
Part II

Acoustic Phonetics

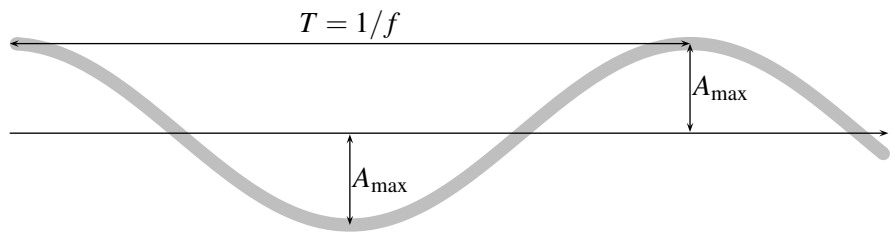
2.1 Sound Waves and Resonance

Pressure is the essence of sound. Pressure variations are called **compressions** (increases in pressure) and **rarefactions** (decreases in pressure). When the air pressure inside our ear changes, a thin membrane called the **ear drum** responds by vibrating in and out. These vibrations are transmitted to a chain of small bones called the **hammer**, **anvil**, and **stirrup**. The vibrations from the ear drum travel through the bones, causing the stirrup to bang against the **cochlea** (or **inner ear**), which is filled with fluid and tiny hairs. As the cochlea is jostled by the stirrup, the cochlear fluid sloshes around, causing the hairs to sway back and forth. These hairs are connected to the **auditory nerve**, which transforms the vibrations of the hairs into electrical impulses, which are then passed along to the brain for processing.

Any pattern of compression and rarefaction is called a **compression wave**. In particular, a compression wave that passes through an acoustic medium like air or water is called a **sound wave**. If the pattern of a wave repeats, it is called a **periodic wave**. Otherwise, it is called an **aperiodic wave**. Aperiodic waves come in two types: **white noise**, which is a continuous aperiodic wave that has a “hissing” sound, and a **transient**, which is an instantaneous aperiodic wave that has a “pop” or “click” sound. Sound waves found in speech can be either periodic or aperiodic, though more frequently, they are some combination of the two: technically aperiodic with some nearly periodic properties.

The amount of time it takes for a periodic wave to repeat itself is called its **period**, usually symbolized by T . The number of waves that pass by a given point in a specific amount of time is called the **frequency** of the wave, usually symbolized by f . Frequency is usually measured in **Hertz** (Hz), which is equal to 1 wave per second. The frequency of a sound wave corresponds to its pitch. Humans can hear in the range of about 20–20,000 Hz. The period and frequency of a wave are related by the formula $f = 1/T$.

The **amplitude** of a wave, symbolized by A , is the magnitude of variation in the wave’s pattern. For sound waves, the amplitude corresponds to loudness or volume. Since waves are patterns of variation by definition, the amplitude of a wave changes over time. The formula for a **simple wave** can be written as $A(t) = A_{\max} \sin(2\pi \cdot ft)$, where $A(t)$ is the amplitude at time t , A_{\max} is the maximum amplitude of the wave, and f is the frequency.

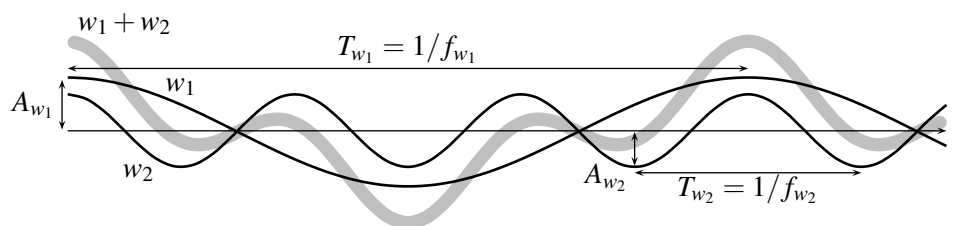


The physical space separating repetitions of a periodic wave is called the **wavelength** of the wave, symbolized by λ . Though frequency is usually the more important property of sound waves that we care about for phonetics, wavelength can play an important role. The **speed** of a wave, symbolized by s , is the rate at which a wave travels. Speed, frequency, and wavelength are related by the formula $s = f\lambda$. The speed of a wave depends on the nature of the medium it is traveling through: density, temperature, elasticity, etc. For warm, moist air, as is typically found in the human mouth, the speed of sound is about 35,000 cm/sec.

Waves may not be simple however, and most speech sounds are actually **complex waves**, combinations of multiple simple waves. The component waves of a complex wave each have their own frequencies and amplitudes, and they are added together to yield the complex wave. The general formula for a complex wave is:

$$A(t) = A_{w_1} \sin(2\pi \cdot f_{w_1} t) + \dots + A_{w_n} \sin(2\pi \cdot f_{w_n} t)$$

where $A(t)$ is the amplitude of the complex wave at time t , A_{w_1} and f_{w_1} are the maximum amplitude and the frequency of the first simple component wave w_1 , up through A_{w_n} and f_{w_n} , which are the maximum amplitude and the frequency of the n th simple component wave w_n .



In addition to their component frequencies, complex waves also have a **fundamental frequency**, usually symbolized as f_0 . (Note: for simple waves, we often say that their frequency f is their fundamental frequency f_0 ; i.e., $f_0 = f$.) The fundamental frequency of a complex wave is calculated by finding the **greatest common factor** of the set of all of the component frequencies. That is, $f_0 = \text{gcf}(f_{w_1}, \dots, f_{w_n})$. The gcf is the largest number that divides all of the components frequencies evenly. Use traditional methods from middle school mathematics to find the gcf of frequencies that are integers. If the frequencies are not integers, multiply them all by the same integer N to turn them into integers, then find the gcf of the new numbers, and finally, divide that result by N to find the actual gcf. For example, if the component frequencies are 44.4 Hz and 55.5 Hz, multiply both frequencies by 10 to get 444 and 555. The gcf of 444 and 555 is 111. Divide this by 10 to undo the previous multiplication, and the result is 11.1 Hz as the fundamental frequency.

All sound waves naturally produce extra sound waves at higher frequencies called **harmonics**. The first harmonic is the sound wave with a frequency equal to the fundamental frequency of the original sound wave. In the case of a simple wave, the simple wave is its own first harmonic. The n th harmonic has a frequency equal to n times the fundamental frequency. The amplitude of the harmonics falls off rather quickly. The maximum amplitude of the first harmonic is equal to the maximum amplitude of the original sound wave, while the maximum amplitude of the n th harmonic is equal to $1/n^2$ times the maximum amplitude of the original sound wave. Putting this all together, the formula for the n th harmonic is $A_n = \frac{1}{n^2} A_0 \sin(2\pi \cdot n f_0 t)$.

Sound waves can be generated by any vibrating object. The natural frequencies at which an object vibrates are called its **resonant frequencies**, which depend on the shape of the object and the material it's made of. For a body of air in a **tube closed at both ends** with a length L , the first resonant frequency is $s/2L$, and in general, the n th resonant frequency is $ns/2L$. For a body of air in a **tube closed at one end and open at the other** with a length L , the first resonant frequency is $s/4L$, and in general, the n th resonant frequency is $(2n - 1)s/4L$.

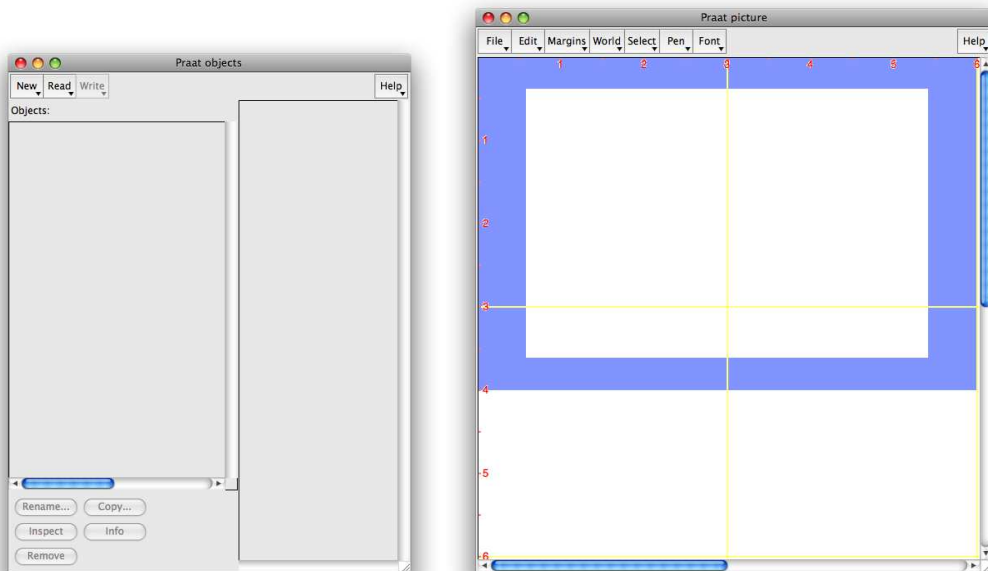
Objects can act as **filters** for sound waves by enhancing waves with particular frequencies and dampening others. Generally, objects acting as filters enhance frequencies near their own resonant frequencies and dampen frequencies farther away from their resonant frequencies. The frequencies

that are enhanced the most by a filter are called its **center frequencies**. Filters usually have a **bandwidth**, which is the size of the range of frequencies around the center frequency that are also significantly enhanced. For example, if a filter has a center frequency of 500 Hz and a bandwidth of 100 Hz, then frequencies between 450 Hz and 550 Hz will be enhanced, while all others will be dampened.

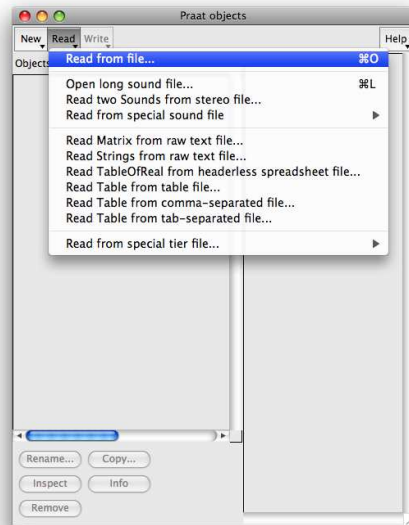
The idea of filtering plays an important role in acoustically analyzing speech, since the vocal tract acts as a filter. By approximating the various cavities in the vocal tract as tubes, we can give a rough but accurate estimate for the acoustic properties of different speech sounds based on how they are articulated. See the discussion on vowel acoustics beginning on page 22 for an implementation of tube models for vowels.

2.2 Praat Basics

When you open Praat, you'll get two windows, the Praat objects window and the Praat picture window:

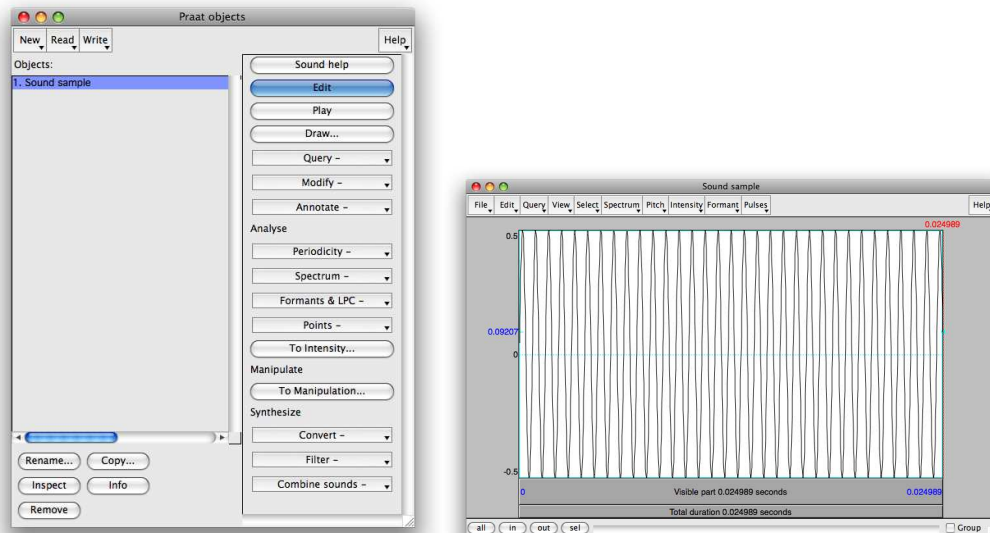


For most purposes in this course, you will only need the Praat objects window, so close the Praat picture window. From the Praat objects window, you can open previously recorded sound files, such as those for labs or recorded by you for your final project, by selecting Read from file... from the Read menu:

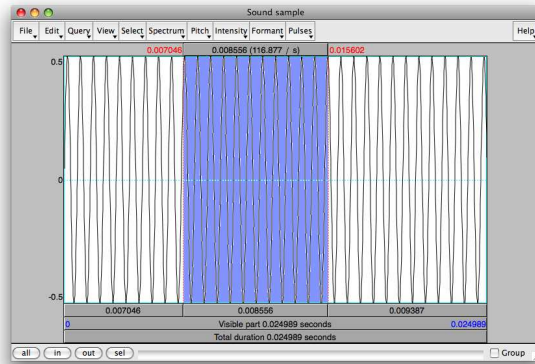


This will open up a window allowing you to select a file from your computer. Select the file you want and click Choose. This places the file in the list of objects in the Praat objects window and gives you a lot of new functions in the right panel of the window.

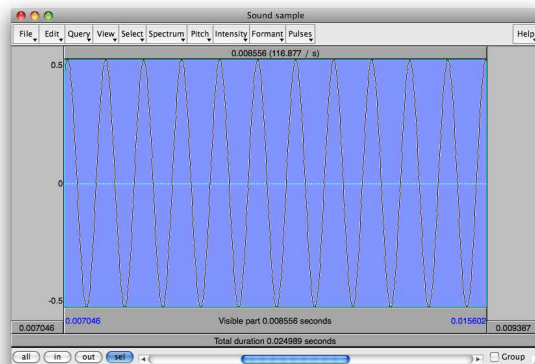
One of the most important functions now available is Edit, which will open up a new window showing the the selected sound as a waveform:



From this new window, you can manipulate the sound, zoom in and out, and most importantly, take measurements of various aspects of the sound. You can select a portion of the wave simply by clicking and dragging the cursor to highlight the portion you want to select.

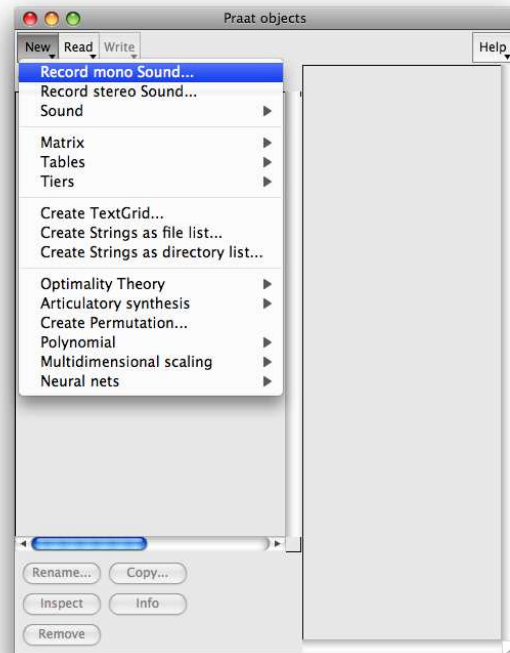


Note how Praat tells you the duration of the selected portion (here, 0.008556 sec). This is how you can measure how long any piece of a sound is. Once you've selected a portion you want to examine, you can click `sel` to zoom in on that portion:

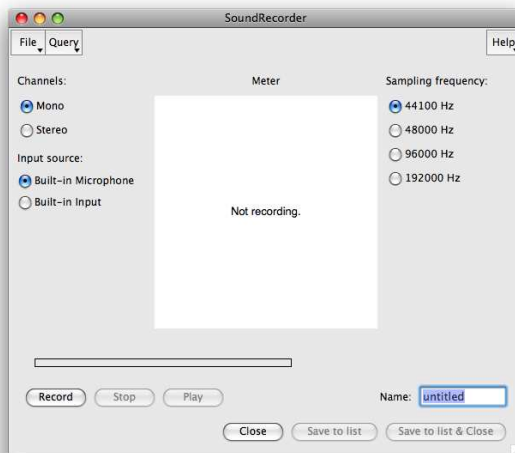


You'll learn how to use many of the options in this window as part of your labs.

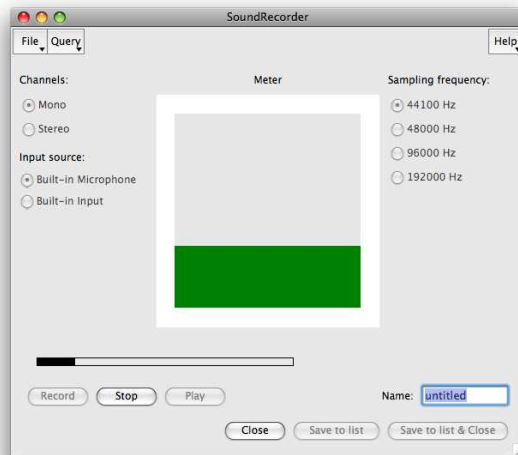
You can also create your own new sounds directly in Praat, rather than having to have access to a previously recorded sound file. To record someone speaking, be sure your computer has some sort of audio input, such as a built-in microphone. Select `Record mono Sound...` (you'll have no need for recording stereo sounds):



This will open up a new window, the Praat SoundRecorder:

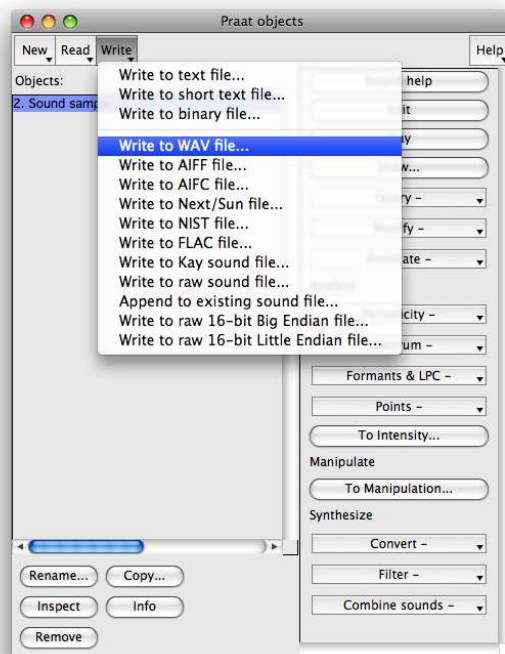


When you are ready to record, press **Record**, which will start recording all sound directed at the microphone. In the middle of the window, you can monitor the microphone's response to the incoming sound. If your sound source is too loud or too close to the microphone, the meter may get into the red, which will likely mean your file will be distorted. As long as you stay in the green and yellow, your sound file should be fine:



When you are finished recording, press **Stop**, and the recording will stop. Clicking **Play** will let you review the recording, and if it's satisfactory (note that you can easily cut out unwanted portions later), you can give it a name and press **Save to list** to create a new sound object in the Praat objects window. From there, you can use the **Edit** function to open the sound up in its own window and examine, edit, and measure it.

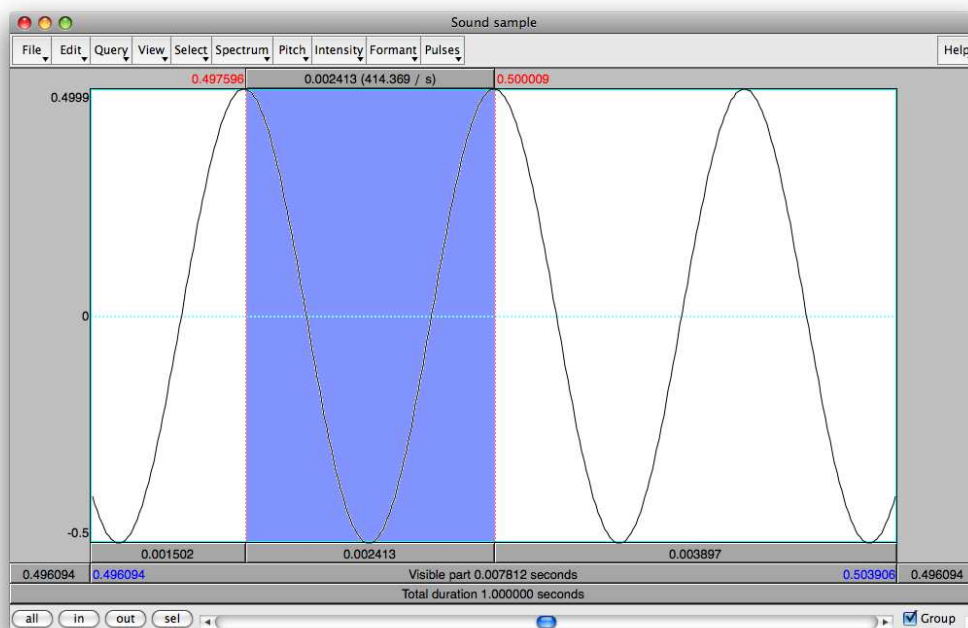
Finally, when you are finished and ready to quit Praat, if you want to save your newly recorded and edited sounds to a file, select the sound you want to save and then select **Write to WAV file...** (or any other format of your choice), which will open up a window allowing you to choose a location (and a new name if you want) for your sound file:



2.3 Sidebar: Measurement and Uncertainty

Any time we collect data, there will always be some limits to the precision and accuracy of our results due to a number of factors, including our own abilities, the nature of the equipment we're using, and the natural random variation inherent to the physical universe. It is important to be aware of these factors and to report our results in a way that properly represents our conclusions, by not only giving an answer, but also an estimate of how far off we think that answer could be.

For example, suppose we wish to estimate the period T of a given sound wave from a WAV file. We would begin by opening the WAV file in Praat and measuring T as the time between some pair of adjacent peaks, as in the picture below, in which T has been measured to be 0.002413 sec:



However, as stated above, there are many ways in which this measurement could be wrong. We might not have placed the cursor exactly on the center of each of the two peaks (in fact, given the nature of the physical world, it's pretty much guaranteed that we never can). Or, when the sound file was created, there could be random fluctuations, file corruption, or other anomalies during the portion we measured. Or, the sound wave simply might not be consistent throughout its duration (which is generally going to be the case when we make measurements of actual human speech).

Thus, to get a better estimate for the true value of T , we should take multiple measurements. The more measurements we make, the more convincing our final result will be. Suppose we take the following five measurements (do not round any raw data!):

i	1	2	3	4	5
T_i (sec)	0.002413	0.002424	0.002439	0.002411	0.002434

In ordinary circumstances, the reported result should be the **mean** of these values. The mean is usually notated by putting a bar over the symbol for the measured value; in this case, the mean is \bar{T} . The mean is calculated simply with the ordinary average of the measurements, i.e., the sum of all of

the measurements, divided by how many total measurements were made. In Excel, the mean of an array of cells like C31:C67 can be calculated with the formula `=AVERAGE(C31:C67)`. The mean of the five measurements in the table above would be:

$$\bar{T} = \frac{0.002413 \text{ sec} + 0.002424 \text{ sec} + 0.002439 \text{ sec} + 0.002411 \text{ sec} + 0.002434 \text{ sec}}{5} = 0.0024242 \text{ sec}$$

The mean is a better estimate of T than one individual measurement by itself, because using multiple measurements helps reduce the effects of some sources of error, giving us more certainty that our estimate of T is close to its true value. To report our certainty of how accurate our estimate is, we use the **standard error of the mean** (notated here as $\text{s.e.}(\bar{x})$ for a mean \bar{x}), calculated as follows:

$$\text{s.e.}(\bar{x}) = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n \cdot (n - 1)}} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n \cdot (n - 1)}}$$

where x_i is the i th individual measurement, \bar{x} is the mean of the measurements, and n is the total number of measurements. The standard error of the mean provides us with the upper and lower bounds on a specific kind of certainty of our estimate, with $\bar{x} - \text{s.e.}(\bar{x})$ being the lower bound and $\bar{x} + \text{s.e.}(\bar{x})$ being the upper bound. Statistically, we are about 68% confident that the true value of the quantity we are measuring is in this interval, and 95% confident that it is in $\bar{x} \pm 2 \cdot \text{s.e.}(\bar{x})$. The standard error of the mean is equivalent to dividing the sample standard deviation (the unbiased estimator using Bessel's $n - 1$ correction) by \sqrt{n} . In Excel, the standard error of the mean of an array of cells like C31:C67 can be calculated with the formula `=STDEV(C31:C67)/SQRT(COUNT(C31:C67))`. Note that we do not report the standard deviation by itself. Our measurements are an attempt to look at a larger reality: when we take measurements of tokens of a vowel, we aren't interested in just those five measurements on their own, but rather, we want to know what those measurements can tell us about the pronunciation of the vowel in general, beyond what we measured directly. In this case, our measurements are not the entire set of possible pronunciations (which is when the standard deviation would be appropriate), but are a random sample from a larger set of possible measurements (which is when the standard error of the mean of the measurements is appropriate, as it characterizes the variance of the larger set). The standard error of the mean of the five measurements from before would be:

$$\begin{aligned} \text{s.e.}(\bar{T}) &= \sqrt{\frac{\sum_{i=1}^5 (T_i - \bar{T})^2}{5 \cdot 4}} = \sqrt{\frac{(T_1 - \bar{T})^2 + \cdots + (T_5 - \bar{T})^2}{20}} \\ &= \sqrt{\frac{\overbrace{(0.002413 \text{ sec} - 0.0024242 \text{ sec})^2}^{T_1} + \cdots + \overbrace{(0.002434 \text{ sec} - 0.0024242 \text{ sec})^2}^{T_5}}{\bar{T}}}{20}} \\ &= 0.00000554437 \dots \text{ sec} \end{aligned}$$

Of course, this statistic is itself only an estimate and therefore has its own inherent uncertainty, so we don't need such a precise value in our final report of our results. As a general rule, for the type of work we do in this course, the standard error of the mean can safely be rounded to the first non-zero digit, which yields $\text{s.e.}(\bar{T}) = 0.000006 \text{ sec}$ in this case. Correspondingly, we report the mean itself rounded to the same decimal position as the rounded standard error of the mean (in this case, the 6th position), giving us $\bar{T} = 0.002424 \text{ sec}$.

Finally, we report both the mean of our measurements and the standard error of the mean together, in the form $\bar{x} \pm \text{s.e.}(\bar{x})$. For our example here, we would report that our estimate for T is $0.002424 \pm 0.000006 \text{ sec}$ (or $2.424 \pm 0.006 \text{ msec}$).

If we need to plug this complex value into some other formula in order to calculate a different quantity, the uncertainty represented by the standard error of the mean **propagates** through to the final derived result. Various formulas for calculating propagation of uncertainty are given below (a is any constant with no uncertainty of its own, and \bar{x} and \bar{y} are means with standard errors $\text{s.e.}(\bar{x})$ and $\text{s.e.}(\bar{y})$; note that propagated uncertainty is always positive, so you can ignore sign changes):

function	propagated uncertainty	function	propagated uncertainty
$a \cdot \bar{x}$	$a \cdot \text{s.e.}(\bar{x})$	\bar{x}^a	$a \cdot \bar{x}^{a-1} \cdot \text{s.e.}(\bar{x})$
$\bar{x} + \bar{y}$	$\sqrt{(\text{s.e.}(\bar{x}))^2 + (\text{s.e.}(\bar{y}))^2}$	$a^{\bar{x}}$	$a^{\bar{x}} \cdot \ln a \cdot \text{s.e.}(\bar{x})$
$\bar{x} - \bar{y}$	$\sqrt{(\text{s.e.}(\bar{x}))^2 + (\text{s.e.}(\bar{y}))^2}$	$e^{\bar{x}}$	$e^{\bar{x}} \cdot \text{s.e.}(\bar{x})$
$\bar{x} \cdot \bar{y}$	$\bar{x} \cdot \bar{y} \cdot \sqrt{\left(\frac{\text{s.e.}(\bar{x})}{\bar{x}}\right)^2 + \left(\frac{\text{s.e.}(\bar{y})}{\bar{y}}\right)^2}$	$\log_a \bar{x}$	$\frac{\text{s.e.}(\bar{x})}{\bar{x} \cdot \ln a}$
\bar{x}/\bar{y}	$\frac{\bar{x}}{\bar{y}} \cdot \sqrt{\left(\frac{\text{s.e.}(\bar{x})}{\bar{x}}\right)^2 + \left(\frac{\text{s.e.}(\bar{y})}{\bar{y}}\right)^2}$	$\ln \bar{x}$	$\frac{\text{s.e.}(\bar{x})}{\bar{x}}$
		$\sin \bar{x}$	$(\cos \bar{x}) \cdot \text{s.e.}(\bar{x})$
		$\cos \bar{x}$	$(\sin \bar{x}) \cdot \text{s.e.}(\bar{x})$
		$\tan \bar{x}$	$(1 + (\tan \bar{x})^2) \cdot \text{s.e.}(\bar{x})$

For example, using our estimated value for T , we can estimate the frequency f of the sound wave by plugging \bar{T} into the formula for frequency: $f = 1/T$. Since $1/\bar{T}$ is equivalent to \bar{T}^{-1} , we need to propagate the uncertainty for \bar{T}^a , with $a = -1$. This means that the propagated uncertainty is $-1 \cdot \bar{T}^{-2} \cdot \text{s.e.}(\bar{T})$ (or simply $\bar{T}^{-2} \cdot \text{s.e.}(\bar{T})$, since uncertainty is always a positive number):

$$\begin{aligned}
 f &= \frac{1}{\bar{T}} \pm \overbrace{\bar{T}^{-2} \cdot \text{s.e.}(\bar{T})}^{\text{uncertainty for } 1/\bar{T}} \\
 &= \frac{1}{0.0024242 \text{ sec}} \pm (0.0024242 \text{ sec})^{-2} \cdot 0.00000554437 \text{ sec} \\
 &= 412.5072189 \pm 0.943379268 \text{ Hz (raw)} \\
 &= 412.5 \pm 0.9 \text{ Hz (reported)}
 \end{aligned}$$

Note that the original full raw values for \bar{T} and $\text{s.e.}(\bar{T})$ are plugged in, not the rounded values. Similarly, if we want to calculate the wavelength of this sound wave, we would plug \bar{T} into the formula $\lambda = s \cdot T$, where s is the constant 35,000 cm/sec, and propagate the uncertainty for $a \cdot \bar{T}$, with $a = s$. This means that the uncertainty for $35,000 \text{ cm/sec} \cdot \bar{T}$ is $35,000 \text{ cm/sec} \cdot \text{s.e.}(\bar{T})$:

$$\begin{aligned}
 \lambda &= 35,000 \text{ cm/sec} \cdot \bar{T} \pm \overbrace{35,000 \text{ cm/sec} \cdot \text{s.e.}(\bar{T})}^{\text{uncertainty for } 35,000 \text{ cm/sec} \cdot \bar{T}} \\
 &= 35,000 \text{ cm/sec} \cdot 0.0024242 \text{ sec} \pm 35,000 \text{ cm/sec} \cdot 0.00000554437 \text{ sec} \\
 &= 84.847 \pm 0.19404 \text{ cm (raw)} \\
 &= 84.8 \pm 0.2 \text{ cm (reported)}
 \end{aligned}$$

It's also important to keep in mind the **relative standard error**, which is how big the standard error of the mean is in relation to the mean, calculated by dividing the standard error of the mean by the mean itself, expressed as a percentage. For example, even though both $500 \text{ Hz} \pm 100 \text{ Hz}$ and $5000 \text{ Hz} \pm 100 \text{ Hz}$ have the same standard error of the mean, in the first case, this error is very significant (20%), while in the second case, it is much less significant (2%). When you need to compare how accurate two separate values are, it is better to compare the relative standard error.

But why bother with all of this?!? What's wrong with just using the old-fashioned significant digit rules that we learned in high school? Because real scientists are intellectually honest, and they report their actual, real uncertainty. If we want to do real science, we should report our real uncertainty, too. Significant digit rules are useful for quick and easy estimates, but for serious, professional quality scientific research, we should report our results correctly and accurately, which includes reporting our uncertainty about our results as rigorously as possible.

Note that significant digit rules are much less informative than standard errors of the mean and propagated uncertainty. Consider a value reported simply as 11.0 cm, arrived at through significant digit rules. This could represent any number from 10.95 cm to almost 11.05 cm, implying a result of 11.00 ± 0.05 cm. But this could be much more or much less certain than we are actually justified in reporting. We could be drastically overstating our certainty (maybe the result based on the data we collected is really 11.0 ± 0.5) or drastically understating it (11.00 ± 0.01). Using significant digit rules completely masks the true precision of our results, and since we have access to the information necessary to calculate our precision, we have an obligation to report it.

Furthermore, the significant digit rules oversimplify the effects of uncertainty propagation. With larger sets of data or deriving values from more complex formulas, we could very well end up with wildly incorrect results if we relied solely on significant digit rules and plugging measured values into formulas, without using the proper formulas to calculate the propagated uncertainty.

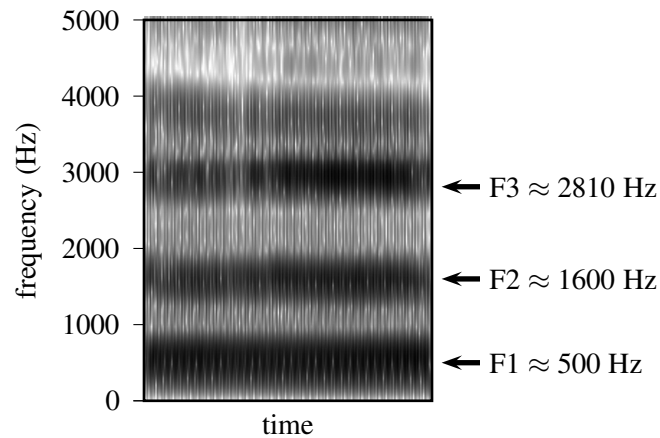
2.4 Vowel Acoustics

Since the vocal tract is roughly tube-shaped, we can predict the acoustic properties of speech sounds by creating a **tube model** of the mouth, making assumptions and approximations of the shape of the vocal tract based on articulation. For a mid-central vowel like [ə], the tube is basically uniform from the lips to the glottis. Since any sufficiently small opening can be approximated as a closure, this means that the articulation of [ə] can be approximated as a tube that is open at one end (the lips) and closed at the other (the glottis):

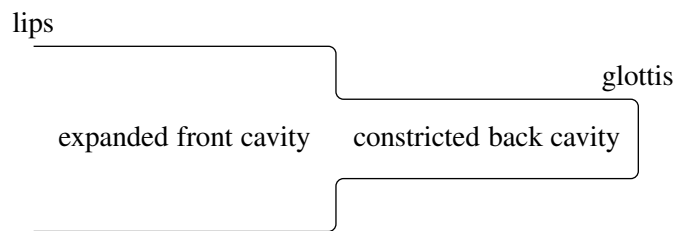


The various resonant frequencies of such a tube are given by the formula $(2n - 1)s/4L$. A value of 17.5 cm for L is typical of adult males, which means the first three resonant frequencies of this model of the mouth are 500 Hz, 1500 Hz, and 2500 Hz. For vowels, resonant frequencies are called **formants** and are usually abbreviated as F1, F2, F3, etc.

Formants can be visualized in a special type of three-dimensional graph of the sound wave. If we plot time on the x -axis versus frequency on the y -axis, then showing intensity as darkness results in a **spectrogram**. Vertical striations in a spectrogram of speech correspond to glottal vibrations; the closer these striations are, the higher the pitch. Formants show up in spectrograms as dark bands, as in the following sample spectrogram for one pronunciation of [ə]:

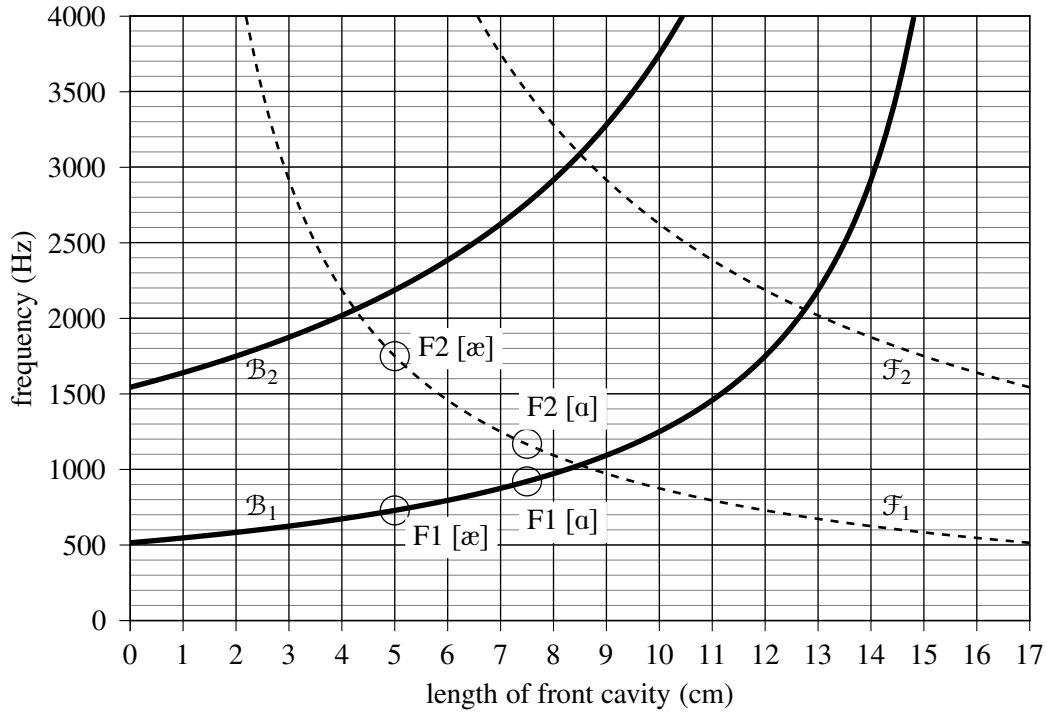


Vowels other than [ə] involve constrictions in the oral tract, so our simple tube model must be made more complex to account for these constrictions. When a low vowel like [æ] or [ɑ] is made, the tongue lowers in the front of the mouth, creating an expanded front cavity, in contrast to the back cavity, which is narrower, partly because of its comparative size with the front cavity, but also because as the tongue lowers, it bunches up in the back, which constricts the throat:



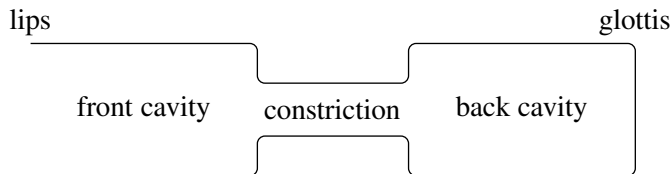
The difference between [æ] and [ɑ] lies primarily in the lengths of the cavities. For a front vowel like [æ], the length of the front cavity is short and the back cavity is long. In comparison, [ɑ] is a back vowel, so the front cavity is longer and the back cavity is shorter. Each of these two tubes affect the sound wave coming from the glottis. At a rough level of approximation, they each separately filter the sound wave with their own resonant frequencies. The final sound wave will have formants that are a simple mixture of those it would get from each tube individually. F1 for the vowel will be the lowest of all of the resonant frequencies for both tubes, F2 will be the second lowest of all of the resonant frequencies, and so on.

We can see this visually in a graph called a **nomogram**, which shows the acoustic effects of a particular articulation. In the following nomogram for a 17 cm vocal tract, the length of the front cavity is measured on the x -axis (this marks the transition between the front and back cavities), while the resulting resonant frequencies are on the y -axis. Each line on the graph represents one of the resonant frequencies of one of the two tubes. The resonant frequencies of the front cavity are drawn in thin, dotted lines and labeled \mathcal{F}_1 and \mathcal{F}_2 , while the resonant frequencies of the back cavity are drawn in thick, solid lines and labeled \mathcal{B}_1 and \mathcal{B}_2 . Note that both tubes are open at one end, and essentially closed at the other, so their resonant frequencies are given by the formula $(2n - 1)s/4L$:



The constriction for [æ] occurs about 5 cm into the mouth. As we can see on the graph above and plugging the appropriate values into the correct formulas, this results in $F1 \approx 729$ Hz and $F2 \approx 1750$ Hz. For [ɑ], the constriction is about 7.5 cm into the mouth, resulting in $F1 \approx 921$ Hz and $F2 \approx 1170$ Hz. Note that these values are only approximations; the mouth is much more complex, and the interactions between the two tubes are not this simple (e.g., acoustic coupling occurs alters the true frequencies somewhat). However, for yielding a rough estimate, this model works well and can give us general information about the formants for low vowels.

Modeling a high vowel results in a very different configuration of tubes. When the tongue raises to create the vowel constriction, the front cavity remains relatively wide, but so does the back cavity. This results in a three-tube model, with a small tube corresponding to the constriction itself:



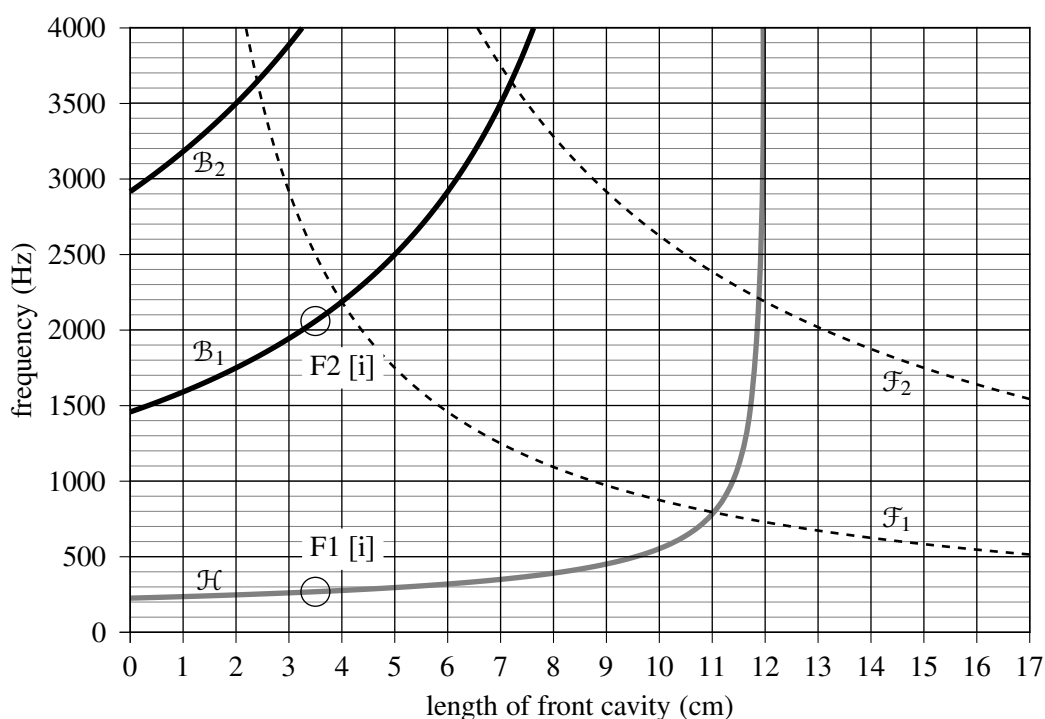
When a thin tube is attached to one end of a large tube that is otherwise closed at both ends, the resulting structure is called a **Helmholtz resonator**. The back cavity combined with the constriction can be approximated as a Helmholtz resonator. A Helmholtz resonator has its own resonant frequency independent of its component tubes. This frequency is given by the following equation:

$$f_H = \frac{s}{2\pi} \sqrt{\frac{A_1}{A_2 \ell_1 \ell_2}}$$

In the Helmholtz formula, A_1 is the cross-sectional area of the thin tube, A_2 is the cross-sectional area of the large tube, and ℓ_1 and ℓ_2 are their lengths. Like the other tubes in the model, the

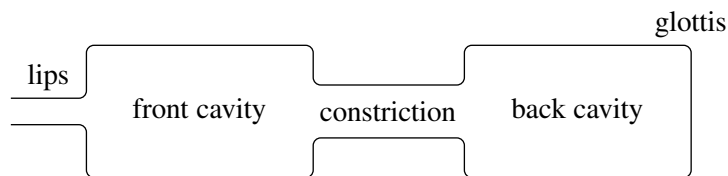
Helmholtz resonator also contributes to the filtering properties of the mouth, and thus, to the creation of formants. For our models of the mouth, we can simplify this formula to $f_H = \alpha s / \sqrt{\ell_1 \ell_2}$, where ℓ_1 and ℓ_2 are as before, and α is a parameter that, for our purposes, ranges approximately between 0 and 0.16. When the constriction is more open, α is larger, so f_H is higher. Conversely, for a narrower constriction, α is smaller, so f_H is lower.

In the following nomogram, the resonant frequencies of the front cavity are drawn in thin, dotted lines and labeled \mathcal{F}_1 and \mathcal{F}_2 , while the resonant frequencies of the back cavity are drawn in thick, solid lines and labeled \mathcal{B}_1 and \mathcal{B}_2 . The Helmholtz frequency is drawn in a thick grey line and labeled \mathcal{H} . Note that front cavity is open at one end, and essentially closed at the other, so its resonant frequencies are given by the formula $(2n - 1)s/4L$, but unlike for the low vowels, the back cavity for high vowels is essentially closed at both ends, so its resonant frequencies are given by the formula $ns/2L$. In this nomograph, the length of the constriction is set to 5 cm, and α is set to 0.05:

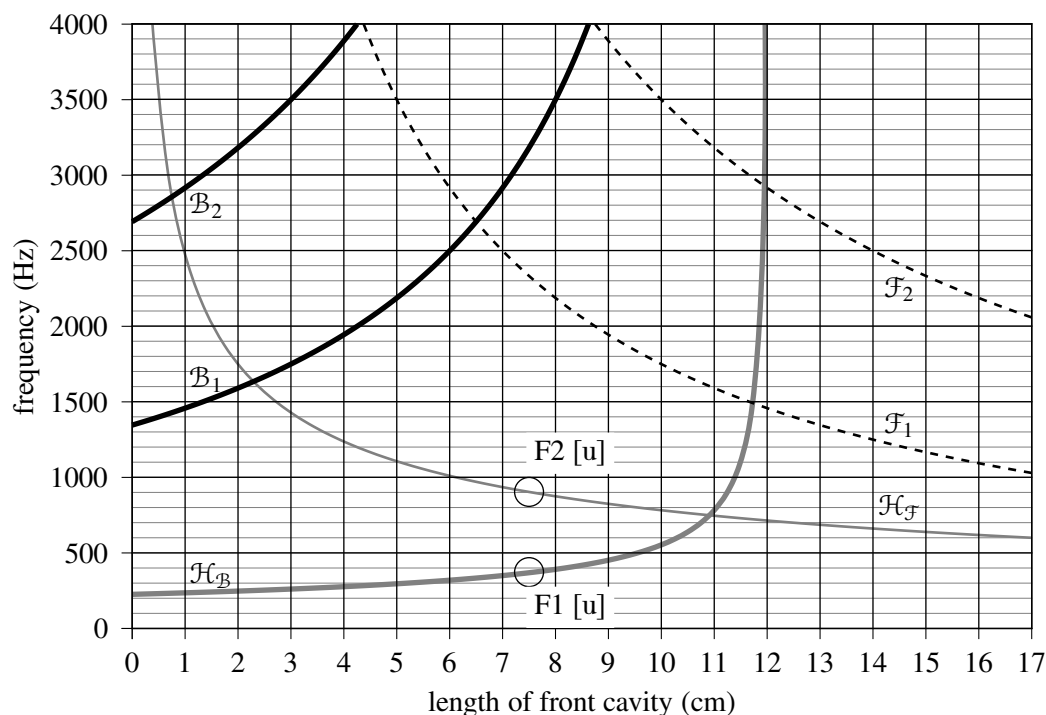


The constriction for [i] begins about 3.5 cm into the mouth. Using the nomogram and the correct formulas, we obtain $F1 \approx 268$ Hz and $F2 \approx 2060$ Hz.

For [u], the picture becomes more complicated. The lips are rounded, which has two effects: it turns the front cavity into a fully closed tube, and it creates a second Helmholtz resonator in the front of the mouth:



The following nomogram shows graphs for the resonant frequencies of the front cavity (thin, dotted lines, \mathcal{F}_1 and \mathcal{F}_2), the resonant frequencies of the back cavity (thick, black lines, \mathcal{B}_1 and \mathcal{B}_2), the Helmholtz frequency from the coupling of the back cavity with the tongue constriction (thick grey line, \mathcal{H}_B), and the Helmholtz frequency from the coupling of the front cavity with the lip constriction (thin grey line, \mathcal{H}_F), with the tongue constriction fixed at 5 cm, the lip constriction fixed at 0.5 cm, and α for both Helmholtz resonators fixed at 0.05:



With [u]’s constriction starting around 7.5 cm, we get $F1 \approx 373$ Hz and $F2 \approx 904$ Hz.

From all of these nomograms, we can see a few apparent patterns and make some predictions. Vowel backness is very closely linked to F2: front vowels have a higher F2, while back vowels have a lower F2. Vowel height is very closely linked to F1: low vowels have a higher F1, while high vowels have a lower F1. These predictions are borne out when we measure actual vowels, as in the following typical sample values for F1 and F2 for some vowels of English:

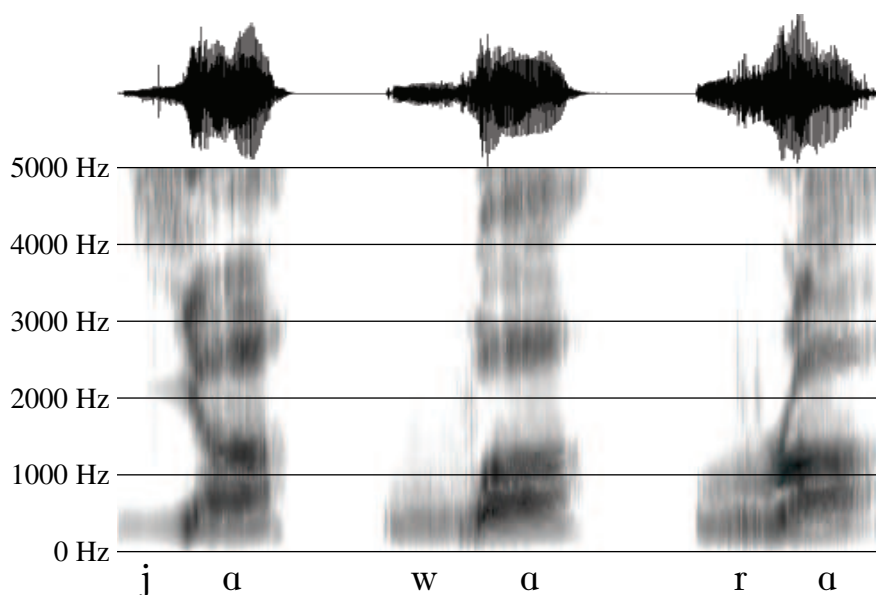
	[i]	[ɪ]	[ε]	[æ]	[ɑ]	[ɔ]	[ʊ]	[u]
F1	300	400	550	700	700	600	450	300
F2	2200	1900	1750	1700	1100	900	1000	850

F3 doesn’t vary much in the English vowel system, so it isn’t very useful for distinguishing vowels in English. However, rhotacized vowels have a characteristically low F3. This is not a fact that is easy to obtain from tube models, but can be derived more easily by other models. See discussion of perturbation theory beginning on page 37 for more information.

2.5 Consonant Acoustics: English

Acoustically, central approximants look much like their corresponding vowels but with a shorter duration and lower amplitude:

[j] ≈ [i]	low F1, high F2, high F3
[w] ≈ [u]	low F1, low F2, low F3
[r] ≈ [ɹ]	low F1, low F2, very low F3

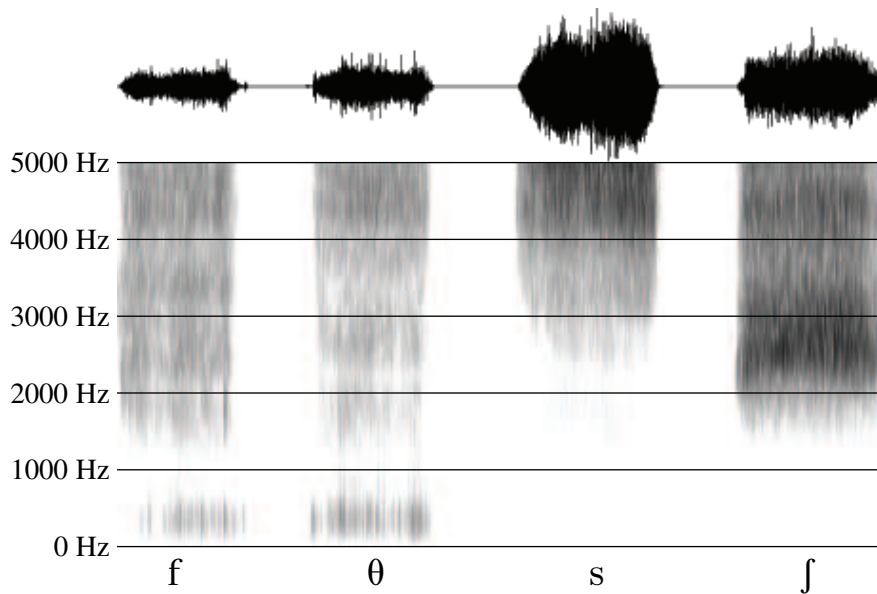


Diffuse fricatives have turbulent airflow scattered as it hits an obstruction (for [f] and [θ], this obstruction is the teeth and/or outside air), causing mostly uniform white noise over a wide range of frequencies; there is very little obvious acoustic difference between diffuse fricatives [f] and [θ]. Diffuse fricatives tend to have low amplitudes, because they lose a lot of acoustic energy from hitting the obstruction.

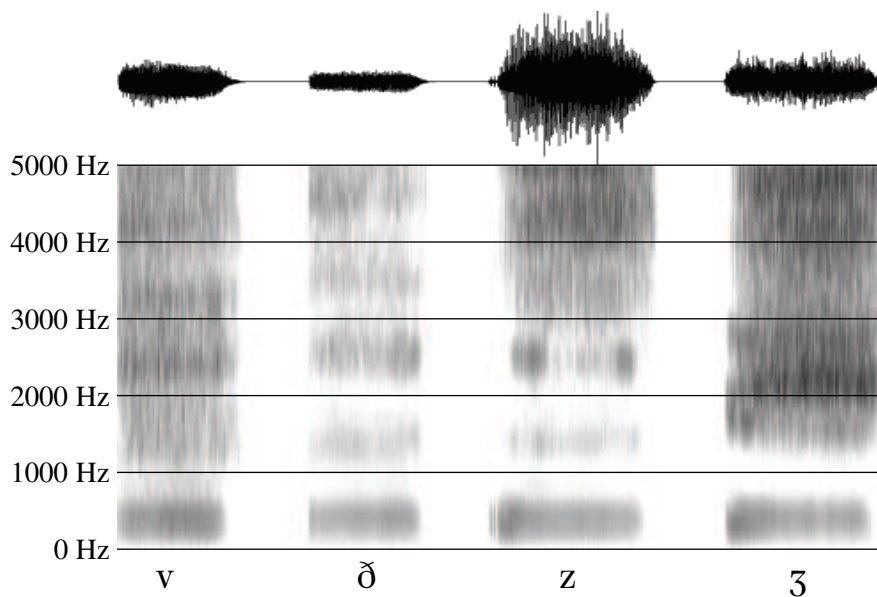
Compact (or sibilant) fricatives have turbulent airflow focused through a narrow channel, causing high frequency (≥ 2000 Hz) white noise dependent on the length of the channel:

[s]	short channel	higher frequencies (> 3000 Hz)
[ʃ]	long channel	lower frequencies (2000–3000 Hz)

Because compact fricatives have focused airflow and no significant obstruction, their acoustic energy remains high, so they tend to have high amplitudes, making them noticeably louder than diffuse fricatives; this is the reason we use [ʃ] rather than [f] to get the attention of a noisy room.



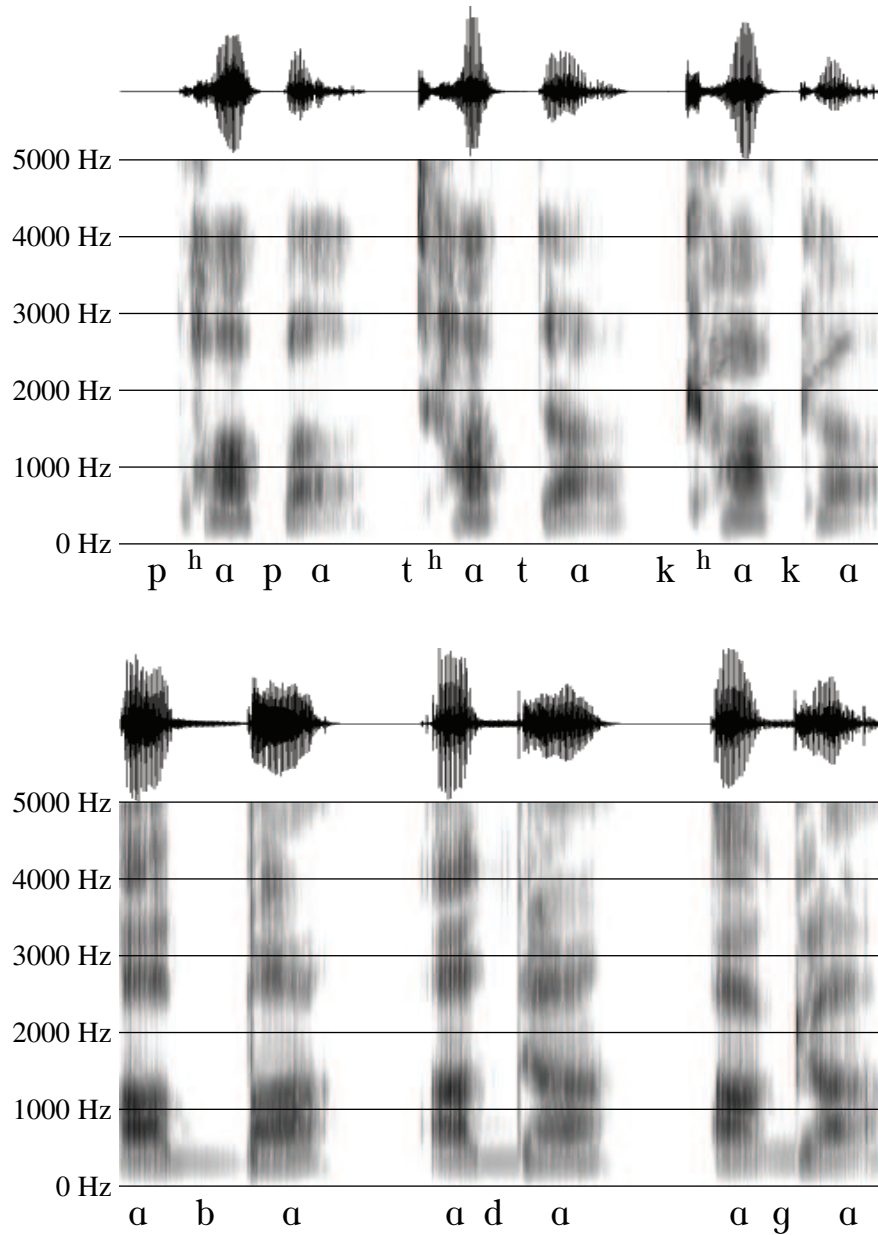
Voicing in fricatives creates a buzzing that behaves like the resonance of a vowel, making voiced fricatives rather complex sounds with both formant-like periodicity and aperiodic white noise.



Oral stops have a full **stop closure** followed by a **release burst**. The closure is silent for voiceless stops, while for voiced stops, some or all of the closure has low amplitude and low frequency periodic vibration, which can be seen as a **voicing bar** in a spectrogram. The duration of a stop closure can depend on place of articulation, especially for voiced stops: [g] has a closure far back in mouth, so air pressure behind the closure builds faster than for a stop made more forward in the mouth, so it's harder to maintain the closure. The release burst is a transient that resonates through the remaining tube and can acquire some formants. A longer remaining tube (e.g., for [k g]) creates lower formants in the release burst than a shorter tube does (e.g., for [t d]). There are typically no formants in the release burst for [p b], since there is no remaining tube for the burst to resonate in.

Stop constrictions also cause **formant transitions** for the formants of neighboring vowels. All stops have F1 transitions that decrease away from the vowel and into the stop closure. For a bilabial stop constriction, F2 and F3 of neighboring vowels are also lower near the stop closure. For a velar stop constriction, F2 of neighboring vowels is higher near the stop closure, while F3 is lower, creating the so-called **velar pinch** as F2 and F3 approach each other (and sometimes merge). An alveolar stop constriction generally causes F3 of neighboring vowels to be higher or steady near the stop closure and F2 of neighboring vowels to approach 1750 Hz near the stop closure.

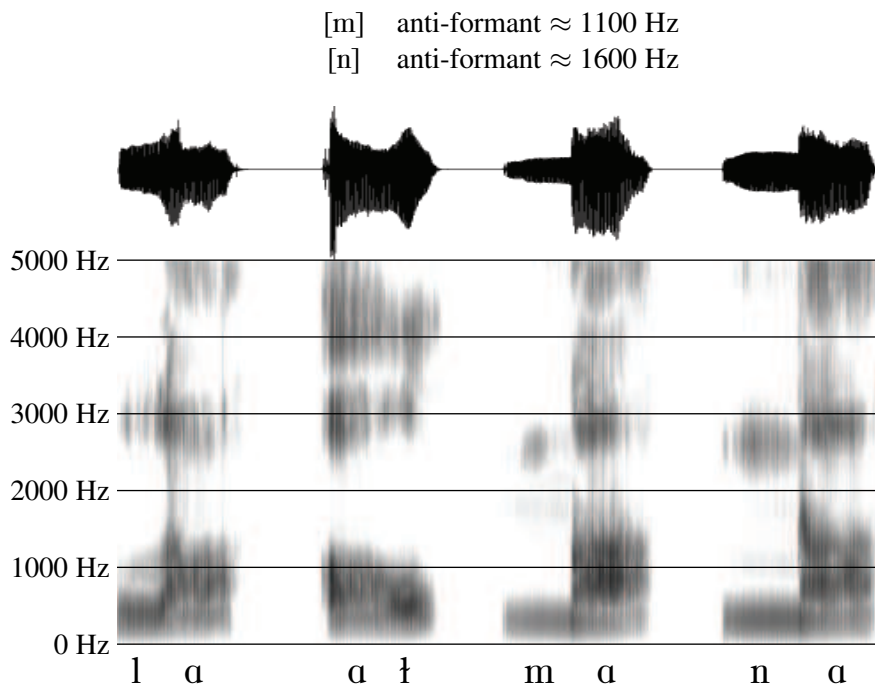
Voiceless stops in certain positions have **aspiration**, which is a period of voicelessness after the release burst. Aspiration looks like white noise overlaying faint formants for the following sound.



An affricate has essentially the same acoustic properties as a stop followed by a fricative, except the total duration is about the same as a stop by itself (i.e., affricate duration is shorter than the duration of a true stop followed by a true fricative).

A lateral approximant has [ə]-like formants, but they can be more [u]-like, especially with a very low F2 when velarized. Lateral approximants also have an **anti-formant** around 1900 Hz. An anti-formant is a negative resonance (dampening) created by side tubes. It appears as noticeable white space in a spectrogram.

Nasal stops have faint vowel-like formants around 300 Hz, 1100 Hz, and 1900 Hz. Nasals may have formant transitions similar to oral stops, but not always. In addition, nasals often have anti-formants based on place of articulation:



2.6 Consonant Acoustics: Beyond English

The **airstream mechanic** of a sound is the source of its airflow and the direction it flows. The default airstream mechanic in speech is **pulmonic egressive**: the lungs are the source of the airflow, and the air flows outward from the lungs. The vast majority of sounds in the world's languages are made with a pulmonic egressive airstream mechanic. It is possible to have a pulmonic **ingressive** airstream mechanic, with airflow being sucked into the lungs, but this is rarely used in speech.

There are two major sources of non-pulmonic airflow that are possible in human languages: **glottalic** (with closure at the glottis, along with the main closure elsewhere, creating a pocket of air as the airflow) and **velaric** (with closure at the velum creating a pocket of air). Physically, both of these kinds of airstream mechanics could be either egressive or ingressive, though velaric airstream mechanics tend to be ingressive only in speech; both egressive and ingressive glottalic airstream mechanics are found in speech.

Recall from the discussion of phonation on page 3 that phonation is the state of the vocal cords during a sound, based on how abducted (open) they are. The following table summarizes the five phonation types, and gives IPA symbols for sounds made with each:

	glottal stop	creaky voice	modal voicing	breathy (murmured)	voiceless
glottal opening	fully closed	mostly closed	medium	mostly open	maximally open
vibration	none	some	maximal	some	none
IPA symbol	ʔ	ɓ	b	ɓ or b ^{fi}	p

Oral stops are divided into different types, based on their airstream mechanic. **Plosives** are egressive pulmonic oral stops, which have a single closure in the oral tract; these are the most common type of oral stops in the world's languages. As discussed earlier (beginning on page 28), English oral stops, which are all plosives, are characterized by a closure (while air flows into the oral tract from the lungs) and a release burst (after the closure is released, and the built-up air pressure is quickly equalized as the air flows out of the mouth). All plosives are thus characterized by a period of silence or near silence during the closure, and a short transient burst when the stop is released.

The acoustic properties of the closure of a plosive depend primarily on its phonation type. If the plosive is voiceless (unaspirated or aspirated), then the closure is silent. In the waveform of a sound, the closure will just be a flat line, while on a spectrogram, it will be essentially blank. If the plosive is voiced (whether creaky, modal, or breathy), the vocal cords will vibrate during the closure, which will show up as a voicing bar. Because modal voiced sounds involve maximum vibration, their closure voicing will typically be louder than for murmured or creaky sounds, which shows up in the waveform as higher amplitude vibration and in the spectrogram as a darker voicing bar. Due to resonance in the mouth, the glottal vibrations during the closure of a voiced sound can create very faint formant-like resonance, if they are said particularly loudly or recorded with very sensitive equipment.

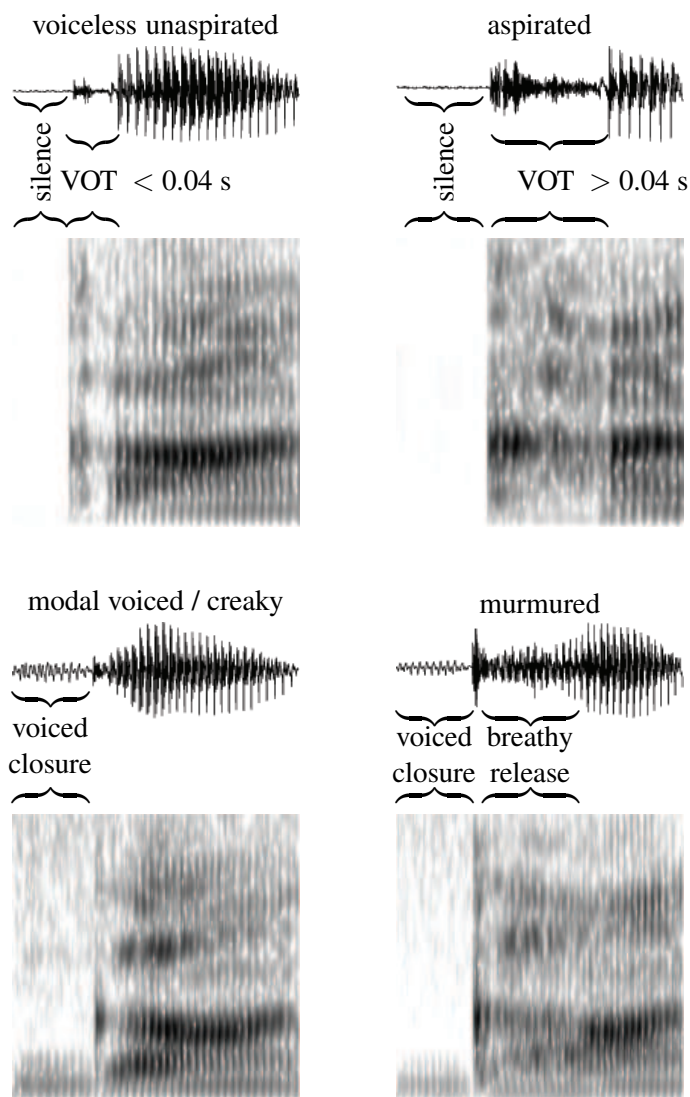
As also discussed for English plosives, the closure duration for a plosive can depend on its place of articulation: the more forward the plosive closure is, the longer it can be kept closed while air pressure increasingly builds up behind it. However, plosives are typically so short that there is usually no opportunity for air pressure build-up to make a difference in difficulty in keeping the closure, so this is not a reliable acoustic property.

A plosive's release burst can also be affected by place of articulation: the farther back the plosive closure is, the more quickly air pressure builds up behind it, and the higher the pressure is, the louder the release will be as the air pressure is equalized. Thus, a velar plosive will generally have a louder release burst than an alveolar plosive, which in turn will generally have a louder release burst than a bilabial plosive.

The release burst also resonates in the remaining portion of the mouth, which enhances certain frequency ranges within the burst. The farther back in the mouth the plosive closure is, the longer the tube is that the release burst will resonate in. Longer tubes have lower frequencies, so a velar plosive's release will generally be louder at lower frequencies than an alveolar plosive's release. Bilabial plosives have no remaining tube to cause resonance, but the release can be reflected back into the mouth, causing some [ə]-like resonance.

Voiceless plosives can be further distinguished by their **voice onset time (VOT)**, which is how long it takes from the release for the vocal cords to move close enough together to allow voicing to occur. If the VOT is long (typically greater than about 0.04 sec), the sound is **aspirated**. During this period of time, the sound produced is very turbulent, creating aperiodic white noise. If the VOT is shorter than about 0.04 sec, **unaspirated** or **plain**, which means that it has very little noticeable white noise. If VOT is negative, voicing starts before the release, which means the sound is voiced. It is possible for voicing to only occur during part of the closure, in which case, the sound is **partially voiced**. Partially voiced sounds are sometimes more precisely described by giving the percentage of the closure duration that has identifiable voicing.

Murmured plosives also usually have a long, noisy release, but it coincides with voicing from vocal cord vibration. This breathy release looks like a combination of the following vowel with the white noise seen in the aspirated release of an aspirated plosive. However, the white noise in a breathy release is softer, and the formants of the following vowel are darker and easier to see. Creaky plosives look much like modal voiced plosives, though the frequency of the closure voicing tends to have a lower frequency and a lower amplitude.



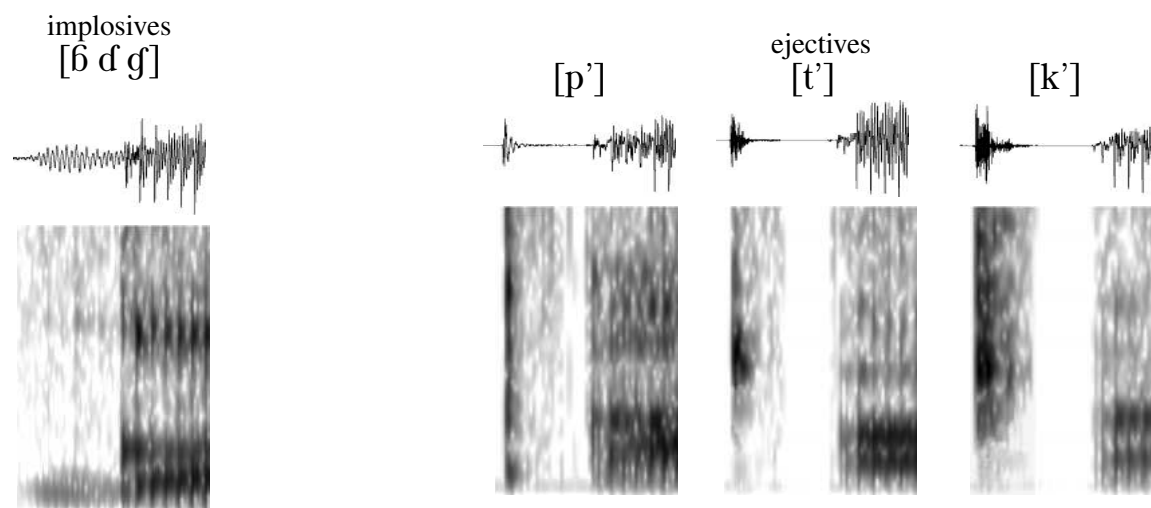
place	bilabial	alveolar	velar
burst intensity	softest	medium	loud
burst frequencies	none / [ə]	high	low
closure duration	longest	medium	short

Implosives [ɓ ɗ ɠ] are ingressive glottalic stops, which are made by trapping air between two closures: one at the primary place of articulation and one at the glottis. The larynx then lowers, which decreases the pressure of the trapped air. However, it is difficult to keep the glottis completely closed during this maneuver, because air naturally flows out of the lungs as we speak. Thus, some air tends to leak through the glottis, causing it to spontaneously vibrate as it lowers, making implosives naturally voiced (though they can be forced to be voiceless).

Once the larynx has lowered, the closures are released, and air flows into the mouth from outside before the airstream shifts to its normal pulmonic egressive flow. Because the air pressure does not change as much as for plosives, an implosive release is much softer than for plosives. Acoustically, implosives sound similar to voiced plosives, but have a weaker release, and an “odd” sound because of the ingressive airflow.

Ejectives [pʼ tʼ kʼ] are egressive glottalic stops, which are made by trapping air between two closures: one at the primary place of articulation and one at the glottis. The larynx then raises along with the normal pulmonic airstream behind it. This raising of the glottal closure increases the air pressure of the trapped air. The primary stop closure is then released, and the high pressure air rushes out of the mouth (the glottal closure remains closed to prevent the air from traveling backwards). Then the glottis opened, and the pulmonic airstream behind it is released.

Ejectives are characterized acoustically by two transients: one for the oral release and one for the glottal release. The silence separating the releases can be 0.05 sec or longer. The intensity and resonance of the first release are the same as for plosives, while the second (glottal) release is much weaker and shows no resonance or formant transitions.



Clicks are ingressive velaric stops, which are made by trapping air between two closures: one at the primary place of articulation and one at the velum. The tongue then lowers, which decreases the pressure of the trapped air. Because the trapped air is so small, it is very easy to change its pressure significantly. Then the closures are released, and air rushes into the low pressure area.

Clicks are characterized acoustically by a very loud transient (due to the very high pressure differential) followed by a period of silence as the tongue moves into position for the following vowel and the airflow switches from ingressive to egressive. Acoustically, they are somewhat similar to ejectives, but clicks only have a single release burst, not two.

Bilabial clicks [ɔ̥] are less intense than vowels, because the soft cheeks absorb much of the sound. It doesn't take long for air to fill the cavity, because the lips can move quickly and create a large opening, so the transient for a bilabial click is short (~ 0.01 sec). It also doesn't take long for the tongue to move into position for the following vowel, because the front of the tongue is not used in producing a bilabial click and can be in position already, so bilabial clicks also have a short silence after the transient ($\leq \sim 0.02$ sec). Because of the dampening effects of the cheeks, bilabial clicks have spectral peaks over a range of high and low frequencies.

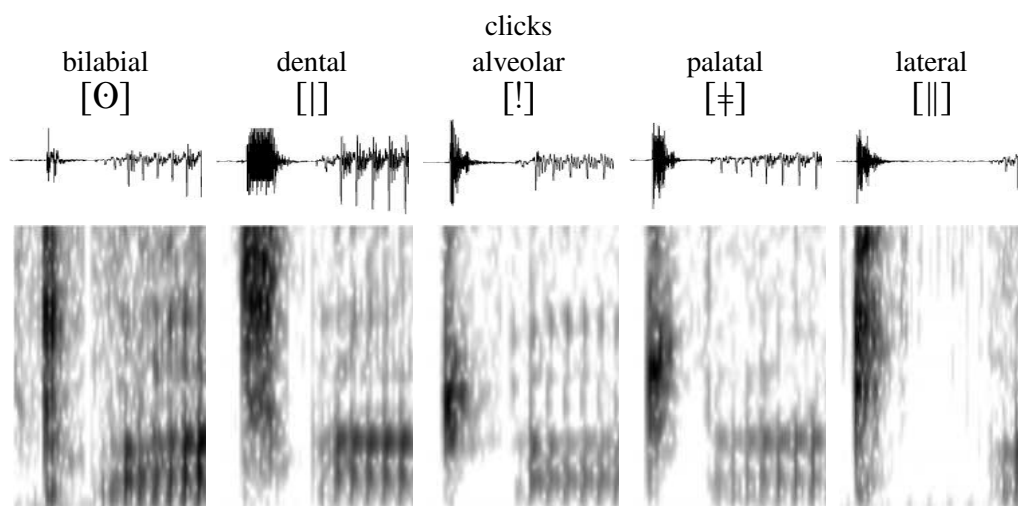
Dental clicks [t̪̥] are less intense than vowels, because the teeth diffuse the incoming air. It takes longer for air to fill the cavity, because it must travel around the teeth, so the transient for a dental click is long ($\geq \sim 0.02$ sec). It also doesn't take long for the tongue to move into position for the following vowel, because the front of the tongue is released almost directly into a vowel-like position, so dental clicks have a short silence after the transient ($\leq \sim 0.02$ sec). Because the click cavity is very flat, dental clicks have high frequency spectral peaks ($> \sim 2000$ Hz).

Alveolar clicks [t̪̥] are more intense than vowels, because the sound echoes off the hard palate instead of the cheeks. It doesn't take long for air to fill the cavity, because the tongue tip can move out of the way quickly, so the transient for an alveolar click is short (~ 0.01 sec). It also doesn't take long for the tongue to move into position for the following vowel, because the front of the tongue is released almost directly into a vowel-like position, so alveolar clicks have a short silence after the transient ($\leq \sim 0.02$ sec). Because the click cavity is very deep, alveolar clicks have low frequency spectral peaks ($< \sim 2000$ Hz).

Palatal clicks [t̪̥] are more intense than vowels, because the sound echoes off the hard palate instead of the cheeks. It doesn't take long for air to fill the cavity, because the tongue tip can move out of the way quickly, so the transient for a palatal click is short (~ 0.01 sec). It also doesn't take long for the tongue to move into position for the following vowel, because the front of the tongue is released almost directly into a vowel-like position, so palatal clicks have a short silence after the transient ($\leq \sim 0.02$ sec). Because the click cavity is very small, palatal clicks have high frequency spectral peaks ($> \sim 2000$ Hz).

Lateral clicks [ɬ̪̥] are more intense than vowels, because the sound echoes off the hard palate instead of the cheeks. It takes longer for air to fill the cavity, because it must travel around the tongue, so the transient for a lateral click is long ($\geq \sim 0.02$ sec). Unlike other clicks, it also takes comparatively longer for the tongue to get into position for the following vowel, because after the initial lateral release, the tongue tip is still touching the alveolar ridge, so lateral clicks have a long silence after the transient ($\geq \sim 0.05$ sec). Because the click cavity is very flat, lateral clicks have high frequency spectral peaks ($> \sim 2000$ Hz).

	bilabial [ɸ]	dental [θ]	alveolar [t]	palatal [ç]	lateral [ɬ]
transient intensity	soft		loud		
transient duration	~0.01 s	≥ 0.02 s	~0.01 s	≥ 0.02 s	
post-transient silence	≤ 0.02 s				≥ 0.05 s
spectral peaks	varied	> 2000 Hz	< 2000 Hz	> 2000 Hz	



Fricatives can be either compact or diffuse, or some combination of both. In a spectrogram, fricatives look like white noise, often with regions of higher intensity, known as **spectral peaks**.

The turbulence in compact fricatives is created by focusing the airstream through a narrow channel, creating high intensity white noise with spectral peak frequencies that depend on the length of the resonance tube formed by the remaining portion of the mouth after the end of the fricative channel. Shorter resonance tubes result in higher frequency spectral peaks. This tube can be lengthened (thus, causing the spectral peak frequencies to lower) by raising the front of the tongue to create a **sub-lingual cavity**, forcing the airstream to travel a longer distance before leaving the mouth.

Alveolar fricatives [s z] are mostly compact and a very short resonance tube (alveolar ridge to the lips), resulting in high frequency spectral peaks.

Post-alveolar fricatives [ʃ ʒ] are mostly compact and have a longer resonance tube (behind the alveolar ridge to the lips), resulting in lower frequency spectral peaks than for alveolar fricatives. In addition, the raised tongue tip creates a sub-lingual cavity, so the spectral peaks are at even lower frequencies than expected for a post-alveolar place of articulation.

Retroflex fricatives [ʂ ʐ] are much like post-alveolar fricatives, with a longer resonance tube (behind the alveolar ridge to the lips), and a raised tongue tip creating a sub-lingual cavity. In addition, retroflex fricatives have a curled tongue, which lengthens the sub-lingual cavity even more, so the spectral peaks are at even lower frequencies than for post-alveolars.

Palatal fricatives [ç ʝ] are mostly compact and have a longer resonance tube (hard palate to the lips), resulting in lower frequency spectral peaks than for alveolar fricatives. However, palatal fricatives lack a sub-lingual cavity, so their spectral peak frequencies are higher than for post-alveolar fricatives, despite being formed farther back in the mouth.

Uvular fricatives [χ ʁ] are mostly compact and have a long resonance tube (uvula to the lips), resulting in lower frequency spectral peaks than even for retroflex fricatives. They are somewhat more dampened than more forward fricatives because of energy loss in the airstream as it travels further.

For diffuse fricatives, turbulence is made by focusing the airstream against an obstruction that reduces its overall intensity, flattening the spectral peaks. In contrast with compact fricatives, spectral peak frequencies for diffuse fricatives depend largely on the length of the fricative channel, rather than the length of the remaining mouth, because there is little resonance after the channel, because of the severe dampening effect of the obstruction.

Bilabial fricatives [ɸ β] are mostly diffuse, with the open air outside the mouth acting like a wall. The fricative channel formed by the lips creates mild spectral peaks across a range of frequencies.

Labiodental fricatives [f v] are mostly diffuse, with the teeth and open air acting like a wall. The fricative channel formed by the teeth and lower lip is shorter than the bilabial channel because the teeth are thinner than the upper lips, resulting in higher frequency spectral peaks than for bilabial fricatives.

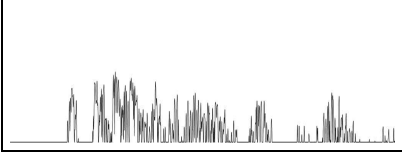
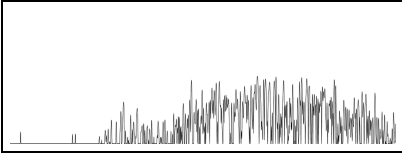
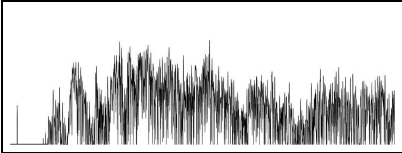
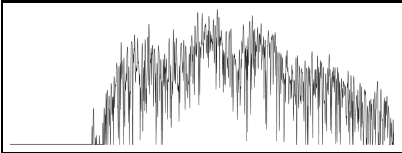
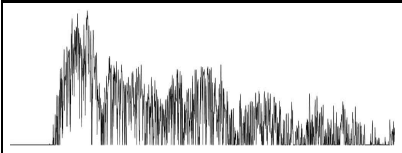
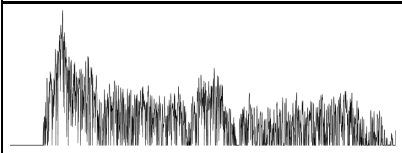
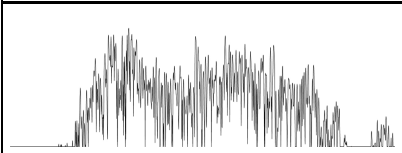
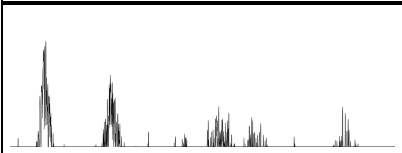
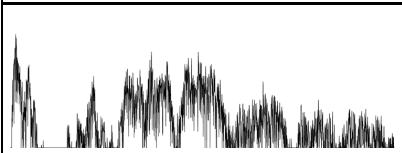
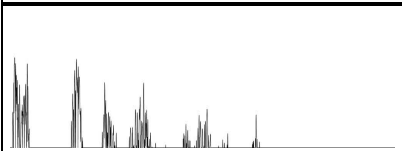
Interdental fricatives [θ ð] are mostly diffuse, with the teeth and open air acting like a wall. The fricative channel formed by the tongue along the alveolar ridge is longer than the bilabial channel, resulting in lower frequency spectral peaks than for bilabial fricatives.

Some fricatives do not fit neatly into either category. For example, velar fricatives [x ɣ] have both compact and diffuse properties. They have a long resonance tube (velum to the lips), resulting in lower frequency spectral peaks than even for retroflex fricatives, but higher than for uvular fricatives. In addition, the airstream flows directly to the back of the alveolar ridge, which acts like a wall and dampens the overall intensity. Because the airstream travels so far in comparison to frontier fricatives, it loses energy along the way, and the dampening effect on the non-peak frequencies is even greater.

Pharyngeal fricatives [ħ ʕ] also have both compact and diffuse properties. They have a very long resonance tube (pharynx to the lips), resulting in the lowest frequency spectral peaks of all fricatives. In addition, the airstream flows through at a right angle formed by the throat and mouth, with this right angle acting like a wall and dampening the overall intensity. As with velars, because the airstream travels so far in comparison to frontier fricatives, it loses energy along the way, and the dampening effect on the non-peak frequencies is even greater.

Glottal fricatives [h fi] aren't really like other fricatives (and in fact, some phoneticians do not even classify them as fricatives at all), because they have no inherent tongue position. Instead, the shape of the resonance tube depends on the adjacent sound, with the glottal fricatives just acting as an overlaid phonation. For example, [hɑ] is acoustically [qɑ], while [hi] is acoustically [ji].

Note that these are general trends, and that for a given fricative in a given language, the acoustic properties may be somewhat different due to variation in the way it is articulated. For example, the English post-alveolars are often rounded, which lengthens the resonance tube and lowers the spectral peak frequencies below what would ordinarily be expected.

bilabial	Φ		heavily diffused by outside air
labiodental	f		smaller channel than for bilabial; heavily diffused by teeth and air
interdental	θ		longer channel than for bilabial; heavily diffused by teeth and air
alveolar	s		small resonance tube
post-alveolar	\int		longer resonance tube, lengthened more by sub-lingual cavity
retroflex	\S		longer resonance tube, lengthened more by both sub-lingual cavity and tongue curling
palatal	\C		longer resonance tube, no sub-lingual cavity
velar	x		longer resonance tube; severely diffused by back of alveolar ridge
uvular	χ		longer resonance tube
pharyngeal	h		longest resonance tube; severely diffused by right angle at velum

2.7 Perturbation Theory (NOT YET INCLUDED)