Síntesis articulatoria del habla humana

Josep Martí
Departamento de Acústica de Ingeniería La Salle.
Universidad RAMON LLULL

Modelo de la producción del habla

El modelo más utilizado para explicar la generación del habla responde al conocido esquema generador-filtro, según el cual la señal de voz se interpreta como una convolución entre la señal producida por un generador y la respuesta impulsional de un filtro variable en el tiempo (Fig. 1). El generador corresponde a la señal periódica y casi triangular producida por las cuerdas vocales y/o la señal aleatoria generada por la turbulencia del aire al pasar por una constricción del conducto bucal. El filtro variable es la cavidad resonante constituida por el tracto bucal (y nasal en el caso de los sonidos nasales) que determina la articulación de la cadena de sonidos constituyentes del habla humana. Por efecto de radiación hay que añadir además un realce de las frecuencias más altas, con una pendiente aproximada de + 6 dB/octava.

Este esquema es fácilmente implementable mediante sistemas digitales, que pueden simular el proceso en la medida que se conozcan los datos necesarios para su actualización en tiempo real. En este artículo presentamos las últimas tecnologías basadas en este método para la transmisión de voz codificada y su aplicación a los conversores texto - habla.

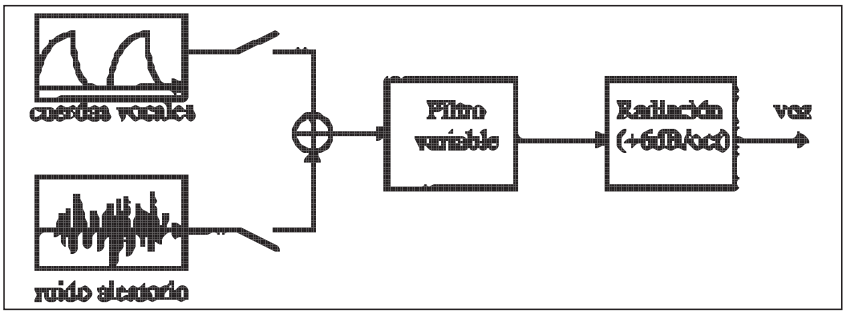

Fig nº 1.- Modelo básico del sistema fonatorio.

Nº	Descripción	Límite inferior	Límite superior
1	radio de la lengua (cm)	0,5	3,5
2	ángulo de la mandíbula (grados)	10,9	28,0
3	centro de la lengua (cm)	4,0	10,0
4	paletilla de la lengua (cm)	2,0	4,4
5	avance de los labios (cm)	-0,36	1,86
6	posición x del hioides (cm)	5,20	7,60
7	centro de la lengua (cm)	1,48	7,48
8	elevación de la lengua (grados)	68,2	96,8
9	altura del labio inferior	-0,93	1,32
10	posición y del hioides (cm)	7,43	9,83
11	abertura del velo (cm^2)	0	1,0

Tabla 1. Modelo articulatorio de Mermelstein.

Modelo de tracto bucal

La caracterización del filtro variable se realiza a través de los parámetros que definen la posición de los órganos articulatorios. Según el modelo propuesto por Mermelstein (MERMELSTEIN 1973) estos parámetros son los que se indican en la tabla I. Coker propuso un modelo más simplificado (Tabla II) y que, en algunos casos, parece que ha dado mejores resultados (GUPTA 1993).

El conocimiento preciso de estos parámetros permite realizar un trazado de la sección sagital del tracto bucal según puede verse en la Fig. 2. A partir de esta representación y por un sistema de segmentación como el que se indica en la misma figura se determinan los segmentos obtenidos de la proyección de las sucesivas áreas coronarias sobre el plano sagital. El valor de la sucesión de áreas se determina a partir de elipses cuyos ejes verticales son estos segmentos y cuyos ejes horizontales se determinan matemáticamente por funciones deducidas de diferentes estudios fisiológicos, muchos de ellos basados en imágenes de rayos X (HEINZ 1964) (LADEFOGET 1971) (LINDBLOM 1971). Últimamente están apareciendo estudios mucho más precisos basados en imágenes obtenidas por resonancia magnética (NA-

Nº	Descripción	Límite inferior	Límite superior
1	posición x de la lengua	-1,3	1,2
2	posición y de la lengua	-1,3	2
3	retroceso de la punta de la lengua	-0,6	2,6
4	altura de la punta de la lengua	0,0	1,5
5	apertura de los labios	0,0	2,0
6	redondeo de los labios	-0,6	2,6
7	área de la faringe	1,0	3,0
8	parámetro de constricción	0.0	0,6
9	apertura del velo (cm^2)	0,0	1,0

Tabla 2. Modelo articulatorio de Coker.

RAYANAN 1995, 1997) (BAER 1991) (DANG 1993) y por métodos electromagnéticos (PERKERELL 1992). Igualmente se están ensayando métodos para la obtención de la forma del tracto a partir del mismo sonido por el denominado "proceso inverso" que incorpora técnicas de cuantificación vectorial (LARAR 1988) y redes neuronales (PAPCUN, 1992) (McGOWAN, 1995) (SCHROETER, 1994). Se trata de un sistema de parametrización en tiempo real para aplicaciones de transmisión de voz comprimida.

A partir de esta información se puede construir un modelo discretizado de tracto bucal a base de tubos cilíndricos de longitud constante y secciones A1, A2, ... (Fig. 3); que no es más que un filtro acústico con una función de transferencia fácilmente deducible.

Los conversores texto-habla

La filosofía de un convesor texto-habla es distinta a la de un sistema de transmisión de voz, ya que no precisa de una parametrización en tiempo real a partir de la señal acústica. La información básica para obtener la dinámica del sistema es la posición de los articuladores para cada sonido en la lengua que se pretende sintetizar. Después, por un proceso de interpolación, se construye la sucesión temporal de tractos bucales que dan todo el proceso de articulación y coarticulación entre sonidos contiguos de forma idéntica a la articulación natural. La interpolación se puede realizar de forma puramente lineal o bien a partir de la función arco tangente, que puede añadir una ligera mejora al sistema (GUPTA, 1993).

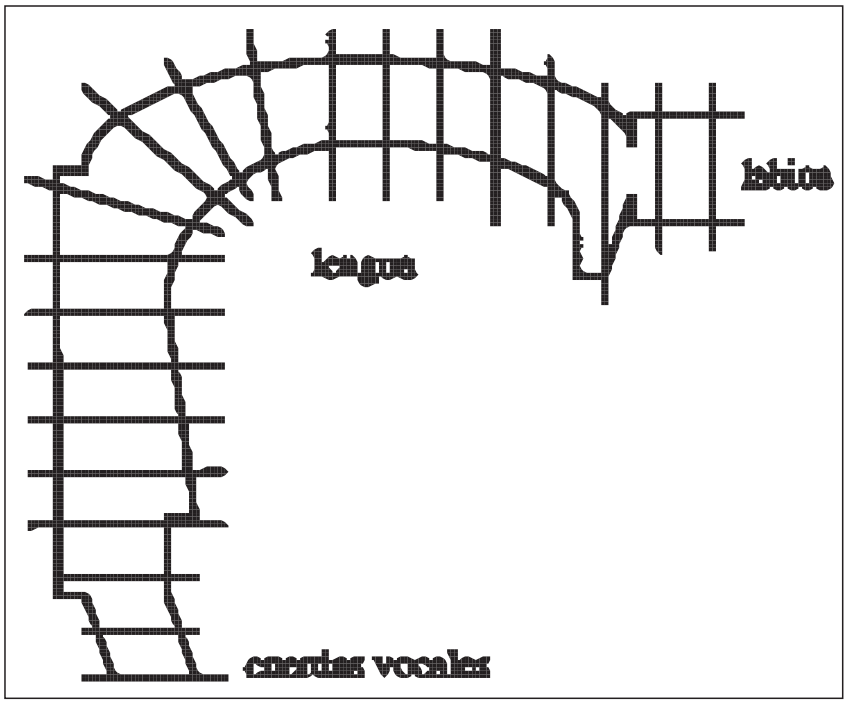

Fig. 2. Sección sagital del tracto bucal dividida en 20 segmentos.

Estos tubos acústicos, incluidos los que corresponden a la derivación nasal (de tamaño fijo), permiten simular la propagación de ondas acústicas con sus reflexiones y transmisiones en forma de filtro digital (Fig. 4). Los parámetros del filtro dependen de la sección de cada tubo, de su longitud y de las sección del tubo posterior, según la estructura de Kelly-Lochbaum (KELLY 1962). El coeficiente de reflexión de velocidad volumétrica entre un tubo y el siguiente es: $r_i = (A_{i+1} - A_i) / (A_{i+1} + A_i)$. Si tenemos la precaución de igualar la longitud de todos los tubos entre sí, el paso de los frentes acústicos por las discontinuidades entre tubos sucesivos se realizará de forma sincrónica, de manera que obtendremos muestras de la onda de salida con un período de muestreo igual al doble del tiempo necesario para la propagación en cada tubo. El funcionamiento del tracto nasal se controla mediante la apertura o cierre del velo del paladar. En cada tubo se puede introducir un factor de pérdidas según la expresión:

$\alpha^{1/2} = 1 - 0{,}007 \cdot A^{1/2}$

(A = área del tubo en cm^2)

(FLANAGAN, 1972)

La excitación se realiza mediante una señal aproximadamente triangular inyectada desde el principio del conducto para el caso de los sonidos periódicos; y mediante un generador de ruido situado inmediatamente después de la constricción del tracto bucal para los sonidos aleatorios. Todo ello se realiza mientras se producen las fluctuaciones de la articulación de la cadena de sonidos. Finalmente sumando la radiación bucal y la radiación nasal se obtiene el sonido resultante en cada momento con un buen efecto de continuidad y de naturalidad.

En definitiva, la conversión texto-habla precisa de una base de datos relativamente reducida, que contiene la posición de los órganos articulatorios correspondiente a cada alófono en su parte estacionaria. En el caso de fonemas poco estacionarios habrá que disponer de las posiciones sucesivas del

tracto bucal mientras dura la realización del sonido. El sistema está dando unos resultados muy satisfactorios por lo que se refiere a la continuidad de la onda acústica y a su nivel de naturalidad. La calidad de las unidades acústicas generadas depende básicamente de la precisión con que se conozcan los datos geométricos del tracto bucal. Con todo, no hay que olvidar que la contribución más importante a la naturalidad del habla sintética proviene de una buena definición de los parámetros prosódicos (melodía, ritmo y energía) que obedecen a un proceso suprasegmental independiente de la forma concreta de generación y concatenación de unidades básicas.

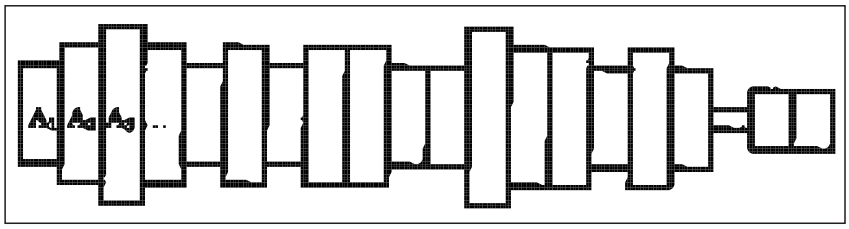

Fig. 3. Modelo de tracto bucal con tubos cilíndricos concatenados.

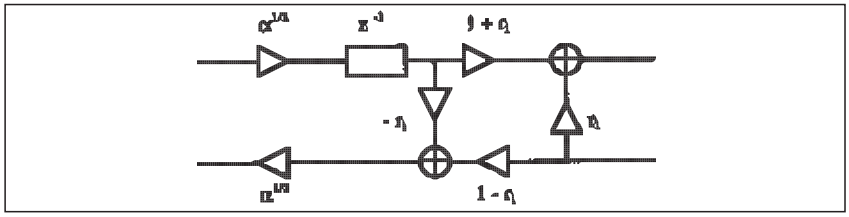

Fig. 4. Modelo digital de cada uno de los tubos del tracto bucal.

Referencias

BEAR, T.; GORE, J. C.; GRACCO, L. C.; NYE, P. W. "Analysis of vocal tract shape and dimensions using magnetic resonance imaging. Vowels". JASA 90 (2), pp. 799-828 (1991).

COKER, C. H. "A model of articulatory dynamics and control". Proc. IEEE, vol 64, pp. 452-460 (1976).

DANG, J.; HONDA, K.; SUZUKI, H. "MRI measurements and acoustic investigation of the nasal and paranasal cavities" JASA 94, pp. 1765 (A) (1993).

FLANAGAN, J. L. "Speech analysis, synthesis and perception". Springer - Verlag. New York (1972).

GUPTA, S. K.; SCHROETER, J. "Pitch-synchronous frame-by-frame and segment-based articulatory analysis by synthesis". JASA 94 (5), pp. 2517 - 2530 (1993).

HEINZ, J. M.; STEVENZ, K. N. "On the derivation of area functions and acoustic spectra from cineradiographic films of speech". JASA, 36, pp. 1037 ss. (1964).

KELLY, J. L.; LOCHBAUM, C. C. "Speech synthesis". Proc. 4 Int. Congr. Acoustic, Paper G-42, pp. 1-4 (1962).

LADEFOGED, P.; ANTHONY, J.; RILEY, D. "Direct measurements of the vocal tract". JASA, 49, pp. 104 ss. (1971).

LARAR, J. N.; SCHROETER, J.; SONDHI, M. M. "Vector quantization of the articulatory space". IEEE Trans. on ASSP vol 36, nº 12, pp. 1812 - 1818 (1988).

McGOWAN, R., KOENING, L., LˆFQUIST, A. "Vocal tract aerodynamics in /aCa/ utterances: Simulations". Speech Communication 16, pp. 67-68 (1995).

MERMELSTEIN, P. "Articulatory model for the study of speech production". JASA, 53, pp. 1070-1082 (1973).

NARAYANAN, S. S.; ALWAN, A. A.; HAKER, K. "An articulatory study of fricative consonants using magnetic resonance imaging". JASA 98 (3), pp. 1325-1347 (1995).

NARAYANAN, S. S.; ALWAN, A. A.; HAKER, K. "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The laterals". JASA 101 (2), pp. 1064-1077 (1997).

NARAYANAN, S. S.; ALWAN, A. A.; HAKER, K. "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics". JASA 101 (2), pp. 1078-1089 (1997).

PAPCUN, G.; HOCHBERG, J.; THOMAS, T. R.; LAROCHE, F.; ZACKS, J. "Inferring articulation and recognising gestures from acoustics with a neural network trained on X-ray microbeam data". JASA 92 (2) pp. 688 - 700 (1992).

PERKELL J. S.; COHEN, M. H.; SVIRSKY, M. A.; MATTHIES, M. L.. "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements". JASA 92 (6), pp. 3078 - 3096 (1992).

SCHROETER, J.; SONDHI, M. M. "Techniques for estimating vocal-tract shapes from speech signal". IEEE Transactions on speech and audio processing, vol 2, n. 1, pp. 133 - 150 (1994).