Universidad de la República

Facultad de Ingeniería

Procesamiento digital de señales de audio

Práctico 3

Autor:
Federico Bello

29 de octubre de 2024

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Índice

1. Ejercicio 1

En este ejercicio se estudia la relación entre la transformada de Fourier de tiempo corto (STFT) y la función de autocorrelación de tiempo corto. Si se define la densidad espectral de potencia en tiempo corto de una señal ($x[n]$) en función de su transformada de Fourier en tiempo corto como

$$S_n(e^{jw}) = |X_n(e^{jw})|^2 \tag{1}$$

y la función de autocorrelación de tiempo corto de la señal ($x[n]$) como

$$R_n[k] = \sum_{m=-\infty}^{\infty} w[n-m]x[m]w[n-k-m]x[m+k], \tag{2}$$

se vera que si

$$X_n(e^{jw}) = \sum_{m=-\infty}^{\infty} x[m]w[n-m]e^{-jwm} \tag{3}$$

$R_n[k]$ y $S_n(e^{jw})$ son un par de transformadas, i.e. $S_n(e^{jw})$ es la transformada de Fourier de $R_n[k]$.
De 1 se desprende que

$$\begin{aligned} S_n(e^{j\omega}) &= |X_n(e^{j\omega})|^2 && (4)\\ &= X(e^{j\omega})X^*(e^{j\omega}) && (5)\\ &= \mathrm{DTFT}(x[m]w[n-m])\,\mathrm{DTFT}(x[m]w[n-m])^* && (6) \end{aligned}$$

Llamando $u[m] = x[m]w[n-m]$ se llega a que

$$IDTFT(S_n(e^{j\omega})) = u[m] * u[-m] \tag{7}$$

.

Finalmente, utilizando que $f_n[k] * f_n[-k] = \sum_{m=-\infty}^{\infty} f_n[m]f_n[m+k]$ se tiene que

$$\begin{aligned} \mathrm{IDTFT}(S_n(e^{j\omega})) &= \sum_{m=-\infty}^{\infty} w[n-m]x[m]w[n-k-m]x[m+k] && (8)\\ &= R_n[k] && (9) \end{aligned}$$

Por lo tanto, $S_n(e^{jw})$ es la transformada de Fourier de $R_n[k]$.

2. Ejercicio 2

Detección de *pitch* basado en la transofrmada de Fourier de tiempo corto

2.1. Parte 1

El producto armónico espectral -*Harmonic Product Spectrum*, HPS está dado por

$$P_n(e^{jw}) = \prod_{r=1}^{K} |X_n(e^{jwr})|^2 \tag{10}$$

Tomando el logaritmo se obtiene (*log-Harmonic Product Spectrum*, log-HPS),

$$\hat{P}_n(e^{jw}) = 2\sum_{r=1}^{K} \log |X_n(e^{jwr})| \tag{11}$$

El **Harmonic Product Spectrum (HPS)** es eficaz para la detección de pitch debido a su capacidad para resaltar las frecuencias fundamentales de una señal armónica. En el contexto de una señal de audio monofónica (una única fuente armónica), la frecuencia fundamental f_0 y sus armónicos $2f_0, 3f_0, \ldots$ están presentes en el espectro de frecuencia. El HPS multiplica las magnitudes espectrales de las frecuencias en diferentes escalas (o compresiones de frecuencia) para enfatizar las componentes que son múltiplos enteros de la frecuencia fundamental. Esto significa que cuando se toma el producto armónico, los picos resultantes en el espectro corresponden a la frecuencia fundamental y sus armónicos, lo que facilita su identificación.

El uso del **log-HPS** presenta varias ventajas: en primer lugar, la reducción de la varianza es significativa, ya que al aplicar el logaritmo a las magnitudes espectrales, se minimizan las grandes diferencias de amplitud entre las componentes espectrales. Esto ayuda a suavizar las fluctuaciones y hace que el análisis sea menos sensible a variaciones en la amplitud de la señal. Además, el log-HPS proporciona una mejor respuesta a frecuencias bajas, permitiendo que estas se analicen sin ser dominadas por componentes de mayor energía.

Por último, el log-HPS mejora la robustez del método frente al ruido. En presencia de ruido, el log-HPS puede mantener una mejor relación señal-ruido en comparación con el HPS. Asimismo, la transformación logarítmica convierte la multiplicación de espectros en una suma, lo que permite una combinación más efectiva de las distintas escalas en el análisis. En cuanto a las señales cuya frecuencia fundamental está ausente, como aquellas que han sido filtradas por un filtro pasa-altos, el HPS podría no detectar picos significativos, lo que dificultaría la identificación de la frecuencia fundamental.

2.2. Parte 2 y 3

Otro algoritmo para la estimación de la frecuencia fundamental es el espectro logarítmico acumulado (ó GLogS por sus siglas en inglés), se calcula como el promedio de logaritmo de la magnitud del espectro en posiciones armónicas de una frecuencia fundamental f_0, como

$$\rho_n(f_0) = \frac{1}{n_H}\sum_{i=1}^{n_H} \log |X_n(if_0)| \tag{12}$$

siendo n_H la cantidad de armónicos de f_0 cuya frecuencia es menor a cierta frecuencia máxima $f_{\text{máx}}$.

En este caso, se implemento un algoritmo de detección de *pitch* que calcule el GLogS para valores de f_0 distribuidos de forma logarítmica entre $55Hz$ ($A1$) y $1046{,}5Hz$ ($C6$) con un paso de cuarto de tono, y $f_{\text{máx}} = 5000Hz$.

Para evaluar el algoritmo implementado, se tomo como referencia el archivo *LP-mem-6-a*, un audio de una mujer cantando el cual ya se había analizado previamente en la práctica 1.

En la figura 1 se observa la estimación de aplicar el algoritmo sin ningun pos-procesado. Se observa como, si bien de a momentos la frecuencia estimada se asemeja a la real, hay notorios errores. En particular, se ve como cuando no hay voz se presentan errores, posiblemente debido a la presencia del ruido.

Para mejorar esta estimación, se aplicó un pos-procesado. Este pos-procesado se basa en seleccionar la primera frecuencia que supera un determinado umbral, en lugar de tomar todas las frecuencias obtenidas directamente del algoritmo. Específicamente, se selecciona la primera frecuencia que supera el 85 % del valor máximo de GLogS , lo cual permite descartar los armónicos de mayor energía en la señal y elegir la frecuencia más baja como la frecuencia fundamental *frame*.

Este enfoque aporta dos ventajas importantes:

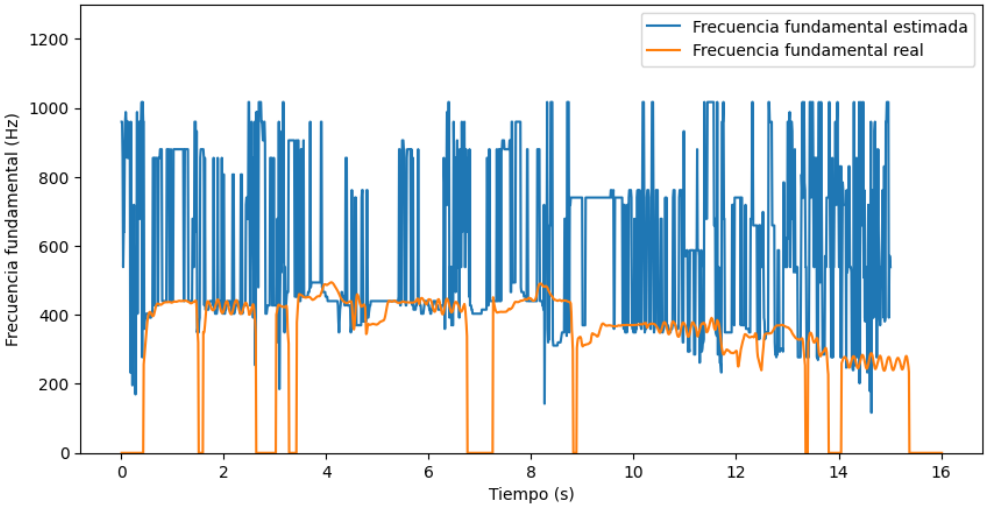


Figura 1: Estimación con GLogS de la frecuencia fundamental (azul) y frecuencia fundamental de referencia (naranja)

- Reducción de armónicos: Al seleccionar solo la primera frecuencia que supera el umbral, se evita tomar armónicos que podrían desviar la estimación de la frecuencia fundamental. Así, el pos-procesado se centra en la frecuencia básica de la señal.
- Minimización de ruido: En las zonas donde no hay actividad de la señal (es decir, cuando no hay voz), el ruido tiende a generar frecuencias de alta energía que pueden ser incorrectamente detectadas como fundamentales. El umbral del 85 % del valor máximo ayuda a evitar estas falsas detecciones en ausencia de señal real, mejorando la precisión de la estimación.

El resultado de aplicar este pos-procesado se puede observar en la figura 2. Aunque todavía existen errores en algunas zonas, la calidad de la estimación mejora considerablemente en comparación con la versión sin pos-procesado, proporcionando una representación más fiel de la frecuencia fundamental en la señal de voz.

En la figura 3 se observa el espectrograma del audio original. Se ve claramente como la señal se atenúa rapidamente luego de una frecuencia levemente mayor a $4kHz$, lo cual indica que tiene sentido utilizar $f_{max} = 5kHz$

En la figura 4 se observa la comparación entre el espectrograma del audio original y el f_0 grama realizado. Se observa como ambos presentan una señal notoria en el entorno de los $400Hz$ y otra en el entorno de los $800Hz$, indicando que el algoritmo capturo de forma razonable las características deseadas de la señal original.

3. Ejercicio 3

En este ejercicio teórico, se recuerda la condición la síntesis de la STFT discreta mediante el método Overlap-Add (OLA), de que la suma de las ventanas en el tiempo debe ser igual a una constante, y se probara que esto se cumple bajo ciertas condiciones para el caso de las ventanas de Hann.

Las ventanas de Hann, comúnmente usadas en análisis y síntesis mediante OLA, se definen para el caso de una ventana de largo 2M+1 de la manera siguiente:

$$w_{Hann}[n] = [0{,}5 + 0{,}5cos(\pi n/M)]w_r[n] \tag{13}$$

donde $w_r[n]$ es una ventana rectangular que representa el requerimiento de $w_{Hann}[n] = 0$ cuando $|n| > M$. Esta ventana podría ser de la forma

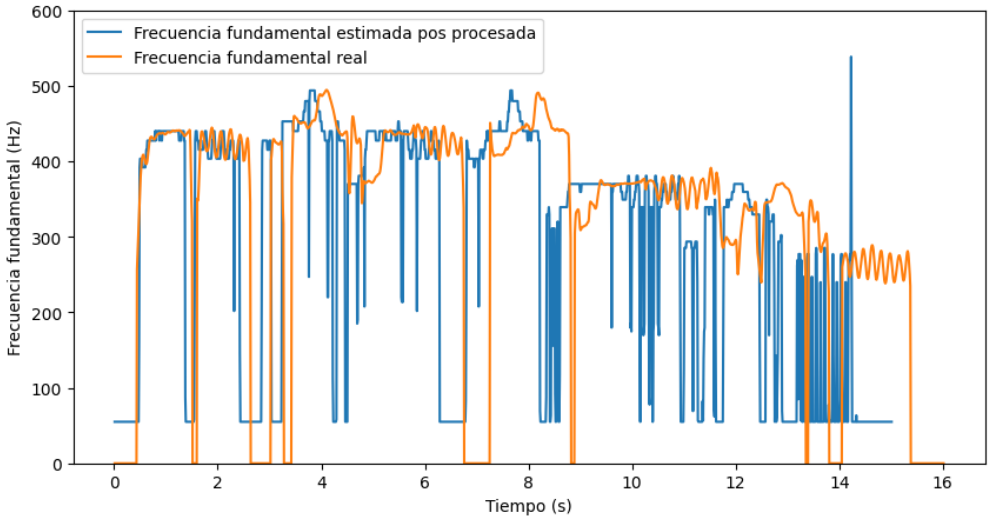


Figura 2: Estimación pos-procesada con GLogS de la frecuencia fundamental (azul) y frecuencia fundamental de referencia (naranja)

$$w_r[n] = \begin{cases} 1, & -M \leq n \leq M-1 \\ 0, & \text{en otro caso.} \end{cases} \tag{14}$$

3.1. Parte 1

Se comenzara probando que la DTFT de $w_r[n]$ es:

$$W_r(e^{j\omega}) = \left(\frac{1 - e^{-j\omega 2M}}{1 - e^{-j\omega}}\right) e^{j\omega M} \tag{15}$$

Para esto, observar que:

$$X(e^{jw}) \overbrace{=}^{(1)} \sum_{n=-\infty}^{+\infty} x[n] e^{-jw} \tag{16}$$

$$\overbrace{=}^{(2)} \sum_{n=-M}^{M-1} e^{-jwn} \tag{17}$$

$$\overbrace{=}^{(3)} \frac{e^{-jw(-M)} - e^{-jw(M)}}{1 - e^{-jw}} \tag{18}$$

$$\overbrace{=}^{(4)} \left(\frac{1 - e^{-j\omega 2M}}{1 - e^{-j\omega}}\right) e^{j\omega M} \tag{19}$$

Donde (1) es de aplicar la DTFT, (2) de imponer la forma de la ventana, (3) de utilizar la formula de la serie geométrica y (4) simplemente de reordenar los términos.

3.2. Parte 2

Ahora, se probara que:

$$W_{Hann}(e^{j\omega}) = 0{,}5 W_r(e^{j\omega}) + 0{,}25 W_r(e^{j(\omega - \pi/M)}) + 0{,}25 W_r(e^{j(\omega + \pi/M)}) \tag{20}$$

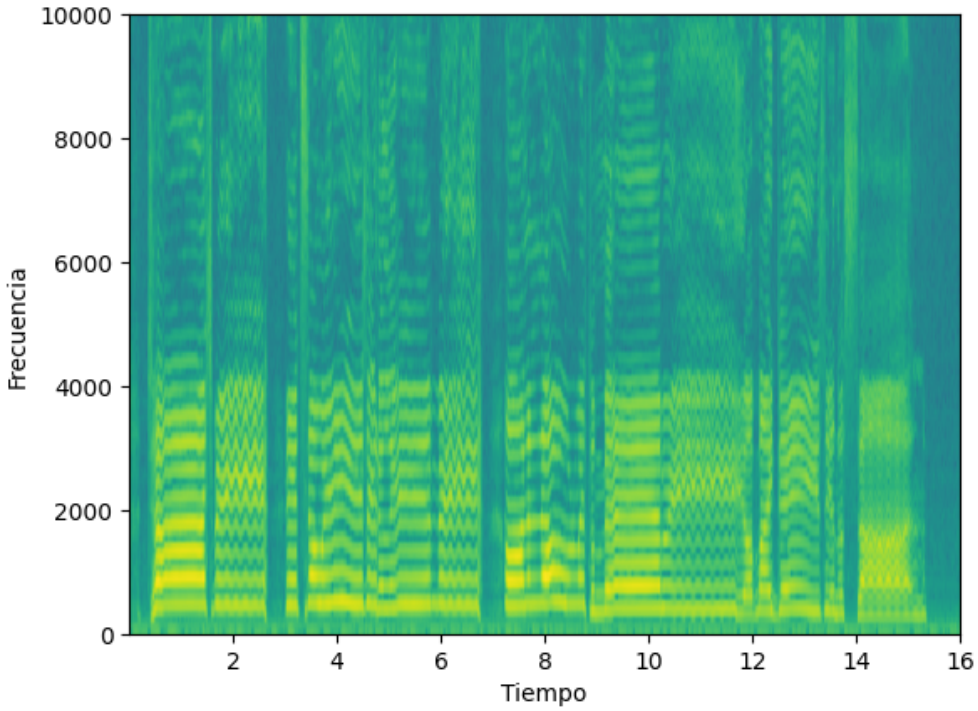


Figura 3: Espectro completo de la señal original

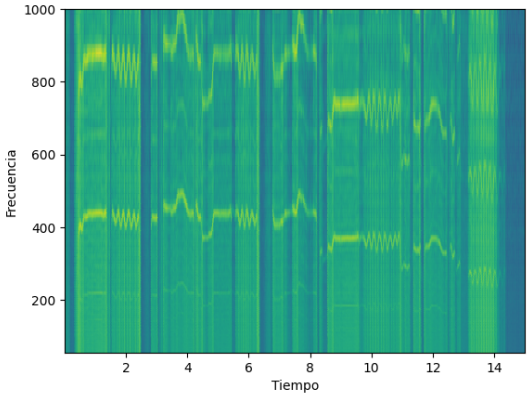
Para esto, se utilizara que la ventana de Hann sigue la siguiente ecuacion de recurrencia:

$$w_{Hann}[n] = [0{,}5 + 0{,}5cos(\pi n/M)]w_r[n] \tag{21}$$

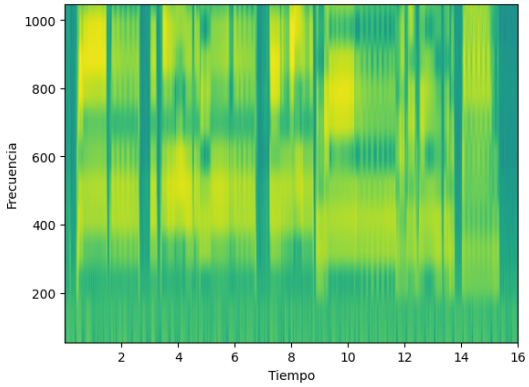
$$= \left(0{,}5 + 0{,}5\frac{1}{2}[e^{j(\pi n/M)} + ^{-j(\pi n/M)}]\right) w_r[n] \tag{22}$$

donde la segunda igualdad resulta de descomponer el coseno en exponenciales complejas. Utilizando la propiedad de que: $x[n]e^{nw_0} = X(e^{jw-w_0})$ se obtiene que

$$W_{Hann}(e^{j\omega}) = 0{,}5W_r(e^{j\omega}) + 0{,}25W_r(e^{j(\omega-\pi/M)}) + 0{,}25W_r(e^{j(\omega+\pi/M)}) \tag{23}$$



(a) f_0 grama



(b) Espectrograma

Figura 4: Espectrograma y f_0 grama

Operando a partir de esto y suplantando por la expresión calculada previamente, se puede llegar a que:

$$W_{Hann}(e^{j\omega}) = 2j\sin(\omega M)\left[\frac{1}{2}\left(\frac{1}{1-e^{-j\omega}}\right) - \frac{1}{4}\left(\frac{1}{1-e^{-j(\omega-\pi/M)}}\right) - \frac{1}{4}\left(\frac{1}{1-e^{-j(\omega+\pi/M)}}\right)\right] \tag{24}$$

3.3. Parte 3

Se demostrará que $W_{\text{Hann}}(e^{j\omega_k}) = 0$ para todo $k = 1, 2, \ldots, M-1$, donde $\omega_k = \frac{2\pi k}{M}$. Esto implica que es posible una reconstrucción perfecta si $R = M$ o $R = \frac{M}{2}$ (si $\frac{M}{2}$ es un entero), siendo R el período de muestreo (en muestras) en el tiempo de la Transformada de Fourier de Tiempo Corto (STFT).

Se comenzará observando que $W_{\text{Hann}}(e^{j\omega}) = 0$ para todo $k = 1, 2, \ldots, M-1$. Este resultado se deduce del análisis anterior, donde se observa que las raíces de $W_r(e^{j\omega})$ se encuentran en múltiplos de $\frac{\pi}{M}$ para $k = 2, 4, \ldots, 2M-2$. Al desplazar la ventana $\pm\pi$, estas raíces vuelven a coincidir, lo que conduce a la anulación en las frecuencias $\frac{2\pi k}{M}$ para $k = 1, 2, \ldots, M-1$, tal como se quería demostrar.

La condición establecida en la parte teórica para la reconstrucción perfecta es que

$$\sum_{r=-\infty}^{+\infty} w[rR - n] = C, \tag{25}$$

donde C es una constante. En este caso, w es la ventana de Hann. Se puede observar que

$$\sum_{r=-\infty}^{+\infty} w[rR - n] = \frac{1}{R}\sum_{k=0}^{R-1} W^*(e^{j(2\pi k)/R})e^{j(2\pi k)/R} \tag{26}$$

ya que $\sum_{r=-\infty}^{+\infty} w[rR-n]$ es periódica con periodo R. Por lo tanto, se puede representar como la DFT inversa de $W^*(e^{j(2\pi k)/R})$, que corresponde a la DTFT de $w[-n]$ muestreada en $\frac{2\pi k}{R}$ para $k = 0, 1, 2, \ldots, R-1$. Para lograr la reconstrucción perfecta, es necesario que $|W \ldots$

3.4. Parte 4

Se desea evaluar $W(e^{jw}) = \sum_{m=-\infty}^{+\infty} w[m]e^{-jwm}$ en $w = 0$, por lo que se quiere encontrar $W(e^{j0}) = \sum_{m=-\infty}^{+\infty} w[m]$. En este caso se tiene que

$$W(e^{j0}) = \sum_{m=-M}^{M-1}(0{,}5 + 0{,}5cos(\pi m/M)) + \sum_{m=-M}^{M-1} 0{,}5 + 0{,}5\sum_{m=-M}^{M-1}\frac{1}{2}\left(e^{j\pi/Mm} + e^{-j\pi/Mm}\right) \tag{27}$$

$$= M \tag{28}$$

donde el resultado sale de aplicar la sumatoria de geométricas y realizar cuentas.

y utilizando el resultado de la parte anterior para la reconstrucción perfecta se tiene que

$$C = \frac{W(e^{j0})}{R} = \frac{M}{R} = \frac{M}{\frac{M}{2}} = 2 \tag{29}$$

Por lo que se puede reconstruir perfectamente la señal con un factor de ganancia de 2.

4. Ejercicio 4

Ejercicio 4

En este ejercicio se implementa la técnica de phase-vocoder y se la utiliza para generar transformaciones de la señal de audio.

En la etapa de análisis se calcula la transformada de Fourier de tiempo corto, como

$$X_{n_a^u}(e^{j\omega_k}) = \sum_{m=-\infty}^{\infty} w_a[n_a^u - m]\, x[m]\, e^{-j\omega_k n}$$

en donde, $w_a[n]$ es la ventana de análisis, $\omega_k = \frac{2\pi}{N}k$, con N la cantidad de puntos de la DFT, y $n_a^u = u\, R_a$, con R_a el hop de análisis en muestras y u el índice de la trama temporal, de valor inicial 0.

En la etapa de síntesis se reconstruye la señal en el dominio del tiempo mediante la antitransformada de Fourier de cada trama temporal y el procedimiento de solapamiento y suma (overlap-add), como

$$y[n] = \sum_{u=-\infty}^{\infty} w_s[n - n_s^u] y_u[n - n_s^u]$$

con

$$y_u[n] = \frac{1}{N} \sum_{k=0}^{N-1} Y_{n_s^u}(e^{j\omega_k})\, e^{j\omega_k n}$$

en donde, $w_s[n]$ es la ventana de síntesis, y $n_s^u = u\, R_s$, siendo R_s el hop de síntesis en muestras. Notar que $y_u[n]$ es la transformada inversa de Fourier de una trama de la STFT. Cuando no hay modificaciones entre la etapa de análisis y síntesis, $Y_{n_s^u}(e^{j\omega_k}) = X_{n_a^u}(e^{j\omega_k})$ y $R_s = R_a$. En ese caso la ventana de síntesis $w_s[n]$ es opcional, pero se hace importante si se aplican modificaciones, por ejemplo cuando $R_s \neq R_a$.

4.1. Parte 1

Como se mostro anteriormente, para tener una reconstruccion perfecta es necesario elegir un R tal que $R = M$ o $R = \frac{M}{2}$. En este caso, se selecciono un $R = \frac{M}{2} = \frac{L}{4}$, donde L es el largo de la ventana en muestras. El resultado de aplicar el algoritmo con estos parametros se puede observar en la figura 5, donde se ve claramente como la señal reconstruida coincide con la original[1].

Esto se puede verificar calculando la suma solapada de las ventanas. Como se ve en la figura 6, al calcular la suma de las ventanas con los parámetros antes mencionados se obtiene una constante de valor 2, indicando que la señal reconstruida debería ser idéntica a la original con una ganancia de 2.

4.2. Parte 2

En esta sección, se introducirán modificaciones en la escala temporal mediante la variación del parámetro R durante la síntesis. Al tomar un valor de R_S igual al doble del anterior, resulta en un aumento de la escala temporal del doble. Esto se observa en la figura 7, donde la señal reconstruida (abajo) es idéntica solo que en el doble de tiempo. Sin embargo, se observan cambios auditivos notables, incluyendo la aparición de artefactos de carácter robótico". Este efecto se origina por el solapamiento de tramas en las que la fase no es coherente; en otras palabras, aunque se experimenta un salto temporal, la fase de la señal se mantiene inalterada. Como la ventana empleada es lo suficientemente amplia, la percepción de la altura de la señal se conserva, lo que permite seguir reconociendo las características tonales del audio a pesar de los cambios introducidos.

Al disminuir el valor de R_S a la mitad, sucede lo análogo. La escala temporal se ve reducida a la mitad, y, si bien la estructura del audio es la misma se mantienen los artefactos provocados por las incoherencias

[1] Acá no me quedo algo claro. Si el factor de ganancia es 2, no me tuvo que haber quedado el doble de amplitud?

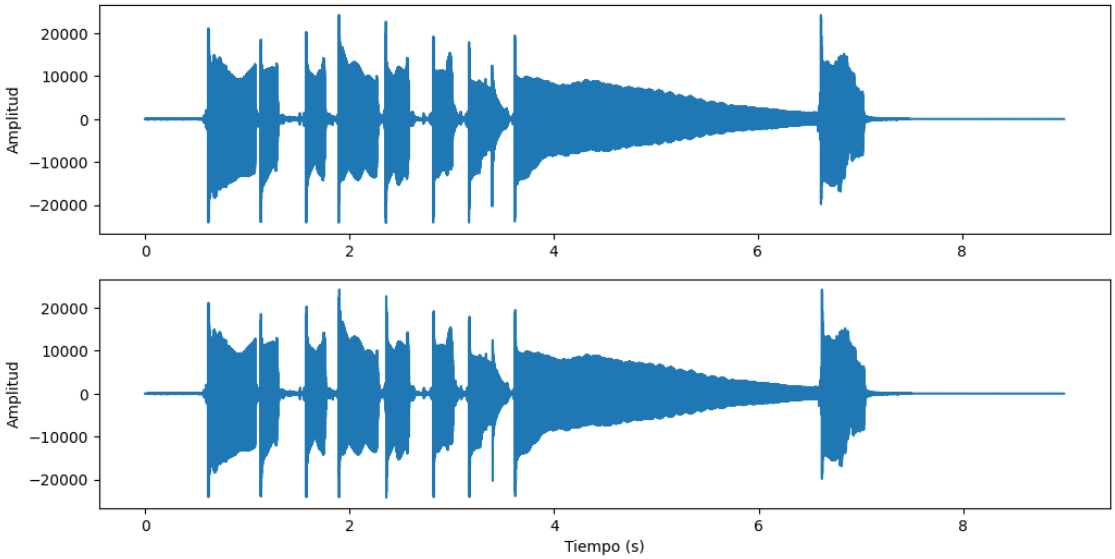


Figura 5: Señal original (arriba) y señal luego del análisis y síntesis (abajo)

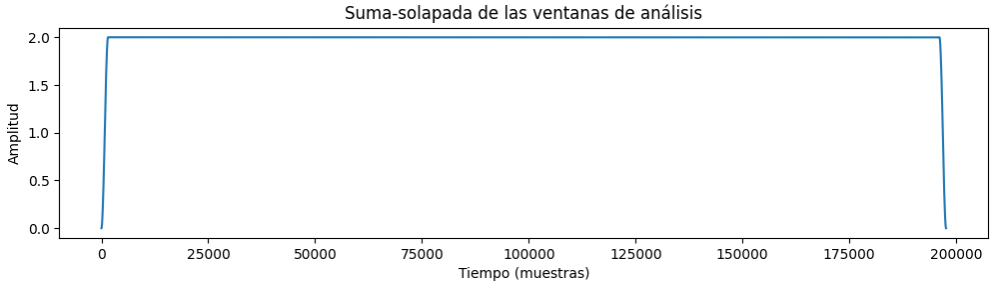


Figura 6: Suma Solapada de la ventana resultante

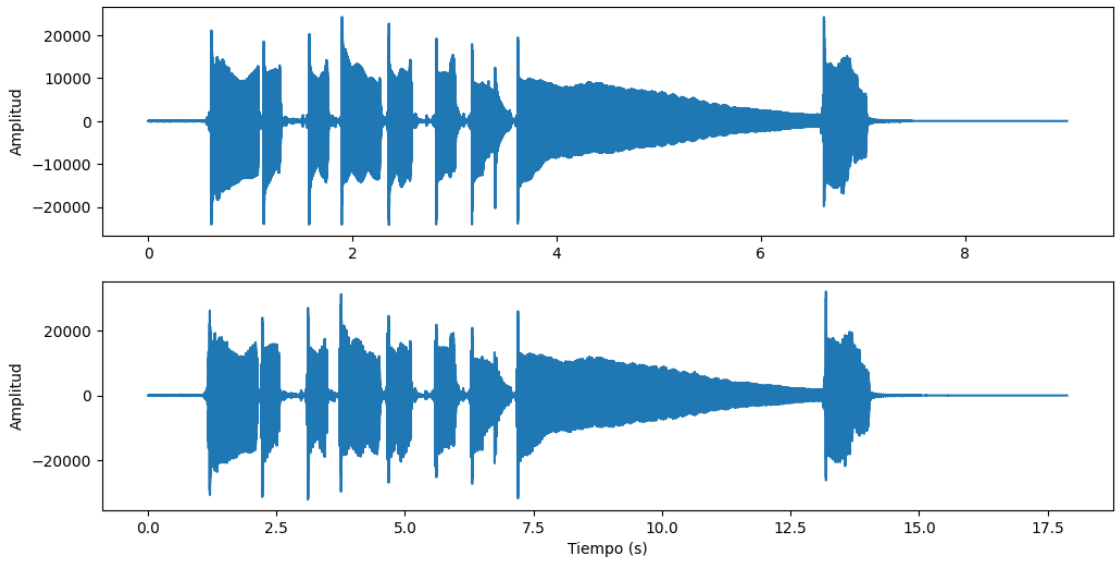


Figura 7: Señal original (arriba) y señal luego del análisis y síntesis (abajo) con doble R_S

de fase.

4.3. Parte 3

Para evitar la mayoría de los problemas introducidos debido a la inconsistencia de fase, se utilizara el procedimiento de desdoblamiento de fase (phase unwrapping).

Asumiendo que existe un solo componente sinusoidal por bin de la DFT, se puede plantear las siguientes ecuaciones para estimar la fase de $Y_{n_s^u}(e^{j\omega_k})$, cuando se transforma la escala temporal utilizando un hop de síntesis $R_s \neq R_a$.

Se calcula el incremento de fase heterodino, a partir del incremento de fase de tramas sucesivas

$$\Delta\Phi_k^u = \angle X_{n_a^u}(e^{j\omega_k}) - \angle X_{n_a^{u-1}}(e^{j\omega_k}) - R_a\,\omega_k \tag{30}$$

Notar que el término $R_a\ \omega_k$ es el incremento de fase que cabría esperar si la frecuencia del componente sinusoidal correspondiera exactamente a la frecuencia de análisis.

Se toma el argumento principal de $\Delta\Phi_k^u$ entre $(-\pi, \pi)$, denominado $\Delta_p\Phi_k^u$.

Luego se calcula la estimación de la frecuencia instantánea

$$\hat{\omega}_k[n_a^u] = \omega_k + \frac{1}{R_a}\,\Delta_p\Phi_k^u \tag{31}$$

Finalmente se calcula la fase de $Y_{n_s^u}(e^{j\omega_k})$ utilizando la fórmula de propagación de fase

$$\angle Y_{n_s^u}(e^{j\omega_k}) = \angle Y_{n_s^{u-1}}(e^{j\omega_k}) + R_s\,\hat{\omega}_k[n_a^u] \tag{32}$$

Luego de aplicado el algoritmo, el resultado mejora considerablemente respecto a la parte anterior. Auditivamente, se dejan de escuchar los artefactos de altas frecuencias antes mencionados. Luego, cuando se paso a analizar los espectrogramas resultantes, se observo el comportamiento de la figura 9. El ruido disminuye de forma considerable al aplicar esta técnica, logrando mantener de forma mas clara la señal original.

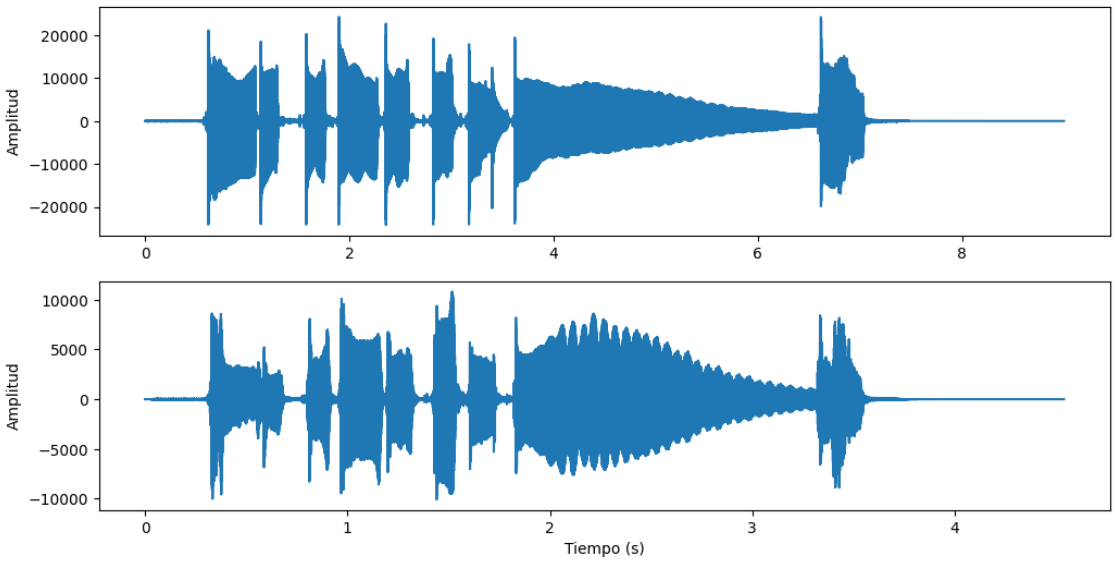
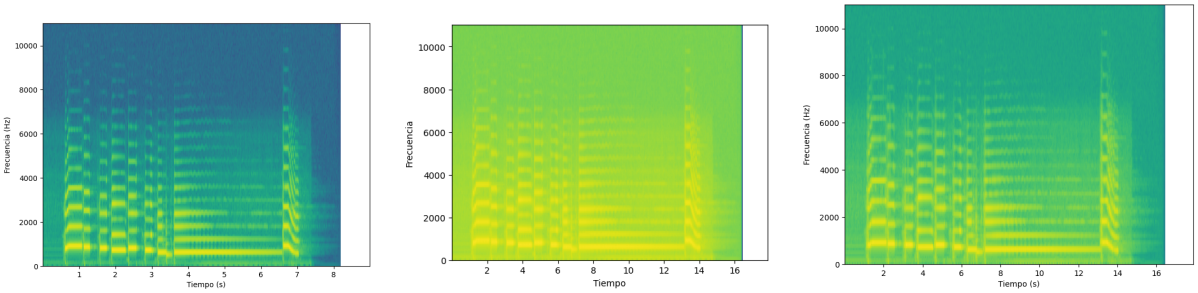


Figura 8: Señal original (arriba) y señal luego del análisis y síntesis (abajo) con mitad R_S



(a) Espectrograma Original (b) Sin Phase Unwrapping (c) Con Phase Unwrapping

Figura 9: Espectrogramas para señal real y reconstruidas

4.4. Parte 4

Se utilizo el *phase-vocoder* para cambiarle el *pitch* a una señal de audio. Al analizar el audio obtenido, se nota claramente como cambia el sonido, volviéndose mas agudo, indicando que el *phase-vocoder* fue correctamente implementado.

Por ultimo, se utilizo esta herramienta para generar un efecto de coro en la señal. Esto se realiza mediante varios cambios de *pitch* cercanos, agregando un pequeño desfajase temporal para darle mas naturalidad al coro.