

Reconocimiento Automático del Habla

Producción de la voz

María José Castro

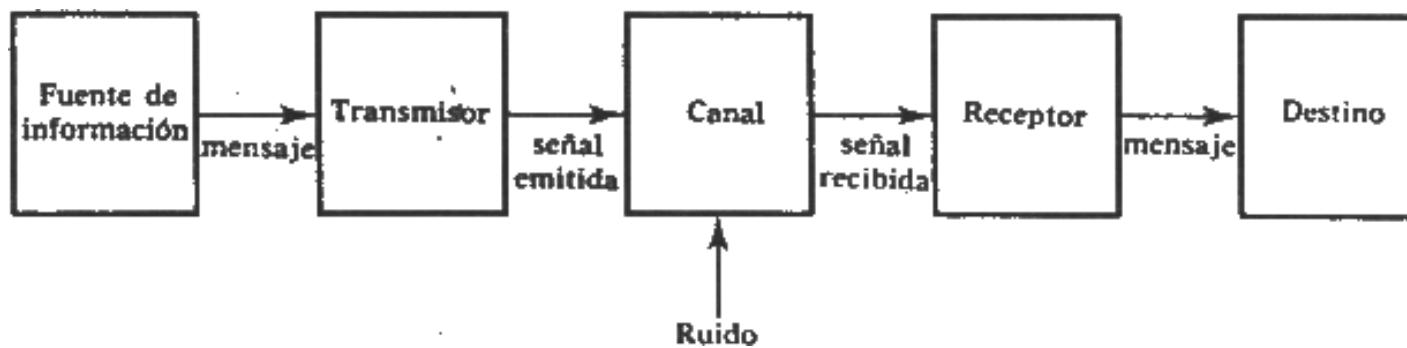
mcastro@dsic.upv.es

El proceso de comunicación

Sistemas de comunicación

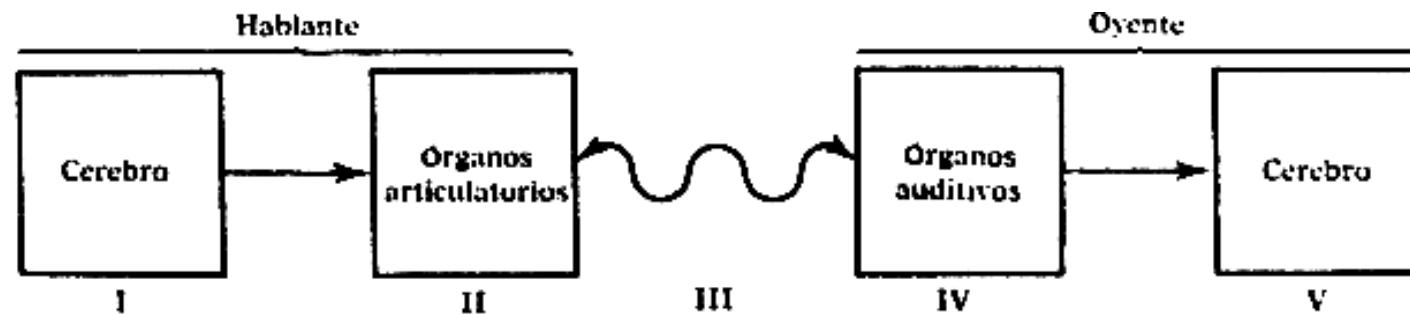
Un sistema de comunicación consta de:

- Un **emisor** o fuente de información: selecciona *signos* de un *alfabeto* para formar un *mensaje*;
- Un **transmisor**: *codifica* el mensaje siguiendo un conjunto de reglas transformándolo de la representación original a otra apta para su transmisión.
- Un **canal**: medio material usado para la transmisión de la información. Todo canal puede venir afectado por *ruido*, esto es, puede introducir defectos que ocasionan una pérdida de información.
- Un **receptor**: *descodifica* el mensaje para devolverlo a su representación original;
- Un **destino**, que recibe el mensaje.



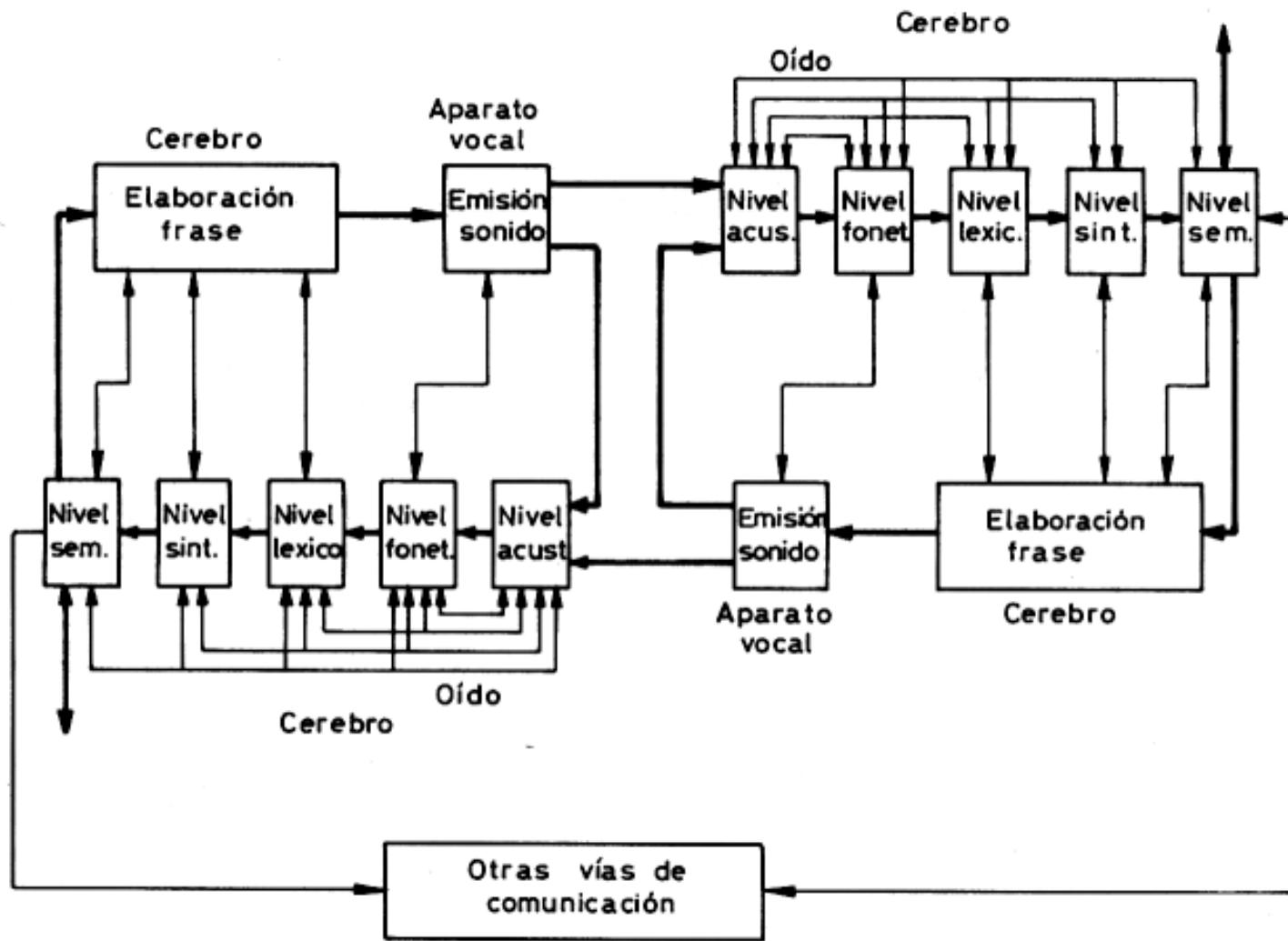
Etapas en el proceso de comunicación.

En la comunicación lingüística tenemos:



Etapas en el proceso de comunicación hablada.

- Emisor: cerebro del hablante.
- Transmisor: órganos articulatorios del hablante que generan ondas sonoras.
- Canal: aire (y, posiblemente, otras etapas).
- Receptor: aparato auditivo del receptor.
- Destino: cerebro oyente.



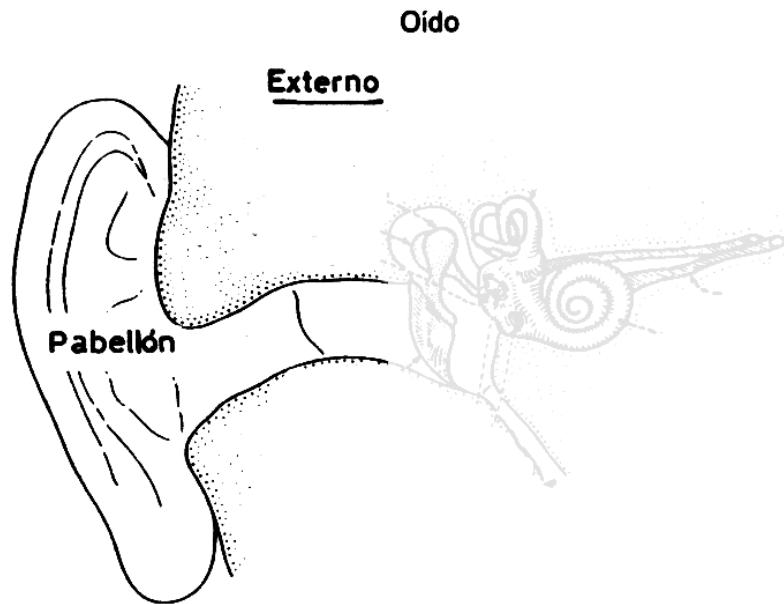
Una visión más detallada.

El oído humano

Un poco de anatomía

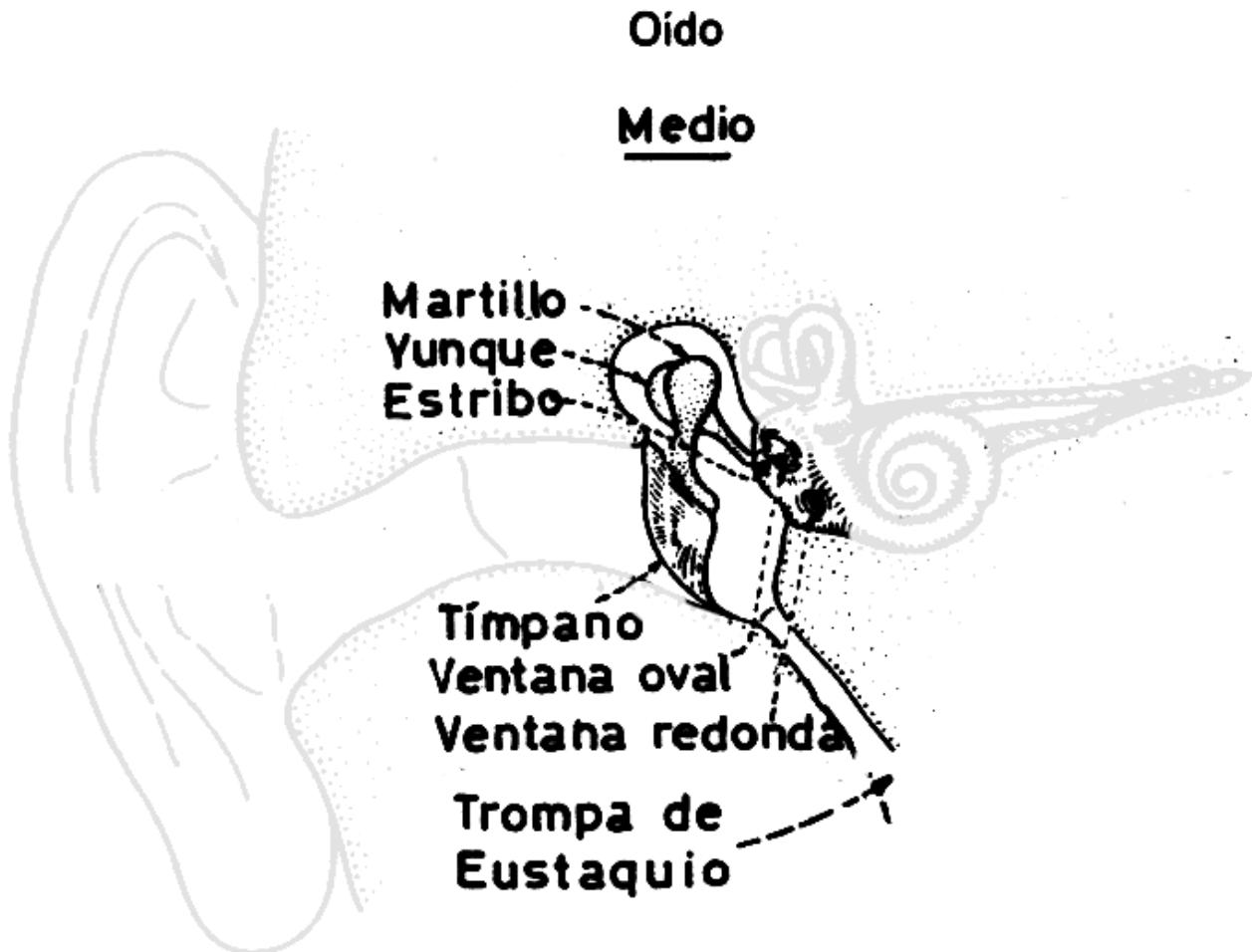
El oído se divide en tres partes:

- El **oído externo**. Es el pabellón auditivo y el conducto acústico externo, que recogen y canalizan la onda acústica.



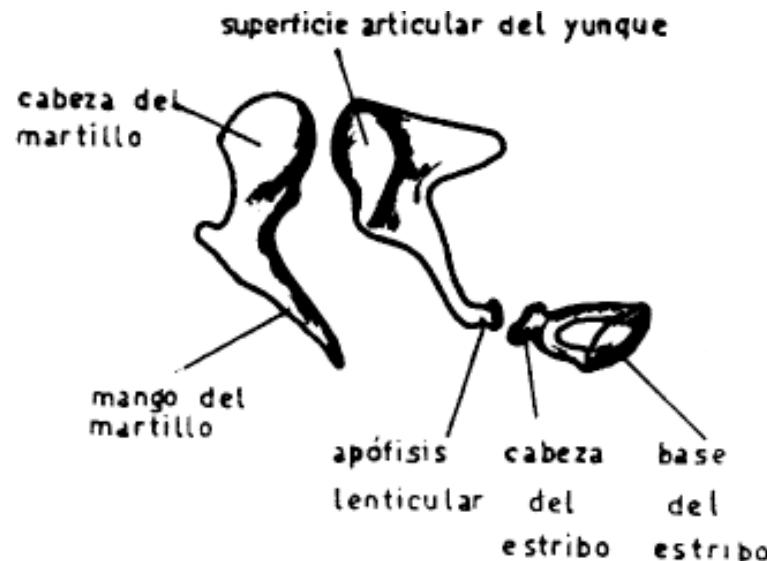
El conducto actúa como un resonador que refuerza el sonido de entre 3 y 4 kHz.

- El **oído medio**. Empieza en el tímpano (membrana elástica con forma de cono) que transforma la vibración aérea en vibración sólida. Es una cavidad con huesecillos (martillo, yunque y estribo). Termina en la ventana oval y la ventana redonda.



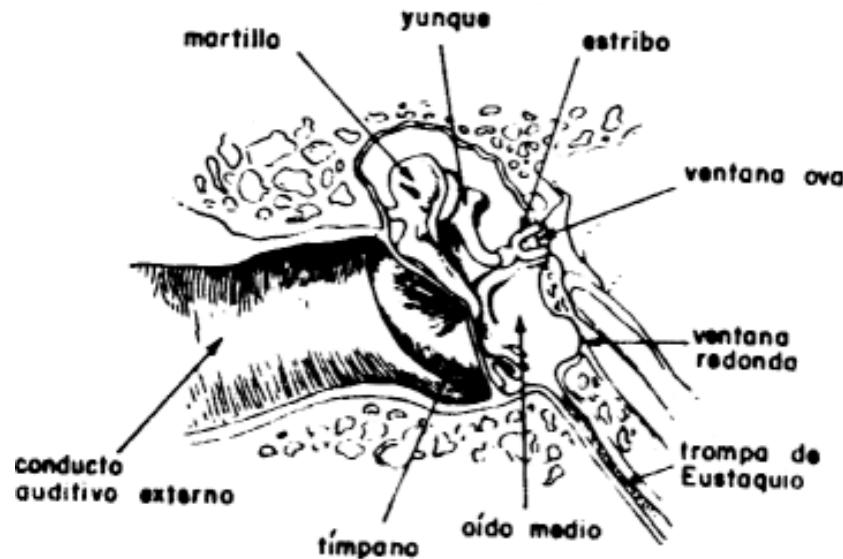
Oído medio.

La presión del tímpano se comunica al martillo, que se mueve sobre la superficie articular del yunque. La apófisis lenticular del yunque enlaza con la cabeza del estribo. La base del estribo cierra la ventana oval.



Huesecillos del oído medio.

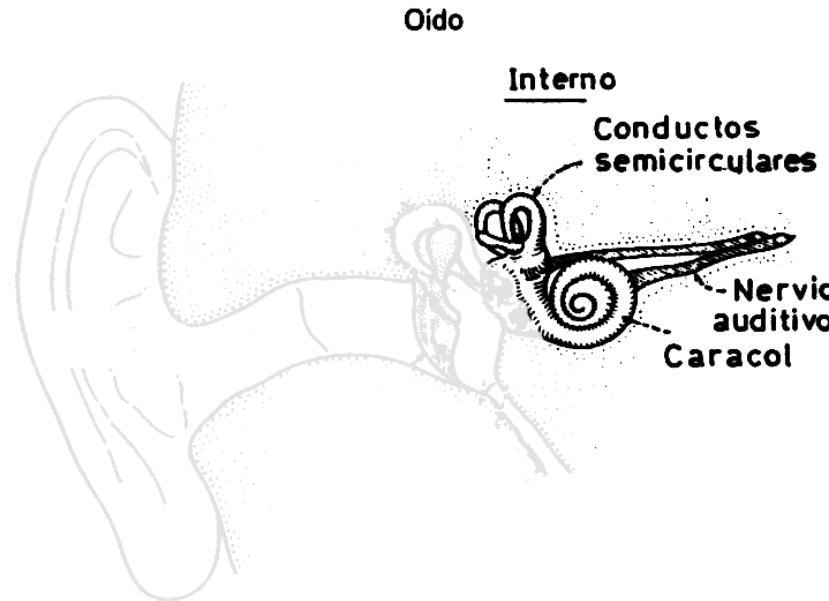
La *trompa de Eustaquio* comunica el oído medio con la faringe para que la presión de oído medio se iguale a la presión del exterior y el tímpano vibre óptimamente.



Oído interno.

El oído medio es un amplificador: los huesecillos aumentan de 25 a 30 veces la presión que llega al tímpano. Además, protege al oído medio de ruidos fuertes mediante un músculo tensor del tímpano que lo hace rígido y el músculo del estribo que neutraliza las vibraciones (acciones reflejas y voluntarias).

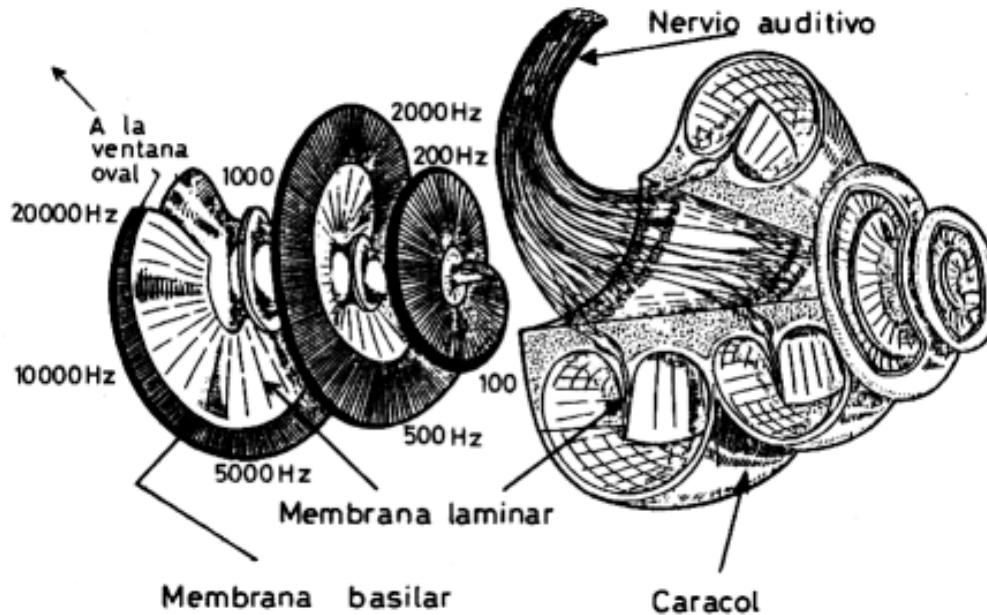
- El oído interno (también llamado laberinto).



Oído interno.

Se divide en laberinto óseo y, dentro de él, laberinto membranoso

- Vestíbulo: comunica con la caja del tímpano por medio de la ventana oval.
 - Canales semicirculares: están abiertos a la parte posterior del vestíbulo y les llegan ramificaciones del nervio vestibular. Son los responsables, junto a dicho nervio, del equilibrio.
 - Caracol óseo: transforma vibraciones mecánicas en impulsos nerviosos.



Caracol.

La lámina espiral es ósea en la parte interna y membranosa en la parte externa y divide el caracol en dos zonas: rampa timpánica y rampa vestibular. La zona final se denomina helicotrema. El caracol está lleno de líquido. Los movimientos mecánicos en la membrana basilar se convierten en señales que llegan al cerebro. Por encima de la membrana basilar está el órgano de Corti, con unas 25000 células ciliadas de las que parten las fibras nerviosas que, reunidas en haz, forman el **nervio auditivo** o *colear* (unas 30000 neuronas).

Sensibilidad del oído

La intensidad del sonido se percibe en una escala logarítmica: un crecimiento de intensidad en progresión *geométrica* se manifiesta en crecimiento en progresión *aritmética* de la sensación.

La presión del aire se mide en Pascales (Pa). La presión normal del aire es de 10^5 Pa. Podemos oír modulaciones de presión de 10^{-6} Pa.

La diferencia entre dos presiones P_1 e P_2 se mide en *decibelios*:

$$n \text{ (dB)} = 20 \log_{10} \frac{P_1}{P_2}.$$

Sensibilidad del oído

La intensidad del sonido se percibe en una escala logarítmica: un crecimiento de intensidad en progresión *geométrica* se manifiesta en crecimiento en progresión *aritmética* de la sensación.

La presión del aire se mide en Pascales (Pa). La presión normal del aire es de 10^5 Pa. Podemos oír modulaciones de presión de 10^{-6} Pa.

La diferencia entre dos presiones P_1 e P_2 se mide en *decibelios*:

$$n \text{ (dB)} = 20 \log_{10} \frac{P_1}{P_2}.$$

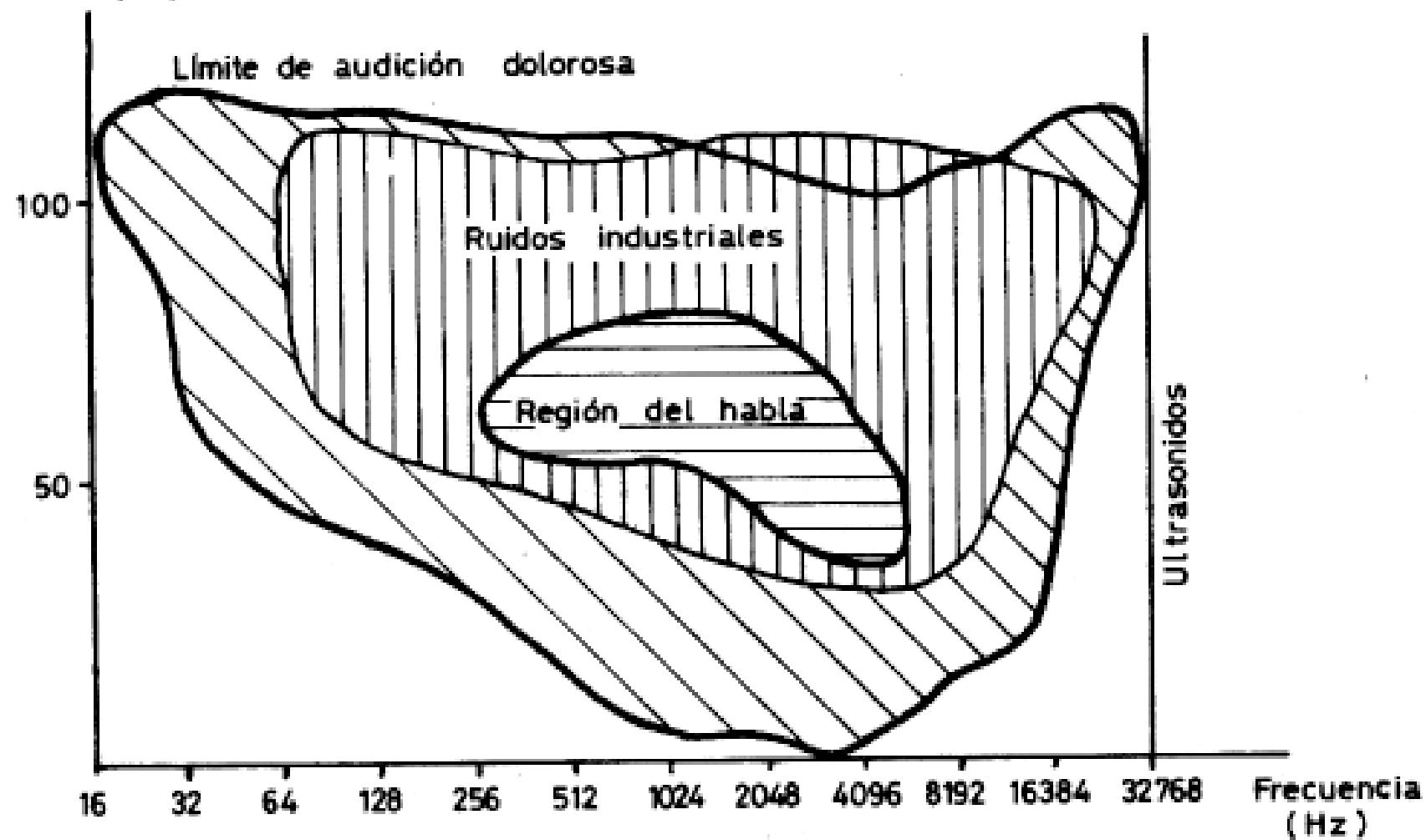
No es medida absoluta de presión, sino relativa: si una canal amplifica 3 dB hace que la salida sea 3 dB más fuerte que la entrada:

- +20 dB significa un incremento en un factor 10.
- -6 dB significa un decremento en un factor 2.

El cero de referencia son $2 \cdot 10^{-5}$ Pa.

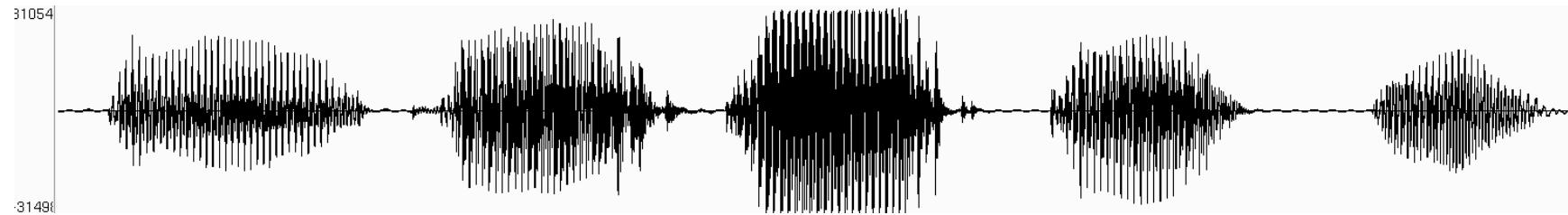
130–140 dB	umbral del dolor
120 dB	tormenta, concierto de rock
110 dB	avión a pocos metros
100 dB	taller de chapa, martillo neumático
90 dB	motocicleta
70–80 dB	calle con mucho ruido
60 dB	conversación
50 dB	automóvil
40 dB	calle en calma, refrigerador
20 dB	cuchicheo
10 dB	hojas movidas por la brisa
6 dB	límite de la audición

Nivel (dB)



La voz humana

¿Cómo se producen las ondas sonoras del habla?



Señal acústica: amplitud a lo largo del tiempo (pronunciación de "ieaou").

La señal acústica es producida por el aparato fonador y se transmite mediante ondas de presión propagadas por el aire.

El aparato fonador consta de tres elementos:

- Un **generador de energía**: los pulmones.

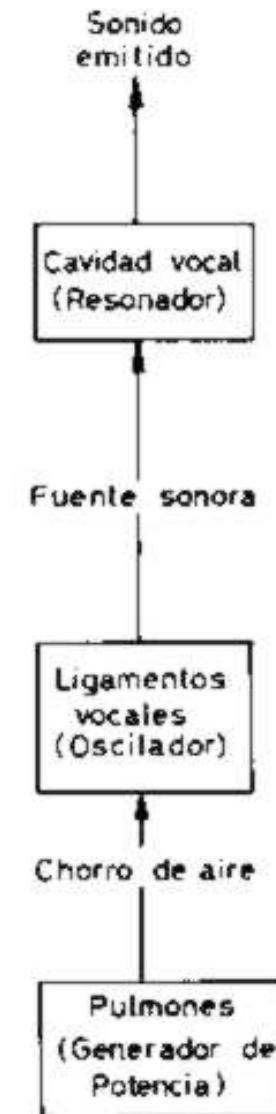
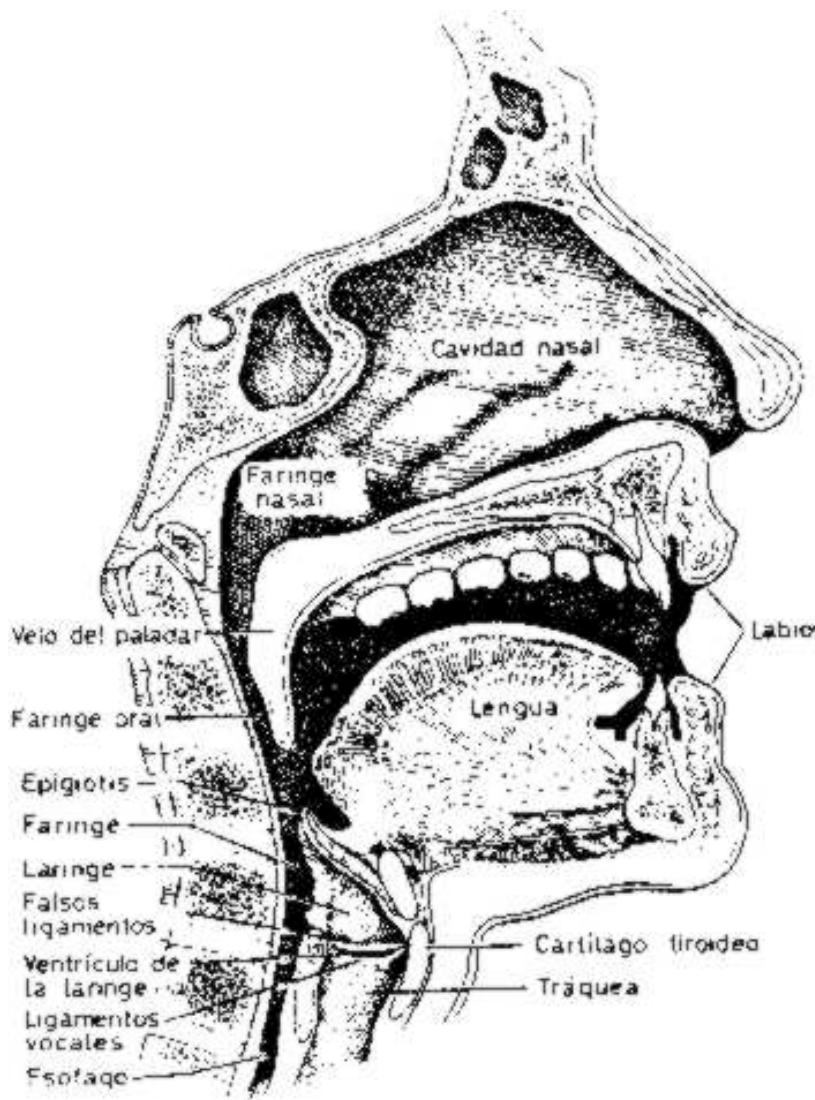
Produce la diferencia de presión que crea el flujo de aire que activa la siguiente etapa.

- Un **sistema vibrante**: laringe y cuerdas vocales.

Crea una onda rica en armónicos que es modulada en la siguiente etapa. La frecuencia de vibración se denomina frecuencia fundamental (o *pitch*). (Las mujeres tiene un pitch más agudo.)

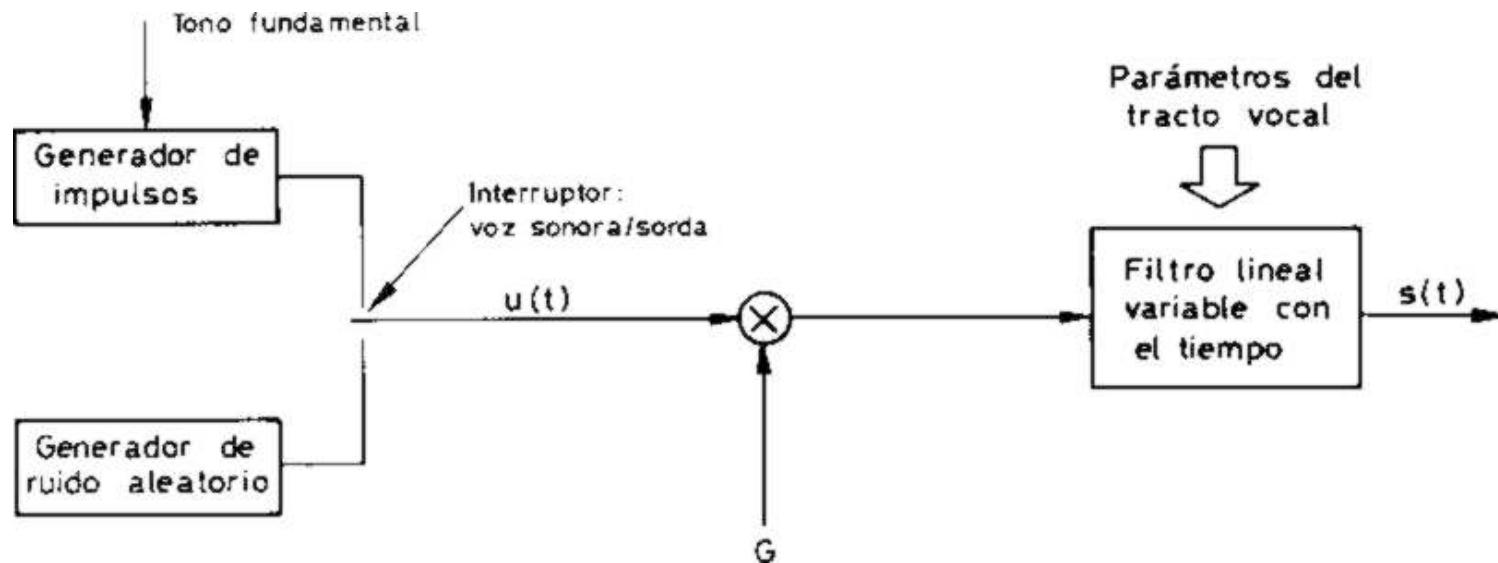
- Una **cavidad resonante**: tracto vocal.

Formado por cavidades (faringe, la boca y la nariz) deformables por elementos articulatorios (lengua, labios, mandíbulas, velo del paladar).



Aparato fonador.

Podemos modelar así el aparato fonador.



Modelo del aparato fonador.

La frecuencia del sonido humano se encuentra en el rango 300–16000 Hz, aunque la mayor parte se encuentra por debajo de los 7000 Hz. (Si la señal es telefónica, la máxima frecuencia es de 3400 Hz).

Fonética

La fonética es la ciencia que estudia los sonidos del habla humana.

Los sonidos producidos en el habla pueden transcribirse usando un alfabeto fonético.

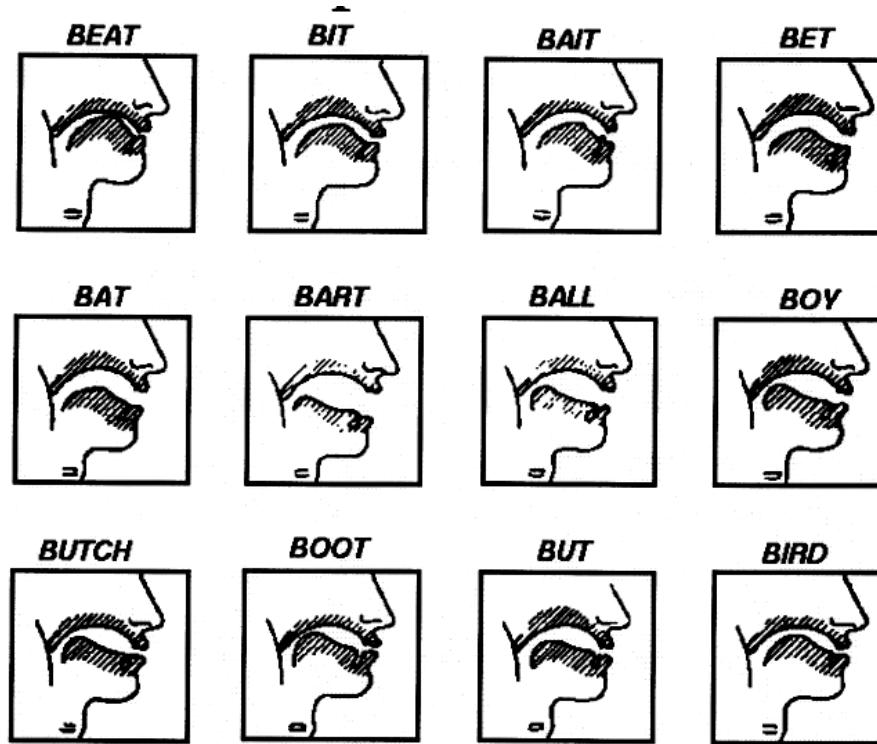
Vocales	abocinadas			anterior		central		posterior	
cerradas	(y	θ	u)	i	y	i	θ	w	u
semicerradas	(ø	o)	e	ø			x	ø	o
cerradas						θ			
semiabiertas	(œ	ɔ)		ɛ	œ			æ	ʌ
abiertas	(ɒ)				a		d	ɒ	

	labial	labio- dental	dental, alveolar o post-alveolar	retrofleja	palato- alveolar	palatal	velar	uvular	labio- palatal	labio- palatal	faringea glotal
nasales	m	n		n	ɳ	jŋ	ŋ	N			
occlusivas	p b	t d	t̪ d̪		c ɟ	k g	q G		kpgb		?
fricativas:											
centrales	ɸ β	f v	θ ð s z	ʂ ʐ	ʃ ʒ	x ɣ	χ ʁ		m w	h f	h f̪
aproximantes		v	x	ɿ	j	w	ɥ				
con aire				ɬ ɭ							
expirado				i l							
fricativas											
aproximantes											
laterales:											
vibrantes:											
múltiples				r				R			
simples				r̪	ɻ			R̪			
sin aire	eyectivas	p'		t'			k'				
expirado	inyectivas	θ		d			g				
clítes:											
centrales	ʘ	ʈ	ɖ								
laterales				ʂ							

¿Cómo se pronuncian los fonemas? ¿Cómo se perciben los fonemas? Hay dos aproximaciones a la ciencia fonética: fonética articulatoria y fonética acústica.

Fonética articulatoria

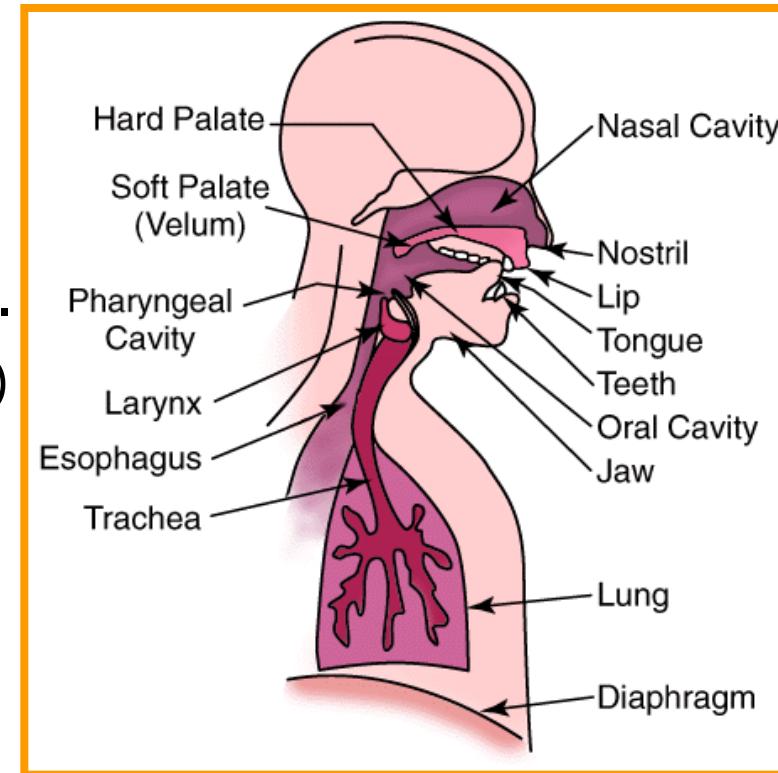
Estudia las propiedades del mecanismo fisiológico de producción de los sonidos y caracteriza los sonidos mediante descripciones del estado del aparato fonador durante la emisión del sonido.



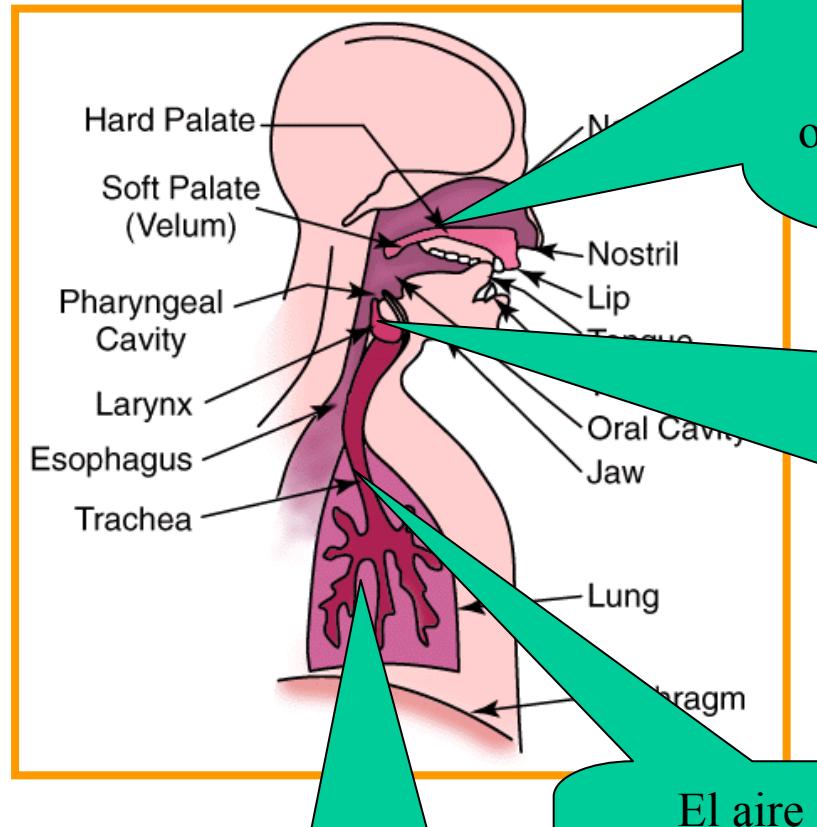
Cada posición de los órganos articulares da un lugar a un sonido determinado. Cualquier modificación origina otro diferente.

Aparato Fonador Humano

- Órganos Respiratorios. (Sistema Subglotal):
 - Pulmones, Bronquios y Traquea.
 - Fuente de Energía.
- Órganos Fonadores (Cavidades Glóticas):
 - Laringe, Cuerdas Vocales y Resonadores (nasal, bucal y faríngeo).
- Órganos Articulatorios (Sistema Supraglotal)
 - Paladar, Lengua, Dientes y Labios.
- Voz: Onda acústica radiada cuando los pulmones expulsan el aire y el flujo resultante es perturbado por alguna constricción en el tracto vocal.



Mecanismo de Producción del Sonido



El aire es expulsado con fuerza desde los pulmones

El aire gana velocidad en la tráquea

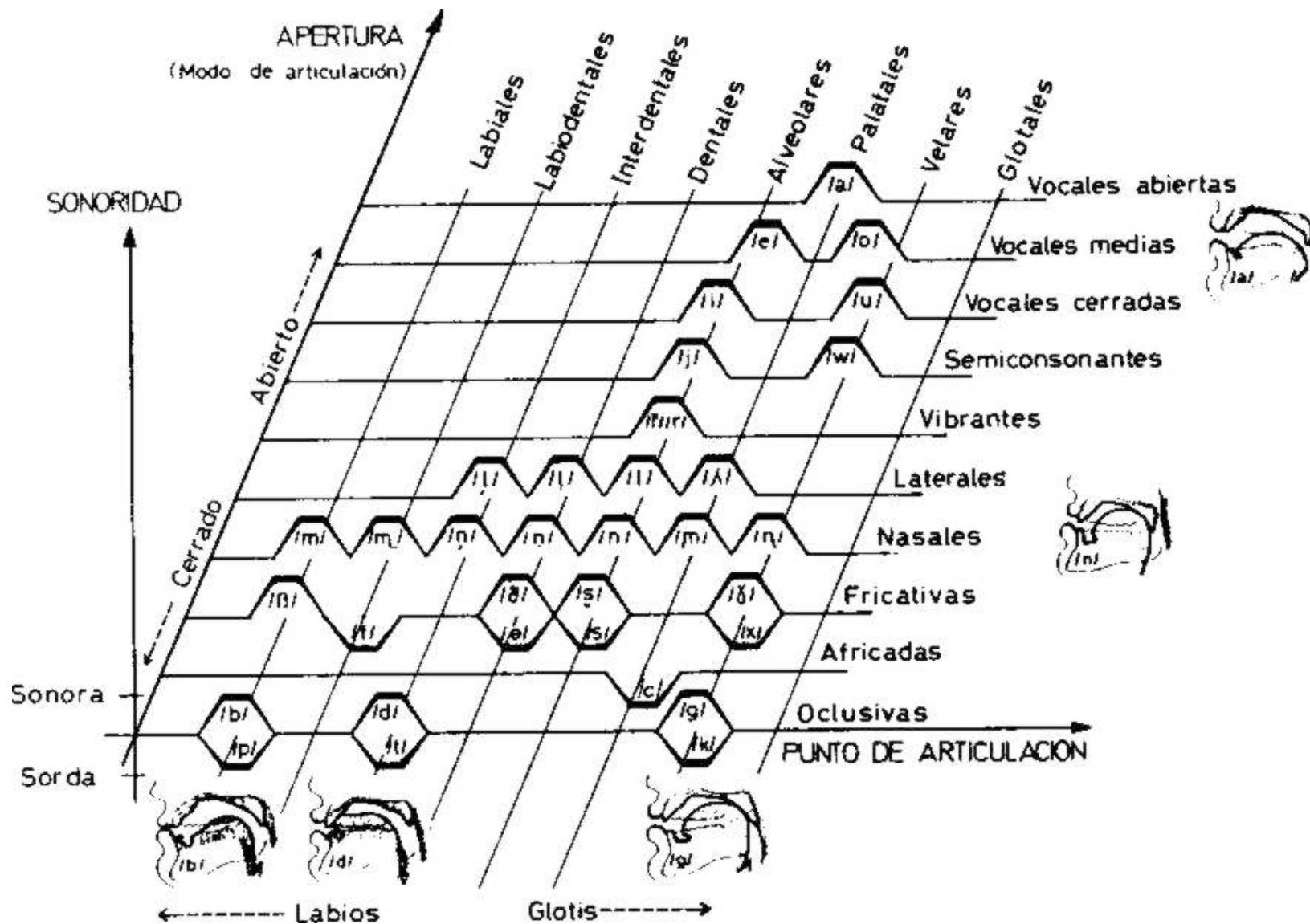
Sonidos Sonoros: Los Órganos articulatorios, paladar, lengua, dientes, etc. conforman espectralmente el sonido al adoptar la posición apropiada.

Sonidos Sordos: El aire encuentra alguna oposición a su paso en algún punto del tracto vocal.

Sonidos Sonoros: El aire a su paso hace vibrar las cuerdas vocales. Éstas se encuentran tensas cerrando la glotis.

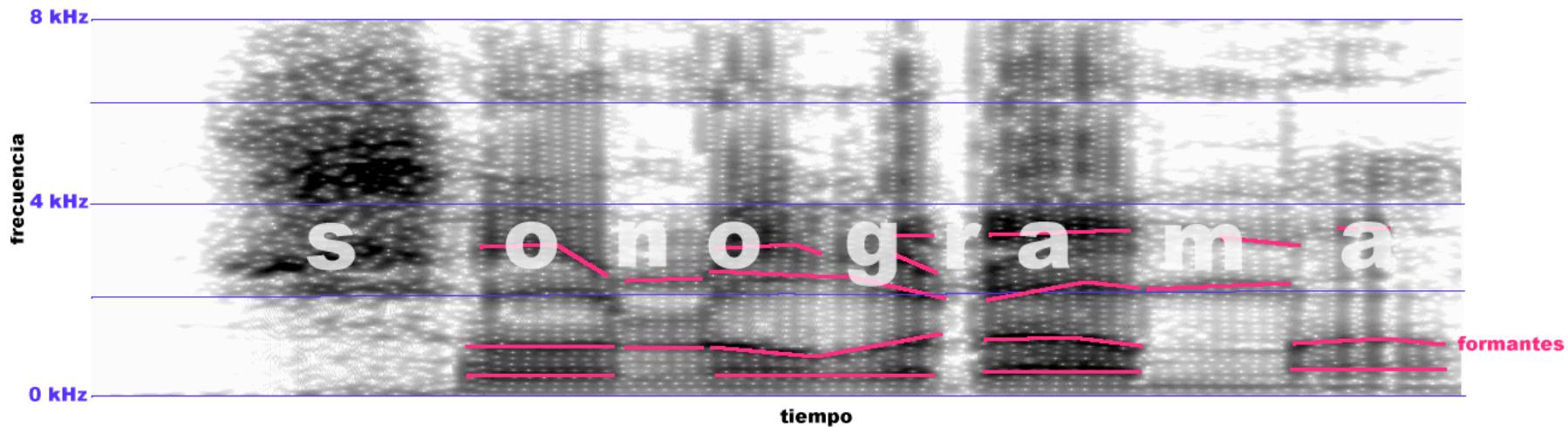
Sonidos Sordos: El aire fluye sin obstáculos a través de la glotis. Ésta se encuentra abierta con las cuerdas vocales relajadas.

He aquí la relación entre sonidos y configuraciones del aparato fonador para el español.



Fonética acústica

Estudia las propiedades de la forma de onda producida: duración, energía, etc. Requiere de instrumental de análisis para obtener medidas de esas propiedades.



Los fonéticos encuentran rasgos distintivos en los sonidos.

- Rasgos prosódicos.
 - Físicos:
 - * De frecuencia fundamental (tono): alto ↔ bajo.
 - * Intensidad (fuerza): acento dinámico (intersilábico) y *stosston* (intrasilábico).
 - * Cantidad.
 - Psicológicos:
 - * Tonía (altura de la voz).
 - * Sonía (estrépito).
 - * Duración (duración subjetiva o protensidad).
- Rasgos intrínsecos.
 - De sonoridad.
 - * Vocálico-no vocálico.
Presencia/ausencia de estructura formántica.
 - * Consonántico-no consonántico.
Disminución/aumento de energía y presencia/ausencia de zonas no resonantes.
 - * Compacto-difuso.

Concentración/ausencia de energía en zona estrecha.

- * Tenso-laxo.
Definición/indefinición de zonas resonantes y aumento/disminución de energía total en el tiempo.
- * Sonoro-sordo.
Superposición o no de fuente armónica sonora.
- * Nasal-oral.
Reducción o no de frecuencia en el primer formante de vocales y aparición de formantes en ciertas frecuencias de las consonantes.
- * Interrupto-continuo.
Presencia/ausencia de momento de silencio seguido de difusión de energía en una banda ancha.
- * Estridente-mate.
Propios sólo de las consonantes. Irregularidad/regularidad de las áreas de frecuencias.
- * Bloquedad/no bloquedad. Alta/baja proporción de descarga de energía en breve/largo intervalo de tiempo.

– De tonalidad.

* Grave/agudo.

Predominio de la zona baja/alta del espectro.

* Bemolizado/no bemolizado (o normal).

Descenso o no de la línea de frecuencias de algunos o todos los formantes.

* Sostenido/no sostenido.

Elevación o no del segundo formante.

La determinación de qué rasgos se manifiestan en cada momento es difícil. Las definiciones de los rasgos son un tanto laxas.

Estudiaremos el sonograma de las pronunciaciones para ver qué características presentan al pronunciar cada fonema (o secuencia de fonemas) y cómo evolucionan éstas.

Vocales

Las vocales en español son fácilmente diferenciables (al menos más fácilmente que en otras lenguas)... pero no aportan tanta información como la consonantes:

- e_ _e__o _i_ _o_a_e_ _o e_ _i_i_i_ _e e__e_e_

Vocales

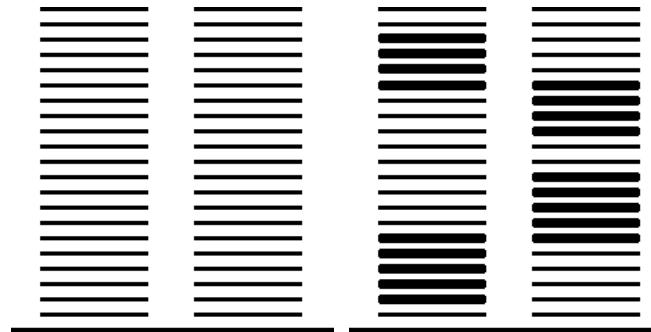
Las vocales en español son fácilmente diferenciables (al menos más fácilmente que en otras lenguas)... pero no aportan tanta información como la consonantes:

- e_ _e__o _i_ _o_a_e_ _o e_ _i_i_i_ _e e__e__e_
- _l t_xt_ s_n v_c_l_s n_ _s d_f_c_l d_ _nt_nd_r

Se producen por la vibración de las cuerdas vocales, que es filtrada por la boca/nariz en una posición prácticamente fija.

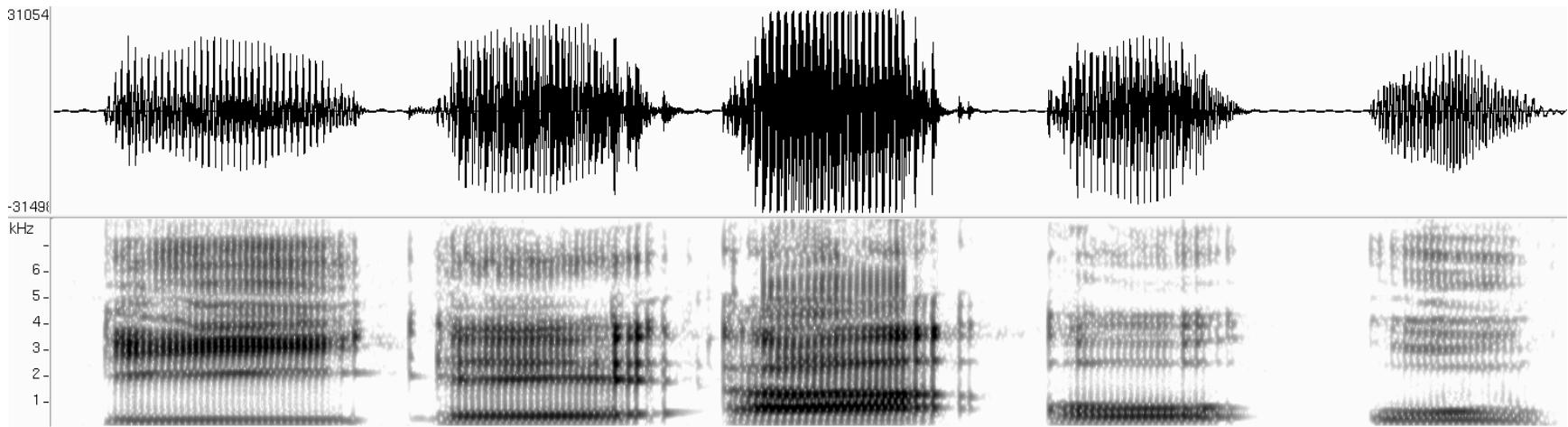
Su duración es larga en comparación a otros sonidos.

En principio, la vibración es una onda compuesta con una frecuencia fundamental (o *pitch*) y una serie de armónicos (frecuencias múltiplo del pitch) con igual potencia:



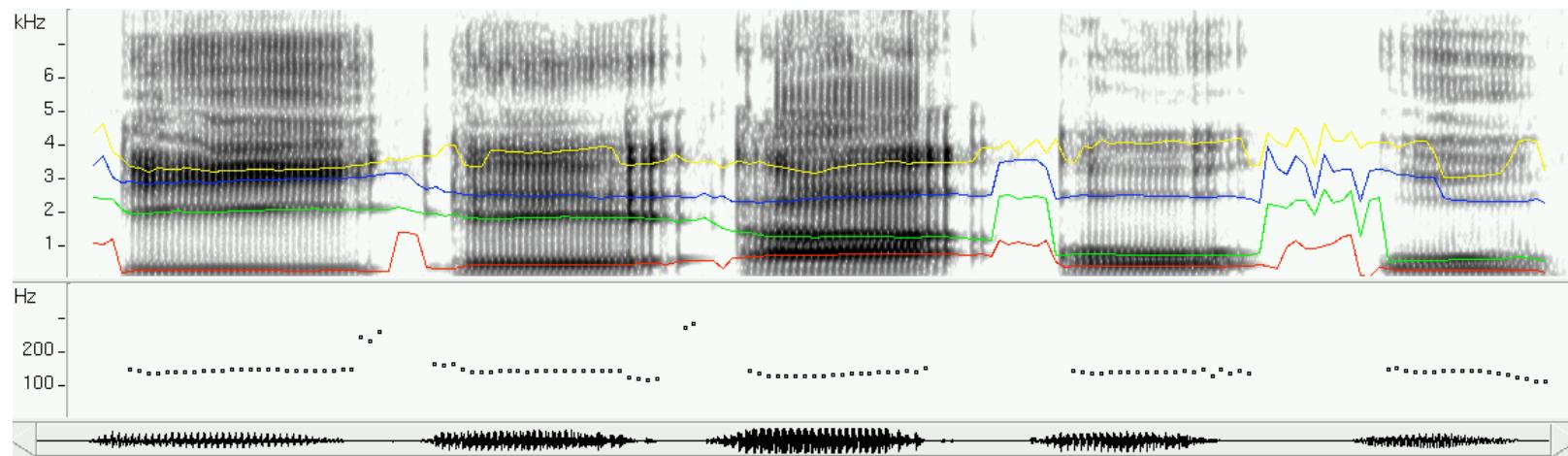
La boca actúa como resonador y realza ciertas frecuencias.

Los formantes son bandas oscuras en el sonograma:



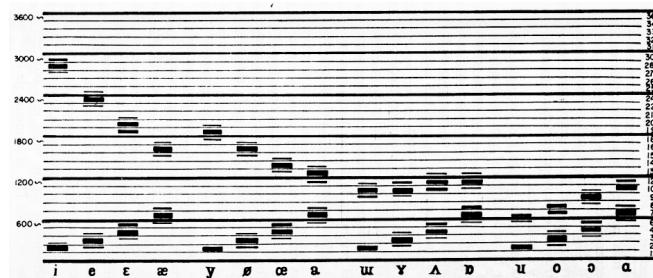
Pronunciación de “ieaou”.

Cada conjunto de armónicos (múltiplos del pitch) reforzados es un **formante**.



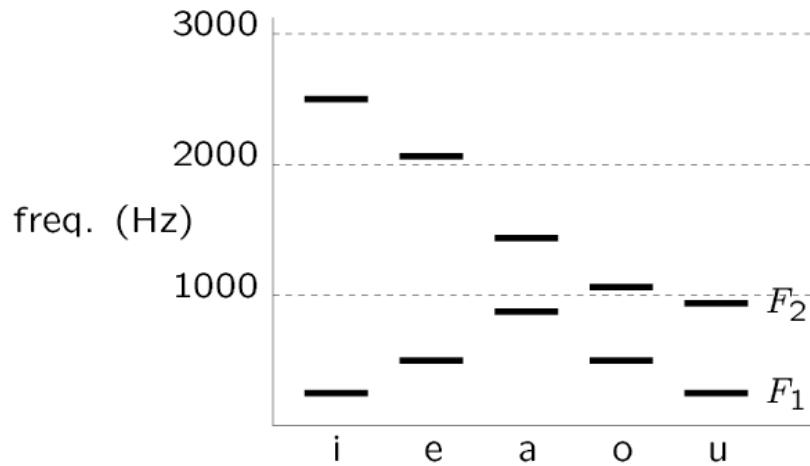
Formantes (arriba) y pitch (abajo) en la pronunciación de “ieaou”.

Los dos primeros formantes (los de más baja frecuencia) son críticos para la percepción de la vocal.



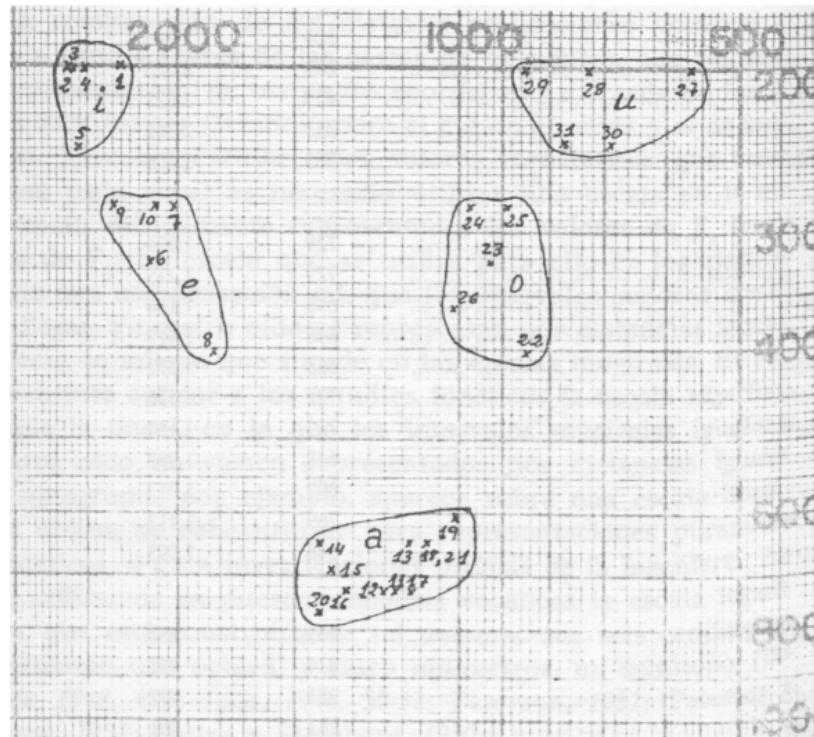
Primeros dos formantes de diferentes vocales.

El tercer formante aporta información cuando el segundo es alto.



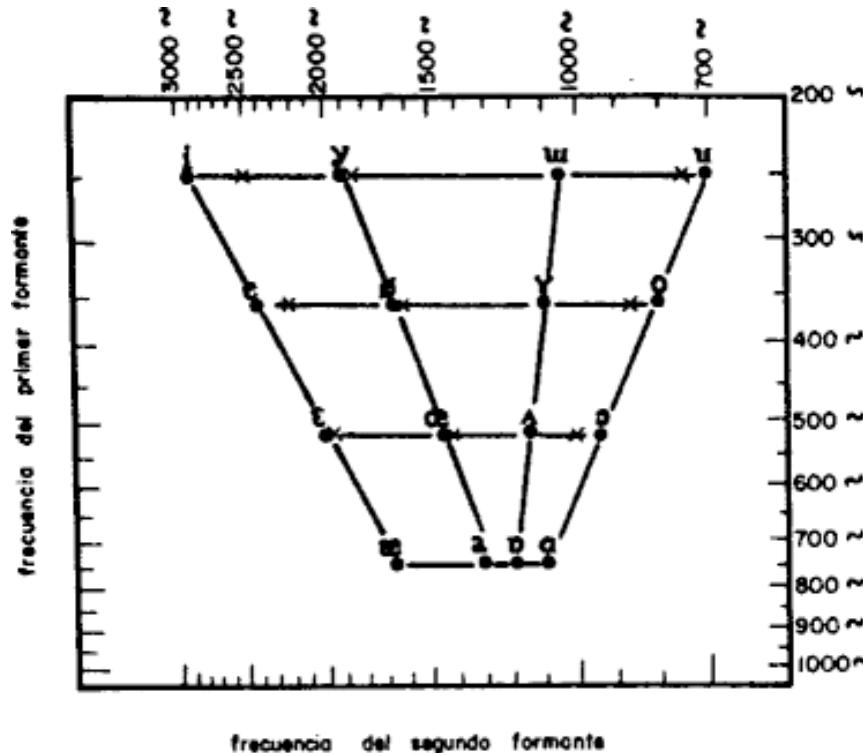
Los dos primeros formantes en las vocales españolas.

Podemos hacer un “mapa” de las vocales utilizando la frecuencia de sus dos primeros formantes como sistema de coordenadas:



Mapa vocálico del español a partir de los dos primeros formantes.

Este mapa da lugar al denominado *triángulo acústico de las vocales*.

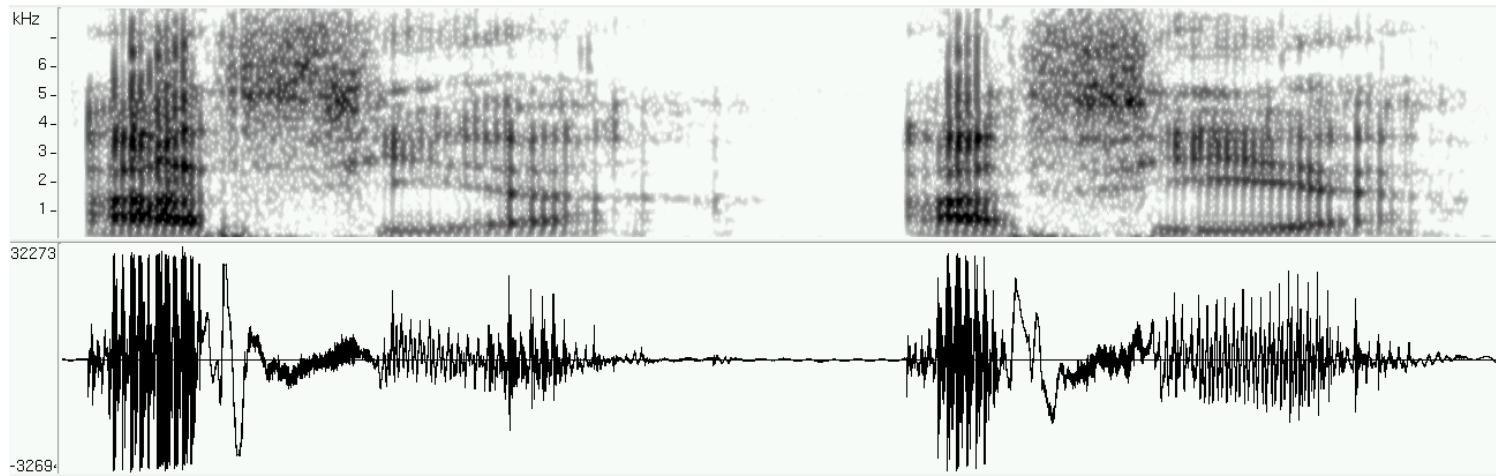


En español, las vocales pueden aparecer en secuencias. Los diptongos son transiciones suaves de una vocal a otra. Normativamente:

- Diptongos crecientes: /i, u/ + /e, a, o/.
La /i/ y la /u/ son márgenes silábicos.
- Diptongos decrecientes: /e, a, o/ + /i, u/.
La /i/ y la /u/ son márgenes silábicos.
- /i/ + /u/.
- /u/ + /i/.

Además, aparecen otras secuencias: /e, a, o/ + /e, a, o/.

Las secuencias de vocales se caracterizan por presentar sonogramas con un “deslizamiento” de formantes de la configuración de una vocal (o similar) a otra (o similar):



La /i/ y la /u/ dan lugar a las *semiconsonantes* /j/ y /w/, respectivamente.

Consonantes explosivas orales

Son sonidos de transición, no continuos, con presión sobre una constricción total (/b/, /p/ en labios; /d/, /t/ tras los dientes; /g/, /k/ cerca del velo).

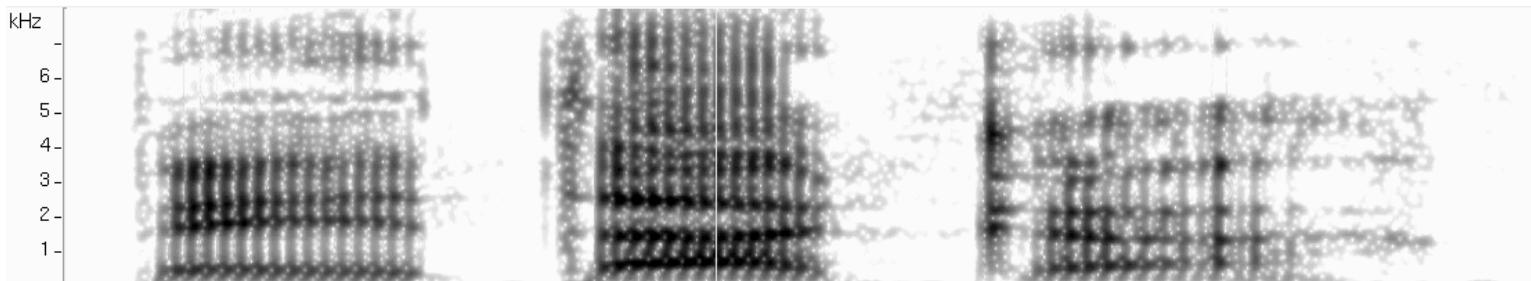
p, b, t, d, k, g

El término se debe a que el momento más audible es el de una explosión. Son difíciles de distinguir.

Desde el punto de vista articulatorio se las denomina *occlusivas orales*, pues se caracterizan por un cierre momentáneo del canal bucal.

Los fonemas explosivos se dividen acústicamente en **sordos** y **sonoros**.

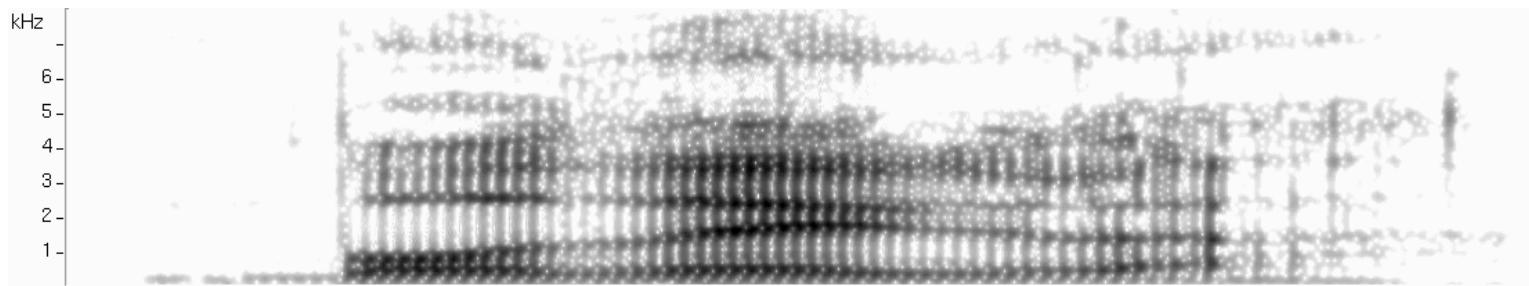
- sordos: p, t, k.



Pronunciación de “petaca” /petaka/.

En el sonograma se aprecia una zona de silencio seguida de un estallido (burst) claramente diferenciado del segmento vocálico. Toda la información discriminativa está en el burst. Al pronunciar una frase, las regiones de silencio suelen corresponder a los fonemas explosivos, y no a pausas entre palabras (que no suele haberlas).

- sonoros: b, d, g.



Pronunciación de “bodega” /bodega/.

En el sonograma no se aprecian regiones de silencio, aunque sí de disminución de energía.

En el segmento consonántico se aprecia el trazo de la transición de formantes de las vocales que flanquean al fonema explosivo.

Desde el punto de vista articulatorio se dividen en labiales (/p, b/), dentales (/t, d/) y velares (/k, g/).

En posición *prenuclear* (al principio de una sílaba, antes de vocal) se manifiestan plenamente.

En forma *postnuclear* (al final de sílaba) se manifiestan de formas muy diversas, llegando incluso a desaparecer:

/doktór/ /dogtór/ /doγtór/ /doutór/ /dotór/

En posición *prenuclear* (al principio de una sílaba, antes de vocal) se manifiestan plenamente.

En forma *postnuclear* (al final de sílaba) se manifiestan de formas muy diversas, llegando incluso a desaparecer:

/doktór/ /dogtór/ /doγtór/ /doutór/ /dotór/

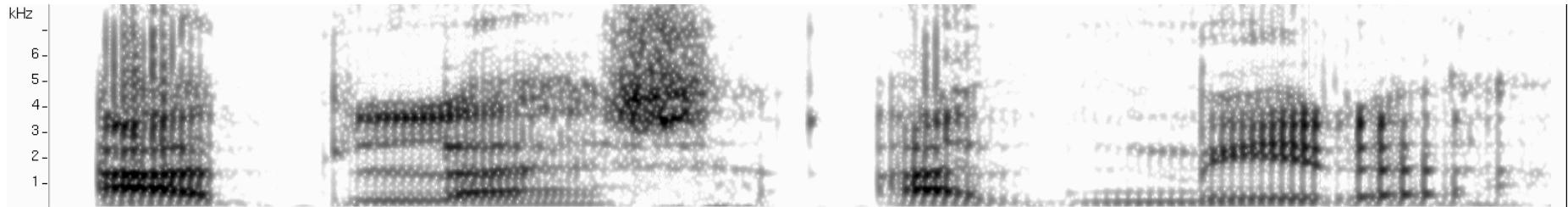
Es habitual que en posición postnuclear se manifiesten como fricativos. El resultado de la neutralización de las explosivas son los archifonemas /B, D, G/. (Un **archifonema** se produce cuando dos fonemas pierden sus rasgos distintivos).

apnea /áBnea/ ábside /áBside/



Pronunciación de “apnea ábside”.

atlas /áDlas/ admira /aDmíra/



Pronunciación de “atlas admira”.

acta /áGta/ signo /síGno/



Pronunciación de “acta signo”.

Consonantes explosivas nasales

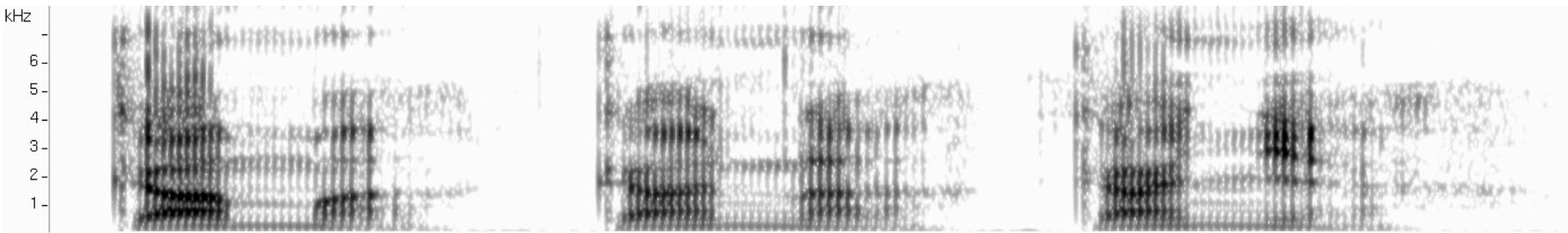
La cavidad oral está ocluida (en /m/, los labios; en /n/ tras los dientes). El velo desciende para que el aire pase por la nariz.

Hay tres explosivas nasales en español: /m, n, ñ/.

cama /káma/

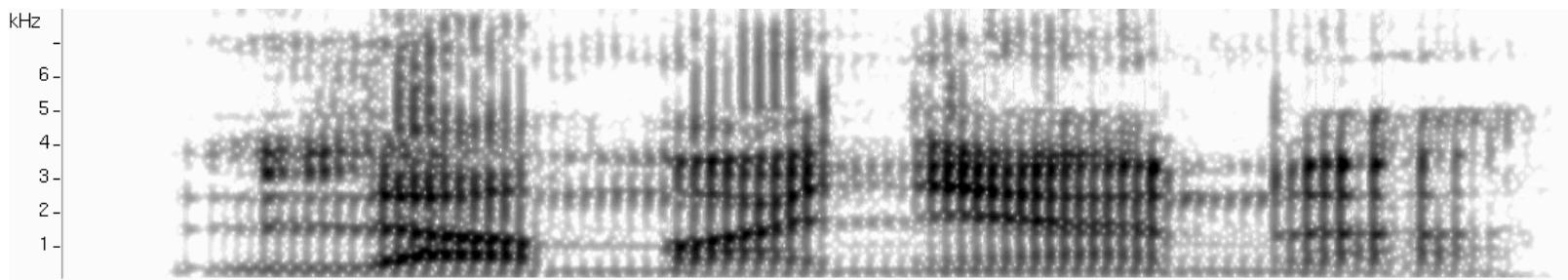
cana /kána/

caña /káña/



En posición prenuclear se manifiestan plenamente.

- La /m/ presenta tres formantes con independencia de la vocal silábica.
- La /n/ presenta los formantes 1 y 3 y pocas veces el formante 2.
- La /ŋ/ suele presentar sólo el primer formante y aparecer en blanco la zona de altas frecuencias..



Pronunciación de “la mañana”.

El posición postnuclear, /m/ y /n/ se manifiestan como el archifonema /N/ y se asimila al sonido siguiente:

- /-N/ + /n/ (consonante alveolar): un lado /unláðo/.
- /-N/ + /m/ (consonante bilabial): un pan /unpán/.
- /-N/ + /ŋ/ (consonante labiodental): un farol /umfaról/.
- /-N/ + /ɳ/ (consonante dental): un tomo /uɳtómø/.
- /-N/ + /ɳ/ (consonante interdental): un cero /uɳθérø/.
- /-N/ + /ɳ/ (consonante palatal): un chico /unɳfíko/.
- /-N/ + /ŋ/ (consonante velar): un caso /uɳkáso/.

Consonantes fricativas

Se produce una fricción del aire al pasar por una estrechez entre órganos articulatorios. Se caracterizan por el ruido de fricción y modifican los formantes vocálicos contiguos.

- Fricativas de resonancias bajas (sonoras).

- /β/ (labial)

Se diferencia de /b/ en la presencia de zonas que se aproximan a los formantes vocálicos.

bomba /bómba/ boba /bóβa/

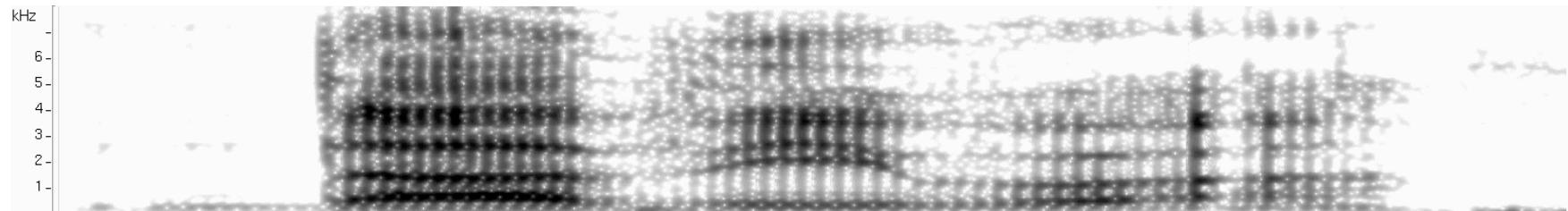


Pronunciación de “bomba boba”.

- /ð/ (dental)

Se diferencia de /d/ en la presencia de zonas que se aproximan a los formantes vocálicos.

dádiva /dáðiβa/



Pronunciación de “dádiva”.

- velar /γ/

Se diferencia de /g/ en la presencia de zonas que se aproximan a los formantes vocálicos.

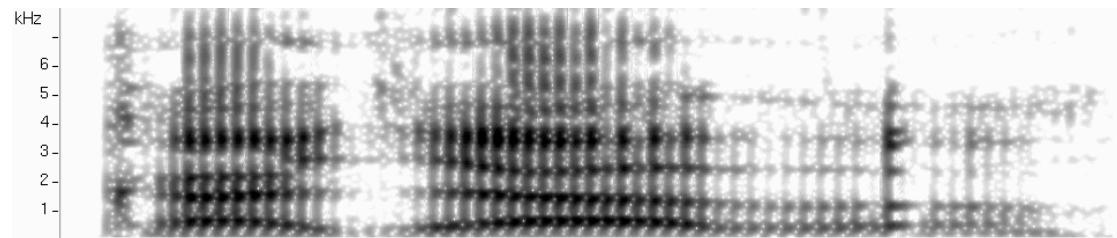
venga /bén̪ga/ vega /béγa/



Pronunciación de “venga vega”.

- /j/ (palatal)

cayado /cajádo/



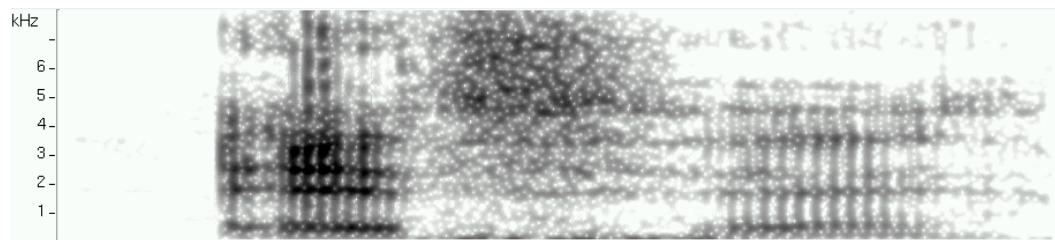
Pronunciación de “cayado”.

- Fricativas de resonancias altas (sordas).

Un flujo estable de aire se vuelve turbulento en una región constrictiva. La energía se concentra en frecuencias altas y es de naturaleza no periódica.

- /f/ (labial)

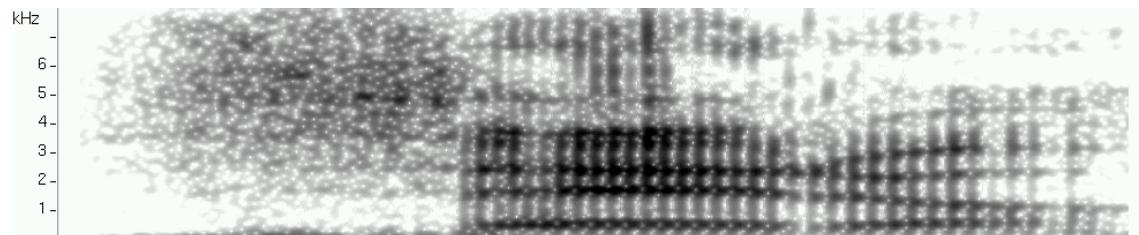
Intensidad débil.



Pronunciación de “efe”.

- /θ/ (dental)

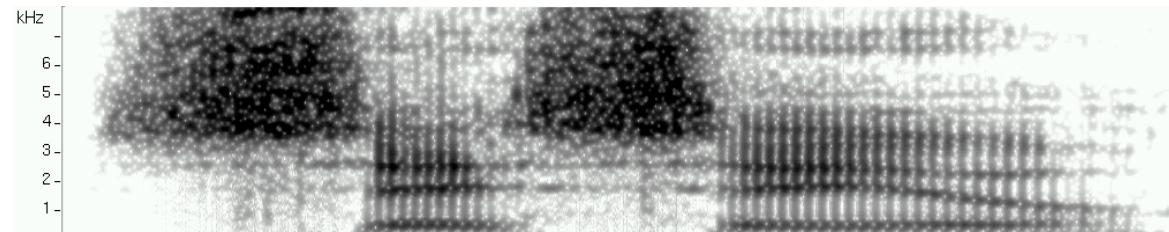
Su frecuencia varía mucho (/θa/ presenta un rango de frecuencias muy diferente de /θu/). Intensidad débil.



Pronunciación de “cero”.

- /s/ (alveolar)

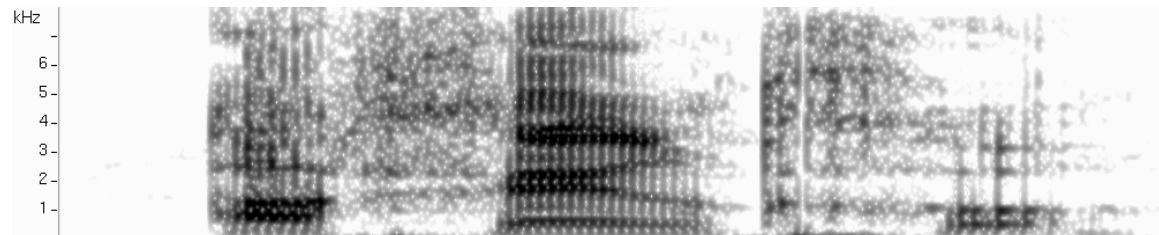
Ruido muy intenso en alta frecuencia. Presenta muchas realizaciones diferentes.



Pronunciación de “seseo”.

- /χ/ (velar)

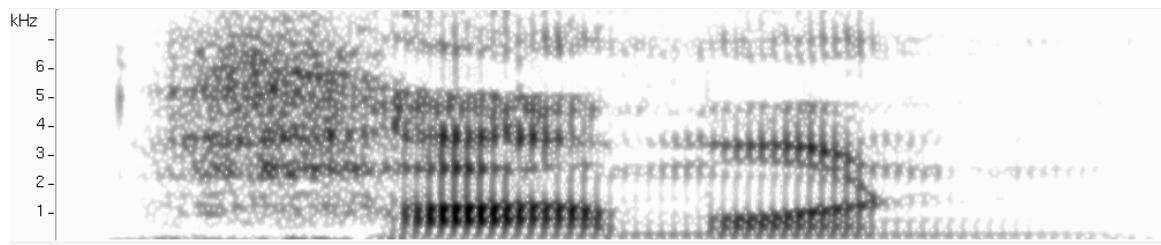
Es vibrante, por lo que presenta estriaciones verticales en el espectro.



Pronunciación de “ajenjo”.

(En Chile hay un alófono de /χ/ ante /i/ y /e/: /ç/.)

- /h/



Pronunciación de “jamón” (como /hamón/).

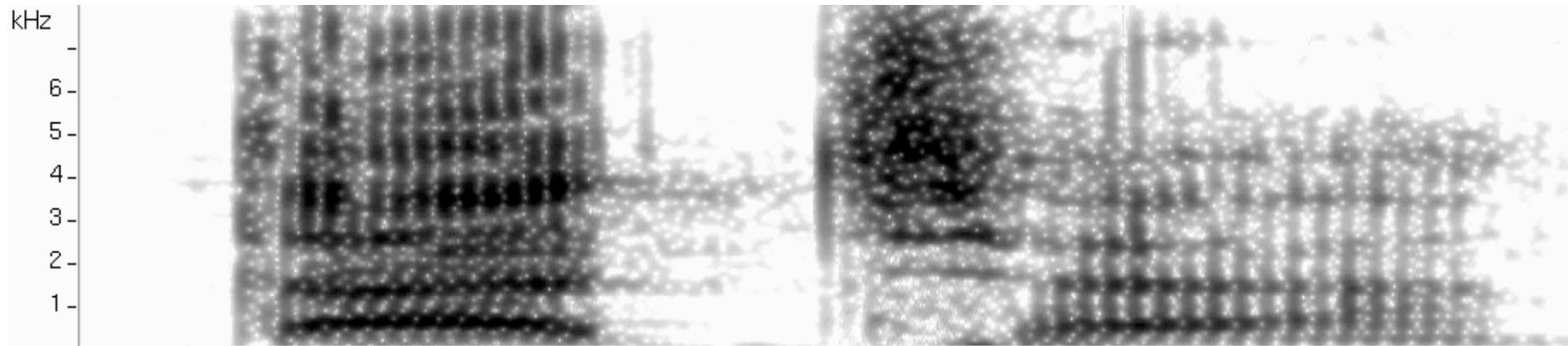
(Propio del acento andaluz y de algunos países latinoamericanos.) Breve sonido turbulento. Muy débil acústicamente.

Africadas

Aparece un momento interrupto y otro constrictivo en su realización.

- Africadas sordas.

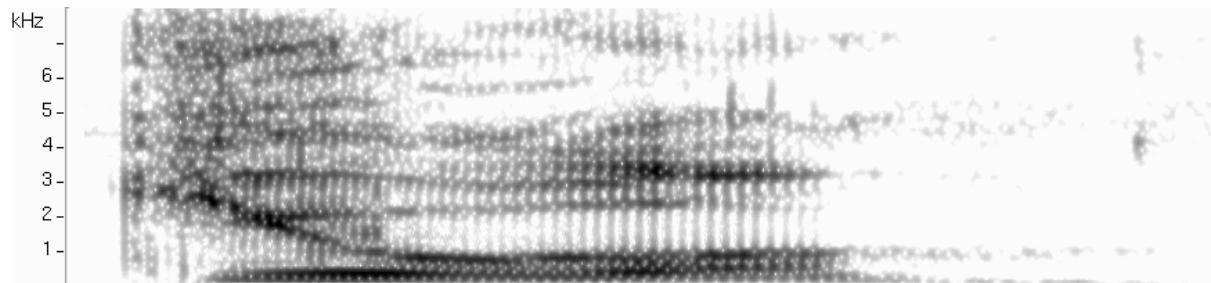
/tʃ/ (tacha /táʃa/).



Pronunciación de “tacha”.

- Africadas sonoras.

/χ/ (yugo /χúγo/).



Pronunciación de “yugo”.

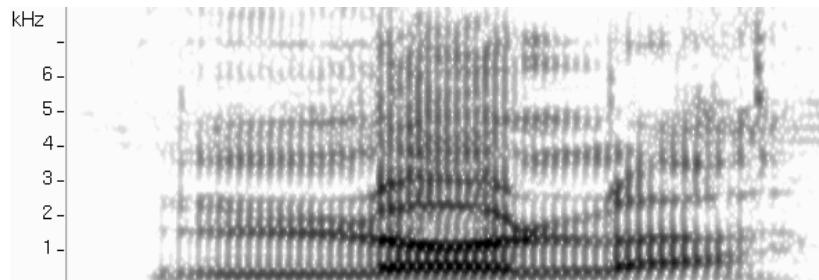
Muchas diferencias en su realización (acentos).

En ocasiones (tras consonante nasal, como en “cónyuge” (/kónχuxε/), no presenta momento interrupto.

Líquidas

- Laterales.

- /l/



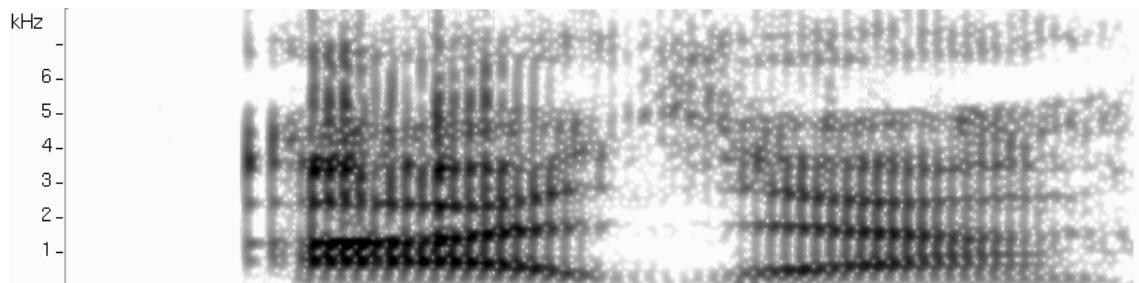
Pronunciación de “lola”.

Aparecen formantes similares a los vocálicos.

Presenta muchos alófonos:

- * /l/ alveolar (ala /ála/),
- * /l/ linguodental (alto /álto/),
- * /l/ linguointerdental (alza /á!θa/),
- * /l/ linguopalatalizada (colcha /kó!tʃa/).

$$-|\lambda|$$

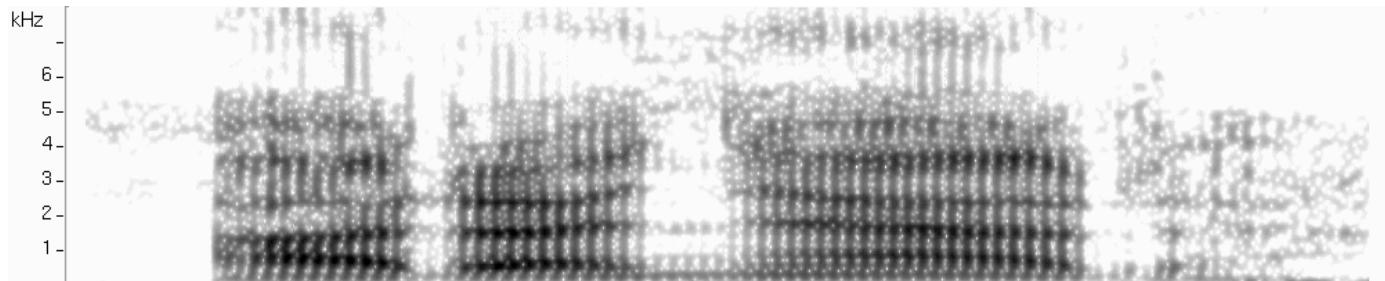


Pronunciación de “allá”.

Los dos primeros formantes que presenta son de frecuencia ligeramente inferior a los de /l/ y el tercero algo superior.

- Vibrantes.

- /r/ simple.



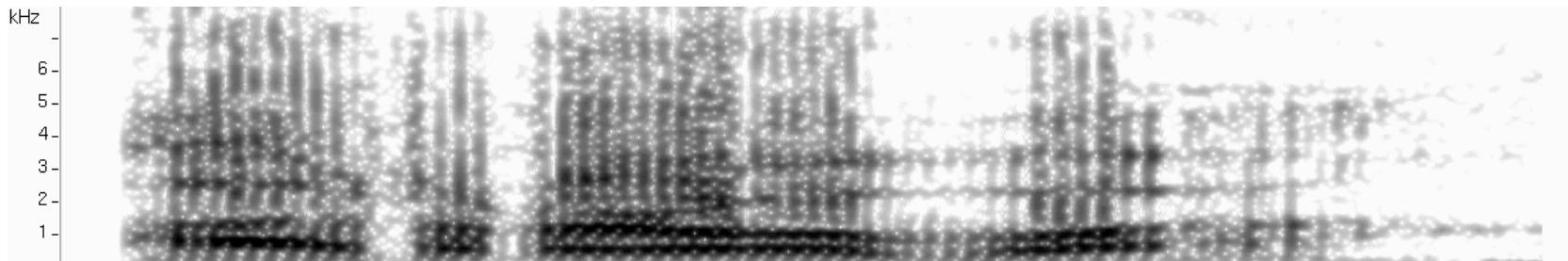
Pronunciación de “arañar”.

Interrupto muy breve.

Presenta una variante africada /ɾ/.

- /r̚/ múltiple.

Más larga que /r/. Una media de tres interrupciones.



Pronunciación de “arroba”.

Presenta una variante asibilada /r̚/ y una variante fricativa /ɹ/ (y otra faríngea en Puerto Rico).

El elemento esvarabático es la presencia de un elemento vocálico que hace que “prado”, por ejemplo, se realice como “parado”.

Otros fenómenos fonológicos

Omitimos de esta introducción el estudio de fenómenos como el acento prosódico y la entonación, aunque, naturalmente, influyen en la percepción.

La prosodia informa sobre:

- intención de la pronunciación (pregunta, orden, sentencia),
- énfasis (centra la atención en una parte específica),
- resolución de ambigüedades sintácticas/semánticas,
- estado anímico del hablante.

La información prosódica enriquece la pronunciación de las frases con:

- entonación,
- pausas,
- acento,
- ritmo.

Transcriptor ortográfico-fonético para el castellano

El programa `ort2fon.py` es una herramienta (script Python) que permite obtener la transcripción fonética de textos ortográficos en español.

La transcripción ortográfico-fonética es relativamente sencilla para el español.

Representación de los fonemas

Vocales			
	Anterior	Central	Posterior
Semiconsonantes	j		w
Cerrada	i I		u U
Media	e E		o O
Abierta		a A	

Nota: las versiones en mayúscula corresponden a vocales tónicas.

Consonantes

	Bilabial	Labio dental	Linguo-dental	Linguo-interdental	Linguo-alveolar	Linguo-palatal	Linguo-velar
	0 1	0 1	0 1	0 1	0 1	0 1	0 1
Oclusivas	p b		t d				k g
Fricativas		f		z	s	y	x
Africadas						c	
Nasales		m			n	h	
Laterales					l		H
Vibrante simple					r		
Vibrante múltiple					@		

Nota: 0 es "sorda" y 1 "sonora".

Existe, además, una marca para las pausas: un punto.

(Equivalencias con el alfabeto fonético: /θ/ ≡ z, /n/ ≡ h, /λ/ ≡ H, /r̄/ ≡ @, /j̄/ ≡ y.

Algunas características

- Utiliza reglas (extraídas del tratado de Quilis y Fernández) que indican cómo se deben transcribir los grafemas en unidades de tipo fonético atendiendo al contexto en que se presentan.
- Se ha incluido la opción de pronunciasniones múltiples (posibilidad de que un sonido no se pronuncie o se pronuncie de diferentes formas).
- La utilidad de este transcriptor ortográfico-fonético es, principalmente, el entrenamiento de sistemas de reconocimiento de voz: un sistema de reconocimiento de voz necesita la pronunciación de cada palabra que debe reconocer.
- Finalmente, se ha desarrollado una herramienta para crear modelos léxicos de un sistema de reconocimiento automático del habla.

Funcionamiento

1. Módulo de acentuación

Este módulo devuelve el texto acentuado a partir de la representación ortográfica.

La página del idioma Español <http://www.el-castellano.com/acentos.html>

Encuentra la sílaba tónica en cada palabra y sustituye su vocal tónica por su “versión mayúscula”.

Ejemplo: *barco* → bArko.

El proceso de acentuación se ha realizado como si fuéramos hablantes no nativos y tuviéramos que leer una palabra que desconocemos: sólo disponemos de información gráfica, es decir la presencia o ausencia de tilde.

→ Se aplica la inversa de las reglas de acentuación (suficientes para poder “leer” una palabra desconocida).

Ejemplo, regla para detectar las palabras agudas: “una palabra que acaba en consonante, excepto en *n* o *s* se acentúa en la última sílaba”.

2. Conversión grafema-fonema

- Muchas de las reglas son una correspondencia 1 a 1.

Ejemplo: “*b* siempre se pronuncia como *b*”, *barco* → bArko.

- Pero la mayoría son dependientes del contexto.

Ejemplo: “reglas asociadas a *c*”: *casa* → kAsa, *cero* → zEro, *chino* → cIno.

- Elimina la concurrencia de vocales.

Ejemplo: *la alameda* → lalamEda.

- Modos de transcripción intra-palabra (palabra-a-palabra) e inter-palabra (frase-a-frase).

Ejemplo: *costa azahar* → kOsta azAr o kOstazAr.

- Permite obtener múltiples pronunciaciones para un mismo texto:
 - Borrado de sonidos por pronunciación relajada (*abogado* → es abogAdo o abogAo).
 - Inserción de sonidos por pronunciación enfática (*psicólogo* → sikOlogo o psikOlogo).
 - Sustitución de fonemas por acentos regionales (*azul* → azUI o asUI).
 - Fenómenos de concurrencia (*ciudad de* → ziudAde o ziudAd.de).
- Representa variaciones de pronunciación y fonemas de pronunciación opcional:
 - elección ($\text{opt}_1 \mid \text{opt}_2 \mid \dots \mid \text{opt}_n$) y
 - opción [opt].
- Ejemplo: *abogado* → abogA[d]o, *calle* → ka(H|y)e.
- Puede etiquetar el tipo de fenómeno asociado a la salida múltiple.
Ejemplo: *azul* → a(z|<sso>s)UI, con la etiqueta <sso> para denotar el fenómeno de seseo.

Ejemplos de transcripción de palabras

Original	-p	-m	-mr
papel	papEl	papEl	papEl
barco	bArko	bArko	bA(r @)ko
casa	kAsa	kAsa	kAsa
cero	zEro	(z s)Ero	(z s)Ero
chino	clno	clno	clno
che	ce	ce	ce
abogado	abogAdo	abogA[d]o	abogA[d]o
psicólogo	sikOlogo	[p]sikOlogo	[p]sikOlogo
azul	azUl	a(z s)Ul	a(z s)Ul
calle	kAHe	kA(H y)e	kA(H y)e
cielo	zjElo	(z s)jElo	(z s)jElo
cerilla	zerlHa	(z s)erl(H y)a	(z s)erl(H y)a
Madrid	madrld	madrl[(d t z)]	madrl[(d t z)]
apto	Apto	Apto	A[(p b)]to
atlas	Atlas	Atlas	Atlas
acta	Akta	Akta	A[(k g)]ta
casa	kAsa	kAsa	kAsa
caza	kAza	kA(z s)a	kA(z s)a
Israel	is@aEl	is@aEl	i[s]@aEl

Ejemplos de transcripción de frases

Salida estándar python ort2fon.py

Insultad directamente, sin tapujos sucios ni mentiras.

insultAdirEktamEnte.sintapUxosUzjosnimentIras.

La abeja picó al abogado sin gran éxito pues es peor que un áspid.

labExapikOalabogAdosingranEksitopwesespeOrkeunAspid.

Ata la jaca a la reja.

AtalaxAkala@Exa.

Salida palabra a palabra python ort2fon.py -p

Insultad directamente, sin tapujos sucios ni mentiras.

insultAd dirEktamEntE . sin tapUxos sUzjos ni mentIras .

La abeja picó al abogado sin gran éxito pues es peor que un áspid.

la abExa pikO al abogAdo sin gran Eksito pwes es peOr ke un Aspid .

Ata la jaca a la reja.

Ata la xAka a la @Exa .

Salida con pronunciacin mltiples python ort2fon.py -m

Insultad directamente, sin tapujos sucios ni mentiras.

insultAd[.]d]irEktamEnte[.]sin[.]tapUxos[.]U(z|s)jos[.]ni[.] mentIras[.]

La abeja picó al abogado sin gran éxito pues es peor que un áspid.

I[a.]abExa[.]pikO[.]al[.]abogA[d]o[.]sin[.]gran[.]Eksito[.]pwes[.]es[.]peOr[.]ke[.]un[.]Aspid[.]

Ata la jaca a la reja.

Ata[.]la[.]xAk[a.]a[.]la[.]@Exa[.]

-mr

Insultad directamente, sin tapujos sucios ni mentiras.

insultAd[.]d]irE[(k|g)]tamEnte[.]sin[.]tapUxos[.]U(z|s)jos[.]ni[.] mentIras[.]

La abeja picó al abogado sin gran éxito pues es peor que un áspid.

I[a.]abExa[.]pikO[.]al[.]abogA[d]o[.]sin[.]gran[.]E[(k|g)]sito[.]pwes[.]es[.]peOr[.]ke[.]un[.]Aspid[.]

Ata la jaca a la reja.

Ata[.]la[.]xAk[a.]a[.]la[.]@Exa[.]

Salida con pronunciaciões múltiples, relajación y etiquetas en las variaciones

python ort2fon.py -mre

Insultad directamente, sin tapujos sucios ni mentiras.

insultAd[.]d]irE[(k|g)]tamEnte[<pe>.]sin[.]tapUxos[.]s]U (z|<sso>s)jos[.]ni[.]mentIras[<pe>.]

La abeja picó al abogado sin gran éxito pues es peor que un áspid.

I[a.]abExa[.]pikO[.]al[.]abogA[<mr>d]o[.]sin[.]gran[.]E [(k|g)]sito[.]pwes[.]es[.]peOr[.]ke[.]un[.]As

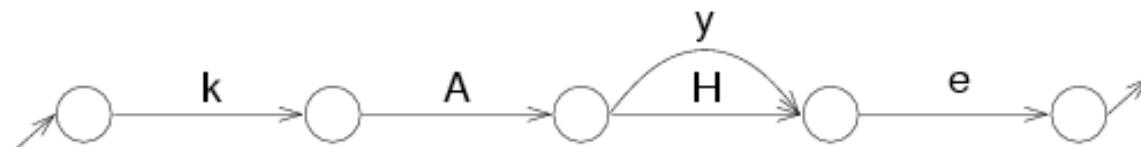
Ata la jaca a la reja.

Ata[.]la[.]xAk[a.]a[.]la[.]@Exa[<pe>.]

fon2lex: Conversión transcripción-autómata

Para generar un autómata a partir de la salida del transcriptor ortográfico-fonético se ha implementado un traductor (fon2lex), también programado en Python.

Esta herramienta facilita el uso de la construcción del léxico de una cierta tarea en sistemas de reconocimiento automático del habla.



Ejemplo de autómata generado por fon2lex a partir de la salida múltiple $kA(H|y)e$. La palabra calle se puede pronunciar como kAHe o kAye.

Trabajos relacionados

Aproximaciones a la conversión de grafemas en fonemas:

- los sistemas basados en reglas y
- los métodos inductivos, que intentan aprender automáticamente las reglas fonológicas a partir de ejemplos.

Para el castellano.....

Aunque el castellano es un idioma relativamente transparente en su relación ortográfica-fonética, se necesita algún tipo de conversión tanto en los sistemas de síntesis de voz como en sistemas de reconocimiento.

En la tesis de Ríos [Ríos99, capítulo 2] se realiza una revisión de los trabajos realizados para el castellano.

Para otras lenguas.....

Más atención en otras lenguas (inglés).

Los sistemas de síntesis utilizan normalmente transcripciones basadas en reglas y un diccionario de excepciones (MITalk), aunque también se han propuesto aproximaciones inductivas (NETtalk).

Bibliografía

- Antonio Quilis: *Fonética acústica de la lengua española*. Biblioteca Románica Hispánica. Editorial Gredos. Madrid. 1981.
- Antonio Quilis, José A. Fernández: *Curso de fonética y fonología españolas para estudiantes angloamericanos*. CSIC, Instituto Miguel de Cervantes. 1979.
- Lawrence Rabiner, Biing-Hwang Juang: *Fundamentals of speech recognition*. Prentice Hall. 1993.
- Francisco Casacuberta, Enrique Vidal: *Reconocimiento automático del habla*. Marcombo. 1987.
- María José Castro, Salvador España, Andrés Marzal, Ismael Salvador. Transcriptor ortográfico-fonético para el castellano. *Procesamiento del Lenguaje Natural*, 27:241–245, septiembre 2001.
- Antonio Ríos Mestre. La transcripción fonética automática del diccionario electrónico de formas simples flexivas del español: estudio fonológico en el léxico. *Estudios de Lingüística Española*, 4, 1999.