Modelos de Producción de Voz

Juan Carlos Gómez

Presentación basada en las siguientes **Referencias**:

[1] Rabiner, L. & Juang, B-H.. *Fundamentals of Speech Recognition,* Prentice Hall, N.J., 1993.

[2] Rabiner, L. & Juang, B-H.. *Speech Recognition by Machine*, Chap. 47 in *The Digital Signal Processing Handbook*, CRC Press, IEEE Press, 1998.

La señal de voz se produce cuando una columna de aire desde los pulmones excita el tracto vocal, que se comporta como una cavidad resonante.

Los distintos sonidos emitidos pueden clasificarse en:

Tonales (voiced): se produce la oscilación relajada de las cuerdas y puede pensarse que el tracto vocal es excitado por pulsos de aire cuasi-periódicos. Ejemplos son las vocales y varias consonantes.

No tonales (unvoiced): aquí las cuerdas permanecen abiertas y se insufla aire a una velocidad tal que se produce turbulencia. En este caso puede pensarse que el tracto vocal es excitado por una fuente de ruido aleatorio. Ejemplos son algunas consonantes.

Cuando se emiten sonidos tonales, las cuerdas vocales forman un oscilador de relajación. El aire fluye desde los pulmones, se incrementa la presión y las cuerdas se separan, permitiendo que fluya a traves de ellas, disminuyendo la presión y permitiendo que se vuelvan a cerrar, repitiéndose el ciclo.

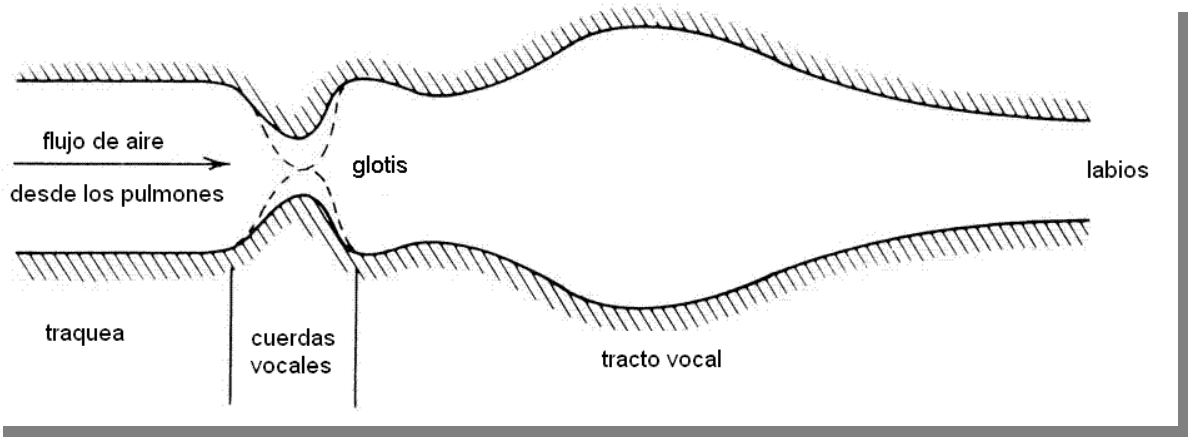
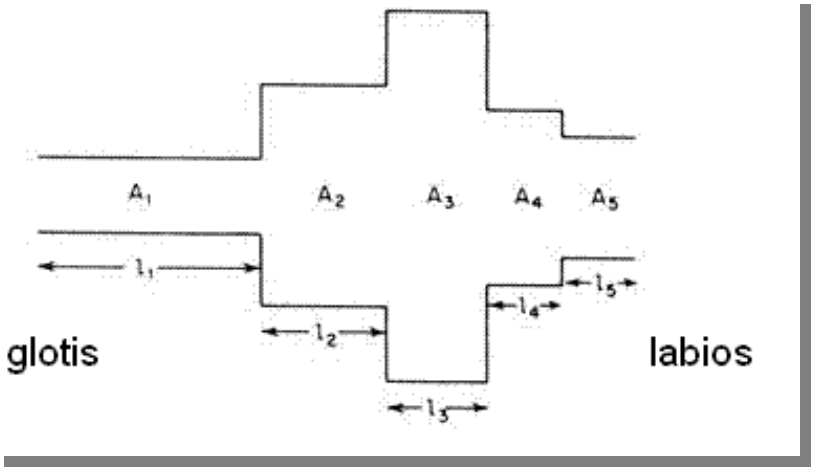
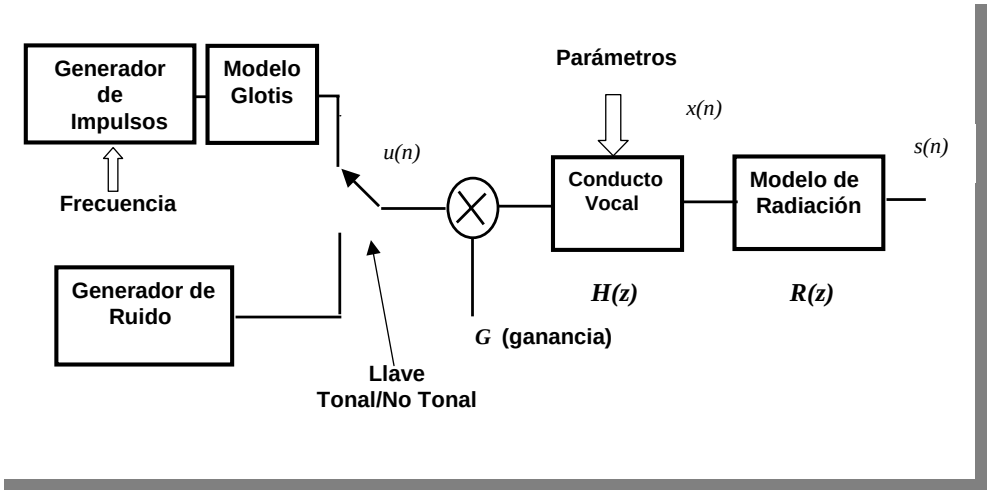


Diagrama esquemático del conducto traqueo-laringeo-vocal

El tracto vocal es modelado como la concatenación de tubos acústicos de distinto diámetro (con o sin pérdidas). Esto deriva en un **modelo lineal inestacionario** (ya que las secciones de los tubos van cambiando de acuerdo al fonema que se está emitiendo).



Modelo del tracto vocal



Modelo de Producción/Síntesis de Voz

Modelo de Radiación

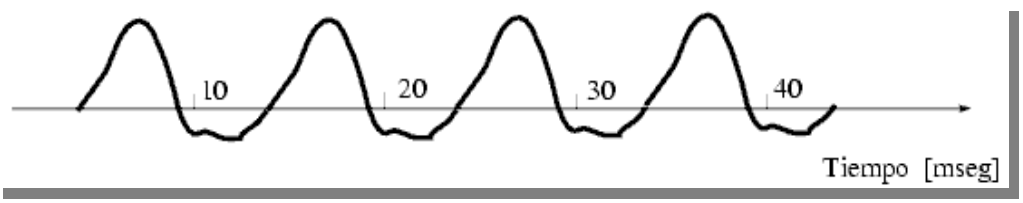
$$R(z) = \left(1 - \alpha\, z^{-1}\right)$$

El **modelo de radiación** describe la impedancia de radiación vista por la presión de aire cuando abandona los labios.

Modelo de Glotis

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

Existen diferentes modelos de la respuesta de la glotis, cuando es excitada por un tren de pulsos.



Respuesta típica de la glotis a un tren de pulsos

Modelo Predictivo Lineal del Tracto Vocal

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

Para intervalos cortos de tiempo (10 a 30 mseg) puede pensarse que el tracto vocal es un sistema estacionario y que se puede predecir la señal de voz en el instante *n* a partir de los valores de la señal en *p* instantes anteriores.

Si se incluye un término de excitación se obtiene el **Modelo Predictivo Lineal**

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

donde α_k son los denominados **Coeficientes de Predicción Lineal** (LPC: Linear Predictive Coefficients).

El modelo LPC puede derivarse discretizando un modelo continuo de transmisión acústica basado en la concatenación de tubos acústicos sin pérdidas.

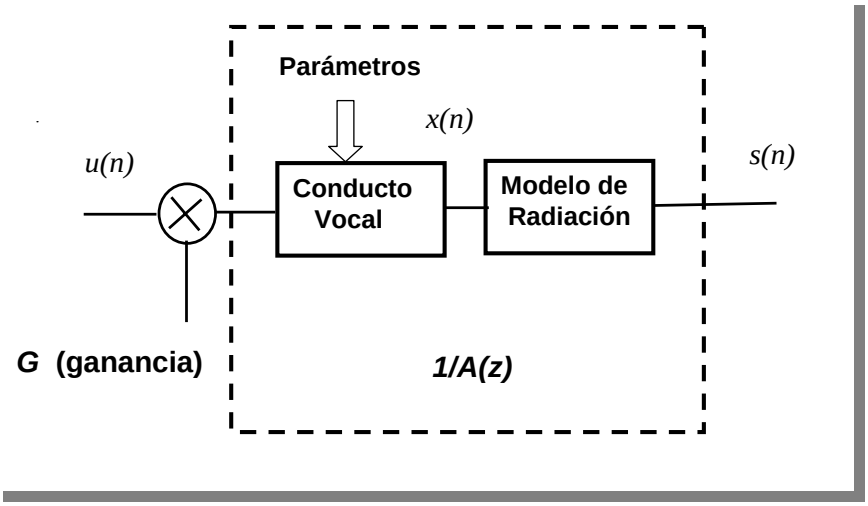
En el dominio *Z* se tiene

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

que conduce a la función transferencia

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

que se denomina sólo polo (**all pole**).



Modelo LPC del Tracto Vocal + Radiación

Estimación de los LPC

Un **predictor** de la señal de salida en el instante *n* puede calcularse como

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

por lo que el **error de predicción** $\varepsilon(n) = s(n) - \hat{s}(n)$ resulta

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

La idea, para síntesis de voz, es **estimar** los LPC a partir de muestras de la señal que se quiere sintetizar, minimizando una **función de costo** o **criterio**, que es una función definida positiva del error de predicción. Típicamente es una función cuadrática del error, por lo que la estima de los coeficientes resulta la denominada **Estima de Mínimos Cuadrados.**

Definiendo

$$\alpha = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_p \end{bmatrix}^T$$
$$\varphi^T(n) = \begin{bmatrix} -s(n-1) & -s(n-2) & \cdots & -s(n-p) \end{bmatrix}$$

Coeficientes LPC

regresor

resulta

$$\varepsilon(n) = s(n) - \varphi^T(n)\alpha$$

Definiendo un criterio cuadrático en el error de predicción y asumiendo que se tienen *N* muestras, resulta

$$V_N(\alpha) = \frac{1}{N}\sum_{n=1}^{N}\varepsilon^2(n) = \frac{1}{N}\sum_{n=1}^{N}\left[s(n) - \varphi^T(n)\alpha\right]^T\left[s(n) - \varphi^T(n)\alpha\right]$$

la estima de mínimos cuadrados se obtiene igualando a cero la derivada respecto a α del criterio, es decir

$$\hat{\alpha}^{LS} = \underset{\alpha}{\operatorname{argmin}}\, V_N(\alpha) \quad \text{tal que} \quad \frac{\partial V_N(\alpha)}{\partial \alpha} = 0$$

que resulta en

$$\hat{\alpha}^{LS} = \left[\frac{1}{N}\sum_{n=1}^{N}\varphi(n)\varphi^T(n)\right]^{-1}\left[\frac{1}{N}\sum_{n=1}^{N}\varphi(n)s(n)\right]$$

Estima de Mínimos Cuadrados

Considerando que la señal está dividida en segmentos (frames) de corta duración (10 a 30 mseg) donde se asume que no cambian las características del tracto vocal, puede escribirse en forma matricial para un frame genérico

$$\underbrace{\begin{bmatrix} s(n) \\ s(n+1) \\ \vdots \\ s(n+N-1) \end{bmatrix}}_{S_{N,n}} = \underbrace{\begin{bmatrix} -s(n-1) & \cdots & -s(n-p) \\ -s(n) & \cdots & -s(n-p+1) \\ \vdots & \vdots & \vdots \\ -s(n+N-2) & \cdots & -s(n+N-2-p) \end{bmatrix}}_{\Phi^{T}_{N,n}} \underbrace{\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix}}_{\alpha}$$

$$S_{N,n} = \Phi_{N,n}{}^{T}\,\alpha$$

y la estima de mínimos cuadrados resulta

$$\hat{\alpha}^{LS} = \left(\Phi_{N,n}\Phi_{N,n}{}^{T}\right)^{-1}\Phi_{N,n}S_{N,n}$$

$$\Phi^{+}{}_{N,n} = \left(\Phi_{N,n}\Phi_{N,n}{}^{T}\right)^{-1}\Phi_{N,n}$$

Pseudo-inversa izquierda de Moore-Penrose

La pseudo inversa existe siempre que los regresores sean **persistentes** (el concepto de persistencia de excitación está relacionado con la riqueza armónica de la señal).