Biological Signal Processing

At first it may seem a bit unusual to find a chapter on *biological* signal processing in a book dedicated to *digital* signal processing; yet this is in reality no more peculiar than motivating DSP by starting with the analogous principles of *analog* signal processing. Indeed the biological motivation should be somewhat closer to our hearts (or eyes, ears and brains). In this book we have chosen to introduce analog and digital signal processing together, but have confined our discussion of biological signal processing to this chapter.

In the first two sections we examine how we map external signal parameters into internal (biological/psychological) representations. This question belongs to the realm of psychophysics, the birth of which we describe. Our senses are highly sensitive and yet have a remarkably large dynamic range; we would like to understand and emulate this ability. We will see that a form of universal compression is employed, one that is useful in many DSP contexts.

The majority of the signals we acquire from the outside world and process in our brains are visual, and much interesting signal processing takes place in our visual system. Much has been discovered about the functioning of this system but here we concentrate on audio biological mechanisms since the focus of this book is one-dimensional signals. Hearing is the sense with the second largest bandwidth, and speech is our primary method of communications. We will devote a section each to speech production and perception mechanisms. In a later chapter we will study a DSP model of speech production that is based on this simplified biology.

After studying the signal input and output mechanisms we proceed to the processing apparatus, namely the brain. We discuss the basic processor, the neuron, and compare its architecture with that of processors with which we are more familiar. We introduce a simple model neuron and the concept of a neural network, and conclude with a performance comparison of man vs. machine.

11.1 Weber's Discovery

Ernst Weber was professor of physiology and anatomy at the university of Leipzig in the first half of the nineteenth century. His investigations involved the sensitivity of the senses. His initial studies dealt with the tactile sense, for example, the effect of temperature, pressure and location on the sense of touch. One of his discoveries was that cold objects felt subjectively heavier than hot objects of the same weight.

In his laboratory Weber would study the effect of different stimuli on human subjects. In order to measure subjective sensitivity he invented the idea of the Just Noticeable Difference (JND), which is the minimal change in the physical world that produces a noticeable difference to the subject's senses. For example, he studied the minimal separation required between two points of contact with the skin, in order to be noticeable. He found that this varied widely, with large separations required on the back while very small separations could be distinguished on the fingertips. From this he could infer the relative densities of neural coverage.

In order to study the subjective feeling of weight he defined the JND to be the minimal weight that must be added in order for a subject to perceive them as different. In a typical experiment (from about 1830) a subject would be given two bags of coins to hold, one placed on each hand. Let's assume that there were 29 coins on the left hand and 30 coins on the right. If most subjects could reliably report the right-hand bag as heavier than the left, Weber would be able to conclude that the threshold was equal or less than the weight of a single coin.

Weber's most important discovery that the JND varied with total weight. Adding a single coin to 29 coins produced a discernible difference, but 59 coins were indistinguishable from 58. Albeit subjects could reliably and repeatably distinguish between 58 and 60 coins. Likewise, most subjects could not reliably feel the difference between 116 coins in one hand and 118 or 119 in the other, only the addition of 4 coins caused a reliably distinguishable effect. Thus the JND definitely increased with increasing total weight.

Upon closer examination Weber noticed something even more significant. The threshold was a single coin when the total weight was that of 29 coins, two coins for 58, 4 coins for 116. The conclusion was obvious—the ratios (1:29, 2:58, 4:116) were all the same. Weber stated this result as 'the sensitivity of a subject to weight is in direct proportion to the weight itself', which translated into mathematics looks like this. This means that in order for a change in weight to be noticeable, one has to add a specific *percentage* of the present weight, *not* an *absolute* weight value.

This radically changed the way Weber understood the JND. He set out to check the dependence of other sensitivity thresholds on total stimulus intensity and found similar relationships.

$$\Delta I = K_I I \tag{11.1}$$

In each case the ratio K_I , called Weber's constant, was different, but the linear dependence of the JND on total stimulus was universal.

Although this relationship surprised Weber it really is quite familiar. Have you ever lain awake in the middle of the night and heard the ticking of a clock or the barking of a distant dog? These sounds are not heard during the day when the ambient noise is higher, but seem quite loud at night when the total stimulus is low. Yet they *must* be there during the day even if not discernible. It is simply that the addition of the ticking or distant barking to the other sounds does not increase the total sound by a sufficient percentage.

You get out of bed and open the window. You remember how the stars were so bright when you were a child, yet seem so dim now. During the day you can't see them at all. Yet they *must* be there during the day even if not discernible. It is simply that with no light from the sun the starlight is a more significant fraction of the total stimulus. With the expansion of cities and the resulting 'light pollution' the stars are disappearing, and one has to go further and further out into the countryside in order to see them. You close the window and strike a match in the dark room. The entire room seems to light up, yet had you struck the same match during the day no change in illumination would have been noticed.

Let's now consider the sequence of physical values that are perceivably different. Think of turning on the radio and slowly increasing the volume until you just begin to hear something. You then turn a bit more until you notice that the sound has definitely grown louder. Continuing this way we can mark the points on the volume control where the sound has become noticeably louder. A direct application of Weber's law tells us that these marks will not be evenly spaced.

Assume for the purpose of argument that the particular stimulus we are studying just becomes detectable at one physical unit $I_0 = 1$ and that Weber's constant for this stimulus is a whopping 100%. Then the second distinguishable level will be $I_1 = 2$ because any value of I that adds less than one unit is indistinguishable from I_0 . Continuing, we must now add $K_I I = 2$ units to the existing two in order to obtain the third distinguishable level $I_3 = 4$. It is easy to see that $I_l = 2^l$, i.e., that the levels of Just Noticeable Differences (JNDs) form a geometric progression. Similarly, the distinguishable intensity levels for a stimulus that just becomes detectable at I_0 physical units, and for which Weber's constant is K_I , obey

$$I_l = I_0 (1 + K_I)^l \tag{11.2}$$

which is an alternative statement of Weber's law.

Weber's law, equation (11.1) or (11.2), has been found to hold, at least approximately, for hundreds of different stimuli. Scientists have measured the required increase in the length of lines, the amount of salt that must be added to soup, and even the extra potency perfume requires. At extremely low and high stimuli there *are* deviations from Weber's law, but over most of the range the linear relationship between threshold and stimulus holds astonishingly well.

EXERCISES

- 11.1.1 Try Weber's coin experiment. Can you measure Weber's constant?
- 11.1.2 Write a computer program that presents a random rectangle on one part of the graphics screen, and allows subjects to reproduce it as closely as possible somewhere else on the screen. What is K here?
- 11.1.3 Allow a subject to listen for a few seconds to a pure sinusoid of constant frequency and then attempt to adjust the frequency of a sinusoid to match it. What is K here? Repeat the experiment with amplitude instead of frequency.
- 11.1.4 Patterns of dots can be hidden by randomly placing large numbers of dots around them. The original pictures stand out if the dots are of different color or size, are made to slowly move, etc. Devise an experiment to determine different people's thresholds for detecting patterns in random dot pictures.

11.2 The Birth of Psychophysics

Psychophysics is precisely what its name implies, the subject that combines psychology and physics. At first, such a combination sounds ridiculous, how could there possibly be any relationship between *physics*, the queen of the rationalistic empirical sciences, and *psychology*, the most subjective and hard to predict study? On second thought scientists learn everything they know by observing the world with their senses. So even scientists are completely dependent on the subjective in order to arrive at the objective.

The English philosopher Berkeley was fond of saying 'esse est percipi', that is, 'existence is being perceived'. We have all heard the famous conundrum about a tree falling in a forest not making a sound if there is no one around to hear it. A physical signal that is not captured by our senses might as well not exist. This capturing of physical signals and their translation into internal representations is called *perception*.

The connection between physical signals and psychological manifestations is by no means simple. The cover of this book looks the same in direct sunlight, under a fluorescent lamp, and by the light of a candle. Your mother's voice sounds the same outside, in a train car, and over the phone. Your friend seems the same height when he is standing close to you, when he has walked across the street, and even on television. In all these cases the physical signals varied widely but the internal psychological representation remained the same. Our perception of quite different physical phenomena may be the nearly the same.

Is it possible to say anything quantitative about internal psychological representations? Can feelings be measured? Surely our perceptions and thoughts are personal and unobservable to the outside world. How then can we talk about representing them quantitatively? Although consideration of such questions has convinced many sages to completely reject psychophysics, these very same questions can be raised regarding much of modern science. We cannot directly observe quarks, electrons, protons, or even atoms, but we become convinced of their existence by indirectly perceiving their effects. Individual cells cannot be seen, but biologists are convinced of their existence. We cannot hold the Milky way galaxy in our hand, yet astronomers have deduced its existence. Feelings may not be openly witnessed, but their existence may be inferred from psychophysical experiments.

Notwithstanding the importance and wide applicability of Weber's law, it is not a true psychophysical law. Psychophysical laws should relate external physical signals to internal psychological representations. Weber's law relates the intensity threshold ΔI to the total stimulus I, both of which are physical entities. Yet another step is needed to make a true psychophysical law.

The first direct attempt to quantitatively pin down feelings was made by one of Weber's students, Gustav Theodor Fechner. Fechner initially studied medicine, but after graduation was more involved in physics. Weber's discoveries retriggered his interest in psychophysics. Fechner started studying color perception, and later performed a series of experiments on the persistence of color after a bright light has been removed. One series of experiments involved viewing sunlight filtered through colored lenses. Fechner, who acted as his own subject, was tragically blinded from the prolonged exposure to direct sunlight. Without his eyesight his promising scientific career was finished. Fechner became depressed and took up the study of philosophy, religion, and mysticism. His main interest was in the so-called 'body and mind' problem. Unlike many of his contemporaries, Fechner believed that the external physical world and the world as viewed internally by the mind were two aspects of one entity.

Then, in 1850, his eyesight miraculously returned. Fechner was convinced that this was a sign that he was to complete the solution to the body and mind problem once and for all. His unique background, combining medicine, physics, and philosophy, allowed him to make a mental leap that his contemporaries were not able or willing to achieve. The solution came to him in what is called a 'Eureka experience' while lying in bed on the morning of October 22, 1850. The anniversaries of this day are celebrated the world over as 'Fechner day'.

Fechner's solution was made up of two parts, a physical part and a psychological part. For the physical part Fechner assumed that Weber's law was correct, namely that equation (11.2) regarding the geometric progression of JND levels holds. For the psychological part Fechner made the simple assumption that all just noticeable changes were somehow equivalent. When we feel that the music has become noticeably louder, or that the light has become brighter, or the soup just a little saltier, or the joke just noticeably funnier, these all indicate an internal change of one unit.

Fechner invented three different methods of experimentally determining the connection between physical and psychological variables. We will demonstrate one by considering a scientist sitting on a mountaintop waiting for the sun to rise. The scientist has brought along nothing save a light meter (which measures physical units I) and a pair of eyes (which register psychological units Y). Sometime before the scientist notices anything happening the light meter shows an increase in the illumination. Suddenly the scientist perceives the light and records that Y = 0 corresponds to the physical reading I_0 . When the light becomes just noticeably brighter the scientist records that Y = 1 corresponds to $I_1 = I_0(1 + K_I)$. The next event is recorded as Y = 2, which corresponds to $I_2 = I_0(1 + K_I)^2$. In general we see that the scientist's personal feeling of Y corresponds to a physical reading of $I_p = I_0(1+K_I)^Y$. We are more interested in knowing the converse connection—given the physical event of intensity I, what is the psychological intensity Y? It is easy to show that

$$Y = A \log I + B \tag{11.3}$$

i.e., that apart from an additive constant that derives from the minimum biological sensitivity, the psychological intensity is proportional to the logarithm of the physical intensity.

We know that the logarithm is an extremely compressive function. A logarithmic psychophysical connection would explain the fantastic ranges that our senses can handle. Under proper conditions we can hear a sound that corresponds to our ear drum moving less than the diameter of a hydrogen atom, and we can actually see single photons. Yet we can also tolerate the sound of a jet engine corresponding to 10^{12} times the minimum intensity and see (for short periods of time as Fechner learned) direct sunlight 15 orders of magnitude stronger. In order to quantitatively compare two signals that may differ by such large amounts we introduce the *Bel* (named after Alexander Graham), defined as the base 10 logarithm of the ratio of the powers of the two signals. In other words, if the power of the second signal is greater than that of the first by a factor of ten, we say that it is one Bel (1 B) stronger. In turns out that the Bel is a bit too large a unit for most purposes, and so we usually use the decibel (dB), which is ten times smaller.

$$d(dB) = 10\log\frac{P_1}{P_2}$$
(11.4)

Since power is the integral of the square of the signal values, if we know RMS signal values we can directly compute the difference between two signals.

$$d(dB) = 20\log\frac{S_1}{S_2}$$
(11.5)

The JND for strong sounds is about 0.5 dB, while at the threshold of hearing about 3 dB is needed.

An audio signal's amplitude is not the only characteristic that is perceived approximately logarithmically. Humans can hear from about 20 Hz (lower than that is felt rather than heard) to over 20 KHz (the precise upper limit depending on age). This corresponds to about 10 octaves, each octave being a doubling of frequency. Sinusoids separated by whole octaves sound similar to us, this fact being the principle behind the musical scale. Inside each octave the conventional western ('well-tempered') division is into twelve chromatic keys, each having frequency $\sqrt[12]{2}$ higher than the previous one. These keys sound to us approximately equally spaced, pointing once again to a logarithmic perception scale. The *mel* (from 'melody') frequency scale is designed to correspond to the subjective psychophysical sensation of a tone's pitch. The perceived pitch of a 1 KHz tone at 40 dB above the hearing threshold is defined to be 1000 mels. Equal mel intervals correspond to equal pitch perception differences; under about 1 KHz the mel scale is approximately linear in frequency, but at higher frequencies it is approximately logarithmic.

 $M \approx 1000 \log_2(f_{KHz} + 1)$

The Bark (named after the acoustician H.G. Barkhausen) scale approximates the natural frequency scale of the auditory system. Psychophysically, signals heard simultaneously are perceived as separate sounds when separated by one Bark or more since they excite different basilar membrane regions. A Bark is about 100 Hz for frequencies under 500 Hz, is about 150 Hz at 1 KHz, and a full KHz at about 5 KHz.

$$1 \operatorname{Bark}_{Hz} \approx 25 + 75(1 + 1.4 f_{\kappa Hz}^2)^{0.69}$$

If we divide the entire audio range into nonoverlapping regions of one Bark bandwidth we get 24 'critical bands'. Both the mel and Bark scales are approximately logarithmic in frequency.

EXERCISES

- 11.2.1 Derive equation (11.3). What is the meaning of A and B? What should be the base of the logarithm?
- 11.2.2 How long does a tone have to be on for its frequency to be identifiable? Experiment!
- 11.2.3 The well-tempered scale is a relatively recent invention, having become popular with the invention of keyboard-based instruments such as the piano. Using a computer with a programmable sound generator, test the difference between a linearly divided scale and a well-tempered one. Play a series of notes each higher than the previous one by 50 Hz. Do the differences sound to same? Play a simple tune on the well-tempered scale and on a linearly divided octave scale. Can you hear the difference? Can you describe it?
- 11.2.4 Since we perceive sound amplitudes logarithmically, we should quantize them on a logarithmic scale as well. Compare the μ -law and A-law quantizations prevalent in the public telephone system (equations (19.3) and (19.4)) with logarithmic response. How are negative values handled? Can you guess why these particular forms are used?

11.2.5 Two approximations to the Bark warping of frequency are

$$B \approx 13 \tan^{-1}(0.76 f_{KHz}) + 3.5 \tan^{-1}\left(\frac{f_{KHz}}{7.5}\right)^2$$
$$\approx 7 \sinh^{-1}(f_{KHz}/0.65)$$

while the Mel warping was given in the text. Compare these three empirical formulas with true logarithmic behavior $\alpha \ln(1+x)$ in the range from 50 Hz to 5 KHz.

- 11.2.6 Recent research has shown that Fechner's law is only correct over a certain range, failing when the stimuli are either very weak or very strong. Stevens proposed a power law $Y = kI^n$ where k and n are parameters dependent on the sense being described. Research Stevens' law. For what cases does Stevens' law fit the empirical data better than Fechner's law?
- 11.2.7 Toward the end of his life Fechner studied æsthetically pleasing shapes. Write a program that allows the user to vary the ratio of the sides of a rectangle and allow a large number of people to find the 'nicest' rectangle that is not a square. What ratio do people like? (This ratio has been employed in architecture since the Greeks.)

11.3 Speech Production

In this section we introduce the biological generation mechanism for one of the most important signals we process, namely human speech. We give a quick overview of how we use our lungs, throats, and mouths to produce speech signals. The next section will describe speech perception, i.e., how we use our ears, cochlea, and auditory nerves to detect speech.

It is a curious fact that although we can input and process much more visual information than acoustic, the main mode of communications between humans is speech. Wouldn't it have been more efficient for us to communicate via some elaborate sign language or perhaps by creating rapidly changing color patterns on our skin? Apparently the main reason for our preferring acoustic waves is their long wavelengths and thus their diffraction around obstacles. We can broadcast our speech to many people in different places; we can hear someone talking without looking at the mouth and indeed without even being in the same room. These advantages are so great that we are willing to give up bandwidth for them; and speech is so crucial to the human race that we are even willing to risk our lives for it. To understand this risk we have to compare our mouth and throat regions with those of the other primates. Comparing the profile of a human with that of a chimpanzee reveals that the chimpanzee's muzzle protrudes much further, while the human has a longer pharynx (throat) and a lower larynx (voice box). These changes make it easy for the human to change the resonances of the vocal cavity, but at the expense of causing the respiratory and alimentary tracts to overlap. Thus food can 'go down the wrong way', impeding breathing and possibly even leading to death by choking. However, despite this importance of spoken communication, the speech generation mechanism is still basically an adapted breathing and eating apparatus, and the speech acquisition mechanism is still essentially the acoustic predator/prey detection apparatus.

It is convenient to think of speech as being composed of a sequence of basic units called *phonemes*. A phoneme is supposed to be the smallest unit of speech that has independent meaning, and thus can be operationally defined as the minimal amount of speech that if replaced could change the meaning of what has been said. Thus **b** and **k** are distinct phonemes in English (e.g., 'book' and 'cook' have different meanings), while **l** and **r** are indistinguishable to speakers of many oriental languages, **b** and **p** are the same in Arabic, and various gutturals and clicks are not recognized by speakers of Latin-based languages. English speakers replace the French or Spanish **r** with their own because the originals do not exist in English and are thus not properly distinguished. Different sources claim that there are between 42 and 64 phonemes in spoken English, with other languages having typically between 25 and 100. Although the concept of a phoneme is an approximation to the whole story, we will posit speech generation and perception to be the production and detection of sequences of phonemes.

Speech generation commences with air being exhaled from the lungs through the 'trachea' (windpipe) to the 'larynx' (voice box). The 'vocal cords' are situated in the larynx. While simply breathing these folds of tissue are held open and air passes through them unimpeded, but when the laryngeal muscles stretch them taut air must pass through the narrow opening between the cords known as the 'glottis'. The air flow is interrupted by the opening and closing of the glottis, producing a periodic series of pulses, the basic pulse rate being between 2.5 and 20 milliseconds. The frequency corresponding to this pulse interval is called the *pitch*. The tighter the cords are stretched, the faster the cycle of opening the cords, releasing the air, and reclosing, and so the higher the pitch. Voice intensities result from the pressure with which the expelled air is forced through the vocal cords. The roughly triangular-shaped pulses of air then pass into the *vocal tract* consisting of the 'pharynx' (throat), mouth cavity, tongue, lips, and nose and are finally expelled. There are two door-like mechanisms that prohibit or allow air to flow. Two passages proceed downward from the pharynx, the 'esophagus' (food pipe) and trachea. The 'epiglottis' separates the two by closing the air passage during swallowing. In addition, air can enter the nasal tract only when the 'velum' is open.

The air pulses exiting the vibrating vocal cords can be thought of as a signal with a basic periodicity of between 50 and 400 Hz (typically 50–250 for men, 150–400 for women) but rich in harmonics. Thus the spectrum of this signal consists of a set of equally spaced lines, typically decreasing in amplitude between 6 and 12 dB per octave. Because of its physical dimensions, the vocal tract resonates at various frequencies called formants, corresponding to the length of the throat (between 200 and 800 Hz), length of the nasal passage (500–1500 Hz), and size of the mouth between throat and teeth (1000–3000 Hz). These resonances enhance applicable frequencies in the glottal signal, in the manner of a set of filters. The result is the complex waveform that carries the speech information. The spectrum thus consists of a set of lines at harmonics of the pitch frequency, with amplitudes dependent on the phoneme being spoken.

The vocal cords do not vibrate for all speech sounds. We call phonemes for which they vibrate *voiced* while the others are *unvoiced*. Vowels (e.g., \mathbf{a} , \mathbf{e} , \mathbf{i} , \mathbf{o} , \mathbf{u}) are always voiced unless spoken in a whisper, while some consonants are voiced while others are not. You can tell when a sound is voiced by placing your fingers on your larynx and feeling the vibration. For example, the sound \mathbf{s} is unvoiced while the sound \mathbf{z} is voiced. The vocal tract is the same in both cases, and thus the formant frequencies are identical, but \mathbf{z} has a pitch frequency while \mathbf{s} doesn't. Similarly the sounds \mathbf{t} and \mathbf{d} share vocal tract positions and hence formants, but the former is unvoiced and the latter voiced. When there is no voicing the excitation of the vocal tract is created by restricting the air flow at some point. Such an excitation is noise-like, and hence the spectrum of unvoiced sounds is continuous rather than discrete. The filtering of a noise-like signal by vocal tract resonances results in a continuous spectrum with peaks at the formant frequencies.

The unvoiced fricatives \mathbf{f} , \mathbf{s} , and \mathbf{h} are good examples of this; \mathbf{f} is generated by constricting the air flow between the teeth and lip, \mathbf{s} by constricting the air flow between the tongue and back of the teeth, and \mathbf{h} results from a glottal constriction. The \mathbf{h} spectrum contains all formants since the excitation is at the beginning of the vocal tract, while other fricatives only excite part of the tract and thus do not exhibit all the formants. Nasal phonemes, such as \mathbf{m} and \mathbf{n} , are generated by closing the mouth and forcing voiced excitation through the nose. They are weaker than the vowels because the nasal tract is smaller in cross sectional area than the mouth. The closed mouth also results in a spectral zero, but this is not well detected by the human speech recognition apparatus. Glides and liquids, such as \mathbf{w} and \mathbf{l} , are also voiced but weaker than vowels, this time because the vocal tract is more closed than for vowels. They also tend to be shorter in duration than vowels. Stops, such as \mathbf{b} and \mathbf{t} , may be voiced or unvoiced, and are created by first completely blocking the vocal tract and then suddenly opening it. Thus recognition of stops requires observing the signal in the time domain.

We have seen that all phonemes, and thus all speech, can be created by using a relatively small number of basic building blocks. We need to create an excitation signal, either voiced or unvoiced, and to filter this signal in order to create formants. In 1791, Wolfgang von Kempelen described a mechanical mechanism that could produce speech in this fashion, and Charles Wheatstone built such a device in the early 1800s. A bellows represented the lungs, a vibrating reed simulated the vocal cords, and leather pipes performed as mouth and nasal passages. By placing and removing the reed, varying the cross-sectional area of the pipes, constricting it in various places, blocking it and releasing, etc., Wheatstone was able to create intelligible short sentences. Bell Labs demonstrated an electronic synthesizer at the 1939 World's Fair in New York. Modern speech synthesizers are electronic and computerized, digitally creating the excitation and filtering using methods of DSP. We will return to this subject in Section 19.1.

EXERCISES

- 11.3.1 What are the main differences between normal speaking on the one hand and whispering, singing, and shouting on the other?
- 11.3.2 Why do some boys' voices change during adolescence?
- 11.3.3 Match the following unvoiced consonants with their voiced counterparts: t, s, k, p, f, ch, sh, th (as in think), wh.
- 11.3.4 Simulate the speech production mechanism by creating a triangle pulse train of variable pitch and filtering with a 3-4 pole AR filter. Can you produce signals that sound natural?
- 11.3.5 Experiment with a more sophisticated software speech synthesizer (source code may be found on the Internet). How difficult is it to produce natural-sounding sentences?

11.4 Speech Perception

The human ear along with the human brain are a most impressive sound receiver. We can actually detect sounds that are so weak that the air pressure density fluctuations are less than one billionth of the average density. These sounds are so weak that the ear drum moves only about the diameter of a single hydrogen atom! But we can also hear very strong sounds, sounds so strong that the ear drum moves a millimeter. The frequency range of the ear is also quite remarkable. Not only can we hear over ten octaves (our visual system is sensitive over only about one octave), most people can distinguish between 998 Hz and 1002 Hz, a difference of a few parts per thousand. Piano tuners tune to within much better than this by using beat frequencies. Even the most tone deaf can easily distinguish a great variety of timbres, which are effects of lack of sinusoidality.

Sound perception commences with sound waves impinging on the outer ear, and being funneled into the 'auditory canal' toward the middle ear. The sound waves are amplified as they progress along the somewhat narrowing canal, and at its end hit the 'tympanic membrane' or eardrum and set it into vibration. The physical dimensions of the outer ear also tend to band-pass the sound waves, enhancing frequencies in the range required for speech. The eardrum separates the outer ear from the middle ear, which is a small air-filled space, with an opening called the 'Eustachian tube' that leads to the nasal tract. The Eustachian tube equalizes the air pressure on both sides of the eardrum, thus allowing it to vibrate unimpeded. A chain of three movable bones called 'ossicles' (and further named the 'hammer', 'anvil' and 'stirrup') traverses the middle ear connecting the eardrum with the inner ear. The vibrations of the eardrum set the hammer ossicle into motion, and that in turn moves the anvil and it the stirrup. The vibrations are eventually transmitted to a second membrane, called the 'oval window', which forms the boundary between the middle and inner ear. Since the base of the stirrup is much smaller than the surface of the eardrum, the overall effect of this chain of relay stations is once again to amplify the sound signal.

From the oval window the vibrations are transmitted into a liquid-filled tube, coiled up like a snail, called the 'cochlea'. Were the cochlear tube to be straightened out it would be about 3 centimeters in length, but coiled up as a $2\frac{1}{2}$ - to 3-turn spiral it is only about 0.5 cm. The cochlea is divided in half along its length by the 'basilar membrane', and contains the organ of Corti; both the basilar membrane and the 'organ of Corti' spiral the length of the cochlea. Vibrations of the oval window excite waves in the liquid in the cochlea setting the basilar membrane into mechanical vibration. Were we to straighten the cochlea out we would observe that its width tapers from about a half-centimeter near the oval window to very small at its apex; however, the basilar membrane is stiff near the oval window and more flexible near the apex. Combined, these two characteristics make the basilar membrane frequency selective. High frequencies cause the basilar membrane to vibrate most strongly near the oval window, and as the frequency is lowered the point of strongest vibration moves along the length of the basilar membrane toward the apex of the cochlea.

The organ of Corti transduces the mechanical vibrations into electric signals. It has about 15,000 sensory receptors called 'hair cells' that contact the basilar membrane and stimulate over 30,000 motion sensitive neurons that create electric pulses that are transmitted along the auditory nerve to the brain. There are two types of hair cells, three rows of 'outer' hair cells and one row of 'inner' hair cells. Motion of the basilar membrane moves the hair cells back and forth causing them to release neurotransmitter chemicals that cause auditory neurons to fire. Since different parts of the membrane respond to different frequencies, auditory neurons that are activated by inner hair cells that contact a particular location on the basilar membrane respond mainly to the frequency appropriate to that location. Complex sounds activate the basilar membrane to different degrees along its entire length, thus creating an entire pattern of electric auditory response. Similarly the outer hair cells are intensity selective, different sound intensities stimulate different hair cells and create different neuron activity patterns.

We can roughly describe the operation of the cochlea as a bank of filters spectral decomposition with separate gain measurement. As different sounds arrive at the inner ear the hair cell response changes creating a varying spatial representation. The neural outputs are passed along the auditory nerve toward the cortex without disturbing this representation; the spatial layout of the neurons in the nuclei (groups of nerve cells that work together) closely resembles that of the hair cells in the inner ear. Indeed in all nuclei along this path tonotopic organization is observed; this means that nearby neurons respond to similar frequencies, and as one moves across the nucleus the frequency of optimal response smoothly varies.

The auditory nerve from each ear feeds a cochlear nucleus in the auditory brainstem for that ear. From both cochlear nuclei signals are sent both upward toward the primary auditory cortex and sideways to the superior olivary complex, from which they proceed to the pathway belonging to the opposite ear. This pathway mixing enables binaural hearing as well as mechanisms for location and focus. What about the auditory cortex itself? We started the previous section by contrasting the vocal tracts of the human with those of other primates, yet the difference in our brain structure between ourselves and the apes is even more remarkable. The human brain is not the most massive of any animal's, but our brain mass divided by body mass is truly extraordinary, and our neocortex is much larger than that of any other animal. There are two cortical regions that deal specifically with speech, Broca's area and Wernicke's area, and these areas are much more highly developed in humans than in other species. Broca's area is connected with motor control of speech production apparatus, while Wernicke's area is somehow involved in speech comprehension.

To summarize, the early stages of the biological auditory system perform a highly overlapped bank of filters spectral analysis, and it is this representation that is passed on to the auditory cortex. This seems to be a rather general-purpose system, and is not necessarily the optimal match to the speech generation mechanism. For example, there is no low-level extraction of pitch or formants, and these features have to be derived based on the spectral representation. While the biology of speech generation has historically had a profound influence on speech synthesis systems, we are only now beginning to explore how to exploit knowledge of the hearing system in speech recognition systems.

EXERCISES

- 11.4.1 Experiment to find if the ear is sensitive to phase. Generate combinations of evenly spaced sines with different phase differences. Do they sound the same?
- 11.4.2 Masking in the context of hearing refers to the psychophysical phenomenon whereby weak sounds are covered up by stronger ones at nearby frequencies. Generate a strong tone at 1 KHz and a weaker one with variable frequency. How far removed in frequency does the tone have to be for detection? Attenuate the weaker signal further and repeat the experiment.
- 11.4.3 Sit in a room with a constant background noise (e.g., an air-conditioner) and perform some simple task (e.g., read this book). How much time elapses until you no longer notice the noise?
- 11.4.4 Go to a (cocktail or non-drinking) party and listen to people speaking around the room. What effects your ability to separate different voices (e.g. physical separation, pitch, gender, topic discussed)?
- 11.4.5 Have someone who speaks a language with which you are unfamiliar speak a few sentences. Listen carefully and try to transcribe what is being said as accurately as you can. How well did you do?

- 11.4.6 Talk to someone about speech recognition and then quickly ask 'Do you know how to wreck a nice peach?'. Ask your partner to repeat what you just said. What does this prove?
- 11.4.7 Most of the time and energy of the speech signal is spent in the vowels, and hence the speech perception mechanism performs best in them. But do vowels carry most of the information? You can find out by performing the following experiment. Select several sentences at random from this book. From each sentence create two character strings, one in which all consonants are replaced by question marks, and one in which all vowels are. Now present these strings to subjects and ask them to fill in the blanks. What are your findings?
- 11.4.8 Explain the possible mechanisms for acoustic source location. Take into account the width of the head, the fact that localization is most accurate for high-frequency sounds with sharp attack times, and the idea that the head will absorb some sounds casting an acoustic shadow (about 3 dB at 500 Hz, 20 dB at 6 KHz). How is height determined? Devise a neurobiologically plausible model for a time-of-arrival crosscorrelator.
- 11.4.9 Simulate the sound perception mechanism by building a bank of overlapping band-pass filters (at least 100 are needed) and graphically displaying the output power with time as horizontal axis and filter number as vertical axis. Test by inputting a sinusoid with slowly varying frequency. Input speech and try to segment the words on the graphic display.

11.5 Brains and Neurons

The human brain is certainly a remarkable computer and signal processor. We have seen above how it can communicate with other brains using audio frequency waves in the air by coercing the mouth (an organ developed for eating and breathing) to broadcast messages and obliging the ear (originally for detecting predators and prey) to capture these messages. It can also communicate by using the hands to write and eyes to read; it can process visual information at high speed, recognizing human faces and familiar objects in real time; it can instruct the hands to manipulate objects, and enable the body to avoid obstacles and navigating in order to get to wherever it wants. The brain can use tools, create new tools, find rules in complex phenomena; it can write music and poetry, do mathematical calculations, learn to do things it didn't know how to do previously. It can even create new thinking machines and signal processing machines that excel in areas where it itself is limited.

This section contains a short introduction to the brain's hardware architecture. Our purpose is not the study the brain's physiology per se; rather we wish to understand its prominent features in order to gain inspiration that may lead to the building of better signal processing and computing machines. In the following section we will introduce *artificial neural networks*, which are models that attempt to capture the essential properties of the brain's computational architecture and are used both to explain the functioning of the biological brain and to solve practical problems.

The brain as an organ had been studied by the ancients, and by the mid-nineteenth century it was known that certain well-defined areas of the brain were responsible for specific functions such as speech. It was Santiago Ramón y Cajal who first convincingly demonstrated, in the latter half of the 1880s, that the brain is not simply a large mass of fibers, but a vast number of richly interconnected brain cells, which we call 'neurons'. Using a cell staining technique earlier developed by Camillo Golgi (with whom he shared the 1906 Nobel prize for Physiology or Medicine) he both observed neurons and mapped their anatomy.

Neurons are a specialized type of cell, of which there are over 10 billion (10^{10}) in the human brain. From a functional point of view we can roughly categorize neurons into three classes, namely 'sensory neurons' (such as those in the retina of the eye that are sensitive to light), 'motor neurons' (e.g., those which activate and control the motion of our fingers), and 'higher processing neurons' (those in the neocortex). We will discuss mainly the last of these categories, but even of these neurobiologists have discovered many different varieties, such as pyramidal cells, Golgi cells, spiny stellate cells, smooth stellate cells, interneurons, etc. Our description will be so simplified that the differences between these various varieties will be unimportant.

The classical processing neuron, depicted schematically in Figure 11.1, is made up of three anatomical structures—the 'soma', the 'axons', and the 'dendritic tree'. The soma is the cell body and is responsible for the processing itself. The dendrites supply inputs to the neuron, while the axon carries the neuron's output. We can think of the neuron as a simple processing element that inputs multiple signals (about 10^4 to 10^5 is typical) and outputs a single signal.

What kind of signals are input and output? The interior of a neuron is usually electrically negative relative to the outside, due to the cell membrane selectively passing ions from inside the cell outward and from outside inward. The membrane's electric potential is not constant however, and its behavior as a function of time can be viewed as a signal. Perhaps the most significant type of behavior is the 'action potential'. This is a very fast event,

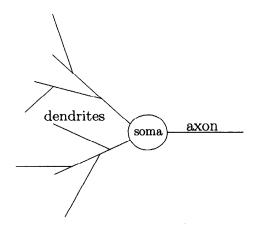


Figure 11.1: A highly schematic diagram of the classic higher processing neuron. The dendrites at the left are the inputs, the axon at the right is the output, and the processing is performed by the soma.

occupying about a millisecond, although afterward slow oscillations can occupy a further 100 milliseconds. The event itself starts with the membrane potential becoming even more negative than usual followed by short sign reversal. This spike can travel along the neuron's axon without decrease in amplitude; this propagation is not like electric current in a conductor, rather it is due to the axon being made of active material with each section exciting the next to spike. Due to the all-or-none nature of the action potential we will usually refer to the neuron as 'firing' if it has developed an action potential, or 'quiescent' if it has not. After a spike there is an 'absolute refractory period' during which the neuron cannot fire again, and a 'relative refractory period' during which the neuron is less susceptible to spiking. Although rates of several hundred spikes per second are possible, more typical frequencies are on the order of 10 Hz.

A 'synapse' is formed where one neuron touches and influences another. While there are other possibilities we will discuss synapses formed by the axon of the presynaptic neuron touching a dendrite of the postsynaptic neuron. At the synapse the two cell membranes touch but cellular material does not indiscriminately flow between the cells, the influence of the presynaptic potential being indirect. For example, in chemical synapses a presynaptic spike causes a transmitter substance to be released that changes the postsynaptic membrane permeability. Synapses may be 'excitatory', meaning that the firing of the presynaptic neuron increases the probability of the postsynaptic neuron firing as well; or 'inhibitory' if that firing decreases the postsynaptic neuron's chance of firing. No single neuron is really that significant; only the 'network' of myriad neurons is of consequence. Each neuron's dendrites are contacted by axons of other neurons, and its own axon contacts, in turn, many other neurons. Thus the emerging picture is that of a huge number (about 10^{10}) of neurons, each firing or quiescent. The decision whether to fire is made in the soma, based on the input from all the (about 10^5) neurons that make synapses upon its dendrites. Once fired the action potential rapidly propagates from the soma down the axon to influence the firing of yet more (about 10^5) neurons.

Neurons in contact need not be physically close, nor do physically close neurons need to be in direct contact. Thus in our quest to understand brain function we are led to consider large areas of brain matter. The brain is highly organized, with specific locales responsible for specific functions, and various task geometries mapped onto brain geometries. As early as 1861, Paul Broca described a patient who had lost his ability to speak although he did understand spoken language. After the patient's death he tied this to a lesion in a specific position in the brain, now called Broca's area. Hubel and Wiesel earned the 1981 Nobel prize in medicine for their description of the early visual system. They discovered formations they called hypercolumns. The neurons in each hypercolumn respond to lines in certain areas of the visual field, with nearby hypercolumns responding to lines in nearby locations. As one travels along a hypercolumn the angle of the detected line slowly rotates. These facts and more lead us to conclude that the entire brain is a large interconnected network of neurons that can be broken down into task-specific subnetworks that are tightly connected with other subnetworks.

Since the days of Cajal and Golgi, neurobiologists have studied in depth the characteristics of single isolated neurons, and although much progress has been made, this study does not seem to lead to any deep explanation of brain function. Others have studied the larger-scale structure of the brain and discovered the mappings between function and specific areas of the brain, but even this immensely valuable information explains *where* but not *how* or *why*. In order to gain insight into the connection between the brain's anatomy and its function it is necessary to simplify things.

EXERCISES

11.5.1 The central nervous system is composed of the spinal cord, the brain stem, the cerebellum, the midbrain, and the left and right hemispheres of the neocortex. What are the functions of these different components? How do we know? What creatures (reptiles, mammals, primates) have each component?

- 11.5.2 In addition to neurons the brain also contains glia cells; in fact there are more glia than neurons. Why did we focus on neurons and neglect glia cells in our explanation?
- 11.5.3 There are many morphologies other than the classical model described in this section. For example, there are neurons with no axon or many axons, synapses may occur between two axons, or between two dendrites, or even on the cell body. Research several such variations. What is the function of these cells?
- 11.5.4 Research the Nobel prize-winning contribution of Hubel and Wiesel to the understanding of the neurons in the mammalian visual system. What are simple cells, complex cells, and hypercomplex cells? What is a hypercolumn and how are the cells arranged geometrically?

11.6 The Essential Neural Network

The single neuron does not perform any significant amount of computation; computation is performed by large collections of neurons organized into 'neural networks'. The term 'neural network' is actually misleading; the concept is not that of a network that has neural characteristics, but simply a network of neurons. Perhaps the term should be 'neuron network', but the original term has become entrenched. By association, other collections of interconnected processors, including ones we can make ourselves, are often called neural networks as well. However, the term is only fitting when the collection of processors is somehow inspired by the brain. A LAN of personal computers is a collection of interconnected processors that would not normally be considered a neural network.

When does a collection of processors become a neural network? Any definition we give will be subjective, and probably the number of different definitions equals the number of people working in the field. However, there are a number of requirements that most researchers would agree upon. My own definition goes something like this.

Definition: neural network

A *neural network* is a large set of simple, richly interconnected processing units that exhibits collective behavior after learning.

There must be a *large* number of processors, at least in the hundreds, before we leave the more conventional 'parallel processing' and reach the

regime where collective behavior is meaningful. The individual neuron must be *simple*, performing one basic operation like calculating the dot product of its input with stored weights, or finding the distance between its input and a stored pattern. We definitely don't want to depend on multifunctional, highly precise, processors here. Some people would require nonlinearity of the neuron's operation, but we will be lax in this regard. To make up for the simplicity of the individual processor, and to exploit the large number of processors, we want them to be *richly interconnected*. Conventional parallel processing techniques prefer to connect processors to nearest neighbors on a grid or with hypercubic geometry. Biological networks may not be fully interconnected, but the connectivity is quite high.

The robustness to failure of conventional computers is infinitesimal. Were one to remove a randomly selected circuit from a personal computer or even simply cut a randomly selected conductor, the probability of total system failure is very high. This should be contrasted with the brain which loses large numbers of neurons daily without serious performance degradation. How is this robustness obtained?

A clue is the fundamentally different methods of storing information in the two competing architectures. The conventional computer uses Location Addressable Memory (LAM) wherein information is stored in a particular location. In order to retrieve this information the location must be known. The brain uses *content addressable memory* (CAM). For example, once an image is stored we can present it and ask whether it is a known image. A generalization of this idea is *associative memory*, by which we mean that we can present an image and ask if there is a stored picture that is similar (the association). In this fashion we can recognize a friend's face even with sunglasses and a different hair cut.

We can now try to piece the puzzle together. The real motive for the high connectivity of neural networks is to obtain *collective behavior*, also called *self-organization* and related to *distributed representations*. Were each memory to be stored, as in an LAM, in a specific neuron or definite small set of neurons, then failure of that neuron would wipe out that memory. Instead it seems plausible that memories are stored as eigenstates of the entire network. The mechanism that brings this about is spontaneous collective behavior, or self-organization.

Learning refers to the method of introducing memories and storing procedures. Conventional computers must be laboriously programmed; each new task requires expensive and time-consuming outside intervention. Brains learn from experience, automatically adapt to changing environments, and tend to be much more forgiving to 'bugs'.

EXERCISES

- 11.6.1 Which of the following are neural networks according to the definition given in the text?
 - 1. transistors on the substrate of an integrated circuit
 - 2. arithmetic registers in a microprocessor
 - 3. CPUs in a parallel processing environment
 - 4. cells in the spinal column
 - 5. neurons in an aplysia
- 11.6.2 By introspection, make a rough order-of-magnitude estimate of the amount of information (in bits) passed to the brain by the various senses. For vision, for example, estimate the size of the main field of vision, the pixel density, the dynamic range, and the number of pictures transferred per second. Based on the above estimates, how much information must the brain store in a day? A year? A lifetime? The brain contains about 10¹⁰ neurons. Does the above estimate make sense?

11.7 The Simplest Model Neuron

In this section we will consider a simple model neuron. This model does not do justice to the real biological neuron. Even using a single model, no matter how complex, is a gross simplification. Real neurons have complex timedependent properties that we will completely ignore in this simple model; and the functioning of our model will be a mere caricature of the real thing.

So why should we attempt to model the neuron? An analogy is useful here. The reader will remember the *ideal gas law* $PV = nk_BT$, which reliably relates the pressure, volume and temperature for a large number of gases. This law is only approximate, and indeed it breaks down at very high pressures or a temperatures close to the condensation temperature of the particular gas. However, it is a good approximation for a very large number of gases over a large regions of P, V, and T, and furthermore corrections can be added to better approximate the actual behavior. The ideal gas law can be derived in statistical physics from the microscopic behavior of the gas molecules under the assumption that they are essentially ping pong balls. By this we mean that the gas molecules are assumed to be small spheres of definite size, which only interact with other molecules by colliding with them. Upon collision the molecules change their velocities as colliding ping pong balls would. Using techniques of statistical physics, which is a mathematical formalism designed to derive macroscopic 'average' laws from the behavior of huge numbers of simple particles, the ideal gas can be derived.

No-one really believes that the gas molecules are ping pong balls. They are definitely not spheres of well-defined radius—they are composed of a nucleus with protons and neutrons surrounded by electron 'clouds'. They definitely do not interacting like ping pong balls—there are electromagnetic fields that act at a distance and the dynamics is inherently quantum mechanical. So why does the ideal gas law work?

The answer is that it doesn't. When the pressure is high or the temperature low, the molecules are close together and the model breaks down miserably. The gas condenses into a liquid, and the temperature at which this happens is different for different gases. But for a large range of parameters the most important thing is that there are a very large number of molecules that interact only weakly with the others, except for short periods of time when they are close. Thus any model that obeys these constraints will give approximately the same behavior, so we might as well pick the easiest model to work with. Since the ping pong ball model is the simplest to handle mathematically, it is the natural starting point.

Let's return to the neuron. There is a large variety of types, and each is an extremely complex entity; but we believe that as a first approximation the most important features are the huge number of neurons, and the fact that these are so richly interconnected. In the spirit of statistical physics we search for the simplest 'ping pong ball' model of a neuron. This is the McCulloch-Pitts model, first proposed in the early 1940s.

The McCulloch-Pitts neuron was originally designed to show that a simple neuron-like device could calculate logical functions such as AND and OR. The neuronal output is calculated by comparing a weighted linear combination (convolution) of the inputs to a threshold. Only if the linear combination is above the threshold will the neuron fire. Such a function is often called a linear threshold function.

In order to state this description mathematically, we must introduce some notation. The output of the neuron under consideration will be called O, while its N inputs will be called I_j with $j = 1 \dots N$. At this early stage the neural computation community already divides into two rival camps. Both camps represent the neuron firing as O = +1, but one uses O = 0for quiescence, while the other prefers O = -1. The synaptic efficacy of the connection from input j will be represented by a real number W_j . For excitatory synapses $W_j > 0$, while for inhibitory ones $W_j < 0$. The absolute value of W_j is also important since not all inputs affect the output in the same measure. If $W_j = 0$ then the input does not affect the output at all (there is no synapse). If $|W_j|$ is large then the effect of this input is significant, while small $|W_j|$ means the input only marginally affects the output.

The linear combination of the McCulloch-Pitts neuron means that the cell's potential is assumed to be

$$h = \sum_{j=1}^{N} W_j I_j \tag{11.6}$$

and the neuron will fire when this potential surpasses the threshold potential required to excite an action potential θ . Using the 0, 1 representation we can write

$$O = \Theta(h - \theta) = \Theta\left(\sum_{j=1}^{N} W_j I_j - \theta\right)$$
(11.7)

where Θ is the step function. With the ± 1 representation we write

$$O = \operatorname{sgn}(h - \theta) = \operatorname{sgn}\left(\sum_{j=1}^{N} W_j I_j - \theta\right)$$
(11.8)

where the signum function sgn returns the sign of its argument. Note that the meaning of this latter equation is somewhat different from the previous one; here neurons that are *not* firing also contribute to the sum.

The McCulloch-Pitts neuron is the simplest possible neuron model. The inputs are combined in a simple deterministic way. The decision is made based on a simple deterministic comparison. The summation and comparison are instantaneous and time-independent. Yet this completely nonbiological formal neuron is already strong enough to perform nontrivial computations. For example, consider the following image processing problem. We are presented with a black-and-white image, such as a fax, that has been degraded by noise. The classic DSP approach would be to filter this binary-valued image with a two-dimensional low-pass filter, which could be implemented by averaging neighboring pixels with appropriate coefficients. This would leave a gray-scale image that could be converted back to a black-and-white image by thresholding. This combination of the linear combination of input pixels in neighborhoods followed by thresholding can be implemented by a two-dimensional array of McCulloch-Pitts neurons. This same architecture can implement a large variety of other image processing operators.

What is the most general operation a single linear threshold function can implement? For every possible input vector the McCulloch-Pitts neuron outputs 0 or 1 (or ± 1). Such a function is called a 'decision function' or a dichotomy. Thinking of the 2^N possible input configurations as points in N-dimensional space, the linear combination is obviously the equation for a N-1-dimensional hyperplane, and the thresholding separates the inputs that cause positive output (which are all on one side of the hyperplane) from the others. Thus we see that the threshold linear function implements linearly separable dichotomies.

This is only a tiny fraction of all the dichotomizations we might need to use. If this were all neuron models could do, they would find little practical use. The modern reincarnation of neural networks exploits such architectures as the multilayer perceptron (see Section 8.4), which broaden the scope of implementable dichotomies. In fact, feedforward networks of neurons can implement arbitrarily complex functions.

How does the brain learn the weights it needs to function? Hebb proposed a principle that can be interpreted at the neuron level in the following way.

Theorem: Hebb's Principle

The synaptic weight increases when the input to a neuron and its output tend to fire simultaneously. Using the notation W_{ij} for the weight that connects presynaptic neuron s_j with postsynaptic neuron s_i ,

$$W_{ij} \leftarrow W_{ij} + \lambda s_i s_j \tag{11.9}$$

where the s_i are either 0, 1 or ± 1 .

This form, where weights are updated accordingly to a constant times the product of the input and output, strongly reminds us of the LMS rule, only there the product is of the input and error. However, this difference is only apparent since if the postsynaptic neuron fires $s_i = +1$ when it shouldn't have the error is 1 - (-1) = 2 while in the opposite case the error is -1 - (+1) = -2 and the difference is only a factor of two that can be absorbed into λ . The true difference is that the desired s_i can only take on the discrete values $(0, 1 \text{ or } \pm 1)$; so rather than converging to the true answer like LMS, we expect a neuron-motivated adaptation algorithm to eventually attain precisely the right answer. The first such algorithm, the 'perceptron learning algorithm' was actually discovered before the LMS algorithm. It can be shown to converge to an answer in a finite number of steps, assuming there is an answer to the problem at hand.

The problem with the perceptron learning algorithm is that it does not readily generalize to the more capable architectures, such as the multilayer perceptron. The most popular of the modern algorithms is based on a variant of LMS.

EXERCISES

- 11.7.1 One can convert 0, 1 neurons into ± 1 neurons by using the transformation $S \rightarrow 2S 1$. Show that equations (11.7) and (11.8) are equivalent by thus transforming the I_j and finding transformations for W_j and θ .
- 11.7.2 Show that AND and OR gates of in-degree N can be implemented using McCulloch-Pitts neurons. That is, specify W_j and θ such that the neuron will fire only if all inputs fire, or if at least one input fires.
- 11.7.3 Draw and label the possible input configurations of a three-input linear threshold function as vertices of a cube. Show graphically which dichotomies can be implemented with zero threshold.
- 11.7.4 Extend the results of the previous problem to general McCulloch-Pitts neurons (nonzero threshold).
- 11.7.5 The McCulloch-Pitts neuron can be used as a signal detection mechanism. Assume we wish to detect a signal s_n of N samples and report the time that it appears. One builds a neuron with N continuous inputs and weights precisely $W_j = s_j$. We then have the input flow past the neuron, so that at any time the neuron sees N consecutive inputs. Consider the ± 1 representation and show that the PSP will be maximal when the required signal is precisely lined up with the neural inputs. How should the threshold be set (take noise and false alarms into account)? To what signal processing technique does this correspond?
- 11.7.6 Extend the results of the previous exercise to image recognition. Would such an approach be useful for recognition of printed characters on a page? If not why not? What about cursive handwriting?
- 11.7.7 Discuss the use of McCulloch-Pitts neurons for associative memory.
- 11.7.8 What is the difference between Hebb's principle for 0,1 neurons and ± 1 neurons?

11.8 Man vs. Machine

Now that we have a basic understanding of the brain's computational architecture we can attempt a quantitative comparison between the brain and the conventional computer. The pioneers of the modern computer were aware of the basic facts of the previous section, and were interested in eventually building a brain-like device. However, the prospect of 10^{10} parallel processing elements was quite daunting to these early computer engineers, who thus compromised on a single processing element as a kind of first approximation.

Surprisingly, by making this one processing element faster and more powerful, this computer evolved into a completely different architecture, quite powerful in its own right. Only now, with computer speeds approaching the absolute limits that physics imposes, is parallel processing being once again seriously considered; but even now when computer engineers talk about parallel processing they are referring to small numbers of CPUs, such as two, four, or eight. A comparison of the human brain to a conventional computer, based on the information of the last few sections, is to be found in Table 11.1.

	Brain	Computer
number of	≈ 10 billion <i>neurons</i>	1 CPU
processors	(massively parallel)	(intrinsically serial)
processor	simple	complex
complexity	inaccurate	accurate
processor	slow	fast
speed	(millisec)	(nanosec)
inter-processor	fast	slow
communications	(μsec)	(millisec)
learning mode	learn from experience	manual programming
failure	many neurons die	single fault often
robustness	without drastic effect	leads to system failure
memory	content addressable	location addressable
organization	(CAM)	(LAM)

 Table 11.1: A quantitative and functional comparison of the human brain and a serial processing computer.

The term *architecture* as applied to computers was invented to describe all the aspects of the computer's hardware that software must take into account. Two computers of identical architectures but different speeds can be uniquely compared as to strength—if the clock speed of one is twice that of the other, every program will run on it twice as fast. Two computers of similar, but not identical architectures are not uniquely comparable, and thus different programs will run at slightly different speed-up ratios. The more the architectures differ, the greater will be the divergence of the benchmarking results. This does not mean that we cannot say that a supercomputer is stronger than a desktop computer! There is another way to define the concept of 'stronger'. Many engineering workstations today come with software emulations of personal computer environments. These emulations can run actual applications designed for the personal computer by emulating that entire platform in software. When a PC program is input to the emulation, all the PC opcodes must be read and interpreted and the required operation precisely simulated by an appropriate workstation command or routine. Were the program to run on the workstation emulation faster than on the target PC, we would be justified in concluding that the workstation is stronger than the PC, even though their architectures are dissimilar. Can we make such a comparison between the conventional computer and the brain?

To answer this question definitively we require an estimate as to the number of computers required to emulate a human brain at the hardware level. From the previous section we know that to emulate the simplest possible neuron, we would have to carry out N multiplications and accumulate operations (see Section 17.1) every 'clock' period. Here the number of synapses $N \approx 10^5$ and the clock period is about 5 milliseconds, and so a single neuron would require at least 210^7 MACs per second, and 10^{10} neurons would require over 210^{17} MAC/sec. Assuming even an extremely fast CPU that could carry out a MAC in 5 nanoseconds, we would require 10^9 such computers in parallel to simulate a single human brain!

So the brain is equivalent to 1 gigacomputer! This sounds quite impressive, even without taking the small physical size and low power requirements into account. Now let's ask the converse question. How many humans would be required to emulate this same 5 nanosecond computer? Even assuming that a human could carry out the average operation in five seconds (and it is doubtful that many of us can perform an arbitrary 16-bit multiplication in this time, let alone 32-bit divisions) the computer would have carried out 10^9 operations in this same 5 seconds, and so we would need 10^9 humans in parallel to emulate the computer! So the computer is equivalent to a gigahuman as well.

How could these comparisons have turned out so perverse? The reason is that the underlying architectures are so very different. In such a case cross-emulation is extremely inefficient and direct comparison essentially meaningless. Certain benchmarking programs will run much faster on one machine while others will demonstrate the reverse behavior. The concept of 'stronger' must be replaced with the idea of 'best suited'.

Thus when two quite different computational architectures are available and a new problem presents itself, the would-be solver must first ask 'Which architecture is more suited to this problem?'. Although it may indeed be *possible* to solve the problem using either architecture, choosing the wrong one may make the solution extremely inefficient or even unsolvable in practice. For example, were we required to calculate the 137th root of a 50digit number, I believe that most readers would agree that the conventional number-crunching computer is more suited to the task than the human (or chimpanzee). However, when the problem is the understanding of spoken language, the reading of cursive handwriting, or the recognition of faces, the architecture of the brain has proved the more viable. Indeed for these tasks there is at least an existence proof for a neural network solution, but none has yet been proffered for the serial computer!

Despite the above argument for neural computation, the neural network approach has had only limited success so far. Optical Character Recognition (OCR) engines based on neural networks have indeed eclipsed other technologies, yet progress on speech recognition has been slow. At least part of the fault lies in the size of network we can presently build, see Figure 11.2. Our largest systems seem to be on the level of a mentally retarded mosquito! We are forced to conclude that our previous 'existence proof' for neural solutions to ASR, OCR, face recognition, and other problems is contrived at best. The only way our present-day artificial neural networks will be able to solve practical problems is by being more efficient than biology by many orders of magnitude.

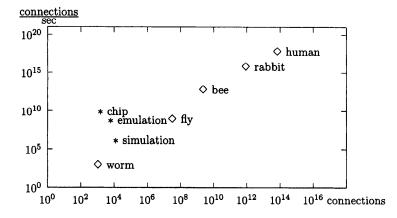


Figure 11.2: The speed and complexity of various neural networks. The horizontal axis is the number of synapses in the network, a number that determines both the information capacity and the complexity of processing attainable. The vertical axis is the number of synaptic calculations that must be performed per second to emulate the network on a serial computer, and is an estimate of the network's speed.

EXERCISES

- 11.8.1 The 'expert system' approach to artificial intelligence is based on the tenet that anything a human (or other rational agent) can do a standard computer can be programmed to do. Discuss the veracity and relevance of this hypothesis. What types of problems have been solved using expert systems? What AI problems have eluded prolonged attempts at solution?
- 11.8.2 In the Hopfield model there are N neurons $\{s_i\}_{i=1...N}$ each of which can take on the values ± 1 , where $s_i = +1$ means that the i^{th} neurons is firing. The synapse from presynaptic neuron j to postsynaptic neuron i is called W_{ij} , and the matrix of synaptic efficacies has zeros on the diagonal $W_{ii} = 0$ and is symmetric $W_{ij} = W_{ji}$. At any given time only one neuron updates its state; the updating of the i^{th} neuron is according to

$$s_i(t+1) = \operatorname{sgn} \sum_{j=1}^N W_{ij} s_j(t)$$

after which some other neuron updates. Write a program that generates a random symmetric zero diagonal synaptic matrix, starts at random initial states, and implements this dynamics. Display the state of the network graphically as a rectangular image, with $s_i = \pm 1$ represented as different colored pixels. What can you say about the behavior of the network after a long enough time? What happens if you update all the neurons simultaneously based on the previous values of all the neurons? What happens in both of these cases if the synaptic matrix is antisymmetric? General asymmetric?

11.8.3 Given P N-bit memories $\{\xi_i^{\mu}\}_{i=1...N}^{\mu=1...P}$ to be stored, the Hebbian synaptic matrix is defined as

$$W_{ij} = \sum_{\mu=1}^P \xi_i^\mu \xi_j^\mu$$

which is the sum of outer products of the memories. Enhance the program written for the previous exercise by adding a routine that inputs desired memory images and computes the Hebbian matrix. Store P < 0.1N memories and run the dynamics starting near one of the memories. What happens? What happens if you start far from any of the memories? Store P > 0.2N memories and run the dynamics again. What happens now?

Bibliographical Notes

Good general introductions to psychophysics can be found in [50, 71].

Alexander Graham Bell's original vocation was speech production and much of the early research on speech generation mechanisms was performed at Bell Labs [54, 55]. The classic formant tables [193] and the ear sensitivity curves [62] also originated there.

Speech production mechanisms are presented in many books on speech processing, such as [211] and in more depth in [253]. Speech perception is covered in [253, 195, 129]. The Bark scale is formally presented in [290, 232, 6] and the mel scale was defined in [254]. Cochlear modeling is reviewed in [5]. The application of the psychophysical principle of masking to speech compression is discussed in [232].

The McCulloch-Pitts neuron was introduced in [171]. In 1957 a team led by Frank Rosenblatt and Charles Wightman built an electronic neural network, which they called the *Mark I Perceptron*. This device was designed to perform character recognition. It was Rosenblatt who discovered and popularized the perceptron learning algorithm [224]. Minsky and Papert's charming book [174] both thoroughly analyzed the algorithm and dampened all interest in neural network research by its gloomy predictions regarding the possibility of algorithms for more capable networks.

For a light introduction to the functioning of the brain, I suggest [32], while a more complete treatment can be found in [180]. [8] is a thorough introduction to neuron modeling from a physicist's point of view. [168] is a seminal work on neural network modeling and [169] is the companion book of computer exercises.