

ANÁLISIS DE SERIES TEMPORALES (AST1)

Trabajo Práctico: Linear predictive Coding

1. Introducción

La codificación por predicción lineal (Linear predictive Coding, o abreviadamente LPC) es una técnica de codificación basada en predicción lineal. Este tipo de método suele ser utilizados para representar la misma información “dispersa” de la señal no solo de una manera más compacta sino también más conceptual. Como ya veremos más en detalle, LPC está especialmente adaptado a señales de habla ya que tiene una relación muy cercana con la manera en que la que este tipo de señales es generada. Cada una de las palabras que decimos está asociada a una serie de movimientos nuestro aparato fonador. Cada movimiento particular genera sonidos elementales que en cada lenguaje tienen significación como unidad. Los sonidos elementales (el correspondiente a la /a/, a la /s/, a la /n/, etc.) podrían describirse desde un punto de vista físico (en promedio) como una cierta configuración del aparato fonador para cada uno de ellos. LPC logra asociar a cada una de esas configuraciones un código diferente. Por lo tanto se la considera no sólo una técnica para codificar una señal sin importar su contenido, sino una metodología para conceptualizar el mensaje contenido en la señal. Los coeficientes obtenidos por LPC no sólo constituyen un ahorro de coeficientes de representación, sino que de alguna manera podrían simular los mismos procesos que hace nuestro cuerpo cuando genera la señal de habla, y por lo tanto es posible preguntarse si esa representación no es adecuada para entender el mensaje inmerso en la señal de habla.

2. Producción del habla

En la Fig. 1 puede verse un esquema anatómico que utilizaremos para explicar como funciona la producción de habla. Hay tres tipos de sonidos (en el lenguaje español) básicos que deben distinguirse para este análisis.

- Fonemas sonoros. Son los que son generados por las cuerdas vocales; incluye tanto las vocales (/a/, /e/, etc.) como constantes sonoras (/n/, /m/, etc.). Estos sonidos son producidos por el humano en una manera muy similar a la que los instrumentos de viento. El elemento vibrador son las cuerdas vocales, ubicadas en la laringe, que producen excitaciones cuasiperiodicas por una combinación del paso del aire proveniente de los pulmones y la tensión de éstas.
- Fonemas sordos. Son los generados de forma ruidosa o aleatoria como /f/, /s/ o /j/. En este caso las cuerdas vocales están totalmente abiertas, dejando pasar el aire de

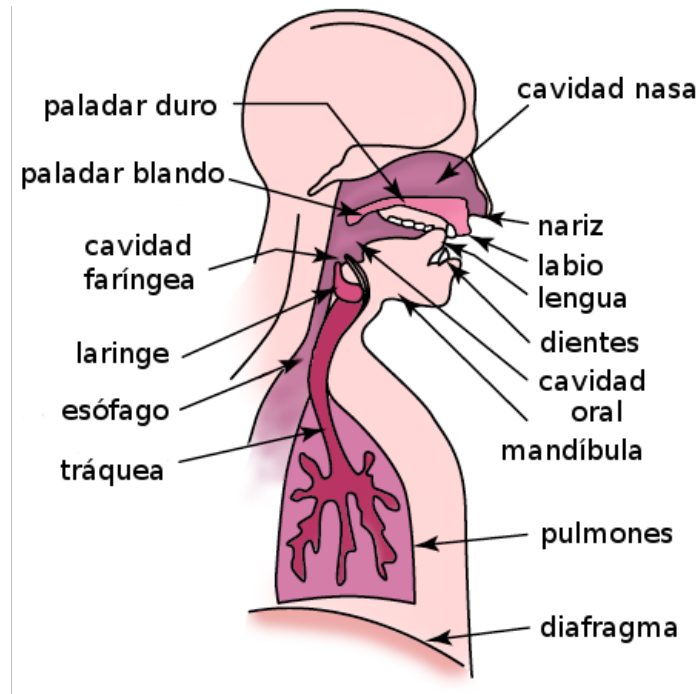


Figura 1: Sistema de producción del habla.

manera libre, es decir que no son las generadoras del sonido. El proceso de excitación posee una característica ruidosa (cuasi-blanco) tal que puede ser considerado estacionario en sentido amplio (ESA).

- **Fonemas Explosivos.** Los fonemas explosivos son generados de forma similar a los sordos pero concentrando todo el aire en un solo impulso. Ejemplos de estos fonemas son /t/ o /c/. Debido a su corta duración no son de gran relevancia para este trabajo práctico.

La excitación sonora puede poseer distintas características, pero siempre es aire impulsado a atravesar diversas cavidades en la cabeza (boca, nariz, lengua) que hacen las veces de cámara de resonancia hasta su salida exterior. El conjunto de estas cavidades se denomina tracto vocal. La excitación sonora a la entrada de una cámara resonadora y la onda de salida producida se relacionan entre sí mediante una ecuación diferencial. Si denominamos $x[t]$ a la excitación sonora, e $y[t]$ a la onda de salida producida, el sistema puede aproximarse con la siguiente modelo autoregresivo:

$$y[t] = x[t] + \sum_{m=1}^{\infty} a_m y[t - m] \quad (1)$$

El proceso de producción del habla entonces se puede describir mediante un sistema y una señal de excitación. El sistema descrito por los coeficientes a_k representan como

colocamos el aparato fonador para emitir los diferentes sonidos. La señal de excitación $x[t]$ representa la vibración de las cuerdas vocales (impulsos cuasiperiódicos) para los sonidos sonoros, y ruido estocástico para los sordos. En el caso de los sonidos sonoros la frecuencia representará el tono en que la persona está hablando y la magnitud el volumen. El objetivo de este trabajo práctico será estimar, para cada porción de audio, el sistema representativo del aparato fonador y la señal de excitación $x[t]$ correspondiente.

Ejercicio 1 Utilizando `load` (librosa), cargar el archivo de audio `estocastico.wav` y escucharlo utilizando `Audio` (IPython.display).

La función `load` devuelve dos variables: la señal y la *frecuencia de muestreo*. La frecuencia de muestreo representa cuantas muestras se toman en 1 segundo. En este trabajo será utilizada para convertir los índices en tiempo real a la hora de efectuar los gráficos.

Ejercicio 2 Determinar aproximadamente los instantes de tiempo donde se pronuncian las vocales en el audio `estocastico.wav`. La función `recortar` de `recorte.py` puede ser útil.

La función `recortar` es una herramienta interactiva que ayuda a encontrar los momentos de comienzan y terminan el sonido de las vocales, mediante dos perillas. A su vez posee dos botones, uno que reproduce el segmento seleccionado y otro que guarda los valores donde inició y finalizó dicho segmento. Se utiliza de la siguiente manera:

```
from recorte import recortar
x,fs = librosa.load("estocastico.wav")
markers = recortar(x,fs)
```

3. Linear predictive Coding (LPC)

Sea $\tilde{y}[t] = \sum_{k=1}^M h_k y[t-k]$, se desea estimar los M -coeficientes que predicen el próximo valor de la señal por mínimos cuadrados. Es decir,

$$\mathbf{h} = \begin{pmatrix} h_1 \\ \vdots \\ h_M \end{pmatrix} = \mathbf{R}^{-1} \mathbf{r} \quad (2)$$

con

$$\mathbf{r} = \begin{pmatrix} \Gamma_y[1] \\ \vdots \\ \Gamma_y[M] \end{pmatrix}, \quad \mathbf{R} = \begin{pmatrix} \Gamma_y[0] & \Gamma_y[1] & \cdots & \Gamma_y[M-1] \\ \Gamma_y[1] & \Gamma_y[0] & \cdots & \Gamma_y[M-2] \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_y[M-1] & \Gamma_y[M-2] & \cdots & \Gamma_y[0] \end{pmatrix} \quad (3)$$

donde $\Gamma_y[k]$ es la autocorrelación estimada a través de su estimador sesgado. En el caso de sonidos sordos, la señal de habla $y[t]$ será ESA (ya que es la salida de un sistema LTI

con excitación ESA) y por lo tanto esta operación representa una predicción lineal que minimiza el error cuadrático medio $\mathbb{E}[(y[t] - \tilde{y}[t])^2]$. Para sonidos sonoros, si bien la señal $y[t]$ será determinística, el procedimiento sigue minimizando el error cuadrático promedio: $\frac{1}{N} \sum_{t=0}^{N-1} (y[t] - \tilde{y}[t])^2$, donde N es el largo de la señal.

Ejercicio 3 Utilizando `correlate` (numpy) implementar una función que calcule los coeficientes LPC de una señal (representativa de un solo sonido) para $N > M$.

En el caso donde las señales de habla, el modelo esté caracterizado por (1), el error de predicción $\hat{x}[t]$ será de la forma

$$\hat{x}[t] = y[t] - \tilde{y}[t] = x[t] + \sum_{m=1}^M (a_m - h_m)y[t-m] + \sum_{m=M+1}^{\infty} a_m y[t-m] \approx x[t] \quad (4)$$

El método de LPC posee dos hipótesis fundamentales. Por un lado se asume que a partir de un determinado coeficiente estos a_m serán muy pequeños ($a_m \approx 0$ para $m > M$). Por el otro lado se postula que la señal de excitación sea impredecible a partir del pasado de la señal de habla, volviendo a $\tilde{y}[t]$ una predicción razonable de la señal de habla ($a_m \approx h_m$ para $m \leq M$). Es ese sentido es que se propone el error de predicción como una estimación de la señal de excitación. De esta manera, mediante una predicción lineal podemos caracterizar tanto el aparato fonador (a partir de los h_k) como la señal de excitación $\hat{x}[t] = y[t] - \tilde{y}[t]$.

Ejercicio 4 Asumiendo conocidos los coeficientes LPC, hallar analíticamente la respuesta impulsiva de un sistema LTI cuya entrada sea $y[t]$ y su salida $\hat{x}[t]$. Es decir, hallar $h_f[t]$ tal que $\hat{x}[t] = (y * h_f)[t]$.


Ejercicio 5 Utilizando `convolve` (numpy) implementar una función que estime la frecuencia de excitación a partir de la señal de un sonido y sus coeficientes LPC.

Ejercicio 6 Asumiendo conocidos los coeficientes LPC, hallar una solución analítica para $H_f(e^{j\omega})$ (la respuesta en frecuencia del filtro del Ej. 4).

Si tanto la señal de excitación como los coeficientes LPC fueron correctamente estimados, se debe poder reconstruirse la señal de habla original como función de estos. Es decir, se define la reconstrucción $\hat{y}[t]$ como:

$$\hat{y}[t] = \hat{x}[t] + \sum_{m=1}^M h_m \hat{y}[t-m] \quad (5)$$

Ejercicio 7 Asumiendo conocidos los coeficientes LPC, hallar una solución analítica para $H_i(e^{j\omega})$ (la respuesta en frecuencia del filtro definido en (5), donde $\hat{x}[t]$ es la entrada y $\hat{y}[t]$ la salida).

Ejercicio 8 Implementar una función que represente el sistema descrito en (5) (entrada $\hat{x}[t]$ y salida $\hat{y}[t]$) a partir de su ecuación en diferencias. : No trate encontrar la respuesta impulsiva $h_i[t]$ definida en el Ej. 7 porque la misma posee duración infinita.

4. Ventaneo

Las señales de habla no se caracterizan por contener un solo fonema, sino que están formados combinando de muchos silencios y sonidos que representan el lenguaje. Esto implica que hay que hacer una actualización de los coeficientes cada determinado tiempo. Esta actualización debería coincidir con los tramos donde la señal tiene cierta uniformidad espectral, pero en la práctica la actualización se hace cada un determinado tiempo sincrónicamente, supuesto suficientemente pequeño como para no perder ningún fonema, y no tan chico como para que el método sea impráctico. El uso estándar es considerar ventanas de señal de longitud 0.025 seg, actualizadas cada 0.010 seg para calcular los coeficientes LPC. Es recomendable utilizar la ventana de Hamming que reduzca la inclusión de altas frecuencias debido al proceso de ventaneo. Se considera que para la mayor parte de las secciones de la señal de habla, 20 coeficientes es suficiente.

Ejercicio 9 Utilizando la función desarrollada en el Ej. 3, calcular los coeficientes LPC del audio *estocastico.wav* para cada ventana de tiempo de la señal. Utilizar la ventana de *hamming* (numpy).

Ejercicio 10 Utilizando *specgram* (matplotlib) graficar un espectrograma del audio *estocastico.wav*, configurando sus parámetros según lo discutido previamente.

Ejercicio 11 Elegir los coeficientes LPC correspondientes a una ventana de tiempo de la letra /e/. Graficar $|H_f(e^{j\omega})|$ y $|H_i(e^{j\omega})|$.

Ejercicio 12 Utilizando la función desarrollada en el Ej. 5, estimar la función de excitación $\hat{x}[t]$ del audio *estocastico.wav* cuidando de evitar los transitorios del filtro. Esto puede hacerse conservando los transitorios finales de cada sección, y sumándolos al siguiente segmento.

Ejercicio 13 *Utilizando la función desarrollada en el Ej. 8, reconstruir la señal de habla $\hat{y}[t]$ y escucharla utilizando **Audio** (`IPython.display`).*

5. Optativos

Como se mencionó anteriormente, el hecho de utilizar una codificación representativa del proceso de producción del habla nos permite modificar el audio a gusto. Los siguientes ejercicios buscan modificar, por separado, el modelado del aparato fonador y la señal de excitación.

Ejercicio 14 *Se desea reemplazar todas las vocales del audio con la letra /e/. Para ello, defina unos nuevos coeficientes LPC a partir de los originales. Reemplace todos los coeficientes correspondientes a vocales con los utilizados en el Ej. 11. Reconstruir la señal de habla (utilizando la señal $\hat{x}[t]$ del Ej. 12) y escucharla utilizando **Audio** (`IPython.display`).*

Ejercicio 15 *Se desea cambiar la frecuencia glótica del audio. Para ello se modificará $\hat{x}[t]$ en los segmentos correspondientes a los fonemas sonoros (vocales en este caso). Cada segmento completo correspondiente a una vocal será duplicado (concatenando uno a continuación del otro) y quedándose con una de cada dos muestras (duplicando la frecuencia). Reconstruir la señal de habla (utilizando los coeficientes LPC del Ej. 9) y escucharla utilizando **Audio** (`IPython.display`).*