SEÑALES Y SISTEMAS

(66.74 - 85.05)

TRABAJO PRÁCTICO ESPECIAL

Análisis y procesamiento de la señal de habla

Segundo cuatrimestre de 2016

Facultad de Ingeniería - Universidad de Buenos Aires

Índice

Objetivo

El presente proyecto especial tiene como objetivo hacer uso de técnicas y herramientas de análisis de señales y sistemas, aplicándolas al análisis y procesamiento de la señal de habla.

Requisitos para la aprobación

El proyecto especial tendrá una fecha límite de presentación electrónica y su posterior evaluación está establecida en el calendario de la materia, día en el cual el alumno deberá presentarse indefectiblemente con el informe del proyecto en forma impresa.

Habrá un rango de fechas anterior a la fecha definitiva de entrega en el cual el alumno podrá hacer una pre-entrega del proyecto especial en versión electrónica. Durante ese rango de fechas el docente puede aconsejar al alumno la revisión de ciertos puntos en el proyecto. Luego del cierre del período de pre-entrega, se habilitará un período de entrega definitiva, donde el alumno debe depositar su versión electrónica del informe y los algoritmos correspondientes.

Luego del vencimiento del período de entrega definitivo no se admitirán más entregas y el alumno que no cumpla este requisito quedará libre.

Luego el docente de cada curso evaluará el mismo en tiempo y forma utilizando la versión electrónica o la impresa, y asentará en la versión impresa la nota del proyecto. La modalidad de la evaluación se realizará según el docente lo crea conveniente (oral, escrita, el día de la entrega, otro día, etc.), de modo de asegurar el conocimiento del tema desarrollado y la realización individual del trabajo por parte del alumno. El trabajo sólo podrá ser presentado y evaluado en el curso en el cual el alumno se haya inscripto.

La evaluación final puede incluir preguntas sobre:

- Ítems particulares sobre los ejercicios de esta guía y su implementación en Matlab/Octave.
- Conceptos teóricos necesarios para realizar los ejercicios. Puede requerirse también al alumno que implemente alguno de los ejercicios similares en la computadora en el momento de la evaluación.

Por lo tanto el alumno debe presentarse el día de la evaluación con:

- Esta guía.
- Las soluciones a los problemas planteados: Cuando el problema requiera una implementación, la misma debe estar adecuadamente descripta y debidamente justificada. Es decir, si es necesario justificación teórica, ésta debe estar desarrollada. Si se pide una implementación práctica la misma debe estar adecuadamente documentada de modo que el docente pueda constatar que las especificaciones requeridas se cumplan. Esto incluye la presentación del programa de Matlab/Octave utilizado, y los gráficos necesarios para mostrar los resultados obtenidos en formato electrónico e impresos. Se sugiere que el formato electrónico no dependa de que funcione internet para poder verse, para evitar inconvenientes. Todos los gráficos deberán tener título, comentarios en ambos ejes sobre la unidad a representar y el eje de abscisas debe estar en unidades de tiempo o frecuencia según corresponda.

Introducción

El habla es el medio natural de comunicación entre personas y es una de las principales características que nos distinguen del resto de los animales. Desde tiempos inmemoriales el hombre necesitó comunicarse, siendo el habla una de las herramientas más importantes para su desarrollo intelectual. Es por eso que en este afán de comprender y reproducir dicha capacidad humana se han realizado diversos estudios originando varias disciplinas tales como el reconocimiento automático del habla (ASR, del inglés Automatic Speech Recognition) o la conversión de texto en habla (TTS, del inglés Text-To-Speech), reconocimiento del locutor, detección de patologías orales, entre otras. El análisis de habla en general conforma un área de investigación muy importante dentro del procesamiento de señales debido a sus aplicaciones en multimedia, telecomunicaciones, entre otros.

Sistema de producción de voz

El aparato fonador se puede considerar como un sistema que transforma energía muscular en energía acústica. La voz es una onda de presión sonora que es originada por los movimientos de las estructuras anatómicas que forman el sistema humano de producción de la voz, al cual denominamos aparato fonador. En la Figura 1 se puede observar un esquema del corte sagital medio del aparato fonador.

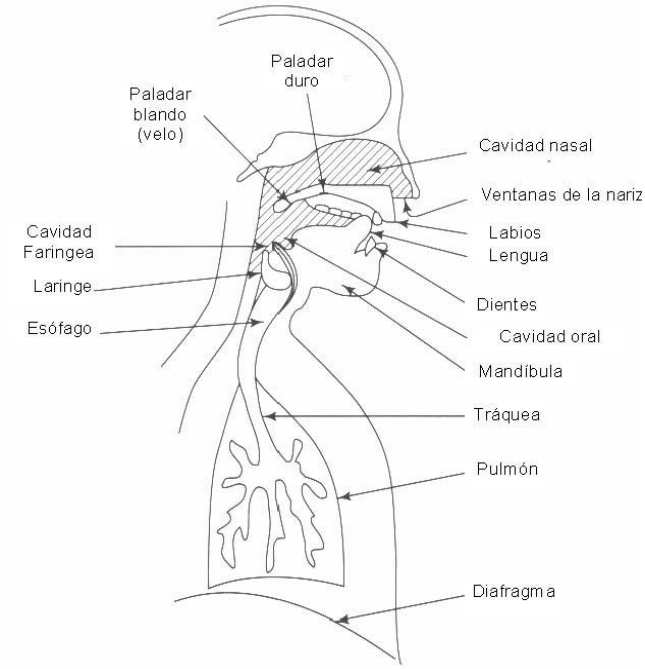

Figura 1: Esquema del corte sagital medio del tronco superior, donde se puede observar el aparato fonador y respiratorio. Adaptado de Deller et al. 1993.

Los pulmones cumplen la función de fuente de energía. El accionar del músculo diafragma expulsa el aire contenido en los pulmones en la dirección del tracto vocal. El tracto vocal es el conducto que se extiende desde las cuerdas vocales a los labios, con una derivación a la cavidad nasal. Su longitud es de aproximadamente 17 cm, aunque esto puede variar ligeramente.

La laringe contiene a los pliegues vocales, mal conocidos como cuerdas vocales, que están formadas por pliegues de ligamento que se extienden desde el cartílago tiroides a los cartílagos aritenoides (Ver Figura 5). La apertura en forma de V entre las cuerdas vocales, llamada glotis, es la fuente de sonido más importante en el sistema vocal. Las cuerdas vocales pueden actuar de varios modos diferentes durante el habla. Su función más importante es modular el flujo de aire abriendo y cerrándose, causando el sonido que produce las vocales y las consonantes sonoras.

La faringe une la laringe con la cavidad bucal. Tiene dimensiones cuasi fijas, pero su longitud puede variar ligeramente levantando o bajando la laringe en uno de sus extremos y el paladar blando en el otro. Éste último también aísla o une la ruta de la cavidad nasal a la faringe. En el extremo de la faringe se encuentran la epiglotis y los pliegues vestibulares, también conocidos como falsas cuerdas vocales, que son las encargadas de prevenir que el alimento alcance la laringe y aíslan acústicamente el esófago del tracto vocal. La epiglotis, las falsas cuerdas vocales y las cuerdas vocales se cierran al tragar y se abren durante la respiración normal (Ver Figura 2).

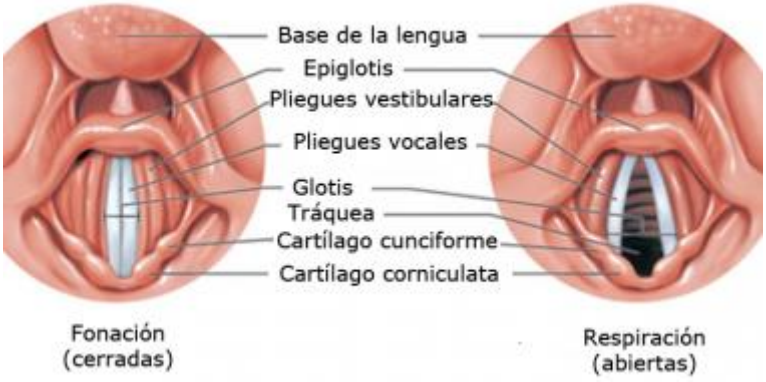

Figura 2: Posición de las cuerdas vocales para hablar y respirar.

La cavidad bucal es una de las partes más importantes del tracto vocal. Su tamaño, forma y acústica pueden ser variados por los movimientos del paladar, la lengua, los labios, las mejillas y los dientes. A este conjunto de estructuras móviles se las denomina articuladores. La extremidad blanda del velo es la úvula. Los componentes anatómicos finos se mueven en diferentes posiciones para producir los diferentes sonidos de la voz. Los labios controlan el tamaño y la forma de la apertura de la boca por la cual el sonido es irradiado. La mandíbula también es considerada un articulador, dado que es la responsable de los movimientos finos y gruesos que afectan el tamaño y la forma del tracto vocal, así como también la posición de los demás articuladores.

A diferencia de la cavidad bucal, la cavidad nasal tiene dimensiones y forma fija. Su longitud es de aproximadamente 12 cm y su volumen es de 60 cm^3. El tracto nasal comienza en el velo (paladar blando) y termina en los orificios de la nariz.

La producción de voz y su modelización

La producción de la voz puede ser modelada en términos de una operación de filtrado acústico, la cual asocia las partes anatómicas con un modelo teórico. Las cavidades del sistema fonador (tracto vocal y nasal) comprenden el filtro acústico principal. El filtro es excitado por los órganos que se encuentran por debajo de este, y caracterizan su salida las propiedades de sistema, la forma de excitación y su evolución temporal. Un esquema de este modelo acústico se muestra en la Figura 3.

El tracto vocal adopta diferentes configuraciones para generar los distintos sonidos del habla. Estas distintas configuraciones cambian las propiedades de los filtros acústicos. En la Figura 4 se puede observar un esquema de la posición de los articuladores para las vocales [i], [a], y [u].

En el adulto masculino (femenino) medio, la longitud total del tracto vocal está cercano a 17 (14) cm, y para un niño, es de aproximadamente 10 cm. El movimiento de los articuladores del tracto vocal hace que la sección transversal de éste varíe desde 0 (completamente cerrado) a 20 cm^2. El tracto nasal constituye un camino auxiliar para la salida del sonido. Para un adulto su longitud media es de alrededor de 12 cm. La unión acústica entre los tractos vocal y nasal está controlada por el tamaño de la abertura del velo. En general, la unión nasal puede influenciar sustancialmente la frecuencia característica del sonido irradiado por la boca. Si el velo está en una posición baja, el tracto nasal está acústicamente unido para producir los sonidos nasales de la voz. La apertura del velo puede

variar desde 0 a 5 cm^2 para un adulto masculino. Para producir sonidos no nasales el velo bloquea la unión del tracto vocal con el tracto nasal.

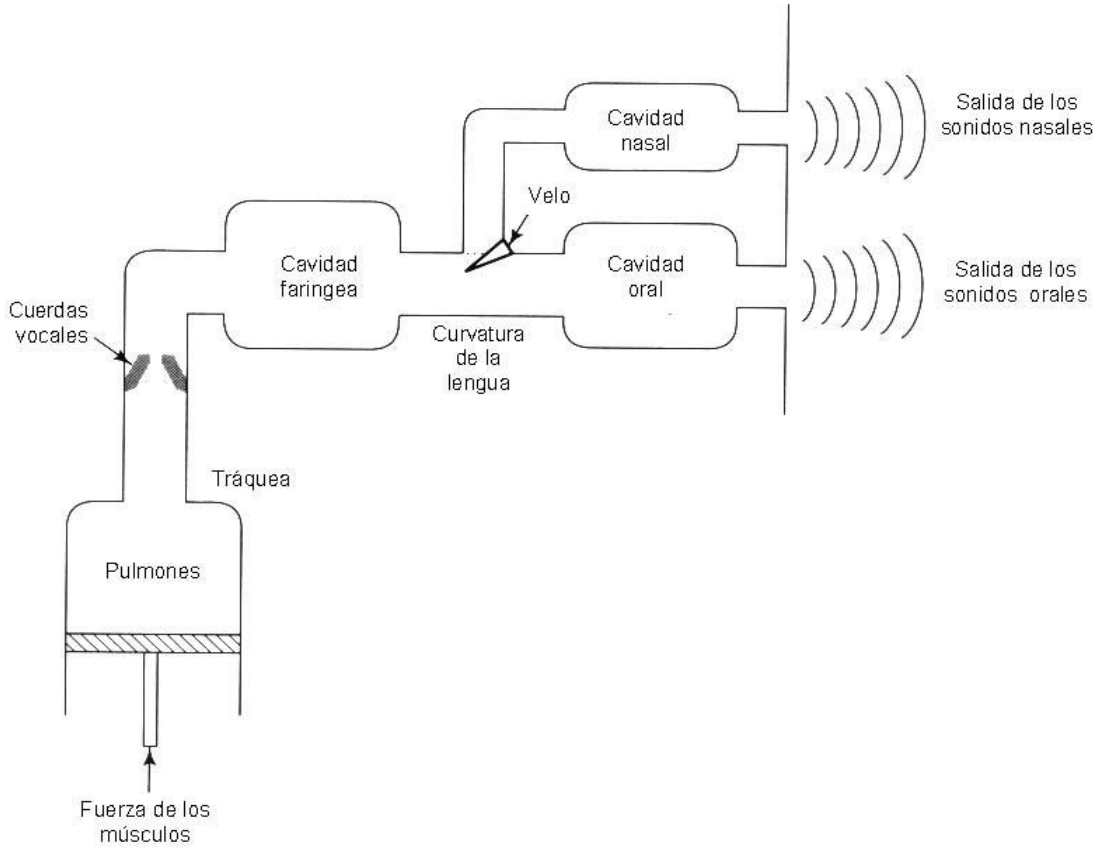

Figura 3: Diagrama en bloques de la producción del habla. Adaptado de Deller et al. 1993.

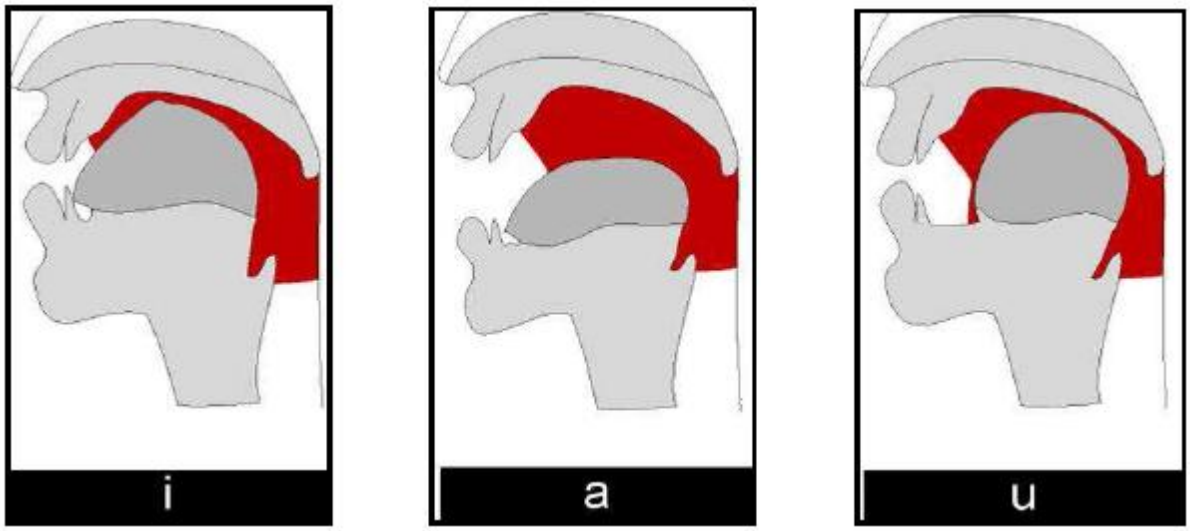

Figura 4: Esquema de la posición de los articuladores para las vocales [i], [a], y [u].

Desde un punto de vista técnico la laringe tiene un rol simple, pero muy significativo, en la producción de la voz. Esta función es producir una excitación periódica del sistema para los sonidos que nosotros conocemos como sonoros. Desde un punto de vista anatómico (y fisiológico), sin embargo, la laringe es un intrincado y complejo órgano. En la Figura 5 se observa un esquema de la laringe. La armazón de la laringe está constituida por cartílagos. El cartílago tiroides está formado por dos platos, que a su

vez, forman la pared anterior y lateral de la laringe. Su proyección anterior es comúnmente llamada nuez de Adán. El otro par de cartílagos importantes, los cuales no son visibles en la Figura 5, son los aritenoides, que se unen a las cuerdas vocales en la parte posterior de la laringe. Las cuerdas vocales se extienden enfrente del cartílago tiroides en la parte anterior, y de la aritenoides en la parte posterior.

El proceso de vibración de las cuerdas vocales se genera por las diferencias de presión a ambos lados de la glotis. Cuando el aire es impulsado desde los pulmones, y las cuerdas vocales están juntas, cerrando la glotis, la presión subglotal se incrementa hasta lograr separar las cuerdas vocales, abriendo la glotis. Este permite el paso del aire, disminuyendo la presión subglotal. Cuando la presión llega a un umbral las cuerdas vocales vuelven a juntarse, cerrando la glotis. Esto genera un nuevo aumento de la presión subglotal, reiniciando el ciclo, generando pulsos cuasi-periódicos por encima de la glotis. La frecuencia de este pulso se la denomina frecuencia fundamental, o F0, y toma valores de desde los 60 Hz hasta 500 Hz. La fluctuación del F0 puede ser voluntaria, modificando la tensión a la cual están sometidas las cuerdas vocales, o pueden depender de la anatomía y/o fisiología del locutor.

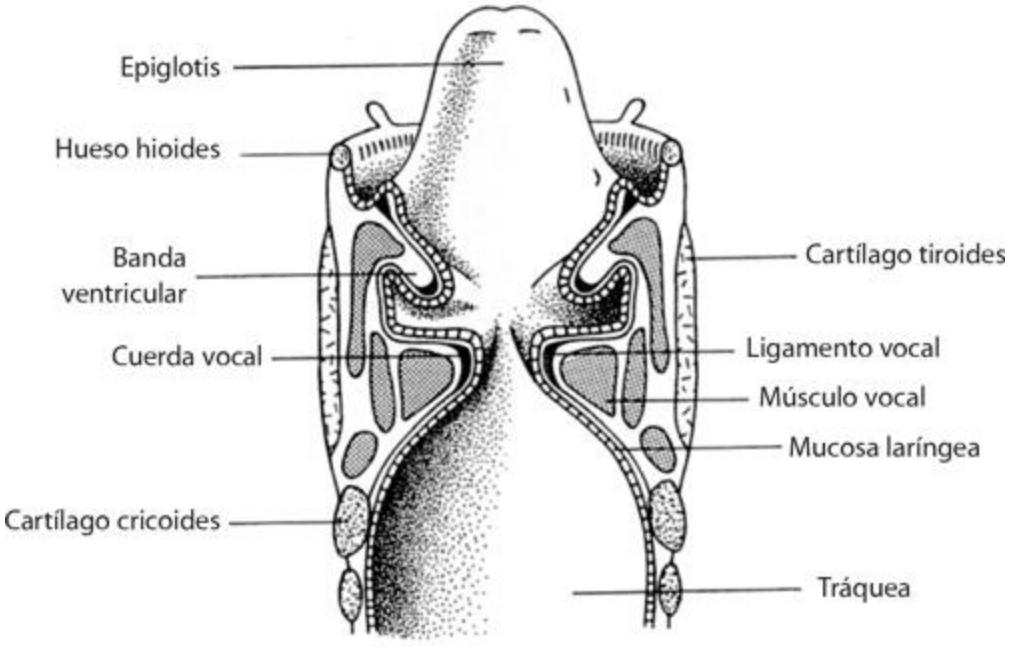

Figura 5: Sección transversal de la laringe, vista de frente. Adaptado de Borden and Harris, 1980.

Fonética acústica

Los fonemas son las unidades teóricas básicas postuladas para estudiar el nivel fónico-fonológico de una lengua humana. Los fonemas no son sonidos con entidad física, sino abstracciones mentales o abstracciones formales de los sonidos del habla. En este sentido, un fonema puede ser representado por una familia o clase de equivalencia de sonidos, que los hablantes asocian a un sonido específico durante la producción o la percepción del habla.

Las características espectrales de la señal de voz varían en el tiempo (son no-estacionarias) dado que el sistema físico cambia rápidamente en el tiempo (Flanagan 1972, Rabiner and Juang1995, Rabiner and Schafer 1978). Como resultado, la voz puede ser dividida en segmentos de sonidos de corta duración que poseen propiedades acústicas similares. Inicialmente, los sonidos de la voz son típicamente divididos en dos grandes categorías: (1) las vocales, las cuales no contienen grandes restricciones en el flujo de aire a través del tracto vocal, y (2) las consonantes que tienen una importante restricción y poseen una amplitud menor y frecuentemente son más ruidosas que las vocales.

Las vocales tienen una amplitud bastante más alta que las consonantes y son también más estables y más fáciles para analizar y describir acústicamente. Las consonantes muchas veces implican cambios rápidos de la posición de los elementos del aparato fonador, falta de periodicidad, o asordinamiento de la periodicidad (como en el caso de la [m], [n], [l]), entre otras. Para lograr algunas consonantes las cuerdas vocales pasan de una posición donde cortan completamente el flujo de aire, a una posición totalmente abierta que produce una tos ligera o una oclusión glotal (por ejemplo [p] [t] y [k]). Para lograr consonantes como la [s], o la [f], las cuerdas se abren completamente, y por lo tanto el sonido que las identifica es de naturaleza ruidosa. También existen posiciones intermedias que dan lugar a sonidos soplados, como la [x][1].

Otra manera posible de clasificar los fonemas que componen la señal de habla es desde el punto de vista de la excitación que los genera (Parsons 1978). Se pueden identificar dos tipos elementales de excitación, que dan lugar a los fonemas: sonoros y sordos.

Los sonidos sonoros son producidos forzando el aire a través de la glotis o a través las cuerdas vocales. La tensión de las cuerdas vocales se ajusta de manera tal que vibre en forma oscilatoria. La interrupción periódica del flujo de aire subglotal resulta en un soplido casi periódico de aire que excita el tracto vocal. El sonido producido por la laringe es llamado sonoro o con fonación. Este tipo de sonido consiste en una frecuencia fundamental (F0) y sus componentes armónicos producidos por las cuerdas vocales. El tracto vocal modifica esta señal de excitación que causa el formante y a veces el antiformante. El término formante se utiliza para indicar el centro de estas frecuencias de resonancia, en la cual la concentración de energía es mayor. Las formantes en el espectro son usualmente llamadas F1, F2, F3,...., comenzando con la de menor frecuencia. La frecuencia fundamental y las formantes son probablemente los conceptos más importantes en la producción y análisis del habla en general.

Los sonidos sordos son generados formando una constricción en algún punto a lo largo del tracto vocal, y forzando el aire a través de esta obstrucción, formando turbulencias que de hecho pueden ocurrir en varios sitios entre la glotis y la boca. Con sonidos puramente sordos, no hay ninguna frecuencia fundamental en la señal de excitación y por lo tanto ninguna estructura armónica, pudiendo considerar a la excitación como ruido blanco. Los sonidos sordos son también por lo general más silenciosos y menos estables que los sonoros.

Un sonido puede ser simultáneamente sonoro y sordo. Además, algunos sonidos de voz están compuestos por una corta región de silencio, seguido por una región sonora o sorda, o ambas. Estos son llamados sonidos oclusivos, y se logran cerrando en algún punto el tracto vocal, seguido por una descarga de aire.

En la se Tabla 1 se presenta los códigos SAMPA (del inglés Speech Assesment Methods: Phonetic Alphabet) para el español de la Argentina (Gurlekian et al 2014), junto con su descripción articulatoria y acústica.

Tabla 1: Denominaciones derivadas de la fonética articulatoria y acústica de los sonidos de habla.

SAMPA	Articulatoria	Acústica
p, t ,k	Oclusivos sordos: corto, medio y largo	Silencio + ruido
b, d, g	Oclusivos sonoros	Un primer formante de baja frecuencia
B, D, G	Aproximantes (no fricativos)	Formantes transicionales de baja amplitud
s, f, x, h, C	Fricativos sordos	Bandas de ruido
H	Africada sorda	Silencio + ruido
r	Liquida vibrante simple	Una interrupción entre dos vocales
R	Liquida vibrante múltiple	Múltiples interrupciones
l	Liquida Lateral Pre-palatal lateral	Formantes de amplitud media

[1] Según SAMPA, se representa con [x] al fonema fricativo, dorso palatal, sordo. Por ejemplo, al leer la palabra “juez”, el primer fonema se etiqueta con [x].

Z	Fricativa sonora; Palatal y pre-palatal; Africada sonora	Periodicidad + banda de ruido de alta frecuencia.
i, e, a, o, u	Vocales	Periódicos de amplitud alta con formantes de alta intensidad
w, j	Semivocales	Formantes transicionales de amplitud media
m, n, N, J	Nasales	Formantes fijos de baja frecuencia y amplitud

Representación y análisis de una señal de habla

Una de las más importante propiedades de la voz es que no está formada por una cadena de sonidos discretos, pero sí estaría formada por una serie de estados articulados mediante transiciones. El sonido precedente o el que sucede en una cadena puede afectar fuertemente finos detalles del sonido actual. Esta interrelación entre sonidos en una oración es llamada coarticulación. Los cambios en forma de onda de la señal de voz son una consecuencia directa de los movimientos del sistema articulador de la voz, el cual raramente permanece fijo durante algún período de tiempo. La incapacidad del sistema de producción de la voz de cambiar instantáneamente se debe a la necesidad de los movimientos del sistema articulador para producir un sonido.

A diferencia del sistema auditivo, el cual ha sido desarrollado solamente para el propósito de oír, los órganos utilizados en la producción de la voz son utilizados para otras funciones, como la alimentación, la respiración y el olfato. El rol múltiple de estos órganos sugiere que ellos no presentan una forma óptima para la comunicación. El sistema de producción de la voz está limitado en banda a 7-8 KHz.

En ingeniería, se puede inferir información acerca de un fenómeno físico desde gráficas en el dominio de la frecuencia derivada de la señal temporal. En el espectro también se puede notar, especialmente en un sonido vocálico, regiones enfatizadas (resonancia) y regiones des-enfatizadas (anti-resonancia). La localización de esta resonancia en el dominio de la frecuencia depende de la forma y dimensiones físicas del tracto vocal. Recíprocamente, cada forma de tracto vocal es caracterizada por un conjunto de frecuencias de resonancia. Como se mencionó anteriormente el centro de cada una de estas frecuencias se denomina formante. Para un determinado sonido existe un número infinito de formantes, pero en la práctica, usualmente, se toman entre 3 y 5 formantes para trabajar con señales de habla.

Un método muy útil para describir una señal de voz es el espectrograma, que es un gráfico de las variaciones temporales de la energía en bandas de frecuencias mediante niveles de colores. Las amplitudes más altas son representadas con niveles de colores más intensos, por lo tanto las frecuencias de formante y los armónicos son fáciles de percibir. A este tipo de análisis se los denomina tiempo frecuencia, y es una representación gráfica de la transformada de Fourier de corto plazo. Un parámetro extra a definir es la ventana de análisis, cuya forma y longitud modificaran apreciablemente el resultado. Las ventanas de Hamming y Hanning son usuales en procesamiento del habla. Ventanas de corta duración permiten mayor resolución temporal, en decremento de la resolución en frecuencia, y viceversa. Al utilizar una ventana de corta duración se obtiene un espectrograma de banda ancha, donde podemos observar la evolución de los formantes. Con las ventanas de larga duración obtenemos el espectrograma de banda angosta, que es útil para ver la estructura armónica de la señal.

A continuación, en las Figuras 6, 7 y 8, se dan algunos ejemplos de forma de ondas y espectros correspondientes a una frase, a una palabra y a un fonema.

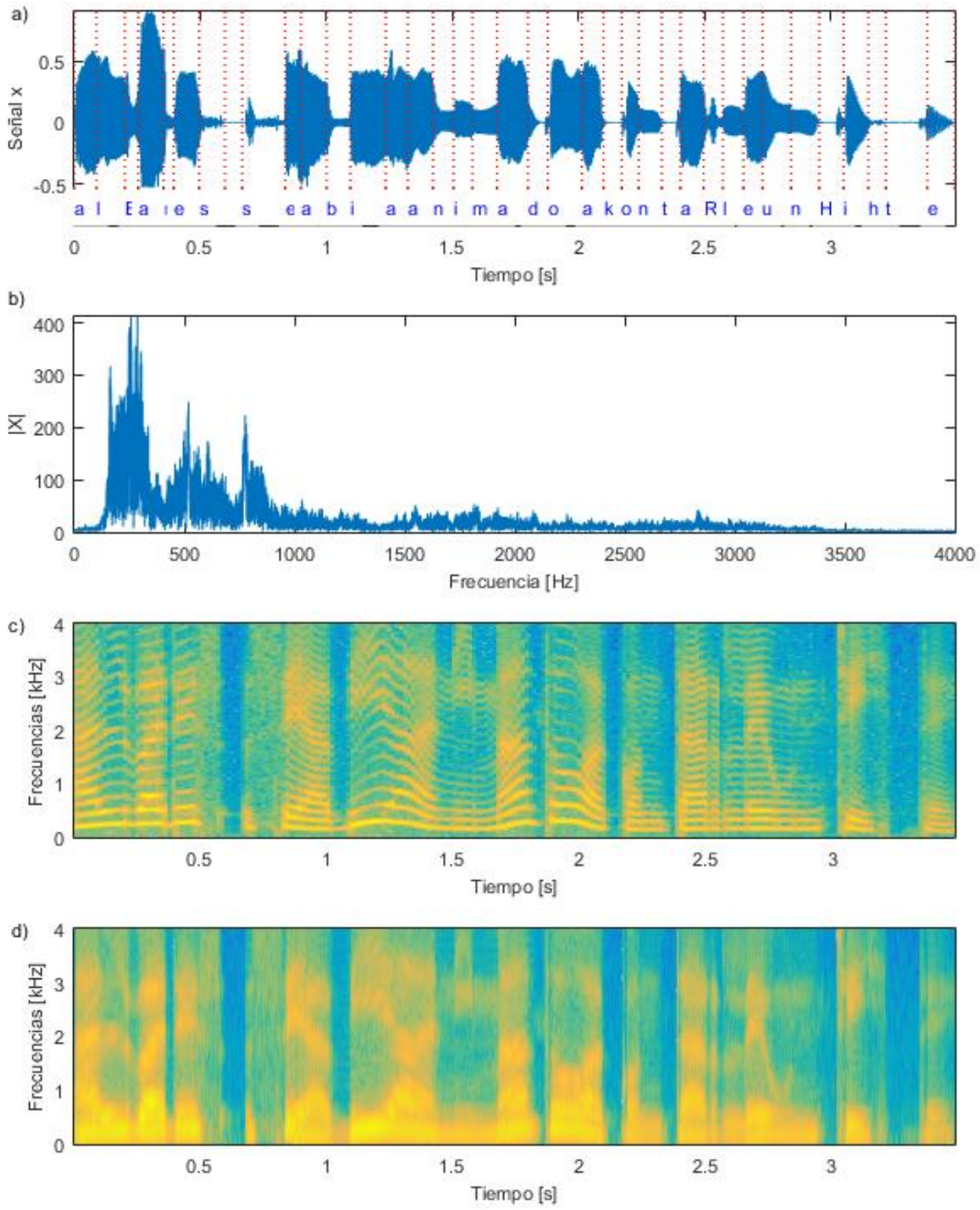

Figura 6: Señal correspondiente a la frase "Álbares se había animado a contarle un chiste", emitido por una locutora. a) Forma de onda y la segmentación fonética, con las etiqueta SAMPA; b) modulo de su transformada de Fourier; c) espectrograma de banda angosta, donde se puede observar su estructura armónica; y d) espectrograma de banda ancha, donde se pueden observar los formantes.

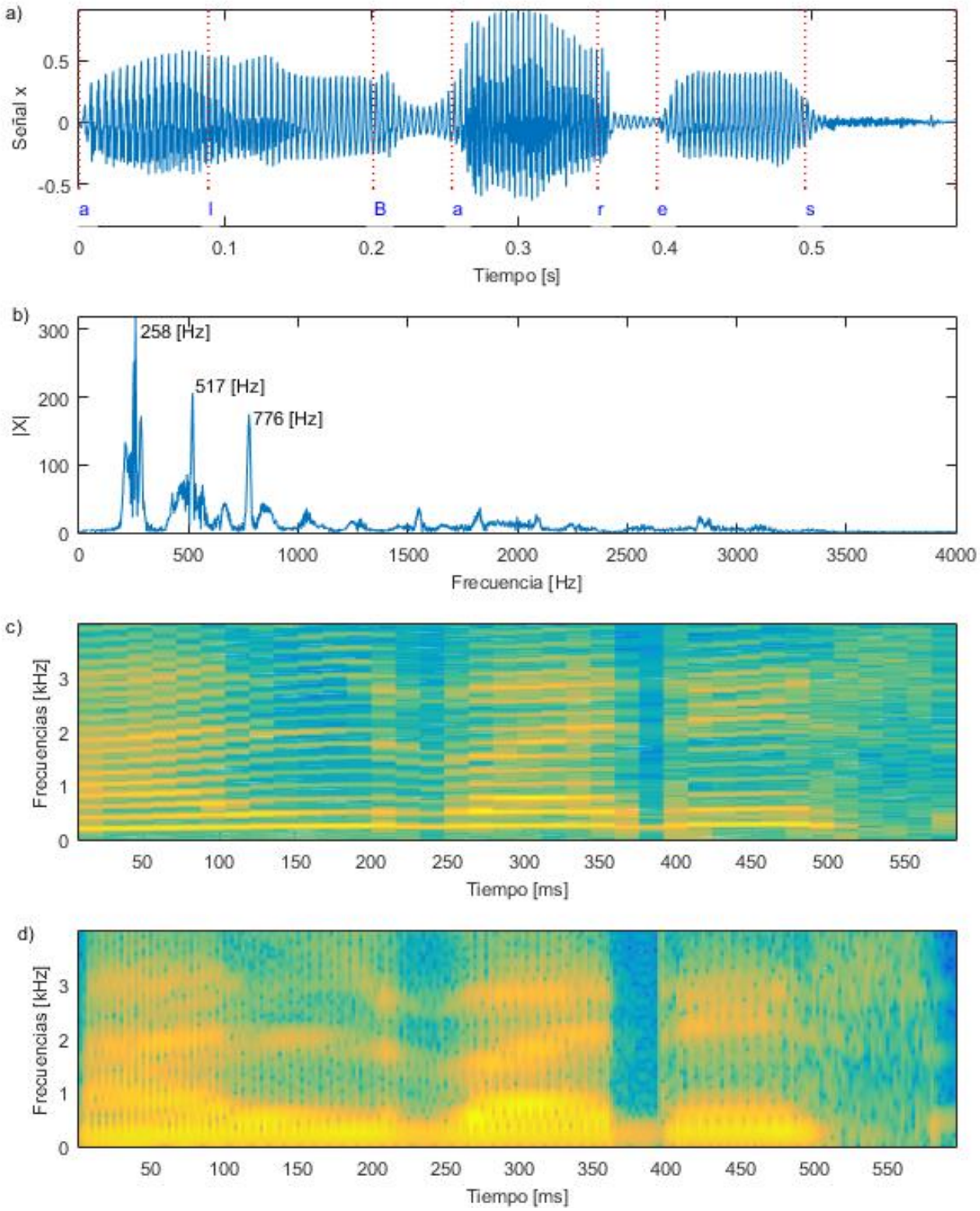

Figura 7: Señal correspondiente a la palabra "Álbares", emitido por una locutora. a) Forma de onda y la segmentación fonética; b) modulo de su transformada de Fourier donde se han marcado los tres primero picos para confirmar su carácter armónico; c) espectrograma de banda angosta, donde se puede observar su estructura armónica; y d) espectrograma de banda ancha, donde se pueden observar los formantes.

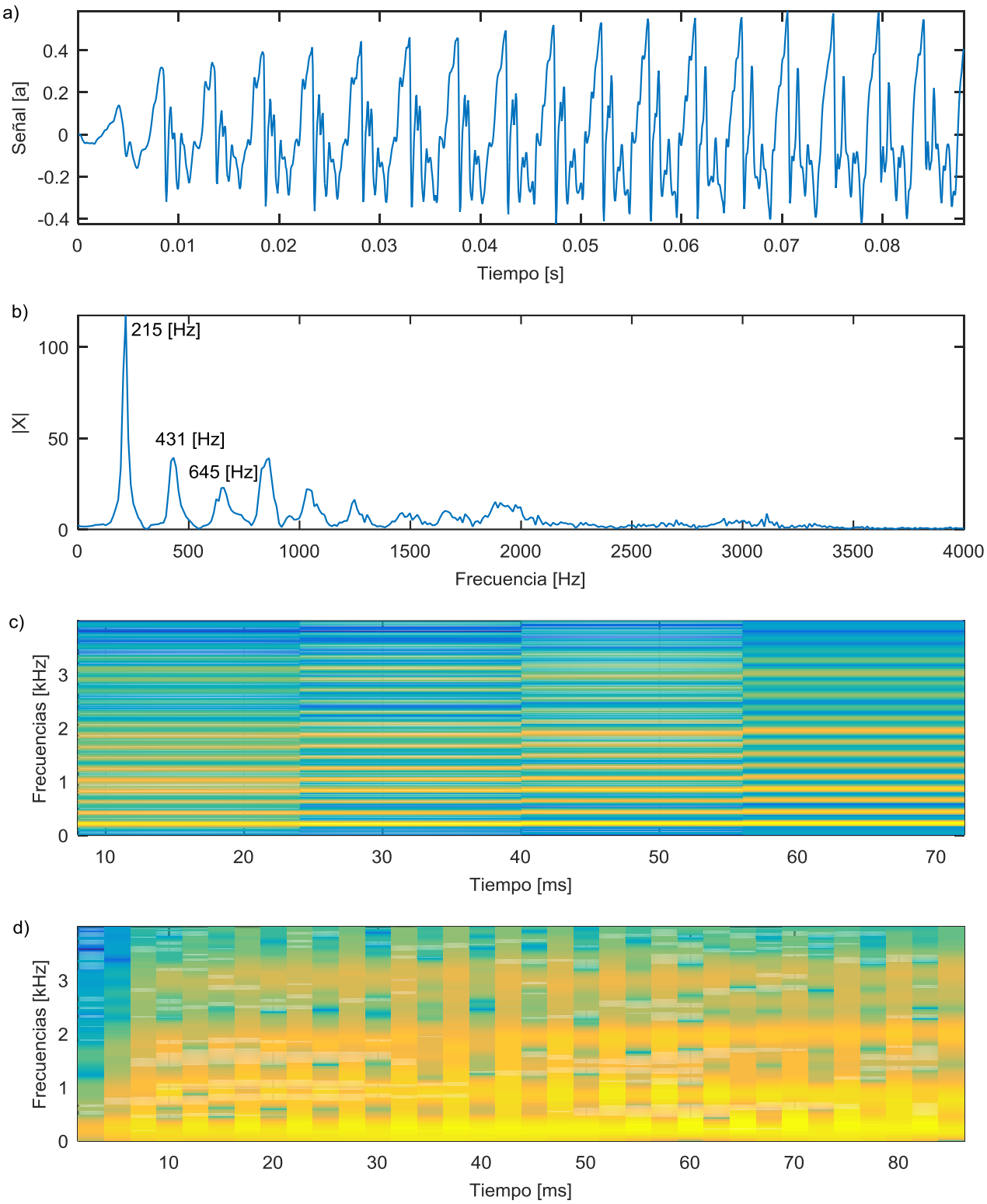

Figura 8: Señal correspondiente al fono [a], emitido por una locutora. a) Forma de onda donde se puede observar la cuasi-periodicidad de la señal; b) modulo de su transformada de Fourier donde se han marcado los tres primero picos para confirmar su carácter armónico; c) espectrograma de banda angosta, donde se puede observar su estructura armónica; y d) espectrograma de banda ancha, donde se pueden observar los formantes.

Pulso glótico

El pulso glotal es el flujo de aire que atraviesa la abertura superior de la glotis durante la producción de sonidos sonoros y se caracteriza por un comportamiento pulsátil cuasiperiódico. Es por lo tanto la excitación física al sistema de producción de la voz. Sus características dependen del locutor y del tipo de voz producida. Según lo descripto en secciones anteriores, en un pulso glótico se pueden observar tres fases: de apertura, de cierre y de cerrada. Esta última puede estar ausente, dependiendo de la frecuencia del pulso.

Existen varias propuestas para modelar el pulso glótico, entre las que podemos mencionar las realizadas por Rosenberg (1971) y la de Liljencrants y Fant (LF) (Fant et al. 1985). En el modelo de Rosenberg, un período del pulso glótico se puede expresar como:

$$P(t) = \begin{cases} \frac{P_0}{2}\left[1 - cos\left(\frac{t}{T_p}\pi\right)\right] & si\ \ 0 \le t\ \le\ T_p \\ P_0 cos\left(\frac{t-T_p}{T_n}\frac{\pi}{2}\right) & si\ \ T_p < t\ \le\ T_p + T_n \\ 0 & en\ otro\ caso \end{cases} \qquad (1)$$

Donde T_p es la duración de la fase de apertura, T_n es la duración de la fase de cierre, y P_0 es la amplitud máxima de la presión subglotal.

Transferencia del tracto vocal

Teniendo en cuenta lo dicho anteriormente, podemos inferir que las propiedades espectrales de la señal de voz (en especial las frecuencias donde hay más concentración de energía o formantes) depende de la anatomía del sistema de la producción de voz. Este sistema, para producir los distintos sonidos, varía su configuración a cada instante. Por otra parte, la forma, volumen, longitud, y otras propiedades anatómicas de los órganos involucrados en la producción de voz, son intrínsecas a cada hablante. Esto hace que se encuentren diferencias en la señal de voz, llamadas variaciones interhablantes, las cuales hacen posible diferenciar un sonido pronunciado por distintos locutores (McDonough et al. 1998). Existen también diferencias llamadas variaciones intrahablante, las cuales suceden en las señales de voz de un mismo individuo, y se deben, por ejemplo, al estado de ánimo y/o contexto de la oración.

Una forma de modelar el tracto vocal (Deller et al. 1993) es a través de dos tubos, como se esquematiza en la Figura 9. El primer tubo, de longitud l_1 y área transversal A_1, representa la cavidad faríngea, y el segundo tubo, de longitud l_2 y área transversal A_2, representa la cavidad oral. Además, se asume que la fuente de excitación es sinusoidal, representado por un pistón.

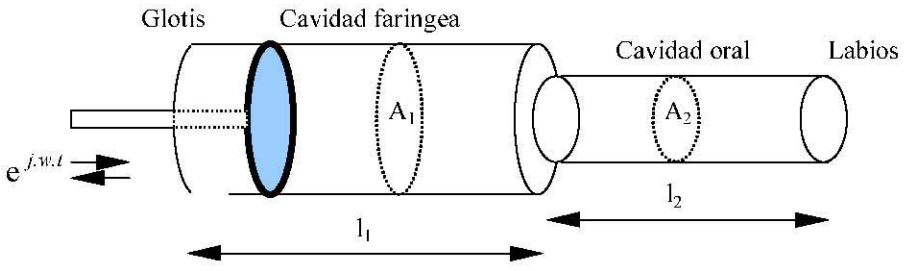

Figura 9: Modelo de dos tubos para el tracto vocal humano. Adaptado de Deller et al. 1993.

La función de transferencia de este modelo está dada por:

$$H(\omega) = \frac{A_2}{A_1\ sinh\left(\frac{j\,\omega\,l_1}{c}\right) sinh\left(\frac{j\,\omega\,l_2}{c}\right) + A_2\ cosh\left(\frac{j\,\omega\,l_1}{c}\right) cosh\left(\frac{j\,\omega\,l_2}{c}\right)} \qquad (2)$$

donde $\omega = 2\ \pi\ f$, f es la frecuencia, y c es la velocidad del sonido en el aire.

En la Figura 10, se muestran las respuestas en frecuencia para distintas longitudes de la cavidad faríngea: l_1 = 10 cm, l_1 = 11 cm y l_1 = 12 cm, respectivamente.

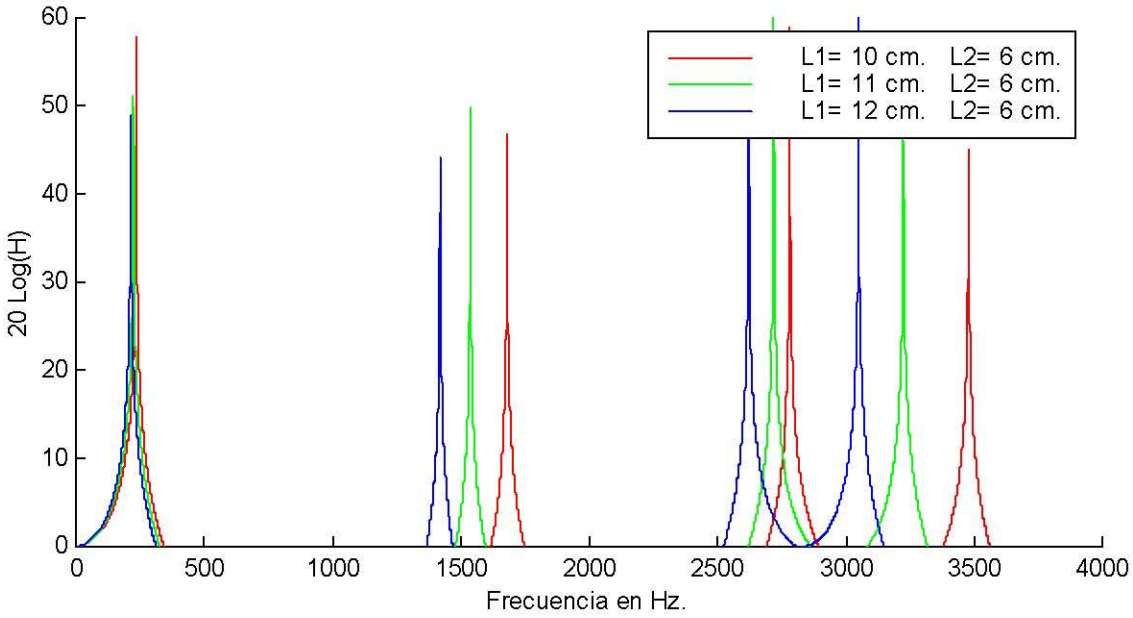

Figura 10: Respuestas en frecuencia de tres conjuntos de dos tubos acústicos uniformes de distintos diámetros acoplados.

Otra aproximación modela el tracto vocal como la conexión en cascada de filtros IIR de segundo orden, donde cada filtro modela una resonancia. Así, para el formante n-ésimo, su filtro puede expresarse como:

$$H_n(z) = \frac{1}{(1-p_n\ z^{-1})(1-p_n^*\ z^{-1})} \qquad (3)$$

Donde p_n es el polo para el formante n-ésimo,y p_n^* es su conjugado. Los polos pueden ser estimados en función de la frecuencia de resonancia y de su ancho de banda, como:

$$p_n = e^{\frac{-2\pi B}{F_s}} e^{j\frac{2\pi F_n}{F_s}} \qquad (4)$$

Frecuencia fundamental

La frecuencia fundamental, o F0, es uno de los rasgos de la señal de voz más estudiado, dada la cantidad e importancia de la información que conlleva: define el modo de la oración, marca los acentos, es una característica del locutor, contiene información sobre el estado emocional del locutor, entre otros. Para su estimación se han propuesto diferentes algoritmos, tanto directos como indirectos. Una forma indirecta es a partir de la señal de voz. Por ejemplo, Noll (1967) presentó un algoritmo basado en el cepstrum.

El cesptrum $c[n]$ de una señal $x[n]$, se define como[2]:

$$c[n] = \mathcal{F}^{-1}\{log|\mathcal{F}\{x[n]\}|\}$$

Si *x[n]* es la señal de voz, según la teoría de producción de voz fuente-filtro, *x[n]* es el resultado del filtrado del pulso glótico *s[n]* por el tracto vocal *h[n]*, que se puede escribir como:

$$x[n] = h[n] * s[n]$$

$$\mathcal{F}\{x[n]\} = \mathcal{F}\{h[n]\}\mathcal{F}\{s[n]\}$$

$$log|\mathcal{F}\{x[n]\}| = log|\mathcal{F}\{h[n]\}\mathcal{F}\{s[n]\}|$$

[2] Existen definiciones alternativas, pero en el presente trabajo nos resulta útil esta aproximación.

$$log|\mathcal{F}\{x[n]\}| = log|\mathcal{F}\{h[n]\}| + log|\mathcal{F}\{s[n]\}|$$

$$\mathcal{F}^{-1}\{log|\mathcal{F}\{x[n]\}|\} = \mathcal{F}^{-1}\{log|\mathcal{F}\{h[n]\}|\} + \mathcal{F}^{-1}\{log|\mathcal{F}\{s[n]\}|\}$$

Vemos que en el dominio del cepstrum, la convolución del pulso glótico con el filtro del tracto vocal se mapea como una suma. Además, el tracto vocal genera una onda de baja frecuencias en el espectro logarítmico, mientras que el pulso glótico se manifiesta como altas frecuencias. Es decir que en el dominio del cepstrum el efecto del tracto vocal y el pulso glótico están separados. Para evitar confusión con la variable independiente, se utiliza el término quefrencia, y está expresada en segundos. Por lo tanto, el período de la frecuencia fundamental se representa en el cepstrum como un pico en altas frecuencias, expresado en segundos. Un algoritmo para detectar el período de la frecuencia fundamental debería buscar el pico del cepstrum en el rango de 1/500 s a 1/50 s, que es el intervalo de períodos usual de la voz humana.

En la Figura 11.a se puede observar nuevamente la vocal [a] de la Figura 8.a, con la parte real del cepstrum y una ampliación en el rango correspondiente de 50 a 500 Hz. Claramente se puede observar el pico del cepstrum en 0.004783 s, que daría un F0 de 209 Hz.

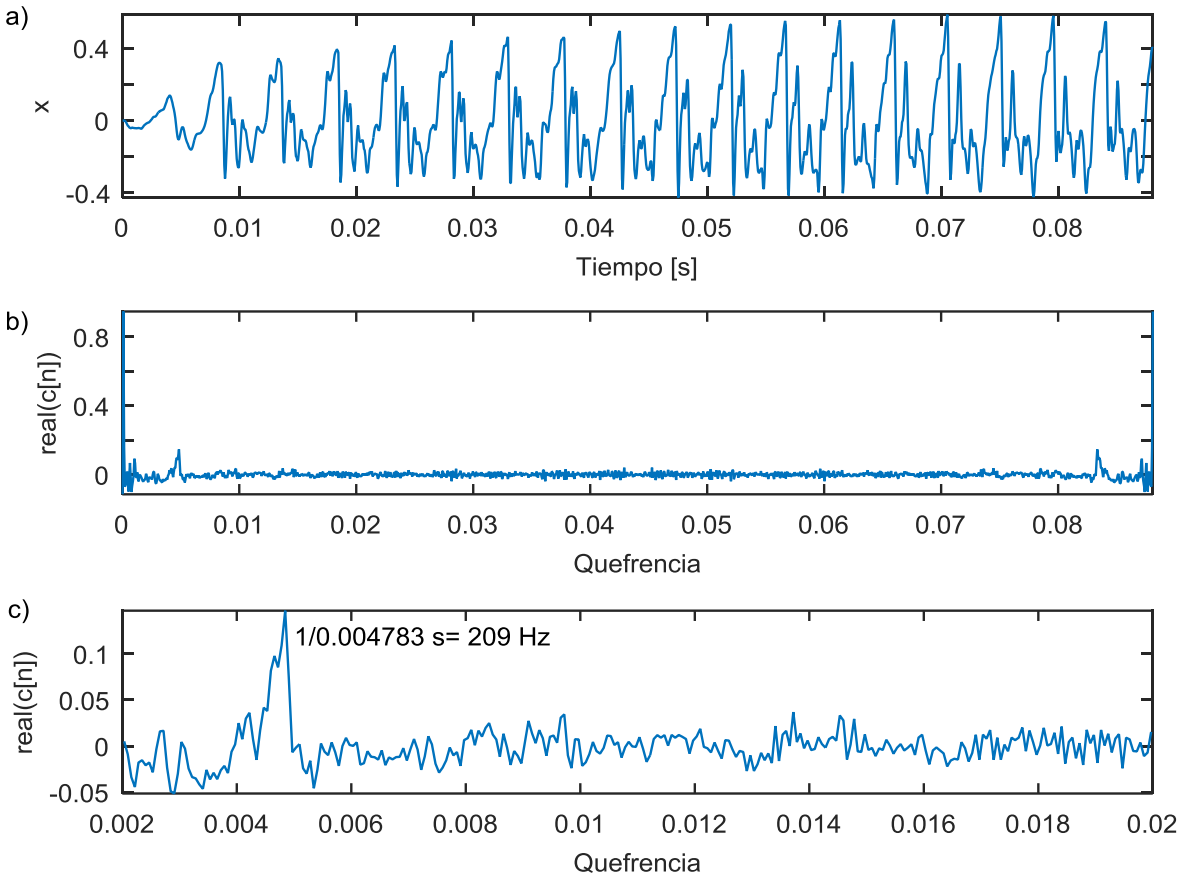

Figura 11: Señal correspondiente al fono [a], emitido por una locutora. a) Forma de onda donde se puede observar la cuasi-periodicidad de la señal; b) pate real del cepstrum; y c) el segmento del cepstrum en el rango de 0.002 s a 0.02 s donde pude observar el pico correspondiente a un F0 de 209 Hz.

Como el F0 varia en el tiempo, es conveniente realizar un análisis por ventanas, similar al realizado en la transformada de Fourier de corto plazo o su representación en el espectrograma. En la Figura 12.a se puede observar nuevamente la señal de la Figura 7 y una representación del análisis tiempo-quefrencia. Claramente se puede observar la variación temporal del contorno de F0.

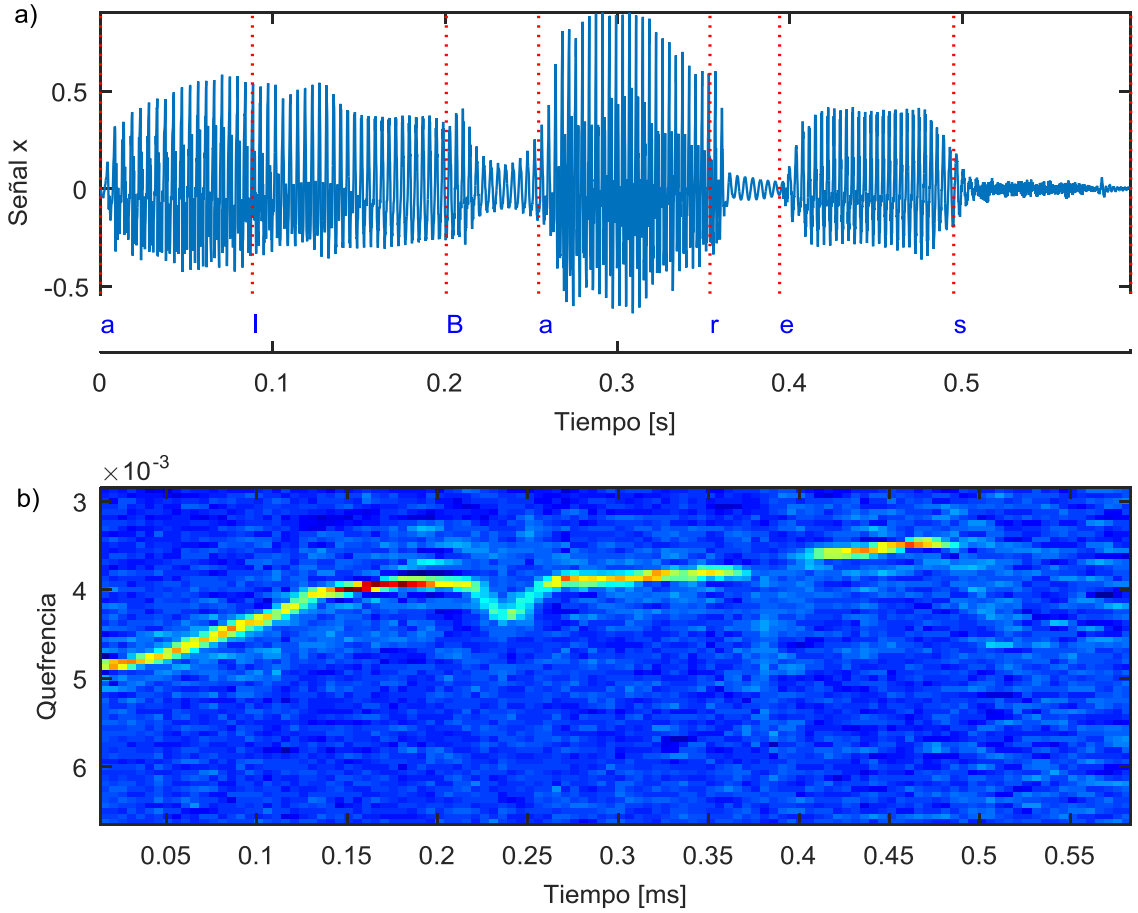

Figura 12: Señal correspondiente a la palabra "Álbares", emitido por una locutora. a) Forma de onda y la segmentación fonética; b) y una representación tiempo-quefrencia donde se puede observar la evolución temporal del contorno de la frecuencia fundamental.

El control de la prosodia es una de las tareas de interés dentro del área de procesamiento de señales de voz. Para esto se han diseñado una variedad de métodos, entre los cuales se destaca el conocido como Suma Solapada Sincrónica con la Frecuencia Fundamental (PSOLA, del inglés Pitch Synchronous Overlap Add), y de la cual existen varias versiones, por ejemplo la denominada PSOLA en el dominio temporal (TD-PSOLA, del inglés Time Domain PSOLA) (Moulines. and Charpentier, 1990). El método consiste en tomar porciones de la señal de análisis, multiplicarla por una ventana temporal sincrónica con la frecuencia fundamental para luego recombinarlas sincrónicamente con una nueva frecuencia fundamental (Figura 13). Además, los segmentos pueden ser repetidos o eliminados, según se desee aumentar o disminuir la duración de la emisión de voz.

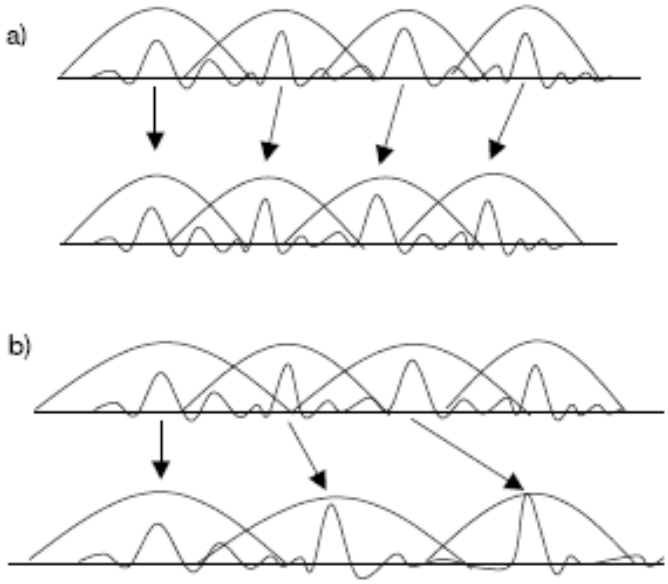

Figura 13: Modificación de la frecuencia fundamental de segmentos sonoros de la señal de voz, utilizando el método de PSOLA. a) Aumento y b) disminución de la frecuencia fundamental.

El algoritmo de modificación de la frecuencia fundamental mediante el algoritmo TD-PSOLA puede resumirse como:

a) Detectar los segmentos de la señal que se corresponden con sonidos sonoros.
b) Localizar los picos de cada ciclo que conforman los segmentos sonoros.
c) Aplicar una ventana centrada en cada pico, desplazarla temporalmente y sumarla para obtener la señal resultante.

Ejercicios

1. Grafique la señal de voz del archivo hh2.wav, ubicando en ella porciones de señales que se o correspondan con fonemas sonoros y sordos. Segmentar y etiquetar en forma aproximada cada uno del los fonemas presentes en la señal.
2. Con la segmentación realizada en el ejercicio 1 de la señal hh2.wav, encuentre los coeficientes de Fourier de un período del segmento de señal correspondiente a un fono [a]. Repetir el cálculo para varios períodos de la vocal.
3. Reconstruya la señal temporal a partir de los coeficientes calculados. Escuche y compare las distintas reconstrucciones correspondientes a coeficientes de Fourier tomados de distintos períodos. Compárelas también con la señal original. ¿Qué observación se puede hacer sobre la periodicidad de los fonemas vocálicos?
4. Grabe la misma frase del ejercicio 1. Mencionar las diferencias entre ambas señales.
5. Grafique los espectrogramas de banda angosta de los segmentos de señal correspondientes a tres vocales presentes en la señal hh2.wav. Compare y analice las diferencias.
6. Genere diez ciclos del tren de pulsos glóticos según los modelos de Rosenberg. Tomar una frecuencia F0 = 200 Hz, y fases de apertura y cierre de 40% y 16%, respectivamente, de la duración de un pulso. Considerar una amplitud máxima de 1. A los efectos de la simulación, considerar una frecuencia de muestreo de 16 kHz. Estimar su espectro de amplitud y explicar su contenido.
7. Utilizando las ec. 3 y 4, generar un modelo de tracto vocal para cada uno de los siguientes conjuntos de valores de parámetros, que se corresponde con una vocal emitida por una locutora.

	F_1	B_1	F_2	B_2	F_3	B_3	F_4	B_4
a	830	110	1400	160	2890	210	3930	230
e	500	80	2000	156	3130	190	4150	220
i	330	70	2765	130	3740	178	4366	200
o	546	97	934	130	2966	185	3930	240
u	382	74	740	150	2760	210	3380	180

   Grafique diagrama de polos y ceros, y la respuesta en frecuencia del sistema para cada vocal, y compare.
8. Utilizando los resultados de los dos último ejercicios, sintetice un segundo de las cinco vocales. Escuche y grafique. Haga un análisis en frecuencia, y en tiempo-frecuencia.
9. Aplicando el método descripto en la introducción, estime el contorno de la frecuencia fundamental de la voz en el archivo hh2.wav. Grafique en forma sincrónica con la onda.
10. Con el resultado del ejercicio 9, aplique el método PSOLA para aumentar y disminuir un 10%, 20% y 30% los valores de frecuencia en el contorno de F0 de la voz en el archivo hh2.wav.
11. Repita el ejercicio 10 pero aplicando el método para ajustar la velocidad del habla.
12. Repita el ejercicio 8, pero variando la frecuencia fundamental desde 200 Hz a 300 Hz en forma lineal. Escuche la onda resultante, ¿cómo se percibe el cambio en la frecuencia fundamental? Estime el F0 resultante y compárelo con el teórico.
13. Repita el ejercicio 12, pero variando la frecuencia fundamental desde 200 Hz a 100 Hz.
14. A la señal de vocales generada en el ejercicio 8 realícele un filtrado con el objetivo de eliminar su frecuencia fundamental. Puede utilizar la herramienta fdatool para diseñar el filtro. Justifique el filtro implementado. Grafique ambas señales, haga un análisis en frecuencia y compare. ¿Perceptualmente se percibe alguna diferencia? ¿Porqué?

Bibliografía

Aronson, L., Rufiner, H. L., Furmanski, H. y Estienne, P. Características Acústicas de las Vocales del Español Rioplatense" Fonoaudiológica, Vol. 46, No. 2, pp. 12-20. Julio 2000.

Borden, G. J. and Harris, K. S. Speech science primer. Physiology, acoustics and perception of speech. Baltimore: Williams & Wilkins. 1980.

Deller, J. R., Proakis, J. G. and Hansen, J. H.L. Discrete Time Processing of Speech Signals. Prentice-Hall, New York, 1993.

Fant, G., Liljencrants, J. and Lin, Q. G. A four-parameter model of glottal flow. STL-QPSR, 4(1985):1–13, 1985.

Flanagan, J. L. Speech Analysis Synthesis and Perception. Springer-Verlag, 2 edición, 1972.

Gurlekian, J., Torres, H. y Evin, D. Guía para la segmentación y transcripción fonética para las tecnologías del habla. Fonoaudiológica, vol. 61 (2), pp. 24-47. Diciembre de 2014.

Klatt, D. Software for a Cascade/Parallel Formant Synthesizer. JASA, vol. 67, pp.971-995. 1980.

Martínez Celdrán, Eugenio. Fonética. Barcelona, Teide, 1984.

McDonough, J., Byrne, B. y Xiaoqiang Luo. Speaker Normalization with All Pass Transforms. In Proc. of the 5th International Conference on Spoken Language Processing (ICSLP) Vol. 6, pp. 2307-2310, Sydney, Australia, November 1998. http://citeseer.ist.psu.edu/article/mcdonough98speaker.html.

Moulines, E. and Charpentier, F. Pitch-SynchronousWaveform Processing Techniques for Text-to-Speech Synthesis Using Diphones. Speech Communication, Vol. 9, pp. 453-467. 1990.

Noll, A. M. Cepstrum Pitch Determination. J. Acoust. Soc. Am. 41, 293. 1967. http://dx.doi.org/10.1121/1.1910339

Parsons, T. Voice and Speech Processing. McGraw-Hill, 1978.

Rabiner, L. and Juang, B.-H.. Fundamentals of Speech Recognition. Prentice Hal, 1995.

Rabiner, L. and Schafer, R. W. Digital Processing of Speech Signals. Prentice-Hall, 1978.

Rosenberg, A. E.. Effect of glottal pulse shape on the quality of natural vowels. The Journal of the Acoustical Society of America, 49(2B):583–590, 1971.