Chapter 1

Introduction

Je pense, donc je suis—René Descartes.

The brain is indeed the most complex computational device we have ever had. The brain not only controls the internal states and coordinated movements of our bodies, but also processes sensory signals for appropriately responding to the outside world, and above all, it is the brain that makes us human (or at least, makes us scientists).

This amazing organ, occupying only $\sim$2% of body mass but yet consuming $\sim$20% of total energy, consists of several hundred billion of neurons and even more glial cells in humans (Abeles, 1991; Kandel et al., 2000; Smith, 2000). The former works as a computational unit that sends and receives electro-chemical signals—or "spikes" in essentially a binary manner—between each other, whereas the latter provides physical, functional, and nutritional support for nearby neurons. Therefore neurons are the main players of the brain from the computational viewpoint ("neuron doctrine;" McCulloch and Pitts, 1943; Barlow, 1972). Aside from the plastic properties and developmental aspects, however, the basic performance of neurons themselves is not necessarily superior to that of the state-of-the-art transistors that work as computational units in very large-scale integration (VLSI) systems. In fact, (1) the switching speed of a neuron—or the firing rate—is limited up to $\sim 10^3$ Hz, whereas that of transistors

can be several orders of magnitude faster (e.g., $\sim 10^{15}$ Hz for the leading-edge supercomputers); and (2) the conduction velocity of action potentials is at most $\sim 10^2$ m/sec in myelinated axons, whereas that of electronic circuits can be $\sim 10^8$ m/sec. Considering that the number of constituent units is comparable between the two systems ($\sim 10^{10}$), there is then no reason to imagine that we cannot build an "artificial brain." But the truth is that currently available computers can outperform the nervous system only in some "tedious routines," and in most computations, the brain is second to none at this moment.

Among many problems we confront in our everyday lives, some require huge efforts for the brain to solve, whereas many others do not. (Ironically, the former tasks include the "tedious routines" that an artificial system is good at handling.) It should however be emphasized that the fact that the brain can do effortlessly does not necessarily mean that the underlying computations are readily tractable. One such type of the problems is sensory inference. Taking acoustic signal processing problems for example, the brain can easily separate, localize, and identify the sound sources from a mixture of sounds coming from multiple sources, but no artificial system can do these tasks in general settings.

Section 1.1 will review the organization of mammalian auditory systems, specifically focusing on what computations would take place at each level. Then in Section 1.2, by comparing to other sensory modalities (especially to the visual systems), I will describe some characteristics and challenges of auditory signal processing problems, as well as the motivations of this study.

1.1 Auditory System Organization

The goal of the auditory system is to "feel" rapid changes of local air pressure in a certain range (e.g., the dynamic range of human hearing is 20–$20{,}000$ Hz; Bregman, 1990; Kandel et al., 2000; Smith, 2000), and make sense of the acoustic environment by extracting behaviorally meaningful information such as communication calls. For this purpose, nature has developed

(1) a special mechanical device to receive and transform sounds into an appropriate signal format for the nervous system (i.e., a series of action potentials); (2) faithful ways to transmit such acoustic information to downstream subcortical and cortical stations for further (serial and parallel) processing; and (3) neural circuits to represent and perceive sounds, eventually leading to appropriate motor actions. Below I will briefly overview the ascending auditory pathway—from hair cells at the cochlea up to the (primary) auditory cortex—focusing on the computations and the current working models at each station (for more comprehensive reviews on the anatomical and physiological characters, see e.g., Schreiner et al., 2000; Read et al., 2002; Malmierca, 2003; Malmierca and Irvine, 2005; Winer et al., 2005).

1.1.1 Hair Cells and Auditory Nerve Fibers

The first—and a fundamental—auditory processing is the decomposition of acoustic signals into frequency components (Fletcher, 1940; Hudspeth, 1989, 1997). This transformation from the time domain to appropriate time-frequency representations depends on the mechanical properties and the anatomical organization of the cochlea, where each frequency element leads to the vibration of the basilar membrane only at specific locations, due to resonance effects. Therefore each mechanosensory hair cell on the membrane responds only to a limited range of frequencies—or the frequency bandwidth—and such frequency analysis leads to the tonotopical organization of the cochlea as manifested in the tuning curve properties of the ascending auditory nerve fibers (Kiang et al., 1967; Kiang and Moxon, 1974).

From a computational viewpoint, the cochlea can thus be considered as a bank of bandpass filters. Conventional choices of time-frequency analysis are the short-term Fourier transform (STFT or spectrogram; see Eq.(3.25) on page 102; Cohen, 1995) or a Gammatone filter bank (Irino and Patterson, 2001), but more sophisticated cochlear models have been developed by incorporating psychoacoustic knowledge and physiological hair cell dynamics such as adaptations (Patterson, 1974; Lyon, 1982; Meddis, 1986; Seneff, 1988; Patterson and Holdsworth, 1996). Note that certain acoustic signals will become less overlapped in the frequency domain,

which in turn makes it easier to separate the sources (Bregman, 1990). Also note that, as in many other sensory systems, sound perception generally follows the Weber-Fechner law (Weber, 1846; Fechner, 1860), and thus the perceptual scale of frequencies (or pitches) is often chosen in the logarithmic scales (e.g., in units of octave as in "piano keys" or in units of Mel as measured in psychophysical studies; Stevens et al., 1937); likewise, the sound pressure level (SPL) is also often expressed in the logarithmic scale where 0 dB SPL approximately corresponds to the threshold of human hearing (the sound pressure of ~ 20 μPa or $\sim 10^{-16}$ watt$/$cm^2; Smith, 2000), and approximately related to the perceptual "loudness" by a power law with a coefficient around 0.6 (Stevens, 1956).

1.1.2 Cochlear Nucleus

Cochlear nucleus is the first "relay station" that receives inputs mainly from the ipsilateral auditory nerve fibers and sends outputs to both ipsilateral and contralateral superior olivary nuclei (Kandel et al., 2000; Smith, 2000). The tonotopical organization is preserved in the cochlear nucleus, and neurons in the cochlear nucleus can be classified into several types based on their morphological features and their "response maps," showing areas of excitation and inhibition plotted on the sound-level vs. frequency coordinates (Young, 1984). In addition, spiking patterns can also be classified into several categories on the basis of the shape of peristimulus time histograms (PSTHs; Kiang, 1975). These observations suggest that spectro-temporal, sound levels, and some other forms of coding schemes are already employed at this level (Oertel, 1991), but it remains to be addressed what kind of computations the cochlear nucleus performs.

1.1.3 Superior Olivary Nucleus

Binaural inputs first meet each other at superior olivary nucleus, a group of nuclei in the pons, which receives inputs from (anterior ventral) cochlear nuclei bilaterally and sends outputs to

inferior colliculus through lateral lemniscus (Kandel et al., 2000; Konishi, 2003). It thus plays a very important role in sound localization by exploiting binaural cues, and is in fact one of the best characterized auditory stations from functional viewpoints.

The lateral superior olive exploits the interaural level differences (ILD)—a major cue in localizing high frequency sounds—where some cells are excited by ipsilateral sounds and inhibited by contralateral sounds (E-I cells), leading to faithfully encoding the intensity differences as small as 10 dB SPL. Other cells are responsive to similar variations in the sound intensities at both ears (E-E cells; Caird and Klinke, 1983).

The medial superior olive is involved in detecting the interaural time differences (ITD), which is particularly useful for locating low frequency sounds. It is believed to be the site of the coincidence detection originally proposed by Jeffress (1948), where a spatial array of cells receives inputs from both ipsilateral and contralateral sides but with different "wire length" from the two sides. This causes a certain internal delay for signals to reach, and thus the convergence of the signals from the two ears coincides only when the difference in these latencies matches exactly to the arrival time difference of the sounds between the two ears (see also Joris et al., 1998; Palmer, 2004).

Note that sound sources can be localized using monaural cues, although it is more difficult—and less efficient—than using binaural cues (Bregman, 1990). One such monaural cue is the spectral cue, where the detailed shape of the pinnae, head, and torso acts as a differential filter imposed on a source in a location-dependent manner (head-related transfer function; see also Section 2.1.2). This spectral cue would help determine the sources in vertical as well as horizontal axes, whereas the interaural time or intensity differences would mainly help localize the signals in the azimuthal plane. Although the use of spectral cues has been studied well in (human) psychophysics (Wightman and Kistler, 1989; Bregman, 1990; Hofman and van Opstal, 2002), it is not clear yet where and how they are encoded in the neural circuits and exploited—binaurally or monaurally—for auditory signal processing including sound localization (Knudsen and Konishi, 1979; Wenzel et al., 1993; Carlile et al., 2005).

1.1.4 Inferior Colliculus

Inferior colliculus is located in the midbrain where a tonotopical map is also preserved, and many cells show strong binaural responses and preferences to rather complex stimuli (such as amplitude- or frequency-modulations) but not to steady tones (Ryan and Miller, 1978; Kandel et al., 2000; Smith, 2000). Major ascending auditory pathway converges to this principal midbrain station of the auditory pathway, and thus it would play important roles in auditory scene analysis (Caird and Klinke, 1987; Davis, 2005). For example, inferior colliculus is responsive to interaural delay (Skottun et al., 2001) and may form a topographic as well as tonotopic map of the acoustic environment (FitzPatrick, 1975; Langner et al., 2002). In addition, some neurons in the inferior colliculus show sensitivities to spectral changes, which would help recognize specific phonemes and intonations in speech (Fitch et al., 1997). But it is unknown yet what exactly the inferior colliculus does and how it contributes to auditory scene analysis.

1.1.5 Medial Geniculate Nucleus

Medial geniculate nucleus is the final subcortical station before the auditory signals reach the (primary) auditory cortex (Clarey et al., 1992; de Ribaupierre, 1997; Suga et al., 2002; Jones, 2003; Suga and Ma, 2003). This auditory thalamus consists of many sub-nucleus types, distinguished on the basis of anatomical features and coding properties (Calford and Webster, 1981; Calford, 1983; Winer, 1985; Winer et al., 1999; He and Hu, 2002; Read et al., 2004). A tonotopical map is organized as in the inferior colliculus (Aitkin and Webster, 1971), and some cells show binaural responses whereas others show monaural responses, mainly to inputs from the contralateral side (Altman et al., 1970a,b; Cetas et al., 2002). Cells in the medial geniculate nucleus show a wide variety of response types; both broadly and narrowly tuned cells are observed (Aitkin et al., 1966; He and Hu, 2002), and some cells respond only to complex stimuli whereas others even show multi-modal responses, receiving visual and somatosensory inputs as well as auditory signals (Bordi and LeDoux, 1994a,b; Komura et al., 2005). Note

however that virtually nothing is known about computational aspects of this auditory signal processing station, even though extensive physiological research has been conducted over the past decades (for the earliest electophysiological studies, see e.g., Rose and Galambos, 1952; Galambos et al., 1952; Galambos, 1952).

1.1.6 Primary Auditory Cortex

Primary auditory cortex is the first cortical station that receives inputs from the auditory thalamus and further processes the acoustic signals to make sense of them (for review, see e.g., Ehret, 1997; Schreiner et al., 2000; Read et al., 2002; Winer et al., 2005; King and Schnupp, 2007; for review on auditory processing in higher cortices, see e.g., Kaas et al., 1999; Griffiths and Warren, 2004; Griffiths et al., 2004). Representations of the signals would be most likely in the time-frequency domain, as suggested by the tonotopic organization (Merzenich et al., 1973, 1975), but auditory cortical neurons have a wide variety in receptive field sizes and show highly heterogeneous response patterns (Hromádka, 2007), and it is currently unclear what acoustic features these neurons respond to, or what stimulus would be optimal for exciting auditory neurons (but see approaches in deCharms et al., 1998; Machens et al., 2005; O'Connor et al., 2005; and also Sections 2.4.1 and 2.A).

Note however that there is no consensus on the definition of the "auditory cortex" because many functional subregions seem to exist. Among many areas—more than 10 divisions in cats as well as in monkeys—seemingly involved in the auditory signal processing, several "primary" areas have a tonotopic map (the primary auditory cortex, the anterior auditory field, and the posterior, ventral, and ventroposterior auditory areas; Phillips and Irvine, 1981; Schreiner and Mendelson, 1990; Schreiner et al., 1992; Mendelson et al., 1993, 1997; Eggermont, 1998), suggesting their roles in spectro-temporal—and binaural—analysis based on the information directly inherited from subcortical areas (Miller et al., 2001). In contrast, some non-tonotopic areas (the secondary auditory cortex and the suprasylvian fringe) typically have broader tuning curves and longer durations, and are thought to be involved in process-

ing communication signals and non-spectral stimuli (Schreiner and Cynader, 1984; He et al., 1997; Rauschecker and Tian, 2000), whereas others (the limbic-related fields and the posterior ectosylvian fields) receive not only auditory but also visceral and visual inputs, and would process multi-modal information (Kelly, 1973, 1974; Colavita et al., 1974; Bowman and Olson, 1988a,b; Kayser et al., 2005). Note that the rodent auditory cortex seems to be organized in a slightly simpler manner (Kelly and Sally, 1988; Sally and Kelly, 1988; Doron et al., 2002; Polley et al., 2007), and rats were used as experimental animals in this thesis (Chapter 3).

With respect to computations, the primary auditory cortex has traditionally been considered as stimulus feature—or, spectro-temporal "edge" (Fishbach et al., 2001, 2003)—detectors (Nelken et al., 2003), and linear models have been widely used to characterize the response patterns in the auditory cortex (e.g., Eggermont et al., 1983; Kowalski et al., 1996; Klein et al., 2000; Depireux et al., 2001; Theunissen et al., 2001; Escabí and Schreiner, 2002; Linden et al., 2003; Machens et al., 2004). However, such oversimplified models do not fully capture the computational properties of auditory cortical processing in general (see also Chapter 3).

1.2 Characteristics of Auditory Signal Processing

Computational problems faced by different sensory modalities share many characteristics; e.g., signals are processed by neural circuits in units of spikes (McCulloch and Pitts, 1943; Barlow, 1972), be they visual, tactile/somatosensory, or auditory signals. In addition, the fact that the auditory cortex is able to process and "see" visual information after rewiring visual projections from the retina to the ascending auditory pathway (Sharma et al., 2000) suggests that the principles underlying the computations would most likely be the same; i.e., complex input signals are transformed into rather simple context-invariant features, which in turn collectively form biologically meaningful representations and interpretable objects (for the auditory system, see e.g., Nelken et al., 2003; Nelken, 2004; for the visual system, see e.g., Carandini et al., 2005; Rust and Movshon, 2005). This common framework then gives *a* justification of

the conventional systems neuroscience approaches—even though we often treat the brain as a "black box" that transforms input signals into output behavioral responses, the signals must be transformed in such a way that the representations become somehow more easily interpreted (or *decoded*) by us experimenters (or by the "homunculus;" Penfield and Rasmussen, 1950) as the signals are transmitted to "higher" processing stages (e.g., Bialek et al., 1991; Rieke et al., 1997; deCharms and Zador, 2000; Dayan and Abbott, 2001; Shamma, 2001; Pouget et al., 2000, 2003).

The difficulties in such sensory signal processing problems reside in; (1) animals are bombarded with continuously changing high-dimensional inputs, and thus must deal with the "curse of dimensionality" (Bellman, 1961) and the "frame problems" (McCarthy and Hayes, 1969); and (2) associated computations to extract meaningful information are often ill-posed, and thus appropriate constraints must be imposed to guarantee the compatibility, uniqueness, and continuity of the solutions (Marr, 1982). To decipher how the brain performs sensory inferences, as Marr (1982) pointed out in his seminal work, (at least) three levels of the understandings would be required;

(1) computational theory to identify the logics and goals of computations;

(2) representation and algorithm to achieve the computations and appropriate transformations of input signals; and

(3) hardware and implementation to physically realize the algorithm in neural circuits.

In this thesis, on the one hand, the theoretical approach works on the middle level and Chapter 2 demonstrates how auditory streams can be segregated by exploiting the idea of sparse overcomplete representations as a working principle. On the other hand, the experimental approach works on the bottom level, trying to characterize the properties of neural dynamics in the auditory cortex for building a plausible encoding model at the single-cell level (Chapter 3).

Since Marr (1982) formulated the basic framework of theoretical analysis on sensory signal processing problems, theoretical and computational research on the visual systems has

been expanded and leading ahead of the analysis on any other modality, including the auditory systems. This would be because (1) we humans are visually-oriented animals; (2) the visual systems research has attracted more scientists; and (3) we have accumulated more physiological and psychological evidences on the visual systems for historical reasons (Palmer, 1999; Kandel et al., 2000; Smith, 2000). Consequently, our current knowledge on the sensory signal processing in the brain comes mainly from the extensive research on the visual system, which potentially results in a biased view on the underlying mechanisms. We thus have to be aware that each sensory system has its own specific problems, even though the strategy or the general principles could be the same among all modalities in a broad sense (see also Chapter 4).

The auditory signal processing problems are distinct at least in the following three respects. First, an acoustic environment we encounter often consists of sounds from a rich combination of sounds, and behaviorally relevant information can be masked by the many irrelevant sounds—or background noise—that may even constitute the majority of the acoustic energy received. Therefore, the auditory system must *segregate* such overlapped signals in time into individual streams, as opposed to the visual system that must *infer* occluded parts of superimposed objects in space.

Second, sensory processing would be facilitated by transforming signals into an appropriate feature domain—e.g., time-frequency representations for acoustic signals as is most likely achieved at the cochlea (Section 1.1.1)—but distinct neural circuit structures at the subcortical level suggest that each sensory modality has its own ways for pre-processing the signal to achieve a seemingly common goal; i.e., based on the information from the receptor surface activity, subcortical stations compute several aspects of the received signals that are (1) in some sense optimized for representing and analyzing the environment, and (2) readily used for subsequent processing in the cortex to make sense of the external world. For example, the fact that the auditory system has many stations at the brainstem would reflect the computational complexity of the pre-processing and a large size of feature dimensions required to extract before reaching the auditory cortex—e.g., interaural time, level, and spectral differences are exploited

at the subcortical level for sound source localization (Joris et al., 1998; Konishi, 2003; Carlile et al., 2005; see also Section 1.1.3). In contrast, the many receptor types in the somatosensory system (Kandel et al., 2000; Smith, 2000) and the complex local retinal processing schemes in vision (Masland, 2001) seem to accomplish the task of generating sufficient feature dimensions without equally extensive processing at the brainstem itself.

The third feature of auditory signal processing is that sound pressure waves evolve in time and thus there is (almost) no information on "instantaneous" signals, suggesting that the temporal integrations of the received signals are required. What is more, auditory signals inevitably disappear shortly due to their physical properties, and thus precise and quick processing would be needed to extract meaningful information out of such transient signals. In contrast, visual objects are often robust in time, and thus spatial processing would be more critical than temporal processing for the visual systems, even though temporal information is also important to perform some visual tasks (e.g., motion detection).

These characteristics of auditory signal processing in fact underlie the motivations of this thesis. The theoretical work in Chapter 2 is motivated foremost by the computational demands of source separation—or, how to extract features from sound ensembles and use them for computations—and the experimental approach in Chapter 3 aims at deciphering the temporal dynamics of auditory cortical neurons especially because psychophysical studies show that stimulus integration over time is critical for acoustic signal processing (Bregman, 1990); for details of the specific motivations, see the introdcutory sections in each chapter. Although in this thesis I have concentrated on a restricted form of the challenges in the auditory signal processing problems and the two approaches hardly met each other yet, I believe that such complementary approaches are required for better understanding computations in the brain; see Chapter 4 for more general discussions on this topic and future challenges.