The Network

Dan Auerbach

2015

Part I

The Telephone

The 1876 World's Fair in Philadelphia was a steampunk fantasy: displays of steam engines, the world's Remington Typographic Machine (the first typewriter), and precision watches. Yet while mechanical attractions dominated the event, the age of electricity was just around the corner, evident in a modest area across the Machinery Hall, where a young Alexander Graham Bell was demonstrating the transfer of human voice through wires.

People viewed early telephone prototypes like the one on display by Bell as a kind of toy, not a serious contender to the business-oriented telegraph empire of Western Union. But as the viability of the telephone grew, that perception changed quickly and it did not take long for commercial telephony to develop, seeded by Bell and his collaborators. Once off the ground, there was a rapid shift in power away from the well-established Western Union and instead towards the upstart corporation that would become AT&T, a telecommunications giant that would come to define the industry in the United States for a century.

In this part of the book, we will examine how the telephone works: what is sound? how is an analog signal transmitted over a wire? To understand telephony, we will go step by step through the process of building a telephone from scratch, and then outline the key insights required to build a telephone network. In addition to learning about the technology, we will also examine some historical and social issues related to the emergence of the modern telecommunications industry.

Given that our ultimate goal is understanding the Internet and not the telephone network, we will NOT attempt to follow each step of the complex historical evolution of the telephone, nor describe every aspect of a real-world telephone system, which as you might guess is quite complex. For readers interested in developing deeper knowledge of telephony, I recommend [INSERT BOOK REC].

Chapter 1

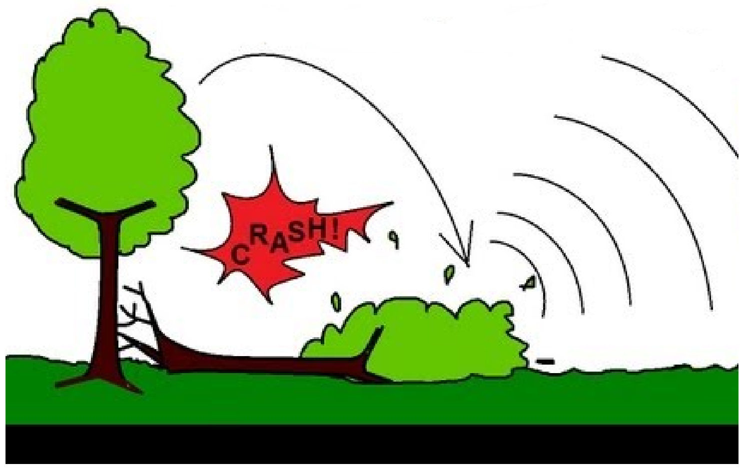
Signals

In the previous chapter, we studied symbols quite extensively, emphasizing the transmission of symbols from one location to another as a *digital* telecommunication system. In contrast to representing information as a series of discrete symbols, information can also be presented in an *analog* fashion via *signals*, in which a variable might take on an infinite number of values over time. Recall that such a variable is called *continuous*. A few examples of continuous variables are time, pitch, temperature, human height.

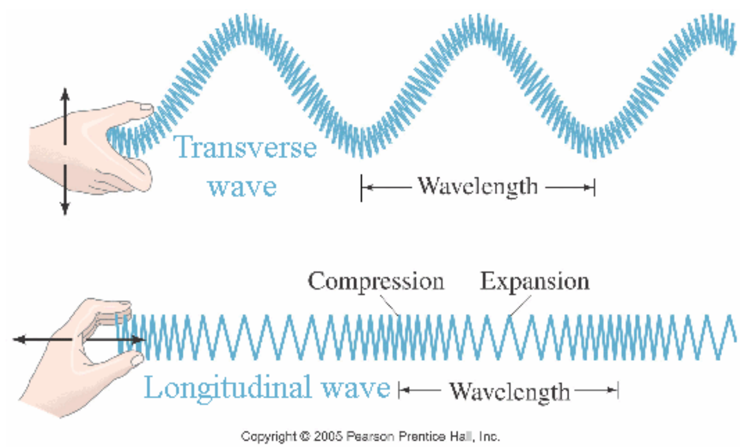
Early telephone systems represented human society's first foray into an analog telecommunication technology, allowing sound to be transmitted from one location to another, and so we will use sound as a natural jumping off point to understand analog signals.

1.1 Sound

When a tree falls and hits the ground, some of the energy upon impact is transferred to nearby air particles, pushing them outwards in a wave.



The wave is a *longitudinal* wave, meaning it does not have the peaks and troughs of an ocean wave perpendicular to the direction of travel, but instead expands and contracts along a single dimension as it travels.

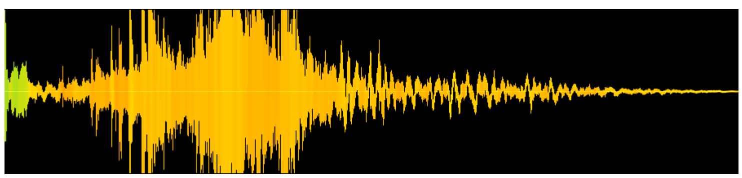


As a first approximation towards understanding sound, imagine a large nuclear explosion right where the tree strikes the ground, except that instead of an outwardly expanding sphere of fire, there is an outwardly expanding sphere where air particles are more tightly bunched together. This momentary compression of air particles constitutes a change in *air pressure* in that location.

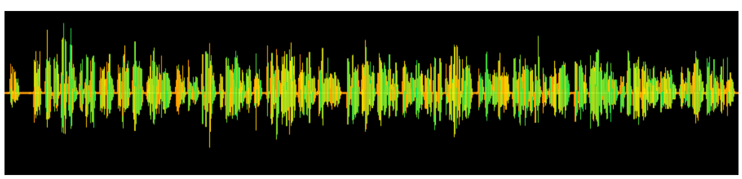
As objects on Earth interact with each other, waves of pressure expand outward in the surrounding medium (in the example above, this medium is the air). These pressure changes often signal events that are important to humans and other biological organisms, like when something is approaching an organism that could be either a predator or food. Given the importance of this information for survival, many species including humans evolved auditory systems to make sense of the constantly and subtly changing air pressure all around us; we call our experience of this phenomena *hearing*. And in parallel to the development of our sense of hearing, we evolved increasingly sophisticated ways to create sound ourselves, culminating in vocal chords and speaking.

The simplified description above of a tree falling might lead one who is unfamiliar with hearing to believe that there is only a single, uniform experience of a sound, but of course that's not the case. Sound has many different qualities, like pitch, intensity, duration, and timbre. It is quite easy to tell the soulful barotone of a jazz singer from the sound of a tree hitting the ground. This differentiation is possible because sound waves can be rather sophisticated, the miniscule air particles compressing and un-compressing in non-uniform ways over very small distance and time scales.

To illustrate these differences, here is a representation of the sound of a tree falling:



And here is a representation of human speech:

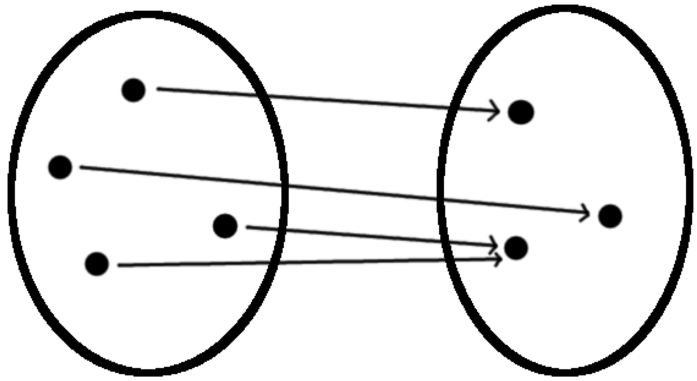


The above figures are called *time domain representations*: the x axis is time, and the y axis represents the instantaneous change in air pressure. Large values on the y axis mean that particles are tightly compressed at that instant. Of course these representations have to be taken from a single source point: two people standing in different places may have different auditory experiences of the same tree-falling event.

1.2 Functions

The two figures above representing various sounds are examples of *functions*, a central concept in mathematics that's worth covering carefully, since we will be returning to it quite often.

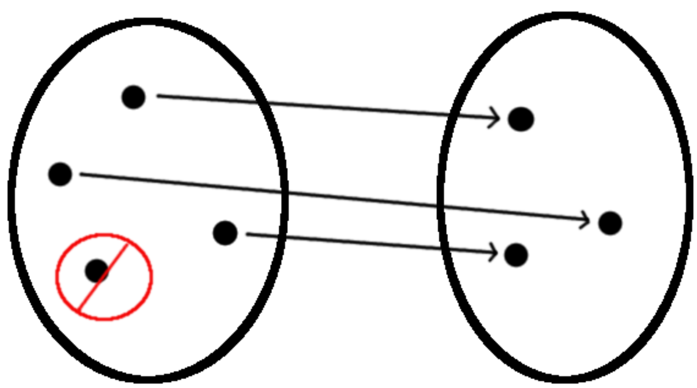
A function is a mapping from one set of objects to another:



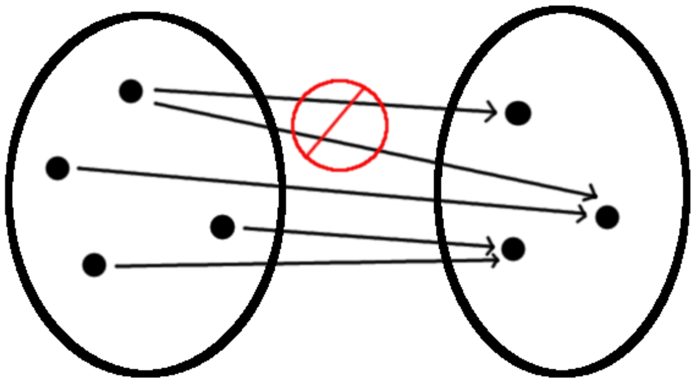
We say that a function has a *domain*, which is the set of objects on the left in the figure above, and a *co-domain* which is the set of objects on the right. We call the objects here *elements*. So a domain and a co-domain both are composed of elements, and a function is a mapping from the elements of the domain to the elements of the co-domain. There is one important rule that must be followed for something to count as a function:

- For each element in the domain, a function MUST map that to EXACTLY ONE element in the co-domain

In other words, since functions must be defined on the entire domain, this isn't allowed:

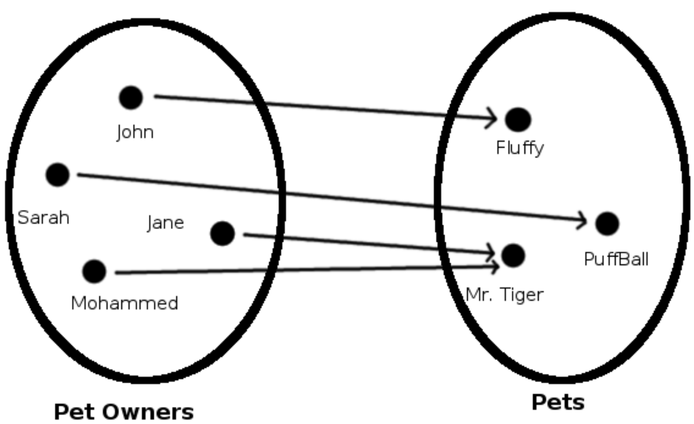


And neither is this:



Note that the converse of the second rule above does not hold; it is perfectly acceptable for several elements in the domain to be mapped to a single element in the co-domain, as is the case in the first diagram above.

So far, the domain has been 4 blobs, and the co-domain has been 3 blobs, but this is just one example. There are no restrictions on the domain or co-domain. You can have a function where the domain is pet owners on Earth, and the co-domain is the set of pets:



We might call this function "FavoritePet", and write $FavoritePet(John) = Fluffy$ to mean that the function "FavoritePet" maps the element "John" in the domain of pet owners to the element "Fluffy" in the domain of pets, a formal way of writing that Fluffy is John's favorite pet. (*Exercise*: would "Owns" with the same domain and co-domain constitute a function? why or why not?)

To indicate that a function f has domain A and co-domain B, we use the notation $f : A \to B$. In this case the function itself is f and the *value* of the function at x is denoted by $f(x)$. The value of a function at a particular element x of the domain is just the element of the co-domain where x is mapped.

The domain and co-domain of a function can be finite (like Pet Owners) or infinite (like natural numbers: $\mathbb{N} = 0, 1, 2, 3, etc.$). The domain and co-domain could be the same, but they of course do not have to be.

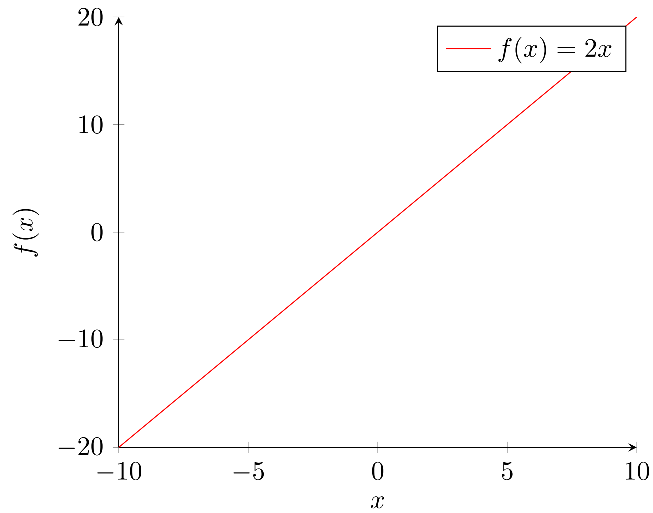
Sometimes functions can be described easily, like “FavoritePet” or the function $f : \mathbb{N} \to \mathbb{N}$ that takes a natural number and doubles it. We could call this function “Doubles”, but mathematicians prefer a more compact and precise description: $f(x) = 2x$. Here x is a variable and so the description of f above works for the entire domain, although the expression does not specify the domain explicitly.

Other times functions cannot be described easily at all but instead are simply a random-looking mapping from the domain to the co-domain. In this case, the only way to give a full description of the function is to give it’s value for every single input object, a task that might be (provably) impossible when there are an infinite number of input objects. [FOOTNOTE] But even in these cases, even if they defy easy description or understanding, the functions themselves are valid in a mathematical sense.

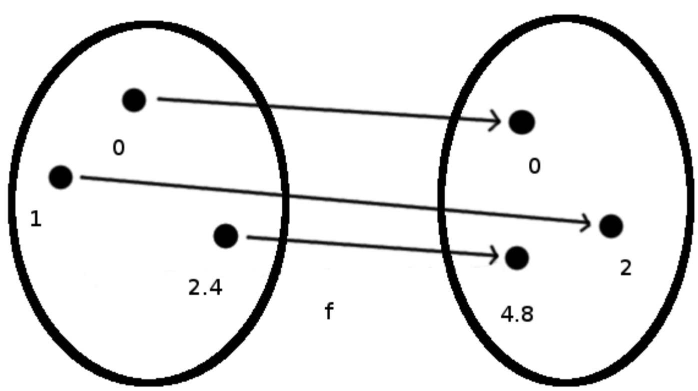
Domains or co-domains might be multi-dimensional spaces. For example, a function $f : \mathbb{N}^3 \to \mathbb{N}$ would map triplets like $(4, 5, 1)$ to natural numbers like 17774. Domains and co-domains can also be discrete, like the natural numbers. Or they can be continuous, in that they range over a continuous variable such as time. [FOOTNOTE: the mathematical definition of a continuous variable is subtle, but a very rough approximation is the notion of *density*: for a variable A, and any values x, y that can be taken on by A such that $x < y$, there is a value z that the variable A can take on and $x < z < y$. Exercise: show that this property does NOT hold for the natural numbers $\mathbb{N}$.]

How does this discussion of functions relate to the plots of sound waves above? It turns out that that familiar plot along an x and y axis is a representation of a function, where the x axis is the domain and the y axis is the co-domain.

After all, a function is an abstract object, and as such it has many representations. For example, the function $f(x) = 2x$ can be described by plotting it on a two dimensional axis, where the x-axis is the input of the function and the y-axis is the output:



It could also be represented with the primary visualization that we used to introduce the notion of a function:

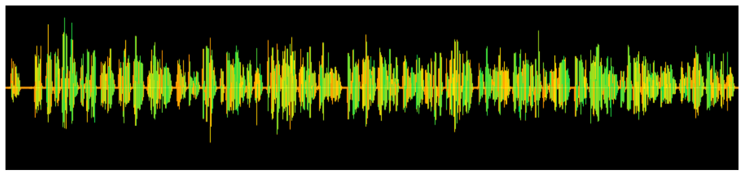


We will return to functions quite often, so it is important to develop a sharp understanding of this mathematical object.

Exercise: write out a mathematical description for a function f that takes two numbers, and raises the first to the power of the second, then subtracts the second. What is the value of $f(3, 4)$? As a bonus, use graphing software to create a graphical representation of this function as a 2-dimensional surface sitting in a three-dimensional space.

1.3 Speech on a wire: an analog approach

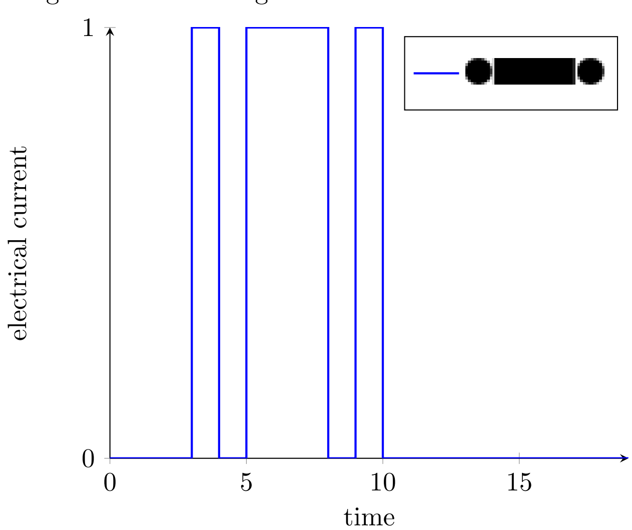
Above, we saw examples of functions that represented sound. Here again is our time domain representation of speech, also called a *waveform* of speech:



As we have seen, this is just a function in which the domain is time and the co-domain is air pressure. The information in the figure above is enough to create a sound that we would recognize as speech. In order for telephony to be possible, we need a way to continuously transmit the information of the sound waveform from one location to another, farther and faster than we could via sound waves (in other words, by just shouting).

An electrical wire does not contain air particles that can be compressed and uncompressed, but it does have other properties that can be varied continuously over time. In particular, the current flowing through a wire can increase and decrease over time.

Let's go back to the telegraph for a moment. In the case of the telegraph, the current was on or off, resulting in a function of electrical signal over time that might look something like this:



(Note that this figure is an idealized representation. In reality, the telegraph signal was distorted by the time it was received and so did not look quite so clean, but this serves as an idealized representation of the signal being sent and received. We will discuss distortions later on.)

The telephone, in contrast, has to solve the more difficult problem of translating an incoming speech waveform into an electrical signal and then back into a sound signal on the other end of the line.

[CHART OF CONCEPTUAL OUTLINE OF TELEPHONE]

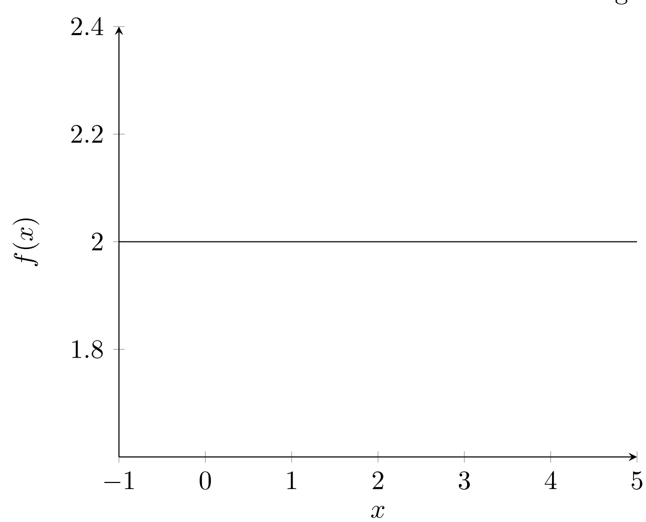
1.4 Basics of signals

We will see how the telephone accomplishes this shortly, but first let us turn our attention to the idea of a *signal*.

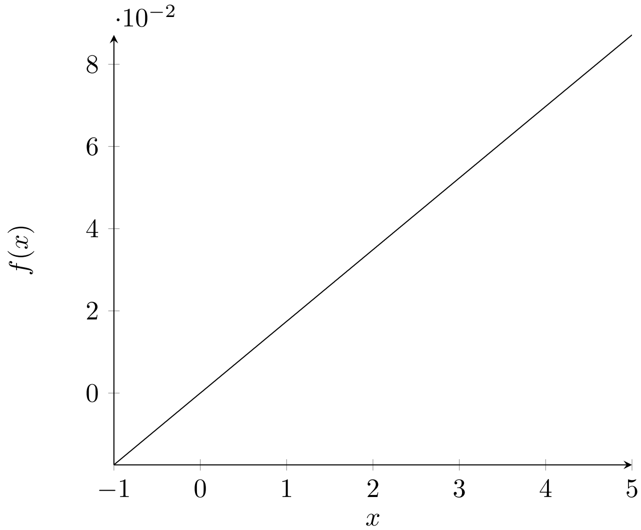
We have already covered basic symbolic systems, and the fact that any symbolic representation of information is interchangeable [ADD NAME FOR THIS THEOREM?]. But what happens when the information we care about is an analog function over time like the human voice that cannot be easily represented as a set of symbols?

In this case, we are instead dealing with a *signal*. While there is no universally agreed upon definition of a signal, for our purposes let us consider it to be a function in which the domain is time, and the co-domain is a continuous variable that bears information. This definition allows us to abstract away the information from the particular medium in which it resides. So, for example, we can consider the speech waveform above and its translation into electrical current to both be different representations of the same signal.

So far we have looked at a speech signal, and the signal produced by a tree hitting the ground. What about the sound of a C note that does not vary in pitch or intensity? Given the uniformity of experience associated with the monotone sound of a C note, you might expect at first blush that the signal associated with such a uniform sound to look like the function that graphs pitch over time:



However, a single C note in fact is better represented by a signal that looks like this:

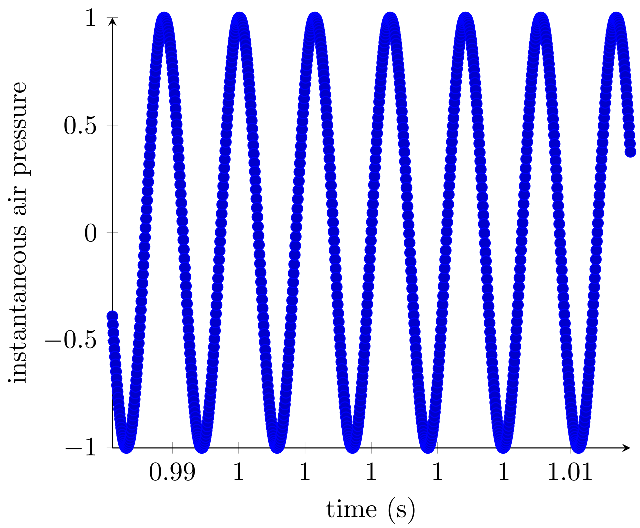


This signal is *periodic*. This very simple periodic function is sometimes called a *sine wave*, corresponding to the trigonometric sinusoidal function. [ADD FOOTNOTE EXPLANATION OF SINE?]. Most sound signals, including signals corresponding to voices, are actually a composition of many such periodic signals.

Periodic signals have a *frequency*, indicating the number of cycles per second, and an *amplitude* indicating the magnitude of the peaks and troughs of the signal. In the case of a sound signal, it turns out that changes in frequency correspond to changes in pitch. To give a concrete example, the note A above middle C in a piano creates the following signal:

$$f(x) = sin(880\pi x)$$

Here the domain is time in seconds, and we can graph the function:



The frequency of the function $sin(x)$ is 2π since that comprises a full rotation around the unit circle. Hence, the function $f(x)$ above corresponding to our A note has a frequency of $1/440$, meaning that there are 440 oscillations per second. The unit "per second" has a name, which is *hertz*, abbreviated *Hz*, and so we say that the frequency of an A note is 440Hz. If you are a musician with a basic knowledge of music theory, you probably know this fact.

So why do we hear an A or C note as a uniform tone? Keep in mind that the oscillations are whizzing by rather fast – 440 per second. Our auditory system is not designed to discern the individual peaks and troughs at such a low level of granularity, but rather to pick up on the frequency of oscillations. There is no particular reason for this, it just turned out to be more useful and practical from an evolutionary standpoint to detect patterns in how frequencies change. If our evolutionary history played out differently, we might have evolved instead to react in a completely different way to the constant changes in air pressure and the experience of "sound" might be radically different. Instead, as it turns out we became sensitized to rapid oscillating vibrations, the pitches they produce, and to surrounding phenomena like the changes from one rapid oscillating vibration to another, the combining of different vibrations, the slow change in frequency or amplitude of a vibration.

Moreover, given the sensitivity of the human auditory system to small oscillating vibrations, our vocal chords evolved to make sound by creating vibrating sinusoidal sound waves. But we are not single-note rookies. We have mastery over creating these sinusoidal waves and can start, stop, and with unconscious ease are able to quickly shift and alter the frequencies and amplitudes of the waves, resulting in the incredible feat of human speech.

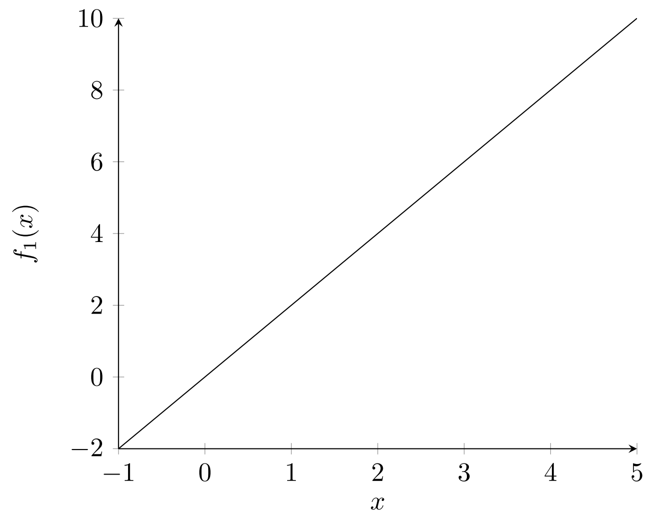
It may not seem obvious that the disorderly-looking speech waveform displayed above is made up of sinusoidal components. But that is only because as speakers and listeners we have a sophisticated innate mastery over producing and deciphering sounds, combining and adjusting different frequencies. Here is a zoomed in look at a section of the waveform over a very small time slice:

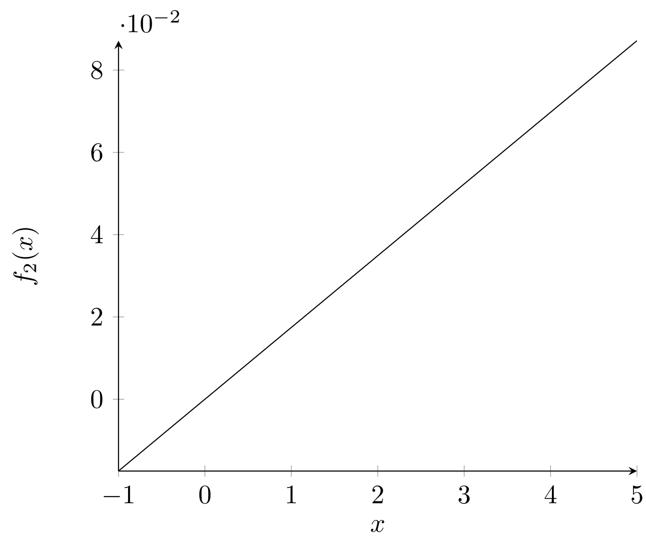
[SHOW SECTION OVER SMALL TIME SLICE]

As it happens, the mathematics behind combining these sinusoidal signals is quite important for telecommunication technology, and is a very straightforward extension of the addition operator you learned in elementary school called *point-wise addition*. Let's start with an illustration. Suppose you have the following two functions:

$$f_1 : \mathbb{N} \to \mathbb{N} f_2 : \mathbb{N} \to [-1, 1]$$

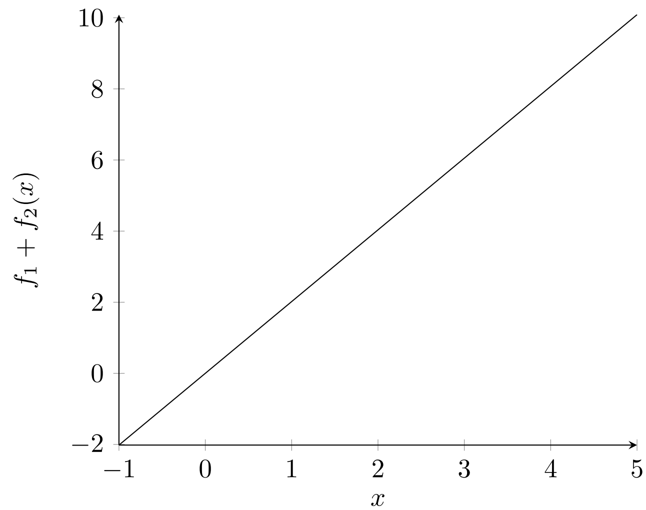
Let f_1 be defined by $f_1(x) = 2x$ and let $f_2(x) = sin(x)$.





So $f_1(4) = 8$, and $f_2(4) = sin(4) = -0.7568$ (rounding some decimal places). So what is $f_1 + f_2(4)$? Well let's just add the two values at that point: $f_1 +$

$f_2(4) = f_1(4) + f_2(4) = 8 + (-0.7568) = 7.2432$. By taking every element of the shared domain for f_1 and f_2 and adding the values for that element at f_1 and f_2 just as we have done for the element 4, we can construct the new function $f_1 + f_2$.



[TODO explicitly point out the value of this graph at 4?]

What does this have to do with sound waves or human speech? Well if there are two distinct vibrations happening from the same source, let's say a C note at XXX Hz and an A note at YYY Hz, then the resulting sound wave will be added together as above.

[ADD IMAGE]

This forms the basis for how signals can be combined, shifted, and generally manipulated to form speech waveforms.

[MORE HERE?]

Chapter 2

How A Telephone Works

The telephone transforms a sound signal into an electrical signal and back again.

But it was not the only game in town when it came to experimenting with sound. For example, the phonograph was invented in 1877 and is the ancestor of record players, cassette tapes players, CD players, and MP3 players. [FOOTNOTE: it seems this line of technology may have been subsumed by more general purpose desktop and mobile computers in the last 5 years or so]. The phonograph was aimed at solving a dual problem to the telephone. Both devices needed to capture sound. The difference is that once captured, the phonograph created a physical object from which the sound could be later retrieved, a record of the sound. Early phonographs used a circular disk as the record, and a sound signal was captured by deviations in a spiral groove. The telephone, in contrast, needed to quickly transport the sound to far away location, but did not need to make a record while doing so.

[DIAGRAM OF TELEPHONE VS PHONOGRAPH]

In this chapter, we will work towards building a telephone. It is worth mentioning that the telephone we build is not a historical replica, nor does it do justice to the incredible complexity of the telephone system in even the early 20th century. Instead, deliberate choices have been made to stick to the simple core concepts, borrowing freely from different historical periods to offer the simplest version for readers who are new to telephony, in keeping with the goal of developing a solid conceptual understanding of the technology.

2.1 Ohm's Law

Before embarking on a project to build any electrical device, it's important to understand a bit more about electricity. The most important principle for the purpose of building simple electrical components and devices is called Ohm's Law, an empirical law discovered in the early 19th century.

To motivate Ohm's law, note that there are two critically important electrical quantities that we can measure whenever electricity is flowing through a system.

The first quantity is the electrical *current*, commonly abbreviated I in the field of electrical engineering. This is a measure of the "flow" of the electricity through the circuit, analogous to the amount of water that might flow through a pipe system per second. Electrical current can be measured using a device called a *galvanometer*: a magnetic compass needle is deflected more or less depending on how much current is flowing through a wire. Just as we might choose a unit such as liters per second to describe the flow of water, a popular unit of measurement for electrical current is *amperes (amps)*.

The second quantity of interest is called *voltage*, abbreviated V. Borrowing again from the well-worn analogy of electricity to water flowing through a pipe, the voltage would be analogous to the pressure of the water in the pipe. The voltage is measured in *volts*, and can be measured independently of current, with the details of measurement depending on the type of circuit being used.

What is the relationship between current and voltage? Just as increasing the water pressure would increase the number of gallons per second flowing through the pipe system, so too an increase in voltage corresponds to a proportional increase in current. That much was well known as soon as people started tinkering with electricity. But until experiments are performed, the nature of this relationship is unclear. If we doubled the voltage of a circuit, would the current be doubled? Quadrupled? Or perhaps only increase slightly?

That is the question that Georg Ohm set out to answer through experimentation. He formed a circuit with some copper wire, varied the voltage, and measured the current. He did this with different wires of different thicknesses (the *gauge* of a wire is a standard measurement of diameter or, equivalently, cross sectional area). What he determined was that for any given wire, I was proportional to V. In other words, for a given wire $I = CV$ for some wire-specific constant C. This constant changed from wire to wire.

Here is a possible example illustrating the sorts of calculations of these quantities that Ohm performed almost 200 years ago that led us to Ohm's law:

[INSERT CHARTS]

How do we make sense of this data? Once we posit that each length of wire and each electrical component has a specific internal electrical *resistance* (which can be measured in the unit *Ohms*) then ALL of the data above can be condensed into a simple equation, which we call Ohm's Law:

$I = V/R$

The final term R stands for resistance, and is the property of a particular circuit that depends on physical properties like the type and thickness of wire, the physical properties of each electrical component along the path that the current travels, and the temperature of the system. This law is quite simple but also very powerful, and is an essential conceptual tool we'll use when designing electrical systems.

2.2 Electrical components

Before we build a telephone, let's take stock of what electrical components we have at our disposal and formalize our electrical circuit schematic diagrams a bit. We have already met the *relay*, the *switch* and the *power source* (e.g. a battery). Here is how these components are typically represented in circuit diagrams:

IAMHERE

[INSERT IMAGES]

A resistor is a component that provides electrical resistance and is written like so:

[RESISTOR]

Using this more formal electrical engineering notation, we would write a simple circuit with a switch like this:

[ADD IMAGE]

2.3 A one-way telephone

Now let's build a

A one-way telephone has a microphone at one end capable of translating an analog sound signal into an electrical one, and a receiver at the other end capable of translating the electrical signal back into an audible sound signal:

[DIAGRAM]

2.3.1 Microphone

A *microphone* serves as the input mechanism, translating our sound signal into a signal of electrical currents. The main type of microphone for years was the carbon microphone and it worked like this:

[INSERT IMAGE]

Much like our eardrums, a thin metal diaphragm is used that vibrates in response to sound waves. When it is pushed inward, as in the diagram above, the carbon granules are squeezed closer together, decreasing the resistance of the granules, and thus increasing the current flowing through the wire. This means that the wire now can carry the same signal as the sound waves.

You might think that the carbon granules have to be specially placed for this scheme to work, and imagine an engineer hunched over a telephone receiver carefully moving granules of carbon around with the precision of a caretaker working on a bonsai tree. But thankfully, this is not the case. We do not need to know the exact path that electrons might take for the macroscopic principle to hold: as the instantaneous air pressure increases, the diaphragm is pushed inward, packing the granules more tightly and decreasing their electrical resistance proportionally to the increase in air pressure.

2.3.2 Receiver

A *receiver* is a diaphragm attached to an electromagnet. As the electrical current goes up, the electromagnet has a stronger magnetic pull, causing the diaphragm to vibrate more. As it vibrates, it generates changes in air pressure which are emanated from the receiver as sound.

2.4 Telephones of the early 1900s

In the above conceptual diagram, the electrical circuit is always closed, meaning that current is flowing continuously through the circuit without any means to stop it (aside from cutting the cord). In practice, a mechanism is needed to open the circuit when it is not in use, which helps to save battery. It is also desirable for phones to work in the familiar two-way fashion in which a person can talk and listen at the same time. These engineering challenges are significant and interesting, but unfortunately beyond the scope of this text. Readers interested in learning more about telephony should consult one of the excellent historical texts or reference manuals that spare no detail [CITATIONS]

Chapter 3

The Telephone Network

The first telephones connected a pair of adjacent rooms, and so knowing who was at the other end of the line did not require much detective work. But the dream of industrialist Theodore Vail, president of AT&T from 1885 to 1889 and again from 1907 to 1919, was far more ambitious: universal telephone service for all.

[QUOTE FROM VAIL]

A proud and unabashed monopolist, and genuinely believed that a telecommunications monopoly was in the best interest of the country.

[QUOTE FROM VAIL]

This created a host of challenges: overcoming competition, getting the government to agree with the monopolistic approach to telephony, and establishing service in the most remote and rural parts of America. But alongside these business related challenges, a couple of technical issues loomed large when it came to building out a telephone network. First, how are people going to be identified? Second, how is person A going to be able to talk to person B?

3.1 Telephone Numbers

Naming was becoming a problem for the telegraph in the late 1800s. Until that point, if a telegram was addressed to John Smith in Boston, the message could properly be routed to the Boston office but then what? A lot of responsibility for getting this right fell on the sender of the message to appropriately specify a recipient (*John Smith of Commonwealth Ave in Back Bay*) and on the local courier boys who worked for the telegraph office to have some familiarity with the neighborhood and where people or businesses were located.

As cities grew and telegram volume grew, this ad hoc system was becoming unsustainable.

Chapter 4

Modern Digital Telephone

4.1 Information content and analog signals

How much information does a sound signal contain? Of course it's not fair to compare a hundred years of sounds to a millisecond, so to start to get a handle on the information content of a sound, let's fix our metric to be information *per second*.

But even now the answer is far from straight forward. A completely uniform sound intuitively has no information

[IAMHERE]