TB8606

Signals and Systems

Analysis and Characterization of the Voice Signal

Author:
Falcon Luciana B.

Student ID:
107316

Date:
05/07/2025

Contents

1 Analysis of the Speech Signal

Speech can be considered as the output of a system, and its variations can be attributed to two causes: changes in the excitation or changes in the configuration of the vocal tract, that is, changes in the system itself. If the input behaves like a quasi-periodic impulse train, the output will be one of the possible vowel-like sounds (/a/, /e/, /i/, /o/, /u/, /m/, /n/, /l/). On the other hand, if the input is a white noise generator, the resulting sound will be a fricative phoneme (/s/, /f/, /sh/).

The distinction between phonemes of the same class is produced by the shape that the vocal tract takes for each one. The variation of the system's transfer function is assumed to be slow enough to consider that the speech signal is the concatenation of segments that originate as the output of an LTI (Linear Time-Invariant) system. For this reason, they appear well represented in a spectrogram. Explosive sounds (/p/, /k/, /t/), however, have a different nature and are more similar to a transient than to a stationary sound. A schematic of the voice production model is shown in Figure 1.

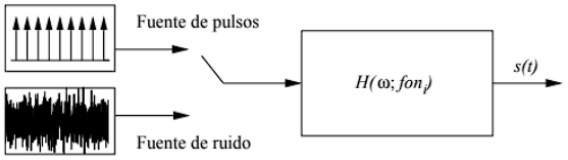


Figure 1: Voice Production Model.

1.1 Periodic and Aperiodic Signals

Initially, the periodicity of the signals produced when pronouncing the word "Picasso" both slowly and quickly was analyzed. The corresponding plots are shown below in Figures 2 and 3.

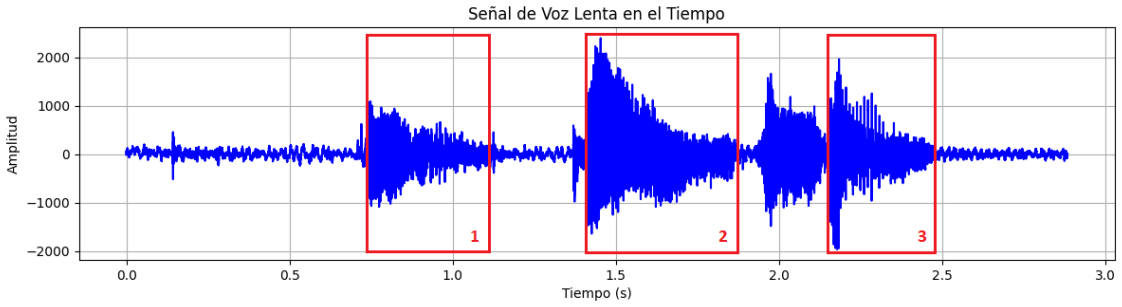


Figure 2: Voice Signal Plot *Slow Picasso*.

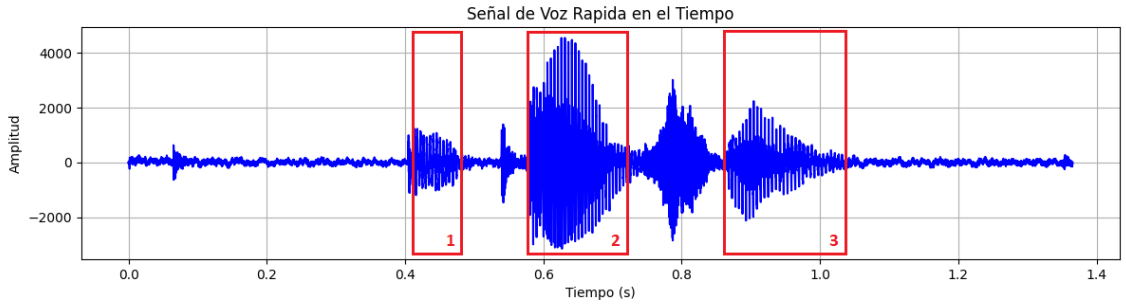


Figure 3: Voice Signal Plot *Fast Picasso*.

Points 1, 2, and 3 marked in the plots correspond to the quasi-periodic sounds generated by the vocal cords: /i/, /a/, and /o/. In contrast, the phonemes /p/, /c/, and /s/ correspond to noise.

Specifically:
/p/ and /c/ are plosive sounds: they are characterized as transient sounds with a rapid release of air.
/s/ is a fricative sound: it is generated by air turbulence, which results in a longer duration compared to plosive sounds.

As can be visually observed in the plots, fricatives like /s/ tend to extend over a longer period of time due to the continuous air friction, whereas plosive sounds occur in a fraction of a second.

When the speech signal is produced more slowly, the transitions between vowels become more prolonged, and non-vocalic sounds (plosives and fricatives) may also be extended over time. This generates more noise between vowels due to the elongation of phonemes and pauses during plosive sounds. Additionally, there are more pronounced transitions between phonemes, which makes the noise between vowels more noticeable.

1.2 Segmentation: Period and Frequency

Segments /a/ and /s/:

A segmentation of the voice signal was performed to locate the segments corresponding to the phonemes /a/ and /s/. In the following Figures 4 and 5, these segments are highlighted in the audio signal:

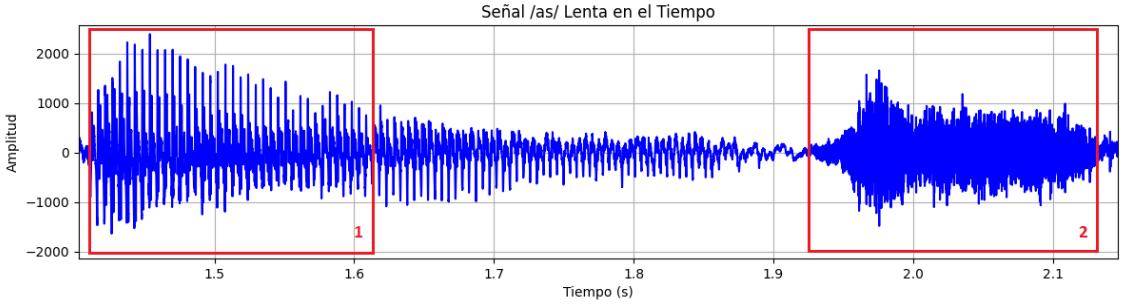


Figure 4: Plot of the voice signal for the *slow /as/* segment.

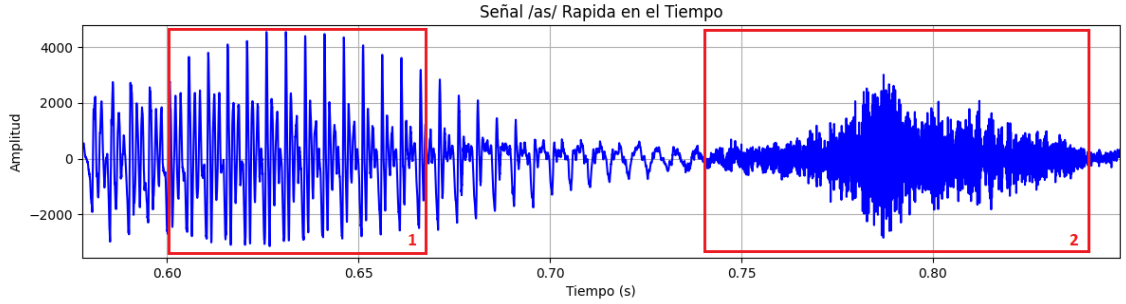


Figure 5: Plot of the voice signal for the *fast /as/* segment.

The main difference between these phonemes is that /a/ is a quasi-periodic and resonant sound, while /s/ is noisy and high-frequency due to its fricative nature.

/a/ → It is found in region 1 with higher amplitude and periodicity. It produces a quasi-periodic sound generated by the vibration of the vocal cords.
/s/ → It appears in region 2 with lower periodicity and more energy dispersion. It produces a non-periodic sound generated by the turbulence of air passing through a constriction in the vocal tract. It extends over a longer time than a plosive sound.

Segments /i/:

In the following Figures 6 and 7, the segment corresponding to the vowel /i/ is highlighted.

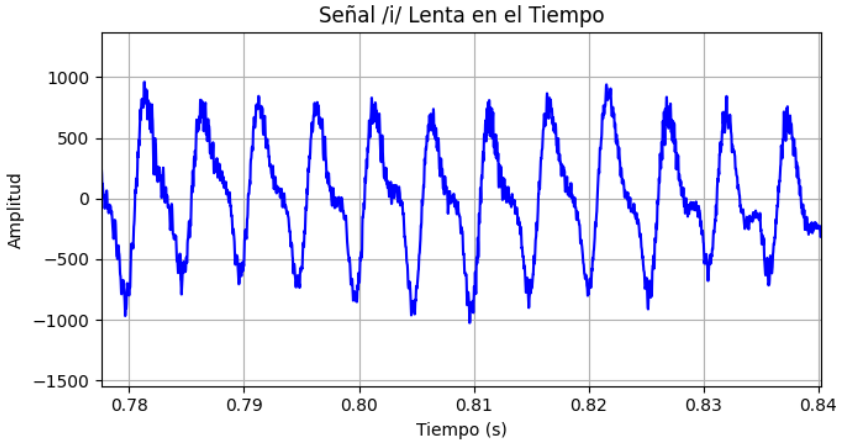


Figure 6: Plot of the vowel *slow /i/*.

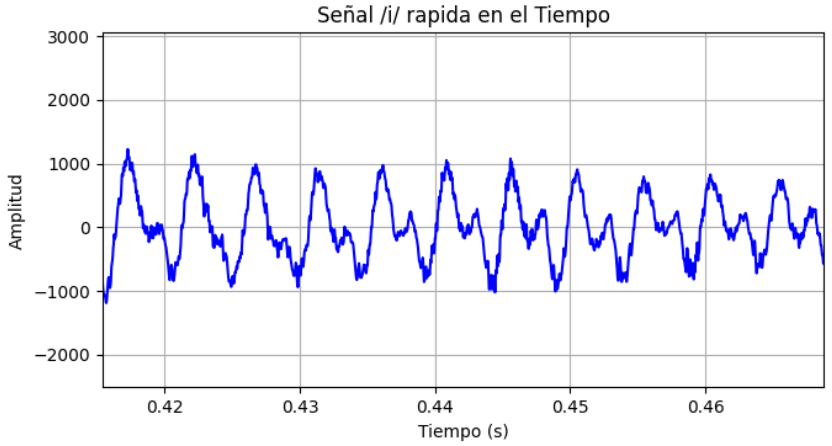


Figure 7: Plot of the vowel *fast /i/*.

Segment /a/:

In the following Figures 8 and 9, the segment corresponding to the vowel /a/ is highlighted:

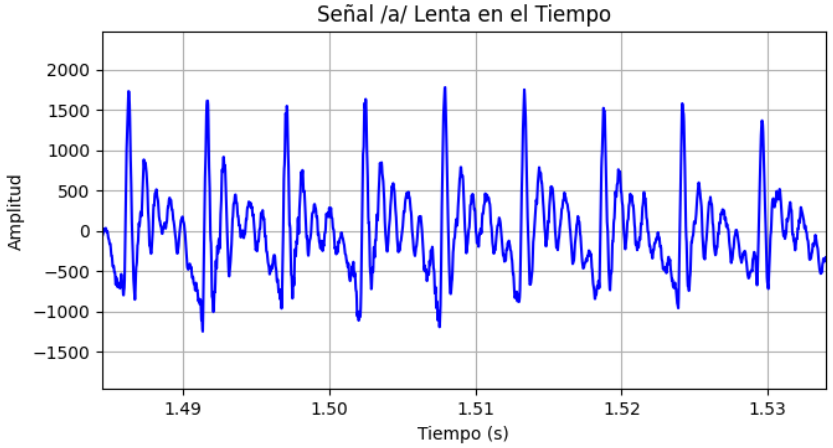


Figure 8: Plot of the vowel *slow /a/*.

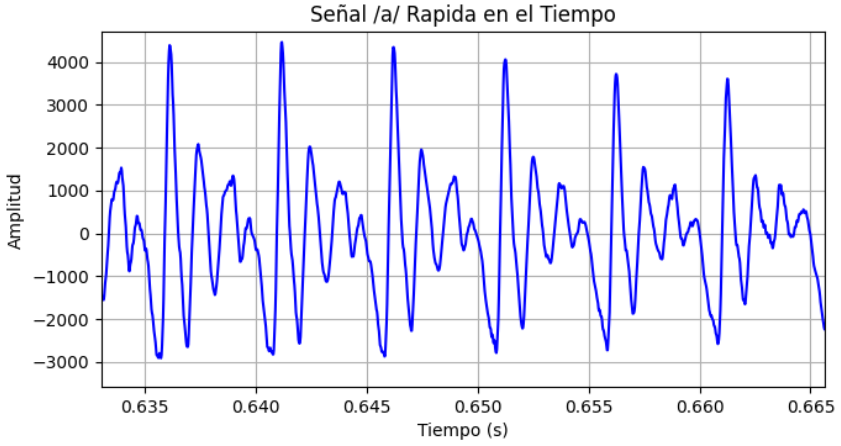


Figure 9: Plot of the vowel *fast /a/*.

Segment /o/:

In the following Figures 10 and 11, the segment corresponding to the vowel /o/ is highlighted:

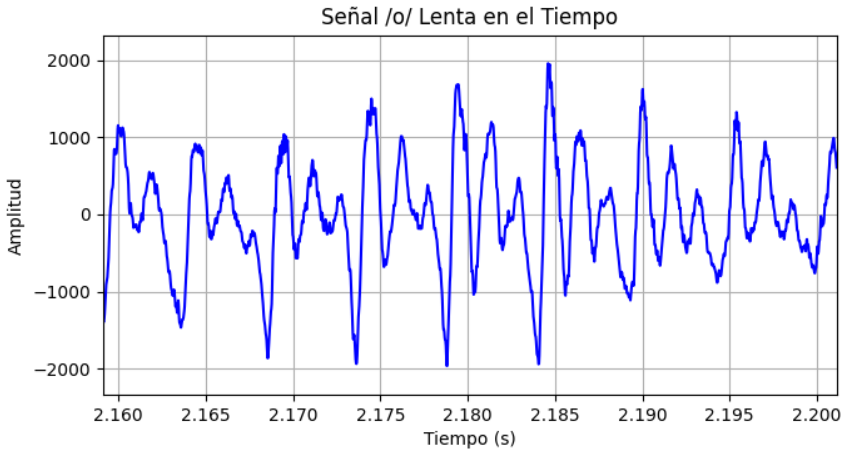


Figure 10: Plot of the vowel *slow /o/*.

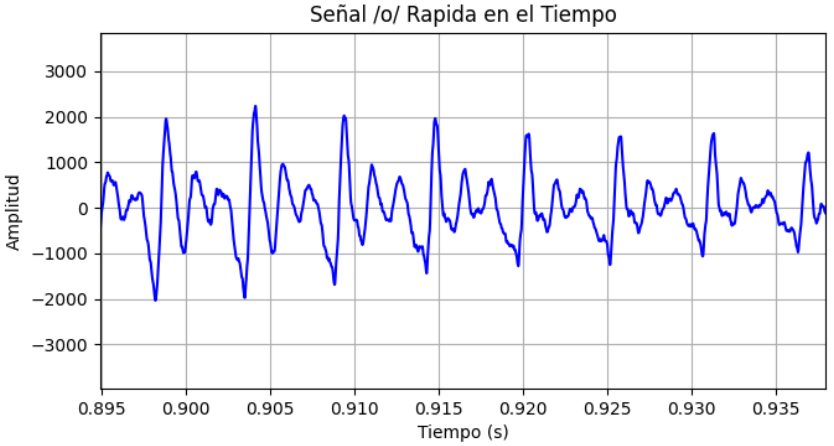


Figure 11: Plot of the vowel *fast /o/*.

The period (T) of a signal can be calculated as the total time t divided by the number of periods N within that interval. Then, the frequency f is obtained as the inverse of the calculated period.

$$f = \frac{1}{T} \tag{1}$$

Based on the plots, the periods and frequencies were estimated and are shown in Table 1:

Vocal	Períodos [ms]	Frecuencia [Hz]
/i/ lenta	5.45	183.33
/i/ rapida	5	200
/a/ lenta	5.71	175
/a/ rapida	5	200
/o/ lenta	5.71	175
/o/ rapida	5.71	175

Table 1: Estimated Periods and Frequencies of the Vowels.

The small differences in frequency (less than 0.2 Hz) were due to rounding and do not significantly affect the accuracy of the data.

1.3 FFT Analysis

Voiced sounds are produced by forcing air through the glottis or vocal cords. The tension of the vocal cords is adjusted so that they vibrate in an oscillatory manner. The periodic interruption of the subglottal airflow results in an almost periodic puff of air that excites the vocal tract. The sound produced by the larynx is called voiced or phonated. This type of sound consists of a fundamental frequency (F0) and its harmonic components produced by the vocal cords. The vocal tract modifies this excitation signal causing the formants. The term *formant* is used to indicate the center of these resonance frequencies, where the energy concentration is highest. Formants are the resonance frequencies of the spectrum, that is, the peaks of the envelope of the voice signal spectrum representing the resonance frequencies of the vocal tract. Each formant has a central frequency, amplitude, and bandwidth, and they are usually denoted F1, F2, F3, ..., starting with the lowest frequency. The frequencies at which the first formants occur are very important for voice recognition or synthesis. In Figure 12, the first three formants of a voice signal are shown, and in Figure 13, a table with the formant frequencies in Spanish is presented.

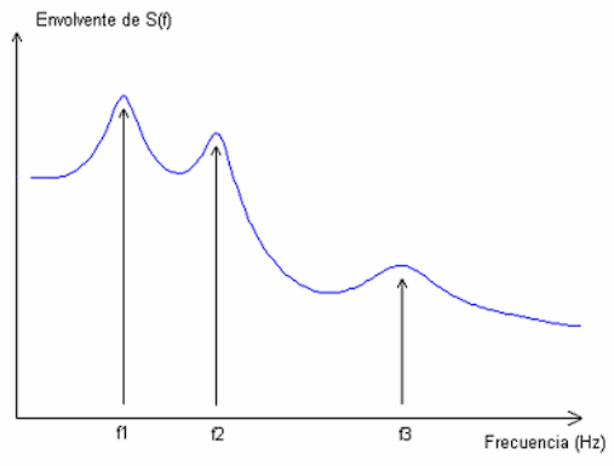


Figure 12: Envelope of the spectrum of a vowel.

Vocal	F1 (Hz)	F2 (Hz)
/i/	250 - 350	2200 - 3000
/e/	400 - 600	1900 - 2300
/a/	600 - 900	900 - 1300
/o/	400 - 600	800 - 1000
/u/	250 - 450	600 - 900

Figure 13: Formants in Spanish.

The spectra of the signal were plotted, Figures 14 and 15, in order to obtain the Fourier coefficients of the segments corresponding to the vowels present in the signal and perform the calculation using several periods of the vowel as well as using a single period for both signals.

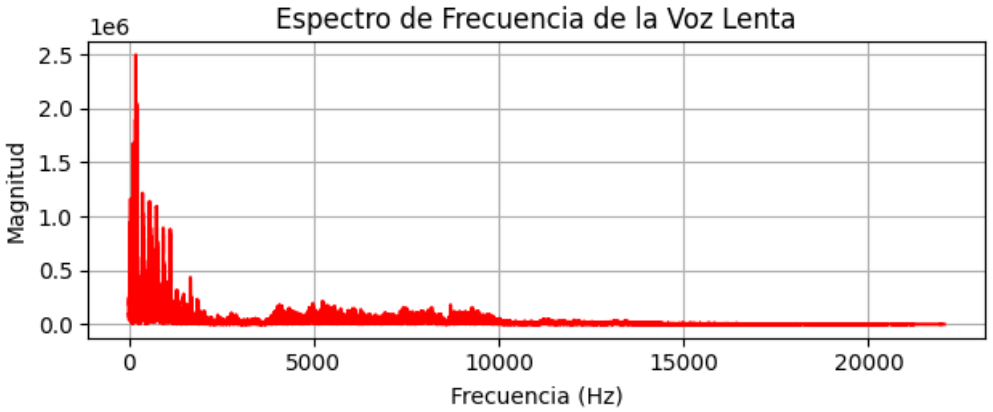


Figure 14: Plot of the spectrum of the *Slow Picasso* voice signal.

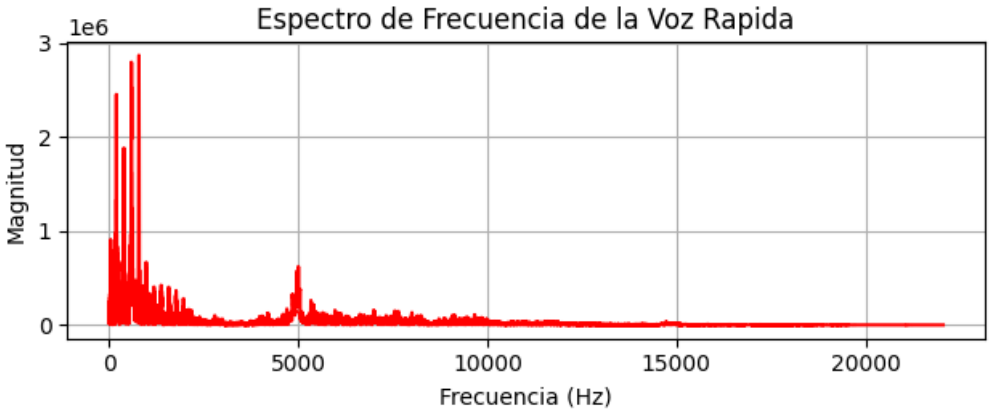


Figure 15: Plot of the spectrum of the *Fast Picasso* voice signal.

Segment /i/:

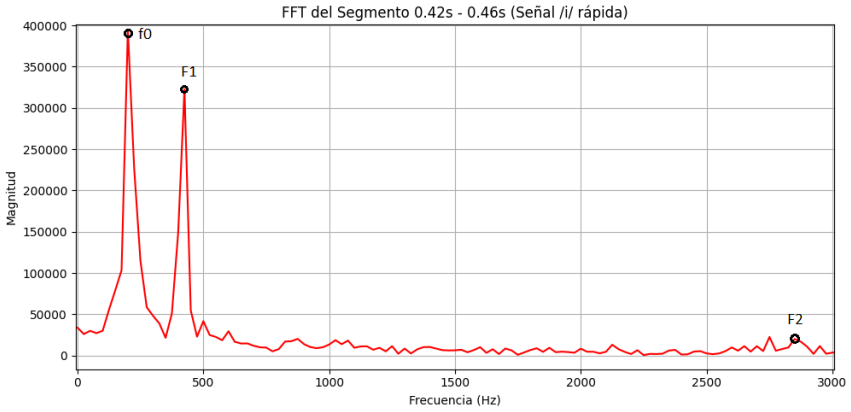


Figure 16: FFT of the *fast /i/* vowel.

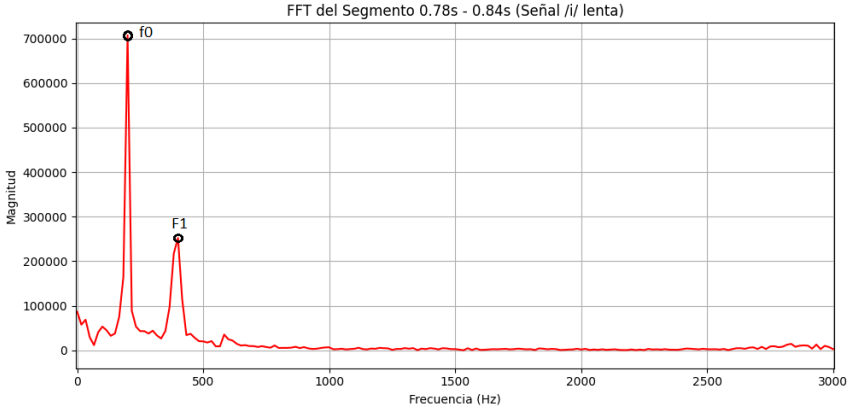


Figure 17: FFT of the *slow /i/* vowel.

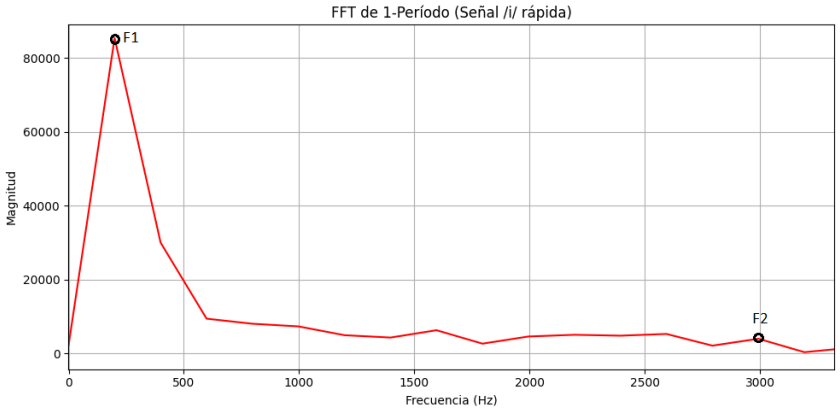


Figure 18: FFT of the *fast /i/* vowel, 1 period.

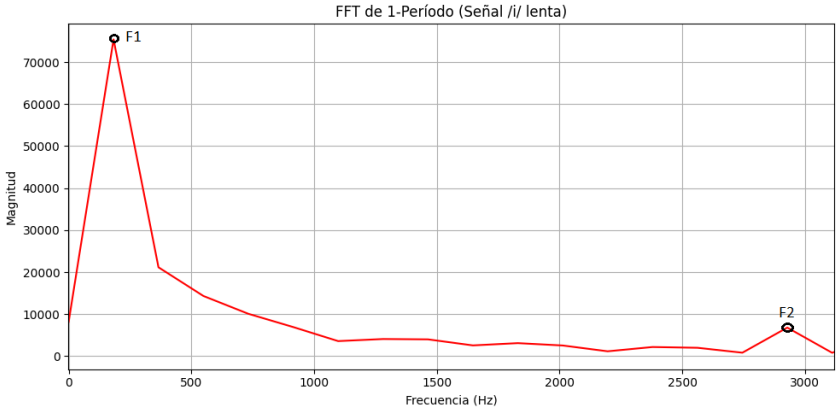


Figure 19: FFT of the *slow /i/* vowel, 1 period.

Segment /a/:

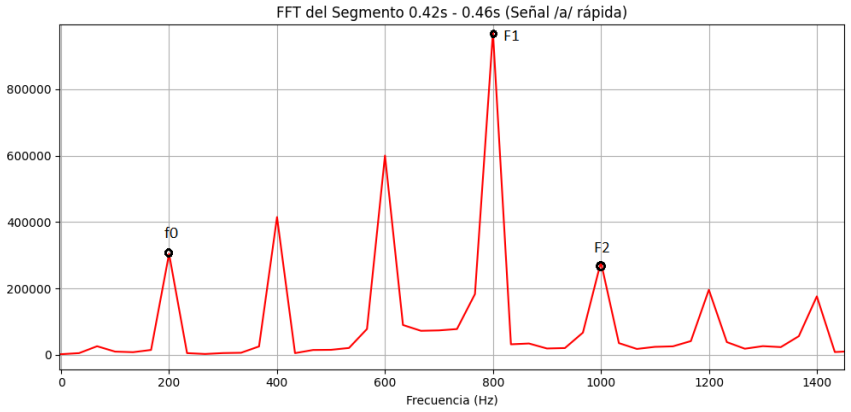


Figure 20: FFT of the *fast /a/* vowel.

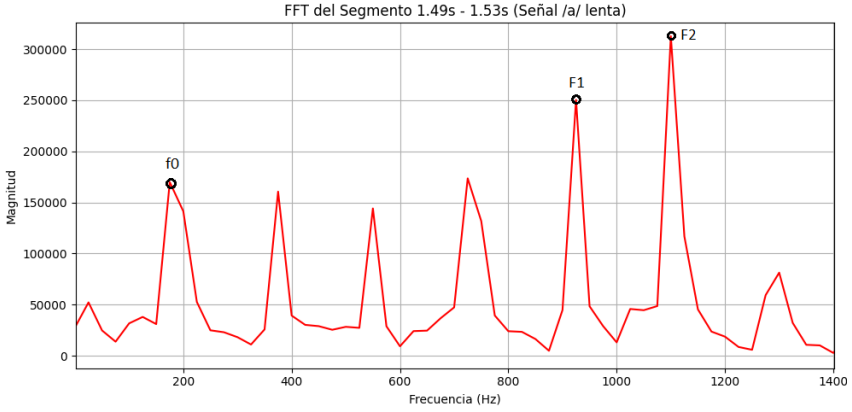


Figure 21: FFT of the *slow /a/* vowel.

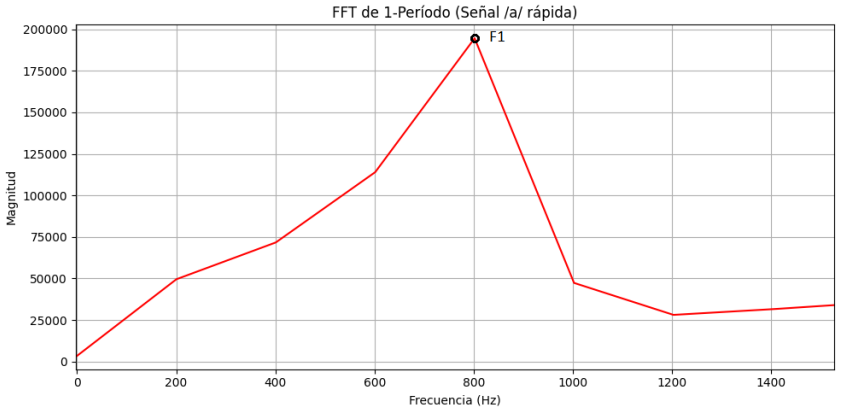


Figure 22: FFT of the *fast /a/* vowel, 1 period.

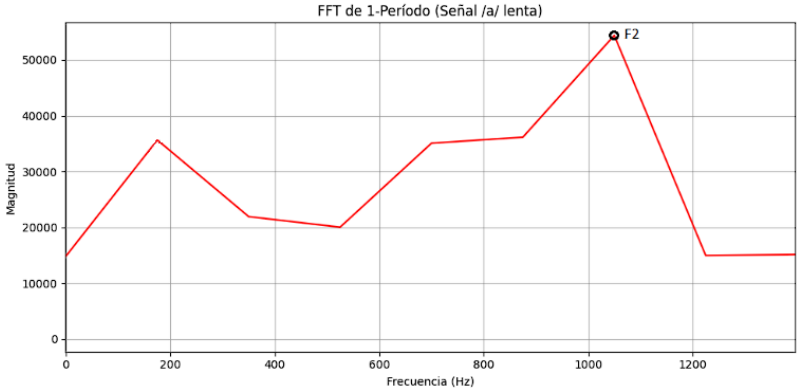


Figure 23: FFT of the *slow /a/* vowel, 1 period.

Segment /o/:

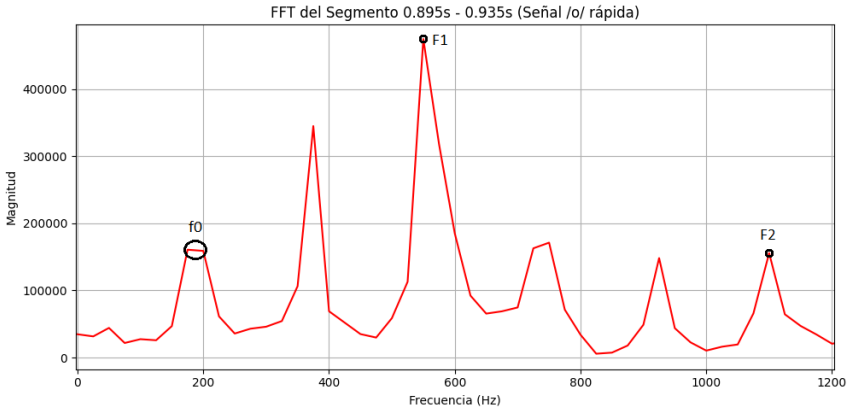


Figure 24: FFT of the *fast /o/* vowel.

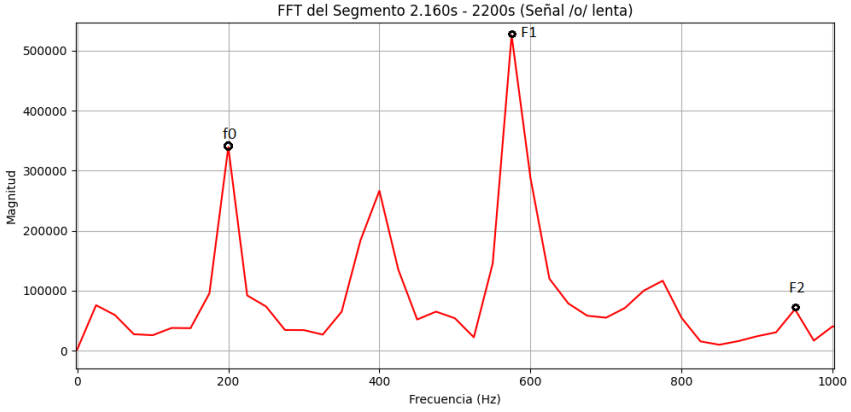


Figure 25: FFT of the *slow /o/* vowel.

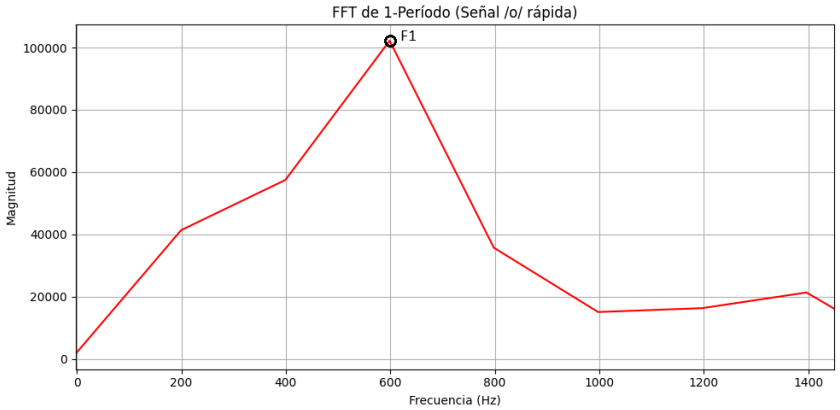


Figure 26: FFT of the *fast /o/* vowel, 1 period.

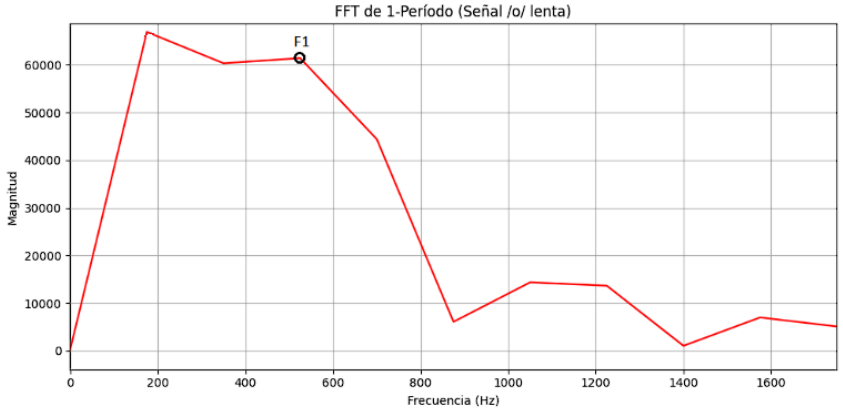


Figure 27: FFT of the *slow /o/* vowel, 1 period.

Analyzing the FFTs of the fast and slow vowels, corresponding to Figures 16 through 27, it was observed that all exhibit a fundamental frequency close to 200 Hz, which falls within the expected values calculated and shown in Table 1.

Regarding formants F1 and F2, it was observed that when analyzing the vowels using a single period, it is not possible to identify the fundamental frequency, but the formants are still distinguishable. However, when considering the graphs corresponding to multiple periods, F0 becomes visible and its value for each vowel was confirmed to be within the expected range, the band of which is shown in Figure 13. This suggests that a greater number of periods improves spectral resolution, favoring the precise identification of the formants, and thus the fundamental frequency can also be seen.

It is worth noting that the first formant of the vowel /i/ was the only one that presented a discrepancy of up to 14.29% compared to the expected value. This deviation is attributed to the sharpness of the voice used during recording, since a higher-pitched phonation tends to raise the resonance frequencies of the vocal tract, thereby altering the expected position of the formants.

2 Short-time Fourier transform

The Short-Time Fourier Transform (STFT)[1] is a Fourier transform based on the DFT. In practice, there are many applications where the properties of the signal under analysis change over time. For example, this occurs with non-stationary signals such as radar, sonar, voice, and communication signals. In these cases, calculating a single DFT for the entire signal is insufficient, in addition to the added difficulty that the signal could be extremely long and impossible to handle in practice, since digital computers usually have limited computational power and storage capacity. All of this leads us to the concept of the Short-Time Fourier Transform (STFT). The STFT of a signal $s(n)$ is defined as:

$$S(n,\omega) = \sum_{m} s(m)\, w(n-m)\, e^{-j\omega n} \tag{2}$$

where $w(n)$ is the window function. In the STFT, the one-dimensional sequence $s(n)$, a function of a discrete variable, is transformed into a two-dimensional function of the discrete variable n and the continuous frequency ω. Note that the STFT is periodic in ω with period 2π, and therefore only values in the interval $0 \leq \omega \leq 2\pi$ or any other interval of length 2π need to be considered. Taking into account the symmetry of the windows, the previous equation can be rewritten as:

$$S(n,\omega) = \sum_{m} s(m+n)\, w(m)\, e^{-j\omega n} \tag{3}$$

Thus, the STFT can be interpreted as the Fourier transform of the shifted signal $s(m+n)$ viewed through the window $w(n)$. The window has a fixed origin, and as n changes, the signal slides through the window so that for each value of n, a different portion of the signal is observed.

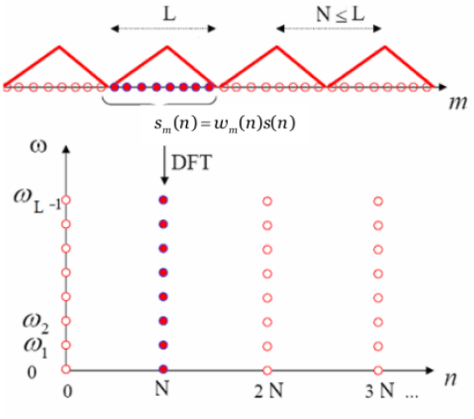


Figure 28: Spectrogram.

The spectrogram is a very useful tool for analyzing phonemes and their transitions. A spectrogram of a signal over time is a special two-dimensional representation, where the horizontal axis represents time and the vertical axis represents frequency. Usually, a grayscale scale is used to indicate the energy at each point (t, f), representing low energy with white and high energy with black. The spectrogram is obtained from the STFT. The spectrogram represents only the energy and not the phase of the STFT. The energy is calculated as:

$$\log |X(k)|^2 = \log\left(X_r^2(k) + X_i^2(k)\right) \tag{4}$$

The value from the above equation is converted to grayscale. Pixels whose values are not calculated are obtained by interpolation.

In Python, we can use `scipy.signal.spectrogram` or `scipy.signal.ShortTimeFFT`.

2.1 Narrowband and wideband spectrograms of the entire word

2.1.1 Narrowband spectrograms

At this point, the spectrogram of the word "picasso" was plotted using a 2048-sample Hamming window to apply the Short-Time Fourier Transform (STFT), resulting in a narrowband spectrogram. This type of spectrogram provides high frequency resolution and allows clear observation of the fundamental frequency (F) and its harmonics during vowel production. The thin, regularly spaced horizontal lines

visible in the plot represent the harmonics generated by the periodic vibration of the vocal cords.

The Hamming window was chosen for its favorable properties in spectral analysis. This window smooths the edges of each time segment, reducing spectral leakage effects and enabling a more accurate spectrum by minimizing sidelobes without excessively compromising resolution. Compared to a rectangular window, the Hamming window offers a better trade-off between frequency resolution and suppression of spurious components, making the harmonics stand out more sharply in the spectrogram.

In this narrowband spectrogram shown in Figure 29, generated with the aforementioned 2048-sample window, the presence of periodic components characteristic of the vowel phonemes in the word "picasso" is clearly observed. These components appear as thin, regularly spaced horizontal lines over time, especially visible in the approximate intervals between 0.4s–0.5s, 0.6s–0.75s, and 0.9s–1.1s.

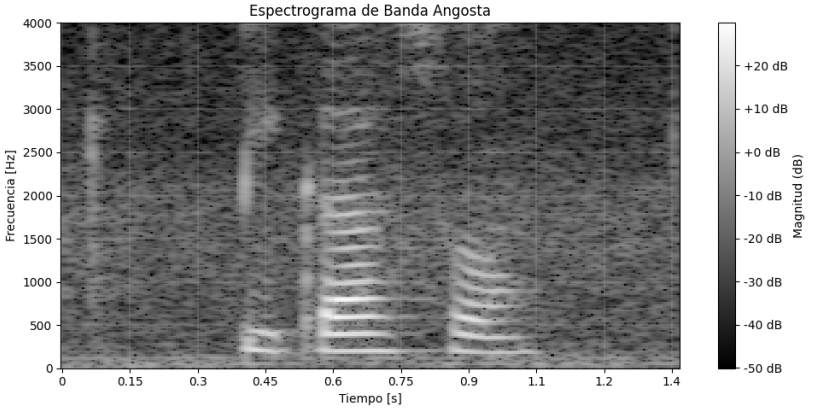


Figure 29: Narrowband spectrogram of the *Picasso fast* signal.

These lines represent the harmonics, which are integer multiples of the fundamental frequency (F) produced by the periodic vibration of the vocal cords. The clear and sustained appearance of these structures indicates the presence of quasi-periodic voiced signals, characteristic of vowels, while their absence or dispersion in other regions suggests the presence of voiceless consonants or fricatives, such as /p/ or /s/. Similarly, this description of the narrowband spectrogram can be seen in Figure 30, corresponding to the slow audio of the word "picasso".

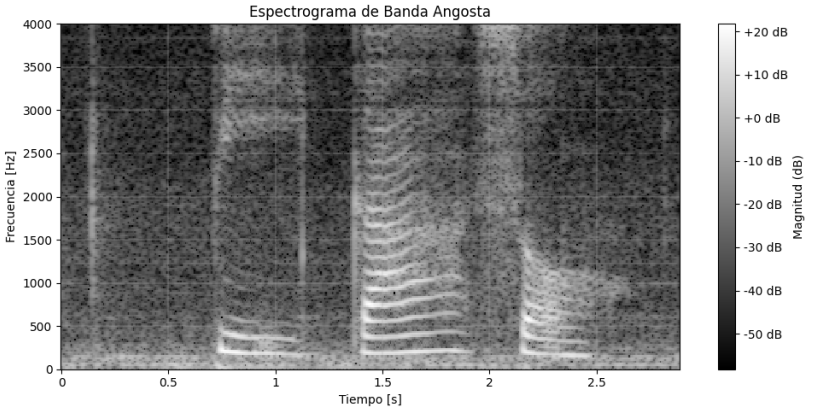


Figure 30: Narrowband spectrogram of the *Picasso slow* signal.

This type of narrowband spectrogram, with high frequency resolution, is ideal for studying the harmonic structure of the voice and allows for precise estimation of the fundamental frequency and its harmonics, which are essential for analyzing intonation and the tonal content of the speech signal.

2.1.2 Espectrogramas de banda ancha

At this point, the spectrogram of the word "picasso" was plotted using a 512-sample Hamming window to apply the Short-Time Fourier Transform (STFT), resulting in a wideband spectrogram. This type of spectrogram provides high temporal resolution, allowing clearer visualization of the vocal formants, i.e., the resonance bands of the vocal tract during the production of voiced sounds.

The choice of the Hamming window is due to its ability to reduce spectral leakage effects without excessively distorting the temporal envelope. With a size of 512 samples, this window achieves an appropriate balance between temporal resolution and spectral accuracy, enabling the observation of rapid changes in the formants that characterize vowel phonemes in fast speech.

In Figure 31, the wideband spectrogram of the fast audio of the complete word "picasso" is shown. At least three high-energy regions associated with the formants of the vowel phonemes can be distinguished:

Between 0.4s and 0.5s, the vowel [i] is observed, with a low first formant (F1) (300–400 Hz) and a high second formant (F2) (2200–2500 Hz), characteristic of this close vowel.

Between 0.6s and 0.75s, the vowel [a] appears, with F1 near 700 Hz and F2 around 1200 Hz, matching the typical pattern of an open vowel.

Between 0.85s and 1.1s, the vowel [o] is visualized, with F1 around 500 Hz and a lower F2 (900–1000 Hz), indicating a back, rounded vowel.

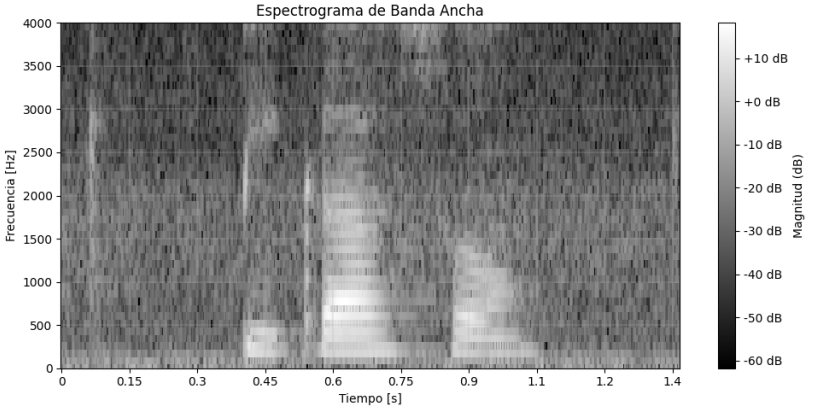


Figure 31: Wideband spectrogram of the *Picasso fast* signal.

Similarly, this description of the wideband spectrogram can be seen in Figure 32, corresponding to the slow audio of the word "picasso." These wide horizontal bands in red and yellow correspond to the formants F1, F2, and F3, and remain relatively stable throughout each vowel segment. Their presence allows for the phonetic identification of vowel sounds even under fast speech conditions, in contrast to consonants, which are represented by regions with more diffuse energy or unstructured noise.

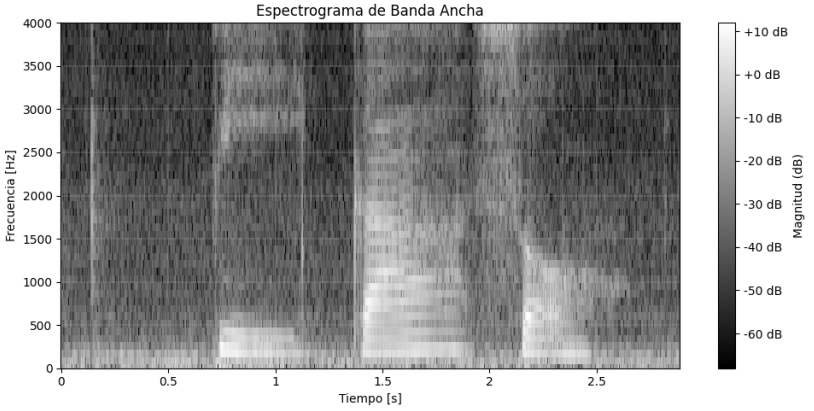


Figure 32: Wideband spectrogram of the *Picasso slow* signal.

It can be observed that this type of wideband spectrogram, with high temporal resolution and the use of a well-adjusted Hamming window, is ideal for the articulatory

and acoustic study of speech, as it allows tracking the evolution of formants in real time and accurately analyzing the differences between vowels in natural contexts.

2.2 Narrowband and Wideband Spectrograms of Vowels

The **narrowband spectrogram** involves a frequency analysis using a large window size (in this case, 1024 points). This smooths the spectrogram and reduces noise, highlighting clearer and more detailed frequency components, which is especially useful for distinguishing nearby resonances and formants.

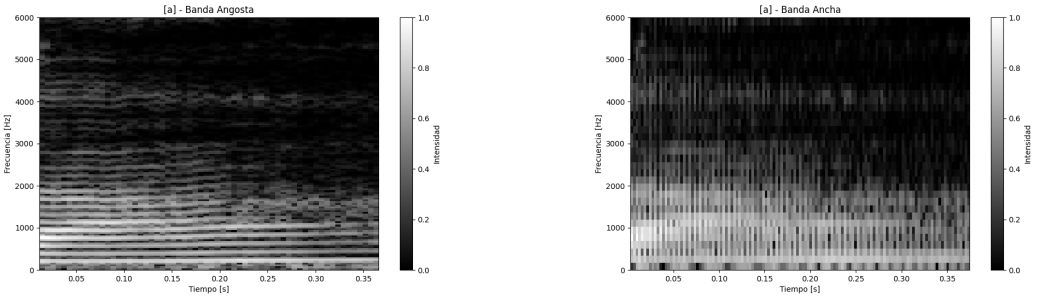
On the other hand, the **wideband spectrogram** uses a shorter window (in this case, 256 points), which provides greater temporal resolution but lower frequency resolution. As a result, the harmonic structure is less clear, and it is less precise for detecting closely spaced formants.

Using a narrowband spectrogram in vowel analysis allows for detailed observation of the evolution of formants over time, making it easier to identify transitions, variations in articulation, and vocal resonance.

2.2.1 Spectrograms of /a/

In Figure 33a, bright areas can be observed, indicating regions with higher energy at those frequencies, which typically correspond to the fundamental frequency and the first formants of the vowel. The region between 1000 and 2000 Hz also shows bright areas, although with lower intensity than in the low-frequency range, indicating the main formants in that band. Above 2000 Hz, the spectrogram appears darker, indicating lower energy in those frequencies.

On the other hand, in Figure 33b, which corresponds to the wideband spectrogram, clear vertical lines stand out. These lines reflect good temporal resolution, allowing the detection of rapid changes in the signal and the presence of harmonics. However, the formants appear less defined and more clustered, making it harder to distinguish between closely spaced frequencies. In this spectrogram, the first two formants (up to approximately 2000 Hz) can be clearly observed, but the frequency resolution is lower compared to the narrowband version.



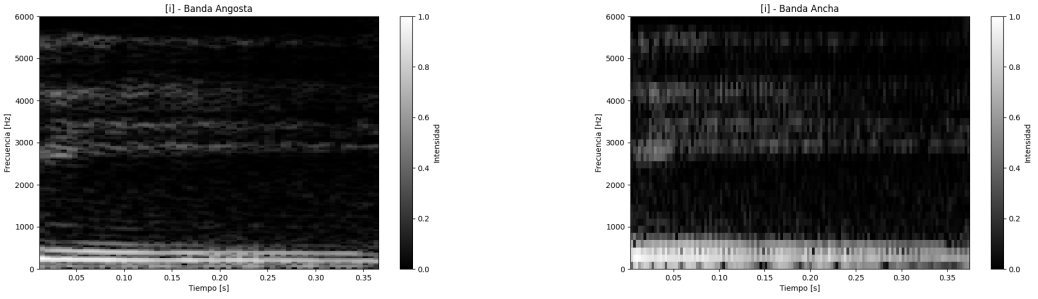
(a) Narrowband Spectrogram. (b) Wideband Spectrogram.

Figure 33: Spectrograms of the vowel /a/.

2.2.2 Spectrograms of /i/

In Figure 34a, it can be observed that in the lower part of the spectrogram, between 0 and 1000 Hz, there are bright areas indicating higher energy. This corresponds to the first formant of the vowel /i/, which, being a close and front vowel, typically appears in the range of 200 to 400 Hz. The second formant, characteristic of this vowel, appears much higher, generally between 2500 and 3500 Hz, but with lower amplitude, reflected in the thinner lines on the spectrogram. Above 3500 Hz, the energy decreases significantly, resulting in a darker and less prominent region.

In contrast, Figure 34b shows a greater spread of energy in the low frequencies, with a clearly present first formant below 900 Hz. Above 1000 Hz, a darker region is observed, which could indicate a higher density of energy at these frequencies, possibly due to the overlapping of harmonics.



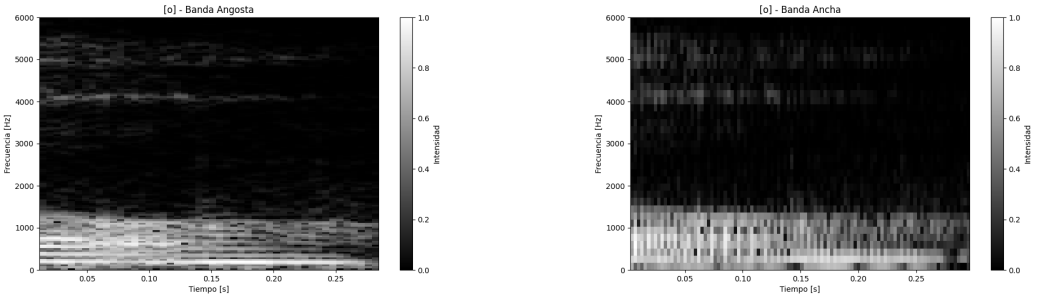
(a) Narrowband Spectrogram. (b) Wideband Spectrogram.

Figure 34: Spectrograms of the vowel /i/.

2.2.3 Spectrograms of /o/

In Figure 35a, the first formant is observed as a brighter region in the lower part of the spectrogram, generally between 400 and 700 Hz, well defined and separated from the background noise. The second formant appears between 800 and 1200 Hz, with harmonics extending up to approximately 4000 Hz. Above 1500 Hz, the harmonic lines become fainter, which is characteristic of back vowels like /o/, reflecting a lower concentration of energy in the higher frequencies.

In contrast, in Figure 35b, most of the energy is concentrated below 1000 Hz, with broader and less defined formants due to the lower frequency resolution. Higher frequencies show significantly lower energy density, evidenced by the darker regions of the spectrogram.



(a) Narrowband Spectrogram. (b) Wideband Spectrogram.

Figure 35: Spectrograms of the vowel /o/.

2.3 Pitch Modification Using TD-PSOLA

In this part of the project, we will modify the voice signal so that speech spoken by a woman sounds like it is spoken by a man, or vice versa. To do this, we will modify the fundamental frequency of the signal, which is the pitch frequency. If we have a recording of a man (85–170 Hz), we will multiply the fundamental frequency by 1.4, and if it is a woman's recording, by 0.7. The method we will use is PSOLA (Pitch Synchronous Overlap and Add) to change the pitch without changing the speech rate. There are several versions of PSOLA; we will use the one called Time Domain PSOLA (TD-PSOLA). The method consists of taking portions of the signal, multiplying them by a temporal window synchronous with the fundamental frequency, and then recombining them synchronously with a new fundamental frequency (Fig. 36). Segments can be repeated or removed to maintain the original duration of the speech.

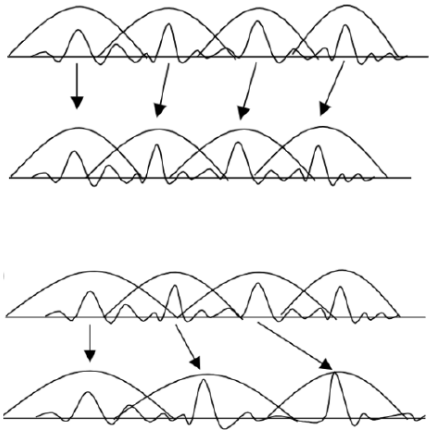


Figure 36: TD-PSOLA, Time Domain PSOLA.

The fundamental frequency modification algorithm using TD-PSOLA can be summarized as follows:

- Detect the segments of the signal that correspond to voiced sounds, which are the only ones that need to be modified.
- Locate the peaks of each cycle that make up the voiced segments.
- Apply a window centered on each peak, shift it temporally, and sum them to obtain the resulting signal.

Figure 37 shows how the TD-PSOLA algorithm allows the transformation of a voice signal, generating a version with a lower pitch from the original signal. It can be observed that the original signal, represented in blue, and the modified signal, in red, overlap on the same time axis.

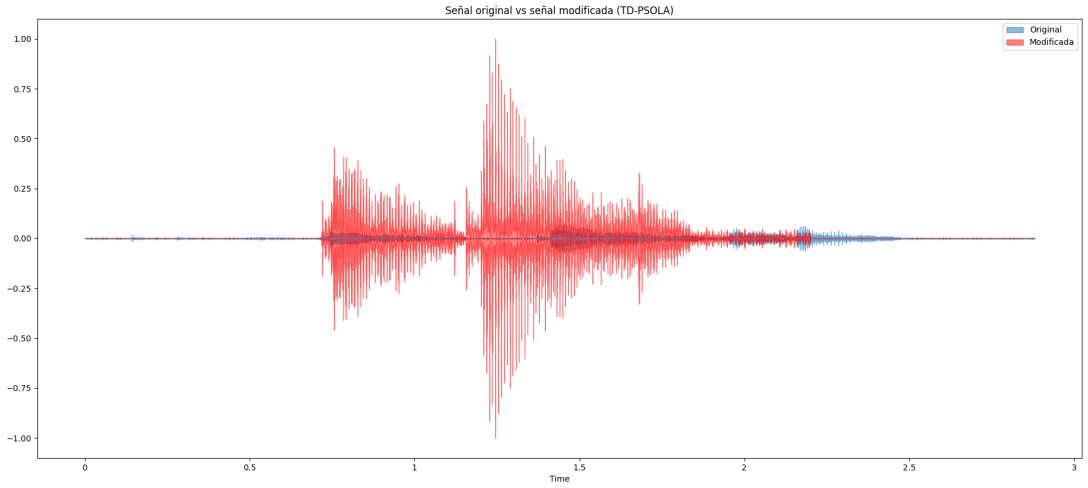


Figure 37: Plot of the slowed voice signal using TD-PSOLA.

Regarding the envelope of the signal, it is maintained, indicating that the total duration of the audio was not modified, thus fulfilling one of the objectives of the algorithm. However, the internal structure is modified; this is due to the change in the fundamental frequency, which produces an alteration in the voice pitch.

Finally, the plot allows visual confirmation that the pitch was altered without distorting the signal or changing its duration.

In this second module of the work, voice signals were analyzed using the Short-Time Fourier Transform (STFT) to obtain spectrograms that allowed observing the temporal and frequency evolution of phonemes. Narrowband and wideband spectrograms were applied, with different temporal and frequency resolutions, which enabled precise identification of both the harmonics and the formants of the vowels.

The narrowband spectrograms, with higher frequency resolution, facilitated the observation of the harmonic components of the signal, while the wideband spectrograms, with better temporal resolution, allowed analysis of the formant dynamics. Additionally, a pitch modification technique using TD-PSOLA was implemented, successfully altering the fundamental frequency of the voice without changing its duration. This analysis demonstrates the usefulness of spectro-temporal processing tools in the study of non-stationary signals such as voice, providing a detailed view of its structure and facilitating practical transformations such as pitch conversion.

3 Changes in the Speech Signal Speed

There are many applications that require changing the speed of the speech signal. A simple scheme that only interpolates or decimates the speech signal would change the speed but also alter fundamental characteristics such as the glottal frequency or the position of the formants. The signal altered in this way could even become unintelligible. Ideally, what is required is to vary the speed of the signal while preserving its frequency characteristics.

3.1 Decimation of the Slow Signal

If the sequence $x[n]$ is the result of sampling a continuous signal $x_c(t)$, and it is represented by an expanded sequence $x_p[n]$ with interleaved zeros (zero-insertion interpolation model), then the process of **decimation** can be interpreted as the removal of those zero samples, which leads to an effective reduction of the sampling rate T of $x_c(t)$ by a factor of N, such that the total sampling interval is NT.

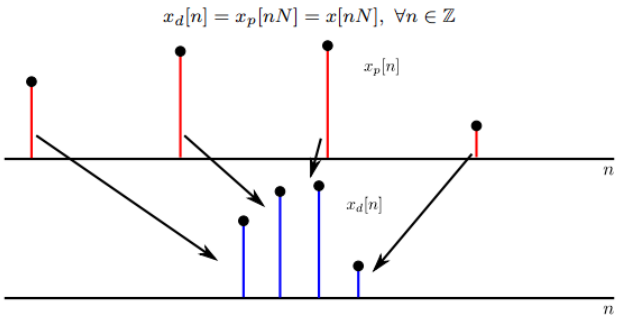


Figure 38: Decimation in the Time Domain.

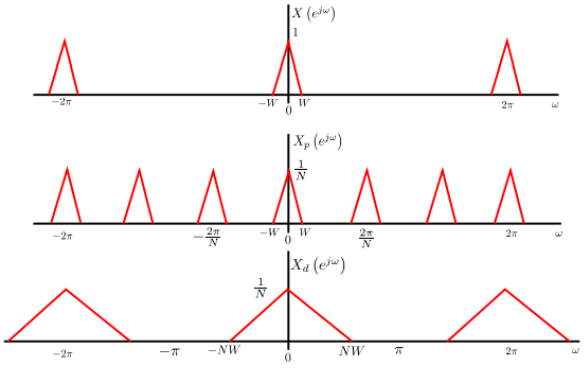


Figure 39: Decimation in the Spectral Domain.

It is clear that, to perform decimation by a factor $N > 1$ without incurring aliasing, the continuous signal $x_c(t)$ must have been sampled at a frequency $\omega_s > 2NW$, where W is the bandwidth of $x_c(t)$.

If there is a discrete signal $x[n]$ whose spectral content does not allow direct decimation without aliasing, it will be necessary to reduce its bandwidth using an ideal discrete-time low-pass filter with unity gain and cutoff frequency $\omega_c \leq \frac{\pi}{N}$, which eliminates frequency components above ω_c.

The impulse response of this ideal filter corresponds to a `sinc` function, which is infinite in duration and non-causal. Therefore, it cannot be implemented directly in a real system. To obtain a realizable version of the filter, this sinc is truncated by multiplying it by a temporal window. This procedure is known as the *windowing method*, and produces a finite impulse response filter:

$$h[n] = h_d[n] \cdot w[n] \tag{5}$$

Although a rectangular window (abrupt truncation) could be used, it generates undesired oscillations in the frequency response, a phenomenon known as the *Gibbs effect*. Instead, windows such as Hamming or Blackman smooth this truncation and reduce the ripples in the transition band, improving the out-of-band attenuation of the filter.

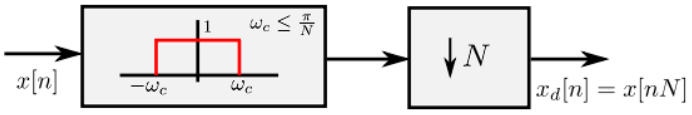


Figure 40: Decimation Process.

Following the decimation process for the *slow Picasso* signal by a factor of $N = 2$, the temporal speed was doubled without incurring aliasing due to the application of the anti-aliasing filter designed using the windowing method.

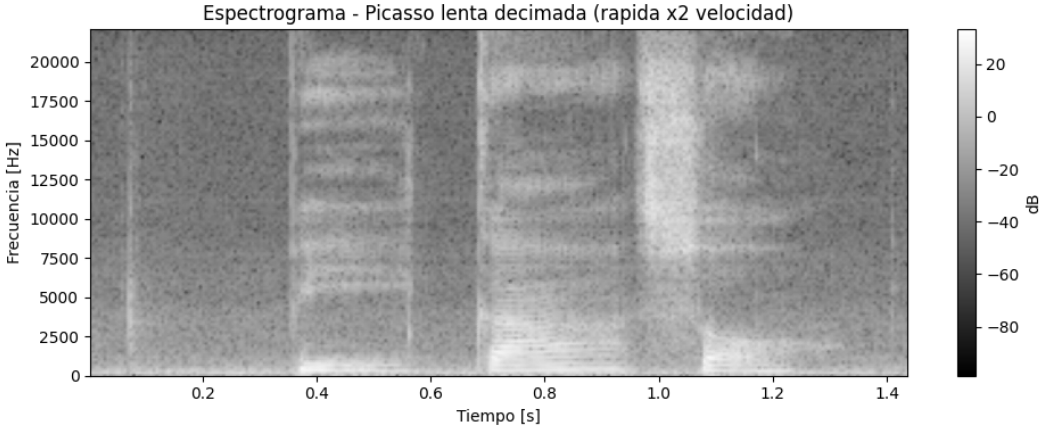


Figure 41: Spectrogram of the Decimated *Slow Picasso* Signal.

In Figure 41, it can be observed that no aliasing occurred (no crossed lines or unexpected spectral content appear), and that vocal characteristics, such as the main formants and harmonics, were preserved.

The spectral components can be compared with the spectrograms of the *original fast* signal, shown in Figure 31 of the wideband type, whose resulting spectral content presented bands shifted more to the left compared to the decimated signal.

Regarding the frequency content, it is also modified, since the original signal reached up to 4 kHz, while the decimated one reaches up to 20 kHz.

3.2 Interpolation of the Fast Signal

Now consider the possibility of *increasing* the sampling rate of a discrete signal $x[n]$.

This process is performed using an **expander**, a device that inserts $L-1$ zeros between each sample of the original signal. The output of this expander, although it has a higher sampling rate, is not yet a useful signal since it contains many artificial zeros.

$$x_e[n] = \begin{cases} x[n/L] & n \text{ es múltiplo de } L \\ 0 & \text{en otro caso} \end{cases}$$

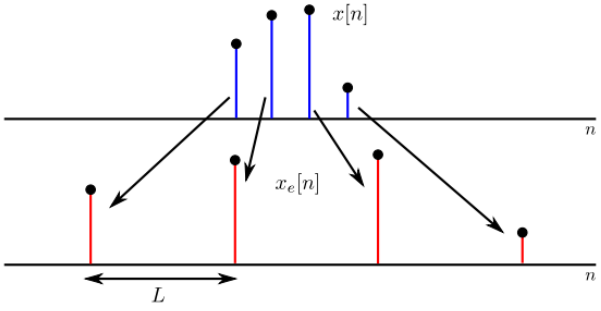


Figure 42: Interpolation in the Time Domain.

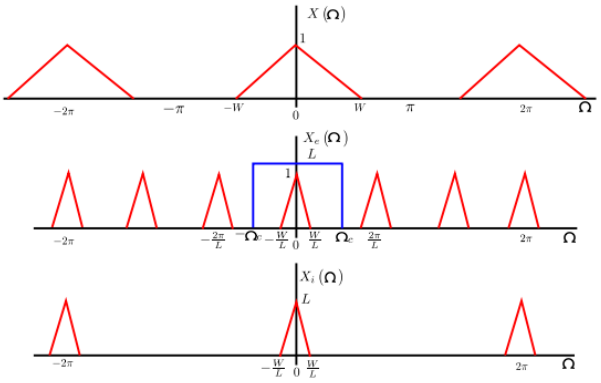


Figure 43: Interpolation in the Spectral Domain.

The Fourier transform of this expanded signal is equal to the transform of the original signal, **compressed in frequency by a factor of** L, and periodically replicated over the interval $[-\pi, \pi]$. To recover a useful interpolated signal, a subsequent low-pass filter is required to eliminate the replicas.

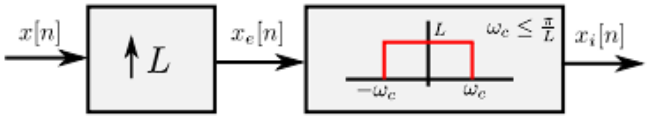


Figure 44: Interpolation Structure.

The interpolation method was applied to modify the temporal duration of the *fast*

Picasso speech signal in such a way as to reduce its speed by half and obtain a longer signal in time.

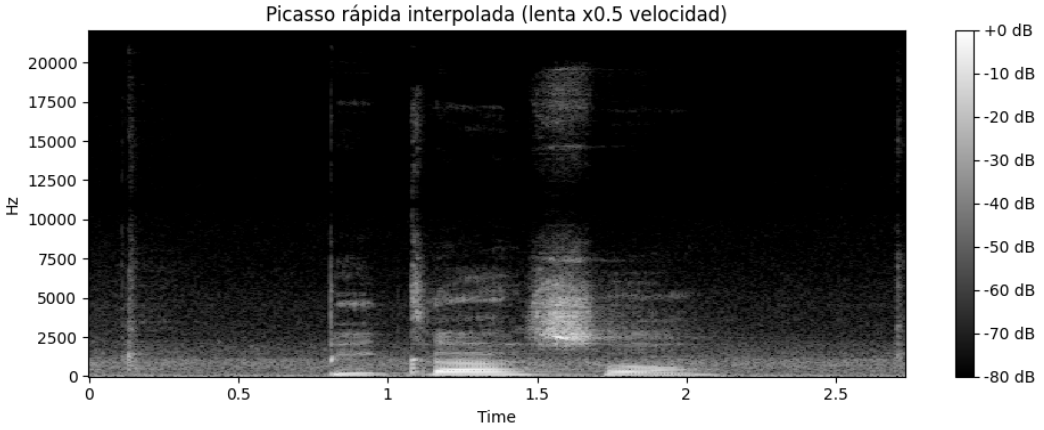


Figure 45: Spectrogram of the Interpolated *Fast Picasso* Signal.

In Figure 45, the spectral structure appears stretched along the time axis, as expected when doubling the duration.

The spectral components can be compared with the spectrograms of the *original slow* signal, shown in Figure 32 of the wideband type, whose formants, harmonics, and bands are shifted more to the left compared to the spectrogram corresponding to the interpolated fast signal. The frequencies reach up to 4 kHz, while the interpolated one increases to 20 kHz.

3.3 The Phase Vocoder Method

The decimation and interpolation methods change the temporal duration of a speech signal but also shift the formants and harmonics. To reduce this effect, several methods exist. One of them is PSOLA, applied in the second part of the project (which works in the time domain), and another is the **phase vocoder**, which works with the STFT (Short-Time Fourier Transform):

1. STFT: The STFT is applied to the signal.
2. Interpolation: An interpolation or resampling of the "rows" of the STFT matrix is performed, adding or removing columns to adjust to the desired duration.
3. Time Synthesis: The inverse STFT (iSTFT) is performed. This can consist, for example, of performing the inverse DFT column-wise, compensating for the

windowing effect used in the STFT, and combining the results into a single signal.

3.3.1 Speed Increase of the Slow Signal Using STFT

The process of doubling the speed of the signal by modifying its Short-Time Fourier Transform (STFT) involves working in the frequency domain. Specifically, one out of every two columns of the spectrogram is removed, which is equivalent to reducing the temporal resolution of the STFT representation.

This operation can be interpreted as a form of subsampling along the temporal axis of the STFT. By removing columns, only half of the temporal segments are preserved, generating a compressed version of the original signal. When reconstructing the signal using the inverse STFT (iSTFT), a shorter-duration signal is obtained, that is, with a higher playback speed.

This procedure is not based on directly removing samples in the time domain, as occurs in traditional decimation, but modifies the temporal structure of the signal through its spectro-temporal representation. Although it does not guarantee perfect preservation of the original content, it can be useful in contexts where an efficient perceptual transformation is desired.

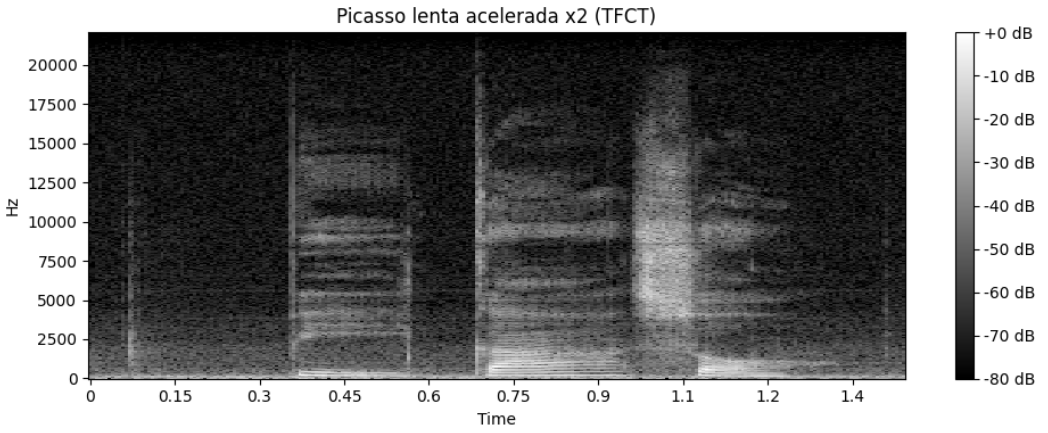


Figure 46: Spectrogram of the *Slow Picasso* Signal Accelerated Using STFT.

The spectrogram in Figure 46 is the result of modifying the Short-Time Fourier Transform (STFT) by removing one out of every two columns of the *slow Picasso* signal and subsequently recovering the time-domain signal using the inverse STFT (iSTFT). It corresponds closely to Figure 30, the narrowband spectrogram with higher resolution, as the defined harmonic lines can be seen in the plot, except for

the frequency content, which reaches up to 20 kHz, whereas the original narrowband spectrogram reaches 4 kHz.

3.3.2 Speed Reduction of the Fast Signal Using STFT

To halve the speed of the fast Picasso signal, its Short-Time Fourier Transform (STFT) was modified by interpolating columns. This process consists of estimating intermediate columns between consecutive pairs of the original STFT, calculating both the magnitude and spectral phase from their neighboring columns.

By doubling the number of columns in the time-frequency matrix, the resulting spectrogram represents a signal that varies more slowly over time, thus achieving a slowed-down version.

Finally, the inverse STFT (iSTFT) was applied to reconstruct the signal in the time domain. The result is a signal with approximately twice the duration of the original, preserving its main spectral components.

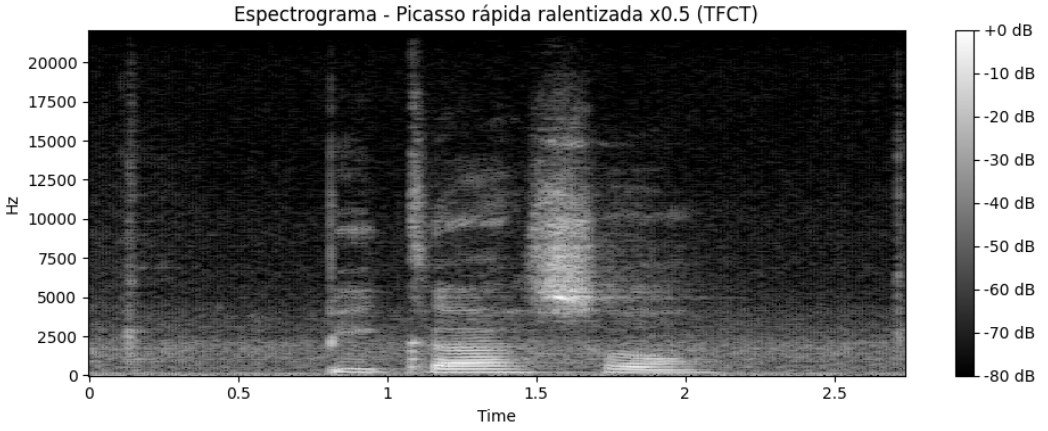


Figure 47: Spectrogram of the *Fast Picasso* Signal Slowed Down Using STFT.

The spectrogram in Figure 47, obtained after modifying its Short-Time Fourier Transform (STFT) by interpolation and subsequent reconstruction via the inverse STFT (iSTFT), matches Figure 30, corresponding to the narrowband spectrogram with higher resolution.

As was the case with the spectrogram of the slow signal accelerated by STFT, we observe that the frequencies are higher than those of the original signal.

4 Conclusions

- The first section analyzed the periodicity of the speech signal of the word *Picasso* in its slow and fast versions, along with its analysis using FFT. It was verified that the vowel frequencies match the theoretical tabulated formants.

  Although the speaking rate affects the time domain (duration, spacing), the spectral content *especially of the vowels* remains stable. This highlights the importance of formants as robust features for speech recognition and demonstrates the value of combined time-frequency analysis.

- The second module used spectrograms based on the STFT to observe the temporal and spectral evolution of phonemes, applying narrowband and wideband variants to identify harmonics and formants.

  Narrowband spectrograms facilitated the visualization of harmonic components, while wideband spectrograms allowed analysis of the temporal dynamics of formants. Additionally, TD-PSOLA was implemented to modify pitch without altering duration, demonstrating the usefulness of these tools for voice signal analysis and transformation.

- The last section compared four methods to modify the speed of a speech signal: decimation and interpolation in the time domain, and modifications via STFT. While both time-domain and spectral-domain techniques allow similar speed modifications, it was observed that signals processed through the Short-Time Fourier Transform (STFT) present higher spectral resolution compared to those modified by decimation or interpolation. The latter alter the pitch during acceleration or slowing, causing the signals to sound higher or lower in pitch.

  In contrast, signals modified using vocoder-type techniques, such as the phase vocoder, maintain the original spectral content, thus preserving the timbre and acoustic characteristics of the voice. This demonstrates a clear advantage of the spectro-temporal approach for speed manipulation, as it preserves the perceptual quality of the signal without introducing audible artifacts or pitch distortions. Finally, it is noted that when modifying signal duration via spectral techniques, the resulting spectrum shows a wider frequency range.