Universidad de la República

Facultad de Ingeniería

Procesamiento Digital de Señales de Audio

Práctica 3

Autor:
Julian O'Flaherty

30 de mayo de 2022

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Índice

1. Ejercicio 1 - Relación entre SFT y autocorrelación de tiempo corto

Se estudiará la relación entre la short time fourier transform (STF) y función auto-correlación de tiempo corto. Se define la densidad espectral de potencia de tiempo corto de una señal como:

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

y la autocorrelación de tiempo corto como:

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

En la ecuación 1 se define la densidad espectral de potencia en función de la SFT de la señal:

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

por lo que podemos reescribir la ecuación como:

$$S_n(e^{jw}) = X_n(e^{jw})X_n^*(e^{jw}) = \sum_{m=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} x[m]w[n-m]e^{-jwm}x^*[l]w^*[n-l]e^{jwl}$$

Como se trabaja con señales reales:

$$S_n(e^{jw}) = \sum_{m=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} x[m]x[l]w[n-m]w[n-l]e^{jw(l-m)}$$

Tomando $k = m - l$:

$$S_n(e^{jw}) = \sum_{m=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} x[m]x[m-k]w[n-m]w[n-m+k]e^{-jwk} = \mathcal{F}\{R_n[-k]\}(e^{jw})$$

La auto-correlación es una función par, por lo $R_n[k] = R_n[-k]$. Llegamos entonces a:

$$S_n(e^{jw}) = \mathcal{F}\{R_n[k]\}(e^{jw}) \tag{4}$$

2. Ejercicio 2 - espectro logarítmico acumulado

2.1. Producto armónico espectral

El producto armónico espectral (HPS) se define como:

$$P_n(e^{jw}) = \prod_{r=1}^{K}\left|X_n(e^{jwr})\right|^2 \tag{5}$$

En esencia, esta fórmula hace el producto de varios espectros comprimidos en valores enteros en frecuencia, por lo que permite comparar la energía que carga un frecuencia y sus armónicos. ¿Cómo se puede usar para detectar frecuencias fundamentales? Suponiendo una señal monofónica, la frecuencia fundamental y sus armónicos es donde se acumula la mayor parte de la energía de la señal, por lo que el producto armónico espectral tenga un máximo en esa frecuencia.

Podemos tomar el logaritmo de esta expresión para simplificar el cálculo y el computo necesario, manteniendo la propiedad antes mencionada, obteniendo el espectro logarítmico acumulado(LHPS):

$$\hat{P}_n(e^{jw}) = 2\sum_{r=1}^{K} \log \left| X_n(e^{jwr}) \right| \tag{6}$$

Como el HPS y el LHPS tienen en cuenta la energía de los armónicos además de la de la frecuencia fundamental, en caso de perderse la frecuencia fundamental (un filtrado pasa-altos por ejemplo) aún se mantendrá el máximo antes mencionado. Esto permite estimar la frecuencia fundamental aún cuando no es parte de la señal a estudiar.

2.2. Espectro logarítmico acumulado

A partir de la ecuación 6, podemos definir el espectro logarítmico acumulado como el promedio de la magnitud espectro en escala logarítmica de una frecuencia y sus armónicos sucesivos,

$$\rho_n(f_0) = \frac{1}{n_H} \sum_{i=1}^{n_H} \log |X_n(if_0)| \tag{7}$$

donde $n_H = \lfloor \frac{f_{max}}{f_0} \rfloor$ es la cantidad de armónicos menores a f_{max} que tiene la frecuencia f_0, siendo f_{max} el tope de frecuencia considerado para el cálculo.

De esta forma, la función alcanza el máximo cuando se la evalúa en la frecuencia fundamental. Los sub-armónicos, pese a sumar más frecuencias, se ven divididos por un n_H mayor. Los armónicos mayores, están divididos por un n_H menor, pero suman menos frecuencias, por lo que siguen siendo menor que el valor del gLogs para la frecuencia fundamental.

Podemos entonces utilizar el GLogS para estimar el tono de la señal, evaluando la gLogs para distintas frecuencias en cada trama de una señal y tomando la frecuencia donde se da el máximo[1]. Se toma un grilla con frecuencias en la escala temperada entre $55Hz$(A1) y $1046{,}5Hz$(C6) con un paso de cuarto tono para evaluar la GLogS.

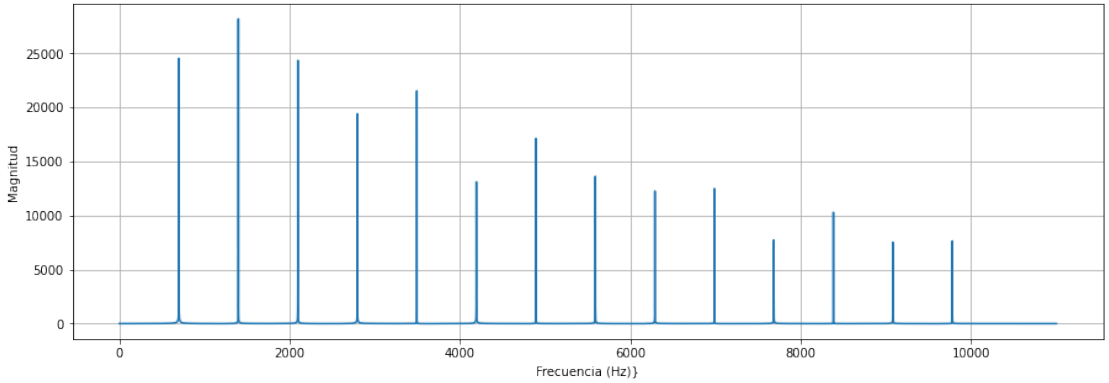
Insertando un tono sintético de frecuencia $f_0 = 698{,}46Hz$, con el espectro mostrado en la figura 1a. El resultado del GLogS se muestra en al figura 1b. Se observa como se tienen varios picos en frecuencias armónicas, siendo el máximo absoluto en $f \approx 698{,}46Hz$. No es exactamente esta por la apreciación de la grilla utilizada. La apreciación de la grilla tiene un segundo problema cuando se usa una frecuencia que no esta cerca de los valores de la grilla, especialmente cuando se tiene un valor sintético donde los picos en frecuencia aproximan mucho una delta.

2.3. Cálculo del f0-grama

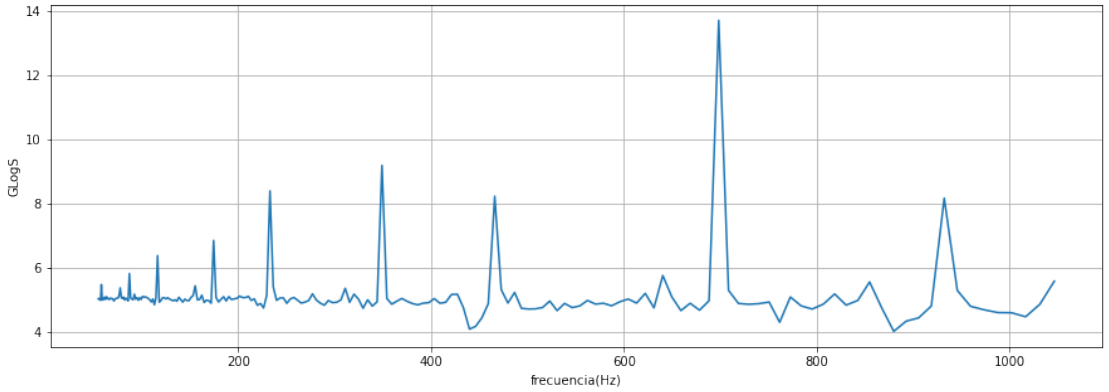
Utilizando la función GLogS diseñada en la parte anterior se puede realizar un f0-grama. Haciendo la STFT obtenemos un espectro para cada trama de la señal de audio. Pasando ese espectro por el GLogS con la grilla antes descrita obtenemos una estimación de la energía que tiene el espectro para cada frecuencia y sus armónicos.

El resultado de calcular el f0-grama del audio se muestra en la figura 2. Se observa como la frecuencia fundamental y sus armónicos aparecen bien marcadas, y se observan "fantasmas.[en] sub-armónicos de la frecuencia (llamamos fantasmas a marcas con la misma forma pero con menor trazo).

[1]Se verá más adelante que no basta solo con esto, sino que hay que agregar un pos-procesado a la estimación.



(a) Espectro de la señal sintética



(b) Glogs señal sintética

Figura 1: Estudio del GLogS implementado para una señal sintética

2.4. Estimación de f_0

En la sección anterior se explicó como obtener el f0-grama de la señal a estudiar. A partir de este f0-grama se puede estimar la frecuencia fundamental de la señal en cada trama. Un acercamiento inicial sería tomar la frecuencia donde se da el máximo, pero el resultado obtenido es muy malo (figura 3).

Se tomaron 3 decisiones para mejorar la estimación de f_0:

1. **Reducir grilla f_0:** reducir el tamaño de la grilla evita el problema de tener máximos en armónicos mayores donde el valor de n_H es bajo.

2. **Estimar cero cuando la desviación estándar de GLogS es pequeña:** esto resuelve el problema cuando el audio tiene sonidos sordos o silencios. Cuando no se tiene una frecuencia fundamental clara los valores del GLogS son aproximadamente constantes para las frecuencias, por lo que se puede asumir que si el desvío estándar es menor a 0.11[2] estamos ante una frecuencia fundamental nula.

3. **Filtro de mediana movil sobre resultado:** aplicando un filtro de mediana móvil con

[2]este resultado es halló experimentalmente, no tiene una deducción teórica

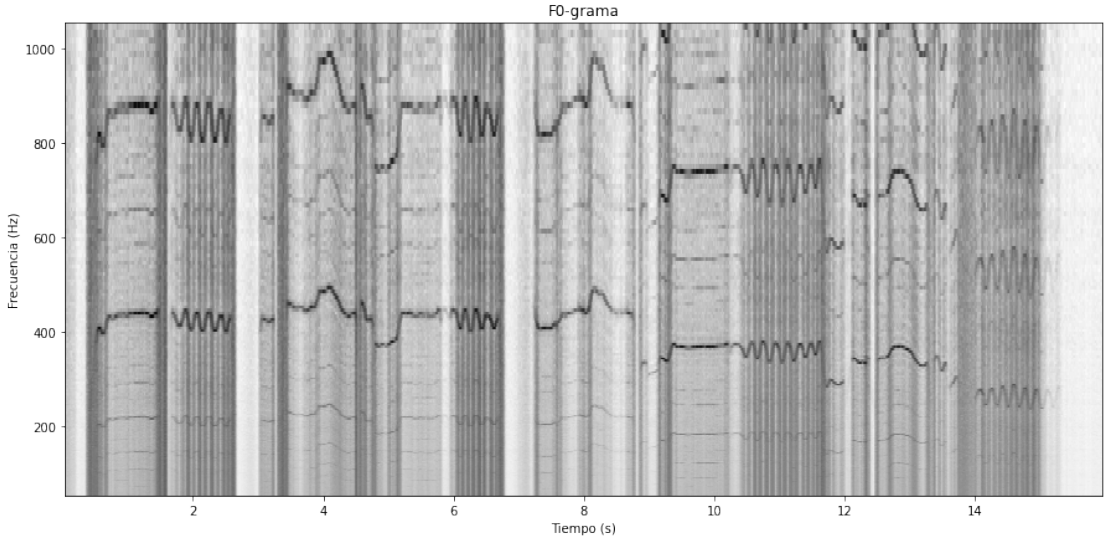


Figura 2: f0-grama del audio a estudiar

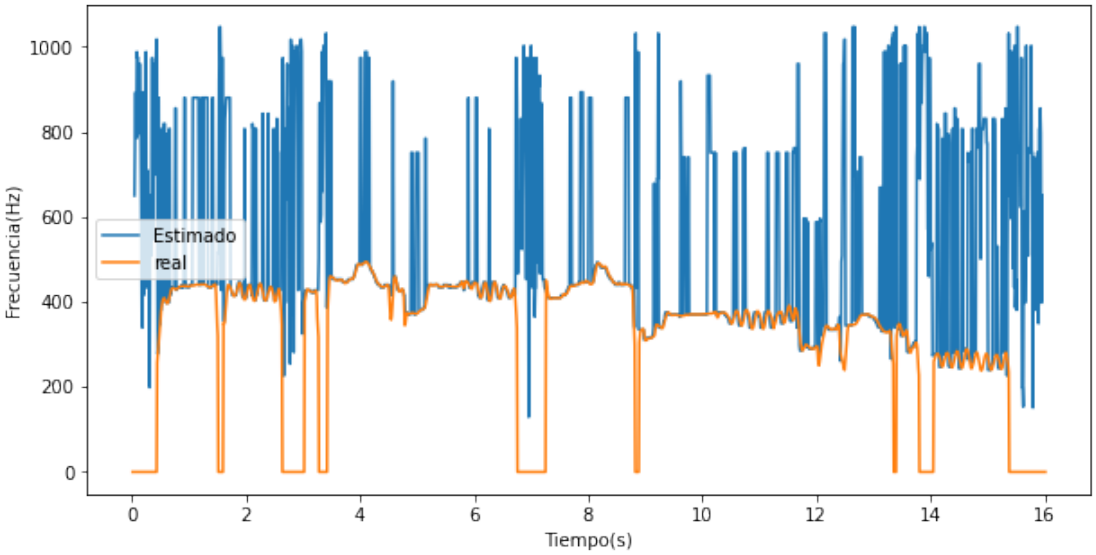


Figura 3: Estimación de f_0 tomando el máximo del GLogS para cada trama

kernel 5 se pueden eliminar los picos hacia arriba que se ven sobre el final, actuando como un filtro pasabajos que preserva los cambios[3].

El resultado de estos cambios se muestra en la figura 4, donde se observa una gran mejoría de la estimación. Se siguen teniendo algunos picos, pero aumentar el orden del filtro lleva a que se pierda detalle en algunas zonas, además que aumenta el retardo y empieza a ser notorio (ya se puede observar un pequeño retardo pero a efectos de esta aplicación no impacta negativamente).

[3]Este filtro introduce un retardo pero es despreciable en esta aplicación

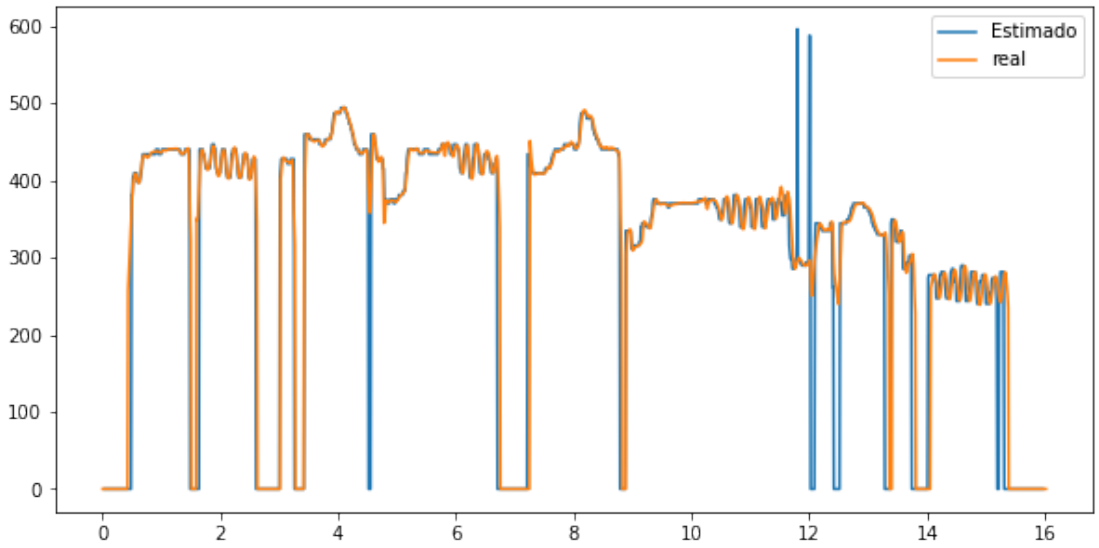


Figura 4: estimación de f_0 con el procesamiento extra.

3. Ejercicio 3 - Síntesis STFT

Se estudiará la síntesis mediante Overlap-Add(OLA), utilizando una ventana de Hann. Definimos la ventana de Hann de largo 2M como:

$$w_{Hann}[n] = (0{,}5 + 0{,}5\cos(\pi\frac{n}{M}))w_r[n] \tag{8}$$

donde $w_r[n]$ es una ventana rectangular que hace que $w_{Hann}[n] = 0$ cuando $|n| > M$.

$$w_r[n] = \begin{cases} 1, & -M \leq n \leq M-1 \\ 0, & \text{en otro caso} \end{cases} \tag{9}$$

3.1. DTFT de $w_r[n]$

Para hallar la DTFT lo hacemos con su definición.

$$\mathcal{F}(e^{jw}) = W_r(e^{jw}) = \sum_{n=-\infty}^{\infty} w_r[n]e^{-jwn} \overset{(9)}{=} \sum_{n=-M}^{M-1} e^{-jwn} = e^{jwM}\sum_{n=0}^{2M-1}(e^{-jw})^n$$

La suma geométrica parcial es $\sum_{n=0}^{N} x^n = \frac{1+x^{N+1}}{1-x}$ con $|x| > 0$. Por lo tanto, la DTFT de la ventana rectangular es:

$$W_r(e^{jw}) = \begin{cases} \left(\frac{1-e^{-jw2M}}{1-e^{-jw}}\right)e^{jwM} & \text{si } |w| > 0 \\ 2M & \text{si } w = 0 \end{cases} \tag{10}$$

3.2. DTFT ventana de Hann

Podemos reescribir la ecuación 8 como:

$$w_{Hann}(e^{jw}) = (0{,}5 + 0{,}25(e^{jw\pi\frac{n}{M}} + e^{-jw\pi\frac{n}{M}}))w_r[n]$$

Recordando que $\mathcal{F}\{e^{jwnN}x[n]\} = X(e^{j(w-N)})$, al tomar la DTFT de la ventana de Hann obtenemos:

$$W_{Hann}(e^{jw}) = 0{,}5W_r(e^{jw}) + 0{,}25W_r(e^{j(w-\frac{\pi}{M})}) + 0{,}25W_r(e^{j(w+\frac{\pi}{M})})$$

Sustituyendo con la DTFT de la ventana rectangular hallada en 10:

$$W_{Hann}(e^{jw}) = 0{,}5\left(\frac{1-e^{-jw2M}}{1-e^{-jw}}\right)e^{jwM} - 0{,}5\left(\frac{(1-\cos(\frac{\pi}{M}))e^{-jw}\left(1-e^{-jw(2M)}\right)}{1-\cos(\frac{\pi}{M})e^{-jw}+e^{-j2w}}\right)e^{jwM} \quad (11)$$

3.3. Condición para reconstrucción perfecta

A diferencia de la DTFT, en la STFT la tranformada de fourier también es discreta, por lo que además de un muestreo temporal se tiene un muestreo frecuencial. Tenemos entonces que $w_k = \frac{\pi k}{M}$ para $k \in \mathbb{N} \cap [1, M-1]$ es el muestreo en frecuencia de la DFT, donde M es el la mitad del largo de la ventana.

$$\begin{aligned} W_{Hann}(e^{jw_k}) &= 0{,}5\left(\frac{1-e^{-j2\pi k}}{1-e^{-j\frac{\pi k}{M}}}\right)e^{j\pi k} - 0{,}5\left(\frac{2(1-\cos(\frac{\pi}{M}))e^{-j\frac{\pi k}{M}}\left(1-e^{-jw(2\pi k)}\right)}{1-\cos(\frac{\pi}{M})e^{-j\frac{k\pi}{M}}+e^{-j\frac{2\pi k}{M}}}\right)e^{j\pi} \\ &= (1-e^{-j2\pi k})\left[\left(\frac{0{,}5}{1-e^{-j\frac{\pi k}{M}}}\right)e^{j\pi k} + 0{,}5\left(\frac{2(1-\cos(\frac{\pi}{M}))e^{-j\frac{\pi k}{M}}}{1-\cos(\frac{\pi}{M})e^{-j\frac{k\pi}{M}}+e^{-j\frac{2\pi k}{M}}}\right)\right] \\ &= 0 \end{aligned}$$

Es decir:

$$W_{Hann}\left(e^{j\frac{\pi k}{M}}\right) = 0 \qquad \forall k = 1, 2, \ldots, M-1 \quad (12)$$

La ecuación de síntesis para el OVA es con un decimador de factor R es:

$$y[n] = \sum_{r=-\infty}^{\infty}\left(\frac{1}{2M}\sum_{k=0}^{2M-1}X_{Rr}(e^{jw_k})e^{jw_kn}\right) = x[n]\sum_{r=-\infty}^{\infty}w[rR-n] = x[n]\tilde{w}[n]$$

Se requiere que $y[n] = Cx[n] \Rightarrow \sum_r w[rR-n] = C$. La constante multiplicativa se puede eliminar dándole ganancia C^{-1} a la ventana o como un bloque de ganancia en cascada a la reconstrucción. Se puede representar $\tilde{w}[n]$ como la inversa de la DFT de la ventana de Hann de largo L (muestreo en frecuencia de la DTFT).

$$\tilde{w}[n] = \sum_{r=-\infty}^{\infty}w[rR-n] = \frac{1}{R}\sum_{k=0}^{R-1}W^*_{Hann}\left(e^{j\frac{2\pi k}{R}}\right)e^{j\frac{2\pi k}{R}n}$$

donde W^*_{Hann} es la DTFT de la ventana reflejada ($w_{Hann}[-n]$) muestreada en frecuencia. Aplicando el resultado obtenido en 12, si tomamos $R = \frac{2M}{i}$ con $i = 1, 2, \ldots, 2M-1$ que sean enteros.

$$\tilde{w}[n] = \frac{i}{2M}\sum_{k=0}^{R-1}W^*_{Hann}\left(e^{j\frac{\pi ki}{M}}\right)e^{j\frac{\pi ki}{M}n} = \frac{i}{2M}W_{Hann}(e^0) = \frac{W_{Hann}(e^j0)}{R} = C \quad (13)$$

3.4. Calculo $W_{Hann}(e^{j0})$

Evaluando $w = 0$ en la ecuación 11 obtenemos:

$$W_{Hann}(e^{j0}) = 0{,}5W_r(e^{j0}) + 0{,}25W_r(e^{j(0-\frac{\pi}{M})}) + 0{,}25W_r(e^{j(0+\frac{\pi}{M})}) = 0{,}5W(e^{j0}) = M$$

Por lo tanto, sustituyendo en la ecuación 13:

$$\frac{W_{Hann}(e^{j}0)}{R} = \frac{M}{R} = C \tag{14}$$

4. Ejercicio 4 - Phase Vocoder

4.1. Reconstrucción

En el ejercicio anterior se demostró la condición que deben cumplir la ventana de análisis para una reconstrucción perfecta, suponiendo que $R_a = R_s$. Aplicando una ventana de largo 2048 muestras y un hop de 256 ($\frac{L}{8}$), el resultado de la reconstrucción es auditiva mente perfecto. En la figura 5 se muestran el espectrograma original y el de la reconstrucción, donde se aprecia la reconstrucción perfecta.

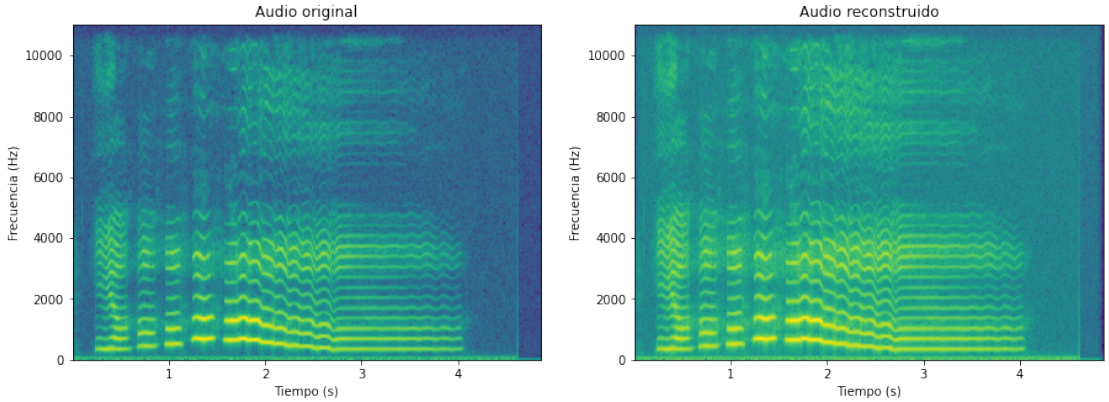
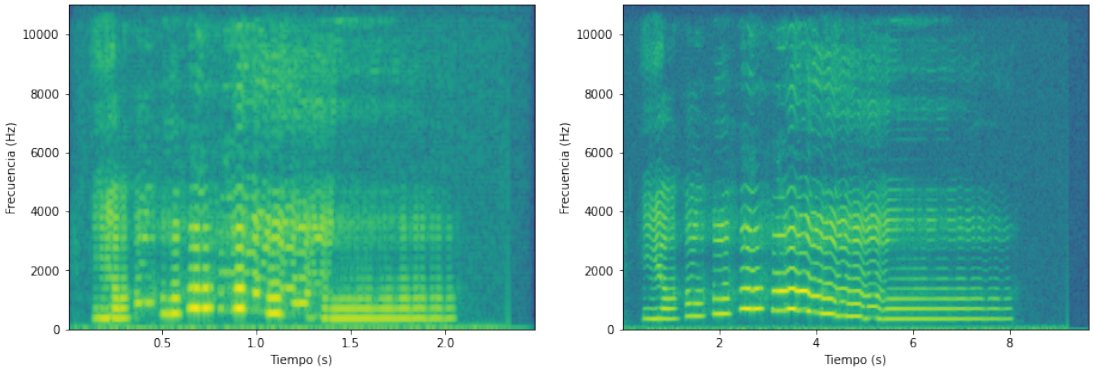


Figura 5: comparación espectrograma original y reconstruido

4.2. Reconstrucción con diferente ventana de síntesis

Si tomamos una ventana de síntesis distinta a la utilizada en el análisis resulta en una modificación de la señal de audio. Si aumentamos la el largo de la ventana de síntesis ($R_s > R_a$), obtenemos una señal de mayor duración en muestras pero la mismas características, por lo que se hace un dilatamiento temporal de la señal. Si la ventana de síntesis es menor, el resultado es una señal de menor duración, por lo que se tiene una compresión temporal. En la figura 6 se observan los espectrogramas resultantes de la reconstrucción del audio con un tamaño de síntesis distinto al de análisis, con L 2048 y R_a 256.

Desde un punto de vista auditivo, pareciera que el audio acelerado sube su tono, mientras que el enlentecido disminuye. Sin embargo, se escuchan unos artefactos, una especie de sonido metálico, a causa de que no se tiene encuentra la fase en la reconstrucción. En el espectrograma del audio acelerado (figura 6a) se observa como la frecuencia fundamental esta de a tramos.



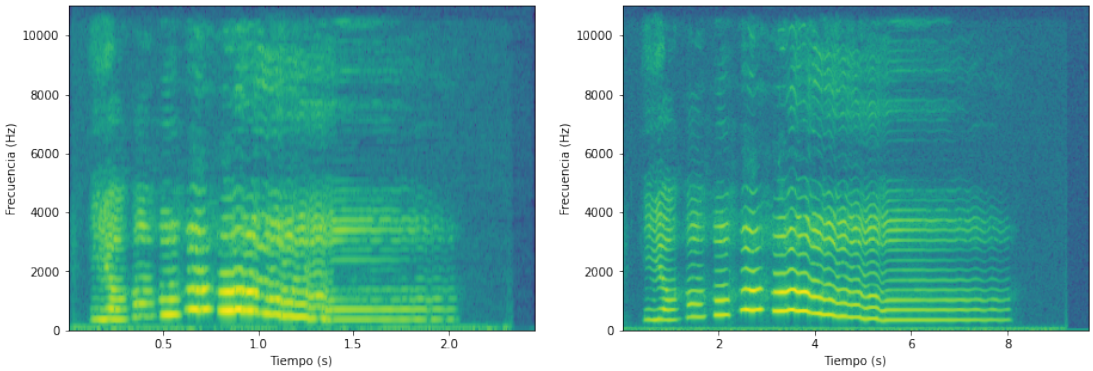
(a) reconstrucción con $R_s = 128$ (b) reconstrucción con $R_s = 512$

Figura 6: resultados de la reconstrucción con ventana de análisis $L = 2048$ y $R_a = 256$

4.3. Phase-vocoder

Para estimar la fase se sigue la estimación presentada en [A. Gotzen and Arfib, 2000] y [Laroche and Dolson, 1999]. El método presentado permite hacer una estimación de la fase para cada valor de la STFT, lo cual mejora notablemente el resultado auditivo de la señal. Como para dar idea, La estimación se basa en incremento de fase de una trama a la siguiente y la deformación que introduce la diferencia de las ventanas.

Aplicando el mismo análisis que en la parte anterior, en la figura 7 se realiza la reconstrucción que en la figura 6 utilizando el phase-vocoder.



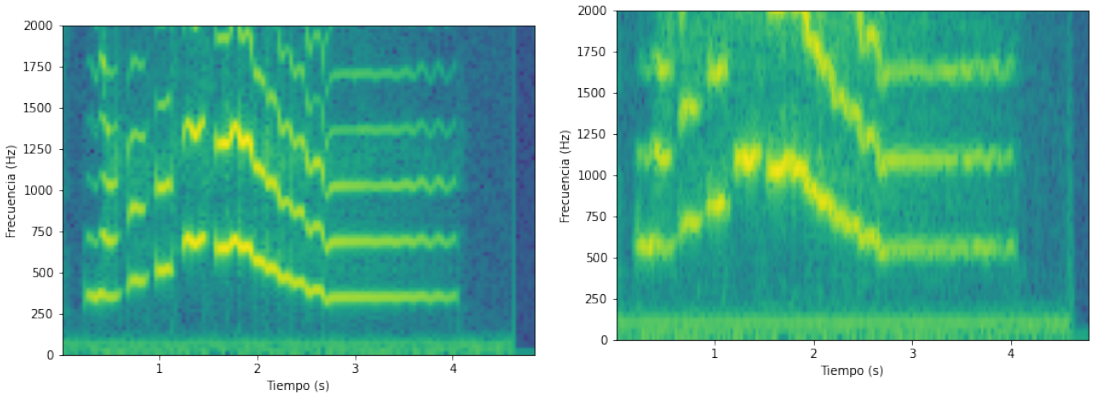
(a) reconstrucción con $R_s = 128$ (b) reconstrucción con $R_s = 512$

Figura 7: resultados de la reconstrucción con phase-vocoder con ventana de análisis $L = 2048$ y $R_a = 256$

Se observa que ahora el espectro es más continuo y con menos ruido armónico, en particular para la ventana de menor tamaño (figuras 7a y 6a). En la ventana de mayor tamaño se observa una mejoría también, manteniéndose la potencia de la frecuencia fundamental (fig 7b) en vez de oscilar como en el caso anterior (fig 6b).

4.4. Transposición en frecuencia

Se puede aprovechar los efectos antes estudiados para hacer transformaciones deseables sobre la señal. En particular, podemos hacer un desplazamiento en frecuencia cambiando el tamaño de la ventana de síntesis y cambiando la frecuencia de muestreo de la señal para obtener que mantenga el largo original.



(a) audio original

(b) audio subido aumentado 8 semitonos (escala temperada)

Figura 8: resultado de la transposición en frecuencia.

En la figura 8 se muestra el resultado de la transposición en frecuencia, subiendo 8 semitonos la señal de estudio. Se redujo el espectro observado a las frecuencias 0Hz a 2000Hz para que pueda observarse el resultado. Una transposición de 8 semitonos equivale a un aumento de las frecuencias por un factor de $2^{\frac{2}{3}} \approx 1{,}6$.

Este efecto se puede aprovechar para generar efectos digitales sobre el audio. Por ejemplo, podemos crear un armonizador de 5^{ta} sumando a la señal original la señal traspuesta una 5ta. Para que estén a la misma frecuencia de muestreo se debe sub-muestrear la señal transpuesta.

El resultado del armonizador se muestra en la figura 9. Se observa como se tienen nuevas componentes armónicas en el espectro resultantes de la señal transpuesta una quinta.

Otro efecto que se puede generar es el chorus, que simula tener un coro de personas cantando. Para esto se suma a la señal muchas realizaciones transpuestas menos de un cuarto de semitono, que emula muchas personas cantando una misma canción con ciertas imperfecciones.

En el espectrograma resultante (figura 10) se observa que las bandas de frecuencias son un poco más gruesas, y en altas frecuencias se logra distinguir las frecuencias que la componen más claramente.

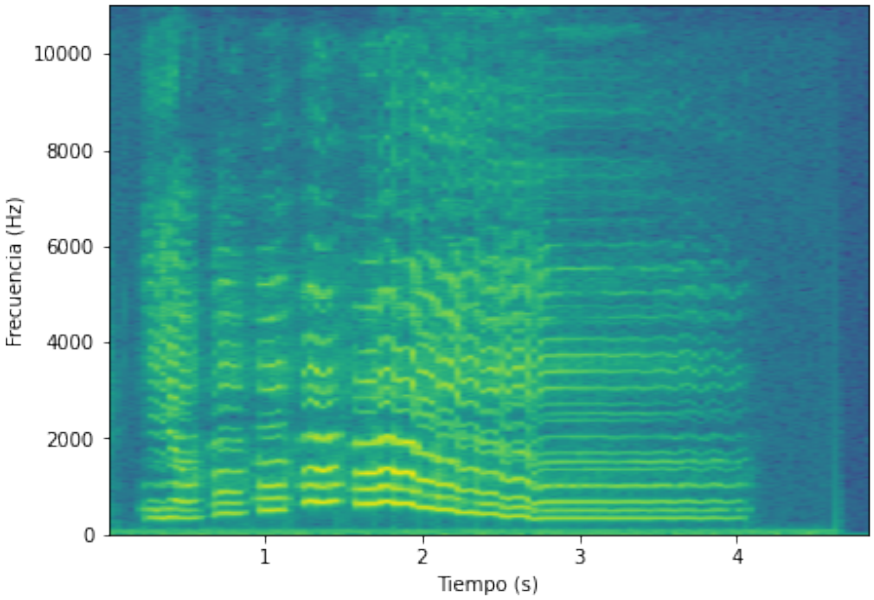


Figura 9: audio armonizado con la señal transpuesta una quinta

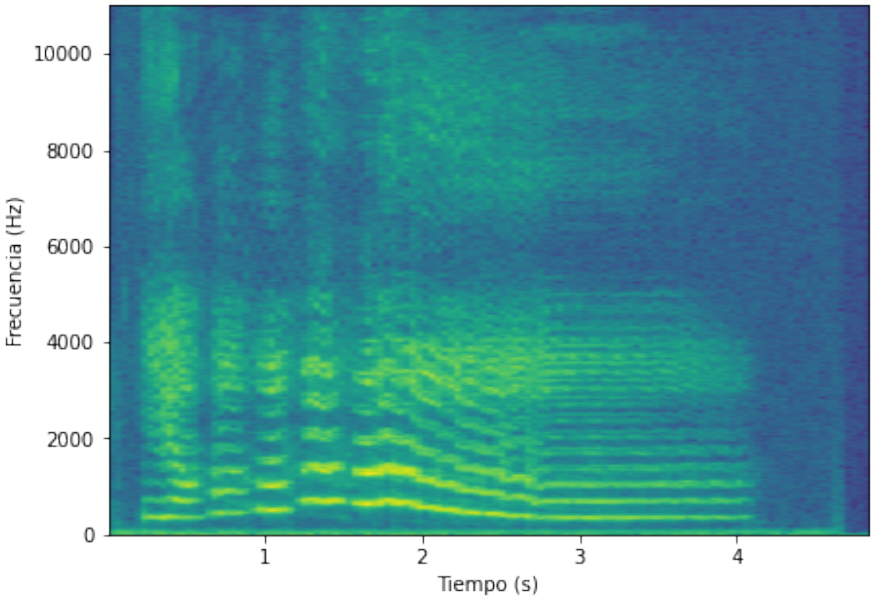


Figura 10: espectro del audio con el efecto del chorus

Referencias

[A. Gotzen and Arfib, 2000] A. Gotzen, N. B. and Arfib, D. (2000). Traditional implementations of a phase-vocoder: The tricks of the trade. *International Conference on Digital Audio Effects*,

pages 286–295.

[Laroche and Dolson, 1999] Laroche, J. and Dolson, M. (1999). Improved phase vocoder time-scale modification of audio. *IEEE Transactions on Speech and Audio processing*, 7(3):323–332.