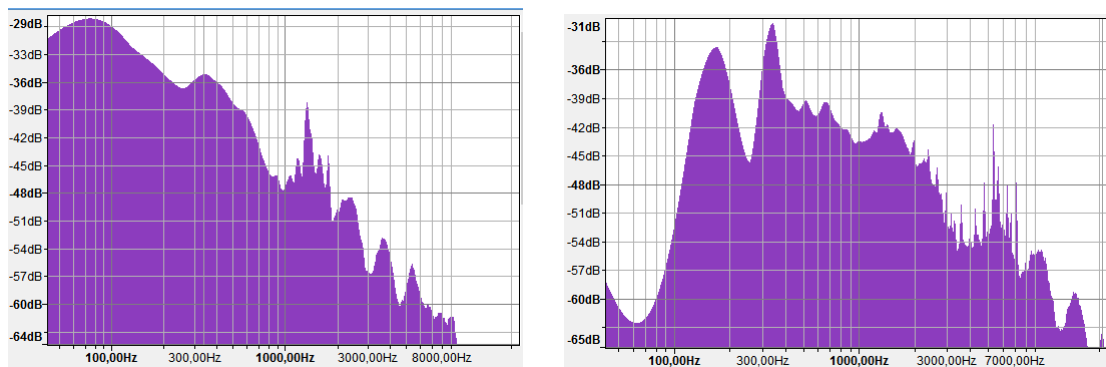


## 0.1. Síntesis por muestras

La Síntesis por muestras tiene sus orígenes en los estudios de grabación. Previo a las técnicas de Síntesis se debían grabar nuevamente aquellos sonidos que no satisfacían por completo a los músicos o a los ingenieros de sonido obligándolos a utilizar más horas el estudio de grabación lo cual, por supuesto, llevaba varios gastos asociados y retrasaba la culminación del proyecto.

### 0.1.1. Time Stretching

Supongamos que deseamos grabar un comercial de televisión y por cuestiones legales debemos añadir una pequeñas aclaraciones al final del mismo. Sin embargo, el tiempo de aire es altamente costoso entonces tenemos que poder incluir toda la información necesaria en un espacio de tiempo muy pequeño. Entonces podríamos suponer que basta con grabar el mensaje a velocidad de habla normal y luego reproducirlo más rápido para que ocupe menos tiempo. Es decir reproducir el sonido salteándose segmentos para poder acortar la duración.



(a) Espectro de un discurso hablado a velocidad normal (b) Espectro del discurso con una velocidad 4 veces mayor

Figura 1: Consecuencias de la compresión temporal

Como se puede observar, la figura 1a representa el espectro de habla de una persona hablando a una velocidad promedio. Notamos que la zona de mayor contenido armónico se halla aproximadamente entre los  $80\text{Hz}$  y  $200\text{Hz}$ . Sin embargo, si miramos al espectro del mismo discurso pero ahora con 4 veces menos duración, figura 1b, podemos observar que el espectro no se preserva. De hecho se puede notar que hay mayor potencia espectral a frecuencias más altas que en el espectro original. Esto se traduce en un sonido chillón que poco nos recuerda al discurso original. De forma análoga si por alguna razón se quisiera aumentar o reducir el pitch de una pista de audio se debería recurrir al método de acortar la pista o en el caso contrario alargarla. En la siguiente sección se explora como es posible controlar la duración y el pitch de una pista de audio de manera independiente.

### 0.1.2. TD-PSOLA

**TD-PSOLA** son las siglas para Time Domain Pitch-Synchronous Overlap-Add.

Los algoritmos basados en **PSOLA** se basan en la reutilización de pequeños segmentos llamados **short term signals** que son el resultado de aplicar una ventana de **Hanning**, que se extiende hacia los pitch-marks vecinos, sobre cada **pitch-mark** (más acerca de pitch-marks en la siguiente sección) y solaparlos convenientemente.

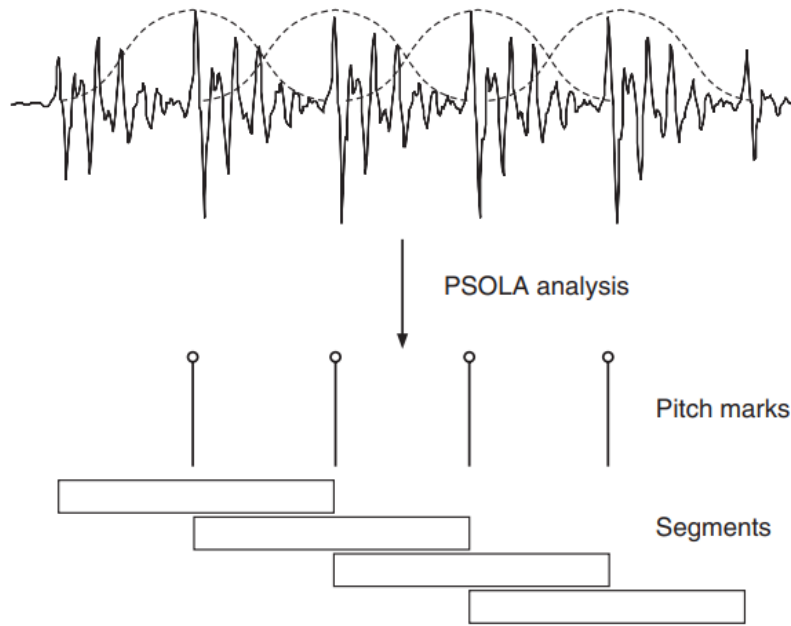


Figura 2: Segmentación del sonido de acuerdo a los pitch-marks

Para poder conseguir alargar la duración de los sonidos sin alterar así el **pitch** del mismo es necesario conservar la forma de los segmentos que la conforman.

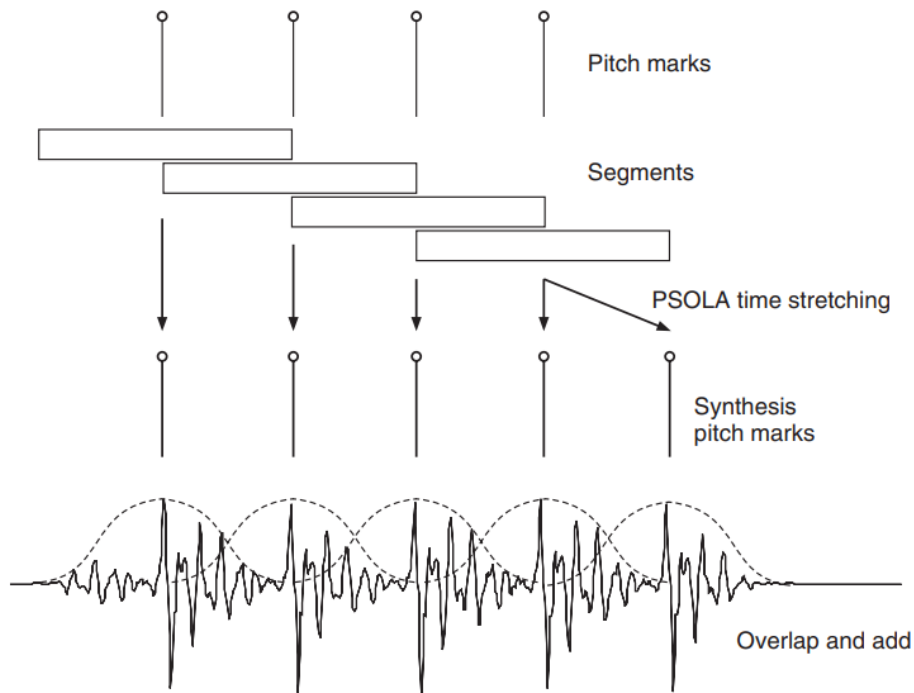
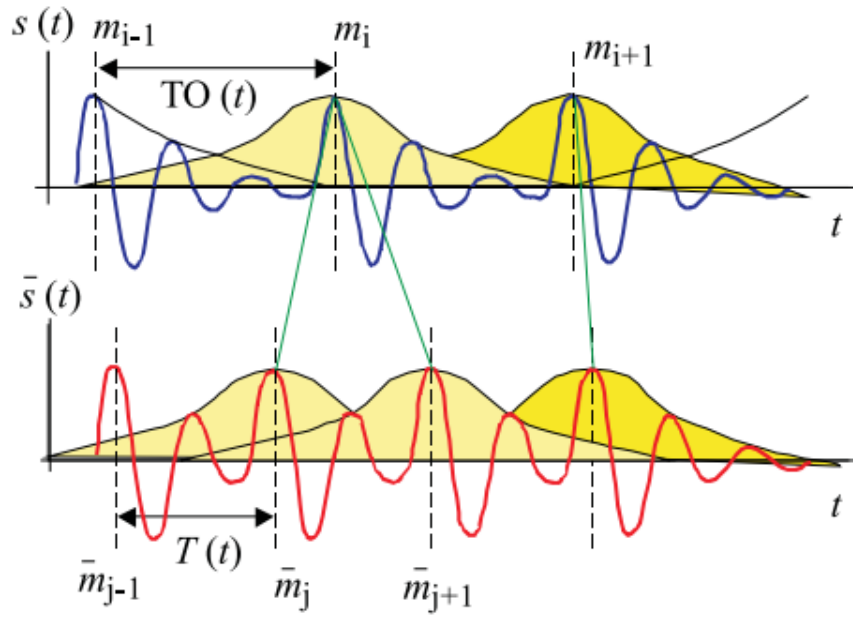


Figura 3: Ilustración del estiramiento temporal usando PSOLA

Esto se expresa dentro del algoritmo de estiramiento como un factor  $\alpha = \frac{d_{nueva}}{d_{original}}$  que nos permite mapear la posición de los pitch-marks dentro de nuestro sonido original hacia el sonido de llegada. Las ventanas de Hanning son de gran utilidad porque permiten transiciones más suaves entre cada segmento.

Para poder tener control sobre el **pitch** del sonido debemos cambiar la distancia entre los pitch-marks para así afectar a la frecuencia fundamental  $f_0$  del sonido. Análogamente, este escalado en frecuencia viene denominado como un factor  $\beta = \frac{f_{nueva}}{f_{original}}$ .

Estos dos factores en combinación nos permiten regular tanto el **pitch** como la duración del sonido.



**Fig. 2.** Example of pitch-shifting and time stretching using PSOLA

Figura 4

### 0.1.3. Estimación de los Pitch-Marks

Los **pitch-marks** representan momentos del sonido en el que su amplitud es máxima en un entorno, en realidad este termino se emplea cuando se habla de **glottal pulses** dentro del campo de Voice Processing y hacen referencia a ciertos impulsos de aire genera nuestra habla. Estas posiciones son centrales para la estructura del sonido. Poder localizarlos con precisión son un punto clave para los algoritmos que se basan en ellos para sintetizar nuevos sonidos.

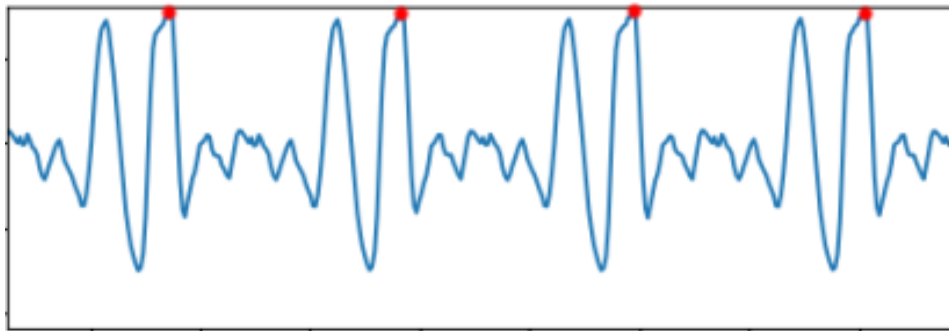


Figura 5: Imagen aumentada de una nota musical

En la figura 4 se pueden apreciar los pitch-marks estimados para la nota **A4**. En este caso podemos decir, mediante inspección visual, que se la estimación ha tenido un éxito considerable. Sin embargo, estimarlos para una sonido más complejo como puede ser la voz humana o una pista de audio no es una tarea sencilla y se deben tener en cuenta varios factores. Para la estimación de los pitch-marks de las notas se utilizo el programa *Audacity* para visualizar el espectro y obtener la frecuencia fundamental de la nota  $f_0 = 395Hz$ . A partir de la frecuencia fundamental y el previo conocimiento del **sample rate** de la muestra se realizaran las siguientes hipótesis.

$$P \approx 1/f_0 = 2.53ms$$

Es decir que  $P$ , el **pitch-period**(tiempo entre pitch-marks) tendrá aproximadamente esa duración.

$$M_p = \frac{\text{sampling rate}}{f_0} = \frac{44100}{395} \approx 112 \text{ muestras}$$

Este dato nos indica que entre cada pitch-mark tenemos 112 puntos de espacio. Con este dato es posible utilizar el paquete científico *Scipy* y calcular los máximos de la señal que estén equiespaciados por lo menos 112 muestras entre sí.

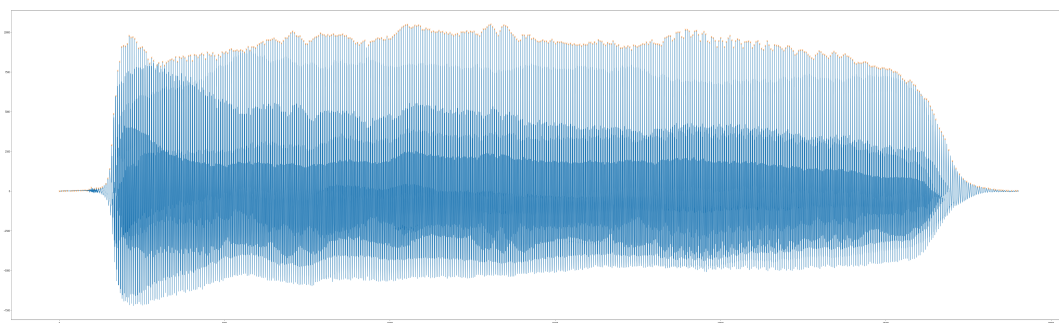


Figura 6: Pitch Marks en una Nota A4 de Saxofón,  $f_0 = 395Hz$  (poner el zoom al maximo y mirar en la parte superior)

Como se puede apreciar en la figura ?? este método funciona bien para la nota A4. Sin embargo al probar este método sobre la nota A#6

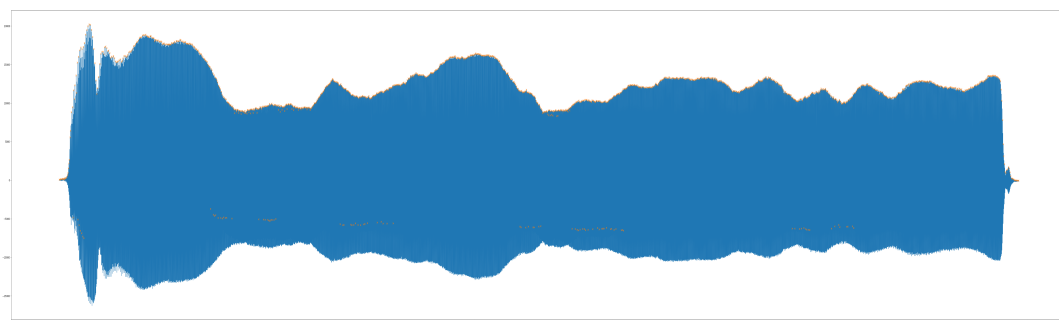


Figura 7: Pitch Marks en una Nota A6 de Saxofón(poner el zoom al maximo y mirar en la parte superior)

no se obtiene resultados óptimos.