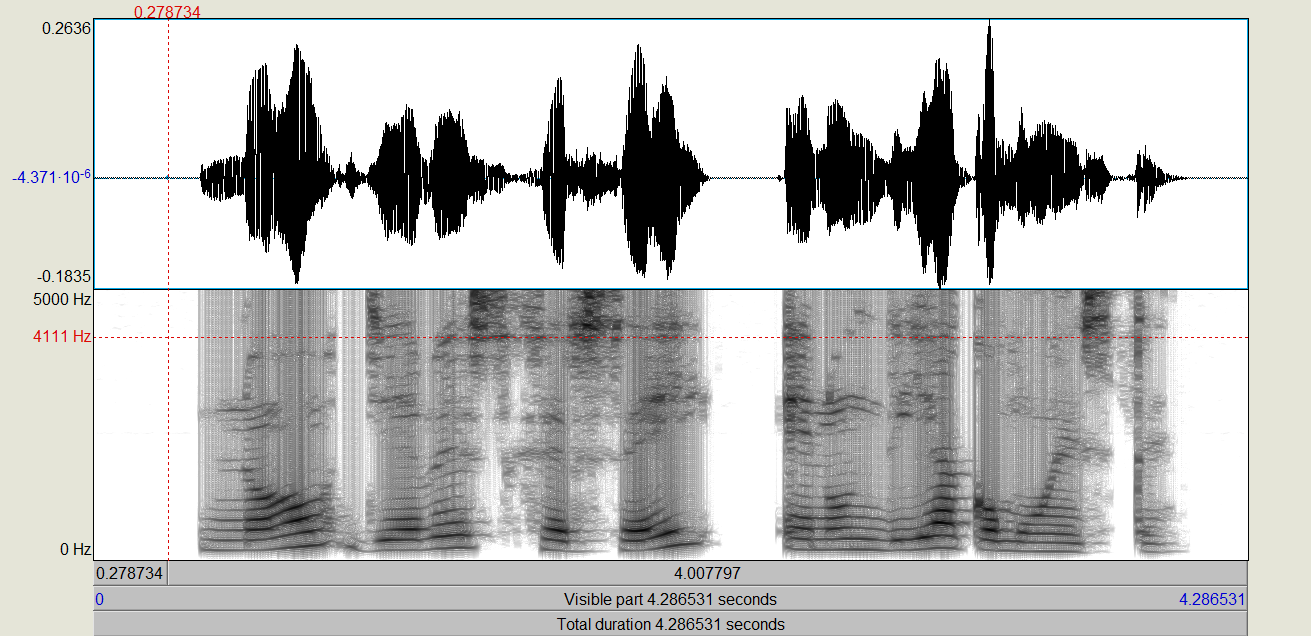
BASIC CONCEPTS ABOUT SPEECH

SPEECH SIGNAL ANALYSIS (PART I)

1. INTRO

2. BANDWIDTH OF THE SPEECH SIGNAL

For this section, we are going to use the recorded sentence “No lleves jersey en un día como este”. The first step in order to develop the exercise is opening the file:  
(FIGURE 1)



In this case, the signal is a narrowband signal with a total duration of 4.29 seconds. In Figure 2, we can observe the pitch and the formants of the recorded sentence.

(FIGURE 2)

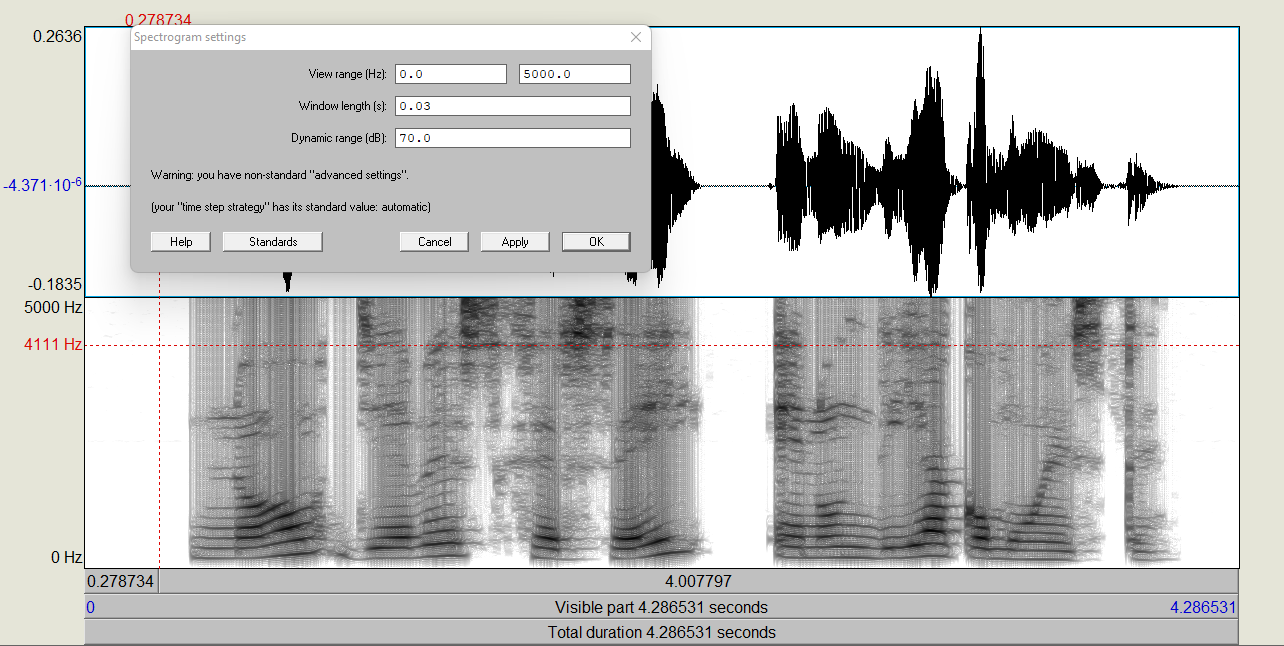
Imagen que contiene luz, colgando, tráfico, reloj

Descripción generada automáticamente

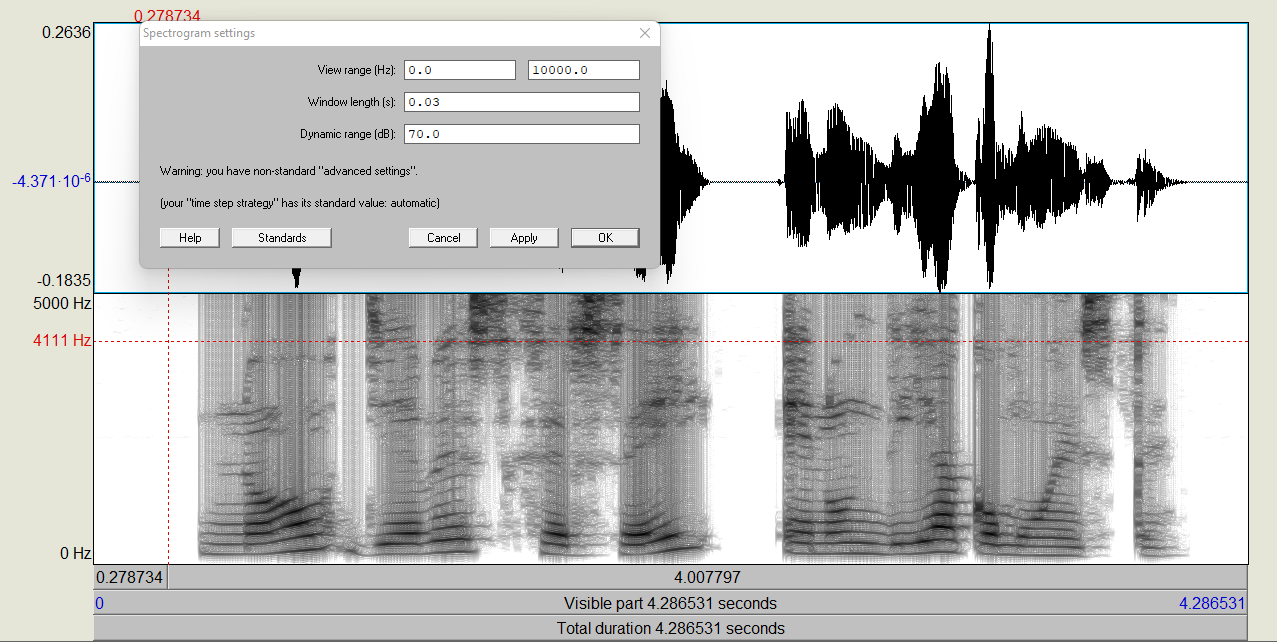
Now, we are going to focus on the spectrogram of our recorded signal, and we are going to measure the approximate bandwidth of this voice signal. The spectrogram could also be referred as heatmap, because it represents the accumulation of frequencies in a specific point in time. In Praat the spectrogram is grayscale, hence the points in time with more frequency are going to be represented in darker tones. Consequently, the absence or the lowest accumulation points are going to be lighter, with a tendency to white.

Exploring the bandwidth of our recorded signal is the first brick for building our house. Therefore, the first step is escalating the spectrogram, in other words, we are going to increase the range frequency until we cannot observe more heatmap. In Figure 3, we have clicked in Spectrum > Spectrogram settings… and we have some parameters. The first option is the frequency range, this means we are going to settle the initial and final point for overview our signal in terms of frequency. In Figure 4, we have changed the value for the end of the range to 10000.

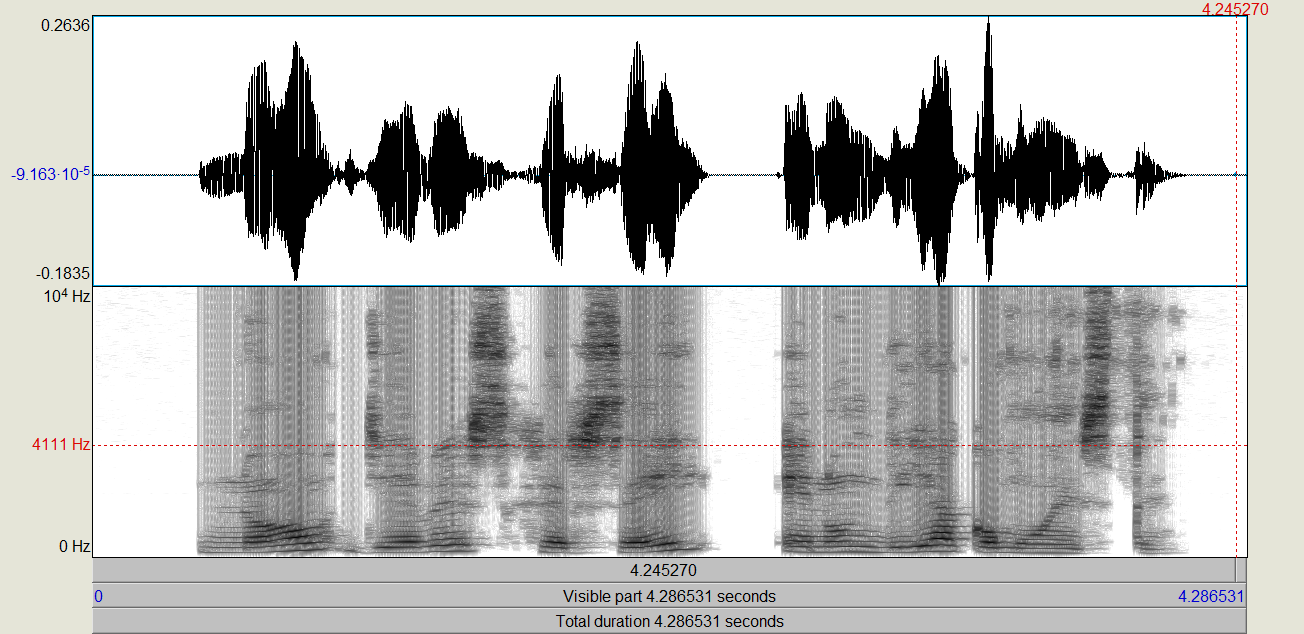
(FIGURE 3)



(FIGURE 4)

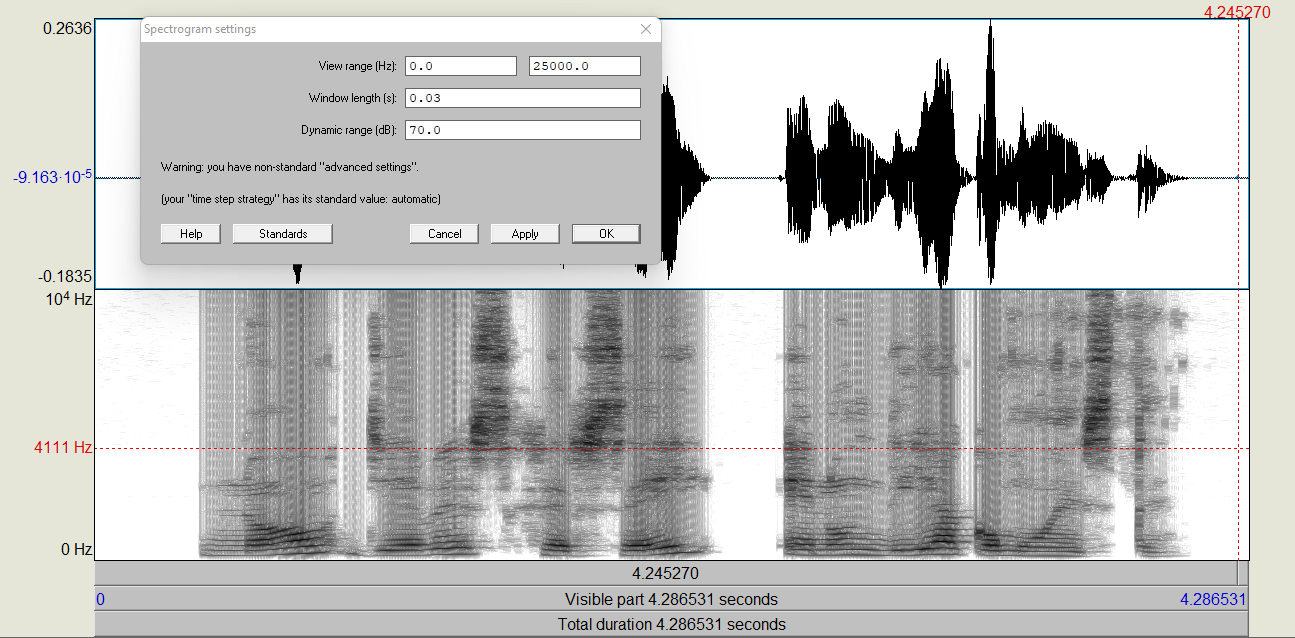


(FIGURE 5)

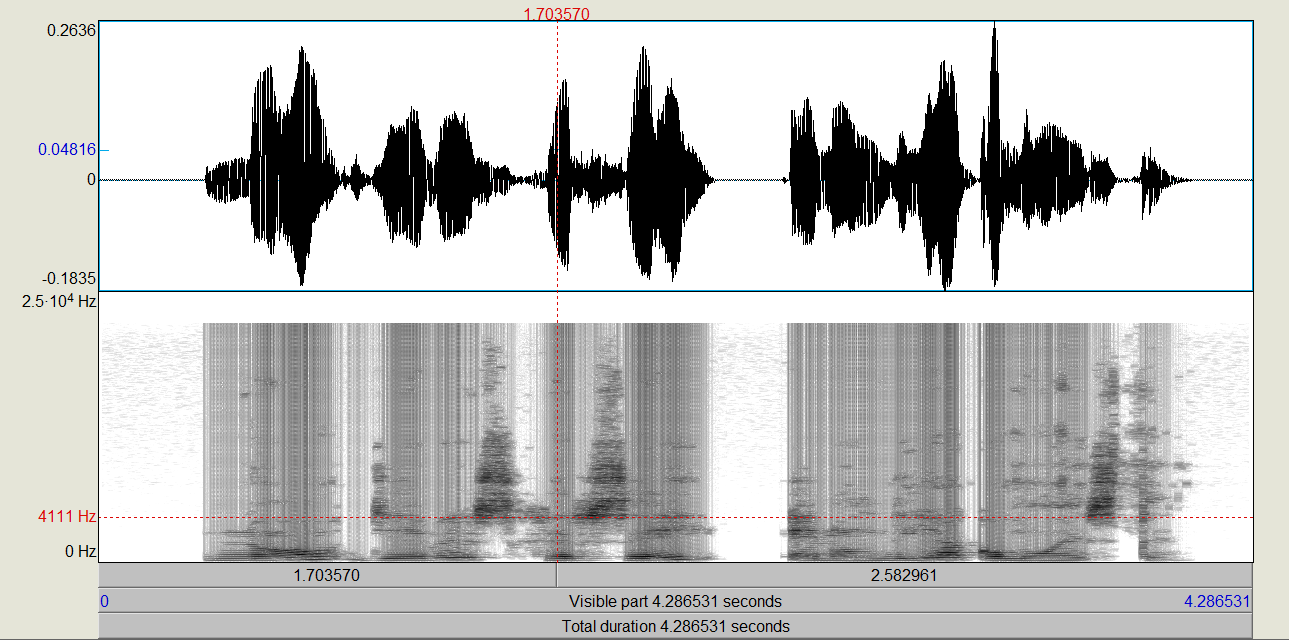


The result of changing the frequency range could be analysed in Figure 5, where we can appreciate darker colours in our heatmap. However, we have not arrived yet to the limit of the bandwidth. Therefore, we have increased more the frequency range as it is shown below in Figure 6, and the result of this change in Figure 7.

(FIGURE 6)

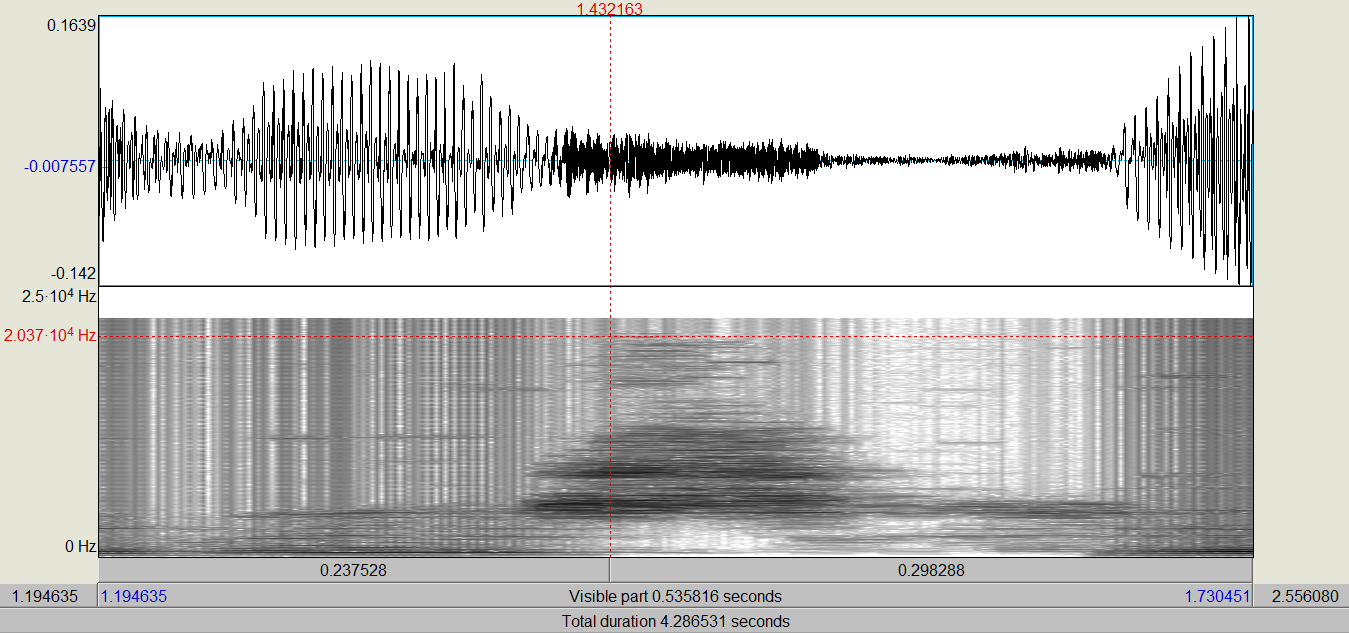


(FIGURE 7)



This last figure shows the limit of the bandwidth at some point around 22.000 Hz. Moreover, there are some frequencies that have really high values which leads us to the second part of this first assignment: what kind of sounds are those reaching highest frequencies in your sentence? Figure 8 shows one of the pieces of signal that gets to frequencies around 20.000 Hz.

(FIGURE 8)



Normally, voiced sounds reach the highest frequencies but, in this case, due to demographic particularities of the speaker is the letter ‘s’. These high frequencies of approximately 20.000 Hz could be spotted three times in our recording: the first from 1.42 to 1.50 seconds, the second from 1.84 to 1.94 seconds and the third from 3.68 to 3.82 seconds.

It is highly important having a sampling frequency that can show the specifications of the waves. This leads us to the issue about sampling frequency: which would be the minimum sampling frequency? When do we lose quality of the sound? After some experiments, we have considered that 2000 Hz is the minimum sampling frequency required in order to continue having an understandable record signal, although the highest frequencies have been deleted, and the speech seems like someone talking under the water.

Observing the spectrogram of the recorded signal we can assure that the maximum bandwidth has been occupied by the sound ‘s’ in the three different moments in time that have been pointed out before.

The last process of Section 2. is how to use a lowpass filter and analyse at which *cutoff* *frequency* the sentence becomes unintelligible describing all the intermediate steps. Firstly, we should define what is an acoustic filter and, more precisely, the lowpass filter. An acoustic filter is “used to filter unwanted frequencies of oscillations from a power system”[[1]](#footnote-1). Therefore, an acoustic filter allows or blocks the flow of components of sound with different frequencies. In this case, a lowpass filter is going to reject high-frequency elements of a wave, allowing only the low-frequency ones.

We should go to our panel of Praat Objects where we have our recorded sentence. Then, on the right part of the window we click Filter > Filter (pass Hand band)… and we get the window shown in Figure 9.

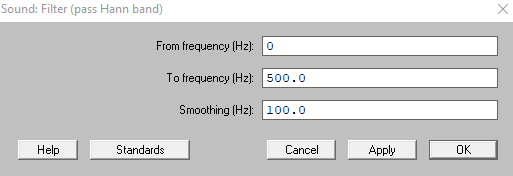
(FIGURE 9)

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

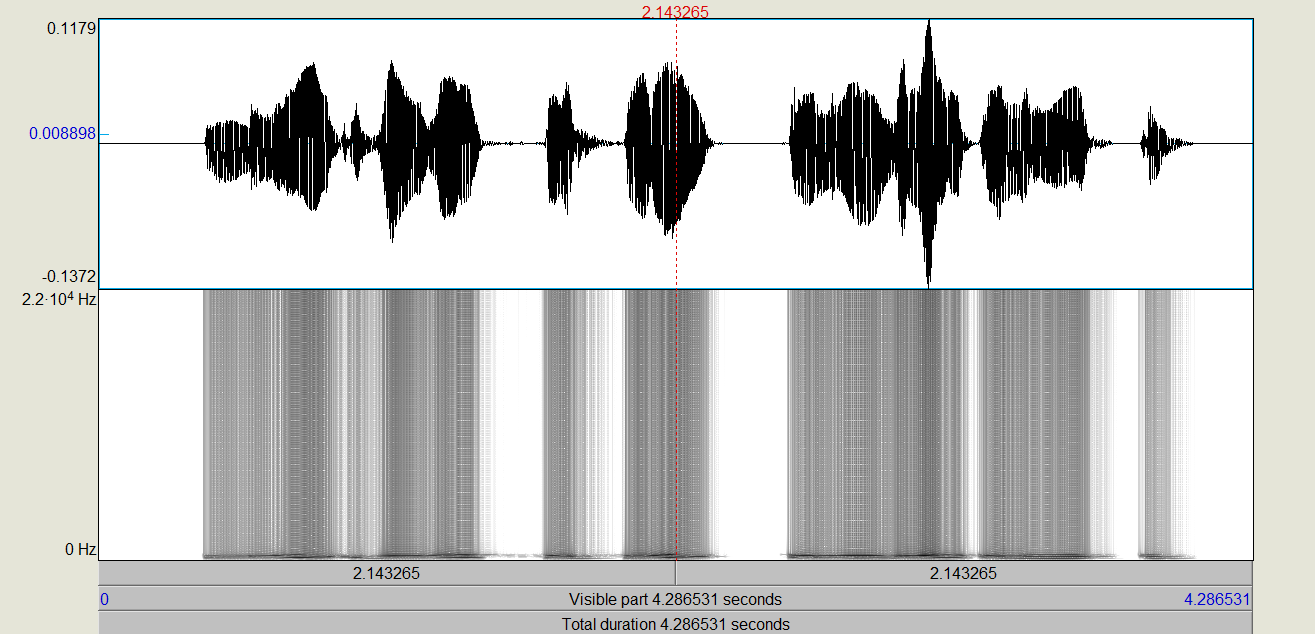
We should change some parameters in order to apply properly the lowpass filter. For instance, the initial frequency (From frequency (Hz)) should be zero in order to allow the lowest frequencies to pass the filter. Then, the second parameter is the one we are going to change continuously to reach the aforementioned *cutoff frequency* that is going to create an unintelligible signal. In Figure 10 the conclusive value for the *cutoff frequency* is 500 Hz. After progressively reducing the bandwidth of our signal, we noticed that around 500Hz the sentence was not comprehensible.

(FIGURE 10)



In addition, we see several changes in the spectrogram of the signal (Figure 11). Almost all the frequencies have disappeared.

(FIGURE 11)

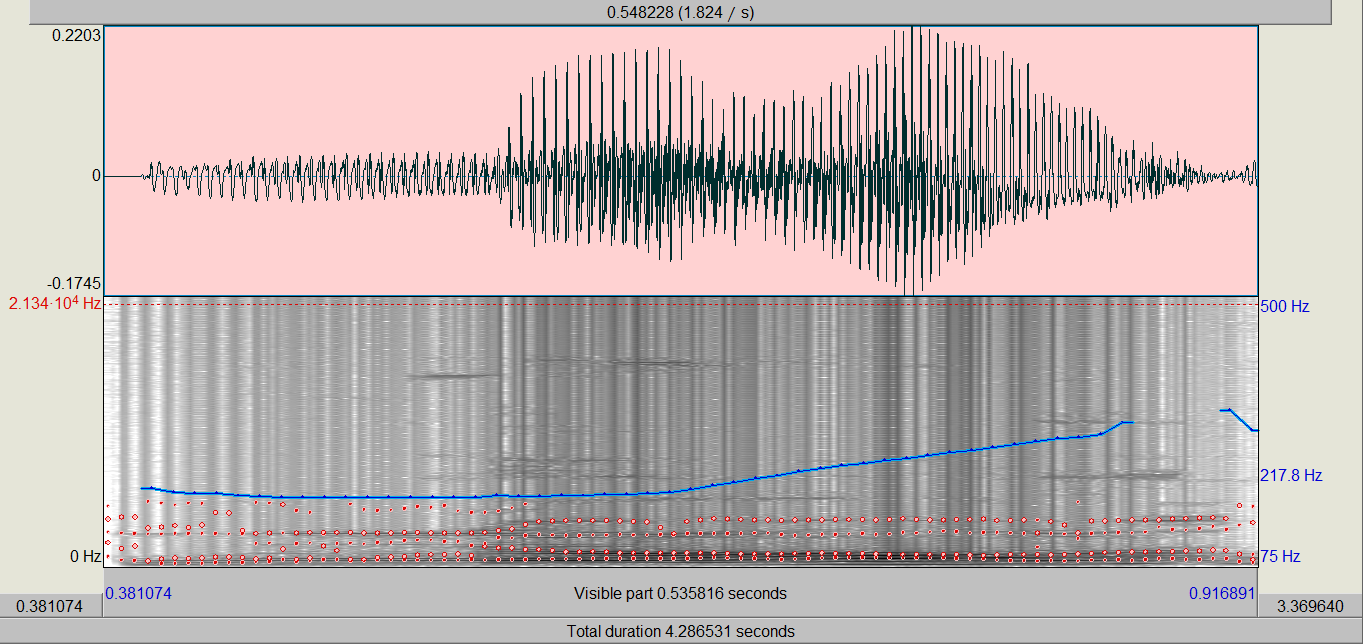


3. WAVEFORM, SPECTROGRAM AND SPECTRUM OF VOICED AND UNVOICED SOUNDS

In this part of the practice we are going to analyse the syllables of our recorded sentence and detect the voiced and unvoiced sounds. We are going to take into account both the waveform and the spectrogram for providing and accurate and specific description of the differences between voiced and unvoiced segments.

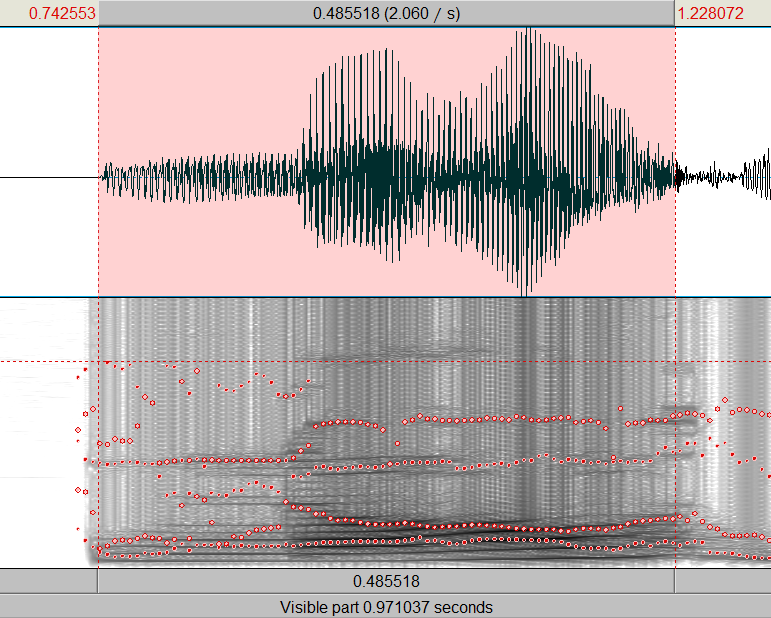
The first syllable of our recorded sentence (“No lleves jersey en un día como este.”) is the monosyllabic negative particle “no” in Spanish. In Figure 12 we have both the waveform of the first syllable, “no”. The phoneme /n/ is an alveolar nasal voiced sound. In the waveform is represented with less amplitude than the next sound because the vibration of the vocal cords is smaller than in vowels. In the spectrogram the frequency line is lower compared to the vowel, the initial frequency of the syllable is around 180Hz and it gets to 210Hz. The phoneme /o/ is a vowel, hence a voiced sound. The amplitude of the vowel is way bigger than the consonant, getting its high peak at 0.22.

(FIGURE 12)



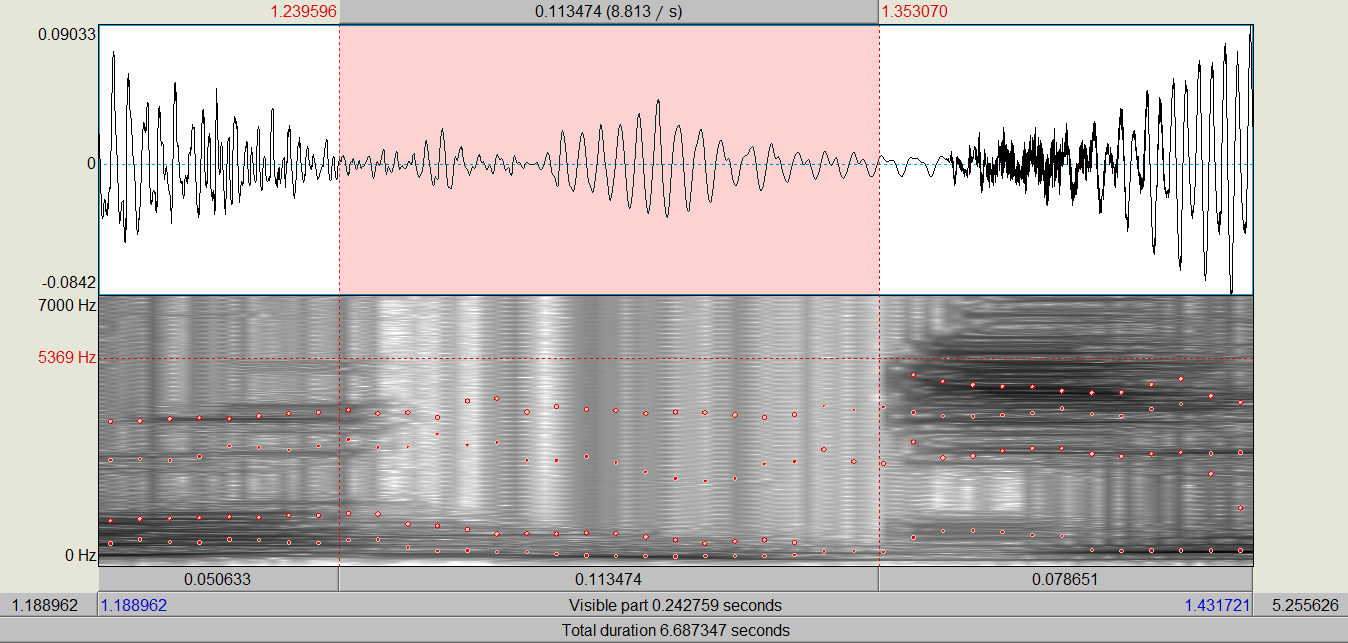
Moreover, in Figure 13 we can appreciate that the accumulation of frequency is higher in the vowel /o/ than in the consonant /n/:

(FIGURE 13)



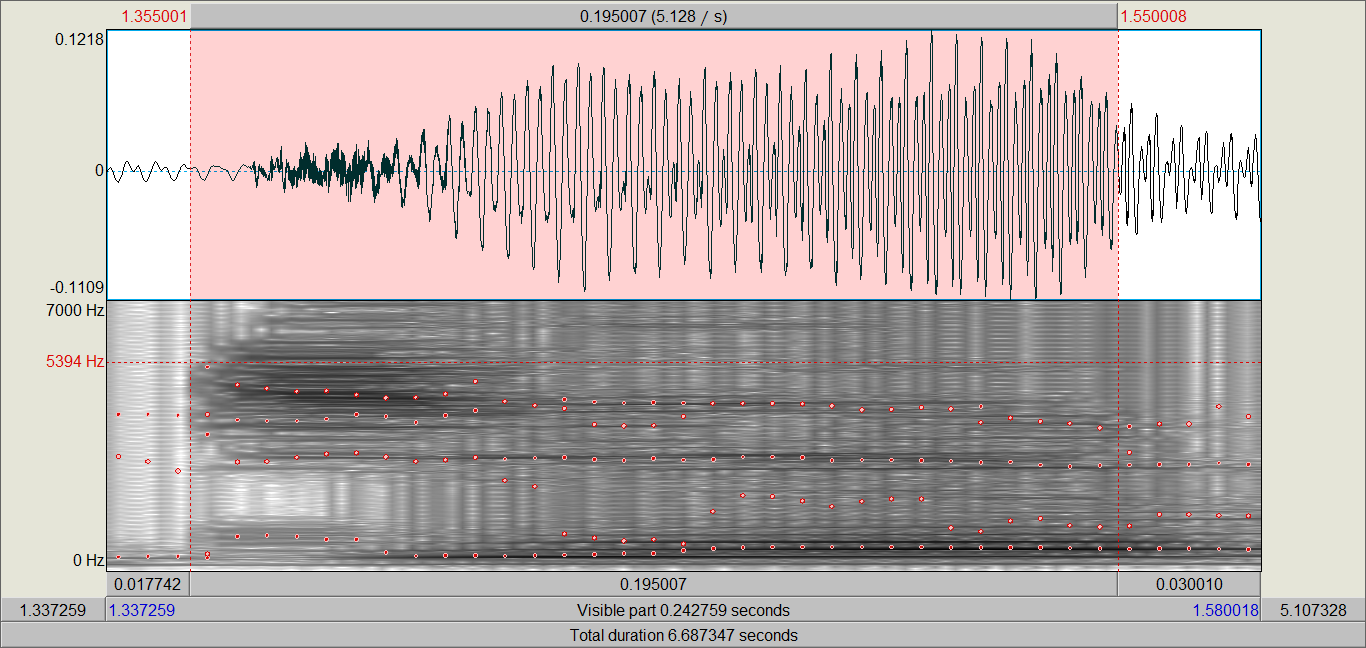
The following syllable is “lle” from the word “lleves” (in English ‘wears’). The first phoneme of the syllable is the palatal lateral voiced consonant, /ʎ/, and the second element is the vowel /e/, it is also a voiced sound. In Figure 14, we observe the formant transition from the vowel /o/ to the consonant /ʎ/:

(FIGURE 14)



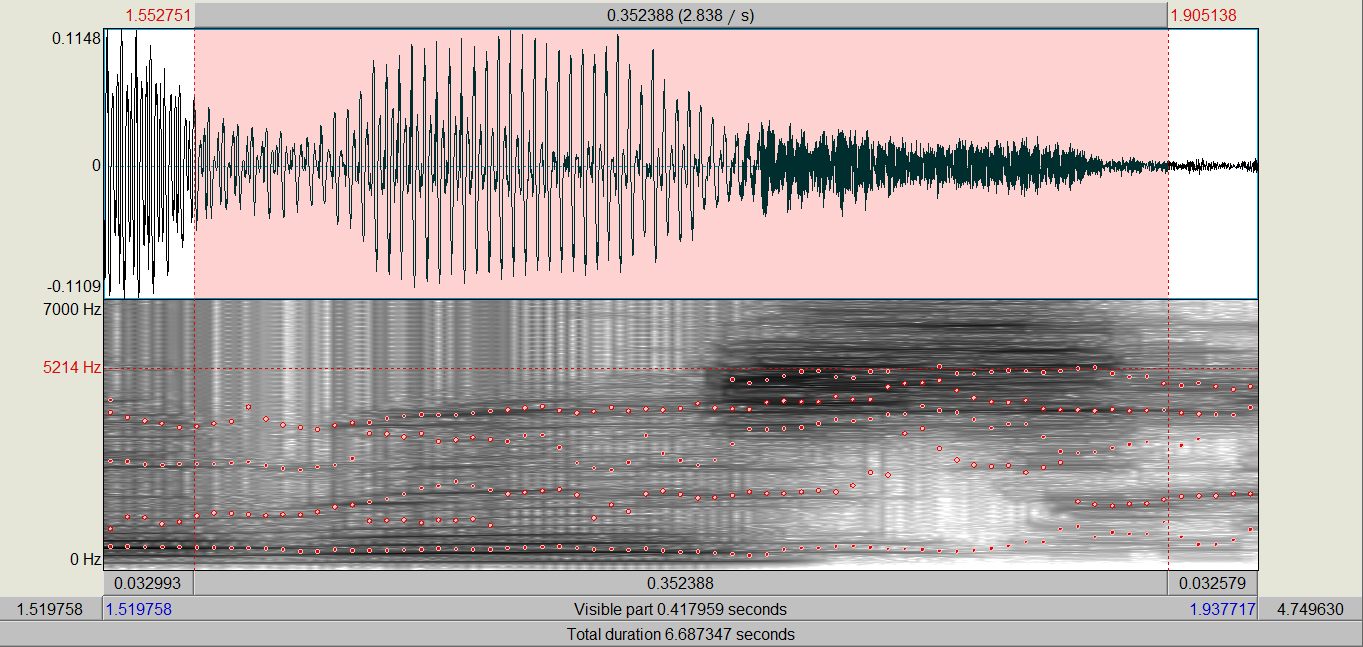
In Figure 15, our first impression is that the consonant sound is smaller in terms of amplitude in the waveform compared to the vowel. Moreover, in the spectrum the first part of the syllable, corresponding to the consonant, is lighter in colour than the vowel at the bottom (lower frequencies), and darker in high frequencies. We can appreciate also more concentration of darker colours in higher frequencies, approximately until 7000 Hz. Moreover, the beginning of the waveform of this syllable (the consonant, /ʎ/) is more irregular than when the wave changes into the vowel /e/.

(FIGURE 15)



Then, we get to the syllable “ves” which composition is CVC (consonant-vowel-consonant). The first element is a voiced bilabial fricative consonant, /β/, the vowel /e/ and the alveolar fricative unvoiced consonant, /s/. In Figure 16, we appreciate the differences between the parts of the syllable:

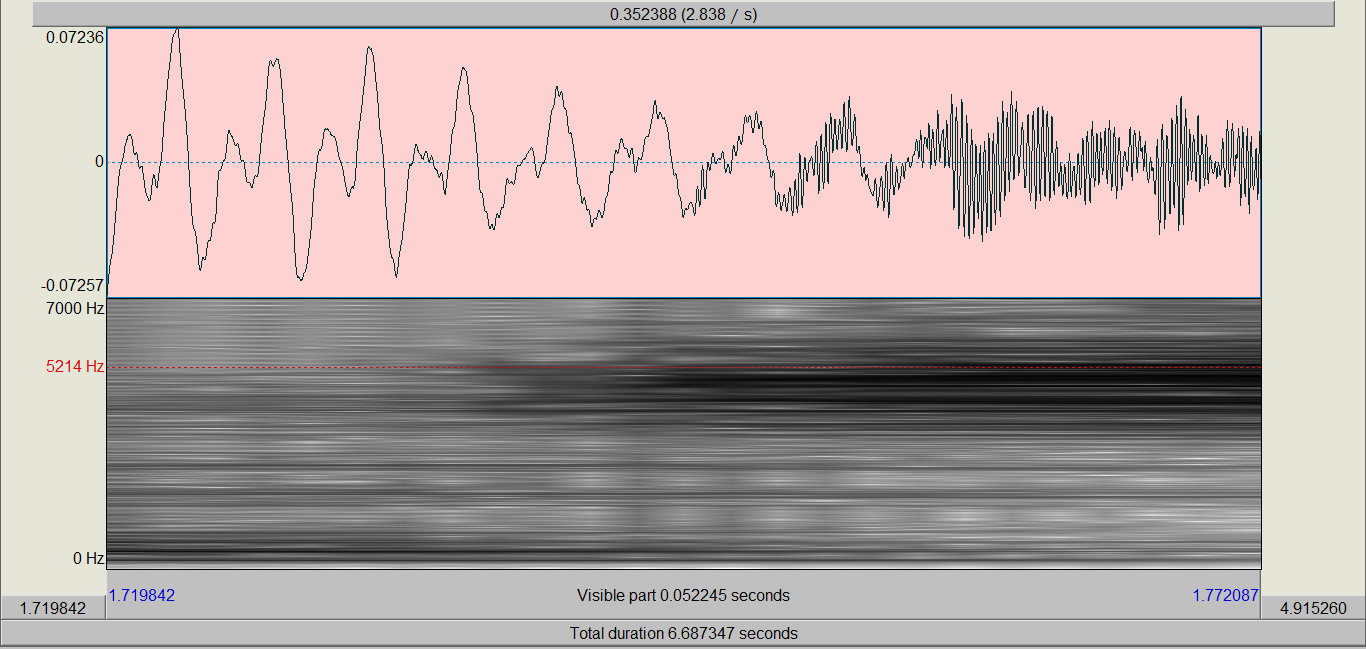
(FIGURE 16)



The first sound /β/ presents a lower amplitude in the waveform compared to the following sound, the vowel /e/. Moreover, this consonant has not much dark colours in the spectrogram. The frequency of /β/ is quite low and also the time for producing the consonant is quite short.

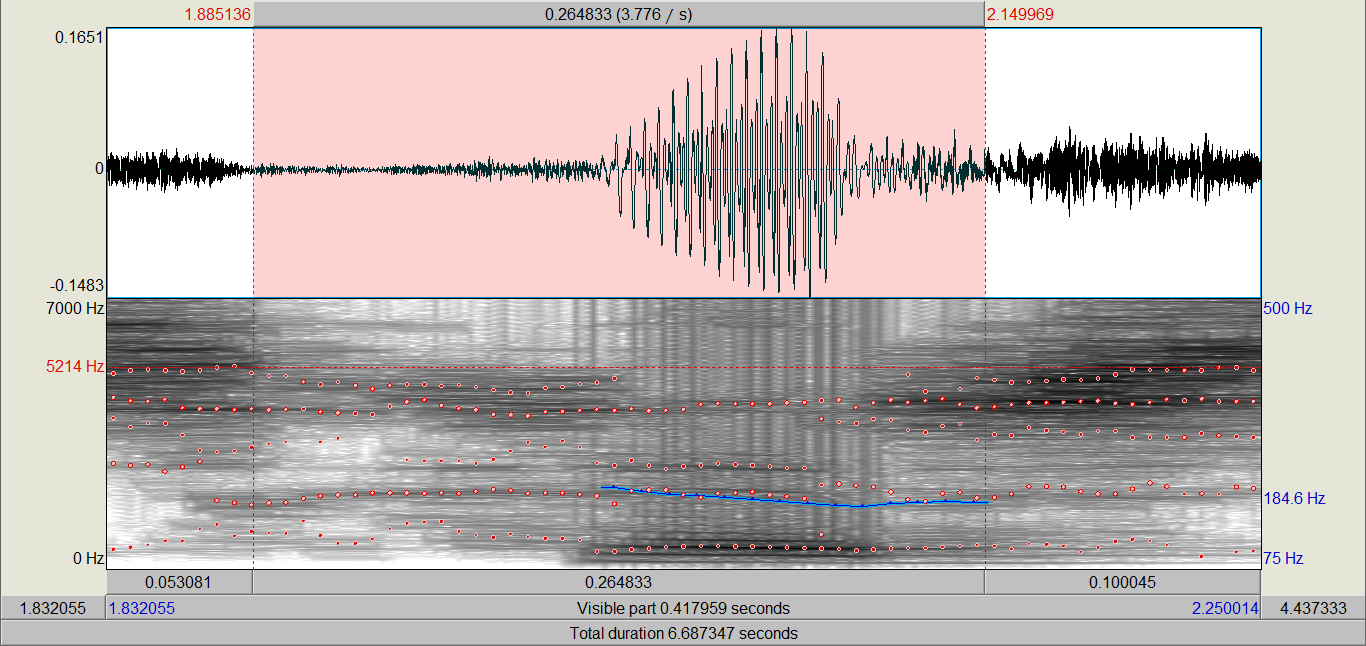
The vowel /e/, as well as in other cases, it has bigger amplitude in the syllable and more frequency shown in the spectrogram. It is interesting because in relation with the following sound, the /s/, the structure of the waveform changes from regular to random. We can see this change in the structure in Figure 17:

(FIGURE 17)



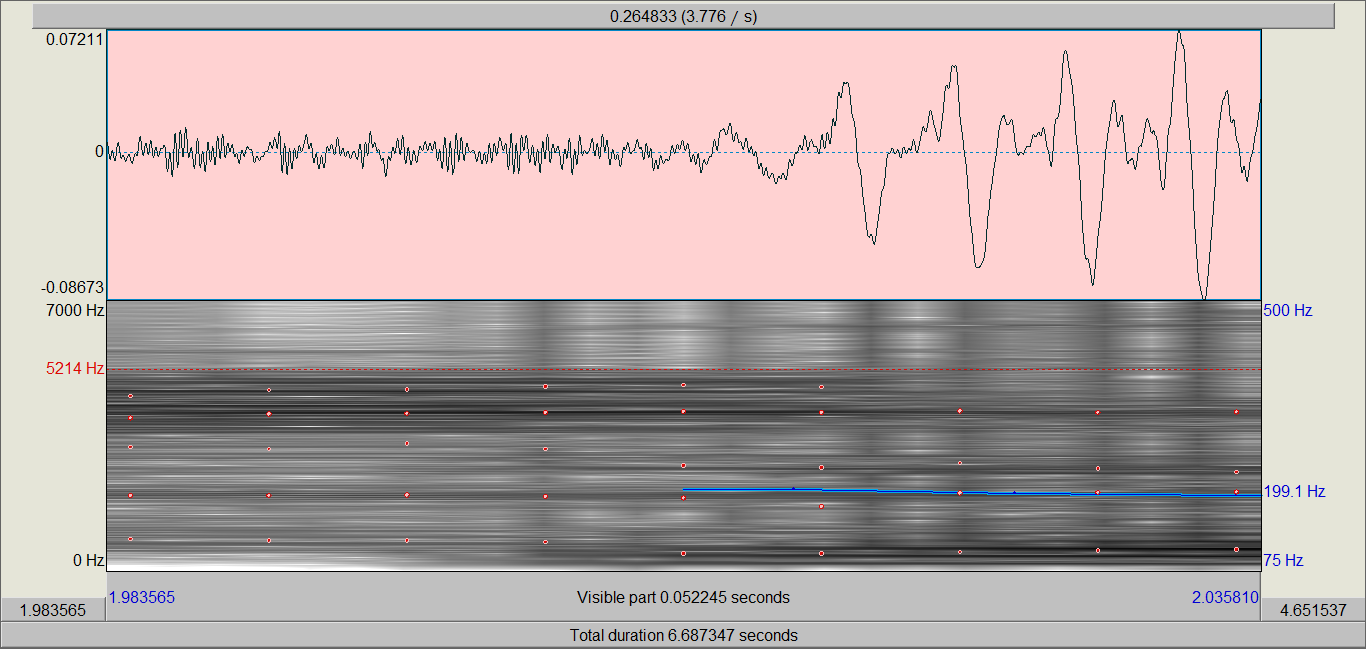
The syllable that we are going to analyse now is the first syllable “jer” from the word “jersey”. The first component of the syllable (which has a structure CVC) is the voiceless velar fricative /x/.

(FIGURE 18)



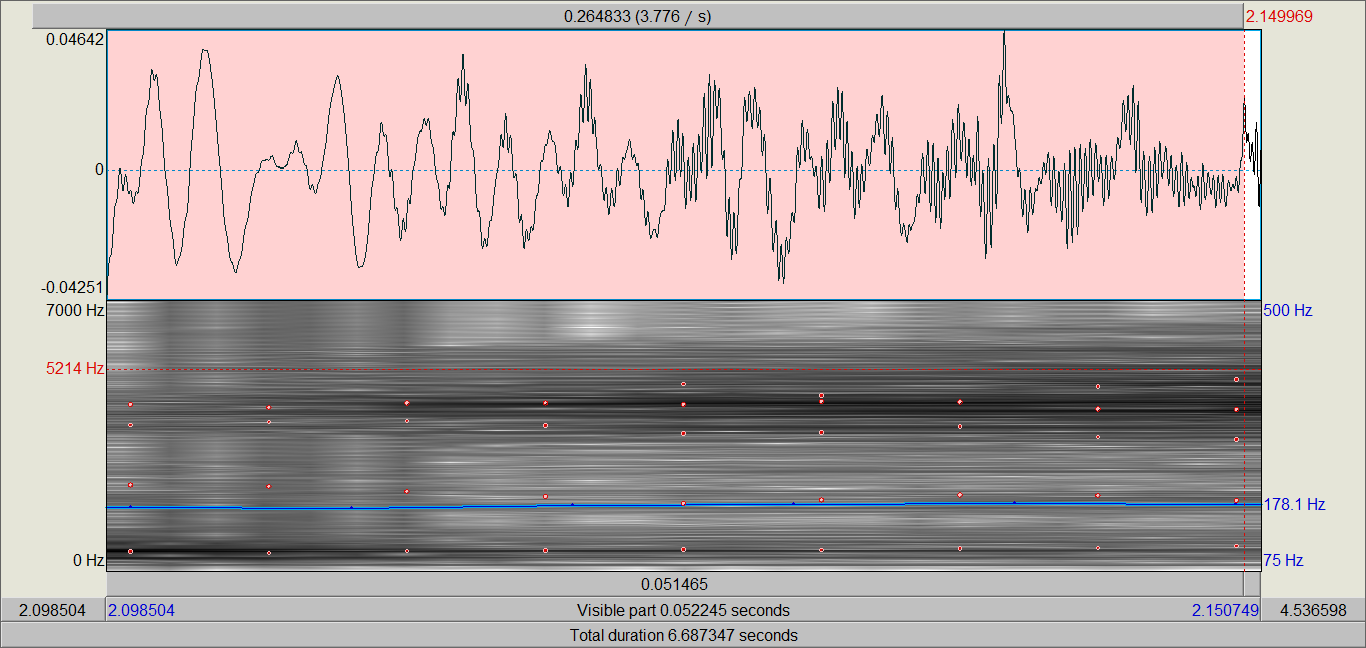
In Figure 18 we can observe how the bottom part of the spectrogram is almost void (it is almost white, because fricatives don’t have fundamental frequency) and the waveform is presented random and almost flat in terms of amplitude. Furthermore, in the spectrogram we see that when it changes from /x/ to /e/, some vertical lines (glottal pulse) appear when /e/ is being produced. In Figure 19 we can recognize the change in the wave from /x/ to /e/:

(FIGURE 19)



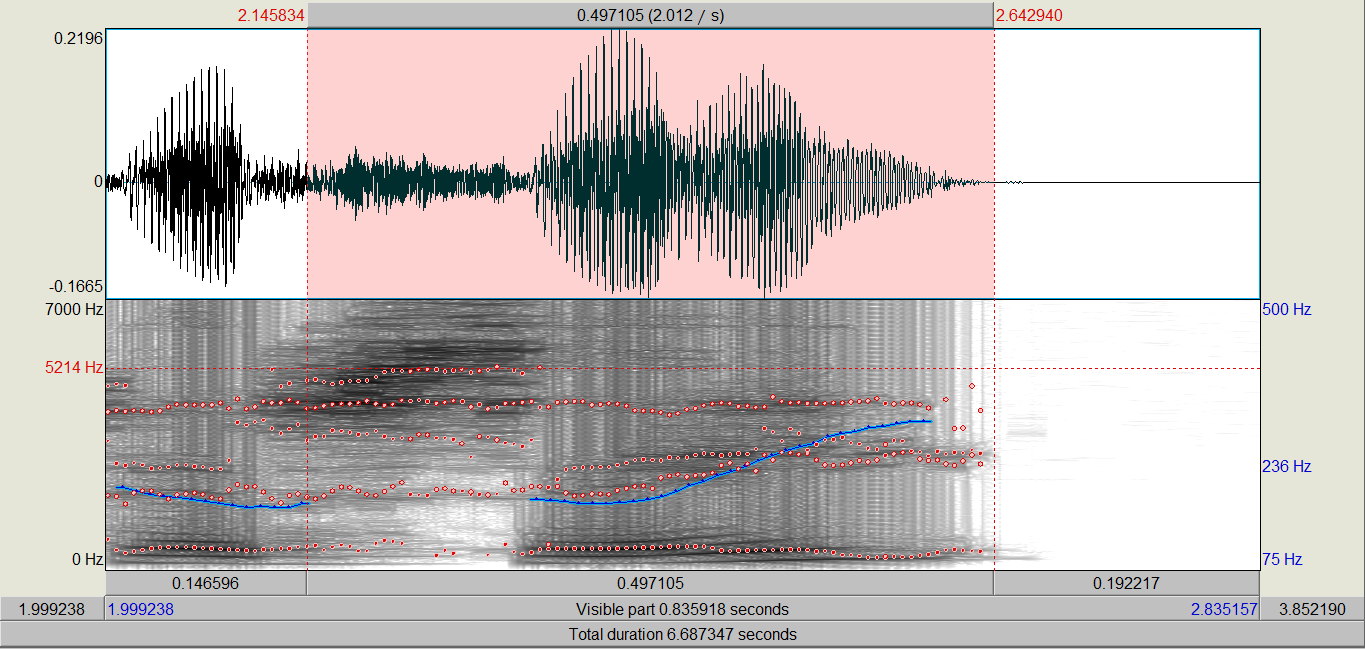
The vowel /e/ is voiced sound that has bigger amplitude in the syllable. Moreover, in the spectrogram it shows darker areas on the bottom part indicating that is a voiced sound and the vibration of the vocal cords. The last element of the syllable is the voiced alveolar tap /ɾ/, shown in Figure 20, which the top part of the spectrogram presents more friction noise than the vowel, but small wave amplitude.

(FIGURE 20)



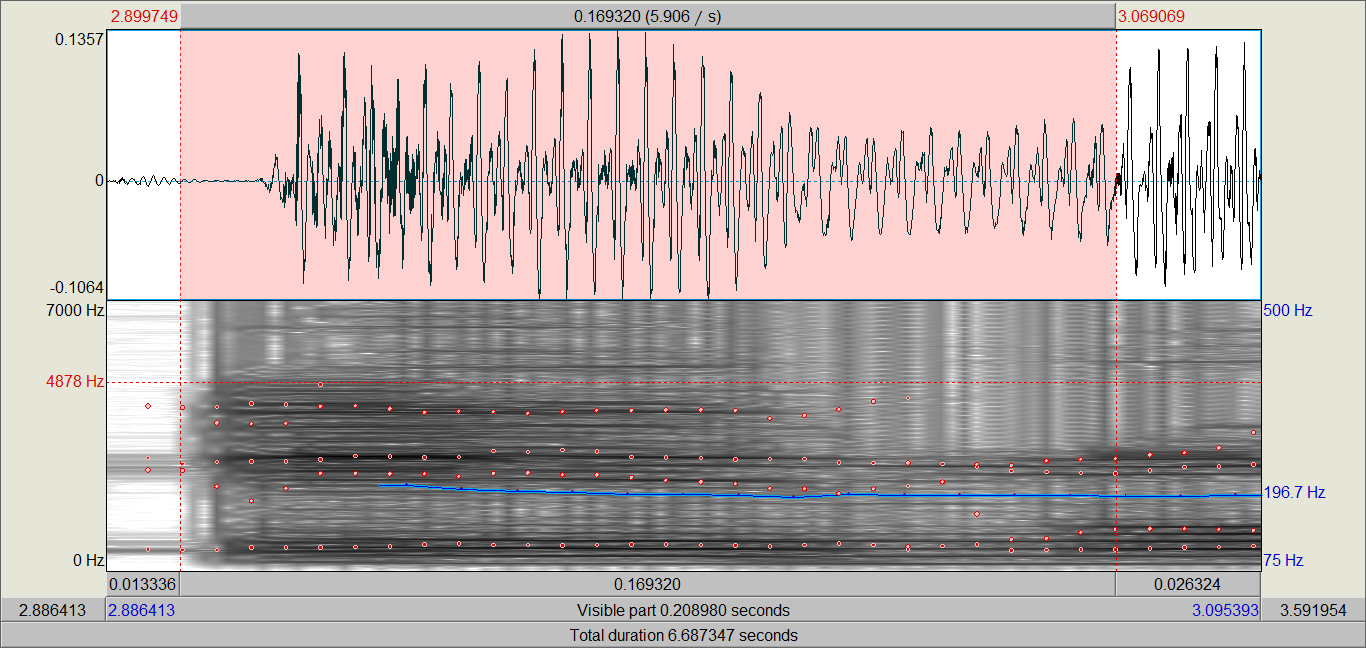
The second part of the word “jersey” is “sey” and phonetically is formed by /sei/, see in Figure 21. The first component is the alveolar fricative unvoiced consonant, /s/. The /s/ has little amplitude (as we can observe in the waveform) and also a random structure. Then we have the union of two voiced sounds, the vowels /e/ and /i/, which shows a bigger amplitude than the former sound. Furthermore, the spectrogram shows vertical lines that are the representation of glottal pulses. In addition, along the spectrogram we observe some horizontal dark lines that show the harmonics of the vocalic part /ei/.

(FIGURE 21)



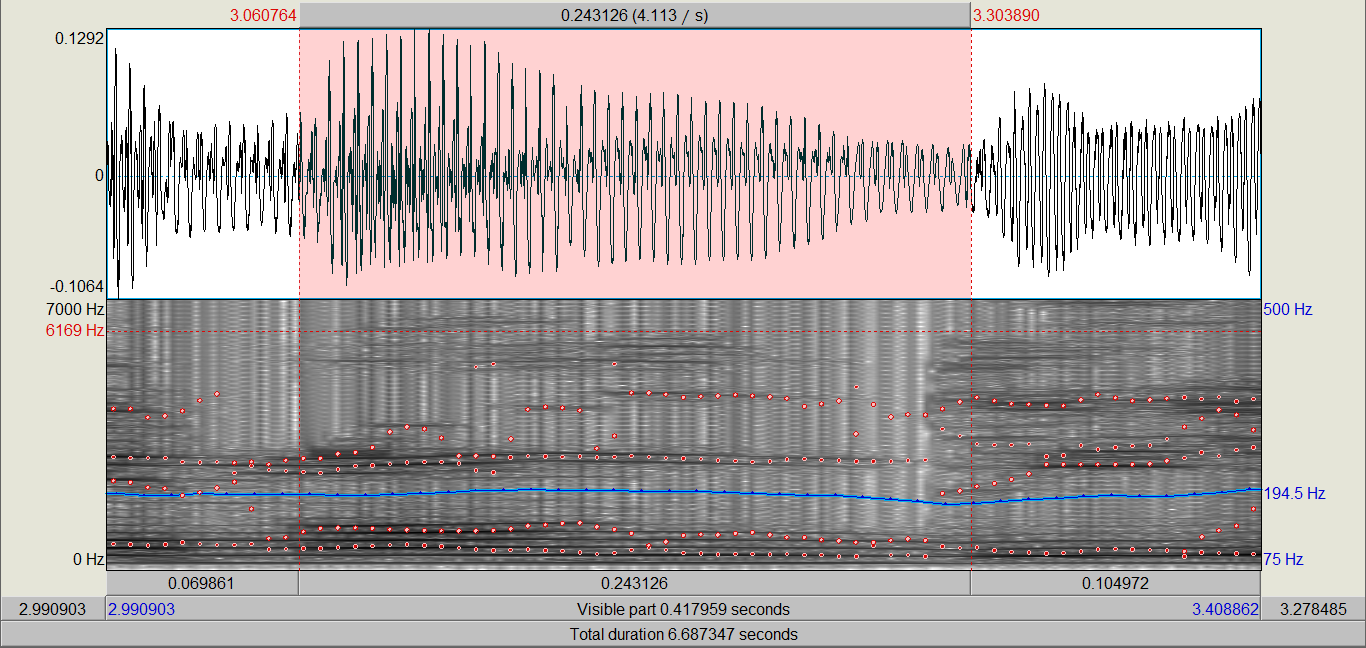
We continue with the recorded sentence with the preposition “en”, show in Figure 22. In this segment we have a vocalic voiced sound, /e/, and the alveolar nasal voiced sound, /n/. The vowel is represented in the waveform with more amplitude than the nasal consonant. Moreover, the bottom part of the spectrogram reveals high accumulation of frequency due to its dark colours. In the spectrogram part related to the /n/ sound, we could see darker areas on the bottom and just glottal pulses on the top part.

(FIGURE 22)



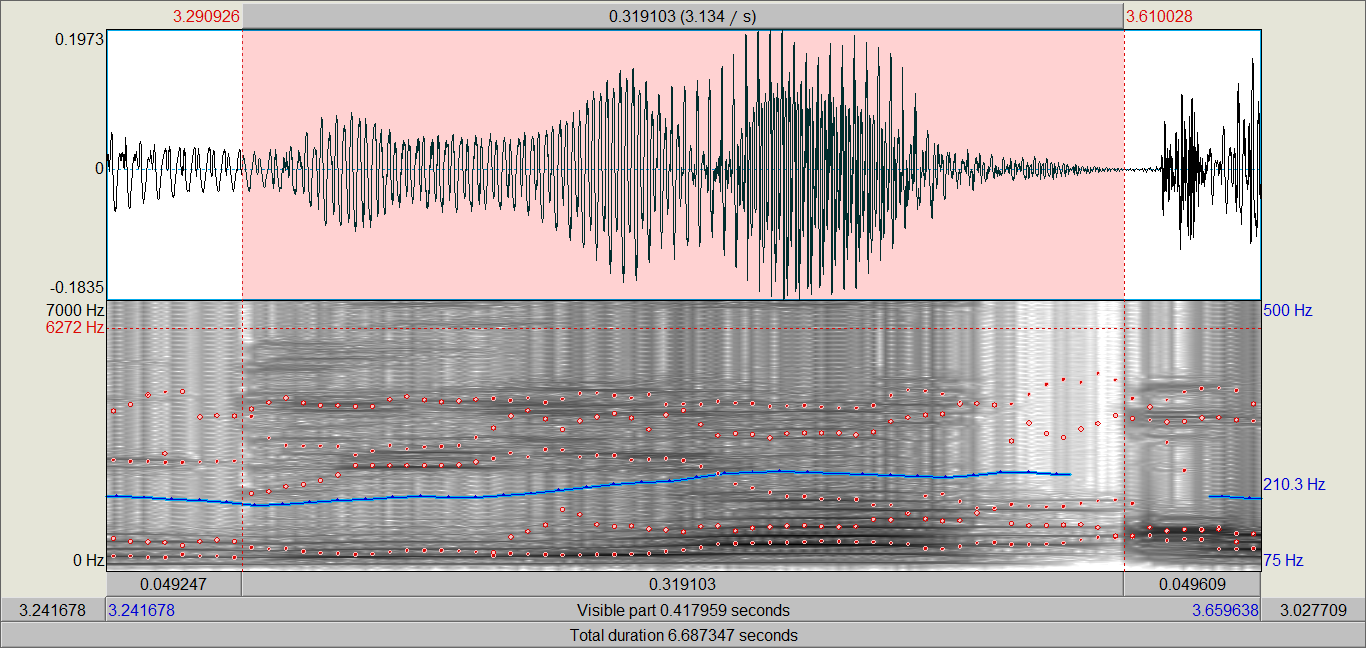
In Figure 23, we examine the article “un” of the recorded sentence. It is formed by the voiced sound, the vowel /u/, and the alveolar nasal voiced sound, the consonant /n/. In this case, the amplitude is being reduced progressively during the production of the particle. The pitch remains constant during the whole syllable. The bottom part of the spectrogram has horizontal dark lines that shows the accumulation of frequency for both the vocalic and consonant sound.

(FIGURE 23)



The following syllable presented in Figure 24, is the word “día” that is monosyllabic. The first element is the voiced dental postalveolar plosive consonant /d/, and the combination of the two voiced vowels /ia/. We can observe that in the waveform, the amplitude is higher when it begins the vowel part, and the pitch (shown in the spectrogram) also increases. The vibration of the vocal cords is also represented in black on the bottom part of the spectrogram, showing the darkest areas in the production of the vowels of this syllable.

(FIGURE 24)



4. MEASURING THE FUNDAMENTAL FREQUENCY (F0)

4.1. MEASURE THE FUNDAMENTAL FREQUENCY ON THE WAVEFORM

4.2. MEASURE THE FUNDAMENTAL FREQUENCY ON THE SPECTROGRAM

1. https://www.intechopen.com/chapters/64590 [↑](#footnote-ref-1)