Trabajo práctico 3

Sintetización de fonemas mediante procesos AR

1. Introducción

Un enfoque para producir la síntesis de voz artificial es imitar el proceso de generación del habla humana. Considerando el modelo de resonancia de la Figura 1, la producción de sonidos de habla se puede pensar como una fuente de señal que excita un sistema que resonará más en ciertas frecuencias que en otras, dependiendo la configuración que tenga el tracto vocal para cada fonema particular. Uno de los métodos más conocidos para resolver este problema se basa en la técnica LPC (Linear Predictive Coding), con la que se podrá modelar la producción de habla a través de un sistema LTI excitado por un proceso aleatorio.

1.1. Descripción del modelo

El modelo utilizado por LPC para la generación habla se representa en la Figura 2 [2, 3], en donde el sistema $H(z)$ es utilizado para representar la resonancia del tracto vocal de un determinado fonema, asumiéndolo LTI (para intervalos de tiempo de corta duración). Existen dos clases de sonidos dependiendo la fuente de excitación: de tipo *sonora*, para representar fonemas de vocales (ej: "a", "e", "i", etc) o consonantes sonoras (ej: "b", "l", "m", etc), y de tipo *sorda*, para fonemas de consonantes no sonoras (ej: "s", "f","t", etc.). Para el habla sonora la fuente de excitación más apropiada es un tren de impulsos periódico $\sum_{k=-\infty}^{\infty} \sqrt{T f_s}\, \delta(t-kT)$ de periodo T, siendo $f = 1/T$ el "*pitch*" o frecuencia fundamental de la excitación. Por su parte, el habla sorda recibe como entrada un proceso de ruido blanco gaussiano $\sim N(0,1)$, que resulta más adecuado para este caso. El sistema LTI utilizado para modelar a $H(z)$ en este problema es del tipo *all-pole* (de solo polos) [4], como se describe en la ecuación (1), caracterizado por los coeficientes $\{a_1, a_2, ..., a_P\}$ y la ganancia G. Asumiendo la entrada $u(n)$ como un proceso aleatorio ESA, la salida del sistema representa un proceso autorregresivo $x(n)$ de orden P que puede expresarse con su ecuación en diferencias indicada en (2).

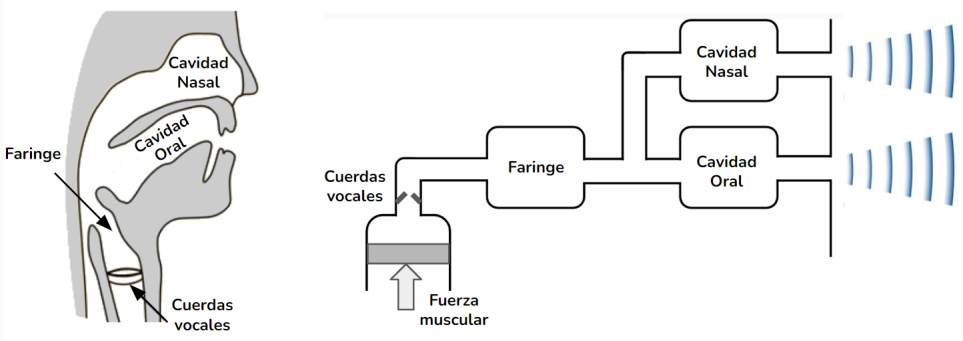

Figura 1: Modelo de resonancia del sistema de generación de habla. A la izquierda modelo anatómico. A la derecha, modelo idealizado.

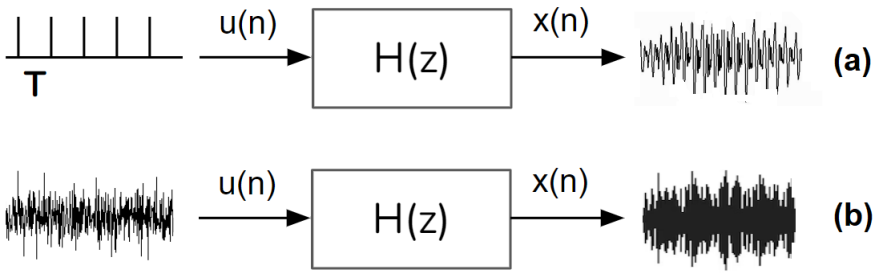

Figura 2: Modelos de generación del habla mediante un sistema LTI. (a) señales sonoras. (b) señales sordas.

$$H(z) = \frac{X(z)}{U(z)} = \frac{G}{1 - \sum_{i=1}^{P} a_i z^{-1}} \tag{1}$$

$$x(n) = \sum_{i=1}^{P} a_i \, x(n-i) + G \, u(n) \tag{2}$$

1.2. Estimación de parámetros del modelo

Suponiendo que el proceso $x(n)$ es un proceso AR que se genera como salida de un sistema all-pole excitado por un proceso blanco $u(n) \sim N(0,1)$. Nos interesa encontrar el conjunto de coeficientes a_1, a_2, ..., a_P y el factor de ganancia G (de la ecuación 1) de modo que éstos se ajusten lo mejor posible a la señal de habla que queremos simular. Una forma de abordar este problema es resolviendo las ecuaciones de Yule-Walker [5]. Partiendo de la ecuación (2) se puede comprobar que $R_X(k) = \sum_{i=1}^{P} a_i R_X(k-i) + G^2\delta(k)$, donde $\delta(k)$ es una delta de Dirac discreta. Esta ecuación se puede separar en una solución para $k = 0$, resultando en $R_X(0) = \sum_{i=1}^{P} a_i R_X(i) + G^2$ y otra para $k \neq 0$ que conduce a $R_X(k) = \sum_{i=1}^{P} a_i R_X(k-i)$. De estas últimas dos expresiones se puede plantear un sistema matricial para estimar los parámetros de $H(z)$ utilizando (3) para los coeficientes y (4) para la ganancia, donde $\mathbf{a} = [a_1 \, a_2 \, ... \, a_P]^T$, $\mathbf{r} = [R_X(1) \, R_X(2) \, ... \, R_X(P)]^T$ y $\mathbf{R}$ como se define en (5). Como se puede observar, todos los parámetros estimados del modelo dependerán de la autocorrelación $R_X(k)$, que a su vez deberá ser estimada a partir de las observaciones de una realización del proceso $x(n)$.

$$\widehat{\mathbf{a}} = \mathbf{R}^{-1}\mathbf{r} \tag{3}$$

$$\widehat{G} = \left(R_X(0) - \sum_{i=1}^{P} a_i R_X(i) \right)^{1/2} \tag{4}$$

$$\mathbf{R} = \begin{bmatrix} R_X(0) & R_X(1) & \dots & R_X(P-1) \\ R_X(-1) & R_X(0) & \dots & R_X(P-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_X(-P+1) & R_X(-P+2) & \dots & R_X(0) \end{bmatrix} \tag{5}$$

2. Desarrollo

Ejercicio 1

Tomando como referencia el proceso AR-P que se describe en la ecuación (2):

(a) Considerando un proceso blanco de entrada $u(n)$. Demuestre que se cumple la siguiente ecuación:

$$R_X(k) = \sum_{i=1}^{P} a_i R_X(k-i) + G^2\delta(k) \tag{6}$$

(b) Demuestre que dada la función de autocorrelación del proceso AR, $R_X(k)$, para $k = 0$ se puede obtener la ganancia G como se indica en la ecuación (4).

(c) Demuestre que para el caso de $k \neq 0$, se cumple la ecuación (3), con la que se pueden hallar los coeficientes del modelo.

(d) Teniendo en cuenta que el proceso de entrada $u(n)$ es un proceso blanco, demuestre que la densidad espectral de potencia (PSD) $S_X(\omega)$ del proceso de salida resulta:

$$S_X(\omega) = \frac{G^2}{|1 - \sum_{i=1}^{P} a_i e^{-j\omega\, i}|^2} \tag{7}$$

Ejercicio 2

En el campus se proveen un conjunto de archivos de audio WAV (frecuencia de muestreo $f_s = 14700\ Hz$) que contienen pequeños tramos de señal (que asumimos aproximadamente ESA[1]) con la pronunciación de diferentes fonemas. Considere las *vocales* (excitación sonora) "`a.wav`", "`e.wav`", "`i.wav`", "`o.wav`" y "`u.wav`", y las *consonantes fricativas* (excitación sorda) "`sh.wav`", "`f.wav`" , "`s.wav`", "`j.wav`":

(a) Defina una función con prototipo `param_ar(x, P)`, donde `x` es una realización del proceso $x(n)$ y `P` el orden del modelo. La función debe retornar los coeficientes LPC (en un vector **a**) y la ganancia G.

(b) Para cada pista de audio, estime los parámetros LPC suponiendo un órden $P = 20$. Grafique la respuesta temporal de los audios utilizados, la estimación de su autocorrelación y su periodograma (en decibeles y para el rango $\omega \in [0, \pi)$) superpuesto a su respectiva PSD teórica obtenida de la ecuación (7).

Ejercicio 3

(a) Considere los parámetros LPC hallados en el ejercicio anterior. Para una vocal y una consonante (por ejemplo "e" y "sh"), grafique la PSD teórica superpuesta a la PSD estimada con el método de Welch [6]. Utilice una ventana $v(n)$ de Hamming con un solapamiento del 50 % considerando los siguientes largos de ventana: $M = 10$, $M = 100$, $M = 1000$.

(b) ¿Qué puede decir acerca del tamaño de la ventana? En qué aspectos mejora o empeora variar este parámetro?

[1] Si se usaran muestras de audio mas largas, deberían tomarse segmentos suficientemente chicos (30-100 ms) para garantizar la estacionariedad por tramos.

Ejercicio 4

(a) Con los parámetros LPC obtenidos, sintetice las señales de cada fonema para una duración de 500 ms y una frecuencia de pitch de 100 Hz para las vocales. Grafique los periodogramas de las señales sintéticas superpuestos a la PSD teórica de cada una.

(b) Genere una señal concatenando los nueve fonemas sintéticos (las cinco vocales y las cuatro consonantes) obtenidos en el punto anterior. Utilice la función `suavizar_bordes()`, disponible en el campus, para modular la amplitud de cada fonema suavizando los extremos (pruebe con `fade=30`). Reproduzca la secuencia resultante para evaluar subjetivamente los sonidos percibidos. Repita la consigna pero generando los fonemas con un pitch diferente para cada vocal (Ej: *a*-100 Hz, *e*-125 Hz, *i*-150 Hz, *o*-125 Hz y *u*-100 Hz).

3. Conclusiones

Como conclusiones, elabore un resumen breve y conciso comentando características que considere relevantes del método propuesto en este trabajo y los resultados obtenidos, así como dificultades encontradas (si fuera el caso) y cómo fueron abordadas.

4. Normas y material entregable

- **Informe**: El informe debe entregarse en formato PDF (**no se aceptarán otros formatos**) y con nombre: `TP3_GXX.PDF` (donde `XX` es el número de grupo). El código fuente (de MATLAB, OCTAVE o PYTHON) debe presentarse en un archivo aparte.
- **Código**: Los archivos de código fuente desarrollados deben estar en formato `.m` (si se usa MATLAB/OCTAVE) o formato `.py` (si usa PYTHON). El código debe incluirse junto al informe en un archivo ZIP (con mismo nombre que el informe) que deberá subirse al campus.

Referencias

[1] M Olmo R Nave. Resonancia del Tracto Vocal. http://hyperphysics.phy-astr.gsu.edu/hbasees/Music/vocres.html

[2] Steven M. Kay. Intuitive probability random processes using MATLAB. Springer 1951.

[3] Rabiner, L.R., R.W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, Englewood Cliffs, NJ, 1978.

[4] Makhoul, J., "Linear Prediction: A Tutorial Review,ÏEEE Proceedings, Vol. 63, pp. 561-580, 1975.

[5] S. Haykin, Adaptive filter theory, Prentice-Hall, 1996.

[6] Stoica, Petre, and Randolph L. Moses , 2005. Spectral analysis of signals.