

## Speech Processing Labs: PHON 2: Exploring speech acoustics

# Exploring speech acoustics

---

## Comments and Answers

This is the module 2 lab worksheet annotated with comments and answers (generally in red).

## Learning Outcomes

- Understand what spectrograms represent with respect to an input audio signal
- Understand what wide-band and narrow-band spectrograms are
- Understand what the fundamental frequency of a periodic signal is and how it can be estimated from the waveform and from a spectrogram
- Observe the relationship between fundamental frequency and its harmonics
- Observe the relationship between vocal tract shape and vowel formants (i.e., resonances of the vocal tract)
- Explore variation in the vowel spaces of different speakers

## Before the lab

- Go through the [Module 2 videos and readings](#)
  - See [Module 1 lab](#) for an introduction to using Praat
- You may wish to record the word list and "Mary" sentence in the last sections of this lab in a quiet spot before the lab.

*You don't need to submit anything for this lab*

In this lab we look at bit more in detail at what's in a spectrogram. We go through the Discrete Fourier Transform in more detail in module 3. So, far we've only really talked about the Fourier Transform as a sort of black box way of determining what frequency components are in a window of audio.

## Wide-band and Narrow-band Spectrograms

---

The main point of this bit is to see time vs frequency trade off as we change the DFT length of the window that the DFT is applied over.

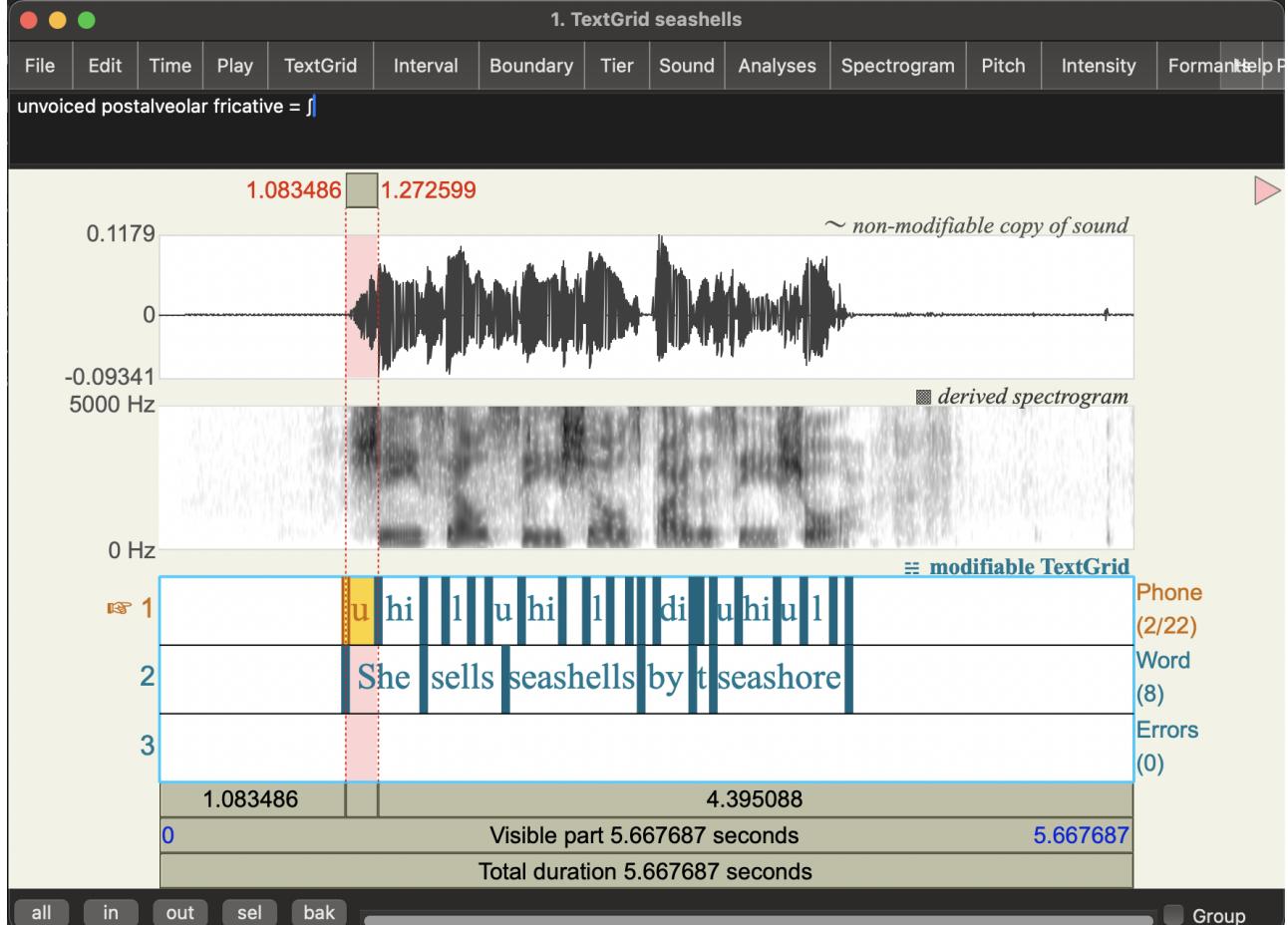
Spectrograms give us a plot of what frequencies are present in a recording through time: The vertical axis represents different frequencies, the horizontal axis represents time, the darkness of each coordinate (i.e., (time, frequency) point) in Praat represents the magnitude of that frequency in the input signal (waveform). The darker the spot, the higher the amplitude of that specific frequency (you can think of this as the volume or loudness of that specific frequency as a pure tone).

As discussed in class, applications like Praat generate spectrograms by applying the (Discrete) Fourier Transform to short windows in time. Changing the window size of that window changes what information you get out.

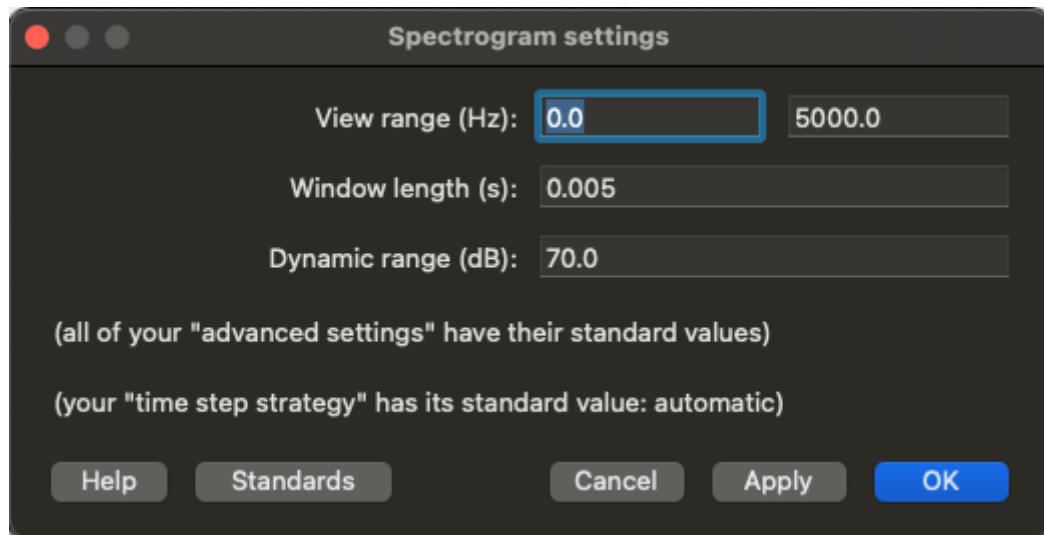
# Changing Spectrogram Settings

1. Open the [seashells.wav](#) recording from the module 1 lab.
  - o You may find it handy to use your consonant and vowel annotations from last week to get oriented.
2. View the [seashells](#) Sound object (and TextGrid annotations from last week if you have it)

You should see something like this:



3. In the sound viewer, Click on the [Spectrogram](#) menu and go to [Spectrogram settings...](#)
4. You should see a [Spectrogram settings](#) pop up window with various parameters you can change:
  - o **View range (Hz):** what frequencies are shown
    - default: 0 to 5000 Hz
  - o **Window length (s):** The size of the windows we apply the Fourier Transform to
    - default: 0.005 (seconds)
  - o **Dynamic range (dB):** This basically tells you what loudness the colour white represents in the Praat spectrogram (frequencies with amplitude lower than a certain threshold all appear white, i.e. too quiet to notice)
    - default: 70 (decibels)



We're going to focus on the \_window length.

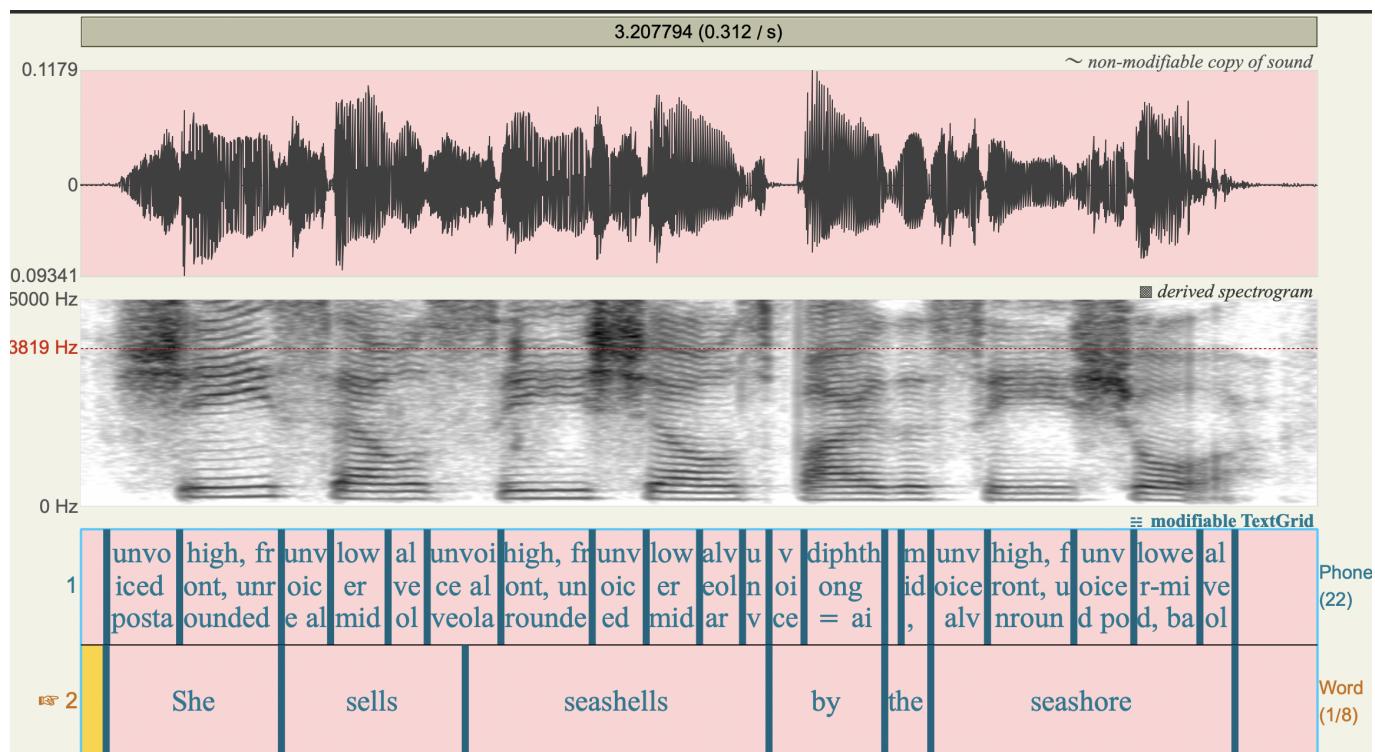
5. Change the **Window Length** value to 0.05 (i.e., bigger window)

6. Click on **Apply**

**Question:** What changes do you see in the spectrogram?

- Do you see more detail in terms of time or frequency under these settings compared to the default (Window length = 0.005)? That is, is it easier to distinguish which specific frequencies are in the speech signal with the longer window setting or the shorter one?

**Answer:** It should be easier to distinguish specific frequencies for the longer window. You (window length = 0.05). Specifically, you'll see more frequency bands at regular intervals (these are the harmonics of F0) which vary in their amplitude. "On top" of this you'll see the formant (vocal tract resonance structure).

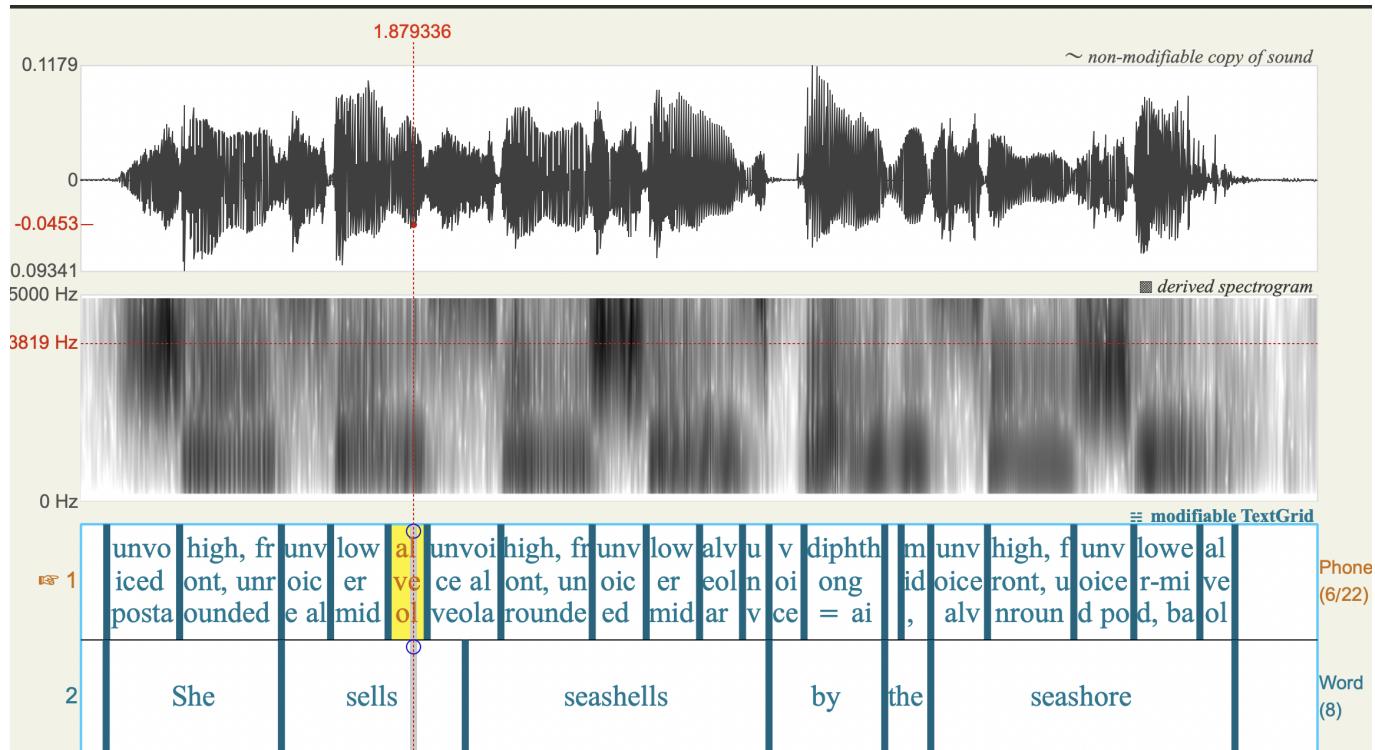


Let's look at some other (extreme) window lengths

**Question:** What happens when you change the window length to 0.001?

- Can you still distinguish different which frequencies are present in the speech signal?
- What do the *vertical* stripes that appear in this setting correspond to in human speech

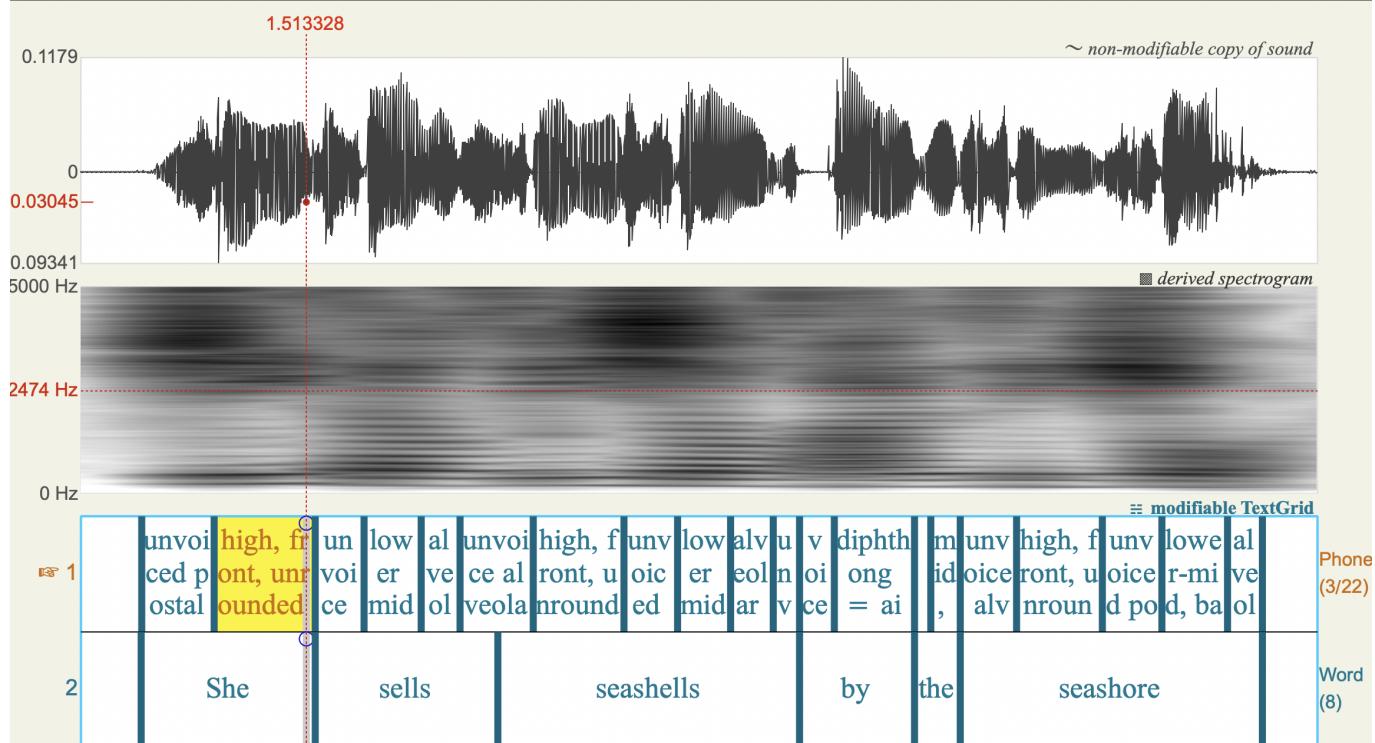
**Answer:** Here the window length is very small, so you get good time resolution: the vertical lines basically correspond to the individual pulses of coming through the vocal folds (if you zoom in, you should see these line up quite well with the cycles in the waveform). But we can no longer really see the frequencies in the signal (it's very blurry). You can still see some differences between the fricatives and the vowels, but you can't see what the specific formants or harmonics are.



**Question:** What happens when you change the window length to 0.5? Can you distinguish different manners of articulation?

**Answer:** Here we get the opposite sort of blurriness: You can see the differences in which frequencies are present, but this is blurred in the time dimension. You can sort of see there are some different sounds here, but it's not really possible to see the acoustic cues for different manners of articulation. For example, we

can no longer clearly see acoustic characteristics of fricatives and vowels, or if there are any oral stops.



**Question:** Looking at all of these different settings, what seems to be the trade-off between time and frequency resolution?

**Answer:** We saw that with bigger windows, we can see frequency details but time details get blurred out. With smaller windows, we can see more time detail, but less frequency detail. So the trade-off is that if we increase time resolution, we lose frequency resolution and vice-versa: if we increase frequency resolution we lose time resolution. The Praat defaults are set to be useful for humans doing phonetic annotation. However, this may not be optimal for automatic speech recognition. The window lengths (i.e., level of frequency resolution) you see in applications are generally tuned to best fit the task at hand.

What you should see is if the window is too big, we don't see any detail of how the sound is changing in time: the spectrogram is blurred on the horizontal (time) axis. If the window is too small, we don't see the frequency details: the spectrogram is blurred on the vertical axis. This is an unavoidable issue in generating spectrograms: if we increase the **time resolution** we reduce the **frequency resolution** (and vice versa)

When we talk about **narrow-band** and **wide-band** spectra, we are talking about the amount of blurriness in the *frequency*.

- Narrow-band spectra have higher resolution in frequency (possibly blurry in time)
- Wide-band spectra have higher resolution in time (possibly blurry in frequency)

**Comment:** I've never been very good at keeping these terms straight in my head. But my own key to remembering is that they're terms about frequency resolution, not window size!

## Different ways to measure the fundamental frequency

This next bit on F0 estimation is really about relating the periodic structure of voiced speech to harmonics in the spectrogram. We covered the source-filter model of speech a bit in class, but we'll come back to this in module 4. It's discussed in the Ladefoged & Johnson textbook from a phonetics perspective, but I think it's

actually easier to see what's going on with the computational model we'll see in module 4. We haven't covered why vocal fold vibration causes harmonics as it's quite complicated (out of scope for this course).

Perceived pitch of (voiced) human speech corresponds to the rate of vocal fold vibration. You can think of every opening and closing cycle of the vocal folds as shooting a "pulse" of air through the vocal tract. The rate of these pulses determine the **fundamental frequency (F0)** of the speech at that point in time.

**Terminology warning:** We call F0 the *acoustic correlate* of perceived pitch. That is, F0 is the thing we measure from the waveform and it correlates pretty well with what we hear as changes in pitch. The mapping isn't perfect though as speech perception is quite complicated in reality (the brain does a lot of complicated things with the information it gets from your ears!). You may see pitch and F0 used interchangeably in speech technology papers, but for the most part when we talk about pitch from an acoustic phonetics perspective, we're actually talking about F0.

We can measure F0 directly from the waveform (i.e., the time domain) and from the frequency spectra that make up a spectrogram (i.e., the frequency domain).

## Measuring fundamental frequency in the time domain

We can measure the fundamental frequency (F0) of a waveform by counting the number of cycles it makes in a given period of time.

**Task:** Calculate the F0 at the middle of the [i] vowel of the word "She" in the [seashells](#) recording.

1. Select an [i] vowel in the seashells recording (i.e., the vowel in "She")
2. Zoom in so that you can see at least 5 of the beginning and ends of cycles of the waveform clearly for the [i] vowel
3. Calculate the fundamental frequency over 5 cycles:
  - o Use the highlighter to select 5 cycles
  - o The number in the bar above the pink highlighted section will be the duration of your selection, let's call this  $d$
  - o The number to the left of selection at the top is the start time of your selection. Note this down for later!
  - o You can now calculate the *fundamental period (T0)* of this wave as  $d/5$
  - o The fundamental frequency is calculated as:  $F0 = 1/T0$

**Comments:** You can see here, I've zoomed in and selected 5 cycles starting at time 1.392063. The number at the top of the selected bit give the selection duration:  $d=0.022033$  seconds. 5 cycles in  $d$  seconds gives us  $T0 = d/5 = 0.0044066$  as the fundamental period. So we have:  $F0 = 1/T0 = 1/0.0044066 = 226.9323$

**Hz.** Of course, you could just simplify and calculate F0 directly:  $F0 = 1/(d/5) = 5/d = \text{number of cycles/length of time}$



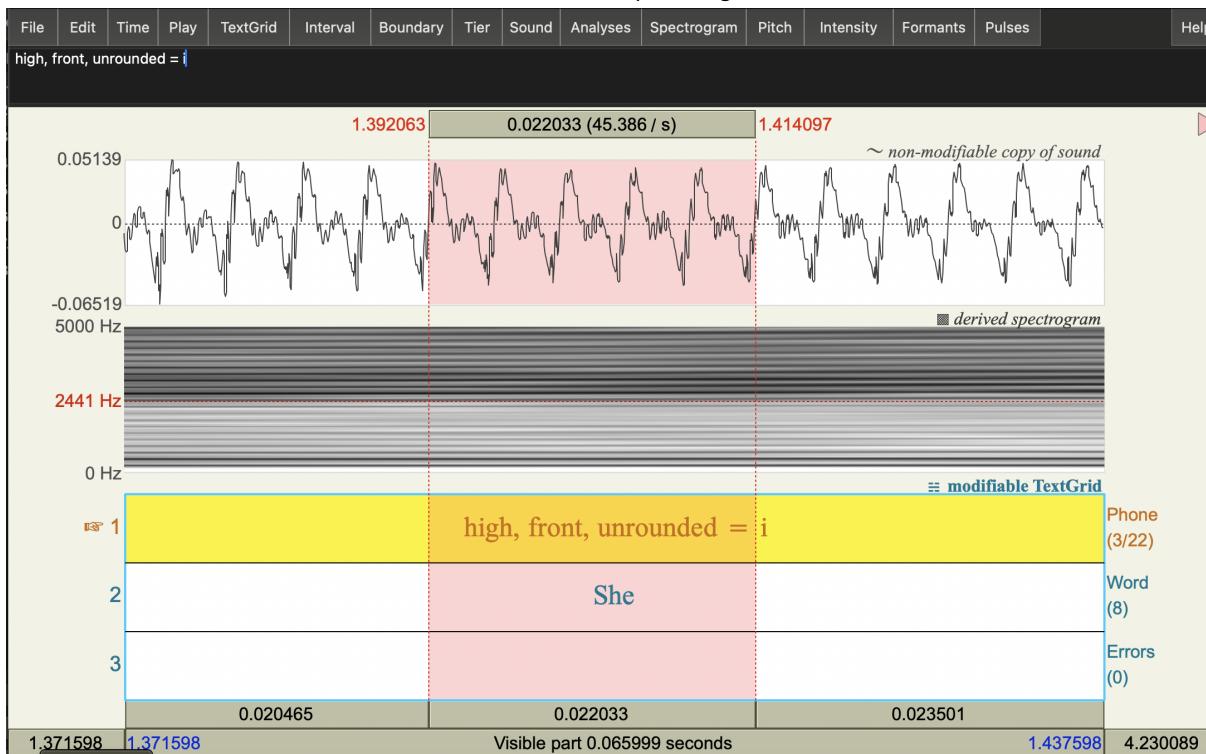
## Measuring fundamental frequency with a narrow-band spectrogram

If the frequency resolution is good enough, you can also measure F0 directly from the spectrogram: since F0 should be a prominent frequency component of the complex sound wave, so you should be able to see it as the first dark frequency band in a *narrow-band* spectrogram.

**Task:** Estimate F0 from a spectrogram and a spectral slice.

1. Change the Spectrogram window length setting to 0.05 seconds (a bigger window than the standard).

- You should see a bunch of horizontal lines in the spectrogram.

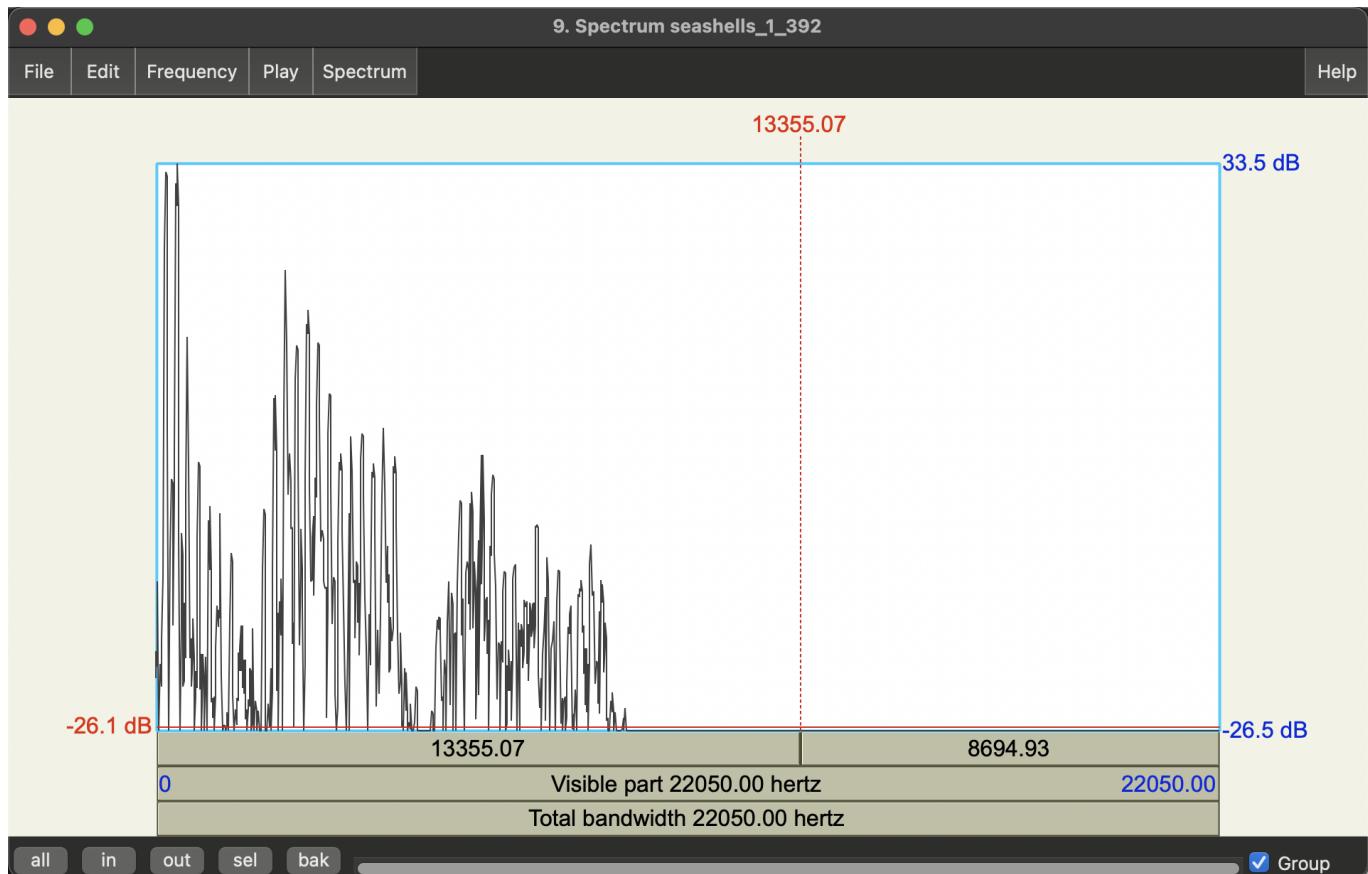


- Place your cursor on the lowest frequency dark horizontal line on the spectrogram at the beginning point of your 5 cycle selection (as above).
- Note the frequency of this dark band at that time: it's the number on the *left* side of the spectrogram.

This is also known as the **first harmonic** of the voice source. But it's bit hard to get accurate frequency measures on the spectrogram itself. So, let's take a spectral slice instead.

**Comment:** This should give you something in the same ball park, but it's actually quite hard to get a very precise frequency measure clicking on the spectrogram!

- With your cursor in the same point in time, get a Spectral Slice:
  - Spectrogram > View spectral slice**
  - This should pop up a new figure. If not have a look in your objects list for a new **Spectrum** object and view that.



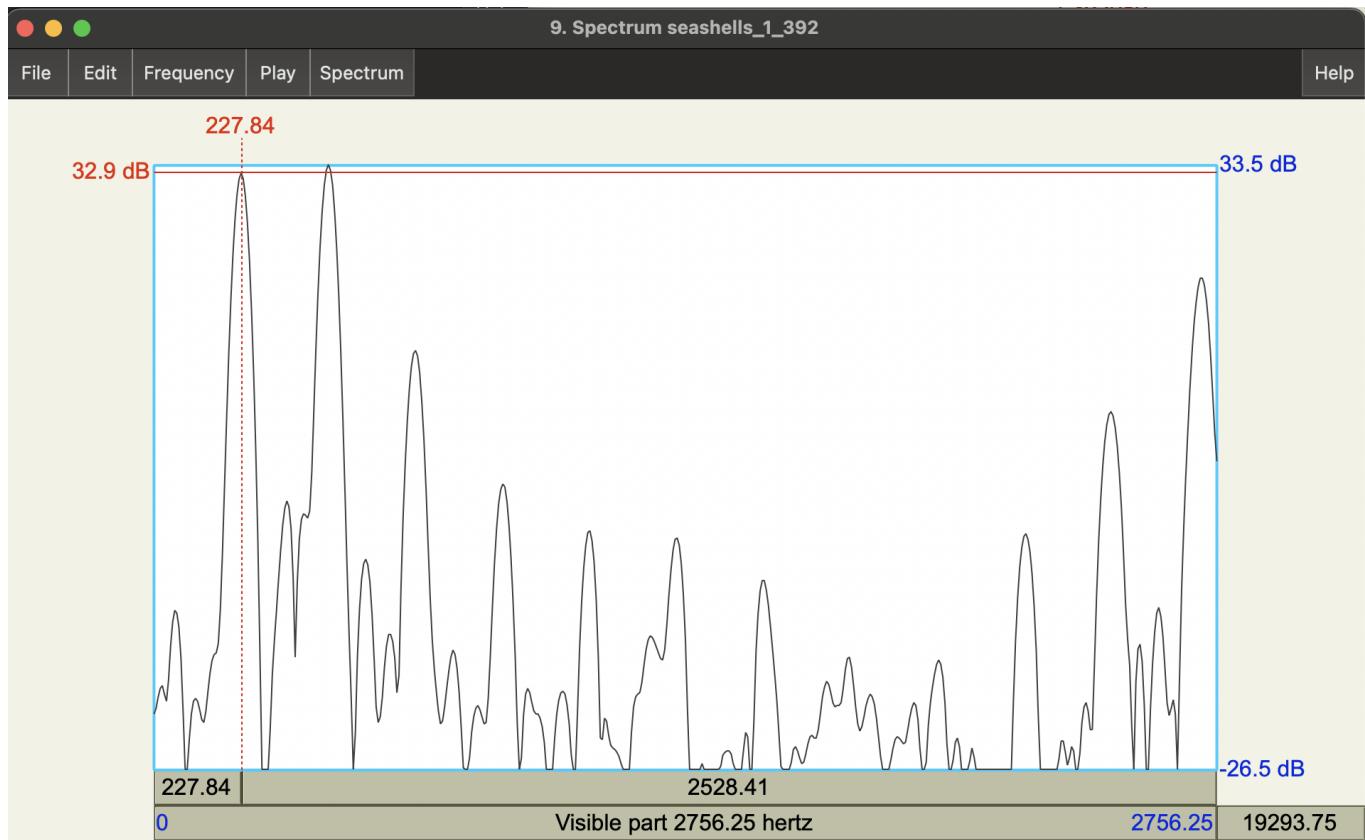
This is the frequency spectrum at the time corresponding to your cursor selection (more correctly a window of time around that point, where the length of that window is what you chose in the Spectrogram settings). You should see it's quite spiky!

5. Get an F0 estimate by selecting the top-most point of the first big spike.

- o Note that spectral slices (aka frequency spectrum plots) have frequency on horizontal axis, and magnitude (amplitude) on the vertical axis.

**Question:** Does this method give the same F0 value as you get from measuring the period of the waveform?

**Answer:** You'll probably need to zoom in a bit to get the top of the spike with your mouse, but yes it's pretty close! The cursor shows 227 Hz and is as close as I can get the peak, though you'll see it's still a bit off - close enough though! Note, the small peak to the left of this is background noise.



In the spectral slice itself, you should also see more spikes (i.e. boosted frequencies). These are resonances of the vocal source and we call these harmonics.

**Question:** What is the relationship (roughly) between the F0 of the vocal source and its harmonics?

- Hint: what's the difference between consecutive spikes?

**Answer:** You should see that the main spikes in the spectral slice are fairly evenly spaced. The distance between each is roughly our F0 estimate, i.e., the harmonics of the source are multiples of F0. This harmonic structure is an important part of what makes voiced speech sound like it does!

We associate F0 with the speech source, specifically vocal fold vibrations. We can see this directly from the waveform, but we can also see the effect of these vibrations in the frequency spectrum.

## Measuring fundamental frequency with a pitch tracker

As you may have noticed last week, we can also check our estimate against Praat's pitch (i.e., F0) tracker:

1. Go to the **Pitch** menu and turn on the pitch track (**Show pitch**).
  - You should see a blue line appear the spectrogram
2. Place the cursor on at the beginning point of the 5 cycle selection from above
3. Note the frequency value of the blue line at that time: it's the number on the *right* side of the spectrogram.

**Comment:** This also give you 226 Hz! Usually, if we're just interested in F0, we would just use the pitch tracker to extract estimates. But it's worth noting pitch trackers can also make use of time-domain and frequency domain information in making estimates (there's a lot of different pitch tracking algorithms out there! See more in the speech synthesis course...)

We're not going to go over F0 tracking methods in this course, but you can find more in the [Speech Synthesis \(sem 2\) course materials on speech.zone](#)

## Vowel Formants

---

Narrow-band spectrograms allow us to see the frequency harmonics due to the voice source. But when it comes to identifying vowels in English, we want to focus on the resonances of the vocal tract, i.e., the *filter*.

In this section, we'll look at how changes in the vocal tract shape are visible in the spectrogram. We'll do this by looking at the estimated **formants** (i.e., vocal tract resonances) change for different vowels in a recording of a US English speaker. You can then record yourself and see how your vowel space compares!

**It's important here to make the distinction between harmonics of the source and resonances of the vocal tract.** Both are basically frequencies that are amplified to different degrees by the combination of source and filter. In this section, we're focusing more on the vocal tract resonances = formants. In general, since F0/source features aren't contrastive in English for word recognition, automatic speech recognition of English ignores source features. However, it's worth noting that they are important for distinguishing words in tone languages though (and prosody of course!).

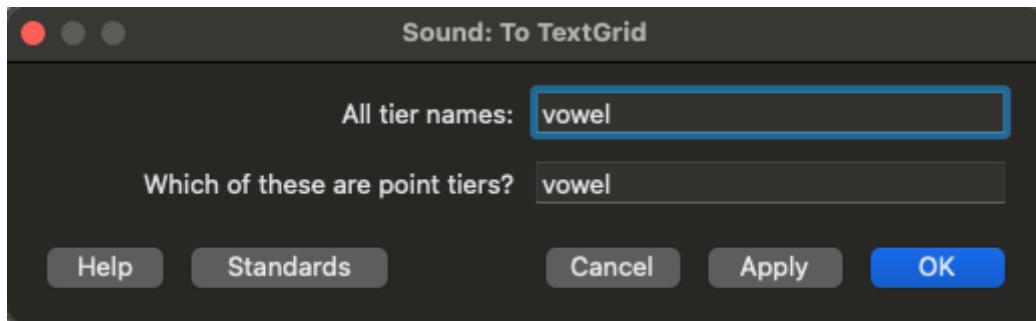
### Vowel Space (US English speaker)

Since we want to focus on the resonances of the filter, we're interested in the overall shape of the frequency spectrum (i.e., the **spectral envelope**) rather than the fine frequency detail of the source harmonics. This means it can be more useful to use a wider band spectrogram (losing some harmonic detail).

It turns out the Praat defaults are pretty much tuned to this sort of analysis, so let's change the spectrogram window settings back to the default.

**Task:** Extract the first and second formant values (F1 and F2) for a series of English words which differ only by vowel.

1. Download and open the following sound file in Praat: [speechproc\\_phonlab2\\_rvowels.wav](#)
2. Create a TextGrid for that sound object with a single *Point Tier* called vowel:



**We're going to extract F1 and F2 for only one point in the vowel. This is for simplicity here, but you can think about/discuss whether this really is enough to characterise a vowel!**

3. Open the corresponding Sound and TextGrid objects together ([View & Edit](#))
4. Change the spectrogram window length back to default: 0.005 seconds

## 5. Turn on the automatic formant tracker: **Formants > Show formants**

Now let's get some vowel formant measurements.

The words in the recording are:

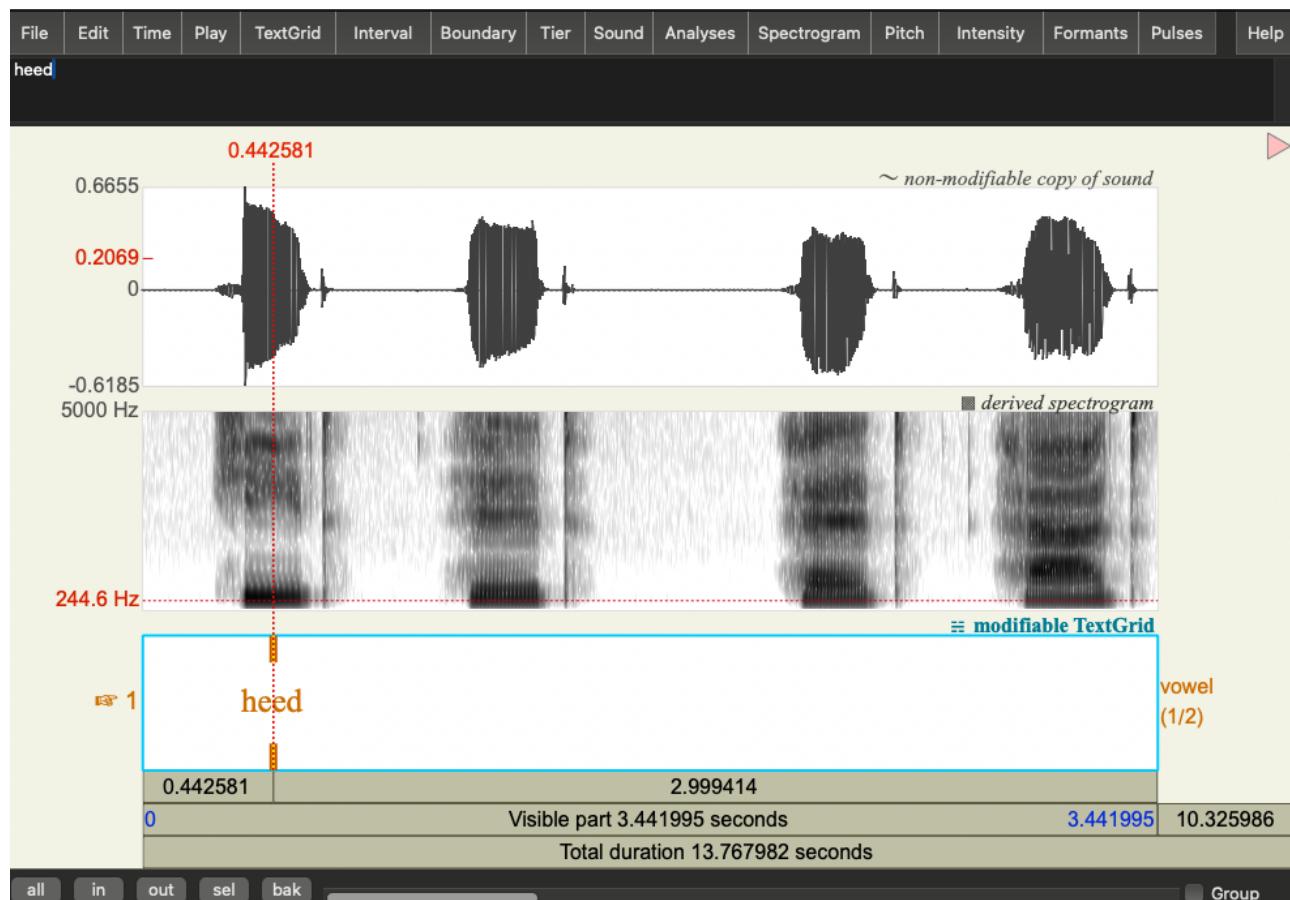
- Heed
- Hid
- Head
- Had
- Hod
- Hawed
- Hud
- Hood
- Who'd
- Heard
- Hide
- How'd
- Hayed
- Hoed

## 6. For each of the words in the recording:

- Click on middle-point of the vowel (in terms of time)
- Add a point boundary at that point in time (click on the circle at the top of the annotation tier)
- label that boundary with the corresponding word

You'll want to zoom in enough that you can see the spectrogram of the word clearly.

Here's the an example for the first word:



**Question:** Does taking the middle point of the vowel really make sense if the vowel is a diphthong (e.g. in the word "hide")? What might be a better way of representing this?

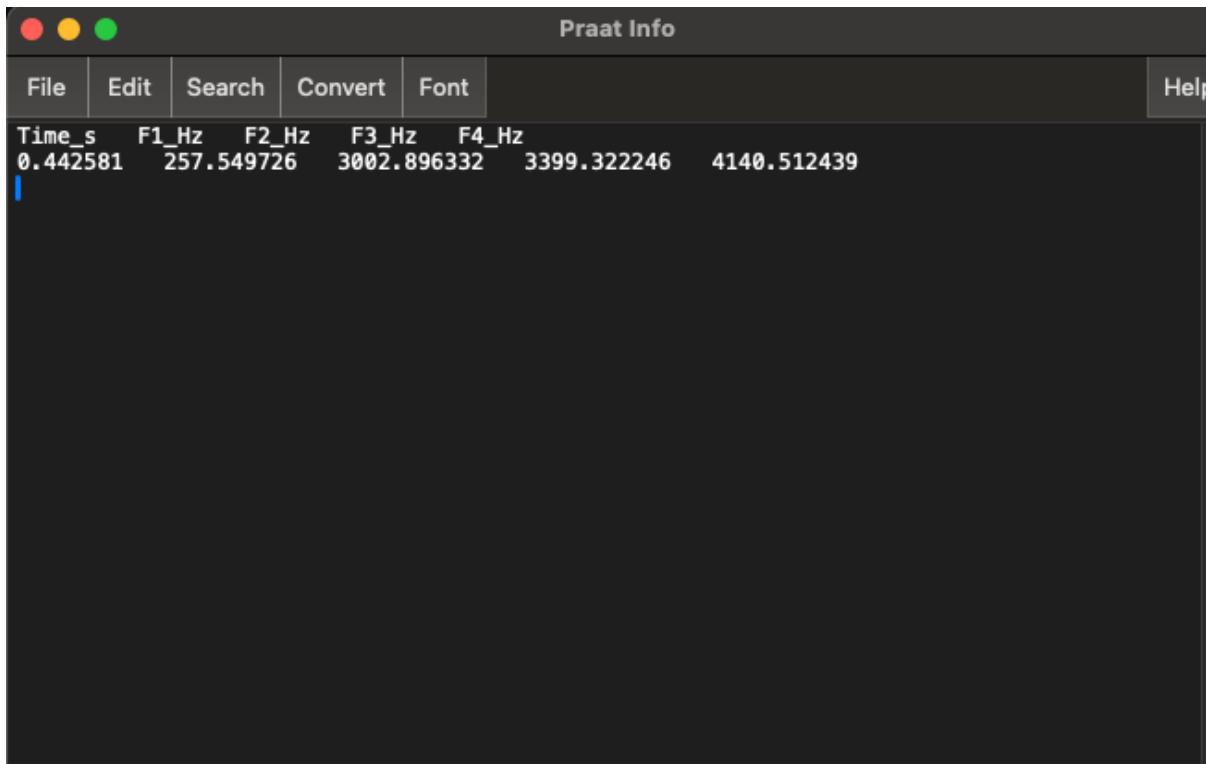
No, diphthongs are characterised by the smooth movement from one vowel to another. The diphthong in "hide" starts in the [a] position and ends at [i]. So if we take a point in the middle, it's really just a transitory point between those two articulatory targets. It would be better to take a measurement near the beginning of the diphthong and towards the end of the diphthong. Then we could draw an arrow between the start and end to show the transition.

For now, let's just take the middle points but if you have extra time, you can think about you might better measure and plot diphthongs.

#### 7. For each of the vowel annotations, get estimates of the 1st and 2nd formants (F1 and F2)

- Click on **Formants > Formant Listing**

This will pop up a window with estimated formant values. The first row is the variable name (Time\_s, F1\_Hz, F2\_Hz, F3\_Hz, F4\_Hz). The second row shows the corresponding values. Here's an example:



The formants values should correspond the spectral envelope peaks when you look at a Spectral Slice at the same time point. Check this for yourself!

- Copy and paste the values to a spreadsheet (e.g. excel, google sheets)
  - We'll want to make a table with the following columns:
    - word
    - F1
    - F2

## Plot the vowel space

**Task:** now that you have formant values for each of the words in the recording, you can make a plot of the vowel space!

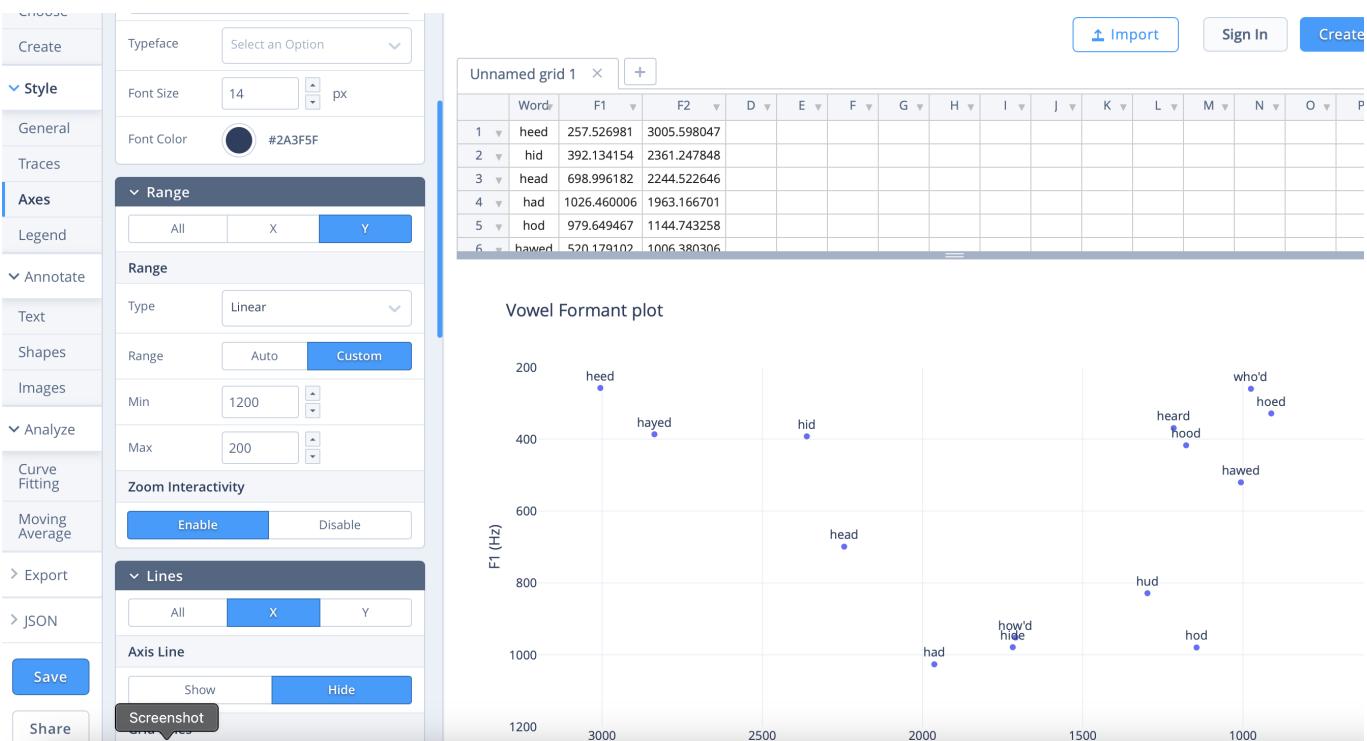
You can use whatever method you like for this: e.g., Excel, R, python, online plotting tools (e.g., [plotly chart studio](#)), or pen and paper.

If you want to give the plot the same orientation as the IPA vowel chart, you'll need to:

- Plot F2 on the horizontal axis
- Plot F1 on the vertical axis
- Orient the axes so that (0,0) is in the top right hand corner of your graph

From past experience, students seem to know how to plot using excel, but I find using that quite complicated. Below is one I made using [plotly chart studio](#), which is more straightforward (to me!). The main things to adjust are the axes: switching max and min values reverses the direction so that (0,0) is in the top right corner. You can change these in Axes > Range. In the traces menu, you can check "text" in the Display options to show the words (if you have the format I had above). You can rename the columns by clicking on the existing column names in the spreadsheet. There are lots of other options to fiddle with!

The plot does broadly match the IPA vowel chart but you should see that some vowels are quite close together, and some further apart. This is partially an artefact of not doing any log-scaling here, but it's also due to the pronunciation of this specific speaker. Note, the diphthongs are in a weird middle place, as we'd predicted.



## Plot your own vowel space

**Task:** Plot F1 and F2 measures for your own speech.

1. Record yourself saying the word list above
  - o See the module 1 lab for some instructions on how to record in Praat.
2. Repeat the procedure for annotating the vowels and getting F1 and F2 values.
3. Plot your own vowel formant values and compare them to our US English speaker.

### Questions:

- Are there any major differences in your vowels relative to the US English speaker?
- Does your vowel formant chart match the positioning on the IPA chart?

Individual vowel plots should broadly match the IPA chart, but we'd expect to see a lot of variation in the specific measurements. A specific difference you might notice between yourself and the US English speaker above is in whether or not you pronounce the "r" in "heard". English speakers from England and Australia generally don't pronounce the "r", while Scottish and US English speakers probably will. This is actually quite variable, so the goal here is just to look at the chart and see if you can account for any differences. There are some pronunciation variants given in the Ladefoged and Johnson chapter on transcription that might be helpful to look at.

## Vowel variation: *marry, Mary, merry*

One of the major issues for speech technologies is that pronunciation varies a lot! Depending on your language background, you may find that certain words are pronounced the same or different. Your ability to hear differences in pronunciation in others may also be affected by your own pronunciation.

**Task:** Compare your pronunciation of the words "marry", "Mary", and "merry" and see whether they are the same or different.

1. Record yourself saying the phrase "Marry me Mary, make me merry"
2. Look at the spectrograms of the words "Marry", "Mary" and "merry".
  - **Question:** Do they look the same or different?
3. Get the formant values (F1 and F2) for the first vowel in each of these words.
  - **Question:** If you plot F1 and F2 for these vowels are the values close together or spread apart?

Again, this is quite exploratory so there's no right or wrong answer here! You may only realise that you pronounce these words the same (or different) by listening and looking here. If you do pronounce them the same, you should see that they look similar on the spectrogram. You should then be able to see the similarities or differences when you plot F1 and F2. Note: Relying on your own perception can be quite tricky here! For example, if you have a vowel merger here you may not hear differences that other people do hear! The main thing to take away is that there's a lot of variation in pronunciation. Within the class we usually have people who pronounce them all the same, all different, and only some the same! Here's a cool site that maps other pronunciation variation in the UK:

<https://www.ourdialects.uk/phonological/>

**Extra:** Compare your Marry/Mary/merry to some other speakers. Here's some examples to start you off:

- [english\\_mary-sk.wav](#)
- [english\\_mary-kr.wav](#)

## End notes

---

The goal of this lab was to start exploring how differences in speech acoustics can reflect how human speech is actually produced. We can get a lot of information from spectrograms about both the speech source (e.g., vocal fold vibrations) and the vocal tract resonances (e.g., the shapes you make with your articulators). We saw two types of resonances in speech that interact: F0 and its harmonics which are driven by vocal source, and formants which are driven by the vocal tract shape.

We can see the effects of these on the spectrogram. However, there is a tradeoff between time and frequency resolution. When we apply Fourier transform to increasingly shorter analysis windows, we lose frequency resolution. When we use longer windows, we get better frequency resolution but lose our ability to see the details in time.

We'll go into more detail about what the Fourier Transform actually is in Module 3. We'll then come back to look at the source filter model of speech through a more computational viewpoint in module 4.

The marry-merry-Mary merger is an example of tense-lax neutralisation. I first learned about this from [Aaron Dinkin](#). You can learn about it more in his paper: ([Dinkin 2004](#)). No doubt though that a lot can change in people's speech patterns in 20 years! This isn't a course about language variation and change,

but if you are going to do further in speech technology, you will need to deal with individual and group variation in speech.

## Acknowledgements

This lab builds on a previous lab designed by Rebekka Puderbaugh.