Capítulo 2

Bases fisiológicas de la comunicación

"Pero ellos no entendían nada de esto, eran cosas ininteligibles para ellos, no entendían lo que les decía."

(Lucas 18,34)

Contenido

2.1. Introducción

LA comunicación verbal, tanto escrita como oral, diferencia claramente al hombre del resto de las criaturas. El habla constituye además nuestra forma de comunicación más importante. Las sorprendentes características de este sistema natural no han podido aún ser emuladas por medios artificiales. A los efectos de encontrar una representación de la señal óptima para el diseño de nuevos dispositivos tecnológicos se debe comprender la naturaleza del habla y su forma de producción. Así mismo, es necesario interpretar

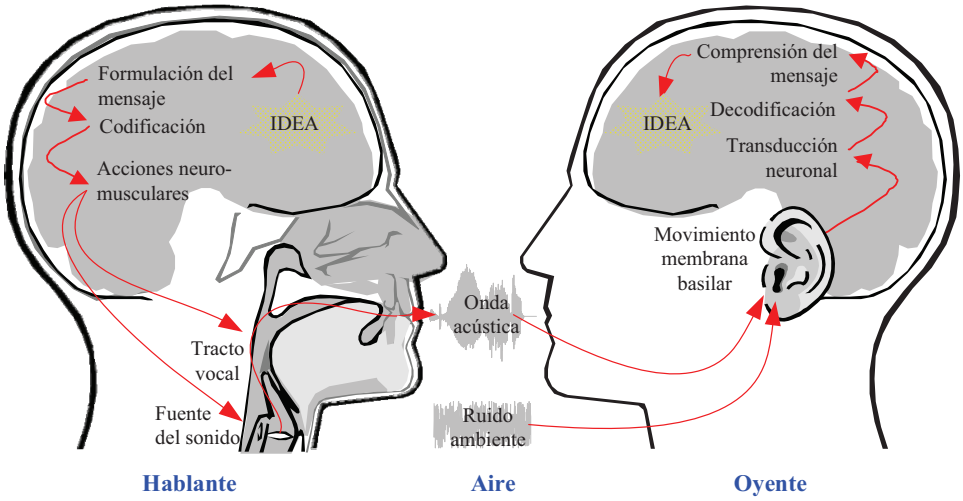

Figura 2.1: *Diagrama simplificado del proceso de comunicación oral de un mensaje en el hombre. Se resaltan solo las etapas y órganos intervinientes más importantes del proceso, en un único sentido.*

los aspectos fundamentales del procesamiento llevado a cabo por el sistema auditivo que permiten extraer las características significativas de la señal de voz. Es posible entonces discernir cuales son los parámetros relevantes que deberían preservarse en esta representación y bajo que principios correspondería codificar los mismos. Por todo esto se requiere el estudio de los fundamentos anatómicos y fisiológicos involucrados en el proceso de comunicación oral humana.

En este punto surge claramente la cuestión acerca de cuanto debe acercarse un mecanismo diseñado por el hombre a este proceso natural, para intentar resolver el problema planteado en este trabajo. Se puede decir que el criterio aquí será acercarse lo necesario como para capturar en los dispositivos artificiales aquellos aspectos esenciales que permitan asegurar algunas capacidades de utilidad práctica. Entre estas capacidades deseables es posible mencionar el lograr la independencia de su desempeño bajo diferentes condiciones como ser cambios en el volumen y la velocidad de pronunciación, en la identidad del hablante (por cambios de rasgos particulares o cuestiones regionales), o en las interferencias del ambiente acústico circundante.

Se denomina comunicación al proceso de transmisión y recepción de información. En el hombre el habla es utilizada para transmitir información de un hablante a un oyente. En la Figura 2.1 se aprecia una diagrama simplificado del proceso de comunicación oral humano. Se puede resumir este proceso de la siguiente forma. El mismo comienza con una idea o pensamiento que el hablante desea transmitir al oyente [32]. El hablante traduce este pensamiento a través de una serie de procesos neurológicos y movimientos musculares

para producir una onda de presión sonora. Esta señal es recibida por el sistema auditivo del oyente, procesada, y convertida nuevamente en una señal neurológica. A partir de ello el oyente forma una idea del mensaje recibido.

Esta explicación resumida esconde algunos aspectos importantes. Para realizar su tarea el hablante convierte la idea a transmitir en una estructura lingüística. Esto se realiza mediante la selección de las palabras y el orden de las mismas que mejor representen la idea, basada en reglas asociadas con el lenguaje en particular. Se agregan también algunas características adicionales como por ejemplo la entonación. En estas primeras etapas se incluye una redundancia importante en el sentido explicado anteriormente. A continuación, el cerebro produce una serie de comandos motores que mueven diversos músculos del sistema vocal para producir la onda de presión sonora deseada. Esta onda acústica es recibida por el sistema auditivo del hablante y convertida nuevamente en una secuencia de pulsos neurológicos. Esto produce la realimentación necesaria para controlar su propia producción de voz. El proceso de percepción en el oyente comienza cuando recibe la onda de presión sonora en el oído externo y la convierte en impulsos neurológicos al pasar por el oído medio e interno. Finalmente interpreta estos pulsos en la corteza auditiva del cerebro para determinar cuál fue el mensaje (lo que implica también la comprensión del significado del mensaje).

Todo este complejo proceso tiene sus bases en los órganos del aparato fonador, el sistema auditivo, y el procesamiento realizado a nivel cerebral en ambos sentidos, requiriendo también para su comprensión una perspectiva lingüística. El aparato fonador y el sistema auditivo no pueden tratarse tampoco de manera aislada. Según Greenberg [52] el aparato vocal humano está probablemente optimizado para producir la comunicación de señales, con propiedades que aprovechan la habilidad del sistema auditivo de codificar la información de una manera robusta, o tolerante a fallas. El espectro del habla está sesgado hacia las bajas frecuencias, que son particularmente resistentes a alteraciones debidas al ruido de fondo. El nivel de presión sonoro de la mayor parte del habla es suficientemente alto como para asegurar que esa información espectral de baja frecuencia se extienda por una amplia serie de canales de frecuencia auditiva. La periodicidad glótica asegura que el sistema pueda seguir o rescatar el habla en condiciones de ruido, acústicamente adversas, y la modulación de la longitud de las sílabas ayuda al cerebro a juntar entidades espectrales dispares en unidades más significativas. Dentro de este marco, la importancia del sistema auditivo para el discurso, está en que precondiciona la representación nerviosa para maximizar la fiabilidad y la tasa de transmisión de información. El cerebro por consiguiente necesita sólo seguir el rastro de estas características

en la señal, "confiando" en que son sólo estos rasgos los que codifican la información importante.

Durante el desarrollo de este capítulo se explicarán con mayor detalle todos estos mecanismos para poder dilucidar aquellos aspectos que se deberían preservar en la representación de la señal de voz. El enfoque pretende ser integrador, incluyendo esquemas y diagramas que faciliten la comprensión de las funciones y su relación con las estructuras anatómicas involucradas. Para un estudio más detallado el lector se deberá remitir a la extensa bibliografía específica disponible para cada área (por ejemplo [22, 127, 108, 79]).

Este capítulo se organizará siguiendo un orden similar al de la exposición anterior acerca del proceso de comunicación oral humana. Los aspectos funcionales del proceso son relativamente independientes del idioma considerado, aunque este análisis se limitara al idioma español (principalmente en su versión *argentina rioplatense* [108]). En primer lugar se describirán el mecanismo de producción del habla y los órganos involucrados. Esto incluye la descripción de los principales tipos de sonidos o fonemas que es posible generar mediante el aparato fonador. Luego se presentarán aspectos relacionados con la señal de voz propiamente dicha mostrando algunos ejemplos típicos. Posteriormente se esbozarán los principios y elementos que intervienen en la percepción de los sonidos del habla y la audición. Se enfatizarán aquí los fundamentos de la codificación de la señal de voz a nivel neurosensorial por considerarse de importancia para los objetivos planteados.

2.2. Mecanismo de producción del habla

Para comenzar se esbozarán brevemente los mecanismos involucrados en la producción del habla. Como se mencionó en la sección anterior el proceso de comunicación comienza en el hablante con la traducción de una idea a patrones de variación de la presión sonora en la señal de voz. Para ello el primer paso se realiza principalmente en la corteza cerebral involucrando varias áreas de manera simultánea o alternada. Este proceso es bastante complejo ya que el cerebro debe enviar las ordenes adecuadas al aparato fonador para codificar la información acústica a transmitir por medio de una serie de reglas lingüísticas a diferentes niveles[1]. Cada uno de estos niveles impone ciertas restricciones y "estructura" que forman parte del "código" compartido entre el hablante y el oyente [36, 127, 99] :

Fonológico: se encarga de la representación o modelado de las características físicas de

[1]Existe información que se codifica simultáneamente en varios niveles para proveer la necesaria redundancia para aumentar la robustez de la comunicación.

los sonidos utilizados para la producción del habla (fonemas). *No todos los sonidos posibles de generar constituyen fonemas.*

Fonético: se ocupa de la descripción de las variaciones en la pronunciación de los fonemas que aparecen dentro de una palabra o cuando las palabras son dichas juntas en una frase (coarticulación, fusión de sílabas, etc.). *La realización particular de un fonema depende principalmente de su contexto.*

Morfológico: realiza una descripción del modo en que los morfemas (unidades de significación) son combinados para formar palabras. (formación de plurales, conjugación de verbos, etc.). *No todas las combinaciones de morfemas son admitidas.*

Léxico: se ocupa de definir las palabras válidas y el sentido que estas poseen. *No todos las combinaciones de fonemas constituyen palabras permitidas.*

Sintáctico: consiste en las reglas de formación de frases, dando lugar a una limitación del numero de frases. *No todas las combinaciones de palabras son frases autorizadas.*

Prosódico: consiste en una descripción de la fluctuación en la acentuación y entonación durante el transcurso de una frase. *No se admite cualquier patrón de fluctuación.*

Semántico: se ocupa del significado de las palabras y las frases que puede ser visto también como una restricción sobre el alcance del mensaje. *No todas las frases gramaticalmente válidas tienen significado.*

Pragmático: se ocupa de las reglas de conversación. *La respuesta de un interlocutor no debe ser solamente una frase con significado sino también una respuesta razonable acerca de lo que se esta diciendo.*

En la mayoría de las personas las funciones más importantes asociadas con el lenguaje se localizan en el hemisferio izquierdo. A pesar de este predominio del lado izquierdo, el contenido emocional del lenguaje está gobernado principalmente por el hemisferio derecho. En la Figura 2.2 se puede apreciar un diagrama de las diferentes partes funcionales de la corteza relacionadas con la producción y la comprensión del habla. Dos áreas conocidas como el área de Wernicke y el área de Broca son las más importantes y están involucradas en el almacenamiento de información relacionada con el habla [124]. Ambas áreas se comunican mediante una vía bidireccional denominada *fascículo arqueado*. El área de Wernicke guarda información necesaria para colocar las palabras de un vocabulario previamente aprendido en forma de una conversación con sentido. El área de

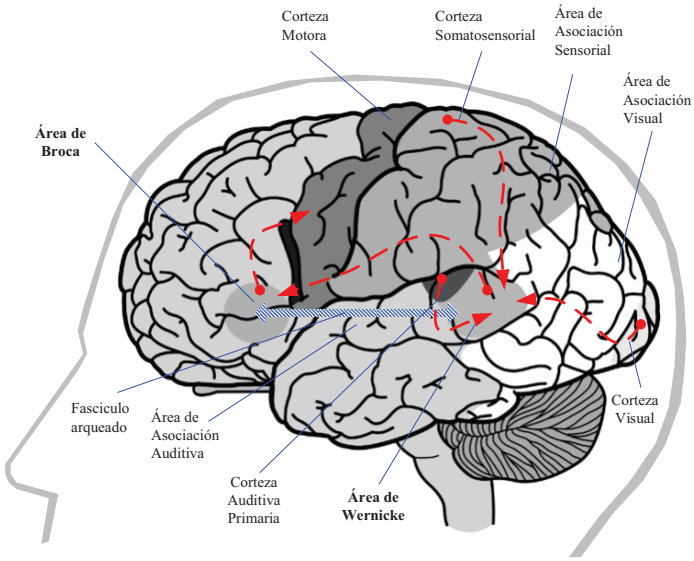

Figura 2.2: *Diagrama de las principales áreas cerebrales implicadas en la producción y comprensión del habla. Las cortezas sensorial, auditiva, visual y motora primarias muestran la relación del las áreas del lenguaje de Broca y de Wernicke con las áreas menos especializadas que, no obstante están incluidas en el proceso.*

Broca almacena información necesaria para la producción del habla. Esta última es precisamente la responsable de la programación de la corteza motora para mover la lengua, los labios y los músculos del aparato fonador para articular las diferentes palabras. A continuación la corteza ejecuta este programa que permite coordinar adecuadamente los distintos órganos y partes del aparato fonador para producir la señal sonora requerida. La percepción de su propia voz, en conjunto con la del ruido ambiente, le permite al hablante un continuo monitoreo y control de su fonación. Los cambios producidos en la misma debido a la presencia de ruido se denominan *efecto Lombard* [100, 88, 77] y tienen por objeto minimizar los efectos del ruido. Todo esto conlleva también la necesaria activación de la corteza auditiva en el proceso de producción del habla.

Se debe aclarar que en una conversación normal, además de la comunicación por medio del habla, se utilizan otros medios de transmisión de información no verbales. Un ejemplo de ello son los gestos. Sin embargo estos medios alternativos no se incluirán en este desarrollo. Para la percepción de la señal de voz en condiciones adversas otra información visual como la del movimiento de los labios puede mejorar la inteligibilidad. Este aspecto es procesado en zonas de integración sensorial de la corteza y tampoco será analizado en este trabajo.

2.2.1. Aparato fonador

La forma en la que los cambios en la configuración del aparato fonador modifican las características de la señal acústica serán examinados a continuación. En la Figura 2.3 se observa un esquema simplificado del aparato fonador en conjunto con una sección sagital del mismo (que no incluye a los pulmones). La zona comprendida entre la laringe (glotis) y los labios constituye el *tracto vocal* propiamente dicho. Este está formado por las cavidades supraglóticas, faríngeas, oral y nasal. El aparato fonador se puede considerar como un sistema que transforma energía muscular en energía acústica. La teoría acústica de producción del habla describe este proceso como la respuesta de un sistema de filtros a una o más fuentes de sonidos. En la representación simbólica, y suponiendo linealidad, si $H(f)$ es la función de transferencia del filtro que representa el tracto vocal en un instante dado y $X(f)$ la fuente de excitación, el producto $Y(f) = H(f).X(f)$ representa el sonido resultante. La fuente $X(f)$ indica la perturbación acústica de la corriente de aire proveniente de los pulmones. A veces suele agregarse a este modelo la función transferencia $L(f)$ del fenómeno de radiación a la salida de los labios. Es decir que los sonidos del habla son el resultado de la excitación acústica del tracto vocal, el cual varía constantemente sus características. En este proceso los órganos fonatorios desarrollan distintos tipos de actividades, tales como movimientos de pistón que inician una corriente de aire, movimientos o posiciones de válvula que regulan el flujo de aire, y al hacerlo generan sonidos o en algunos casos simplemente modulan las ondas generadas por otros movimientos.

Para comprender la forma en la que el tracto vocal varía sus características muchas veces se utiliza un modelo sencillo de dos tubos uniformes sin pérdida que varían su ancho o su longitud. Esto permite explicar no solo las diferencias entre los sonidos producidos por un mismo hablante, sino también las existentes entre los sonidos de diferentes hablantes, debido a sus diferencias anatómicas.

El sistema respiratorio constituye la principal fuente de energía para producir sonidos en el aparato fonador humano. La energía es proporcionada en forma de flujo o corriente de aire y presiones que, a partir de las distintas perturbaciones, generan los diferentes sonidos. De esta forma se pueden identificar tres mecanismos generales en la excitación del tracto vocal:

1. Las cuerdas vocales modulan un flujo de aire que proviene de los pulmones dando como resultado la generación de pulsos cuasiperiódicos.

2. Al pasar el flujo de aire proveniente de los pulmones por una constricción en el

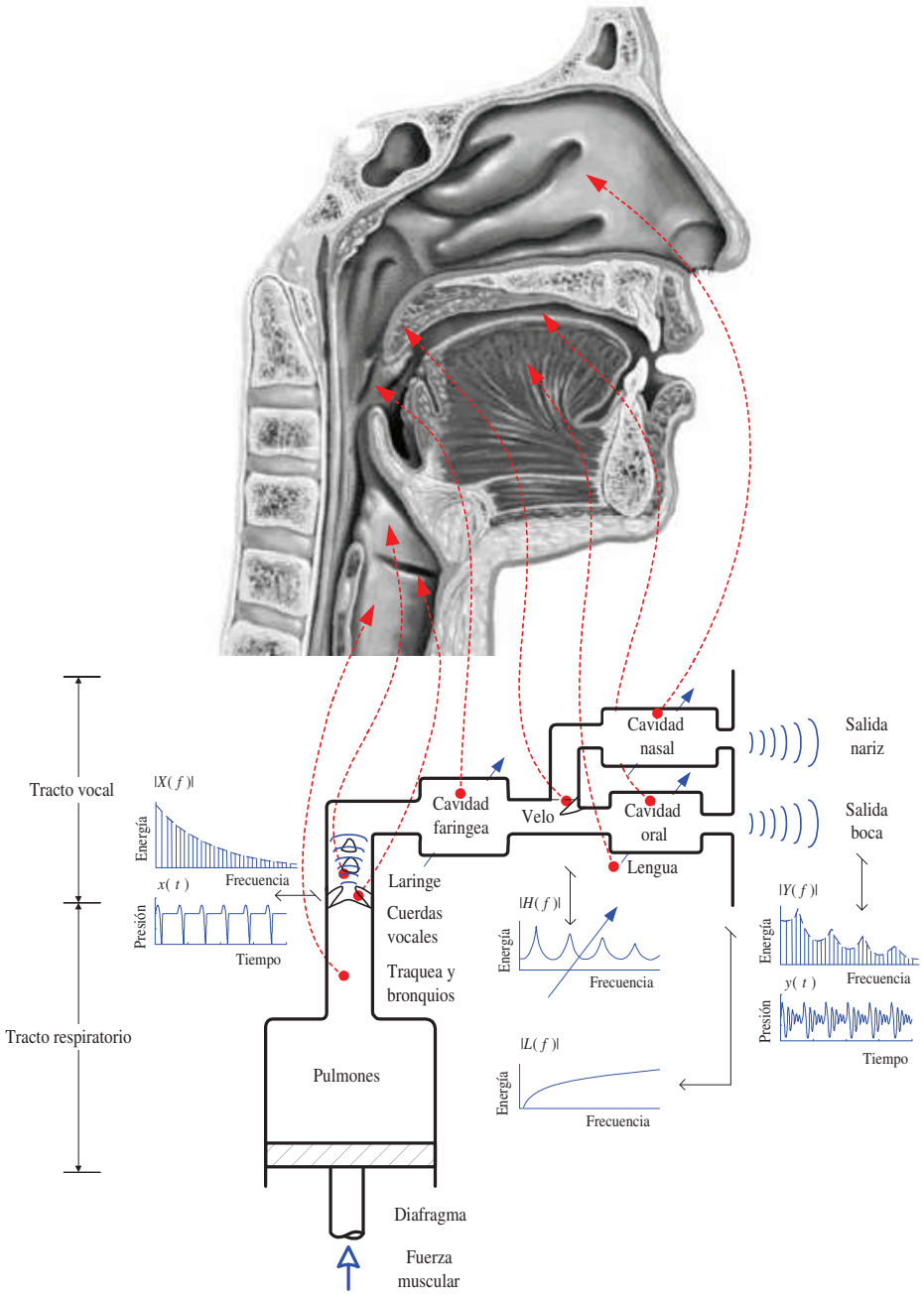

Figura 2.3: *Corte sagital anatómico del aparato fonador (arriba) y diagrama esquemático del mismo que ilustra su funcionamiento (abajo). En el diagrama se ejemplifican las señales temporales, sus correspondientes espectros y sus funciones de transferencia espectrales, para el caso de producción de un fonema sonoro. La suposición subyacente es que se trata de un sistema lineal.*

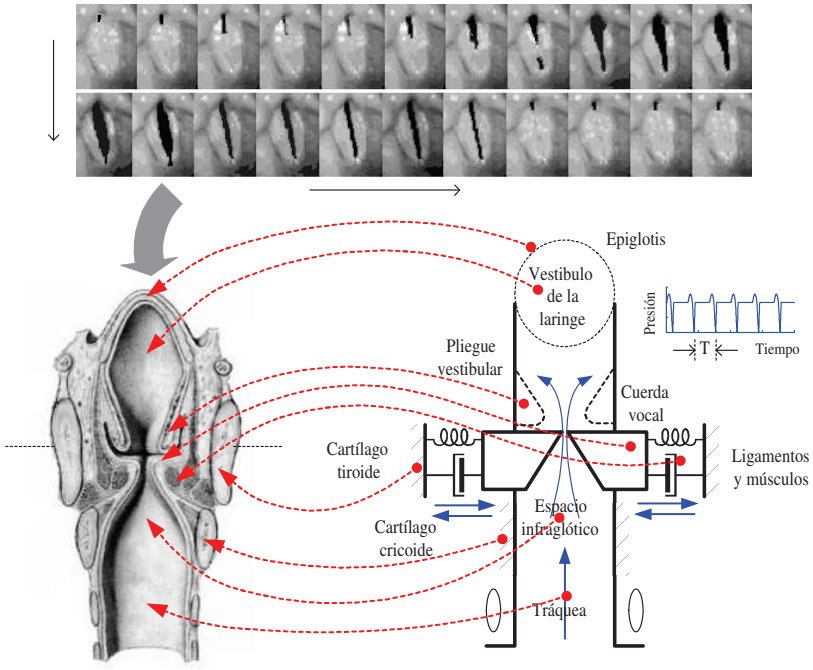

Figura 2.4: *Corte longitudinal de la laringe (abajo izquierda) junto con el diagrama funcional correspondiente (abajo derecha). Se muestra también el aspecto de la glotis en diferentes instantes de la vibración de las cuerdas vocales (arriba). Durante esta secuencia de apertura y cierre de las cuerdas vocales se producen variaciones bruscas en la presión sonora a la salida de las mismas, lo que puede representarse a partir de una señal periódica de período T.*

tracto vocal se presenta la generación de ruido de banda ancha.

3. El flujo de aire produce una presión en un punto de oclusión total en el tracto vocal; la rápida liberación de esta presión, por la apertura de la constricción, causa una excitación de tipo plosivo, intrínsecamente transitoria.

El aparato respiratorio actúa también en la regulación de parámetros tan importantes como la energía (intensidad), la frecuencia fundamental de la fuente cuasiperiódica, el énfasis y la división del habla en varias unidades (sílabas, palabras, frases).

La laringe juega un papel fundamental en el proceso de producción del habla. En la Figura 2.4 se aprecia un corte longitudinal de la misma junto con un diagrama funcional. La función fonatoria de la laringe se realiza mediante un mecanismo en el que intervienen las cuerdas vocales, los cartílagos en los que se insertan y los músculos laríngeos intrínsecos, y que depende también de las características del flujo de aire proveniente de los pulmones. La forma de onda de los pulsos generados puede representarse en forma simplificada como una onda triangular. En el hombre, la frecuencia de esta onda de vibración de las cuerdas vocales varía entre 100 y 170 Hz, en las mujeres entre 180 y 280 Hz y en los niños puede superar los 300 Hz. Los valores de esta vibración glótica (o fre-

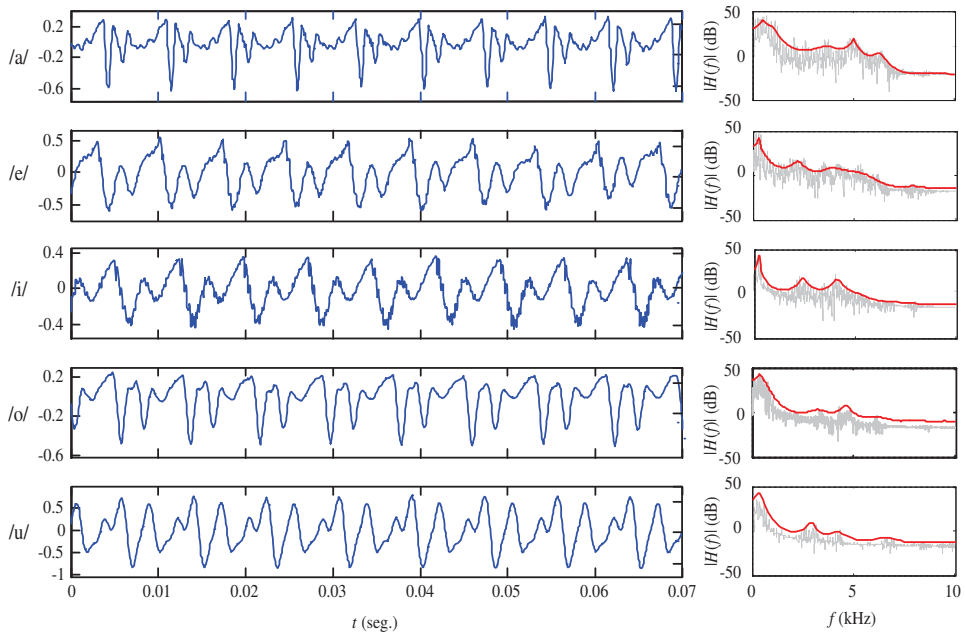

Figura 2.5: *Ejemplos de sonogramas (izquierda) y espectros (derecha) de las vocales del español pronunciadas en forma sostenida y aislada por un hablante masculino nativo. A pesar de la similitud de algunas de sus formas de onda temporales es posible discriminarlas a partir de las resonancias o picos espectrales.*

cuencia glótica) se modifican en forma voluntaria y son los responsables de la frecuencia fundamental (denominada F_0) producida al hablar (ver Figura 2.7 más adelante).

El tracto vocal puede mantener una configuración relativamente abierta y actuar sólo como modulador del tono glótico o estrechar o cerrar el paso de la corriente de aire en una zona específica. El tracto actúa como filtro acústico, principalmente en los sonidos con componente glótica, pudiendo modificar sus parámetros en forma continua. Si se observan los espectros de los sonidos vocálicos, éstos proporcionan información sobre todos los aspectos relevantes de la configuración del tracto en ese instante. Es decir, todas las resonancias del tracto, resultantes de su configuración, pueden observarse directamente en el espectro del sonido vocálico. En la Figura 2.5 pueden observarse los sonogramas[2] de las cinco vocales del español junto con sus respectivas envolventes espectrales donde se pueden apreciar claramente estas resonancias a través de los picos espectrales.

2.2.2. Sonidos y fonemas

Como se ha mencionado las unidades lingüísticas básicas del habla son los fonemas. En realidad los fonemas son modelos de los sonidos que pueden diferir luego en su expre-

[2]En este trabajo se denominará "sonograma" a las gráficas de variación de la presión sonora en función del tiempo.

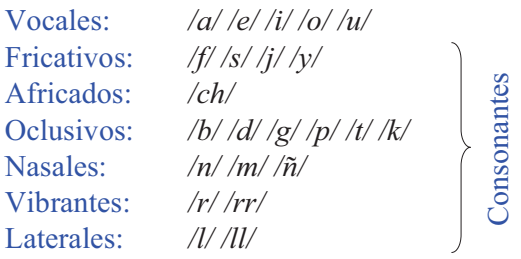

Figura 2.6: *Cuadro simplificado de clasificación de los fonemas del español* rioplatense. *De acuerdo con las características acústicas y los gestos articulatorios que dan lugar a cada tipo de sonido la principal división se da entre las vocales y las consonantes.*

sión acústica[3]. Se los puede definir como el conjunto mínimo de unidades que permite decir cualquier palabra en un idioma determinado. Dos fonemas son distintos si el cambio de uno por otro cambia la palabra (por ejemplo *boda* vs. *moda*). En la Figura 2.6 puede apreciarse un cuadro que muestra los fonemas de uso corriente en nuestro idioma[4].

Se consideraran ahora las configuraciones del tracto que corresponden a cada fonema ya que –como se dijo antes– toda configuración presenta características propias de resonancia que, junto con la fuente de excitación actuante, dan al sonido su peculiar cualidad fonética. Por ello los fonemas se agrupan en vocálicos y consonánticos. Esta división se sustenta tanto en las características acústicas como en los gestos articulatorios que dan lugar a cada tipo de sonido. La duración temporal de los fonemas no es uniforme. Para dar una idea general se puede decir que las vocales son más largas (en el orden de los 100 mseg promedio) que las consonantes (en el orden de los 20 mseg promedio).

Vocales

En la articulación de vocales y sonidos tipo vocálicos, el tracto presenta una configuración relativamente abierta y la fuente de excitación es siempre glótica. Las propiedades de estos sonidos persisten por un tiempo apreciable o cambian muy lentamente mientras se mantenga la configuración del tracto.

Los pulsos glóticos estimulan el tracto vocal que actúa como sistema resonador. Este

[3]Se denominan alófonos a las diferentes realizaciones de un mismo fonema. También se utiliza el término *fono* como sinónimo de alófono.

[4]Existen alfabetos fonéticos para aplicaciones tecnológicas con adaptaciones particulares para el español rioplatense [56], tales como:

- SAMPA: `http://www.phon.ucl.ac.uk/home/sampa/spanish.htm`
- Worldbet: `http://www.ling.gu.se/~jimh/courses/ipa.ps`

Sin embargo, por razones de sencillez y salvo que se indique lo contrario, para hacer referencia a los fonemas se utilizará la grafía más cercana (a su pronunciación) encerrada entre /•/.

puede modificar su configuración y con ello sus frecuencias de resonancia como una especie de filtro acústico adaptativo. Esta posibilidad de variación es la que permite al hablante producir muchos sonidos diferentes. La forma del tracto en la producción de las vocales esta controlada principalmente por la posición de la lengua, de la mandíbula y de los labios. Los sonidos vocálicos se pueden clasificar por sus distintas características acústicas[5] [99]:

Zonas de estrechamiento: Por estudios sistemáticos de radiografías de articulaciones vocálicas se han localizado tres zonas principales de producción de la constricción. Esto depende de la posición de la lengua, los labios, y la boca. De esta manera los sonidos vocálicos se agrupan en *anteriores* (*/i/*, */e/*), *medios* (*/a/*), y *posteriores* (*/o/*, */u/*) según la posición de la constricción.

Abertura de la boca: Esta abertura cuya configuración y grado están determinadas por la acción de los labios y del maxilar inferior, da lugar a importantes diferenciaciones acústicas y fonéticas. Así se tienen en forma relativa a las vocales *abiertas* (*/a/*), *medias* (*/e/*, */o/*) y *cerradas* (*/i/*, */u/*).

Grado de estrechamiento: De esta manera se describen los sonidos vocálicos según el grado de estrechamiento en la región de menor área o constricción máxima, en *estrechos* (*/i/*, */u/*, */o/*) y *amplios* (*/e/*, */a/*).

Longitud del tracto: La longitud del tracto se modifica redondeando los labios, subiendo y bajando la posición de la laringe. Así se tienen las vocales *labializadas* (*/o/*, */u/*) y *deliabializadas* (*/a/*).

Consonantes

Los sonidos consonánticos se producen con una configuración relativamente cerrada del tracto vocal. El cierre o estrechamiento del canal se realiza en zonas especificas del tracto vocal por acción de partes especificas de las estructuras articulatorias. Entre los factores que determinan la cualidad del sonido resultante, se deben distinguir aquellos que hacen al modo de articulación (cierre o estrechamiento) de los que señalan la zona o lugar de articulación (lugar donde se produce cierre o estrechamiento). La participación de la fuente glótica, la naturaleza del cierre o estrechamiento y la transmisión a través de la cavidad oral y/o nasal, constituyen los principales factores del modo de articulación.

[5]En el español las dos primeras características son las más importantes para diferenciar entre las vocales.

Las consonantes, por otro lado, pueden ser agrupadas en los siguientes tipos articulatorios:

Fricativas: se caracterizan por ser ruidos aleatorios generados por la turbulencia que produce el flujo de aire al pasar por un estrechamiento del tracto. Pueden ser sonoros como */y/* si hay componente glótica o sordos como */f/*, */s/* o */j/* (también */z/* en otras versiones del español) si no la hay.

Africadas: si los fonemas comienzan como oclusivos y la liberación del aire es fricativa se denominan africados. Por ejemplo la */ch/*.

Oclusivas: se producen por el cierre momentáneo total o parcial del tracto vocal seguido de una liberación más o menos abrupta del aire retenido. Por ejemplo las totales */p/*, */t/*, */k/* o las parciales */b/*, */d/*, */g/*. Estas últimas son sonoras.

Nasales: son producidas a partir de excitación glótica combinada con la constricción del tracto vocal en algún punto del mismo. Por ejemplo */m/*, */n/* o */ñ/*.

Vibrantes: éstas son producidos al pasar el aire por la punta de la lengua y producir su vibración. Tienen componente glótica. Por ejemplo */r/* y */rr/*.

Laterales: estas se producen cuando se hace pasar la señal sonora glótica por los costados de la lengua. Por ejemplo */l/* y */ll/*.

Semivocales: están formadas por la unión de dos de los anteriores hasta el punto de convertirse en otro sonido (por ejemplo dos vocales). Algunos consideran en este grupo a las vibrantes (*/rr/*) y las laterales (*/ll/*).

2.2.3. Segmentos, suprasegmentos y sílabas

De lo dicho anteriormente, se podría inferir que el habla es, de alguna manera, un fenómeno secuencial "discreto", es decir una sucesión de fonemas. De hecho, como se verá más adelante, es posible asignar *etiquetas* a los diferentes trozos de señal asociados con estos fonemas. Sin embargo si se observa la señal de la voz, la representación acústica de una frase, se verán muy pocas pausas o intervalos entre los sonidos. De esta forma el habla constituye un continuo acústico, producido por un movimiento ininterrumpido de los órganos del aparato fonador. A pesar de la naturaleza continua de la voz los oyentes pueden segmentarla en sonidos.

Aquellas características de la voz de una escala temporal superior al fonema se denominan suprasegmentales. Estas características están determinadas principalmente por la entonación, la cual determina la prosodia. Las variables que intervienen en la entonación son las variaciones de frecuencia fundamental o F_0, la duración y variaciones de energía y sonoridad.

La prosodia en las uniones puede ser caracterizada por silencios, duración en las vocales, o por formas como puede ser la presencia de sonoridad o aspiración. Por ejemplo en la frase "perdonar, no matar" existe una pausa después de "perdonar" pero si la coma cambia de lugar "perdonar no, matar" el silencio se produce después de "no" cambiando totalmente el significado del mensaje.

La sílaba constituye una unidad lingüística de escala temporal mayor que la del fonema. Si bien para una lengua la cantidad de sílabas es muy superior a la de fonemas, en general la variabilidad acústica de estas unidades es también mucho menor. Por ello algunos investigadores prefieren su utilización como unidad de modelado del habla.

2.3. Señal de voz

Hasta ahora se han descripto los distintos tipos de fonemas y la forma en la que se originan en el aparato fonador. Sin embargo se han hecho pocas referencias a los aspectos relacionados con la señal de voz propiamente dicha, que constituye el substrato del que se obtendrá una representación adecuada. Los aspectos discutidos en la presente sección están más relacionados con la fonética acústica que con la fonología.

Se comenzará por analizar las vocales, por constituir el caso más sencillo. En la Figura 2.5 pueden observarse el sonograma de las vocales del español pronunciadas en forma sostenida y aislada junto con sus respectivos espectros. En este caso se aprecia un cierto parecido entre */o/* y */u/* o entre */e/* y */i/*, lo cual es de suponer porque se puede decir que son vocales 'cercanas' según se verá a continuación. Como ya se mencionó en los espectros de los sonidos vocálicos pueden observarse todas las resonancias del tracto. Estas resonancias aparecen como picos en el espectro y se denominan formantes. Las formantes se numeran a partir del 1. Las formantes, principalmente F_1 y F_2, constituyen un medio para caracterizar a las vocales. De hecho, la presencia de formantes, y en particular de F_0 evidencia si se trata de un trozo sonoro o sordo (con o sin componente glótica). A pesar de la notación F_0 no constituye estrictamente una formante sino, como ya se indicó, la frecuencia fundamental que está directamente relacionada con la ento-

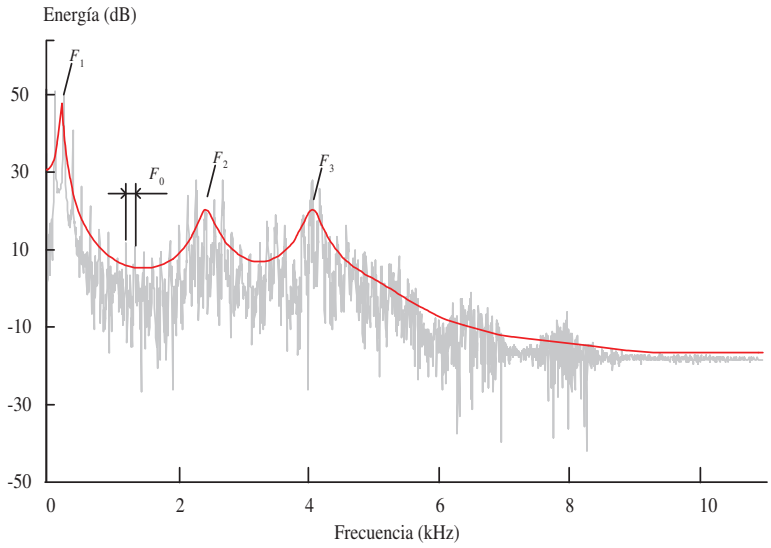

Figura 2.7: *Espectro de una vocal /i/ pronunciada en forma sostenida y su envolvente, donde se resaltan las frecuencias formantes (F_1, F_2 F_3 y la frecuencia fundamental F_0). F_0 corresponde a la frecuencia glótica y es uno de las componentes de la entonación del habla, mientras que el resto constituyen las formantes que permiten discriminar entre las vocales. Su variación temporal permite también diferenciar entre los diferentes fonemas sonoros.*

nación de una frase o emisión[6]. En la Figura 2.7 aparece el espectro de una */i/* y su correspondiente envolvente espectral (estimada mediante un modelo autoregresivo) donde se aprecian claramente los picos y se muestran las distintas formantes. En la Figura 2.8 se puede apreciar un gráfico de la distribución de las vocales del español –o mapa de formantes– para hablantes masculinos en función de F_1 y F_2. Se puede observar que mediante estas características es posible separar o modelar fácilmente a las diferentes vocales. En el gráfico se muestra también la relación del valor de las formantes con los atributos articulatorios discutidos en la Sección 2.2.2 y el denominado *triángulo de las vocales*. Las formantes de esta figura han sido obtenidas de vocales aisladas pronunciadas en forma sostenida. En el caso del discurso continuo las formantes siguen siendo un rasgo distintivo importante para las vocales. Sin embargo en este caso es preciso seguir también la evolución de los patrones formánticos debido a que las clases no se encuentran tan bien separadas [64]. Este fenómeno está relacionado con el hecho, explicado anteriormente, que la voz constituye en realidad un fenómeno continuo. A lo largo de una frase las variaciones en la morfología del tracto vocal y las características de la excitación dan como resultado un cambio permanente del espectro de la señal resultante. En el caso más general estos patrones espectrales permiten caracterizar a los distintos fonemas mediante determinadas *pistas acústicas* que son requeridas para poder diferenciarlos.

[6]En el modelo lienal de producción de la voz fuente-filtro discutido en la Sección 2.2.1 F_0 es una caracterítica de la fuente mientras que F_1 y F_2 corresponden a características del filtro.

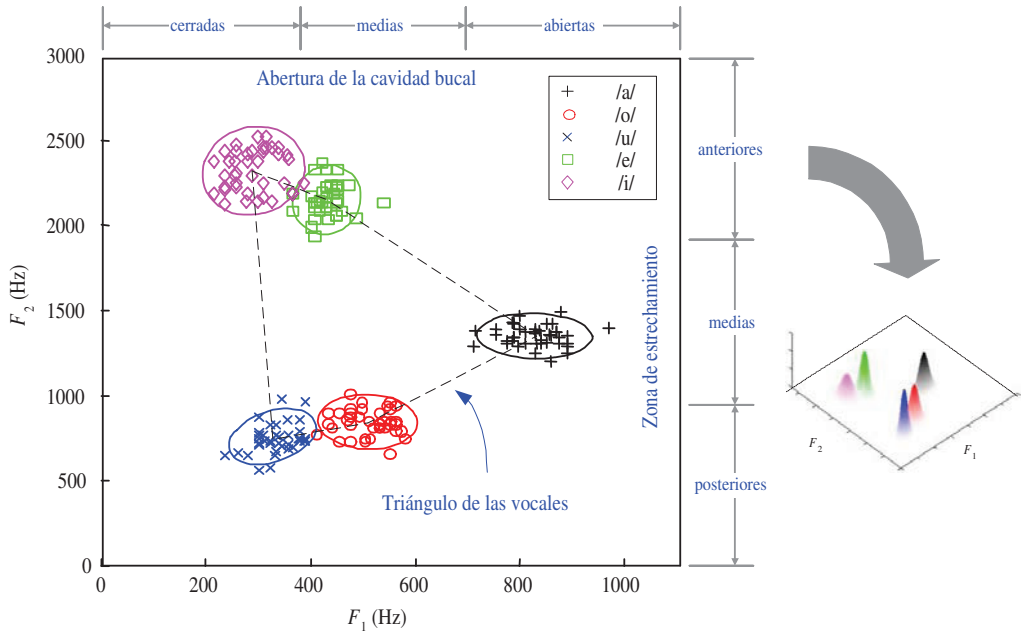

Figura 2.8: *Mapa de las formantes obtenido a partir de datos experimentales para las vocales del español pronunciadas en forma sostenida por un conjunto de hablantes masculinos. Para dibujar las elipses se ha supuesto una distribución gaussiana bidimensional para cada clase [5]. Sobre el mapa se ha superpuesto el clásico triángulo de las vocales del español, mostrando además sobre ambos ejes la relación de F_1 con la abertura de la boca y de F_2 con las zonas de estrechamiento del tracto vocal.*

Existen algunas características de la señal de voz que se pueden manifestar mediante análisis relativamente simples como ser la *energía de corta duración* y la *cantidad de cruces por cero* (Cx0). Estos análisis tienen la ventaja de ser sencillos en su implementación digital y muy rápidos. La energía da una idea de la intensidad de la señal en función del tiempo y constituye un parámetro de suma importancia ya que permite diferenciar entre varios tipos de fonemas. Es también una parte esencial de la entonación (junto con F_0). Los cruces por cero constituyen una medida indirecta del contenido frecuencial de la señal. En la Figura 2.9 se observa una sección ampliada de la frase "¿Cómo se llama el mar...". En ella se muestran el espectrograma, las formantes y las curvas derivadas de estos análisis temporales. Se pueden destacar algunas pistas acústicas presentes en el espectrograma de esta figura. Se observa la corta duración y la *explosión* de la oclusiva */k/*. La estructura formántica de las vocales está evidenciada por las regiones más oscuras de conjuntos equiespaciados de lineas paralelas en dirección horizontal, producto de su carácter sonoro cuasiperiódico. Se puede observar también el contenido de alta frecuencia de la */s/* y la ausencia de sonoridad. En general otro rasgo distintivo de los fonemas sonoros consiste en que poseen una menor cantidad relativa de Cx0 que de energía (ver por ejemplo */o/* y */a/*). La situación inversa puede apreciarse en los fonemas sordos (no

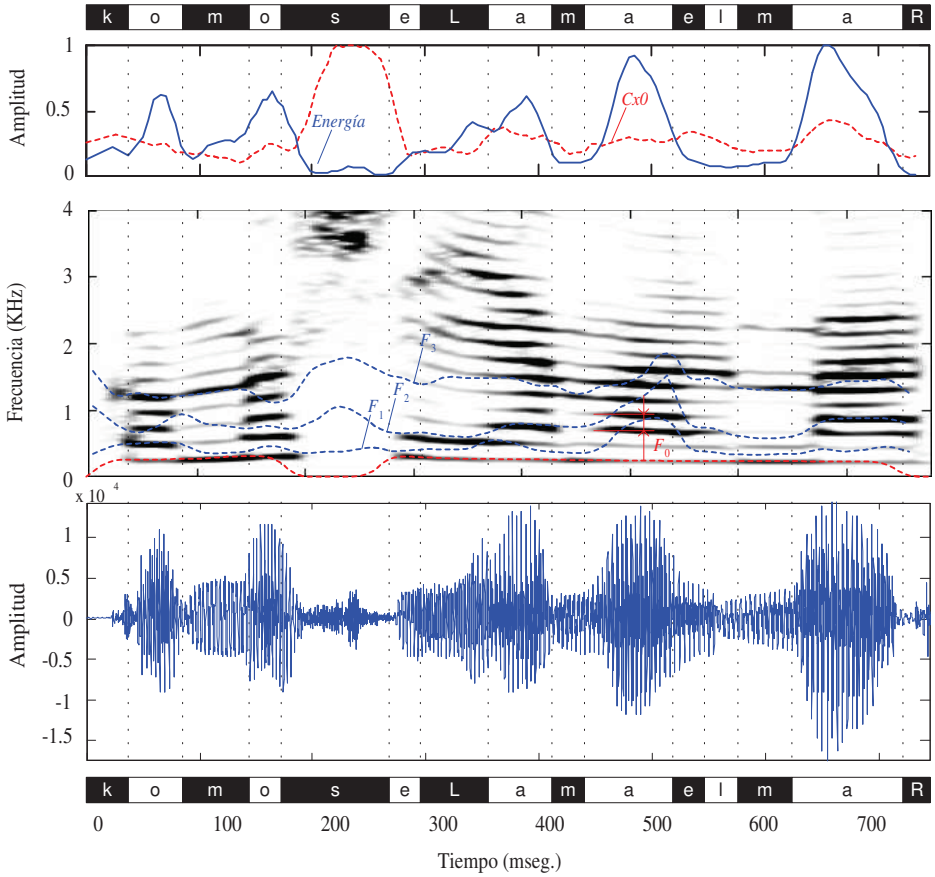

Figura 2.9: *Sonograma, espectrograma, formantes, energía y cruces por cero simultáneos de un trozo de la frase "¿Cómo se llama el mar...?", segmentada y etiquetada . La combinación simultánea de estos análisis permite la rápida caracterización de los diferentes fonemas (etiquetas de acuerdo al alfabeto fonético* Worldbet, *frase tomada de la base de datos de habla española Albayzin [15]).*

sonoros), como la /*s*/, debido a que poseen poca energía y distribuida en las frecuencias altas. De esta manera es posible distinguir rápidamente entre ambas clases. En el caso de los fonemas sordos puede apreciarse también la pérdida de la sonoridad por la anulación de F_0 (otra vez como en /*s*/).

Pueden destacarse también otras pistas acústicas que permiten discriminar entre los diferentes fonemas, generalmente visibles en su representación espectral. En la Figura 2.10 pueden observarse algunos ejemplos de estas pistas que permiten discriminar entre /*s*/, /*f*/, /*m*/, /*n*/, /*l*/ y /*r*/ [13]. La /*s*/ suele ser fácil de reconocer y distinguir de la /*f*/. Ninguna de la dos posee componente glótica. En el caso de la /*s*/ aparece un área de fricación de mayor energía en la zona de las altas frecuencias (entre los 3000 y los 8000 Hz) En el caso de la /*f*/ el área de mayor energía suele ser un triángulo alrededor de los 1200 Hz. También puede existir alguna coarticulación con los fonemas adyacentes. De forma similar pueden establecerse algunas pistas para discriminar entre los fonemas

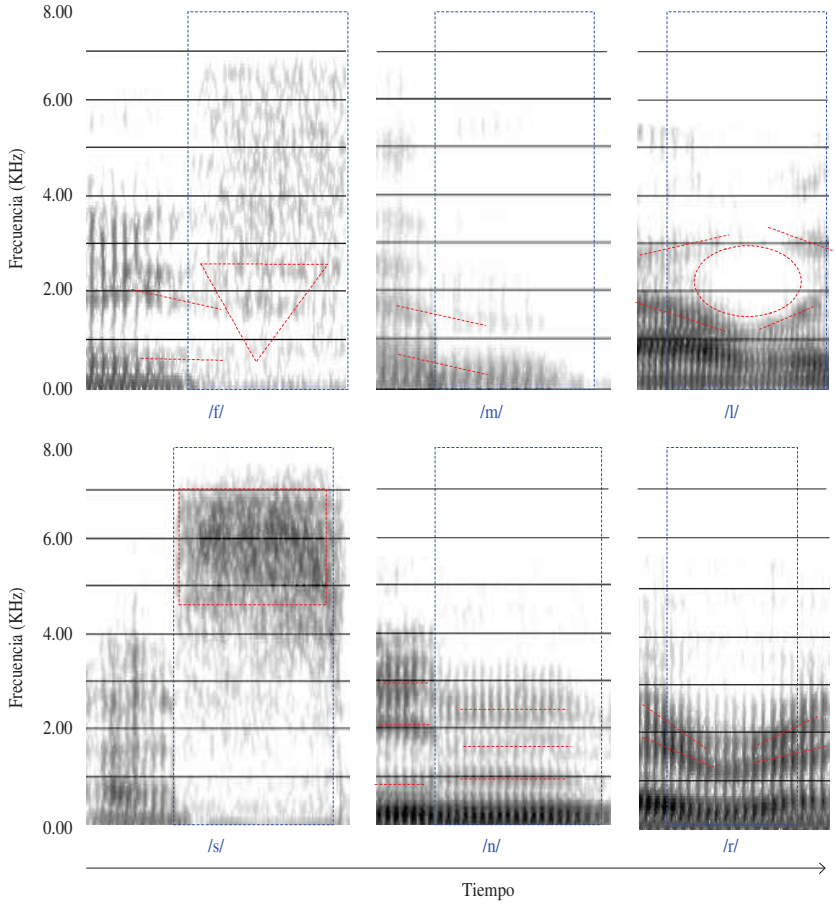

Figura 2.10: *Pistas acústicas correspondientes a ejemplos típicos de varios de los fonemas explicados en el texto resaltadas en los correspondientes espectrogramas de banda ancha (espectrogramas tomados de [13]). Estas pistas o rasgos acústicos permiten discriminar entre los diferentes fonemas (o alófonos de los mismos).*

sonoros */m/* y */n/*. En el caso de */m/* las formantes generalmente se "sumergen" dentro del fonema y luego se elevan cuando este termina, excepto cuando las frecuencias de las mismas ya son bajas. En */n/* el cambio suele ser más abrupto. El nivel de frecuencia al que tiende F_2 para */m/* está entre 900 a 1400 Hz, mientras que para */n/* está entre 1650 a 1800 Hz. Para el fonema */l/* es posible notar un "hueco" (cero o anti-resonancia) en el espectro, aproximadamente entre 1500 y 2000 Hz. A ambos lados de este hueco F_2 y F_3 al principio divergen y posteriormente se juntan. En algunos casos */l/* sólo se puede distinguir como una disminución en la energía de F_2 y F_3. En el caso del fonema */r/* se puede apreciar que F_3 y F_2 se acercan, o inclusive se combinan, siempre se fuerza F_3 por debajo de 2000 Hz.

Podrían llenarse muchas páginas con gráficos y análisis de los distintos fonemas. Sin embargo el interés aquí no es presentar este material de manera exhaustiva sino más

bien, y como ya se mencionó, mostrar unos pocos ejemplos que permitan comprender mejor la naturaleza de la señal de voz y sus rasgos más significativos.

Como consideraciones finales de esta sección se debe remarcar el hecho ya discutido acerca de que la realización acústica de un fonema depende mucho de su contexto inmediato. Por otra parte muchas veces, especialmente en el caso del habla espontánea, los fonemas no están articulados adecuadamente o no se parecen tanto a lo que se esperaba idealmente. El hecho que el habla sea una secuencia continua de fonemas sin pausas acústicas explícitas entre las palabras constituye un problema adicional.

2.4. Fisiología de la audición

En este trabajo, resulta de interés comprender cómo se realiza el procesamiento de la señal de habla en el sistema auditivo. Se debe tener en cuenta que este sistema realiza una enorme cantidad de procesamiento para que la señal llegue hasta nuestro cerebro, pero es realmente allí donde se produce el fenómeno de la audición. Se podría decir entonces que en realidad "escuchamos" con el cerebro. Por ello es importante comprender que rasgos significativos se preservan en las representaciones internas de la corteza cerebral, y cuales son los principios que orientan la formación de estas representaciones. Se podría realizar la siguiente pregunta: ¿Que características del sistema auditivo son particularmente apropiadas para codificar la voz?. La respuesta, en parte, se encuentra en la magnífica capacidad de este sistema para resolver simultáneamente tanto las características espectrales como temporales de los estímulos de banda ancha que constituyen el habla humana. Por otra parte esta capacidad se mantiene aún en condiciones acústicas muy desfavorables, con relativa independencia de cambios en el canal (presencia de ruido o ambiente reverberante) o la fuente del mensaje (velocidad de pronunciación o identidad del hablante).

En la Figura 2.11 puede apreciarse un corte transversal del oído, junto con un diagrama esquemático que ilustra su funcionamiento. En el mismo se observan sus tres secciones principales: el oído externo, el medio y el interno. Se podría decir que las dos primeras partes se encargan de la recepción y adecuación del sonido para su posterior procesamiento en la sección siguiente. Las funciones más importantes, como la transducción del sonido a impulsos nerviosos, se realizan en el oído interno. Se describirán a continuación estas partes del oído y sus funciones con mayor detalle.

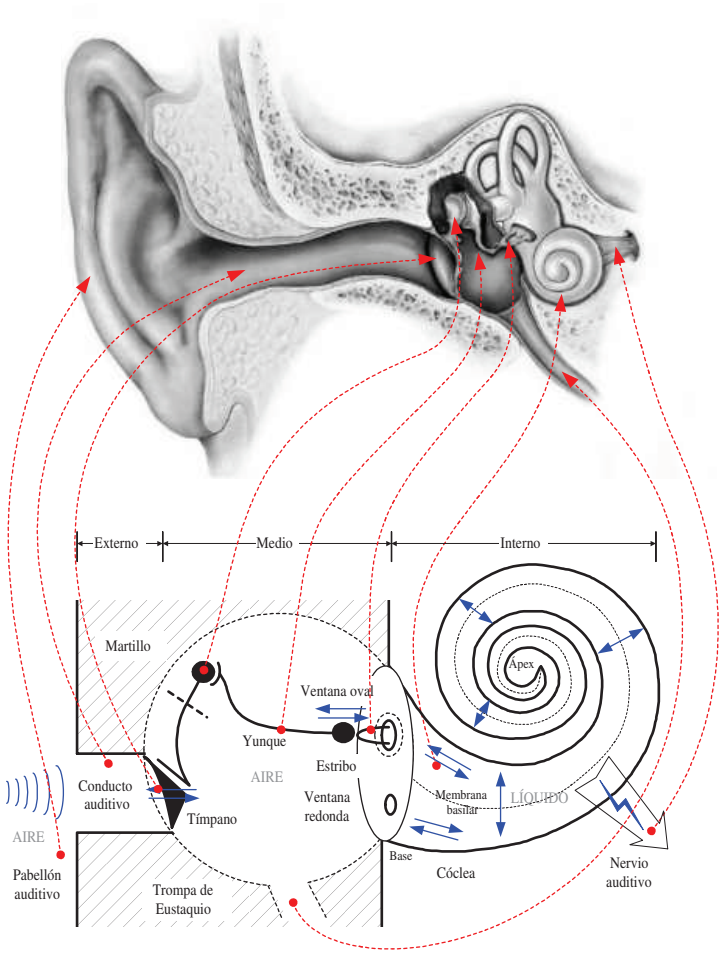

Figura 2.11: *Corte sagital anatómico del oído (arriba) y diagrama esquemático que ilustra su funcionamiento (abajo). El oído es el encargado de la recepción y adecuación del sonido y de su transducción a impulsos nerviosos. En el diagrama se resaltan sus secciones principales: el oído externo, el medio y el interno, que son las que realizan cada una de estas tareas.*

2.4.1. Recepción y adecuación acústica

El oído humano funciona en un medio aéreo y por ello necesita cierta eficiencia para la recepción de sonidos transmitidos por el aire. La parte más externa es el *pabellón auditivo* que está encargado de captar el sonido y enfocarlo hacia el *conducto auditivo*. Las ondas de presión siguen el conducto auditivo hasta el *tímpano* que separa oído externo del oído medio. Este último está constituido por una cámara ocupada por aire (que se comunica con la faringe a través de la *trompa de Estaquio*) y un conjunto de huesecillos: el *martillo*, el *yunque* y el *estribo*. El sonido se transmite entonces desde la membrana del tímpano a través de la cadena de huesecillos, cuya función principal es la de adaptación de impedancias acústicas [80]. El *estribo*, el más interno de estos huesecillos, establece contacto con la *ventana oval* que está ubicada en la base de la *cóclea*, en lo que

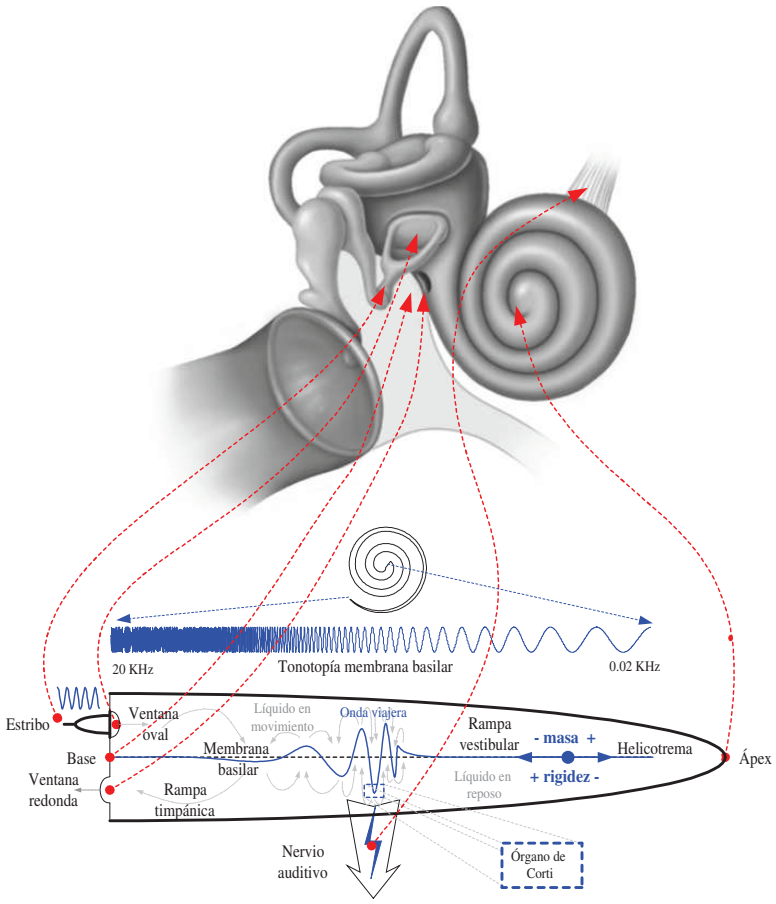

Figura 2.12: *Cóclea aislada (arriba) y diagrama esquemático que ilustra su funcionamiento (abajo). En el diagrama la cóclea se halla desplegada para mayor claridad. Se muestra también la forma de una* onda viajera *típica (cuya amplitud se ha exagerado) y se resaltan los aspectos relativos a la tonotopía de la membrana.*

constituye el oído interno. La amplificación de las vibraciones producidas en el tímpano está limitada, en condiciones de cambios abruptos, por el *reflejo estapedial* para proteger al oído interno[7].

2.4.2. Transducción mecánico-eléctrica

El órgano principal del oído interno es la cóclea. La cóclea puede describirse como un tubo cónico lleno de líquido (*perilinfa*) y enrollado en forma de caracol. En la Figura 2.12 puede apreciarse una versión aislada y ampliada de la misma con su correspondiente diagrama esquemático. En este diagrama la cóclea se muestra desenrollada para mayor claridad. Una vez excitada la ventana oval el sonido se transmite a través del líquido de la *rampa vestibular* en la cóclea, atraviesa el *helicotrema* y sigue su recorrido en la *rampa*

[7]Esto funciona en la práctica como un *control automático de ganancia mecánica.*

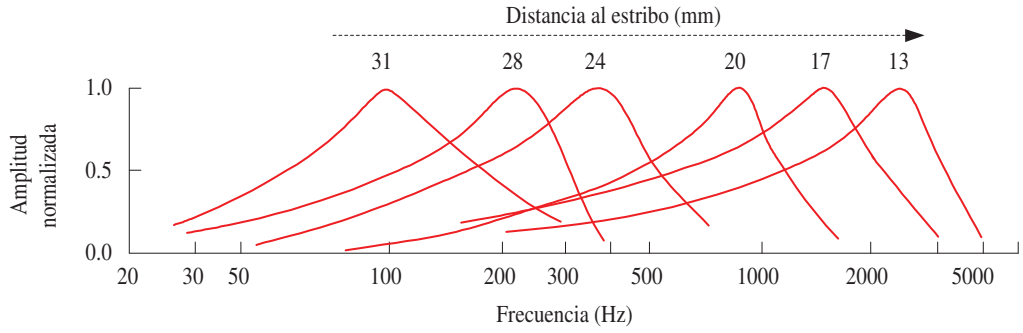

Figura 2.13: *Curvas de Resonancia: amplitudes relativas de las excursiones de la membrana basilar como función de la frecuencia de estimulación, para seis puntos a lo largo de la membrana. El estudio se realizó con cadáveres por lo cual algunos mecanismos activos no están presentes (adaptado de [9]).*

timpánica hasta la *ventana redonda*. La ventana oval y la redonda trabajan de forma tal que cuando una se comba hacia adentro la otra se comba hacia afuera y viceversa. El movimiento hacia adentro y afuera se repite con la misma frecuencia del estímulo sonoro.

Es en la *membrana basilar* donde tiene lugar la transducción, de manera selectiva, en base a la relación de las características del estímulo y la zona de vibración de la misma [60, 9, 41]. La membrana basilar varía sus propiedades mecánicas de forma continua a lo largo de su eje longitudinal. La membrana es más rígida en su base, cerca de la ventana oval, donde su ancho es mínimo. Por lo tanto tiene allí menor cantidad de masa por unidad de longitud. Esto hace que la región de la base región vibre con preferencia ante un estímulo de alta frecuencia. De esta forma, las vibraciones de frecuencias altas tienen su máxima amplitud cerca del lugar donde las ondas comienzan a desplazarse, luego disipan la mayor parte de su energía y se desvanecen en el camino, no alcanzando nunca el ápex. Las vibraciones de baja frecuencia, por el contrario, comienzan con una amplitud pequeña cerca de la base y la aumentan a medida que se acercan al ápex. De esta manera están representadas todas las frecuencias audibles a lo largo de toda la cóclea. A esta característica se la denomina *tonotopía de la membrana*.

Se han registrado las excursiones máximas de la membrana basilar en función de la distancia al estribo (envolventes de la onda de desplazamiento), para tonos de igual intensidad pero distintas frecuencias. Empleando estos datos se pueden dibujar las curvas de resonancia o sintonía mecánica, esto es las amplitudes relativas de las excursiones para los distintos puntos sobre la membrana basilar como una función de la frecuencia del estímulo (Figura 2.13). De estas curvas de sintonía resulta ser que la relación entre la distancia al estribo y la frecuencia de vibración máxima no es lineal, sino más bien de tipo logarítmica. Esta es una de las causas por las que la resolución frecuencial y la

percepción de las frecuencias no es uniforme en toda la cóclea. A la escala psicoacústica que da cuenta de la relación entre la frecuencia física del sonido y la percibida se la denomina *escala de mel* (Ver Figura 3.16 más adelante). Los experimentos psicofísicos demuestran también una escala similar de carácter logarítmico en la percepción de la intensidad de los sonidos, cuya unidad es el *fono*[8].

La transducción mecánico-eléctrica se produce en el denominado *órgano de Corti* ubicado a lo largo de toda la membrana basilar (Ver Figura 2.14). Ésta tiene lugar como respuesta a una curvatura de las *cilias* de las *células ciliadas*. Esta curvatura produce una variación en el potencial de membrana de las células; si las cilias se curvan hacia el cuerpo basal se produce una despolarización, mientras que si se curvan en el otro sentido se produce una hiper-polarización.

La excitación de las células ciliadas está determinada, en gran medida, por las excursiones de la membrana basilar. Sobre ella actúan las ondas de presión oscilatorias resultantes de la transmisión del sonido en las rampas vestibular y timpánica. De esta manera –dado que la amplitud de las vibraciones en distintos puntos de la cóclea varía con la frecuencia del estímulo– el grado en el cual es excitada una determinada célula ciliada es una función conjunta de su posición en la membrana basilar y de la amplitud del estímulo.

La curva de resonancia de la membrana basilar de la Figura 2.13 describiría con precisión la excitación de las células ciliadas en función de la frecuencia, si éste fuera el único factor que influyera en la vibración de las células ciliadas. Sin embargo, las propiedades mecánicas de las cilias y de la *membrana tectoria* que las cubre también influyen en la vibración de las células ciliadas. De hecho, la rigidez de las cilias, la masa y la elasticidad de la membrana tectoria también varían de un extremo al otro de la cóclea. Se ha registrado también cierto comportamiento "activo" de algunas células ciliadas[9]. Además de ello el penacho ciliar posee propiedades mecánicas especiales que derivan en un comportamiento no-lineal. Ésto parece explicar el conocido efecto de "oír" un tercer tono cuando solo se estimulo con dos [132]. Estas características del complejo célula-membrana tectoria tiene el efecto de limitar la sintonía de las células ciliadas a un ancho de banda de frecuencias más estrecho que el del punto de la membrana basilar donde se encuentra la célula. Se debe mencionar también que las células ciliadas se despolarizan solo durante la fase positiva de los estímulos sonoros produciendo un efecto

[8]Por ejemplo, si un sonido complejo con muchas componentes, parece igualmente intenso que un tono puro de 1000 Hz con un nivel de presión de 80 dB (SPL), aquel tendrá un nivel de sonoridad de 80 fonos, independientemente del nivel de presión "real" que tenga.

[9]Cuando éstas son estimuladas eléctricamente cambian su longitud.

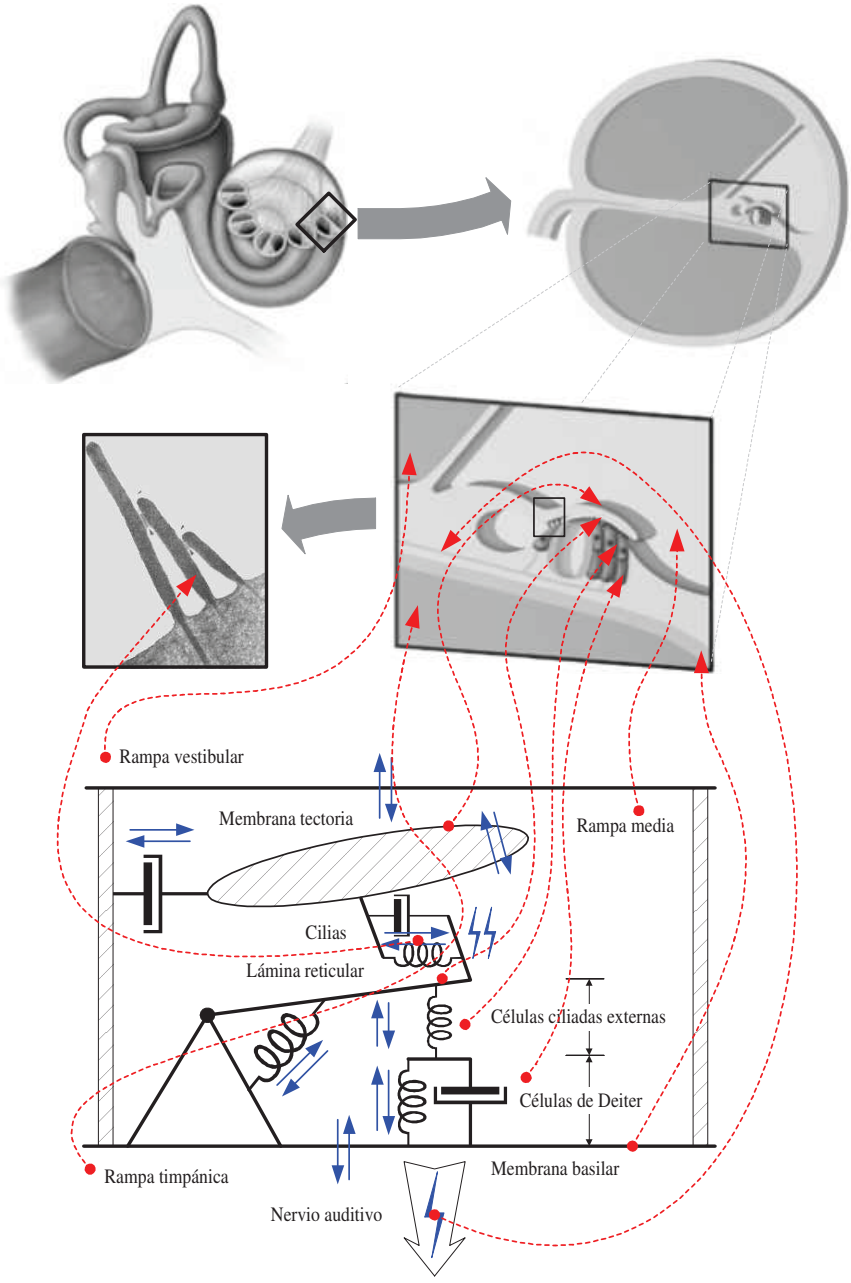

Figura 2.14: *Detalle del órgano de Corti y las células ciliadas (arriba). Diagrama esquemático que ilustra su funcionamiento (abajo). En el órgano de Corti, ubicado a lo largo de toda la membrana basilar, se produce la transducción mecánico-eléctrica.*

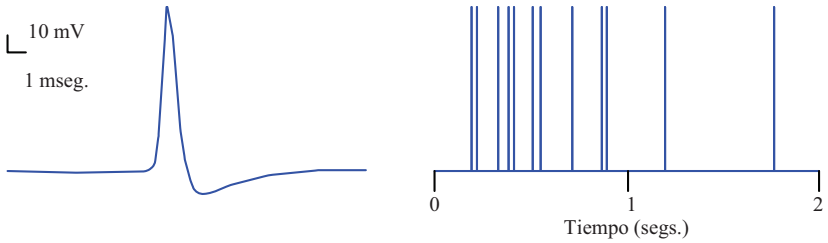

Figura 2.15: *Potencial de acción o pulso típico producido por la despolarización de una célula nerviosa o neurona (izquierda). Tren de pulsos característico producido por la despolarización repetida de una neurona como respuesta ante distintos estímulos de entrada (cada tiempo de disparo esta señalado con una barra vertical, derecha). Todo el código de comunicación neuronal está basado en estos trenes de pulsos (adaptado de [101]).*

de *rectificación de media onda* sobre las respuestas del nervio auditivo [124].

2.4.3. Nervio auditivo y codificación nerviosa

Una pregunta fundamental en neurociencias está relacionada con la comprensión del código neuronal que se utiliza para organizar las distintas señales dentro del sistema nervioso. Este código está basado en la utilización de trenes de pulsos como el mostrado en la Figura 2.15. Estos trenes de pulsos se encuentran a todos los niveles del sistema, desde los transductores sensoriales hasta la corteza cerebral. En esta sección se discuten distintos aspectos que permiten explicar la codificación de los sonidos a nivel del nervio auditivo, mientras que en las secciones siguientes se describe lo ocurrido a lo largo del resto de la vía auditiva hasta llegar a la corteza.

El nervio auditivo está formado por la colección de axones periféricos correspondientes a las neuronas aferentes y eferentes que inervan a las células ciliadas. Aquí el interés principal se pondrá en la parte aferente, es decir aquellas fibras que llevan información desde la periferia auditiva en la dirección del sistema nervioso central. La respuesta de una fibra aislada puede describirse en términos de la frecuencia del correspondiente tren de pulsos, su fase y su patrón temporal de activación. Se considera que la respuesta de una fibra es estocástica, en el sentido que el patrón de disparo está relacionado de manera probabilística con las características del estímulo [159]. Aún sin estimulación acústica muchas fibras poseen respuesta espontánea, y ésta varía de fibra a fibra. Para el caso de tonos puros es posible suponer que existen tres características del estímulo que se deberían codificar a nivel nervioso: la intensidad, la frecuencia y la fase. La codificación de la fase es directa y tiene importancia principalmente en cuestiones de ubicación espacial de la fuente sonora. De acuerdo con lo presentado en la sección anterior se podría pensar que la frecuencia se codifica en términos de cuál es la fibra individual que dispara, y la

intensidad en la tasa de disparo de los pulsos. Sin embargo, aunque ésto puede representar una primera aproximación, la codificación de los diferentes sonidos puede ser bastante más compleja y utilizar estrategias "mixtas" como se discutirá a continuación.

Respuesta a estímulos simples

Como se ha visto, la membrana basilar está mecánicamente sintonizada con la frecuencia del sonido aplicado; por esta razón se puede pensar que las descargas nerviosas provenientes de zonas determinadas de la membrana basilar ya poseen la información de la frecuencia del estímulo. A esta forma de codificación de la frecuencia del estímulo se la denomina *mecanismo de la localización*. Los estudios fisiológicos iniciales de trenes de pulsos en fibras únicas del nervio auditivo brindaron información importante acerca de estos aspectos [81]. Estos estudios se realizaron en animales, principalmente en gatos, debido a la dificultad para realizarlos en humanos. Para ello se utilizaron fundamentalmente tonos puros. Una vez aislada una fibra se pudieron registrar impulsos de esa fibra única. De esta forma se obtenía una *curva de sintonía nerviosa* que trazaba los umbrales de respuesta en función de la frecuencia (Ver Figura 2.16). El mínimo de esta curva de sintonía (*frecuencia característica* o FC) indica el lugar a lo largo de la cóclea que ocupa la célula ciliada que excita la fibra. Ésto quiere decir que FC es la frecuencia para la cual la intensidad de estímulo necesaria para excitar la fibra es la mínima. Para estas fibras si estimulamos a la FC la intensidad del estímulo se codifica en la frecuencia o tasa de disparo (siempre por encima de su frecuencia espontánea). Se debe recalcar el hecho de que las fibras no responden a una única frecuencia, aunque requieren una mayor intensidad para ser excitadas fuera de su FC[10]. Ésto también sirve para codificar información acerca de la intensidad del estímulo (de acuerdo a la cantidad de fibras que responden).

En la Figura 2.17 se observa la curva de resonancia mecánica en un punto de la membrana basilar y la curva de sintonía de una fibra nerviosa que inerva a la célula ciliada en ese punto. La curva de resonancia muestra los niveles de presión sonora relativos requeridos para hacer vibrar la membrana en ese punto a una amplitud dada para varias frecuencias de sonido. La curva de sintonía muestra el umbral de la fibra nerviosa en función de la frecuencia del estímulo sonoro. Nótese que ambas curvas tienen frecuencias de corte similares, pero del lado de las bajas frecuencias la curva de sintonía posee una subida mucho más abrupta que la de resonancia. Se propusieron varios mecanismos para explicar esta aparente discrepancia entre las curvas de sintonía mecánicas y nerviosas.

[10]Esto se conoce como el problema del rango dinámico de una fibra nerviosa auditiva.

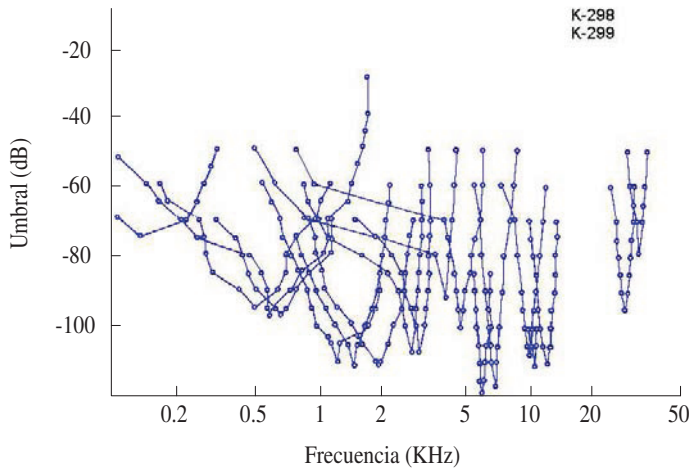

Figura 2.16: *Curvas de sintonía nerviosa: umbral de respuesta en función de la frecuencia de estimulación para varias fibras individuales del nervio auditivo de gato (adaptado de [81]). La frecuencia característica de una fibra es el mínimo de esta curva de sintonía y está relacionada con el lugar a lo largo de la cóclea que ocupa la célula ciliada que excita la fibra en cuestión.*

Estudios de la mecánica de la membrana basilar utilizando métodos más refinados mostraron una agudeza de sintonía mecánica bastante parecida a la de la sintonía neural [144].

Además de la percepción de la frecuencia de acuerdo a la posición de la fibra, para tonos de baja frecuencia (< 1 KHz) e intensidad moderada, las descargas nerviosas de una fibra determinada pueden "seguir" a los estímulos en frecuencia con una relación uno a uno. Ésto quiere decir que la información de la frecuencia se codifica también en la tasa de disparos. Sin embargo para tonos de frecuencias mayores ya no es posible seguir el "ritmo" tan de cerca. Entonces se recurre al fenómeno de excitación de varias fibras simultáneas, cada una con una fase diferente pero invariante. Este fenómeno, denominado respuesta *enganchada en fase*, permite la codificación de la frecuencia del estímulo en forma "distribuida" entre varias fibras. Este mecanismo funciona de manera confiable aproximadamente hasta los 3 KHz [124]. Este último modelo para la codificación de la frecuencia del estímulo se denomina *mecanismo temporal*.

Como resumen podríamos decir que existe acuerdo de que para la codificación de la frecuencia coexisten los dos mecanismos expuestos. Ésto es que para las bajas frecuencias se utiliza principalmente el temporal y para altas frecuencias principalmente el de localización. Sin embargo hay discrepancia acerca de la frecuencia a la cual comienza a reemplazarse uno por el otro [31]. Para la codificación de la intensidad también existe coincidencia acerca de un mecanismo mixto entre las tasas de disparo individuales y la cantidad de fibras que responden, según se ha explicado en esta sección.

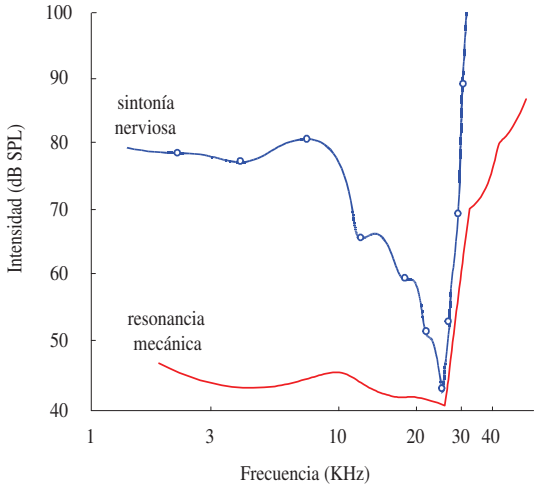

Figura 2.17: *Comparación entre resonancia mecánica y sintonía nerviosa en un punto de la membrana basilar (adaptado de [144]). Es posible observar que ambas curvas poseen frecuencias de corte similares, pero del lado de las bajas frecuencias la curva de sintonía nerviosa posee una subida mucho más abrupta que la de resonancia mecánica.*

Respuesta a estímulos complejos

La distinción entre estímulos simples y complejos es algo arbitraria. Se puede hablar de "complejo" en el sentido espectral cuando se tiene más de un tono puro. Complejidad también puede referirse al caso de señales no periódicas o aleatorias. A veces puede relacionarse con la cantidad de parámetros necesarios para una descripción matemática completa. Bajo este punto de vista todas las señales "naturales" son complejas. El interés aquí está puesto en aquellos estímulos sonoros similares al habla humana.

El estudio detallado de la respuesta del nervio auditivo a este tipo de estímulos requirió de algún tiempo. Los estudios iniciales con tonos puros daban solo una aproximación lineal para el estudio de un sistema "bastante" no lineal. Estas no linealidades no solo se dan a nivel nervioso sino inclusive a nivel de la mecánica coclear[11](como ya se ha visto en la Sección 2.4.2). Por ello no es posible comprender el comportamiento frente a estímulos complejos por la simple adición de los efectos producidos por sus componentes sinusoidales.

La continuación "natural" en este sentido de los estudios con tonos puros fue la utilización de tonos múltiples y señales de voz sintéticas basadas en modelos de producción del habla [112]. Con posterioridad se comenzó a trabajar con señales de voz reales [14]. Los estudios se continuaron realizando en animales[12].

[11] A pesar de ello muchas de las pruebas clínicas para valorar la audición de uso habitual en la actualidad continúan utilizando principios lineales debido a su simplicidad.

[12] Aunque es posible realizar extrapolaciones al caso del hombre, debe tenerse en cuenta que el procesamiento de sonidos como el habla puede ser diferente ya que se trata de criaturas que no poseen

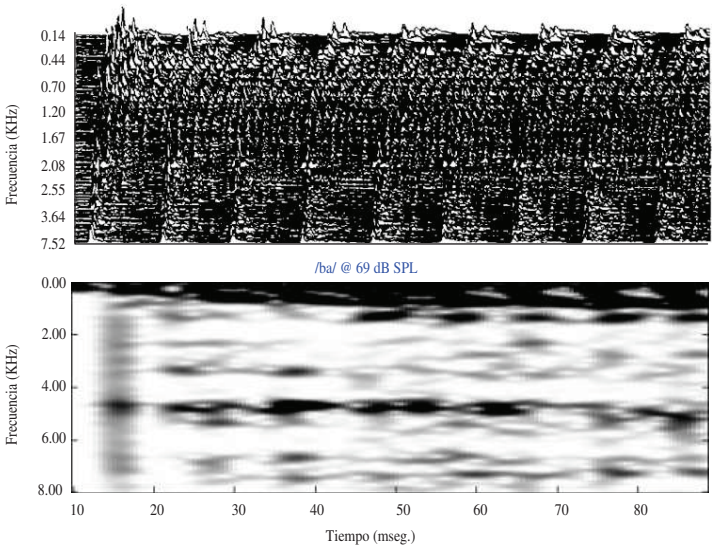

Figura 2.18: *Neurograma: tasas de disparo instantáneas promedio de las fibras del nervio auditivo del gato como respuesta a la estimulación acústica mediante la sílaba /ba/ sintética (arriba, tomado de [151]). Espectrograma de la misma sílaba pronunciada por un hablante masculino (abajo, nótese que el eje de frecuencias está invertido para facilitar la comparación).*

En general se asume que la representación de la señal del habla en el nervio auditivo está compuesta por un numero finito de elementos (aproximadamente 30.000 fibras del nervio auditivo en el hombre) y las respuestas de cada elemento están determinadas por una secuencia compleja de estados distribuidos e iterativos que preceden la iniciación de los pulsos de descarga. El nervio auditivo puede considerarse una disposición ordenada de elementos arreglados de acuerdo a la FC. Las fibras en esta disposición responderán incrementando su probabilidad de descarga cuando el nivel del estimulo supera el umbral. El neurograma [151] es una representación directa de la información experimental de la estimulación del nervio auditivo, ordenada de acuerdo con la FC de las fibras individuales. En la Figura 2.18 se muestra un neurograma basado en las respuestas fisiológicas al sonido */ba/* sintetizado. Cada línea del neurograma representa tasa de disparo instantánea promedio de una fibra nerviosa. La FC de la fibra está dada a la izquierda. A pesar del parecido con el clásico espectrograma, el neurograma presenta información de manera distinta, utilizando otra forma de codificar los patrones generados "más a la medida del sistema auditivo". En todos estos trabajos se pudo encontrar las fibras nerviosas respondían como detectores de características sencillas, como ser la ubicación y seguimiento de las frecuencias formantes o la detección del *tiempo de ataque de la sonoridad*[13] (TAS, en inglés *voiced onset time* o VOT) [14]. Como se ha visto éstas cons-

un lenguaje hablado (aunque pueden reconocer palabras). El problema parece ser mayor a medida que avanzamos en la vía auditiva hacia centros más especializados.

[13]Se denomina así al tiempo transcurrido entre la liberación de la presión sonora posterior a una

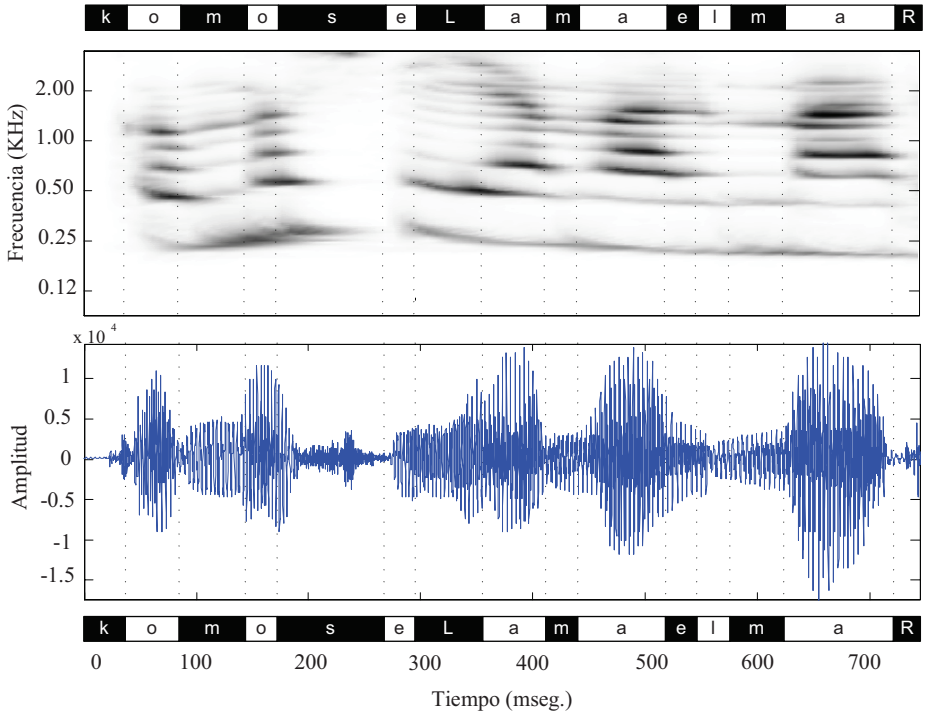

Figura 2.19: *Sonograma (abajo), espectrograma (centro) y espectrograma auditivo (arriba) del trozo de la frase de la Figura 2.9. Los diferentes tonos de gris expresan la actividad neuronal de cada fibra del nervio auditivo ordenadas de acuerdo a su frecuencia característica.*

tituyen pistas acústicas importantes para la discriminación de los fonemas. El examen fino de los patrones temporales de descarga de las fibras reveló además la codificación de otras características espectrales simples (como la representación directa de F_0). El efecto de enganche de fase discutido anteriormente resalta los picos espectrales en los sonidos complejos. La redundancia asociada a este mecanismo provee cierta robustez en la codificación y ésta es una de las razones por las cuales la información más importante del habla se concentra en las bajas frecuencias [52]. También se corroboraron algunos efectos de enmascaramiento de frecuencias en la presencia de estímulos simultáneos y no simultáneos. Para tener una idea del tipo de representación a nivel del nervio auditivo[14] se han desarrollado varios modelos que incluyen los principales aspectos discutidos en esta sección y las anteriores, y que se han validado mediante experimentos fisiológicos [172, 153]. La salida de estos modelos sería equivalente al neurograma ya descripto y suele denominarse *espectrograma auditivo*. En la Figura 2.19 puede apreciarse un espectrograma auditivo para un trozo de una oración, en comparación con el sonograma

consonante plosiva, es decir el momento de apertura de los labios, y el comienzo de la vibración de las cuerdas vocales en el fonema sonoro subsiguiente.

[14] A este tipo de representaciones se las refiere como *representaciones auditivas tempranas*.

y el espectrograma tradicional correspondientes. Se puede notar fácilmente la mayor resolución frecuencial en la zona de las bajas frecuencias.

2.4.4. Vía auditiva

El nervio auditivo constituye sólo la primera parte de la denominada *via auditiva* (ver Figura 2.20). A lo largo de este camino, que lleva a la *corteza auditiva*, las señales nerviosas atraviesan una serie compleja de etapas de procesamiento en el tronco cerebral a través del *núcleo coclear*, el *núcleo olivar superior*, el *colículo inferior* y el *núcleo geniculado medio*. Las 30.000 fibras del nervio auditivo humano, se convierten en unos 100 millones de neuronas en cada lado de la corteza auditiva[15]. La organización tonotópica de la cóclea se mantiene en diversas partes de la vía auditiva, incluyendo la propia corteza. En el núcleo coclear se detectan algunos eventos acústicos simples, como comienzos y finales de fonemas y algunas transiciones. Ésto ha llevado a conjeturar que juega el papel de un modelo articulatorio inverso aproximado [113]. En el núcleo olivar superior se realiza la integración de la información proveniente de ambos oídos, cuyo objetivo principal es el de proveer la localización espacial de las fuentes de sonido. A partir de allí se continua en forma ascendente principalmente con información biaural, aunque existen centros que continúan procesando en forma monoaural [124]. La integración de las diferentes vías continúa en el colículo inferior. Allí se procesan y analizan principalmente aquellos con patrones temporales especiales, como ser los modulados en frecuencia o con una duración específica. Antes de llegar a la corteza la información auditiva pasa por el núcleo geniculado medio, que es el primer lugar donde se generan respuestas específicas para ciertas combinaciones espectrales. Estas respuestas incluyen no solo la detección de combinaciones de frecuencias simultáneas, sino también de intervalos específicos entre dichas frecuencias. De esta manera a medida que se avanza hacia la corteza aparecen detectores de características cada vez más complejas [162].

2.4.5. Corteza auditiva

La corteza auditiva es la encargada de procesar los estímulos nerviosos para convertirlos en diferentes representaciones internas. Un dato neurobiológico importante es la arquitectura neuronal de la corteza auditiva. La corteza está formada por varias capas de células nerviosas, cada una de las cuales está constituida por tipos específicos de neuro-

[15] Se verá más adelante que esta sobre-representación es una característica importante para la robustez del sistema.

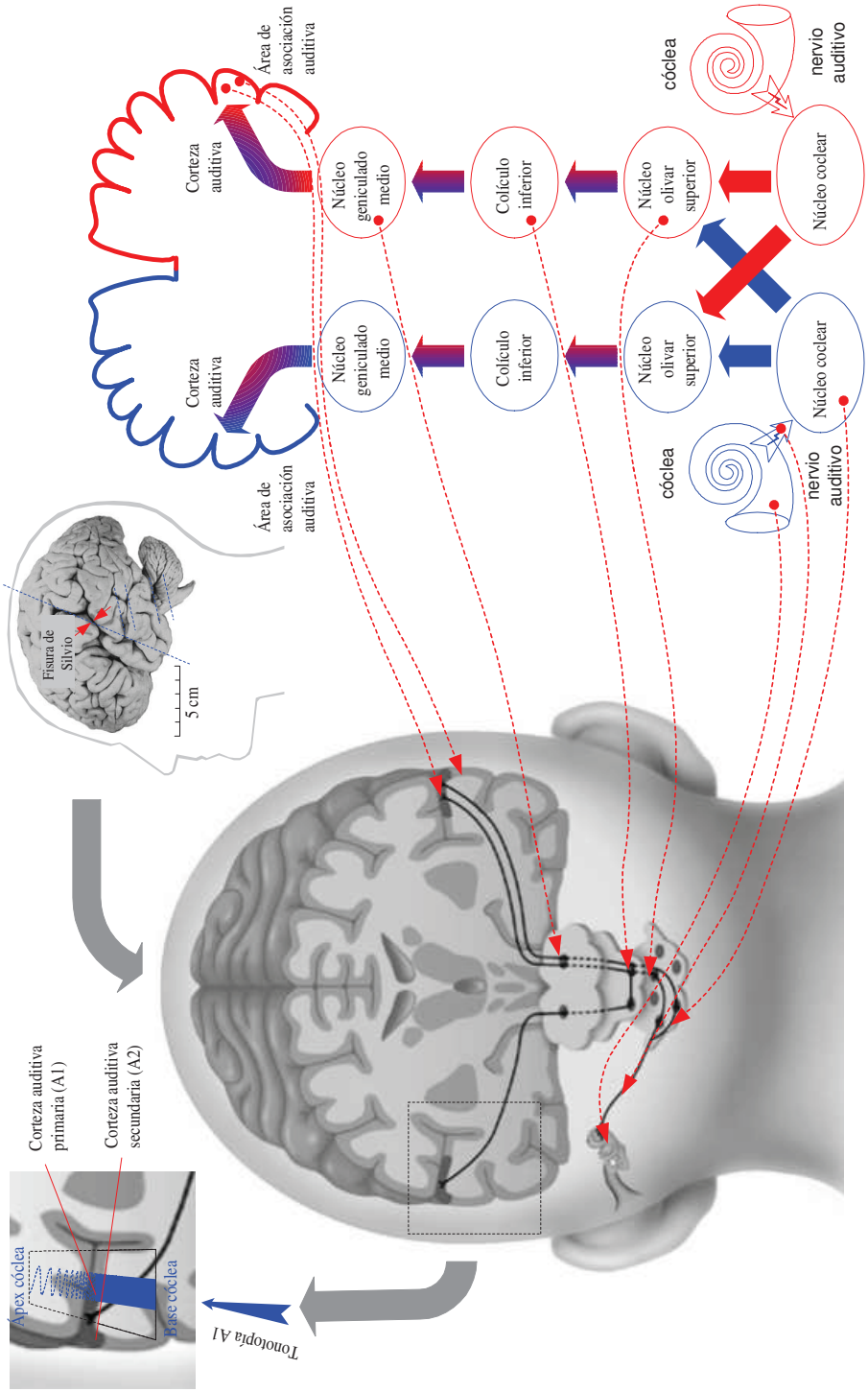

Figura 2.20: *Diferentes secciones de la vía auditiva y detalle de la corteza auditiva (izquierda). Diagrama esquemático que ilustra las conexiones y el trayecto seguido por la información en la vía, desde el nervio auditivo hasta la corteza (derecha).*

nas. Las capas corticales superiores, en el caso del hombre, poseen una gran proporción de la totalidad de las neuronas [152, 79]. La actividad neuronal sigue, en general, un patrón vertical que da lugar a la formación de columnas que a su vez están relacionadas lateralmente entre sí. Dentro de cada columna, una neurona perteneciente a una capa hace sinapsis directas sobre neuronas de la siguiente capa, o bien indirectamente, a través de interneuronas [79]. Ésto da lugar –teniendo en cuenta los retardos sinápticos– a que una neurona cualquiera de las capas más altas reciba simultáneamente información que fue generada en instantes distintos en la periferia, lo que permite establecer relaciones temporales complejas.

Gracias a las técnicas relativamente recientes de generación de imágenes funcionales como la *resonancia magnética funcional* [10] o la localización de dipolos mediante *potenciales evocados auditivos* (PEA) [16], es posible el estudio no invasivo de algunas funciones corticales en el hombre. Ésto ha permitido la identificación de las zonas que intervienen en el procesamiento del habla. A pesar de ésto solo se conocen unas pocas características organizacionales de la corteza auditiva [153]. Ésta puede dividirse principalmente en dos áreas funcionales: la *corteza auditiva primaria* (AI) y la *corteza auditiva secundaria* (AII) (Ver detalle en Figura 2.20). La zona AI recibe información directa del núcleo geniculado medio y por lo tanto posee un mapa tonotópico preciso [124]. Se puede decir que AI posee un mapa topográfico de la cóclea, por lo que a veces se lo denomina también mapa *cocleotópico*. En disposición ortogonal al mapa tonotópico existe una organización en bandas de las propiedades biaurales. La zona AII posee una organización tonotópica menos precisa y posiblemente analiza sonidos más complejos. El área de Wernicke (Ver Sección 2.2) se ubica en el interior de AII [124].

Representación cortical

En la Sección 2.4.3 se discutió acerca de la codificación neuronal a nivel del nervio auditivo. En esta sección se pretende introducir algunos conceptos que permitan comprender los aspectos sobresalientes de la codificación más complejos que se da a nivel de la corteza. En particular es de interés conocer como se codifican los rasgos distintivos del habla a nivel de la corteza auditiva. Se cree que el sistema auditivo ha aplicado principios de codificación eficiente para procesar a los sonidos naturales, especialmente el habla. Ésto parece muy razonable si se piensa que éstos son los sonidos "más importantes" de nuestro entorno. La teoría de la información provee conceptos generales que permiten abordar el tratamiento de los problemas de comunicación mediante señales. Entre estos conceptos aparece el de eficiencia de la codificación. Hace ya un tiempo que estos prin-

cipios se han tratado de aplicar al código neuronal, pero es más recientemente cuando se ha obtenido cierto éxito [52]. Una versión neuronal de esta hipótesis de eficiencia en la codificación establece que el rol de los sistemas sensoriales "tempranos" es remover la redundancia estadística o aumentar la independencia entre las respuestas neuronales a estímulos naturales. A esta hipótesis suele agregarse otra que asegura que estos sistemas tienden a crear representaciones internas sumamente ralas, es decir teniendo en cuenta una cantidad importante de rasgos significativos de manera explícita (esto tiene su correlato en la sobre-representación de características a nivel cortical). De esta forma el cerebro crea un código eficiente mediante una representación rala e independiente de la señal, consistente principalmente en detectores de cambios en los picos espectrales y en los parámetros temporales (representaciones tiempo-frecuencia). Para llegar a validar estas hipótesis un posible camino consiste en armar un modelo sensorial que se base en ellas y tratar de contrastar las predicciones realizadas mediante este modelo con las respuestas reales. Entre las predicciones que han logrado validarse mediante estos modelos se puede mencionar la representación sensorial interna a nivel cortical a partir de los denominados *campos receptivos espectro-temporales* (STRF)[16].

Campos receptivos espectro-temporales Como se mencionó anteriormente el enfoque tradicional para caracterizar la respuesta a nivel cortical basada en la utilización de tonos puros es inaplicable para un sistema como éste. Para que ésto funcione adecuadamente el sistema, con entrada a nivel sensorial y salida en la corteza, debería ser lineal e invariante en el tiempo. Por ello la respuesta frente a tonos puros constituye solo una primera aproximación al problema. A pesar de ello la mayoría de los estudios y experimentos tradicionales utilizan este tipo de estímulos (incluyendo por supuestos aquellos que permitieron caracterizar las diversas organizaciones tonotópicas) [153]. Ésto se agrava si se tiene en cuenta que la no-linealidad intrínseca de todo este sistema no es un mero accidente de la implementación biológica, sino que constituye un aspecto fundamental que le otorga características funcionales especiales (como su robustez al ruido, entre otras) [150]. La mayoría de las neuronas sensoriales de los niveles superiores poseen respuestas no lineales con propiedades complejas por lo que la caracterización completa de las mismas constituye un desafío importante aún sin resolver. Varios estudios recientes utilizando estímulos complejos combinados con análisis lineal y no-lineal han provisto una nueva visión acerca de las propiedades de estas respuestas en varios

[16]Esta predicción se ha validado inicialmente para el sentido de la visión y más recientemente para el caso de la audición.