PRÁCTICA 5: PROCESAMIENTO DIGITAL DE VOZ

1. Introducción: Modelo Digital de Producción de Voz

En la figura 1 se muestra un modelo digital de producción de voz usualmente conocido como modelo LPC (linear prediction coding). Mediante este modelo se supone que la señal de voz $s(n)$ es producida por un filtro digital todo-polos $h(k)$ (representando al tracto vocal) excitado por una señal $u(n)$, que puede ser de dos tipos:

1. Un ruido blanco (imitando flujo de aire procedente de los pulmones) para los sonidos sordos (s, f, ...).
2. Un tren de impulsos (imitando la vibración de las cuerdas vocales) para los sonidos sonoros (a,m,b,...).

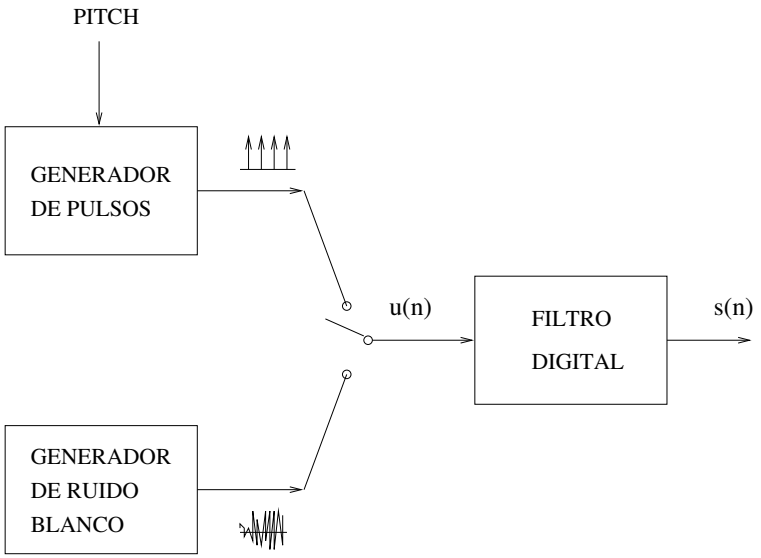

Figura 1: *Modelo digital LPC de producción de voz.*

En el caso de sonidos sonoros, el tren de impulsos tendra una frecuencia igual a la de vibración de las cuerdas vocales, conocida como *frecuencia de pitch*. Su inversa se denomina *periodo de pitch* o, simplemente, *pitch*.

2. Estimación Espectral de Voz

2.1. Periodograma

En muchas aplicaciones es necesario obtener una estimación del espectro de la señal de voz. En concreto, se trata de obtener la función *Densidad Espectral de Potencia* (función PSD),

definida como,

$$P_x(\omega) = \lim_{N\to\infty} \frac{1}{2N+1} \left| \sum_{n=-N}^{N} x(n) e^{-j\omega n} \right|^2 \quad (1)$$

El problema de la expresión anterior es que se requeriría una señal con infinito número de muestras para obtener la forma exacta del espectro. Esto es, obviamente, imposible, por lo que tenemos que conformarnos con una estimación del mismo. La forma más sencilla de obtener esta estimación es mediante el *Periodograma*, que se obtiene a partir de DFT de un segmento de señal de N muestras como:

$$P_x(\omega_k) = |DFT[x(n)]|^2/N \quad (2)$$

donde $\omega_k = 2k\pi/N$ con $k = 0, 1, \ldots, N-1$.

En esta parte de la práctica se aplicará el método del periodograma para la estimación de un segmento de voz. Este segmento está formado por 256 muestras de la vocal ”e” muestreada a 8 KHz. Para cargar y visualizar la señal se ejecutará:

```
load e.asc
plot(e)
```

La señal (almacenada en la variable e) presenta un aspecto periódico. El periodo de la señal vocal coincide con el pitch, anteriormente definido. Medir el pitch de la vocal directamente sobre la señal.

Posteriormente se obtendrá el periodograma de la vocal aplicando la expresión (2). Para su visualización es conveniente usar la escala decibélica ($10\log_{10} P_x$). El espectro proporciona información sobre la señal que sería imposible extraer en el dominio del tiempo:

- Podemos observar una estructura fina o rizado que es debida a la excitación. Podemos medir la frecuencia de pitch como la diferencia entre dos picos consecutivos del rizado y comparar con el valor del pitch obtenido directamente sobre la señal.
- Obviando el rizado, se pueden distinguir una serie de picos suaves que corresponden a las frecuencias de resonancia del tracto vocal, también conocidas como *formantes*. Medir de forma aproximada dichos formantes.

2.2. Espectro LPC/AR

Diversas aplicaciones requieren la determinación del espectro de voz entendido como la respuesta en frecuencia del filtro todo-polos del modelo de voz (es decir, se prescinde de la excitación). Para sonidos sonoros, este espectro corresponde aproximadamente a la respuesta en frecuencia de un modelo determinista todos-polos, mientras que en el caso de sonidos sordos se trataría de un espectro basado en un modelo AR de proceso aleatorio. En ambos casos, la solución es idéntica y es normalmente conocida como *espectro AR o LPC* de voz.

En este apartado se obtendrá y comparará el espectro LPC con el periodograma. Se obtendrá la PSD de la señal por los tres métodos que se mencionan a continuación.

1. Periodograma:
   Se calculará de nuevo, pero a partir de la de la autocorrelación. Ésta puede obtenerse mediante la función *xcorr*, de la siguiente manera.

```
rx = xcorr(e,'biased');
plot(rx);
```

El argumento "biased" indica que se esta realizando una normalización del tipo $1/N$, por lo que se está usando la estimación $\hat{r}_x$ (periodograma), mientras que si se usase "unbiased" se estaría normalizando con $1/(N-k)$ (estimación $\tilde{r}_x$). El periodograma es $10\log_{10}|RX|$, donde RX es la FFT de la autocorrelación.

2. Espectro AR(12) por el método de autocorrelación:
   Aplicar la expresión vista en teoría. La matriz de autocorrelación Toeplitz puede obtenerse usando el comando *toeplitz*. Para realizar la inversión de la matriz de autocorrelación puede usarse el comando *inv*. Comprobar el resultado obtenido con el que se tiene usando el comando *lpc*.

3. Espectro AR(12) por el método de covarianza:
   Aplicar las expresiones vistas en teoría, y proceder de modo similar al apartado anterior. Comparar el resultado con el obtenido por el método anterior.

Comentar las diferencias entre la PSD no paramétrica del periodograma y las dos PSDs paramétricas obtenidas (superponer gráficas). Comparar los valores de los formantes obtenidos en cada método.

3. Análisis Homomórfico

Hemos visto que el periodograma nos proporciona información acerca de los parámetros del modelo de producción de voz, tanto de la excitación (pitch) como del filtro digital que representa el tracto vocal (formantes). Sin embargo, estas informaciones aparecen bastante entremezcladas. Mediante el *análisis homomórfico* es posible obtener una separación más clara de la excitación $U(\omega)$ y el filtro $H(\omega)$. Para ello basta aplicar una función logarítmica sobre el periodograma $P_x(\omega)$ de la señal,

$$X(\omega_k) = H(\omega_k)U(\omega_k) \tag{3}$$

$$P_x(\omega_k) = |H(\omega)|^2 P_u(\omega_k) = P_h(\omega_k)P_u(\omega_k) \tag{4}$$

$$\log P_x(\omega_k) = \log P_h(\omega_k) + \log P_u(\omega_k) \tag{5}$$

Vemos que la aplicación del logaritmo descompone el espectro en dos sumandos, uno debido al tracto vocal y otro a la excitación. El análisis homomórfico se completa con la aplicación de la DFT inversa:

$$c_x(n) = IDFT[\log P_x(\omega_k)] = IDFT[\log P_h(\omega_k)] + IDFT[\log P_u(\omega_k)] = c_h(n) + c_u(n) \tag{6}$$

El resultado de esta operación sobre el periodograma es una nueva señal que recibe el nombre de *cepstrum FFT*. El dominio n del cepstrum es de nuevo el tiempo (ya que hemos aplicado una IDFT), pero se suele denominar *cuefrencia* para diferenciar claramente el cepstrum de la señal original. El cepstrum $c_x(n)$ tiene dos componentes, $c_h(n)$ y $c_u(n)$, debidas al filtro vocal (correlaciones de retardo corto) y a la excitación (correlaciones de retardo largo), respectivamente. En la

práctica es sencillo separar ambas componentes, ya que la excitación contribuye esencialmente a componentes de alta cuefrencia, mientras que el filtro contribuye a las bajas.

El cepstrum también puede obtenerse a partir del espectro LPC (AR), en lugar del periodograma. En este caso recibe el nomre de *cepstrum LPC*, y solo contiene información relativa al tracto vocal.

Obtener y visualizar el cepstrum FFT de la señal anterior. Determinar de nuevo el pitch de la misma midiendo el valor de cuefrencia al que se produce el primer pico del cepstrum para valores superiores a $n = 20$.

4. Reconocimiento de Vocales

Para el reconocimiento de señales de voz la información relevante es la relativa al tracto vocal, ya que es la que define el tipo de sonido que se ha emitido. Por el contrario, la información relativa a la excitación no es útil, ya depende de factores altamente variables como la entonación, sexo del locutor, estado emocional del locutor, etc... Por ello, una buena manera de representar la información relativa exclusivamente al tracto vocal es mediante un vector de parámetros que contenga los primeros L coeficientes cepstrales $(c(1), c(2), \ldots, c(L))$, siendo L un número pequeño (típicamente entre 12 y 20). El primer coeficiente cepstral $c(0)$ tampoco se suele incluir en el vector, ya que está relacionado con la energía de la señal, que es también un parámetro sometido a una alta variabilidad. El **cepstrum LPC** proporciona mejores resultados que el cepstrum FFT, por lo que usaremos esta variante en lo que sigue. Este cepstrum puede obtenerse indistintamente por el método anteriormente descrito, o mediante la siguiente recursión:

$$c(n) = \begin{cases} 0 & n \leq 0 \\ -a_1 & n = 1 \\ -a_n - \sum_{k=1}^{n-1} \frac{k}{n} c(k) a_{n-k} & n > 1 \end{cases}$$

Para la realización práctica de un reconocedor de vocales simple, se dispone de 5 segmentos vocálicos de 256 muestras, correspondientes a los 5 fonemas vocálicos (ficheros a.asc, e.asc, i.asc, o.asc, u.asc). Un diagrama de bloques del reconocedor se muestra en la figura 2. Utilizando el vector de coeficientes cepstrales LPC ($L = 12$) de cada una de esta vocales como referencia, ha de determinarse a qué vocal corresponde un segmento incógnita (x.asc), mediante la determinación de las distancias euclideas correspondientes entre este segmento y cada una de las referencias. La distancia euclidea entre un vector de referencia $\mathbf{c_r}$ y otro incógnita $\mathbf{c_x}$, se obtiene como,

$$d_C(\mathbf{c}_x, \mathbf{c}_r) = \sum_{n=1}^{L} (c_x(n) - c_r(n))^2 = \|\mathbf{c}_x - \mathbf{c}_r\|^2 \tag{7}$$

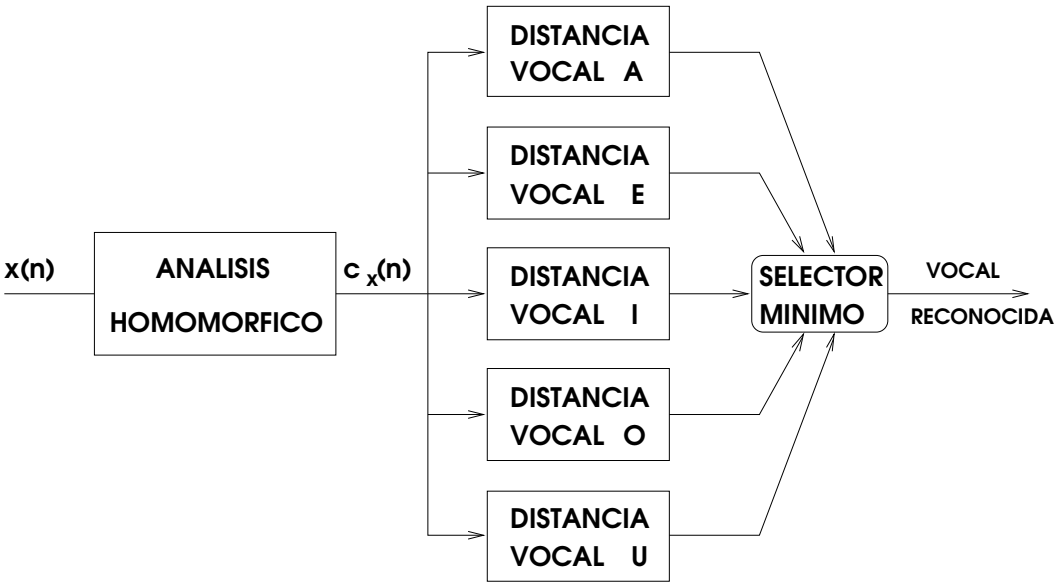

Figura 2: *Diagrama de bloques de un reconocedor de segmentos vocálicos*