TB8606

Señales y Sistemas

Análisis y caracterización de la señal de voz

Autora:
Falcon Luciana B.

Padrón:
107316

Fecha:
05/07/2025

Índice

1. Análisis de la señal de habla

La señal de habla como la salida de un sistema, podemos atribuir dichas variaciones a dos causas: cambios en la excitación o cambios en la configuración del tracto vocal, es decir en el sistema. Si la entrada se comporta como un tren de impulsos cuasi-periódicos, la salida será uno de los posibles sonidos vocálicos (/a/,/e/,/i/,/o/,/u/,/m/,/n/, /l/). Si la entrada en cambio es un generador de ruido blanco, el sonido obtenido será un fonema fricativo (/s/,/f/,/sh/).

La distinción entre los fonemas de la misma clase se produce por la forma que va tomando el tracto vocal para cada uno de ellos. La variación de la transferencia del sistema se supone que es suficientemente lenta como para considerar que la señal de habla es la concatenación de porciones de señales que se originan como salida de un sistema LTI. Por esto aparecerán bien representados en un espectrograma. Los sonidos explosivos (/p/,/k/,/t/) en cambio tienen una naturaleza distinta, y son más parecidos a un transitorio que a un sonido estacionario. Un esquema del modelo de producción de la voz se muestra en la Figura 1.

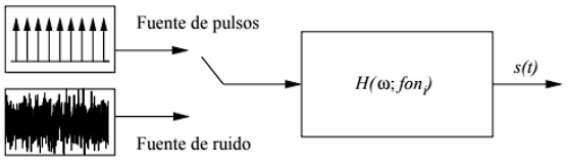


Figura 1: Modelo de reproducción de la voz.

1.1. Señales Periódicas y No Periódicas

Inicialmente, se analizó la periodicidad de las señales producidas al pronunciar la palabra 'Picasso' en forma rápida y lenta. A continuación, en las Figuras 2 y 3 se presentan los gráficos correspondientes.

Los puntos 1, 2 y 3 marcados en los gráficos corresponden a los sonidos cuasi-periódicos generados por las cuerdas vocales /i/, /a/ y /o/. Mientras que los fonemas /p/, /c/ y /s/ corresponden a ruido.

En particular: /p/,/c/ son explosivos: se caracterizan por ser sonidos transitorios con una liberación rápida de aire. /s/ es fricativo: se genera por turbulencia del aire, lo que hace que su duración sea mayor en comparación con los sonidos explosivos. Es normal como visualmente se ve en los gráficos que los fricativos como /s/ se extienden más en el tiempo debido a la fricción continua del aire, en cambio con los sonidos explosivos ocurren en una fracción de segundo.

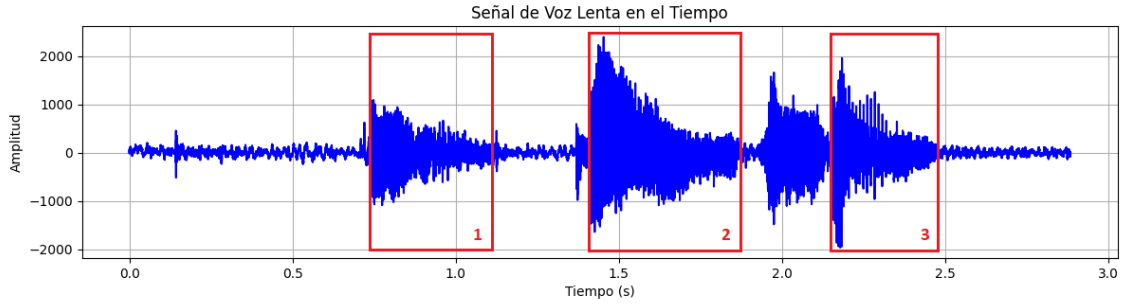


Figura 2: Gráfico de la señal de voz *Picasso lenta*.

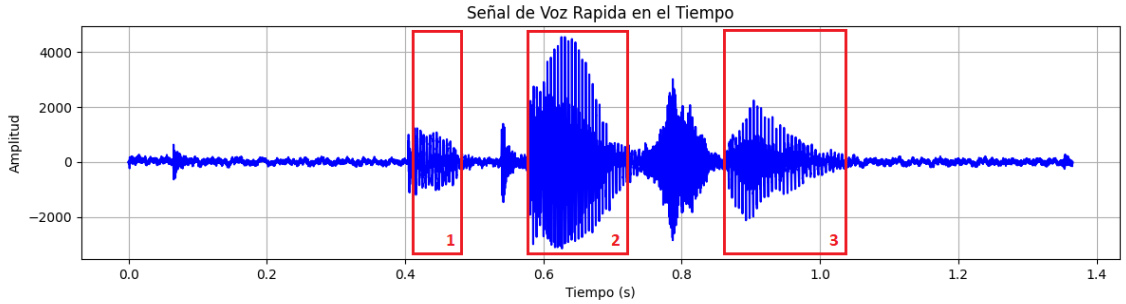


Figura 3: Gráfico de la señal de voz *Picasso rápida*.

Comparativamente cuando la señal es más lenta, las transiciones entre vocales se alargan y los sonidos no vocálicos (explosivos y fricativos) pueden prolongarse más en el tiempo. Esto genera más ruido entre vocales debido a la prolongación de fonemas y pausas en sonidos explosivos. También se producen transiciones más marcadas entre fonemas, lo que hace que se note más el ruido entre vocales.

1.2. Segmentación: Período y Frecuencia

Segmentos /a/ y /s/:

Se realizó una segmentación de la señal de voz para localizar los segmentos correspondientes a los fonemas /a/ y /s/. En las siguientes Figuras 4 y 5, se destacan estos segmentos en la señal de audio:

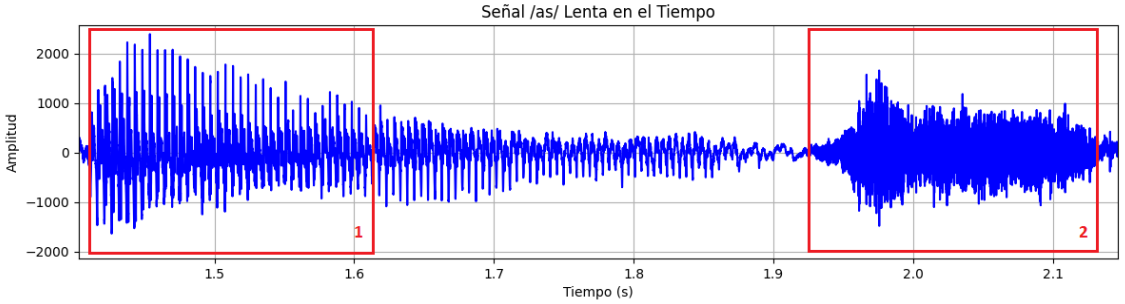


Figura 4: Gráfico de la señal de voz */as/ lenta*.

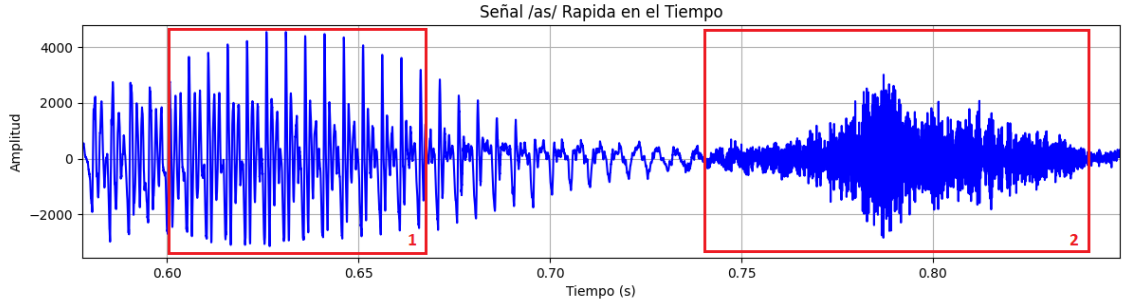


Figura 5: Gráfico de la señal de voz */as/ rápida*.

La principal diferencia entre estos fonemas es que /a/ es un sonido cuasi-periódico y resonante, mientras que /s/ es ruidoso y de alta frecuencia debido a su naturaleza fricativa.

/a/ → Se encuentra en una región 1 con mayor amplitud y periodicidad. Genera un sonido cuasi-periódico, generado por la vibración de las cuerdas vocales.
/s/ → Aparece en una zona 2 con menor periodicidad y más dispersión de energía. Genera un sonido no periódico, generado por la turbulencia del aire al pasar por una constricción en el tracto vocal. Se extiende más en el tiempo que un sonido explosivo.

Segmentos /i/:

En los siguientes gráficos 6 y 7 se destaca el segmento de la vocal /i/.

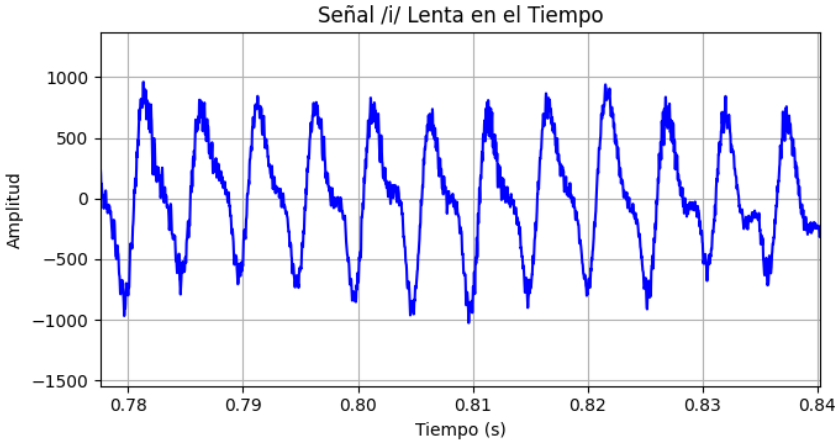


Figura 6: Gráfico de la vocal */i/ lenta*.

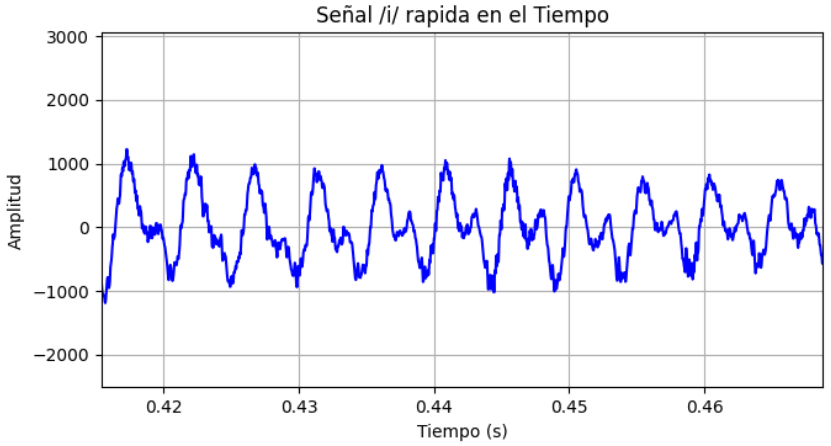


Figura 7: Gráfico de la vocal */i/ rápida*.

Segmento /a/:

En los siguientes gráficos 8 y 9 se destaca el segmento de la vocal /a/:

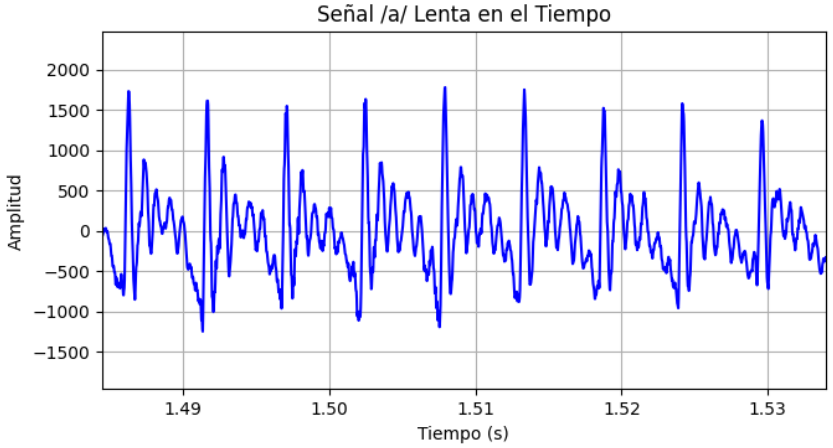


Figura 8: Gráfico de la vocal */a/ lenta*.

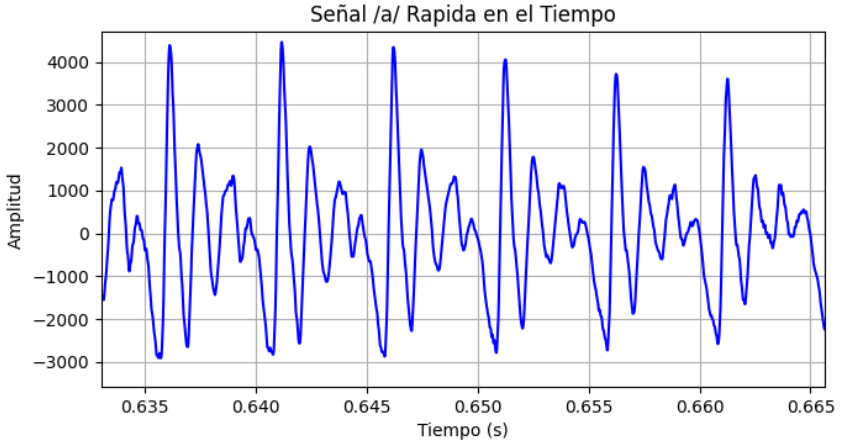


Figura 9: Gráfico de la vocal */a/ rápida*.

Segmento /o/:

En los siguientes gráficos 10 y 11 se destaca el segmento de la vocal /o/:

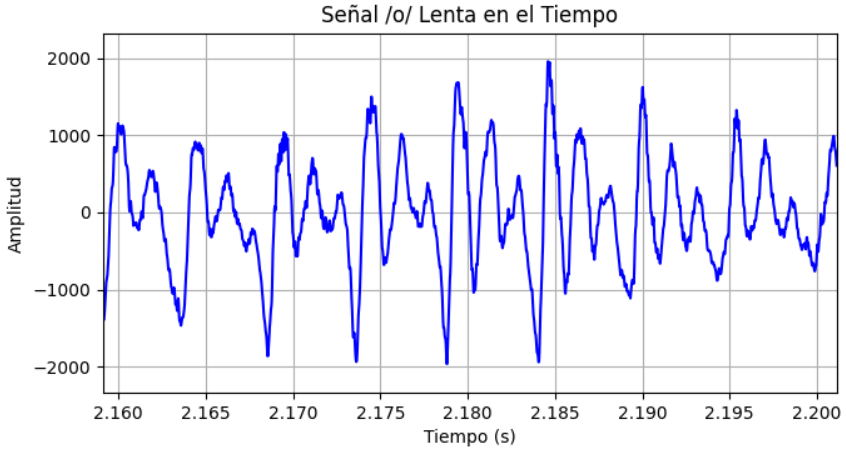


Figura 10: Gráfico de la vocal*/o/ lenta*.

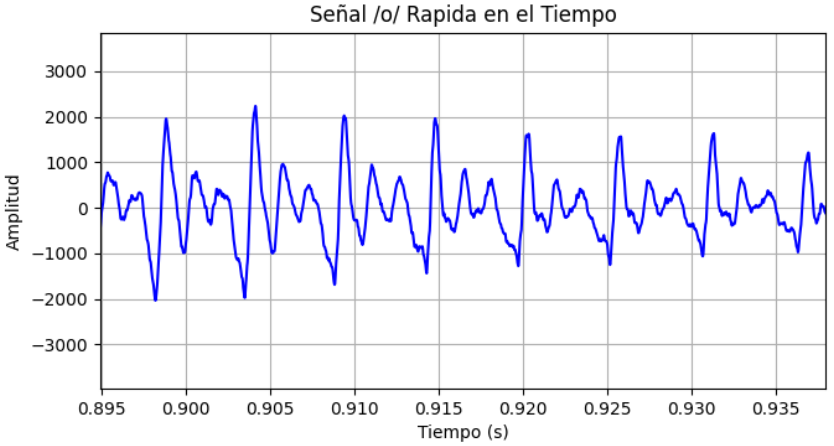


Figura 11: Gráfico de la vocal */o/ rápida*.

El período (T) de una señal se puede calcular como el tiempo total t dividido por el número de períodos N en ese intervalo. Luego, la frecuencia (f), que es el inverso del período obtenido.

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

En base a los gráficos se estimaron los períodos y frecuencias y se muestran en la tabla 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

Tabla 1: Períodos y Frecuencias estimados de las vocales.

Las pequeñas diferencias en la frecuencia (menos de 0.2 Hz) fueron debido al redondeo y no afectan significativamente la precisión de los datos.

1.3. Análisis FFT

Los sonidos sonoros son producidos forzando el aire a través de la glotis o a través de las cuerdas vocales. La tensión de las cuerdas vocales se ajusta de manera tal que vibre en forma oscilatoria. La interrupción periódica del flujo de aire subglotal resulta en un soplido casi periódico de aire que excita el tracto vocal. El sonido producido por la laringe es llamado sonoro o con fonación. Este tipo de sonido consiste en una frecuencia fundamental (F0) y sus componentes armónicos producidos por las cuerdas vocales. El tracto vocal modifica esta señal de excitación causando el formante. El termino formante se utiliza para indicar el centro de estas frecuencias de resonancia, en donde la concentración de energía es mayor. Los formantes son las frecuencias de resonancia del espectro, es decir, los picos de la envolvente del espectro de la señal de voz que representan las frecuencias de resonancia del tracto vocal. Cada formante tiene una frecuencia central, amplitud y un ancho de banda, y son usualmente denotadas F1, F2, F3,...., comenzando con la menor frecuencia. Las frecuencias a las que se producen los primeros formantes son muy importantes para reconocer o sintetizar la voz. En la siguiente Figura 12 pueden verse representados los 3 primeros formantes de una señal de voz y en la Figura 13 se muestra una tabla con las frecuencias de los formantes en español.

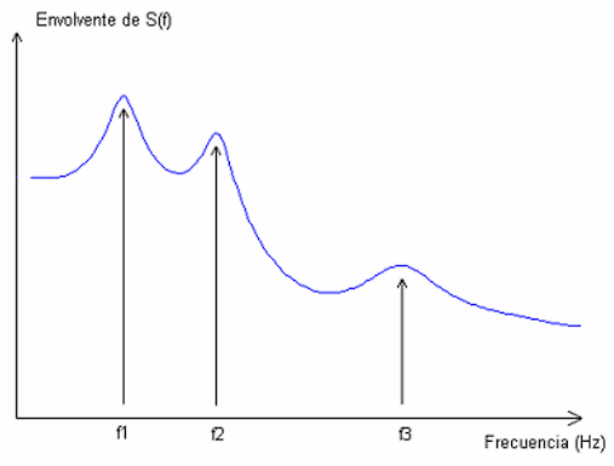


Figura 12: Envolvente del espectro de una vocal.

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

Figura 13: Formantes en español.

Se graficaron los espectros de la señal, Figuras 14 y 15 para obtener los coeficientes de Fourier de las porciones correspondientes a las vocales que hay en la señal y hacer el calculo tomando varios períodos de la vocal y también tomando un solo período para las 2 señales.

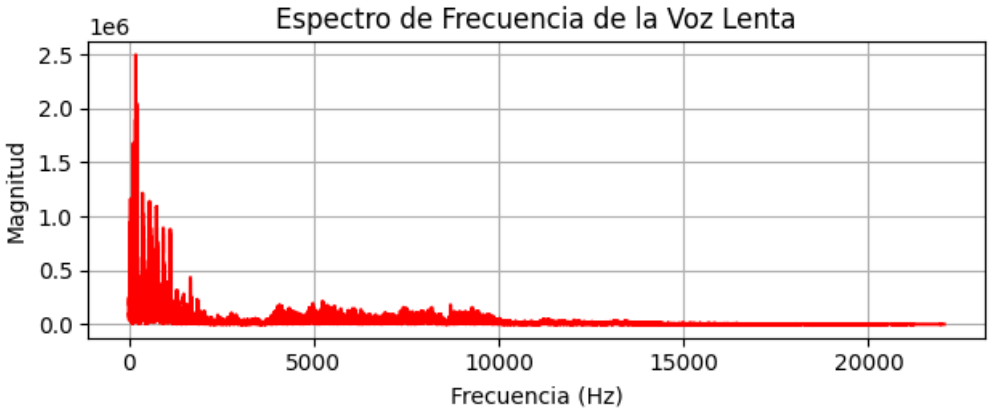


Figura 14: Gráfico del espectro la señal de voz *Picasso lenta*.

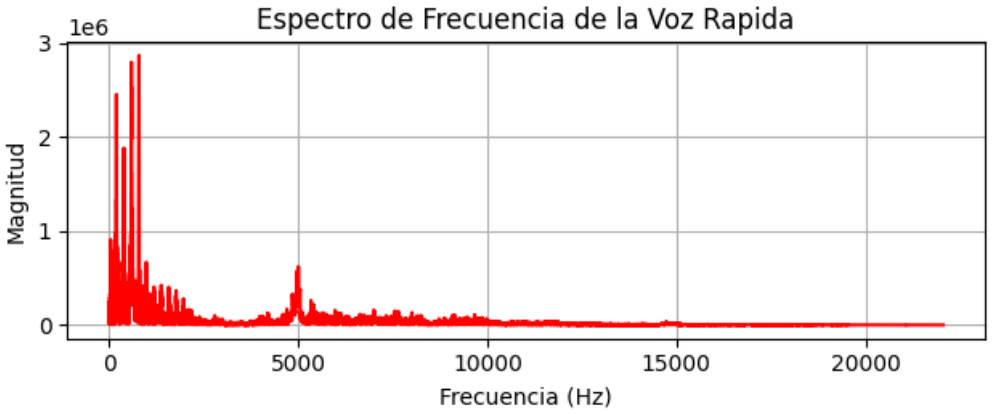


Figura 15: Gráfico del espectro la señal de voz *Picasso rápida*.

Segmento /i/:

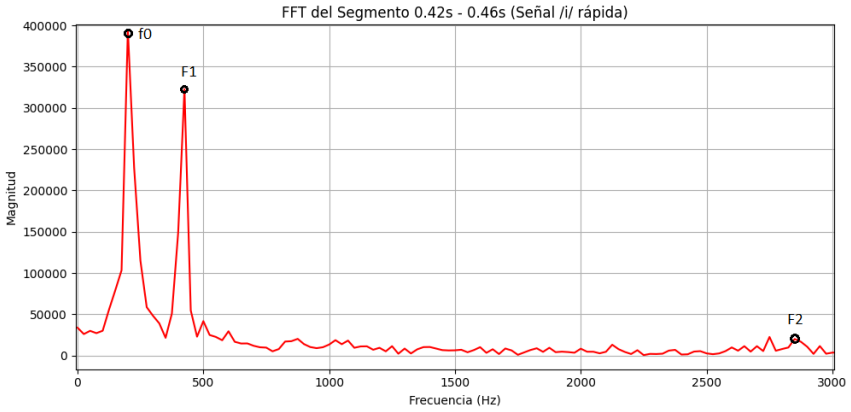


Figura 16: FFT */i/ rápida*.

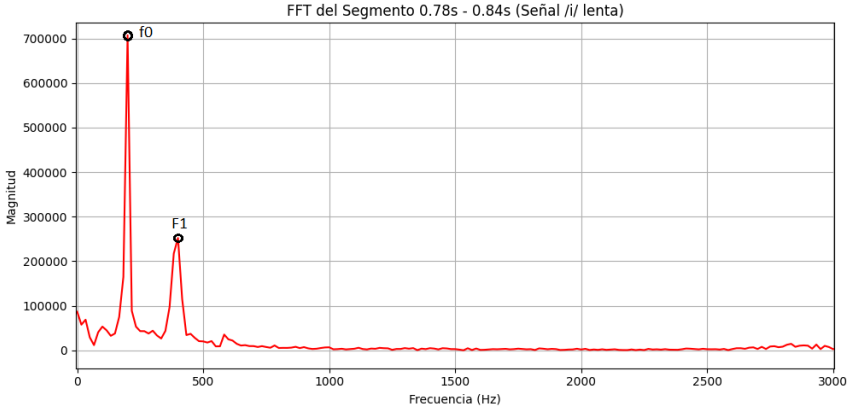


Figura 17: FFT */i/ lenta*.

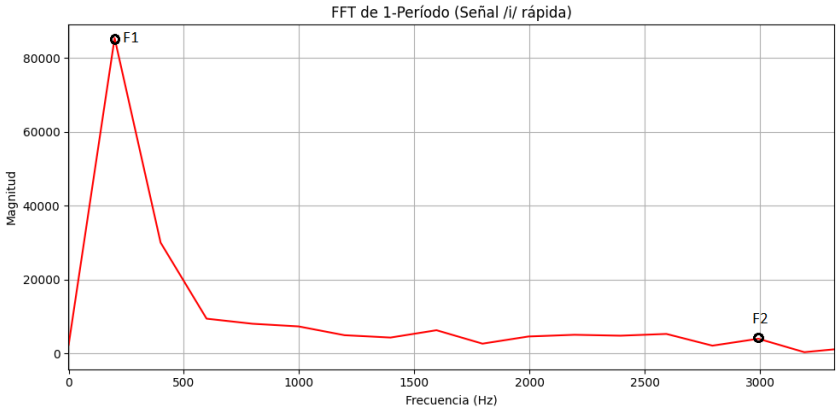


Figura 18: FFT */i/ rápida* 1 período.

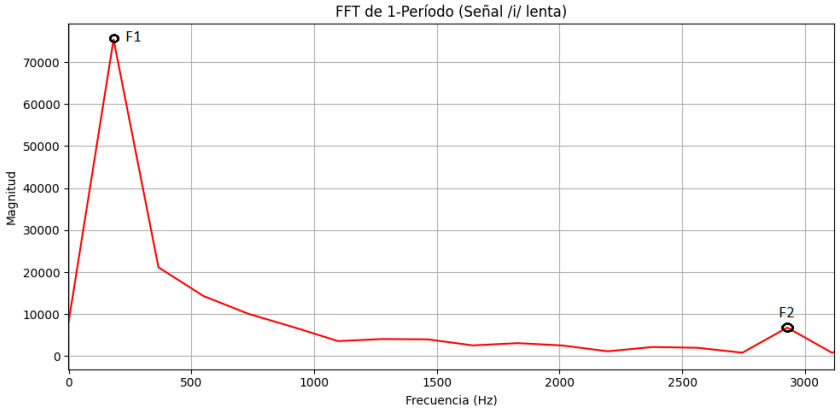


Figura 19: FFT */i/ lenta* 1 período.

Segmento /a/:

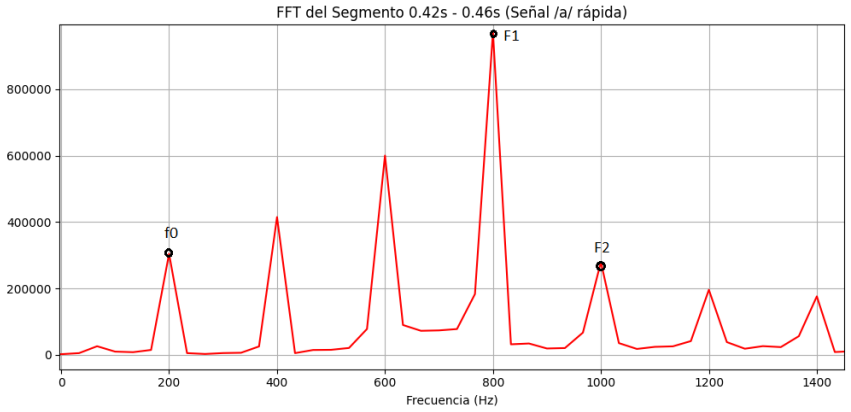


Figura 20: FFT */a/ rápida.*

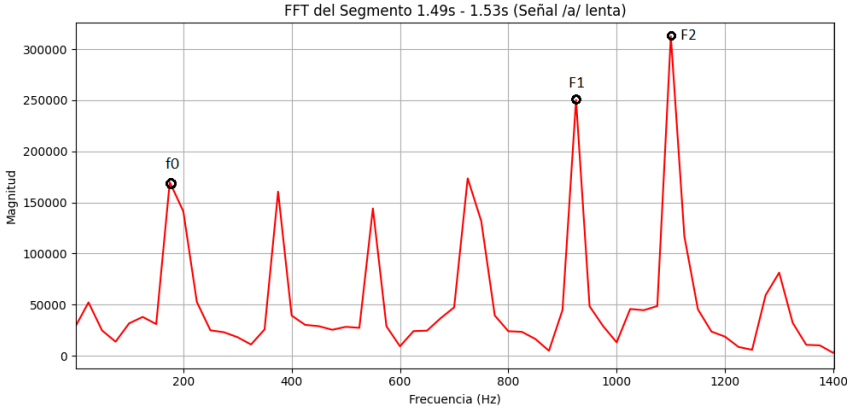


Figura 21: FFT */a/ lenta.*

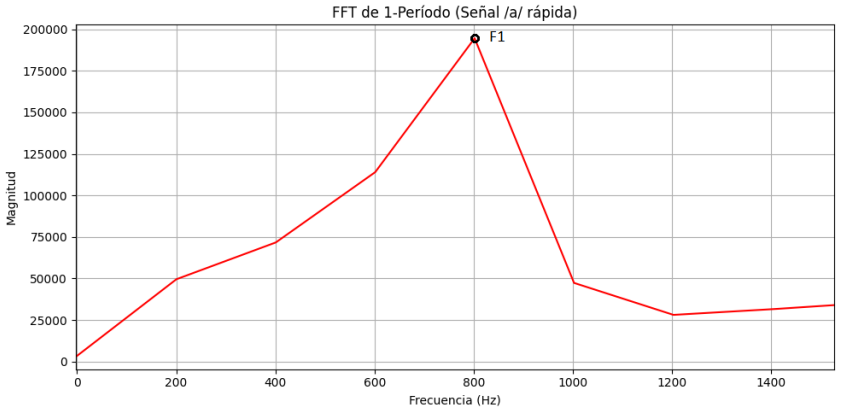


Figura 22: FFT */a/ rápida* 1 período.

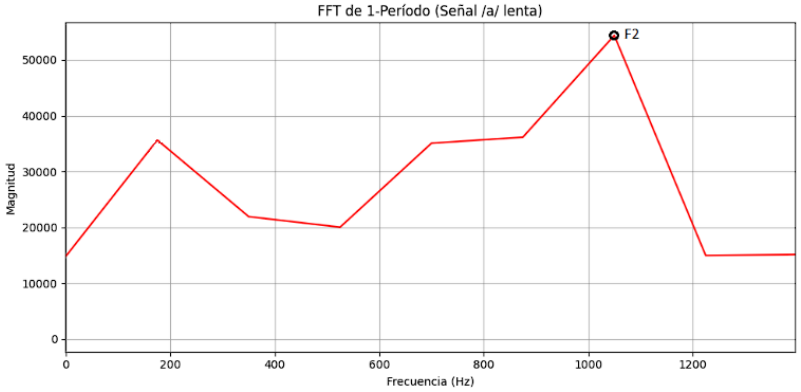


Figura 23: FFT */a/ lenta* 1 período.

Segmento /o/:

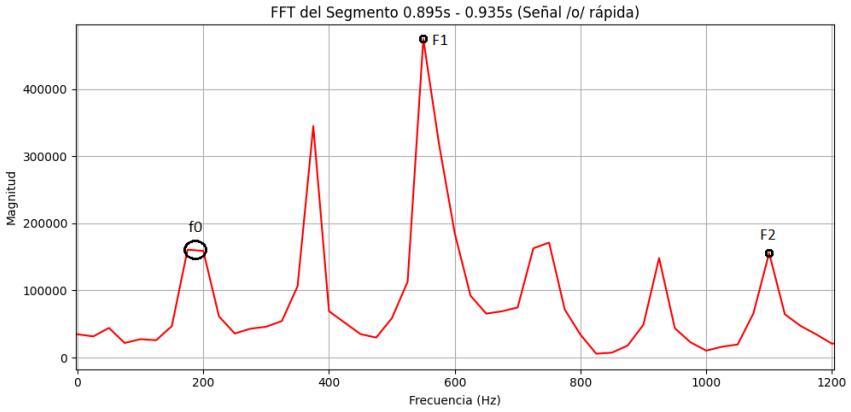


Figura 24: FFT */o/ rápida.*

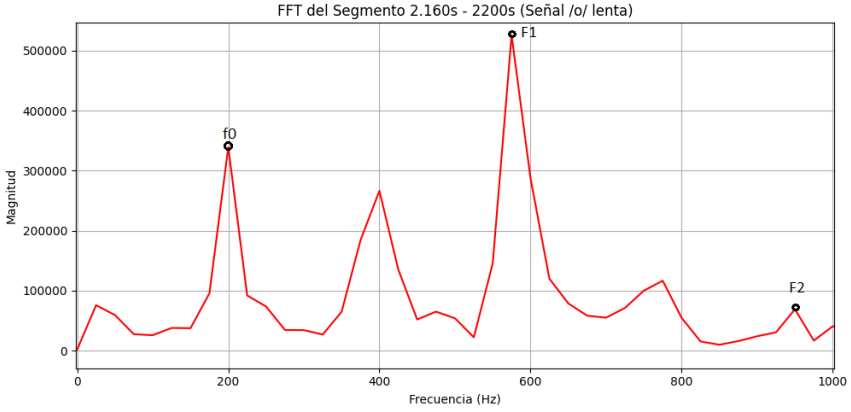


Figura 25: FFT */o/ lenta.*

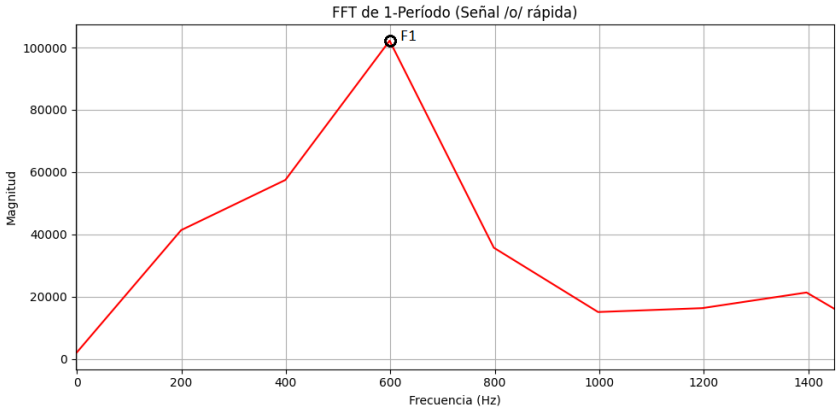


Figura 26: FFT */o/ rápida* 1 período.

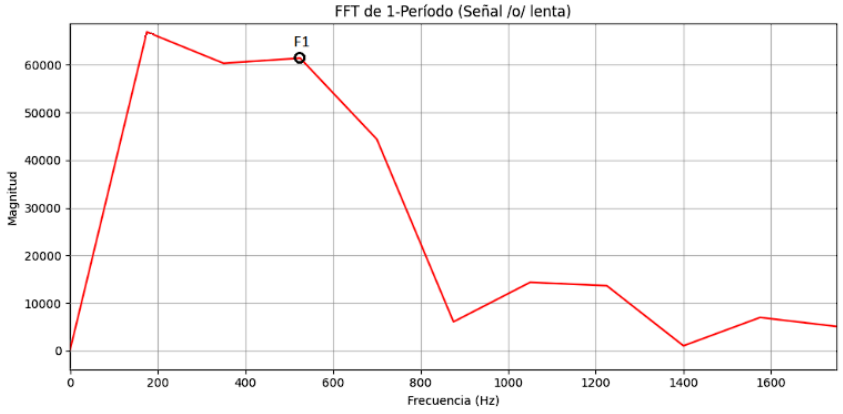


Figura 27: FFT */o/ lenta* 1 período.

Analizando las FFT de las vocales rápidas y lentas, correspondientes a las figuras 16 a 27, se observó que todas presentan una frecuencia fundamental cercana a 200 Hz, lo cual se encuentra dentro de los valores esperados calculados y mostrados en la tabla 1.

En relación con los formantes F1 y F2, se observó que, al analizar las vocales utilizando un solo período, no es posible identificar la frecuencia fundamental, pero si los formantes. Sin embargo, al considerar los gráficos correspondientes a varios períodos, F0 se hace visible y su valor para cada vocal pudo corroborarse que esta dentro del rango esperado cuya franja de valores se mostraron en la figura 13. Esto sugiere que a mayor cantidad de períodos mejora la resolución espectral, favoreciendo la identificación precisa de los formantes por eso también se puede ver la frecuencia fundamental.

Cabe destacar que el primer formante de la vocal /i/ fue el único que presentó una discrepancia $\leq 14,29\,\%$ con respecto al esperado. Esta desviación se atribuye a la agudeza de la voz utilizada durante la grabación, ya que una fonación más aguda tiende a elevar las frecuencias de resonancia del tracto vocal, modificando así la posición esperada de los formantes.

2. Transformada de corto tiempo

La TFCT (Transformada de Fourier de Corto Tiempo)[1] es una transformada de Fourier basada en la DFT. En la práctica, hay muchas aplicaciones en las que las propiedades de la señal que se trata cambian con el tiempo. Por ejemplo, esto sucede con señales no estacionarias tales como las de radar, sonar, voz y señales de comunicaciones. Pues bien, en estos casos calcular una única DFT para toda la señal no es suficiente, además de la dificultad añadida de que ´esta podría ser larguísima siendo imposible de tratar en la práctica, ya que suelen usarse computadores digitales con una capacidad de cálculo y almacenamiento limitados. Todo ello nos guía hacia el concepto de transformada de Fourier de corta duración o TFCT. La TFCT de una señal s(n) se define como:

$$S(n, w) = \sum s(m)w(n - m)e^{-jwn} \tag{2}$$

Donde ω(n) es la ventana. En la TFCT, la secuencia unidimensional s(n), función de una variable discreta, es transformada en una función bidimensional de la variable n, que es discreta, y de la frecuencia ω, que es continua. Hay que ver de que la TFCT es periódica en ω con periodo 2π, y por lo tanto sólo tendremos que considerar los valores incluidos en $0 \leq \omega \leq 2\pi$, o cualquier otro intervalo de longitud 2π. Teniendo en cuenta la simetría de las ventanas, la ecuación anterior puede reescribirse como:

$$S(n, w) = \sum s(m + n)w(m)e^{-jwn} \tag{3}$$

De esta forma, la TFCT puede interpretarse como la transformada de Fourier de la señal desplazada s(m+n), y vista a través de la ventana w(n). La ventana tendría un origen fijo, y según n va cambiando, la señal se desliza pasando a través de la ventana de forma que para cada valor de n vemos una porción diferente de la señal.

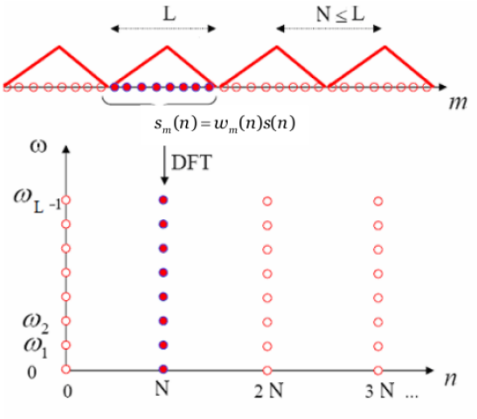


Figura 28: Espectograma.

El espectrograma es una herramienta muy útil para analizar los fonemas y sus transiciones. Un espectrograma de una señal en el tiempo es una representación especial en dos dimensiones, en el eje horizontal representa el tiempo y en el vertical representa la frecuencia. Normalmente se utiliza la escala de grises para indicar la energía en cada punto (t, f) representando con blanco las bajas energías y con negro las altas. El espectrograma se obtiene a partir de la TFCT. El espectrograma solamente representa la energía y no lo fase de la TFCT. La energía la calculamos como:

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

El valor de la ecuación anterior lo convertimos a escala de grises. Aquellos píxeles, cuyo valor no es calculado, se obtienen interpolando.
En Python podemos usar scipy.signal.spectrogram o scipy.signal.ShortTimeFFT.

2.1. Espectrogramas de banda angosta y ancha de la palabra completa

2.1.1. Espectrogramas de banda angosta

En este punto se graficó el espectrograma de la palabra "picassoütilizando una ventana de Hamming de 2048 muestras para aplicar la Transformada de Fourier de Corto Tiempo (STFT), lo que genera un espectrograma de banda angosta. Este tipo de espectrograma proporciona alta resolución en frecuencia y permite observar con claridad la frecuencia fundamental (F) y sus armónicos durante la emisión de vocales. Las líneas horizontales finas y regularmente espaciadas visibles en el gráfico representan los armónicos generados por la vibración periódica de las cuerdas

vocales. La ventana de Hamming fue elegida por sus propiedades favorables para el análisis espectral. Esta ventana suaviza los bordes de cada segmento temporal, lo que reduce el efecto de fugas espectrales (leakage) y permite obtener un espectro más preciso al minimizar los lóbulos laterales sin comprometer demasiado la resolución. En comparación con una ventana rectangular, la ventana de Hamming ofrece un mejor compromiso entre resolución en frecuencia y supresión de componentes espurias, haciendo que los armónicos se destaquen con mayor nitidez en el espectrograma. En este espectrograma de banda angosta de la Figura 29, generado con dicha ventana de 2048 muestras, se observa con claridad la presencia de componentes periódicas propias de los fonemas vocálicos de la palabra "picasso". Estas componentes se manifiestan como líneas horizontales finas y regularmente espaciadas a lo largo del tiempo, especialmente visibles en los intervalos aproximados entre 0.4s–0.5s, 0.6s–0.75s y 0.9s–1.1s.

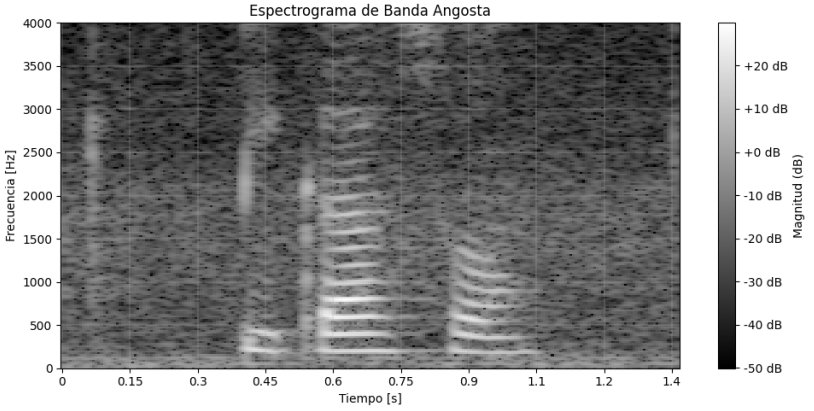


Figura 29: Espectrograma de banda angosta de la señal *Picasso rápido*.

Estas líneas representan los armónicos, que son múltiplos enteros de la frecuencia fundamental (F) producida por la vibración periódica de las cuerdas vocales. La aparición clara y sostenida de estas estructuras indica la presencia de señales sonoras cuasiperiódicas, características de las vocales, mientras que su ausencia o dispersión en otras regiones sugiere la presencia de consonantes sordas o fricativas, como /p/ o /s/. De forma análoga se puede ver esta descripción del espectrograma de banda angosta en la Figura 30 correspondiente al audio lento de la palabra "picasso".

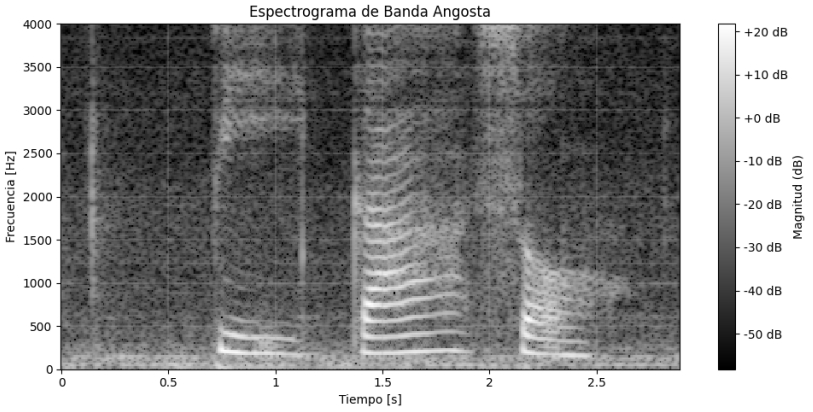


Figura 30: Espectrograma de banda angosta de la señal *Picasso lento*.

Este tipo de espectrograma de banda angosta, con alta resolución en frecuencia, es ideal para estudiar la estructura armónica de la voz y permite estimar con precisión la frecuencia fundamental y sus armónicos, fundamentales para el análisis de la entonación y del contenido tonal de la señal de habla.

2.1.2. Espectrogramas de banda ancha

En este punto se graficó el espectrograma de la palabra "picassoütilizando una ventana de Hamming de 512 muestras para aplicar la Transformada de Fourier de Corto Tiempo (STFT), lo que genera un espectrograma de banda ancha. Este tipo de espectrograma proporciona alta resolución temporal, lo que permite visualizar con mayor claridad los formantes vocálicos, es decir, las bandas de resonancia del tracto vocal durante la producción de sonidos sonoros. La elección de la ventana de Hamming responde a su capacidad para reducir los efectos de fugas espectrales sin distorsionar excesivamente la envolvente temporal. Con un tamaño de 512 muestras, esta ventana logra un equilibrio adecuado entre resolución temporal y precisión espectral, permitiendo observar los cambios rápidos en los formantes que caracterizan a los fonemas vocálicos en habla rápida. En la Figura 31 se observa el espectrograma de banda ancha del audio rápido de la palabra completa "picasso", generado con dicha ventana, se pueden distinguir al menos tres regiones de alta energía asociadas a los formantes de los fonemas vocálicos: Entre 0.4s y 0.5s, se observa la vocal [i], con un primer formante (F1) bajo (300–400 Hz) y un segundo formante (F2) alto (2200–2500 Hz), característico de esta vocal cerrada.

Entre 0.6s y 0.75s, aparece la vocal [a], con F1 cerca de 700 Hz y F2 alrededor de 1200 Hz, lo que concuerda con el patrón típico de una vocal abierta.

Entre 0.85s y 1.1s, se visualiza la vocal [o], con F1 en torno a 500 Hz y F2 más bajo (900–1000 Hz), lo que indica una vocal posterior y redondeada.

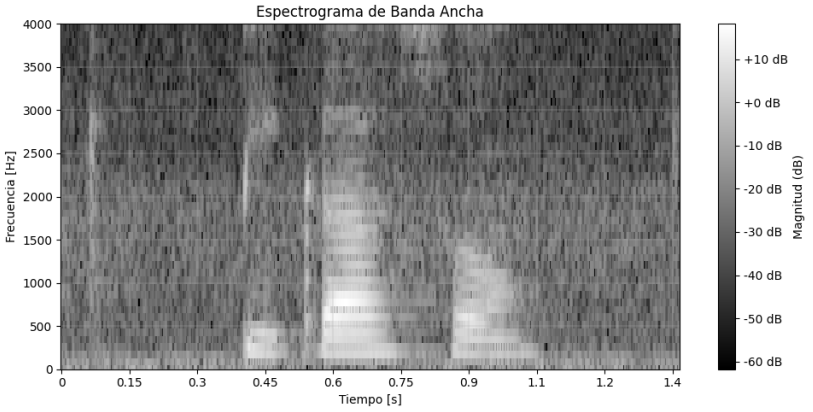


Figura 31: Espectrograma de banda ancha de la señal *Picasso rápido*.

De forma similar se puede ver esta descripción del espectrograma de banda ancha en la Figura 32 correspondiente al audio lento de la palabra "picasso". Estas bandas o franjas horizontales anchas de color rojo y amarillo corresponden a los formantes F1, F2 y F3, y se mantienen relativamente estables durante cada tramo vocálico. Su presencia permite identificar fonéticamente los sonidos vocálicos incluso en condiciones de habla rápida, a diferencia de las consonantes, que se representan por regiones con energía más difusa o ruido no estructurado.

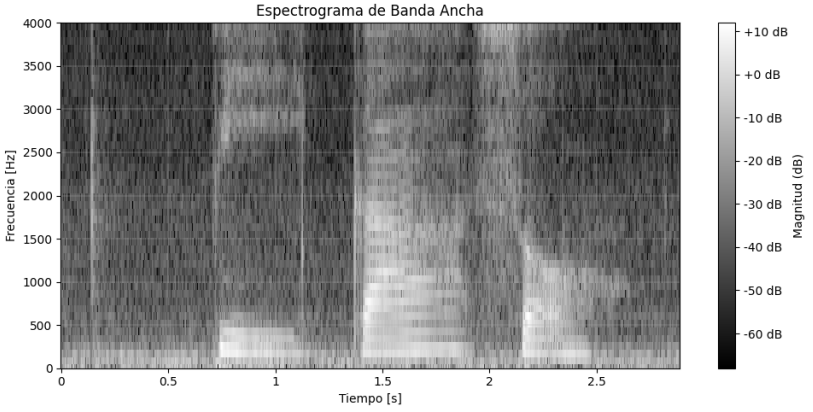


Figura 32: Espectrograma de banda ancha de la señal *Picasso lento*.

Se observa que en este tipo de espectrograma de banda ancha, con alta resolución en tiempo y uso de una ventana de Hamming bien ajustada, es ideal para el estudio

articulatorio y acústico del habla, ya que permite seguir la evolución de los formantes en tiempo real y analizar con precisión las diferencias entre vocales en contextos naturales.

2.2. Espectogramas de banda angosta y banda ancha de las vocales

El espectrograma de banda angosta consiste en un análisis en frecuencia que emplea una ventana de tamaño alto (en este caso, 1024 puntos). Esto suaviza el espectrograma y reduce el ruido, mostrando componentes de frecuencias más claras y detalladas, especialmente útil para distinguir resonancias y formantes cercanas.

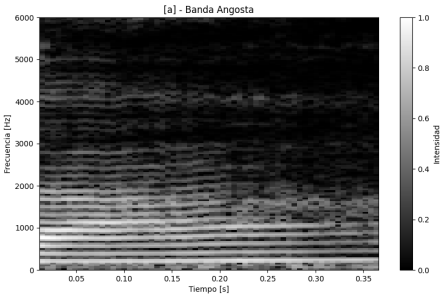
Por otro lado, el espectrograma de banda ancha utiliza una ventana más corta (en este caso, 256 puntos), lo que proporciona mayor resolución temporal pero menor resolución en frecuencia. Como resultado, se obtiene menos claridad en la estructura de armónicos y es menos preciso para detectar formantes cercanas.

Al emplear un espectrograma de banda angosta en el análisis de las vocales, se facilita la observación detallada de la evolución de las formantes a lo largo del tiempo, permitiendo identificar mejor las transiciones, variaciones en la articulación y resonancia vocal.

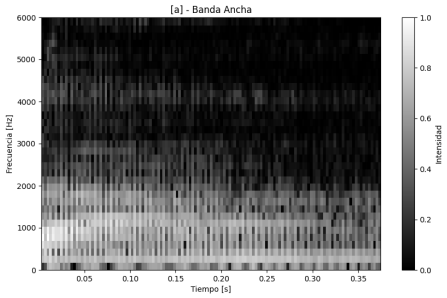
2.2.1. Espectogramas de /a/

En la figura 33a se observan áreas claras que indican regiones con mayor energía en esas frecuencias, lo cual suele corresponder a la fundamental y a los primeros formantes de la vocal. La región de 1000 a 2000 Hz también presenta áreas claras, aunque en menor intensidad que en las frecuencias bajas, señalando los principales formantes en esa banda. Por encima de los 2000 Hz, la gráfica se ve más oscura, indicando menor energía en esas frecuencias.

Por otro lado, en la figura 33b, que corresponde al espectrograma de banda ancha, se destacan líneas claras verticales. Estas líneas reflejan una buena resolución temporal, que permite detectar cambios rápidos en la señal y la presencia de armónicos. Sin embargo, los formantes aparecen menos definidos y agrupados, lo que hace más difícil distinguir entre frecuencias cercanas en la resolución en frecuencia. En este espectrograma, las primeras dos formantes (hasta los 2000 Hz aproximadamente) se perciben con claridad, pero la resolución en frecuencia es menor en comparación con la de banda angosta.



(a) Espectograma de banda angosta.



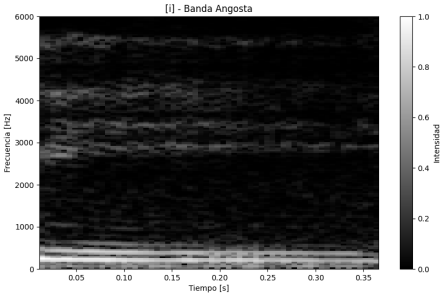
(b) Espectograma de banda ancha.

Figura 33: Espectogramas de la vocal /a/.

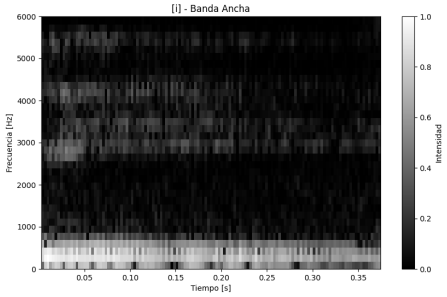
2.2.2. Espectogramas de /i/

En la figura 34a, se observa que en la parte inferior del espectrograma, entre 0 y 1000 Hz, hay áreas claras que indican mayor energía. Esto corresponde a la primera formante de la vocal /i/, que, al ser una vocal cerrada y anterior, típicamente se presenta en el rango de 200 a 400 Hz. La segunda formante, característica de esta vocal, aparece mucho más alta, generalmente entre 2500 y 3500 Hz, pero con menor amplitud, reflejada en las líneas más delgadas del espectrograma. Por encima de los 3500 Hz, la energía disminuye significativamente, resultando en una región más oscura y menos prominente.

En contraste, la figura 34b muestra una mayor difusión de energía en las frecuencias bajas, con una primera formante claramente presente por debajo de los 900 Hz. Por encima de los 1000 Hz, se observa una región más oscura, lo que podría indicar una mayor densidad de energía en estas frecuencias, posiblemente debido a la superposición de armónicos.



(a) Espectograma de banda angosta.



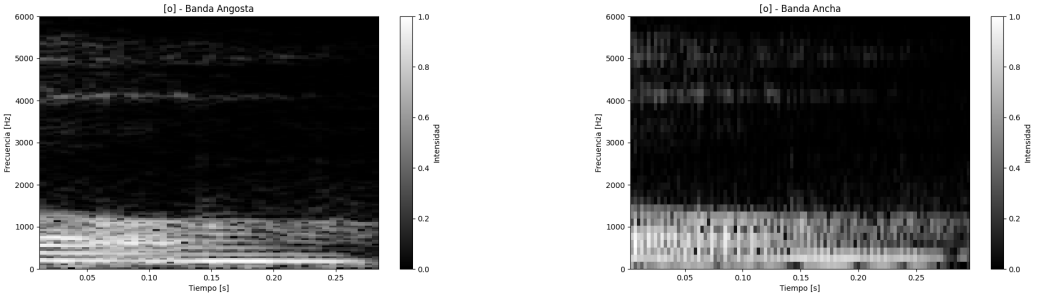
(b) Espectograma de banda ancha.

Figura 34: Espectogramas de la vocal /i/.

2.2.3. Espectogramas de /o/

En la figura 35a se observa el primer formante como una región más clara en la parte baja del espectrograma, generalmente entre 400 y 700 Hz, bien definida y separada del ruido de fondo. El segundo formante aparece entre 800 y 1200 Hz, con armónicos que se extienden hasta aproximadamente 4000 Hz. Por encima de los 1500 Hz, las líneas armónicas se vuelven más tenues, lo que es característico de las vocales posteriores como /o/, reflejando una menor concentración de energía en las frecuencias superiores.

En contraste, en la figura 35b, la mayor parte de la energía se concentra por debajo de los 1000 Hz, con formantes más anchos y menos definidos debido a la menor resolución en frecuencia. Las frecuencias más altas se presentan con una densidad energética significativamente menor, evidenciada por las regiones más oscuras del espectrograma.



(a) Espectograma de banda angosta. (b) Espectograma de banda ancha.

Figura 35: Espectogramas de la vocal /o/.

2.3. Modificación del Pitch usando TD-PSOLA

En esta parte del Proyecto vamos a modificar la señal de voz para que lo dicho por una mujer suene como dicho por un hombre o viceversa. Para esto vamos a modificar la frecuencia fundamental de la señal que es la frecuencia de pitch. Si tenemos la grabación de un hombre 85- 170 Hz,vamos multiplicar la frecuencia fundamental por 1.4 y si es de mujer por 0.7. El método que vamos a usar es el PSOLA (Pitch Synchronous Overlap and Add) para cambiar tono sin cambiar velocidad. Existen varias versiones de PSOLA, usaremos la denominada PSOLA en el dominio temporal(TD-PSOLA, del inglés Time Domain PSOLA). El método consiste en tomar porciones de la señal, multiplicarla por una ventana temporal sincrónica con la frecuencia fundamental para luego combinarlas sincrónicamente con una nueva

frecuencia fundamental (Fig. 36). Los segmentos pueden ser repetidos o eliminados para no aumentar ni disminuir la duración de la emisión de voz.

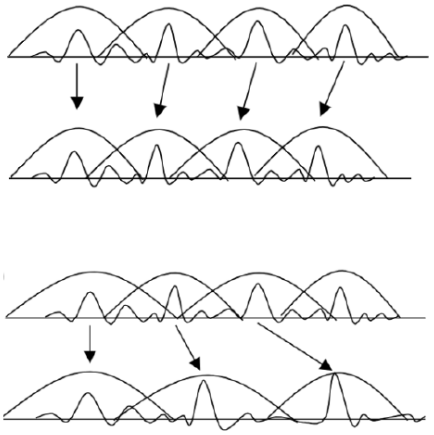


Figura 36: TD-PSOLA, Time Domain PSOLA.

El algoritmo de modificación de la frecuencia fundamental mediante el TD-PSOLA puede resumirse como:

- Detectar los segmentos de la señal que se corresponden con sonidos sonoros. Que son los únicos que hay que modificar.
- Localizar los picos de cada ciclo que conforman los segmentos sonoros.
- Aplicar una ventana centrada en cada pico, desplazarla temporalmente y sumarla para obtener la señal resultante.

La Figura 37 muestra como el algoritmo TD-PSOLA permite transformar una señal de voz, generando una versión con tono más grave a partir de la señal original. Se puede observar que la señal original, representada en color azul, y la señal modificada, en color rojo, se superponen en el mismo eje temporal.

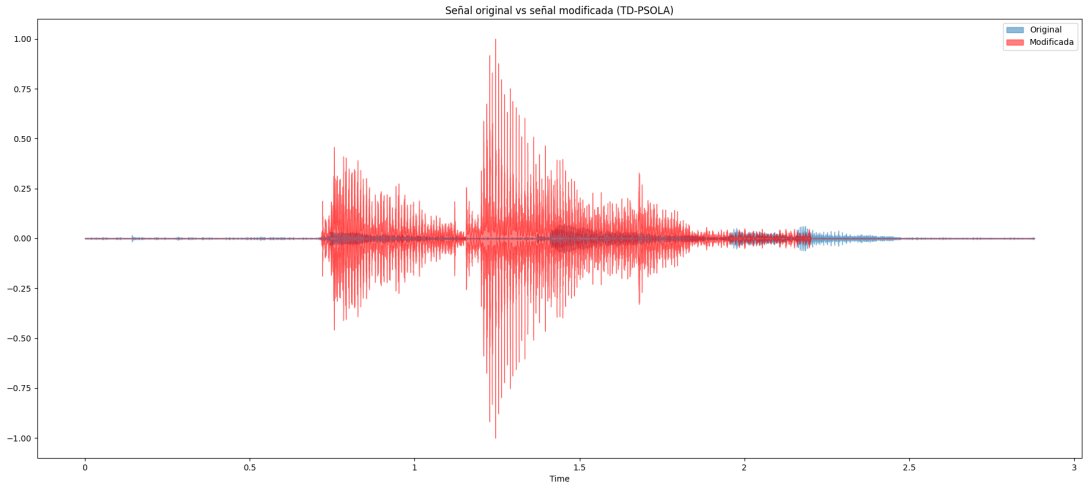


Figura 37: Gráfico de la señal de voz lenta con TD-PSOLA.

En cuanto a la envolvente de la señal, esta se mantiene, lo que indica que la duración total del audio no fue modificada, cumpliendo así con uno de los objetivos del algoritmo. Sin embargo, la estructura interna sí se modifica; esto se debe al cambio en la frecuencia fundamental, lo cual produce una alteración en el tono de la voz.

Finalmente, el gráfico permite confirmar visualmente que el pitch fue alterado, pero sin distorsionar la señal ni modificar su duración.

En este segundo módulo del trabajo, se analizaron señales de voz utilizando la Transformada de Fourier de Corto Tiempo (TFCT) para obtener espectrogramas que permitieran observar la evolución temporal y frecuencial de los fonemas. Se aplicaron espectrogramas de banda angosta y banda ancha, con distintas resoluciones temporales y frecuenciales, lo cual permitió identificar de manera precisa tanto los armónicos como los formantes de las vocales.

Los espectrogramas de banda angosta, con mayor resolución en frecuencia, facilitaron la observación de las componentes armónicas de la señal, mientras que los de banda ancha, con mejor resolución temporal, permitieron analizar la dinámica de los formantes. Además, se implementó una técnica de modificación de pitch mediante TD-PSOLA, logrando alterar la frecuencia fundamental de la voz sin modificar su duración. Este análisis demuestra la utilidad de las herramientas de procesamiento espectro-temporal en el estudio de señales no estacionarias como la voz, aportando una visión detallada de su estructura y facilitando transformaciones prácticas como la conversión de tono.

3. Cambios de velocidad de la señal de habla

Hay muchas aplicaciones que requieren un cambio de la velocidad de la señal de habla. Un esquema simple que solo interpole o decime la señal de habla cambiaría la velocidad pero también alteraría las características fundamentales como la frecuencia glótica o la posición de los formantes. La señal alterada de esta manera podría incluso resultar ininteligible. Idealmente, lo que se requiere es variar la velocidad de la señal preservando sus características en frecuencia.

3.1. Decimación de la señal lenta

Si la secuencia $x[n]$ es el resultado del muestreo de una señal continua $x_c(t)$, y se representa mediante una secuencia expandida $x_p[n]$ con ceros intercalados (modelo de interpolación por inserción de ceros), entonces el proceso de **decimación** puede interpretarse como la eliminación de dichas muestras nulas, lo que conduce a una reducción efectiva de la velocidad de muestreo T de $x_c(t)$ por un factor N, de forma tal que la velocidad de muestreo total es NT.

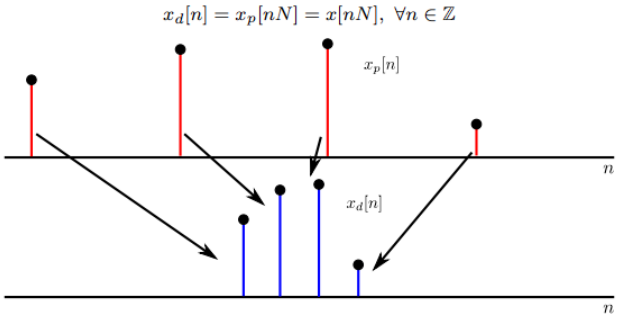


Figura 38: Decimación en el dominio temporal.

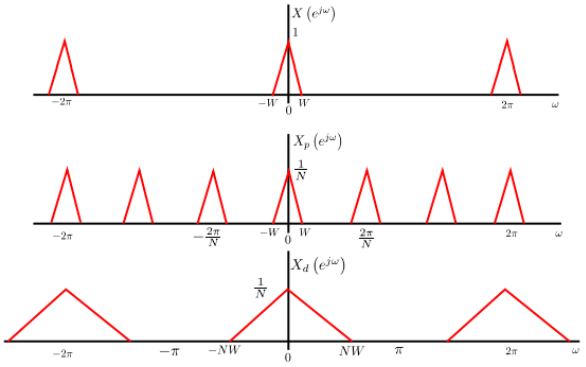


Figura 39: Decimación en el dominio espectral.

Es claro que, para poder realizar la decimación por un factor $N > 1$ sin incurrir en aliasing, la señal continua $x_c(t)$ debe haber sido muestreada con una frecuencia $\omega_s > 2NW$, donde W es el ancho de banda de $x_c(t)$.

Si se tiene una señal discreta $x[n]$ cuyo contenido espectral no permite la decimación directa sin aliasing, será necesario reducir su ancho de banda mediante un filtro pasa-bajos ideal en tiempo discreto, con ganancia unitaria y frecuencia de corte $\omega_c \leq \frac{\pi}{N}$, que elimine las componentes de frecuencia por encima de ω_c.

La respuesta al impulso de este filtro ideal corresponde a una función `sinc`, la cual es infinita en duración y no causal. Por lo tanto, no puede implementarse directamente en un sistema real. Para obtener una versión realizable del filtro, se recorta esta sinc mediante la multiplicación por una ventana temporal. Este procedimiento se conoce como el *método del ventaneo*, y produce un filtro de respuesta finita:

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

Si bien podría utilizarse una ventana rectangular (recorte abrupto), esto genera oscilaciones no deseadas en la respuesta en frecuencia, fenómeno conocido como el *efecto Gibbs*. En cambio, ventanas como Hamming o Blackman suavizan este recorte y reducen los "ripples" en la banda de transición, mejorando la atenuación fuera de banda del filtro.

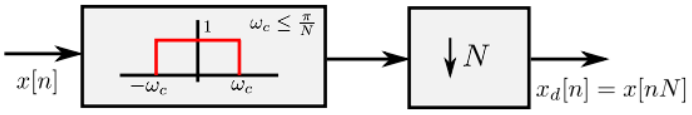


Figura 40: Proceso de decimación.

Siguiendo el proceso de decimación para la señal *Picasso lenta* por un factor $N = 2$, se duplicó la velocidad temporal sin incurrir en aliasing debido a la aplicación del filtro antialiasing diseñado con el método del ventaneo.

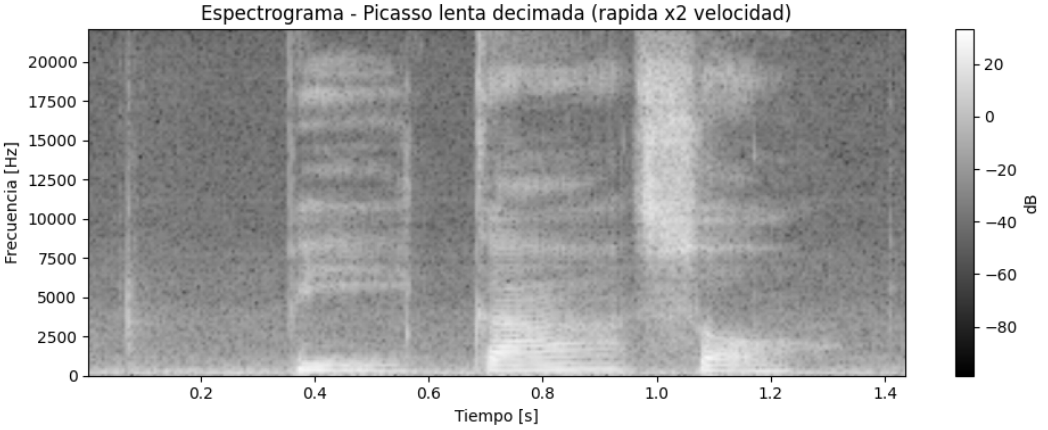


Figura 41: Espectrograma de la señal *Picasso lenta* decimada.

En la figura 41 se observa que no se produjo aliasing (no aparecen líneas entrecruzadas ni contenido espectral inesperado), y que se conservaron las características vocales, como los formantes y armónicos principales.

Se pueden comparar las componentes espectrales con los espectrogramas de la señal *rápida original*, mostrados en la figura 31 de banda ancha, cuyo contenido espectral resultante presentó bandas mas a la izquierda con respecto a la señal decimada. Con respecto al contenido de frecuencias, también se ve modificada, ya que el de la señal original llegaba a los 4KHz y el decimado a los 20KHz.

3.2. Interpolación de la señal rápida

Se considera ahora la posibilidad de *aumentar* la frecuencia de muestreo de una señal discreta $x[n]$.

Este proceso se realiza mediante un **expansor**, un dispositivo que inserta $L-1$ ceros entre cada muestra de la señal original. La salida de este expansor, aunque tiene una mayor tasa de muestreo, aún no es una señal útil, ya que contiene muchos ceros artificiales.

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

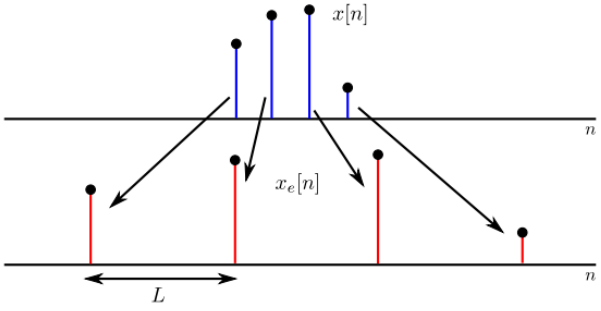


Figura 42: Interpolación en el dominio temporal.

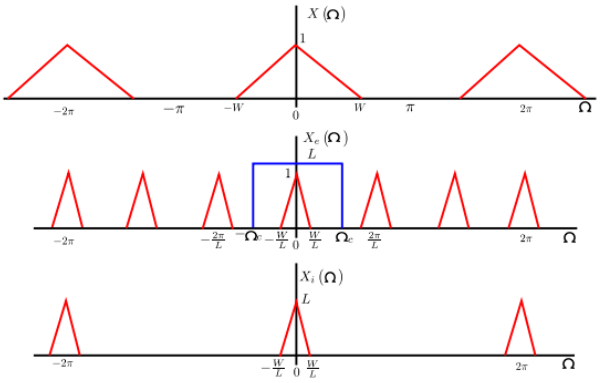


Figura 43: Interpolación en el dominio espectral.

La transformada de Fourier de esta señal expandida es igual a la transformada de la señal original, **comprimida en frecuencia por un factor** L, y replicada periódicamente en el intervalo $[-\pi, \pi]$. Para recuperar una señal interpolada útil, se requiere un filtro pasa-bajos posterior que elimine las réplicas.

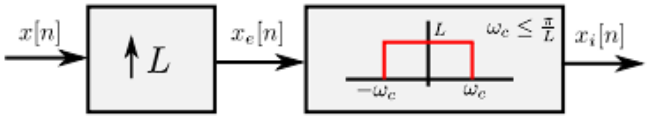


Figura 44: Estructura interpolación.

Se procedió a utilizar dicho método de interpolación para modificar la duración

temporal de la señal de habla *Picasso rápida* de manera tal de disminuir su velocidad a la mitad y de obtener una señal más prolongada en el tiempo.

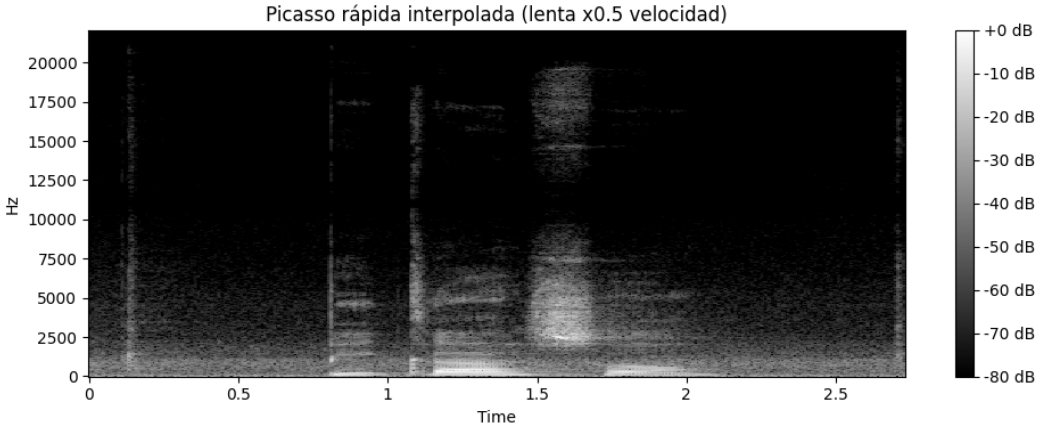


Figura 45: Espectrograma de la señal *Picasso rápida* interpolada.

En la figura 45 se observa la estructura espectral se ve estirada en el eje temporal, como es de esperarse al duplicar la duración.

Se pueden comparar las componentes espectrales con los espectrogramas de la señal *lenta original*, mostrados en la figura 32 de banda ancha, cuyos formantes, armónicos y bandas están mas a la izquierda que el espectrograma correspondiente a la señal rápida interpolada, y las frecuencias alcanzan hasta los 4KHz mientras que el interpolado aumenta a 20KHz.

3.3. El método phase vocoder

Los métodos de decimación e interpolación cambian la duración temporal de una señal de habla pero también desplazaron los formantes y armónicos. Por lo que para reducir este efecto hay varios métodos. Uno de ellos es el PSOLA aplicado en la segunda parte del proyecto (que trabaja en el dominio temporal) y otra es el **phase vocoder** que trabaja con la TFCT (Transformada de Fourier de corto tiempo):

1. TFCT: Se aplica la TFCT a la señal.
2. Interpolación: Se realiza una interpolación o remuestreo de las "filas" de la matriz de la TFCT, agregando o quitando columnas para ajustar a la duración deseada.
3. Síntesis Temporal: Se realiza la inversa de la TFCT (iTFCT). Esto puede consistir, por ejemplo, en hacer la iDFT por columnas, compensando el efecto

de la ventana usada para la TFCT, y combinando los resultados en una única señal.

3.3.1. Aumento de la velocidad de la señal lenta por TFCT

El proceso de aumentar al doble la velocidad de la señal mediante la modificación de su Transformada de Fourier de Corto Tiempo (TFCT) consiste en trabajar en el dominio de la frecuencia. Específicamente, se eliminan una de cada dos columnas del espectrograma, lo que equivale a reducir la resolución temporal de la representación TFCT.

Esta operación puede interpretarse como una forma de submuestreo en el eje temporal de la TFCT. Al eliminar columnas, se conservan solo la mitad de los segmentos temporales, lo que genera una versión comprimida de la señal original. Al reconstruir la señal utilizando la inversa de la TFCT (iTFCT), se obtiene una señal más corta en duración, es decir, con una mayor velocidad de reproducción.

Este procedimiento no se basa en la eliminación directa de muestras en el tiempo, como ocurre en la decimación tradicional, sino que modifica la estructura temporal de la señal a través de su representación espectro-temporal. Aunque no garantiza la preservación perfecta del contenido original, puede ser útil en contextos donde se busca una transformación perceptual eficiente.

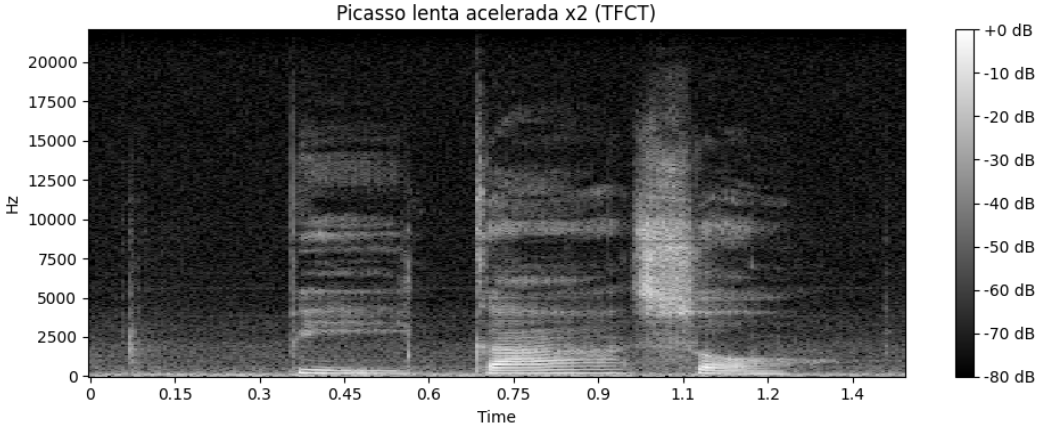


Figura 46: Espectrograma de la señal *Picasso lenta* acelerada por TFCT.

El espectrograma de la figura 46, es el resultado de la modificación de su Transformada de Fourier de Corto Tiempo (TFCT) mediante la eliminación de una de cada 2 columnas de la señal *Picasso lenta* y posterior recuperación de la señal temporal

utilizando la iTFCT, el mismo coincide con la figura 30, correspondiente espectrograma de banda angosta con mayor resolución, ya que en el gráfico se pueden ver las lineas definidas de los armónicos, salvo por el contenido en frecuencias ya que éste llega a los 20KHz mientras que el original de banda angosta a los 4KHz.

3.3.2. Reducción de la velocidad de la señal rápida por TFCT

Para reducir a la mitad la velocidad de la señal Picasso rápida, se modificó su Transformada de Fourier de Corto Tiempo (TFCT) mediante interpolación de columnas. Este proceso consiste en estimar columnas intermedias entre pares consecutivos de la TFCT original, calculando tanto la magnitud como la fase espectral a partir de sus columnas vecinas.

Al duplicar la cantidad de columnas en la matriz tiempo-frecuencia, el espectrograma resultante representa una señal que varía más lentamente en el tiempo, logrando así una versión ralentizada.

Finalmente, se aplicó la inversa de la TFCT (iTFCT) para reconstruir la señal en el dominio temporal. El resultado es una señal con una duración aproximadamente doble respecto de la original, conservando sus componentes espectrales principales.

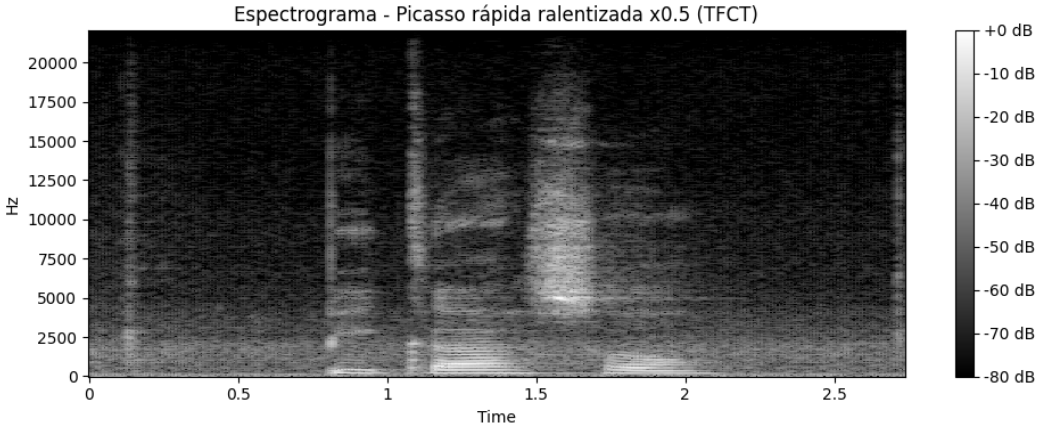


Figura 47: Espectrograma de la señal *Picasso rápida* ralentizada por TFCT.

El espectrograma de la figura 47, obtenido tras la modificación de su Transformada de Fourier de Corto Tiempo (TFCT) mediante interpolación y posterior reconstrucción mediante la iTFCT, coincide con la figura 30 correspondiente al espectrograma de banda angosta con mayor resolución. Tal como ocurrió con el espectrograma de la señal lenta acelerada por TFCT vemos como las frecuencias son mayores a los de la señal original.

4. Conclusiones

- En la primera sección se analizó la periodicidad de la señal de voz de la palabra *Picasso* en sus versiones lenta y rápida, junto con su análisis mediante FFT. Se comprobó que las frecuencias de las vocales coinciden con los formantes teóricos tabulados.

  Aunque la velocidad de pronunciación afecta el dominio temporal (duración, espaciado), el contenido espectral *especialmente de las vocales* se mantiene estable. Esto resalta la importancia de los formantes como rasgos robustos para el reconocimiento de voz y evidencia el valor del análisis combinado temporal-frecuencial.

- En el segundo módulo se utilizaron espectrogramas basados en la TFCT para observar la evolución temporal y frecuencial de los fonemas, aplicando variantes de banda angosta y ancha para identificar armónicos y formantes.

  Los espectrogramas de banda angosta facilitaron la visualización de componentes armónicas, mientras que los de banda ancha permitieron analizar la dinámica temporal de los formantes. Además, se implementó TD-PSOLA para modificar el pitch sin alterar la duración, demostrando la utilidad de estas herramientas para el análisis y transformación de señales de voz.

- En la última sección se compararon cuatro métodos para modificar la velocidad de una señal de voz: decimación e interpolación en el dominio temporal, y modificaciones mediante TFCT. Si bien tanto las técnicas basadas en el *dominio temporal* como las *espectrales* permiten modificar la velocidad de una señal de manera similar, se observó que las señales procesadas mediante la Transformada de Fourier de Corto Tiempo (TFCT) presentan una mayor resolución espectral en comparación con aquellas modificadas mediante decimación o interpolación. Estas últimas alteran el pitch de la señal durante el proceso de aceleración o ralentización, provocando que suenen más agudas o más graves.

  En cambio, las señales modificadas mediante técnicas de tipo vocoder, como el vocoder de fase, logran mantener el contenido espectral original, preservando así el timbre y las características acústicas de la voz. Esto evidencia una clara ventaja del enfoque espectro-temporal para la manipulación de la velocidad, ya que permite conservar la calidad perceptual de la señal sin introducir artefactos audibles o distorsiones en la altura tonal. Finalmente, se destaca que al modificar la duración de la señal mediante técnicas espectrales, el espectro resultante muestra un rango de frecuencias mayor.