Procesamiento Digital de Señales de Voz

Modelos de Producción de Voz

Juan Carlos Gómez *(versión 01/10/01)*

1. Introducción

Las señales de voz son producidas cuando una columna de aire desde los pulmones excita el conducto vocal, que se comporta como una cavidad resonante. El conducto vocal es usualmente modelado como la concatenación de tubos acústicos sin pérdidas, con distintas secciones transversales, que comienza en las cuerdas vocales y termina en los labios. La apertura de las cuerdas vocales se denomina *glotis.* Los diferentes sonidos pueden genéricamente ser clasificados en: sonidos *tonales* (en inglés *voiced,* como el de las vocales) y sonidos *no tonales* (en inglés *unvoiced,* como por ejemplo el de una 's' final de palabra).

Los sonidos *tonales* son producidos al forzar aire a través de la glotis con las cuerdas vocales tensadas de manera que se produce la oscilación relajada de las mismas, excitando de esa forma el conducto vocal con pulsos de aire cuasi-periódicos. Cuanto más grande es la tensión de las cuerdas, más alta es la frecuencia fundamental de la voz producida. Los sonidos *no tonales,* en tanto, son generados manteniendo las cuerdas abiertas, formando una constricción del conducto vocal, y forzando aire a través de la constricción a una velocidad lo suficientemente alta como para producir turbulencia. En este caso, puede pensarse que el conducto vocal es excitado por una fuente de ruido aleatorio.

La Figura 1 representa un modelo (en tiempo discreto) del sistema de producción de voz. El conducto vocal se representa por un sistema lineal (en general inestacionario) que es excitado a traves de una llave que selecciona entre una fuente de impulsos cuasi-periódicos para el caso de sonidos *tonales*, o una fuente de ruido aleatorio para el caso de sonidos *no tonales*. La ganancia apropiada de la fuente, *G*, es estimada a partir de la señal de voz, y la señal escalada es usada como entrada del modelo del conducto vocal.

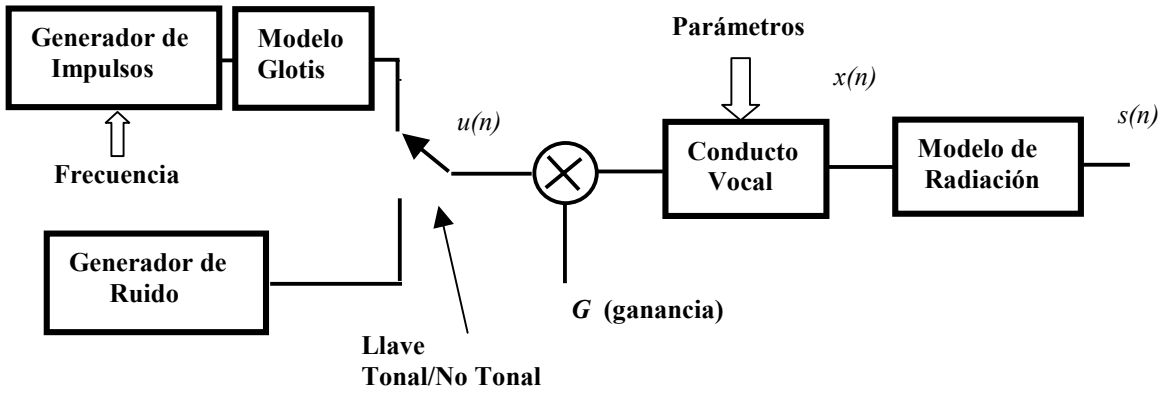

Figura 1: Modelo de Producción de voz.

2. Modelo de radiación

En la Figura 1, el **modelo de radiación** describe la impedancia de radiación vista por la presión de aire cuando abandona los labios, que puede ser razonablemente aproximada por una ecuación en diferencias de primer orden, o equivalentemente por una función transferencia de la forma

$$R(z) = \left(1 - z^{-1}\right) \tag{1}$$

3. Modelo de glotis

Existen diferentes modelos de la glotis, que han sido propuestos en la literatura, para el caso en que es excitada por pulsos. Un modelo simple es el denominado *modelo exponencial* representado por una función transferencia *Z* de la forma

$$G(z) = \frac{-ae\ln(a)z^{-1}}{\left(1 - az^{-1}\right)^2} \tag{2}$$

donde *e* es la base de los logaritmos neperianos. El numerador en (2) se selecciona de manera que $g(n) = Z^{-1}\{G(z)\}$ tenga un valor máximo aproximadamente igual a 1. El modelo está inspirado en mediciones de la respuesta de la glotis a impulsos, que se asemejan a la respuesta de un sistema de segundo orden. Una respuesta típica se representa en la figura 2.

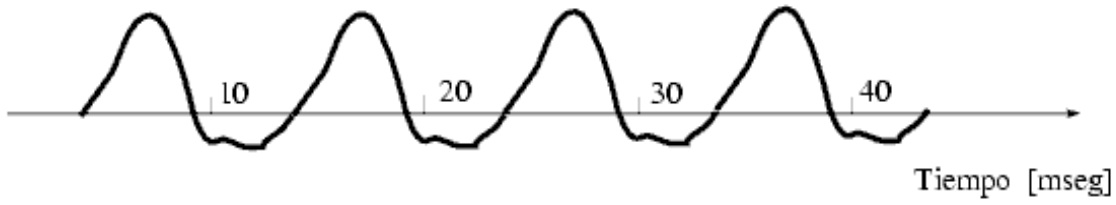

Figura 2: Respuesta típica de la glotis a una excitación con tren de impulsos.

4. Modelo Predictivo Lineal del Conducto Vocal

Un modelo matemático frecuentemente usado para el **conducto vocal** + **radiación** es una ecuación en diferencias (modelo *auto-regresivo*) que se obtiene asumiendo que cada muestra de la señal de voz está estrechamente relacionada con las muestras anteriores, de manera que el valor presente de la señal se puede obtener como una combinación lineal de, por ejemplo, *p* muestras anteriores, es decir

$$s(n) \approx -\sum_{k=1}^{p} \alpha_k s(n-k) \tag{3}$$

Incluyendo un término de excitación *Gu(n)*, la ecuación (1) puede escribirse como una igualdad de la forma

$$s(n) = -\sum_{k=1}^{p} \alpha_k s(n-k) + Gu(n) \tag{4}$$

Este modelo se denomina **Modelo de Predicción Lineal** (LPM: Linear Predictive Model) para producción/síntesis de voz, siendo los coeficientes α_k los denominados **Coeficientes de Predicción Lineal** (LPC), y G, la ganancia de excitación. El modelo LPC puede derivarse discretizando un modelo continuo de transmisión acústica basado en la concatenación de tubos acústicos sin pérdidas.
En el dominio *Z*, la ecuación (2) puede escribirse como

$$S(z) = -\sum_{k=1}^{p} \alpha_k z^{-k} S(z) + GU(z) \ , \tag{5}$$

lo que conduce a una función transferencia

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1 + \sum_{k=1}^{p} \alpha_k z^{-k}} = \frac{1}{A(z)} , \tag{6}$$

que es del tipo ***all-pole***. Una interpretación de esta ecuación está dada en la Figura 3, que es una versión simplificada (donde no se explicita el tipo de excitación) del modelo de Figura 1.

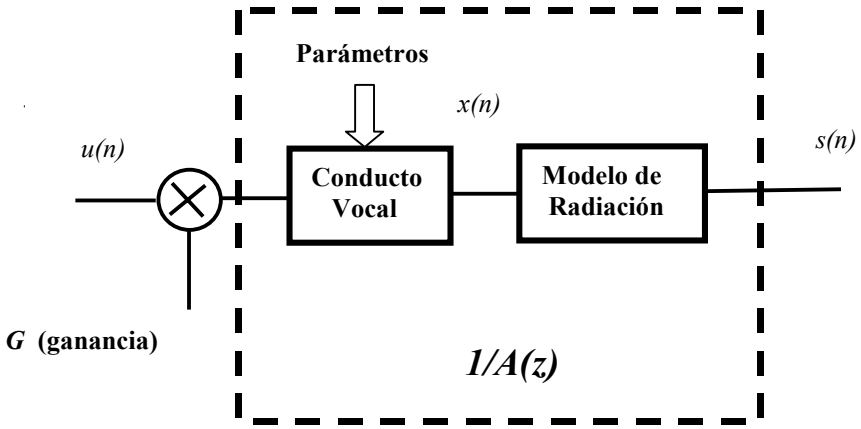

Figura 3: Modelo de Producción de voz basado en LPC.

Básicamente entonces, *H(z)* representa la función transferencia de un modelo lineal del **conducto vocal** + **radiación.** Los parámetros del filtro digital *H(z)*, son controlados por la señal de voz que está siendo producida.
Los parámetros del modelo completo de Figura 1 son entonces: la clasificación entre sonidos tonales y no tonales, la frecuencia de los sonidos tonales, la ganancia *G*, y los coeficientes del filtro digital *H(z)* (viz., los LPC).

3.4 Estimación de los LPC

De la ecuación (2) vemos que una estima (o predicción) de $s(n)$ basada en *p* muestras anteriores, puede calcularse como

$$\hat{s}(n) = -\sum_{k=1}^{p} \alpha_k s(n-k), \tag{5}$$

y el error de estimación (predicción) puede entonces definirse como

$$\varepsilon(n) = s(n) - \hat{s}(n), \tag{6}$$

resultando,

$$\varepsilon(n) = s(n) + \sum_{k=1}^{p} \alpha_k s(n-k). \tag{7}$$

Claramente, cuando la señal $s(n)$ es realmente generada por un sistema lineal como el de la Figura 3, entonces el error de estimación $\varepsilon(n)$ deberá igualar al término de excitación $Gu(n)$.

El modelo del predictor (5) se usa en telecomunicaciones para incrementar el número de señales de voz que pueden transmitirse por un canal. Si los coeficientes α_k son conocidos en el transmisor y en el receptor, entonces sólo necesita transmitirse el error y la señal de voz puede ser reconstruída en el receptor usando la ecuación en diferencias (7). En el transmisor, *s(n)* es la entrada al filtro de predicción en tanto que $\varepsilon(n)$ es la salida del filtro. Transmitir sólo la señal de error resulta en un ahorro substancial de ancho de banda del canal.

El modelo de predicción descripto puede modificarse para su uso en **síntesis de voz**. En este caso el problema básico se reduce al cálculo de los parámetros del modelo, es decir de los coeficientes de predicción lineal y de la ganancia de excitación. En la práctica, los coeficientes de predicción deben ser computados a partir de muestras de la señal de voz que se quiere sintetizar. Como la señal es inestacionaria, en el sentido que la configuración del conducto vocal cambia con el tiempo (de acuerdo al sonido que se está emitiendo), el conjunto de coeficientes se debe estimar en forma adaptable sobre cortos intervalos (típicamente de 10 ms a 30 ms de duración) denominados cuadros (en inglés ***frames***) donde se asume que la señal es estacionaria, y los LPC son constantes. Tipicamente los LPC se obtienen minimizando un criterio cuadrático en los errores de predicción $\varepsilon(n)$, para cada cuadro en que es dividido el segmento de voz.
Suponiendo que en cada cuadro hay $m+1 >> p$ muestras, la ecuación (5) puede escribirse en forma matricial como

$$\begin{bmatrix} s(n) \\ s(n+1) \\ \vdots \\ s(n+m) \end{bmatrix} = \begin{bmatrix} -s(n-1) & \cdots & -s(n-p) \\ -s(n) & \cdots & -s(n-p+1) \\ \vdots & \vdots & \vdots \\ -s(n+m-1) & \cdots & -s(n+m-1-p) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix} \tag{8}$$

o equivalentemente,

$$S_m(n) = \Phi^T(n)\alpha, \tag{9}$$

con las obvias definiciones para la **matriz de regresión** $\Phi(n)$, y el vector de **coeficientes de predicción lineal** α, y donde sin pérdida de generalidad se ha supuesto que el cuadro que se está analizando comienza en el instante n. La ecuación (9) tiene la forma de una **regresión lineal** (la linealidad es respecto al vector de parámetros α). La estima $\hat{\alpha}$ del vector de coeficientes LPC, α, que minimiza un criterio cuadrático en los

errores de predicción (7) (es decir, la **estima de mínimos cuadrados),** viene dada por (ver Apéndice)

$$\hat{\alpha} = \left(\Phi\Phi^T\right)^{-1}\Phi\mathbf{S}_m(n), \tag{10}$$

que depende únicamente de las muestras de la señal de voz.

Para poder sintetizar el segmento de señal de voz resta seleccionar la excitación correspondiente a cada cuadro donde los LPC fueron estimados. Como ya se mencionó, para el caso de sonidos tonales la excitación es un tren de impulsos periódicos, cuya frecuencia deberá también determinarse a partir de las muestras de la señal, en tanto que para sonidos no tonales es una señal de ruido aleatorio.

Apéndice

A. Estimación de Mínimos Cuadrados para Estructura de Regresor Lineal

- Se asume un modelo en tiempo discreto ya que en general los datos son recolectados por muestreo, por lo que es más simple relacionar los datos medidos a modelos en tiempo discreto.
- Por simplicidad, la notación $y(n)$ se usará para indicar la señal de salida en el instante nT (i.e., $y(nT)$), donde T es el período de muestreo (asumiendo muestreo periódico).
- Se asume que la relación entrada-salida puede ser descripta por una estructura de **regresor lineal** de la forma

$$y(n) = \varphi^T(n)\theta + v(n) \qquad \text{(A.1)}$$

donde

$\varphi(n)$: vector de regresión. Depende de los datos de entrada-salida pasados hasta el instante $n-1$.

$\theta \in D_M \subset \Re^p$: vector de parámetros a estimar.

$v(n)$: es una perturbación que trata de modelar lo que de los datos no puede ser explicado por el término $\varphi^T(n)\theta$. En general se da a este término una caracterización estadística, es decir se asume que es un proceso aleatorio, pero no necesariamente debe ser así.

- El apelativo **lineal**, refiere a que depende linealmente del vector de parámetros θ.

- Un **ejemplo típico** de estructura de modelo que puede ponerse en la forma de regresor lineal es una estructura del tipo **ARX** (**A**uto **R**egressive with e**X**ogenous inputs), que es una ecuación lineal en diferencias de la forma

$$y(n)+a_1y(n-1)+\cdots+a_Py(n-P)=b_1u(n-1)+\cdots+b_Mu(n-M) \qquad \text{(A.2)}$$

o equivalentemente

$$y(n)=-a_1y(n-1)-\cdots-a_Py(n-P)+b_1u(n-1)+\cdots+b_Mu(n-M) \qquad \text{(A.3)}$$

Definiendo el **vector de parámetros**

$$\theta = [a_1 \cdots a_P b_1 \cdots b_M]^T,$$

y el **vector de regresión**

$$\varphi(n) = [-y(n-1)\cdots-y(n-P)u(n-1)\cdots u(n-M)]^T,$$

la ecuación (A.3) puede escribirse en la forma (A.1) (i.e., una regresión lineal).

- En general, los parámetros son desconocidos y lo que se pretende es estimarlos a partir de datos de medición, de manera que se pueda predecir la salida. Basado en la estructura de modelo (A.1), se define un predictor de la salida que permite computar

una estima de la salida en el instante *n* basado en los datos hasta el instante *(n-1)*, y en la estima del vector de parámetros. El predictor se suele denotar

$$\hat{y}(n \mid n-1,\theta) = \varphi^T(n)\theta \qquad \text{(predictor)}$$

- Se asume que para la estimación está disponible un conjunto $\{u(n), y(n)\}_{n=1}^{N}$ de N datos de entrada-salida.
- Se define el **error de predicción**

$$\varepsilon(n,\theta) = y(n) - \varphi^T(n)\theta$$

- Se propone un **criterio cuadrático** en el error de predicción, de la forma

$$V_N(\theta) = \frac{1}{N}\sum_{k=1}^{N} Tr\left\{\left[y(k) - \varphi^T(k)\theta\right]\left[y(k) - \varphi^T(k)\theta\right]^T\right\}$$

- La estimación consiste en hallar la estima $\hat{\theta}_N$ que minimiza el criterio $V_N(\theta)$.
- $\hat{\theta}_N$ se denomina **estima de mínimos cuadrados**, y viene dada por

$$\hat{\theta}_N = \underset{\theta}{argmin}\{V_N(\theta)\} = \left[\sum_{k=1}^{N}\varphi(k)\varphi^T(k)\right]^{-1}\left[\sum_{k=1}^{N}\varphi(k)y(k)\right]$$

que se obtiene igualando a zero la derivada de $V_N(\theta)$ con respecto a θ, i.e.

$$0 = \frac{d}{d\theta}V_N(\theta) = 2\sum_{k=1}^{N}\varphi(k)\left(y(k) - \varphi^T(k)\theta\right)$$

- Vectorizando

$$Y = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix}; \qquad \Phi^T = \begin{bmatrix} \varphi^T(1) \\ \varphi^T(2) \\ \vdots \\ \varphi^T(N) \end{bmatrix};$$

el modelo (A.1) puede escribirse:

$$Y = \Phi^T\theta$$

y la estima de mínimos cuadrados resulta:

$$\hat{\theta}_N = \left(\Phi\Phi^T\right)^{-1}\Phi Y = \Phi^+ Y$$

donde Φ^+ es la denominada **Moore-Penrose pseudo inversa**.

- Para que exista la estima de mínimos cuadrados, la matriz $\Phi\Phi^T$ debe ser **no singular**. Esto exige condiciones de **persistencia de excitación (PE) de los regresores** $\varphi(n)$**.** Se dice que los regresores son **PE** si existe un entero ℓ_0 y constantes positivas α_1 y α_2 tal que

$$\alpha_2 I \geq \sum_{n=n_0}^{n_0+\ell_0} \varphi(n)\varphi^T(n) \geq \alpha_1 I > 0$$

- Como en general los regresores dependen de las entradas pasadas, entonces la PE de los regresores requiere la PE de las entradas. Intuitivamente es claro que sólo pueden identificarse aquellos modos del sistema que son observables a la salida y que son sufientemente excitados desde la entrada.
- En el caso en que la perturbación $v(n)$ sea caracterizada como un proceso estocástico, entonces la estima $\hat{\theta}_N$ resulta una variable aleatoria, y bajo ciertas condiciones sobre los regresores y la perturbación, puede probarse que es **consistente** en el sentido que

$$\hat{\theta}_N \xrightarrow{a.s} \theta_0 \qquad \text{(léase converge con probabilidad 1)}$$

cuando $N \to \infty$, donde θ_0 es el verdadero valor del vector de parámetros.

B. Obtención del modelo LPC a partir de las ecuaciones de propagación del sonido en el tracto vocal

Las ecuaciones que describen la propagación de la onda sonora en el tracto vocal, modelado como un tubo de sección variable sin pérdidas, son (ver [2]):

$$\frac{\partial p}{\partial x} = -\frac{\rho}{A}\frac{\partial u}{\partial t} \tag{B.1}$$

$$\frac{\partial u}{\partial x} = -\frac{A}{\rho c^2}\frac{\partial p}{\partial t} \tag{B.2}$$

donde

$p(x,t)$	presión
$u(x,t)$	velocidad de volumen
$A(x)$	áerea transversal del tracto vocal
c	velocidad del sonido
ρ	peso específico del aire

De las ecuaciones anteriores es posible eliminar u , resultando

$$\frac{\partial}{\partial x}\left(A\frac{\partial p}{\partial x}\right) = \frac{A}{c^2}\frac{\partial^2 p}{\partial t^2} \tag{B.3}$$

que es conocida como la *ecuación de la corneta* o *ecuación de la bocina,* de Webster. Tomando las transformadas de Laplace, las ecuaciones (B.1), (B.2), y (B.3) resultan

$$\frac{\partial P}{\partial x} = -\frac{\rho s}{A}U \tag{B.4}$$

$$\frac{\partial U}{\partial x} = -\frac{As}{\rho c^2}P \tag{B.5}$$

$$\frac{\partial}{\partial x}\left(A\frac{\partial P}{\partial x}\right) = \frac{As^2}{c^2}P \tag{B.6}$$

donde $P(x,s)$ and $U(x,s)$ son las transformadas de Laplace de $p(x,t)$ y $u(x,t)$, respectivamente.

Consideraremos ahora una aproximación del modelo del tubo de sección transversal variable como la concatenación de tubos de sección uniforme, entonces el area A en cada tubo es independiente de *x,* y la ecuación (B.6) para un tubo puede escribirse como:

$$\frac{\partial^2 P}{\partial x^2} = \frac{s^2}{c^2}P \tag{B.7}$$

que tiene dos soluciones independientes de la forma $e^{\frac{s}{c}x}$ y $e^{-\frac{s}{c}x}$, por lo que puede escribirse

$$P(x,s) = ae^{\frac{s}{c}x} + be^{-\frac{s}{c}x}. \qquad \text{(B.8)}$$

Considerando ahora la ecuación (B.4), resulta entonces

$$U(x,s) = -\frac{Aa}{\rho c}e^{\frac{s}{c}x} + \frac{Ab}{\rho c}e^{-\frac{s}{c}x} \qquad \text{(B.9)}$$

Las ecuaciones (B.8) y (B.9) permiten relacionar las presiones y velocidades volumétricas a la entrada y salida de un tubo de sección uniforme. A la entrada del tubo es $x = 0$, en tanto que a la salida es $x = L$, resultando

$$P_{in} = a + b, \qquad P_{out} = ae^{\frac{s}{c}L} + be^{-\frac{s}{c}L},$$

$$U_{in} = -\frac{Aa}{\rho c} + \frac{Ab}{\rho c}, \qquad U_{out} = -\frac{Aa}{\rho c}e^{\frac{s}{c}L} + \frac{Ab}{\rho c}e^{-\frac{s}{c}L}$$

Estas cuatro ecuaciones pueden escribirse en forma matricial como

$$\begin{bmatrix} P_{in} \\ U_{in} \end{bmatrix} = \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} \begin{bmatrix} P_{out} \\ U_{out} \end{bmatrix} = K \begin{bmatrix} P_{out} \\ U_{out} \end{bmatrix},$$

donde K es la denominada matriz *ABCD* del tubo de sección uniforme.