

TB065-86.05 - SEÑALES Y SISTEMAS

TRABAJO PRÁCTICO ESPECIAL 1: ANÁLISIS DE LA SEÑAL DE HABLA

☞ = Python.

La señal de habla está compuesta de secciones de propiedades cambiantes. Si pensamos a 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.

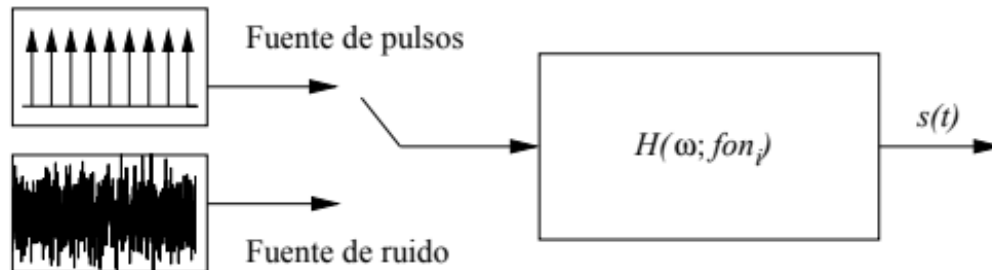


Figura 1: Modelo de producción de la voz

1. Grabar 2 muestras de su propia voz diciendo la palabra “Picasso”. Una rápida y otra lenta, de forma que una tenga aproximadamente el doble de duración que la otra. Graficar en ☞ las señales de voz ubicando porciones de señales periódicas y no periódicas.

2. Realizar una segmentación ☞ de la señal lenta localizando los segmentos de muestras donde aparece una [a] y una [s]. Describir qué diferencias hay entre estos fonemas. Graficar ☞ los segmentos cuasi periódicos de la señal lenta y estimar el periodo y la frecuencia. Calcular lo mismo para la señal rápida.

Los sonidos sonoros son producidos forzando el aire a través de la glotis o a través 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 (F_0) y sus componentes armónicos producidos por las cuerdas vocales. El tracto vocal modifica esta señal de excitación causando el formante. El término formante se utiliza para indicar el centro de estas frecuencias de resonancia, 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 F_1 , F_2 , F_3 ,..., 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 pueden verse representados los 3 primeros formantes de una señal de voz.

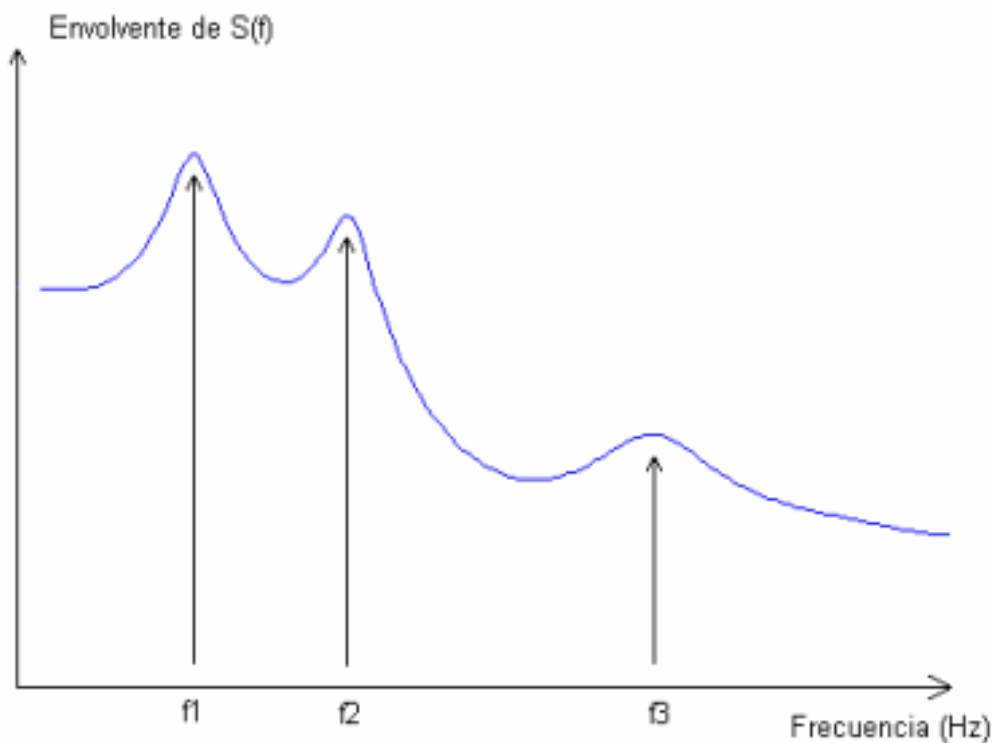



Figura 2: Envolvente del espectro de una vocal.

3. Utilizando la FFT  grafique los coeficientes de Fourier de las porciones correspondientes a las vocales que hay en la señal. Hacer el cálculo tomando varios períodos de la vocal y también tomando un solo período para las 2 señales. Identificar los primeros máximos en las envolventes de estos espectros y estimar los valores de frecuencia en que se producen.

TRABAJO PRÁCTICO ESPECIAL 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 ésta 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:

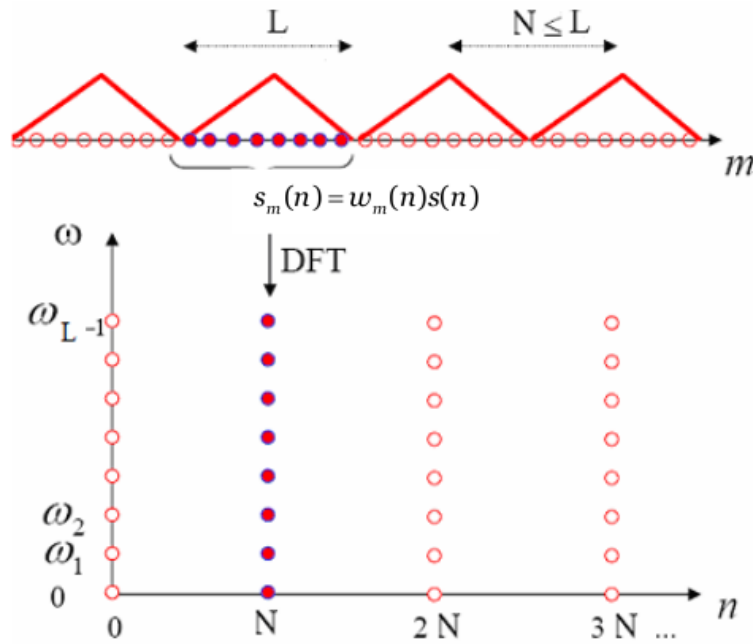
$$S(n, \omega) = \sum_{m=-\infty}^{\infty} s(m)w(n-m)e^{-j\omega m}$$

Donde $w(n)$ 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:

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

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.

¹También llamada STFT, del inglés Short-Time Fourier Transform






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 la fase de la TFCT. La energía la calculamos como:

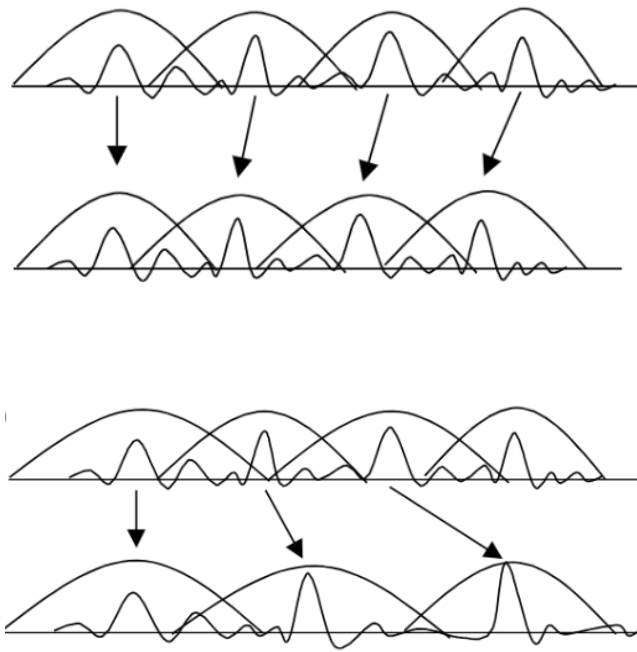
$$\log |X(k)|^2 = \log (X_r^2(k) + X_i^2(k))$$

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`

-
4. Graficar  el espectrograma de banda angosta de la palabra completa de forma que se pueda observar la frecuencia fundamental y sus armónicos.
-
5. Graficar  el espectrograma de banda ancha de la palabra completa de forma que se puedan observar los formantes de los sonidos vocálicos.
-
6. Graficar  el espectrograma de banda angosta y de banda ancha de las vocales presentes en la palabra.

7. 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 a 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.). Los segmentos pueden ser repetidos o eliminados para no aumentar ni disminuir la duración de la emisión de voz.



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.