Speech Processing Labs: PHON 1: Analysing speech articulations

Analysing speech articulation

Comments and Answers

This is the module 1 lab worksheet annotated with comments and answers (generally in red)

Learning Outcomes

- Learn to use phonetics software, Praat, to inspect and annotate audio files
- Practice the mapping between speech articulations and the International Phonetic Alphabet symbols and layout (IPA)
- Practice using phonetics terms related to voicing, place and manner of articulation.
- Observe complexity of in physical speech production by examining common articulation errors

Before the lab

- Go through the Module 1 videos and readings
    - These focus on speech articulation and the IPA
- Make sure you have a handy copy of the IPA chart. You may find it useful to use version with linked audio.

You don't need to submit anything for this lab

Comment: This lab is really about exploring articulation and having a go at using Praat. You don't need to learn everything about Praat, though there are some resources linked below if you want to find out more. Similarly, our goal isn't to do close phonetic transcription here, but rather to have some practice linking articulation to the phonetic terms and the IPA.

Inspecting speech with Praat

Get started with Praat

1. Download Praat from https://www.fon.hum.uva.nl/praat/
2. Follow the install instructions linked from the Praat homepage for your operating system (i.e., Mac, Windows, Linux)
3. Open Praat by double clicking on the icon
    - If it's not obvious where the App was installed, you should be able to find it from the Start menu in Windows, or Spotlight on a Mac.

After opening Praat you should be able to see two windows:

- *Praat Objects*: This is the main one we'll work with for loading and manipulating recordings
- *Praat Picture*: This one is used for plotting.
    - We won't go into this in this lab, but you can find out more about this (and much more!) in these Praat tutorials by Will Styler: tutorials.

Open a sound file

Download the following sound file to your computer: seashells.wav

Open the sound in Praat:

1. Click `Open` on the *Objects* windows to show a drop down menu
2. Select `Read from file...`
3. Select the file you want to open from the pop-up window and press `open`

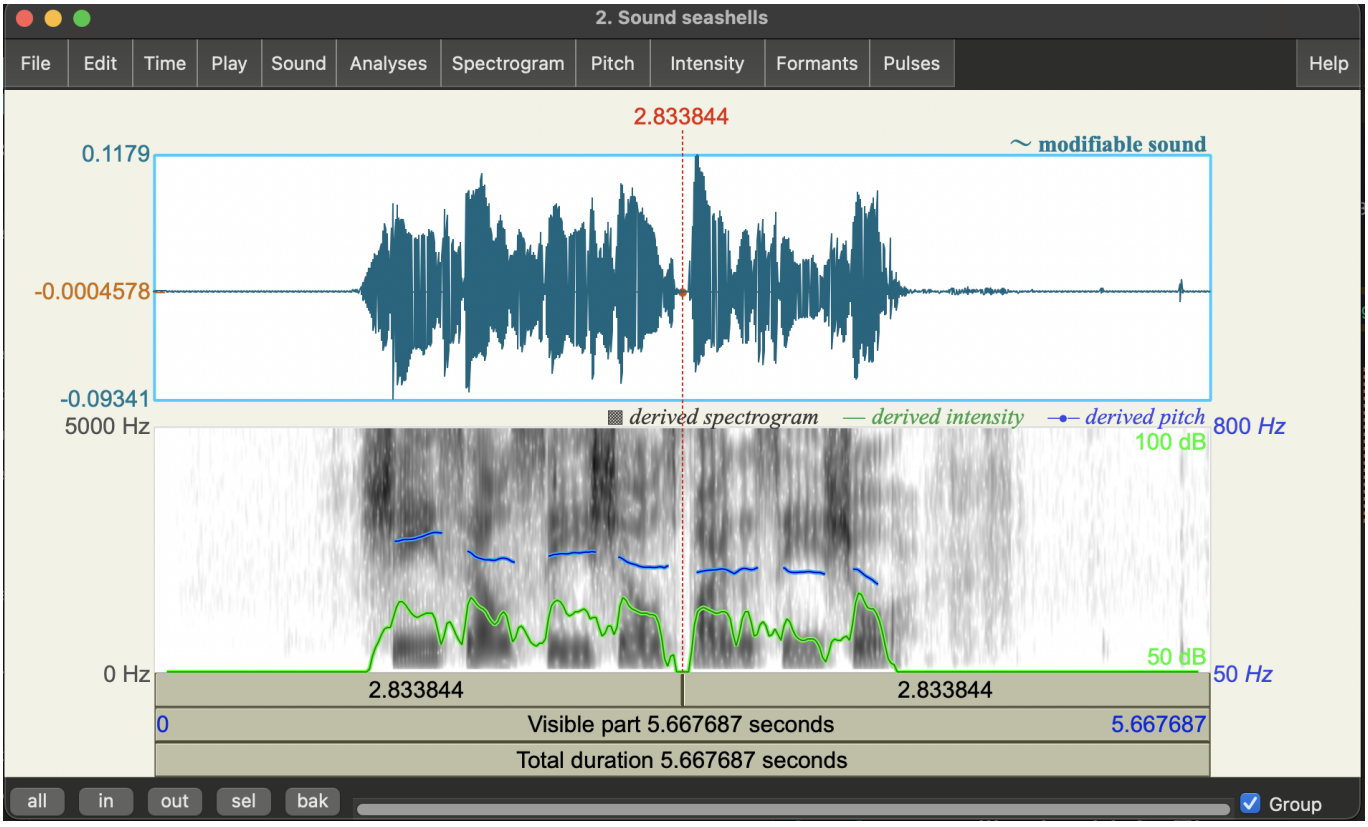
You should now see a line in the Objects list called: *Sound seashells*.

Examine the Sound

Praat will allow you to do many things with this new Sound object , but for now let's just open it and have a listen.

1. Select the *Sound* object in the object list.
2. Click on `View & Edit`

You should see a new window that looks like this:



The horizontal axis shows different points in time. The panels show different representations of the recording.

- The top panel shows the waveform, i.e. variation in air pressure in time.
- The middle panel shows the spectrogram: a representation of the sound frequencies that are present in the recording
    - We properly discuss these in Module 2-3, but for now just consider them as a way to see changes in the speech.

The bottom three bars give some play back controls, clicking on the bar labelled:

- `Total duration`: will play the entire recording
- `Visible part`: will just play the bit you can see (e.g., if you've zoomed in)
- The top bar will pay up until the point of the cursor.

On the bottom left corner you'll see some buttons (`all`, `in`, `out`, `sel`, `bak`). You can play with them to zoom in and out. You can move around time but scrolling left or right with your mouse.

Clicking on different spots will show you different information about the audio at that point in time. For example, in the image above we see red vertical and a horizontal dashed lines.

- The vertical line represents a point in time: 2.833844 seconds into the recording
- The red horizontal line on the waveform panel gives the waveform amplitude at that point in time: -0.0004578

If you click on the spectrogram panel, you'll also get a red horizontal line that gives you the coordinates of the point you clicked on in the spectrogram where the x-axis (horizontal) is time, and the y-axis (vertical) is frequency (in Hertz). We won't get into it this week, but it can be helpful for measuring different properties of the spectrogram by hand.

Pitch (the blue line)

You may see a blue bar superimposed on the spectrogram (as in the screenshot above). This represents the estimated pitch at a point in time (more accurately: Fundamental Frequency - we'll talk about this more in Module 2!). This is estimated using a different algorithm from the spectrogram so the fact that you see it here is really just a design choice from the makers of this software.

You can turn the pitch track on and off by clicking on the `Pitch` menu at the top of the window and checking/unchecking `Show Pitch`. The default method used here is "filtered autocorrelation" which you can see from the check mark in the menu.

Automation Warning: If the pitch tracking is good (as it is in the example above) you should be able to see a relatively smooth contour that matches your perception of when pitch goes up and down through the speech. Unfortunately, pitch tracking can be quite prone to error. Almost all pitch trackers are sensitive to the range settings (i.e. expected minimum and maximum pitch values in Hertz). If the expected range is doesn't really match the speaker's actually range you can get errors like octave doubling and halving. You will also get errors if the phonation is "non-modal", e.g. creaky or breathy. Sometimes data driven studies don't bother to check this and then end up with spurious results.

To change the range settings click on `Pitch settings` from the `Pitch` menu. The default range for Praat (50-800Hz) is ok, but you can often do better if you tweak this (e.g., see this paper).

Intensity (the green line)

Another common overlay is the Intensity estimate (green or yellow). You can turn this on by going to the `Intensity` Menu and clicking on `Show Intensity`. This will essentially give you a measure of the loudness in time (based on the amplitude of the wave). You should see that the peak structure in this contour broadly represent syllables in the speech.

Other functions

There are many other menus at the top of the window that will show other overlays (e.g. `Formants`, `Pulses`). Feel free to click around and see what they do. You can even use the functions in the `Edit` to cut and paste speech segments!

For the moment, we will just press on and learn what we need to as we go, so we can get started analysing some speech. If you want to learn more about Praat, there are several tutorials linked from the Praat website (which also hosts a lot of documentation): tutorials page. You may also find this video based guide by Richard Ogden (University of York) helpful: video guide.

Analysing Tongue twisters

Now that we've got the basics of Praat, let's go back to thinking about speech articulation. Specifically, we're going to use Praat to visualise and analyse what's going on in some tongue twisters! These are phrases that are difficult to say properly. Thinking about why they are difficult to articulate will hopefully help better understand difference in place and manner of articulation.

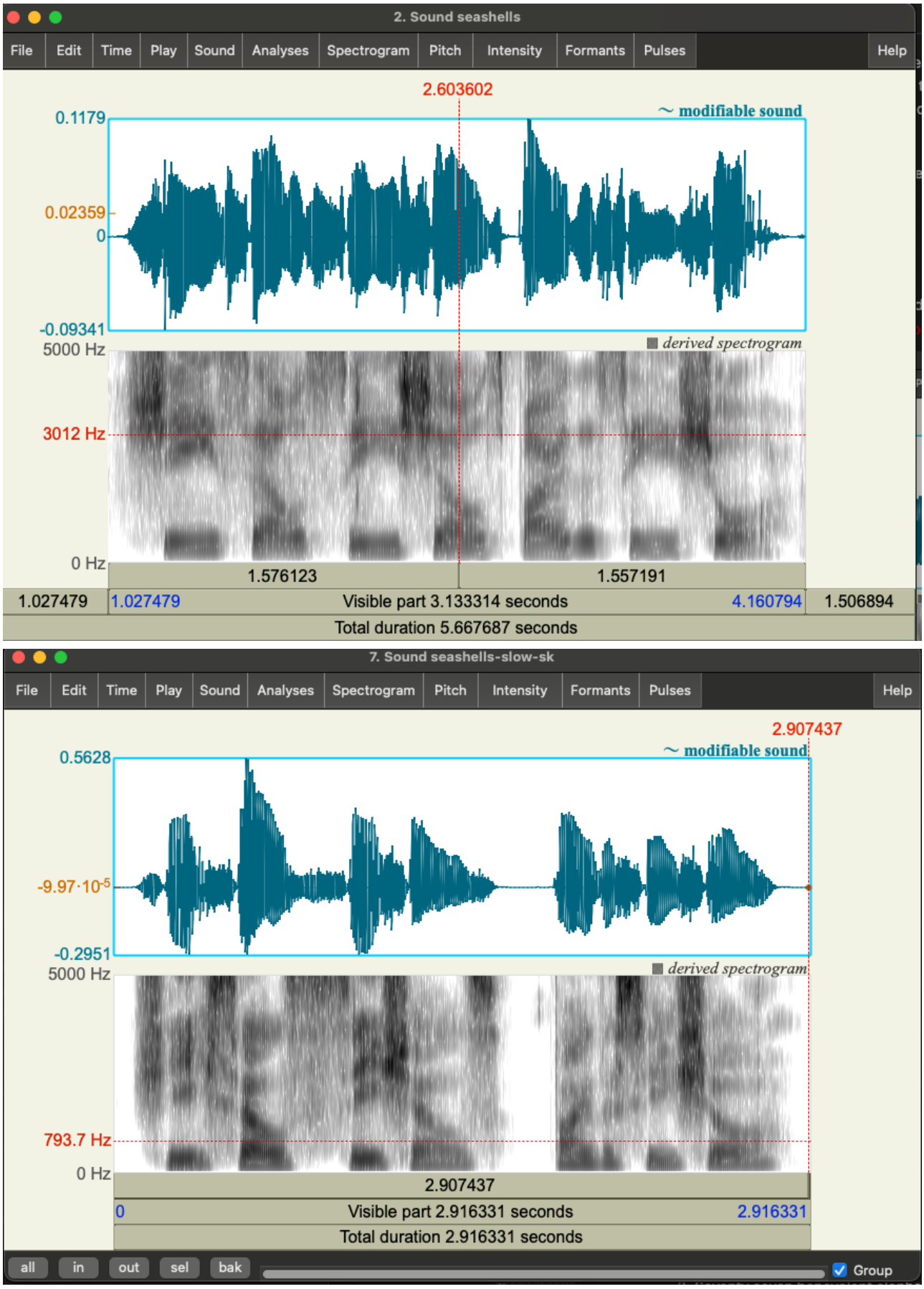
Let's start with some classic English ones, recorded with fast and slow speaking rates:

1. She sells sea shells by the sea shore

    - english_seashells-fast-sk.wav (fast)
    - english_seashells-slow-sk.wav (slow)

    This one is, of course, hte same tongue twister as the one we looked at above by spoken by a different speaker - can you tell just by looking at the waveform or spectrogram?

    Comment: You should be able to see some differences in the following two visualisations: the first one is Catherine's recording, the second one is the link above (slow), recorded by Simon King. You should see similarities in the spectrogram but they aren't completely the same. Part of the difference hear is the difference in their voices, but some of this is also just recording conditions (i.e., noise). The waveforms also look different, but you can't really tell that one recording is by another speaker just by looking at it. Various other factors could be changing the waveform and spectrogram.



2. Peter Piper picked a peck of pickled peppers. Where's the peck of pickled peppers Peter Piper picked?

    - english_peter_fast_pb.wav (fast)
    - english_peter_slow_pb.wav (slow)

3. Seventy seven benevolent elephants

    - english_seventy_fast_kr.wav (fast)
    - english_seventy_slow_kr.wav (slow)

Please note, for this lab it really doesn't matter if you can say these correctly! In fact, errors will probably be more useful!

Articulating tongue twisters

Task: Before we start analysing these in Praat, try saying each of these phrases out aloud. You may wish to take turns with the people next to you in the lab and then discuss the following questions (but it's totally fine to do this on your own).

Questions

- What parts of these phrases are difficult to say? What words tend be be said incorrectly? Are there specific phones that are difficult?

  Comment: This part of the lab is really to get you thinking about your articulators, but here's some observations (which will be repeated in the later tasks)

    1. Seashells: the difficult here is usually mixing up "she" and "sea", i.e. the place of articulation of the syllable initial fricative. I tend to say "by the she shore". The difference in place between "s" [s] and "sh" [ʃ] is quite small, so it's easy to hit the wrong target.

    2. Peter piper: In this case you have alternation between "p" (1st syllable) and other oral stops "t", "p", "k" (second syllables). The vowel variation can also trip people up - the first 3 words have high front vowels in the first syllable, you then sort of expect the last three to be lower front [e], but you also have the "pickled" in there. The addition of the "l" in "pickled" also trips me up.

    3. benevolent elephants: I always want to say "benelovent elephants". This seems to be a planning/rhyme thing: You have "**s**e**v**enty **s**e**v**en be-" but then "-nevolent elephants", which breaks the pattern of the first two onsets pattern matching in place ([n]-[l], vs [l]-[f]). You probably get priming for the "elephants" pattern because it's a more common word.

- Do you need to speak slower than you usually would to say these correctly?

    1. I sure do! Generally people need more time to make sure their articulations are correct with tongue twisters because the issues come with trying to say the words in connected speech.

- What happens when you try to say them faster?

    1. Errors occur, usually from hitting the wrong place of articulation, but you can also get manner of articulation errors. As you get faster, coordination of your articulators becomes more difficult and, without practice, your brain may direct your tongue to a "likely" place rather than the correct on.

We'll do some analysis on these one by one.

Example 1: Seashells

Comment: Some more practice with Praat. This time adding TextGrids for annotations.

Download and open one of the recordings of "She sells sea shells by the sea shore". In the following, I'll just use the first example (`seashells.wav`, spoken by Catherine) but you can use one of the others (spoken by Simon) if you prefer.

A big reason Praat is so popular with phoneticians is that it's convenient for annotation. Let's add some textgrids to do annotations now.

1. Click on the `Sound seashells` object in the *Objects* window
2. Click on the `Annotation` button to the right
3. Select `To TextGrid...`

You should see a little popup window named *Sound: To TextGrid* which you can use to set the annotation parameters. Edit the parameters there as follows:

4. `All Tier Names`: delete "Mary John bell" and replace it with "Phone Word Errors"
5. `Which of these are point tiers`: write Errors
6. Click `Ok`

You should now see a new `TextGrid seashells` object in the *Objects* window.

7. Select both the `Sound seashells` and `TextGrid seashells` objects so both are highlighted and then click on `View & edit`.

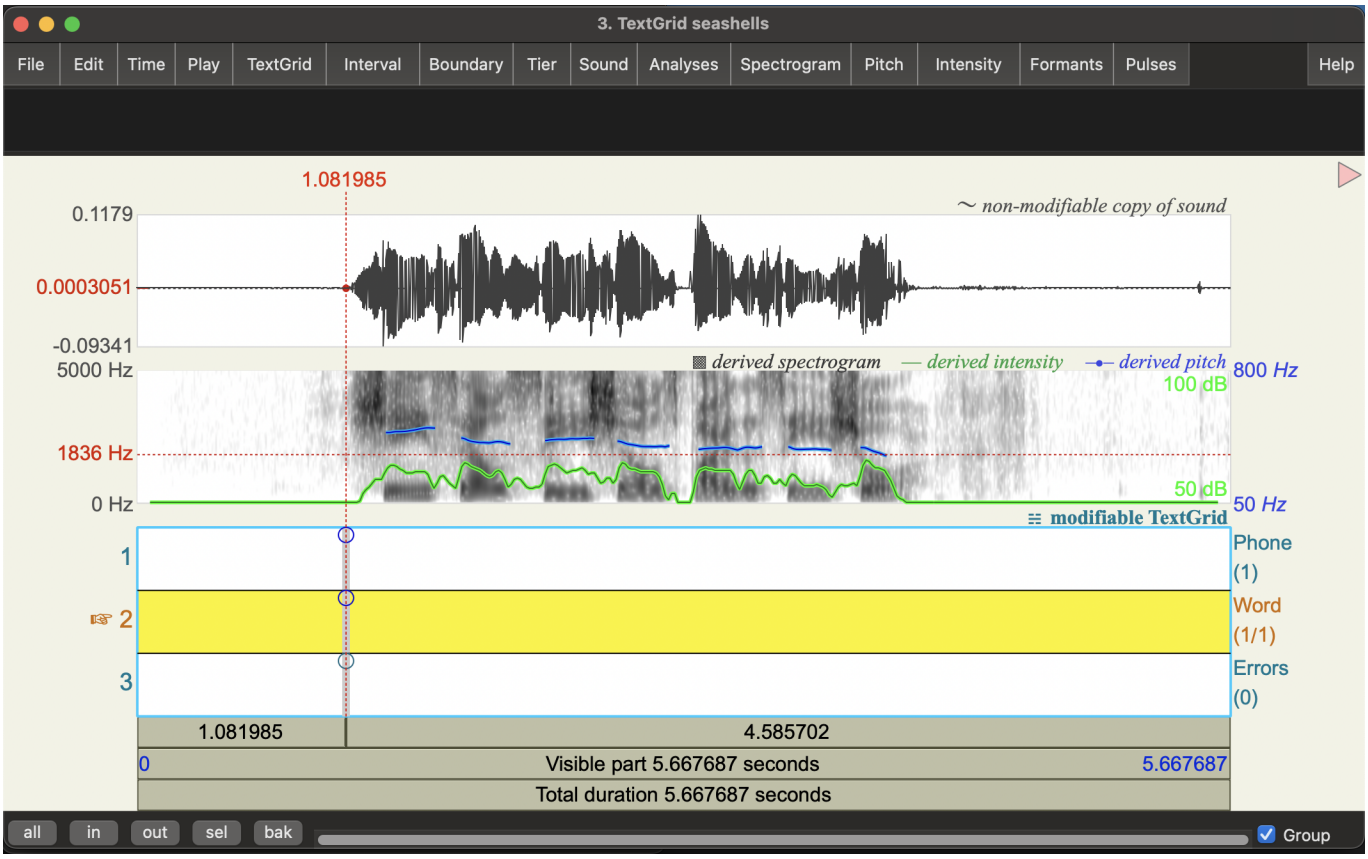
You should now see the sound viewer with the waveform and spectrogram up top, but now also 3 blank annotation tiers: `Phone`, `Word`, and `Errors`. The first two are *interval tiers*, while the last is a *point tier*. As the name suggests, we use interval tiers to annotate spans of time (intervals!), and point tiers to annotate specific points in time. The choice to make the `Errors` tier a point tier here is a bit abitrary and just for illustrating what you can do with Praat.

Toggling the IPA symbol selector: You'll probably see a large table of IPA symbols on the right of the viewer. You can use this to add IPA symbols into annotations, but it takes up a lot of space. So, for the moment, let's just hide this by clicking the pink crossed boxed at the top right of this. You should see it turns to a pink triangle - clicking on this will show the IPA symbol table again.

Annotate the words

1. Click on any point in the word tier. You should see it turns yellow.
2. Find the start of the first word on the waveform (or spectrogram) and click there. You can do this by listening, but you should also be able to see where the sound starts in the waveform.

You should see a vertical line with some circles on the text tiers.



3. Click on the circle at the top of the `Word` tier to make a boundary. You should now see a vertical red line on the Word tier.
4. Do the same at the end of the first word. You now have a word interval.
5. Click in that word interval and type the word in ("She")

6. Continue through the recording an annotate all the words intervals.

Some things to note:

- There aren't really any gaps between words: words flow seamlessly into one another.
- You can see a very brief silence at the beginning of "by" in the waveform, but you won't really hear a pause in the speech.
    - **Question**: what's happening here in articulatory terms?
        - Answer: This is the closure portion of the [b] phone
- You need some context and knowledge about a language's written form (orthography) to place word boundaries.
    - **Question**: When you are transcribing speech how to do you know whether "sea" and "shells" are should transcribed as separate words or as a compound word ("seashells")?

      * Answer: You need some knowledge of the language and it's writing conventions. In this case, "seashell" as a noun is a single word. This is maybe a good time to point out that an ASR system that only recognises words, but doesn't know anything else about English wouldn't be able to distinguish "sea shell" and "seashell". So, this is some foreshadowing of the importance of a language model in automatic speech recognition.

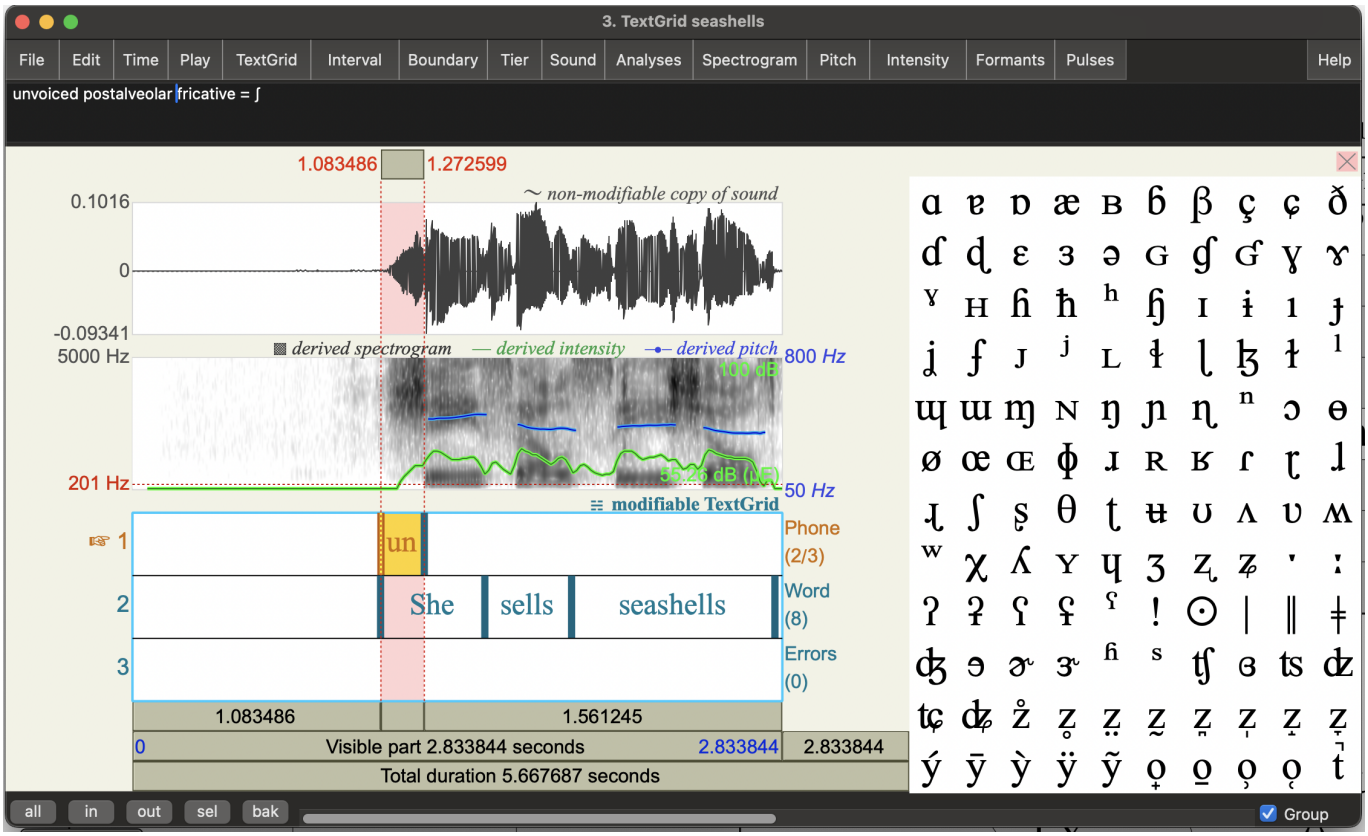
Annotate syllable initial consonants

Comment: The goal here is to break down articulations in terms of manner and place (hence the annotation task). We also start to see the relationship between speech sounds, i.e., *phones*, the waveform and the spectrogram. Even without much knowledge of what a spectrogram is, you should be able to see that there are some consistent patterns associated with specific types of speech sounds.

The tricksy bit of this tongue twister is the syllable initial consonants (aka syllable *onsets*). Let's see what's going on by annotating the first phone in each syllable for place and manner, in the phone tier. You may find it useful to say the phrase yourself and to determine what your articulators are doing.

1. Add boundaries for the start and end of the first phone in each of the syllables in the recording.
2. Using the IPA chart, annotate each the syllable initial phone interval with:
    - Voicing
    - Place
    - Manner
    - the IPA symbol

The interval box itself will be too small to see the full anotation, but you can see and edit the full thing up the top of the window. Here's the first one as an example (re-expanding the IPA symbol selector):



Annotate the vowels

Let's now look at the pattern of movement for vowels. Again, you may find it useful to say the phrase yourself and to determine what your articulators are doing.

1. Add boundaries for the vowels in this recording
2. Using the IPA chart, annotate each vowel with
    - height
    - backness
    - lip rounding
    - the IPA symbol

Where does your tongue twist?

1. Try to say the phrase as fast as you can until you start making errors.
2. Annotate the points where *you* made errors on the `Error` tier.
    - This is a point tier, so you'll be creating annotations for specific times rather than intervals.

Questions:

- What type of mistake are people likely to make with this tongue twister? Is it in articulation of voicing, place or manner?
    - Answer: Errors are usually to do with place of articulation: alveolar "s" [s] vs post-alveolar "sh" [ʃ]
- What other factors may be contributing to pronunciation difficulties here? (Hint: does saying words that rhyme make it easier or harder?)
    - Answer: The rhyming can also cause speakers to anticipate a specific articulation following the rhyming pattern. I don't have too much trouble with this for the seashells example, but I do with

other tongue twisters.

Waveform vs spectrogram: a preview of acoustic phonetics

Even without any training in spectrogram interpretation (i.e. acoustic phonetics), you should be able to see that fricatives are quite distinctive from other consonants in the spectrogram!

Questions:

- Based on what you see (i.e. don't worry about technical terms for now): how would you describe what fricatives look like in the spectrogram?
    - Fricatives are fairly distinctive for being quite solid blocks of grey/black on the spectrogram, usually with clear beginning and end points. This reflects the fact that turbulent flow ("white noise") contains a range of frequencies.
- What's the differences between [s] and [ʃ] sounds? How about [s] vs [ð]?
    - In my seashells example, [ʃ] has a clear dark band around 3500 Hz, which you don't see for [s]. In general, higher frequencies are more prominent for [s] than [ʃ], but you'll probably need to extend the frequency range on the spectrogram to see this for my recording (default is 5000 Hz).
- Can you tell these fricatives apart by just looking at the waveform (i.e. without the spectrogram)?
    - No, it's very hard to tell anything just from the waveform as both have a general "white noise" shape (i.e. lacking other cues like periodic structure or closures)
- The boundaries of fricatives are relatively clear. Is this the case for all phones? How about the vowel [ɛ] and following [l] sound in "sells"?
    - No, the change from [ɛ] to [l] is fairly continuous. You can basically track the continuous motion of the tongue from the vowel to the consonant. This makes the actual boundary quite hard to determine.

Example 2: Peter Piper

Comment: The recording sounds pretty clear if you just listen to it. The spectrogram is a good example of how reduced some vowels can get! The stops are fairly clear.

Usually we'd mark the start of each stop at the point of closure (when the tongue or lips actually stops air going through your mouth). But it's not that clear here where the stop segments should start and end because he's speaking at such a slow rate! In this case a little bit arbitrary. If you were doing this for a fine-grained phonetics analysis you would need to make some decisions on how to be consistent with this. For automatic word transcription, we'll see that we don't need to be too worried about what the exact boundary is as long as it is consistent.

Tongue-twister: "Peter Piper picked a peck of pickled peppers"

1. Create similar annotations for the first sentence of the "Peter Piper" tongue twister:
    - english_peter_slow_pb.wav
    - This time you'll want to focus on word internal consonants.

2. What sorts of manner and place variations cause difficulty in this example?
    - As mentioned above, we have sequences of [p] vowel [p, t, k]. The vowels in the first three "p" words are high front ([i], [al], [I]) in the first syllable while the last three are front but vary in height [ɛ] (low-mid), [I] (high), [ɛ] (low-mid)
3. What sorts of errors are speakers likely to make?
    - mixing up the word internal stops [p, t, k] and/or swapping a vowel in the last words (e.g., [I] to [ɛ])
4. What cues are you using to identify *plosives* (aka oral stops) in the example in terms of waveform and spectrogram. an you identify them just from the waveform?

    - plosives are one of the easier phones to recognize from the waveform. They are characterised by a closure (often seen as a flat line in the waveform) followed by a burst (large spike). You can see these in the waveform, but you won't be able to identify which plosive it is just from the waveform.

    - We see this in the spectrogram a light/grey colored section (closure) followed by a dark vertical line (burst). After the burst there is often a period of *aspiration:* turbulent air flow from the burst. In the spectrogram this will look like a fading away of the burst. In module 2 we'll see that we can guess which plosive from the spectrogram, often by the way it affects the spectrum of the following vowel.

Record and analyse your own tongue twister

Now let's try recording a tongue twister yourself and analysing it. If you speak a language other than English, you might like to try one in another language.

After recording yourself in Praat (see instructions below), try to identify the articulation patterns that cause difficulty. Again, think about whether the confusions/errors that arise are in terms of placement of articulators. Describe this in terms of voicing, place and manner for consonants. For vowels, think about tongue frontness, height, and rounding. We've focused mostly on consonants in this lab, but don't worry we'll do more on vowels next week.

Comment: This is part of the lab is more of an extension and that you can have a go at recording yourself. I think it's also fun to see that tongue twisters are something common across languages!

I wouldn't expect full analyses of these examples. The main thing is to think about what your articulators are doing if you try to say some of these.

If you do the recording in a noisy environment, you'll see that the noise makes the spectrogram less clear.

Examples

Some more English example:

- Seventy seven benevolent elephants
    - See links to recordings above!
- Should such as shapely sash such shabby stitches show

    - english_sash_slow_pb.wav (slow)
    - english_sash_fast_pb.wav (fast)
        - This is another s vs sh one

- Red lorry, yellow lorry, red lorry, yellow lorry

    - english_lorry_slow_pb.wav (slow)
    - english_lorry_fast_pb.wav (fast)
        - Distinctions between approximants are hard to maintain here: i.e., "r" v "l" v "y"

- I'm not a pheasant plucker, I'm a pheasant plucker's son.

    - english_pheasant_slow_kr.wav (slow)
    - english_pheasant_fast_kr.wav (fast)
        - This one is slightly NSFW, but a good example of "ph" -> [f] and that getting transposed to the following "pl" word. Both have lips as place of articulation, but the manner makes a lot of difference!

You can find many more on the internet!

And for inspiration, here are a some tongue twisters in other languages offered up by members of the Centre for Speech Technology Research (including the lab tutors):

- Catalan:

    - Setze jutges d'un jutjat mengen fetge d'un penjat
    - "Sixteen judges of a court eat liver off a hangman"
        - catalan_slow_asc.wav
        - catalan_fast_asc.wav

- Spanish:

    - Tres tristes tigres comen trigo en un trigal
    - "Three sad tigers eat wheat in a wheat field"
        - spanish_slow_asc.wav
        - spanish_fast_asc.wav

- Icelandic:

    - hnoðri í norðri verður að veðri þótt síðar verði.
    - "A small cotton ball (cloud) in the north becomes weather sooner or later"
        - icelandic_slow_as.mp3
        - icelandic_fast_as.mp3

- Czech:

    - Strč prst skrz krk
    - "Stick a finger through the neck"
        - czech-slow-ok.mp3
        - czech-fast-ok.mp3

- Mandarin Chinese:

    - 四是四，十是十，十四是十四，四十是四十，四十四是四十四
    - "Four is four, ten is ten, fourteen is fourteen, forty is forty, forty-four is forty-four."
        - Mandarin_slow_yl.wav
        - Mandarin_fast_yl.wav

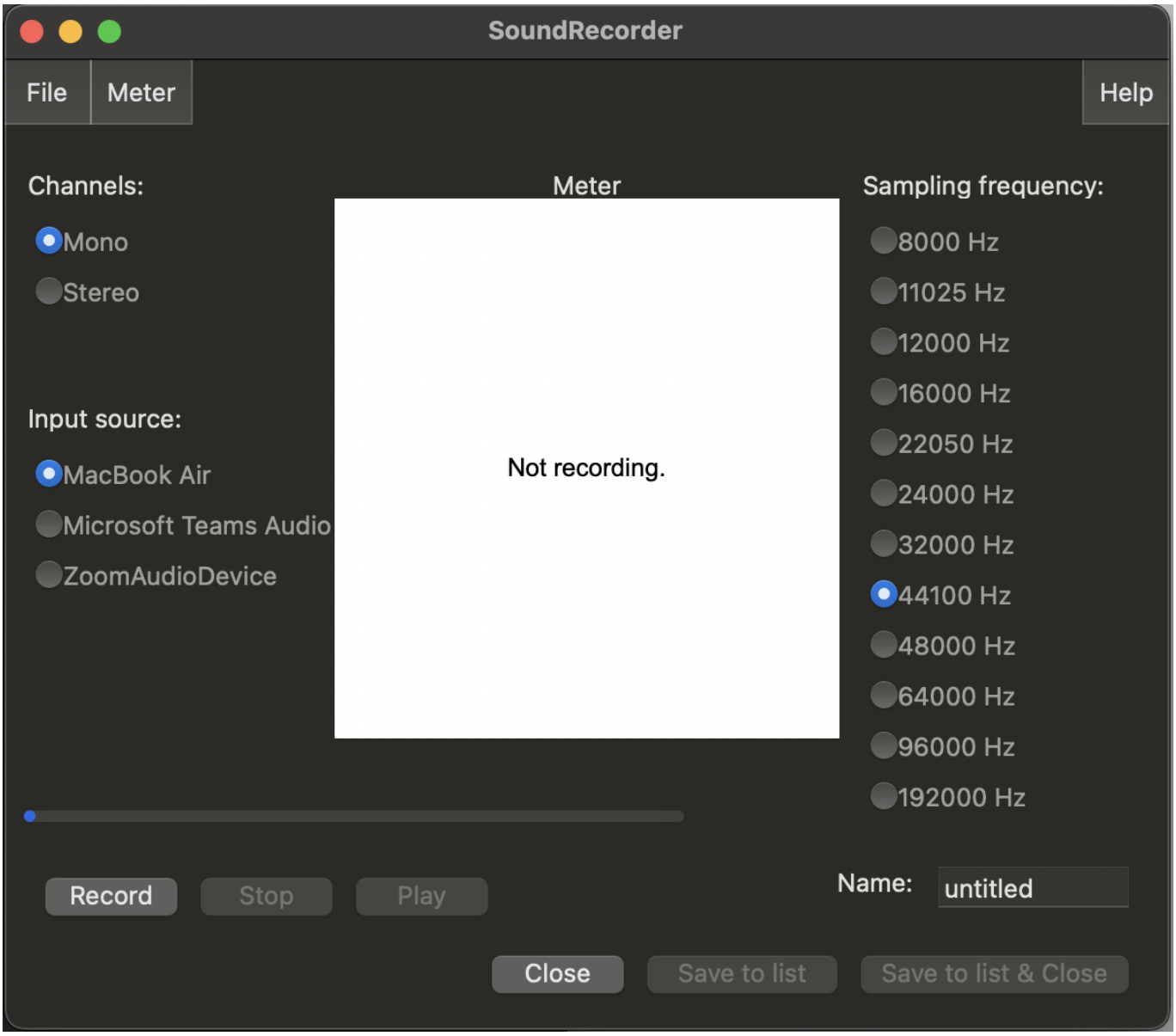
- Japanese
    - 生麦、生米、生卵
    - "Raw wheat, raw rice, raw egg"
        - Japanese_slow_yl.wav
        - Japanese_fast_yl.wav

- Vietnamese
    - Tâm tưởng tôi tỏ tình tới Tú từ tháng tư, thú thật, tôi thương Tâm thì tôi thì thầm thử Tâm thế thôị
        - vietnamese-slow_md.mp3
        - vietnamese-fast_md.mp3

You can also find many others linked in the description of this video by Hank Green: Tongue twisters. This also has a nice discussion of why tongue twisters are hard!

Make a recording in Praat

1. Click on New in the *Praat Objects* window
2. Select Record mono sound...
3. You should see a *SoundRecorder* pop up window like the following



There are 3 main parameters you can change:

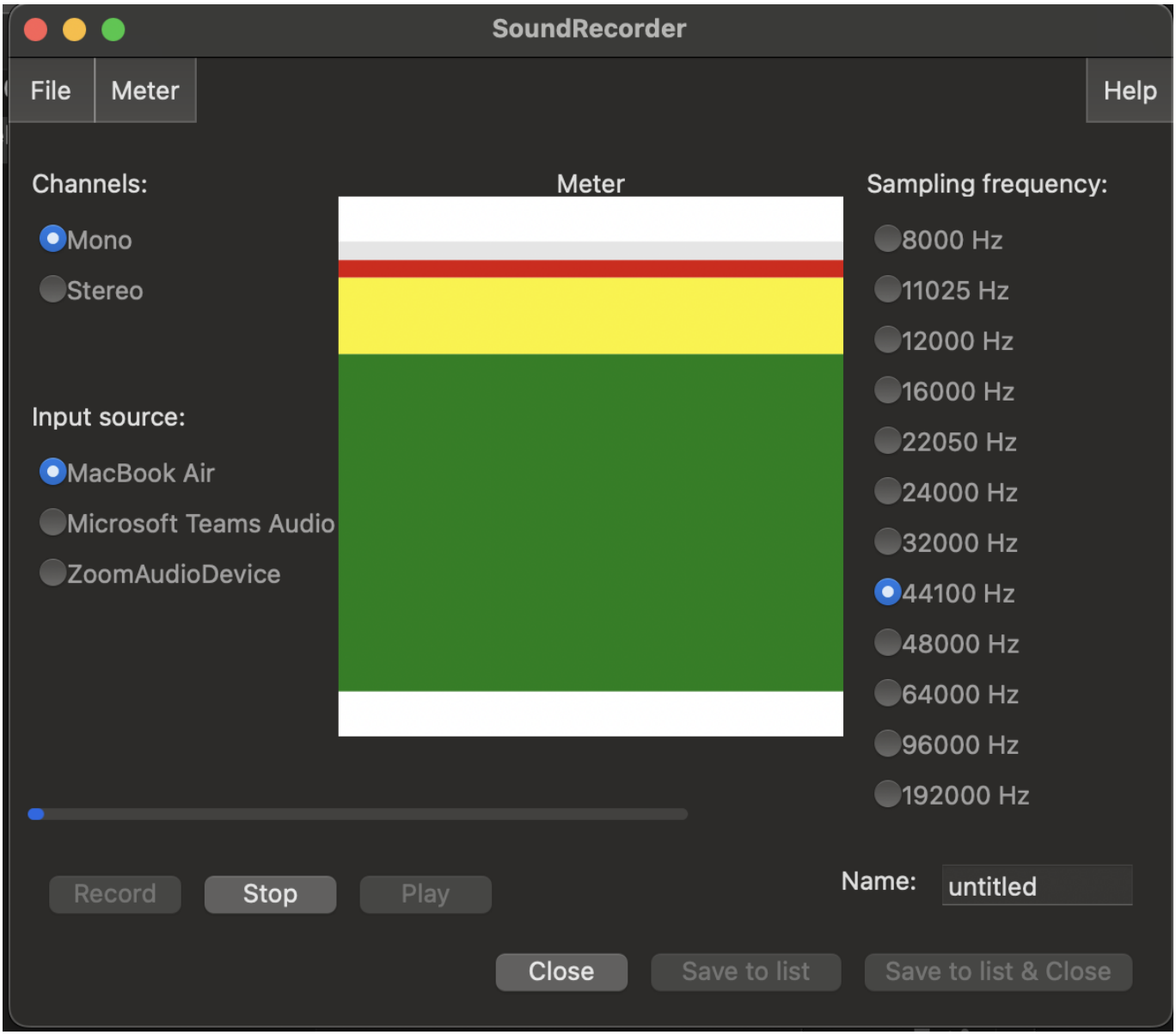
- Channels: Mono or Stereo
    - Mono basically records all sound into one channel (i.e.waveform). But if you have a stereo microphone, you can try it to see what the difference is.
- Input source: which microphone to use
    - If you don't pick up anything on recording, this is the first place to check.
- Sampling frequency: How often to sample sound (i.e. air pressure) at the microphone. This is measured in Hertz (Hz) which you can also read as samples per second.
    - We'll stick with the default 44100 Hz for now. We'll come back to the difference this makes in module 3, but feel free to play around and see for your self.

4. Change the *Name* of the recording from `untitled` to whatever you'd like it to be.
5. Press `Record` to start recording and `Stop` to stop the recording.

When you start speaking you'll see some colours appear in the the *Meter* box. This will tell you if the sound level is at an appropriate level. If you see some green movement, you should fine. If you see the meter go into yellow up into red, the sound is too loud to capture faithfully and you likely get distortion in the

recording. This usually happens if your microphone volume is too high and/or you're too close to the microphone (we'll come back to this in module 3).

The following shows the meter going into the red (produced by clapping several times with the microphone set to high volume):



6. You can play back your recording using the `play` button. When you're happy with it click `Save to list & Close`.

You should now see a new `Sound` object in the *Objects* window with the name you gave your recording.

Analyse your recording

As for the other examples, add a TextGrid for annotations and inspect the audio.

- Explain where and why your chosen text is difficult to say in terms of your articulators or any other relevant factors.
- Annotate the recording to show what is going on. It may be helpful to make a second recording where you make errors to contrast this to the correct pronunciation!

End notes

The goal of this lab was to get you thinking about how people create speech using actual physical articulators in our vocal tracts. Tongue twisters show that this process, between thinking and speaking, is actually very complicated.

Speaking is also constrained by the physicality of our actual articulators and respitory systems. It's actually proven very hard to reproduce human speech in purely physical models. You can get an idea of how difficult this problem is by looking at the work from the lab of Prof. Takayuki Arai (Sophia University, Japan). See this recent paper, for example:

- Arai, T., Suzuki, R., Earp, C., Tsuji, S., Ochi, K. (2024) Production of phrases by mechanical models of the human vocal tract. Proc. Interspeech 2024, 987-988

There are several other very interesting demos on the lab's youtube page: Acoustic-phonetics demonstrations

You're probably now getting to understand why most humanoid robots don't even attempt include vocal tracts! Instead, we generally synthesize speech waveforms using non-physical means on computers and play them out some speakers. To do this we'll need to understand how we can "see speech" just from the waveform: i.e., acoustic phonetics. This is the focus of module 2.