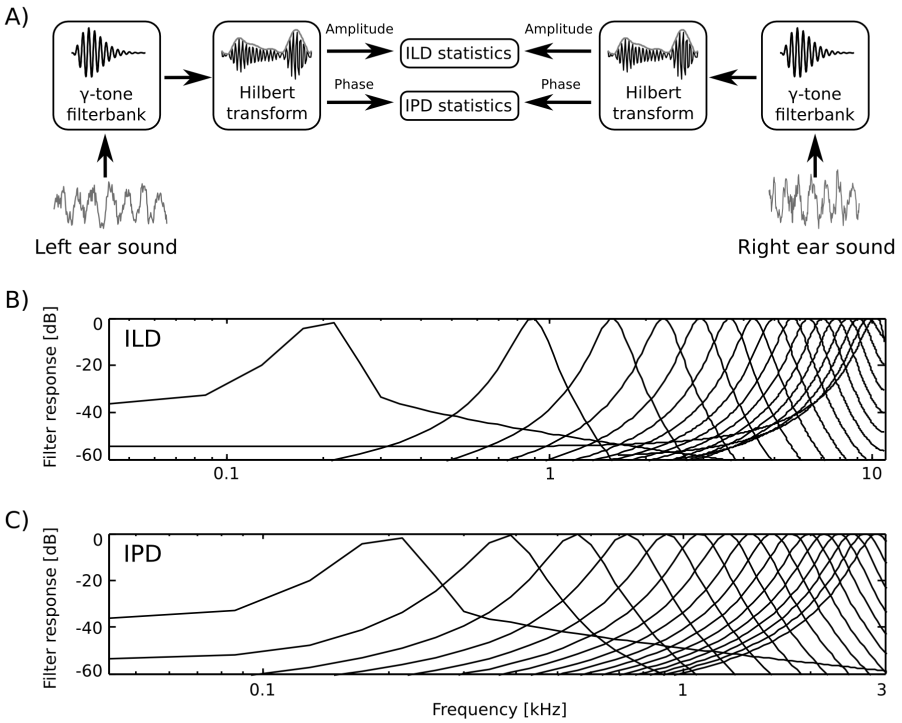

Figure 4.2: Preprocessing and cue extraction pipeline. A) Preprocessing scheme. Raw sounds in each ear were transformed using a cochleotopic filterbank. In the next step the Hilbert transform was computed to separate amplitude from phase. Finally IPDs and ILDs were extracted. B) Filter response spectra of 16 out of 64 filters used to extract interaural level differences. C) Filter response spectra of 16 out of 64 filters used to extract interaural phase differences.

equations:

$$ILD(\omega, t) = 10 \times \log_{10} \frac{A_L(\omega, t)}{A_R(\omega, t)} \tag{4.1}$$

$$IPD(\omega, t) = \phi_L(\omega, t) - \phi_R(\omega, t) \tag{4.2}$$

IPDs with absolute value exceeding Π were wrapped to a $[-\Pi, \Pi]$ interval. Time series of IPD and ILD cues obtained in this way in each frequency channel were subjected to further analysis.

4.2.4 Computation of the "maximal" IPD value

In each frequency channel ω, the maximal IPD value constrained by the head size ($IPD_{\omega,max}$) was computed in the following way. The head shape was assumed

to be spherical. Given this assumption, the time period required by the sound wave to travel the distance between the ears is equal to:

$$ITD = \frac{R_{head}}{v_{snd}}(\Theta + \sin(\Theta)) \tag{4.3}$$

where R_{head} is the head radius, v_{snd} the speed of sound and Θ the angular position of the sound source measured in radians from the midline. The ITD is maximized for sounds located directly oposite to one of the ears, deviating from the midline by $\Theta = \frac{\pi}{2}$. ITD_{max} becomes

$$ITD_{max} = \frac{R_{head}}{v_{snd}}(\frac{\pi}{2} + 1) \tag{4.4}$$

The maximal IPD was computed separately in each frequency channel ω

$$IPD_{\omega,max} = 2\pi\omega ITD_{max} \tag{4.5}$$

The above calculations assume a spherical head shape, which is a major simplification. It was, however, satisfactory for the sake of the current analysis.

4.2.5 Independent component analysis of binaural waveforms

To analyze mutual dependence of monaural waveforms Independent Component Analysis of short recording intervals was performed. The maximum-likelihood ICA variant described in section 2.3.1 was utilized.

Prior to ICA learning, the recordings were downsampled to 14700 Hz sampling rate. A training dataset was created by randomly drawing 100000 intervals each 128 samples long (corresponding to 8.7 ms). The sampling rate and the length of the time interval were equal to those used in [78].

After learning, we rejected spectrally non-localized independent components as they typically reflect noise, not data structures [133]. All basis functions for which the sum of two spectral maxima in each ear constituted less than 15% total power were removed. This resulted in 0, 41 and 5 components rejected from the nocturnal, forest and city scenes respectively.

4.2.6 Generation of artificial data

Two artificial datasets corresponding to extreme cases of binaural redundance were generated using sounds from each recorded scene. Binaural recordings were transformed to a single channel by averaging sound in both ears. Point-source datasets were created by drawing random intervals of the mono recording and convolving them with Head Related Transfer Functions (HRTFs) corresponding to one of the 24 positions (15 degree spacing) on a circle surrounding the head. Human HRTFs were taken from the publicly available LISTEN database [151]. Maximally independent datasets were created by independently sampling two

epochs of sound and treating each of them as an input to one of the ears. Each dataset consisted of $1e5$ samples of binaural sound, each 8.7 ms long. Recorded and simulated datasets had the same Fourier spectra, but a very different dependence structure.

4.3 Results

4.3.1 Recorded scenes

In this chapter three 12 minute recordings of different auditory scenes - nocturnal nature, forest walk and city center were analyzed (the analysis pipeline is depicted on figure 4.2). The scenes were selected as representative examples of a broad range of possible acoustic environments. In each scene multiple sound sources positioned at a diverse set of locations were present. Sound types and spatial configuration of sources however, varied from scene to scene. In the nocturnal nature recording, the recording subject was static, and the scene was dominated by grasshopper calls (which do not move while generating sound). This recording was an example of an environment, where many non-moving sources are present, and their joint activity results in an ambient sound. The forest walk scene was much less stationary - the subject was freely moving in a wooded area while talking to another person. The scene included speech, ambient environmental sound sources (wind, leaves, stream) as well as transient ones (wood cracks, steps, etc.). This case was used as an example of a scene, where binaural information is affected by the motion and speech of the listening subject. In the third scene - city center - the subject was again listening passively, and the sensory input was rapidly changing due to the presence and the constant motion of multiple human speakers. This recording exemplified very dynamic auditory scenes with numerous moving sources.

Auditory environments chosen for recording were different from each other. We attempted to obtain representative samples of three classes of auditory scenes categorized by spatial configurations - static sources (nocturnal nature), moving sources (city center), and moving subject (forest walk). A statistical variation among examples analyzed here should therefore capture variability across numerous other cases.

Scene selection in this study did not include all possible cases. For instance no recording was performed in an enclosed, highly reverberant environment. Additionally all recordings were done in similar weather conditions, which may have narrowed the range of stimulus properties. The nocturnal nature and city center scenes consisted of constantly active sources - no periods of silence were present, which happens in natural hearing conditions. Despite those limitations current data should be heterogeneous enough to draw general conclusions.

While auditory scenes were selected as a representation of diverse environments, recordings of each scene were performed in an unbiased way. Position of the subject was chosen at random in the two static recordings, and his motion was not constrained or pre-designed while walking. In this way, samples of a typical sensory input were collected. By refraining from recording in carefully controlled settings, where some feature (for instance loudness) in each ear would be the same, the selection bias has been reduced. Natural auditory scenes are rarely spatially symmetric and stimuli analyzed here provide examples of what one typically hears. Understanding the structure of unbiased rather than fine-tuned stimuli should give better insights into the functioning of the nervous system in natural conditions [107].

4.3.2 Sound spectra

Frequency spectra of recorded sounds are displayed on figure 4.3. Strong differences in spectrum across all recorded auditory scenes was present. In two of them - the forest walk scene and the city center scene, frequency spectrum had an exponential (power-law) shape, which is a characteristic signature of natural sounds [149]. Since the nocturnal nature scene was dominated by the grasshopper sounds, its spectrum had two dominant peaks around 7 and 10 kHz.

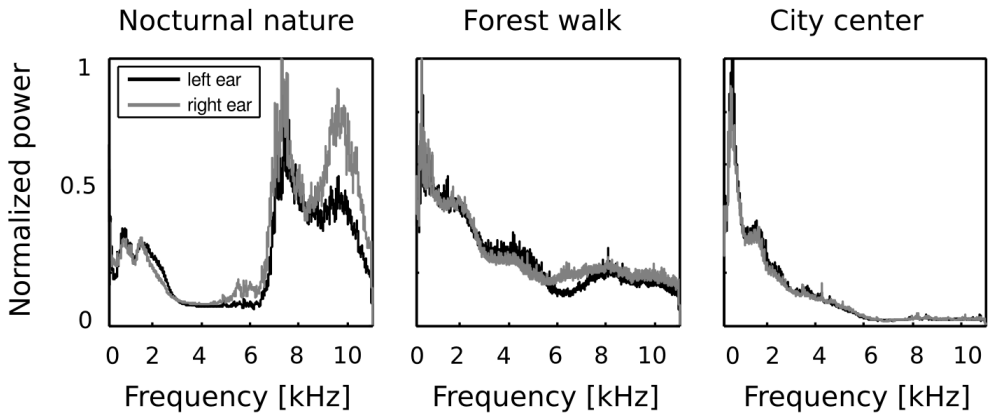

Figure 4.3: Frequency spectra of binaural recordings.In the forest walk and the city center scenes spectra of sounds in the left and in the right ear (black and gray lines respectively) were approximately the same. In the nocturnal nature scene, a sound source was constantly present on the right side of the head, therefore more power was present in high frequencies in the right ear.

Sounds in both ears contained similar amount of energy in lower frequencies (below 4 kHz) - which is reflected by a good overlap of monaural spectra on

the plots. In higher frequencies though, the spectral power was not equally distributed in both ears. This difference is most strongly visible in the spectrum of the nocturnal nature scene. There, due to a persistent presence of a sound source (a grasshopper) closer to the right ear, corresponding frequencies were amplified with respect to the contralateral ear. Since the spatial configuration of the scene was static, this effect was not balanced by being averaged out in time. Monaural spectra of the forest walk scene overlapped to a much higher degree. A small notch in the left ear spectrum is visible around 6 kHz. The city center scene, has almost identical monaural spectra. This is a reflection of its rapidly changing spatial configuration - sound sources of similar quality (mostly human speakers) were present in all positions during the time of the recording.

4.3.3 Interaural level difference statistics

An example joint amplitude distribution in the left and the right ear is depicted on figure 4.4 A. It is not easily described by any parametric probability density function (pdf), however monaural amplitudes reveal a strong linear correlation. Correlation coefficient can be therefore used as a simple measure of interaural redundancy by indicating how similar the amplitude signal in both ears is, at a particular frequency channel. Interaural amplitude correlations for all recorded scenes are plotted as a function of frequency on figure 4.4 C. A general trend across the scenes is that correlations among low frequency channels (below 1 kHz) are strong (larger than 0.5) and decay with increasing frequency. Such trend is expected due to the filtering properties of the head, which attenuates low frequencies much less than higher ones. The spatial structure of the scene also finds reflection in binaural correlation - for instance, a peak is visible in the nocturnal nature scene at 7 kHz. This is due to a presence of a spatially fixed source generating a sound at this frequency (see figure 4.3). The most dynamic scene - city center - reveals, as expected, lowest correlations across most of the spectrum.

Interaural level differences ILD were computed separately in each frequency channel. Figure 4.4 B displays an example ILD distribution (black line) together with a best fitting Gaussian (blue dotted line) and logistic distribution (red dashed line). Logistic distributions provided the best fit to ILD distributions across all frequencies and recorded scenes, as confirmed by the KS-test (data not shown). ILD distribution at frequency ω was therefore defined as

$$p(ILD_\omega|\mu_\omega,\sigma_\omega) = \frac{\exp(-\frac{ILD_\omega-\mu_\omega}{\sigma_\omega})}{\sigma_\omega(1+\exp(-\frac{ILD_\omega-\mu_\omega}{\sigma_\omega}))^2} \tag{4.6}$$

where μ_ω and σ_ω are frequency specific mean and scale parameters of the logistic pdf respectively. Variance of the logistic distribution is fully determined by the scale parameter.

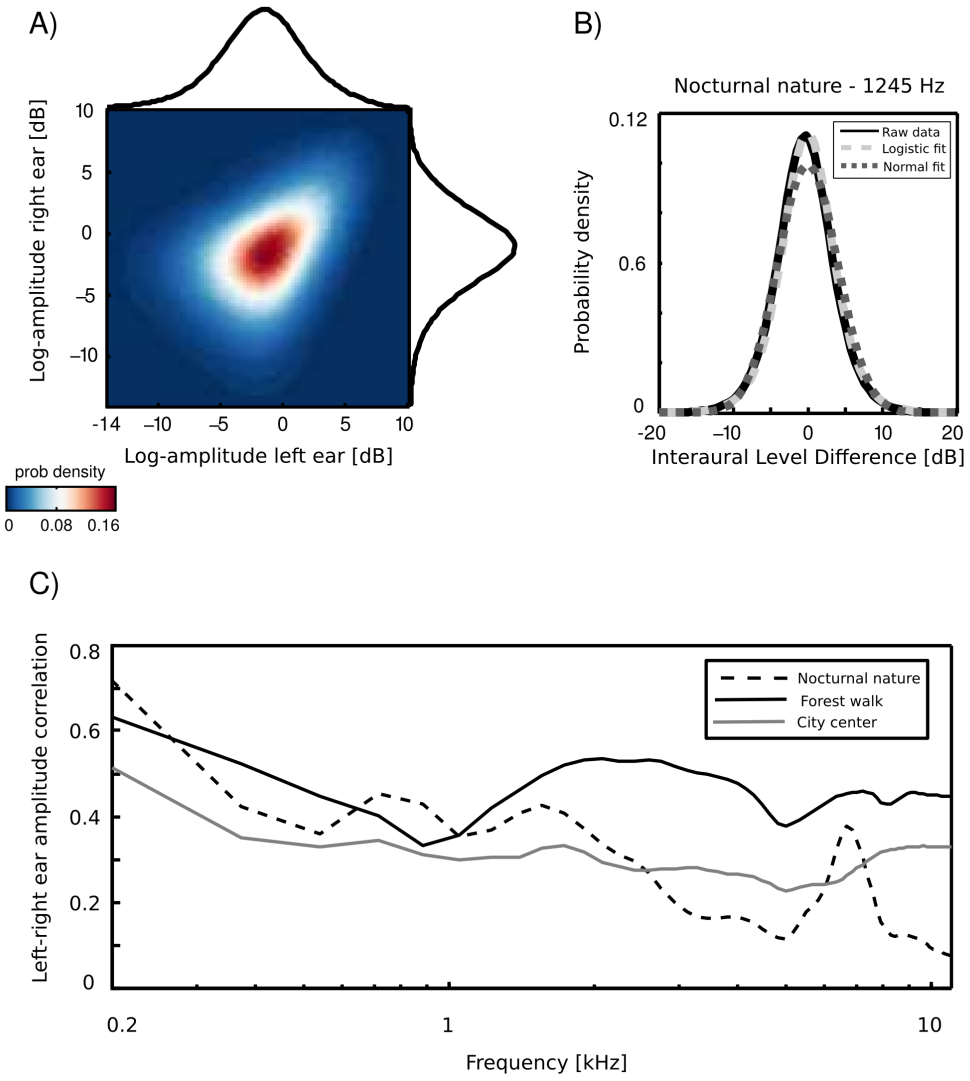

Figure 4.4: Binaural amplitude statistics. A) An exemplary joint distribution of monaural amplitudes at 1245 Hz. Exemplary data were taken from the nocturnal nature recording. B) An ILD distribution of the same data, plottedtoghether with a Gaussian and a logistic fit (blue and red dotted lines respectively) C) Interaural amplitude correlations across frequency channels

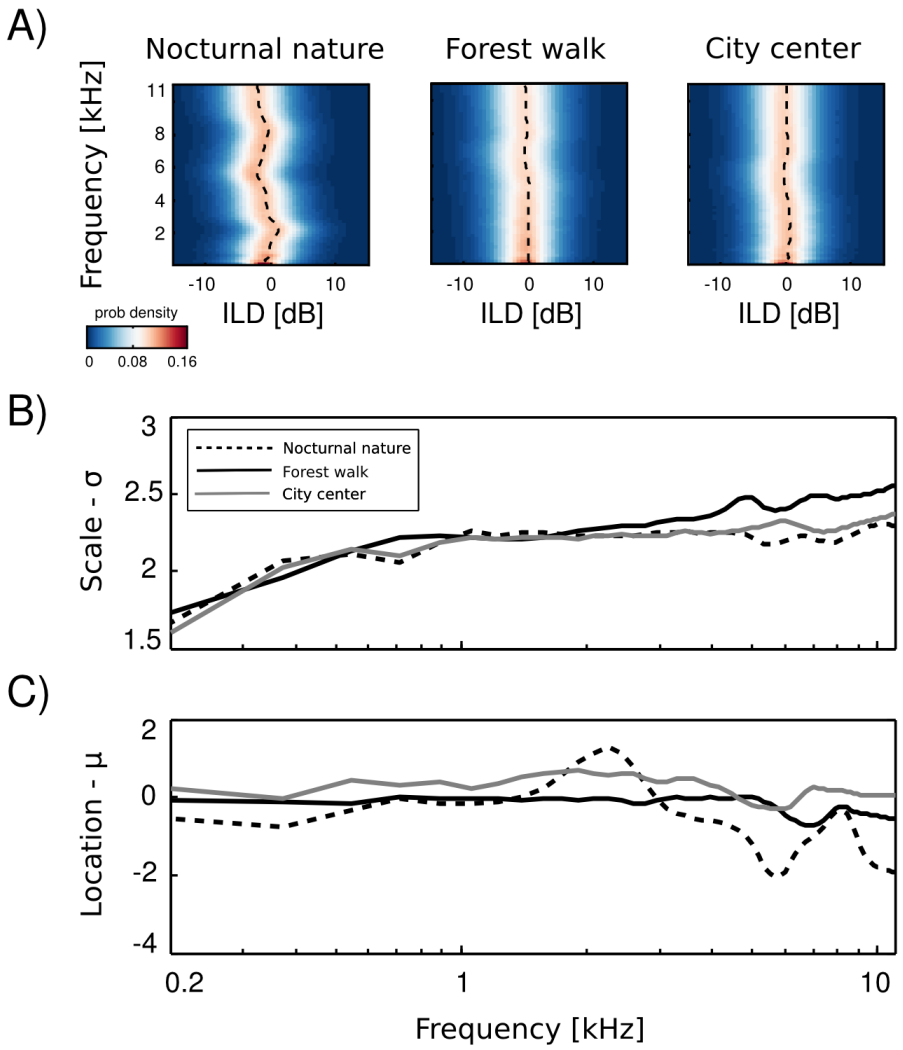

Figure 4.5: Interaural level difference distributions. A) Histograms plotted as a function of frequency - a strong homogeneity of distributions is visible across recorded scenes and frequency channels. B) The scale parameter σ_ω of fitted logistic distributions plotted as a function of frequency C) The location parameter μ_ω plotted as a function of frequency

Empirical ILD distributions are plotted on figure 4.5 A. As can be immediately observed, they preserve similar shape in all frequency channels and auditory scenes regardless of their type. Scale (σ_ω) and mean (or location - μ_ω) parameters of fitted distributions are plotted as a function of frequency on figures 4.5 B and C respectively. The mean of all distributions is very close to 0 dB in most cases. In two non-static scenes i.e. forest walk and city center deviations from 0 are very small. Marginal ILD distributions of the spatially non-changing scene - nocturnal nature - were slightly shifted away from zero for frequencies generated by a sound source of a fixed position. The difference, however was weak. The scale parameter behaved differently than the mean. In all auditory scenes it grew monotonically with the increasing frequency. The increase was quite rapid for frequencies below 1 kHz - from 1.5 to 2. For higher frequencies the change was much smaller and in the $1 - 11$ kHz interval σ did not exceed the value of 2.5. What may be a surprising observation is the relatively small change in ILD distribution, when comparing high and low frequencies. It is known that level differences become much more pronounced in high frequency channels [68], and one could expect a strong difference with a frequency increase. At least partial explanation can be made, when one observes a close relationship between Fourier spectra of binaural sounds and means of ILD distributions. In a typical, natural setting sound sources on the left side of the head are qualitatively (spectrally) similar to those on the other side, therefore spectral power in the same frequency bands remains similar in both ears. Average ILDs deviate from 0 if a sound source was present at a fixed position during the averaged time period. Increase in the ILD variance (defined by the scale parameter σ) with increasing frequency, can be explained by the filtering properties of the head. While for lower frequencies a range of possible ILDs is low, since large spatial displacements generate weak ILD changes, in higher frequency regimes ILDs become more sensitive to the sound source position hence their variability grows. On the other hand, objects on both sides of the head reveal similar motion patterns and, in this way, reduce the ILD variability, which may account for the small rate of change.

Observed ILD distributions revealed very small variation across different frequencies. The variability was much weaker than what can be predicted from known head filtering properties. Additionally, ILD distributions were quite homogenous across different auditory scenes. This means that neuronal codes for ILDs can optimally represent this cue in very different acoustic environment without necessity of a strong adaptation.

4.3.4 Interaural phase difference statistics

Marginal distribution of a univariate, monaural phase variable over a long time period is uniform, since it periodically assumes all values on a unit circle. An interesting structure appears in a joint distribution of monaural phases (an ex-

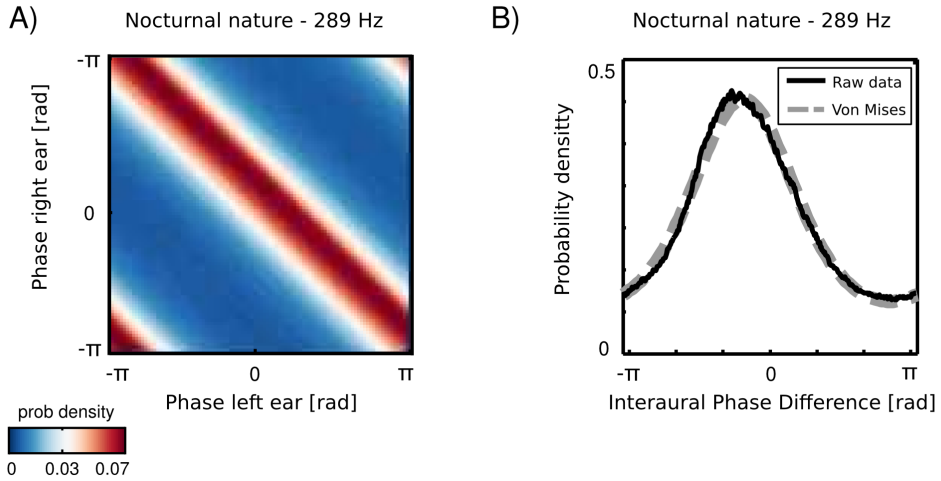

Figure 4.6: Binaural phase statistics A) An exemplary joint probability distribution of monaural phases at 289 Hz. Data were taken from the nocturnal nature scene. B) An empirical IPD distribution of the same data (black line) plotted with a fitted von-Mises distribution (blue dashed line)

ample is plotted on figure 4.6 A). Monaural phases reveal dependence in their difference i.e. they become conditionally independent given the IPD value. Their joint probability is therefore determined by the probability of the IPD [22] :

$$p(\phi_L, \phi_R) \propto p(\phi_L - \phi_R) \tag{4.7}$$

where ϕ_L and ϕ_R are instantaneous phase values in the left and the right ear respectively.

To obtain a parametric description, IPD histograms were fitted with the von Mises distribution as visible in figure 4.6 B (additional structure was present in IPDs from the forest walk scene - see the following subsection). A distribution of an interaural phase difference in the frequency channel ω ($IPD_\omega = \phi_{L,\omega} - \phi_{R,\omega}$), was then given by:

$$p(IPD_\omega|\kappa_\omega, \mu_\omega) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(IPD_\omega - \mu_\omega)} \tag{4.8}$$

where μ_ω and κ_ω are frequency specific mean and concentration parameters and I_0 is the modified Bessel function of order 0. In such case, the concentration parameter κ controls mutual dependence of monaural phases [23]. For large κ_ω values $\phi_{L,\omega}$ and $\phi_{R,\omega}$ are strongly dependent and the dependence vanishes for $\kappa = 0$.

Figure 4.7 A depicts IPD histograms in all scenes depending on the frequency channel. Thick black lines mark $IPD_{\omega,max}$ - the "maximal IPD" value i.e. the

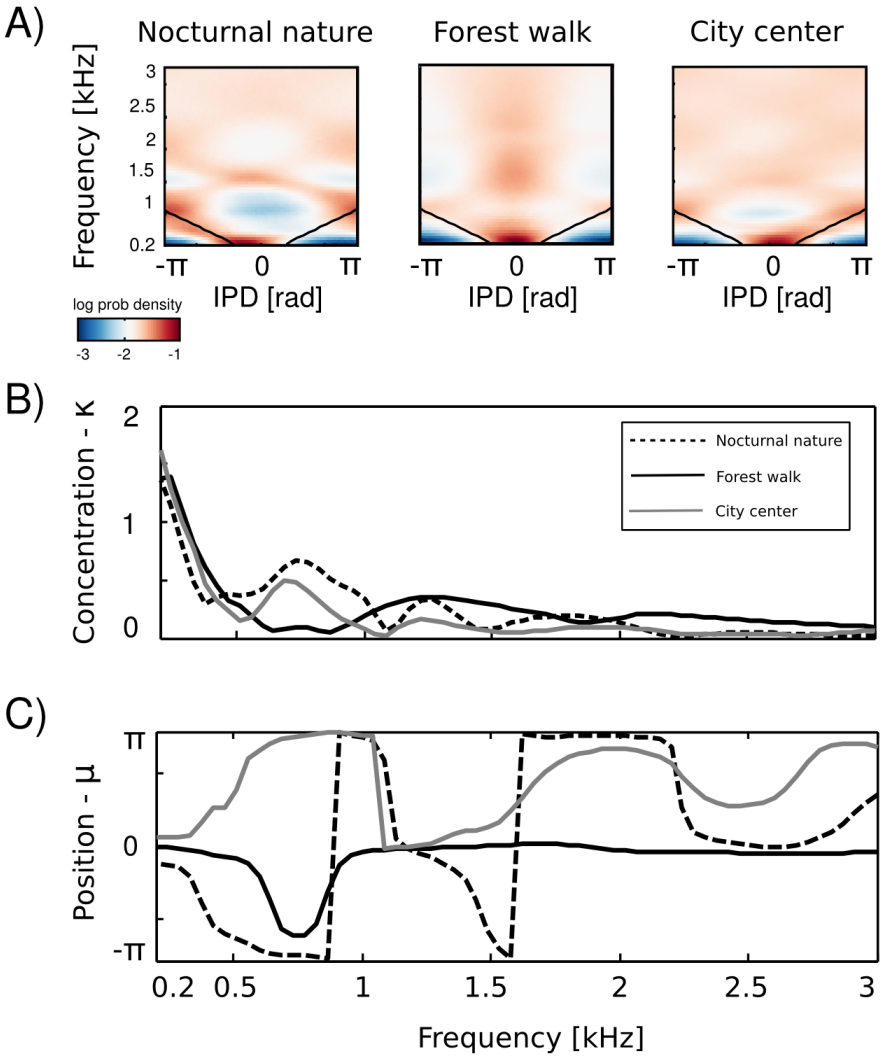

Figure 4.7: IPD distributions. A) Log-histograms plotted as a function of frequency. Black lines mark the "maximal" IPD limit. B) The concentration parameter κ_ω of fitted von-Mises distributions plotted as a function of frequency C) The position parameter μ_ω plotted as a function of frequency

phase shift corresponding to a time interval required for a sound to travel the entire interaural distance equal to the head diameter (for details see the Materials and Methods section). At low frequencies (below 1 kHz), histograms had a triangular shape. This is a common tendency in IPD distributions, visible across all auditory scenes. Additionally, due to phase wrapping, for frequencies where $\pi \leq |IPD_{max}| \leq 2\pi$ the probability mass is shifted away from the center of the unit circle towards the $-\pi$ and π values, which is visible as blue, circular regions. This trend is not present in the forest walk scene, where a clear peak at 0 radians is visible for almost all frequencies. Two panels below i.e. figures 4.7 B and C display plots of κ and μ parameters of von Mises distributions as a function of frequency. The concentration parameter κ decreased in all three scenes from a value close to 1.5 (strong concentration) to below 0.5 in the interval between 200 Hz and 500 Hz. This seemed to be a robust property in all environments. Afterwards, small κ rebounds were visible. For auditory scenes recorded by a static subject i.e. nocturnal nature and city center rebounds occur at frequencies, where IPD_{max} corresponds to π multiplicities (this is again an effect of phase wrapping). The κ value is higher for a more static scene - nocturnal nature - reflecting a lower IPD variance. For frequencies above 2 kHz, concentration converges to 0 in all three scenes. This means that IPD distributions become uniform and monaural phases mutually independent. The frequency dependence of the position parameter μ is visible on figure 4.7 C. Again, division may be made between statically and dynamically recorded scenes. For the latter one, IPD distributions were centered at the 0 value with an exception at 700 Hz. For two former ones, distribution peaks were roughly aligned along the IPD_{max} as long as it did not exceed $-\pi$ or π value. One has to note, that for distributions close to uniform ($\kappa \to 0$), position of the peak becomes an ill defined and arbitrary parameter.

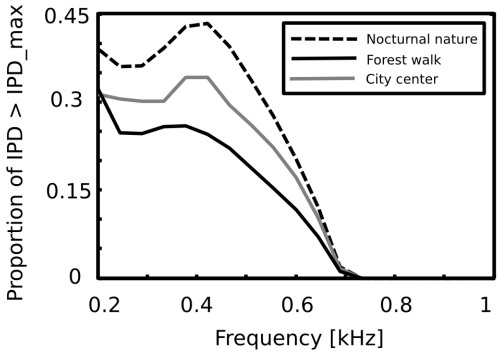

Figure 4.8: Proportion of IPDs exceeding the "maximal IPD" threshold plotted as a function of frequency. In each auditory environment a substantial amount (up to 45% in the 400 Hz channel of the nocturnal scene) of low-frequency IPDs exceeded the limit imposed by the size of the head. While such IPDs can carry relevant information, they can not be used to identify sound source position without additional transformations.

Thick black lines on figure 4.7 A mark the "maximal" IPD value (IPD_{max}), constrained by the head size. A single, point sound source in an anechoic environment would never generate an IPD exceeding IPD_{max}. In natural hearing conditions however, such IPDs occur due to the presence of two (or more) sound sources at both sides of the head or due to acoustic reflections [49]. The presence of IPDs exceeding the IPD_{max} limit is visible on figure 4.7 as a probability mass lying outside of the black lines. Figure 4.8 displays a proportion of IPDs larger than the one defined by the head size plotted against frequency. Lines corresponding to three recorded auditory environments lay in parallel to each other, displaying almost the same trend up to a vertical shift. The highest proportion of IPDs exceeding the "maximal" value was present in the nocturnal nature scene. This was most probably caused by a largest number of very similar sound sources (grasshoppers) at each side of head. They generated non-synchronized and strongly overlapping waveforms. Phase information in each ear resulted therefore from acoustic summation of multiple sources, hence instantaneous IPD was not directly related to a single source position and often exceeded the IPD_{max} value. Surprisingly, IPDs in the most dynamic scene - city center - did not exceed the IPD_{max} limit as often. This may be due to a smaller number of sound sources present and may indicate that the proportion of "forbidden" IPDs is a signature of a numerosity of sound sources present in the scene. For nocturnal nature and city center scenes the proportion peaked at 400 Hz achieving values of 0.45 and 0.35 respectively. For a forest walk scene, the peak at 400 Hz did not exceed the value of 0.31 at 200 Hz. All proportion curves converged

to 0 at 734 Hz frequency, where $IPD_{max} = \pi$.

The percentage of IPDs larger than the value constrained by the head size is another property of auditory scenes, which can not be predicted from head filtering properties or from physics of sound. Our data suggest that since this proportion can be large (up to 45%), many naturally encountered IPDs do not correspond to single sound sources. This in turn implies that they can not be utilized to identify the sound position in the simplest way suggested by the duplex theory.

IPDs of self-generated sounds

As already mentioned before, IPD distributions at most of frequency channels in the forest walk scene revealed an additional property, namely a clear, sharp peak at 0 radians. This feature was not present in two other scenes. As an example, IPD distribution at 561 Hz is depicted on figure 4.9 A. The histogram has a sharp peak close to 0, which implies presence of many equal monaural phase values. Zero IPDs can be generated either by sources located at the midline (directly in front or directly in the back) or self-produced sounds such as speech, breathing or loud footsteps.

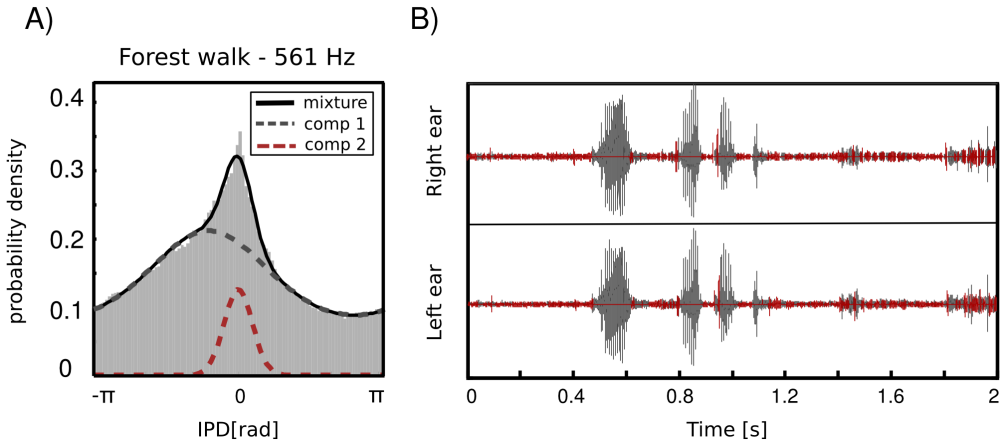

Figure 4.9: IPD distributions in an auditory scene including self-generated speech. A) An exemplary IPD distribution in the forest walk scene. In addition to a broad "background" component a peak centered at 0 radians is visible. Dashed lines mark components of a fitted von-Mises mixture distribution. B) Results of a sample classification using the fitted mixture model. Intervals were assigned by the algorithm to mixture components of the same color plotted on panel A. Blue intervals include utterances generated by the recording subject.

As visible on figure 4.9 two components contributed to the structure of the

marginal IPD distribution - the sharp "speech component" and the broad "background". IPD distributions of the forest walk scene were well suited to be modelled by a mixture model. This means that their pdf could be represented as a linear combination of two von Mises distributions in the following way

$$p(IPD_\omega|\kappa_\omega, \mu_\omega) = \sum_{i=1}^{2} p(C_i)p(IPD_\omega|\kappa_{\omega,i}, \mu_{\omega,i}) \tag{4.9}$$

where $\kappa_\omega \in \mathbb{R}^2$ and $\mu_\omega \in \mathbb{R}^2$ are parameter vectors, $C_i \in \{1, 2\}$ are class labels, $p(C_i)$ are prior probabilities of a class membership and $p(IPD_\omega|\kappa_{\omega,i}, \mu_{\omega,i})$ are von Mises distributions defined by equation 4.8. A fitted mixture of von Mises distributions is also visible in figure 4.9 A, where dashed lines are mixture components and the continuous black line is the marginal distribution. It is clearly visible that a two-component mixture fits the data much better than a plain von Mises distribution. There is also an additional advantage of fitting such a mixture model, namely it allows to perform classification problem and assign each IPD sample (and therefore each associated sound sample) to one of two classes defined by mixture components. Since prior over class labels is assumed to be uniform, this procedure is equivalent to finding a maximum-likelihood estimate $\hat{C}$ of C

$$\hat{C} = \arg\max_C p(IPD_\omega|C) \tag{4.10}$$

In this way, a separation of self generated sounds from background can be performed using information from a single frequency channel (if no other sound source is present at the midline). Exemplary results of self-generated speech separation are displayed in figure 4.9 B. A two-second binaural sound chunk included two self-spoken words with a background consisting of a flowing stream. Each sample was classified basing on an associated IPD value at 561 Hz. Samples belonging to the second, sharp component are coloured blue and background ones are red. It can be observed that the algorithm has successfully separated spoken words from the environmental noise.

IPDs are usually considered as cues generated by external sound sources. Our data demonstrate that self-generated sounds such as speech or footsteps, often constitute a dominant component of a natural acoustic scene. They also possess a characteristic statistical signature, which reflects itself in IPD distributions.

4.3.5 Independent components of binaural waveforms

In previous sections statistics of precomputed stimulus features - IPDs and ILDs were analyzed. In this way low-order properties of the natural input to binaural circuits in the auditory system were characterized. However these results do not allow to draw strong conclusions about mutual dependence of binaural waveforms. This is an important property of the stimulus, since it is informative about the

difficulty of the sound localization task in natural environments. If sounds in each ear are highly dependent - it is very likely they are generated by the same source, which can be simply localized using binaural cues. If, however, sound in the left ear is independent from the one in the right ear - this means that each of them is dominated by a different source. In such a case, instantaneous cue values can not be directly mapped to a spatial position, and sound localization becomes a complex inference process.

This section attempts to estimate the difficulty of sound localization in natural auditory scenes by analyzing mutual dependence of monaural sounds in each scene. In order to do so, Independent Component Analysis (ICA) - a statistical model which optimizes a general-purpose objective - coding efficiency [12] was employed.

In the ICA model, short (8.7 ms) epochs of binaural sounds were represented by a linear superposition of basis functions (or independent components - ICs) multiplied by linear coefficients s (see figure 4.10 A). Linear coefficients were assumed to be independent and *sparse* i.e. close to 0 for most of data samples in the training dataset. Basis functions learned by ICA can be interpreted as patterns of correlated inter- and intra-aural variability present in a dataset.

Figure 4.10 B depicts exemplary basis functions learned from each recording. Each feature consists of two parts, representing signal in the left and in the right ear (black and red colours respectively). Importantly, monaural parts of almost all trained basis functions were well localized in frequency i.e. their Fourier spectra had a prominent peak, in agreement with results presented in [78, 133, 1] (few non-localized features were excluded from the analysis - see Methods 4.2). Features trained on different recordings have characteristic shapes determined by the spectrotemporal composition of auditory scenes. On one hand, the city center scene is modelled by time extended and frequency-localized basis functions (capturing mostly the harmonics of human speech), while on the other the representation of the forest walk scene included temporally localized, instantaneous features (induced by transient sounds like wood cracks etc). Spectrotemporal characteristics of learned basis functions (depicted on figure 2 in the supplementary material) constitute a characteristic property of each auditory scene [1, 78]. Here however they are not analyzed in detail, since this is not the main focus of the current study.

In order to measure how strongly information from each ear contributed to features encoded by each of the independent components, the peak power ratio (PPR) was computed as follows:

$$PPR = 10 \log_{10}(\frac{A_{max,L}}{A_{max,R}}) \tag{4.11}$$

where $A_{max,L}, A_{max,R}$ are maximal spectrum values of left and right ear parts of each IC respectively. A large positive PPR value implies a dominance of a left

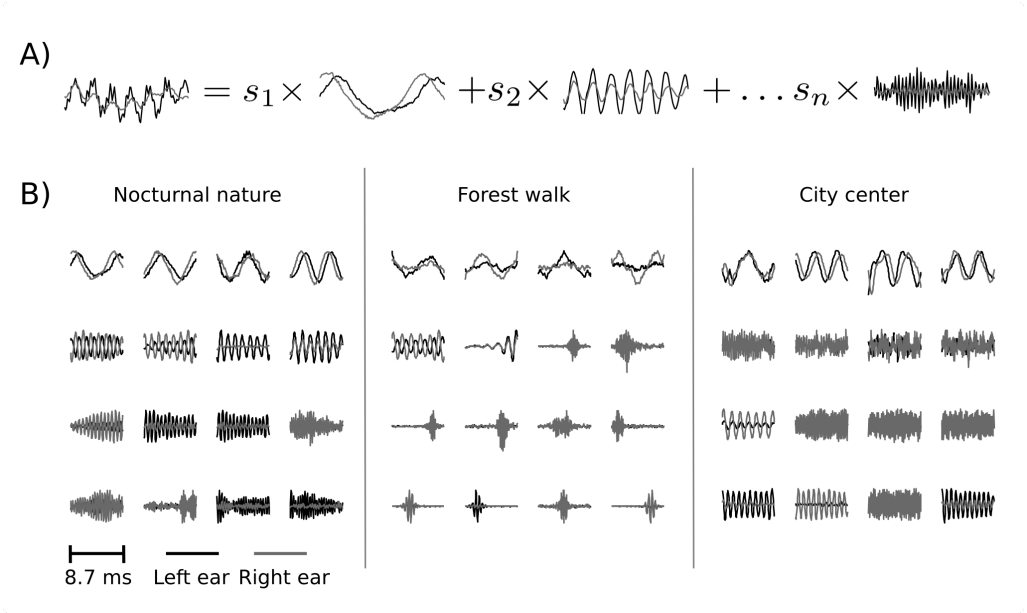

Figure 4.10: Independent components of natural binaural sounds. A) An explanation of the ICA model. Each epoch of binaural sound (left hand side of the equation) is represented by a linear combination of basis functions (or independent components). Coefficients s_i are assumed to be sparse and their joint distribution is equal to the product of marginals. B) Exemplary ICA basis functions from each recorded scene. Nocturnal nature and city center scenes consisted mostly of harmonic sounds and are mostly represented by ICs resembling Fourier bases. The forest walk scene included multiple transient sounds, which gave rise to wavelet-like features.

ear sound, while when the PPR is negative the right ear dominates. Values close to 0 imply a balanced power in each ear. This index is conceptually similar to the binocularity index used to quantify the ocular dominance of real and model visual receptive fields [63, 60].

Figure 4.11 depicts binaural properties of learned independent components. Each circle represents a single IC. Its vertical and horizontal coordinates are monaural peak frequencies and its color encodes the PPR value. Features which lie along the diagonal can be considered as a representation of "classical" ILDs, since they encoded a feature of the same frequency in each ear and differed only in level. ICs lying away from the diagonal coupled information from different frequency channels in both ears.

Pronounced differences among IC representations of the three auditory scenes are visible on figure 4.11. Majority (161) of ICs learned from the nocturnal nature scene cluster closely to the diagonal and encode the same frequency in each ear. The basis function set trained on the mostly dynamic scene (city center) separated

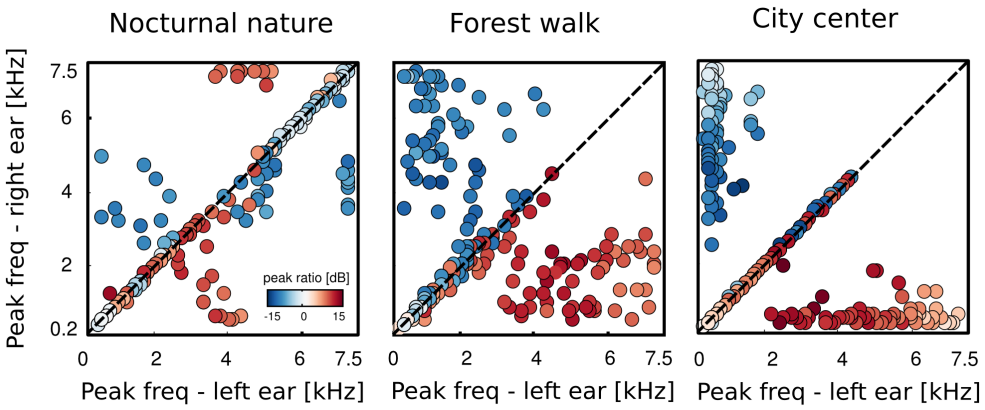

Figure 4.11: Binaural composition of independent components. Each circle corresponds to a single IC. Horizontal and vertical coordinates are spectral maxima of the left and the right ear parts respectively. Colors encode the peak power ratio. Each pannel depicts one of the recorded scenes.

into three clear subpopulations. Two of them (including 140 features in total) were monaural. Monaural basis functions were dominated mostly by a single ear, and the contralateral part was of a very low frequency, close to a flat line (a DC component). The binaural subpopulation contained 111 basis functions perfectly aligned with the diagonal. Such separation suggests that waveforms in both ears were highly independent and modelled by a large separate sets of monaural events. ICA trained on the forest walk scene also yielded a set of basis functions, separable into two populations. Here, the highest number of features - 165 lied off the diagonal and coupled separate frequency channels in each ear. A clear division into two monaural subsets was apparent - almost no IC was characterized by a PPR close to 0.

As data displayed in figure 4.11 suggest, there is a relationship between interaural redundance and PPR values. In dynamic scenes, where monaural waveforms are generated mostly by independent causes, stereo sounds are best represented by ICs of large absolute PPR values (dominated by a single ear). In order to get a better understanding of this effect, for each recorded scene, two artificial datasets of opposite properties were generated. The first dataset consisted of single, point sources presented in anechoic conditions with zero background noise. It was created by convolving chunks of a recording with human head related transfer functions (HRTFs) from the LISTEN database [151]. This dataset constituted a specific case, where sounds in each ear were maximally dependent given the head filter. In the second dataset the binaural signal was created by drawing two independent sound intervals and treating each of them as an input to a separate ear.

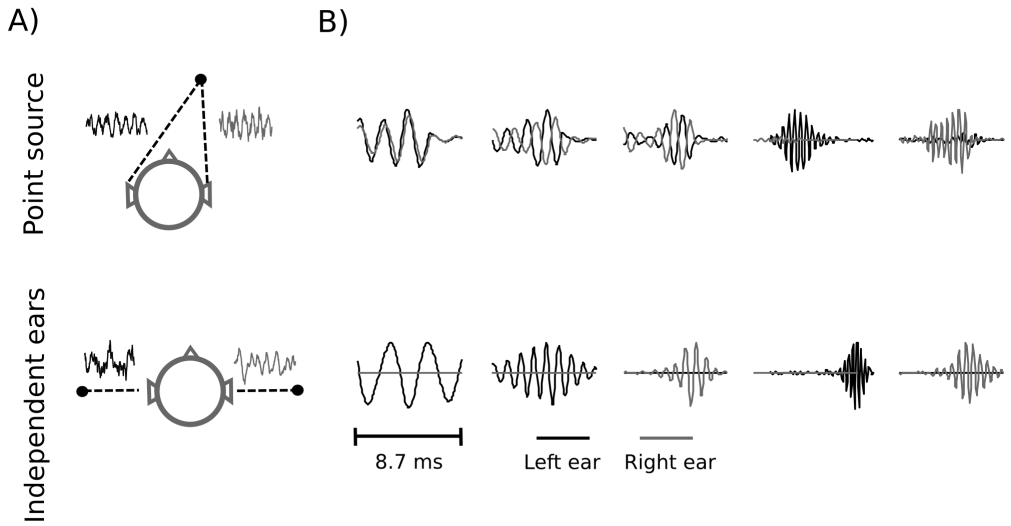

Figure 4.12: Independent components of simulated data. A) Cartoon illustrations of the generation process of the maximally dependent (top) and the maximally independent (bottom) datasets. B) Exemplary ICs trained on simulated data (point source data - top row, independent ears data - bottom row). If binaural sounds had the same underlying cause, vectors corresponding to each ear captured the signal structure. In the independent ear setting, one of the monaural parts of every IC was always flat.

The interaural dependence was therefore minimized and emulated a situation, in which sounds in each ear originate from separate sources. A cartoon illustration of those two simulations is depicted in figure 4.12 A.

Both - the point source as well as the independent ears dataset were extreme, opposite settings, which do not occur naturally. While in the first one, binaural cues could be directly mapped to a source position, in the second they were spurious and carried no spatial information. Recordings of natural scenes should lay in the space spanned by those two.

ICA was performed on each artificial dataset. Exemplary basis functions learned using sounds from the forest scene are depicted on figure 4.12 B. Top and bottom rows present ICs trained on point source and independent ears data respectively. Low frequency basis functions representing maximally dependent data (first row) had a very similar value of the spectral peak in each ear, and some of them were shifted in time (encoding an ITD). The power difference increased with frequency growth, due to the head attenuation. ICs encoding independent sounds in each ear, were almost completely monaural i.e. one of the single-ear parts was flat and equal to zero.

In the next step of the analysis, histograms of the PPR value for each learned IC dictionary were computed. They are depicted in figure 4.13. A clear, repetitive structure is visible in PPR distributions of ICs trained on artificial datasets. Histograms of point source data (first column) have three peaks - first one at 0 dB and two shorter ones, symmetrically located on either side. The middle peak, located at 0 dB corresponds to low-frequency features, which were weakly attenuated by the HRTF, and carried similar power in each ear. High frequency ICs, where sound in one ear was strongly suppressed by the head can account for the two symmetric peaks located between $\pm 10 - 15$ dB. A very different structure is visible in peak ratio histograms of ICs trained on datasets where monaural sounds were independent (middle column). There, two modes were present at extreme PPR values, close to ± 20 dB. Basis functions learned from those data were dominated by a single ear, while signal in the opposite ear was equivalent to noise fluctuations, giving rise to large absolute PPR values.

Histograms of binaural dominance of natural scene ICs are presented in the third column of figure 4.13. As expected, they fell in between extremes established by artificial datasets. Both dynamic scenes (recorded in the forest and in the city center) were characterized by PPR distributions highly similar to those obtained from independent ears data. Corresponding histograms consisted of two sharply separated peaks, located away from the 0 dB point. The distance between the peaks was, however, not as large as for the maximally independent dataset, which implied existence of some binaural dependencies. Importantly, the peak at 0 dB visible in maximally dependent datasets was absent in natural scenes. Some binaural features emerged from natural data, however in proportion to monaural ICs their amount was low. This means that monaural sounds were much less redundant than in the simplistic, simulated case. The nocturnal scene, where multiple static sources were recorded by a non-moving subject gave rise to a different PPR distribution. While the 0 dB maximum was absent as well, the positive and negative peaks were not very sharply separated. Additionally, a clear bias towards the right ear (negative PPRs) was visible. This can be accounted by the fact that this recording was performed in a static environment with a non-moving sound source present close to the right ear. Despite the almost complete lack of motion, even this scene was very different from the simulated point-source one.

The above analysis points to the fact that in a typical auditory environment, sounds in each ear are much stronger dominated by independent acoustic events that can be predicted from considerations of solitary point sources. In such conditions sound localization requires a sophisticated computational strategy and becomes itself a scene-analysis task.

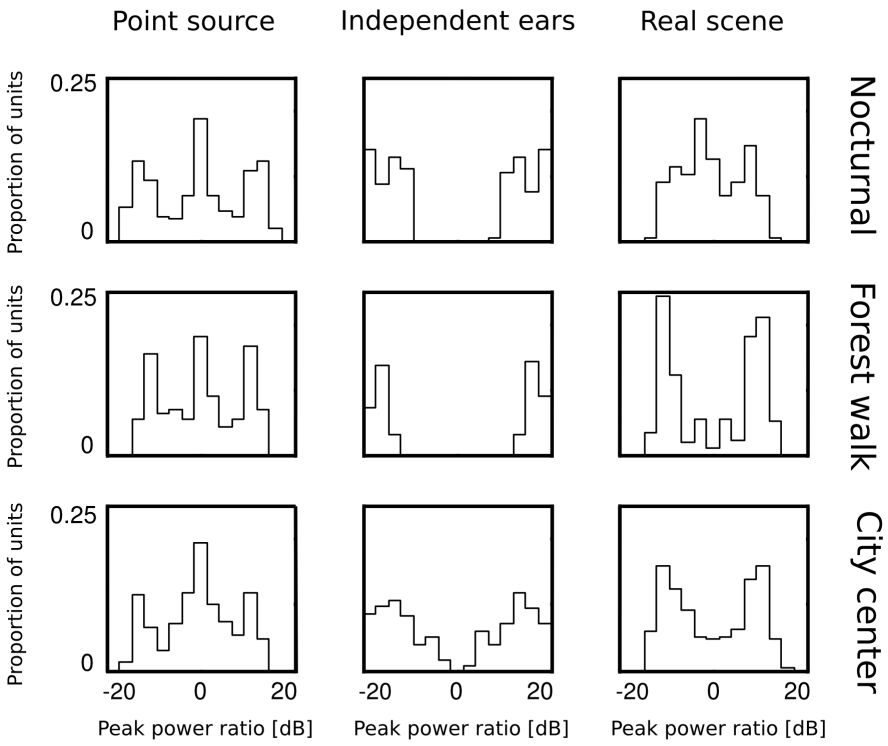

Figure 4.13: Distributions of peak power ratios of independent components trained on simulated and natural sounds. Columns correspond to datasets (point source, independent ears, natural scene) and rows to recorded environments (nocturnal nature, forest walk, city center). Simulated data gave rise to stereotypical and repetitive PPR distributions. Natural scenes, while being a compromise between simulated environments, were more similar to the independent ear data.

4.4 Discussion

Binaural cues are usually studied in a relationship to the angular position of the generating stimulus [40, 39, 55]. In probabilistic terms this corresponds to modelling a conditional probability distribution of a cue, given a sound position. Analysis of this relationship in natural environments is a very hard task, since a full knowledge about the spatial configuration of the scene (i.e. position and trajectory of every object) is required in addition to the recorded sound. Research discussed in this section approaches binaural hearing from a different perspective - it focuses on marginal distributions of naturally encountered binaural sounds.

As a representation of the real sensory world three auditory scenes were recorded and analyzed. They varied in terms of spatial configuration as well

as sound quality. This diversity increased the likelihood that any other auditory scene typically encountered by a human listener would resemble one of those recorded in the present study. Selected scenes were not free from limitations. Inspection of sound spectra as well as cue statistics revealed slight biases towards the right ear in nocturnal and forest scenes, which may not be the case in all realistic conditions. Moreover, one could invision analyzing a larger amount of recordings performed also in interior, reverberant environments, which are often encountered by humans. Such analysis should allow to draw stronger conclusions about general properties of natural binaural sounds. Despite their differences and limitations, analyzed scenes revealed common features such as the shape of ILD distributions for instance. If all analyzed cases share some statistical property, one may conclude that it should not change strongly in different hearing conditions.

4.4.1 Binaural cue distributions in natural auditory scenes

Our current understanding of how the nervous system may localize sound sources was primarily derived from considerations of solitary, point sources of pure frequency sound in noiseless and non-reverberant listening conditions. In such case, knowledge of head filtering properties and analysis based on physics of sound suffices to predict the range of possible binaural cues and their relationship to the position of a generating source.

When considering natural environments, the analytical approach very quickly becomes intractable. In a typical auditory scene, a number of objects unknown to the organism generates interfering sound waves affected by motion and reverberation. Additionally, the number of sources at each side of the head is different. Under such conditions, binaural cues become highly stochastic, and as such should be characterized in statistical terms. In this work low-order statistics of naturally encountered binaural cues were characterized. In many aspects, empirical distributions of natural stimuli deviated from reductionist, analytical predictions.

Interaural level differences

The human head strongly attenuates high frequency tones, acting as a low-pass filter [16]. For this reason, intensity differences between the ears do not carry much information about the position of a low-frequency sound. An ILD becomes informative about the location of a point-source, when the tone frequency exceeds 4 kHz [68]. Based on those observations, one could expect that naturally encountered ILDs are also strongly frequency dependent. This was however, not the case. Empirical ILD distributions were strikingly homogenous across almost entire measured frequency spectrum. Distribution at each frequency was approx-

imately logistic and centered at 0 dB. The ILD invariance to a frequency channel is not predictable by the HRTF analysis (although it has been demonstrated before that sound sources proximal to the listener can generate pronounced ILDs also below 1.5 kHz [20, 129]). Weak frequency dependence of natural ILD distributions implies that binaural circuits computing and encoding this cue are exposed to similar patterns of stimulation across large parts of the cochleotopic axis. This allows to make a prediction that similarly tuned neurons encoding both high and low frequency ILDs should be present in the early auditory system. ILD sensitive cells characterized by low best frequencies have been found in the Lateral Superior Olive (LSO) of the cat [146]. Their presence may constitute a manifestation of an adaptation of the binaural auditory system to natural ILD statistics.

A neuron maximizes its coding efficiency (defined by the amount of the stimulus information it conveys), if its tuning curve is equivalent to the cumulative distribution function (CDF) of the naturally encountered stimulus [13]. Since natural ILD distributions are logistic, one can speculate that ILD tuning curves of neurons in the early auditory system should be well approximated by a CDF of this distribution i.e. the logistic function.

In addition to the frequency invariance, ILDs revealed only a small variability across recorded auditory scenes. Despite strong differences between spatial configurations of each scene, ILD distribution parameters fluctuated very weakly. In the nocturnal nature scene, centers of some ILD distributions were slightly shifted away from 0 dB, but their shapes were the same. This observation suggests that a very similar tuning curve suffices to efficiently convey the ILD information in various listening conditions. One may conclude that ILD coding neurons do not have to strongly adapt their tuning properties, when an auditory scene changes from one to another. This does not exclude the possibility that adaptation on time scales shorter than analyzed here may still occur. Experimental evidence of a rapid adaptation to fast changes of a cue distribution has been delivered for ILDs [30] (similar effects for ITDs have also been demonstrated in [86]).

Interaural phase differences

In anechoic environments, point sources of sound generate interaural time disparities constrained by the head size of the listener - no IPD value should exceed the frequency dependent, physiological threshold. In more complex listening situations larger values can appear, either due to a sound reflection or to a presence of two (or more) desynchronized sound sources [49]. Even though large IPDs can not be directly mapped to a source position, they still may be of high value to the organism. Sound reflections generate reproducible cues and carry information about the spatial properties of the scene [46]. If a large IPD did not arise as a result of a reflection, it means that at least two sound sources contribute to the

stimulus at the same frequency. In the latter case, IPDs become a strong source separation cue.

The amount of IPDs larger than the head-imposed threshold is another property of an auditory scene, which can not be derived by the analysis of the head filtering - it has to be estimated from empirical measurements. Present results demonstrate that in low frequency channels large proportions of IPDs exceed the "maximal" value. This was true for to up to 45% of cues at around 500 Hz. It means that a large amount of potentially useful signal falls outside of the range predicted by analysis of point sources in echo-free conditions. IPD coding circuits are often exposed to cue values exceeding the threshold when the organism explores the natural environment. In order to retain this information, the auditory system should be adapted to encode IPDs larger than the physiological limit. Interestingly, this notion converges with experimental data. In many mammalian species, tuning curve peaks of IPD sensitive neurons are located outside of the head size constrained range [49]. Moreover, the observed proportion of large IPDs decreased with the frequency increase (since the maximal IPD limit increases with frequency). This observation agrees with the experimental data showing that neurons characterized by the low best frequency are predominantly tuned to IPDs lying outside of the head limit [89, 17, 50, 73]. Based on the above considerations, one can conjecture that tuning to large phase disparities could be also understood as a form of adaptation to the natural distribution of this cue.

The natural auditory stimulus consists not only of external sounds generated by environmental sources, but also of self-generated sounds such as speech. We have found that speech alters the IPD distribution by increasing the number of disparities equal to 0 radians. Distribution structure different than in scenes where no self-speech was present implies that binaural stimuli perceived by humans and other vocalizing animals are strongly affected by self generated sounds. This in turn influences activity of cue-coding neurons, since they have to represent IPDs close to 0 more often. Prior to localizing a source using binaural cues, it has to be determined, whether it is an external source or is it a self-generated one. To a limited extent this can be perfomed using instantenous, single channel IPD values as has been demonstrated here by using a simple mixture model to separate speech from background sounds. The proposed model suggests a possible abstract algorithm, which could be implemented by the nervous system to differentiate between self generated sounds and sounds of the environment. This is a behaviorally relevant task which has to be routinely performed by many animals. One should note that the separation of acoustic sources using binaural cues is a well-known paradigm of computational scene analysis and substantial research has been devoted to it in other contexts (see [18] for an exemplary review).

4.4.2 Binaural hearing in complex auditory environments

Interaural cues can be directly mapped to a stimulus position only if no other sources of sound overlap with the signal of interest. A natural question to ask is - how often does this happen in the natural environment? This is equivalent to asking - how useful are instantaneous, one-dimensional cues to localize typical, real world sources?

Since a direct estimation of a number of auditory objects in real environments is technically very difficult, present work approached this problem indirectly. By performing Independent Component Analysis, redundant patterns of natural binaural stimulus were learned. If signals in each ear originated typically from the same source - their dependence was maximized and independent components captured a signal structure in both ears. However, if sounds in each ear were dominated by independent sources, they were best represented by monaural basis functions, where the signal power in one ear was greatly exceeding power in the other one. In order to obtain a frame of reference, the same analysis using simulated datasets was performed. One of them consisted solely of solitary point-sources. Monaural sounds were therefore maximally dependent given the head filter, and sound localization could have been easily performed using simple cues. In the second dataset, sound waves in each ear were completely independent, and binaural cues carried no spatial information.

Basis functions trained on natural auditory scenes had a very different binaural composition than those trained on simulated point sources. In two out of three environments analyzed here, two equinumerous, clearly separated subsets of independent components emerged (in the third one the separation was not so prominent). Each of them was dominated by the signal in only one of the ears. This structure was rather reminiscent of basis functions trained on the artificial, maximally independent data.

These results allow to conclude that in the real-world hearing conditions binaural sound is rarely generated by a single object. Actually, sounds in each ear seem to be dominated by independent environmental causes. In such settings, an inversion of a binaural cue to a sound source position becomes an ill-posed problem. This is because multiple scene configurations can give rise to the same cue value (for instance an ILD equal to 0 can be generated by a single source located at the midline, or two identical sources symmetricaly located on both sides of the head). A mere extraction of the instantenous cue (as performed by the brainstem nuclei MSO and LSO) is not equivalent to the identification of the sound position. Computation of binaural cues is only a beginning of a complex inference process, whose purpose is to estimate the spatial configuration of an auditory scene [80].

The ICA analysis has yielded a large amount of monaural and a smaller number of binaural features. One can interpret them as model neuronal receptive

fields [78, 133, 25], and ask which role could neurons of such response characteristics play. One possible answer is that while binaural neurons may subserve localization tasks, monaural ones could be used for the purpose of the "better ear listening" i.e. encoding ipsilateral sound sources. On the other hand also monaural sound features similar to ones described here can be utilized in further stages of the auditory processing to recover spatial information.

4.5 Conclusions

Properties of naturally encountered binaural sounds deviate from predictions formulated in limited, experimental settings. Many aspects of cue distributions such as an ILD frequency invariance, or a proportion of IPDs larger than the "physiological" head-imposed limit can not be predicted from the analysis of simple stimuli. This is an example showing that even low-order properties of the natural sensory input are hard to be predicted from analytical, physics-based considerations.

An often repeated statement is that the function of MSO and LSO - binaural comparators located in the brainstem is to localize sound sources [49]. While those structures most surely compute interaural time and level differences, the ICA based analysis presented here has demonstrated that under natural conditions the extraction of a cue does not immediately correspond to an estimation of the source position. The function played by substructures of the olivary complex in spatial hearing may be more transformative i.e. to preprocess the signal and extract cues, which subserve further scene-analysis processing.

The first point I argue for in this thesis states that without analyzing the structure of a natural sensory input processed by neural circuits it is nearly impossible to explain algorithms they implement and the function they play in sensory computations. Results presented in this chapter seem to confirm that this statement holds in the case of binaural hearing.

Chapter 5

Sparse Representation of Natural Stereo Sounds Reproduces Neuronal Codes in the Auditory Cortex

5.1 Overview

When considering the notion of function in the nervous system, the auditory cortex provides a particularly mysterious example. Despite its obvious importance, the precise role played by this area in hearing remains unclear. Before reaching the cortex, raw sounds undergo numerous transformations in the brainstem and the thalamus. The subcortical processing seems to be more substantial than in other senses and constitutes a specific property of the auditory system. What are the computations performed by the cortex on the output generated by lower auditory regions is a question far from being answered.

One of the issues making functional characterization of the auditory cortex a conceptually difficult task, is an apparent lack of specificity. Spiking activity of cortical auditory neurons is modulated by multiple sound features such as pitch, timbre and spatial location [15, 53]. Responses invariant to any of those aspects seem to be rare. This interdependence is especially puzzling in the context of extracting spatial information. Despite efforts to identify "what" and "where" streams in the auditory system (e.g. [120, 84]), no clear signature of a sharp separation has been found [100, 28].

Neurons reveal sensitivity to sound position in most parts of the mammalian auditory cortex [14]. Their spatial tuning is quite broad - neural firing can be

modulated by sounds located on the entire azimuthal plane. While activity of single units does not carry information sufficient to accurately localize sounds, larger numbers of neurons seem to form a population code for sound location [137, 93, 139, 155]. These observations stand against initial expectations of finding a topographic cortical map of the auditory space, where neighboring units would encode presence of a sound source at proximal positions in the area surrounding the animal [92].

From a theoretical perspective one question seems to be particularly important - is there any general principle behind functioning of the auditory cortex, or does it carry out computations which are purely task- or modality-specific and are therefore not performed in other parts of the nervous system? A growing body of evidence seems to point to efficient coding as an abstract computational mechanism implemented by the auditory system. To date however, the connection between natural stimulus statistics and auditory spatial receptive fields remains unexplained. It is therefore unclear if spatial computations performed by the auditory cortex are unique to this brain area or whether they can be also predicted in a principled way from a broader theoretical perspective.

Work described in this chapter attempts to connect spatial computations carried by the auditory cortex with statistics of the natural stimulus. Here, a hierarchical model of stereo sounds recorded in a real auditory environment is proposed. Based on principles of sparse coding the model learns the spectrotemporal and interaural structure of the stimulus. In the next step, it is demonstrated that when probed with spatially localized sounds, higher level units reveal spatial tuning which very well matches spatial tuning of neurons in the mammalian auditory cortex. Additionally, the learned code forms an interdependent representation of spatial information and spectrotemporal quality of a sound. Activity of higher units is therefore modulated by sound's position and identity, as observed in the auditory system.

Results I describe here suggest that the function of the auditory cortex is to reduce redundancy of the stimulus representation preprocessed by the brainstem. Representation obtained in this way can be hard to be described in terms of selectivity for abstract features of sound such as pitch, timbre or location. At the same time, they may facilitate tasks performed by higher brain areas such as sound localization

5.2 Methods and Models

5.2.1 Overview of the hierarchical model

In this chapter a hierarchical statistical model of binaural sounds, which captures binaural and spectrotemporal structure present in natural stimuli is proposed. The architecture of the model is shown in figure 5.1. It consists of the input layer

and two hidden layers. The input to the model were N samples long epochs of binaural sound: from the left ear - x_L and from the right ear - x_R. The role of the first layer was to extract and separate phase and amplitude information from each ear by encoding them in an efficient manner. Monaural sounds were transformed into phase (ϕ_L, ϕ_R) and amplitude (a_L, a_R) vectors. This layer can be thought of as a statistical analogy to cochlear filtering. Phase vectors were further modified by computing Interaural Phase Differences (IPDs) - a major sound localization cue [49].

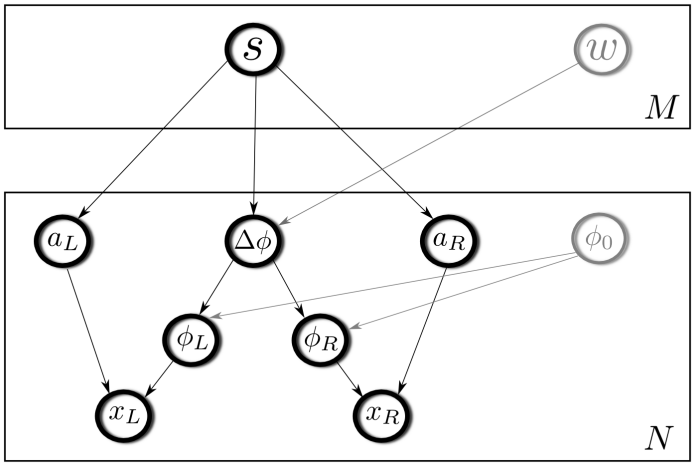

Figure 5.1: The graphical model representing variable dependencies. The lowest layer represents sound epochs perceived by the left and the right ear x_L and x_R. They are decomposed by a sparse coding algorithm into phase and amplitude vectors ϕ_L, ϕ_R and a_L, a_R. Phases are further subtracted from each other in order to obtain an IPD vector $\Delta\phi$. The second layer encodes jointly monaural amplitudes and IPDs. Auxiliary variables (phase offset and the scaling factor w) are depicted in gray.

The second layer of the model learned a joint sparse representation of monaural amplitudes (a_L, a_R) and phase differences ($\Delta\phi$). Level (amplitude) and temporal (phase) information from each ear was jointly encoded by a population of M units. Each of them was therefore capturing higher-order spectrotemporal patterns of sound in each ear. Additionally by combining monaural information into single units higher level representation achieved spatial tuning not present in the first layer. The second hidden layer was constructed as a model of cortical auditory neurons, which receive converging monaural input. An additional assumption was that they jointly operate on phases and amplitudes - two kinds of information, which is known to be important for spatial hearing.

5.2.2 First layer - sparse, complex-valued representations of natural sounds

As demonstrated in previous work, filtering properties of the auditory nerve can be explained by sparse coding models of natural sounds [78]. There, short epochs of natural sounds are modelled as a linear combination of real-valued basis functions multiplied by sparse (i.e. of highly curtotic marginal distributions), independent coefficients. Adapted to sets of natural sound chunks, basis functions become localized in time and/or frequency matching properties of cochlear filters.

While being capable of capturing interesting properties of the data, real valued representations are not well suited for modelling binaural sounds. This is because binaural hearing mechanisms utilize interaural level and time differences (ILDs and ITDs respectively). In pure frequency channels, differences in time correspond to phase displacements known as interaural phase differences (IPDs). Therefore a desired representation should both be adapted to the data (i.e. non-redundant) and separate amplitude from phase (where phase is understood as a temporal shift smaller than the oscillatory cycle of a particular frequency).

The present work addresses this twofold constraints with the complex-valued sparse coding. Each data vector $x \in \mathbb{R}^N$ is represented as:

$$x_t = \sum_{i=1}^{N} \Re\{z_i^* A_{i,t}\} + \eta \tag{5.1}$$

where $z_i \in \mathbb{C}$ are complex coefficients, $*$ denotes a complex conjugation, $A_i \in \mathbb{C}^T$ are complex basis functions and $\eta \sim \mathcal{N}(0, \sigma)$ is additive Gaussian noise. Complex coefficients in Euler's form become $z_i = a_i e^{j\phi_i}$ (where $j = \sqrt{-1}$) therefore equation (5.1) can be rewritten to explicitly represent phase ϕ and amplitude a as separate variables:

$$x_t = \sum_{i=1}^{N} a_i (\cos\phi_i A_{i,t}^{\Re} + \sin\phi_i A_{i,t}^{\Im}) + \eta \tag{5.2}$$

Real and imaginary parts $A_i^{\Re}$ and $A_i^{\Im}$ of basis functions $\{A_i\}_{i=1}^N$ span a subspace within which the position of a data sample is determined by amplitude a_i and phase ϕ_i. Depending on number of basis functions N (each of them is formed by a pair of vectors), the representation can be complete ($N/2 = T$) or overcomplete ($N/2 > T$).

In a probabilistic formulation, equations (5.1) (5.2) can be understood as a likelihood model of the data, given coefficients z and basis functions A:

$$p(x|z, A) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^T} \prod_{t=1}^{T} e^{-\frac{(x_t - \hat{x}_t)^2}{2\sigma^2}} \tag{5.3}$$

where $\hat{x}_t = \sum_{i=1}^{N} \Re\{z_i^* A_{i,t}\}$. A prior over complex coefficients applied here assumes independence between subspaces and promotes sparse solutions i.e. solutions with most amplitudes close to 0:

$$p(z) = \frac{1}{Z} \prod_{i=1}^{N} e^{-\lambda S(a_i)} \tag{5.4}$$

where Z is a normalizing constant. Function $S(a_i)$ promotes sparsity by penalizing large amplitude values (the above equation has the same form as the factorial coefficient prior 2.18). Here, a Cauchy prior on amplitudes is assumed i.e. $S(a_i) = log(1 + a_i^2)$. One should note however that amplitudes are always non-negative and that in general the Cauchy distribution is defined over the entire real domain. The model attempts to form a data representation keeping complex amplitudes maximally independent across subspaces, while still allowing dependence between coordinates $z^{\Re}, z^{\Im}$ which determine position within each subspace. Inference of coefficients z which represent data vector x in the basis A is performed by minimizing the following energy function

$$E_1(z, x, A) \propto \frac{1}{2\sigma^2} \sum_{t=1}^{T} (\hat{x}_t - x_t)^2 + \lambda \sum_{i=1}^{N} S(a_i) \tag{5.5}$$

which corresponds to the negative log-posterior $p(z|x, A)$. This model was introduced in [24] and used to learn motion and form invariances from short chunks of natural movies. Assuming $N = T/2$ and $\sigma = 0$, it is equivalent to 2-dimensional Independent Subspace Analysis(ISA) [61].

When trained on natural image patches, real and imaginary parts of an overwhelming majority of basis functions A form pairs of Gabor-like filters, which have the same frequency, position, scale and orientation (see figure 5.2). The only differing factor is phase - real and imaginary vectors are typically in a quadrature-phase relationship (shifted by $\frac{\pi}{2}$). By extension, one may expect that the same model trained on natural sounds should form a set of frequency localized phase-invariant subspaces, where imaginary vector is equal to the real one shifted a quarter of a cycle in time. Somewhat surprisingly such representation does not emerge, and learned subspaces capture different data aspects - bandwidth, frequency or time invariance [150, 96] as depicted on figure 5.3.

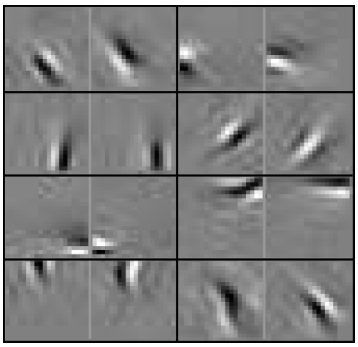

Figure 5.2: Complex basis functions trained on natural image patches. With one exception, they are formed by pairs of quadrature phase vectors. Other parameters such as scale, frequency, location and orientation remain the same. They are therefore invariant to spatial shifts.

Natural sounds possess strong cross-frequency correlations [144] and other highly non-local features. Reflecting this structure, sparse, complex codes of natural acoustic stimuli capture frequency and bandwidth invariances. Only a small fraction is phase-shift (or time) invariant [150]. Figure 5.3 depicts four examples of complex basis functions learned by from natural, speech sounds. In addition to temporal plots in Cartesian (first row) and polar (second row) coordinates each basis function is also depicted in the frequency domain (third row). Real ($A_i^{\Re}$ - black lines) and imaginary ($A_i^{\Im}$ - gray lines) parts of basis functions do not resemble each other and are not temporally localized, capturing the non-local structure of speech sounds.

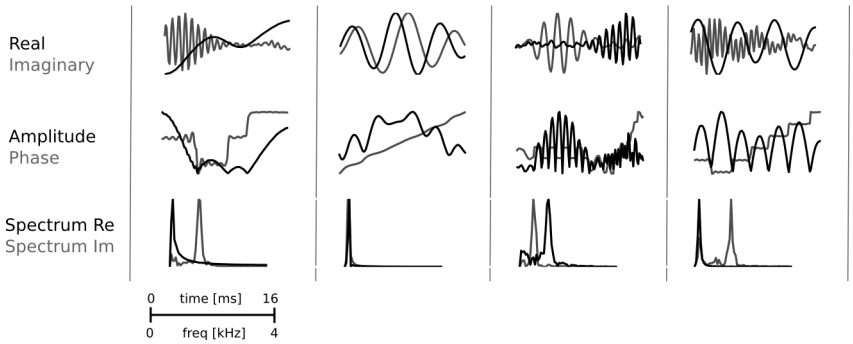

Figure 5.3: Complex basis functions trained on speech sounds. Representations in cartesian and polar coordinates are depicted in the first and second row respectively. The third row depicts Fourier spectra of real and imaginary parts. As visible, they do not capture phase invariance.

Phase variable within such subspaces does not correspond to the temporal shift. Therefore sound representations learned with the basic version of the

complex-valued sparse coding are not suitable for modeling spatial hearing, where interaural phase difference is imposed by time delay. In two following subsections I describe alternative solutions to finding a non-redundant data representation, which preserves desired property of phase-invariance.

Phase and amplitude continuity priors

In order to learn from the statistics of the data a representation that preserves a desired property such as phase invariance, one could select a parametric form of basis functions and adapt the parameter set [147]. This method has been applied before to audio data by adapting a gammatone dictionary [156]. Despite many advantages of this solution, there exists a possibility, that the parametric form of dictionary elements is not flexible enough to efficiently span the data space. To alleviate this problem in this section I propose to learn a sparse and complex representation of natural sounds with the phase-invariance promoting priors. Proposed priors induce temporal continuity, i.e. *slowness* [41, 154] of both phase and amplitude, which turns out to be a correct assumption for learning phase-invariant features.

The sparse coding model described in the previous subsection, does not constrain the basis functions in any way. They are allowed to vary freely during the learning process. As visible on figure 5.3, an unconstrained adaptation to natural sound corpus yields complex basis functions invariant to numerous stimulus aspects such as frequency or time shifts, not necessarily phase.

Learning a structured dictionary requires therefore placing priors over basis functions, which favour solutions of desired properties such as phase-invariance. Real and imaginary parts of a phase-shift invariant basis function, have equal, unimodal frequency spectra and both span the same temporal interval. Additionally, the imaginary part should be shifted in time a quarter of the cycle with respect to the real one.

Before proposing a prior promoting such solutions, it should reminded that each temporal basis function $A_{i,t}$ can be represented in polar coordinates in the following way:

$$A_{i,t} = a^A_{i,t}\Big(\cos\phi^A_{i,t} + j\sin\phi^A_{i,t}\Big) \tag{5.6}$$

In such representation variables $a^A_{i,t}$ and $\phi^A_{i,t}$ denote instantaneous phase and amplitude respectively. Angular frequency can be defined as a temporal derivative of instantaneous phase. If phase dynamics are highly variable and non-monotonic over time, real and imaginary components of this signal have non-identical spectra and/or their frequencies change in time (see figure 5.3, second and third rows). On the other hand, by enforcing phase $\phi^A_{i,t}$ to change smoothly and monotonically, one should obtain real and imaginary parts with matching frequency spectra. In

the limiting case, when phase is a linear function of time, real and imaginary parts oscillate in the same frequency and are in a quadrature phase relationship. Furthermore, vectors which span a phase-shift invariant subspace should have the same temporal support, implying that the complex amplitude should also vary slowly in time.

In order to learn a phase-shift invariant representation of natural sounds, the present section proposes a prior over basis functions of the following form:

$$p(A_i) = p_\phi(A_i)p_a(A_i) = \frac{1}{Z}e^{-(\gamma S_\phi(A_i)+\beta S_a(A_i))} \tag{5.7}$$

Function $S_a(A_i)$ introduces the penalty proportional to the variance of amplitude's temporal derivative:

$$S_a(A_i) = \sum_{t>1}^{T}\left(\Delta a_{i,t}^A\right)^2 \tag{5.8}$$

where $\Delta a_{i,t}^A = a_{i,t}^A - a_{i,t-1}^A$. It promotes basis functions with a slowly-varying envelope, highly correlated between consecutive time steps. Phase prior S_ϕ is defined by function $S_\phi(A_i)$ of the following form:

$$S_\phi(A_i) = -\sum_{t>1}^{T} sgn\left(\Delta\phi_{i,t}\right)\left(\Delta\phi_{i,t}\right)^2 \tag{5.9}$$

where

$$\Delta\phi_{i,t}^\phi = \phi_{i,t}^A - \phi_{i,t-1}^A \tag{5.10}$$

and sgn denotes the sign function. Similarly to $S_a(A_i)$ it promotes temporal slowness of phase. The additional factor $-sgn(\Delta\phi_{i,t})$ enforces $\phi_{i,t}$ to be larger than $\phi_{i,t-1}$. In this way, it prevents phase from changing direction and causes it to be a non-increasing function of time. One could also enforce this by bounding the phase derivative from above: $\Delta\phi_{i,t} < \Theta$. This method would however require the hand tuning of the Θ parameter. The posterior over basis functions given a data sample x and its representation s becomes:

$$p(A|x,z) \propto p(x|A,z)p(A) \tag{5.11}$$

where the likelihood model $p(x|A,z)$ is defined by equation (5.3). Taken together, prior $p(A_i)$ biases the learning process towards temporally localized basis functions with real and imaginary parts of the same instantaneous frequency.

Exemplary complex features learned with introduced priors are depicted on figure 5.4. Compared with unconstrained subspaces from figure 5.3 , their amplitudes are smooth, and their phases change monotonically. Moreover, frequency spectra of $A_i^{\Re}$ and $A_i^{\Im}$ align much better. Such bases form a phase invariant representation of the data.

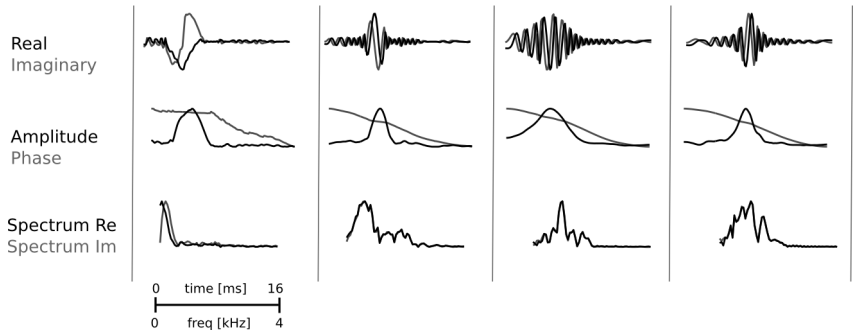

Figure 5.4: Exemplary complex basis functions trained on speech sounds with phase and amplitude continuity priors. Representations in cartesian and polar coordinates are depicted in the first and second row respectively. The third row depicts Fourier spectra of real and imaginary parts.

However, due to the statistics of the sound ensemble their spectra become much broader and non-localized. Instantaneous phase is a monotonic function of time, however its temporal derivative is not constant. This manifests as an increase of frequency, which positively correlates with amplitude. The tradeoff between prior and likelihood terms yields basis functions of not easily interpretable structure.

Sparse code extended with Hilbert transform

As described in sections above, unconstrained complex sparse coding leads to emergence of features, which are predominantly not phase invariant. On the other hand, dictionaries learned with continuity promoting priors loose frequency precision, which makes them difficult to interpret and compare with cochlear filters.

This section describes a different, "semi-supervised" approach to learn an appropriate signal representation. Firstly a real-valued sparse code was trained on chunks of natural sound (see section Sparse Coding). Learned basis functions were well localized in time or frequency and tiled the time-frequency plane in a uniform and non-overlapping manner (figure 5.5 B). They were taken as real vectors $A^{\Re}$ of complex basis functions A. In the second step, imaginary parts were created by performing the Hilbert transform of real vectors. The Hilbert transform of a time varying signal $y(t)$ is defined as follows:

$$H(y(t)) = \frac{1}{\pi} p.v. \int_{-\infty}^{\infty} \frac{y(\tau)}{t-\tau} d\tau \tag{5.12}$$

Where $p.v.$ stands for Cauchy principal value. In such a way every real vector $A_i^{\Re}$ was paired with its Hilbert transform $A_i^{\Im} = H(A_i^{\Re})$ i.e. a vector which complex Fourier's coefficients are all shifted by $\frac{\pi}{4}$ in phase. The obtained dictionary is

adapted to the stimulus ensemble, hence provides a non-redundant data representation yet it makes phase clearly interpretable as a temporal displacement.

5.2.3 Second layer - intermediate level representation of binaural sound

In an approach to model the cochlear coding of sound, monaural sound epochs x_L and x_R were encoded independently using the same dictionary of complex basis functions A described in the previous section. Signal from both ears converged in the second hidden layer, which role was to form a joint, higher-order representation of the entire stimulus processed by the auditory system.

The celebrated Duplex Theory of spatial hearing specifies two kinds of cues used to solve the sound-localization task: Interaural Level and Time (or Phase) differences [143]. While IPDs are supposed to be mostly used in localizing low-frequency sounds, ILDs are a cue, which (at least in the laboratory conditions) becomes useful to identify the position of high frequency sources. Phase and level cues are known to be computed in Lateral and Medial Superior Olive (LSO and MSO respectively) - separated anatomical regions in the brainstem [49]. However, an assumption made here was that neurons in the auditory cortex receive converging input from subcortical structures. This would enable them to form their spatial sensitivity using both fine structure phase and amplitude information. One can take also the inverse perspective: a single object (a "cause") in the environment generates level and phase cues at the same time. Its identification has therefore to rely on observing dependencies between those features of the stimulus.

The second layer formed a joint representation of monaural amplitudes and interaural phase differences. However, not all IPDs were modelled in that stage. Humans stop to utilize fine structure IPDs in higher frequency regimes (roughly above 1.3 kHz), since this cue becomes ambiguous [49]. Additionally, cues above around 700 Hz become ambiguous (a single cue value does not correspond to a unique source position). For those reasons and in order to reduce the number of data dimensions, 20 out of 128 IPD values were selected. The selection criteria were the following: (i) an associated basis function should have the peak of the Fourier spectrum below 0.75 kHz (which provided the upper frequency bound), and (ii) it should have at least one full cycle (which provided the lower bound). All basis function fulfilling these criteria were non-localized in time (they spanned entire 16 ms interval). In result, the second layer of the model was jointly encoding $T = 128$ log-amplitude values from each ear and $P = 20$ phase differences.

Monaural log-amplitude vectors $a_L, a_R \in \mathbb{R}^T$ where concatenated into a single vector $a \in \mathbb{R}^{2\times T}$, and encoded using a dictionary of amplitude basis functions B. Representation of IPDs ($\Delta\phi$) was formed using a separate feature dictionary ξ. Both - phase and amplitude basis functions (B and ξ), were coupled by associated

sparse coefficients s_i. The overall generative model of phases and amplitudes was defined in the following way:

$$a = \sum_{i=1}^{M} s_i B_i + \eta \tag{5.13}$$

$$\Delta\phi = |w| \sum_{i=1}^{M} s_i \xi_i + \epsilon \tag{5.14}$$

The amplitude noise was assumed to be gaussian ($\eta \sim \mathcal{N}(0, \sigma_2)$) with σ_2 variance. Since phase is a circular variable its noise ϵ was modelled by the von Mises with concentration parameter κ.

The second layer was encoding two different physical quantities - phases, which are circular values, and log-amplitudes, which are real numbers. The goal was to form a joint representation of both parameters and learn their dependencies from the data. A simple, linear sparse coding model could be in principle used to achieve this task. However, if a single set of sparse coefficients s_i was used to model both quantities, scaling problems could arise, namely a coefficient value which explains well the amplitude vector may be too large or too small to explain the concomitant IPD vector. For this reason an additional phase multiplier w was introduced. It enters equation 5.13 as a scaling factor, which gives the model additional flexibility required to learn joint probability distribution of amplitudes and IPDs. Figure 5.1 depicts it in gray as an auxiliary variable. In this way, amplitude values and phase differences were modelled by variables sharing a common, sparse support (coefficients s), with a sufficient flexibility.

Seeking analogies between the higher-level representation and auditory neurons, coefficients s can be interpreted as neuronal activity (e.g. firing rate) and pairs of basis functions B_i, ξ_i as receptive fields. An $i-th$ second-layer unit was activated ($s_i \neq 0$) whenever a pattern of IPDs represented by the basis function ξ_i or a pattern of amplitudes represented by B_i was present in its receptive field.

The likelihood of amplitudes and phase differences defined by the second layer was given by:

$$p(a, \Delta\phi | s, w, B, \xi) = \frac{1}{\left(\sigma_2\sqrt{2\pi}\right)^{2\times T}} \prod_{n=1}^{2\times T} e^{-\frac{(a(n)-\hat{a}(n))^2}{2\sigma_2^2}} \frac{1}{\left(2\pi I_0(\kappa)\right)^P} \prod_{m=1}^{P} e^{\kappa\cos(\Delta\phi_m - \widehat{\Delta\phi}_m)} \tag{5.15}$$

where $\hat{a} = \sum_{i=1}^{M} s_i B_i$, $\widehat{\Delta\phi} = |w| \sum_{i=1}^{M} s_i \xi_i$, and I_0 is the modified Bessel function of order 0. The joint distribution of coefficients s was assumed to be equal to the product of marginals:

$$p(s) = \frac{1}{Z} \prod_{i=1}^{M} e^{-\lambda_2 S(s_i)} \tag{5.16}$$

where λ_2 is a sparsity controlling parameter. A Cauchy distribution was assumed as a prior over marginal coefficients (i.e. $S(s_i) = \log(1+s_i^2)$). To prevent degenerate solutions, where sparse coefficients s are very small and the scaling coefficient w grows unbounded, a prior $p(w)$ constraining it from above and from below was placed. A generalized Gaussian distribution of the following form was used:

$$p(w) = \frac{\beta}{2\alpha\Gamma(\frac{1}{\beta})} e^{-\left(\frac{|w-\mu|}{\alpha}\right)^{\beta}} \tag{5.17}$$

Γ denotes tha gamma function, α, β and μ denote the scale, shape and location parameters respectively. When the shape parameter β is set to a large value (here $\beta = 8$), the distribution approximates a uniform distribution. Varying the scale parameter α changes the upper and the lower limit of the interval.

Taken together the negative log-posterior over the second layer coefficients was defined by the energy function:

$$\begin{aligned} E_2(s, w, B, \xi) \propto \frac{1}{\sigma_2^2} \sum_{n=1}^{2\times T} (a_n - \hat{a}_n)^2 + \kappa \sum_{m=1}^{P} \cos(\Delta\phi_m - \widehat{\Delta\phi}_m) \\ + \lambda_2 \sum_{i=1}^{M} S(s_i) + \lambda_w \left(\frac{|w-\mu|}{\alpha}\right)^{\beta} \end{aligned} \tag{5.18}$$

the λ_w coefficient was introduced to control the strength of the prior on the scaling coefficient w. Similarly as in the first model layer, learning of basis functions and inference of coefficients was performed using gradient descent (see Appendix). The total number M of basis function pairs was set to 256.

5.2.4 Simulation details and analysis methods

Sound data

Altogether 75000 epochs of binaural sound randomly drawn from a 60 second-long excerpt from the forest walk recording described in chapter 4 were used to train the model. Each of them was $T = 128$ samples long, which corresponded to 16 ms. Both layers were trained separately. Before training the first layer, Principal Component Analysis was performed and 18 out of 128 principal components were rejected, which corresponded to low pass filtering the data. Left and right ear sound epochs were shuffled together to create a 150000 sample training dataset for the first layer. The first layer sparsity coefficient λ was set to 0.2. Noise variance λ^2 was equal to 2. The sparse coding algorithm converged after 200000 iterations.

A complex-valued dictionary was created by extending the real valued one with Hilbert-transformed basis functions. Amplitude and phase vectors a and ϕ were inferred for each sample using 20 gradient steps. Amplitude vectors were

concatenated and transformed with a logarithmic function, and IPD vectors $\Delta\phi$ were computed by substracting left ear phase vectors ϕ_L from right ear ones ϕ_R. The second layer was trained by performing 250000 gradient updates on basis functions B and ξ. The amplitude sparsity coefficient λ_2 was set to 1. The λ_w parameter was set to 0.01 and the noise variance λ_2^2 as well as the von Mises concentration parameter κ were set to 2.

Test recordings used to map the spatial tuning of second-layer units was performed in an anechoic chamber at the Department of Biology, University of Leipzig. The same recording subject was seated in the middle of the chamber. A female speaker walked in a constant pace following a circular path surrounding the recording subject. While walking she was counting out loud. This was repeated four times. The second test recording was performed in a similar fashion, however instead of speaking the walking person was rubbing two pieces of carton against each other, generating a broad-band sound. To estimate conditional distribution of sparse coefficients given the position and identity of the sound, test recordings were divided into 18 intervals, each corresponding to the same position on a circle.

All recordings were registered in an uncompressed wave format at 44100 Hz sampling rate. Prior to training the model, sounds were downsampled to 8000 Hz. Test recordings are available in the supplementary material.

Computation of modulation spectra of second-layer basis functions

Spectrograms of amplitude basis functions B_i were computed by combining spectrograms of real, first layer basis functions $A_n^{\Re}$, linearly weighted by a corresponding weight $\exp(B_{i,n})$. First layer spectrograms were computed using $T = 29$ windows, each 16 samples (0.002 second) long, with a 12 sample overlap. Altogether, $F = 128$ logarithmically-spaced frequencies were sampled. A two-dimensional Fourier transform of each spectrogram was computed using the Matlab built-in function fft2. The amplitude spectrum of obtained transform is called the Modulation Transfer Function (MTF) of each second layer feature [131]. The center of mass i.e. the point $(C_{S,i}^f, C_{S,i}^t)$ of each monaural part ($S \in \{L, R\}$) of basis functions B_i was computed in the following way:

$$C_{S_i}^t = \sum_t t \sum_f MTF(B_{S,i}) \tag{5.19}$$

$$C_{S_i}^f = \sum_f f \sum_t MTF(B_{S,i}) \tag{5.20}$$

where t and f are time and frequency respectively.

Estimation of spatial tuning curves

To estimate conditional distribution of sparse coefficients given the position and identity of the sound, test recordings of a sound source (either speech, or rubbed

paper) moving around the recording subject were used. Each source circumvented the recording person 4 times resulting in 4 recordings. Each of them was divided into 18 intervals. Intervals corresponding to the same area on the circle were pooled together across all recordings. For each out of 18 sound positions 3000 random sound chunks were drawn and encoded by the model. Position-conditional ensembles were then used to compute conditional histograms. Conditional mean vectors $\mu_{i,\theta}$ were computed by averaging all values of coefficient s_i at position θ. Mean vectors were mapped to a $[0, 1]$ interval by adding the absolute value of a minimal entry and dividing it by the value of the maximum. For plotting purposes of figure 5.13, endings of tuning curves were connected if values at $-180°$ and $180°$ were not equal.

Decoding of stimulus position

The decoding analysis was performed using K second-layer sparse coefficients s averaged over D of samples. The response vectors $d \in \mathbb{R}^K$ were therefore formed as:

$$d = \frac{1}{D} \sum_{i=1}^{D} s_{\{1,\ldots,K\}} \tag{5.21}$$

. Such averaging procedure can be interpreted as an analogy to computation of firing rates in real neurons.

The marginal distribution of response coefficients d over all 18 sound positions $\theta \in \{-180°, -160°, \ldots, 160°, 180°\}$ was equal to:

$$p(d) = \sum_{\theta} p(d|\theta)p(\theta) \tag{5.22}$$

where each conditional $p(d|\theta)$ was a K-dimensional Gaussian distribution with class specific mean vector μ_θ and covariance matrix C_θ:

$$p(d|\theta) = \mathcal{N}(\mu_\theta, C_\theta) \tag{5.23}$$

The prior over class labels $p(\theta)$ was uniformly distributed i.e. $p(\theta_i) = \frac{1}{18}$ for each i.

The decoding procedure iterated over all class labels and returned the one, which maximized the likelihood of the observed data vector. Out of the entire dataset, 80% was used to train the model and remaining 20% to test and estimate the confusion matrix.

Confusion matrix M was a joint histogram of a decoded and true sound position $\hat{\theta}$ and θ. After normalization, it was an estimate of a joint probability mass function $p(\hat{\theta}, \theta)$. Mutual information was estimated from each confusion matrix as:

$$MI(\hat{\theta}\ \theta) = \sum_{\hat{\theta}} \sum_{\theta} p(\hat{\theta}, \theta) \log_2 \left(\frac{p(\hat{\theta}, \theta)}{p(\hat{\theta}) p(\theta)} \right) \tag{5.24}$$

5.3 Results

5.3.1 Properties of the first layer representation

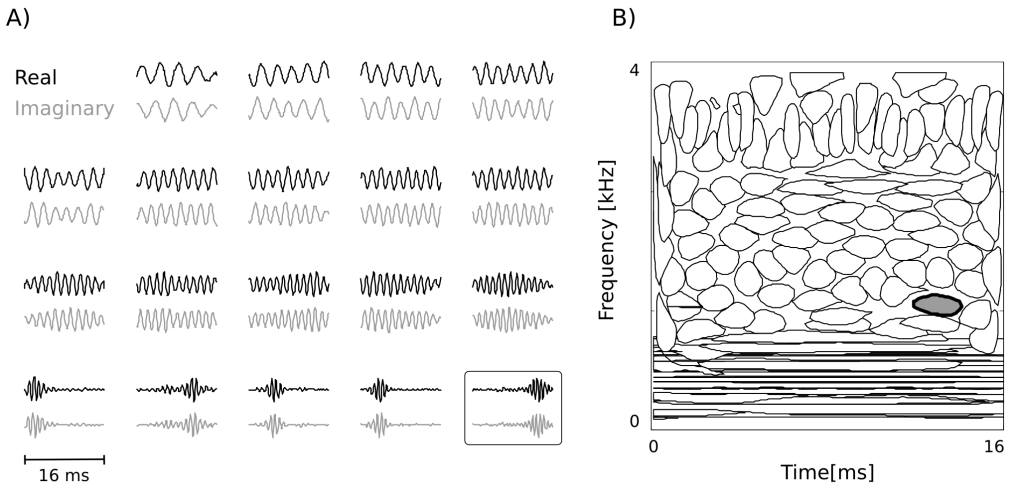

Figure 5.5: First layer basis. A) Exemplary real (black) and imaginary (gray) vectors. B) Isoprobability contours of Wigner-Ville distributions associated with each real vector. Time - frequency plane is tiled uniformly with a weak overlap. Gray-filled oval corresponds to the framed basis function on panel A.

The model was trained using $T = 128$ samples long chunks of sound sampled at 8 kHz, which corresponds to 16 ms time. The complete representation of 128 real basis functions was trained, and each of them was paired with its Hilbert transform, resulting in the total number of 256 basis vectors. Selected basis functions are displayed on figure 5.5 A. Real vectors are plotted in black together with associated imaginary ones plotted in gray. Panel B of the same figure displays isoprobability contours of Wigner-Ville distributions associated with the 256 basis functions. This form of representation localizes each temporal feature on a time-frequency plane [1] (one should note that real and imaginary vectors within each pair are represented by the same contour on that plot). A clear separation into two classes is visible. Low frequency (below 1 kHz) basis function are non-localized in time (they span the entire 16 ms interval), while in higher frequency region their temporal precision increases. An interesting bandwidth reversal is visible around 3 kHz, where temporal accuracy is traded against

frequency precision. Interestingly, a sharp separation into frequency and time localized basis functions, which emerged in this study was not clearly visible in other studies which performed sparse coding of sound [78, 1]. Time-frequency properties observed here reflect the statistical structure of the recorded auditory scene, which mostly consisted of non-harmonic environmental sounds sparsely interspersed with human speech.

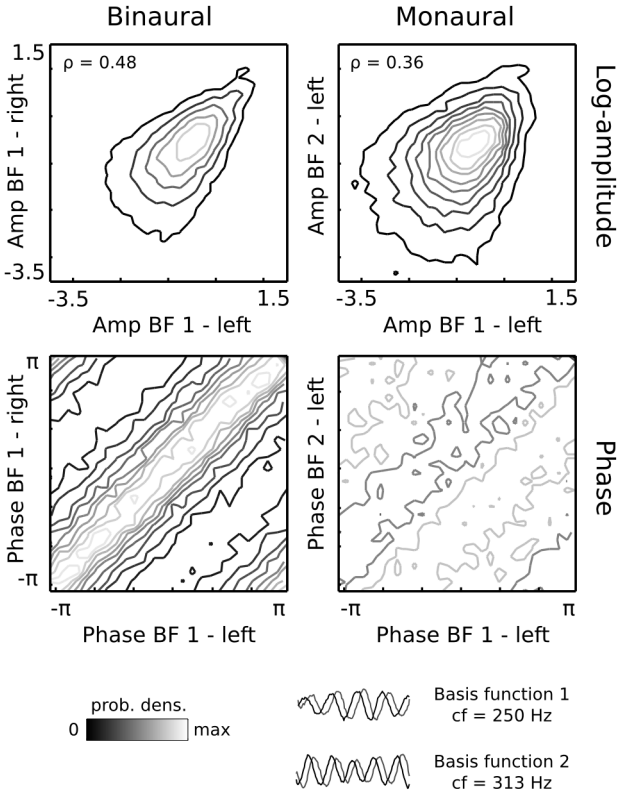

Figure 5.6: Intra and interaural pairwise distributions of phases and amplitudes. First row panels depict exemplary pairwise log-amplitude distributions. Sample phase distributions are depicted in the second row. Panels in the left column display joint coefficient distributions of the same basis function (BF1) in different ears. The right column depicts distributions of coefficients associated with two different basis functions (BF1 and BF2) within the left ear channel. Visible interaural dependencies are stronger than intraural ones.

Figure 5.6 depicts exemplary, pairwise distributions of first layer coefficients a_i and ϕ_i. Amplitudes (a_i) (first row) were transformed with a logarithm function. This transformation spreads positive values concentrated close to zero more broadly along the real line. Additionally, it has been demonstrated in a study of natural image statistics that logarithm linearizes correlations between sparse amplitudes [24]. This effect was also visible here. Amplitudes of the basis func-

tion 1 (plotted below the histograms on figure 5.6) in each ear (upper-left panel) revealed a pronounced linear dependency - Pearson's correlation was equal to 0.48. Correlation of two different basis functions (number 1 and 2) in the same ear (upper-right panel) was weaker ($\rho = 0.36$), but a strong linear relationship was still visible. The strong interaural correlation can be explained by the filtering properties of the head, which only weakly attenuates low frequencies. For this reason, interaural amplitude correlation decreased with increasing central frequency of the associated basis function.

An exemplary joint distribution of phases in the same ear is depicted on the lower-right panel of figure 5.6. As can be seen, intra-aural phase values are typically very weakly dependent. This is not the case for binaural phase relationship. A typical distribution of binaural phase is visible on the lower-left panel of figure 5.6. Phases of the same basis function in each ear reveal dependence in their difference. This means that the joint probability of monaural phases depends solely on the IPD:

$$p(\phi_{i,L}, \phi_{i,R}) \propto p(\Delta\phi_i) \tag{5.25}$$

where $\Delta\phi_i = \phi_{i,L} - \phi_{i,R}$ is the IPD. This property is a straightforward consequence of physics of sound - sounds arrive to each ear with a varying delay giving a rise to positive and negative phase shifts. From the point of view of statistics, this means that monaural phases become conditionally independent given their difference and a phase offset $\phi_{i,O}$:

$$\phi_{i,L} \perp \phi_{i,R} | \Delta\phi_i, \phi_{i,O} \tag{5.26}$$

The phase offset $\phi_{i,O}$ is the absolute phase value - indicating the time from the beginning of the oscillatory cycle. It therefore satisfies the following property:

$$\phi_{i,L} = \phi_{i,O} + \frac{\Delta\phi_i}{2} \tag{5.27}$$

$$\phi_{i,R} = \phi_{i,O} - \frac{\Delta\phi_i}{2} \tag{5.28}$$

This particular statistical property allows to understand IPDs not as an ad-hoc computed feature, but as an inherent property of a probability distribution underlying the data. It is reflected in the structure of the graphical model (see figure 5.1). Since the phase offset $\phi_{i,O}$ does not carry spatial information, for the purposes of current study it is treated as an auxiliary variable and therefore marked in gray.

5.3.2 Properties of the second layer representation

The second layer learned cooccuring phase and amplitude patterns forming a sparse, combinatorial code of the first layer output. Figure 5.7 displays 10 repre-

sentative examples of basis function pairs ξ_i and B_i, which encoded amplitudes and IPDs respectively.

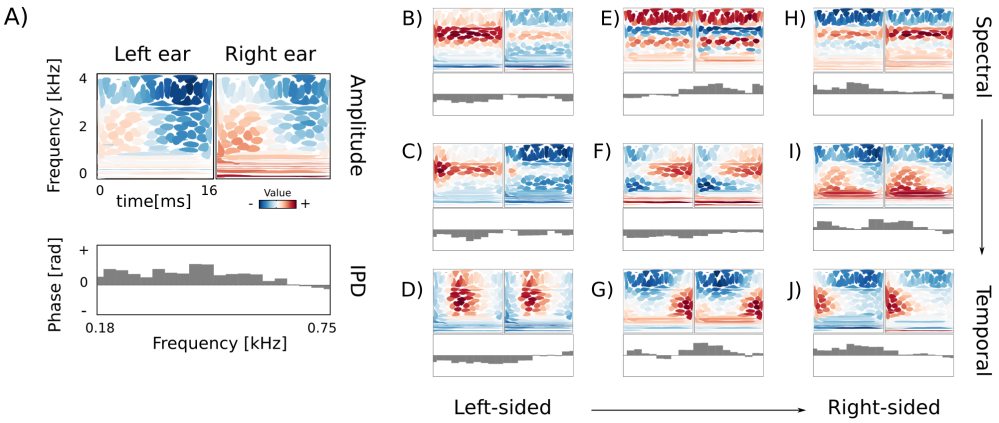

Figure 5.7: Higher-layer basis functions. A) Explanation of the visualization of second layer basis functions. Top two panels depict the binaural amplitude basis function B_i. Spectrotemporal information in each ear is represented using isoprobability contours of Wigner-Ville distributions of first-layer basis functions (see figure 5.5). Colors correspond to the log-amplitude weight. The bottom panel represents the IPD basis function ξ_i. Each gray bar represents one of 20 selected low-layer basis functions. Here almost all values are positive (the bars point upwards), which corresponds to the right-ear precedence. B)-J) Basis functions ordered vertically by spectral modulation and horizontally by the dominating side.

Each amplitude basis function consisted of two monaural parts corresponding to the left and the right ear. First-layer, temporal features were visualized using contours of Wigner-Ville distribution and colored according to the relative weight. Entries of IPD basis functions were values (marked by gray bars) modelling interaural phase disparities in each of selected 25 frequency channels.

The subset of 9 basis functions depicted on subpanels B-J constitutes a good representation of the entire dictionary. Their vertical ordering corresponds to spectrotemporal properties of B_i basis functions. Amplitude features displayed in the first row (B, E, H) reveal pronounced spectral modulation, while the last row (D, G, J) are features which are strongly temporaly modulated. Columns are ordered according to the ear each basis function pair preferred. Left column (B, C, D) are left-sided basis functions. Higher amplitude values are visible in the left ear parts (although differences are rather subtle), while associated IPD features are all negative. IPDs smaller than 0 imply, that the encoded waveform was delayed in the right ear, hence the sound source was closer to the left ear.

The last column (H, I, J) depicts more right-sided basis functions. Features displayed in the middle column (E, F, G) weight binaural amplitudes equally, however entries of associated phase vectors are either mostly negative or mostly positive.

As figure 5.7 shows, higher level representation learned spectrotemporal properties of the auditory scene, which was reflected in shapes of amplitude basis functions B_i. Binaural relations were captured by relative weighting of amplitudes in both ears and the shape of the IPD basis function.

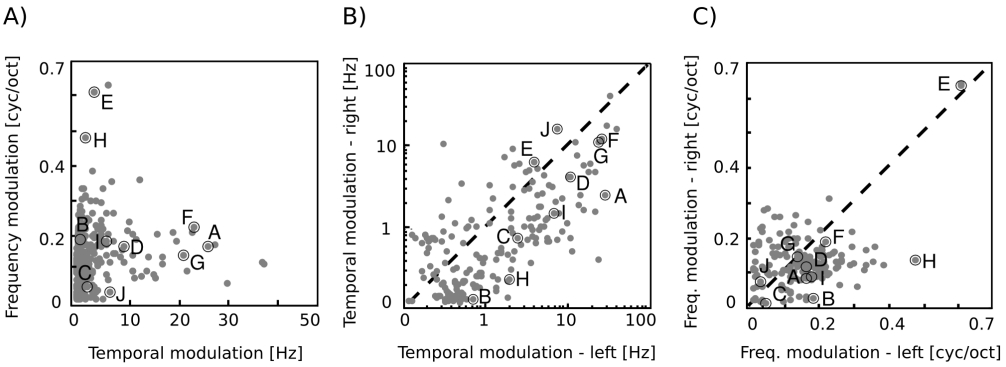

Figure 5.8: Spectrotemporal properties of the representation. A) Centers of mass of monaural modulation spectra. B) Centers of mass of temporal modulation in monaural parts of B_i basis functions plotted C) Centers of mass of spectral modulation in monaural parts of B_i basis functions plotted. Letters correspond to panels on figure 5.7.

To get a more detailed understanding of the spectrotemporal features captured by the representation, analysis of modulation spectra was performed. A modulation spectrum is a 2D Fourier transform of the spectrotemporal representation of a signal. It is known that modulation spectra of natural sounds posses specific structure [131]. Here, modulation spectrum was computed separately for monaural parts of amplitude basis functions B_i (see Methods). In the next step a center of mass of each of the modulation spectra was computed. Centers of mass are represented by single points on figure 5.8 A).

A clear tradeoff between spectral and temporal modulation was visible. Basis functions which were strongly temporally modulated revealed simultaneously weak temporal modulation (and vice versa). It was visible as a "triangular" shape of the point distribution on figure 5.8 A). This seems to be a robust property of natural sounds [131] and was shown to be captured by sparse coding models [25, 35]. Interestingly, spectrotemporal receptive fields of auditory neurons share this property [94, 58].

Average temporal modulation in the left ear is plotted against the right ear on panel B). Generally, a linear trend was present - temporal variation of monaural parts was correlated. The amplitude modulation of basis functions B varied between 0 and 40 Hz.

Spectral amplitude modulation revealed a different interaural dependency pattern. It was slightly negatively correlated, which is visible on figure5.8 C). If a left ear part was strongly modulated, the modulation in the right ear was weaker. This property can be explained by the head filtering characteristics. Head acts as a low-pass filter and attenuates higher frequencies. Therefore fine spectral information above 1.5 kHz was typically more pronounced in a single ear. This may be considered as an example of how stimulus statistics are determined not only by the environmental properties, but also by the anatomy of the organism. Majority of basis functions revealed the spectral modulation smaller than 0.4 cycle per octave, with only a single one exceeding this value.

In the following analysis step, the goal was to analyze how similar were monaural spectrotemporal patterns encoded by each second-layer unit. To this end binaural similarity index (BSI) of each amplitude basis functions [94] was computed. The BSI is a correlation coefficient between the left and the right parts of a binaural, spectrotemporal feature. If the BSI was close to 0, the corresponding unit was representing different spectrotemporal patterns in each ear, while values close to 1 implied their high similarity. BSIs are plotted on figure 5.9 A).

Clearly, overwhelming majority of basis functions revealed high interaural similarity ($BSI > 0.8$, see the histogram at the inset). BSI of only one basis function was slightly below 0. If information encoded by amplitude basis functions in each ear would be independent, the BSI distribution should peak at 0. This observation allows one to state that most of the second-layer units captured the same "cause" underlying the stimulus i.e. a binaurally redundant spectrotemporal pattern. While the BSI index measures similarity of encoded monaural sound features, it is not informative about the side-preference of each unit. To asess whether amplitude basis functions were biased more towards the left or towards the right ear, another statistic - a binaural amplitude dominance (BAD) was computed. The amplitude dominance was defined in the following way:

$$BAD(B_i) = \log\left(\frac{\|\exp(B_{i,L})\|}{\|\exp(B_{i,R})\|}\right) \tag{5.29}$$

where $B_{i,L} = B_{i,(1,\ldots,T)}, B_{i,R} = B_{i,(T+1,\ldots,2\times T)}$ are left and rigth ear parts of an amplitude basis function B_i. Each of them was pointwise exponentiated to map the entries from real log-amplitude values to the positive amplitude domain. The BAD index value larger than 0 means that the left-ear amplitude vector had a larger norm i.e. it dominated the input to the particular unit. Balanced units had a BAD value close to 0 while right-ear dominance was indicated by negative values. Two histograms of dominance scores are displayed on panel B) of figure

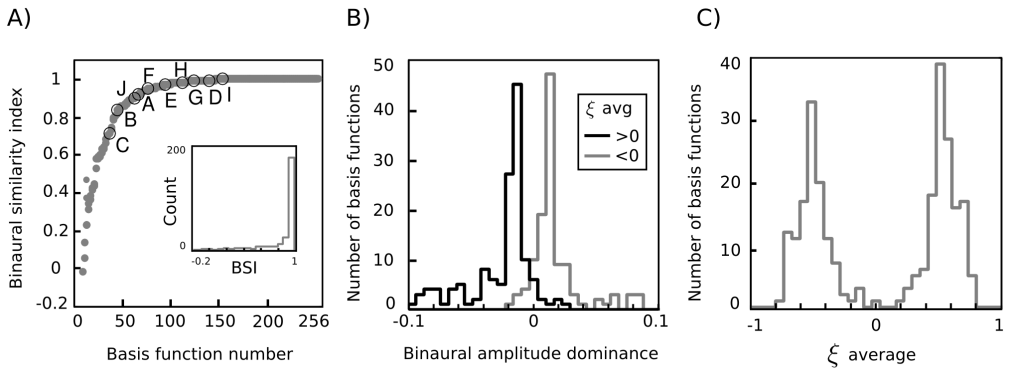

Figure 5.9: Binaural properties of the representation. A) Binaural Similarity Index of amplitude basis functions B_i. The BSI is a correlation coefficient between left and right ear subvectors. The inset depicte the BSI histogram. B) Distribution of binaural amplitude dominance. Values above 0 imply domination of the left, and below 0 of the right ear. Histograms of BAD values of amplitude basis functions associated with negative IPD basis functions are colored gray and those associated with positive ξ_i values are colored black. C) Distribution of averages of normalized ξ_i basis functions.

5.9. The black one is an empirical distribution of BAD values of amplitude basis functions associated with IPD features of a negative average value (left-side preferring). The gray one in turn, corresponds to amplitude features matched with right-side biased phase basis functions. Both distributions are roughly symmetric with their modes located quite close to 0. Such bimodal distribution of the amplitude dominance score implies that amplitude basis functions could be divided into two opposite populations - each preferring input from a different ear. Moreover, amplitude and phase information modelled by basis functions B_i and ξ_i was dependent - amplitude features dominated by information from one ear were associated with IPD features biased towards the same ear.

While amplitude representation encoded the quality of the sound together with binaural differences, the IPD dictionary was representing solely spatial aspects of the stimulus i.e. the temporal difference between the ears. In almost all cases, single entries of each of the phase difference basis functions ξ_i had all the same sign. Negative phase differences corresponded to the left-side bias (it meant that the soundwave arrived first to the left-ear generating a smaller phase value) and positive to the right-side one. These two properties allowed to asess the spatial preference of IPD basis functions simply by computing the average of their entries. The histogram of averages of vectors ξ_i (normalized to have the maximal absolute value of 1) is depicted on figure 5.9 C). A clear bimodality is visible

in the distribution. The positive peak corresponds to right-sided basis functions and the negative one to the left-sided subpopulation. Almost no balanced features (close to 0) were present in the dictionary. This dichotomy is visible also in figure 5.7 - binaurally balanced amplitude basis functions (middle column) were associated with phase vectors biased towards either side. This result may be related to a previous study, which have shown that a representation of natural IPD distribution designed to maximize stimulus discriminability (Fisher information) has also a form of two distinct channels [52]. Each of the channels preferred IPDs of an opposite sign.

5.3.3 Broad spatial tuning of high-level units

The second layer of the model learned a distributed representation of sound features accesible to neurons in the auditory cortex. Assuming that the cortical auditory code indeed develops driven by principles of efficiency and sparsity, one can interpret second layer basis functions as neuronal receptive fields and sparse coefficients s as a measure of neuronal activity (e.g. firing rates). The model can be then probed using spatial auditory stimuli. If it indeed provides an approximation to real neuronal computations, its responses should be comparable with spatial tuning properties of the auditory cortex.

In order to verify whether this was true, a test recording was performed. As a test sound - the hiss of two pieces of paper rubbed against each other was used. It was a broadband signal, reminiscent of white noise used in physiological experiments, yet possessing a natural structure. A recording was performed in an echo-free chamber, where a person walked around the recording subject while rubbing two pieces of paper. The recording was divided into 18 windows, each corresponding to a 20 degree part of a full circle. The number of windows was selected to match experimental parameters in [137, 139]. From each window 3000 sound epochs were drawn and each of them was encoded using the model. Computing histograms of coefficients s at each angular position θ, provided an estimate of conditional distributions $p(s_i|\theta)$. Panel A) on figure 5.10 displays a conditional histogram of coefficient s corresponding to the basis function pair depicted on figure 5.7 A).

Distributions of sparse coefficients revealed a strong dependence on the position of the sound source. As visible on the figure, the conditional mean of the distribution $p(s_i|\theta)$ traced by the red line varied in a pronounced way across all positions. Since mean was the only moment, which revealed such strong dependence, and by analogy to averaged firing rates of neurons measured in physiological studies, average responses at each position were further studied to understand spatial sensitivity of basis functions. Mean vectors $\mu_{i,\theta}$ were constructed for each second-layer unit by taking its average response at the sound source position θ. Each mean vector was shifted and scaled such that its minimum value was equal

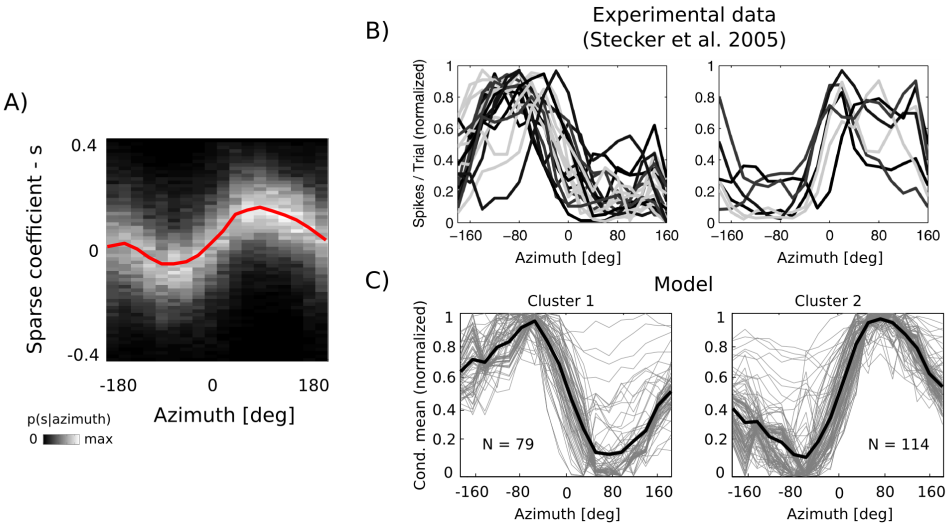

Figure 5.10: Spatial tuning curves of second-layer units. A) A conditional distribution of the coefficient s_i corresponding to basis functions B_i, ξ_i depicted on figure 5.7 A. The red line depicts the average value conditioned on sound position. B) Experimentally measured spatial tuning curves measured in the A1 area of a cat. The left panel depicts contra- and the right panel ipsi- laterally tuned units. Figure modified from [137] C) All position-modulated tuning curves belonging to each of the two clusters. Thin gray lines are single tuning curves, while thick black lines depict cluster averages.

to 0 and the maximal one to 1. Such transformation allowed for comparison to experimentally measured spatial tuning curves of auditory neurons, and for this reason scaled vectors μ_i will be referred to as model tuning curves in the remainder of the thesis. In order to identify spatial tuning preferences, the population of model tuning curves was grouped into two clusters using k-means algorithm. Obtained clusters consisted of 118 and 138 similar vectors. Tuning curves belonging to both clusters and revealing a strong correlation ($|\rho| > 0.75$). with the sound position are plotted on figure 5.10 C) as gray lines. Cluster centroids (averages of all tuning curves belonging to a cluster) are plotted in black. Second layer units were tuned broadly - most of them were modulated by sound located at all positions surrounding the subject's head. A clear spatial preference is visible - members of cluster 1 were most highly activated (on average) by sounds localized close to the left ear ($\theta \approx -90°$), while cluster 2 consisted of units tuned to the right ear ($\theta \approx 90°$). Very similar tuning properties of auditory neurons were identified in the cat's auditory cortex [137]. Data from this study is plotted for comparison in the subfigure B) of figure 5.10. Neuronal recordings were performed in the right hemisphere and two panels depict two subpopulations of

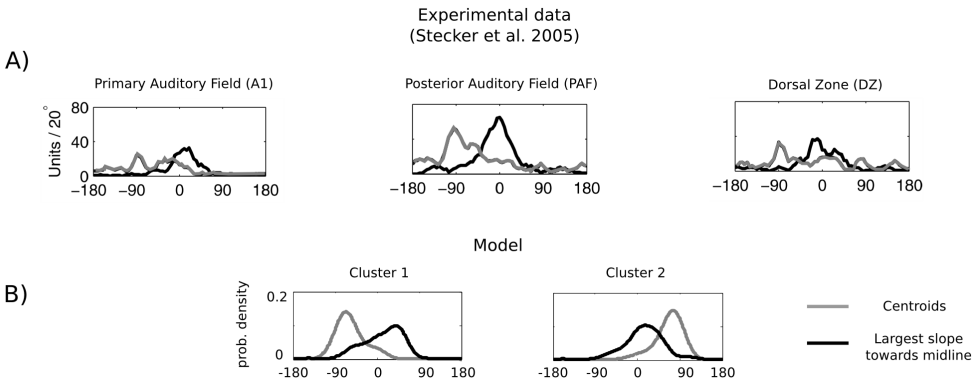

Figure 5.11: Distribution of tuning curve centroids and maximal slope positions in the model and experimental data. A) Histograms of positions of tuning curve centroids (gray) and maximal slopes towards the midline (black) measured experimentally in the auditory cortical areas from the cat. Figure modified from [137]. B) Distribution of the same features computed for model tuning curves belonging to each cluster.

neurons. The larger contra- and the smaller ipsi-lateral one. It is important to note, that the notion of ipsi, and contra laterality is not meaningful in the proposed model, therefore one should compare shapes of the model and experimental tuning curves, not the numerosity of units in each population or cluster.

Two major features of cortical auditory neurons responsive to sound position were observed experimentally: (i) tuning curve peaks were localized mostly at extremely lateral positions (opposite to each ear) and (ii) slopes of tuning curves were steepest close to the auditory midline. Both properties are visible in model tuning curves on figure 5.10. However, in order to perform a more direct comparison between the model and experimental data, analysis analogous to the one described in [137] was performed. Firstly tuning curve centroids were computed. A centroid was defined as an average position, where the unit activation was equal to 0.75 or larger (see Methods). In the following step, position of a maximal slope towards midline was identified for each unit. This means that for units tuned to the left hemifield (cluster 1) the position of the minimal slope value was taken, while the position of the maximal one was taken for units tuned to the right hemifield (cluster 2). In this way, a position of maximal sensitivity to changes of the sound location were identified. Distributions of model centroids and maximal slope positions are depicted on figure 5.11 B). Centroids were distributed close to lateral positions, opposite in each cluster (-90° cluster 1, $+90^\circ$ cluster 2). Distribution peaks were located at positions close to each ear. No uniform

tiling of the space by centroid values was present. At the same time, maximal slope values were tightly packed around the midline - peaks of their distributions were located precisely at, or very close to 0 degrees. This means that while the maximal response was on average triggered by lateral stimuli, the largest changes were triggered by sounds located close to the midline. Both properties were in good agreement with the experimental data reported in [137]. Figure 5.11 A) depicts on three panels centroid and slopes distributions measured in three different regions of cat's auditory cortex - Primary Auditory Field (A1), Posterior Auditory Field (PAF) and Dorsal Zone (DZ). A close resemblance between the model and physiological data was present.

5.3.4 Population coding of sound source position

It has been argued that while single neurons in the auditory cortex provide coarse spatial information, their populations form a distributed code for sound localization [139, 93, 93, 137]. Here, a decoding analysis was performed to verify whether similar statement can be made about the proposed model.

A gaussian mixture model (GMM) was utilized as a decoder. The GMM modelled the marginal distribution of sparse coefficients as a linear combination of 18 gaussian components, each corresponding to a particular position of a sound source (i.e. the θ value). In the first part of the decoding analysis, single coefficients were used to identify the sound position. The GMM was fitted using the training dataset consisting of coefficient values s_i and associated position labels θ. In the testing stage, position estimates $\hat{\theta}$ were estimated (decoded) using unlabeled coefficients from the test dataset. For each of the coefficients, a confusion matrix was computed. A confusion matrix is a two-dimensional histogram of θ and $\hat{\theta}$ and can be understood as an estimate of the joint probability distribution of these two variables. Using a confusion matrix, an estimate of mutual information i.e. the number of bits shared between the position estimate $\hat{\theta}$ and its actual value θ was obtained. Figure 5.12 B depicts histograms of information carried by each coefficient s_i about the sound source position, estimated as described above. A general observation is that single coefficients carried a very small amount of information about the sound location. The histogram peaks at a value close to 0.1 bits. Only few units coded approximately 1 bit of positional information. Even 1 bit, however, suffices merely to identify a hemifield, not mentioning the precise sound position. As can be predicted from the broad shapes of the tuning curves, single second-layer units carried a little amount of spatial information. A similar result was obtained for neurons in different areas of the cats auditory cortex [92]. Figure 5.12 A) depicts histograms of the information amount about the sound position encoded by spike count of neurons in A1 and PAF regions (figure reproduced from [138]). Spike count (which essentially corresponds to a firing rate) is a feature of a neuronal response most directly corresponding to coefficients s in

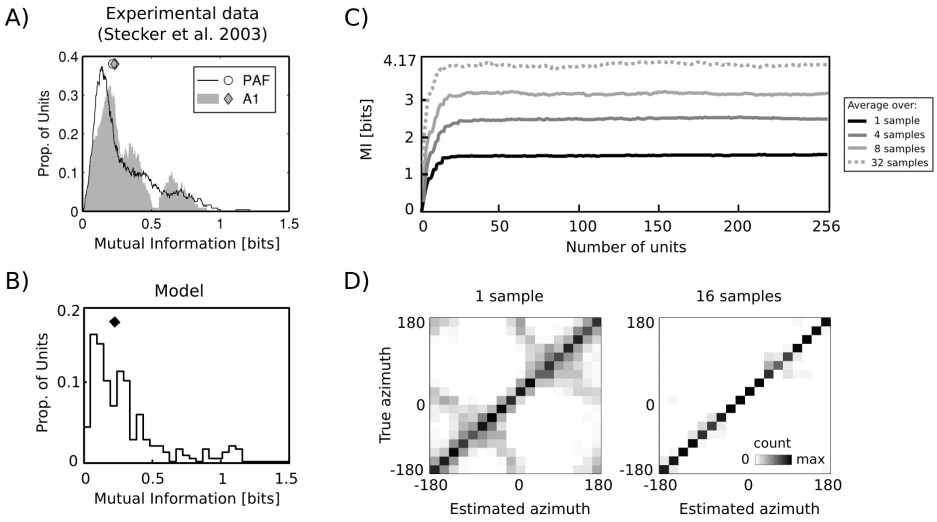

Figure 5.12: Population decoding analysis A) A histogram of mutual information carried by firing rates of single neurons about the position of a sound source estimated from confusion matrix. Figure reproduced from [138] B) Histograms of position-specific information carried by second layer sparse coefficients s. The diamond symbols in panels A and B mark distribution medians. C) Mutual information plotted as a function of the number of units used to decode the position. Colors of lines correspond to data averaged over different number of samples. The scale ends at 4.17 bits, which is the amount of information required to perform errorless decoding ($\log_2(18) = 4.17$) D) Exemplary population confusion matrices for 1 and 16 samples.

the model described here. Maximal peaks of all histograms were close to 0.1 bits, followed by a long, decaying tail. Medians of mutual information distributions (marked by diamond symbols on panels A and B of figure 5.12) estimated from neuronal data and the sparse coefficients aligned well, close to 0.2 bits. Overall, a strong similarity between physiological measurements and the behavior of the model was visible.

While single neurons do not carry much spatial information, the joint population activity was sufficient to decode the sound position [137, 93, 139, 138]. Therefore in the second step of the decoding analysis, multiple coefficients s were used to train and test the GMM decoder. Results of the population decoding are plotted on figure 5.12 C). The decoder was trained with a progressively larger number of second-layer units (from 1 to 256) and the mutual information was estimated from obtained confusion matrices. Each line on the plot depicts the number of bits as a function of the number of units used to perform decoding. Line colors correspond to the number of samples over which the average activ-

ity was computed. Broadly speaking, larger populations of second-layer units allowed for a more precise position decoding. As in the case of single units, averages over larger amounts of samples were also more informative - population activity averaged over 32 samples saturated amount of bits required to perform errorless decoding (4.17). Two confusion matrices obtained from raw population activity and an average over 16 samples are displayed on subfigure 5.12 D). In the former case, the decoder was misclassifying mostly sound positions within each hemifield. Averaging over 16 sound samples yielded an almost diagonal (errorless) confusion matrix. The decoding analysis allowed to draw the conclusion that while single units carried very little spatial information, their population encoded source location accurately, consistently with experimental data.

5.3.5 Interdependent encoding of sound position and identity

Second layer units achieved spatial tuning by assigning different weights to amplitudes in each ear, and to IPD values in different frequency channels. At the same time they encoded spectrotemporal features of sound, as depicted on figure 5.7. Their activity should therefore be modulated by both - sound position as well as its quality. Such comodulation is a prominent feature of the majority of cortical auditory neurons [15, 14]. In order to verify whether this was true, model spatial tuning curves were estimated with a second sound source, very different from a hiss created by rubbing paper - human speech. Frequency spectra of both test stimuli are depicted on figure 5.13 D).

Test sounds distributed their energy over non-overlapping parts of the frequency spectrum. While speech consisted mostly of harmonic peaks below 1.5 kHz, the paper sound was much more broadband and its energy was uniformely distributed between 1.5 and 4 kHz. Panels A)-C) of figure 5.13 depict three amplitude/IPD basis function pairs together with their spatial tuning curves estimated using different sounds. The spatial preference of depicted units (left or right hemifield) was predictable from their binaural composition. Each of them, however, was activated stronger by a stimulus, which spectrum matched better amplitude basis functions. Basis functions visible on panels A) and C) had a lot of energy accumulated in higher frequencies, therefore the paper sound activated them stronger (on average). Basis function B) was spectrally better corresponding to speech sounds, therefore speech was a preferred class of stimuli. This observation means that tuning curves i.e. position-conditional means $\mu_{i,\theta}$ should be understood not as averages of coefficient ensembles conditioned only on the sound position θ but also on spectral properties of the sound. The second-layer representation encoded two aspects of the auditory stimulus - position and identity interdependently.

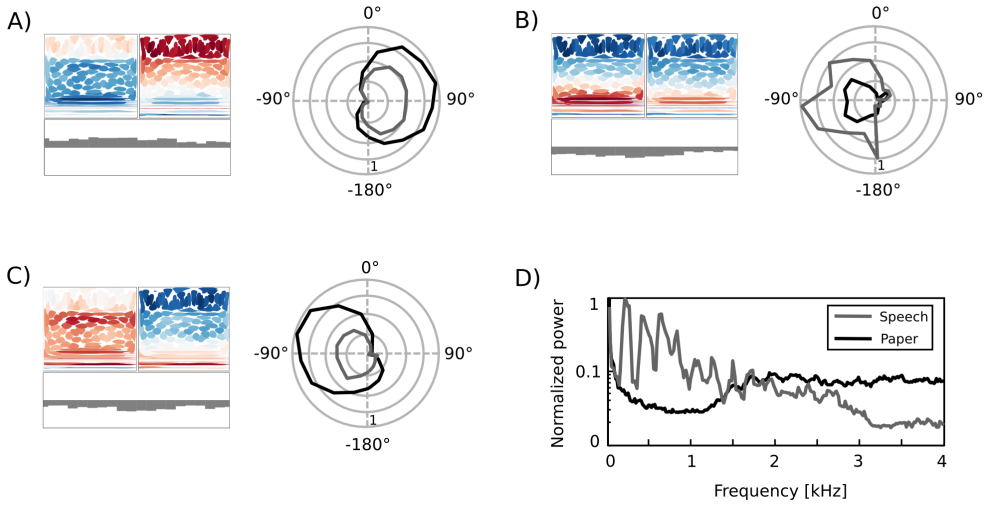

Figure 5.13: Comodulation of unit responses by sound position and identity. A)-C) Three exemplary second layer basis functions plotted with spatial tuning curves obtained using two different sounds - female speech (gray) and paper noise (black). D) Frequency spectra of both test sounds.

5.4 Discussion

Previously proposed statistical models of natural acoustic stimuli focused predominantly on monaural sounds [78, 133, 25, 1, 140, 3]. Studies modelling binaural stimuli constrained to a limited representation - either IPDs [52] or spectrograms [97]. In contrast, the assumption behind the present work was that spatial sensitivity of cortical neurons is formed by fusing different cues. Therefore, in order to understand the role played by the auditory cortex in spatial hearing, the entire natural input processed by the auditory system was analyzed.

To this end, a novel probabilistic model of natural stereo sounds has been proposed. The model is based on principles of sparse, efficient coding - its task was to learn progressively less redundant representations of natural signal. It consisted of two hidden layers, each of them could be interpreted as an analogy to different stages of sound processing in the nervous system. The purpose of the first layer was to form a sparse, non-redundant representation of natural sound in each ear. By analogy to the cochlea, the encoding was supposed to extract and separate temporal information i.e. phase from the amplitude of the signal. In order to do so, a dictionary of complex-valued basis functions was adapted to short sound epochs. On top of the first model layer, which encoded sound in each ear independently, the second layer was trained. Its goal was to encode jointly amplitude and phase - two kinds of information crucial for sound

localization, which may be fused together in higher stages of the auditory system. The higher-order representation captured spectrotemporal composition of the signal, by learning amplitude patterns of the first layer output as well as interaural disparities present in form of interaural phase and amplitude differences. It is important to stress that the model was learned in a fully unsupervised setting - at no point information about positions of sounds sources or the spatial configuration of the environment was accessible. Yet, when tested with a set of spatial sounds, activity of second layer units revealed strong dependence on sound position. Tuning curves describing relation between the sound position and model activity were in good correspondence with experimentally measured spatial tuning properties of cortical auditory neurons.

The data used for comparisons originated from studies of cat's auditory cortex ([137, 138]). Since statistics of the binaural signal are affected by the geometry of ears and the head of the organisms, one could argue that model trained on binaural recordings performed by a human, should not be compared with cats physiology. As long as detailed features of neuronal tuning to a sound position may vary across those species, tuning patterns highly similar to those of a cat have been observed in the auditory cortex of primates [155, 95]. Overall, the cortical representation of sound position seems to be highly similar across mammals [49].

5.4.1 A sparse representation of natural binaural sounds forms a panoramic population code for sound location

In mammals, the location of a sound is encoded by two populations of broadly tuned, spatially non-specific units [49]. This finding stood against initial expectations of finding a "labelled-line code" i.e. a topographic map of neurons narrowly tuned to small areas of space. The "spatiotopic map" was expected to be observed by analogy to the tonotopic structure of the cortex as well as high localisation accuracy of humans and animals. Instead, it has been found that auditory cortical neurons within each hemisphere are predominantly tuned to far, contralateral positions. Peaks of observed tuning curves did not tile the auditory space uniformly, rather they were clustered around the two lateral positions. A prominent observed feature of cortical representation of sound location were slopes of the tuning curves. Regardless of the position of the tuning curve peak, slopes were steepest close to the interaural midline - the area where behavioral localisation acuity is highest [49]. From described observations, two prominent conclusions were drawn. Firstly, that the slope of tuning curves, not the distribution of their peaks determines spatial acuity [137, 49, 21, 45]. Secondly that sound position is encoded by distributed patterns of population activity, not single neurons [93, 137, 139]. It has been argued that these properties are a manifestation of a coding mechanism which evolved to specifically meet the demand of binaural hearing tasks [137, 49]. Here it is shown that crucial properties of

cortical spatial tuning emerge in an unsupervised learning model, which learns a sparse representation of natural binaural sounds. The objective of the model was to code the stimulus efficiently (i.e. with a minimal redundancy within limits of applied transformations), while minimizing unit activity. Properties of the learned representation are therefore a reflection of stimulus statistics, not of any task-specific coding strategy (required for instance to localize sounds with the highest accuracy at the midline).

The position of the sound-generating object is a latent variable for the auditory system. It means that its value is not explicitly present in the raw stimulus - it has to be estimated. This estimation, (or inference) is a complex and non-trivial task in the real acoustic environment, where sounds reaching ear membranes are a reflection of intricate auditory scenes. Sensory neurons perform transformations of those sound waveforms in an attempt to reconstruct the spatial configuration of the scene. Therefore, in an attempt to understand cortical representation of space, it may be helpful to think what is the statistical structure of the naturally encountered binaural stimulus that the auditory system operates on. Sounds reaching ear membranes consist information about their generating sources, spatial configuration of the scene, position motion of the organism and the geometry of its head and outer ears.

Results obtained here, suggest that shapes of the model spatial tuning curves constitute a reflection of regularities imposed on the sensory data by the filtering properties of the head. At lateral positions (directly next to the left or the right ear) there is no acoustic attenuation by the skull, hence sounds are loudest and least delayed. This in turn, elicits strongest response in units preferring that particular side. When the sound is at a contralateral position, response is much weaker, due to the maximal head attenuation and largest delay. The curve connecting those two extrema is steepest in the transition area - at the midline. Since the auditory environment was uniformly sampled at both sides of the head, model units were clustered into two roughly equal subpopulations, basing on the shapes of their tuning curves. Clusters were symmetric with respect to each other - one tuned to to the left and the other to the right hemifield. This groupping is reminiscent of the "opponent-channel" representation of the auditory space, which has been postulated before [137, 49]. Present results provide a theoretical interpretation of this tuning pattern. They suggest that neuronal population which forms a sparse, efficient representation of natural stimuli would reveal two broadly tuned channels, when probed with sounds located at different position.

5.4.2 Interdependent coding of spatial information and other features of the sound

There is an ongoing debate about presence (or lack of thereof) of two-separate "what" and "where" streams in the auditory cortex [100]. The streams would

separate spatial information from other sound features, which determine its identity. An important prediction formed by this dual-stream hypothesis is that there should exist neurons selective to sound position and invariant to other aspects in the auditory cortex. While some evidence has been found supporting this notion [120, 84] it seems that at least in vast parts of the auditory cortex neural activity can be modulated by multiple features of sound such as pitch, timbre and location [15]. Neurons are sensitive to sound position (i.e. changing position affects their firing patterns), but not selective nor invariant to it. The majority of studies analyzing spatial sensitivity in the auditory cortex uses a single class of sound and the source position is the only varying parameter. Therefore, despite initial efforts, the influence jointly exerted by sound quality and position on neuronal activity is not yet well understood.

The statistical model proposed here suggests that no dissociation of spatial and non-spatial information is necessary to either reconstruct the sound source or identify its position. The learned second-layer representation carries both kinds of information - about the sound quality (contained in the spectrotemporal structure of basis functions) and about spatial aspects (contained in the binaural amplitude weighting and IPD vectors). The learned code forms a "what is where" representation of the stimulus i.e. those two aspects are represented interdependently. A manifestation of this fact is visible in different scaling of spatial tuning curves, when probed with two different sound sources. Such comodulation of neuronal activity by sound position and quality has been observed experimentally [15], which may suggest that recorded neurons form a sparse, efficient representation of binaural sound. An advantage of an interdependent "what is where" representation is the absence of the "feature binding problem", which has to be solved if spatial information is processed independently. After separating location of a source from its identity in the auditory cortex, they would have to be fused at higher processing stages. A code similar to the one described here does not create such a problem.

5.5 Conclusion

Results presented in this chapter are strongly related to understanding function of sensory representations in the natural environment.

Firstly, they suggest that the spatial tuning of cortical auditory neurons is a result of an adaptation to natural stimuli. From this point of view modulation of spiking activity by changing the stimulus position is not a manifestation of a computation, which is specifically designed to extract spatial information. It is rather a reflection of a change in the structure of incoming stimulus. The function of neurons in A1 and surrounding areas may be therefore to form an efficient representation of incoming stimuli, rather than making any physical properties

of the environment (such as a source location) explicit.

Secondly, activity of model units as well as of cortical auditory neurons is largely non-specific to high-level properties of sound. Despite this fact, information they carry still allows to compute the position of the source and reconstruct its spectrotemporal structure. This constitutes a strong suggestion that sensory representations do not have to encode a single stimulus parameter exclusively, and be invariant to all other aspects. It is often expected that within the auditory systems different, non-overlapping neuronal populations exclusively encode properties such as pitch, timbre and location. As demonstrated here, this must not be the case. The function of sensory representations may be to encode stimulus structure as such, without separating aspects pre-defined by the human observer.

Chapter 6

Efficient Coding Can Lead to Formation of Auditory Invariances

6.1 Overview

The previous chapter has focused mostly on the *postdictive* approach i.e. it attempted to explain known properties of the auditory system as a form of adaptation to natural stimulus statistics. Contents of this chapter are of a more *predictive* nature. Here I attempt to verify, whether applying the principles of efficient coding can lead to formation of long-postulated invariant auditory representations (the "what" and "where" pathways). Until now no conclusive experimental evidence in support of such separation has been delivered.

As originally proposed by [10], the efficient coding hypothesis suggests that sensory systems adapt to the statistical structure of the natural environment in order to maximize the amount of conveyed information. However, having a sole representation of the stimulus is not enough for the organism to interact with the environment. In order to perform actions, the nervous system has to extract relevant information from the raw sensory data and then segregate it according to its functional meaning, determined by the task at hand. For example the auditory system must extract position invariant information regardless of sound quality, separating "what" and "where" information. In a more recent paper [9] Barlow proposed that behaviorally relevant stimulus features (i.e. ones supporting informed decisions) may be learned by redundancy reduction. In other words, functional segregation of neurons can be achieved by efficient coding of sensory inputs. The evidence in support of this notion is still sparse.

Among different sensory mechanisms, spatial hearing provides a good example for the extraction and separation of behaviorally vital information from the

sensory signal. Even though temporal differences on the order of microseconds are of a substantial importance for sound localization, binaural neurons in the higher areas of the auditory pathway can be characterized with Spectrotemporal Receptive Fields (STRFS), which have much more coarse temporal resolution (ms) [44]. Despite such loss of temporal accuracy, many of those neurons reveal sharp spatial selectivity [125] encoding the position of the sound source in space. What is the neural computation underlying this process remains an open question.

In this chapter I use spatial hearing as an example of a sensory task, to show how information of different meaning ("what" and "where") can be clearly separated. The work described here provides computational evidence that redundancy reduction can lead to the separation of spatial information from the representation of the sound spectrogram. This means that formation of the neural auditory space representation can be achieved without the need of any task-specific computations but solely by applying the general principle of redundancy reduction. It is demonstrated that Independent Component Analysis (ICA) - a linear efficient coding transform trained on a dataset of spectrograms of simulated as well as natural binaural speech sounds, extracts sound position invariant features separating them from the representation of the sound position itself. Learned structures can be understood as model spatial and spectrotemporal receptive fields of auditory neurons which encode different kinds of behaviorally relevant information.

6.2 Methods

High order statistics of natural auditory signal were studied by performing Independent Component Analysis (ICA) on a time-frequency representation of binaural sounds.

As a proxy for natural sounds, speech was used in the present study. Speech comprises a rich variety of acoustic structures and has been successfully used to learn statistical models predicting properties of the auditory system [133, 25, 69]. Additionally, it has been suggested that speech may have evolved to match existing neural representations, which are optimizing information transmission of environmental sounds [133].

Spatial sounds were obtained in two ways. Firstly, the efficient coding algorithm was trained using simulated naturalistic binaural sounds. Simulation gave the advantage of labelling each sound with its spatial position. Secondly a natural auditory scene was recorded with binaural microphones. The signal obtained in this way was less controlled, however it contained more complex and fully natural spatial information. Training datasets were obtained by drawing 70000 random intervals 216 ms long from each dataset separately. The data generation process together with its interpretation is displayed on figure 6.1.

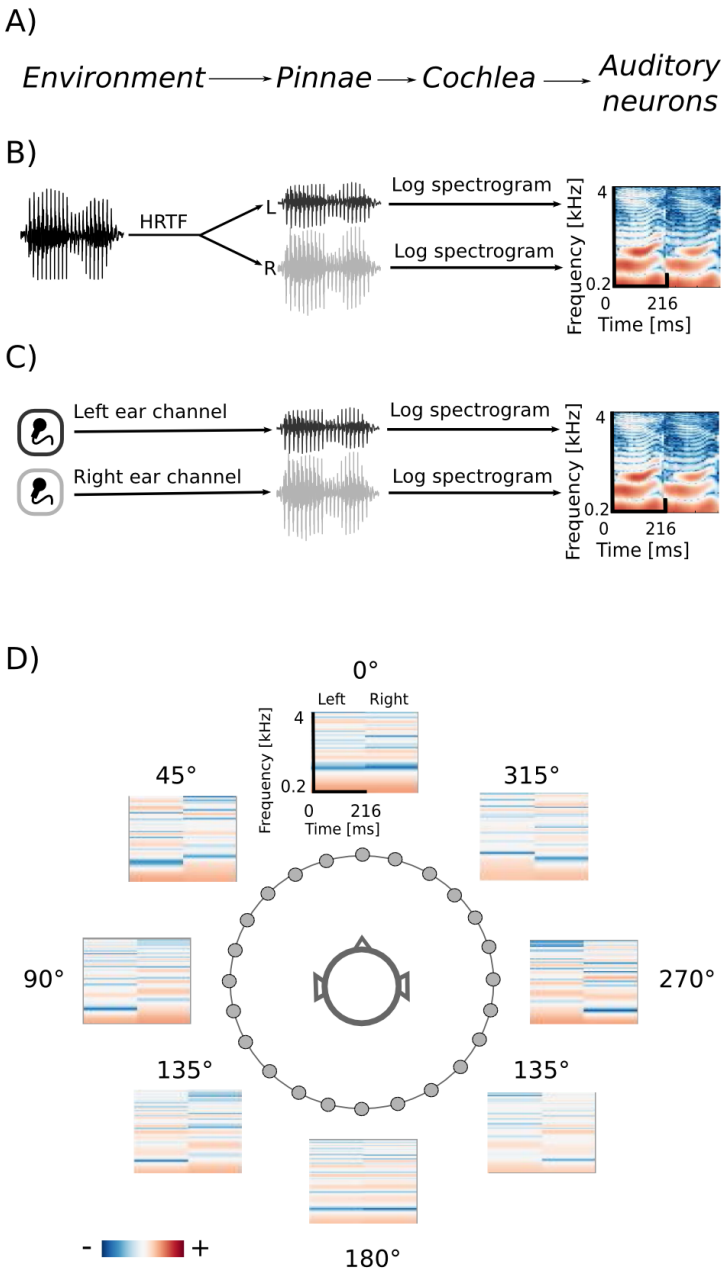

Figure 6.1: Data generation process. (A) Interpretation of consecutive stages of data generation. The acoustic environment is either simulated (B) or recorded with binaural microphones (C). Further stages of the processing include frequency decomposition and transformation with a logarithmic nonlinearity, which emulates cochlear filtering (D) Positions of HRTFs around the head are marked with circles

6.2.1 Simulated sounds

As a corpus of natural sounds, data from the International Phonetic Association Handbook [4] were used. The database contains speech sounds of a narrative told by male and female speakers in 29 languages. All sounds were downsampled to 16000 Hz from their original sampling rate and bandpass filtered between 200 and 6000 Hz. The training dataset was created by drawing random intervals of 216 ms from the speech corpus data. Spatial sounds were simulated by convolving sampled speech chunks with human Head Related Transfer Functions (HRTFs). HRTF fully describe the sound distortion due to the filtering by the pinnae and therefore contain entire spatial information available to the organism. Given an angular sound source position θ, HRTF is defined by a pair of linear filters:

$$HRTF(\theta) = \{h_{L,\theta}(t), h_{R,\theta}(t)\} \tag{6.1}$$

where L, R subscrpits denote left and right ear respectively, and t denotes time sample. One should note that in the temporal domain, HRTFs are often called Head Related Impulse Response (HRIR). A set of HRTFs was taken from the LISTEN database [151]. The database contains human HRTFs recorded for 187 positions in the three-dimensional space surrounding the subject's head. HRTFs from a single random subject were selected and further limited to positions lying on the azimuthal plane with 15 degree spacing (24 positions in total). Monaural stimulus vectors $x_E(t)$ ($E \in \{R, L\}$ denotes the ear) were created by drawing random chunks $g(t)$ of speech sounds and convolving them with $HRTF(\theta)$ corresponding to an azimuthal position θ, which was also randomly drawn:

$$x_E(t) = (g * h_E)(t) = \int_{-\infty}^{\infty} h_E(\tau) g(t - \tau) d\tau \tag{6.2}$$

where $*$ denotes the convolution operator. In this data, spatial and identity information constitute independent factors.

6.2.2 Natural sounds

In order to obtain a dataset of natural binaural sounds a complex auditory scene was recorded using binaural microphones. The recording consisted of three people (two males and one female) engaged in a conversation while moving freely in an echo-free chamber. Such an environment without reflections and echoes reduced the number of factors modifying sound waveforms. One of the male speakers was recording the audio signal with Soundman OKM-II binaural microphones placed in the ear channels. In total 20 minutes were recorded and included moving and stationary, often overlapping sound sources. To test the spatial sensitivity of learned features a recording with a single male speaker was performed. He walked around the head of the recording subject with a constant speed following a circular trajectory while reading a book out loud, twice in the clockwise and twice anti-clockwise direction. The length of the testing dataset was $54s$.

6.2.3 Simulated cochlear preprocessing

Before reaching the auditory cortex, where spatial receptive fields (SRFs) were observed [125], sound waveforms undergo a substantial processing. Since the modelling focus of the present study was beyond the auditory periphery, the data were preprocessed to roughly emulate the cochlear filtering (see the scheme on fig 6.1).

Short Time Fourier Transform (STFT) was performed on each sound interval included in the training dataset. Each chunk was divided into 25 overlapping windows each 16 ms long. STFT spanned 256 frequency channels logarithmically spaced between 200 and 4000 Hz (decomposition into arbitrary, non-linearly spaced frequency channels was computed using the Goertzel algorithm). Logarithmic frequency spacing was observed in the mammalian cochlea and seems to be a robust property across species [47, 134]. The spectral power of the resulting spectrograms was transformed with a logarithmic function which emulates the cochlear compressive nonlinearity [117].

Stimuli were 216 ms long in order to match the temporal extent of cortical neurons' STRFs, which were characterized by spatial receptive fields [125]. Besides emulating the cochlear transformation of the air pressure waveform, such spectrograms were reminiscent of the sound representation most effective in mapping spectrotemporal receptive fields in the songbird midbrain [44]. A very similar representation was used in a recent sparse coding study [25].

Spectrograms of left and right ears were concatenated. Such data representation attempts to simulate the input to higher binaural neurons, which operate on spectrotemporal information, simultaneously fed from monaural channels [125, 112, 94]. In principle, we could first train ICA on monaural spectrograms and then model their codependencies. In such way, however, the algorithm could not explicitly model binaural correlations. Additionally, this would require application of a hierarchical model, which lies outside of the scope of this study. Our approach resembles ICA studies, which focused on modelling of visual binocular receptive fields [56, 60]. There, the input to binocular neurons in the visual cortex was modelled by concatenating image patches from the left and the right eye.

The efficient coding algorithm was run on the resulting time-frequency representation of the binaural waveforms. After preprocessing the dimensionality of data vectors was equal to $2 \times (25 \times 256) = 12800$. Both training datasets: simulated and natural one consisted of 70000 samples. Prior to the ICA learning, the data dimensionality was reduced with Principal Component Analysis (PCA) to 324 dimensions, preserving more than 99% of total variance in both cases. Due to memory issues (allocation of a very large covariance matrix) a probabilistic PCA implementation was used [121].

6.2.4 Independent component analysis of spectrograms

To learn a non-redundant representation of binaural spectrograms the Independent Component Analysis was performed on preprocessed sound spectrograms. Using notation similar to chapter 2, each binaural spectrogram frame $x_t \in \mathbb{R}^N$ was modelled as a linear combination of N basis functions $a \in \mathbb{R}^N$:

$$\hat{x}_{t,i} = \sum_{n=1}^{N} s_{t,n} a_{n,i} \tag{6.3}$$

For learning, the maximum-likelihood ICA variant described in section 2.3.1 was utilized.

6.2.5 Analysis of learned basis functions

Similarity between left and right ear parts of learned basis functions was assessed using the Binaural Similarity Index (BSI), as proposed in [94]. The BSI is simply Pearson's correlation coefficient between left and right ear parts of each basis function. BSI equal to -1 means that absolute values at every frequency and time position are equal and have the opposite sign, while BSI equal to 1 means that the basis function represents the same information in both ears

Dictionary of binaural basis functions learned from natural data was classified according to the modulation spectra of their left ear parts. A modulation spectrum is a two-dimensional Fourier transform of a spectrogram. It is informative about spectral and temporal modulation of learned features and it has been applied to study properties of natural sounds [131] and real [94] as well as modelled [123] receptive fields in the auditory system.

Spatial sensitivity of basis functions learned from natural data was further quantified by means of Fisher information. Fisher information is a measure of how accurate one can estimate a hidden parameter θ from an observable s knowing a conditional probability distribution $p(s|\theta)$ [19]. Here, θ corresponds to the angular position of the auditory stimulus and s to one of the sparse coefficients. Assuming a deterministic mapping $s(\theta) = f(\theta) = \mu_\theta$ distorted with a zero-mean stationary Gaussian noise, one obtains:

$$p(s|\theta) = \mathcal{N}(s|\mu_\theta, \sigma) \tag{6.4}$$

. For simplicity σ was assumed to be equal to 1. Fisher information $\mathcal{I}(\theta)$ then becomes [19]:

$$\mathcal{I}(\theta) = (\frac{d}{d\theta} f(\theta))^2 \tag{6.5}$$

Mean values μ_θ were estimated by averaging coefficient activations over four trials during which the speaker walked around the head of the subject. Each activation time course was additionally smoothed with a 20 samples long rectangular window.

6.3 Results

Besides the properties of the sound source itself, natural sounds reaching the ear membrane are also shaped by head-related filtering. The spectrotemporal structure imposed by the filter depends on the spatial configuration of objects. By performing redundancy reduction the auditory system could, in principle, separate those two sources of variability in the data and extract spatial information. One should observe that transformations performed by the cochlea can strongly facilitate this task. The stimulus x_E (where $E \in L, R$ indicates the left or the right ear) is an air pressure waveform $g(t)$ convoluted with an HRTF (or a combination of HRTFs) $h_{E,\theta}(t)$, as defined by equation 6.2. The basilar membrane performs frequency decomposition, emulated here by the Fourier transform:

$$\mathcal{F}(x,\omega) = \int_{-\infty}^{\infty} x_E(t)\exp(-2\pi i\omega t)dt = A^x_\omega(\cos\phi^x_{E,\omega} + i\sin\phi^x_{E,\omega}) \tag{6.6}$$

where ω denotes frequency, $A^x_{E,\omega}$ amplitude and $\phi^x_{E,\omega}$ phase. By the convolution theorem [66], convolution in the temporal domain is equivalent to a pointwise product in the frequency domain, i.e.

$$\begin{aligned}\mathcal{F}((g * h_E),\omega) &= \int_{-\infty}^{\infty} g(t)\exp(-2\pi i\omega t)dt \int_{-\infty}^{\infty} h_E(t)\exp(-2\pi i\omega t)dt = \\ &= A^g_\omega(\cos\phi^g_\omega + i\sin\phi^g_\omega)A^h_{E,\omega}(\cos\phi^h_{E,\omega} + i\sin\phi^h_{E,\omega})\end{aligned}$$

Additionally, the basilar membrane applies a compressive nonlinearity [117] which this study approximates by transforming the spectral power with a logarithmic function. Since the logarithm of the product is equal to the sum of logarithms, the spectral amplitude of the stimulus $A^x_{E,\omega} = A^h_{E,\omega}A^g_\omega$ can be decomposed into the sum:

$$\log(A^h_{E,\omega}A^g_\omega) = \log(A^h_{E,\omega}) + \log(A^g_\omega) \tag{6.7}$$

. This means that the spectrotemporal representation of the signal generated by the cochlea is a sum of the raw sound and HRTF features. One should note, however, that the above analysis applies to an infinite window Fourier transform, and the data used in this study was generated by performing a Short Time Fourier Transform (STFT) with a 16 ms long, overlapping windows. Fourier coefficients were mixed between neighboring windows due to their overlap. For point-source, stationary sounds this effect did not influence the $\log(A^h_{E,\omega})$ term of the equation 6.7, since HRTFs were shorter than the STFT window, hence hear-related filtering was temporally constant. For a dynamic scene, where neighboring STFT windows contained different spatial information, the additive separability of sound and HRTF features (as described by equation 6.7) may have been distorted. Taken together, a linear redundancy reducing transform such as ICA provides a reasonable approach to separate information about object positions from the raw

sound. In an ideal case, ICA trained on stimulus spectrograms A^x_ω could separate representation of HRTF ($A^h_{E,\omega}$) and stimulus ($A^g_{E,\omega}$) amplitudes into two distinct basis functions sets [51]. The difficulty of the separation task depends on the temporal variability of the spatial information which reflects configuration of the environment (i.e. number of sources, their motion patterns and positions). The current study considers two cases of different complexity: (a) simulated dataset consisting of short periods of speech displayed from single positions and (b) a binaural recording of a natural scene with freely moving human speakers.

6.3.1 Simulated sounds

The research goal of the present chapter was to identify high-order statistics of natural sounds informative about positions of the sound source. Association of a sound waveform with its spatial position requires detailed knowledge about source localization i.e. each sound should be labelled with spatial coordinates of its source. For this reason binaural sounds studied in this section were simulated, using speech sounds and human HRTFs. Naturalistic data created in this way resembled binaural input from the natural environment, while making position labelling of sources available.

From the simulated dataset, after reducing data dimensionality with PCA (see section 6.2.3), 324 ICA basis functions were learned. A subset of 100 features is depicted in fig 6.2. It is clearly visible that the learned basis can be divided into two separate subpopulations by the similarity between their left and right ear parts, which is quantified by the Binaural Similarity Index (BSI) (see Materials and Methods). Sorted values of the BSI are displayed on fig 6.6A as black circles. The majority of basis functions (314) exceed the 0.9 threshold and only 10 fall below it. Out of those 8 reveal strong negative interaural correlation and only 2 are close to 0. Basis functions with the BSI below 0.9, were separated from the rest and all ten of them are depicted on fig 6.2A. Since they represent different information in each ear they are going to be called "binaural" through the rest of the chapter. This is in contrast to "monaural" basis functions which encode similar sound features in both ears (see fig 6.2(B))

The binaural sub-dictionary captures signal variability present due to the head-related filtering. Even though the training dataset included sounds displayed from 24 positions, hence 24 different HRTFs were used, only 10 binaural basis functions emerged from the ICA. Out of those, almost all are temporally stable i.e. do not reveal any temporal modulation (except for 2 - positions 5 and 6 on fig 6.2 A). The dominance of temporally constant features was expected, since training sounds were displayed from fixed positions and were convoluted with filters, which did not change in time. Temporally stable basis functions weight spectral power across frequency channels, mostly with opposite sign in both ears (as reflected by negative values of the BSI). Surprisingly, despite the

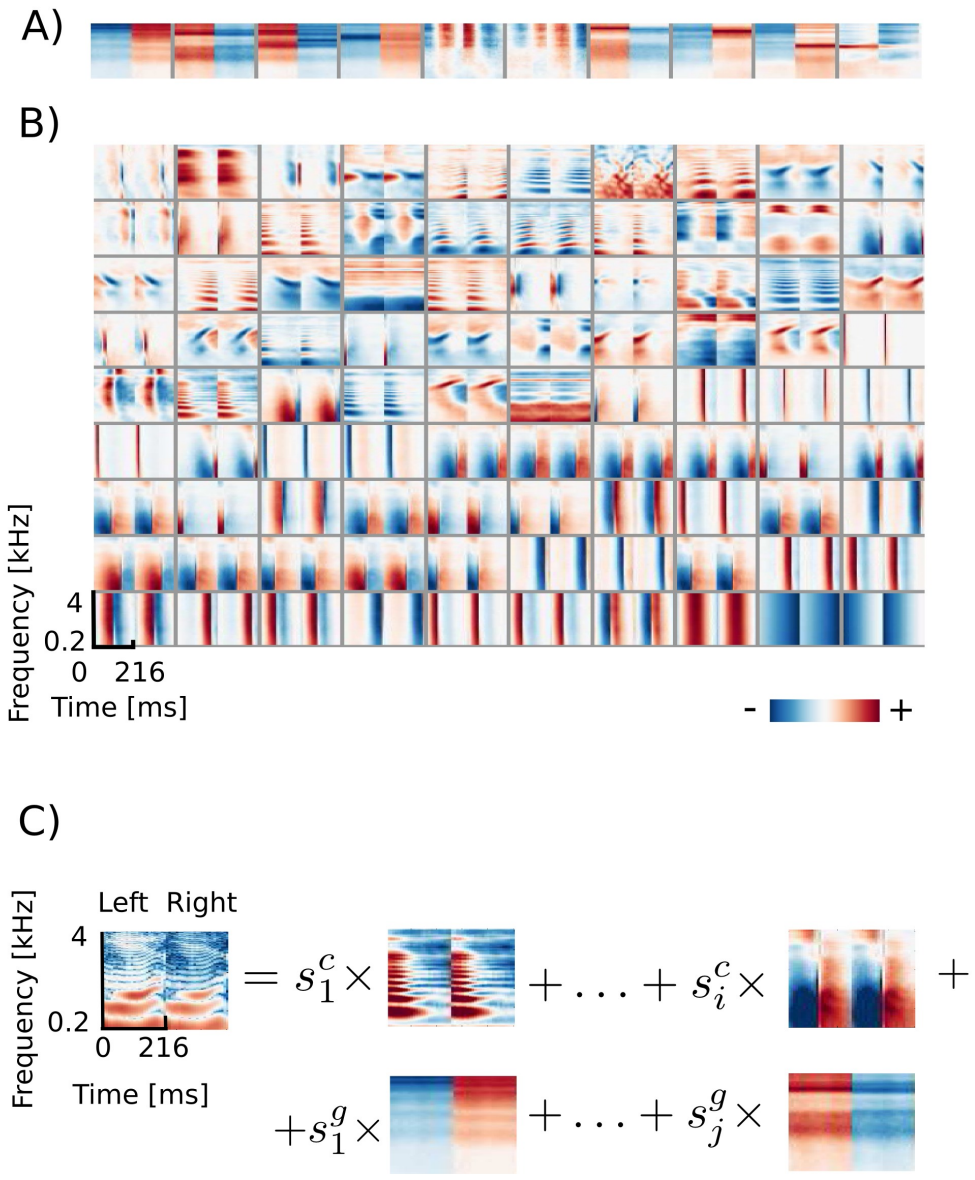

Figure 6.2: ICA basis functions trained on simulated sounds (A) Binaural basis functions a_i^g. Left and right ear parts are dissimilar. (B) Monaural basis functions a_i^k. Left and right ear parts are highly similar. (C) Explanation of the representation. Each stimulus can be decomposed into a linear combination of monaural basis functions (multiplied by their coefficients s_i^c) and binaural ones multiplied by coefficients s_i^g.

lack of moving sounds in the training dataset, two temporally modulated basis functions were also learned by the model. They represent envelope comodulation in high frequencies with an interaural phase shift of π radians.

A representative subset of 90 monaural basis functions is depicted on fig 6.2B. Their left and right ear parts are exactly the same and encode a variety of speech features. Regularities such as harmonic stacks, on- and offsets or formants are visible. Captured monaural patterns essentially reproduce results from a recent study by [25] which shows that efficient coding of speech spectrograms learns features similar to STRFs in the Inferior Colliculus. Monaural basis functions are, however, not a focus of the present study and are not going to be discussed in detail.

A separation of the learned dictionary into two subpopulations of binaural and monaural basis functions (a^g and a^k respectively) allows to represent every sound spectrogram in the training dataset as a linear combination of two isolated factors i.e. representations of speech and HRTF structures (see fig 6.2 (C)). Taking this fact into account, equation 6.3 can be rewritten as:

$$\hat{x}_{t,i} = \sum_{n=1}^{G} s^g_{t,n} a^g_{n,i} + \sum_{m=1}^{K} s^k_{t,m} a^k_{n,i} \tag{6.8}$$

This notation explicitly decomposes the basis into G spatial basis functions a^g and K non-spatial basis functions a^k.

Emergence of model spatial receptive fields

Marginal coefficient histograms conformed rather well to the logistic distribution assumed by the ICA model, although binaural coefficients were typically more sparse (see figure 6.3). In order to understand how informative learned features are about position of sound sources, conditional distributions of the linear coefficients were studied. Histograms conditioned on a location of a sound source reveal whether any spatial information is encoded by learned basis functions.

Fig 6.3 (A)-(F) displays 6 basis functions and corresponding conditional histograms. The horizontal axis of each conditional histogram corresponds to the angular position of the sound source (from 0 to 345 degrees). A vertical cross-section is a normalized histogram of the coefficient values for all sounds displayed in the training dataset from a particular position (around 2900 samples on average).

Three representative monaural basis functions are depicted on fig 6.3 (D)-(F). It is immediately visible that conditional distributions of their coefficients are stationary across spatial positions. The zero-centered logistic pdf with a constant scale parameter (parameters equal to those of the marginal pdf) is preserved across all positions. This implies that coefficients of monaural basis functions are independent from the sound source location. Monaural bases encode speech

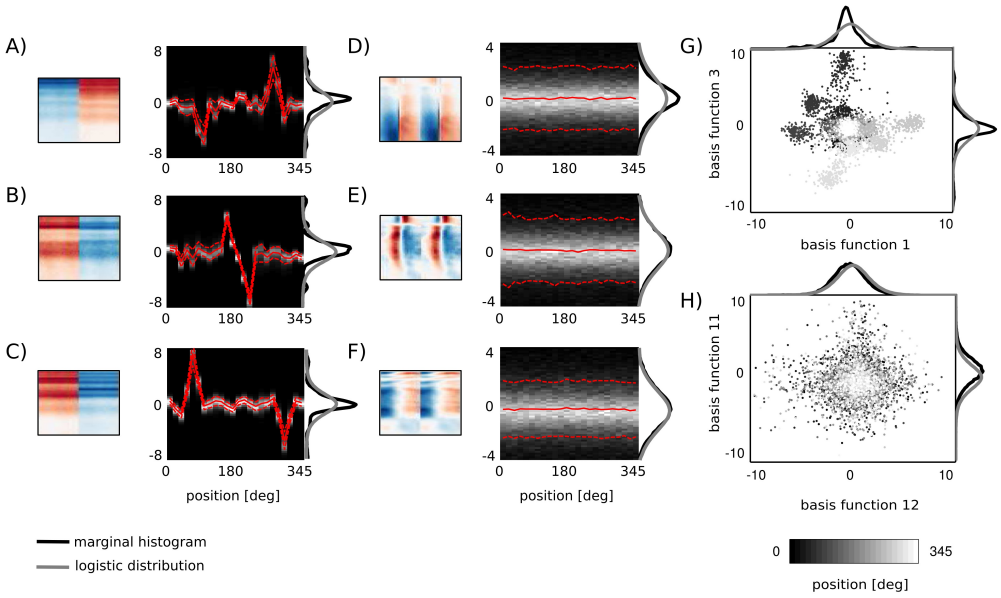

Figure 6.3: Spatial sensitivity of basis functions. (A)-(F) Spectrotemporal basis functions and associated conditional histograms of linear coefficients s. Solid red lines mark means and dashed lines limits of plus/minus standard deviation. (G)-(H) Example pairwise dependencies between monaural and binaural basis functions respectively. Each point is one sound and grayscale corresponds to its spatial position

features and since all speech structures were displayed from all positions in the training data, their activations do not carry spatial information. This property is characteristic for all basis functions with BSI greater than 0.9.

Coefficients of binaural basis functions reveal a very different dependency structure (see fig 6.3 (A)-(C)). Their variance at each spatial position is very low, however, variability across positions is much higher. Activations of binaural features remain close to zero at most angular positions regardless of the sound identity. At few preferred positions they reveal pronounced peaks in activation (positive or negative) reflected by strong shifts in the mean value. This highly non-stationary structure of conditional pdfs is informative about the sound position, while remains almost invariant to the sound's identity (which is reflected by the small standard deviation). Basis function depicted on fig 6.3 A responds with a strong positive activation to sounds originating at 270 degrees (i.e. directly in front of the right ear) and with a strong negative activation to sounds originating from the directly opposite location - at 90 degrees (i.e. in front of the left ear). Sounds at positions deviating $+/-15$ degrees from peaks also modulate basis activations, although activations are weaker. Similar spatial selectivity pattern

is revealed by the basis function on fig 6.3 C, which however responds positively to sounds at 60 and negatively to sounds at 315 degrees. The spectrotemporal feature on fig 6.3 B encodes spatial information of a particularly high behavioral relevance. Its activity significantly deviates from zero, only when sounds are placed behind the head in the interval between 165 and 210 degrees. This region is not visually accessible, therefore position or motion of objects in that area has to be inferred basing on auditory information only. It may appear that conditional histograms are symmetric around the 180 degree point. However, positive and negative peaks of coefficient histograms do not have exactly equal absolute values.

It is important to notice here that each spectrotemporal feature captured by binaural basis functions is an indirect representation of the sound position in the surrounding environment. Therefore if ICA basis functions can be interpreted as STRFs of binaural neurons, the corresponding conditional histograms constitute a theoretical analogy of their spatial receptive fields (SRFs) informing the organism about the position of the sound source within the head-centered frame of reference.

Decoding of the sound position

As described in the previous subsection, linear coefficients of binaural basis functions are informative about the location of the sound source. Spatial selectivity of single basis functions is however not specific enough to reliably localize sounds. Pairwise coefficient activations of two exemplary basis functions are depicted on fig 6.3 G. Each point represents a single sound and its color corresponds to the source's angular position. Strong clustering of same-colored points is strongly visible. They form at least 6 highly separable clusters. This, in turn, shows that the joint distribution of those two coefficients contains more information about the source position than one dimensional conditional pdfs. This is in contrast to fig 6.3 H depicting co-activations of two monaural basis functions. There, points of all colors are strongly mixed, creating a "salt and pepper" pattern, where no clear separation between source positions is visible.

To test, whether reliable decoding of sound position from activations of binaural basis functions is possible, this work employs the Gaussian Mixture Model (GMM). The GMM models the marginal distribution of latent coefficients s^g used for the position decoding as a linear combination of Gaussian distributions, such that:

$$p(s^g) = \sum_{k=1}^{24} p(s^g|C_k)p(C_k) \tag{6.9}$$

$$p(s^g|C_k) = \mathcal{N}(s^g|\mu_k, D_k) \tag{6.10}$$

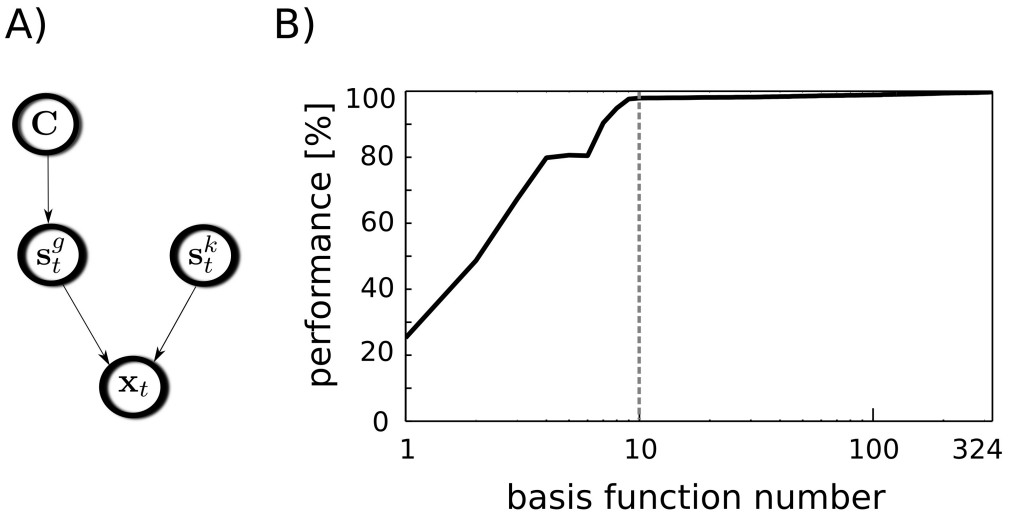

Figure 6.4: Position decoding model.(A) - A graphical model representing variable dependencies. (B) - Decoders performance plotted against the number of used basis functions. Vertical dashed line separates binaural basis functions from monaural ones.

where C_k is a position label ($C_1 = 0\,\text{deg}, C_{24} = 345\,\text{deg}$) and μ_k, D_k denote a position specific mean vector and covariance matrix respectively. The structure of dependencies among random variables is presented in a graphical form in fig 6.4 (A). Since the prior on position labels $p(C_k)$ is assumed to be uniform, the decoding procedure can be recast as a maximum-likelihood estimation:

$$\hat{C} = \arg\max_k p(s^g|C_k) \tag{6.11}$$

where $\hat{C}$ is the decoded position. The resulting procedure iterates over all position labels and returns the one which maximizes the probability of an observed data sample.

The decoding performance relies on the selected subset of basis functions used for this task. To test whether binaural features contribute stronger to the position decoding than monaural ones, all basis functions were sorted according to their BSI. Then, the GMM was trained using incrementally larger number of latent coefficients, starting from a single one corresponding to the basis function with the highly negative BSI and ending using the entire basis function set. In every step, for the GMM training 70% of the data were used, while remaining 30% were used for cross-validation. The average decoder performance is plotted against the number of used features on figure 6.4 B. Binaural features are sep-

arated from the monaural ones with a dashed vertical line. A straightforward observation is that binaural basis functions almost saturate the decoding accuracy. Indeed it reaches the level of 97.9%. Adding remaining 314 monaural basis functions increases the performance to 99.7% which is only 1.8 percentage point. Interestingly, temporally modulated binaural basis functions number 5 and 6 did not contribute to the decoding quality, which is visible as a short plateau on the plot. Saturation of the decoder's performance by binaural basis function activations entails that almost entire spatial information present in the sound is separated from other kinds of information by the ICA model and represented by binaural basis functions. Relating this observation to the nervous system, this means, that the spatial position of natural sound sources can be decoded from the joint activity of a relatively small subpopulation of binaural neurons.

6.3.2 Natural sounds

The previous section described results for simulated sounds. While simulated sounds have the advantage of giving a full control over source positions they are only a very crude approximation to the binaural stimuli occurring in the real natural environment. This section describes results obtained using binaural recordings of a real-world auditory scene, consisting of three speakers moving freely in an echo-free environment.

Binaurality of learned basis functions was again quantified with the BSI. Sorted BSI values are plotted on fig 6.6 A as gray triangles. A strong difference is visible, when compared with values of the dictionary trained on simulated data (black circles). Firstly, 64 natural basis functions lay below the 0.9 threshold - many more compared to only 10 simulated ones. Secondly, natural BSIs vary more smoothly, and are more uniformly distributed between -1 and 0.9 (see the histogram displayed in the inset).

Similarly to the previous case, the learned dictionary was divided into two sub-dictionaries - binaural ones - below and monaural ones - above the 0.9 BSI threshold. The sub-dictionary consisting of binaural basis functions is displayed on fig 6.5 A and fig 6.5 B displays 40 exemplary monaural basis functions. While no qualitative difference is visible between monaural features when compared with results from the previous section (fig 6.2 B), the binaural sub-dictionaries differ strongly. Basis functions trained using natural data, reveal much richer variety of shapes including temporally modulated ones along patterns of strong spectral modulation.

Properties of the learned representation

This subsection presents properties of binaural basis functions trained with the natural binaural data. They were studied in more detail than the dictionary

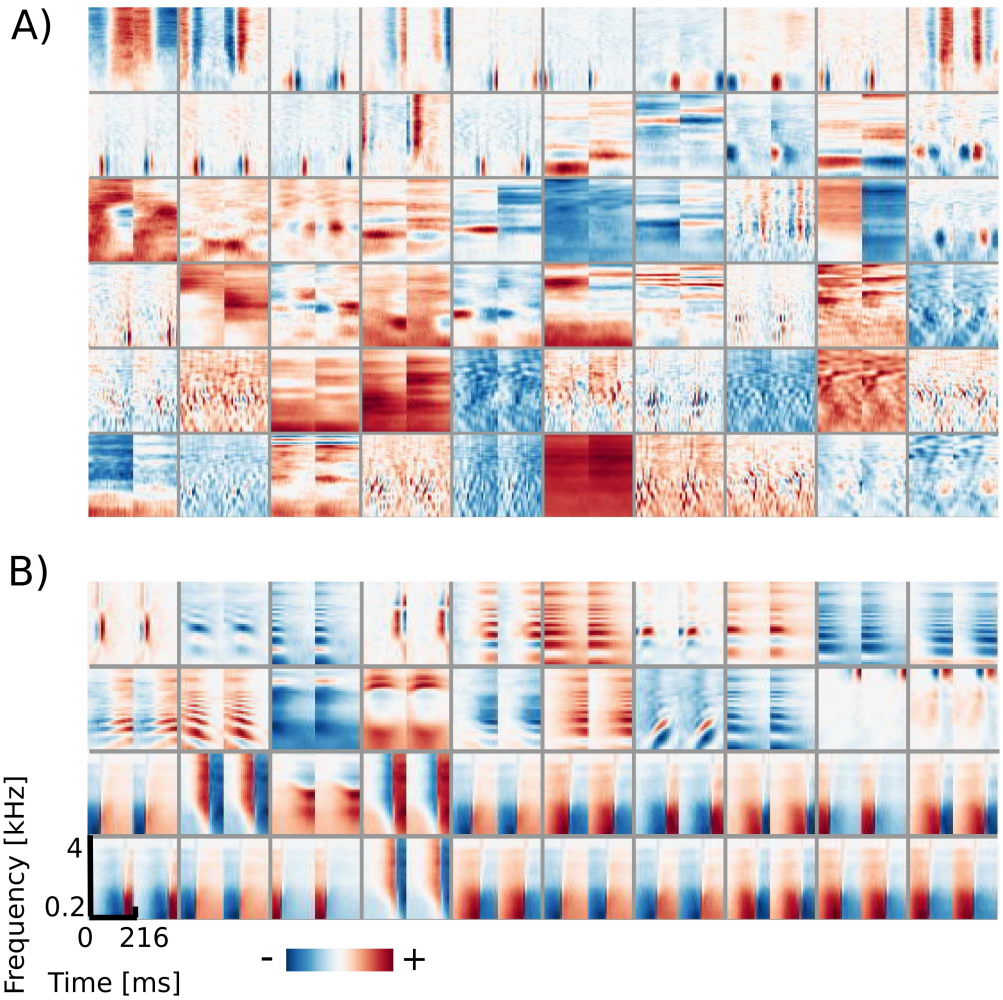

Figure 6.5: Basis functions learned using natural data. (A) - Binaural basis functions (60 out of 64) (B) - Monaural basis functions (40 out of 250)

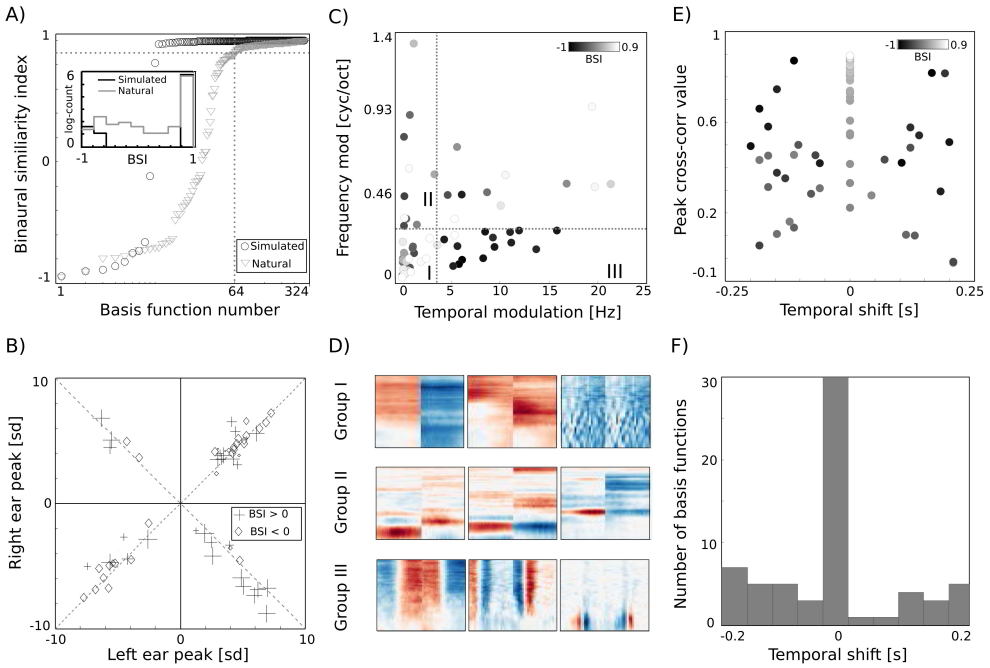

Figure 6.6: Properties of basis functions learned using natural data. (A) - BSI values of natural and simulated bases. (B) - Peak values of binaural bases. (C)- Centers of mass of modulation spectra (D) - Exemplary basis functions belonging to groups I, II and III (E) - Temporal cross-correlation plotted against its peak value. Color marks the BSI (F) - A histogram of temporal shifts maximizing the cross correlation

learned from simulated data since its structure is more complex and may reflect better the properties of binaural neurons. One should note that in neural systems modelling, neural receptive fields correspond better to ICA filters (rows $\mathbf{w}$ of matrix W in equation 6.8). Basis functions, however, constitute optimal stimuli i.e. given basis function $\mathbf{a}_i$ as input the only non-zero coefficient is going to be s_i. Additionally, basis functions are a low-passed version of filters [63], and are more appropriate for plotting, since they represent actual parts of stimulus. For those reasons, this study focuses on basis function statistics.

The binaural dissimilarity of learned features was assessed with two measures. The BSI provides a continuous value quantifying how well the left ear part matches the right ear part. It however does not take into account the dominance of one ear over another. The dominance can be measured by comparing monaural peaks i.e. points of the maximal absolute value of left and right ear parts. Both measures were used by Miller and colleagues [94] to describe receptive fields of binaural neurons in the auditory thalamus and cortex. Monaural peaks (mea-

sured in standard deviation of the basis function dimensions) are compared on fig 6.6 (B). Crosses mark basis functions with the positive and diamonds with the negative BSI. Symbol sizes correspond to the absolute BSI value. Basis functions cluster along the diagonals (marked with dashed lines) which means that left and right ear peaks have similar absolute values and no clear dominance of a single ear is present. Interestingly, while roughly the same number of basis functions lays in upper right and both lower quadrants, only 4 lay in the upper left one, corresponding to basis functions with a negative peak in the left ear and positive in the right ear. Unfortunately, direct comparison of the analysis on fig 6.6 (B) with figure 9 in [94] is not possible, due to the arbitrariness of the sign in the ICA model (coefficients can have positive and negative values, flipping the sign of the basis function). Additionally the notion of ipsi- and contra- laterality is meaningless for ICA basis functions.

Shapes of basis functions belonging to the binaural sub-dictionary were studied by analyzing modulation spectra of their left-ear parts. Even though functions were binaural, classification according to only the single ear part was sufficient to identify subgroups with interesting binaural properties. Centers of mass of modulation spectra (for computation details see Materials and Methods) are plotted as circles on fig 6.6 (C). Circle color corresponds to the BSI value. Left parts of binaural features display a tradeoff between spectral and temporal modulation. This complies with the general trend of natural sound statistics [131]. Dictionary elements were divided into three distinctive groups according to their modulation properties (marked with Roman numerals I, II, III and separated with dotted lines on fig 6.6 (C)). The first group consisted of weakly modulated features with spectral modulation below 0.3 cycles/octave and temporal modulation below 4 Hz. Majority of basis functions belonging to this group had high BSI, close to 0.9. Three representative members of the first group are displayed on fig 6.6 (D) in the first row. Since their spectrotemporal modulation is weak, they capture constant patterns, similar in both ears, up to the sign. The second group consists of basis functions revealing strong spectral modulation - above 0.3 cycles/octave. Three exemplary members are visible in the second row of fig 6.6 (D). Basis functions belonging to the second group resemble majority of ones learned from simulated data. They weight spectral power across frequency channels constantly over time. In contrast to simulated basis functions, their BSIs are mostly close to 0, indicating that channel weights do not necessarily have opposite sign between ears. Additionally, as visible in two out of three displayed examples, low frequencies below $1kHz$ are also weighted.

The third group includes highly temporally modulated features. Their temporal modulation exceeds 4 Hz, while the spectral one stays below 0.3 cycles/octave. Out of 15 members of this group, only one has a positive BSI value - the rest remains close to -1. This implies that when their monaural parts are aligned

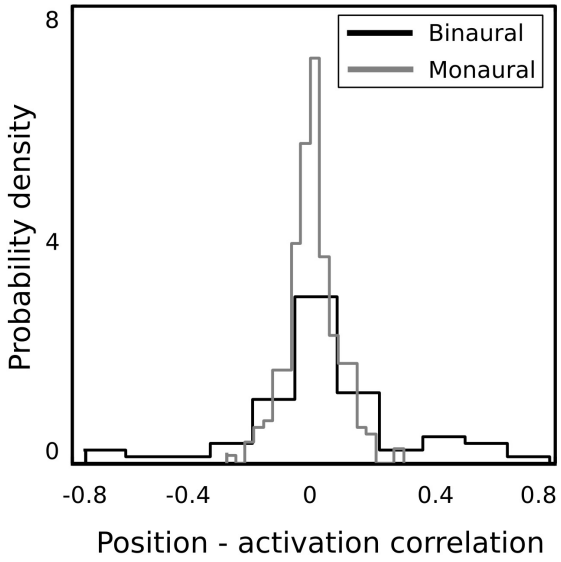

Figure 6.7: Normalized histograms of activation-position correlations

with each other - corresponding dimensions have a similar absolute value and an opposite sign. Three exemplary members of the third group are depicted in the last row of fig 6.6 (D). They are qualitatively similar to two temporal basis functions learned from the simulated data (they represent an envelope comodulation across multiple frequency channels with a π phase difference).

The temporal differences between monaural parts of basis functions were further studied using cross-correlation functions (ccf). Maximal values of the normalized ccf are plotted against maximizing temporal shifts on fig 6.6 (E). As in the fig 6.6 (C) - the color of circles represents the BSI value. The histogram of temporal shifts is depicted on fig 6.6 (F). Cross-correlation of 30 binaural features with a positive BSI, is maximized at 0 temporal shift. In this case, BSI and the peak of cross-correlation have the same value. This is a property of basis functions with a weak temporal modulation, which constitute a major part of the binaural sub-dictionary. Features revealing temporal modulation have a negative BSI value (dark colors) and a non-zero temporal difference, which spanned the range between -0.2 to 0.2 seconds.

Spatial sensitivity of binaural basis functions

In contrast to the simulated dataset, binaural recordings were not labelled with sound source positions. Furthermore, learned features may represent dynamic aspects of the object motion, therefore conditional histograms (constructed as in the previous section) would not be meaningful.

In order to verify whether binaural basis functions reveal tuning to spatial position of sound sources and invariance to their identity, a test recording was performed. One of the male speakers read a book out loud, while walking around the head of the recording subject, following a circular trajectory in a constant pace. This was repeated twice in the anticlockwise and twice in the clockwise direction. In such a way, the angular position of the speaker was made easy to estimate at each time point. The recording was divided into 216 ms overlapping intervals, and each interval was encoded using the learned dictionary. A general trend in the spatial sensitivity of basis functions was measured by computing correlation between estimated speaker's position and time courses of linear coefficients in the following way. Firstly, activation time courses were standarized to have mean equal to 0 and variance equal to 1. In the next step, time intervals where the coefficient's absolute value exceeded 1 were extracted. This was done, since highly sensitive coefficients remained close to 0 most of the time, and correlated with the speaker's position only in a narrow part of the space (i.e. their receptive field). Elements of the binaural sub-dictionary correlated stronger with the estimated position than elements of the monaural one. Normalized histograms of linear correlations between the position of the sound source and sparse coefficients are presented on fig 6.7. Monaural basis functions correlate much weaker with the sound position, which is reflected in the strong histogram peak around 0. Binaural coefficients in turn, reveal strong correlations of the absolute value of 0.8 in extreme cases. Linear correlation is however not a perfect way to assess relationship between sparse coefficients and the source position, since spatial selectivity of basis function may be limited to a narrow spatial area (as in fig 6.8 A and B). This results in correlations of low absolute values, even though spatial sensitivity of a basis function may be quite high. To show spatial selectivity of learned features, their activations were plotted. Resulting time courses of basis function activations are displayed as black continuous lines on fig 6.8. Gray dashed lines mark approximated angular position of the speaker at every time point.

Subfigures (F)-(J) display activations of 5 representative monaural basis functions. As expected, their activity correlates very weakly with the speaker's trajectory. Monaural basis functions encode features of speech and are invariant to the position of the speaker. In contrary, activations of binaural basis functions visible on subfigures (A)-(E), reveal strong dependence on subjects position and direction of motion. Basis function A remains non-activated for most of positions and deviates from zero when the speaker is crossing the area behind the head of the recording subject. The slope of activation time courses is informative about the direction of speaker's motion. Similar, however noisier, spatial tuning is revealed by the basis function D. Basis function B displays broader spatial sensitivity, and its activation varies smoothly along the circle surrounding the

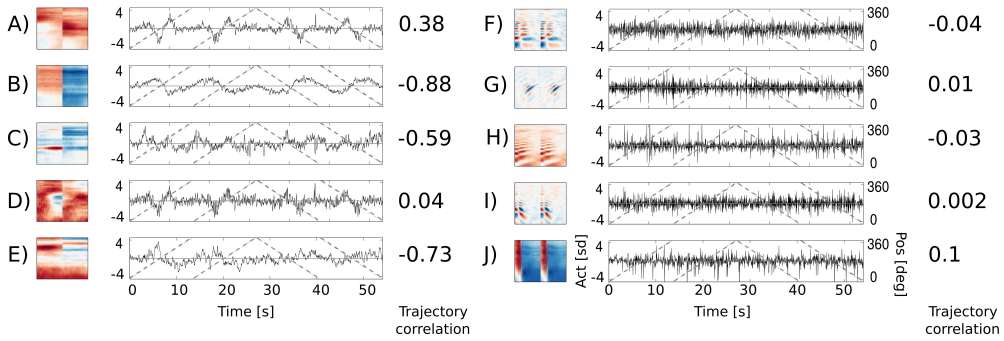

Figure 6.8: Activation time course of basis functions learned using natural data.An audio-video version is available in the supplementary material. Subfigures (A)-(E) depict binaural basis functions with their activation time courses, while subfigures (F)-(J) monaural ones. Black continuous lines mark standardized activation values, gray dashed lines mark speaker's angular position.

subject's head. Spatial information represented by the spectrally modulated basis functions C and E does not have such a clear interpretation, however they display pronounced covariation with sound source's position (feature C for instance is strongly positively activated, when the speaker crosses directly opposite to the left ear).

Spatial sensitivity of basis functions can be further quantified using Fisher information (for computation details please see Materials and Methods). Figure 6.9 shows Fisher information estimates as a function of spatial position for features displayed on figure 6.8. Each binaural basis function reveals a preferred region in space where source's position is encoded with higher accuracy. For this reason, histograms depicted on figures 6.9A-E can be interpreted as an abstract descriptions of auditory spatial receptive fields. Basis function (A), is most strongly informative about position of the sound source behind the head (around 180 degrees), which is also reflected in the time course of its activation. The Fisher information peaks in visually inaccessible areas also in other, depicted basis functions (subfigures (B), (C), (E)). There, however, the peak is not as pronounced as in the first basis function, and sensitivity to frontal positions is also visible. Fisher information of monaural basis functions (subfigures (F)-(J)) does not reveal spatial selectivity, is order of magnitude smaller and would most probably vanish in the limit of more samples.

All binaural basis functions presented on fig 6.8 are weakly temporally modulated. Temporally modulated basis functions, do not correlate strongly with the speaker's position (they also did not contribute to the position decoding, as described in the previous section).

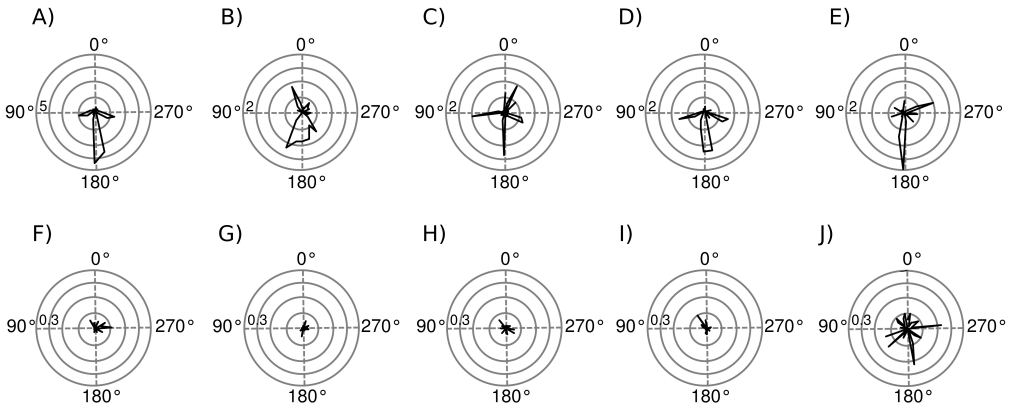

Figure 6.9: Spatial sensitivity quantified with Fisher information. Polar plots represent area surrounding the listener, black lines mark Fisher information $\mathcal{I}(\theta)$ at each angular position. Each subfigure corresponds to a basis function on the previous figure marked with the same letter. Please note different scales of the plots.

6.4 Discussion

The auditory system has to infer the spatial arrangement of the surrounding space by analyzing spectrotemporal patterns of binaural sound. Auditory spatial receptive fields are formed, by extracting signal features which correlate well with environment's spatial states and result from the head related filtering. Both sound datasets used in the present study included two, categorically different variability sources: spatial information carried by binaural differences resulting from the HRTF filtering and the raw sound waveform. Application of ICA - a simple redundancy reducing transform led to a separation of those information sources and formation of distinct model neuron sub-populations with specific spatial and spectrotemporal sensitivity.

6.4.1 Linear processing of spectrotemporal binaural cues

Emulation of the cochlear processing by performing spectral decomposition and application of the logarithmic nonlinearity produces a data representation well adapted for the position decoding task. While it is usually argued that the logarithmic nonlinearity implemented by mechanical response of the cochlear membrane is useful for reducing the dynamical range of the signal [117] it provides an additional advantage. Since in the frequency domain convolution is equivalent to a pointwise product of the signal and the filter [66], a logarithm transforms it to a simple addition. A linear operation on the "cochlear" data representation

suffices to extract features imposed by the pinnae filtering [51]. One should note, however, that in complex listening situations involving more than a single, stationary sound source, this simple relationship (as described by equation 6.7) may be distorted and extracted features can be mixing different aspects of the signal.

It has been observed that a linear approximation of spectrotemporal receptive fields in the auditory cortex predicts their spatial selectivity [125]. This result may be surprising given that sound localization is a non-linear operation [64] and that in a general case, linear STRF models do not explain firing patterns of auditory neurons [36, 27]. Results described in this paper suggest that a linear-redundancy reducing transform applied to log-spectrograms suffices to create model spatial receptive fields, providing a candidate computational mechanism explaining results provided by [125]. Localization of a natural sound source involves information included in multiple frequency channels. Binaural cues such as ILD are computed in each channel separately and have to be fused together at a later stage. This is exemplified by temporally constant basis functions learned using simulated and natural datasets. They linearly weight levels in frequency channel of both ears and in this way form their spatial selectivity. Interestingly the weighting is often asymmetric (which is reflected by BSI values different from -1). Such patterns represent binaural level differences coupled across multiple frequency channels. A recent study has shown that a similar computational strategy underlies spatial tuning of binaural neurons in the nucleus of the brachium of inferior colliculus (IC) in monkeys [132]. Since it has already been suggested that IC neurons code natural sounds efficiently [25], present results extend evidence in support of this hypothesis.

6.4.2 Complex shapes of binaural STRFs

Early binaural neurons localized in the auditory brainstem can be classified according to kinds of input they receive from each ear (inhibitory-excitatory - IE and excitatory-excitatory - EE) [49]. At the higher stages of auditory processing (Inferior Colliculus, Auditory Cortex), binaural neurons respond also to complex spectrotemporal excitation-inhibition patterns [94, 112, 125]. This chapter suggests, which kinds of binaural features may be encoded and used for spatial hearing tasks by higher binaural neurons. It demonstrates that the reconstruction of natural binaural sounds requires basis functions representing various spectrotemporal patterns in each ear. The dictionary of learned binaural features is best described by a continuous binaural similarity value (in this case Pearson's correlation coefficient - BSI) and not by a classification into non-overlapping IE-EE groups. Temporally modulated basis functions constitute a particularly interesting subset of all binaural ones. Many of them represent a single cycle of envelope modulation, in opposite phase in each ear (see figure 6.6 (D)). The time interval

corresponding to such phase shift is, however, much larger than the one required for the soundwave to travel between the ears. Their emergence and aspects of the environment they represent remain to be explained. Coding of different spectrotemporal features in each ear is useful not only for sound localization and tracking, but may be also applied for separation of sources while parsing natural auditory scenes (i.e. solving the "cocktail party problem").

6.4.3 The role of HRTF structure

Spatial information is created when the sound waveform becomes convoluted with the head and pinnae filter - HRTF. By taking into account that this convolution is equivalent to addition of the log-spectral representation of the sound and the HRTF, one may conclude that the ICA recovers exact HRTF forms. A subset of basis functions learned by the ICA model from the simulated data could, in principle, contain 24 elements, which would constitute an exactly recovered set of HRTFs used to generate the training data (see figure 6.1 D). The other basis function subset would contain features modelling speech variability. This is, however, not the case. Firstly - in the simulated dataset - HRTFs corresponding to 24 positions were used, 10 basis functions emerged and only 8 were temporally non-modulated, as HRTFs are. Despite such dimensionality reduction, information included in the 8 basis functions was sufficient to perform the position decoding with 15 deg spatial resolution. This implies that binaural basis functions did not recover HRTF shapes but rather formed their compressed representation. It is important to note here that learned binaural features were much smoother and did not include all spectral detail included in HRTFs themselves (compare basis binaural basis functions from figures 6.2 A and 6.5 A with HRTFs from figure 6.1 D). The fact that coarse spectral information suffices to perform position decoding stands in accord with human psychophysical studies. It has been demonstrated that HRTFs can be significantly smoothed without influencing human performance in spatial auditory tasks [72].

In humans and many other species, the area behind the listener's head is inaccessible to vision and information about the presence or motion of objects there can be obtained only by listening. This particular spatial information is of high survival value since it may inform about an approaching predator. Interestingly, in both used datasets features providing pronounced information about presence of sound sources behind the head clearly emerged (see figs 6.3 (B) and 6.8 (A)). Their sensitivity to sound position quantified with Fisher information is highest for the area roughly between 160 to 230 degrees. Since those basis functions reflect the HRTF structure, one could speculate that the outer ear shape (which determines the HRTF) was adapted to make this valuable spatial information explicit. It is interesting to think that one of the factors in pinnae evolution, was to provide spectral filters, highly informative about sound positions behind the

head. This, however, can not be verified within the current setup and remains a subject of the future research.

6.5 Conclusion

Taken together, results described in this chapter demonstrate that a theoretical principle of efficient coding can explain the emergence of functionally separate neural populations. Using an exemplary task of binaural hearing I have demonstrated, that a linear redundancy-reducing transform is capable of learning informative signal features, which belong to two classes - position and identity invariant. As long as such invariances have not yet been conclusively identified in the auditory system, their existance is theoretically possible and can be accounted for by efficient coding principles.

Chapter 7

Conclusions and Outlook

Being a theoretician and thinking about a specific part of the brain (such as the binaural hearing system) one is exposed to a strong temptation of seeking for generalities. Insights into the detailed functioning of a particular neuronal subsystem could be hopefully generalized and provide hints of abstract principles which describe functioning of other parts of the brain. In this chapter I will summarize specific auditory findings of this work and relate them to two general statments about sensory information processing proposed in the introduction.

7.1 Neuronal function in the natural environment - lessons from spatial hearing

In this thesis I described my attempts to analyze the binaural auditory system through the lens of natural stimulus statistics. Chapters $4-6$ describe results of three such studies and discuss them in the context of auditory physiology. In the first chapter I have introduced two general tenets which relate the function of sensory neurons (or what the external observer may call their "role" or "purpose") to natural stimuli. Below I state them again, and discuss them in light of results described in this work.

1. The function of sensory neurons can not be fully elucidated without understanding statistics of natural stimuli they process

Since the early days of research into the binaural auditory system it has been observed that neurons of the superior olivary complex reveal sensitivity to binaural cues - IPDs and ILDs. Considering a very simple scenario (similar to the one studied by Lord Rayleigh) extraction of a cue is equivalent to the localization of a sound on the horizontal plane. In cases where this is true it can be said that the nervous system at a surprisingly early processing stage computes the localization

of a sound source. Neurons which perform this function for high-frequency sounds by computing ILDs are located in LSO and low-frequency sounds are localized by IPD extractors in the MSO.

This chain of reasoning raises a number of questions. If sounds are localized so early, what is the function of further binaural processing? Is the auditory cortex necessary for the sound localization? Those issues are often mentioned and discussed in the literature [124, 125, 100, 101]. A satisfactory answer has not yet been provided.

While the analysis of natural sound statistics presented in chapter 4 may not give immediate answers to those questions, it may provide useful hints. Firstly, as the ICA analysis has demonstrated, natural sounds in each ear seem to be dominated by independent acoustic events. If this is the case then the computation of a cue as performed by SOC neurons is not equivalent to the localization of a sound source. The function of those cells is therefore not to "localize sound sources" but to perform a stimulus transformation which constitutes a first step of an intricate scene analysis process. It becomes obvious that numerous other computations are required to understand the scene configuration from the sensory input. The importance is shifted away from the question "what does the auditory cortex do?" towards "how does it do it?".

The second observation is that natural cue distributions deviate (in some aspects quite strongly) from analytical predictions. For instance ILD distributions are almost invariant to the frequency and the scene, and profound level differences are present in low frequency ranges. To make use of such potentially useful information the auditory system should encode low-frequency ILDs. This observation predicts existence of ILD sensitive cells of low best-frequencies, which may be neglected according to the duplex-theory.

More generally, the results described in chapter 4 of this thesis highlight the importance of understanding the structure of natural stimuli for sensory neuroscience. Performing experiments with simple stimuli has without doubt many advantages. Artificial stimuli are well controllable, and research basing on them has led to a large increase of understanding of the auditory system as well as other sensory systems. It has been even argued that it is a most fundamental line of research, and that natural stimuli should be used as benchmark tests of theories derived in reductionist experimental settings [122]. In my view the analysis of natural stimulus statistics is at least as important and should be performed in parallel. After all it is impossible to understand the algorithm implemented by an information processing system without knowing the data it processes. As statistics of binaural sounds show, Nature can be surprisingly complex, and one can rarely predict all features of the stimulus only with pen and paper.

2. Function of sensory representations reflects redundancies present in the natural sensory environment

Finding a tonotopic representation of sound in the auditory cortex has elevated hopes that the brain also forms such clearly interpretable, topographic representations of other stimulus features - for instance sound location [92]. Surprisingly this was not the case. When anaesthetized cats were presented with sounds located at different positions multiple neurons were responding to a broad range of locations [93]. Since they varied their activity with a change of the sound position - the experimenters concluded that the function of studied neurons is to encode this property. The coding strategy was not well understood - it was called "the panoramic code" [93]. In the following years attempts have been made to explain these experimentally observed tuning properties. For instance, it has been argued that spatial tuning curves are "designed" to be highly informative about behaviorally relevant areas of space [137].

Results of chapter 5 show that a sparse representation of natural stereo sounds reveals "spatial tuning" very similar to cortical neurons. It has, however, not been pre-designed to encode a pre-selected aspect of the environment - the position of a point sound source. It rather emerges in a process of adaptation to the natural stimulus via seeking an information efficient encoding. This observation leads to a hypothesis that perhaps the entire nervous system, rather than being a collection of loosely coupled "problem-solvers", follows a single coding strategy. Introduction and chapter 3 discuss that as experimenters we may be under the illusion that neurons we study in an experimental setup encode the chosen parameter. Observed variation in neuronal responses does not necessarily mean that.

As suggested by numerous previous studies, the function of sensory neurons may be to recode the stimulus stream in an efficient form. It may be true even for sensory neurons located away from the sensory periphery. This means that the stimulus structure preferred by a neuron (i.e. its receptive field) can be very hard to interpret using natural language terms. The form of neuronal receptive fields may defy our high-level introspective intuitions, for instance the apparent necessity of "what" and "where" separation. A good example has been provided by a recent study, which questions the perceptual relevance of traditional classification of speech sounds into vowels and consonants [141]. As results of the study indicate perception of speech depends on entropy of the signal not on the presence of sounds belonging to these pre-supposed categories.

As postulated by Barlow [9] maximization of the coding efficiency can extract stimulus regularities which are vital to perform inferences about the environment. As results presented in chapter 6 show, finding a linear efficient encoding of the binaural sound in the log-spectral domain is capable of separating sources of variability in the stimulus. Spatial information imposed by HRTFs is represented

by a distinct subpopulation of units. It is important to stress again that their function is not pre-determined - they emerge in an unsupervised learning process by adaptation to stimulus statistics.

It is interesting to speculate that the general principle of efficient coding guides the formation of sensory representations at all processing stages. The goal of the nervous system may be to remain in the "informational equilibrium" with the environment by absorbing as much information as possible. Categories pre-supposed by human observers may therefore not map directly onto the true function implemented by neuronal circuits. Their understanding may require insights into the structure of the natural sensory world.

7.2 Caveats and limitations

Arguments presented in this thesis rely on the strong assumption that the set of collected sound data is representative for the mammalian (and human in particular) sensory niche. In contrast to natural images, natural sounds seem to have rather inhomogeneous structure. Statistics of image patches, are well reproducible - algorithms such as sparse coding yield similar features when trained on different images. Sparse representations of natural sounds in turn, vary strongly, depending on a sound class [78]. There is an ongoing debate on how a representative dataset of natural sounds would look like [145]. Results obtained here could be strengthened by extending the set of analyzed auditory scenes.

It is important to keep in mind that stimulus statistics are determined not only by the environment but also by the organism. In vision, for instance, it has been demonstrated that local statistics of natural images measured at the center of gaze differ from those of uniformly sampled image patches [115]. In spatial hearing this is especially the case. Shape of the head and pinnae affect properties of the stimulus. While certain features of neuronal space representations (such as the broad tuning of cortical neurons) seem to be replicated across mammalian species [49], some others may be not. Moreover, in birds [49, 88] and reptiles [37] spatial hearing seems to rely on different mechanisms than in mammals.

In chapter 3, two possible roles of efficient coding have been proposed (illustrated on figure 2.7) They both assume the existence of the "raw sensory stream" - unprocessed stimulus from which relevant information has to be extracted. In hearing this may be the activity of all haircells aligned along the cochleotopic axis in the organ of Corti. An important and unanswered question is - how is the "bandwidth" of this raw stream determined in the first place? Is the frequency range available to humans selected by evolutionary mechanisms as behaviorally relevant?

The second closely related problem regards time-scales of adaptation. In this thesis I have analyzed relatively small datasets (at most 12 minutes long) and

modelled short time receptive fields (measured in milliseconds). Efficient coding mechanisms could in principle operate on different time-scales - from milliseconds ([142]) to evolutionary epochs. The work described here remains largely agnostic about this issue.

Considerations discussed above provide possible starting points for extensions of research described in this thesis.

7.3 Coda

Turtles and finches of the Galapagos quickly attracted Mr Darwin's attention. What seemed remarkable was that members of the same species looked very different depending on which of the archipelago's well separated islands they inhabited. Anatomical traits such as the shell color or the shape of the beak seemed to be determined by the animal's surrounding. This observation has led Charles Darwin to reason that organisms are adapted to their environment - it is an idea, which became one of the cornerstones of the evolutionary theory.

Nowadays principles of adaptation to the environmental niche guide our study of not only crude anatomical traits, but also of abstract information processing mechanisms employed by the nervous system. The way to the satisfactory comprehension of the inner workings of this mysterious structure is still long. We seem, however, to know at least the good direction. It has been indicated more than one and a half century ago by a young naturalist on the deck of HMS *Beagle*.

Bibliography

[1] Samer A Abdallah and Mark D Plumbley. If the independent components of natural images are edges, what are the independent components of natural sounds. In *Proceedings of International Conference on Independent Component Analysis and Signal Separation (ICA2001)*, pages 534–539, 2001. [cited at p. 64, 89, 90, 102]

[2] Edgar Douglas Adrian. The basis of sensation. 1928. [cited at p. 5]

[3] Hiroki Asari, Barak A Pearlmutter, and Anthony M Zador. Sparse representations for the cocktail party problem. *The Journal of neuroscience*, 26(28):7477–7490, 2006. [cited at p. 102]

[4] International Phonetic Association. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999. [cited at p. 110]

[5] Joseph J Atick. Could information theory provide an ecological theory of sensory processing? *Network: Computation in neural systems*, 3(2):213–251, 1992. [cited at p. 30]

[6] H Attias and CE Schreiner. Temporal low-order statistics of natural sounds. *Advances in neural information processing systems*, pages 27–33, 1997. [cited at p. 42]

[7] Fred Attneave. Some informational aspects of visual perception. *Psychological review*, 61(3):183, 1954. [cited at p. 13, 15, 18]

[8] Nihat Ay and Daniel Polani. Information flows in causal networks. *Advances in complex systems*, 11(01):17–41, 2008. [cited at p. 18]

[9] Horace Barlow. Redundancy reduction revisited. *Network: computation in neural systems*, 12(3):241–253, 2001. [cited at p. 107, 133]

[10] Horace B Barlow. Possible principles underlying the transformation of sensory messages. *Sensory communication*, pages 217–234, 1961. [cited at p. 6, 13, 15, 18, 107]

[11] Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995. [cited at p. 25, 26]

[12] Anthony J Bell and Terrence J Sejnowski. The independent components of natural scenes are edge filters. *Vision research*, 37(23):3327–3338, 1997. [cited at p. 30, 64]

[13] William Bialek. *Biophysics: searching for principles*. Princeton University Press, 2012. [cited at p. 71]

[14] Jennifer K Bizley and Kerry MM Walker. Sensitivity and selectivity of neurons in auditory cortex to the pitch, timbre, and location of sounds. *The Neuroscientist*, 16(4):453–469, 2010. [cited at p. 10, 44, 75, 101]

[15] Jennifer K Bizley, Kerry MM Walker, Bernard W Silverman, Andrew J King, and Jan WH Schnupp. Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. *The Journal of Neuroscience*, 29(7):2064–2075, 2009. [cited at p. 10, 44, 75, 101, 105]

[16] Jens Blauert. *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997. [cited at p. 70]

[17] Antje Brand, Oliver Behrend, Torsten Marquardt, David McAlpine, and Benedikt Grothe. Precise inhibition is essential for microsecond interaural time difference coding. *Nature*, 417(6888):543–547, 2002. [cited at p. 72]

[18] Guy J Brown and DeLiang Wang. Separation of speech by computational auditory scene analysis. In *Speech enhancement*, pages 371–402. Springer, 2005. [cited at p. 72]

[19] Nicolas Brunel and Jean-Pierre Nadal. Mutual information, fisher information, and population coding. *Neural Computation*, 10(7):1731–1757, 1998. [cited at p. 112]

[20] Douglas S Brungart and William M Rabinowitz. Auditory localization of nearby sources. head-related transfer functions. *The Journal of the Acoustical Society of America*, 106:1465, 1999. [cited at p. 71]

[21] Daniel A Butts and Mark S Goldman. Tuning curves, neuronal variability, and sensory coding. *PLoS biology*, 4(4):e92, 2006. [cited at p. 103]

[22] Charles Cadieu. Probabilistic models of phase variables for visual representation and neural dynamics. *PhD thesis, University of California, Berkeley*, 2009. [cited at p. 58]

[23] Charles F Cadieu and Kilian Koepsell. Phase coupling estimation from multivariate phase statistics. *Neural computation*, 22(12):3107–3126, 2010. [cited at p. 58]

[24] Charles F Cadieu and Bruno A Olshausen. Learning intermediate-level representations of form and motion from natural movies. *Neural computation*, 24(4):827–866, 2012. [cited at p. 79, 90]

[25] Nicole L Carlson, Vivienne L Ming, and Michael Robert DeWeese. Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. *PLoS computational biology*, 8(7):e1002594, 2012. [cited at p. 42, 74, 93, 102, 108, 111, 116, 128]

[26] Gal Chechik, Michael J Anderson, Omer Bar-Yosef, Eric D Young, Naftali Tishby, and Israel Nelken. Reduction of information redundancy in the ascending auditory pathway. *Neuron*, 51(3):359–368, 2006. [cited at p. 41]

[27] G Björn Christianson, Maneesh Sahani, and Jennifer F Linden. The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields. *The Journal of Neuroscience*, 28(2):446–455, 2008. [cited at p. 128]

[28] Yale E Cohen, Brian E Russ, Gordon W Gifford, Ruwan Kiringoda, and Katherine A MacLean. Selectivity for the spatial and nonspatial attributes of auditory stimuli in the ventrolateral prefrontal cortex. *The journal of neuroscience*, 24(50):11307–11316, 2004. [cited at p. 44, 75]

[29] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012. [cited at p. 16, 17, 20]

[30] Johannes C Dahmen, Peter Keating, Fernando R Nodal, Andreas L Schulz, and Andrew J King. Adaptation to stimulus statistics in the perception and neural representation of auditory space. *Neuron*, 66(6):937–948, 2010. [cited at p. 43, 71]

[31] John G Daugman. Entropy reduction and decorrelation in visual coding by oriented neural receptive fields. *Biomedical Engineering, IEEE Transactions on*, 36(1):107–114, 1989. [cited at p. 28, 29]

[32] Robert Desimone. Face-selective cells in the temporal cortex of monkeys. *Journal of cognitive neuroscience*, 3(1):1–8, 1991. [cited at p. 6]

[33] Michael R DeWeese, Michael Wehr, and Anthony M Zador. Binary spiking in auditory cortex. *The Journal of neuroscience*, 23(21):7940–7949, 2003. [cited at p. 41]

[34] Theodosius Dobzhansky. Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher*, 75(2):87–91, 1972. [cited at p. 20]

[35] V Ming E Bumbacher. Pitch-sensitive components emerge from hierarchical sparse coding of natural sounds. *ICPRAM*, 2012. [cited at p. 93]

[36] Monty A Escabı and Christoph E Schreiner. Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. *The Journal of neuroscience*, 22(10):4114–4131, 2002. [cited at p. 128]

[37] Richard R Fay, Arthur N Popper, and Douglas B Webster. *The evolutionary biology of hearing*. Springer-Verlag, 1992. [cited at p. 134]

[38] David Field. What is the goal of sensory coding? *Neural computation*, 6(4):559–601, 1994. [cited at p. 27, 28]

[39] Brian J Fischer. Optimal models of sound localization by barn owls. In *Advances in Neural Information Processing Systems*, pages 449–456, 2007. [cited at p. 69]

[40] Brian J Fischer and José Luis Peña. Owl's behavior and neural representation predicted by bayesian inference. *Nature neuroscience*, 14(8):1061–1066, 2011. [cited at p. 69]

[41] Peter Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200, 1991. [cited at p. 81]

[42] Apostolos P Georgopoulos, John F Kalaska, Roberto Caminiti, and Joe T Massey. On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *The Journal of Neuroscience*, 2(11):1527–1537, 1982. [cited at p. 20]

[43] Wulfram Gerstner, Richard Kempter, J Leo van Hemmen, and Hermann Wagner. A neuronal learning rule for sub-millisecond temporal coding. *Nature*, 383(LCN-ARTICLE-1996-002):76–78, 1996. [cited at p. 40]

[44] Patrick Gill, Junli Zhang, Sarah MN Woolley, Thane Fremouw, and Frédéric E Theunissen. Sound representation methods for spectro-temporal receptive field estimation. *Journal of computational neuroscience*, 21(1):5–20, 2006. [cited at p. 108, 111]

[45] Noam Gordon, Trevor M Shackleton, Alan R Palmer, and Israel Nelken. Responses of neurons in the inferior colliculus to binaural disparities: insights from the use of fisher information and mutual information. *Journal of neuroscience methods*, 169(2):391–404, 2008. [cited at p. 103]

[46] Boris Gourévitch and Romain Brette. The impact of early reflections on binaural cues. *The Journal of the Acoustical Society of America*, 132:9, 2012. [cited at p. 71]

[47] Donald D Greenwood. A cochlear frequency-position function for several species29 years later. *The Journal of the Acoustical Society of America*, 87:2592, 1990. [cited at p. 111]

[48] Charles G Gross. Genealogy of the grandmother cell. *The Neuroscientist*, 8(5):512–518, 2002. [cited at p. 6]

[49] Benedikt Grothe, Michael Pecka, and David McAlpine. Mechanisms of sound localization in mammals. *Physiological Reviews*, 90(3):983–1012, 2010. [cited at p. 37, 38, 40, 61, 71, 72, 74, 77, 84, 103, 104, 128, 134]

[50] Kenneth E Hancock and Bertrand Delgutte. A physiologically based model of interaural time difference discrimination. *The Journal of neuroscience*, 24(32):7110–7117, 2004. [cited at p. 72]

[51] Nicol Harper and Bruno Olshausen. "what" and "where" in the auditory system - an unsupervised learning approach. *COSYNE 2011 Proceedings*, 2011. [cited at p. 114, 128]

[52] Nicol S Harper and David McAlpine. Optimal neural population coding of an auditory spatial cue. *Nature*, 430(7000):682–686, 2004. [cited at p. 43, 96, 102]

[53] Ian A Harrington, G Christopher Stecker, Ewan A Macpherson, and John C Middlebrooks. Spatial sensitivity of neurons in the anterior, posterior, and primary fields of cat auditory cortex. *Hearing research*, 240(1):22–41, 2008. [cited at p. 75]

[54] Hermann LF Helmholtz. *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Cambridge University Press, 2009. [cited at p. 6]

[55] Paul M Hofman and A John Van Opstal. Bayesian reconstruction of sound localization cues from responses to random spectra. *Biological cybernetics*, 86(4):305–316, 2002. [cited at p. 69]

[56] Patrik O Hoyer and Aapo Hyvärinen. Independent component analysis applied to feature extraction from colour and stereo images. *Network: Computation in Neural Systems*, 11(3):191–210, 2000. [cited at p. 111]

[57] Tomáš Hromádka, Michael R DeWeese, and Anthony M Zador. Sparse representation of sounds in the unanesthetized auditory cortex. *PLoS biology*, 6(1):e16, 2008. [cited at p. 41]

[58] Anne Hsu, Sarah MN Woolley, Thane E Fremouw, and Frédéric E Theunissen. Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons. *The Journal of neuroscience*, 24(41):9201–9211, 2004. [cited at p. 42, 93]

[59] David H Hubel and Torsten N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574, 1959. [cited at p. 6]

[60] Jonathan J Hunt, Peter Dayan, and Geoffrey J Goodhill. Sparse coding can predict primary visual cortex receptive field changes induced by abnormal visual input. *PLoS computational biology*, 9(5):e1003005, 2013. [cited at p. 65, 111]

[61] Aapo Hyvärinen and Patrik Hoyer. Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, 2000. [cited at p. 79]

[62] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430, 2000. [cited at p. 25, 26]

[63] Aapo Hyvèarinen, Jarmo Hurri, and Patrick O Hoyer. *Natural Image Statistics*, volume 39. Springer, 2009. [cited at p. 27, 65, 122]

[64] J Yu Jane and Eric D Young. Linear and nonlinear pathways of spectral information transmission in the cochlear nucleus. *Proceedings of the National Academy of Sciences*, 97(22):11780–11786, 2000. [cited at p. 128]

[65] Lloyd A Jeffress. A place theory of sound localization. *Journal of comparative and physiological psychology*, 41(1):35, 1948. [cited at p. 40]

[66] Yitzhak Katznelson. *An introduction to harmonic analysis*. Cambridge University Press, 2004. [cited at p. 113, 127]

[67] Andrew J King and Israel Nelken. Unraveling the principles of auditory cortical processing: can we learn from the visual system? *Nature neuroscience*, 12(6):698–701, 2009. [cited at p. 44]

[68] Andrew J King, Jan WH Schnupp, and Timothy P Doubell. The shape of ears to come: dynamic coding of auditory space. *Trends in cognitive sciences*, 5(6):261–270, 2001. [cited at p. 38, 57, 70]

[69] David J Klein, Peter Konig, and Konrad P Kording. Sparse spectrotemporal coding of sounds. *EURASIP Journal on Applied Signal Processing*, 7:659–667, 2003. [cited at p. 108]

[70] Jerzy Konorski. Integrative activity of the brain. 1967. [cited at p. 6]

[71] Kenneth Kreutz-Delgado, Joseph F Murray, Bhaskar D Rao, Kjersti Engan, Te-Won Lee, and Terrence J Sejnowski. Dictionary learning algorithms for sparse representation. *Neural computation*, 15(2):349–396, 2003. [cited at p. 25]

[72] Abhijit Kulkarni and H Steven Colburn. Role of spectral detail in sound-source localization. *Nature*, 396(6713):747–749, 1998. [cited at p. 129]

[73] Shigeyuki Kuwada and Tom C Yin. Binaural interaction in low-frequency neurons in inferior colliculus of the cat. i. effects of long interaural delays, intensity, and repetition rate on interaural delay function. *Journal of Neurophysiology*, 50(4):981–999, 1983. [cited at p. 72]

[74] Ann B Lee, Kim S Pedersen, and David Mumford. The nonlinear statistics of high-contrast patches in natural images. *International Journal of Computer Vision*, 54(1-3):83–103, 2003. [cited at p. 29]

[75] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y Ng. Efficient sparse coding algorithms. In *Advances in neural information processing systems*, pages 801–808, 2006. [cited at p. 25]

[76] Peter Lennie. The cost of cortical computation. *Current biology*, 13(6):493–497, 2003. [cited at p. 30]

[77] William B Levy and Robert A Baxter. Energy efficient neural codes. *Neural Computation*, 8(3):531–543, 1996. [cited at p. 30]

[78] Michael S Lewicki. Efficient coding of natural sounds. *Nature neuroscience*, 5(4):356–363, 2002. [cited at p. 30, 42, 51, 64, 74, 78, 90, 102, 134]

[79] Michael S Lewicki and Bruno A Olshausen. Probabilistic framework for the adaptation and comparison of image codes. *JOSA A*, 16(7):1587–1601, 1999. [cited at p. 29]

[80] Michael S Lewicki, Bruno A Olshausen, Annemarie Surlykke, and Cynthia F Moss. Scene analysis in the natural environment. *Frontiers in psychology*, 5, 2014. [cited at p. 11, 73]

[81] Michael S Lewicki and Terrence J Sejnowski. Learning overcomplete representations. *Neural computation*, 12(2):337–365, 2000. [cited at p. 23, 25, 29]

[82] Ralph Linsker. An application of the principle of maximum information preservation to linear systems. In *Advances in neural information processing systems*, pages 186–194, 1989. [cited at p. 26]

[83] Timm Lochmann and Sophie Deneve. Neural processing as causal inference. *Current opinion in neurobiology*, 21(5):774–781, 2011. [cited at p. 11, 24]

[84] Stephen G Lomber and Shveta Malhotra. Double dissociation of'what'and'where'processing in auditory cortex. *Nature neuroscience*, 11(5):609–616, 2008. [cited at p. 44, 75, 105]

[85] Christian K Machens, Tim Gollisch, Olga Kolesnikova, and Andreas VM Herz. Testing the efficiency of sensory coding with optimal stimulus ensembles. *Neuron*, 47(3):447–456, 2005. [cited at p. 32]

[86] Julia K Maier, Phillipp Hehrmann, Nicol S Harper, Georg M Klump, Daniel Pressnitzer, and David McAlpine. Adaptive coding is constrained to midline locations in a spatial listening task. *Journal of Neurophysiology*, 108(7):1856–1868, 2012. [cited at p. 43, 71]

[87] David Marr. Vision: A computational investigation into the human representation and processing of visual information, henry holt and co. *Inc., New York, NY*, pages 2–46, 1982. [cited at p. 12]

[88] David McAlpine and Benedikt Grothe. Sound localization and delay lines–do mammals fit the model? *Trends in neurosciences*, 26(7):347–350, 2003. [cited at p. 40, 134]

[89] David McAlpine, Dan Jiang, and Alan R Palmer. A neural code for low-frequency sound localization in mammals. *Nature neuroscience*, 4(4):396–401, 2001. [cited at p. 72]

[90] Josh H McDermott. The cocktail party problem. *Current Biology*, 19(22):R1024–R1027, 2009. [cited at p. 11]

[91] Josh H McDermott and Eero P Simoncelli. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron*, 71(5):926–940, 2011. [cited at p. 42]

[92] John C Middlebrooks. Distributed cortical representation of sound locations. In *Perspectives on Auditory Research*, pages 361–378. Springer, 2014. [cited at p. 76, 99, 133]

[93] John C Middlebrooks, Ann E Clock, Li Xu, and David M Green. A panoramic code for sound location by cortical neurons. *Science*, 264(5160):842–844, 1994. [cited at p. 76, 99, 100, 103, 133]

[94] Lee M Miller, Monty A Escabí, Heather L Read, and Christoph E Schreiner. Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *Journal of Neurophysiology*, 87(1):516–527, 2002. [cited at p. 93, 94, 111, 112, 122, 123, 128]

[95] Lee M Miller and Gregg H Recanzone. Populations of auditory cortical neurons can accurately encode acoustic space across stimulus intensity. *Proceedings of the National Academy of Sciences*, 106(14):5931–5935, 2009. [cited at p. 103]

[96] Wiktor Mlynarski. Sparse, complex-valued representations of natural sounds learned with phase and amplitude continuity priors. *arXiv preprint arXiv:1312.4695*, 2013. [cited at p. 79]

[97] Wiktor Mlynarski. Efficient coding of spectrotemporal binaural sounds leads to emergence of the auditory space representation. *Frontiers in Computational Neuroscience*, 2014. [cited at p. 14, 102]

[98] Wiktor Mlynarski. Sparse coding model of natural stereo sounds reproduces spatial tuning of neurons in the mammalian auditory cortex. *under review*, 2014. [cited at p. 14]

[99] Wiktor Młynarski and Jürgen Jost. Statistics of natural binaural sounds. *PLOS One*, 2014. [cited at p. 14]

[100] Israel Nelken. Processing of complex sounds in the auditory system. *Current opinion in neurobiology*, 18(4):413–417, 2008. [cited at p. 44, 75, 104, 132]

[101] Israel Nelken, Alon Fishbach, Liora Las, Nachum Ulanovsky, and Dina Farkas. Primary auditory cortex of cats: feature detection or something else? *Biological cybernetics*, 89(5):397–406, 2003. [cited at p. 41, 132]

[102] Bruno A Olshausen. Learning linear, sparse, factorial codes. *MIT CBCL Technical Report*, 1996. [cited at p. 25]

[103] Bruno A Olshausen. Perception as an inference problem. *The Cognitive Neurosciences*, 2013. [cited at p. 24]

[104] Bruno A Olshausen et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996. [cited at p. 21, 24, 29, 30]

[105] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997. [cited at p. 23, 24, 28]

[106] Bruno A Olshausen and David J Field. Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4):481–487, 2004. [cited at p. 28, 29]

[107] Bruno A Olshausen and David J Field. What is the other 85% of v1 doing. *Problems in Systems Neuroscience*, 4(5):182–211, 2004. [cited at p. 53]

[108] Cesare V Parise, Katharina Knorre, and Marc O Ernst. Natural auditory scene statistics shapes human spatial hearing. *Proceedings of the National Academy of Sciences*, 111(16):6104–6108, 2014. [cited at p. 43]

[109] Zygmunt Pizlo. Perception viewed as an inverse problem. *Vision Research*, 41(24):3145–3161, 2001. [cited at p. 11]

[110] Tomaso Poggio and Christof Koch. Ill-posed problems in early vision: from computational theory to analogue networks. *Proceedings of the Royal society of London. Series B. Biological sciences*, 226(1244):303–323, 1985. [cited at p. 11]

[111] Carroll C Pratt. The spatial character of high and low tones. *Journal of Experimental Psychology*, 13(3):278, 1930. [cited at p. 44]

[112] Anqi Qiu, Christoph E Schreiner, and Monty A Escabí. Gabor analysis of auditory midbrain receptive fields: spectro-temporal and binaural composition. *Journal of Neurophysiology*, 90(1):456–476, 2003. [cited at p. 40, 111, 128]

[113] R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 2005. [cited at p. 6]

[114] Martin Rehn and Friedrich T Sommer. A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. *Journal of computational neuroscience*, 22(2):135–146, 2007. [cited at p. 30]

[115] Pamela Reinagel and Anthony M Zador. Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems*, 10(4):341–350, 1999. [cited at p. 134]

[116] F Rieke, DA Bodnar, and W Bialek. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 262(1365):259–265, 1995. [cited at p. 42]

[117] Luis Robles and Mario A Ruggero. Mechanics of the mammalian cochlea. *Physiological reviews*, 81(3):1305–1352, 2001. [cited at p. 111, 113, 127]

[118] Suzanne K Roffler and Robert A Butler. Factors that influence the localization of sound in the vertical plane. *The Journal of the Acoustical Society of America*, 43(6):1255–1259, 1968. [cited at p. 44]

[119] Lizabeth M Romanski and Patricia S Goldman-Rakic. An auditory domain in primate prefrontal cortex. *Nature neuroscience*, 5(1):15–16, 2001. [cited at p. 44]

[120] Lizabeth M Romanski, Biao Tian, J Fritz, Mortimer Mishkin, Patricia S Goldman-Rakic, and Josef P Rauschecker. Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nature neuroscience*, 2(12):1131–1136, 1999. [cited at p. 44, 75, 105]

[121] Sam Roweis. Em algorithms for pca and spca. *Advances in neural information processing systems*, pages 626–632, 1998. [cited at p. 111]

[122] Nicole C Rust and J Anthony Movshon. In praise of artifice. *Nature neuroscience*, 8(12):1647–1650, 2005. [cited at p. 33, 132]

[123] Andrew Saxe, Maneesh Bhand, Ritvik Mudur, Bipin Suresh, and Andrew Ng. Unsupervised learning models of primary cortical receptive fields and receptive field plasticity. In *Advances in neural information processing systems*, pages 1971–1979, 2011. [cited at p. 112]

[124] Jan Schnupp, Israel Nelken, and Andrew King. *Auditory neuroscience: Making sense of sound*. MIT Press, 2011. [cited at p. 38, 132]

[125] Jan WH Schnupp, Thomas D Mrsic-Flogel, and Andrew J King. Linear processing of spatial cues in primary auditory cortex. *Nature*, 414(6860):200–204, 2001. [cited at p. 41, 108, 111, 128, 132]

[126] Odelia Schwartz, Jonathan W Pillow, Nicole C Rust, and Eero P Simoncelli. Spike-triggered neural characterization. *Journal of Vision*, 6(4):13, 2006. [cited at p. 10]

[127] Claude E Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27(3):379–423, 1949. [cited at p. 15]

[128] Tatyana Sharpee, Nicole C Rust, and William Bialek. Analyzing neural responses to natural signals: maximally informative dimensions. *Neural computation*, 16(2):223–250, 2004. [cited at p. 10]

[129] Barbara G Shinn-Cunningham, Scott Santarelli, and Norbert Kopco. Tori of confusion: Binaural localization cues for sources within reach of a listener. *The Journal of the Acoustical Society of America*, 107:1627, 2000. [cited at p. 38, 71]

[130] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001. [cited at p. 47]

[131] Nandini C Singh and Frédéric E Theunissen. Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America*, 114:3394, 2003. [cited at p. 42, 87, 93, 112, 123]

[132] Sean J Slee and Eric D Young. Linear processing of interaural level difference underlies spatial tuning in the nucleus of the brachium of the inferior colliculus. *The Journal of Neuroscience*, 33(9):3891–3904, 2013. [cited at p. 128]

[133] Evan C Smith and Michael S Lewicki. Efficient auditory coding. *Nature*, 439(7079):978–982, 2006. [cited at p. 29, 30, 42, 51, 64, 74, 102, 108]

[134] Zachary M Smith, Bertrand Delgutte, and Andrew J Oxenham. Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, 416(6876):87–90, 2002. [cited at p. 111]

[135] Matthew W Spitzer and Malcolm N Semple. Interaural phase coding in auditory midbrain: influence of dynamic stimulus features. *Science*, 254(5032):721–724, 1991. [cited at p. 43]

[136] Matthew W Spitzer and Malcolm N Semple. Transformation of binaural response properties in the ascending auditory pathway: influence of time-varying interaural phase disparity. *Journal of neurophysiology*, 80(6):3062–3076, 1998. [cited at p. 43]

[137] G Christopher Stecker, Ian A Harrington, and John C Middlebrooks. Location coding by opponent neural populations in the auditory cortex. *PLoS biology*, 3(3):e78, 2005. [cited at p. 41, 76, 96, 97, 98, 99, 100, 103, 104, 133]

[138] G Christopher Stecker, Brian J Mickey, Ewan A Macpherson, and John C Middlebrooks. Spatial sensitivity in field paf of cat auditory cortex. *Journal of neurophysiology*, 89(6):2889–2903, 2003. [cited at p. 41, 99, 100, 103]

[139] G Christopher Stecker and John C Middlebrooks. Distributed coding of sound locations in the auditory cortex. *Biological cybernetics*, 89(5):341–349, 2003. [cited at p. 76, 96, 99, 100, 103]

[140] Christian Stilp and Michael Lewicki. Statistical structure of speech sound classes is congruent with cochlear nucleus response properties. *The Journal of the Acoustical Society of America*, 134(5):4229–4229, 2013. [cited at p. 102]

[141] Christian E Stilp and Keith R Kluender. Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility. *Proceedings of the National Academy of Sciences*, 107(27):12387–12392, 2010. [cited at p. 133]

[142] Christian E Stilp, Timothy T Rogers, and Keith R Kluender. Rapid efficient coding of correlated complex acoustic properties. *Proceedings of the national academy of sciences*, 107(50):21914–21919, 2010. [cited at p. 135]

[143] John W Strutt. On our perception of sound direction. *Philosophical Magazine*, 13:214–232, 1907. [cited at p. 37, 84]

[144] Hiroki Terashima and Masato Okada. The topographic unsupervised learning of natural sounds in the auditory cortex. In *Advances in Neural Information Processing Systems 25*, pages 2321–2329, 2012. [cited at p. 80]

[145] Frédéric E Theunissen and Julie E Elie. Neural processing of natural sounds. *Nature Reviews Neuroscience*, 15(6):355–366, 2014. [cited at p. 41, 134]

[146] Daniel J Tollin and Tom CT Yin. Interaural phase and level difference sensitivity in low-frequency neurons in the lateral superior olive. *The Journal of neuroscience*, 25(46):10648–10657, 2005. [cited at p. 71]

[147] Ivana Tosic and Pascal Frossard. Dictionary learning. *Signal Processing Magazine, IEEE*, 28(2):27–38, 2011. [cited at p. 81]

[148] Leslie G Ungerleider and James V Haxby. whatand wherein the human brain. *Current opinion in neurobiology*, 4(2):157–165, 1994. [cited at p. 44]

[149] Richard F Voss and John Clarke. 1/fnoise'in music and speech. *Nature*, 258:317–318, 1975. [cited at p. 53]

[150] Jimmy Wang, , Bruno Olshausen, and Vivienne Ming. A sparse subspace model of higher-level sound structure. *COSYNE Proceedings*, 2008. [cited at p. 79, 80]

[151] O Warfusel. Listen hrtf database. http://recherche.ircam.fr/equipes/salles/listen/index.html. [cited at p. 51, 66, 110]

[152] Benjamin Willmore and David J Tolhurst. Characterizing the sparseness of neural codes. *Network: Computation in Neural Systems*, 12(3):255–270, 2001. [cited at p. 27]

[153] David J Willshaw, O Peter Buneman, and Hugh Christopher Longuet-Higgins. Non-holographic associative memory. *Nature*, 1969. [cited at p. 30]

[154] Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002. [cited at p. 81]

[155] Timothy M Woods, Steve E Lopez, James H Long, Joanne E Rahman, and Gregg H Recanzone. Effects of stimulus azimuth and intensity on the single-neuron activity in the auditory cortex of the alert macaque monkey. *Journal of neurophysiology*, 96(6):3323–3337, 2006. [cited at p. 76, 103]

[156] Mehrdad Yaghoobi, Laurent Daudet, and Mike E Davies. Parametric dictionary design for sparse coding. *Signal Processing, IEEE Transactions on*, 57(12):4800–4810, 2009. [cited at p. 81]

Appendices

Appendix A

Appendix A - Derivations of Gradients for Learning Sparse, Complex-Valued and Hierarchical Models

A.1 First layer - complex-valued basis functions

In this section learning rules i.e. gradients over the first layer linear coefficients and basis functions are derived.

Coefficients gradient

Let us remind that $\hat{x}$ is the reconstruction of the original data vector x using inferred coefficients z and basis functions A:

$$\hat{x}_t = \sum_{i=1}^{n/2} \Re\{z_i^* A_{i,t}\} \tag{A.1}$$

Residue r_t i.e. difference between the data vector and its reconstruction is:

$$r_t = x_t - \hat{x}_t \tag{A.2}$$

Inference of coefficients is equivalent to minimization of the following energy function:

$$E_1(z, x, A) \propto \frac{1}{2\sigma^2} \sum_{t=1}^{T} (\hat{x_t} - x_t)^2 + \lambda \sum_{i=1}^{N} S(a_i) \tag{A.3}$$

In the present work use of function $S(a_i) = a_i$ is equivalent to placing an $L1$ norm penalty on amplitudes $a_i = \|z_i\| = \sqrt{z_i^{\Re 2} + z_i^{\Im 2}}$. The gradient over linear coefficients $z^{\Re}, z^{\Im}$ becomes:

$$\frac{\partial E_z}{\partial z_i^{\mathfrak{S}}} \propto \frac{1}{\sigma^2} \sum_{t=1}^{T} A_{i,t}^{\mathfrak{S}} r_t + \lambda \frac{z_i^{\mathfrak{S}}}{\sqrt{z_i^{\Re 2} + z_i^{\Im 2}}} \tag{A.4}$$

Where $\mathfrak{S} \in \{\Re, \Im\}$ indicates whether the coefficient is real or imaginary.

Basis function gradient

Basis functions are learned by performing a gradient step given inferred z values. The negative log-posterior is given by:

$$E_A = E_{Res} + \gamma E_\phi + \beta E_{Sa} = \frac{1}{2\sigma^2} \left(\sum_{t=1}^{T} r_t^2 \right) + \gamma \sum_{i=1}^{n/2} S_\phi(A_i) + \beta \sum_{i=1}^{n/2} S_a(A_i) \tag{A.5}$$

Functions $S_\phi(A_i)$ and $S_a(A_i)$ are of following forms:

$$S_a(A_i) = \sum_{t>1}^{T} \left(\Delta a_{i,t}^A \right)^2 \tag{A.6}$$

$$S_\phi(A_i) = -\sum_{t>1}^{T} sgn\left(\Delta \phi_{i,t} \right) \left(\Delta \phi_{i,t} \right)^2 \tag{A.7}$$

where $\Delta a_{i,t}^A = a_{i,t}^A - a_{i,t-1}^A$ and $\Delta \phi_{i,t}^A = \phi_{i,t}^A - \phi_{i,t-1}^A$.

Priors defined by S_a and S_ϕ determine temporal phase and amplitude correlations respectively.

Gradient of equation A.5 can be decomposed into three terms:

$$\frac{\partial}{\partial A_{i,t}} E_A = \frac{\partial}{\partial A_{i,t}} E_{Res} + \beta \frac{\partial}{\partial A_{i,t}} E_{Sa} + \gamma \frac{\partial}{\partial A_{i,t}} E_{S\phi} \tag{A.8}$$

representing the reconstruction error term and phase and amplitude priors consecutively. In polar coordinates, for $1 < t < T$ phase prior gradient is:

$$\frac{\partial}{\partial \phi_{i,t}^A} \propto 2\phi_{i,t}^A \Big[sgn\left(\Delta \phi_{i,t+1}^A \right) \phi_{i,t+1}^A - sgn\left(\Delta \phi_{i,t}^A \right) \phi_{i,t}^A \Big] \tag{A.9}$$

For boundary conditions i.e. $t = 1$ and $t = T$, this gradient becomes consecutively:

$$\frac{\partial E_\phi}{\partial \phi_{i,1}^A E_\phi} \propto 2\phi_{i,1}^A sgn\left(\Delta \phi_{i,2}^A \right) \phi_{i,2}^A \tag{A.10}$$

$$\frac{\partial E_\phi}{\partial \phi^A_{i,T}} \propto -2\phi^A_{i,T} sgn\Big(\Delta\phi^A_{i,T}\Big)\phi^A_{i,T} \tag{A.11}$$

In the same way, the amplitude term gradient is defined separately for $1 < t < T$:

$$\frac{\partial E_a}{\partial a^A_{i,t}} \propto 2\Big(\Delta a^A_{i,t} - \Delta a^A_{i,t+1}\Big) \tag{A.12}$$

and separately for the boundary conditions ($t = 1$ and $t = T$):

$$\frac{\partial E_a}{\partial a^A_{i,1}} \propto -2\Delta a^A_{i,2} \tag{A.13}$$

$$\frac{\partial}{\partial a^A_{i,T}} E_a \propto 2\Delta a^A_{i,T} \tag{A.14}$$

The residue term is most conveniently represented in Cartesian coordinates for real and imaginary coefficients $z^{\mathfrak{S}}_i$, where, as previously, $\mathfrak{S} \in \{\mathfrak{R}, \mathfrak{I}\}$, indicates whether coefficient is real or imaginary:

$$\frac{\partial E_{Res}}{\partial A^{\mathfrak{S}}_{i,t}} \propto \frac{z^{\mathfrak{S}}_{i,t}}{\sigma^2} r_t \tag{A.15}$$

A.2 Second layer basis functions

The second layer of the model was trained after the first layer converged, and cofficient values z were inferred for all training data samples. The higher order encoding formed by coefficients s as well as the scaling factor w was inferred via gradient descent on function E_2 (equation 5.15):

$$\begin{aligned}\frac{\partial}{\partial s_i} E_2 \propto -\frac{2}{\sigma_2^2} \sum_{n=1}^{2\times T} B_{i,n}(a_n - \hat{a}_n) + \kappa|w| \sum_{m=1}^{P} \sin(\Delta\phi_m - \widehat{\Delta\phi}_m)\xi_{i,m} \\ + 2\lambda_2 \frac{s_i}{\log(1+s_i^2)}\end{aligned} \tag{A.16}$$

$$\frac{\partial}{\partial w_i} E_2 \propto \kappa \frac{w}{|w|^2} \sum_{m=1}^{P} \widehat{\Delta\phi}_m \sin(\Delta\phi_m - \widehat{\Delta\phi}_m) + \lambda_w \left[\left(\frac{1}{\alpha}\right)^\beta \beta w |w|^{\beta-2}\right] \tag{A.17}$$

The gradients steered sparse coefficients s to explain amplitude and phase vectors a and $\Delta\phi$ while preserving maximal sparsity. Simultaneously the multiplicative factor w was adjusted to appropriately scale the estimated vector $\widehat{\Delta\phi}$.

Finally, learning rules for second-layer dictionaries were given by:

$$\frac{\partial}{\partial B_{i,k}} E_2 \propto -\frac{2}{\sigma_2^2} s_i (a_k - \hat{a}_k) \tag{A.18}$$

$$\frac{\partial}{\partial \xi_{i,k}} E_2 \propto s_i \kappa |w| \sin(\Delta\phi_k - \widehat{\Delta\phi}_k) \tag{A.19}$$

Bibliographische Daten

Functional Sensory Representations of Natural Stimuli: The Case of Spatial Hearing
Młynarski, Wiktor
Universität Leipzig, Dissertation, 2014
154 Seiten, 48 Abbildungen, 156 Referenzen, S-Zahl 3

Selbstständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, den January 23, 2015

.....
(Wiktor Młynarski)

Daten zum Autor

Name:	Wiktor Młynarski
Geburtsdatum:	01.10.1986 in Krakow
2002 - 2005	Liceum Jana III Sobieskiego, Krakow Abschluss Abitur
10/2005 - 06/2010	Studium der Informatik Jagiellonen Universität in Krakow
seit 01/2011	Doktorand am Max-Planck Institut für Mathematik in den Naturwissenschaften .....