FUNDAMENTOS DEL RECONOCIMIENTO AUTOMÁTICO DE LA VOZ

"Fundamentos de la producción y percepción de la señal de voz"

Agustín Álvarez Marquina

Introducción (I)

- **Forma de onda y espectro de una señal de voz. Ejemplo: *"Esto es una señal de voz"***

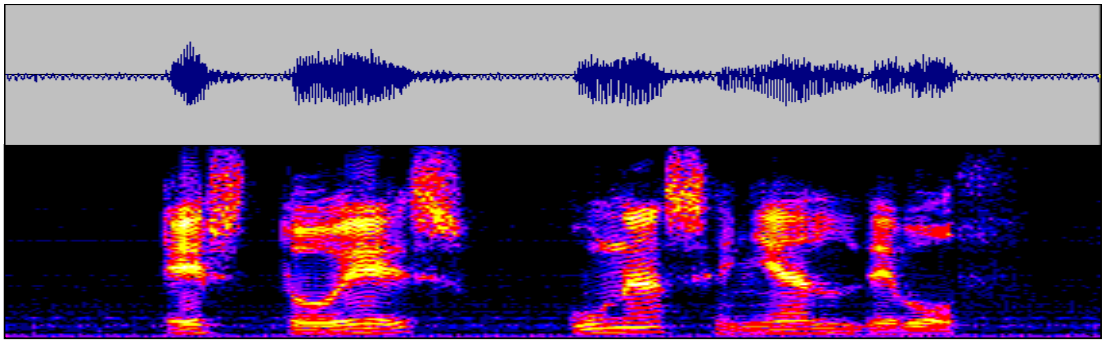

Figura 1. Ejemplo de forma de onda y espectrograma asociado.

Introducción (II)

Esquema de la comunicación.

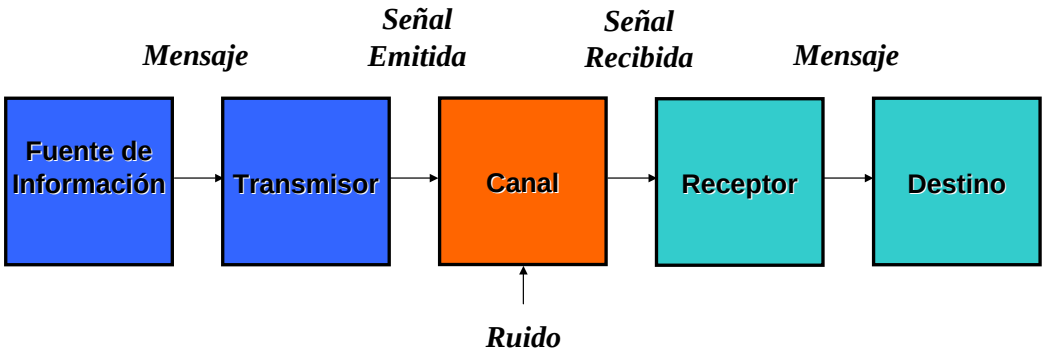

Figura 2. Esquema del proceso de comunicación.

Introducción (III)

Fonética y Fonología.

- Estudian la forma de los sonidos de la lengua.
- Desde el punto de vista de su función en el sistema de la comunicación lingüística ⇒ ***Fonología***.
- Desde el punto de vista de su producción, de su constitución acústica y de su percepción ⇒ ***Fonética***.
- Tradicionalmente se ha estudiado la fonética desde el punto de vista articulatorio sin tener en cuenta el acústico.

Introducción (IV)

- **Fonética acústica y técnicas de tratamiento de la voz.**
  - La fonética acústica estudia los mecanismos por los cuales el oyente es capaz de identificar las unidades de comunicación que constituyen el mensaje hablado.
  - Sin embargo la fonética articulatoria es el origen de la acústica.
  - Aunque ambos enfoques son útiles para estudiar el proceso del habla, nos centraremos más en el primero (acústico).

Introducción (V)

- **Fonemas, rasgos distintivos y archifonemas.**
  - **Fonema**: es la unidad lingüística más pequeña, desprovista de significado formado por un haz simultáneo de rasgos distintivos.
    - Ejemplos: /p/, /t/, /k/
  - **Rasgos distintivos**: son las unidades inferiores al fonema, que pueden definirse en el nivel articulatorio o acústico.
    - Ejemplo: /p/ consonante, oclusivo, bilabial, sordo.
    - Ejemplo: /t/ consonante, oclusivo, dental, sordo.
    - Ejemplo: /k/ consonante, oclusivo, velar, sordo.

Introducción (VI)

- **Archifonema**: es el resultado de una neutralización.
  - La neutralización se produce cuando una oposición fonológica deja de ser pertinente en ciertas posiciones de la cadena hablada.
  - Ejemplo: /ɾ/ frente a /r/ en posición postnuclear o implosiva.

- **Un fonema está constituido por la unión de un conjunto de rasgos distintivos y por otros que no lo son.**

Detección de rasgos fonéticos y reconocimiento de voz (I)

- **La percepción del habla es un tema de la máxima importancia dentro de la comunicación oral, tanto entre humanos como en el caso de hombre y máquina.**
  - La decodificación en el nivel fonológico es fuertemente dependiente de la lengua.
  - Cuando se produce un desajuste en el proceso de la percepción, la asignación de significado puede fallar parcial o totalmente.
    - Problemas auditivos.
    - Lengua materna diferente de la lengua percibida.

Detección de rasgos fonéticos y reconocimiento de voz (II)

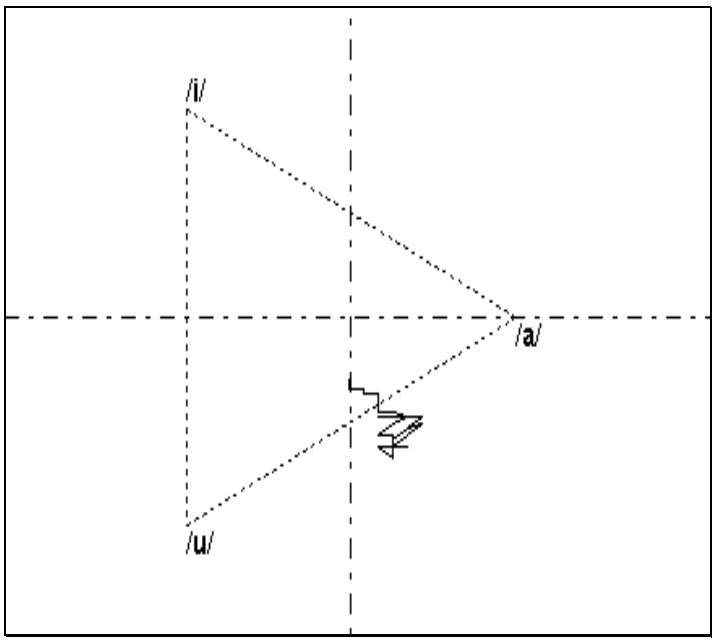

Figura 3. Representación en el triángulo vocálico de la vocal del inglés */aa/*. Ej. *"Cot"*.

Anatomía y fisiología del sistema de producción de habla (I)

- **Forma de onda de la señal de voz = Onda de presión acústica.**
- **Estructuras que la originan.**
  - Pulmones.
  - Tráquea.
  - Laringe (órgano de producción de voz).
  - Tracto vocal (cavidad faríngea, cavidad oral y cavidad nasal).

Anatomía y fisiología del sistema de producción de habla (II)

- **Articuladores.**
  - Cuerdas vocales.
  - Velo.
  - Lengua.
  - Dientes.
  - Labios.
  - Mandíbula.

Anatomía y fisiología del sistema de producción de habla (III)

- **Laringe.**
  - Función: producir una excitación periódica al sistema para los sonidos sonoros.
  - Se compone de:
    - Cuatro cartílagos
    - Un par de bandas elásticas de músculo y mucosas que van de los cartílagos tiroideos a los artenoides (nuez).

El aparato fonador (I)

- **El aparato fonador humano es un sistema fuertemente ligado con la función de respiración, que utiliza recursos comunes con la misma.**
- **Se compone de un conjunto de cavidades:**
  - **Cavidades infraglóticas**: pulmones, bronquios y traquea.
  - **Cavidad glótica**: formada por una serie de cartílagos y músculos
  - **Cavidades supraglóticas**: cavidad faríngea, cavidad nasal y cavidad oral.

El aparato fonador (II)

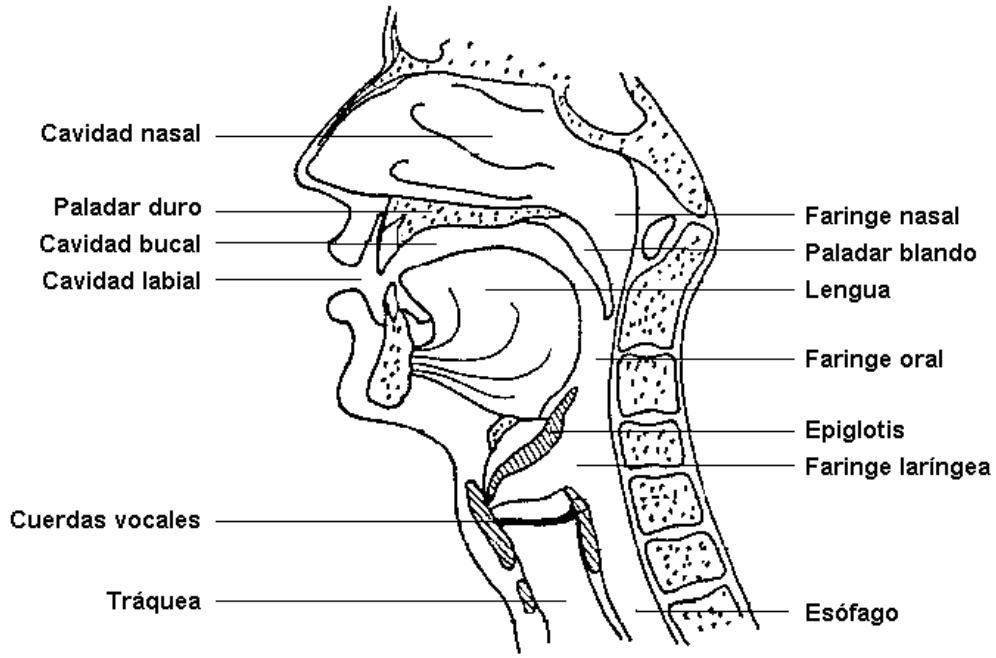

Figura 4. Corte vertical de los órganos fonadores.

Equivalencia producción de habla / filtro (I)

- **La producción de habla se puede asimilar a una operación de filtrado.**
  - Cavidades= Principal filtro acústico.
  - Excitación (cuerdas vocales para sonidos sonoros).
  - Carga a la salida: impedancia de radiación debida a los labios.
  - Articuladores: Cambian las propiedades del sistema.

Equivalencia producción de habla / filtro (II)

- **Tracto vocal.**
  - Hombre adulto: 17 cm.
  - Mujer adulta: 14 cm.
  - Niño: 10 cm.
    - Área del tracto vocal: desde 0 a 20 cm^2.
- **Cavidad nasal: Camino auxiliar de transmisión del sonido (12 cm).**
    - Acoplamiento acústico entre cavidad nasal y resto de cavidades: controlado por el velo (apertura de 0 a 5 cm^2).

Equivalencia producción de habla / filtro (III)

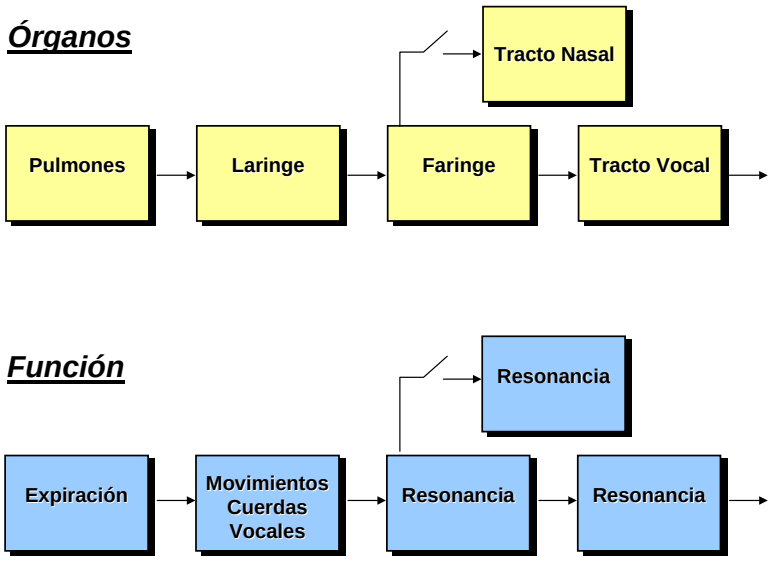

Figura 5.a. Diagrama funcional del aparato fonador [JUN96].

Equivalencia producción de habla / filtro (IV)

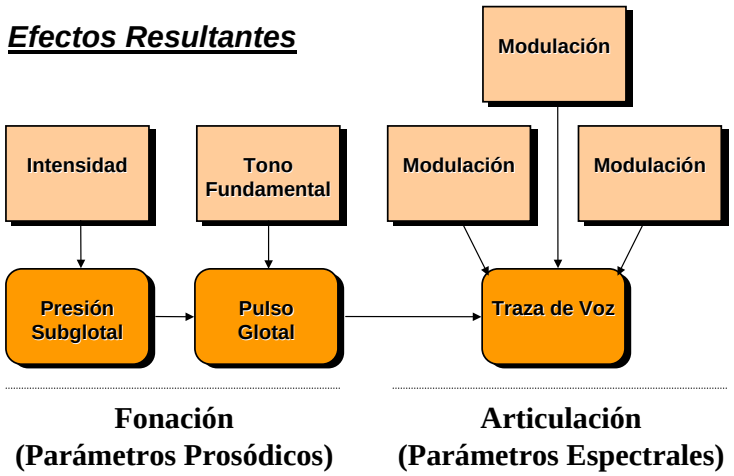

Figura 5.b. Diagrama funcional del aparato fonador [JUN96].

Clasificación de los sonidos de la lengua (I)

Sonidos sordos.

- Se realizan sin el concurso de las cuerdas vocales (posición de respiración).
- El flujo de aire procedente de los pulmones se vuelve más rápido, produciendo fricciones y turbulencias.
- Esto se traduce en vibraciones desordenadas en aquellos puntos en que por producirse un estrechamiento, la velocidad del fluido alcanza máximos.

Clasificación de los sonidos de la lengua (II)

- Si el punto de estrechamiento o articulación se encuentra próximo al exterior (labios, dientes), el resto de la estructura influye escasamente en el sonido emitido.
  - En caso contrario (zona velar o medial), la posición de los restantes órganos puede modular bastante el resultado.
- Su forma de onda presenta un escaso carácter repetitivo, recordando al ruido aleatorio.
- Los sonidos sordos son típicamente consonánticos aunque también hay consonantes sonoras.

Clasificación de los sonidos de la lengua (III)

Sonidos sonoros.

- Se realizan con el concurso de las cuerdas vocales, que obturan el paso del flujo de aire durante breves instantes para ceder expeliendo breves pulsos glotales.
- Este tren de pulsos es modificado por los órganos de articulación que vienen posteriormente añadiendo una cierta codificación.
- Esta codificación consiste en el resalte de determinadas frecuencias contenidas en el pulso glotal original o en su eliminación.

Clasificación de los sonidos de la lengua (IV)

- El sonido así emitido tiene una forma relativamente periódica.
- Un ejemplo de sonido sonoro lo constituyen las vocales.

Clasificación de los sonidos de la lengua (V)

Punto de articulación.

- Son los puntos de estrechamiento en las cavidades supraglóticas, antes mencionados, para los sonidos sordos y sonoros.
- Son controlados por los órganos móviles contra los órganos inmóviles de su misma zona.
- Una lista de posibles puntos de articulación puede ser:

Clasificación de los sonidos de la lengua (VI)

Órgano móvil	Órgano fijo	Nombre de la articulación
Velo	Raíz	Nasal
Uvula	Raíz	Uvular
Medio	Paladar	Palatal
Ápice	Alveolo	Apicoalveolar
Ápice	Corona dental	Apicodental
Ápice	Labios	Apicolabial
Dientes	Dientes	Interdental
Labios	Corona dental	Labiodental
Labios	Labios	Bilabial

Tabla 1. Lista de los puntos de articulación.

Clasificación de los sonidos de la lengua (VII)

Modo de articulación.

- Define la forma concreta en que la articulación puede tener lugar.
- Dos posibilidades: modo espirado y modo no espirado.
- El modo espirado se produce bajo un flujo de aire de origen pulmonar y es el modo habitual en la mayoría de las lenguas.
  - El modo no espirado se produce por compresión o enrarecimiento del aire en la cavidad bucal, que sufre dos cierres, uno de los cuales cede bruscamente con entrada o salida de flujo aéreo.

Clasificación de los sonidos de la lengua (VIII)

Modo de articulación.

- Dentro del modo espirado y dependiendo del tipo de aproximación practicada por los órganos articulatorios podemos tener varias submodalidades: ***consonantes***, ***sonantes*** y ***vocales***.

1. Consonantes.

- Corresponden al máximo cierre del tracto vocal en algún punto concreto. Pueden ser:
  - Oclusivas.
  - Fricativas

Clasificación de los sonidos de la lengua (IX)

- **Oclusivas**.
  - Se produce una obstrucción total al paso del flujo aéreo, pero solamente durante breves instantes, dando lugar a una explosión posterior a la apertura del tracto vocal.
  - Su duración por lo general es de entre 10 a 30 ms.
- **Fricativas**.
  - Se produce una obstrucción incompleta del flujo, que busca los pequeños resquicios residuales que puedan existir, aumentando enormemente la velocidad del fluído en la zona.
  - Pueden tener una duración superior a los 100ms.

Clasificación de los sonidos de la lengua (X)

2. Sonantes.

- Son aquellas articulaciones en las cuales no se produce estrechamiento suficiente para que aparezca fricción, y sólo se realiza una modificación por cierre parcial o derivación.
- Pueden ser:
  - Nasales.
  - Líquidas.
  - Vibrantes.
  - Deslizantes.

Clasificación de los sonidos de la lengua (XI)

- Nasales.
  - La cavidad nasal se abre al paso del flujo aéreo, produciendo una oclusión total o casi total en algún punto de articulación de la cavidad oral.
  - La radiación es bastante menor en amplitud que la que se produciría por el tracto vocal.
- Líquidas.
  - Se basan en una obstrucción incompleta alrededor del tercio anterior de la lengua.
  - El ápice bloquea total o parcialmente al flujo glotal, pero los bordes laterales de la lengua permiten en el paso por el espacio que queda a ambos lados de la lengua.

Clasificación de los sonidos de la lengua (X)

- Vibrantes.
  - Variante de las líquidas, en las que el ápice realiza bruscas aperturas y cierres de su contacto con la zona alveolar o prepalatal, junto con la oclusión total o parcial del flujo lateral por medio de las zonas laterales de la lengua.
- Deslizantes.
  - Son sonidos fuertemente relacionados con la evolución de un sonido a otro en la articulación de un diptongo. En tal caso, uno de los sonidos de tipo vocálico se consonantiza por cierre excesivo.
  - Para su articulación requieren siempre de la presencia de una vocal.

Clasificación de los sonidos de la lengua (XI)

3. Vocales.

- Se corresponden con la mayor apertura del tracto vocal.
- La energía emitida correspondiente a una vocal suele ser mucho mayor que la correspondiente a consonantes o sonantes.
- En ningún caso llegan a presentar una obstrucción o estrechamiento superior a un 70% de su sección media en ningún punto.

Clasificación de los sonidos de la lengua (XII)

- La apertura en las vocales permite clasificarlas en cerradas, semicerradas, semiabiertas y abiertas.
- Otro rasgo de las vocales que permite clasificarlas es su punto de articulación o de máximo estrechamiento: anteriores, medias y posteriores.

Percepción de los sonidos (I)

- **Varios aspectos condicionan la habilidad de los oyentes humanos para percibir y discriminar los sonidos:**
  - Bandas críticas.
  - Altura.
  - Saturación,
  - Adaptación,
  - Enmascaramiento.
  - Supresión.
  - Inhibición lateral.

Percepción de los sonidos (II)

- **Bandas críticas.**
  - El concepto de bandas críticas está asociado al fenómeno del enmascaramiento.
    - Cuando dos tonos adyacentes se escuchan simultáneamente, el umbral de amplitud para percibir el tono más débil aumenta, es decir, el tono se enmascara por el otro, que es perceptualmente dominante.
    - Se comprueba experimentalmente que el umbral aumenta solamente si la diferencia en frecuencia entre los dos tonos está por debajo de un valor crítico.
      - Más allá de este valor crítico la percepción del primer tono no se ve afectado por el segundo.

Percepción de los sonidos (III)

- Un sonido cuya frecuencia esté dentro de una banda crítica puede influir en la percepción de otro sonido en la misma banda, pero no de fuera de ésta.
  - Experimentalmente se comprueba que el ancho de una banda crítica aumenta según sea mayor la frecuencia central de ésta.
    - Así para una frecuencia central del 200 Hz el ancho de banda es de unos 100 Hz y para 5 KHz es de 1 KHz.
- Desde un punto de vista psicológico, los filtros de banda crítica pueden considerarse como filtros de banda de paso cuya frecuencia se corresponde más o menos con las curvas de ajuste de las neuronas del sistema auditivo.

Percepción de los sonidos (IV)

- Dos sonidos en la misma banda excitarán a las mismas neuronas y cada una interferirá en la percepción del otro.
- Una medida perceptual, conocida como ***bark***, asocia la frecuencia absoluta de un sonido y la resolución en frecuencia del oído en términos de banda crítica.
  - Así un bark cubre el rango de frecuencias de una banda crítica.
  - Su expresión analítica puede encontrarse en [ZWI80].

Percepción de los sonidos (V)

- De manera similar, encontramos la escala ***mel***.
  - Esta escala define una relación entre el tono subjetivo y la frecuencia, donde la escala se ajusta para que 1000 mels correspondan a 1 KHz.
  - La correspondencia es lineal por debajo de este valor y logarítmica por encima.

Percepción de los sonidos (VI)

Altura.

- La altura de un sonido depende a la vez de la intensidad y la frecuencia.
  - La altura es el correlato subjetivo de la intensidad, de la misma manera que el tono es correlato subjetivo de la frecuencia.
- La altura es una función no lineal de la intensidad.
- Las curvas de misma altura se establecen comparando sonidos puros de diferentes amplitudes y frecuencias.

Percepción de los sonidos (VII)

- **Saturación.**
  - La respuesta de las fibras nerviosas del sistema auditivo depende de la duración e intensidad del estímulo.
  - Para una duración dada, la respuesta se incrementa con la intensidad o el nivel de presión sonora hasta un cierto umbral. Ése es el fenómeno de saturación.

Percepción de los sonidos (VIII)

- **Adaptación.**
  - Para un nivel de presión sonora, la respuesta decrece regularmente cuando la duración del estímulo crece y se acerca asintóticamente a un valor estable.
  - Éste es el fenómeno de la adaptación a corto plazo, que desempeña un papel importante en la percepción de cambios rápidos en frecuencia y amplitud.

Percepción de los sonidos (IX)

Enmascaramiento.

- La importancia del enmascaramiento se ha mostrado en la presentación de las bandas críticas.
- Otro fenómeno importante es el enmascaramiento hacia adelante de las fibras nerviosas auditivas.
  - Consiste en la disminución en la respuesta a un determinado sonido debido al sonido precedente por lo general más intenso.
- El enmascaramiento hacia atrás también ha sido propuesto [ELL62].

Percepción de los sonidos (X)

Supresión.

- Otro aspecto relacionado con el enmascaramiento es el fenómeno de la supresión de dos tonos [JAV83].
  - Experimentalmente se observa que la respuesta de los filtros neuronales auditivos a un tono puro cuya frecuencia es igual a la frecuencia característica de esos filtros puede disminuir por la presencia de otro tono puro, incluso cuando el tono por separado no produzca excitación en los citado filtros.
- De acuerdo con este fenómeno, los componentes de alta frecuencia de un sonido complejo son los más influyentes en la respuesta de las fibras nerviosas.

Percepción de los sonidos (XI)

- **Inhibición lateral [HOU71].**
  - Se puede describir como la supresión de la actividad de las fibras nerviosas de la membrana basilar causadas por la actividad de las fibras adyacentes.
  - La inhibición lateral puede contribuir a la alta selectividad con respecto a la frecuencia del sistema auditivo.

Análisis acústico elemental (I)

- **Las características espectrales de la señal de voz son no estacionarias.**
  - Razón: el sistema físico cambia rápidamente.
- **El habla se puede dividir en segmentos de sonido con propiedades acústicas similares.**
  - El análisis se realiza para periodos cortos de entre 5 y 50 ms.

Análisis acústico elemental (II)

- **Una primera división:**
  - Vocales: no hay restricción al flujo de aire.
  - Consonantes: las cavidades se ven obstaculizadas por órganos articulatorios.
- **Los sonidos que preceden o siguen a uno dado afectan a dicho sonido (fenómenos de coarticulación).**
- **Las limitaciones físicas en la producción hacen que la comunicación oral se limite a un ancho de banda de 10 KHz.**

Análisis acústico elemental (III)

- **Periodo fundamental: Tiempo entre dos aperturas sucesivas de las cuerdas vocales.**
- **Frecuencia fundamental, F0 o *pitch*: tasa de vibración de las cuerdas vocales.**
  - Rango de variación del pitch: 50-250 Hz (hombres), 120-500 Hz (mujeres).
  - A veces el pitch se define como la frecuencia fundamental percibida (en telefonía la señal está limitada a una banda de 300 - 3000 Hz aproximadamente).

Espectro de una señal de voz (I)

- Sonidos vocálicos: se observa la presencia de una excitación periódica
- Sonidos sordos: no existe tal excitación.
- En los dos casos aparecen regiones enfatizadas (resonancias), y deenfatizadas (antirresonancias).
- Estas resonancias están causadas por las reflexiones del sonido en las cavidades del tracto vocal => Cada tracto vocal esta caracterizado por un conjunto de resonancias.

Espectro de una señal de voz (II)

- Los articuladores determinan las propiedades del filtro del sistema de producción de voz.
- Resonancias -> formantes (forman el espectro).
- Teóricamente existen infinitos formantes de un sonido.
  - En la práctica usamos solamente unos pocos (entre tres y cinco).

Bibliografía

[ELL62] L. L. Elliott, “Backward and Forward Masking of Probe Tones of Different Frequencies”, *Journal of Acoustic Society of America*, Vol. 34, Nº 8, agosto 1962, pp. 1116-1117.

[HOU71] T. Houtgast, “Psychophysical Evidence for Lateral Inhibition in Hearing”, *Journal of Acoustic Society of America*, Vol. 51, Nº 6, Parte 2, abril 1971, pp. 1885-1894.

[JAV83] E. Javel et al., “Suppression of auditory nerve responses. II. Suppression threshold and growth, iso-suppression contours”, *Journal of Acoustic Society of America*, Vol. 74, Nº 3, septiembre 1983, pp. 801-813.

[JUN96] J. C. Junqua and J. P. Haton, *Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, 1996.

[ZWI80] E. Zwicker and E. Terhardt, “Analytical expressions for critical-band rate and critical bandwidth as a function of frequency”, *Journal of Acoustic Society of America*, Vol. 68, Nº 5, febrero 1980, pp. 1523-1525.