Tesis doctoral

Una aportación al procesado neuromórfico de audio basado en modelos pulsantes. Desde la cóclea a la percepción auditiva.

Mª Lourdes Miró Amarante

Sevilla, mayo de 2013

Departamento de Arquitectura y Tecnología de Computadores
Escuela Técnica Superior de Ingeniería Informática
Universidad de Sevilla

Una aportación al procesado neuromórfico de audio basado en modelos pulsantes. Desde la cóclea a la percepción auditiva.

por

Mª Lourdes Miró Amarante

PROPUESTA DE TESIS DOCTORAL
PARA LA OBTENCIÓN DEL GRADO DE
DOCTOR INGENIERO EN INFORMÁTICA
SEVILLA, MAYO 2013

Directores: **Dr. Francisco de Asís Gómez Rodríguez**

y Dr. Ángel Fco. Jiménez Fernández

UNIVERSIDAD DE SEVILLA

Memoria presentada para optar al grado de Doctor Ingeniero en Informática por la Universidad de Sevilla

Autora: **Mª Lourdes Miró Amarante**

Título: **Una aportación al procesado neuromórfico de audio basado en modelos pulsantes. Desde la cóclea a la percepción auditiva.**

Departamento: **Departamento de Arquitectura y Tecnología de Computadores**

Vº Bº Director _____

Francisco de Asís Gómez Rodríguez

Vº Bº Director _____

Ángel Fco. Jiménez Fernández

La autora _____

Mª Lourdes Miró Amarante

A Sergio, Sofía y Laura.

Agradecimientos

Gracias papá y mamá por vuestro amor, paciencia y tiempo.

Gracias Paco por recordarme cada día que todo es posible.

Gracias a todos por haber creído en mí.

Índice

Listado de figuras

Listado de tablas

Capítulo 1

Introducción

En el mundo animal encontramos un gran número de especies que poseen el sentido del oído más desarrollado que los seres humanos. Mientras que el hombre percibe sonidos en el rango de frecuencias de 20 Hz a 20 kHz, los perros captan sonidos entre 40 y 46 kHz; los elefantes advierten incluso los infrasonidos, permitiéndoles comunicarse entre sí a una distancia superior a 4 Km; el delfín y el murciélago han evolucionado hasta hacer del sonido una forma de ver el mundo que los rodea, ambas especies emiten vibraciones sonoras que generan un eco al rebotar contra cuerpos sólidos, lo cual les permite localizar a sus presas. La Naturaleza se convierte en este sentido en una fuente de inspiración para los científicos e ingenieros que aspiran a construir sistemas complejos de procesado del sonido, con el fin de percibir e interactuar con su entorno.

Los sistemas neuroinspirados en el mundo animal se caracterizan por emular propiedades básicas del procesado sensorial. Así tenemos sistemas de visión y olfativos, entre otros, que realizan funciones básicas con propiedades muy parecidas a las que podemos encontrar en la naturaleza. Se podría utilizar el procesado desarrollado en esta tesis para reconocer sonidos y caracterizarlos de forma particular, sin embargo, en este caso se pretende avanzar un poco más hacia niveles cognitivos superiores más propio del ser humano. Este trabajo está enfocado al procesado neuromórfico[1] de la señal acústica, con el objeto de obtener un eficiente sistema automático de reconocimiento del habla. Se han unido dos áreas de investigación: por un lado, las investigaciones relacionadas con el desarrollo de modelos computacionales de la fisiología auditiva aplicados a cócleas artificiales, y por otro lado, las investigaciones sobre el reconocimiento automático del habla.

Actualmente, los reconocedores de voz utilizan directamente la señal percibida por micrófonos, sin aplicarles las transformaciones no lineales que se producen en el oído humano. La mayoría de los sistemas de reconocimiento automático del habla se basan en algoritmos que incluyen una serie de métodos para filtrar las señales de ruido, pero con resultados a día de hoy bastantes limitados. Este trabajo presenta una aportación a un nuevo enfoque de los sistemas de reconocimiento, que aplican un modelo del comportamiento del oído humano en una fase previa al reconocimiento del habla, para mejorar significativamente su calidad. El proceso de la propagación del sonido en el oído interno así como la conversión de la energía acústica en representaciones neuronales se puede modelar como un banco de filtros paso-banda, cuyas frecuencias de corte cubren todo el rango de frecuencias de la voz humana, de 20 Hz a 20 kHz. En este trabajo se ha desarrollado una cóclea artificial neuromórfica basada en una arquitectura paralela de filtros digitales y generadores de pulsos, capaz

[1] Los sistemas neuromórficos son aquellos que tratan de imitar el funcionamiento de los sistemas neuronales biológicos, copiando de ellos su estructura interna y tratando de resolver los mismos problemas. Se describe en el capítulo 2.

de convertir la señal de audio en pulsos, que se van a transmitir usando el protocolo de comunicación neuromórfico AER[2].

Además se han desarrollado tres modelos de neuronas pulsantes encargadas de todo el proceso de reconocimiento: las neuronas de reconocimiento de patrones, las neuronas ganadoras y las neuronas de retraso. En el sistema propuesto, a partir de la información pulsante de la cóclea neuromórfica, las neuronas de reconocimiento de patrones identifican a cada una de los cinco fonemas vocálicos de la lengua española. Para definir el patrón de reconocimiento de cada fonema vocálico se usa el valor del primer y segundo formante de la señal de voz, claves acústicas para la descripción y clasificación de las vocales[3]. Las neuronas ganadoras contribuyen a mejorar la eficacia del proceso de reconocimiento, potenciando la respuesta válida y debilitando otras que aparecen en menor proporción. Con la información de los fonemas vocálicos identificados se inicia la etapa de reconocimiento de palabras bisílabas; en este caso para cada neurona de reconocimiento de patrones se ha definido un patrón en función de las dos vocales presentes en cada palabra bisílaba. La información de los dos fonemas vocálicos reconocidos llega al mismo tiempo a la neurona de reconocimiento, gracias a la acción de las neuronas de retrasos que consiguen la simultaneidad en el tiempo de sucesos temporalmente separados.

Aunque, en este trabajo se presenta un sistema de reconocimiento de palabras bisílabas de la lengua española, el diseño modular de estos tres modelos de neuronas pulsantes permite la construcción de otros sistemas de reconocimiento, capaces de identificar un conjunto arbitrario de fonemas, palabras y frases de cualquier lengua.

Para el sistema completo se ha realizado una implementación hardware en una FPGA, permitiendo el diseño de un sistema neuromórfico caracterizado por una

---

[2] AER, siglas en inglés de *Address Event Representation*. Es un mecanismo neuromórfico para la transmisión de información entre sistemas neuronales artificiales (descrito en el capítulo 2).

[3] La caracterización de la señal de voz se describe en el capítulo 3.

respuesta en tiempo real, un procesado paralelo, y un bajo consumo energético, propiedades inherentes a los sistemas biológicos.

Con el trabajo realizado en esta tesis se ha pretendido construir una base que abre las puertas a un nuevo enfoque en el proceso de reconocimiento del habla, en particular, y al procesado de la señal de audio, en general. Se han desarrollado diferentes elementos basados en el paradigma neuromórfico con el fin último de facilitar la construcción de nuevos sistemas neuromórficos complejos de procesado de audio, con el objetivo de imitar la estructura y el funcionamiento del sistema auditivo humano y de este modo investigar y aprender más sobre cómo el cerebro procesa el sonido; principal objetivo de la ingeniería neuromórfica.

1.1. Motivaciones

Desde los años 80's en el que Carver Mead plantea las bases de lo que hoy se conoce como ingeniería neuromórfica, han sido muchos los investigadores que se han inspirado en el conocimiento de la biología para el desarrollo de diferentes sistemas con el objetivo de mejorar su rendimiento. Con esta premisa, la autora de este trabajo, dentro del grupo de investigación de Robótica y Tecnología de Computadores de la Universidad de Sevilla (RTC-US), ha participado en los siguientes proyectos de investigación relacionados con el desarrollo de sistemas neuromórficos para el procesamiento visual: el proyecto nacional SAMANTA II: Sistema de visión multi-chip Address-Event-Representation para plataforma Robótica II (TEC2006-11730-C03-02) y el proyecto nacional VULCANO: Visión Ultra-Rápida por eventos y sin fotogramas. Aplicación a automoción y Robótica cognitiva antropomorfa (TEC2009-10639-C04-02). En ambos proyectos, se han aplicado los principios de la biología al campo de la robótica, con el objetivo de construir sistemas que se relacionen con su entorno a través de diferentes sensores y

actuadores, con la mínima o ninguna intervención del hombre. Además, los diferentes sistemas neuromórficos desarrollados en estos proyectos, usan una representación basada en pulsos (AER) que facilita la implementación de una comunicación asíncrona entre los diferentes circuitos del sistema.

Hasta la fecha, la construcción de este tipo de sistemas ha requerido la intervención de un computador convencional, como por ejemplo un PC, encargado fundamentalmente del procesado de la información capturada por diferentes sensores y del envío de órdenes a los diferentes actuadores. Esto ha impedido la total autonomía de las diferentes plataformas de robots. Por esta razón, en nuestro grupo de investigación se trabaja con FPGAs para el desarrollo de sistemas neuromórficos, tanto sensoriales como procesadores de la información.

Gracias a los conocimientos adquiridos durante estos años de investigación, ha sido posible el desarrollo de un sistema neuromórfico para el procesado de audio basado en modelos pulsantes que se presenta en esta tesis.

El principal motivo para la elección de la temática de esta tesis ha sido la creación de un nuevo sistema de sensado y procesado de la señal de voz, basado en el paradigma neuromórfico, con el fin último de que este nuevo sistema sirva como plataforma para el estudio del comportamiento del cerebro ante un sonido.

Un segundo motivo ha sido el desarrollar un sistema neuromórfico de sensado auditivo capaz de integrarse mediante fusión sensorial en un sistema neuromórfico más complejo; uno de los objetivos que se pretenden alcanzar en el actual proyecto de investigación BIOSENSE: Sistema Bioinspirado de Fusión Sensorial y Procesamiento Neurocortical Basado en Eventos. Aplicaciones de Alta Velocidad y Bajo Coste en Robótica y Automoción (TEC2012-37868-C04-02). Hasta ahora, los esfuerzos de investigación del grupo estaban orientados al campo visual y motor; con este trabajo se pretende aportar el diseño e implementación de un sistema

neuromórfico basado en modelos pulsantes, que integre tanto el sensor de sonido como el sistema para el procesado de la información auditiva.

Y por último, destaco mi inquietud por diseñar e implementar un sistema completo de sensado y procesado de información codificada en pulsos sobre una FPGA, sin necesidad de recurrir al cómputo realizado por un PC, permitiendo así construir un sistema de bajo consumo, bajo coste y en tiempo real.

1.2. Objetivos

El objetivo principal de esta tesis es abordar un nuevo sistema de procesado neuromórfico de audio basado en la representación pulsante de la información. Para ello se pretende desarrollar un nuevo sensor neuromórfico de audio, que imite la funcionalidad de la cóclea biológica así como la estructura y funcionalidad del sistema nervioso para la transmisión de la información; junto con un nuevo mecanismo de reconocimiento del habla basado en modelos pulsantes. También es objetivo de esta tesis implementar y probar con experimentos reales el nuevo sistema que se propone para verificar la viabilidad del mismo. Para ello se ha elegido una plataforma hardware basada en una FPGA, con el fin de obtener un sistema de bajo coste, bajo consumo y capaz de realizar un procesado paralelo en tiempo real.

A continuación se enumeran los objetivos que se persigue en esta tesis, distinguiendo entre objetivos generales y específicos.

Objetivos generales

La tesis que se presenta, pretende desarrollarse dentro del marco científico de la Ingeniería Neuromórfica. Por tanto, los objetivos generales de la misma están centrados tanto en estudiar las posibilidades que tiene los sistemas neuromórficos

para el procesamiento de audio; como en la posibilidad de obtener un mayor conocimiento sobre los procesos cognitivos relacionados con el procesamiento del habla. Estos objetivos son:

- Analizar la posibilidad de implementar sistemas neuromórficos para el procesado de audio.
- Aportar nuevas evidencias a las ventajas de usar la representación de la información en pulsos.
- Y estudiar la viabilidad de uso de sistemas digitales en la construcción de sistemas neuromórficos, campo dominado por la electrónica analógica en los últimos tiempos.

Objetivos específicos

Para conseguir estos objetivos generales, tan amplios y ambiciosos, se han fijado un conjunto de objetivos específicos, más concretos y realistas en su ejecución. Ellos son:

- Estudiar y diseñar un nuevo modelo de cóclea neuromórfica, basada en una representación en pulsos de la información.
- Estudiar y crear nuevos modelos neuronales artificiales, que a partir de la información en pulsos de una cóclea artificial pulsante, sea capaz de reconocer un fonema vocálico de la lengua española.
- Diseñar una arquitectura, basada en estos nuevos modelos de neuronas artificiales pulsantes, para la construcción de un sistema de reconocimiento automático del habla.
- Implementar estos nuevos modelos sobre una FPGA, con los siguientes requisitos:

  - La implementación no debe incluir ningún computador convencional en el núcleo del procesado.
  - La implementación debe ser realista y realizable, modular y que permita demostrar empíricamente la viabilidad de la construcción de este nuevo sistema neuromórfico, que incluye tanto el sistema de sensado como el de procesado de audio.
- Caracterizar los nuevos modelos desarrollados a partir de pruebas y experimentos sobre estímulos reales.

1.3. Estructura de la tesis

Esta memoria se ha estructurado en cinco partes que se detallan a continuación, con los capítulos que contiene cada una de ellas.

Parte I. Introducción. Presenta todo el documento y contiene el capítulo actual.

Capítulo 1. Introducción. Es el capítulo actual, en el que se presentan las motivaciones, los objetivos y la estructura del documento.

Parte II. Estado del arte. Se hace una exposición de las diferentes materias en las que está centrada esta investigación así como aquellas que son necesarias para entender su desarrollo. Se divide en los siguientes capítulos:

Capítulo 2. Estado de los desarrollos neuromórficos actuales. Se hace un repaso del estado de los desarrollos neuromórficos actuales basados en la representación *AER*.

Capítulo 3. Caracterización de la señal de voz. Se describe el proceso de producción y percepción de la señal sonora; así como las características articulatorias, acústicas y perceptivas del habla.

Capítulo 4. Modelos e implementaciones del Sistema Auditivo. Se hace un repaso exhaustivo de los diferentes modelos e implementaciones de cócleas artificiales.

Capítulo 5. Sistemas de reconocimiento automático del habla. Se hace una introducción al proceso de reconocimiento automático del habla, recogiendo los conceptos básicos, detallando las etapas que lo componen y los métodos tradicionales de implementación existentes.

Parte III. Aportación. Contiene la aportación al procesado neuromórfico de audio basado en modelos pulsantes. Se divide en los siguientes capítulos.

Capítulo 6. Cóclea artificial pulsante. Se describe una cóclea neuromórfica basada en filtros digitales.

Capítulo 7. Sistema de reconocimiento pulsante. Se presenta el sistema de reconocimiento propuesto, capaz de identificar palabras a partir del reconocimiento de los fonemas vocálicos de la lengua española.

Capítulo 8. Experimentos. Se describe las pruebas realizadas sobre el sistema final para la evaluación de la robustez, versatilidad y precisión en el reconocimiento. Se concluye el capítulo con una interpretación de los resultados obtenidos.

Parte IV. Conclusiones. Se recogen las conclusiones de esta investigación. Está formado por un único capítulo.

Capítulo 9. Conclusiones y trabajos futuros. Las aportaciones y las conclusiones a las que se ha llegado en el desarrollo de esta tesis se recogen en este capítulo, así como las líneas de trabajo futuras.

Parte V. Bibliografía y anexos. La última parte del documento contiene las referencias utilizadas en esta investigación, así como los scripts de Matlab usados en las pruebas y experimentos.

Capítulo 2

Estado de los desarrollos neuromórficos actuales

Los seres vivos son sistemas complejos, dotados de una gran variedad de instrumentos de medición, de análisis, de recepción de estímulos y de reacción y respuesta. Los cinco sentidos, los cuales nos conectan con el mundo exterior y a través de los cuales percibimos importante información sobre todo cuanto nos rodea, nos permiten ejercer nuestra capacidad de selección en el proceso de la información. Así, a un ser humano no le cuesta ningún esfuerzo identificar y evaluar las cosas por medio de sus percepciones sensoriales en combinación con su memoria. La vista, el oído, el olfato, el tacto, … trabajando por separado o en combinación constituyen literalmente nuestra conexión con el mundo, una conexión que se erige como el gran misterio a desvelar por la ciencia actual.

Crear sistemas computacionales que se parezcan a cerebros humanos, capacitados para observar un comportamiento inteligente, es el campo de investigación de la robótica y la inteligencia artificial. Dentro de este comportamiento inteligente se encuentran tanto las actividades relacionadas con el raciocinio, es decir, planeamiento y estrategia, como con la percepción y reconocimiento de imágenes, sonidos, olores, etc.

Así, las llamadas tecnologías bioinspiradas nacen de la aplicación de conceptos de inspiración biológica al diseño de sistemas analíticos. El objetivo es comprender e imitar la forma en que los sistemas biológicos aprenden y evolucionan. Para diseñar estos sistemas, además de utilizar la computación tradicional numérico-simbólica, se usan otras metodologías tales como las redes neuronales artificiales, la lógica difusa y la computación evolutiva. Por ello, este intento de emulación del funcionamiento de los seres vivos se debe apoyar en un entorno multidisciplinar que agrupa físicos, biólogos, psicólogos, informáticos, electrónicos, microelectrónicos y áreas de la ingeniería, como la biomédica o la neuromórfica, y aspira a conseguir auténticos sistemas electrónicos dotados de sentidos artificiales que permitan facilitar tareas y resolver problemas hasta ahora no resueltos.

En este capítulo, se estudian los principios por los que se rige la ingeniería neuromórfica; se verán diferentes sistemas sensoriales neuromórficos, muchos de ellos basados en la representación por direcciones de eventos, *AER,* también explicada en este capítulo.

2.1. Ingeniería neuromórfica

La historia de la ingeniería neuromórfica se inició en el Instituto de Tecnología de California (*CalTech*) en los años ochenta con el trabajo de Carver Mead, quien se ha

dedicado al estudio de los sistemas biológicos con un innovador planteamiento: "para entender el funcionamiento del ojo o el oído, es muy útil reproducirlos artificialmente". Su aportación, ha consistido en comprender los sistemas neuronales biológicos por medio de su recreación en silicio, lo que ha impulsado el campo del diseño de circuitos neuromórficos analógicos, (Mead, 1990).

La ingeniería neuromórfica se define como un campo de investigación multidisciplinar dedicado al diseño y a la fabricación de sistemas artificiales de computación, cuyas propiedades físicas, estructuras o representaciones de la información están basadas en el sistema nervioso biológico.

Los sistemas biológicos sobrepasan a cualquier sistema de percepción hecho por el hombre en la forma en que realizan el procesamiento de la información y en el poco consumo de potencia necesario para su funcionamiento. Es por esto, que el diseño de sistemas de procesamiento bioinspirados es una alternativa en el que se puede lograr un mejor resultado en el consumo de potencia, la velocidad de procesamiento y área utilizada, en comparación con los sistemas y técnicas tradicionales.

Los ingenieros neuromórficos siguen los siguientes principios de diseño que han sido tomados de la biología:

1. *Procesamiento paralelo o procesamiento neuronal.* Sin duda, muchas de las asombrosas propiedades de los sistemas naturales son debidas a que son sistemas de computación colectivos, implementando de esta forma modelos masivamente paralelos.
2. *Computación cooperativa.* El modo en que los sistemas biológicos realizan procesamientos robustos y de precisión basándose en unidades (neuronas) de computación imprecisas y estocásticas, con gran nivel de ruido y respuestas impredecibles frente a estímulos similares, constituyen una característica de gran interés.

3. *Capacidad de auto-configuración.* Esta característica presente en los sistemas biológicos es de gran utilidad para adaptar sistemas de computación genéricos a distintas funciones. Además esta propiedad aumenta la efectividad de los sistemas a entornos específicos de cada individuo y permite la adaptación a nuevas condiciones durante su tiempo de vida.
4. *Pre-procesado de la información para incrementar el rango dinámico.* Esta propiedad es posible gracias a los nuevos sensores adaptativos.
5. *Representación de la señal a través de eventos discretos (spikes).* Esta característica facilita una comunicación robusta y eficiente. Además los sistemas pulsantes combinan de un modo más eficiente las características fundamentales de los sistemas analógicos y digitales.

Por tanto, cuando se desarrollan sistemas de computación bioinspirados es importante identificar las características de los sistemas biológicos en las que se basa su capacidad de computación y que pueden ser adaptadas a la tecnología de implementación mediante circuitos electrónicos relativamente sencillos. Además, debe aprovechar características inherentes a los circuitos electrónicos como su mayor ancho de banda, velocidad de respuesta, etc.; y utilizar esquemas como multiplexación temporal, comunicación binaria, comunicación mediante direcciones, etc. La combinación óptima de principios bioinspirados y características explotables de los circuitos electrónicos constituye la esencia de la ingeniería neuromórfica.

En el campo de la ingeniería neuromórfica hay que destacar el *Institute of Neuromorphic Engineering* (INE, 2012), responsable de la organización de uno de los eventos anuales de esta disciplina más importantes, el *Telluride Neuromorphic Cognition Engineering Workshop* (TellurideWorkshop, 2012), que se celebra en Telluride en el estado de Colorado (Estados Unidos) desde mediados de los 90; y que tiene un reflejo en Europa en el *CappoCaccia Neuromorphic Cognition Engineering Workshop* (CappoCacciaWorkshop, 2012). En ambos eventos se dan cita los investigadores

más importantes en la ingeniería neuromórfica, así como una gran cantidad de estudiantes que inician sus pasos en dicho campo científico-técnico. También, es interesante destacar una publicación especializada en este campo, *Frontiers in Neuromorphic Engineering*[4].

2.2. Sistema sensorial neuromórfico

Se ha descrito como la ingeniería neuromórfica se caracteriza por la emulación de funciones biológicas muy específicas, generalmente de tipo sensorial, cuya estructura y funcionalidad biológica han sido estudiadas con gran detalle, dando lugar a modelos, que han propiciado la construcción de sistemas neuromórficos.

Una implementación de un sistema neuronal es definida como neuromórfica si cumple fundamentalmente el requisito de tiempo real. En este sentido, la mayoría de los sistemas neuromórficos son etiquetados como sistemas sensoriales. Así, nos encontramos con una gran cantidad de ejemplos de sistemas neuromórficos en el campo visual, auditivo, olfativo y táctil. Otra característica básica de los sistemas neuromórficos es el procesado paralelo, el cual aporta ventajas a los sistemas sensoriales.

2.2.1. Sistema neuromórfico para el procesado visual

Mead y Mahowald (Mead & Mahowald, 1988) describen uno de los primeros diseños de una retina artificial. Es una retina de 48x48 píxeles[5], cuya salida es la diferencia entre la intensidad del pixel central y la media ponderada de las intensidades de sus

[4] Revista online de acceso libre sobre ingeniería neuromórfica:
http://www.frontiersin.org/neuromorphic_engineering

[5] Pixel, acrónimo del inglés *picture elemnent*: elemento de una imagen.

píxeles vecinos; comportamiento diferente al de una cámara digital. El efecto final es que la respuesta a un borde estático es una derivada en el espacio, tal y como ocurre en una retina biológica. Además, la respuesta a una superficie sin variaciones (fija) es cero, independientemente del brillo.

Posteriores investigadores han usado esta misma idea para el desarrollo de fotoreceptores (Delbrück & Mead, 1996) y retinas (Andreou, Meitzler, Strohbehn, & Boahen, 1995) más eficientes. Además, surgen sistemas basados en estas retinas artificiales capaces de modelar diferentes capacidades visuales como son la detección de colisiones (Indiveri, 1998), detector de movimiento de una mosca (Harrison & Koch, 1998), obtención de la profundidad a partir del movimiento (Yang, Murray, Woergoetter, Cameron, & Boonsobhak, 2006), y más recientemente la estimación de la distancia de un objeto en movimiento por disparidad entre retinas (Domínguez-Morales et al., 2011). En (Lichtsteiner, Posch, & Delbruck, 2008) se presenta un sistema basado en la detección de cambios en la luminosidad, los cuales son transmitidos en serie usando el protocolo de comunicación *AER*. Esto permite la detección rápida de alteraciones en una escena sin la necesidad de procesar y, por tanto, sin enviar todos los fotogramas de la escena.

En la Figura 1 se muestra una foto de la retina artificial a); una microfotografía del chip que implementa la retina b); un esquema básico del circuito de cada uno de los píxeles de la retina c) formado por el foto-receptor, el circuito diferenciador y los comparadores que van a permitir emitir pulsos positivos o negativos dependiendo del sentido del cambio en la luminosidad; y por último se muestra el principio de funcionamiento de cada uno de los píxeles d), en el que dependiendo del sentido de cambio en la luminosidad y de unos umbrales previamente establecidos se producen pulsos positivos o negativos.

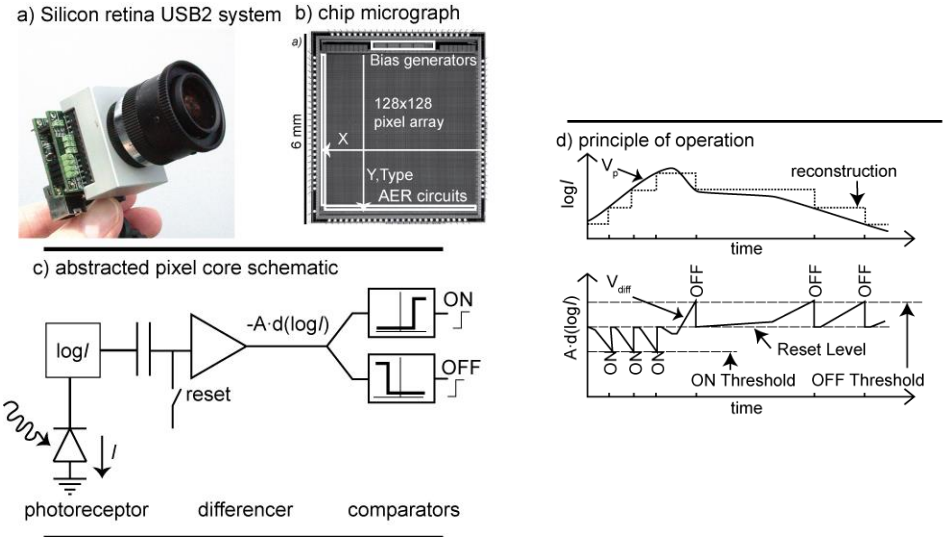
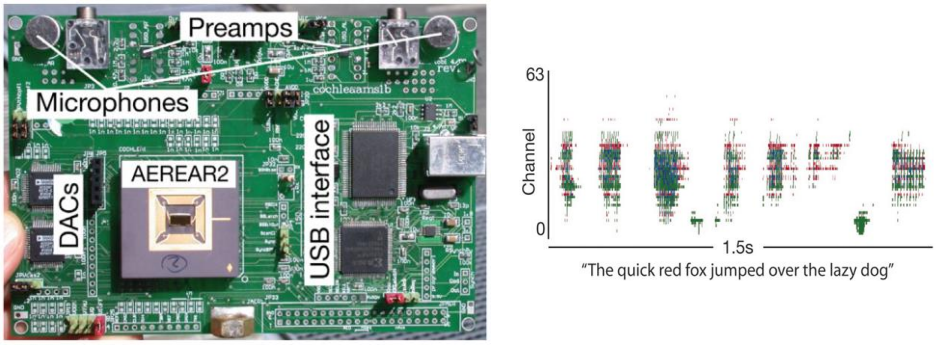


Figura 1. Retina artificial diferencial espacio-temporal.
Imagen tomada de (Lichtsteiner et al., 2008).

2.2.2. Sistema neuromórfico para el procesado del sonido

A partir de los trabajos realizados por Helmholtz y von Bekesy sobre el proceso de transducción de la onda viajera en la cóclea, surge un gran interés por los modelos de cócleas artificiales. Uno de ellos, el propuesto por Lyon y Mead (Lyon & Mead, 1988), basado en una secuencia de filtros, será utilizado en posteriores desarrollos (Lazzaro & Mead, 1989a) y aplicaciones como la localización del sonido (Lazzaro & Mead, 1989b), (Schauer & Paschke, 1999) y (V. Y. Chan, Jin, & Schaik, 2010); identificación del tono de voz (Lazzaro & Mead, 1989c) y (Yu, Schwartz, Harris, Slaney, & Liu, 2009); identificación del hablante (Li, Delbruck, & Liu, 2012); análisis de voz (W. Liu, Andreou, & Goldstein, 1993a), (W. Liu, Andreou, & Goldstein, 1993b) y (Lazzaro & Wawrzynek, 1997); y detector del ritmo (Gómez-Rodríguez et al., 2007).

En la siguiente Figura 2 se muestra un ejemplo de implementación de una cóclea artificial analógica desarrollada por Liu, S.C (S.-C. Liu, v. Schaik, Minch, & Delbruck, 2010). En la placa de prototipo se destaca el chip que implementa el par de cócleas analógicas artificiales de 64 canales, la interfaz USB, y la interfaz con dos micrófonos, (a). A la derecha, se representa la salida del sistema en forma de cocleograma[6], como respuesta a la señal de voz "*The quick red fox jumped over the lazy dog*". Los dos colores, rojo y verde, corresponden a los 64 canales de cada cóclea artificial, que modelan el oído izquierdo y derecho, (b).



(a) Placa prototipo que incorpora el chip *AEREAR2.*

(b) Respuesta a la señal de voz *"The quick red fox jumped over the lazy dog".*

Figura 2. Cóclea artificial AER EAR2. Imagen tomada de (S.-C. Liu et al., 2010).

Hay que tener en cuenta que el proceso de audición no sólo implica el proceso mecánico de transducción de la onda viajera a impulsos nerviosos, también hay que considerar el proceso que tiene lugar en el cerebro. Así, podemos destacar diversos

[6] Un cocleograma es usado para representar en cada instante de tiempo el ritmo de disparo de cada nervio auditivo que sale de la cóclea. Es por tanto, una representación acústica que informa sobre la energía de las diferentes frecuencias de la señal de entrada.

trabajos que modelan el núcleo coclear, (v. Schaik, Fragnière, & Vittoz, 1996), (v. Schaik & Vittoz, 1997) y (Glover, Hamilton, & Smith, 2002).

En los capítulos 3 y 4 de esta tesis se describe con más detalle todo el proceso de audición humano y diferentes modelos e implementaciones de cócleas artificiales, respectivamente.

2.2.3. Otros sistemas sensoriales neuromórficos

También existe un interés por el sentido del olfato: la nariz electrónica es un dispositivo utilizado en diversas industrias, como la perfumería y la fabricación de cerveza, capaz de detectar cambios eléctricos a partir de las moléculas del aire. En (Koickal et al., 2005) se describe una implementación basada en pulsos.

Existen sistemas que emulan el sentido del tacto, los cuales traducen la presión o el movimiento en señales eléctricas. Estos sistemas neuromórficos están basados en el movimiento de los bigotes de un ratón (Pearson et al., 2006), (Argyrakis et al., 2007). Además, existen desarrollos de conjuntos de sensores, como existen en la piel (Roy, 2006) y (Vasarhelyi et al., 2006).

2.2.4. Fusión sensorial

El estudio de la fusión sensorial en el campo de la ingeniería neuromórfica permite probar los principios aprendidos desde la biología, al mismo tiempo que nos da pistas sobre cómo el cerebro realiza esta fusión.

Un sistema sensorial que combina diferentes modalidades sensoriales puede operar en una amplia variedad de entornos beneficiándose de las ventajas que aportan cada uno de los diferentes sentidos de un modo individual o en colaboración con los otros sentidos.

Aunque la fusión sensorial audio-visual ha sido materia de estudio en el campo de la robótica a lo largo del tiempo (Bothe, Persson, Biel, & Rosenholm, 1999) y (W. K. Wong, Neoh, Loo, & Ong, 2008), no encontramos muchos ejemplos de sistemas neuromórficos que combinan sensores de diferentes modalidades.

Existen muchos ejemplos de sistemas sensoriales y robots que incorporan sensores neuromórficos, pero la mayoría de ellos se limitan al uso de un único tipo de sensor neuromórfico (Gómez-Rodríguez et al., 2007), (Linares-Barranco, Gomez-Rodriguez, Jimenez-Fernandez, Delbruck, & Lichtensteiner, 2007) y (Angel Jimenez-Fernandez, Lujan-Martinez, Paz-Viecente, Jimenez, & Civit, 2009). Existen sistemas que combinan múltiples sensores de una única modalidad, por ejemplo de visión (Becanovic, Hosseiny, & Indiveri, 2004).

En (V. Chan, Jin, & v. Schaik, 2012) se describe un ejemplo de fusión sensorial basado en sensores neuromórficos de diferentes modalidades para la localización de la fuente de un sonido. Este sistema combina dos cócleas artificiales y una retina artificial, Figura 3.

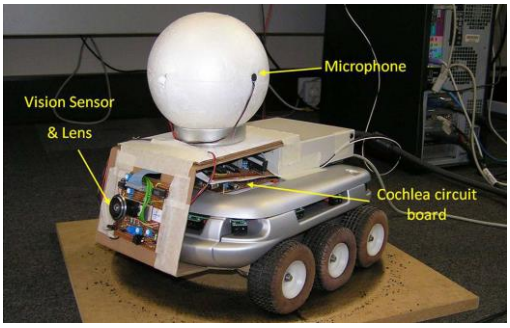


Figura 3. El robot Koala. Imagen tomada de
(V. Chan et al., 2012).

2.3. Bus neuromórfico AER

La habilidad del cerebro humano para procesar la información en paralelo y de un modo distribuido, es posible gracias a su masiva arquitectura de conexiones. En los sistemas neuronales biológicos, el procesamiento está estructurado en capas de neuronas, donde cada neurona de una capa está conectada a un 'campo proyectivo' de neuronas en las siguientes capas. Los pesos de las conexiones obedecen a un cierto patrón, por lo que es como si se implementara un *kernel*[7] de convolución, realizándose una operación de convolución al transmitir la información de una capa a otra sucesiva (la información de una neurona llega ponderada a un vecindario de neuronas de la siguiente capa). Las neuronas se comunican mediante pulsos con constantes de tiempo del orden de los milisegundos, estando cada neurona conectada a varios miles, (Kandel, Schwartz, & Jessell, 2000).

La mayoría de los sistemas neuromórficos están formados por uno o más sensores neuromórficos y una red de neuronas artificiales pulsantes, que tratan de imitar la interconexión de las neuronas biológicas. Desafortunadamente, esta extensiva conectividad del cerebro es imposible implementarla directamente en sistemas *VLSI*[8] debido a las limitaciones físicas de conectividad dentro y entre microchips. Sin embargo, hay que señalar que el tiempo de transición de las tecnologías actuales *VLSI* son del orden de nanosegundos, es decir, un millón de veces más rápido que en el caso de las neuronas biológicas. Basados en esta diferencia temporal y en un alto ancho de banda de los sistemas *VLSI*, se propone, como solución al problema de la conectividad, un sistema de multiplexación en el tiempo sobre un mismo canal, a través del cual se transmite un identificador para cada neurona que emite. A este

---

[7] *Kernel* en inglés núcleo; es el nombre que recibe la matriz usada en la operación matemática de convolución.

[8] *VLSI*, siglas en inglés de *Very Large Scale Integration;* muy alta escala de integración.

esquema se le denomina *Address-Event Representation*, *AER*, (Sivilotti, 1991), (Mahowald, 1992), (Mahowald, 1994) y (K. a. Boahen, 2000), (Morgado Estévez, 2003).

2.3.1. Representación *Address Event*

La representación *address event* (*AER*) es un protocolo de comunicación conducido por eventos (*event-driven*) usado originalmente en implementaciones *VLSI* de redes neuronales para transferir pulsos (*action potentials*) entre neuronas (Mahowald, 1992).

La representación *AER* fue usada primero como un acercamiento a la masiva conectividad de las redes neuronales biológicas, aunque, en general es adecuado para transportar un gran número de valores analógicos, codificados en frecuencia de eventos, a través de un canal de menor capacidad (bus digital asíncrono). Es, por tanto, una técnica de multiplexación digital asíncrona. En principio, no se hace ningún compromiso sobre los detalles de las unidades involucradas: ellas pueden ser circuitos integrados neuromórficos, circuitos digitales, emulaciones software de sistemas neuronales, neuronas biológicas o programas de ordenador estándares.

Su aplicación clásica, ilustrada en la Figura 4, es entre dos circuitos electrónicos, uno emisor y otro receptor. En este sistema *AER*, un conjunto de neuronas (emisor) codifica su actividad en la forma de pulsos que son transmitidos a otro conjunto de neuronas (receptor). Cada vez que una celda emisora genera un pulso, mediante un codificador y un sistema de multiplexación, se escribe su dirección en un bus compartido por todas las celdas, llamado bus *AER*.

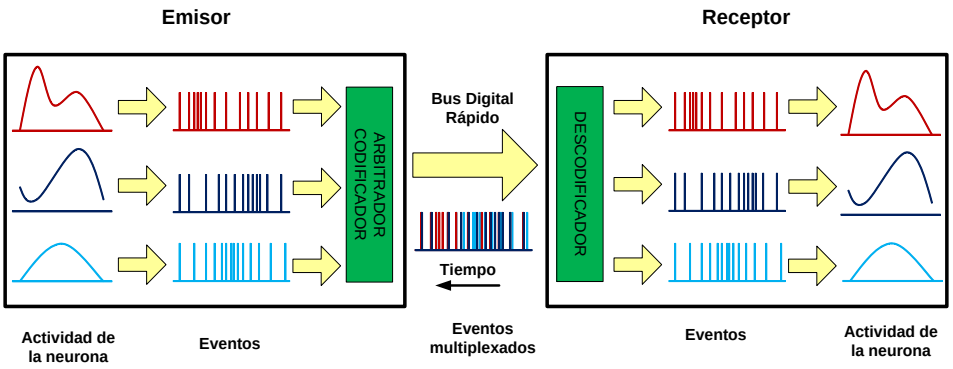


Figura 4. Esquema de comunicación *Address Event Representation* (*AER*).

En el esquema anterior la actividad de la neurona es codificada en frecuencia de pulsos, pero además de la dirección, es posible enviar de forma explícita otra información asociada a la actividad; o bien codificar la información en el tiempo entre dos únicos eventos, teniendo así un esquema de codificación en pulsos (que no en frecuencia de pulsos), comportamiento que parece muy cercano a la forma de trabajo de las neuronas biológicas.

En su formulación original, *AER* implementa una topología de conexión uno-a-uno (muy apropiada para emular los nervios de visión y audio). Si se considera el enfoque de 'fuerza-bruta', para comunicar estas celdas se necesitaría un cable por cada par de neuronas, requiriendo N cables para N pares de celdas. Sin embargo, un sistema *AER* identifica la posición de la neurona pulsante y la codifica como una dirección, la cual es entonces enviada a través de un bus de datos compartido. El conjunto receptor descodifica la dirección y la enruta a la neurona apropiada, reconstruyendo así la actividad del emisor. Este esquema reduce el número de cables requeridos de N a $\log_2 N$.

Debido a que las neuronas operan de forma concurrente, podría darse el caso de que dos o más neuronas emitieran eventos al mismo tiempo; un arbitrador es el

encargado de resolver estos conflictos, preservando en la medida de lo posible la correcta información temporal codificada en los trenes de pulso. De esta manera, toda la información de todas las neuronas es combinada en el tiempo y enviada al circuito receptor. En el receptor los eventos son descodificados y separados, y posteriormente enviado a la neurona correspondiente.

Ventajas de la representación *AER*

La representación *AER* está diseñada para proporcionar una comunicación de alto ancho de banda entre grandes conjuntos de elementos neuronales. La multiplexación en el tiempo, característica del *AER*, es el único modo de transferir datos desde miles de nodos de salida con las limitaciones de pines de la tecnología existente. La premisa esencial de la representación *AER* es que el ancho de banda del canal estaría dedicado a la transmisión de señales significativas. Lo que supone que las celdas del circuito emisor solo deben enviar información cuando ocurra algo importante. Esto, presenta una diferencia con las técnicas de comunicación tradicionales de escaneo (*scanning*) que requieren que cada nodo sea muestreado continuamente y por tanto su estado es emitido por el bus cada vez. El protocolo *AER*, por tanto, es un protocolo guiado por los datos o guiado por eventos. Así pues sólo aquellas celdas que tienen algo que informar generan eventos que serán enviados por el bus. Por lo que las celdas que varían poco no contribuyen a la carga del bus. Una ventaja más importante de las comunicaciones *address-event* es que se minimiza el *aliasing* temporal transmitiendo solo los eventos cuando ellos ocurren. No es necesario introducir el grado de muestreo inherente en una técnica de escaneado secuencial. Para frecuencia de datos baja, el ancho de banda del bus está completamente dedicado a la correcta transmisión de eventos.

La elección de la representación de información para una comunicación entre los distintos circuitos que forman un sistema neural artificial es crítica, porque determina el modo en que el sistema puede fácilmente operar. Se cree que la elección de esta

representación de la información guiada por eventos y el protocolo *AER* puede conducir al desarrollo de sistemas de neuronas artificiales cuyas estrategias de procesamiento de la información son similares a los del sistema nervioso. El protocolo *AER* se ha usado satisfactoriamente en muchos sistemas neuromórficos tanto sensoriales (S.-C. Liu & Delbruck, 2010), como de procesamiento (Gómez-Rodríguez et al., 2007), (Linares-Barranco et al., 2007), (T. Delbruck & Lichtsteiner, 2007) y de actuación (Angel Jimenez-Fernandez et al., 2009). Estos chips pueden ser fácilmente integrados en otros sistemas más complejos colocándolos en un entorno de diseño *AER*. Por lo que se puede afirmar que el *AER* proporciona una organización unificada para la construcción de sistemas multi-chips (R. Serrano-Gotarredona et al., 2005) y (R. Serrano-Gotarredona et al., 2009).

El uso de direcciones digitales para especificar la identidad de la neurona emisora hace el mapeado de señales pre-sypnápticas sobre destinos pos-sypnápticos extremadamente flexible porque el evento (*address event*) lleva su posición de origen en sí mismo. La representación *AER* garantiza el orden temporal de los eventos, el evento puede ser fácilmente descodificado en cualquier ordenación física sobre el circuito receptor. Alternativamente, el patrón de conectividad se puede cambiar dinámicamente, lo que permite cambiar la estructura de la interconexión de las neuronas artificiales tal y como ocurre en los sistemas biológicos. El mapeado de la entrada a la salida es en sí mismo una operación compleja en el sistema nervioso y es una tarea más fácilmente realizable por un ordenador que cableando mano a mano.

Implementación del protocolo *AER*

El protocolo de comunicación *AER* 0.02 sólo hace referencia a la comunicación unidireccional punto a punto de datos asíncronos desde un emisor S a un receptor C, como se muestra en la Figura 5 (a). La conexión entre S y C se realiza a través de dos tipos de cables: cables de control y cables de datos, (Mahowald, 1993).

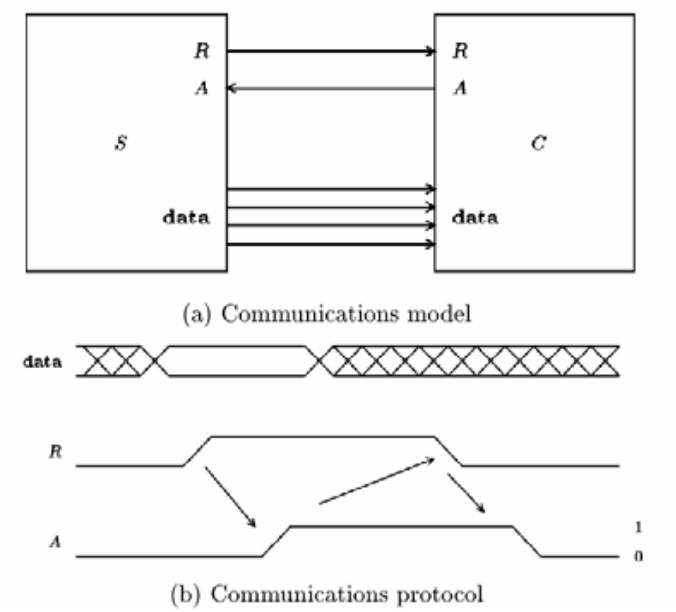


Figura 5. Protocolo *AER*. Imagen tomada de (Mahowald, 1993).

Las líneas de datos son unidireccionales. La codificación de estas líneas no se especifica en *AER* 0.02; el número de líneas, el número de estados de cada línea, etc. son dependientes de la implementación del protocolo. Sólo nos interesa la validez de las líneas de datos, de manera que si S envía una señal estable sobre las líneas de datos y es reconocida por C, las líneas de datos se consideran válidas; en otro caso serán inválidas.

Sin embargo, sí se especifica el comportamiento de las líneas de control: *Request* (R) y *Acknowledge* (A), y la validez de las líneas de datos respecto de las líneas R y A. En la Figura 5 (b) se muestra la secuencia de control para comunicar datos desde S a C.

El protocolo *AER* es un protocolo *handshake* de cuatro fases, Figura 6, entre el emisor y receptor que garantiza la sincronización entre ambos chips; las líneas de

datos comunican la dirección del nodo emisor que solicita la petición al chip receptor. El protocolo permite de forma eficiente que un nodo emisor de un chip comunique pulsos digitales a un nodo receptor. Un chip, el emisor, inicia el proceso con una petición (*request)*. El segundo chip, el receptor, debe contestar a la petición del emisor con un asentimiento (*acknowlegde)*. Para completar la transmisión, el emisor elimina la petición y el receptor el asentimiento. El sistema vuelve así a su estado inicial. Ambas partes están en silencio hasta que algún proceso del emisor inicia otra petición. Así, se dice que el protocolo *addres-event* está conducido por los datos (*data driven*) porque el inicio de la trasmisión depende de los nodos neuronales del emisor que tratan de transmitir un evento. Puesto que las peticiones pueden aparecer en cualquier tiempo desde cualquier nodo, es necesario utilizar un esquema de arbitración para serializar las operaciones del protocolo lo más rápido posible, del orden de nanosegundos.

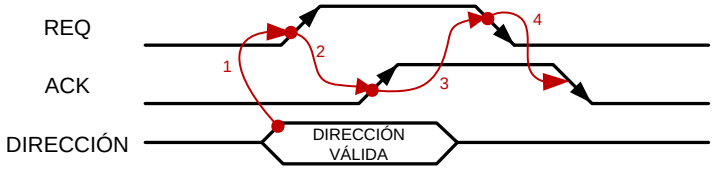


Figura 6. Protocolo *Handshake*

Una trasmisión completa, desde la petición de transmisión hasta la confirmación de la misma es lo que se llama en el entorno del *AER* un evento.

La circuitería necesaria para implementar el protocolo variará según el esquema de conexión. Las primeras implementaciones que se llevaron a cabo fueron la 'punto a punto' y la 'multipunto' ambas unidireccionales (Higgins & Koch, 1999), (K. a. Boahen, 2000) y (Avis, Shihab, Giacomo, & Tim, 2001).

2.3.2. Estado actual de los desarrollos basados en la representación *AER*

A continuación se nombran diferentes grupos de investigación interesados en el *AER*:

- El grupo *Brain in Silicon* de la Universidad de Stanford, liderado por Kwabena Boahen (http://www.stanford.edu/group/brainsinsilicon/).
- El grupo Neuromórfico del IMSE-CNM-CSIC, liderado por Bernabé Linares Barranco (http://www.imse-cnm.csic.es/).
- *The sensors research group* del Instituto de Neuroinformática de la ETH de Zúrich, liderado por Tobias Delbruck (http://sensors.ini.uzh.ch/).
- *The Neuromorphic Cognitive Systems group* del Instituto de Neuroinformática de la ETH de Zúrich, liderado por Giacomo Indiveri (http://ncs.ethz.ch/).
- El *e-Lab* de la Universidad de Yale, liderado por Eugenio Culurciello (https://engineering.purdue.edu/elab/blog/).
- El *Computational NeuroEngineering Lab* (CNEL) de la Universidad de Florida, liderado por John G. Harris (http://www.cnel.ufl.edu/).
- El *Computational Sensory-Motor Systems Lab* de la Universidad Johns Hopkins, liderado por Ralph Etienne-Cummings (http://etienne.ece.jhu.edu/).
- El grupo de Robótica y Tecnología de Computadores de la Universidad de Sevilla, liderado por Antón Cívit Balcells (www.rtc.us.es).

Los desarrollos más importantes que usan *AER* incluyen, módulos de interconexión y testeo, retinas, cócleas, circuitos de convoluciones[9] y redes de aprendizaje.

2.3.3. Uso del *AER* en este trabajo

En esta tesis, el protocolo de comunicación *AER* constituye la base para la transmisión y el procesado de la información auditiva. El protocolo *AER*, usado en este trabajo y en la mayoría de los desarrollos *AER*, fue especificado en el proyecto europeo CAVIAR: *Convolution AER Vision Architecture for Real-Time* (IST-2001-34124) (Caviar Project, 2006). En el documento de especificaciones del *AER* (Häfliger, 2007) se describen las características del conector y el cable (ATA/133 de 40 pines). El significado de cada pin se ilustra en la Figura 7. En la Figura 8 se muestran los parámetros temporales del protocolo *AER*.

[9] Los circuitos de convoluciones son circuitos capaces de realizar la operación matemática de convolución de matrices; técnica matemática muy usada en tratamiento digital de imágenes, para realizar todo tipo de filtrados.

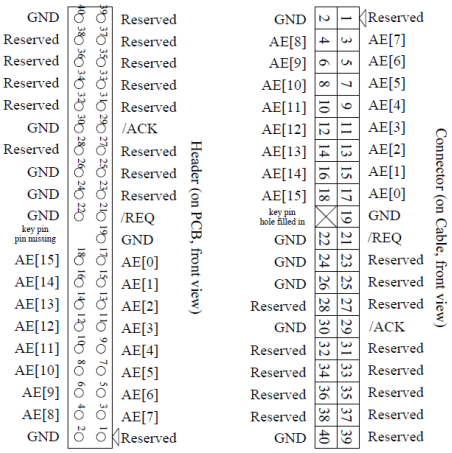


Figura 7. Especificación de pines de los conectores y cables *AER* del proyecto CAVIAR. Imagen tomada de (Häfliger, 2007).

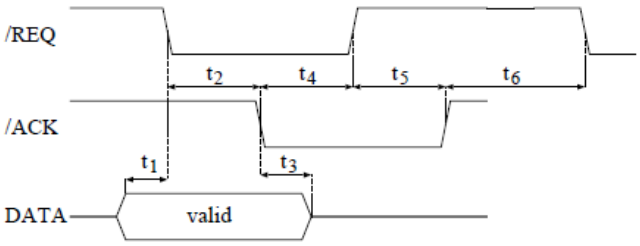


Figure 2: 4 phase handshake. Timing constraints in table 2.

	min	max	avg
t_1	0s	∞	
t_2	0s	∞	$\leq$ 700ns
t_3	0s	∞	
t_4	0s	100ns	
t_5	0s	100ns	
t_6	0s	∞	

Table 2: Timing requirements of 4 phase handshake (figure 2)

Figura 8. Especificaciones temporales del protocolo *AER* del proyecto CAVIAR. Imagen tomada de (Häfliger, 2007).

También hay que destacar el papel desempeñado por la placa *USBAERmini2*, Figura 9, desarrollada por Raphael Berner (R. Berner, 2006), en el marco del proyecto CAVIAR (IST-2001-34124), en el proceso de pruebas y testeo de las diferentes etapas del sistema implementado.

Gracias a esta placa y a los diferentes scripts de *Matlab*, desarrollados para dicha placa, se ha podido capturar de un modo sincronizado todos los eventos *AER* que se han transmitido a través de un bus *AER*. De esta forma se ha obtenido en tiempo real la información sobre la dirección y el instante de tiempo en que se produjo cada evento *AER*. Esta información ha sido fundamental para el diseño, implementación y configuración de los diferentes componentes del sistema.

Figura 9. Placa *USBAERmini2*.

La placa ofrece 4 modos de operación, según se utilice como master o esclavo y se elija entre una resolución temporal de 1μs o 0,033μs, tal como se muestra en la Tabla 1. En nuestro sistema se ha utilizado el modo 1 de operación. Además, es importante resaltar que esta placa sólo reconoce eventos con una diferencia de tiempo inferior a 65ms (*interspike intervals*). Todos estos parámetros han condicionado

la configuración del sistema para evitar la pérdida de eventos y el bloqueo del sistema.

Tabla 1. Modos de operación de la placa *USBAERmini2*.

Mode	**Tick**	**Trigger**
0	1μs	Master (Host)
1	0,033μs	Master (Host)
2	1μs	Slave
3	0,033μs	Slave

Capítulo 3

Caracterización de la señal de voz

En este capítulo se pretende caracterizar la señal que producimos al hablar. Para ello, en primer lugar, se describe cómo el sistema fonador humano es capaz de generar dicha señal; y se hace una breve clasificación desde el punto de vista articulatorio y acústico de los diferentes sonidos que somos capaces de producir. A continuación, se detalla cómo el oído capta, analiza y codifica la información acústica en impulsos nerviosos para que puedan ser procesados por el cerebro. Pero, debido a que el conocimiento sobre el procesado de audio en los centros superiores del cerebro es

muy limitado, es necesario recurrir a la psicoacústica[10] para caracterizar los sonidos desde el punto de vista perceptivo.

3.1. La producción del habla

El sistema de producción del habla no forma parte estricta del sistema sensorial humano, pero su importancia es indudable. Para determinar las operaciones de un sistema de reconocimiento de voz y hablante, es fundamental conocer y determinar los mecanismos que han producido un mensaje hablado. A continuación, se repasan algunos conceptos fundamentales y básicos en el mecanismo de producción del habla, tanto el órgano físico que soporta dichos mecanismos, como la producción propia del mensaje hablado.

3.1.1. Anatomía y Fisiología[11] del aparato fonador humano

El habla, como señal acústica, se produce a partir de las ondas de presión que salen de la boca y las fosas nasales de un locutor. El proceso comienza con la generación de la energía suficiente en los pulmones y la modificación de ese flujo de aire en las cuerdas vocales. A continuación, se encuentra la parte que moldea los sonidos, el tracto vocal formado por las cavidades oral y nasal, que aportan la capacidad de

[10] La psicoacústica es la ciencia que estudia la interconexión entre las propiedades físicas del sonido y la interpretación que el ser humano hace de esas propiedades.

[11] Concepto de anatomía humana: ciencia de carácter práctico y morfológico dedicada al estudio de las estructuras macroscópicas del cuerpo humano. Concepto de fisiología: ciencia que estudia las funciones de los seres multicelulares. La anatomía y fisiología son campos de estudio estrechamente relacionados (forma/función).

entonación. Por último, la parte final de las cavidades oral y nasal que permiten la expulsión del sonido en forma de ondas de presión sonora.

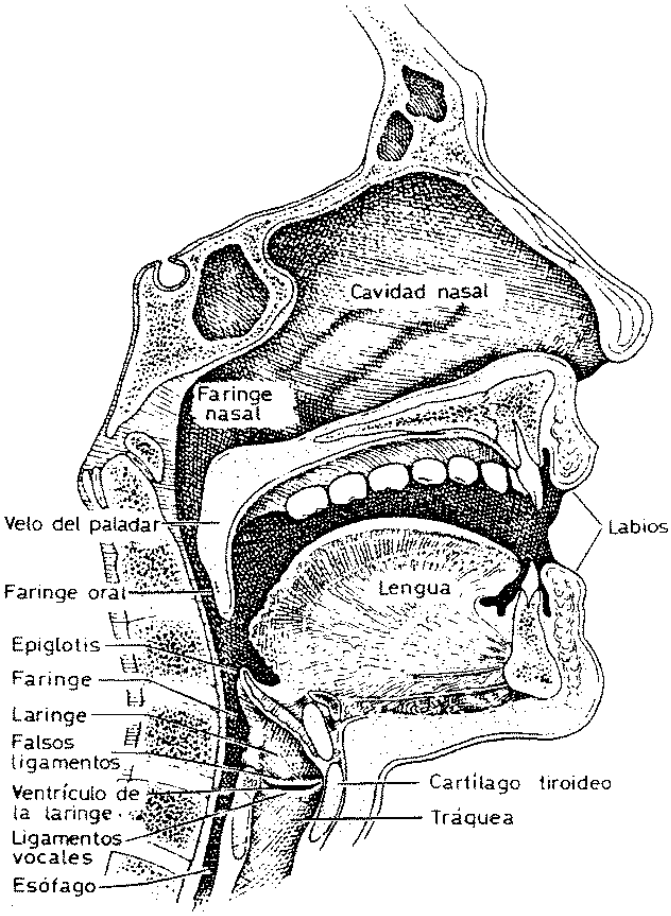


Figura 10. Esquema del aparato fonador humano. Imagen tomada de (Casacuberta & Vidal, 1987).

El conjunto de órganos que intervienen en la fonación, Figura 10, puede dividirse en tres grupos bien delimitados: cavidades infraglóticas u órgano

respiratorio, cavidad laríngea u órgano fonador y cavidades supraglóticas (Casacuberta & Vidal, 1987).

Cavidades infraglóticas

Las cavidades infraglóticas constan de los órganos propios de la respiración (diafragma, pulmones, bronquios y tráquea) que son la fuente de energía para todo el proceso de producción de voz.

En el proceso de inspiración, los pulmones toman aire, bajando el diafragma y agrandando la capacidad torácica. En el momento de la fonación, la espiración, provocada por la contracción de los músculos intercostales y del diafragma, aporta la energía necesaria para generar la onda de presión acústica que atravesará los órganos fonadores superiores. Es así, como el aire sale expulsado hacia la laringe, atravesando la tráquea y la glotis, a diferente presión en función del sonido que se quiere generar.

Cavidad laríngea

La cavidad laríngea es la responsable de modificar el flujo del aire generado por los pulmones y convertirlo en una señal susceptible de excitar adecuadamente las posibles configuraciones de las cavidades supraglóticas.

El último cartílago de la tráquea, el cricoides, forma la base de la laringe, cuyo principal órgano son las cuerdas vocales que son dos pares de repliegues compuestos de ligamentos y músculos, Figura 11. Las cuerdas vocales pueden juntarse o separarse mediante la acción de los músculos crico-aritenoides lateral y posterior, y que están protegidas en su parte anterior por el cartílago tiroides, el más importante de la laringe, abierto por su parte posterior. A la apertura que queda entre las cuerdas vocales se le denomina glotis.

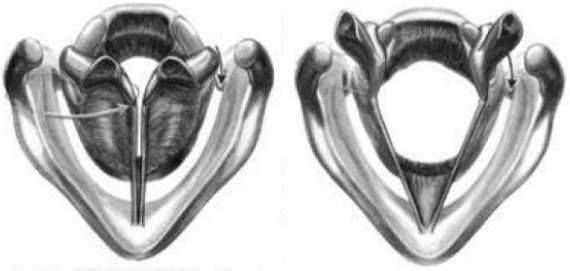


Figura 11. Vista transversal de las cuerdas vocales abiertas y cerradas.

La cavidad laríngea está terminada por la epiglotis, un cartílago en forma de cuchara que permite cerrar la apertura de la laringe en el acto de la deglución.

La cualidad de sonoridad de los sonidos sonoros se produce por la acción vibradora de las cuerdas vocales. El mecanismo de vibración se produce de la siguiente forma: si suponemos que inicialmente las cuerdas vocales están juntas, la presión subglotal se incrementa lo suficiente para forzar a las cuerdas vocales a separarse. Al separarse, el aire pasa a través de ellas y la presión subglotal disminuye, momento en el que la fuerza de los músculos hace que las cuerdas vocales vuelvan a juntarse. Cuando las cuerdas vocales se juntan, el flujo de aire disminuye y la presión subglotal aumenta de nuevo, con lo que se vuelve a repetir el ciclo, Figura 12. Esta vibración de las cuerdas vocales produce pulsos casi periódicos de aire que excitan el sistema por encima de la laringe.

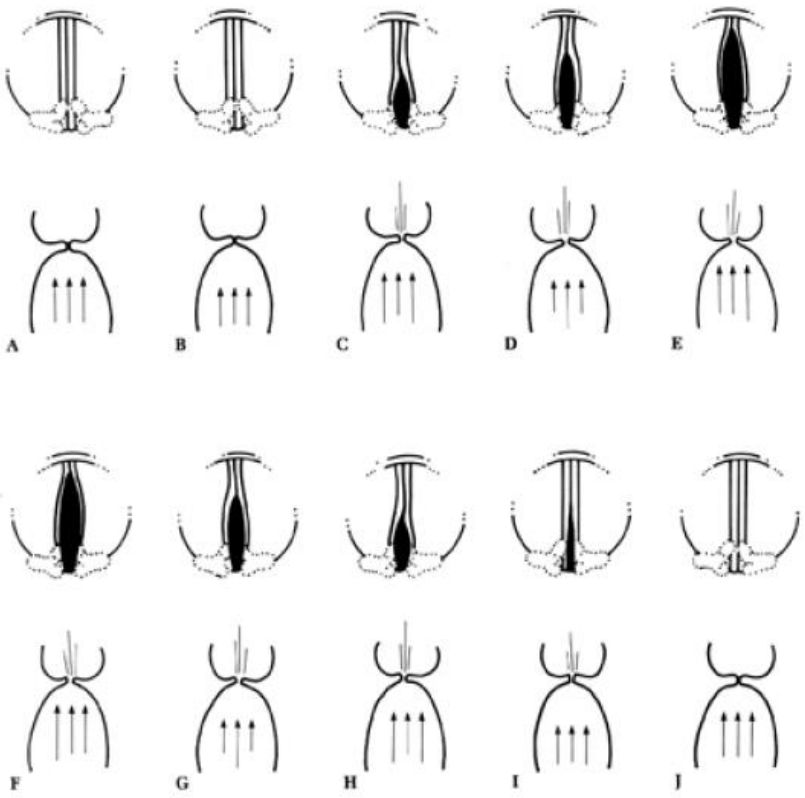


Figura 12. Esquema del proceso de vibración de las cuerdas vocales.

A esta frecuencia de vibración se le denomina frecuencia fundamental, y sus valores típicos oscilan entre los 60 Hz para un hombre voluminoso, y los 300 Hz para una mujer o un niño. La señal generada en las cuerdas vocales puede variar en frecuencia e intensidad según varíe la masa, la longitud y la tensión de las mismas.

Cavidades supraglóticas

Tras superar la glotis el aire se acerca al tracto vocal, que va variando su forma de manera más o menos rápida en función del sonido que se quiere producir. El tracto vocal está integrado por tres cavidades bien diferenciadas: cavidad faríngea, después

de la faringe; cavidad bucal, paladar, lengua, dientes y labios; y cavidad nasal, entre velo del paladar y orificios nasales. Los elementos articulatorios actúan como resonadores que favorecen o neutralizan componentes espectrales de la onda de presión que llega hasta allí. Las resonancias que se producen tienen su energía concentrada alrededor de unas frecuencias, llamadas formantes. Éstas concentran la mayor parte de la información psicoacústica existente en la señal, necesaria para la comprensión del mensaje oral en ellas contenido. Las frecuencias formantes equivalen a los máximos relativos de la envolvente del espectro de la señal. Cada sonido tiene varios máximos relativos (de 4 a 5). Se llama alófono[12] a cada sonido que se diferencia de otros por el número de formantes, posición de los mismos, ancho de banda, nivel espectral, etc.

3.1.2. Propiedades articulatorias y acústicas del sonido del habla

Los sonidos del habla pueden ser estudiados desde diferentes puntos de vista: articulatorio, acústico y perceptual. A continuación, se describen desde el punto de vista articulatorio y acústico, es decir, se hace una descripción sobre cómo se relacionan las características lingüísticas de los sonidos a posiciones y movimientos de los órganos fonatorios, así como la relación entre los fonemas y sus realizaciones acústicas interpretando la señal de voz como la salida del proceso de producción. Las características perceptuales se explican en la siguiente sección: La percepción del habla.

Propiedades articulatorias de los sonidos del español

En esta sección se detalla algunas posibles clasificaciones de los sonidos en base a diferentes criterios, (Flanagan, 1972) y (Quilis & Fernández, 1985).

---

[12] Alófono, son las distintas realizaciones de un mismo fonema según el entorno en que esté situado. Fonema, es la unidad fonológica más pequeña.

1) La acción de las cuerdas vocales:

   a) Sonoros, si las cuerdas vocales se aproximan y comienzan a vibrar. Se distingue entre sonidos vocálicos y sonidos consonánticos. Periódica. La frecuencia fundamental es el llamado tono. Alta energía y estabilidad a corto plazo: /b/, /d/, /g/, /y/, /m/, /n/, /ñ/, /l/, /ll/, /r/, /a/, /e/, /i/, /o/, /u/.

   b) Sordos, si las cuerdas vocales se acercan entre sí pero no llegan a vibrar. Turbulencias. Baja energía, alta frecuencia y poca estabilidad a corto plazo: /p/, /t/, /k/, /f/, /z/, /s/, /j/, /ch/.

2) La acción del velo del paladar:

   a) Orales o bucales, si se encuentra adherido a la pared de la faringe. El aire sale entonces sólo por la cavidad bucal: /p/, /t/, /k/, /f/, /z/, /s/, /ch/, /j/, /b/, /d/, /g/, /y/, /ll/, /r/, /rr/, /a/, /e/, /i/, /o/, /u/.

   b) Nasales, si el conducto nasal está abierto al estar el velo del paladar separado de la pared de la faringe. Parte del aire sale a través de la cavidad nasal: /m/, /n/, /ñ/.

3) El modo de articulación. El modo de articulación se define como la posición de los órganos articulatorios en referencia a su apertura o cierre:

   a) Vocales, se producen con el tracto vocal abierto saliendo el aire libremente a través de la boca. En general, los sonidos vocálicos se caracterizan por su amplitud. Según el grado de abertura se clasifican en:

      i) abierta (lengua totalmente separada del paladar): /a/;

      ii) media (lengua a una distancia intermedia del paladar): /e/, /o/;

      iii) cerrada (lengua muy cerca del paladar): /i/, /u/.

b) Consonantes, se clasifican en:

i) Oclusiva: el sonido se produce en dos fases, cierre del tracto seguido de apertura súbita (explosión): /p/,/b/,/t/,/d/,/k/,/g/,/m/,/n/,/ñ/;

ii) Fricativa: el aire encuentra un cierre parcial o total en algún punto del tracto, provocando una turbulencia: /f/,/s/,/z/,/j/;

iii) Africada: composición de oclusiva seguida de fricativa: /ch/;

iv) Vibrante: el ápice de la lengua se pone en vibración simple o múltiple: /r/,/rr/;

v) Lateral: el aire sale por uno o ambos lados de la lengua: /l/, /ll/.

4) El lugar de articulación:

a) Vocales, a partir de sus formas de onda se puede afirmar que son sonidos casi periódicos, debido principalmente al movimiento cíclico de la glotis (excitación del tracto vocal por las cuerdas vocales). El tracto vocal se mantiene en una configuración relativamente estable durante la mayor parte del sonido vocálico. La posición del tracto vocal para las diferentes vocales varía afectando principalmente a la frecuencia central de sus formantes[13]: observando la posición de los dos primeros formantes se puede obtener una clasificación de los sonidos vocálicos. Los formantes están básicamente influenciados por la longitud del tracto vocal (incluida la faringe), la posición de posibles obstáculos a lo largo del mismo, y el tamaño de dichos obstáculos. El ancho de banda de cada uno de los formantes no suele utilizarse tanto como su frecuencia central para la clasificación de los sonidos vocálicos debido a su mayor variabilidad y su menor consistencia

[13] Las frecuencias formantes equivalen a los máximos relativos de la envolvente del espectro de la señal. Concentran la mayor parte de la información psicoacústica existente en la señal.

entre clases de vocales. Según la posición de la lengua se clasifican en (Tabla 2):

i) Anterior, la lengua se próxima a la región delantera del paladar. Generalmente, son las vocales que tienen los dos primeros formantes más separados: /e/, /i/;

ii) Central, la lengua se encuentra en la parte central del paladar: /a/;

iii) Posterior, la lengua se aproxima a la zona velar. La posición de los dos primeros formantes, siempre pequeños (por debajo de los 1500 Hz) se mantiene muy cercana: /o/, /u/.

Tabla 2. Clasificación articulatoria de las vocales.

		Posición de la lengua		
		Anterior	Central	Posterior
Abertura de la boca	Cerrada	/i/		/u/
	Medio cerrada	/e/		/o/
	Abierta		/a/	

La abertura de la boca (o posición vertical de la lengua), la localización de la constricción (o posición horizontal de la lengua) y la labialización determinan la frecuencia de los formantes. Existe una relación directa entre el grado de abertura de la boca y la frecuencia del primer formante, o F1, cuanto más alta es la frecuencia del F1, la vocal es más abierta; y entre la posición de la lengua y la frecuencia del segundo formante, o F2, cuanto más alta es la frecuencia del F2, más anterior es la vocal (Figura 13 y Figura 14)[14].

[14] Se volverá a hablar de esta relación en el apartado de las propiedades acústicas del sonido del habla. En él se describen cuáles son estos parámetros acústicos: frecuencia, amplitud y tiempo.

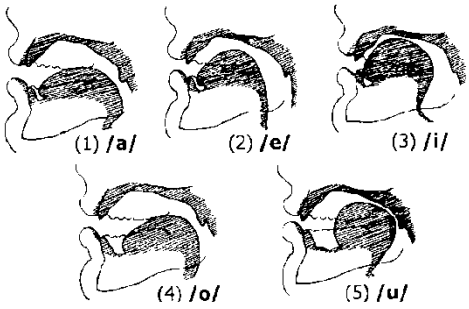


Figura 13. Relaciones articulatorias orales de los sonidos vocálicos.

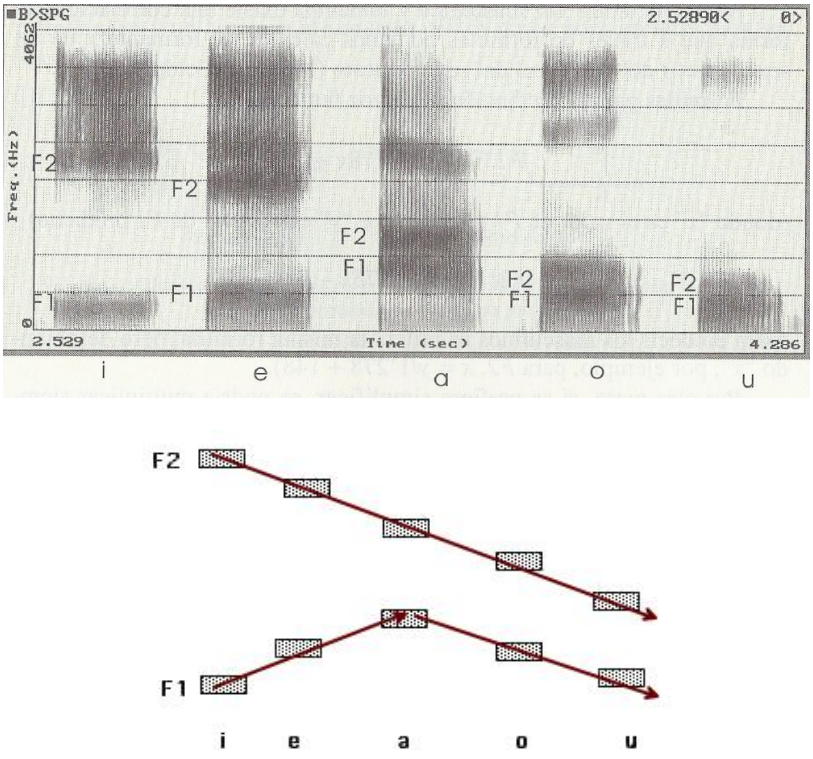


Figura 14. Representación esquemática del primer y del segundo formante de las 5 vocales españolas en un espectograma.

b) Consonantes, son sonidos que no son exclusivamente sonoros, y que no radian únicamente por los labios y con una configuración estable del tracto vocal. Generan fuertes variaciones del tracto vocal. Se clasifican en:

   i) Bilabial: Labios superior e inferior en contacto durante la producción: /p/,/b/,/m/;

   ii) Labiodental: incisivos superiores con labio inferior: /f/;

   iii) Linguodental: ápice de la lengua con incisivos superiores: /t/,/d/;

   iv) Linguointerdental: ápice de la lengua se sitúa en posición interdental: /z/;

   v) Linguoalveolar: ápice de la lengua contacta con los alveolos: /s/,/n/,/l/,/r/,/rr/;

   vi) Linguopalatal: ápice de la lengua contacta con el paladar: /ch/,/ll/,/y/,/ñ/;

   vii) Linguovelar: el postdorso de la lengua contacta con el velo del paladar: /j/, /ga/, /k/.

En el castellano se definen 24 fonemas cuya clasificación, desde el punto de vista de las propiedades articulatorias descritas en esta sección, se resume en la siguiente Tabla 3 (Quilis & Fernández, 1985).

Tabla 3. Clasificación, desde el punto de vista articulatorio, de los 24 fonemas del castellano (SN: sonoro; SR: sordo).

Punto de articulación	Abierto		Labiales				Dentales		Alveolares		Palatales		Velares		Glotales	
Modo de Articulación			Bilabiales		Labiodentales											
	SN	SR	SN	SR	SN	SR	SN	SR	SN	SR	SN	SR	SN	SR	SN	SR
Plosivas			b	p					d	t			g	k		
Nasales			m						n			ñ				
Laterales									l			ll				
Fricativas						f				s	y			x		
Vibr simple									r							
V. Comp.									r							
Africadas												c				
Vocales	a										e,i		o,u			
Semivocales			w								j					

Propiedades acústicas de los sonidos del español

Se ha descrito como la onda sonora producida en la fonación es el resultado del paso del aire por la glotis en la emisión de una serie de sucesivas bocanadas de aire al ritmo de abertura y cierre de las cuerdas vocales. Para la producción de la onda sonora, las moléculas del aire deben entrar en vibración, lo que se consigue por su paso a través de los pliegues vocálicos.

En este sentido, el proceso de producción de los sonidos del habla se puede modelar con una fuente y un filtro (Lieberman & Blumstein, 1988). El aparato fonador puede considerarse en términos de una fuente, lugar donde se produce la corriente de aire indispensable para la producción del sonido; y un filtro, el conjunto de cavidades supraglóticas que por el fenómeno de la resonancia modifican las

características de la fuente (Figura 15 y Figura 16). La fuente puede ser periódica o aperiódica dando lugar a sonidos sonoros (producidos por la vibración de las cuerdas vocales) y sonidos sordos (producidos por una explosión o fricción en algún punto del tracto vocal), respectivamente. El filtro puede clasificarse en fijo o variable y oral o nasal. El filtro fijo no altera la forma de la cavidad de resonancia, por ejemplo en las vocales; mientras que el filtro variable sí produce un cambio, como es el caso de las oclusivas que constan de una fase de oclusión y otra de explosión. Por otro lado, la resonancia puede darse en la cavidad oral (filtro oral) o en la cavidad nasal (filtro nasal), en función de que se cierre o abra el paso de corriente del aire hacia las fosas nasales.

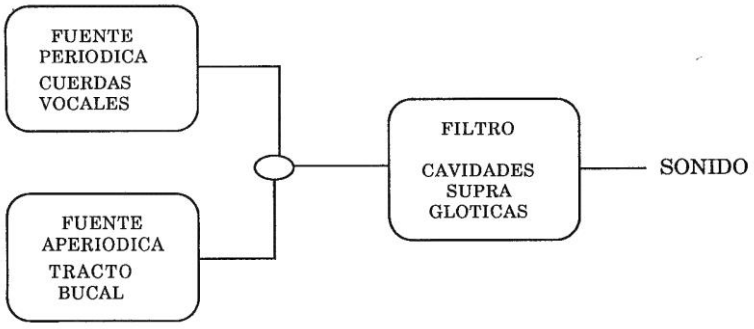


Figura 15. Modelo simplificado de la producción del sonido del habla.

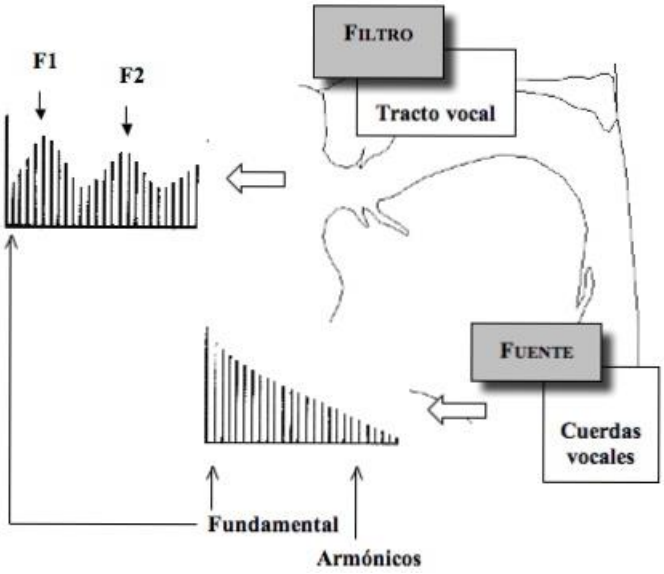


Figura 16. Fuente y filtro en el tracto vocal. Imagen tomada de (Lieberman & Blumstein, 1988).

En la siguiente Tabla 4, se muestra una clasificación acústica de los sonidos del habla en función de este modelo de fuente y filtro (Landercy & Renard, 1977).

Tabla 4. Clasificación acústica de los sonidos del habla, adaptado de (Landercy & Renard, 1977).

Fuente	Filtro	Clase de sonidos
Periódica	Fijo, oral	vocales orales
Periódica	Fijo, oral + nasal	vocales nasales
Periódica	Variable, oral	diptongos
Aperiódica continua	Fijo, oral	fricativas sordas
Aperiódica impulsional	Variable, oral	oclusivas sordas orales
Aperiódica continua + periódica	Fijo, oral	fricativas sonoras
Aperiódica impulsional + periódica	Variable, oral	oclusivas sonoras orales
Aperiódica impulsional + periódica	Variable, oral + nasal	oclusivas sonoras nasales
Aperiódica continua + periódica	Variable, oral	semivocales laterales y vibrantes

La clasificación acústica de los sonidos del habla puede resumirse de la siguiente forma, mostrada en la Tabla 5:

Tabla 5. Clasificación acústica de los sonidos del habla.

Sonidos periódicos compuestos o complejos	Vibración de las cuerdas vocales (frecuencia de la fundamental, F0) y resonancia (armónicos) en el tracto vocal.	Vocales, nasales y laterales
Sonidos aperiódicos impulsionales	Cierre y explosión en el tracto vocal	Oclusivas
Sonidos aperiódicos continuos	Fricción producida por una constricción en el tracto vocal	Fricativas

Los parámetros acústicos que caracterizan una onda sonora son la amplitud[15], frecuencia[16] y el tiempo[17]. A continuación, se va a describir estas características acústicas para el caso concreto de las vocales de la lengua española, ya que son estos fonemas los que se van a utilizar en el sistema implementado. También se presentan, de un modo más general, las características acústicas de las consonantes de la lengua española.

Características acústicas de las vocales

Se ha descrito que las vocales son sonidos periódicos complejos producidos con resonancia en el tracto vocal. Se definen a partir de los siguientes parámetros acústicos:

---

[15] Amplitud: es la distancia existente entre la posición de reposo y el punto de máximo de desplazamiento de una partícula en vibración. Se cuantifica en unidades de intensidad (Decibelios, dB) desde el punto de vista perceptivo.

[16] Frecuencia o número de ciclos por unidad de tiempo. Se mide en Hercio (Hz), equivalente a un ciclo por segundo.

[17] Tiempo o duración de los sonidos del habla. Se cuantifica en milésimas de segundo (ms).

Frecuencia fundamental (F0)

Existe una relación entre el timbre de la vocal y la altura relativa de su frecuencia fundamental. Las vocales cerradas presentan una frecuencia fundamental más elevada que las vocales abiertas. La frecuencia fundamental de las vocales viene modificada por diversos factores. Entre ellos, la sonoridad de las consonantes adyacentes, su aparición en sílaba acentuada o en sílaba no acentuada y su aparición en determinadas posiciones de la curva melódica del enunciado.

Frecuencia de los formantes (F1, F2, F3)

La frecuencia de los tres primeros formantes distingue acústicamente las vocales entre sí, sin embargo, en español es suficiente con la información que nos proporcionan los dos primeros formantes, como así se justifica en el trabajo realizado por Martínez Celdrán (Martínez Celdrán, 1995). Esta característica acústica es la que se usa en el sistema implementado en este trabajo para diferenciar y así identificar a cada uno de los 5 fonemas vocálicos de la lengua española (se describe en el capítulo 7).

Hay que destacar que no existen los valores absolutos para las formantes de las vocales; estos valores puntuales solo sirven como valores de referencia. Lo importante es el campo de dispersión y sobre todo los límites de este campo para cada una de las vocales. En la producción de los sonidos, la coarticulación puede modificar la frecuencia canónica de la formante al adaptarse al contexto; pero este hecho mecánico no tiene consecuencias en la percepción. Es decir, aunque existen variaciones en la articulación, no se van a producir diferencias acústicas notables, y no van a tener ninguna repercusión perceptiva. Por otro lado, en el proceso de percepción tampoco se aprecian las diferencias físicas producidas dentro de una misma categoría. Por lo tanto, se puede concluir con que una vocal, desde la perspectiva acústico-perceptiva, no es un punto en el espacio, sino un dominio con

unos límites amplios. Y lo que realmente tiene importancia es el conocimiento de cada dominio y de sus límites, sobre todo para el reconocimiento de vocales.

En las siguientes Tabla 6 y Tabla 7, se muestran los valores medios, máximo y mínimo en frecuencia (Hz) de las formantes F1 y F2, respectivamente, de las vocales del castellano para una voz masculina. Los datos han sido tomados del estudio sobre las vocales del español realizado por Martínez Celdrán (Martínez Celdrán, 1995). En este estudio se recogen muestras de 5 hablantes masculinos y 5 femeninos universitarios españoles, cuyas edades se sitúan entre los 20 y 30 años. Cada hablante realizó 30 emisiones de cada vocal.

Tabla 6. Valores medios, mínimo y máximo de la formante F1 (Hz) para cada vocal castellana, para una voz masculina, tomada de (Martínez Celdrán, 1995).

F1	**i**	**e**	**a**	**o**	**u**
media	313	457	699	495	349
mínimo	241	381	571	393	277
máximo	414	587	1002	656	449

Tabla 7. Valores medios, mínimo y máximo de la formante F2 (Hz) para cada vocal castellana, para una voz masculina, tomada de (Martínez Celdrán, 1995).

F2	**i**	**e**	**a**	**o**	**u**
media	2200	1926	1471	1070	877
mínimo	1832	1676	1296	793	622
máximo	2523	2212	1642	1313	1175

Las vocales se pueden representar gráficamente. Para ello, se utiliza una carta de formantes que consiste en un eje de coordenadas, en cuya ordenada se colocan los valores de F1 y en cuya abscisa se representan los valores de F2. Este gráfico va a permitir además la comparación de cada uno de los alófonos de una vocal. La escala

que se suele utilizar en estos gráficos no es lineal, sino logarítmica, porque representan mejor la percepción del oído humano.

En la siguiente Figura 17 se muestra una carta de formantes en la que se han situado 31 realizaciones vocálicas de una hablante española (Quilis & Esgueva, 1983). El conjunto de los puntos marcados para cada vocal constituye la zona de dispersión[18] de dicha vocal y determina su campo vocálico.

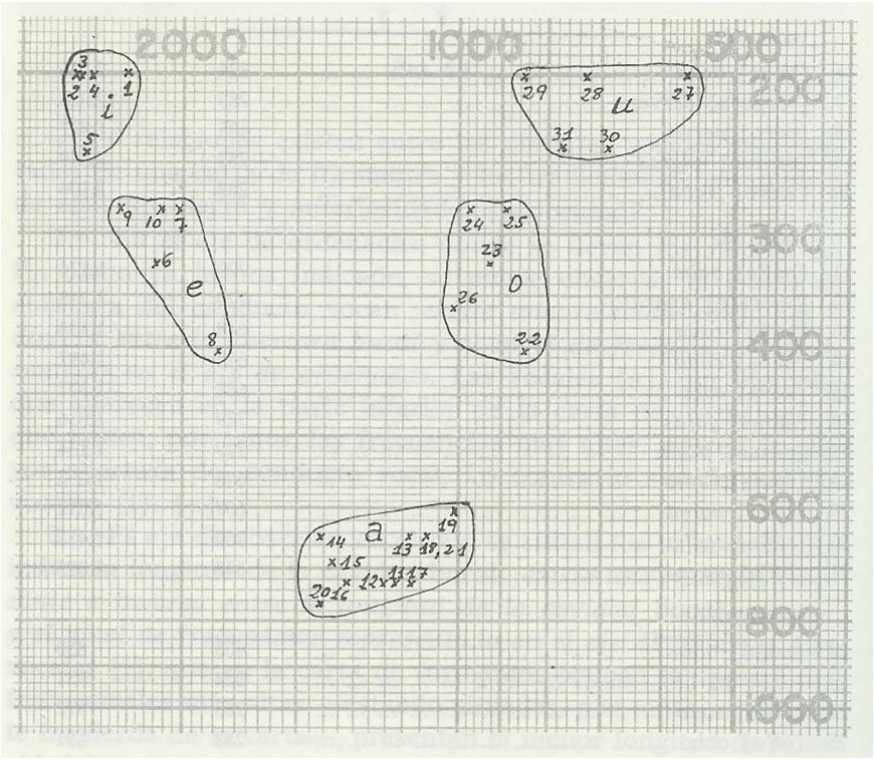


Figura 17. Área de dispersión de las vocales del español.
Imagen tomada de (Quilis & Esgueva, 1983).

[18] El área o campo de dispersión es el espacio que se obtiene al trazar una línea exterior a todos los puntos que representan realizaciones de una misma vocal.

Los valores vocálicos colocados en una carta de formantes muestran la relación entre los parámetros acústicos y las posiciones de los órganos articulatorios. En el eje de ordenadas se observa la abertura vocálica; donde se aprecia una relación directa y constante entre la abertura y el valor de la frecuencia de F1; y en el eje de abscisas puede observarse la localización vocálica (anterioridad-posterioridad), y la relación inversa y constante entre la longitud de la cavidad bucal anterior y el nivel frecuencial de F2.

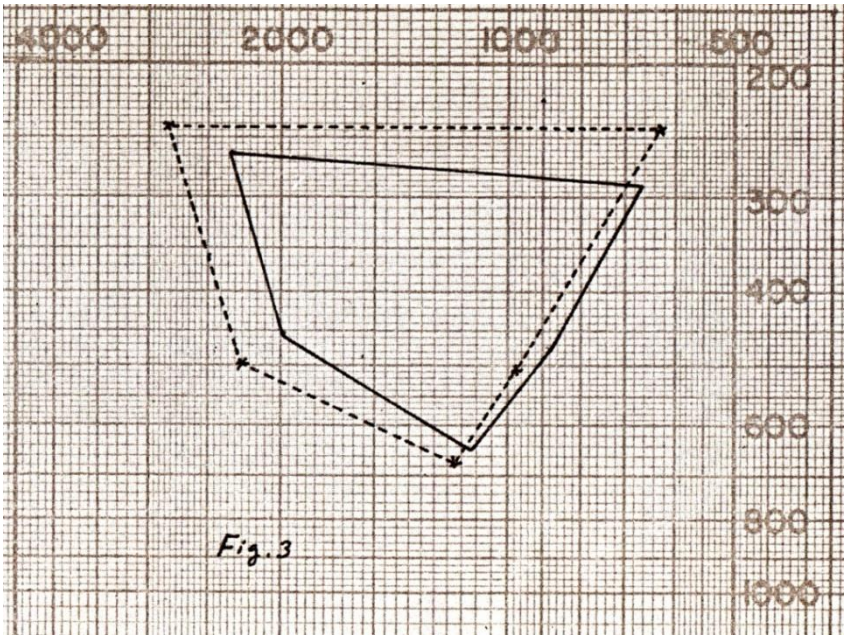


Figura 18. Trapecio vocálico obtenido a partir de una muestra de 16 hablantes masculinos (línea continua) y 6 hablantes femeninas (línea discontinua) hablantes de español. Imagen tomada de (Quilis & Esgueva, 1983).

Si se unen entre sí los puntos representados en una carta de formantes, se obtiene un triángulo o trapecio acústico de las vocales, Figura 18. Esta figura es muy parecida al triángulo articulatorio (que muestra la posición articulatoria de la cavidad bucal en la producción de cada elemento), Figura 19. Por tanto, la información que

ofrecen ambos triángulos es similar, aunque, según explica Quilis (Quilis & Esgueva, 1983), el triángulo articulatorio supone una excesiva simplificación de la realidad articularia que pretende representar; es decir, sólo recoge dos de los muchos parámetros que cabría mostrar: posición lingual anterior/posterior y superior/inferior. Además, hay que tener en cuenta que posiciones articulatorias diferentes pueden dar el mismo resultado acústico.

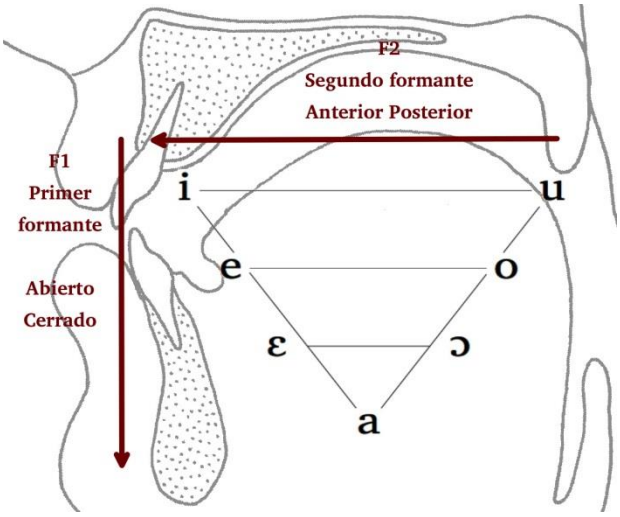


Figura 19. Relación entre las características articulatorias y las características acústicas en el trapecio vocálico.

DURACIÓN

Aunque las vocales abiertas presentan una duración intrínseca[19] mayor que las cerradas, en general, la duración de las vocales viene condicionada por diferentes factores:

[19] Cualidad intrínseca: aquella determinada por sus características fonéticas propias.

- Sonoridad, lugar y modo de articulación de las consonantes adyacentes.
- Situación en sílaba abierta o cerrada.
- Aparición en sílaba acentuada o no acentuada.
- Aparición en palabras monosilábicas o polisilábicas.
- Aparición en posición inicial o final de palabras y en posición pre- o post-pausal.

INTENSIDAD

Las vocales abiertas presentan una intensidad intrínseca mayor que las cerradas. La intensidad de las vocales puede verse afectada por diversos factores:

- Aparición en sílaba acentuada o no acentuada.
- Posición respecto al acento en la palabra.
- Posición en el enunciado.

Características acústicas de las consonantes

En esta sección, se van a describir las principales características acústicas de las consonantes.

Las consonantes oclusivas se originan por un cierre -oclusión- en algún lugar del tracto vocal. Por tanto, la característica acústica durante esa fase es la ausencia de sonido. Aunque, esto no es del todo cierto en las oclusivas sonoras, donde las cuerdas vocales vibran durante la fase de oclusión, generando una banda de frecuencias muy bajas, que se conoce como barra de sonoridad. Tanto para las oclusivas sordas como para las oclusivas sonoras, la presión del aire proveniente de los pulmones rompe esta barrera de sonoridad, originando un ruido característico conocido como explosión. Ésta, acústicamente, se manifiesta como una barra

vertical que ocupa gran parte del espacio frecuencial que se muestra en un sonograma convencional, y recibe el nombre de barra de explosión.

Acústicamente, una consonante fricativa es un puro ruido, es decir, una onda aperiódica continua. En el espectro, se muestra como una mancha, sin las estriaciones ni los formantes característicos del sonido armónico. Los sonidos fricativos se distinguen muy bien del resto de las consonantes, aunque no es sencillo distinguir las diferentes fricativas entre sí. Para identificarlas hay que tener en cuenta la intensidad general, las zonas de mayor concentración de energía a lo largo de la fricación, la presencia o ausencia de sonoridad y las transiciones vocálicas.

Las africadas presentan, en primer lugar, la oclusión y la barra de explosión propias de las consonantes oclusivas. Después, aparece un momento fricativo, es decir, inarmónico.

Los sonidos nasales, debido a sus características articulatorias (paso de aire por la cavidad nasal y constricción en la cavidad oral), presentan una estructura armónica parecida a la vocálica, pero con mucha menos intensidad. La diferencia fundamental de las nasales es que aparecen en el espectro en forma de bloque recto, sin las suaves transiciones características de las vocales.

Las laterales, junto con las nasales y las vocales, presentan una estructura formántica, aunque siempre con menor intensidad que las vocales y más intensidad que las nasales. La diferente altura de F2 y las transiciones permiten distinguir el punto de articulación de las distintas laterales.

Y por último, las vibrantes se caracterizan por presentar periodos oclusivos (interrupciones de energía) muy breves, que corresponden a las pequeñas oclusiones producidas en el tracto vocal. Cada una de esas oclusiones da origen a su propia explosión, seguida de un periodo vocal, todo ello muy breve. Este ciclo iniciado con una brevísima oclusión, seguida de un brevísimo momento vocálico puede producirse una sola vez o varias veces seguidas, dependiendo si la vibrante es simple

o múltiple, respectivamente. Hay que destacar que durante los momentos oclusivos, las cuerdas vocales no dejan de vibrar; de ahí que las vibrantes son sonidos sonoros.

En la siguiente Tabla 8 se muestra la relación entre las características articulatorias (modo de producción), las características acústicas y las características perceptivas de los sonidos del habla, que se describen en la siguiente sección.

Tabla 8. Relación entre características articulatorias, acústicas y perceptivas de los sonidos del habla.

Producción	**Parámetro acústico**	**Parámetro perceptivo**	**Percepción**
Frecuencia de vibración de las cuerdas vocales	Frecuencia Fundamental	Altura tonal Tono (*pitch*)	Grave-agudo
Configuración de las cavidades supraglóticas	Composición espectral: frecuencia y amplitud de los formantes	Timbre (*Quality*)	Oscuro-claro
Fuerza espiratoria	Amplitud	Intensidad subjetiva Sonoridad (*Loudness*)	Fuerte-flojo
Duración de la espiración	Tiempo	Duración	Largo-breve

3.2. La percepción del habla

En esta sección se describe la anatomía y fisiología del aparato auditivo, haciendo énfasis en aquellas partes y estructuras más importantes para el desarrollo de modelos perceptuales. Además se hace una introducción a los conceptos básicos de la psicoacústica los cuales van a permitir caracterizar la respuesta del sistema auditivo humano.

3.2.1. El sentido de la audición y el sistema auditivo

La generación de sensaciones auditivas en el ser humano es un proceso complejo, el cual se desarrolla en tres etapas básicas:

- Captación y procesamiento mecánico de las ondas sonoras.
- Conversión de la señal acústica (mecánica) en impulsos nerviosos, y transmisión de dichos impulsos hasta los centros sensoriales del cerebro.
- Procesamiento neuronal de la información codificada en forma de impulsos nerviosos.

La captación, procesamiento y transducción de los estímulos sonoros se llevan a cabo en el oído propiamente dicho, mientras que la etapa de procesamiento neural, en el cual se producen las diversas sensaciones auditivas, se encuentra ubicada en el cerebro. Así pues, se distinguen dos regiones en el sistema auditivo: la región periférica, en la que los estímulos sonoros conservan sus carácter original de ondas mecánicas hasta el momento de su conversión en señales electroquímicas, y la región central, en la que se transforman dichas señales en sensaciones.

En la región central también intervienen procesos cognitivos, mediante los cuales se asigna un contexto y un significado a los sonidos; es decir, permiten reconocer una palabra o determinar si un sonido dado corresponde a un violín o a un piano.

Región periférica del sistema auditivo

El sistema periférico auditivo, Figura 20, comúnmente conocido como oído, se divide en tres zonas llamadas oído externo, oído medio y oído interno, de acuerdo a su ubicación en el cráneo, (Thibodeau, 1998).

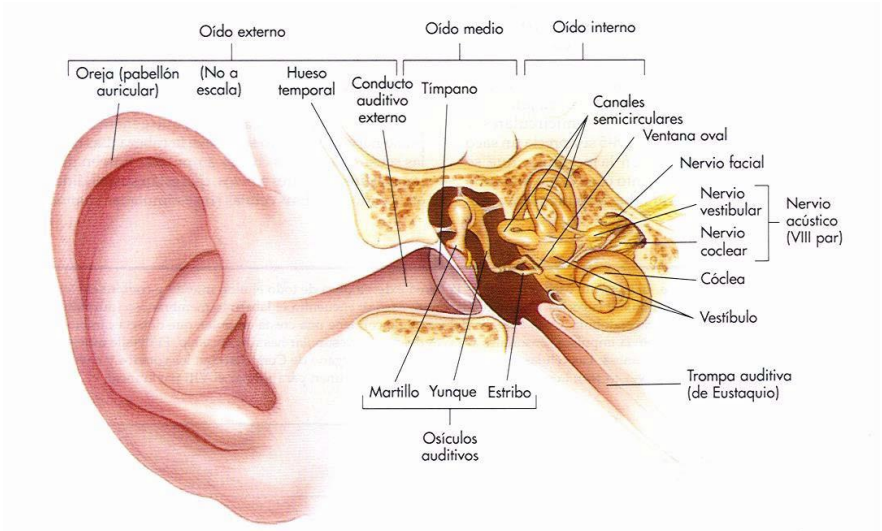


Figura 20. Anatomía del oído. Imagen tomada de (Thibodeau, 1998).

Los estímulos sonoros se propagan a través de estas zonas, sufriendo diversas transformaciones hasta su conversión final en impulsos nerviosos. Tanto el procesamiento mecánico de las ondas sonoras como la conversión de éstas en señales electroquímicas son procesos no lineales (Fastl & Zwicker, 2007), lo cual dificulta la caracterización y modelado de los fenómenos perceptuales.

A continuación, se describe la anatomía y funcionamiento de estas tres zonas del oído, así como la propagación y procesamiento del sonido a través de las mismas.

OÍDO EXTERNO

El primer órgano del aparato auditivo es la oreja, el pabellón auditivo (*pinna,* en latín). Su forma peculiar, llena de huecos y protuberancias mantiene una leve función amplificadora (en frecuencias medias-altas) y, sobre todo, ayuda a la localización de los sonidos. Permite una localización en el eje lateral (izquierda/derecha), según las diferencias temporales y de intensidad entre las señales que provienen de un oído u

otro; y una localización en el eje central (delante/detrás; arriba/abajo) por un efecto de sombra sobre los sonidos que se encuentran detrás de la cabeza. En esencia, se puede afirmar que el pabellón auditivo se comporta como un colector de ondas sonoras.

El oído externo se completa con el conducto auditivo externo, un tubo irregular y no rígido de un diámetro aproximado de 0,7 cm y una longitud aproximada de 2,7 cm, que concluye en el tímpano. Su función es la de proteger la entrada al oído medio y mantener el tímpano y las estructuras del oído medio a una temperatura estable. Además, funciona como un resonador, amplificando las ondas que coinciden con sus frecuencias de resonancia, y amortiguando las restantes. Por sus dimensiones[20], resuena mejor alrededor de 3500 Hz, pero al ser un pasillo ancho, su rango se amplía desde 2000 a 5000 Hz. La presión sonora en estas frecuencias puede llegar a multiplicarse por cuatro o seis (12-15dB *SPL*[21]) desde el exterior hasta su llegada al tímpano.

OÍDO MEDIO

Con la llegada de la onda al tímpano comenzamos la descripción del oído medio. El oído medio es una cavidad localizada entre el oído externo y el oído interno. Su interior está lleno de aire y en él se encuentra el sistema de huesecillos conocidos como el martillo, el yunque y el estribo, los cuales conectan la membrana timpánica con la ventana oval.

---

[20] Dada la velocidad de propagación del sonido en el aire (aprox. 334 m/s), la longitud de 2cm, corresponde a ¼ de la longitud de onda de una señal sonora de unos 4 kHz. Este es uno de los motivos por los cuales el aparato auditivo presenta una mayor sensibilidad a las frecuencias cercanas a los 4 kHz.

[21] La intensidad sonora se expresa en decibelios (dB *SPL*) (*Sound Pressure Level).* El nivel de presión sonora determina la intensidad del sonido. Una intensidad sonora de 0 dB es apenas perceptible, 20 dB equivale a un susurro a 1 m de distancia y es 100 veces más intenso, 60 dB equivale a una conversación normal y es un millón de veces más intenso, y 100 dB equivale a un martillo neumático a 10 m de distancia, siendo 10.000 millones de veces más intenso.

En el oído medio, las ondas sonoras, unos simples cambios en la presión del aire, se convierten en una vibración mecánica. Esta conversión tiene lugar en la membrana timpánica, que se mueve empujada por los cambios de presión que llegan desde el conducto auditivo. Dicho movimiento se transmite a la cadena de huesecillos que aumentarán esas vibraciones, gracias a su disposición en forma de palanca. La base del estribo vibra en la ventana oval; este huesecillo se encuentra en contacto con uno de los fluidos contenidos en el oído interno; por tanto, el tímpano y la cadena de huesecillos actúan como un mecanismo para transformar las vibraciones del aire en vibraciones del fluido. Para conseguirlo, es importante que la presión del aire dentro del oído medio sea igual a la presión atmosférica; se consigue gracias a la trompa de Eustaquio, que lo provee de aire procedente de la faringe.

El oído medio tiene tres funciones. La primera, aumentar la presión recibida del tímpano. Esto es importante porque la cóclea está llena de líquido, no de aire. Y la densidad y compresibilidad del líquido coclear es casi 4000 veces menor que la del aire. Por tanto, si no dispusiéramos de un mecanismo para aumentar la presión, sólo llegaría al interior de la cóclea un 0,1% de la presión timpánica. La segunda función es la de proteger las estructuras del oído interno de ruidos excesivamente fuertes. Gracias al estribo que es capaz de contraerse de forma refleja cuando llega un sonido inferior a 1-2 kHz y con una intensidad superior a 85-90 dB. Este reflejo, conocido como reflejo timpánico o acústico, convierte la acción elevadora que aumenta la presión sobre la ventana oval en una acción rotatoria, lo cual disminuye la presión unos 20 dB. La tercera función es desarrollada por los músculos del oído medio. Se comportan como un filtro paso baja que reducen la transmisión de los sonidos de baja frecuencia, con una atenuación aproximada de 15 dB por octava en la zona de 1000 Hz, disminuyendo así el enmascaramiento que éstos producirían sobre frecuencias más altas.

Y se llega así, con la intensidad multiplicada, al punto donde el estribo se conecta con la ventana oval: la entrada a la cóclea y al oído interno.

OÍDO INTERNO

En el oído interno encontramos, por un lado, los canales semicirculares encargados de controlar el equilibrio y la cóclea, el órgano de la audición por excelencia. Es un tubo rígido en forma de espiral, de unos 32-35 mm de largo y un grosor que va desde 4 mm^2 en la base hasta 1 mm^2 en la punta o ápice, lleno con dos fluidos de distinta composición, Figura 21.

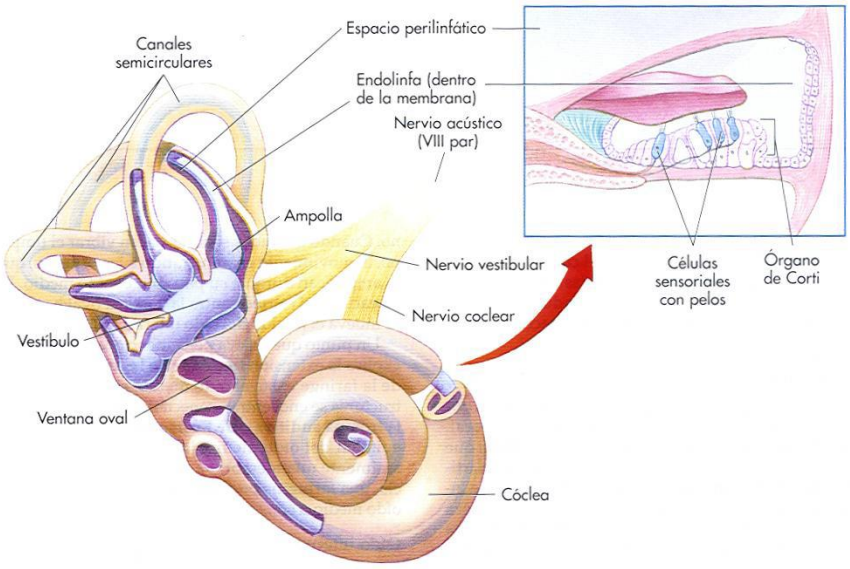


Figura 21. Oído interno. Imagen tomada de (Thibodeau, 1998).

En la cóclea se encuentra una subestructura flexible y hueca, la partición coclear (*cochlear duct*), que la divide en dos rampas o escalas, la vestibular y la timpánica, Figura 22 y Figura 23c. La partición coclear no es plana, sino hueca. Tiene un techo, la membrana Reissner y un suelo, por el lado de la rampa timpánica, la membrana basilar. La rampa vestibular y la rampa timpánica contienen el mismo fluido, perilinfa, (cuyo potencial eléctrico es negativo), puesto que se interconectan por una pequeña abertura situada en el vértice del caracol, llamada helicotrema. Por el

contrario, la partición coclear se encuentra aislada de las otras dos rampas y contiene un líquido de distinta composición a la perilinfa, un líquido viscoso con un alto potencial eléctrico (muy positivo): la endolinfa.

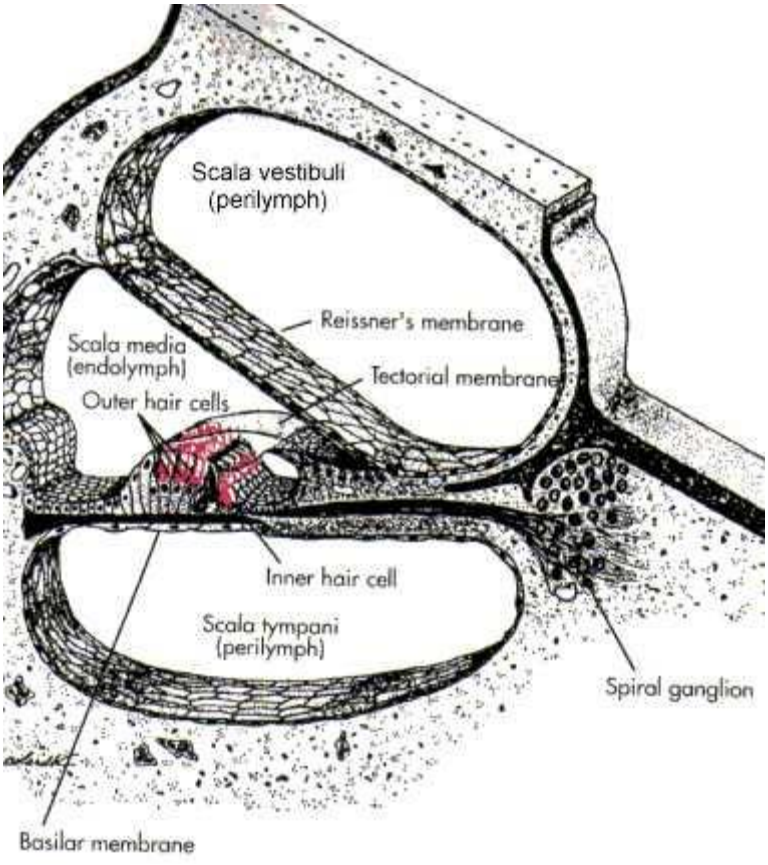


Figura 22. Sección de un canal de la cóclea y el órgano de Corti.

La base del estribo, a través de la ventana oval, está en contacto con el fluido de la rampa vestibular, mientras que la rampa timpánica desemboca en la cavidad del oído medio a través de otra abertura (ventana redonda) sellada por una membrana flexible (membrana timpánica secundaria).

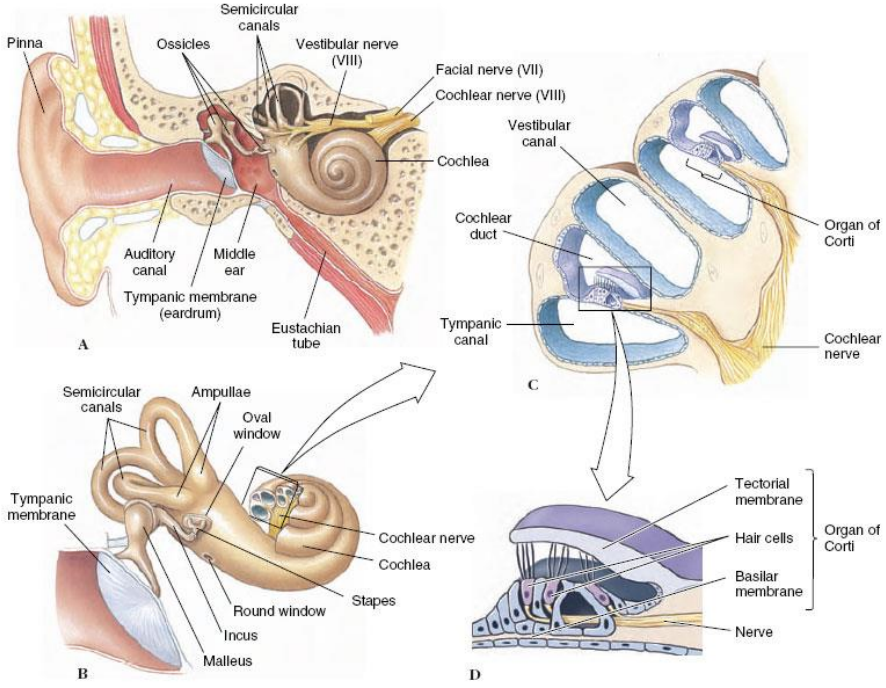


Figura 23. El oído humano. (A) Sección longitudinal del oído externo, oído medio y oído interno. (B) Unión del oído medio al oído interno; el estribo se conecta a la ventana oval. (C) Se realiza un corte en la cóclea para mostrar el interior de sus canales. Esta sección de la cóclea muestra el órgano de Corti. (D) Detalle de la estructura interna del órgano de Corti.

La membrana basilar es una estructura cuyo espesor y rigidez no es constante: cerca de la ventana oval, la membrana es gruesa y rígida, pero a medida que se acerca hacia el vértice de la cóclea se vuelve más delgada y flexible.

La rigidez decae casi exponencialmente con la distancia a la ventana oval; esta variación de la rigidez en función de la posición afecta a la velocidad de propagación de las ondas sonoras a lo largo de ella, y es responsable en gran medida de un fenómeno muy importante: la selectividad en frecuencia del oído interno.

La membrana basilar es el soporte del órgano de Corti, Figura 24, el elemento más importante de la cóclea, encargado de convertir el movimiento en descargas que activen las fibras nerviosas.

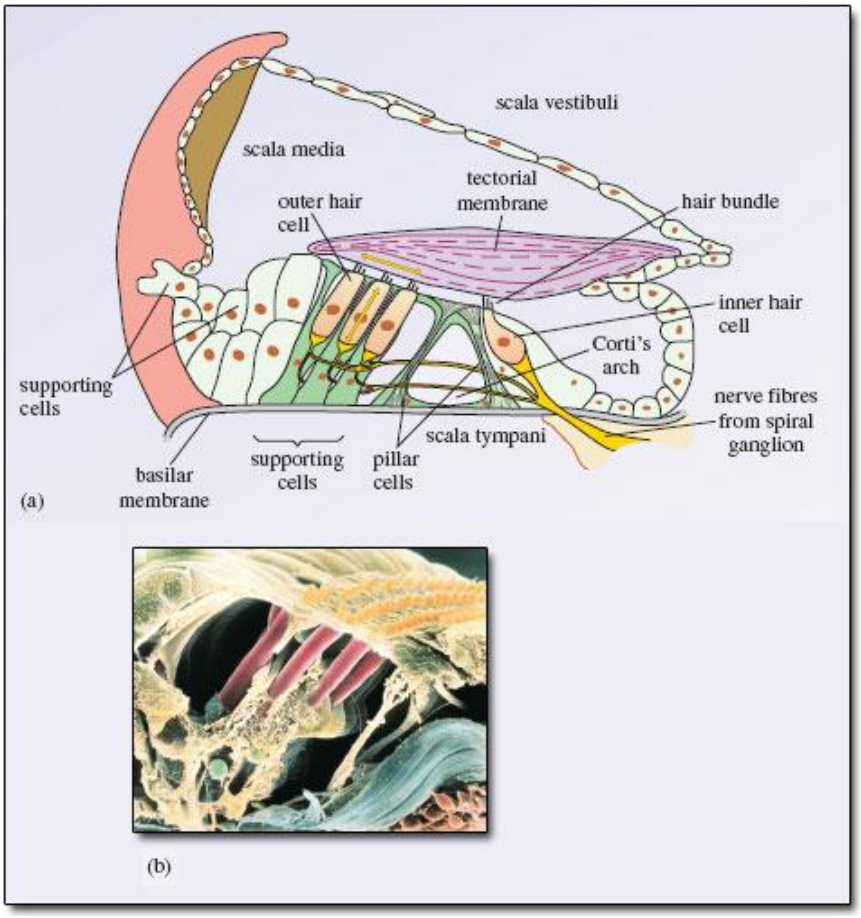


Figura 24. Estructura interna del órgano de Corti.

El órgano de Corti contiene entre 15000 y 30000 receptores del nervio auditivo: las células ciliadas, de las cuales salen los haces de fibras que componen el nervio auditivo o coclear. Cada célula tiene una serie de cilios con capacidad para producir pequeñas descargas eléctricas al rozar la membrana superior, la membrana tectorial.

Dependiendo de su ubicación en el órgano de Corti, se pueden distinguir dos tipos de células ciliadas: internas (*Inner hair cells, IHCs*) y externas (*Outer hair cells, OHCs*). Existen alrededor de 3500 células ciliadas internas y unas 20000 células ciliadas externas. Ambos tipos de células presentan conexiones o sinapsis con las fibras nerviosas aferentes (que aportan impulsos hacia el cerebro) y eferentes (que transportan impulsos provenientes del cerebro), las cuales conforman el nervio auditivo. Sin embargo, la distribución de las fibras es muy desigual: más del 90% de las fibras aferentes inervan a las células ciliadas internas, mientras que la mayoría de las 500 fibras eferentes inervan a las células ciliadas externas.

El funcionamiento de la cóclea comienza como un proceso hidráulico: las oscilaciones del estribo provocan oscilaciones en el fluido de la rampa vestibular (perilinfa). La membrana de Reissner, la cual separa los fluidos de la rampa vestibular y la rampa media, es sumamente delgada y, en consecuencia, los líquidos en ambas rampas pueden tratarse como uno solo desde el punto de vista de la dinámica de los fluidos. Así, las oscilaciones en la perilinfa de la rampa vestibular se transmiten a la endolinfa y de ésta a la membrana basilar; la membrana basilar a su vez, provoca oscilaciones en el fluido de la rampa timpánica, Figura 25.

Puesto que tanto los fluidos como las paredes de la cóclea son incompresibles, es preciso compensar el desplazamiento de los fluidos; esto se lleva a cabo en la membrana de la ventana redonda, la cual permite cerrar el circuito hidráulico.

La propagación de las oscilaciones del fluido en la rampa vestibular a la timpánica no solo se lleva a cabo a través de la membrana basilar; para sonidos de muy baja frecuencia, las vibraciones se transmiten a través de la abertura situada en el vértice de la cóclea (helicotrema).

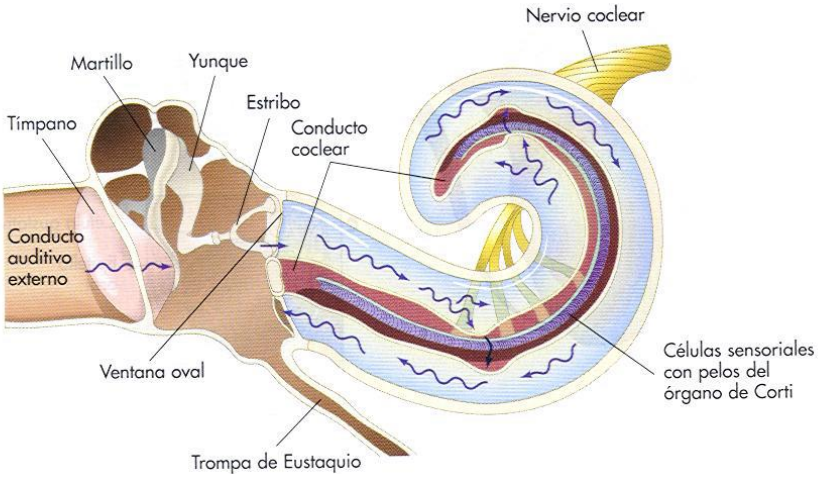


Figura 25. Efecto de las ondas sonoras sobre las estructuras del oído.
Imagen tomada de (Thibodeau, 1998).

En conclusión, el sonido propagado a través del oído externo y medio llega hasta la cóclea, donde las oscilaciones en los fluidos hacen vibrar a la membrana basilar y a todas las estructuras que ésta soporta. En una región específica, esta onda tiene un máximo en su amplitud que depende de la frecuencia del sonido y posteriormente tiende a disminuir rápidamente hacia el ápice de la cóclea. Mientras menor es la frecuencia del tono, mayor es la distancia que viaja la onda a lo largo de la membrana antes de ser atenuada, y viceversa. De esta forma, la membrana basilar dispersa las distintas componentes de una señal de espectro complejo en posiciones bien definidas respecto a la ventana oval, Figura 26 y Figura 27.

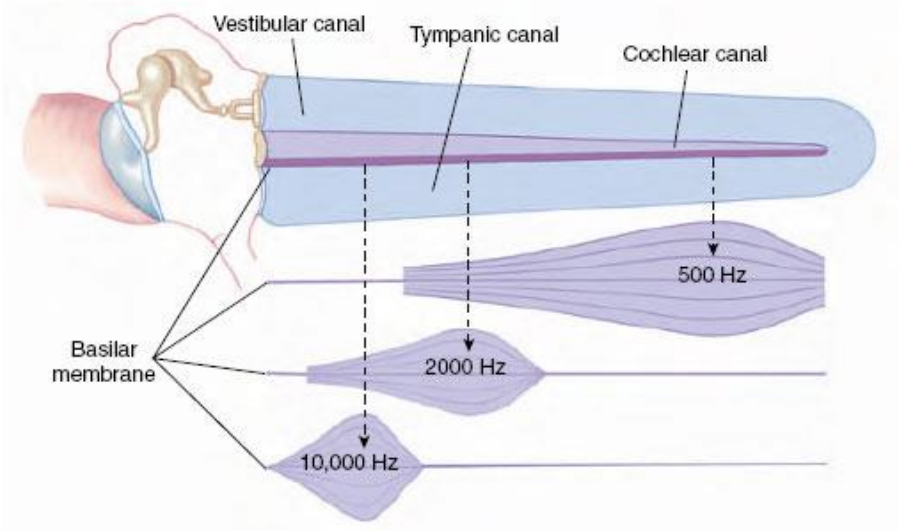


Figura 26. Frecuencia de resonancia de la membrana basilar.

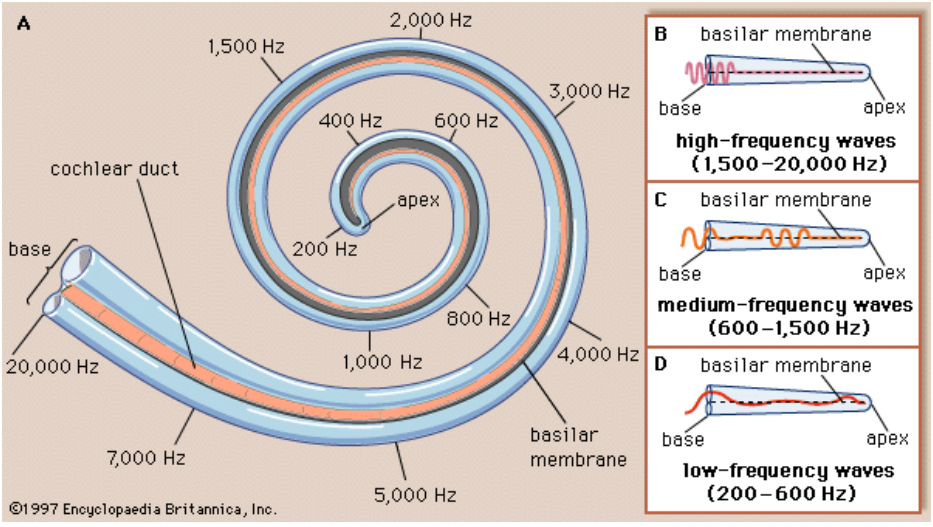


Figura 27. Organización tonotópica de la cóclea. A) Distribución tonotópica de la cóclea. B) Localización de la respuesta coclear a altas frecuencias. C) Localización de la respuesta coclear a frecuencias medias. D) Localización de la respuesta coclear a bajas frecuencias.

Hay que resaltar que no existe una relación lineal entre la distancia desde la ventana oval a un punto sobre la membrana basilar y la frecuencia de resonancia de ese punto, sino que la relación es exponencial.

Lo importante de esta onda que se desplaza no es que llegue hasta el ápice de la cóclea, sino que empuje sobre la partición coclear, tirando de ella arriba y abajo. Este movimiento se extiende al órgano de Corti, y en la subida hace que las células ciliadas rocen la membrana tectorial, Figura 28. Con cada una de esas mínimas flexiones se genera un potencial eléctrico que se propagará a lo largo del nervio auditivo. Serán excitadas o inhibidas dependiendo de la dirección del movimiento. Por tanto, el movimiento de la membrana basilar provoca fuerzas en los cilios de las células ciliadas y desencadena una serie de acontecimientos mecánicos, eléctricos y bioquímicos, responsables de la transducción mecánico-sensorial y del procesamiento inicial de las señales acústicas. La fuerza sobre los cilios provoca la apertura de los canales iónicos existentes en las membranas celulares y modifican la permeabilidad de éstas, permitiendo la entrada de iones potasio en las células. Este flujo de iones potasio hacia el interior causa la despolarización y genera un potencial de acción.

Los neurotransmisores liberados en la unión sináptica por la despolarización de las células ciliadas internas desencadenan impulsos neuronales que se transmiten por las fibras aferentes del nervio auditivo hacia los centros nerviosos superiores. La intensidad de la estimulación auditiva depende del número de potenciales de acción por unidad de tiempo y del número de células estimuladas, mientras que la frecuencia percibida depende de la población específica de fibras nerviosas activadas. Existe una asociación específica entre la frecuencia del estímulo sonoro y la sección de la corteza cerebral estimulada.

La membrana tectorial actúa únicamente como una masa, produciendo una fuerza de desplazamiento horizontal sobre las células ciliadas. Cuanto menor es la

frecuencia de vibración del sonido, más cerca del ápice se produce el máximo desplazamiento de la membrana basilar. Para frecuencias mayores, el máximo desplazamiento se localiza más cerca de la base de la cóclea.

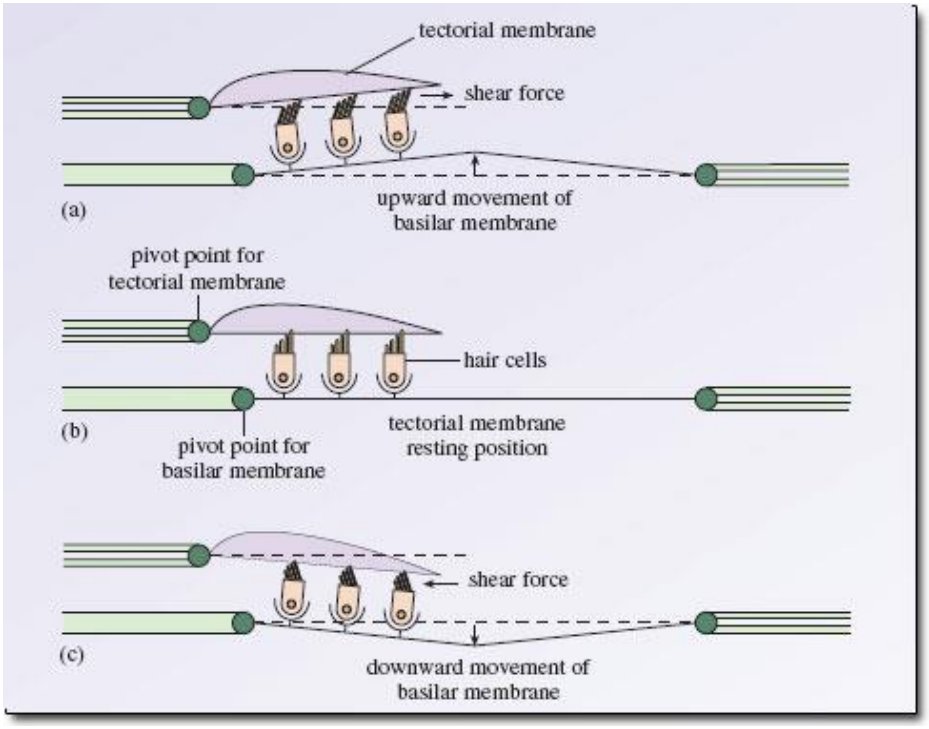


Figura 28. Esquema de la fuerza creada entre las células ciliadas y la membrana tectorial, como consecuencia del desplazamiento de la membrana basilar.

Dependiendo de la región de la membrana basilar que oscila con mayor amplitud, las células ciliadas de esa área se activan en mayor proporción que sus vecinas, excitando subsecuentemente a las neuronas aferentes que hacen sinapsis con ellas. Este proceso ha dado origen al concepto de frecuencia característica para describir la forma en que las neuronas de la vía auditiva responden con un umbral especialmente bajo para los sonidos de cierta frecuencia, y tiene un papel fundamental en la discriminación de los tonos de un sonido. Ante un tono puro muy

próximo a su frecuencia característica, un lugar concreto de la membrana basilar oscilará con una mayor amplitud. A medida que la frecuencia del tono puro se aleja de la frecuencia característica, la respuesta se debilitará. Se dice por tanto que cada zona de la membrana basilar actúa como un filtro auditivo, que responde ante un rango estrecho de frecuencias. La membrana basilar puede ser vista como un banco de filtros auditivos, centrados a diferentes frecuencias características[22]. El rango de frecuencias al que cada filtro responde varía con la frecuencia. Cuando cualquier tono se duplica en frecuencia, se desplaza una octava, la región que resuena de la cóclea se desplaza alrededor de 3.5 a 4 mm, sin importar la diferencia absoluta entre las frecuencias de la octava (por ejemplo, entre 220 y 440, entre 1760 y 3520 o entre 5000 y 10000); siempre que la frecuencia se multiplica, la posición de resonancia en la cóclea no se multiplica, simplemente se desplaza una cierta distancia. Por tanto, son las proporciones de frecuencia y no sus diferencias las que determinan el desplazamiento de la región de resonancia de la cóclea. Una relación de esta clase es obviamente una relación de tipo logarítmica.

Conforme un sonido incrementa su amplitud, aumenta la amplitud de la onda viajera en la membrana basilar incrementándose tanto el número de células ciliadas que excitan, como la cantidad de potenciales de acción que generan en la vía aferente. Los centros cerebrales superiores categorizan los tonos con base a la región de la cóclea que se excita, y las amplitudes según el número de neuronas activas y la intensidad con que éstas descargan.

Se ha descrito que existen dos tipos de células ciliadas: las internas y las externas. La gran mayoría – alrededor del 80% - son externas, pero apenas reciben inervación.

---

[22] Este comportamiento de la membrana basilar puede modelarse como un banco de filtros paso banda. Aunque los parámetros que definen dicho banco de filtros se obtendrán a partir de consideraciones psicoacústicas, y no físicas o fisiológicas, se debe tener en cuenta que tal modelo está basado en las propiedades físicas observables de la membrana basilar y del oído interno en general.

Son las 3000-5000 células ciliadas internas las que reciben el 95% de las fibras del nervio auditivo. Es posible afirmar que estas células ciliadas internas son verdaderos sensores del oído. Las externas no operan como receptores, sino como músculos, es decir, como elementos móviles que pueden modificar las oscilaciones en la membrana basilar. Participan como un mecanismo activo en la selectividad frecuencial: para niveles de señal elevados, el movimiento del fluido que rodea los cilios de las células internas es suficiente para doblarlos, y las células externas se saturan. Sin embargo, cuando los niveles de señal son bajos, los desplazamientos de los cilios de las células internas son muy pequeños para activarlas; en este caso las células externas se alargan, aumentando la magnitud de las oscilaciones hasta que se saturan. Por tanto, se comportan como un segundo filtro que determinan qué es lo que se va a transmitir, ya que incrementan y afinan la selectividad frecuencial, generando un pico de respuesta mucho más fino e intenso, por medio de contracciones rápidas y lentas que facilitan o bloquean la transferencia del estímulo hasta las células ciliadas internas. Este es un proceso no lineal de realimentación positiva de la energía mecánica, de modo que las células ciliadas externas actúan como un control automático de ganancia, aumentando la sensibilidad del oído[23].

La forma descrita de respuesta de la cóclea ante el sonido se conoce como teoría del análisis espectral del sonido. En su forma más general fue propuesta en el siglo XIX por el físico alemán Hermann von Helmholtz (Helmoltz, 1874) quien, inspirado en los descubrimientos anatómicos de Alfonso Corti, estableció que en la cóclea existen una serie de resonadores capaces de descomponer sonidos complejos en sus diversas frecuencias. Tras analizar con detalle la estructura del oído, él identificó a los resonadores como las fibras que atraviesan de lado a lado la membrana basilar. Estas fibras varían en su longitud de manera análoga a las cuerdas de un piano,

[23] Este modelo de mecanismo de transducción indica que el conjunto formado por la membrana basilar y sus estructuras anexas forman un sistema activo, no lineal y con realimentación.

incrementando progresivamente su longitud desde la base hasta el ápice de la cóclea. Aunque esta teoría de la resonancia ya no se acepta debido a que no permite explicar la interacción entre tonos, la idea de que los sonidos de diferentes frecuencias activan diferentes regiones de la membrana basilar es la correcta.

Estudios posteriores realizados por el fisiólogo húngaro Georg von Békésy (Premio Nobel de Fisiología y Medicina, 1961, por sus descubrimientos acerca de la fisiología de la audición, (Békésy, 1960)), demostraron que en la cóclea los tonos se distinguen no con base en una serie de resonadores separados, sino debido a las propiedades físicas de la membrana basilar, cuya rigidez decrece gradualmente desde la base hasta el ápice de la cóclea, determinando que, ante los desplazamientos del estribo, se produzca una onda que recorre la cóclea y que produce un desplazamiento máximo en una región específica de la misma.

La representación de la frecuencia e intensidad en el nervio auditivo

Antes de describir la anatomía y fisiología del sistema nervioso auditivo central, se va a hablar sobre la información transmitida a través del nervio auditivo.

El nervio auditivo está dividido en canales, que extraen información en paralelo sobre intensidad (número de fibras estimuladas), temporalidad (módulo de descarga de cada fibra) y espectro (tasa de descarga). De la suma de las respuestas de todas las fibras se obtiene un neurograma con características comunes al espectro del sonido de procedencia.

Debido a que las células ciliadas internas responden a la vibración en una zona concreta de la membrana basilar, cada una de ellas responderá mejor a un tono puro coincidente con su frecuencia característica. Puesto que cada neurona del nervio auditivo está conectada a una única célula ciliada interna, cada neurona estará también centrada en esa frecuencia característica, y llevará información sobre la vibración en una zona concreta de la membrana basilar. De esta forma, la

correspondencia espacial de frecuencias a lo largo de la membrana basilar de la cóclea se transforma en una correspondencia espacial de frecuencias en el nervio auditivo.

Por tanto, como en la etapa anterior, y en las siguientes, las fibras del nervio auditivo son más sensibles a una determinada frecuencia, ante la cual se activan a intensidades menores. Esta frecuencia característica es la misma que la de la célula ciliada correspondiente en la cóclea.

Además de responder a las mismas frecuencias, algunas fibras nerviosas se sintonizan en fase con la onda de la membrana basilar: solo se activan cuando la onda alcanza un determinado punto en su semiciclo (ocurre en frecuencias bajas). De esta manera se consigue el segundo mecanismo para codificar la información sobre frecuencias.

La sincronización temporal permite un control automático de la intensidad: los estímulos fuertes y suaves se traducen en activaciones a intervalos de tiempo similares. Y así se transmite también información de la intensidad del estímulo. Si incrementa la intensidad del sonido, la amplitud de vibración sobre la membrana basilar aumenta y por tanto el desplazamiento de la misma. Las neuronas conectadas a las células ciliadas producirán más impulsos por segundo. Es decir, la frecuencia de disparo (*rate of firing*) de estas neuronas es mayor conforme aumenta la intensidad del sonido. Pero las neuronas del nervio auditivo no pueden disparar a una frecuencia superior a 200 impulsos por segundo. La mayoría de las neuronas alcanzan esta frecuencia de disparo por debajo de un nivel de 60 dB SPL. Las fibras del nervio auditivo se saturan a intensidades relativamente bajas: 60 dB para tonos puros y 80dB para sonidos complejos. A partir de entonces, la respuesta del nervio se deteriora porque se activan muchas fibras a la vez. Sin embargo, se procesan sonidos más intensos, de más de 100 dB, gracias a la sincronización en fase con la onda

viajera que recorre la cóclea: los patrones temporales resisten bien las altas intensidades.

Cada neurona del nervio auditivo lleva información sobre una parte del espectro del sonido. El nivel de las componentes en frecuencia se representa en términos de frecuencia de disparo de la neurona. Si el sonido tiene componentes de alta frecuencia de mayor intensidad que componentes de baja frecuencia, las neuronas conectadas a la parte más exterior del nervio auditivo dispararán con mayor frecuencia que las neuronas próximas al centro del nervio auditivo. Esto se conoce con el nombre de *rate-place code*, ya que la información del espectro es representada en términos de frecuencia de disparo dependiendo de la zona del nervio auditivo.

Así pues, para las frecuencias bajas en las que la mecánica coclear es menos eficaz (intensidades altas), el nervio auditivo dispone de dos mecanismos que se complementan mutuamente: el frecuencial, la proporción de fibras que se activan ante un determinado estímulo (400 a 500 disparos por segundo), o el temporal, los intervalos de activación entre ellas, que se traducirán en información sobre frecuencias.

Existe otro modo de traducir la información del sonido al nervio auditivo. Se ha descrito que la membrana basilar se mueve hacia arriba o hacia abajo en función de la presión de la onda sonora; y se sabe que los impulsos nerviosos son mayores cuando se mueve en una determinada dirección. Se puede decir, por tanto, que el patrón de disparo es *phase locked* a la vibración de la membrana y por tanto a la variación de presión. Por ejemplo, cuando un tono puro de 100 Hz llega al oído, la membrana basilar vibra hacia arriba y hacia abajo a una frecuencia de 100 Hz, con una vibración máxima próxima al ápice de la cóclea. Las neuronas conectadas a esta parte de la membrana tienden a producir impulsos cada 10 ms. Aunque las neuronas no pueden disparar a frecuencias superiores a 200 impulsos por segundo, continuarán en fase a frecuencias mayores.

Phase locking implica que cada neurona del nervio auditivo no solo lleva información sobre el nivel de actividad que tiene lugar en una zona concreta de la cóclea , sino que también lleva información sobre el patrón de vibración de la membrana basilar en dicha zona. Esta información se usará para la percepción del tono y la localización de los sonidos en el espacio.

En el nervio auditivo, encontramos además un comportamiento no lineal, tanto ante la duración del estímulo (la fibra se cansa y deja de responder a partir de los primeros 15-20ms) como ante las frecuencias (enmascaramiento de un componente de la onda sobre otro). Ambos fenómenos permiten resaltar las diferencias entre sonidos sucesivos: explotar la dinámica de la señal de habla y sacar partido al eje temporal.

Sistema Nervioso Auditivo Central

El sistema nervioso auditivo central está formado por las vías auditivas y los sectores de nuestro cerebro dedicados a la audición.

Las vías auditivas convierten lo que en la cóclea eran activaciones individuales de las células en patrones de actividad neuronal. En los distintos núcleos que conforman estas vías, encargadas de llevar la señal auditiva hasta la corteza del cerebro, se producen una enorme cantidad de interconexiones neuronales.

LAS VÍAS AUDITIVAS

Las vías auditivas (núcleo coclear del Bulbo, complejo olivar superior, lemnisco lateral, colículo inferior y cuerpo geniculado medio) son un complejo sistema de haces neuronales que van y vienen, arriba y abajo, en el tronco encefálico, para acabar en la corteza auditiva, Figura 29. Es el conjunto de interconexiones más complicado de todo el sistema sensorial: el 70% de las vías son contralaterales cruzan

desde un oído a la corteza cerebral del lado contrario; el resto son ipsilaterales[24]. Hay vías nerviosas que atraviesan todas las estaciones intermedias hasta llegar a la corteza, mientras que otras se saltan algunos núcleos. Esta complejidad es necesaria para compensar la inicial pobreza neurológica del sistema auditivo. Si lo comparamos con el visual, se parte de un déficit del 0,01%[25]. Al llegar al cerebro se equilibran: encontramos unos cien millones de neuronas, tanto en el cortex visual como en el auditivo (Handel, 1993).

Las vías auditivas mantienen en todas sus estaciones la organización tonotópica de la cóclea y del nervio auditivo: cada neurona tiene una frecuencia característica, en la cual la intensidad necesaria para activarla es menor, tiene un umbral más bajo.

A medida que avanzamos hacia la corteza cerebral, las neuronas tienden a responder mejor a las partes dinámicas del habla (transiciones, movimientos de los formantes, inicios, finales: puntos de cambio espectral) y a las variaciones de la frecuencia fundamental.

---

[24] Ipsilateral se refiere a elementos en el mismo lado con respecto al plano medio; contralateral a elementos en lados opuestos; y bilateral a elementos presentes a ambos lados.

[25] La retina tiene 130 millones de receptores fotosensibles, frente a los 15000 células ciliadas del oído; el nervio óptico está formado por un millón de fibras, el auditivo solo contiene 40000 – 50000 fibras.

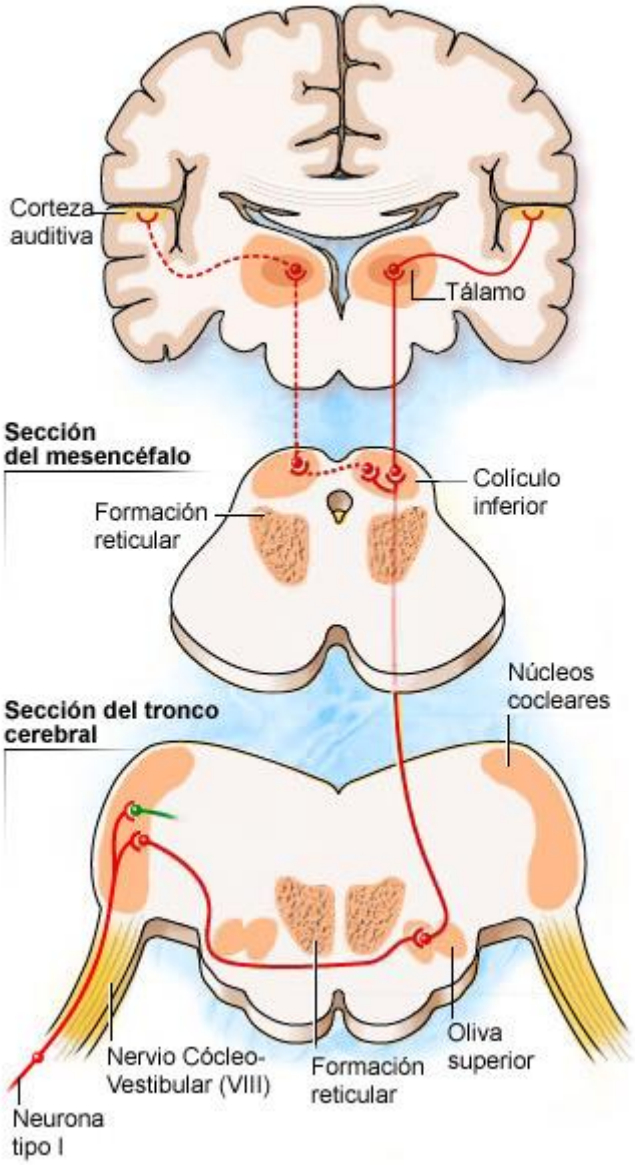


Figura 29. Representación esquemática de la vía auditiva.

NÚCLEO COCLEAR DEL BULBO

En esta estructura ocurre la primera sinapsis del Sistema Nervioso Auditivo Central. Esta se divide en dos ganglios: dorsal y ventral. El primero es el encargado de recoger pequeñas variaciones de frecuencia del sonido, y de analizar la calidad acústica del mismo para disminuir el ruido de fondo. Sus fibras se dirigen directamente al colículo inferior, a través de un tracto nervioso llamado lemnisco lateral. El segundo, con el fin de que el primer ganglio realice su función, conserva la señal auditiva durante microsegundos. Las neuronas de los núcleos cocleares tienen la particularidad de ser excitadas por tonos que estén dentro de un determinado rango e inhibida por tonos que estén fuera de éste. Tienen una distribución tonotópica.

COMPLEJO OLIVAR SUPERIOR

Representa la primera estación en la cual es posible la integración binaural en la localización del sonido. Se divide en dos núcleos: medial y lateral. El primero se relaciona con la localización del sonido, el segundo con la ubicación del estímulo sonoro en base a las diferencias de intensidad interaural.

COLÍCULOS INFERIORES

Son el mayor sitio de integración binaural, sus neuronas realizan el análisis temporal de las estructuras del sonido y el mapeo de los eventos auditivos para lograr su localización. Gran parte de sus fibras aferentes provienen desde el complejo olivar superior, a través del lemnisco lateral.

CUERPO GENICULADO MEDIAL

Es un núcleo talámico que recibe entradas de los colículos inferiores y corresponde a la cuarta sinapsis de la vía auditiva aferente. Sus células son sensibles para combinaciones de frecuencias en intervalos temporales específicos. Se divide en una

porción ventral, que se proyecta hacia la corteza auditiva primaria, y una dorsal, que asciende a la secundaria. La porción ventral es la primera estación de relevo auditivo y tiene una organización tonotópica.

LA CORTEZA CEREBRAL

La corteza auditiva está compuesta por neuronas con una alta conectividad y organizadas en columnas jerárquicas: las del hemisferio izquierdo presentan un diámetro mayor y más espacio entre ellas que las del derecho. Se suelen distinguir en ellas dos zonas: el área auditiva primaria y la secundaria.

El área auditiva primaria se encuentra en el interior de la cisura de Silvio, perteneciente al lóbulo temporal (entre sus funciones destacan la audición, el aprendizaje, la memoria y las emociones). Está organizada tonotópicamente en bandas de isofrecuencias: las frecuencias bajas activan las zonas más externas de la cisura de Silvio, y las altas zonas más profundas. En cuanto al procesamiento temporal, en esta área se encuentran dos poblaciones de neuronas. Las neuronas sincronizadas, que responden a cambios temporales lentos (más de 20 ms), son las encargadas de procesar el tono del sonido. Y las neuronas no sincronizadas, que responden a cambios temporales más rápidos (10-20 ms), que les permite detectar las fronteras entre sonidos (Gil Loyzaga, 2005).

El área auditiva secundaria se localiza en el área de Wernicke, en la superficie lateral del lóbulo temporal izquierdo. Una zona asociativa superior muy importante para el procesamiento del lenguaje, encargada de la descodificación de las unidades lingüísticas. No tiene distribución tonotópica. Se encarga de la localización del sonido y del procesamiento de patrones auditivos complejos.

Alrededor de ambas áreas se encuentra la región periférica, donde la audición se integra con el resto de los sistemas sensoriales, denominada área de integración sensorial polimodal. Es donde la información visual y auditiva entran en relación.

En esencia, el papel de las regiones corticales consiste en realizar las funciones superiores de integración del mensaje oral. El procesamiento de los eventos auditivos forma un patrón de excitación de codificación de la información con carácter definitivo, mucho menos variable que el espectro acústico (Greenberg, 1996). Las neuronas actúan aquí por comparación con patrones aprendidos.

3.2.2. Psicoacústica del habla

Hasta ahora se ha visto que las distintas partes del sistema auditivo son susceptibles de ser modeladas matemáticamente, en términos de su comportamiento como sistemas físicos.

Se podría, por tanto, pensar que el modelo perceptual ideal es aquel que simula, en términos de los procesos físicos y fisiológicos, todas las etapas del sistema auditivo, incluyendo la etapa de procesamiento neural en el cerebro. Sin embargo, la comprensión que se tiene acerca de lo que ocurre en las estructuras cerebrales es muy limitada, especialmente en lo relativo a los centros superiores del cerebro. Por lo tanto, es necesario recurrir a la descripción psicoacústica de los fenómenos perceptuales y de las sensaciones. A continuación se enumeran los objetivos de la psicoacústica:

- Caracterizar la respuesta de nuestro sistema auditivo.
- Obtener el umbral absoluto de la sensación.
- Obtener el umbral diferencial de determinados parámetros de los estímulos, estos umbrales son la mínima variación y mínima diferencia perceptibles.
- Comprender y obtener la capacidad de resolución del sistema auditivo para separar estímulos simultáneos separados para crear sensaciones.

- Entender la variación temporal de la sensación del estímulo

Por tanto, el estudio de audición a través de las respuestas subjetivas a los estímulos acústicos, especialmente en tareas de detección y discriminación, es el objetivo de la psicoacústica o psicofísica auditiva. La señal de habla que aparece en un sonograma no es idéntica al que llega a la corteza auditiva; entre ambas, el sistema auditivo ha podido suprimir algunos elementos o reforzar otros.

Hay que destacar la diferencia entre los conceptos de detección, discriminación e identificación. La detección implica notar la presencia o ausencia de un estímulo, pero sin llegar a identificarlo; se puede detectar estímulos en función de su duración, intensidad y frecuencia. En la discriminación se compara y se buscan diferencias entre estímulos próximos. Y en la identificación, se relaciona el estímulo que se presenta con una representación que tenemos en la memoria, a la cual corresponde una etiqueta determinada. Las tareas de identificación conciernen a la percepción, no a la audición.

Umbrales absolutos

El oído presenta unos límites en su capacidad descodificadora, determinados por la intersección de las tres cualidades físicas principales del sonido: frecuencia, intensidad y duración.

En cuanto a las frecuencias, podemos oír los sonidos entre 20 y 20000 Hz, pero somos especialmente sensibles a los que se sitúan entre 2500 y 5000 Hz. Fuera de estos márgenes están los infrasonidos y los ultrasonidos. Respecto a las intensidades, también existen dos extremos; el umbral de la audición o umbral absoluto (0 dB, la intensidad mínima para distinguir un sonido del silencio) y el umbral del dolor. Cuando se habla de distinguir sonidos, se hace referencia al umbral diferencial, la menor diferencia que puede ser detectada entre dos estímulos (se describe en el siguiente apartado). Y sobre la duración, sólo existe un límite inferior. El sonido

lingüístico más breve perceptible puede oscilar entre 10 y 40 ms. Y la mayor sensibilidad natural aparece en el rango de 40 a 60 ms (López Bascuas, 1997).

Hay que considerar lo que ocurre cuando se entrecruzan estas tres categorías y especialmente las dos primeras, porque no todas las frecuencias requieren la misma intensidad para ser percibidas. Existe una zona, de 500 a 8000 Hz, mucho más sensible que el resto, donde apenas es necesario subir de 0 dB para percibir el sonido; por encima y por debajo se requiere incrementar la intensidad significativamente para alcanzar los mismos resultados. Este efecto se puede apreciar en las curvas de audibilidad de la siguiente Figura 30.

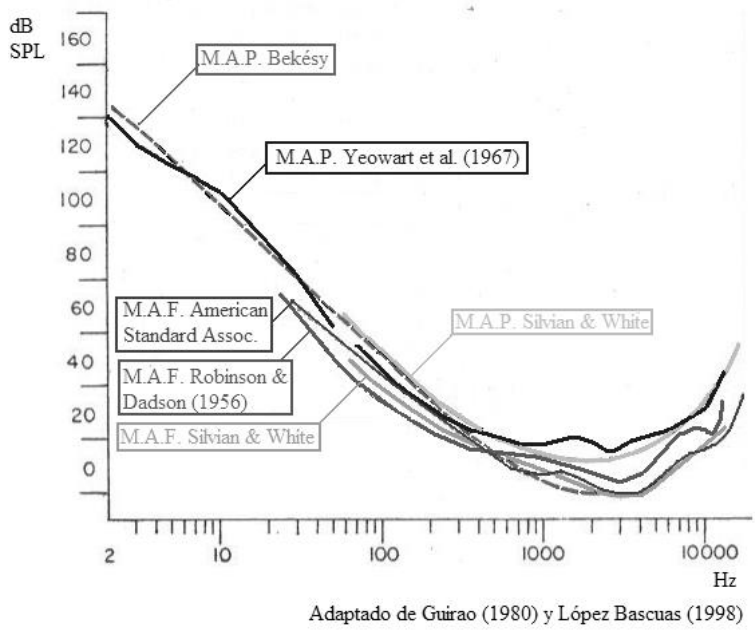


Figura 30. Curvas de audibilidad[26] (*M.A.P., Minimal Audible Pressure*; obtenido mediante auriculares. *M.A.F., Minimal Audible Field*; obtenido en campo libre).

[26] Las curvas de audibilidad de la figura muestran dos formas de medir el umbral de audibilidad: la mínima presión audible (*MAP,* del ínglés *Minimal Audible Pressure*) y el mínimo campo audible (*MAF,* del

En la intersección entre los topes en frecuencias (infrasonidos/ultrasonidos) y los de intensidad (umbral de audición/umbral de dolor) se obtiene el campo de audición o área de respuesta auditiva. La siguiente Figura 31, conocida como curva de Wegel, muestra los umbrales de la audición humana y, dentro de ellos, los márgenes utilizados habitualmente por la música y el lenguaje articulado. En ella, se observa como dos intensidades muy diferentes pueden producir la misma sensación subjetiva de fuerza: un tono de 30 Hz necesitará 65 dB para producir la misma sensación perceptiva que otro de 1000 Hz, con apenas 0-2 dB.

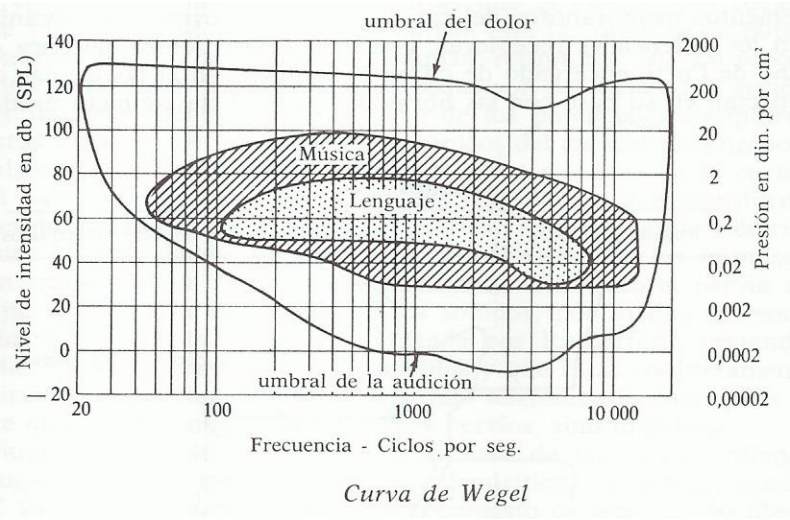


Figura 31. Campo de audición.

En contextos psicofísicos aparece más a menudo otra forma de plasmar esa misma información: las curvas de isofonía que se muestran en la Figura 32, izquierda;

---

inglés *Minimal Audible Field*). El *MAP* se mide colocando pequeños micrófonos dentro del canal auditivo. La información se envía a través de auriculares. El *MAF* se mide en ausencia del sujeto, colocando un micrófono en el lugar que ocupaba la cabeza del sujeto.

una representación más detallada que explica qué ocurre con cada frecuencia dentro del campo de audición aparece en la Figura 32, derecha. Estas curvas de igual sonoridad representan la sensibilidad del oído a distinta frecuencias y niveles. Hay que destacar que la sensibilidad del oído es más lineal a medida que se incrementa el nivel de audio. Es decir que la diferencia de percepción entre la zona de mayor sensibilidad (medios) y las menos sensibles (graves y agudos) se hace más pequeña. Por este motivo si se escucha música a bajo volumen se oye menos los agudos y, especialmente, los graves. Al aumentar el nivel de escucha, las bajas y altas frecuencias aparecen. Esto es porque al aumentar el nivel estamos escuchando en una zona más lineal del oído (cuervas superiores).

La curva inferior se denomina umbral de audición e indica el nivel mínimo percibido para cada frecuencia (por definición, la referencia 0 dB corresponde al umbral de audición para 1000 Hz). La curva superior se llama umbral de dolor y representa la máxima presión sonora que soporta el oído antes de dañarse.

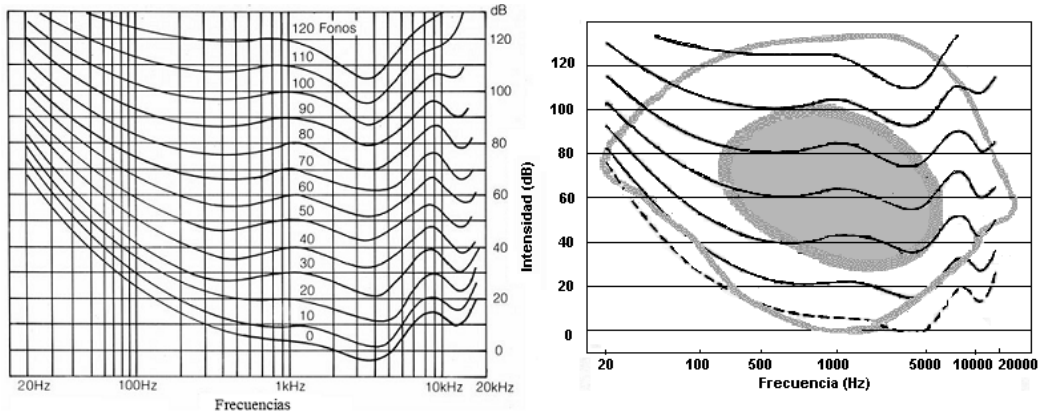


Figura 32. Curvas de isofonía (izquierda). Curvas de isofonía y campo de audición (derecha).

También existe una relación entre el tiempo e intensidad, ya que los sonidos inferiores a 0,3s requieren más energía para ser percibidos (0'Shaughnessy, 1987).

Estos umbrales varían a lo largo de la vida: en la vejez, es muy frecuente la pérdida de capacidad auditiva, mientras que los niños tienen un umbral de audición más alto que los adultos (hasta 15-25 dB), tanto para tonos puros como para el habla. Algunas investigaciones han sugerido que los bebes de 6 a 8 meses también necesitan una relación señal/ruido más alta que los adultos para detectar el habla y para hacer distinciones fonéticas.

Umbrales diferenciales: Las diferencias mínimas perceptibles

Las diferencias mínimas perceptibles, DPM, son unidades utilizadas desde Weber y Fechner para medir los umbrales de audición. Son importantes en los estudios de percepción, porque miden la capacidad de resolución del oído y los límites de audición; y en estudios de ingeniería lingüística, porque definen con cuánta precisión deben cuantificarse los parámetros del habla para la transmisión de datos. Se definen como la menor diferencia que puede ser detectada entre dos estímulos.

A continuación se describen las diferencias mínimas perceptibles en frecuencia, en intensidad y en duración.

En frecuencias somos capaces de percibir diferencias mínimas que van de 0,5 a 2 Hz, aunque se considera que por debajo de 1000 Hz, las DMP son de 1-3 Hz. Es interesante destacar que el sistema auditivo actúa como un conjunto de filtros superpuestos; estos filtros son más estrechos en frecuencias graves y más anchos en frecuencias agudas, y van a definir las llamadas bandas críticas (este concepto se describe en el siguiente apartado).

En cuanto a la intensidad en el campo de audición, de 0 a 110 dB, una persona que oiga normalmente puede detectar más de 100 escalones. Desde que se empieza a oír un sonido, hasta que su intensidad nos hace daño, se incrementa su amplitud en un factor de 10 millones.

Según los estudios iniciales de Flanagan (Flanagan, 1972), el umbral diferencial de intensidad promedio para formantes en estímulos lingüísticos es de 2 dB. Esa media se distribuye de forma desigual en el espectro: el primer formante, F1, requiere tan solo 1,5dB, mientras que el F2 debe duplicar esa cifra para provocar un cambio perceptivo; las frecuencias interformánticas necesitan alcanzar incluso 13 dB.

Respecto a las duraciones, la resolución temporal del oído es buena para estímulos entre 10 y 100 ms, y de banda ancha, característicos del habla. 20 ms es el tiempo característico de integración en el procesamiento auditivo; el sistema auditivo posee una mayor sensibilidad natural para el rango que va de 40 a 60 ms. Aunque somos capaces de percibir en el habla diferencias más breves, entre 10 y 40 ms.

Las duraciones vocálicas del español han sido estudiadas por Pàmies y Fernández Planas, (Pamies & Fernández Planas, 2006). Su resultado proporciona un umbral perceptivo diferencial del 35,9%: el incremento más pequeño que podemos percibir sobre la vocal más breve sería de unos 36 ms. Es un umbral representativo de la media sobre duraciones en habla natural.

Si se combina los tres datos anteriores, la cifra de sonidos potencialmente discriminables por el oído humano es muy elevado. Sin embargo, ninguna lengua natural presenta más de 90 fonemas. Una de las razones es la diferencia entre discriminar e identificar: se puede detectar pequeñas diferencias entre estímulos, pero nuestra capacidad para almacenarlas en la memoria y etiquetarlas como unidades es mucho más limitada.

Enmascaramientos y bandas críticas

En general, cuando la presencia de un estímulo interfiere con la percepción de otro, se dice que el primero está enmascarando al segundo. La definición oficial (de la *American Standards Association*, 1960) de este fenómeno es el proceso por el cual el

umbral de audibilidad de un sonido aumenta debido a la presencia de otro sonido, al que se le llama máscara. Es una respuesta no lineal del sistema auditivo.

En las curvas de enmascaramiento (Figura 33) se observa que el efecto enmascarado es más fuerte en la frecuencia que coincide con la de la máscara. Todo lo que queda por debajo de la línea base es la señal enmascarada, que no puede oírse. Cuanto mayor es la intensidad de la máscara, más amplia es el área enmascarada. Este efecto se incrementa más en las frecuencias agudas que en las graves. Por tanto, los sonidos graves enmascaran a los agudos y no a la inversa.

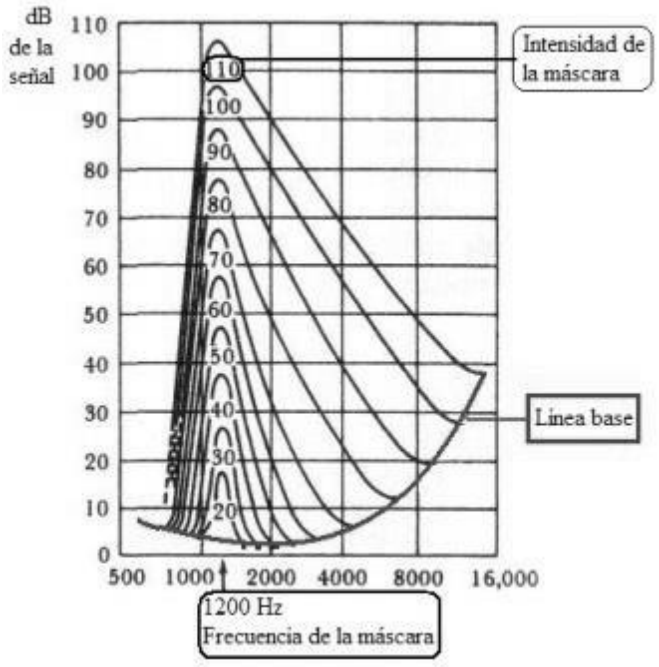


Figura 33. Curvas de enmascaramiento

Se ha descrito en este capítulo la capacidad del sistema auditivo, concretamente la cóclea, para discriminar frecuencias y sonidos. Una característica importante del sistema auditivo es el efecto de enmascaramiento (0'Shaughnessy, 1987), según el

cual la percepción de los sonidos se ve oscurecida o impedida por la presencia de otros sonidos. El enmascaramiento puede ser frecuencial si los dos sonidos se producen simultáneamente: el sonido de menor frecuencia enmascara el sonido de mayor frecuencia. Y por otro lado, está el enmascaramiento temporal, si los dos sonidos se producen con un cierto retraso. El efecto de enmascaramiento es importante por su influencia en la no linealidad del sistema auditivo y perceptivo humano.

Una forma de entender el funcionamiento del sistema auditivo es suponer que contiene una serie o banco de filtros paso banda solapadas conocidos como filtros auditivos (Fletcher, 1940). Estos filtros se producen a lo largo de la membrana basilar y tienen como función aumentar la resolución de frecuencia de la cóclea y así incrementar la habilidad de discriminar entre distintos sonidos. Este banco de filtros no sigue una configuración lineal, y el ancho de banda y morfología de cada filtro depende de su frecuencia central. Los filtros correspondientes al extremo más próximo a la ventana oval y al tímpano responden a altas frecuencias, ya que la membrana es rígida y ligera. Por el contrario, en el extremo más distante, la membrana basilar es pesada y suave, por lo que los filtros correspondientes responden a las bajas frecuencias.

Por ello, los investigadores emprenden trabajos psicoacústicos experimentales para obtener las escalas de frecuencias que modelen la respuesta natural del sistema de percepción humana. El estudio sobre el ancho de banda del enmascaramiento en el oído permitió descubrir el sistema de bandas críticas que caracteriza la audición. Se denomina así a cada una de las fronteras o límites en que el sistema auditivo agrupa las frecuencias para su procesamiento.

Una banda crítica puede ser considerada como un filtro paso banda, cuya respuesta al impulso coincide con la respuesta de enmascaramiento de las neuronas auditivas. Nos definirán un intervalo de frecuencias, en el que la percepción

psicoacústica se ve modificada de manera abrupta. Ante dos sonidos que compiten en una misma banda crítica, el que tenga mayor energía predominará en nuestra percepción y enmascarará a su competidor.

Así, los estímulos que coinciden en el interior de una banda crítica interactúan perceptivamente (el más intenso predomina), pero los que pertenecen a bandas diferentes se analizan de modo independiente.

Cada una de las bandas críticas se corresponde aproximadamente con una región de 1 a 5 mm a lo largo de la membrana basilar. Estos datos implican que con 24 filtros paso banda (con un ancho de banda mayor a medida que la frecuencia central crece), se puede modelar correctamente la membrana basilar y su comportamiento.

Con objeto de aproximarse a la sensibilidad del oído humano, que no tiene una respuesta lineal, existen diferentes escalas para definir cuantas bandas críticas existen en el sistema auditivo y cuál es la frecuencia central de cada una. A continuación se describen la escala de *Bark*, escala de *Mel* y la escala *ERB*.

ESCALA DE BARK.

Existe una escala de medición de las bandas críticas llamada escala de *Bark*, propuesta por Zwicker (Zwicker, 1961). Esta escala se muestra en la Tabla 9. La primera banda crítica se expande desde la frecuencia 0 a 100 Hz, la segunda desde los 100 Hz a 200 Hz y así hasta los 500 Hz, donde el rango de frecuencias de cada banda incrementa. A partir de este punto, con el aumento de la frecuencia, el ancho de banda crítico va aumentando, de manera logarítmica por encima de 1 kHz. Esta es la escala que se va a usar en el sistema presentado en este trabajo.

Tabla 9. Escala de *Bark* para estimación de las bandas críticas del sistema auditivo

Nº de banda crítica (Bark)	**Frec. central (Hz)**	**Frec. superior (Hz)**	**Ancho de la BC (Hz)**
1	50	100	100
2	150	200	100
3	250	300	100
4	350	400	100
5	450	510	110
6	570	630	120
7	700	770	140
8	840	920	150
9	1000	1080	160
10	1170	1270	190
11	1370	1480	210
12	1600	1720	240
13	1850	2000	280
14	2150	2320	320
15	2500	2700	380
16	2900	3150	450
17	3400	3700	550
18	4000	4400	700
19	4800	5300	900
20	5800	6400	1100
21	7000	7700	1300
22	8500	9500	1800
23	10500	12000	2500
24	13500	15500	3500
25	18775	22050	6550

En la siguiente Figura 34 se observa como el rango de frecuencias audible hasta 16kHz se divide en 24 bandas críticas. Su unidad de medida es el *Bark*.

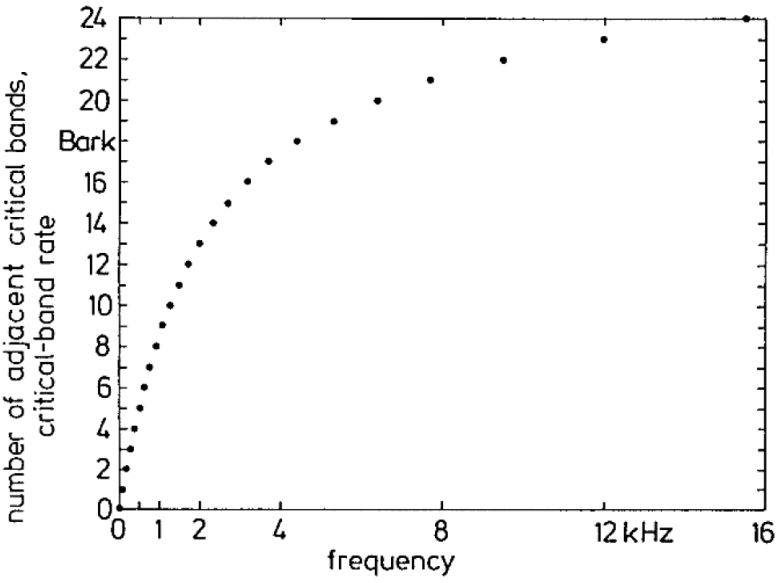


Figura 34. Representación de la escala de *Bark*.

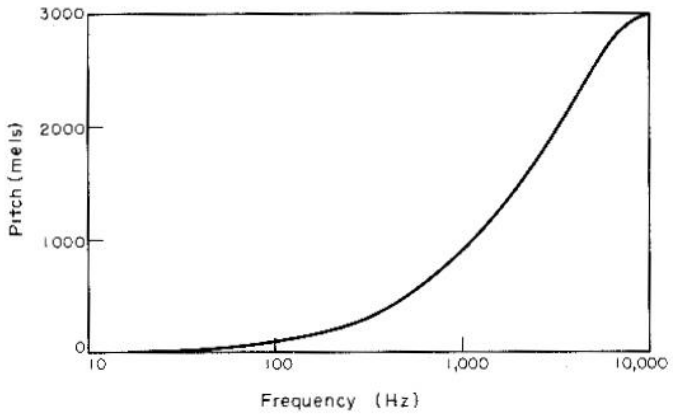
La escala *Bark*, relaciona frecuencias acústicas con frecuencias perceptuales, de tal manera que cada *bark* cubre una banda crítica. En la siguiente (Ecuación 3.1) se expresa una fórmula analítica que relaciona frecuencias ***(f)*** con *barks* ***(b)*** (Fastl & Zwicker, 2007):

$$b = 13\,arctan\left(0.76\,\frac{f}{kHz}\right) + 3.5\,arctan\left(\frac{f}{7.5kHz}\right)^2 \tag{3.1}$$

ESCALA DE *MEL*

La escala *Mel* fue propuesta por (Stevens, Volkman, & Newman, 1937). El nombre *Mel* deriva de melodía, como una forma de justificar que se trata de una escala basada en comparaciones entre frecuencias. La escala *Mel* se construye equiparando un tono de 1000 Hz a 40 dBs, por encima del umbral de audición del oyente, con un tono de 1000 *Mels*. Sobre los 500 Hz, los intervalos de frecuencia espaciados

exponencialmente son percibidos como si estuvieran espaciados linealmente (400Hz-500mels; 1000Hz-1000mels; 3000Hz-2000mels). Esta relación se observa en la Figura 35. En consecuencia, sobre este punto, cuatro octavas en la escala lineal de frecuencias medida en Hz se comprimen alrededor de dos octavas en la escala *Mel*.



Relationship between pitch and frequency as determined by the method of fractionation (Stevens and Volkmann, 1940).

Figura 35. Representación de la escala de *Mel*.

La escala *Mel* puede ser aproximada en función de la frecuencia lineal como se expresa en la (Ecuación 3.2):

$$\hat{f} = 2595\, log_{10}\left(1 + \frac{f}{700}\right) \tag{3.2}$$

ESCALA ERB

Otro concepto asociado a la banda crítica es el equivalente rectangular de ancho de banda o *ERB* (*Equivalent Rectangular Bandwidth*). Esta medida muestra la relación entre el conducto o canal auditivo y el ancho de banda de un filtro auditivo. La idea de esta medida es remplazar una banda crítica por un rectángulo equivalente cuya altura es el máximo de la respuesta en magnitud del filtro y cuya área es igual a la respuesta de

dicho filtro, de manera que permite en su interior la misma cantidad de energía. Esta relación se muestra en la siguiente Figura 36.

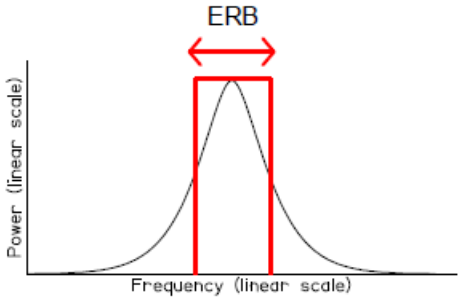


Figura 36. Representación del *ERB*

De esta forma, (Glasberg & Moore, 1990), introdujeron una aproximación al ancho de banda crítica utilizando la escala *ERB*. Para niveles de potencia moderados, podemos aproximar el *ERB* por la siguiente expresión (Ecuación 3.3):

$$ERB(f_c) = 24.7\,(4.37\,\frac{f_c}{1000} + 1) \tag{3.3}$$

donde ERB y f_c están expresadas en Hz. A esta ecuación se le conoce con el nombre de función *ERB* (Figura 37).

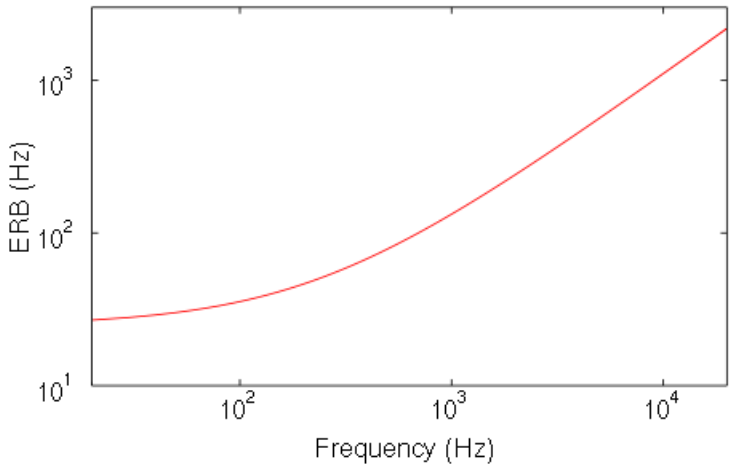


Figura 37. *ERB* relacionado con la frecuencia de acuerdo a la fórmula de Moore y Glassberg.

Una escala derivada del *ERB* y la cual resulta muy útil para tener una aproximación del patrón de excitación que produce una señal en la membrana basilar, es la escala *ERB* (Ecuación 3.4):

$$ERB_{number}(f_c) = 21.4 \log_{10}(4.37 \frac{f_c}{1000} + 1) \quad (3.4)$$

Esto nos indica el número *ERB* (ERB_{number}) en función de la frecuencia fc en Hz. Un incremento de un ERB_{number} se corresponde con un incremento de 0,9 mm en la membrana basilar. Esta escala es similar a otras escalas auditivas como la *Bark* de Zwicker y la *Mel* de Steven. En la siguiente Figura 38 se muestra la escala *ERB* en comparación con la escala de *Bark* y la escala *Mel*. En ella se aprecia la no linealidad respecto a la frecuencia.

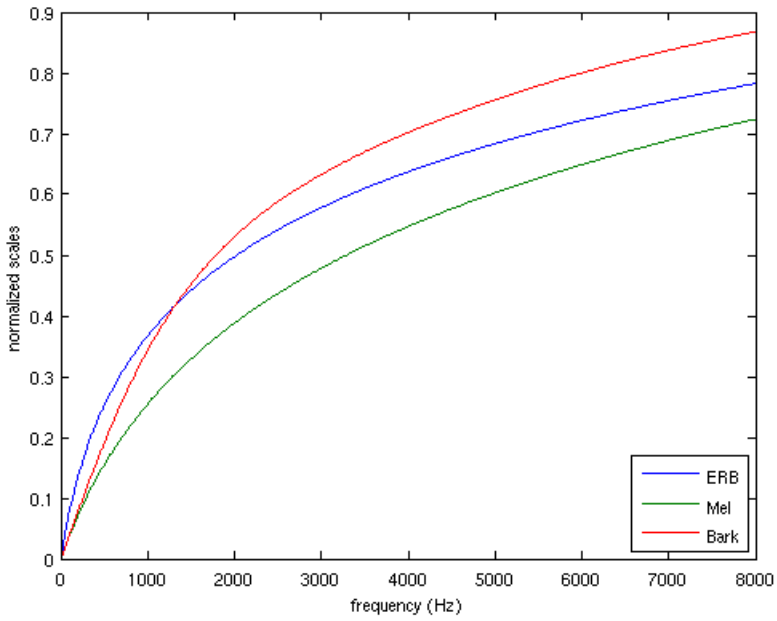


Figura 38. Comparación entre la escala *ERB*, *Mel* y *Bark*.

3.2.3. Los rasgos distintivos

Algunos estudios se han centrado en la percepción de los rangos distintivos articulatorios: modo de articulación, lugar de articulación y sonoridad/sordez (0'Shaughnessy, 1987):

El modo de articulación es el rasgo articulatorio más resistente. Las diferencias entre oclusión, fricación, etc. se han mantenido en circunstancias adversas con más contundencia que las que oponen sonidos sordos a sonoros, o labiales a dentales, palatales o velares. Sus claves perceptivas parecen residir en frecuencias inferiores a 1000 Hz.

El lugar de articulación, presenta la mayor fragilidad. Las frecuencias más relacionadas con su percepción son las situadas por encima de los 1000 Hz, especialmente la zona de los segundos formantes. Se ve afectado por la

superposición de ruido de banda ancha (por su incidencia en F2 y F3) y por la reverberación.

La sonoridad depende de la estructura de los armónicos, más fuerte en frecuencias muy bajas, pero que se mantiene incluso hasta los 3000 Hz.

Sin embargo, es difícil encontrar explicaciones partiendo solo de datos articulatorios, porque la distancia entre la clasificación articulatoria y la percepción es demasiado grande. Así, otros estudios han tomado como referencia las tradicionales clasificaciones acústicas. Concretamente el de la lingüista Vasanta Duggirala, quien establece una relación (Tabla 10) entre los rasgos acústicos y una determinada frecuencia crítica determinante para su percepción (Duggirala, Studebaker, Pavlovic, & Sherbecoe, 1988).

Tabla 10. Relación entre rasgos acústicos y frecuencia.

Nasalidad	472 Hz	Densidad	1618 Hz
Sonoridad	758 Hz	Continuidad	1800 Hz
Gravedad	1290 Hz	Estridencia	2521 Hz

Estos datos han sido confirmados en experimentos sobre el español con oídos patológicos: existe una relación entre cada rasgo distintivo y una banda de frecuencia determinada: a mayor pérdida auditiva en esa frecuencia, peores resultados en el rasgo distintivo correspondiente (Marrero Aguiar, Santos, & Cárdenas, 1993) y (Marrero Aguiar & Martín, 2001).

3.2.4. La percepción de las vocales

Casi un axioma en fonética acústica, originado en los laboratorios Haskins durante la década de los 50, ha sido considerar el primer y segundo formantes como las claves acústicas definitivas para la descripción y clasificación de las vocales. Pero, cada tracto bocal origina modelos espectrales y distribuciones formánticas diferentes. Se

ha descrito que la altura de las formantes y de la frecuencia fundamental varia con la edad y el sexo; la altura de los formantes producidos por los niños decrece con la edad y en las niñas es mayor que en los niños; y los valores de F0 varían de forma inversamente proporcional al incremento de la edad y no existen diferencias significativas respecto al sexo. Por lo tanto, es necesario un proceso de normalización para identificarlas.

Los formantes superiores, F1 y F2, se han considerado tradicionalmente los responsables de las características individuales del habla. Sin embargo, el F3 también puede resultar imprescindible para algunas vocales de lenguas como el inglés, con muchas más unidades que el español.

El español presenta datos excelentes, según Borzone de Manrique: 97% de identificación correcta en vocales naturales aisladas, y del 99% en contexto (Borzone de Manrique, 1979).

La frecuencia fundamental, F0, se ha revelado como un elemento muy importante en la percepción de las vocales. Halberstom y Raphael, (Halberstom & Raphael, 2004), la consideran más relevante que el F3 para normalizar las diferencias entre locutores[27]. Traunmüller, (Traunmüller, 1987), defiende que información sobre la fase en el F0 sería suficiente para diferenciar la cualidad de la vocal si el tono es suficientemente bajo (voz masculina). Las diferencias entre la representación perceptiva y la representación acústica de una vocal pueden estar en la base de esta revalorización del F0: en los espectros auditivos (Figura 39, derecha) la frecuencia fundamental presenta la misma resolución que los formantes, y ocupa un área

[27] Un niño o una mujer, tienen un F0 elevado, que inconscientemente se asocia a formantes más altos que los que corresponden a un F0 masculino.

perceptiva notablemente mayor que la obtenida en una representación acústica, lo cual también ocurre con los dos primeros formantes (Figura 39, izquierda)[28].

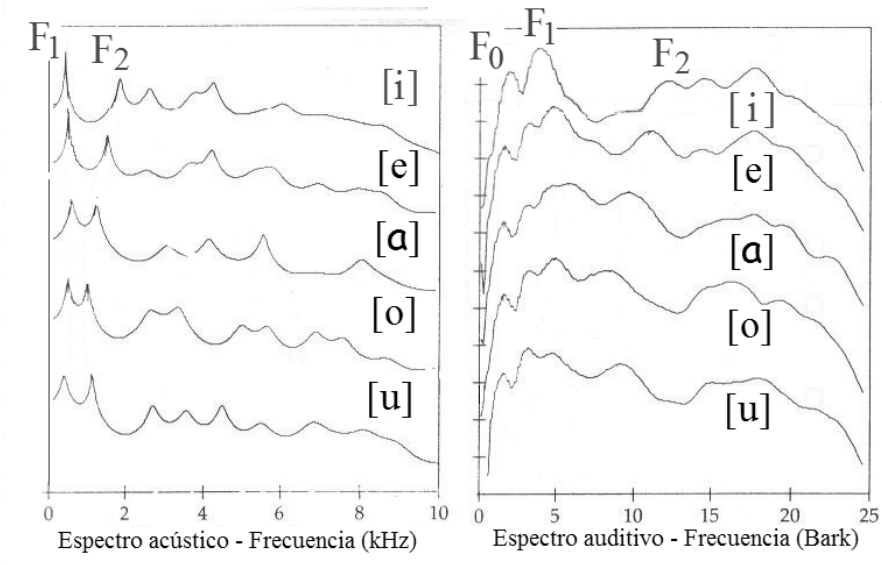


Figura 39. Espectros acústicos (izquierda) y perceptivo (derecha) de las vocales del castellano. Adaptado de (Johnson, 1997).

No se perciben de la misma manera las vocales anteriores o posteriores, abiertas o cerradas. Ludmilla Chistovich, propone la existencia de un centro de gravedad perceptivo determinante para ciertas vocales. Mientras las anteriores (/i/, /e/) dependen del F1 y el F2, las posteriores (/o/, /u/) debido a la cercanía entre ambos (3-3,5 *barks*) conlleva su integración en una sola clave: el centro de gravedad

[28] Las propiedades del sistema auditivo son las responsables de estas diferencias: las bandas críticas, suficientemente estrechas en bajas frecuencias, actúan como un mecanismo de integración frecuencial, fusionando los primeros armónicos del F0 y resolviéndolos como un pico separado. La respuesta no-lineal de la periferia auditiva, refuerza la importancia de las regiones F1 y F2.

perceptivo, situado entre los dos formantes y característico de cada vocal (Chistovich, Sheikin, & Lublinskaja, 1979).

También, en las vocales cerradas el F1 está muy cerca del F0, provocando la integración perceptiva, dominada por el punto más cercano entre ambos elementos: el armónico de F0 más cercano al pico de F1. La existencia de este mecanismo en español ha sido estudiada por López Bascuas. Según sus resultados, la frontera entre vocal alta/baja se establecería cuando la distancia entre F0 y F1 estuviera por debajo o por encima de 2,5*barks*.

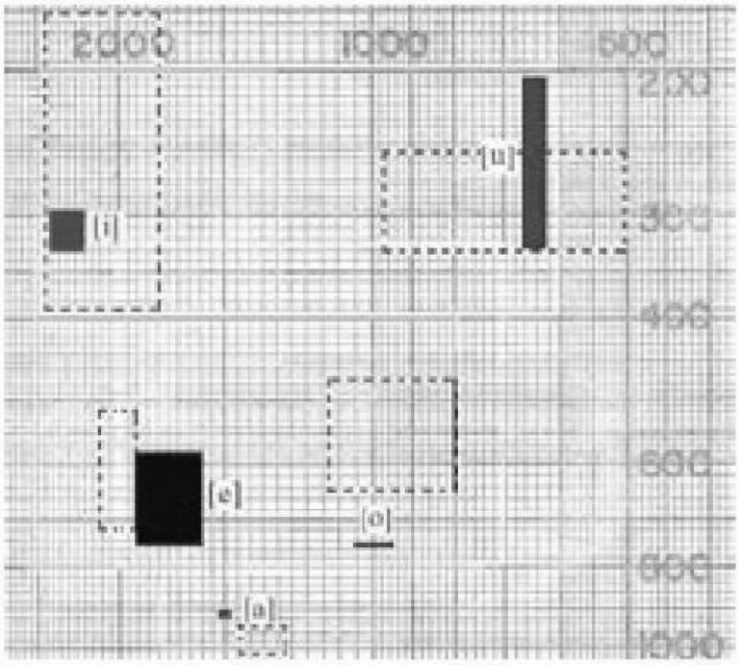


Figura 40. Campos de dispersión perceptivos de las vocales españolas. (Romero, 1989) – rectángulos negros – y (Fernández Planas, 1993) – trazo discontinuo.

En la Figura 40 se muestra el resultado de los estudios de campos de dispersión perceptivos[29] realizados para el español peninsular. En la imagen, se muestran los datos obtenidos por Romero, (Romero, 1989) y Fernández Planas, (Fernández Planas, 1993). Se observa que el espacio delimitado por las vocales extremas del triángulo (/i,a,u/) proporcionarían una zona de referencia para situar las vocales intermedias.

A diferencia de los campos de dispersión acústicos, aparecen unas áreas más amplias, excepto en el caso de la vocal /a/, el elemento con menor variabilidad perceptiva. Las vocales cerradas (/i/, /u/) presentan un campo de dispersión bastante amplio, pero ubicado en los extremos del rectángulo, lejos de las realizaciones vecinas. La vocal /o/ reduce drásticamente su campo a las realizaciones más abiertas interponiendo la mayor de las distancias con /u/. En cambio, la vocal /e/ se permite cierto margen de variación en su rasgo, radicado en el primer formante.

En definitiva, la percepción vocálica se muestra como un mecanismo adaptativo que necesita hacer frente a las fuentes de variabilidad de la señal de habla (sonidos extranjeros, sonidos artificiales, diferentes locutores, etc.), que han de ser normalizadas; de ahí la amplitud de los campos de dispersión perceptivos. Además, el estudio de las características espectrales se complica mucho debido a la variabilidad que se produce en el habla cuando se pronuncian varios sonidos seguidos de forma que existan interacciones entre ellos, que es la situación normal de habla continua.

Una fuente importante de variación en la realización espectral de las vocales viene dada por el contexto de consonantes que puede existir. A la influencia de este fenómeno se le denomina coarticulación. En este caso, la variabilidad que se presenta se debe fundamentalmente a la diversidad de hablantes, hasta tal punto que se puede

[29] No existen valores absolutos para las formantes de las vocales; los datos de cada formante cubren un amplio abanico de valores que se denomina "campo de dispersión".

considerar como un parámetro que contribuya a la identificación del hablante. Por ejemplo, la coarticulación en las nasales (especialmente en la 'm') varía significativamente con las distintas personas, y como consecuencia este factor puede ser tomado en cuenta en el campo del reconocimiento del hablante. Por otra parte, la variación existente entre diversos individuos en el fenómeno de la coarticulación, hace descartar la hipótesis de universalidad en las características de articulación fonética.

En (Mermelstein, 1978), se realiza uno de los estudios primarios destinados a identificar las posiciones de los formantes en vocales con contextos consonánticos. En (Kewley-Port, 1995), se amplía éste con el objetivo de determinar los efectos de los contextos consonánticos en la discriminación de la frecuencia de los formantes.

Examinando los parámetros de estudio empleados, se destacan dos factores discriminantes: la longitud de la porción estable de la vocal y la separación de F1 y F2 en las transiciones de los formantes. Ambos factores se realzan en vocales con contextos consonánticos. Otra conclusión significativa se centra en la mayor influencia que ejercen las consonantes sobre la evolución del formante F2 respecto a F1.

Capítulo 4

Modelos e implementaciones del Sistema Auditivo

En este capítulo se enumeran los desarrollos de cócleas artificiales más relevantes, tanto implementaciones analógicas como digitales. Todos ellos, toman como referencia modelos matemáticos que representan la propagación del sonido a lo largo del oído interno y la conversión de la energía acústica en impulsos nerviosos, para su posterior procesado.

4.1. Modelos del sistema auditivo

El estudio del sistema de audición tiene sus orígenes en los tiempos de Aristóteles (384-322 A.C.), quien creía que su funcionamiento se debía a un aire interno

localizado dentro de una estructura ósea en forma de caracol, a la cual le asignó el nombre de cóclea.

Ya en 1950, Georg von Békésy, premio Nobel de medicina por sus estudios del oído interno, mostró que la membrana basilar dentro de la cóclea es la responsable de separar la señal de sonido en diferentes frecuencias (Békésy, 1960). El conocimiento adquirido hasta ese momento acerca del sistema de audición, permitió el desarrollo de los primeros modelos matemáticos realísticos inspirados en el funcionamiento del oído a partir de los cuales empezaron a surgir sus implementaciones. Empezando con el desarrollo de la primera implementación llevada a cabo en 1950, por Peterson y Bogert (Peterson & Bogert, 1950).

Los modelos del sistema auditivo que se describen a continuación (*ERB*, Lyon, Seneff, Kates y Lyon-Katsiamis) modelan fundamentalmente la cóclea, aunque alguno también modele otras partes del oído. Por tanto, tomaran como entrada la señal auditiva, o una representación de la misma, y darán como salida un cocleograma que es la probabilidad de activación de las neuronas del nervio auditivo a lo largo del tiempo.

4.1.1. Modelo *ERB*

El modelo *ERB* o modelo Patterson-Holdsworth (Slaney, 1993), está basado en los trabajos de Patterson y Holdsworth sobre la cóclea (Patterson et al., 1992). Consiste en una matriz de filtros paso banda en estructura paralela o independientes. Cada uno de ellos sintonizados a una frecuencia diferente. En el modelo de Patterson el ancho de banda de cada filtro coclear está descrito por un Ancho de Banda Rectangular Equivalente, *ERB* (*Equivalent Rectangular Bandwith*), descrito en el capítulo 3. Un filtro de bandas críticas o *ERB* modela la señal que está presente en una única célula del nervio auditivo.

La Figura 41 muestra la respuesta en frecuencia de este modelo según la implementación de Slaney (Slaney, 1993). Es interesante resaltar que el modelo se ha implementado como un banco de filtros en paralelo, todos con la misma señal de entrada. No existe, por tanto, ninguna dependencia entre filtros sucesivos como ocurre en un banco de filtros en cascada, aunque esta dependencia solo tendrá sentido si se considera un modelo no linear.

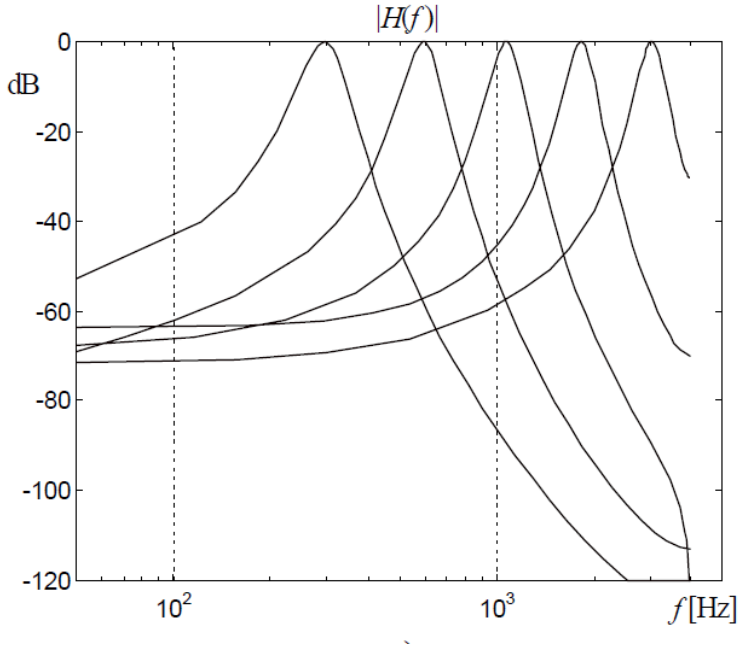


Figura 41. Respuesta en frecuencia de filtros gammatone (de orden 8, N=4) para 5 frecuencias características: 3.03, 1.83, 1.07, 0.6 y 0.3 kHz. Escala ERB.

Este modelo está basado en filtros *gammatone* con respuestas al impulso, como el que se muestra en la Figura 42. La importancia de estos filtros para la audición reside en que pueden generar una respuesta en frecuencia muy parecida a la de los filtros auditivos humanos obtenidos de forma perceptual por Patterson. Es más, son

capaces de indicar como se mueve la membrana basilar frente a un estímulo dado. Un posible inconveniente de este modelo basado en filtros *gammatone* es que las respuestas son muy simétricas, es decir, no existe diferencias entre las pendientes de atenuación ascendente y descendente respecto de cada frecuencia característica.

En este modelo no se especifica la separación frecuencial entre canales. Ni se realiza control de ganancia ni adaptación.

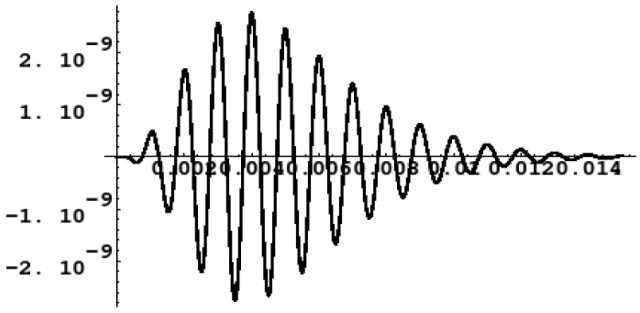


Figura 42. Respuesta de un filtro *gammatone*.

Existen otros modelos auditivos, llamados modelos funcionales, basados en un banco de filtros cuya respuesta imita la respuesta de la cóclea. En ellos, se observa un pico muy pronunciado que coincide con la frecuencia característica y una caída en la respuesta para frecuencias por encima y por debajo de la frecuencia característica, siendo más pronunciada la pendiente en las frecuencias superiores. Este efecto se consigue fundamentalmente con un banco de filtros en cascada. Un banco de filtros paralelos, necesitaría filtros de un mayor orden.

A continuación se describen algunos de estos modelos más significativos.

4.1.2. Modelo de Lyon

Richard F. Lyon desarrolló un modelo coclear basado en el conocimiento del funcionamiento de la cóclea (Lyon, 1982). Este modelo describe la propagación del sonido en el oído interno y la conversión de la energía acústica en representaciones neuronales. Cuando el sonido llega a la cóclea, una onda de presión viaja a lo largo de la membrana basilar. Las propiedades físicas de la membrana basilar cambian desde la base al ápice, de manera que las componentes en frecuencia de la onda alcanzan su máximo en una posición concreta de la membrana basilar.

El modelo coclear descrito por Lyon combina una serie de filtros que modelan la onda de presión viajera, rectificadores de media onda, *HWR* (*Half Wave Rectifiers*) que detectan la energía de la señal actuando como las *IHCs* y distintas etapas de control automático de ganancia, *AGC* (*Automatic Gain Control*), modelando el comportamiento de las *OHCs*. Un esquema de este modelo se muestra en la Figura 43.

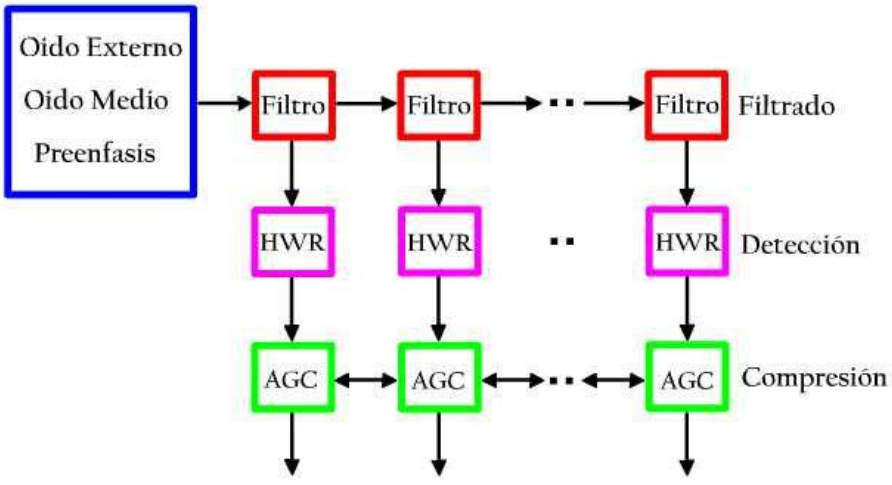


Figura 43. Esquema básico del modelo de Lyon.

En el modelo de Lyon existe una etapa de preénfasis inicial seguida por una cascada de etapas de filtros auditivos. El filtro de preénfasis es usado para modelar los efectos del oído externo y medio. Para ello, se usa un filtro paso alta con una frecuencia de corte de 300Hz. Está seguido por un diferenciador y un compensador de alta frecuencia comunes a todas las etapas.

En cada punto de la cóclea la onda acústica es filtrada por un filtro *notch*[30]. Cada filtro *notch* opera en una frecuencia satisfactoriamente baja, de manera que el efecto global es un flitrado paso baja gradual. Un resonador adicional (filtro paso banda) deja pasar una pequeña parte de la energía de la onda viajera y modela la conversión del movimiento de la membrana basilar que es detectado por las *IHCs*.

La señal de cada etapa de filtrado es una representación paso banda de la señal de audio original. Esta representación transcurre por un rectificador de media onda y una etapa de control de ganancia.

Su salida es un vector proporcional a la tasa de disparo de impulsos eléctricos en cada punto dentro de la cóclea, separando en frecuencia la energía de la onda acústica.

En la Figura 44 se presenta la respuesta de este modelo de 64 secciones para una frecuencia de muestreo de 8 kHz. Es interesante resaltar como existe una diferencia clara entre las pendientes ascendentes y descendentes respecto la frecuencia característica.

[30] Filtro *notch* o filtro de rechazo de banda es un filtro electrónico que no permite el paso de señales cuyas frecuencias se encuentran comprendidas entre las frecuencias de corte superior e inferior.

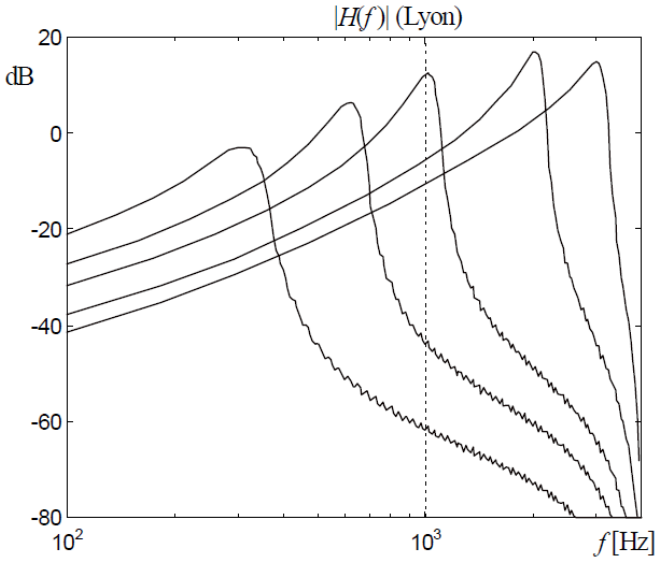


Figura 44. Respuesta en frecuencia del modelo de Lyon (64 secciones) para 5 frecuencias características: 3.0, 2.0, 1.0, 0.6 y 0.3 kHz.

4.1.3. Modelo de Seneff

El modelo de Seneff (Seneff, 1988) es similar al de Lyon y trata de capturar las características fundamentales extraídas de la cóclea.

Consta de tres bloques, como se muestra en la Figura 45. Los dos primeros están relacionados con las transformaciones que tienen lugar en el proceso de audición y el tercero extrae información asociada a la percepción humana.

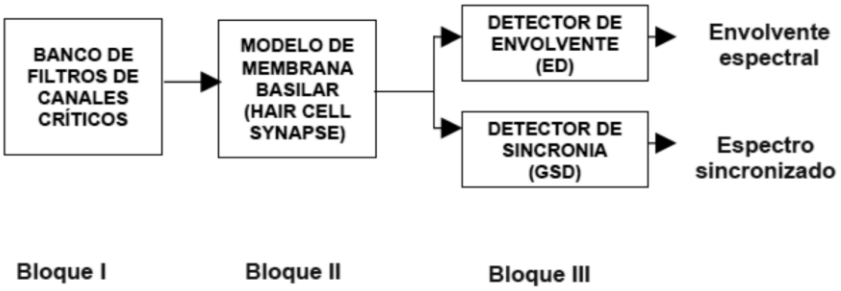


Figura 45. Diagrama de bloques del modelo de Seneff

En primer lugar, la señal de voz es pre-filtrada, para anular sus componentes de muy alta y/o baja frecuencia. Para ello, se usa un filtro *FIR*[31] paso alta de orden 8. A continuación, pasa a través del primer bloque, un banco de filtros lineales de bandas críticas de 40 canales con una frecuencia de muestreo de 16 kHz, cuya estructura se muestra en la Figura 46. Consta de un banco de filtros en cascada de orden 2 con filtros asociados a la cascada de orden 4. Utiliza la escala de frecuencias de *Bark* para la división de las bandas de frecuencia del banco de filtros, que va a cubrir el rango de frecuencias de 130 Hz a 6400 Hz.

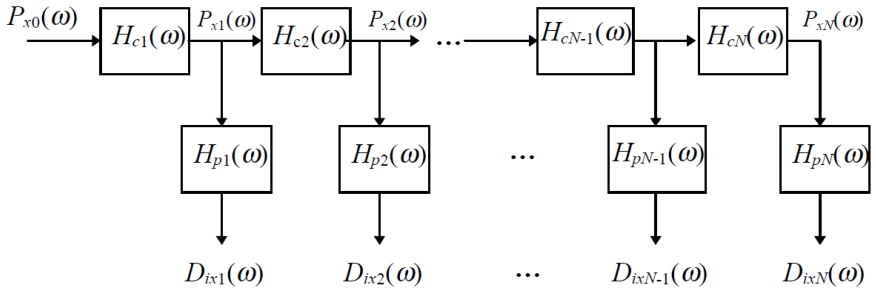


Figura 46. Modelo de cóclea basado en banco
de filtros en cascada y en paralelo.

[31] Filtro *FIR*, siglas en inglés *Finite Impulse response*. Filtro digital con respuesta finita al impulso.

En la Figura 47 se muestra la respuesta en frecuencia de esta primera etapa. Después pasa a través del segundo bloque no lineal, que captura las características predominantes de la vibración de la membrana basilar. La salida de esta etapa representa la probabilidad de disparo de los haces nerviosos en función del tiempo. El tercer y último bloque es un bloque de unidad doble con dos salidas paralelas. La primera unidad, se denomina detector de sincronía (*GSD*), la cual implementa la característica fase de bloqueo de las fibras nerviosas con el objeto de mejorar los picos espectrales creados por el tracto vocal. La segunda unidad se denomina detector de envolvente (*ED*) y se encarga de calcular la envolvente de las señales de la etapa anterior, dando más importancia a la captura dinámica de la palabra. Los resultados de esta unidad son más importantes en los sonidos transitorios.

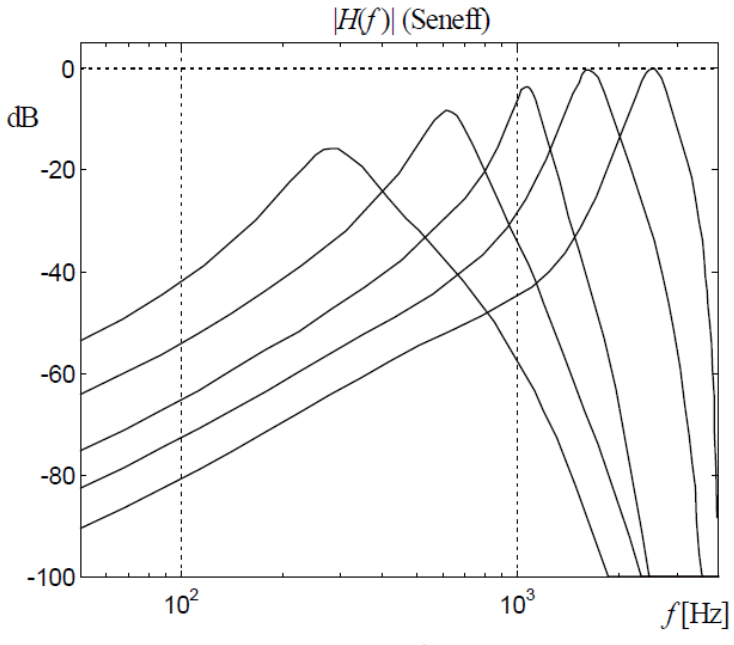


Figura 47. Modelo de Seneff (Bloque I). Las curvas corresponden a la salida de los canales 4,11,18,25 y 32.

4.1.4. Modelo de Kates

El modelo de Kates (Kates, 1991) y (Kates, 1993), no difiere mucho de los dos modelos anteriores en cuanto a su estructura, sigue el mismo esquema de la Figura 46. Los filtros en cascada son paso banda de orden 3. El modelo está formado por 112 secciones. Utiliza una frecuencia de muestreo de 40 kHz. En la Figura 48 se muestra la respuesta en frecuencia de 5 canales en concreto del modelo. Se observa que en este modelo la respuesta en frecuencia es más baja que en los dos anteriores modelos.

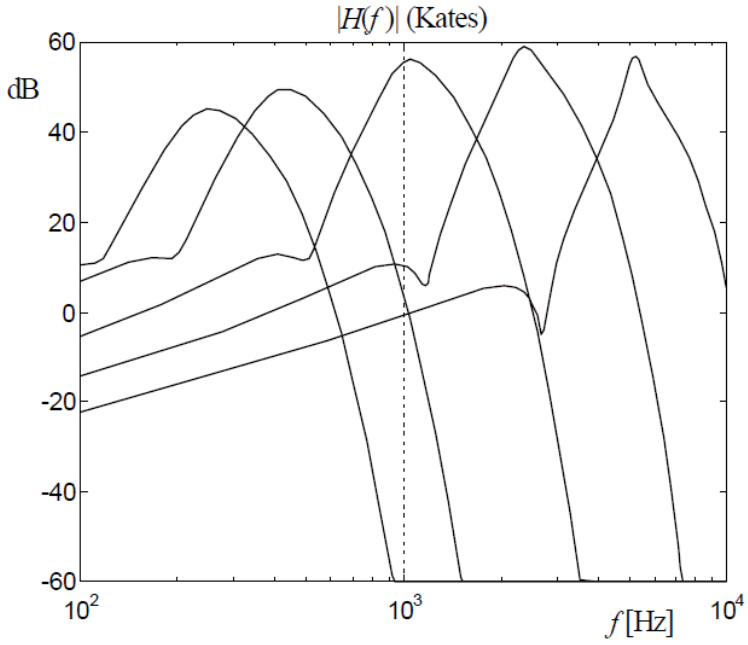


Figura 48. Respuesta en frecuencia del modelo de Kates (sin AGC). Las curvas corresponden a los canales 30, 50, 70, 90 y 100.

4.1.5. Modelo de Lyon y Katsiamis

Richard Lyon junto con Andreas Katsiamis y Emmanuel Drakakis publican un artículo de investigación en el 2007, donde presentan funciones de transferencia en el dominio continuo diseñadas a partir del filtro *gammatone* para el procesamiento auditivo (Katsiamis, Drakakis, & Lyon, 2007). En él se muestra el diseño de dos tipos de filtros, el Filtro *Gammatone* Todos-polos Diferencial, *DAPGF* y el Filtro *Gammatone* Un-Cero, *OZGF*. Estos dos diseños se caracterizan por presentar una arquitectura orientada a la implementación hardware, que cuenta con las mismas propiedades de operación de la cóclea y supera algunas limitaciones del uso de filtros *gammatone*, como son la simetría de su respuesta en frecuencia y su complejidad en la descripción en el dominio de la frecuencia.

A diferencia de la arquitectura en cascada del modelo de Lyon, este modelo se basa en un banco de filtros compuesto por etapas en paralelo las cuales se componen de bloques conectados en cascada. En la Figura 49 se muestra como la membrana basilar puede ser modelada tanto con una arquitectura paralela como con una arquitectura en cascada.

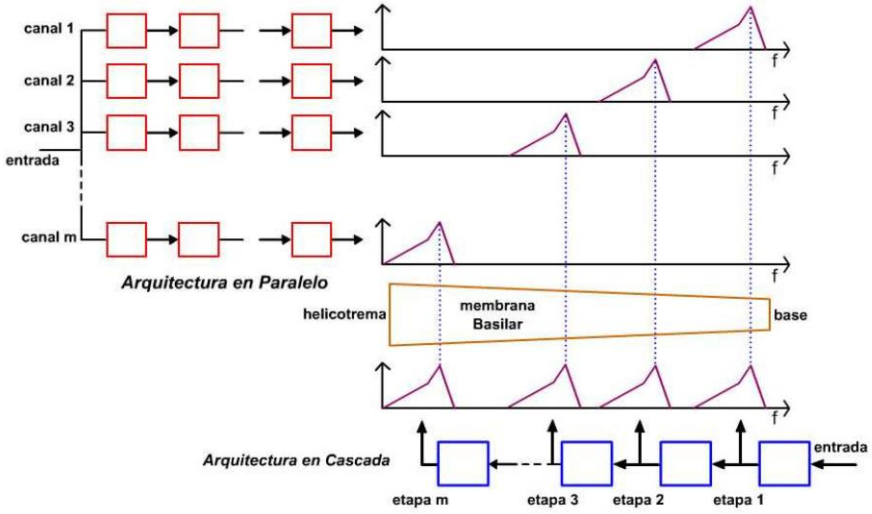


Figura 49. Bancos de filtros en arquitectura de cascada y en arquitectura paralela.

4.1.6. Modelo de las células ciliadas internas, *IHC*

Para crear un completo modelo neuromórfico de la cóclea, se ha añadido al modelo de cóclea básico los elementos que imitarán el comportamiento de las células ciliadas internas (*IHCs*), encargadas de transformar la vibración mecánica de la membrana basilar en señales eléctricas (proceso descrito en el capítulo 3).

Existen varios modelos de las *IHCs*, pero, sin duda, el más conocido y utilizado es el modelo de Meddis. En 1986, Meddis propone un modelo de simulación por computador de la *IHC*, (Meddis, 1986), (Meddis, 1988) y (Meddis, Hewitt, & Shackleton, 1990). Este modelo es ampliamente aceptado y se va a convertir en la base de posteriores modelos implementados en *VLSI*.

4.2. Implementaciones de cócleas artificiales

Idealmente, las características de la cóclea artificial están diseñadas para permitir obtener unos resultados comparables con los de una cóclea biológica. Sin embargo, la complejidad y sobre todo el coste de la implementación de las mismas, han dado lugar a que las cócleas artificiales se diseñen solo con algunas de estas características, dependiendo de la aplicación.

Trabajar con la cóclea artificial en tiempo real ha permitido a los investigadores aislar los componentes individuales del modelo y entender mejor cómo ellos trabajan. De ahí la importancia para los investigadores de usar modelos biológicos realistas: desde el punto de vista de procesado de la señal, la cóclea biológica ha aportado las propiedades de inmunidad al ruido, estabilidad y rango dinámico.

Desde el primer diseño de cóclea artificial (*silicon cochlea*), propuesto por Richard Lyon y Carver Mead en el año 1988 (Lyon & Mead, 1988), han aumentado el número de investigaciones en las cuales se lleva a cabo la implementación de diferentes modelos matemáticos utilizando tecnología *VLSI* analógica, y en menor número, implementaciones digitales utilizando dispositivos lógicos reconfigurables.

Pero a pesar de los 20 años de investigación, las cócleas artificiales están todavía muy lejos de compararse con la cóclea biológica, sobre todo en aspectos de consumo, rango de frecuencias, rango dinámico de entrada o inmunidad al ruido.

En las siguientes secciones se resumen las implementaciones más relevantes, tanto analógicas como digitales, de cócleas artificiales.

4.2.1. Cócleas analógicas implementadas en silicio

Entre los trabajos que usan tecnología *VLSI* analógica, destacan las implementaciones de Lyon y Mead (Lyon & Mead, 1988), Watts (Watts, 1993) y van

Shaik (v. Schaik, Fragnière, & Vittoz, 1995). También, sobresalen los trabajos realizados por Lazzaro en 1991 (Lazzaro, 1991); el desarrollado por Lyon en 1996 (Lyon, 1996) basado en el uso de filtros *gammatone*; el trabajo de Xing en el 2000 (Xing, 2000) y Graham en el 2006 (Graham, 2006), utilizando circuitería analógica programable. En la siguiente Figura 50 se muestra una evolución de las cócleas artificiales neuromórficas implementadas en VLSI analógico.

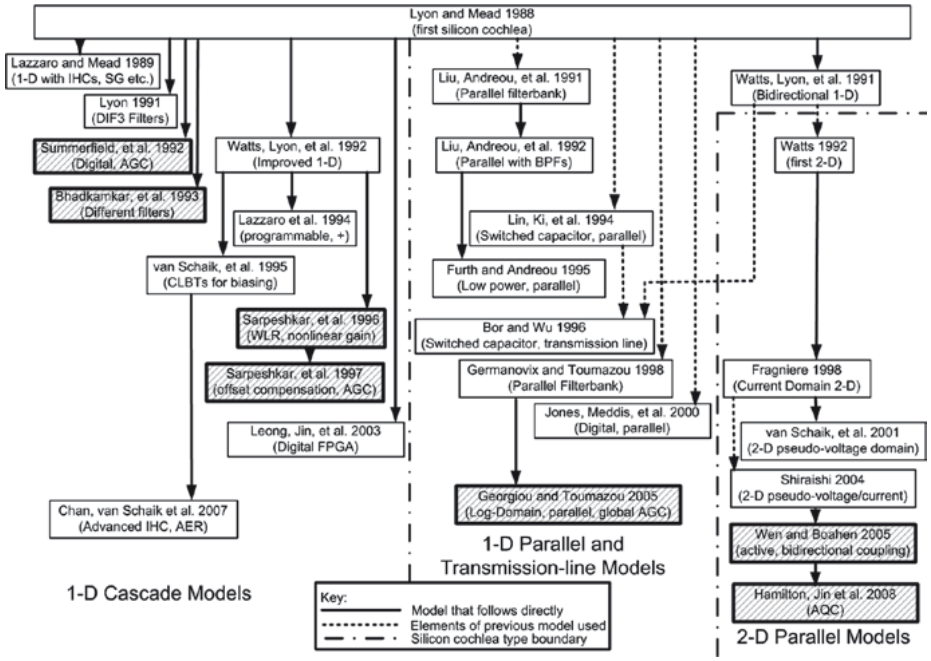


Figura 50. Árbol de la evolución histórica de las cócleas implementadas en silicio.

Se ha descrito que la cóclea artificial modela la membrana basilar usando un conjunto de filtros o resonadores, cuyas frecuencias de corte decrecen exponencialmente (desde la base al ápice), imitando así la distribución de frecuencias a lo largo de la membrana basilar. Las características de flexibilidad y anchura

dependientes de la posición en la membrana basilar se implementan con cambios sistemáticos en los parámetros de los filtros de la cóclea.

En general, la cóclea artificial implementada en silicio, se puede clasificar segun:

1. El factor de acoplamiento existente entre los elementos de los filtros. Se distingue entre cócleas unidimensional (1-D, *one-dimensional silicon cochlea*) y bidimensionales (2-D, *two-dimensional silicon cochlea*).

2. La existencia o no de control de ganancia y frecuencia en los filtros para permitir que se adapten dinámicamente a los cambios de intensidad de la entrada.

Se distinguen por tanto los siguientes tipos de cócleas: unidimensional (1-D) en cascada (Figura 51.b), 1-D paralela (Figura 51.c) o bidimensional (Figura 51.d).

La cóclea artificial unidimensional (1-D) en cascada (Figura 51b) modela la propagación de la onda sonora a través de la membrana basilar en una única dirección (eje x, Figura 51.a), desde la base al ápice de la misma; además, en esta estructura en cascada, cada segmento de la cóclea (*filter element*) realiza su filtrado sobre la salida del anterior elemento. Esto va a permitir que, aunque en general sean filtros de segundo orden, se consigue un efecto de 'pendiente muy pronunciada' (*steep slope*), lo cual favorece la selectividad frecuencial propia de la membrana basilar (Figura 52.b).

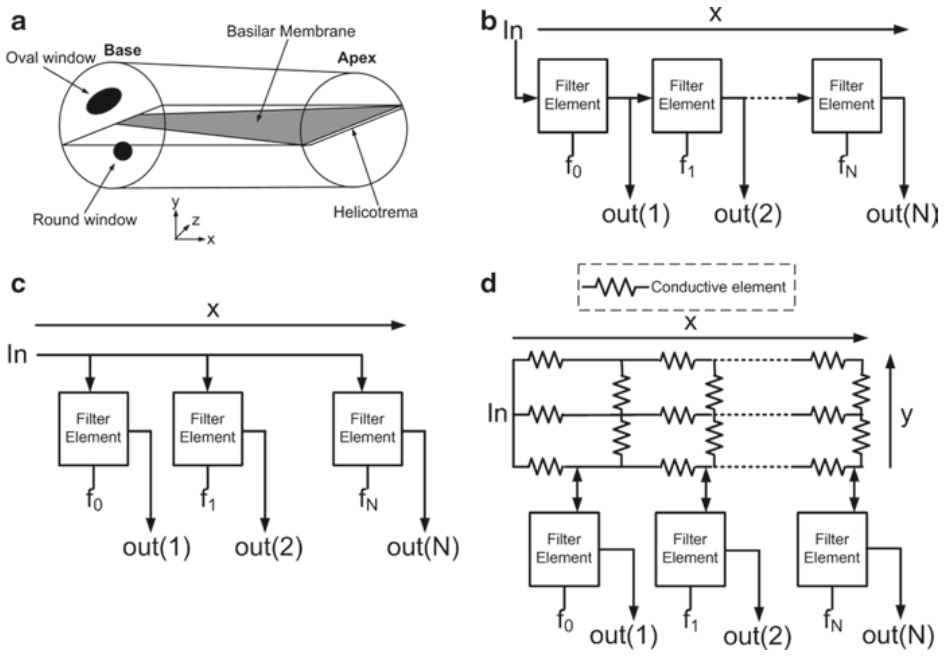


Figura 51. (a) Cóclea desenrollada. (b) Estructura de la cóclea unidimensional en cascada. (c) Estructura 1-D paralela. (d) Estructura bidimensional.

La primera cóclea artificial se basó en esta estructura unidimensional en cascada, considerando únicamente la propiedad de la membrana basilar que varía en función de la posición (Lyon & Mead, 1988). Se ha comprobado, en la cóclea humana, como la frecuencia característica a lo largo de la membrana basilar disminuye exponencialmente en una escala logarítmica: frecuencias altas en la base y frecuencias bajas en el ápice. Para implementar esta cóclea artificial, se ha dividido la membrana basilar en segmentos de igual longitud. Se utiliza un conjunto de filtros organizados en cascada con una frecuencia característica, de acuerdo a la frecuencia característica de cada segmento de la membrana basilar. Todos los filtros son iguales; solo varían en su frecuencia característica y en su respuesta en frecuencia paso-baja o paso banda. Se eliminan las componentes a alta frecuencia a la salida de cada filtro, lo que provoca una fuerte caída en las curvas de respuesta en frecuencia. Esta fuerte caída

también se observa en la cóclea biológica a medida que la onda de presión viaja a lo largo de la membrana basilar. Por tanto, este modelo, a pesar de su simplicidad, proporciona una primera aproximación al procesado de la señal realizado dentro de una cóclea biológica.

En el trabajo presentado por Lyon y Mead, cada etapa de filtrado se implementa con dos filtros separados. Una etapa de filtros conectados en cascada que emulan la propagación de las ondas a través de la cóclea y un conjunto de filtros en paralelo que emulan el movimiento de la membrana basilar, tomando el modelo original el nombre de cascada paralelo. Luego se descubrió que los polos y ceros de los dos filtros para cada etapa podían ser reordenados y cada etapa ser implementada como un filtro de segundo orden conocido como el modelo solo cascada.

El principal inconveniente de este modelo es su poca tolerancia a fallos, inherente a este tipo de estructuras en cascada. Si un elemento falla, este error se propagará al resto de elementos. También, hay que destacar que cada segmento va a añadir un cierto retraso a la señal de entrada, que será inversamente proporcional a la frecuencia central de cada filtro. Esto va a producir un retraso no-real, sobre todo en los segmentos correspondientes a las bajas frecuencias, que son los que se encuentran al final de la estructura (al final de la cóclea). Además, esto implica que el número de secciones (etapas) estará limitado debido a su uso fundamentalmente en sistemas en tiempo real. Por otro lado, el ruido generado internamente por los filtros también se va acumulando a lo largo de la estructura, lo que provocará una reducción en el rango dinámico del sistema. Algunos de estos inconvenientes se resuelven con los otros tipos de organización, arquitectura unidimensional paralela o bien una arquitectura bidimensional.

A menudo se elige la estructura paralela, Figura 51.c, por su fácil implementación. Pero, aunque este tipo de modelos no presenta los inconvenientes propios del modelo unidimensional en cascada, no es el preferido en los desarrollos

de cócleas analógicas porque cada filtro actúa de modo independiente y para crear el mismo efecto de 'pendiente pronunciada' en las altas frecuencias, se necesitaría filtros de un orden mayor, lo cual implica un aumento considerable en el área y consumo del sistema. En la siguiente figura se compara la salida de un único filtro de segundo orden tanto en el modelo 1-D paralelo, Figura 52.a, como en el modelo 1-D en cascada, Figura 52.b.

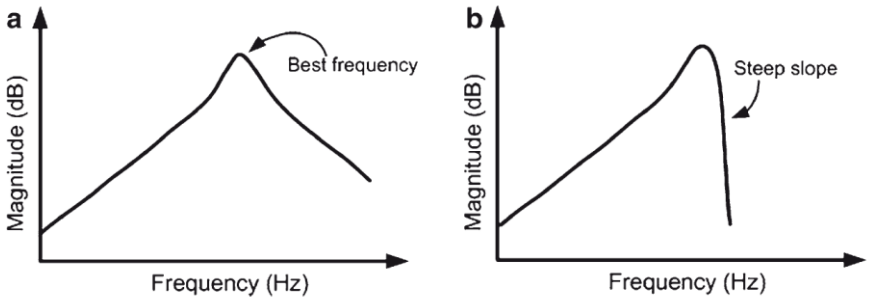


Figura 52. Salida de un filtro de segundo orden, en el modelo de cóclea paralelo (a) y en el modelo de cóclea en cascada (b).

La cóclea bidimensional (2-D) modela tanto la propagación de la onda a lo largo de la membrana basilar como el movimiento del líquido interior de la cóclea y de la membrana basilar. Tiene en cuenta, por tanto, el desplazamiento longitudinal y vertical (ejes x e y). Los filtros vecinos se acoplan a través de un sistema de resistencias que modelan el líquido del interior de la cóclea. Este modelo combina las ventajas de las dos estructuras 1-D anteriores: el acople de los filtros en paralelo permite generar una pendiente pronunciada en las altas frecuencias a pesar de seguir siendo filtros de segundo orden; además se mejora la tolerancia a fallos y se evita la acumulación de retrasos propio de la estructura en cascada. Las implementaciones de este modelo son las más recientes: en 1992, Watts presenta una implementación con 50 etapas que mejora el rango dinámico, estabilidad y errores derivados de los

transistores (Watts, Kerns, Lyon, & Mead, 1992). Con el mismo objetivo, van Schaik (v. Schaik et al., 1995) utilizará en sus diseños transistores bipolares en lugar de los tradicionales MOSFET, consiguiendo una uniformidad en la respuesta en frecuencia y unos buenos valores *Q*[32].

La cóclea artificial es activa cuando tiene control de ganancia y/o los filtros cambian dinámicamente según la intensidad de la entrada. En general, se incrementa la ganancia a baja frecuencias y se disminuye a altas frecuencias. Este comportamiento activo básicamente pretende modelar el funcionamiento de las células ciliadas externas (*OHCs*). En la evolución histórica mostrada en la Figura 50 aparecen sombreadas las cócleas activas. La mayoría de estos modelos usan un control automático de ganancia (*AGC*, *Automatic Gain Control*). Convencionalmente, el *AGC* implica cambiar la ganancia dependiendo de los cambios de la señal de entrada. En este caso, sin embrago, el *AGC* incluye la idea de un control automático del factor de calidad, *Q*. De esta forma, no solo se cambia la ganancia como respuesta a los cambios de la entrada, sino que también se cambia el ancho de banda del filtro.

Desde el primer modelo de *IHC*, propuesto por Lazzaro y Mead (Lazzaro & Mead, 1989a), han surgido distintos circuitos analógicos para implementarlos. Uno de ellos, desarrollado por Mc.Ewan y van Schaik (McEwan & v. Schaik, 2003), será utilizado junto a una cóclea artificial, tal como se muestra en la siguiente Figura 53. La salida de la cóclea se conecta a estos elementos *IHCs*, cada salida de los *IHCs* se usa para estimular una neurona pulsante, con la finalidad de usar la representación

[32] El factor Q, también denominado factor de calidad o factor de selectividad. Se define como la razón entre la frecuencia central y el ancho de banda. En filtros, sirve para determinar lo selectivos que son, es decir, para ver el ancho de banda. En principio, un filtro con menor ancho de banda (mayor Q), será mejor que otro con más ancho de banda.

AER y facilitar de este modo el uso de la cóclea artificial en sistemas neuromórficos, (v. Schaik & Liu, 2005).

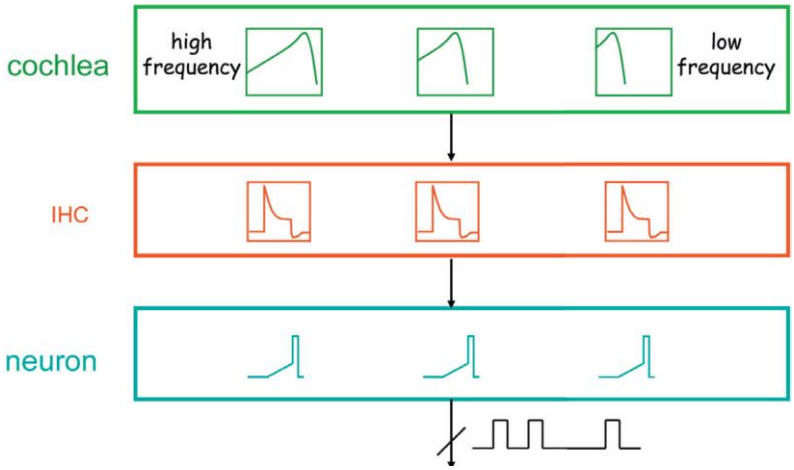


Figura 53. Diagrama de bloques de una cóclea artificial. La salida de cada canal de la cóclea se conecta a un circuito *IHC*, el cual alimenta a un conjunto de neuronas generadoras de pulsos.

4.2.2. Cócleas digitales implementadas en *FPGA*

La necesidad de obtener sistemas eficientes que tengan un bajo consumo de potencia, una menor área y menor coste, han llevado a la investigación y desarrollo de nuevas técnicas de diseño y arquitecturas. Se ha descrito como los sistemas de procesamiento bioinspirados son una alternativa con la que se puede lograr unos buenos resultados en cuanto al consumo de potencia, la velocidad de procesamiento y el área utilizada en comparación con los sistemas y las técnicas tradicionales. La tecnología basada en *FPGA* se presenta como una buena opción para el desarrollo de sistemas neuromórficos. Entre sus ventajas y en comparación con los sistemas analógicos *VLSI*, se destaca por un tiempo más breve de diseño; un tiempo de fabricación más rápido; ser más robusto respecto a cambios en la fuente de alimentación, de temperaturas y de errores en los transistores; un rango dinámico

más amplio; un mayor *SNR*[33]; mejor estabilidad; las placas se pueden reutilizar para diferentes aplicaciones y tiene una interfaz más simple con un PC.

Diseño de filtros en hardware

La respuesta del banco de filtros de una cóclea implementada en hardware, va a depender fundamentalmente del diseño de los filtros y de la configuración en cascada o paralela de las etapas de filtrado.

Tradicionalmente, el diseño de un filtro requiere definir si es paso baja, paso banda o paso alta, lo cual implica especificar sus frecuencias características y sus niveles de ganancia. Sin embargo existe otra alternativa, que es el diseño del filtro a partir de su función de transferencia. Este es el caso del diseño utilizado por Mishra y Hubbard, (Mishra & Hubbard, 2002), que utilizan un algoritmo de diseño *LMSE* (*Least mean square error*) para construir un filtro a partir de una función de transferencia, en este caso la función *invfreqz* de Matlab.

También se elige entre un filtro *FIR* (*Finite Impulse Response*) o un filtro *IIR*[34] (*Infinite Impulse Response*). Para la implementación de la cóclea se usan filtros *IIR*, porque con ellos se consiguen bandas más altas y estrechas con un menor número de coeficientes lo que implica, además una reducción en el número de operaciones aritméticas.

Para llevar a cabo el diseño digital de un filtro *IIR* se parte de un diseño analógico (*Butterworth, Chebyshev* o Elíptico) y se pasa al dominio digital (transformada z bi-linear).

---

[33] *SNR,* siglas en inglés *Signal to Noise Ratio.* La relación señal/ruido se define como el margen que hay entre la potencia de la señal que se transmite y la potencia de ruido que la corrompe. Este margen es medido en decibelios (dB).

[34] Filtro *IIR*, siglas en inglés *Infinite Impulse Response*. Filtro digital con respuesta infinita a un impulso.

La implementación de un filtro *IIR* se puede realizar de diferentes maneras. Para los diseños hardware en aritmética de punto fijo en complemento a dos, destaca la estructura de forma directa I, forma directa II transpuesta y la basada en Aritmética Distribuida (AD). Para la estructura forma directa I, destacamos dos configuraciones de la etapa de filtrado: la primera usa etapas de filtrado independientes (paralela), de tal modo que el procesamiento en cada etapa se realiza de forma concurrente; la segunda está compuesta de una sola etapa, mediante la cual se realiza el cálculo de todas las etapas en serie (cascada). Usando Aritmética Distribuida, encontramos tres tipos de arquitecturas: AD secuencial, AD paralela y AD híbrida entre las dos anteriores.

ESTRUCTURA FORMA DIRECTA

La forma convencional para el diseño en hardware de filtros *IIR* se basa en el uso de estructuras como la forma directa I, *FDI*, utilizando bloques *MAC*s[35] o bloques *DSP*[36]. La estructura *FDI* se construye a partir de la ecuación de diferencias correspondiente al filtro *IIR* de segundo orden (Ecuación 4.1).

$$y(n) = b_0 \cdot x(n) + b_1 \cdot x(n-1) + b_2 \cdot x(n-2) + a_1 \cdot y(n-1) + a_2 \cdot y(n-2) \quad (4.1)$$

El diseño hardware, para este número de coeficientes, requiere el uso de 4 sumadores, 5 multiplicadores y 4 registros para los retardos (Figura 54).

[35] *MAC,* siglas en inglés *multiplier-accumulator unit*. Bloque multiplicador-acumulador.

[36] *DSP,* siglas en inglés *digital signal processing*. Dispositivo para el procesamiento digital de señales.

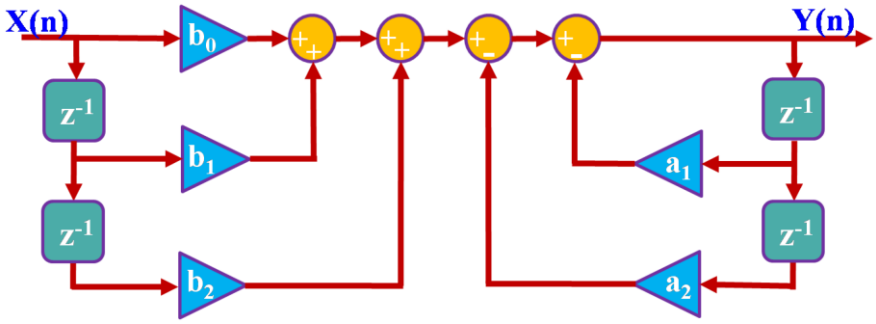


Figura 54. Arquitectura FDI.

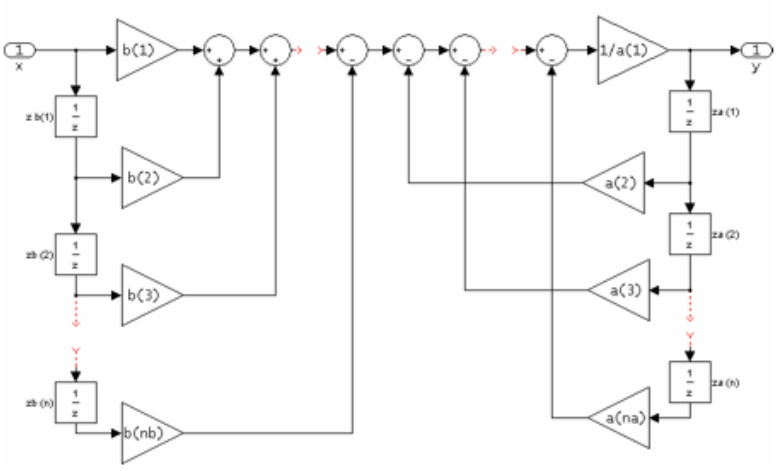


Figura 55. Forma directa I filtro IIR.

Para una arquitectura *FDI* Serial (Figura 56), se usará una sola estructura de *FDI* para el cálculo de varias etapas de segundo orden con diferentes coeficientes. En este caso, es necesario usar memorias *ROM* para almacenar el contenido de los coeficientes de todas las etapas y memorias *RAM* para el almacenamiento temporal de las entradas y salidas correspondientes a los instantes anteriores de todas las etapas.

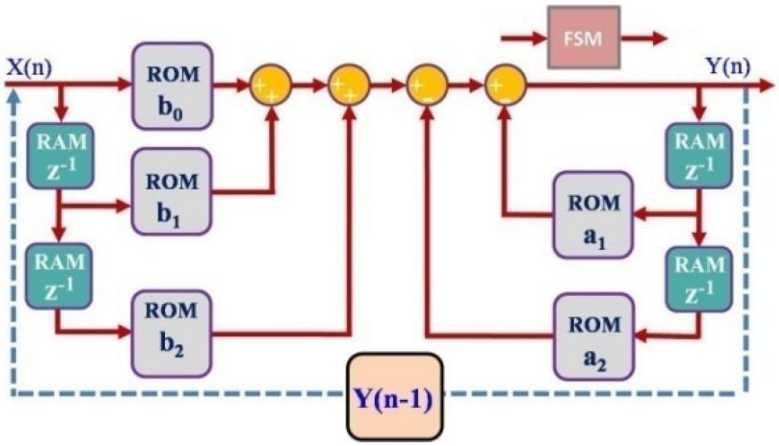


Figura 56. Arquitectura FDI Serial.

Si se usa una organización en cascada del banco de filtros, se debe añadir un bloque de realimentación de la salida a la entrada.

La arquitectura *FDI* es una de las más rápidas, debido al uso de los elementos *DSP*. Sin embargo, la gran cantidad de bloques *DSP* necesarios en la implementación del banco de filtros limita el uso de esta estructura para una implementación concurrente y es necesario, por tanto, recurrir a técnicas como la aritmética distribuida o diseñar sistemas secuenciales.

Se ha descrito que la Forma Directa I, Figura 55, requiere el uso de (N+M) elementos de memoria, (N+M) sumadores y (N+M+1) multiplicadores (siendo N el número de coeficientes del denominador y M el número de coeficientes del numerador). La forma directa II, Figura 57, elimina M elementos de memoria, ya que en esta estructura estos elementos están repetidos.

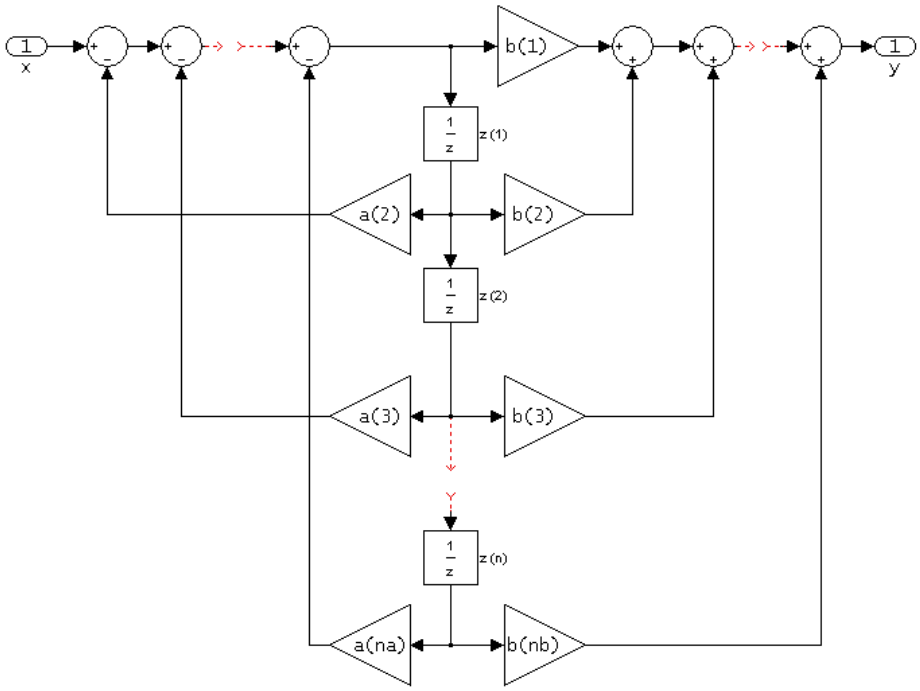


Figura 57. Forma Directa II filtro IIR

Forma Directa II transpuesta, Figura 58, sustituye los nodos por sumas, las sumas por nodos, invierte el sentido de las flechas e intercambia los coeficientes. Esta forma da lugar a una realización con N elementos de memoria, (N+M+1) multiplicadores y N sumadores. Es la forma más utilizada para la implementación.

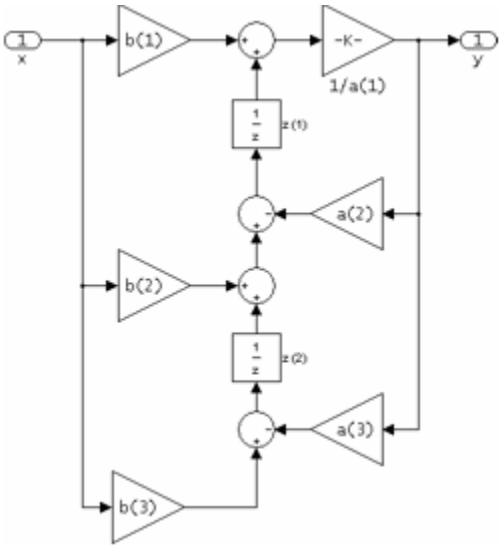


Figura 58. Forma Directa II transpuesta.

ESTRUCTURA BASADA EN ARITMÉTICA DISTRIBUIDA

La aritmética distribuida (AD) es una algorítmica computacional que permite implementar de forma eficiente las sumas de productos, constituyendo una tecnología muy importante en las *FPGAs*. Se basa en almacenar los productos parciales de las multiplicaciones en tablas, evitando así el uso de multiplicadores que reducen la velocidad e incrementan el área. El hecho de conocer a priori el valor de los coeficientes constituye una condición fundamental para poder implementar la AD.

La primera arquitectura propuesta usando AD es una arquitectura serie, Figura 59, donde cada término es calculado en un ciclo de reloj. El resultado se obtiene después de B ciclos de reloj, donde B es el número de bits de la entrada del filtro.

El diseño en hardware de esta estructura requiere el uso de 2 registros de entrada-paralela/salida-serie, 3 registros de desplazamiento serie de B bits, un bloque

acumulador-escalador y una memoria *ROM* para el almacenamiento de los datos de la *LUT*[37].

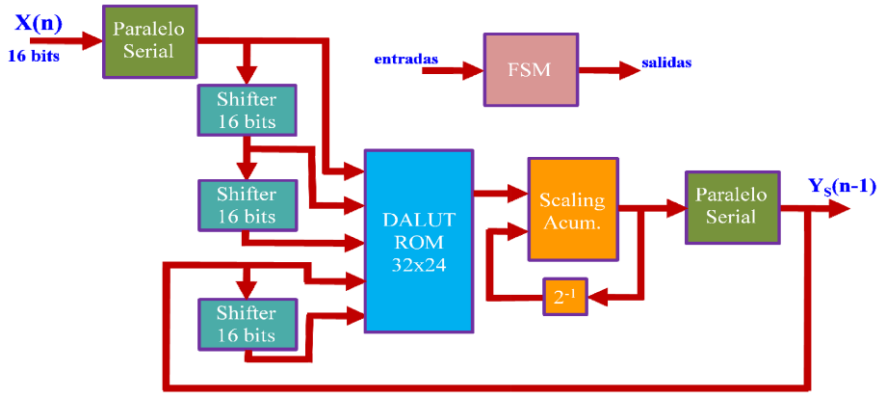


Figura 59. Arquitectura AD secuencial

La arquitectura AD paralela realiza el cálculo de la señal de salida en cada ciclo de reloj mediante procesamiento paralelo. Se necesita B *LUTs* por cada etapa, de manera que se puede calcular el contenido de cada término de la ecuación a la vez. En la Figura 60 se muestra la arquitectura paralela para una señal de entrada de 8 bits.

[37] *LUT,* siglas en inglés *Look-Up Table.* Es una memoria RAM usada para almacenar valores predefinidos. Su uso más general es para implementar funciones combinacionales.

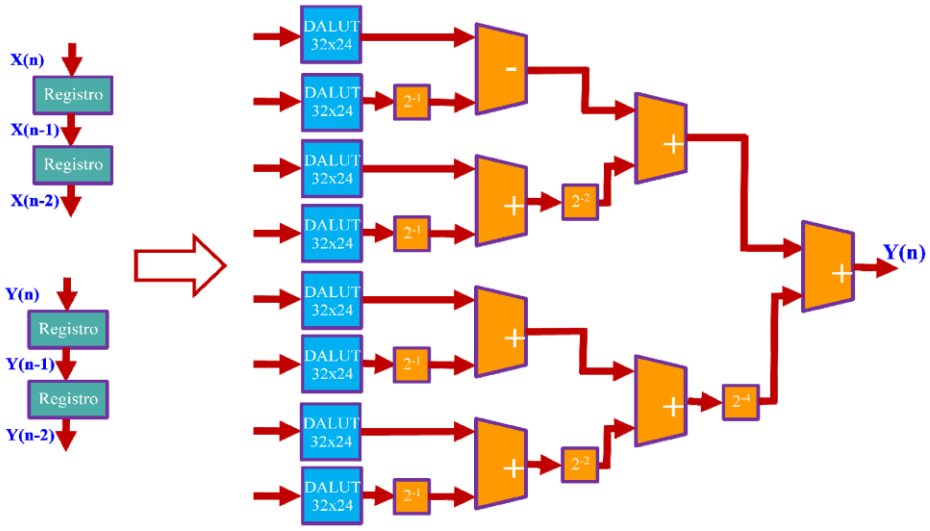


Figura 60. Arquitectura AD paralela.

La arquitectura AD híbrida, Figura 61, usa B/2 ciclos de reloj para el cálculo de la señal de salida.

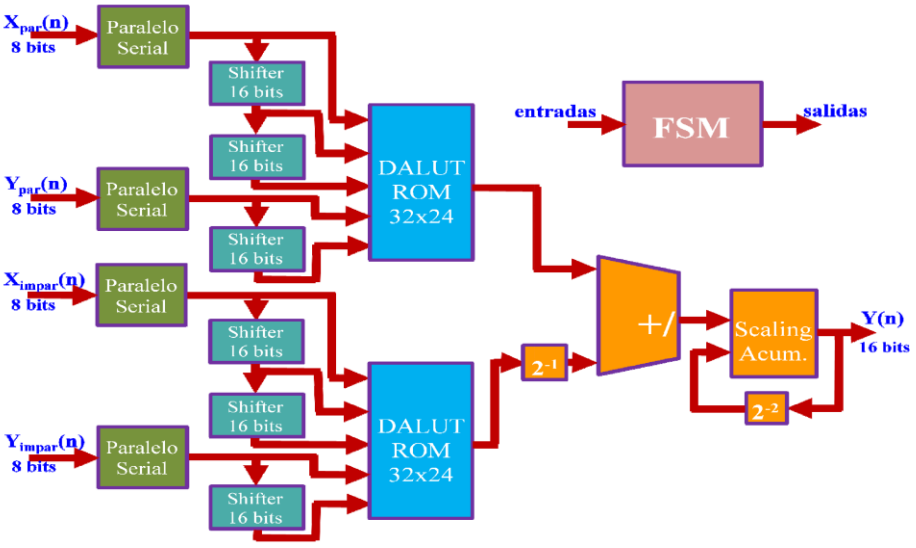


Figura 61. Arquitectura AD Híbrida

El diseño hardware requiere 4 registros de entrada-paralela/salida-serie, 6 registros de desplazamiento seriales, un bloque acumulador-escalador, 2 memoria ROM para el almacenamiento en las dos *LUTs*, un sumador y el diseño de una máquina de estados finita (*FSM*).

La arquitectura AD paralela usa una mayor cantidad de *LUTs* que la arquitectura serie y en general más recursos que la arquitectura híbrida. En recursos de memoria, la cantidad usada por la paralela es mayor que la usada por la serie y la híbrida. Esta diferencia es lo que permite que la arquitectura paralela presente una mayor velocidad de procesamiento.

Sin embargo, la arquitectura híbrida es la que consigue un mayor equilibrio entre la velocidad de procesamiento y recursos de memoria, siendo la más adecuada para la implementación sobre una *FPGA* de menor tamaño y coste.

Cócleas basadas en filtros digitales

En este apartado, es interesante resaltar de nuevo la relevancia que ha tenido el modelo de cóclea propuesto por Lyon y Mead. En la bibliografía, se puede encontrar diferentes investigadores que han usado este modelo para sus implementaciones hardware, ya sea como base de su trabajo o para comparación y verificación de sus resultados.

A continuación se resumen distintas implementaciones digitales.

En 1997, Lim y su equipo describen un sistema de detección de tono utilizando un banco de filtros paso banda *Butterworth* de primer orden, (Lim, Temple, & Jones, 1997).

Se destaca el trabajo realizado por Watts en el 2000 donde implementa un modelo de cóclea de 240 etapas (http://www.lloydwatts.com/neuroscience.shtml).

Mishra y Hubbard en el 2002 (Mishra & Hubbard, 2002), presentan una implementación de un filtro digital de orden 10 basada en la respuesta en frecuencia del modelo TWAmp[38], (A. Hubbard, 1993). El comportamiento de este modelo analógico TWAmp se aproxima al de la cóclea biológica. Ofrece una respuesta en frecuencia con bandas altas y, aunque no son estrechas, su caída es muy pronunciada, característica muy deseable en los filtros de una cóclea. Para la implementación del filtro eligen la Forma Directa I y combinan en un mismo bloque un conjunto de filtros IIR de 2º orden y un único filtro FIR de orden 10; todos en una estructura paralela. Comparan la respuesta en frecuencia de filtros IIR, modelo TWAmp, datos biológicos (Ruggero, Rich, & Robles, 1990) y el modelo de Lyon-Mead. Esta comparación se muestra en la Figura 62.

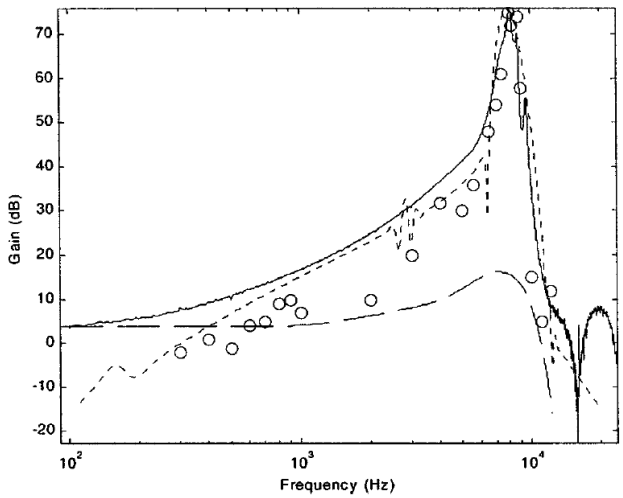


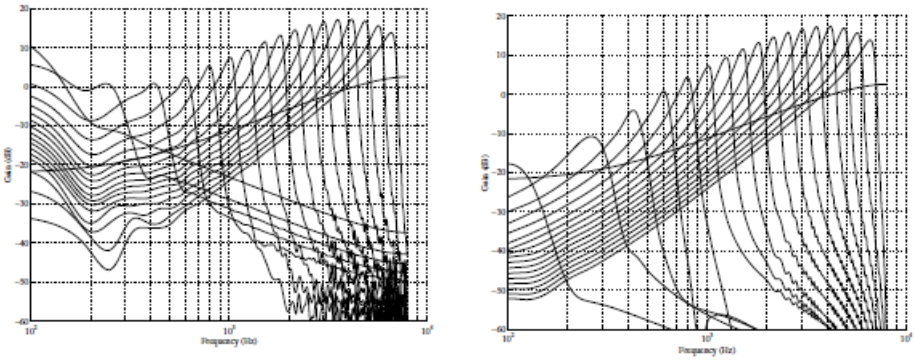
Figura 62. Respuesta de un filtro IIR (línea sólida), datos biológicos (círculos), modelo TWAmp (trazos cortos) y modelo de cóclea Lyon-Mead (trazos largos).

Y llegan a la conclusión de que los resultados, comparado con los datos biológicos, difieren en las frecuencias bajas; también destacan el comportamiento del

[38] *TWAmp model,* siglas en inglés *travelling-wave amplifier model.*

filtro IIR respecto de los otros diseños sobre todo porque tiene una respuesta muy similar con un procesado en tiempo real y con coste bajo.

Leong, a raíz de su trabajo de investigación realizado en el 2003 (Leong & Jin, 2003), implementa sobre *FPGA* una cóclea de 88-etapas basada en el modelo de Lyon-Mead utilizando aritmética distribuida. En este trabajo de investigación realiza un estudio del tamaño óptimo de los valores de entrada (desde 12 bits a 32 bits) y del ancho de la ROM utilizada (desde 12 bits a 24 bits). De los resultados obtenidos, se comprueba como la exactitud de la respuesta de los filtros mejora de un modo gradual con el incremento de la longitud de la palabra y del tamaño de la ROM (Figura 63). Cuando los valores son pequeños aparecen los efectos de cuantización que provocarán una respuesta incorrecta del filtro.



(a) (12 bits, 12 bits)

(b) (32 bits, 24 bits)

Figura 63. Respuesta en frecuencia de diferentes implementaciones de la cóclea (Número de bits de la entrada, Anchura en bits de la ROM).

Wong y Leong, (C. K. Wong & Leong, 2006), presentan una mejora respecto al anterior diseño. Usan decimación[39] para reducir el tiempo de procesado y aritmética doble punto-fijo (*DFX*)[40] para mejorar el rango dinámico. Hay que destacar que el anterior diseño usa una arquitectura paralela de filtros implementados con aritmética distribuida; y en este caso se usan filtros *IIR* de segundo orden. Por tanto, una arquitectura en paralelo sería posible si la *FPGA* dispone de los recursos necesarios (sobre todo multiplicadores) para realizar todas las operaciones en paralelo, en otro caso se debería usar una estructura en cascada. En esta implementación, se usa de nuevo la estrategia de pipeline[41].

Aunque el procesado paralelo proporciona una salida a una mayor velocidad, existe una limitación en el tamaño de la cóclea puesto que los recursos necesarios crecen linealmente con el número de etapas de la cóclea. Sin embargo, la cóclea en cascada implementada en este trabajo puede contener cualquier número de etapas, aunque en este caso la velocidad disminuye linealmente con el número de etapas.

Clarke y Qiang, (Clarke, Qiang, Peremans, & Hernandez, 2004), presentan una eficiente implementación hardware de una cóclea neuromórfica, basada en un banco de filtros paso banda *IIR Butterworth* de 2º orden. Utiliza una arquitectura paralela encadenada (pipeline) para reducir los recursos utilizados en la implementación de cada filtro (3 multiplicadores y 3 sumadores).

---

[39] La decimación es una operación básica que se realiza fundamentalmente en los sistemas multifrecuencia. Consiste en la reducción de la frecuencia de muestreo; para ello se realiza una primera etapa de filtrado y una segunda etapa de muestreo (muestreo hacia abajo o *downsampling*).

[40] *DFX,* siglas en inglés *dual fixed-point*. Es una nueva representación de datos que combina la sencillez del procesamiento de la representación punto-fijo con el amplio rango dinámico que ofrece la representación punto-flotante. Se basa en el uso de un bit para seleccionar entre dos tipos de representaciones punto-fijo, lo que permite escalar de forma dinámica las señales usadas en el sistema (Ewe, Cheung, & Constantinides, 2004).

[41] Pipeline, término inglés ampliamente aceptado para referirse a arquitecturas encadenadas.

En el trabajo de J. Meza-escobar, M. Vera-Lizcano y J. Velasco-Medina, (Velasco-Medina & Hernan-Meza Escobar, 2007), se realiza un análisis comparativo del diseño e implementación en hardware de diferentes arquitecturas para el desarrollo de dos modelos cocleares: el modelo de Lyon-Mead y el modelo propuesto por Lyon-Katsiamis. En este estudio encontramos conclusiones muy interesantes a tener en cuenta en el diseño e implementación de filtros digitales. Una de ellas es que el elevado número de bloques DSP necesarios en la implementación del banco de filtros con estructuras como la FDI, limita el uso de estas estructuras para una implementación totalmente concurrente y es necesario recurrir a técnicas como la aritmética distribuida o diseñar sistemas secuenciales. Resume que la estructura paralela, propia del modelo de Lyon-Katsiamis, presenta mejores resultados debido a los problemas de cuantización que presentan las estructuras en cascada, propia del modelo de Lyon. Además, los modelos basados en estructuras paralelas presentan pocos parámetros para su implementación y suelen estar formados por bloques repetitivos.

En 2008, (Dundur, Latte, Kulkarni, & Venkatesha, 2008), presentan el diseño de una cóclea digital formada por 16 filtros de segundo orden IIR Butterworth paso banda con una estructura de forma directa II transpuesta, asociada a la aritmética de punto fijo en complemento a dos. Este diseño es la base para la implementación de un implante coclear sobre una FPGA.

En (Gambin, Grech, Casha, Gatt, & Micallef, 2010) se presenta la implementación de una cóclea sobre una Spartan-3E FPGA. Debido a la limitación de recursos de la FPGA usada (sólo dispone de 20 multiplicadores), y teniendo en cuenta que se utiliza un filtro IIR paso baja de orden 2 (5 multiplicadores), sólo era posible usar al mismo tiempo 4 filtros. Por esto, plantean usar el esquema de multiplexación en el tiempo para implementar una cadena de filtros de 24 etapas.

También en (Mugliette, Grech, Casha, Gatt, & Micallef, 2011) se presenta una implementación sobre FPGA del modelo activo de cóclea de Lyon. Está basada en un banco de 24 filtros IIR de segundo orden paso bajo en cascada. Utiliza un esquema multiplexado en el tiempo, con el objetivo de sólo utilizar 20 multiplicadores. Un esquema de la arquitectura se muestra en la Figura 64.

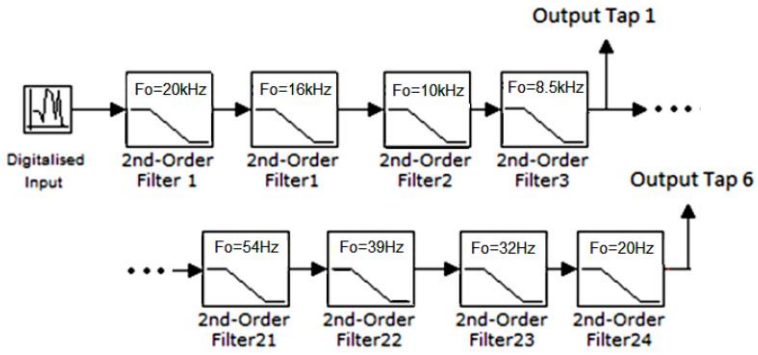


Figura 64. Filtros en cascada con salida cada 4 secciones.

Esta implementación añade además el mecanismo de control de ganancia automático (*AGC*) que imita el comportamiento de las *OHCs*, así como la funcionalidad de las *IHCs*. Se basa en el modelo de Dau (Dau, 1996), el cual tiene una menor complejidad hardware. Puesto que las células *IHCs* son sensibles a la velocidad de la señal de entrada, la salida de los filtros atraviesan una fase de derivación, rectificación y detector de frecuencias de manera que la señal de salida es comparada a un valor anterior de la señal y si es mayor el valor almacenado es actualizado. En el modelo, los valores almacenados decaen exponencialmente, aunque en esta implementación es lineal. Con este proceso se cambia la función paso baja de los filtros implementados a paso banda y se extrae la magnitud de la envolvente de la señal resultante.

Se ha descrito que en el sistema auditivo las *OHCs* amplifican la señal de entrada e incrementan el rango dinámico. El control automático de ganancia implementado ajusta la ganancia para que la señal de entrada sea oída a un nivel constante. Se utiliza un amplificador de ganancia programable (*PGA*[42]) para aumentar el rango dinámico del *ADC* así como el del banco de filtros, en el caso de que la magnitud de la señal de entrada sea baja.

[42] *PGA,* siglas en inglés *Programmable gain amplifier*.

Capítulo 5

Sistemas de reconocimiento automático del habla

En las próximas páginas se describen los conceptos básicos relacionados con el proceso de reconocimiento automático del habla; se enumeran las técnicas clásicas más utilizadas en el proceso de parametrización de la señal de voz así como las diferentes estrategias usadas para la generación de los modelos acústicos. Además se relata la evolución de los sistemas de reconocimiento automático del habla.

5.1. Definición de los sistemas de reconocimiento automático del habla

El objetivo del reconocimiento automático del habla (RAH) es imitar el proceso de reconocimiento que lleva a cabo el receptor en la comunicación oral, interpretando

un mensaje oral como una secuencia de palabras. La dificultad del RAH reside, tanto en la gran variabilidad que presenta la señal vocal (a nivel fonético y acústico entre los distintos locutores), como en las muchas fuentes de conocimiento necesarias para el proceso de reconocimiento (fonéticas, fonológicas, sintácticas, semánticas, pragmáticas). Hay varios niveles de reconocimiento en dicho proceso humano, y los diferentes sistemas de reconocimiento automático implementan todos, algunos o solo los más básicos dependiendo de cuáles sean su aplicación y complejidad. Se pueden distinguir ocho niveles de reconocimiento en orden de complejidad ascendente:

- Nivel acústico: la señal acústica analógica que ha enviado el emisor es recibida y traducida a un conjunto de rasgos relevantes no redundantes. En la comunicación oral, este reconocimiento se hace en el oído. Hay cuatro operaciones incluidas en este nivel, que son total o parcialmente implementadas para automatizarlo en el RAH:
    - *Parametrización:* la señal analógica se transforma en una señal numérica que puede ser tratada por la máquina digital en la que se hace el reconocimiento. Hay varios métodos de parametrización en el dominio del tiempo y en el dominio de la frecuencia, que dan lugar a diferentes parámetros de caracterización.
    - *Segmentación:* determina como separar la señal analógica continua en una cadena de sonidos cuya sucesión es la señal en el tiempo. Se lleva a cabo con métodos basados en las curvas de variación de la energía o de variabilidad de la señal.
    - *Extracción de la información relevante:* se busca retener solo aquellos datos que proporcionen información útil para el

reconocimiento como pueden ser los espectros de los instantes de mayor estabilidad o de los instantes de transición.

- *Información relativa a la prosodia:* estudia la variación del armónico fundamental de la voz, variación de la intensidad, y el ritmo.

- Nivel fonético: la secuencia de información relevante obtenida en el nivel acústico es traducida a una secuencia de fonemas.

- Nivel fonológico: son analizados los fonemas de la lengua que hacen que el contenido fonético de las palabras se modifique en una articulación rápida o por una sucesión de términos léxicos. Las variedades dialécticas son también tratadas.

- Nivel léxico: se identifican las palabras de la lengua en la que se produce la comunicación.

- Nivel sintáctico: se detectan las reglas gramaticales que permiten describir y analizar el lenguaje, y que relacionan las palabras reconocidas a nivel léxico.

- Nivel semántico: analiza el sentido de las palabras, buscando la comprensión del mensaje y eliminando las interpretaciones que no tengan sentido. Es el nivel de conocimiento de las palabras que da un diccionario de la lengua.

- Nivel pragmático: estudia el sentido del mensaje teniendo en cuenta el contexto de su aplicación. Reconoce la información que viene determinada por la situación en la que se produce la comunicación.

- Nivel prosódico: interviene de manera paralela al resto de niveles, sin formar parte de una estructura piramidal como los demás; este nivel detecta la información que el mensaje comunica mediante los modos de

pronunciación: palabras pronunciadas con cierto nivel de insistencia para ponerlas en relieve, fronteras entre grupos de palabras, naturaleza interrogativa o declarativa de una frase, etc.

En particular, la parte del reconocimiento del habla se automatiza implementando, en su totalidad o parcialmente, estos ocho niveles de reconocimiento según la complejidad y necesidades del sistema de reconocimiento. Las posibilidades son muchas.

En la siguiente Tabla 11 (Cole, 1997) se recopila una visión global de las variables que definen un sistema de reconocimiento automático de habla y el rango de valores que pueden tomar. El rango de los valores expresa los límites de complejidad que tiene el sistema para un parámetro dado. Así, el valor de la izquierda indica el caso más simple y el de la derecha la situación más compleja.

Tabla 11. Parámetros típicos empleados en la caracterización de un RAH. (Cole, 1997)

Parámetros	**Rango**
Tipo de discurso	[Palabras aisladas, habla continua]
Dependencia del locutor	[Dependiente del locutor, independiente del locutor]
Tamaño del vocabulario	[Pequeño (< 20 palabras), Grande (> 20000 palabras)]
Estilo de discurso	[Lectura, habla espontanea]
Modelo de lengua	[Contexto explícito, sensible al contexto]
Confusión	[Pequeña (<10), grande (> 100)]
Relación señal ruido	[Alta (> 30 dB), baja (< 10 dB)]
Tipo de transductor	[Micrófono de gradiente, teléfono]

Hay reconocedores de palabras aisladas (el hablante debe realizar pausas entre palabras), de palabras conectadas y de habla continua, lo que supone un orden creciente de complejidad del reconocedor que tiene que delimitar palabras y frases. El habla puede ser no espontánea (leída o dirigida mediante un diálogo de opciones), o puede ser espontánea (contiene cambios bruscos y acusados en la prosodia) con el

consiguiente incremento de la dificultad. El reconocedor de voz puede ser además dependiente del locutor teniendo que discernir la información acústica para un solo hablante, puede ser adaptado al locutor, multilocutor o independiente de locutor con lo que deberá filtrar las distorsiones acústicas debidas a las peculiaridades del hablante. Los fines específicos o generales del reconocedor, y la perplejidad y tamaño del vocabulario que reconoce, son características que aumentan la dificultad o simpleza de la tarea de reconocimiento. La distorsión acústica debida al ruido de canal y al ruido auditivo que acompañe a la voz incrementa la dificultad de la tarea de reconocimiento.

Hay que tener en cuenta, por tanto, que el reconocimiento automático de voz es una tarea inherentemente difícil debido a la variabilidad de las señales de voz. Si la señal de voz se registra en condiciones favorables, se consiguen muy buenas prestaciones en el proceso de reconocimiento. Sin embargo, cuando el sistema funciona en situaciones reales se encuentra con condiciones adversas motivadas fundamentalmente por cambios en el hablante (condiciones fisiológicas, emocionales, cambio en el modo de articulación debido a un fuerte ruido ambiental, etc.) y en el entorno acústico (ruidos, reverberación y ecos) o eléctrico (ruido o distorsiones de la señal provocados por el micrófono o el canal de transmisión). Todas ellas, son irrelevantes desde el punto de vista lingüístico pero degradan en cierta medida la tasa de reconocimiento.

En la mayoría de los sistemas de reconocimiento automático del habla, el proceso se divide en dos bloques claramente diferenciados. En el primero, se realiza una extracción de parámetros característicos de la señal de voz. Se obtiene así, una representación de la señal de voz sensible al contenido fonético y no a las variaciones acústicas de la señal. Y en el segundo se obtiene una medida de similitud entre la muestra obtenida en el bloque anterior y unas muestras de referencia. Esta medida de similitud puede ser tanto un valor determinista como una medida de probabilidad.

(En el apartado 5.3 se describen las técnicas clásicas más usadas en estas dos etapas del proceso de reconocimiento de voz).

Hay que destacar que en los sistemas de reconocimiento de habla complejos, no es suficiente con la información que cada una de las palabras o unidad lingüística aporta de forma aislada. Será necesario integrar información sobre el contexto, la sintaxis y la semántica de las palabras.

La siguiente Figura 65 muestra un ejemplo de estructura de un sistema de reconocimiento de voz. Tras digitalizar la señal sonora, se realiza un análisis acústico para obtener un conjunto de medidas o características útiles. Éstas deben contener toda la información relevante para realizar el proceso de reconocimiento, eliminando el resto de contenidos redundantes o aquellas que no sean útiles para el sistema. A continuación se realiza el proceso de reconocimiento teniendo en cuenta las restricciones impuestas por los modelos acústico, léxico y de lengua (gramática) disponibles. Un modelo acústico recoge las realizaciones dependientes del género del locutor, entonaciones, variantes dialectales, etc. El modelo léxico, contempla las pronunciaciones alternativas de las palabras, con objeto de permitir que los algoritmos de búsqueda encuentren diferentes caminos. Y el modelo del lenguaje, también llamado gramática, permite estimar la frecuencia de ocurrencia de determinadas secuencias de palabras.

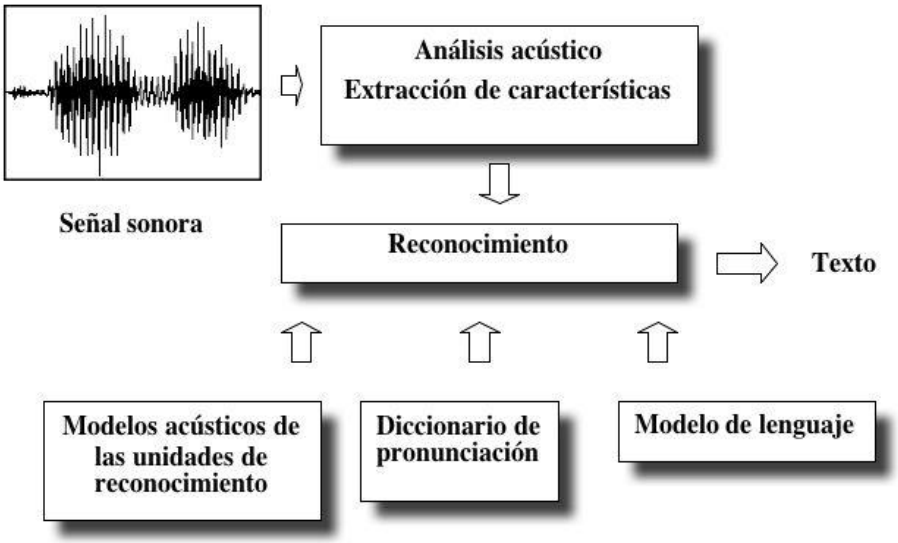


Figura 65. Esquema típico de un sistema automático de reconocimiento de voz

En este trabajo se ha implementado un sistema de reconocimiento capaz de identificar palabras a partir de la identificación de una secuencia de fonemas vocálicos. Se ha llegado, por tanto, al nivel fonológico de reconocimiento. Es un reconocedor de palabras aisladas, las cuales pertenecen a un vocabulario limitado, es independiente del hablante y el hablante ha sido dirigido a través de un programa para la pronunciación y grabación de dichas palabras (proceso descrito en el capítulo de experimentos).

5.2. Antecedentes y desarrollo de los sistemas actuales

El análisis o reconocimiento de una señal de voz es una tarea que presenta una cierta dificultad. Como se ha descrito en el capítulo 3, las principales características y dificultades que presenta una señal de voz a ser analizada son su continuidad (no

existen pausas entre sílabas, palabras, etc.), la redundancia natural al expresarnos normalmente y la gran variabilidad entre distintos locutores. Las razones para esta variabilidad son: anatómicas (longitud del tracto vocal, forma de la cavidad nasal, sexo, edad, etc.), dialectos, hábitos de pronunciación, personalidad del locutor, estilo. Pero esta variabilidad no sólo se observa entre pronunciaciones de distintos locutores, sino que para un locutor resulta imposible pronunciar una misma palabra dos veces iguales. Y es que un mismo locutor puede hablar en un tono bajo o alto, susurrando o gritando, relajado o tenso.

El desarrollo de muchos aspectos de los sistemas de reconocimiento automático del habla es total o parcialmente dependiente de la lengua y de la aplicación. En el estado actual de la tecnología en este campo, no existe una solución general al problema.

Al contrario que en el desarrollo de síntesis de voz en la que primero hubo un gran desarrollo de sistemas mecánicos y eléctricos, en el análisis de voz los principales avances se han logrado con el desarrollo de los computadores en los años setenta. No obstante se construyeron algunos sistemas eléctricos a principios del siglo XX (Fern, 1997).

5.2.1. Primeros dispositivos

J.B. Flowers en 1916 fue quizá el primero en diseñar una máquina para transcribir voz, basándose en sus conocimientos sobre la transmisión de mensajes por cable entre submarinos. Propuso una máquina capaz de convertir los sonidos de las letras en ondas de la misma naturaleza que las que se transmitían entre los submarinos, a las que llamó *alfabeto fonográfico.* Aunque la moderna teoría de formantes aún no se había desarrollado, se dio cuenta de la existencia de tonos a diferentes frecuencias en la señal acústica (100, 200 y 1000 Hz para el caso de un hombre pronunciando la palabra inglesa *go*).

Pero la historia del reconocimiento del habla en sí es más reciente. Comienza en los años cuarenta con el desarrollo de un dispositivo capaz de visualizar la señal acústica sobre el papel, conocido como *espectrógrafo.* Este dispositivo permite obtener un registro de la energía contenida en las diversas bandas de frecuencia de una palabra o frase en función del tiempo, llamado *espectrograma.* A partir de este momento se vislumbra la posibilidad de realizar sistemas para el reconocimiento automático del habla. Es en 1952 cuando K.H. Davis, R. Biddulph, S. Balashek, de los *Laboratorios Bell*, construyen el primer dispositivo de reconocimiento, capaz de discriminar con cierta precisión los diez dígitos ingleses pronunciados de forma aislada por un único locutor, Figura 66 (Davis, Biddulph, & Balashek, 1952). El dispositivo era totalmente electrónico y para la comparación hacia uso de técnicas de intercorrelación entre los parámetros (posiciones de los dos primeros formantes) del dígito pronunciado y los patrones correspondientes a cada una de las diez posibles pronunciaciones esperadas.

La señal de voz era separada por dos filtros dentro de dos bandas de frecuencia. Un filtro paso-alta permite el paso de las componentes frecuenciales por encima de los 900 Hz y un filtro paso baja permitía el paso de las inferiores. La secuencia principal de cada banda se localizaba a partir de la cuenta del número de *cruces por cero.* Estas dos señales eran aplicadas a los canales X e Y de un osciloscopio. Se dibujaba una curva de dos dimensiones en la pantalla del osciloscopio. Estas curvas eran bastantes diferentes para cada uno de los diez dígitos ingleses. Para compararlas automáticamente se cuantificó, se dividió el dibujo en 28 cuadros para representar las diferentes combinaciones de los formantes uno y dos. Los cuadros estaban definidos por un circuito de válvulas que activaba uno de los 28 relés en función de la forma del dibujo. En función de los relés activados se encendía uno de los indicadores correspondiente a un número reconocido.

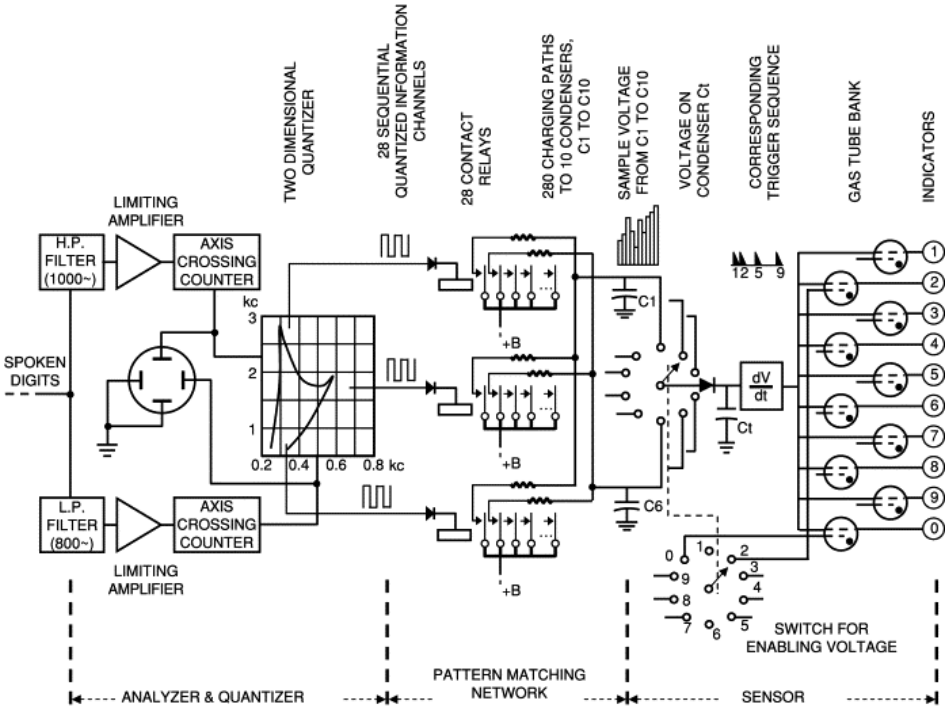


Figura 66. Reconocedor de dígitos desarrollado por Davis, Biddulph y Balashek en 1952 Imagen tomada de (Juang & Rabiner, 2006).

Las medidas de los *patrones* se hicieron con 100 repeticiones de cada uno de los 10 dígitos y se obtuvieron tasas de reconocimiento comprendidos entre un 97% y 99%.

Entre otros trabajos de la época, también basados en dispositivos analógicos, nos encontramos la *máquina de escribir fonética* desarrollada en *los laboratorios RCA* en 1956 por H.F. Olson y H. Bellar. Y el trabajo de J. Wiren y H.L. Stubbs, en 1956, que obtenían información sobre el contenido espectral de la señal usando como criterio de clasificación la frecuencia de resonancia de las vocales. Usaban un *árbol binario* de selección, de modo que en cada hoja se hacía una separación. Por ejemplo,

si en una hoja se separaban fonemas sordos de sonoros, en la siguiente inferior, en la que todas las unidades eran sordas, se separaban en fricativas o no, etc. Los circuitos necesarios para hacer estas separaciones incluían una gran cantidad de válvulas, porque el transistor era un elemento nuevo para esta época. Más tarde, en 1959 fue creado por P. Denes, en la *University College* de Londres, un sistema capaz de reconocer cuatro vocales y nueve consonantes. El aspecto más novedoso de este trabajo fue el uso de información estadística, sobre las secuencias válidas de fonemas en inglés, incorporando así conocimiento lingüístico en este tipo de sistemas.

Se observa como en un primer momento, el reconocimiento automático de voz se fundamenta en los principios de la fonética acústica y se limita al reconocimiento de palabras aisladas de un vocabulario muy reducido, dependiente del hablante y utilizando para ello dispositivos electrónicos.

5.2.2. Era informática

En 1960, P. Deves y M.V. Mathews construyen un reconocedor de dígitos basándose en un computador IBM 704. Introducen el concepto de normalización temporal no lineal o a también llamado alineamiento temporal dinámico (*Dynamic Time Warping*), la cual permite comparar los parámetros de palabras iguales pronunciadas a distinta velocidad. Esta normalización, que no era posible con las técnicas analógicas, implicó un aumento en las tasa de reconocimiento.

En los años 70, ya se habían conseguido grandes avances en los niveles básicos del reconocimiento automático del habla. El principal problema estaba en cómo aplicar los conocimientos de niveles lingüísticos superiores (fonéticos, sintácticos, semánticos y pragmáticos) para ayudar a la comprensión del mensaje hablado.

En este sentido, G. Fant sugiere un modelo de reconocimiento que divide el proceso en una secuencia de 5 pasos: extracción de parámetros, detección de segmentos, transcripción fonética, identificación de fonemas e interpretación

semántica. Por el contrario, R. Reddy propone el sistema *HEARSY*[43] basado en un modelo de reconocimiento en paralelo que incluye tres reconocedores independientes: acústico, sintáctico y semántico.

Empiezan a aparecer reconocedores independientes del hablante para tareas muy concretas. Hay que destacar en 1971, el papel que desempeña el Departamento de Defensa de los EE.UU que lanza el mayor proyecto conocido de la historia del reconocimiento del habla, el ARPA-SUR (*Advanced Research Projects Agency – Speech Understanding System*), con un gran presupuesto y una duración de cinco años. Los ambiciosos objetivos de éste y otros proyectos no llegaron a alcanzarse, pero los estudios realizados han servido para un mejor conocimiento de los mecanismos del habla y de las limitaciones de los sistemas automáticos de reconocimiento.

Otro hito importante es el comienzo de los trabajos del grupo investigador de IBM dedicado al dictado automático por voz para grandes vocabularios. Y al final de la década, en los *AT&T Bell Labs* los investigadores comenzaron una serie de experimentos orientados a conseguir reconocedores independientes del locutor para su uso en aplicaciones telefónicas.

A principio de los años 80, se demuestra la ineficacia de los sistemas basado en conocimiento. Se desarrollan sistemas capaces de extraer conocimiento de forma inductiva, a partir de muestras. Se utiliza un modelo acústico basado en Modelos Ocultos de *Markov* (*HMM, Hidden Markov Models*), discretos y continuos, y se optimizan los algoritmos de aprendizaje para entrenar los sistemas a partir de grandes bases de datos. Se mejoran los sistemas de alineamiento temporal dinámico para el reconocimiento de palabras conectadas, y se desarrollan algoritmos de búsqueda eficientes para determinar la secuencia óptima de patrones para una secuencia de vectores acústicos.

---

[43] HEARSY, de las palabras inglesas HEAR (oír) y SAY(decir).

A partir de este momento, el reconocimiento de habla continua ha mejorado, aumentándose el tamaño de los vocabularios, diversificándose las aplicaciones y enfrentándose a situaciones cada vez más reales, en las que hablante y condiciones de entorno difieren de las usadas para entrenar el reconocedor.

A mediados de los 80, se presenta la aproximación conexionista como alternativa a la aproximación estadístico-probabilístico realizada con los modelos ocultos de Markov. Las redes neuronales artificiales, *RNA* (*Artificial Neural Networks*), comparten con los *HMM* su carácter inductivo, es decir, el aprendizaje a partir de muestras. Pero sus configuraciones clásicas, como el perceptrón multicapa *MLP* (*Multi-Layer Perceptron*), no son capaces de representar fenómenos dinámicos como la señal de voz, por lo que se requiere el desarrollo de arquitecturas recursivas específicas. Otros autores optaron por configuraciones híbridas en las que el perceptrón multicapa se usa para estimar las probabilidades de emisión de un modelo oculto de *Markov*.

Finalmente, en los años 90's, se continúa trabajando con vocabularios cada vez más amplios, los costes disminuyen y las aplicaciones independientes del locutor y flujo continuo empiezan a ser más comunes. En la Figura 67 se muestra esta evolución de los sistemas de reconocimiento automático del habla desde el año 1960.

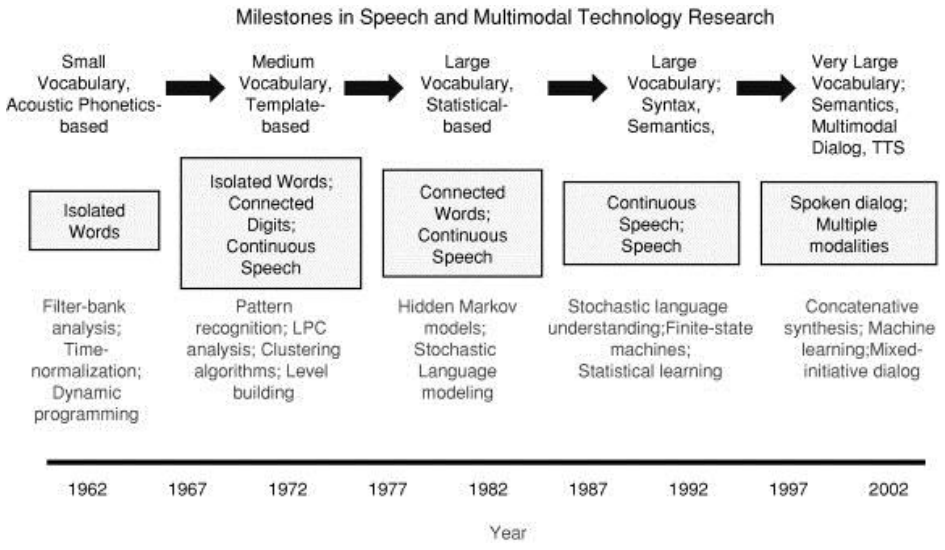


Figura 67. Hitos en las tecnologías del reconocimiento automático del habla.
Imagen tomada de (Juang & Rabiner, 2006).

En la última década, el progreso de las tecnologías de reconocimiento de voz se debe fundamentalmente al hecho de incorporar a las máquinas la capacidad para entender el lenguaje humano. Se ha visto como el rápido ascenso de los dispositivos móviles de gran potencia está haciendo que las interfaces de voz sean cada vez más útiles y omnipresentes. Estos teléfonos 'inteligentes' tienen un gran ancho de banda para las conexiones de datos con la 'nube', donde los servidores hacen el trabajo que precisa el reconocimiento de voz y la comprensión de las consultas orales. En 2010, aparece la aplicación "Búsqueda por voz" (*Voice Search)* de Google para iPhone. Peter Norvig, director de investigación en Google, describe el modelo de reconocimiento de voz usado en su sistema de "Búsqueda por voz"[44]. En este modelo se distinguen

[44] Publicado el 23 de noviembre de 2011 *por LatAm Blog Publisher.* (http://tecnologiayproductosgoogle.blogspot.com.es/2011/11/una-mirada-dentro-de-la-tecnologia-de_23.html)

tres partes: el modelo acústico, el modelo léxico y un modelo de lenguaje. Primero, el modelo acústico mapea todas las maneras posibles en que las ondas sonoras pueden formar fonemas. Se seleccionan las formas más probables para formar fonemas, teniendo en cuenta la gran variedad de representaciones obtenidas debido a la diversidad de dispositivos de grabación, género y edad del hablante, acentos, dialectos, etc. A continuación, los fonemas se agrupan en el modelo léxico, formando así el diccionario de pronunciación de todas las palabras de un idioma. Y por último, las palabras se combinan como parte de un modelo de lenguaje, el cual indica qué palabras suelen venir antes o después de otras palabras. Gracias al uso del sistema de Búsqueda de Google, el modelo se entrena con más de 230 mil millones de palabras incluidas en consultas reales hechas al buscador. Adicionalmente, si el sistema no reconoce una frase y el usuario lo corrige, se lleva a cabo un proceso de aprendizaje automático que mejora el modelo de lenguaje. Este sistema de búsquedas por voz se puede probar en un teléfono Android o en la aplicación de Google *Search* para iOS o BlackBerry.

Otro ejemplo destacado de interfaz de voz móvil es *Siri*, un asistente personal activado por voz incorporado en el iPhone, creado por la empresa *Nuance Communications* (empresa líder en el mercado del reconocimiento de voz con su software *Dragon*).

Esta empresa *Nuance*, inspirada por el éxito de su software de reconocimiento de voz en los teléfonos móviles, inicia el desarrollo de interfaces de voz en el campo de la televisión y los automóviles. Ya existen televisores de la marca *Samsung* que incorporan este software *Dragon TV*; y el sistema de entretenimiento *Sync* de los automóviles *Ford* están basados en la tecnología de *Nuance, Dragon Drive*.

5.3. Fundamentos del Reconocimiento Automático del Habla

Se ha descrito como los primeros sistemas buscaban la identificación de palabras aisladas; para ello, se hacía un análisis localizado de la señal de voz, y luego se generaba un modelo característico para cada una de las palabras a identificar. Para poder analizar una señal, tan no estacionaria como es la voz, se utilizaban las técnicas de análisis de división en ventanas de Fourier, análisis cepstral o análisis por predicción lineal. Una vez analizada la palabra pronunciada, se buscaba la generación de una plantilla de la misma, para poder ser utilizada en el momento de la identificación de la palabra pronunciada. Originalmente, cuando el usuario era uno, el mismo que entrenaba el sistema, se tomaba una pronunciación (repetición) de cada una de las palabras del pequeño vocabulario como referencia. Pronto surge el problema de la no uniformidad en la velocidad de pronunciación de palabras por parte de diferentes locutores e, incluso entre dos repeticiones de la misma palabra por parte de un único locutor. La solución a este problema llegó con la programación dinámica, la cual alinea palabras de duración diferente mediante un seguimiento de su contenido. Así aparecieron todos los sistemas de reconocimiento de palabras aisladas mediante comparación de plantillas. La gran problemática de estos sistemas es su fragilidad frente a ambientes ruidosos.

Cuando el número de palabras aumenta, la programación dinámica es ineficaz; demasiadas palabras complican el sistema de comparación de plantillas, y la semejanza entre palabras hace que las tasas de reconocimiento empiecen a bajar drásticamente. Ante este problema, se presentan tres soluciones clásicas: los Modelos de *Markov* Ocultos, las Redes Neuronales y los Modelos Auditivos.

A continuación se describen las técnicas clásicas más usadas en el proceso de reconocimiento automático del habla. Para ello, se ha considerado, como en la

mayoría de los sistemas de reconocimiento automático del habla, que el proceso se divide en dos etapas: primero, se realiza una extracción de parámetros característicos de la señal de voz, y segundo se obtiene una medida de similitud entre la muestra obtenida en la primera etapa y unas muestras de referencia. Siendo esta medida de similitud un valor determinista o una medida de probabilidad.

5.3.1. Extracción clásica de parámetros

El denominador común de todo sistema de reconocimiento de voz es la etapa inicial de procesamiento de señales, también llamada *front-end*, que convierte la señal de voz en alguna representación paramétrica para su posterior análisis y procesamiento. El objetivo de la fase de parametrización en el proceso de reconocimiento, es la extracción de la información relevante de la señal acústica analógica, eliminando las redundancias y la información asociada a las fuentes de variabilidad que tiene la misma. La información relevante será aquella que permita:

- Diferenciar unos fonemas de otros[45]. Los fonemas están caracterizados por:
    - La envolvente espectral del fonema, determinada por los formantes que los componen. Los formantes se definen como las frecuencias de resonancia del tracto vocal para cada fonema.
    - El tipo de excitación que los produce. Las vocales y consonantes sonoras están generadas mediante una excitación periódica. La frecuencia fundamental de la excitación es también una característica definitoria del fonema, aunque es

[45] En el capítulo 3, se hace una caracterización de los sonidos del habla desde el punto de vista articulatorio, acústico y perceptivo.

variable para los diferentes hablantes y las diferentes entonaciones.

- La energía de la señal. Las vocales y consonantes sonoras tienen mayor energía que las sordas, siendo la energía un buen parámetro de caracterización ya que presenta poca variabilidad para un mismo fonema una vez que ha sido convenientemente normalizada.

- Aportar datos sobre la prosodia de la frase tales como el acento, los tonos y la entonación. Esta información se obtiene analizando:
  - Las variaciones de la frecuencia fundamental.
  - Las variaciones de la duración de los fonemas.
  - La variación en la intensidad de los fonemas diferenciados.

Teniendo en cuenta la información expuesta como necesaria para caracterizar los fonemas y su prosodia, es razonable que la mayor parte de los sistemas de parametrización se basen en el análisis de la potencia espectral en tiempo corto (Rabiner & Juang, 1993). Al hacer este análisis la señal se divide en tramas lo suficientemente cortas como para poder considerar la señal cuasi-estacionaria. Estas tramas se someten a un análisis espectral y quedan caracterizadas por un vector de características. La Figura 68 muestra de manera general el proceso de parametrización con los posibles variantes en cada una de las etapas.

En primer lugar la señal de voz muestreada pasa un filtro de pre-énfasis (típicamente un filtro *FIR* de primer orden) que amplifica las altas frecuencias para compensar el efecto de los pulsos glotales y la impedancia de radiación. A continuación, se realiza un proceso de división en ventanas ya que debido a la naturaleza cambiante de la voz, resulta más conveniente aplicar el análisis a porciones

de ésta. Interesa observar la evolución de los distintos parámetros calculados y por ello se procesan segmentos o *ventanas* de la señal.

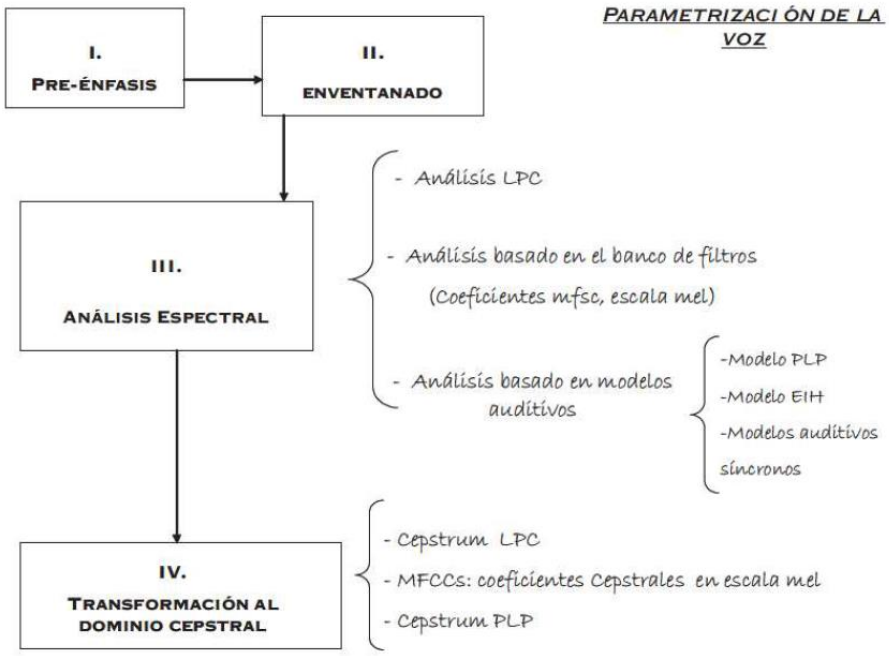


Figura 68. Representación esquemática del proceso de parametrización (Rabiner & Juang, 1993).

La señal es segmentada en tramas de longitudes del orden de 25ms. Para esta trama temporal la señal es cuasi-estacionaria y se puede analizar como tal. Para segmentar se usan funciones *ventana* recibiendo este proceso el nombre de *enventanado*. Cada *ventana* tendrá asociado un *peso*, no todas las secciones tendrán la misma importancia. De esta forma las muestras quedan ponderadas con los valores de la función escogida. La ventana de *Hamming* es la más usada por su compromiso

entre la resolución de frecuencias y distorsión armónica, para un coste computacional medio. Las muestra que se encuentran en los extremos de la ventana tienen un peso mucho menor que las que se hallan en el medio, lo cual es muy adecuado para evitar que las características de los extremos del bloque varíen la interpretación de lo que ocurre en la parte central, que es la más significativa. Existe un pequeño solapamiento en los extremos de las ventanas, esto se usa para mejorar la calidad de los resultados ya que se hace que los valores de las ventanas en sus extremos queden muy reducidos. Pero esto repercute en los tiempos de respuesta de los algoritmos utilizados, ya que los alarga ligeramente.

El análisis de los segmentos de habla obtenidos se puede hacer tanto en el dominio del tiempo como en el dominio de la frecuencia. En el capítulo 3, se ha descrito como en el nervio auditivo se pueden identificar básicamente estos dos tipos de representación de señales, la representación espectral y la temporal. Esta dualidad se debe principalmente al hecho de que las células de la cóclea, que presenta una organización tonotópica, dan una respuesta en función de la amplitud de la señal y de su envolvente temporal. Se ha descrito que el funcionamiento de la cóclea permite una representación espectral de la señal y los *IHCs* ofrecen una representación temporal. Por lo tanto, un tono a una frecuencia dada se representa en el nervio auditivo tanto por su posición, según la posición tonotópica que mejor responde a esa frecuencia, como por la periodicidad de las respuestas de todas las fibras que responden a ese estímulo (representación temporal). En la actualidad existen diferentes esquemas que usan ambos tipos de representaciones, temporal y de posición (frecuencia).

En el dominio del tiempo las magnitudes que se analizan son la energía local (*short-time energy*), la tasa de cruces por cero de la señal (*zero-crossing rate*) y su tasa de

cruce por nivel (*level-crossing rate*). Este dominio aporta un análisis de la señal rápido, sencillo y con una interpretación inmediata.

Sin embargo, el análisis espectral es el utilizado por su mayor potencia para caracterizar la información de la señal de voz. Las parametrizaciones usadas en *RAH* se derivan en su totalidad del análisis de la potencia espectral de las tramas de voz. Existen tres técnicas de análisis espectral. La primera, la técnica basada en bancos de filtros, trata de modelar dos aspectos del sistema auditivo: la sintonización y el aumento del ancho de banda según aumentan las frecuencias características de los haces del nervio auditivo. La segunda, la técnica del Cepstrum que realiza un análisis de la señal de voz teniendo en cuenta cómo ha sido producida. Y la técnica del espectro *LPC*, que se basa en el moldeado que realiza el tracto vocal sobre la señal acústica.

BANCOS DE FILTROS

Un banco de filtros, es una secuencia de filtros paso banda que permite trabajar de forma independiente con distintas porciones del espectro de la señal de voz. Esto es especialmente adecuado ya que no existe la misma cantidad de información útil en todas las bandas de la señal de habla. En el momento de diseñar estos filtros hay que tener en cuenta que el oído humano usa un ancho de banda mayor cuanta más alta es la frecuencia a la que trabaja. Y que las frecuencias fuera del rango (300 Hz, 3000 Hz) son de menor importancia para el oído (descrito en el capítulo 3).

Debido a que el espectro de potencias de la señal se obtiene aplicando la transformada de Fourier a las tramas de voz de ventanas que se solapan, aparecerán armónicos a frecuencias múltiplos de la frecuencia fundamental de las tramas. Este efecto se puede subsanar agrupando los conjuntos de componentes cercanos en unas 20 bandas de frecuencias antes de hacerles el logaritmo de la potencia. Cada filtro hará un promedio pesado de las componentes espectrales presentes en su banda, caracterizando el tracto vocal con la envolvente espectral suavizada. Es común usar

la escala *Mel* (descrita en el apartado 3.2.2 "Psicoacústica del habla" del capítulo 3) o de frecuencias subjetivas. El algoritmo de la energía a la salida de los filtros en escala *Mel* da lugar a los coeficientes *MFSC* (*Mel Frequency Spectral Coefficients*).

CEPSTRUM

Esta técnica realiza un análisis de la señal de voz teniendo en cuenta cómo ésta ha sido formada. La estructura básica de un modelo de producción de voz puede identificarse con la salida de un sistema de resonadores cuya entrada es una excitación (descrito en el apartado "Propiedades acústicas de los sonidos del español" del capítulo 3). La convolución de la excitación con la respuesta al impulso del sistema resonador produciría el modelo de la señal de voz. En este sentido parece lógico realizar un análisis de la señal de voz en el que se separe la fuente (excitación) del filtro (resonadores). A este proceso se le denomina deconvolución y al resultado un análisis *cepstral* de la señal de voz, Figura 69.

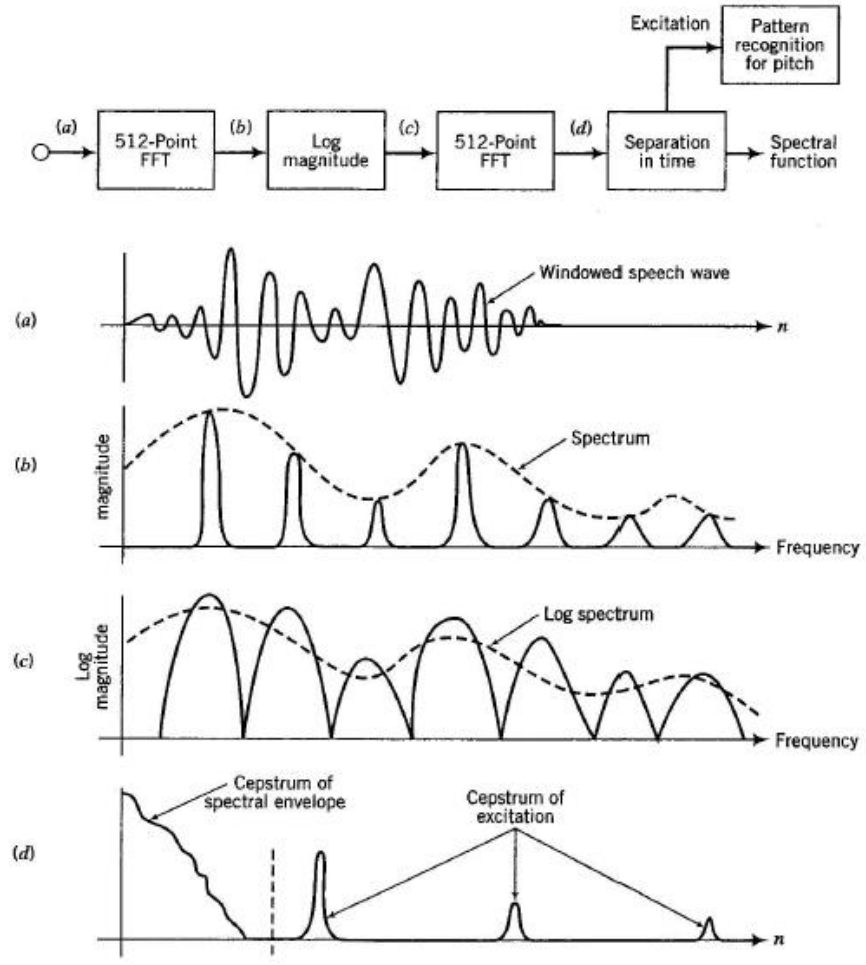


Figura 69. Análisis cepstral por deconvolución. Imagen tomada de (Gold & Morgan, 2000).

Las técnicas de análisis espectral que operan en el dominio de la potencia espectral logarítmica tienen la limitación de que debido a que los espectros de los filtros en bandas adyacentes están bastante correlados, originan coeficientes espectrales también bastante correlados. Es deseable eliminar esa correlación

manteniendo solo la información que sea útil para el reconocimiento. Para ello se utiliza un filtro de decorrelación homomórfica o *Cepstrum* que, mediante la transformada inversa de Fourier del logaritmo del espectro de potencias, lleva los coeficientes espectrales al dominio de la *cuefrencia* (*quefrency*) convirtiéndolos en coeficientes *cepstrales*.

Los coeficientes *ceptrales* representan la señal temporal que corresponde al espectro logarítmico de potencias. El dominio de la *cuefrencia* es un dominio homomórfico del dominio temporal. Esto implica que las convoluciones en el dominio temporal se convierten en sumas en su dominio homomórfico de la *cuefrencia*. Esto será sumamente útil ya que permitirá separar las señales de voz de los ruidos convolucionados con los que estén mezcladas. Las componentes de excitación y envolvente espectral del tracto vocal aparecerán en zonas separadas del dominio transformado de la *cuefrencia*, que se podrán separar mediante ventanas.

Haciendo un juego de paralelismo, los inventores de este operador homomórfico llamado *Cepstrum* (intercambia la posición de las cuatro primeras letras del término *spectrum*), llamaron a ese enventanado en el dominio de la *cuefrencia liftering* (intercambia la posición de las primeras letras del término *filtering*).

Las primeras aplicaciones del análisis *cepstral* fueron en el campo de la determinación del periodo fundamental: se descompone el espectro como producto de dos componentes, una de variación rápida y otra de variación mucho más lenta (la envolvente espectral). Al realizar el logaritmo, el producto se convierte en suma y se puede aplicar sin problemas el operador lineal de la transformada de Fourier.

ESPECTRO LPC

Durante un tiempo existió la preocupación de obtener un modelo matemático del tracto vocal, basado en la concatenación de tubos acústicos sin pérdidas ni bifurcaciones, a través del cual el sonido se va a propagar como una onda plana. En este sentido, los efectos del tracto vocal en la señal de excitación serán la creación de

una serie de resonancias, quedando así el tracto modelado como un filtro todo-polos (aunque pueden incluirse algunos ceros si se desea mejorar la codificación de sonidos nasales).

Este análisis de predicción lineal, *LPC* (*Linear Predictive Coding*) (Makhoul, 1973), es una técnica de parametrización que va a permitir la extracción de los formantes, pues realiza un suavizado del espectro de cada uno de los segmentos de voz considerados (ventanas) con un análisis por transformada de *Fourier*. Se elimina la información asociada a la velocidad de oscilación de los sonidos (*pitch*) manteniendo aquella asociada a la posición (y ancho de banda) de los formantes.

Por tanto, el modelo matemático sugiere que el tracto vocal puede modelarse mediante un filtro digital, siendo los parámetros extraídos los que determinan la función de transferencia. Es decir, dado un segmento de palabra se extraen sus parámetros que en este caso serán los *coeficientes del filtro*.

Uno de los principales inconvenientes es su poca robustez, por lo que se han considerado soluciones alternativas.

Parametrizaciones habituales del habla

En las últimas décadas distintas variantes de los bancos de filtros, la *LPC* y el *cepstrum* han sido usadas en la extracción de parámetros para *RAH*. Recientemente la mayoría de los sistemas ha convergido en el uso de un vector de parámetros *cepstrales* obtenido a partir de un banco de filtros (Gold & Morgan, 2000).

Nos centramos en dos de estas variantes, los *MFCCs* (*Mel Frequency Cpestral Coefficients*), y los coeficientes *PLP* (*Perceptual Linear Prediction*). Ambas técnicas tratan de estimar los parámetros mediante procesos que simulen la forma en que funciona el oído humano y cómo éste percibe sonidos con distintas componentes frecuenciales. Modelan el proceso de activación de las células ciliadas internas (*IHCs*) debido a los movimientos vibratorios de la membrana basilar.

Aunque las características de los *MFCCs* y los coeficientes *PLP* son similares, los coeficientes *MFCC* han demostrado ser los que mejores resultados dan como técnica de parametrización teniendo en cuenta el compromiso entre coste computacional y resultados obtenidos. Ambos tienen baja resolución en altas frecuencias, indicativo de métodos basados en bancos de filtros y proveen salidas ortogonales, típico del análisis espectral. Ambos proporcionan una representación de un espectro a corto plazo alisado que ha sido comprimido y cuantificado al igual que lo hace el oído humano. La principal diferencia entre ambos procedimientos radica en la naturaleza del espectro suavizado (basado en coeficientes *cepstrales* o en coeficientes *LPC*). Además, mientras *MFCC* usa la escala *Mel*, *PLP* usa la escala *Bark*. Aunque es cierto que ambas escalas son escalas de bandas críticas en las que los filtros están distribuidos a lo largo del eje frecuencial de forma lineal hasta los 1000 Hz y de forma logarítmica por encima de esta frecuencia (descrito en la sección "Enmascaramiento y bandas críticas" del capítulo 3).

Es interesante resaltar que estos análisis estiman el espectro localmente como un proceso cuasi-estacionario. Sin embargo, una de las características de la voz es su comportamiento dinámico. Debido a esto, numerosos investigadores utilizan además estimaciones de las variaciones temporales de los *cepstrum* o del espectro a corto plazo. Una de las medidas más comunes es la llamada delta del *cepstrum*, que es una aproximación por mínimos cuadrados de la pendiente local, y como tal es una estimación suavizada de la derivada local de la diferencia entre tramas vecinas del *cepstrum*.

Además, los experimentos han demostrado que las parametrizaciones de voz, hasta ahora descritas, tienen una sensibilidad excesiva al ruido presente en la señal de entrada. Para combatir dicha influencia se han diseñado otras parametrizaciones: Substracción Cepstral de la Media, *CMS* (*Cepstral Mean Substraction*) y *RASTA* (*RelATive SpecTrAl*). La técnica *PLP* se suele complementar con un filtrado *RASTA*, que elimina la influencia no deseada de la respuesta en frecuencia del canal de

comunicación, y recibe el nombre de *RASTA-PLP* (Hermansky, Morgan, Bayya, & Kohn, 1992).

Alrededor de los años 90 surgen nuevas técnicas cuyos resultados son bastante buenos, ligeramente mejores que los de los parámetros *MFCC* del dominio *cepstral*. Una de ellas, la técnica *GSD* (*Generalized Synchrony Detector)*, usada en el bloque III del modelo de *Seneff*, (Seneff, 1986). Este modelo refleja la respuesta de la membrana basilar a los estímulos acústicos usando un banco de filtros y haciendo un promedio de su actividad. A este promedio de las activaciones de cada filtro se le añade este detector de sincronismo que detecta las activaciones síncronas de varios canales adyacentes. La otra técnica, *EIH* (*Ensemble Interval Histogram*) (Ghitza, 1994), da una representación de la voz con alta resolución espectral basada en el cómputo de los histogramas de las frecuencias de activación de los filtros con los que modela la membrana basilar.

Estas técnicas captan cierta información adicional: la información temporal detallada de la señal, la supresión lateral de los canales adyacentes, los contrastes temporales y otras características no lineales del proceso de audición.

Sin embargo, esta línea de investigación no siguió desarrollándose ya que aunque los resultados eran buenos, llevaban asociado un coste computacional y de almacenamiento que no compensaba, no era factible para reconocimientos en tiempo real con coste computacional razonable.

En la actualidad existe un resurgimiento de esta línea de trabajo motivada por las capacidades de computación y almacenamiento superiores a las existentes en los años 80, por la necesidad de encontrar parametrizaciones que mejoren *MFCC* y permitan enfrentarse a los actuales retos de reconocimiento que también han aumentado, y por el descubrimiento de que la información de la sincronía/asincronía temporal de la señal de voz que el oído humano capta es muy útil para caracterizar los formantes y no es capturada por los *MFCCs*. El resultado ha sido el interés en crear algoritmos

que exploren los mecanismos de extracción de la información de sincronía de las salidas de los canales paralelos que modelan el canal auditivo, y capten mejor la relación entre frecuencias y tiempos. Ejemplos de esta línea de trabajo son:

- Uso de la transformada wavelet: el comportamiento del canal periférico auditivo es modelable mediante una transformada wavelet, que captura la información en los dominios de la frecuencia y el tiempo, mejorando algunas de las limitaciones de la transformada de Fourier.
- Mejoras del modelado síncrono de *Seneff* como la aportada con el algoritmo *ALSD* (*Average Localized Synchrony Detection*)(Abdelatty Ali, der Spiegel, & Mueller, 2002).
- Variaciones del *ZPCA* (*Zero-Crossing and Peak Amplitudes*) como las propuestas en (Kim, Lee, & Kil, 1999) y (Ghulam, Horikawa, & Nitta, 2006), que son mejoras del *EIH*, incorporando sincronismo.
- Uso de redes neuronales para capturar las operaciones no lineales en el espectro logarítmico de frecuencias.

5.3.2. Aproximaciones al modelo acústico

Tras la fase de extracción de parámetros característicos de la señal de voz se realiza la comparación de estos parámetros mediante alguna medida de similitud, con ciertos valores previamente almacenados, que se denominan modelos acústicos.

El modelo acústico de un sistema de reconocimiento de voz se puede generar siguiendo diferentes estrategias. La medida de similitud podría estar basada en valores determinísticos o en valores probabilísticos.

En este sentido, un sistema de reconocimiento de voz se puede considerar como un sistema de reconocimiento de patrones dedicado para detectar voz o, en otras

palabras, para identificar un mensaje dentro de una señal de audio adquirida como entrada (Varela Rincón & Loaiza Pulgarín, 2008).

Tanto las redes neuronales (RNA) como los modelos ocultos de *Markov*, *HMM* (*Hidden Markov Model*) se han usado exitosamente para tareas de diferentes complejidades en reconocimiento de voz, desde reconocimiento de palabras aisladas (Alvarez-López, 2004), hasta reconocimiento de voz continua.

La ventaja inmediata del reconocimiento de voz utilizando técnica de comparación de patrones, reside en que no es necesario descubrir características espectrales de la voz a nivel fonético, lo que evita tener que desarrollar etapas complejas de detección de formantes, de rasgos distintivos de los sonidos, de los tonos de voz, etc.

Esto está muy bien para un número finito de palabras, cuyo número no sea muy grande. Si se quiere desarrollar este sistema para un complejo entendimiento del lenguaje sería una completa locura, además de ser casi inútil. Por ejemplo, si se pronuncia la palabra *queso*, se tendría que hacer exactamente igual al momento de grabarla. Además se tendría que pronunciar con la misma velocidad y el mismo tono, si no el sistema no podría identificar la palabra. Para ello, se necesita un sistema que aprenda por sí mismo y que al final intente estipular que palabra es.

De todos modos, existe la necesidad de aplicar este sistema única y exclusivamente a casos donde el vocabulario sea pequeño y limitado.

La bibliografía de clasificación de patrones recoge las tres siguientes aproximaciones: comparación de plantillas o patrones, modelo oculto de Markov y redes neuronales artificiales. Estas tres, conceptualmente, se diferencian en la referencia que se toma para clasificar los patrones.

Comparación de plantillas o patrones

Esta aproximación es la más antigua y toma como referencia plantillas o patrones de los objetos que se clasifican. Para el reconocimiento de patrones se realiza una comparación entre una muestra representativa de la señal de voz, cuyas características están representadas en un determinado dominio paramétrico, con una señal incógnita de entrada que se desea clasificar y que se encuentra representada en el mismo dominio de parámetros que la muestra.

La anterior muestra representativa de la señal de voz se denomina patrón de referencia; y al conjunto de representantes se denomina conjunto de patrones de referencia. La señal de entrada a reconocer se denomina patrón de test. Por tanto, el reconocimiento de patrones consistirá en comparar el patrón de test con cada uno de los elementos del conjunto de patrones de referencia.

La plantilla o patrón es una secuencia de características acústicas ordenadas en el tiempo. La comparación con estas plantillas exige un alineamiento temporal no lineal de las mismas con los datos de entrada, y una medida de distancia. Esta técnica, que trata de modelar la voz de un modo determinista, tiene la limitación de que necesita mucha capacidad de almacenamiento para los patrones de referencia con los que se ha de comparar, y presenta inconvenientes en el habla continua por la dificultad de la lineación temporal: la duración de una palabra no tiene que ser de una duración determinada, por lo que puede que no coincida con la de la plantilla; y el ritmo con el que se realiza la pronunciación no tiene que mantenerse constante por lo que no se ajustará a la plantilla en ese sentido, ya que éste depende de la persona.

Mediante técnicas de programación dinámica basadas en el algoritmo DTW (*Dinamyc Time Warping*, alineamiento temporal dinámico (Deller, Proakis, & Hansen, 1993)), se entrena el sistema obteniendo plantillas o patrones de las unidades que habrá que reconocer. Esta técnica surge precisamente a partir de la problemática inherente a diferentes realizaciones de una misma locución, en las que se observa una

variabilidad interna en la duración de los grupos fónicos que la forman, de modo que no existe una sincronización temporal (alineamiento temporal) Figura 70. Además, esta falta de alineamiento no obedece a una ley fija (p. e., un retardo constante), sino que se da de forma heterogénea, produciéndose así variaciones localizadas que aumentan o disminuyen la duración del tramo de análisis.

La problemática asociada hace referencia a la dificultad añadida en el proceso de medida de distancias entre patrones, puesto que se estarán comparando tramos que pueden corresponder a unidades fónicas distintas. Será necesario alinear temporalmente la locución para proceder a realizar una medida de distancia entre patrones cuyo nuevo eje temporal haya homogeneizado las variaciones iniciales.

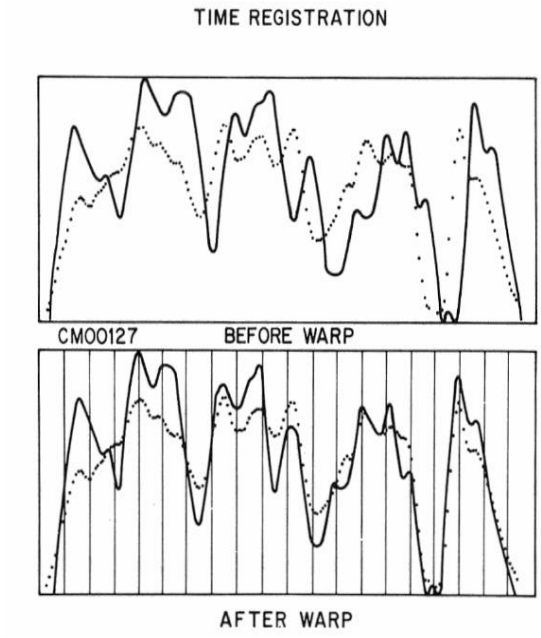


Figura 70. Alineamiento Temporal Dinámico: dos realizaciones de la misma locución antes y después de ser alineadas.

Modelo estadístico: modelos ocultos de *Markov*

Un Modelo Oculto de *Markov* (*HMM*) es una técnica de reconocimiento de plantillas o patrones, basado en un modelado estocástico o aleatorio de dichos patrones, permitiendo una mayor flexibilidad para representar secuencias de duración variable (Rabiner & Juang, 1986).

Un modelo de *Markov* se define como una máquina de estados estocásticos finitos, donde la probabilidad de pasar al estado siguiente depende únicamente del estado actual (proceso de *Markov*), y asociado a cada transición entre estados se produce un vector de observaciones. Por tanto, cambia de estado, siguiendo una distribución probabilística, una vez en cada instante de tiempo. Y por cada instante de tiempo en que se produce un cambio de estado se genera una nueva salida, teniendo en cuenta también ciertas densidades de probabilidad. La particularidad de los *HMMs*, es que el paso de unos estados a otros de la cadena no es directamente observable, está oculto. Por tanto, se puede decir, que un *HMM* está compuesto de 2 procesos estocásticos, el oculto, correspondiente a las transiciones entre estados, y el observable o no oculto, correspondiente a la generación del vector de observaciones que se produce en cada estado, y que representa la plantilla a reconocer. Además, cada estado tiene asociada una distribución de probabilidad sobre los posibles símbolos de salida, con lo que la secuencia de símbolos generada por un *HMM* proporciona cierta información sobre la sucesión de estados. En la siguiente Figura 71, se muestra el esquema de un modelo oculto de *Markov* formado por una cadena de tres estados $Q = \{q^1,q^2,q^3\}$ y las correspondientes transiciones entre los estados con sus probabilidades $P(q'|q)$. Cada estado tiene asociado una función de densidad de probabilidad de emisión, $P(x|q)$. El estado inicial se indica con una flecha que no parte de ningún estado y el estado final con una flecha que no va a ningún estado destino.

Desde el punto de vista del habla, cada nodo representaría una unidad acústica (por ejemplo, fonema) diferente y cada sucesión de unidades acústicas con su probabilidad de transición asociada, representará una palabra determinada. Esta vendrá dada en su totalidad con su probabilidad de emisión del *HMM*. Se hace un alineamiento no lineal con el algoritmo de Viterbi (Forney, 1973), entre los datos de entrada y las palabras de un diccionario cuyos términos son patrones estocásticos. Los alineamientos se establecen como probabilidades de que la secuencia analizada sea generada por los distintos Modelos de *Markov*.

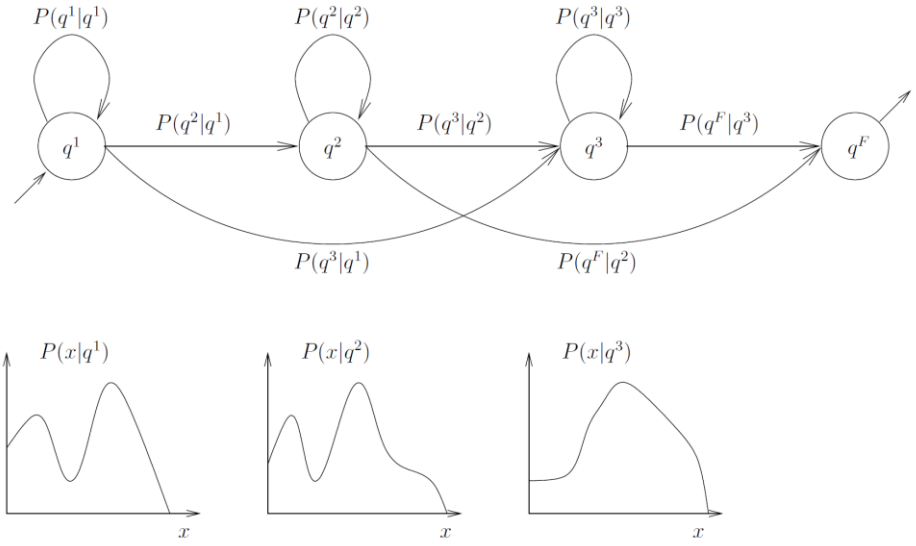


Figura 71. Ejemplo de HMM de 3 estados.

Hasta la fecha este método es el que mejores resultados proporciona y el más utilizado. Las excelentes prestaciones ofrecidas por los *HMMs* se deben, especialmente, a la forma en que modelan las deformaciones temporales de los patrones acústicos y a la existencia de algoritmos eficientes, como son el algoritmo de Baum-Welch (Baum, 1972) y el algoritmo de Viterbi (Jelinek, 1976), que permiten estimar los parámetros de los modelos a partir de un conjunto de muestras de

entrenamiento supervisado. Sin embargo, estos algoritmos presentan un importante inconveniente: la escasa capacidad discriminativa del conjunto de modelos resultante. Otra limitación de esta aproximación es el hecho de que se precisa realizar una serie de asunciones (casi siempre poco realistas) acerca de la naturaleza de las funciones de densidad de probabilidad que se quieren modelar.

En el marco del RAH, los métodos conexionistas se emplean para resolver subtareas en el sistema global, dando lugar a sistemas híbridos con mayor capacidad discriminativa, basados en métodos estructurales y Redes Neuronales Artificiales. A continuación, se describe qué pueden aportan las redes neuronales artificiales a los sistemas actuales de RAH, (Lippmann., 1989).

Redes Neuronales Artificiales

Las redes de neuronas artificiales, *RNA*, son sistemas de procesamiento de información cuya estructura y funcionamiento están inspirados en las redes neuronales biológicas. En todo modelo de *RNA* se tienen cuatro elementos básicos:

- Un conjunto de conexiones, pesos o sinapsis que determinan el comportamiento de la neurona, las cuales pueden ser excitadores, presentan un signo positivo (conexiones positivas), e inhibidoras presentan un signo negativo (conexiones negativas).
- Una función que se encarga de sumar todas las entradas multiplicadas por sus pesos correspondientes.
- Una función de activación que puede ser lineal o no lineal empleada para limitar la amplitud de la salida de la neurona.
- Una ganancia exterior que determina el umbral de activación de la neurona.

La aproximación basada en Redes Neuronales Artificiales, RNA, toma como referencia para el modelado los patrones de actividad. Se define como un modelo computacional paralelo compuesto de unidades procesadoras adaptativas con una alta interconexión entre ellas mediante pesos. Estas unidades se agrupan en diferentes capas, distinguiéndose la capa de entrada y la de salida. En la Figura 72 se muestra un ejemplo de *RNA*. Se caracterizan por ser estructuras intrínsecamente no lineales con capacidad para aprender una determinada tarea a partir de pares de observación-resultado sin hacer asunciones sobre el modelo subyacente.

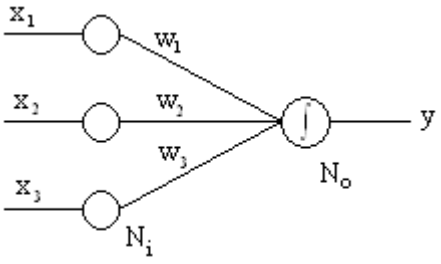


Figura 72. Ejemplo de Red Neuronal Artificial (w_i pesos, x_i vectores de entrada, y vector de salida, f función de activación, N_i capa de entrada, N_o capa de salida).

Desde que Frank Rosenblatt en 1957 introdujo el modelo de perceptrón de una sola capa (Rosenblatt, 1958), las RNAs se convirtieron en una herramienta poderosa para solucionar diversos tipos de problemas relacionados con la clasificación, estimación funcional y optimización del reconocimiento de patrones.

Según la topología de la red existen varios tipos de redes: Perceptrón multicapa (MLP, *Multi-Layer Perceptron*), máquina de Boltzman, Adaline, etc. La más utilizada en los sistemas de reconocimiento de voz es la MLP. Son redes multicapa y de aprendizaje supervisado, es decir, necesitan ser entrenadas con conjuntos de datos

observación-resultado. Además, son de alimentación hacia delante (*feedforward*), y su estructura se caracteriza por introducir entre la capa de entrada (N_i) y la de salida (N_o) una o más capas intermedias u ocultas (N_h), con capacidad de procesamiento y sin otras conexiones con el exterior. En la Figura 73 se muestra un ejemplo de perceptrón multicapa.

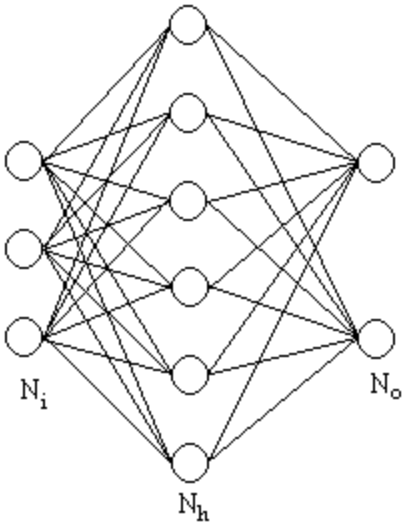


Figura 73. Ejemplo de perceptrón Multicapa
con una única capa oculta (N_h).

La arquitectura del perceptron, llamada mapeo de patrones (*pattern-mapping*), aprende a clasificar modelos mediante un aprendizaje supervisado. Los modelos que clasifica suele ser generalmente vectores con valores binarios y las categorías de la clasificación se expresan mediante vectores binarios.

El perceptron presenta dos capas de unidades procesadoras (PE) y sólo una de ellas presenta la capacidad de adaptar o modificar los pesos de las conexiones. La arquitectura del perceptron admite capas adicionales pero éstas no disponen de la capacidad de modificar sus propias conexiones.

La siguiente Figura 74 muestra la unidad procesadora básica del perceptron. Las entradas ***xi*** llegan por la parte izquierda (constituyen el vector del patrón desconocido), y cada conexión con la neurona tiene asignada un peso de valor ***wi.*** Un aspecto común en las *RNAs* es la existencia de una entrada especial que siempre tiene un valor fijo, +1, el cual puede ser usado como un valor fijo de referencia.

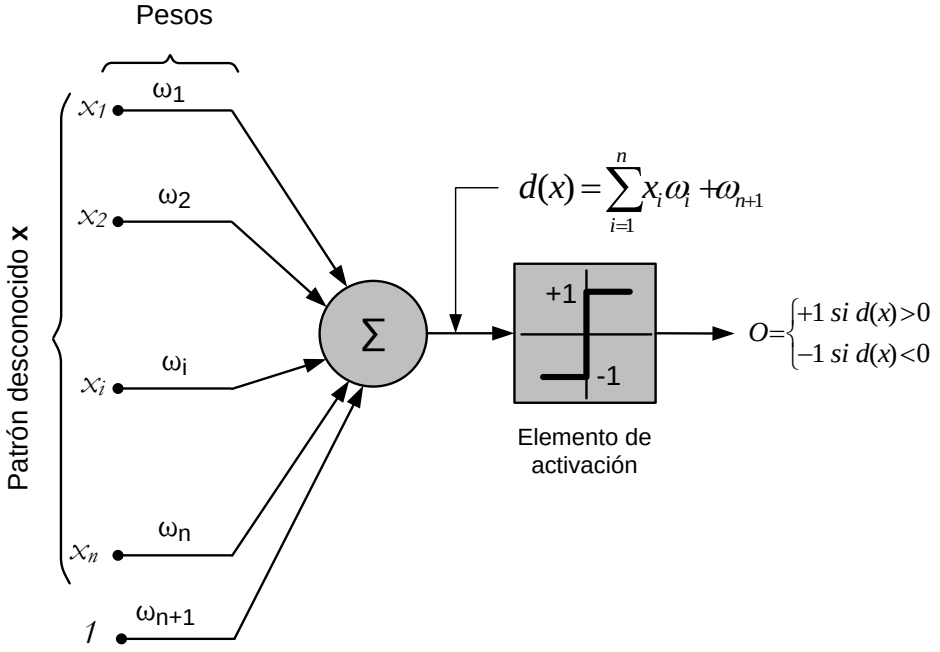


Figura 74. Esquema de unidad procesadora básica del perceptrón.

La unidad procesadora del perceptrón realiza la suma ponderada de las entradas:

$$\mathrm{d}(\mathbf{x}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \omega_{\mathrm{i}} \mathrm{x}_{\mathrm{i}} + \omega_{\mathrm{n}+1} \qquad (5.1)$$

Esta suma ponderada es una función de decisión lineal con respecto a las componentes del vector del patrón. Los coeficientes ω_i, $i = 1,2,\ldots,n,n+1$, llamados pesos, modifican las entradas antes de ser sumadas y alimentar el elemento de umbral. Los pesos realizan el mismo papel que las sinapsis en los sistemas neuronales biológicos. La función que convierte la suma ponderada de las entradas en la salida final del perceptrón se llama función de activación. Ésta comprueba si la suma de las entradas ponderadas es mayor o menor que un cierto umbral. Por tanto, si $d(\mathbf{x}) > 0$, la salida del perceptrón será +1 indicando que el patrón $\mathbf{x}$ es reconocido como perteneciente a la clase ω_1. Por el contrario, cuando $d(\mathbf{x}) < 0$, indica que el patrón $\mathbf{x}$ no pertenece a dicha clase.

Se han conseguido buenos resultados usando redes neuronales para reconocimiento de patrones acústicos debido a su capacidad de clasificación y sobre todo a su capacidad de aprender una determinada tarea a partir de pares de observación-resultado sin hacer ninguna suposición del proceso que modelan, pudiendo encontrar un modelo adaptado a los datos independientemente de la naturaleza de los mismos. Las redes neuronales, por tanto, son capaces de realizar una clasificación subléxica (fonética) adecuada a nivel de vector de características acústicas debido a su gran capacidad discriminativa, pero no existen arquitecturas conexionistas ni algoritmos de aprendizaje asociados que modelen adecuadamente la estructura temporal del habla. Se necesita un post-proceso que interprete la secuencia de salida de la RNA para generar una secuencia de fonemas (o cualquier otra unidad subléxica elegida). A pesar de todo, se han propuesto algunas arquitecturas recurrentes que tratan de realizar un alineamiento temporal, pero solo se han experimentado en tareas pequeñas, debido a que los algoritmos de aprendizaje son extremadamente lentos, lo que los hace impracticables en tareas reales de reconocimiento automático del habla continua (Bengio, 1993). En cambio, para reconocimiento de palabras

aisladas, con pequeños diccionarios, los modelos conexionistas han demostrado ser muy efectivos, superando las prestaciones que ofrecen otras técnicas (Lippmann, 1989), (Waibel, Hanazawa, Hinton, Shikano, & Lang, 1989) y (Sakoe, Isotani, Yoshida, Iso, & Watanabe, 1989).

No obstante, en general, su uso no ha superado al de los *HMMs* porque presentan limitaciones como un tiempo excesivamente elevado para su entrenamiento, o el desconocimiento a priori del número de nodos y capas necesarias para cada problema de clasificación. Además, las RNAs, en general, sólo son capaces de procesar vectores de longitud fija, por lo que no se suelen usar de manera aislada en un sistema de RAH.

Sistemas híbridos

Los sistemas híbridos son el resultado de la integración de la capacidad discriminativa que presentan las redes neuronales con las propiedades de alineamiento temporal de los modelos de Markov. La idea básica es emplear algún tipo de red neuronal para estimar probabilidades que se utilizarán en un alineamiento con HMMs.

Han aparecido numerosos ejemplos de sistemas híbridos, siendo el presentado por el equipo de Bourland (Bourland & Morgan, 1993), (Bourland & Morgan, 1994) y (Bourland, 1995), el sistema que mayor nivel de popularidad ha alcanzado. En esta aproximación se utiliza un perceptrón muticapa como estimador de las probabilidades de emisión de los HMMs. Es el sistema híbrido clásico MLP/HMM. Otros tipos de combinaciones utilizan como modelo de red neuronal, redes recurrentes (T. Robinson, 1994) (Neto et al., 1995) o redes basadas en funciones radiales (Renals, McKelvie, & McInnes, 1991), (Singer & Lippmann., 1992) o redes con retardos ("*TDNN, Time Delay Neural Networks*") (Dugast, Devillers, & Aubert., 1994).

Los sistemas de reconocimiento de voz automático (RAH) están lejos de alcanzar la eficiencia del reconocimiento humano, sobre todo en presencia de ruido. Los investigadores han buscado en la biología pistas sobre la *robustez* del sistema auditivo humano. De hecho, la característica más usada en las aplicaciones RAH son los coeficientes MFCC (*Mel Frequency cepstral coefficient*) que imitan la distribución logarítmica de todas las frecuencias observadas en la cóclea.

Sin embargo los sistemas RAH actuales están basados en el procesamiento de segmentos o *ventanas* de la voz que usan modelos ocultos de Markov (HMMs). Este procesamiento no se parece a la computación neurobiológica. Se cree que los trenes de pulsos (*spikes train*) van a proporcionar las claves para desarrollar sistemas auditivos robustos al ruido.

5.4. Criterios de evaluación del RAH

La evaluación cuantitativa del sistema de reconocimiento provee medidas estandarizadas del funcionamiento del mismo y da la posibilidad de comprarlo con otros, posibilitando así la extracción de conclusiones al respecto. Para obtener una evaluación cuantitativa del reconocedor hay que conocer la probabilidad de cometer errores de reconocimiento, la complejidad computacional y los requisitos de memoria del proceso de reconocimiento, y el tiempo de respuesta del sistema.

5.4.1. Tasa de error y precisión del reconocimiento

La tasa de error se define como la capacidad del reconocedor automático del habla de cometer errores de reconocimiento. En el caso de un reconocimiento de palabras aisladas, la tasa de error de palabra, *WER* (*Word Error Rate*) se define como:

$$WER = \frac{n_e}{n_p} \tag{5.2}$$

Donde np representa el número total de palabras reconocidas y ne el número de palabras clasificadas erróneamente. Para el caso de reconocimiento de habla continua, el reconocimiento se hace frase a frase y se definen tres tipos de errores: errores de inserción ni, errores de borrado nb, y errores de sustitución ns. En este caso la tasa de error de palabra se define como:

$$WER = \frac{n_i + n_s + n_b}{n_p} \tag{5.3}$$

Algunos autores trabajan con la tasa de aciertos de palabras, *WAcc* (*Word Accuracy*) cuya expresión es:

$$WAcc = 1 - WER = 1 - \frac{n_i + n_s + n_b}{n_p} \tag{5.4}$$

5.4.2. Intervalo de confianza de la medida de error

La tasa de error antes definida es una estimación de la probabilidad de error dentro de un determinado intervalo de confianza cuya amplitud dependerá del número total de pruebas con los que se haya obtenido la tasa de error. Si se asume una distribución binomial $B(n,p)$ del número de elementos reconocidos correctamente, siendo p la probabilidad de aciertos y n el número total de ensayos, se puede definir un intervalo de confianza centrado en el valor estimado de la probabilidad de acierto $\hat{p}$ que contendrá con probabilidad $(1-\alpha)$ la probabilidad de acierto. Mediante el teorema del límite central se demuestra que esta probabilidad $\hat{p}$ tiende a la distribución normal $N(0,1)$ y el intervalo de confianza para el valor $\hat{p}$ será:

$$\left[\hat{p} - z_{1-\frac{\propto}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\frac{\propto}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \quad (5.5)$$

Donde $z_{1-\frac{\propto}{2}}$ es el cuantil $(1-\frac{\propto}{2})$ de la distribución normal estándar. En la expresión anterior (5.4), se puede apreciar que cuanto mayor es el número de elementos reconocidos, más estrecho será el intervalo de confianza.

5.4.3. Aspectos computacionales y tiempo de respuesta

El coste computacional del proceso de reconocimiento se puede dividir en dos partes, el coste computacional de la parametrización de la señal de voz, y el coste computacional del proceso de descodificación. Hay que resaltar que el coste computacional de la parametrización unido al coste de almacenamiento, aumentará con los algoritmos más elaborados como son las parametrizaciones basadas en los modelos auditivos. En el proceso de reconocimiento, la descodificación es la que tiene un coste computacional que puede ser bastante elevado, sobre todo para el reconocimiento de habla continua. Las búsquedas en las gramáticas con alta perplejidad hacen que el árbol de búsqueda se expanda. Para controlar esta expansión, y con ello el coste y tiempo de respuesta, se define un umbral de poda y se utilizan algoritmos heurísticos (Ravishankar, 1996) que descartan las opciones menos probables.

Los costes se deben sintonizar según los requisitos de tasa de error de la aplicación, y los medios con los que se cuenta en la implementación.

Capítulo 6

Cóclea artificial pulsante

Uno de los objetivo de esta tesis es implementar un sistema neuromórfico que emule el comportamiento del sistema auditivo humano e identifique una secuencia de fonemas vocálicos en tiempo real.

En este sistema se distinguen dos etapas. En la primera etapa, descrita en este capítulo, una cóclea artificial pulsante modela el comportamiento de la membrana basilar y los *IHCs*[46] de una cóclea biológica. Esta cóclea artificial pulsante es capaz de convertir la señal sonora en un tren de pulsos *AER*[47]. La salida de esta primera etapa es analizada en la segunda etapa del sistema, descrita en el capítulo 7, donde se realiza

---

[46] *IHC,* siglas en inglés *Inner Hair Cell*. Son células ciliadas del órgano de Corti, con capacidad para producir pequeñas descargas eléctricas. De ellas salen los haces de fibras que componen el nervio auditivo. (Descrito en el apartado "Oído interno" de la sección 3.2.1, "El sentido de la audición y el sistema auditivo").

[47] *AER,* siglas en inglés *Address Event Representation*. Protocolo de comunicación conducido por eventos, usado originalmente en implementaciones VLSI de redes neuronales para transferir pulsos entre neuronas. (Descrito en la sección 2.3.1, del capítulo "Estado de los desarrollos neuromórficos actuales").

el proceso de identificación de la secuencia de fonemas vocálicos presentes en dicha señal de voz.

6.1. Diseño e implementación de un modelo de cóclea basada en filtros digitales

Se ha descrito en los anteriores capítulos que un completo modelo neuromórfico de la cóclea debe incluir un mecanismo de selección frecuencial (imitando la membrana basilar) y un mecanismo de transformación de los movimientos mecánicos de la membrana basilar en señales eléctricas (*IHCs*). Ambas funciones son modeladas, en esta cóclea digital pulsante (Figura 75), mediante un banco de filtros digitales combinados con generadores de pulsos, capaces de convertir la señal sonora en un tren de pulsos. La idea es transmitir a través de un bus *AER* una secuencia de eventos *AER* que corresponda con la respuesta de todas las bandas, a partir de la señal *PCM*[48]filtrada. A cada par formado por un filtro y un generador de pulsos se le llama banda.

La señal sonora es digitalizada con el ADC[49] del códec de audio WM8731/WM8731L (Microelectronics, 2009), de la placa de desarrollo EP4CE115F29C7. Esta placa de desarrollo dispone de una *FPGA* de la familia *Cyclone IV E* (Altera, 2010). La salida del ADC, configurado a una frecuencia de muestreo de 48 kHz, es un valor codificado en *PCM*, de 20 bits. En esta implementación, de estos 20 bits de la señal *PCM*, el banco de filtro solo trabaja con los 18 bits más significativos.

---

[48] *PCM,* siglas en inglés *Pulse Code Modulation*.

[49] *ADC,* siglas en inglés *Analogue to Digital Converter*. Dispositivo electrónico capaz de convertir una entrada analógica de voltaje en un valor binario.

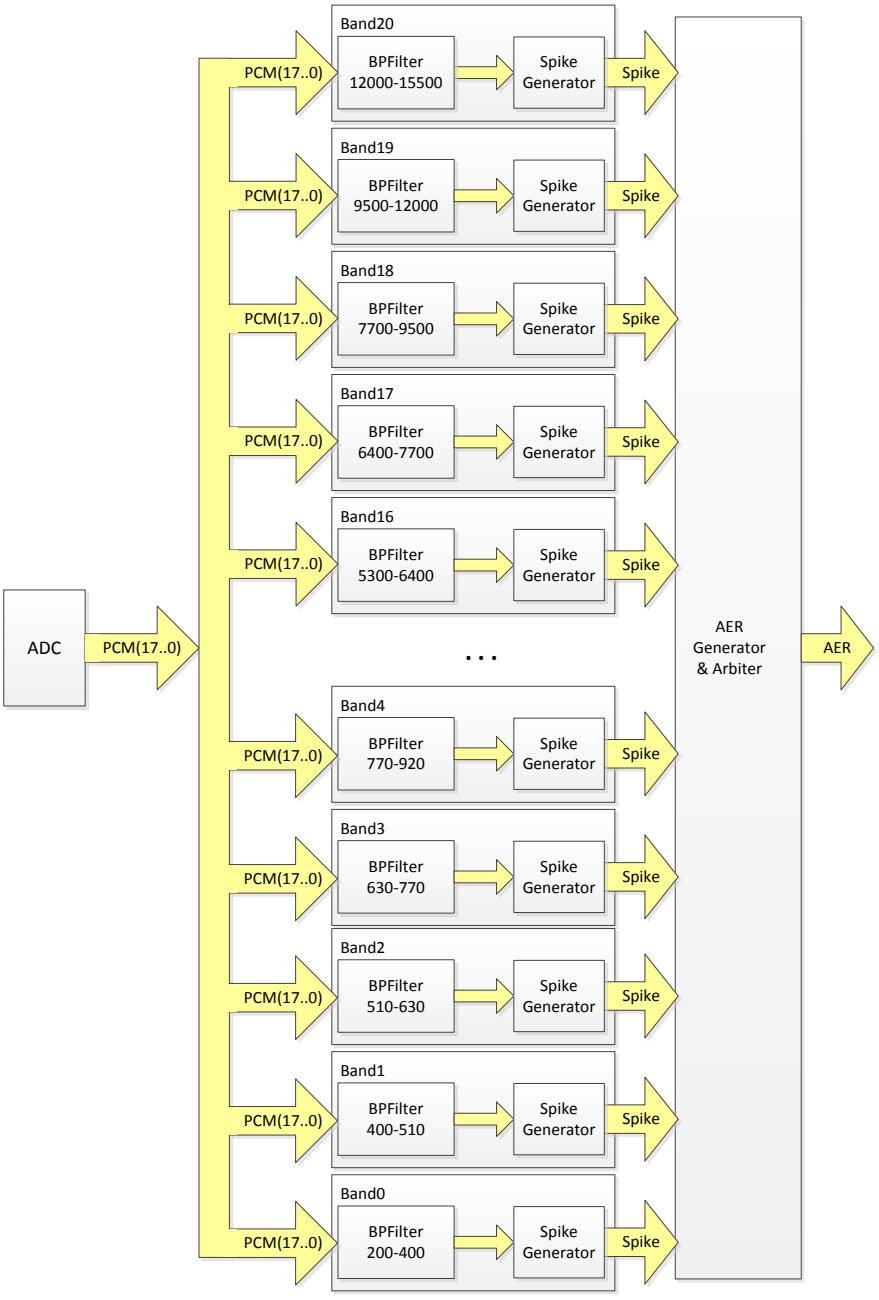


Figura 75. Cóclea digital pulsante de 21 bandas. Cada banda está formada por un filtro paso banda y un generador de pulso.

En el esquema general del sistema que implementa la cóclea digital pulsante, Figura 75, se observa como se está realizando un procesamiento paralelo (propiedad de los sistemas neuromórficos), puesto que cada filtro paso banda y su generador de pulsos asociado trabajan de forma independiente.

6.1.1. Banco de filtros digitales

El banco de filtros digitales, capaz de descomponer la señal de audio en sus diferentes componentes frecuenciales, es una estructura paralela de 21 filtros digitales paso banda. Están organizados tonotópicamente cubriendo todo el rango de frecuencias del sistema auditivo humano, desde altas frecuencias, 20 kHz (base de la cóclea) a bajas frecuencias, 200 Hz (ápice de la cóclea). La subdivisión del rango de frecuencias de este modelo de banco de filtros se basa en la separación en bandas críticas propuesta por Zwicker (Zwicker, 1961), tal como se muestra en la Tabla 12.

Tabla 12. Distribución de las bandas críticas en función de la frecuencia.

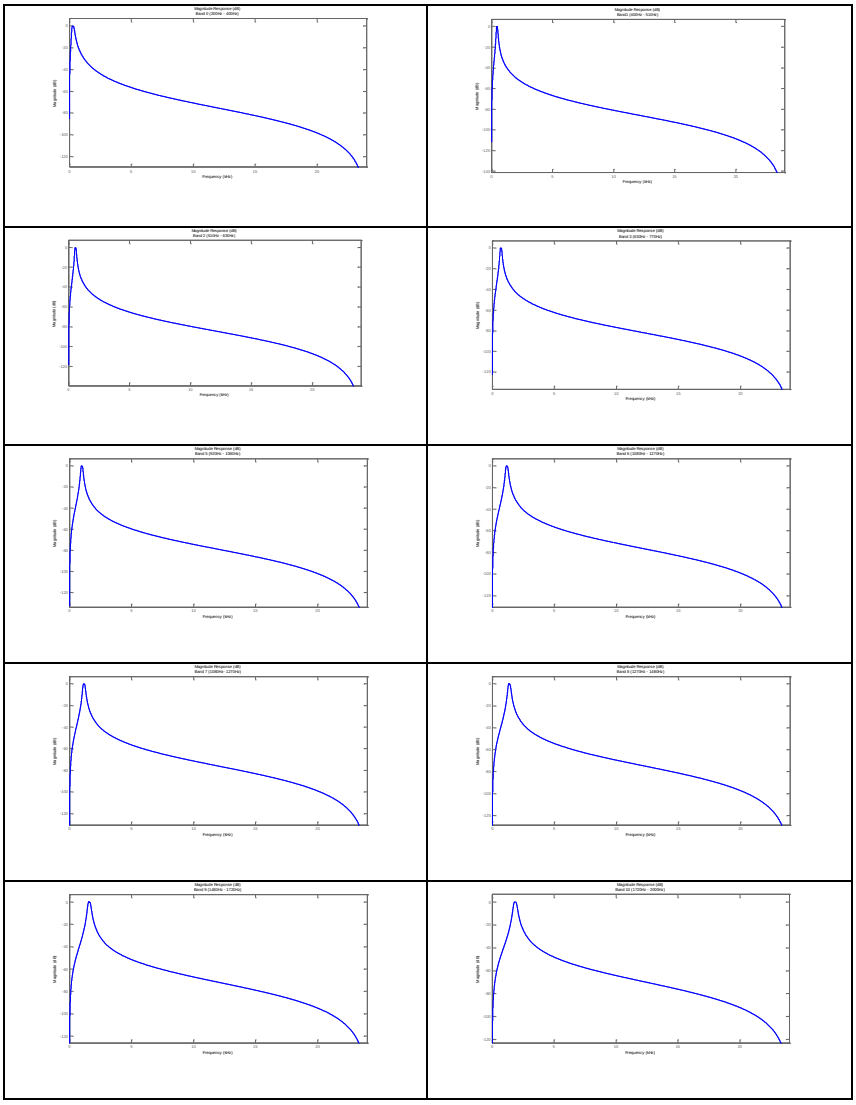
Banda	**Frecuencia central (Hz)**	**Rango de frecuencias del filtro paso banda (Hz)**
0	350	200 - 400
1	450	400 - 510
2	570	510 - 630
3	700	630 - 770
4	840	770 - 920
5	1000	920 - 1080
6	1170	1080 - 1270
7	1370	1270 - 1480
8	1600	1480 - 1720
9	1850	1720 - 2000
10	2150	2000 - 2320
11	2500	2320 - 2700
12	2900	2700 - 3150

13	3400	3150 - 3700
14	4000	3700 - 4400
15	4800	4400 - 5300
16	5800	5300 - 6400
17	7000	6400 - 7700
18	8500	7700- 9500
19	10500	9500 - 12000
20	13500	12000 - 15500

Para modelar la funcionalidad de la cóclea biológica, hemos elegido filtros paso banda. Los filtros paso banda determinan si una señal contiene una componente en frecuencia dentro del rango de frecuencia que define a dicho filtro. Al igual que la cóclea biológica, que actúa como un analizador de frecuencia (descrito en el apartado "Oído interno" del capítulo 3) y es capaz de descomponer un sonido complejo en sus componentes frecuenciales. Son filtros lineales *IIR* de dos secciones de segundo orden. Se ha elegido este tipo de filtro, frente a un filtro lineal *FIR*, porque con ellos se consigue mejores características de calidad para un mismo orden de filtro. Además los filtros *FIR* proporcionan fase lineal, característica que no se usa en los filtros cocleares. Todos los filtros son *Butterworth*, cuya respuesta en magnitud es máximamente plano en la banda de paso. También, se ha optado por una aritmética punto-fijo en complemento a dos, lo cual implica la estructura Forma Directa II transpuesta, con el objetivo de una implementación con un coste más eficiente, menos memoria y menos tiempo de computación. Los filtros digitales han sido descritos en *VHDL*[50] con ayuda de la herramienta *Filter Design HDL Coder* (MathWorks, 2012).

[50] *VHDL*, siglas en inglés formada a partir de la combinación *VHSIC* (*Very High Speed Integrated Circuit*) y *HDL* (*Hardware Description Language*). Es un lenguaje de descripción hardware definido por el IEEE (*Institute of Electrical and Electronics Engineers* – ANSI/IEEE 1076-1993), (IEEE, 2008).

La siguiente Figura 76 muestra el comportamiento de la respuesta en frecuencia obtenida para cada uno de los 21 filtros paso banda que constituyen el banco de filtros de la cóclea digital pulsante.



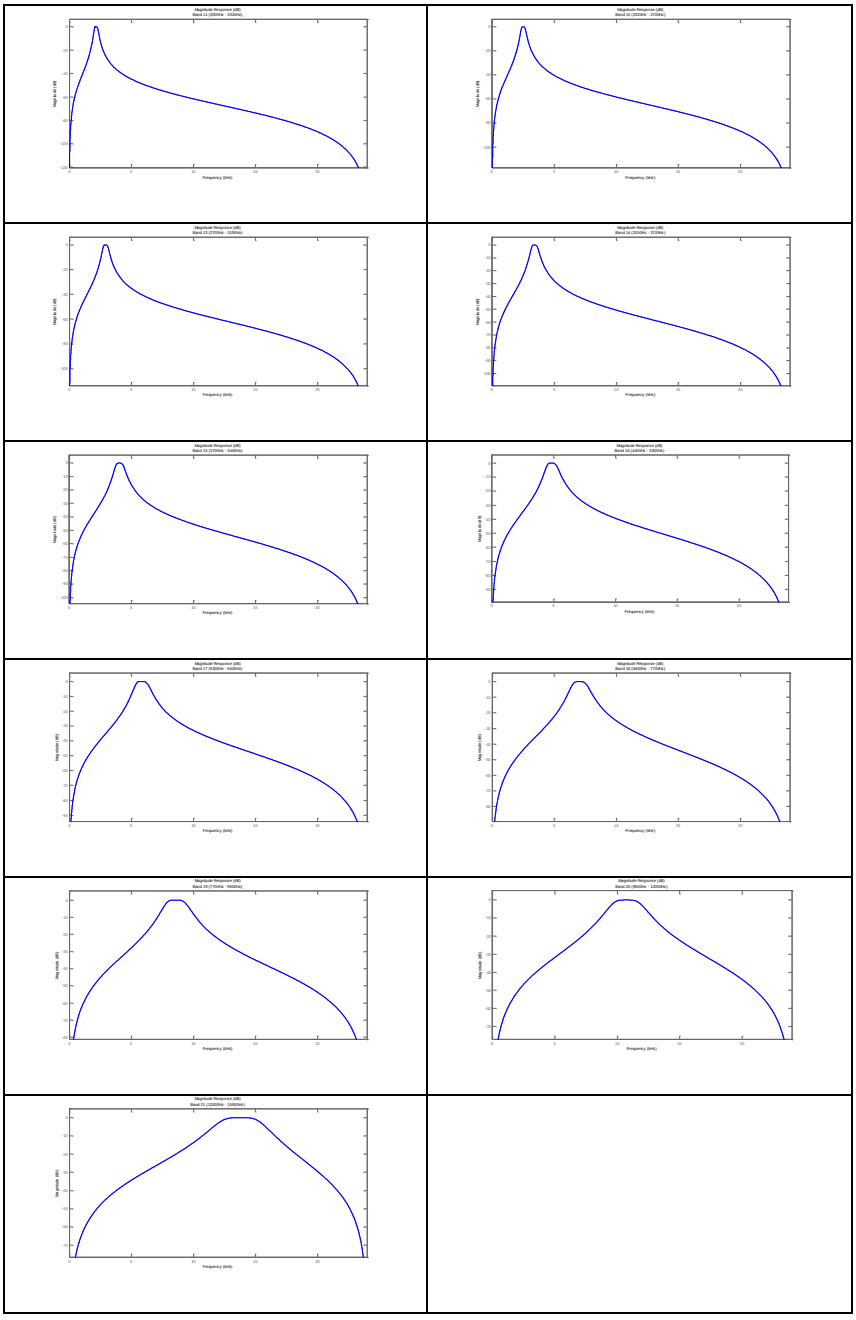


Figura 76. Respuesta en frecuencia o Diagrama de Bode de cada filtro paso banda, correspondiente a las 21 bandas de la cóclea basada en filtros digitales.

En la siguiente Figura 77 se representa la respuesta en frecuencia del banco de filtros de cada una de las 21 etapas implementadas.

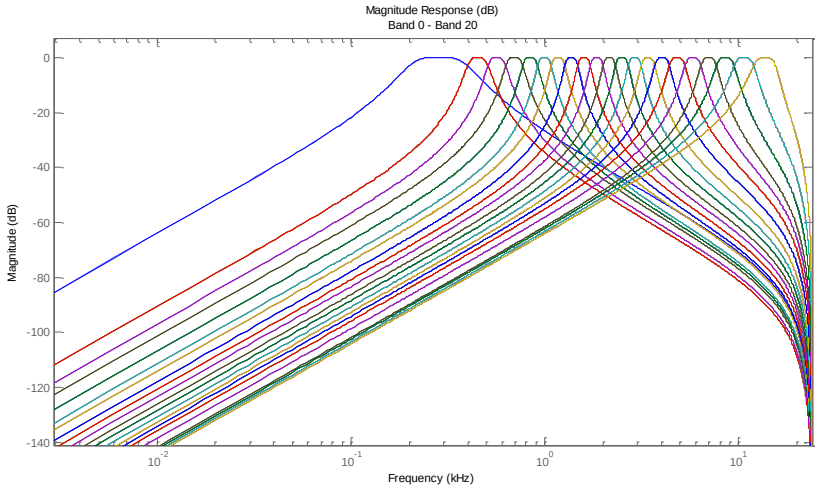


Figura 77. Respuesta en frecuencia o Diagrama de magnitud de Bode de todos los filtros paso banda, correspondiente a las 21 bandas de la cóclea digital.

Ante el estímulo de una señal de audio, cada filtro del banco de filtros responderá de un modo más activo, si las componentes en frecuencia de la señal de entrada coinciden con su frecuencia central.

6.1.2. Generador de pulsos

A la salida de cada filtro, del banco de filtros, se conecta un generador de pulsos, para modelar la funcionalidad de los *IHCs*. En biología, la frecuencia de impulsos eléctricos del nervio auditivo está limitada a 100 pulsos por segundo, pero hay que tener en cuenta que de cada *IHC* sale más de una fibra nerviosa y que más de un *IHC* responde a un rango de frecuencias similar. En nuestro sistema, solo se han considerado 21 filtros para todo el rango de frecuencias y sólo un generador de pulsos por cada filtro (banda), Figura 75.

Por tanto, cada generador de pulsos está asociado a una banda del banco de filtros, de manera que cada generador actuará en función de la actividad de su banda asociada, responde a un único rango de frecuencias.

Estos generadores de pulsos se fundamentan en el *modelo de neurona con sumador* propuesto por Gómez-Rodríguez (Gómez-Rodríguez, Paz, Miro, & Linares-Barranco, 2005). Este modelo de neurona se basa en un registro acumulador que, en cada ciclo de reloj, se le suma el valor *PCM* a su valor. La condición para emitir un pulso será el desbordamiento del registro, es decir, que al realizar la operación de suma el resultado sea menor que el valor *PCM*. Por tanto, el valor *PCM* es usado tanto para la suma en el registro acumulador como para la comparación que determina si hay que emitir un pulso o no. En la Figura 78, se muestra el esquema de la neurona.

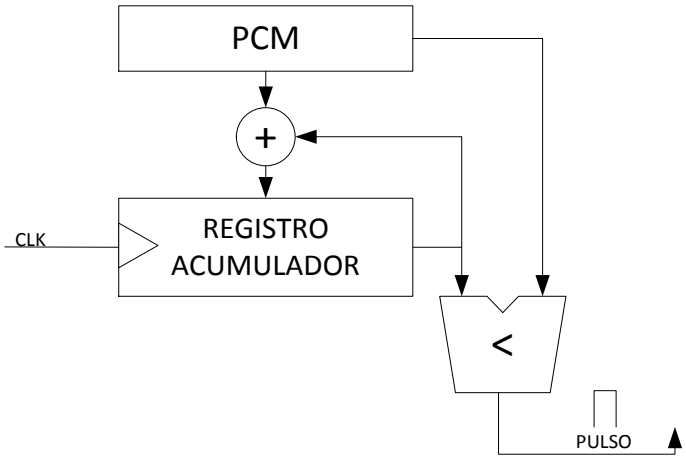


Figura 78. Modelo de neurona con sumador: en cada ciclo de reloj se suma al registro acumulador el valor *PCM*; se emite un pulso si el valor del acumulador es menor que el valor *PCM*.

La salida de este modelo de neurona, una secuencia de pulsos, es proporcional a la amplitud de la señal de audio *PCM* filtrada. Por lo tanto, la frecuencia de los pulsos

de salida cambiará a medida que va cambiando la señal *PCM* filtrada. Gracias a la diferencia de velocidad entre el generador y la señal *PCM* filtrada (del orden de 10^3), se garantiza que por cada valor *PCM* se generarán suficientes pulsos que permitirán a las siguientes etapas procesar esta información. No obstante, dado que la frecuencia de muestreo es 48 kHz y la señal *PCM* varía de acuerdo con esta frecuencia, existen valores por debajo de un determinado umbral para los que el generador no es capaz de emitir pulsos (proporcionales a esos valores), debido a que el acumulador no se desborda en un periodo de la señal *PCM* ($20{,}833 \cdot 10^{-6}$ s) puesto que los valores son muy pequeños. Teniendo en cuenta que el reloj del sistema es de 100 MHz y el registro acumulador es de 18 bits, este umbral es 125,83, según la siguiente fórmula.

$$umbral = \frac{f_m \times 2^n}{f_{CLK}} \tag{6.1}$$

donde:

f_m es la frecuencia de muestreo, 48 kHz;

n es el número de bits del registro acumulador, 18 bits;

f_{CLK} es la frecuencia de reloj del sistema, 100MHz.

Estos generadores de pulsos consideran que el valor de la señal *PCM* de entrada es válido si dicho valor supera el umbral calculado. En la implementación del sistema se ha considerado que todos los generadores tienen el mismo umbral, el valor 127 (valor entero más próximo al calculado a partir de la ecuación 6.1). Este parámetro se puede configurar de forma independiente para cada generador, permitiendo un aumento o disminución del número de pulsos a la salida de cada banda[51]. En la Figura 79.b y Figura 80.b, se observa cómo el número de pulsos aumenta al disminuir a la mitad el valor del umbral de los generadores de pulsos.

[51] El número de eventos también depende del nivel de volumen.

6.1.3. Arbitrador-codificador de eventos *AER*

Para realizar la interconexión de las diferentes etapas del sistema se utiliza un módulo *AERArbitrador*, que permite multiplexar en un único bus *AER* todos los eventos *AER* de salida.

En esta primera etapa del sistema, se va a transmitir a través de un mismo bus *AER* todos los pulsos generados por cada banda. El arbitrador de eventos *AER* codifica la dirección de cada banda (con los valores de 0 a 20) y envía el valor de la banda activa, como un evento *AER*, a través de un bus multiplexado en el tiempo usando el protocolo *AER*.

De esta forma, a la salida de la cóclea tendremos una secuencia de eventos *AER* correspondientes a las diferentes bandas que se han ido disparando a partir del procesado de la señal de entrada. Estos eventos *AER* tienen una dirección perteneciente al rango de 0 a 20, para distinguir cada una de las 21 bandas que componen nuestro banco de filtros.

6.1.4. Respuesta de la cóclea digital pulsante

Este modelo de cóclea digital pulsante ha sido testado usando la herramienta matemática MATLAB y el hardware neuromórfico *USBAERmini2* (descrito en el apartado 2.3.3 "Uso del AER en este trabajo"). En la Figura 79 se observa la salida de la cóclea como respuesta a un barrido en frecuencias en el rango 200Hz – 12000Hz. En la gráfica se muestra el número máximo de eventos emitidos en cada banda del banco de filtro para cada una de las señales senos emitidas. En esta gráfica se dibuja la banda que mayor número de veces dispara para cada una de las señales seno recibidas. Se observa claramente como la banda que más dispara ante un determinado seno, es la banda cuya frecuencia central coincide con la frecuencia de la señal seno.

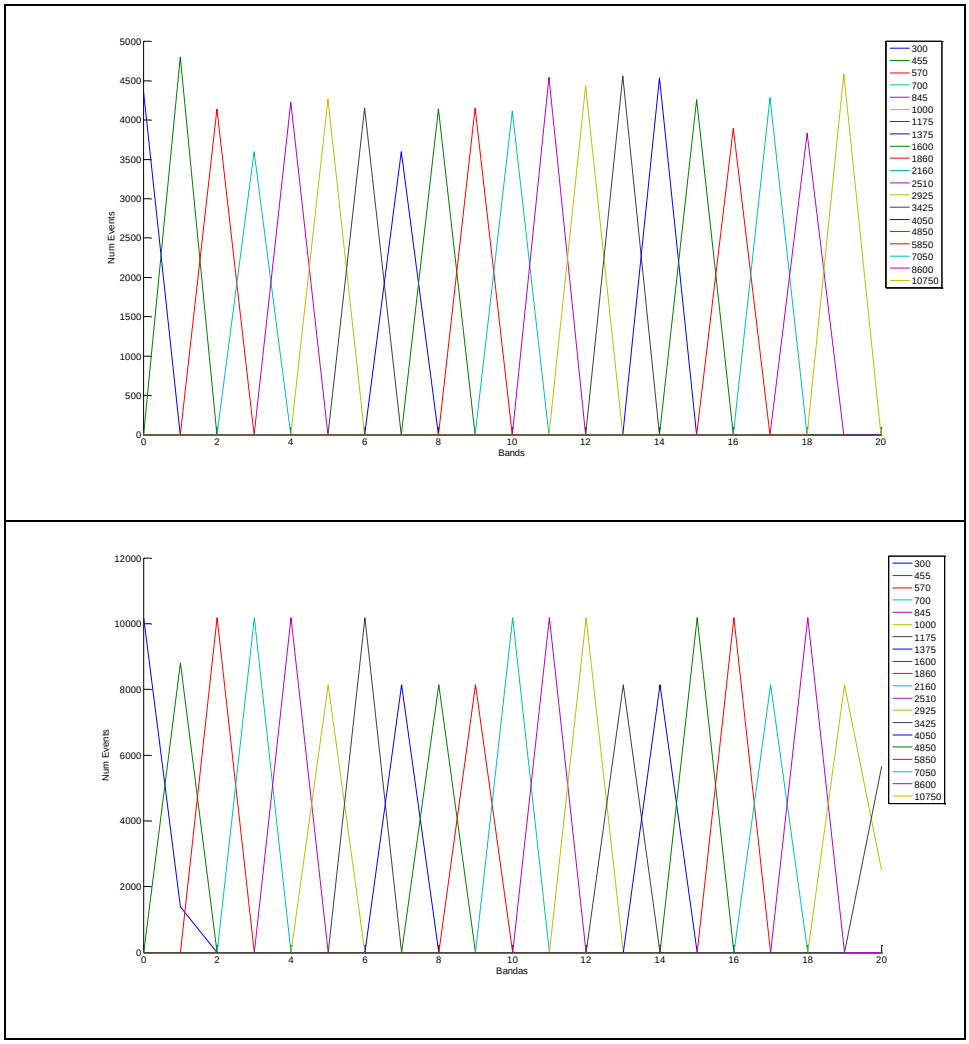


Figura 79. Histograma. Salida del *Script MaxBand.m*[52]. Se ha utilizado dos valores para los umbrales de los generadores de pulsos: a) Umbral: 0x000ff. b) Umbral: 0x0007f. En la leyenda se muestra la frecuencia de las señales seno generadas también en Matlab.

[52] Los diferentes scripts utilizados en este trabajo aparecen al final del documento, en un anexo.

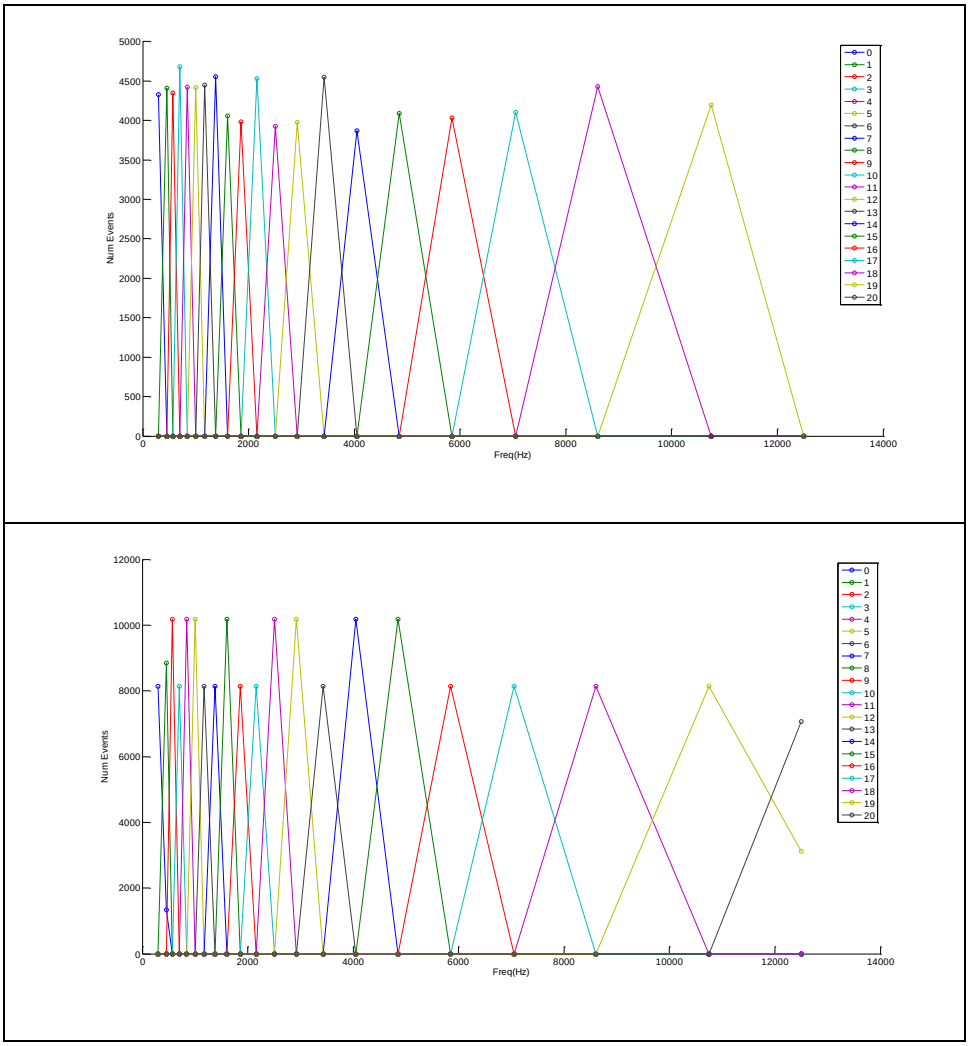


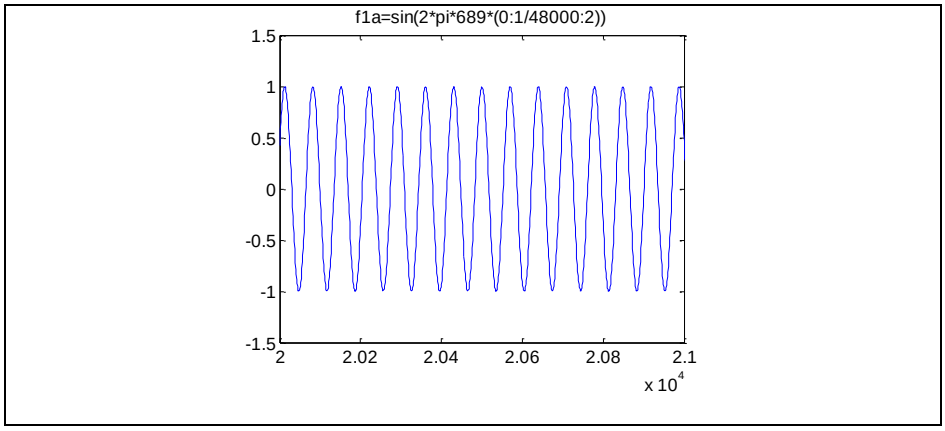
Figura 80. Respuesta de las bandas del banco del filtro a un barrido en frecuencias desde 200 Hz a 20 kHz. Salida del *Script AllBandAllF.m*. Se ha utiliza dos valores para los umbrales de los generadores de pulsos: a) Umbral: 0x000ff. b) Umbral: 0x0007f. En la leyenda se muestra el número de banda.

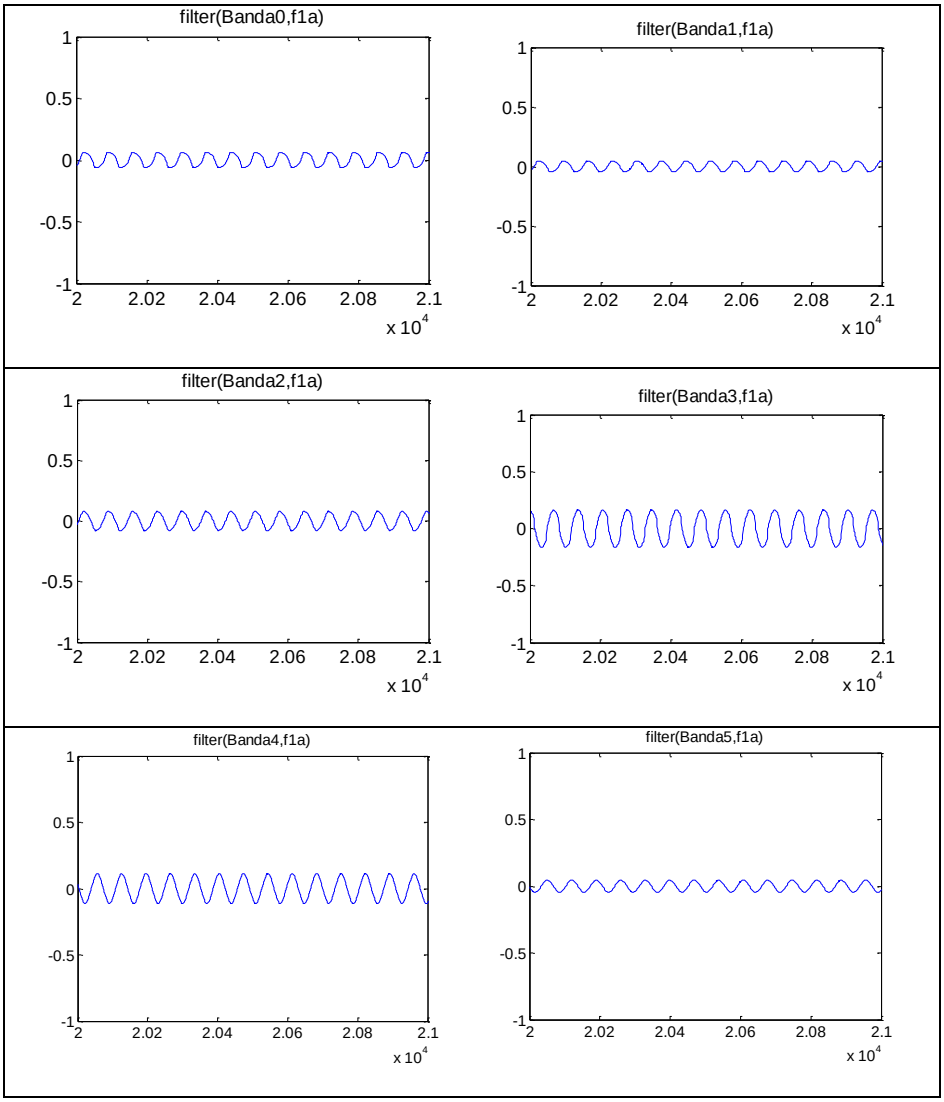
Y en la Figura 80 se muestra la respuesta de todas las bandas del banco de filtro ante un seno de una determinada frecuencia. También se hace un barrido en frecuencias

en el mismo rango 200Hz – 12000Hz, pero en este caso se analiza la respuesta de una banda en concreto ante todas las señales seno.

En las gráficas se observa que el valor máximo de eventos no es igual en todas las bandas. Esto es debido a que los generadores de pulsos no hacen una distribución homogénea de los pulsos, sino que depende del valor *PCM* de cada momento.

Para ilustrar el funcionamiento de la cóclea digital pulsante, en la Figura 81 se muestra la respuesta de la cóclea ante una señal seno de frecuencia 689 Hz (primera fila de la figura). Desde la segunda a la quinta fila se muestra la señal filtrada por los 8 primeros filtros del banco del filtro. En la sexta fila, el cocleograma muestra la salida en forma de tren de pulsos. En la séptima fila se representa el número de eventos que dispara cada una de las 21 bandas de la cóclea digital. En estas dos últimas figuras se observa como la banda 3, cuya frecuencia central es 700 Hz (Tabla 12), es la que más eventos dispara, es la banda más activa.




filter(Banda0,f1a)
filter(Banda1,f1a)
filter(Banda2,f1a)
filter(Banda3,f1a)
filter(Banda4,f1a)
filter(Banda5,f1a)
1
0.5
0
-0.5
-1
2
2.02
2.04
2.06
2.08
2.1
x 10^4

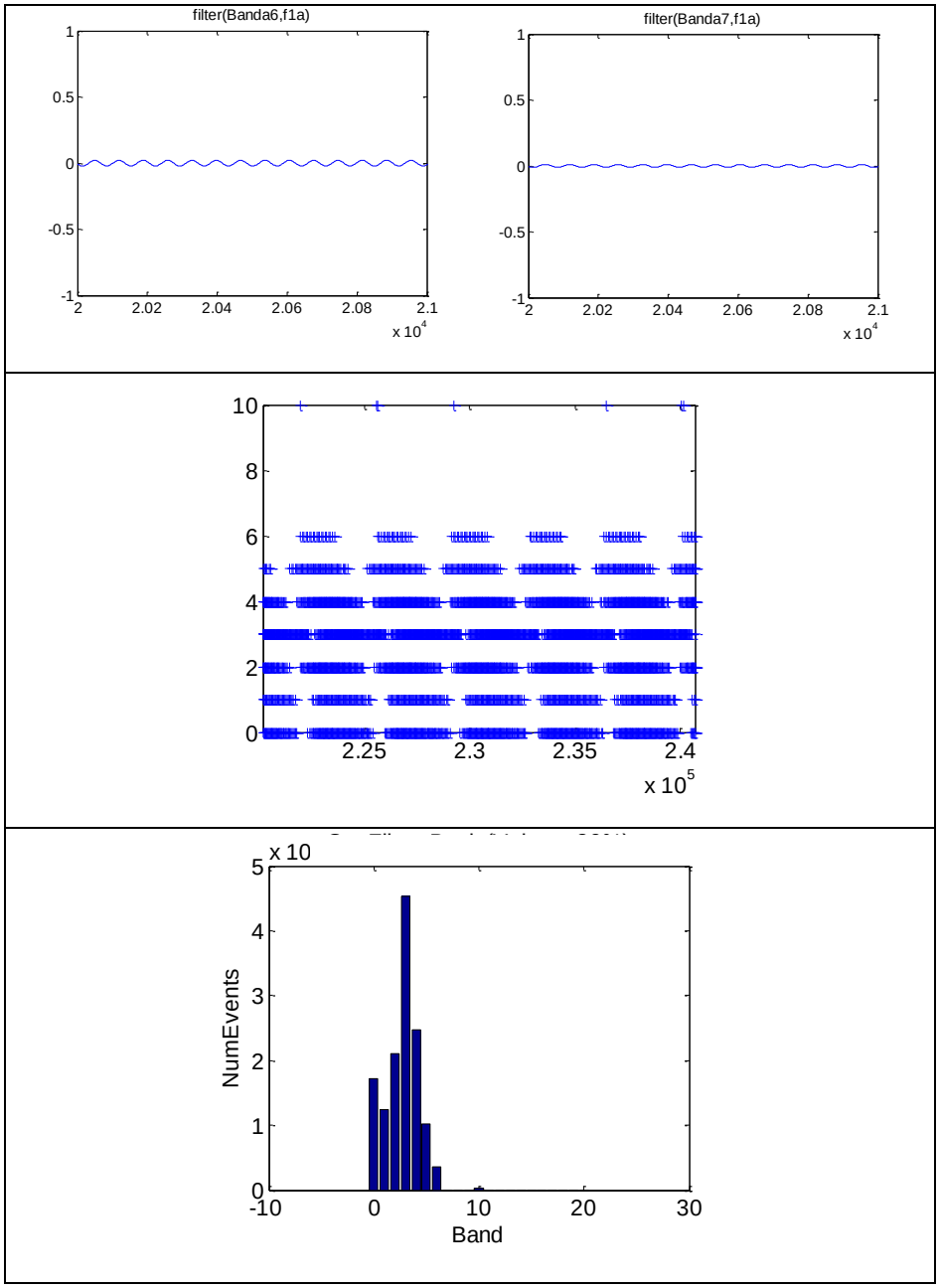


Figura 81. Respuesta de la cóclea digital pulsante ante una señal seno de frecuencia 689Hz.

6.1.5. Recursos hardware

En la siguiente Tabla 13 se resume los recursos de la *FPGA Cyclone IV E* utilizados para la implementación del modelo de cóclea digital pulsante propuesto en este trabajo. La primera columna de la tabla describe los recursos usados en la implementación de un filtro paso banda; la segunda columna los del banco de filtros formado por 21 filtros paso banda y la tercera columna los de la cóclea digital pulsante formada por el banco de filtros, los generadores de pulsos y el arbitrador *AER*.

Tabla 13. Recursos usados en la *FPGA Cyclone IV E* para la implementación de la cóclea digital pulsante.

Filtro paso banda	**21 filtros paso banda**	**21 bandas (filtro paso banda + generador de pulso)**
Total de elementos lógicos: 521 / 114,480 (< 1 %) *LUTs*: 507 / 114,480 (< 1 %) Registros lógicos: 104 / 114,480 (< 1 %)	Total de elementos lógicos: 12,924 / 114,480 (11 %) *LUTs*: 12,852 / 114,480 (11 %) Registros lógicos: 3,098 / 114,480 (3 %)	Total de elementos lógicos: 12,978 / 114,480 (11 %) *LUTs*: 12,906 / 114,480 (11 %) Registros lógicos: 3,099 / 114,480 (3 %)
Multiplicadores de 9-bits: 10 / 532 (2 %)	Multiplicadores de 9-bits: 210 / 532 (39 %)	Multiplicadores de 9-bits: 210 / 532 (39 %)

Capítulo 7

Sistema de reconocimiento pulsante

En este capítulo se describen los componentes del sistema neuromórfico que van a permitir identificar una secuencia de fonemas vocálicos en tiempo real, presentes en la señal de voz. Según se describe en el capítulo 3, la posición del tracto vocal para las diferentes vocales varía afectando principalmente a la frecuencia central de sus formantes. De manera que se ha elegido como método de clasificación de los fonemas vocálicos la posición de sus dos primeros formantes. El sistema implementado es capaz de identificar las cinco vocales de la lengua española a partir del reconocimiento de las dos primeras frecuencias formantes presentes en dichas vocales. Esta información se obtiene a partir de la salida de la primera etapa del sistema, la cóclea digital pulsante, descrita en el capítulo 6.

Para realizar este proceso, se requiere una primera fase, denominada *vowels recognition,* donde se identifica un único fonema vocálico y una segunda fase, llamada *words recognition,* donde se concatenarán los fonemas reconocidos en la anterior fase para identificar la secuencia de fonemas vocálicos consecutivos, Figura 82.

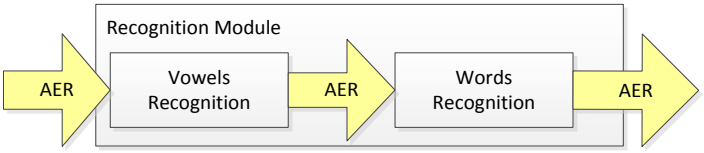


Figura 82. Módulo de reconocimiento, formado por el módulo de identificación de fonemas vocálicos y el módulo de reconocimiento de palabras[53].

La arquitectura de este sistema se basa fundamentalmente en tres modelos de neuronas: neurona de reconocimiento, neurona ganadora y neurona de retraso, que se describen a continuación.

Aunque el sistema propuesto en este trabajo describe una arquitectura que permite reconocer una secuencia de dos fonemas vocálicos, el diseño modular que se presenta, basado en estos tres elementos básicos, ofrece una gran flexibilidad para el desarrollo de otros sistemas de reconocimiento, usando estos mismos elementos básicos. Por ejemplo, la identificación de fonemas de otros idiomas, identificación de fonemas no vocálicos, el reconocimiento de palabras de cualquier longitud (a partir de la concatenación de diferentes fonemas), etc.

7.1. Neurona de reconocimiento, *RNeuron*

La neurona de reconocimiento, *RNeuron*, constituye la base de la etapa de reconocimiento ya que es la encargada de identificar una determinada categoría o patrón. Una colección de estas neuronas va a ser utilizada tanto en la etapa de

[53] En nuestra implementación, el módulo de reconocimiento de palabras identifica una secuencia de dos fonemas vocálicos consecutivos.

reconocimiento de fonemas vocálicos como en la etapa de reconocimiento de palabras. Se pueden añadir al sistema tantas *RNeuron* como categorías se quiera reconocer.

El comportamiento de esta neurona se fundamenta en dos conceptos básicos relacionados con el reconocimiento del habla: la comparación de plantillas o patrones y las redes de neuronas artificiales basadas en el perceptrón y aplicadas al reconocimiento de patrones. Ambos conceptos han sido explicados en el capítulo 5 de esta tesis.

7.1.1. Interfaz del módulo *RNeuron*

La *RNeuron* tiene dos puertos de comunicación *AER*[54], uno de entrada y otro de salida. Por el puerto de entrada recibe una secuencia de eventos *AER* que analizará para determinar si existe o no un patrón previamente definido. Y por el puerto de salida enviará un evento *AER*, cada vez que se detecte un patrón.

7.1.2. Descripción funcional del módulo *RNeuron*

Para entender el funcionamiento de la *RNeuron* es necesario conocer los parámetros que permiten configurar su comportamiento. A continuación se describen cada uno de ellos:

- **Patrón o máscara**: es un conjunto de 21 posiciones que contiene la información del patrón que se quiere detectar. Una posición del conjunto tiene el valor '0', si no se espera recibir eventos con un identificador o dirección igual al de dicha posición; o tendrá valor '1', si se espera recibir eventos con un identificador que coincida con su posición. Por ejemplo, el valor "000000000000000001110" define un

[54] Un puerto de comunicación *AER* consta de las líneas de control *REQ* y *ACK*, y las líneas de datos.

patrón que espera recibir una secuencia de eventos *AER*, cuyas direcciones son 1, 2 y 3 (direcciones que coinciden con las posiciones del conjunto que tienen valor '1').

- **Umbral de potencial**: asociado a cada patrón existe un umbral de potencial que determinará cuando se ha detectado el patrón.

- **Potencial**: se corresponde con el potencial interno de la *RNeuron* y define su comportamiento: por debajo del umbral de potencial, se continúan recibiendo y procesando eventos; o alcanzado el umbral, se generará un pulso indicando que se ha detectado el patrón.

- **Umbral de disparo de las entradas**: es un conjunto de 21 posiciones que definen el número de eventos que han de recibirse con cada uno de los 21 valores de direcciones posibles. Permite saber si una entrada (asociada a una dirección) tiene una actividad relevante y que por tanto será tenida en cuenta para detectar el patrón. Este umbral pretende modelar el hecho de que las conexiones entre las neuronas biológicas están ponderadas.

- **Historial de eventos recibidos**: estructura interna que almacena y actualiza, en cada instante, el número de eventos recibidos por cada una de las entradas. Esta información permite conocer si una entrada ha recibido suficientes eventos y así poder decidir si se incrementa o disminuye su potencial.

- **Tiempo de reinicio o reseteo**: una vez alcanzado este tiempo, si no se ha detectado ningún patrón, se borrará toda la información almacenada, tanto el potencial interno de la *RNeuron* como el historial de eventos recibidos. Este tiempo de reinicio trata de modelar la función de olvido de las neuronas biológicas.

El funcionamiento de la *RNeuron* se describe en Figura 83. Tras el *reset*, la *RNeuron* se encuentra en el estado *INICIO*, donde el historial de eventos recibidos se borra y el potencial interno se coloca en la situación de reposo; se pasa al estado *En espera* donde permanece hasta que recibe un evento.

Cuando se recibe un evento con una determinada dirección, se pasa al estado *actualiza cuenta de eventos*. En este estado, se actualiza el historial de eventos recibidos; es decir, se incrementa en uno el valor de la posición de memoria correspondiente a la dirección del evento recibido. En este estado, también se comprueba si la cantidad de eventos recibidos con esa dirección supera o no el umbral de disparo para esa entrada. Si no se supera el umbral se vuelve al estado *En espera*.

Si el umbral es superado, y por lo tanto se han recibido suficientes eventos de esa entrada, se pasa al estado *comprueba localización en el patrón*. Si la posición está marcada con un '0' en el patrón, significa que no debería haber eventos en esa localización. En este caso se pasa al estado *En espera*. En el caso contrario, la posición del patrón está marcada con un '1', significa que en esa localización debe haber eventos y por tanto se incrementará en 1 el potencial. Una vez incrementado el potencial, si es mayor que el Umbral de potencial significará que el patrón se ha detectado y por tanto se emite un evento para indicarlo y se pasa al estado *INICIO*. En caso contrario, si no se alcanza el Umbral de potencial, se vuelve al estado *En espera*. En este estado, también se comprueba el tiempo de reinicio, de tal modo que si se supera este tiempo y todavía no se ha detectado ningún patrón, el sistema se resetea al volver de nuevo al estado *INICIO*.

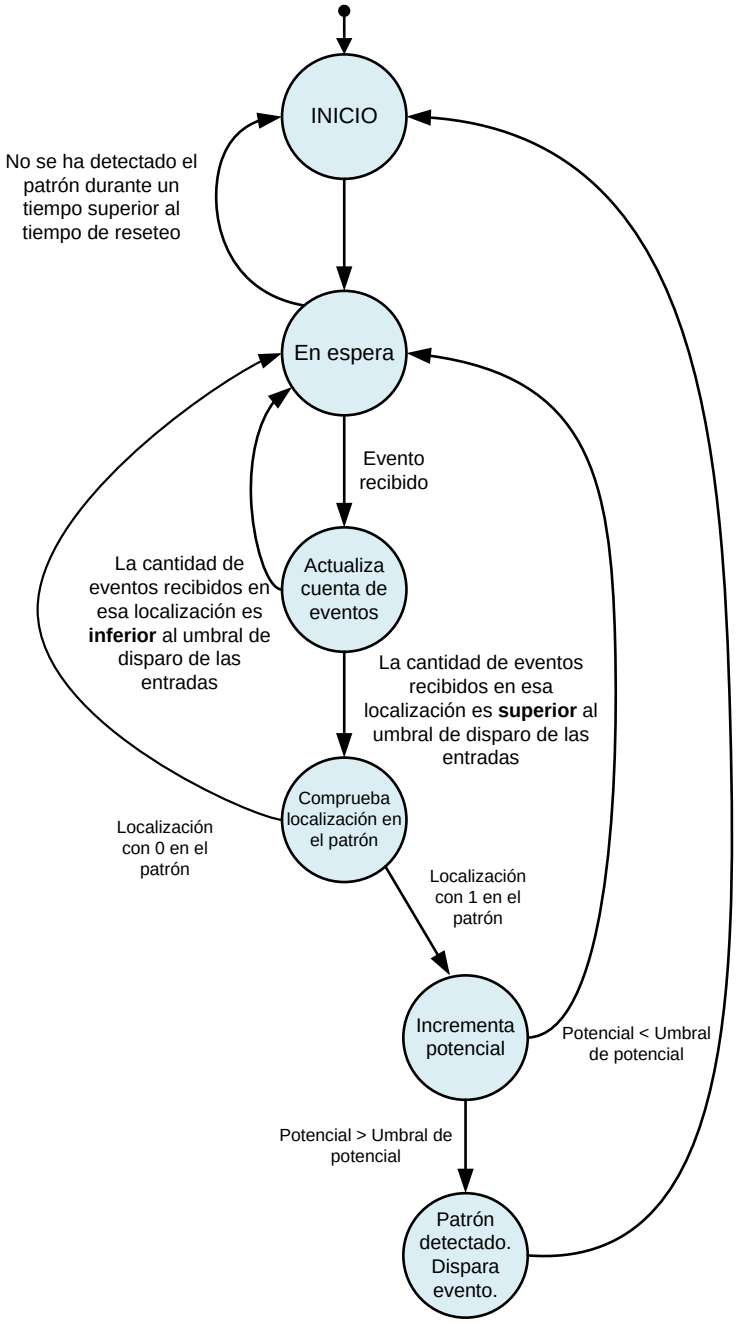


Figura 83. Máquina de estados de *RNeuron*.

En la siguiente Tabla 14 se resume los recursos de la *FPGA Cyclone IV E* utilizados para la implementación del modelo de neurona de reconocimiento, *RNeuron*.

Tabla 14. Recursos usados en la *FPGA Cyclone IV E* para la implementación de neurona *RNeuron*.

Elementos lógicos		**Memoria (bits)**
Registros	***LUTs***	
121/ 114480 (< 1 %)	42/ 114480 (< 1 %)	336/3981312 (< 1 %)

7.2. Neurona ganadora, *WTANeuron*

Las redes competitivas son utilizadas para detectar automáticamente grupos o categorías, dentro de los datos disponibles. Una red neuronal competitiva está formada por una capa de neuronas en la que todas reciben la misma entrada y sólo una neurona de salida está activa en cada momento. Las neuronas de salida compiten entre sí para ser la que se activa como respuesta a una entrada determinada. Este modelo de red competitiva se ha utilizado en nuestro sistema para mejorar el rendimiento del proceso de reconocimiento tanto de fonemas como de palabras. Para ello, se ha incluido en la fase final de ambas etapas de identificación una red de neuronas *WTANeuron*, una por cada categoría que se quiere identificar.

En la definición de las redes competitivas se distinguen dos tipos de competición: la competición dura, en la que sólo una neurona permanece activa y todas las demás están inactivas; y la competición blanda, en la que aunque existe un vencedor claro, la neurona más activa, sus vecinos también aparecen activos, aunque en un menor porcentaje. En neurobiología, la competición blanda está ampliamente extendida y ha sido la inspiración para los modelos desarrollados hasta el momento.

Una red de neuronas competitiva típica consiste en una capa de neuronas en la que todas reciben la misma entrada. La neurona que presenta la mejor salida es declarada vencedora. El concepto de elegir un vencedor a menudo requiere un controlador global que compare cada salida con todas las demás (Figura 84a). Pero, en este caso, se desea construir una red que encuentre la mejor salida sin necesidad de control global. Este sistema se denomina *winner-take-all*[55]. Una red de este tipo se crea usando alimentación lateral, de modo que entre las neuronas de las capas de salida existen conexiones laterales, que va a permitir que cada una de ellas tenga influencia sobre sus vecinas, al producir una excitación en las neuronas más próximas y una inhibición en las más alejadas (Figura 84b).

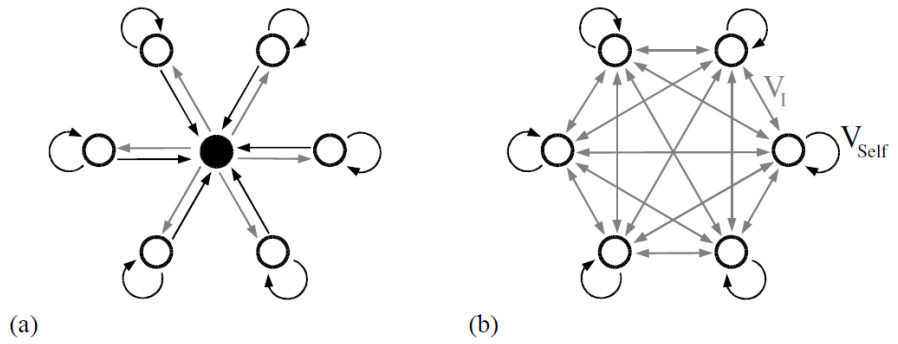


Figura 84. Conectividad en una red neuronal competitiva. (a) Control global. (b) Conexiones laterales.

Oster propone una red *winner-take-all* que implementa una competición dura, solo una neurona es la neurona activa en cada momento (Oster & Liu, 2005). Es una red de neuronas *integrate-and-fire* que reciben como entrada un tren de pulsos a una frecuencia constante, y la neurona ganadora será aquella que recibe los pulsos a una

[55] *Winner-take-all*, cuya traducción es el ganador lo consigue todo. Término utilizado en las redes neuronales competitivas.

mayor frecuencia después de recibir un número determinado de pulsos de entrada, previamente configurado. Cada neurona recibe una entrada externa excitadora y las correspondientes entradas inhibidoras procedentes de todas las demás neuronas de la red (Figura 84b). Cada una de estas entradas provocan un cambio en el valor del potencial interno de la neurona y la neurona emitirá un pulso cuando su potencial interno supere un determinado valor de umbral. En ese momento, su potencial interno se resetea a su valor de auto-excitación previamente configurado. Tan pronto como esta neurona dispara, ninguna otra neurona podrá emitir un pulso, ya que va a recibir un pulso inhibidor desde la neurona ganadora.

El módulo de neurona *WTANeuron*, descrito a continuación, es la base de una red de neuronas *winner-take-all*, que implementa una competición blanda, más próximo al comportamiento de los sistemas biológicos. La gran diferencia respecto al modelo propuesto por Oster se debe a que cuando la neurona ganadora emite un evento, no envía pulsos inhibidores a las otras neuronas que constituyen la red. De esta forma pueden existir otras neuronas activas en la red, aunque en menor porcentaje.

7.2.1. Interfaz del módulo *WTANeuron*

La neurona *WTANeuron* tiene dos puertos de comunicación *AER*, uno de entrada y otro de salida. Por el puerto de entrada recibe una secuencia de eventos *AER* que analizará para determinar si ella es la neurona ganadora. En caso afirmativo, envía esta información a través del puerto de salida a modo de evento *AER*.

7.2.2. Descripción funcional del módulo *WTANeuron*

A continuación se describen los parámetros que permiten configurar el comportamiento de la *WTANeuron*:

- **Identificador o máscara**: es un conjunto que contiene la identificación de la neurona.

- **Umbral de potencial**: asociado a cada neurona existe un umbral de potencial, que una vez alcanzado se considerará que ella es la neurona ganadora.

- **Potencial**: se corresponde con el potencial interno de la *WTANeuron* y define su comportamiento: por debajo del umbral de potencial, se continúan recibiendo y procesando eventos; o alcanzado el umbral, se generará un pulso indicando que es la neurona ganadora.

- **Pesos de las entradas**: valor positivo (excitador) o negativo (inhibidor) asociado a cada entrada de la neurona, que modifican el valor del potencial interno de la neurona. Modelan el carácter excitador o inhibidor de las sinapsis de entrada de una neurona biológica.

- **Umbral de disparo de las entradas**: es un conjunto con el mismo número de posiciones que el identificador o máscara. Para cada posición define el número de eventos que han de recibirse. Permite saber si una entrada (asociada a una dirección) tiene una actividad relevante y que por tanto será tenida en cuenta para elegir la neurona ganadora. Este umbral pretende modelar el hecho de que las conexiones entre las neuronas biológicas están ponderadas.

- **Historial de eventos recibidos**: estructura de datos que almacena y actualiza, en cada instante, el número de eventos recibidos por cada una de las entradas. Esta información permite conocer si una entrada ha recibido suficientes eventos y así poder decidir si se incrementa o disminuye su potencial.

- **Tiempo de reinicio o reseteo**: una vez alcanzado este tiempo, si la neurona no ha sido elegida como ganadora, se borrará toda la información almacenada, tanto el potencial interno de la *WTANeuron* como el historial de eventos recibidos. Este tiempo de reinicio trata de modelar la función de olvido de las neuronas biológicas.

El funcionamiento de la *WTANeuron* se describe en Figura 85. Tras el *reset*, la *WTANeuron* se encuentra en el estado *INICIO*, donde el historial de eventos recibidos se borra y el potencial interno se coloca en la situación de reposo; se pasa al estado *En espera* donde permanece hasta que recibe un evento.

Cuando se recibe un evento con una determinada dirección, se pasa al estado *actualiza cuenta de eventos*. En este estado, se actualiza el historial de eventos recibidos; es decir, se incrementa en uno el valor de la posición de memoria correspondiente a la dirección del evento recibido. En este estado, también se comprueba si la cantidad de eventos recibidos con esa dirección supera o no el umbral de disparo para esa entrada. Si no se supera el umbral se vuelve al estado *En espera*.

Si el umbral es superado, y por lo tanto se han recibido suficientes eventos de esa entrada, se pasa al estado *actualiza y comprueba su potencial*. Si la posición está marcada con un '0' en el identificador o máscara, significa que no debería haber eventos en esa localización. En este caso, se disminuye el potencial con el peso inhibidor y se pasa al estado *En espera*. En el caso contrario, la posición del identificador o máscara está marcada con un '1', significa que en esa localización debe haber eventos y por tanto se incrementará el potencial con el valor del peso excitador. Una vez incrementado el potencial, si es mayor que el Umbral de potencial significará que es la neurona ganadora y por tanto se emite un evento para indicarlo y se pasa al estado *INICIO*. En caso contrario, si no se alcanza el Umbral de potencial, se vuelve al estado *En espera*. En este estado, también se comprueba el tiempo de

reinicio, de tal modo que si se supera este tiempo y todavía no se ha elegido como neurona ganadora, el sistema se resetea al volver de nuevo al estado *INICIO*.

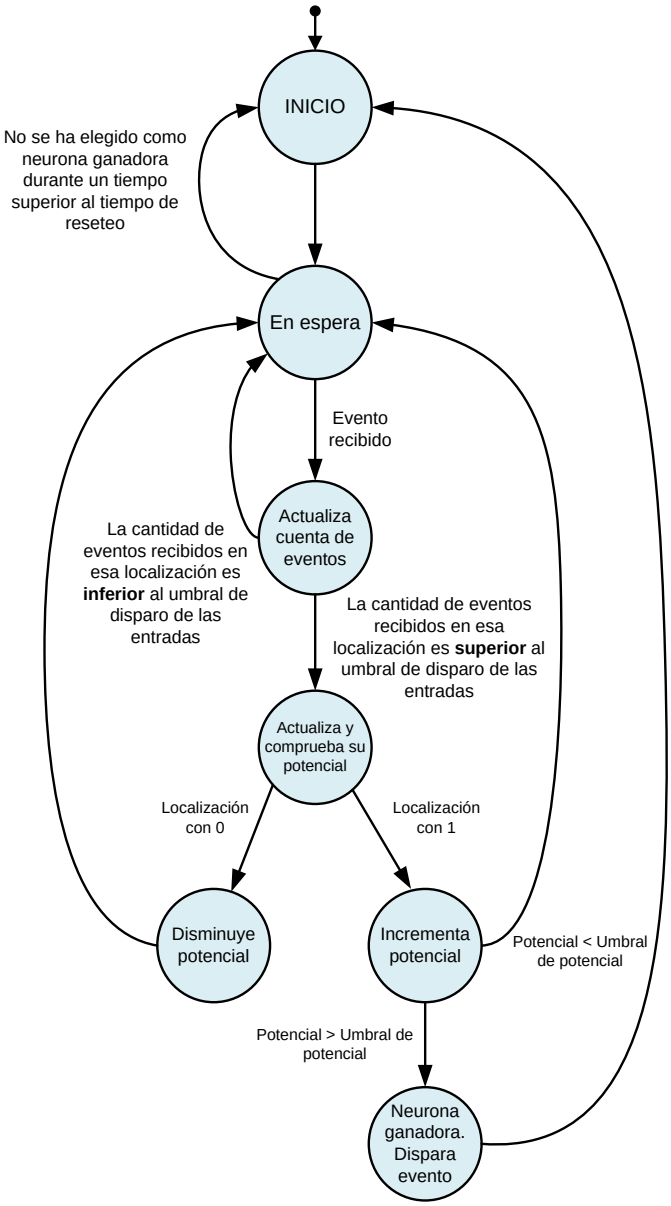


Figura 85. Máquina de estados de *WTANeuron*.

El comportamiento de la neurona *WTANeuron* es muy parecido al de la neurona *RNeuron*, sólo se diferencian en como su potencial interno es modificado cuando reciben un evento con una determinada dirección. En la *RNeuron*, si la dirección del evento recibido coincide con una dirección de su máscara o patrón su potencial interno se incrementa en 1; en caso contrario el valor de su potencial interno no se ve alterado. Sin embargo, el potencial interno de la *WTANeuron* será modificado cada vez que recibe un evento. Si la dirección del evento coincide con una posición de su patrón que está marcado con un '1', su potencial interno se incrementará en la cantidad especificada por el peso de una entrada excitadora; si la dirección del evento coincide con una posición de su patrón que está marcado con un '0', su potencial interno disminuirá en la cantidad especificada por el peso de una entrada inhibidora.

En la siguiente Tabla 15 se resume los recursos de la *FPGA Cyclone IV E* utilizados para la implementación del modelo de neurona ganadora, *WTANeuron*.

Tabla 15. Recursos usados en la *FPGA Cyclone IV E* para la implementación de neurona *WTANeuron*.

Elementos lógicos		**Memoria (bits)**
Registros	***LUTs***	
104/ 114480 (< 1 %)	63/ 114480 (< 1 %)	336/3981312 (< 1 %)

7.3. Neurona de retraso, *delayNeuron*

El sistema, en la etapa *words recognition,* es capaz de reconocer una palabra a partir de la identificación de una secuencia de dos fonemas vocálicos. Por ejemplo, reconoce la palabra /CITA/ si la etapa anterior, *vowels recognition,* ha identificado los fonemas vocálicos /i/ y /a/. Cuando se pronuncia una palabra, sus fonemas se encuentran distanciados a lo largo del tiempo. Del mismo modo, el tren de eventos *AER*, asociado a cada fonema identificado, también se encuentran distanciados un

intervalo de tiempo. Por tanto, el sistema para identificar una palabra va a retrasar el tren de eventos *AER* asociado al primer fonema identificado, con la finalidad de emparejar en el tiempo la secuencia de eventos AER correspondiente a los dos fonemas a partir de los cuales se identificará una palabra. Así, a la entrada de la etapa *words recognition* llegarán al mismo tiempo eventos *AER* de los dos fonemas previamente identificados.

El retraso de eventos *AER* va a ser realizado por la neurona *delayNeuron*.

7.3.1. Interfaz del módulo *delayNeuron*

La *delayNeuron* tiene dos puertos de comunicación *AER*, uno de entrada y otro de salida. Este módulo emite a través del puerto de salida la información que recibe a través del puerto de entrada después de un tiempo de retraso configurable. En este sistema, el tiempo de retraso es 200 ms, valor obtenido experimentalmente.

7.3.2. Descripción funcional del módulo *delayNeuron*

Este módulo se comporta como una *cola*. Va a permitir *encolar* eventos *AER*, que llegan por el puerto de entrada, y *desencolar* eventos *AER* cada vez que se alcanza un tiempo predefinido, consiguiéndose así que los eventos salgan por el puerto de salida con un retraso.

En este módulo se ha definido dos parámetros que van a permitir configurar su comportamiento:

- **Tamaño:** define el número máximo y fijo de elementos de la *cola*. Esta *cola* se ha implementado con una estructura de datos estática, un conjunto de elementos circular. Por lo tanto, requiere definir el número máximo de elementos que se pueden almacenar en ella. En nuestra

implementación se ha elegido el valor de 1023 elementos, valor óptimo según las pruebas realizadas.

- **Tiempo de actualización de índices:** define el tiempo que debe esperar la neurona para *desencolar* un nuevo evento *AER* y emitirlo por el puerto de salida como evento retrasado. En nuestra implementación se ha optado por el valor de 204,8 μs, obtenido experimentalmente.

Los valores elegidos garantizan que no se van a producir pérdidas de eventos, ya que tanto el tamaño de la *cola* como el tiempo de actualización de índices son lo suficientemente grandes para asegurar que un evento *AER* siempre se va a *encolar* en una posición vacía.

Es interesante resaltar que cambiando los parámetros de la neurona *delayNeuron* se va a permitir adaptar el sistema a la velocidad del hablante, haciendo este tiempo de retraso más corto o más largo.

El funcionamiento de la *delayNeuron* se muestra en la Figura 86. Tras el *reset*, la *delayNeuron* se encuentra en el estado *INICIO*, donde se inicializa el contenido y los dos índices de la *cola*: el índice de *entrada* y el índice de *salida*. El índice de *entrada* es una posición anterior al índice de *salida*. Se pasa al estado *En espera* donde permanece hasta que recibe un evento *AER* o hasta que se alcanza el *tiempo de actualización de índices*.

Cuando se recibe un evento se pasa al estado *encola evento*. En este estado, se inserta el elemento en la *cola* en la posición indicada por el índice de *entrada* y se vuelve al estado *En espera*.

En el estado *En espera*, si se alcanza el *tiempo de actualización de índices* se pasa al estado *desencola evento*. En este estado se comprueba si en la *cola* en la posición marcada por *índice de salida* existe un evento. En caso afirmativo se lee el evento y se envía a través del puerto de salida. Además, se borra el contenido de esa posición, se

incrementa el valor del *índice de salida* para que apunte a la siguiente posición y se actualiza el valor del *índice de entrada*. Y se vuelve al estado *En espera*. En caso contrario, si no existe evento en dicha posición, se incrementa el valor del *índice de salida* para que apunte a la siguiente posición y se actualiza el valor del *índice de entrada* y se vuelve al estado *En espera*.

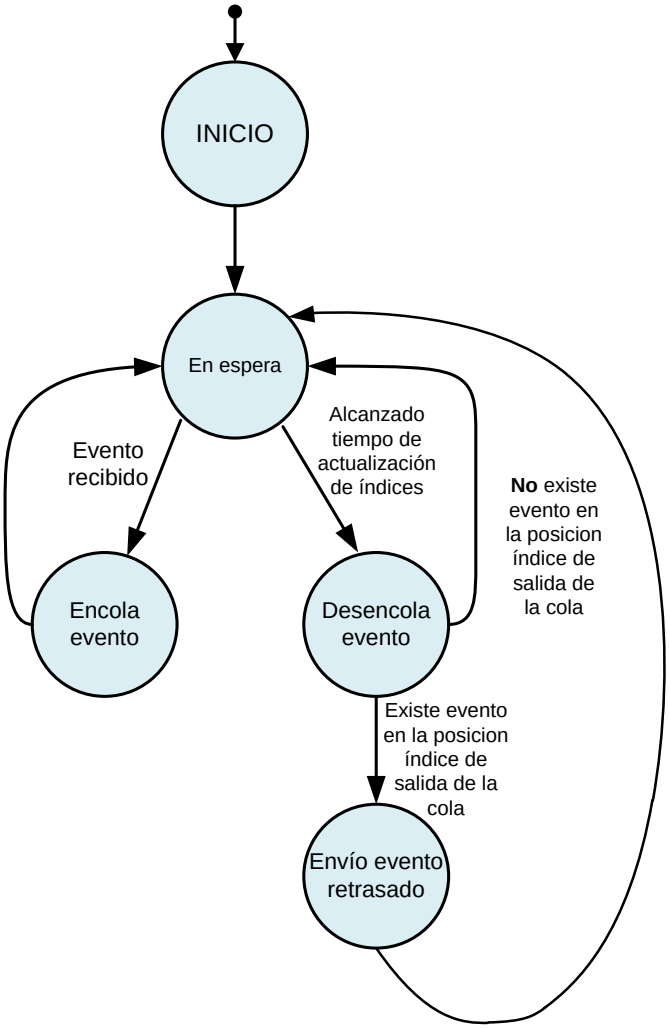


Figura 86. Máquina de estados de *delayNeuron*.

En la siguiente Tabla 16 se resume los recursos de la *FPGA Cyclone IV E* utilizados para la implementación del modelo de neurona de retraso, *delayNeuron*.

Tabla 16. Recursos usados en la *FPGA Cyclone IV E* para la implementación de neurona *delayNeuron*.

Elementos lógicos	
Registros	***LUTs***
3148/ 114480 (3 %)	2788/ 114480 (2 %)

7.4. Sistema de reconocimiento de fonemas

El objetivo de esta etapa *vowels recognition* es la identificación de un fonema vocálico de la lengua española a partir de la salida de la cóclea digital pulsante.

En la Figura 87 se muestra el diagrama de bloques del sistema de reconocimiento de fonemas. En él se distingue una primera fase de identificación del fonema formado por los bloques *VowelNeuronSET* y *VowelWTAsSET*; y una segunda fase, formada fundamentalmente por el bloque *DelayNeuronChain* que implementa una cadena de retraso de los fonemas identificados, para facilitar el reconocimiento de una secuencia de fonemas que se llevará a cabo en la siguiente etapa del sistema de reconocimiento, *words recognition*.

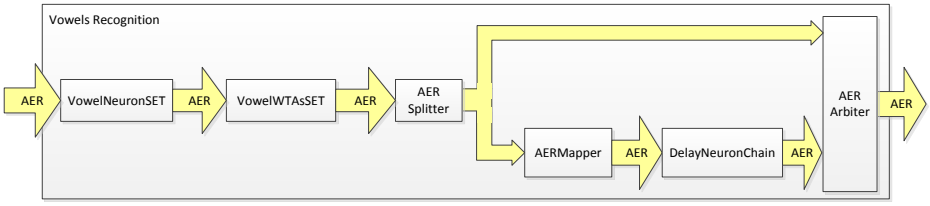


Figura 87. Módulo *VowelsRecognition*.

7.4.1. Bloque VowelNeuronSet

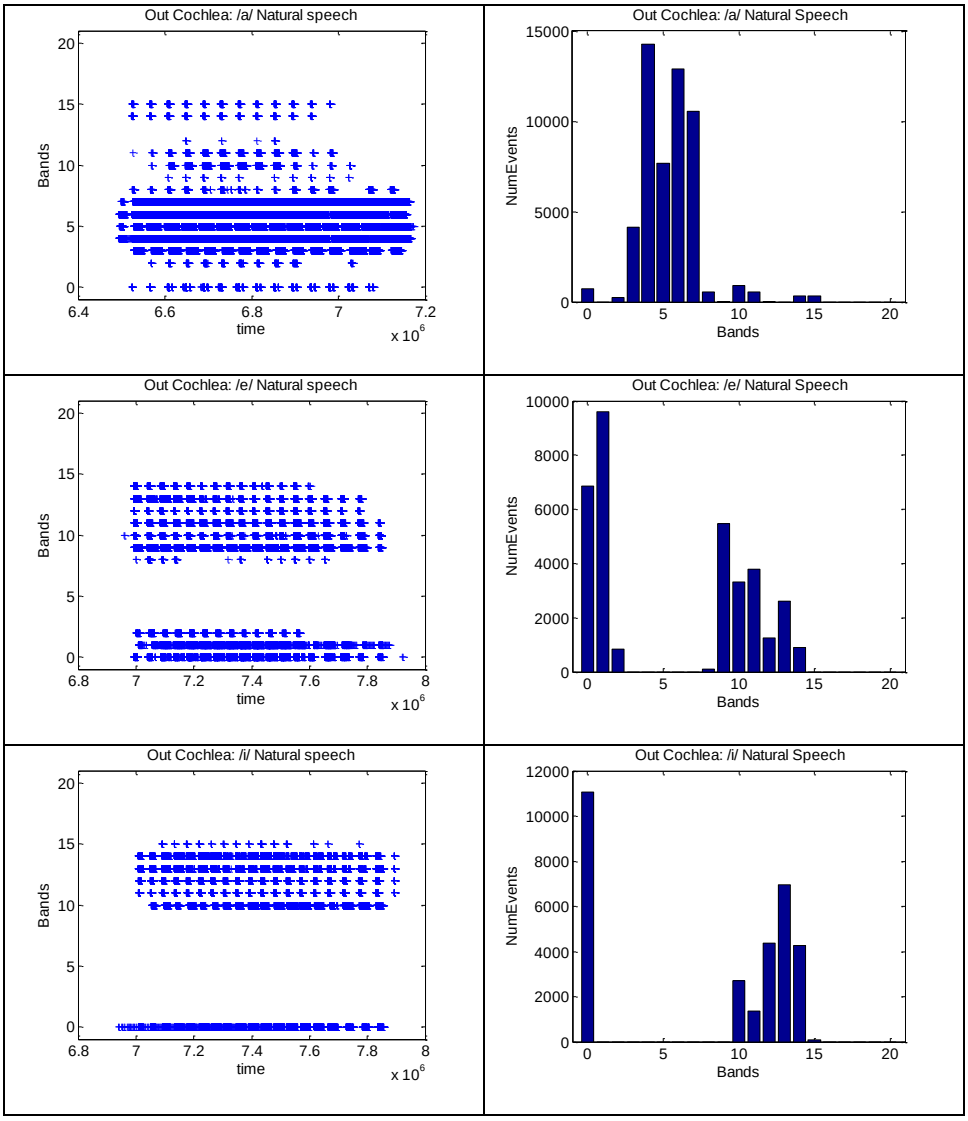
Este bloque está formado por un conjunto de *RNeuron*, y es el encargado de identificar los fonemas vocálicos del castellano.

Cada fonema vocálico del castellano está caracterizado por unos determinados valores del primer y segundo formante (explicado en el capítulo 3). Los valores de estas formantes, F_1 y F_2, pertenecen a una de las 21 bandas del banco de filtros de la cóclea digital pulsante, como se muestra en la siguiente tabla (Tabla 17). Estos valores han sido obtenidos experimentalmente y algunos de ellos coinciden con los encontrados en la bibliografía (Tabla 6 y Tabla 7).

Tabla 17. Primer y segundo formante de los fonemas vocálicos obtenidos experimentalmente para este sistema (dependiente de hablante).

	/i/	**/e/**	**/a/**	**/o/**	**/u/**
F_1	247 (Banda 0)	382 (Banda 1)	789 (Banda 4)	520 (Banda 2)	280 (Banda 0)
F_2	3255 (Banda 13)	1970 (Banda 9)	1200 (Banda 6)	920 (Banda 4)	650 (Banda 3)

Para obtener los valores de las formantes F_1 y F_2 de la tabla anterior, se analizó la salida de las 21 bandas de la cóclea digital pulsante, descrita en el capítulo 6, usando como estímulo el sonido de las 5 vocales. La siguiente Figura 88 representa la salida de la cóclea digital pulsante como respuesta a los 5 fonemas vocálicos. Para cada fonema se muestra el cocleograma (figura izquierda) y el número de eventos por banda (figura derecha). Se observa como las bandas que emiten un mayor número de eventos son las bandas cuyo rango de frecuencias contienen las frecuencias formantes de los fonemas vocálicos. Para el fonema /a/, las bandas 4 y 6 de la cóclea digital pulsante responden con un mayor número de eventos; para el fonema /e/, las bandas 1 y 9; para el fonema /i/, las bandas 0 y 13; para el fonema /o/, las bandas 2 y 4 y para el fonema /u/ las bandas 0 y 3.


Out Cochlea: /a/ Natural speech
Bands
time
x 10^6
Out Cochlea: /a/ Natural Speech
NumEvents
Bands
Out Cochlea: /e/ Natural speech
Bands
time
x 10^6
Out Cochlea: /e/ Natural Speech
NumEvents
Bands
Out Cochlea: /i/ Natural speech
Bands
time
x 10^6
Out Cochlea: /i/ Natural Speech
NumEvents
Bands

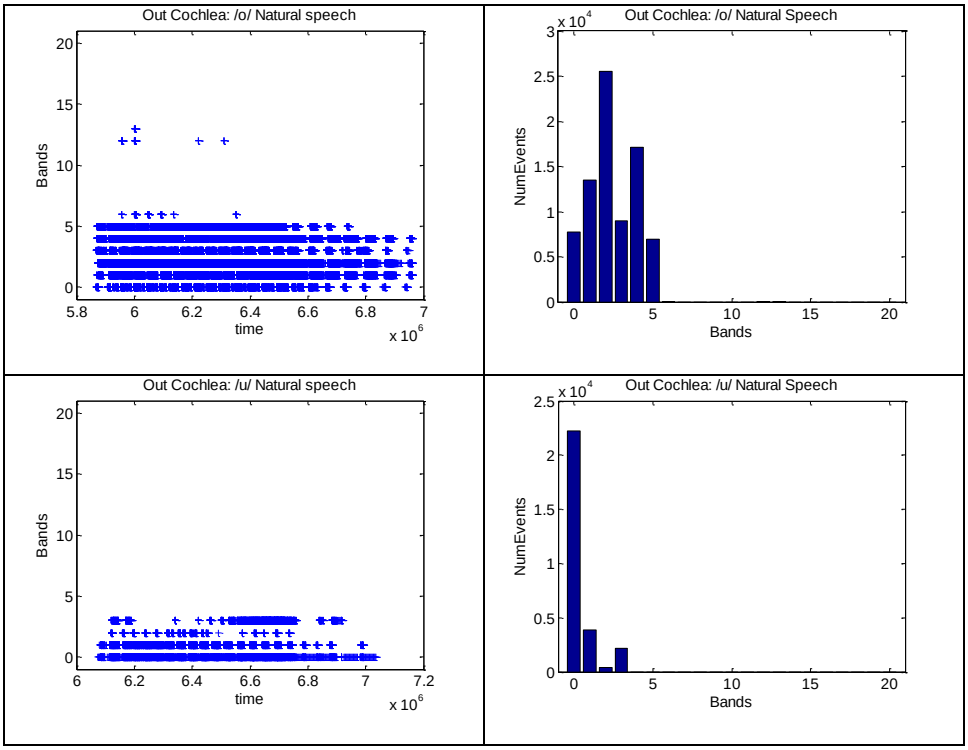


Figura 88. Salida de la cóclea. En estas figuras se representa el número de eventos emitidos por cada banda de la cóclea (bandas 0 a 20), como respuesta a los fonemas /a/, /e/, /i/, /o/ y /u/.

Interfaz del bloque *VowelNeuronSET*

El bloque *VowelNeuronSET* tiene dos puertos de comunicación *AER*, uno de entrada y otro de salida. Este bloque analiza la secuencia de eventos *AER* que recibe a través del puerto de entrada y emite a través del puerto de salida el fonema vocálico reconocido, como una secuencia de eventos *AER*.

Descripción funcional del bloque *VowelNeuronSET*

La secuencia de eventos *AER*, correspondiente a las bandas de la cóclea que se han disparado a partir de la señal de audio filtrada, llega en paralelo a la entrada de 5

neuronas *RNeuron*, una por cada fonema vocálico que se quiere identificar (ver Figura 89).

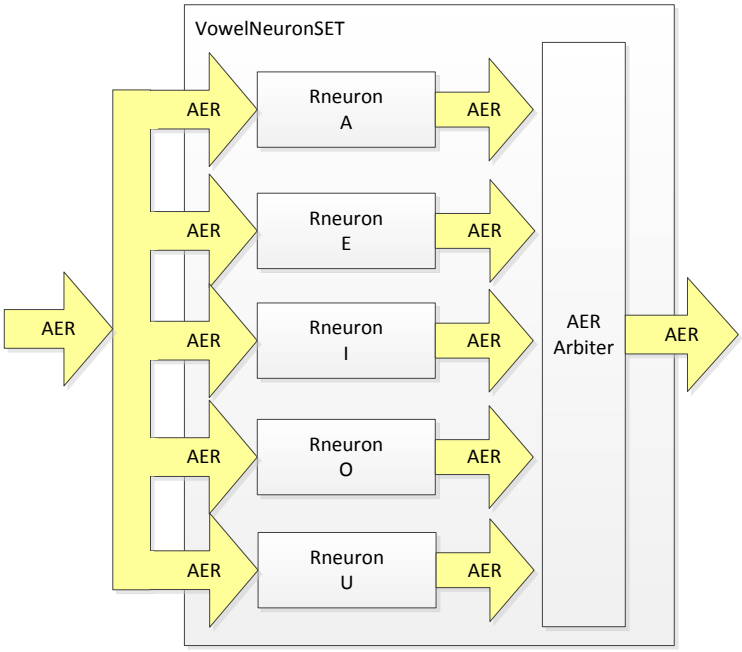


Figura 89. Bloque *VowelNeuronSET*.

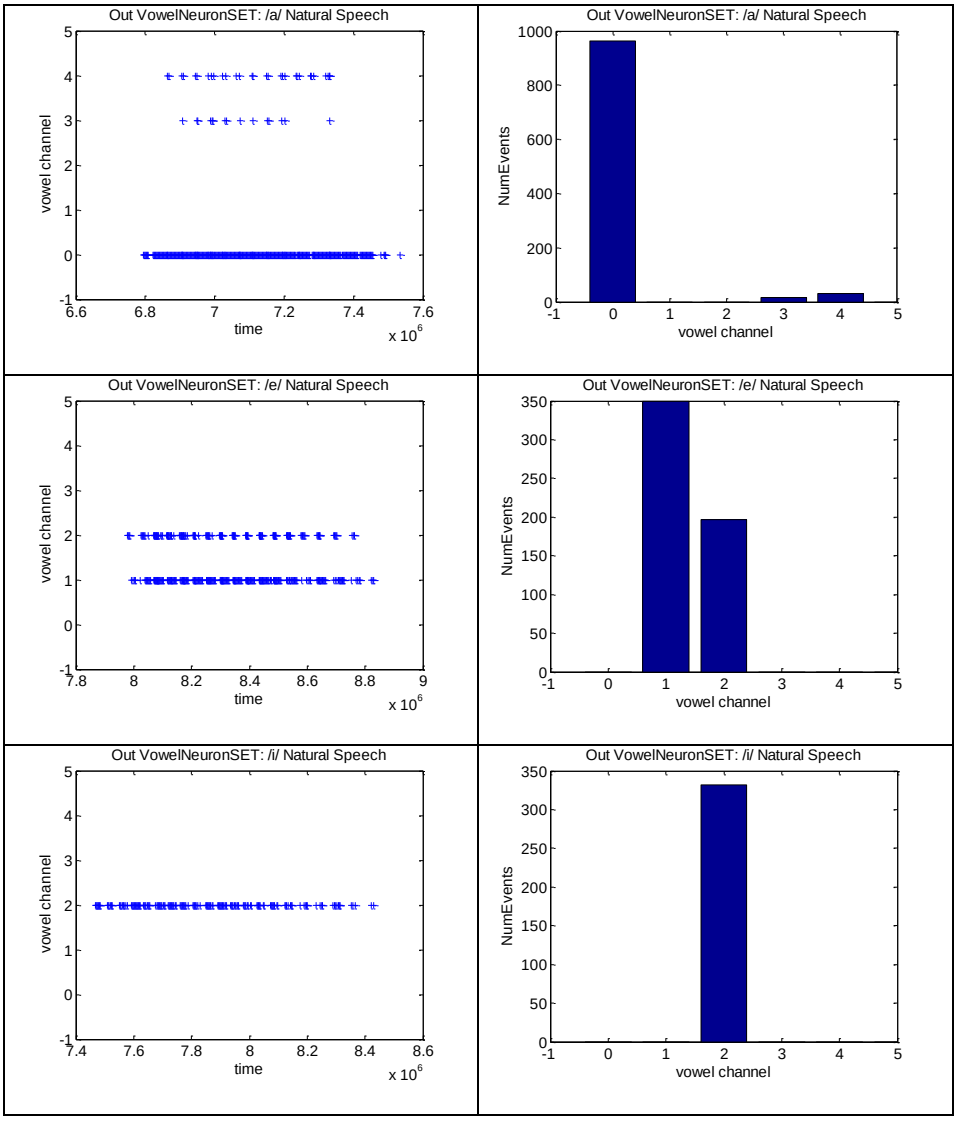
Se ha configurado cada *RNeuron* para identificar un único fonema vocálico. Cada una de ellas tiene un valor de patrón o máscara, un valor de umbral de disparo de las entradas y un tiempo de reinicio. Para definir el valor de cada máscara se ha tenido en cuenta el valor de las dos bandas de la cóclea digital a las que pertenecen las dos primeras formantes del fonema a identificar (ver Tabla 18). Los umbrales de disparo de las entradas se han configurado considerando la respuesta de la cóclea digital pulsante a las frecuencias de las formantes de cada vocal. Los diferentes valores de umbrales tienen su correspondencia biológica en el umbral diferencial de intensidad promedio de cada formante (descrito en el apartado "Las diferencias mínimas perceptibles" del capítulo 3). Se ha descrito como la banda que se corresponde con la

primera frecuencia formante del fonema (F_1) emite un mayor número de eventos que la banda asociada a la frecuencia de la segunda formante (F_2). Por lo tanto, el valor del umbral de la banda asociada a F_1 es mayor que el valor del umbral de la banda asociada a F_2. En cuanto al tiempo de reinicio de estas *RNeuron*, todas tiene el mismo valor, 10 ms (experimentalmente, hemos demostrado que es el mejor valor para el tiempo de reinicio).

Tabla 18. Patrón para identificar a cada uno de los fonemas vocálicos.

	Patrón o máscara	**Bandas asociadas a F_1 y F_2**
RNeuron A	"00000000000001010000"	Bandas 4 y 6.
RNeuron E	"00000000001000000010"	Bandas 1 y 9
RNeuron I	"00000010000000000001"	Bandas 0 y 13
RNeuron O	"00000000000000010100"	Bandas 2 y 4
RNeuron U	"00000000000000001001"	Bandas 0 y 3

La salida de este módulo *VowelNeuronSET*, es una secuencia de eventos *AER* correspondiente al fonema vocálico reconocido. Las direcciones de estos eventos *AER* pertenecen al rango de 0 a 4 (0 para el fonema /a/, 1 para /e/, 2 para /i/, 3 para /o/ y 4 para /u/). En la siguiente Figura 90 se muestra un ejemplo de la salida de esta etapa. Cada fila contiene la información de los eventos *AER* emitidos tras identificar un fonema. En la columna de la izquierda se representan los eventos *AER* con un rango de direcciones de 0 a 4, emitidos a los largo del tiempo y en la columna de la derecha el número total de eventos emitidos para cada dirección.


Out VowelNeuronSET: /a/ Natural Speech
vowel channel
time
x 10^6
Out VowelNeuronSET: /a/ Natural Speech
NumEvents
vowel channel
Out VowelNeuronSET: /e/ Natural Speech
vowel channel
time
x 10^6
Out VowelNeuronSET: /e/ Natural Speech
NumEvents
vowel channel
Out VowelNeuronSET: /i/ Natural Speech
vowel channel
time
x 10^6
Out VowelNeuronSET: /i/ Natural Speech
NumEvents
vowel channel

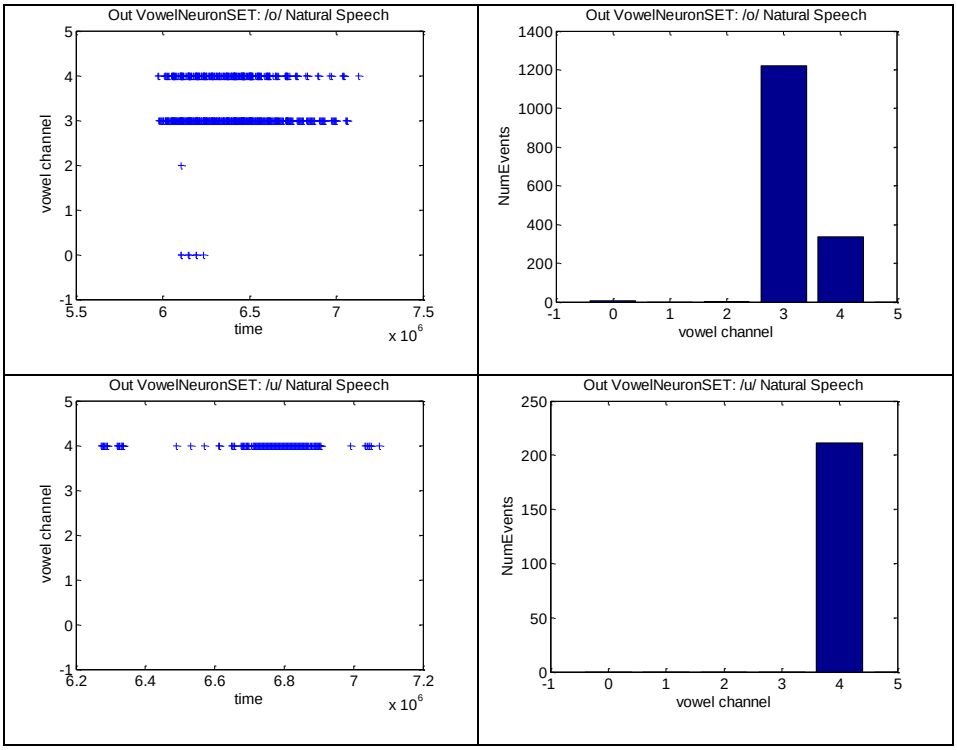


Figura 90. Salida del bloque *vowelNeuronSet*.

En las gráficas anteriores, se observa como los fonemas /i/ y /u/ son identificados de forma unívoca, es decir, sólo se emiten eventos con una única dirección: valor 2 para el fonema /i/y 4 para el fonema/u/; mientras que para los fonemas /a/, /e/ y /o/ aparecen eventos de distintas direcciones. Aunque hay que resaltar también que la diferencia entre el número de eventos con la dirección correcta y el número de eventos con otras direcciones es muy grande y por lo tanto el sistema con esta información es capaz de identificar el fonema correcto.

7.4.2. Bloque VowelWTAsSET

Para una mayor exactitud en el proceso de identificación, se ha usado el bloque *vowelWTAsSET,* con el objetivo de potenciar y de esta manera diferenciar aún más la identificación de un fonema vocálico respecto de otros fonemas vocálicos. Como se ha descrito en la sección anterior, ante el sonido del fonema /a/, el módulo *VowelNeuronSET* identifica el fonema /a/, /o/ y /u/; ante el sonido del fonema /e/, identifica el fonema /e/ y /i/; y ante el sonido del fonema /o/, identifica el fonema /o/ y /u/. Aunque la diferencia de número de eventos entre el fonema correcto y los otros fonemas identificados es significativa, gracias al bloque *VowelWTAsSET* se consigue eliminar los eventos correspondientes a fonemas identificados de manera incorrecta, Figura 92.

Interfaz del bloque *VowelWTAsSET*

El bloque *VowelWTAsSET* tiene dos puertos de comunicación *AER*, uno de entrada y otro de salida. Por el puerto de entrada recibe una secuencia de eventos *AER*, correspondiente a los fonemas vocálicos identificados en el bloque anterior, *VowelNeuronSET*. Esta información es analizada para elegir el fonema vocálico ganador, el cual será enviado a través del puerto de salida como una secuencia de eventos *AER*.

Descripción funcional del bloque *VowelWTAsSET*

El bloque *vowelWTAsSET,* Figura 91, está formado por un conjunto de 5 neuronas *WTANeuron*, una por cada fonema a identificar.

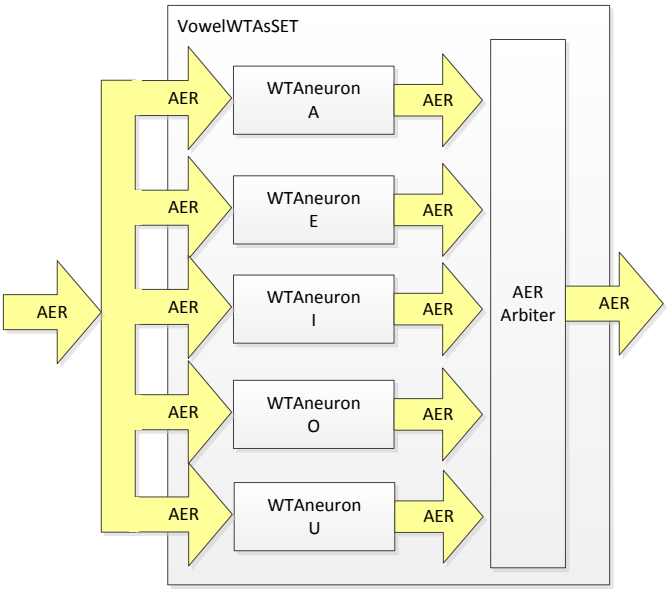


Figura 91. Bloque *VowelWTAsSET*.

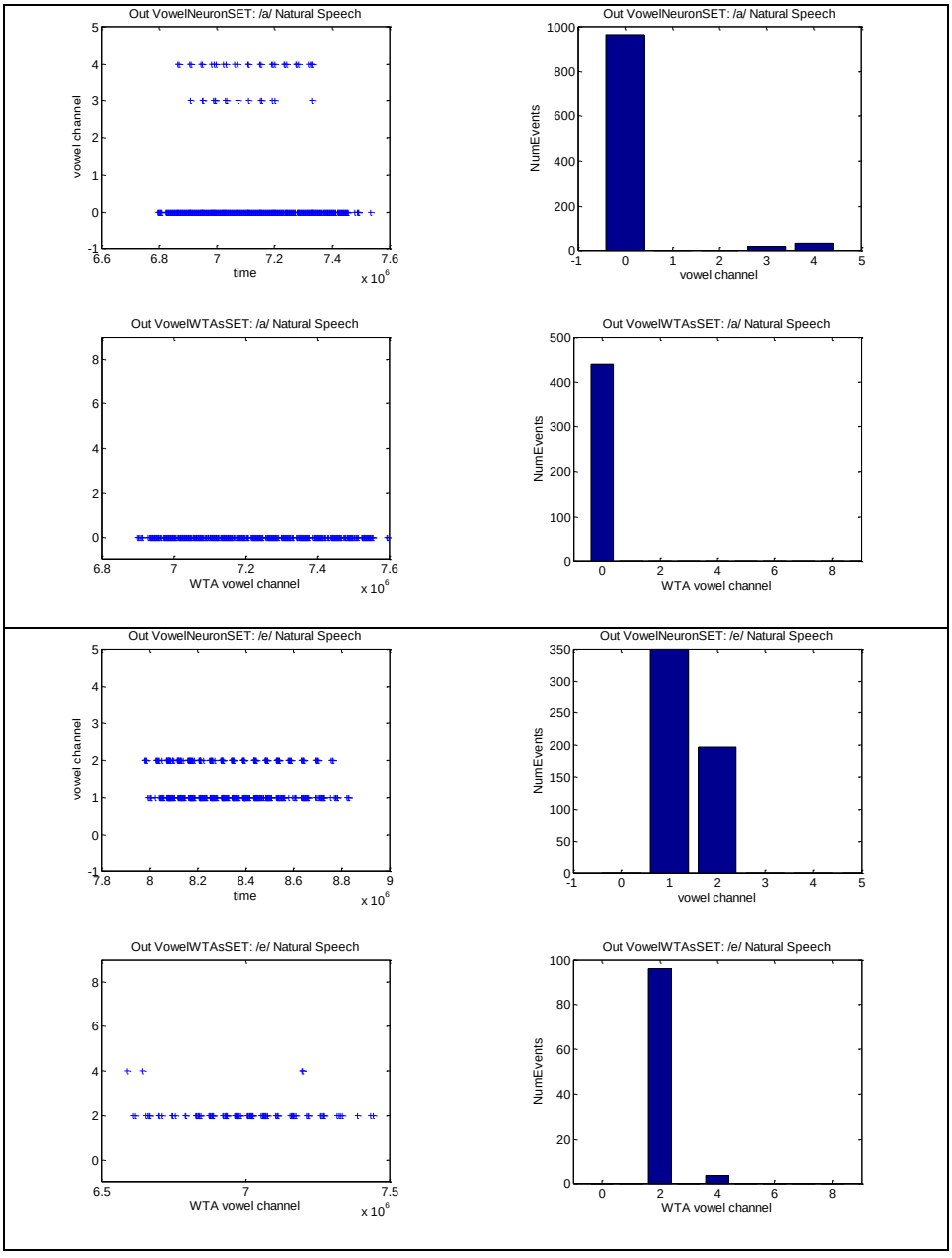
Cada *WTANeuron* va a elegir un fonema vocálico como el ganador, tal como se especifica en su parámetro identificador o máscara. El valor de los pesos excitadores, Ve, e inhibidores, Vi, de las entradas y el del umbral de potencial de cada *WTANeuron* se han elegido experimentalmente y teniendo en cuenta la dificultad o facilidad del sistema para identificar cada uno de los fonemas, Tabla 19. Todas tienen el mismo valor de tiempo de reinicio, 10ms, valor también obtenido a partir de los experimentos.

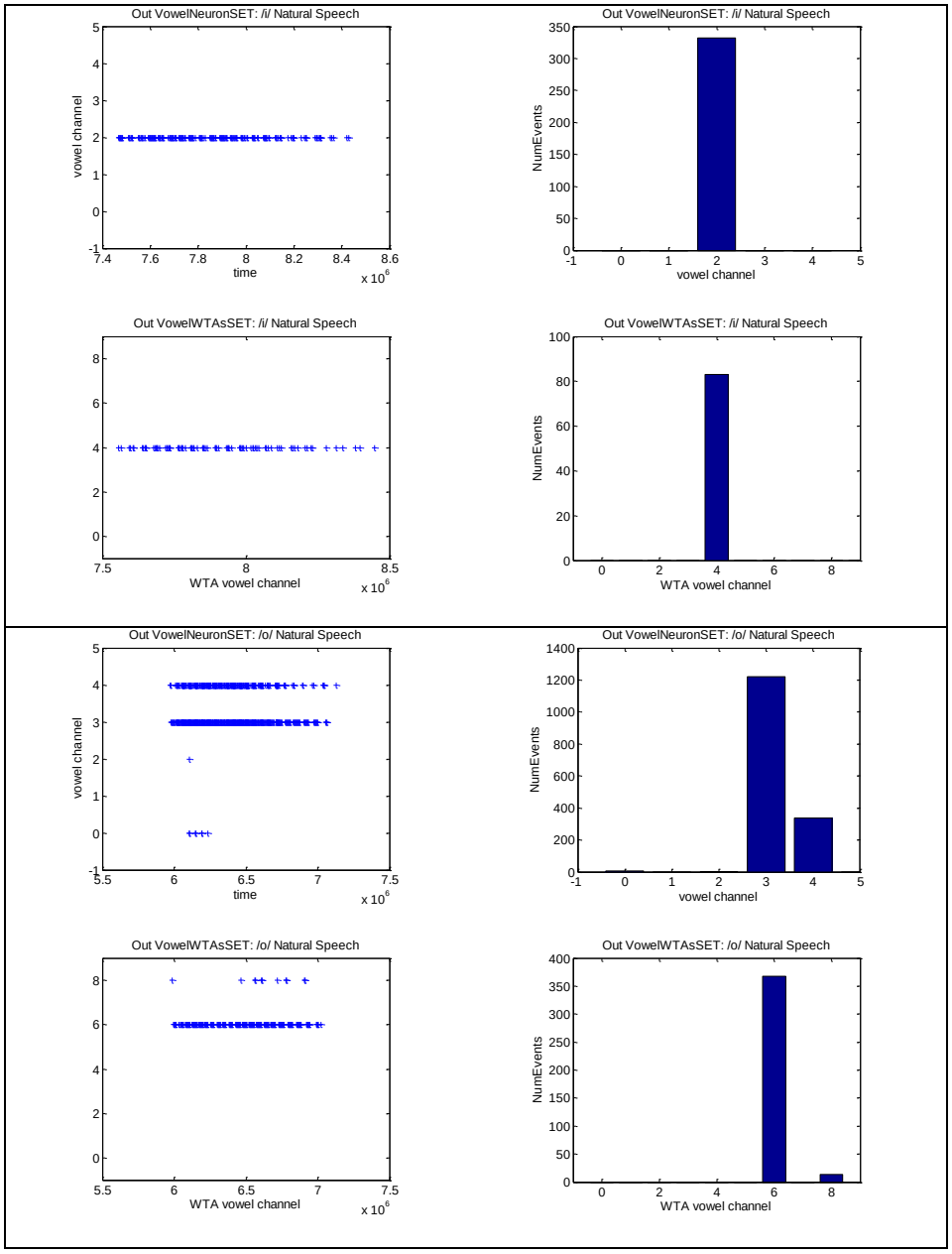
Tabla 19. Parámetros de configuración de cada *WTANeuron* del bloque *VowelsWTAsSET*.

	WTANeuron A	WTANeuron E	WTANeuron I	WTANeuron O	WTANeuron U
Ve	5	5	5	5	5
Vi	10	10	23	23	15
Umbral de potencial	16	16	31	16	16

En la Figura 92 se ilustra la salida de este bloque. Cada recuadro formado por 4 figuras, se corresponde con un fonema vocálico. En cada recuadro, las dos figuras de la primera fila representan la entrada a este bloque, una secuencia de eventos *AER* procedentes del bloque anterior, *VowelNeuronSET*; y las dos figuras de la segunda fila representan la salida de este bloque, también una secuencia de eventos *AER*. Para los fonemas vocálicos /a/, /e/ y /o/, se observa como en la entrada del bloque aparecen eventos cuya dirección no coincide con la dirección que identifica al fonema vocálico que se está procesando. El bloque *VowelWTAsSET* suprime los eventos con una dirección incorrecta, para que en su salida siempre se emita un evento con una dirección correcta. Esto es posible, gracias al uso de los pesos excitadores e inhibidores de cada una de las *WTANeuron*, que van a permitir definir una red neuronal competitiva. Hay que recordar que se ha implementado una competición blanda, de ahí la existencia en la salida de algún evento con una dirección incorrecta, como es el caso de los fonemas vocálicos /e/ y /o/.

Hay que destacar que aunque el bloque *VowelNeuronSET* emite eventos *AER* con direcciones 0 para el fonema /a/, 1 para el fonema /e/, 2 para /i/, 3 para /o/ y 4 para el fonema /u/; el bloque *VowelWTAsSET* emite eventos *AER* con direcciones pares: 0 para el fonema /a/, 2 para el fonema /e/, 4 para /i/, 6 para /o/ y 8 para el fonema /u/. Este cambio de valores de direcciones permite que a la salida de esta etapa, *vowels recognition*, aparezca una secuencia de eventos *AER* con direcciones pares correspondientes al fonema identificado seguido de otra secuencia de eventos *AER* con direcciones impares correspondientes al mismo fonema retrasado.


Out VowelNeuronSET: /a/ Natural Speech
vowel channel
time
x 10^6
Out VowelNeuronSET: /a/ Natural Speech
NumEvents
vowel channel
Out VowelWTAsSET: /a/ Natural Speech
WTA vowel channel
x 10^6
Out VowelWTAsSET: /a/ Natural Speech
NumEvents
WTA vowel channel
Out VowelNeuronSET: /e/ Natural Speech
vowel channel
time
x 10^6
Out VowelNeuronSET: /e/ Natural Speech
NumEvents
vowel channel
Out VowelWTAsSET: /e/ Natural Speech
WTA vowel channel
x 10^6
Out VowelWTAsSET: /e/ Natural Speech
NumEvents
WTA vowel channel


Out VowelNeuronSET: /i/ Natural Speech
vowel channel
time
x 10^6
Out VowelNeuronSET: /i/ Natural Speech
NumEvents
vowel channel
Out VowelWTAsSET: /i/ Natural Speech
WTA vowel channel
x 10^6
Out VowelWTAsSET: /i/ Natural Speech
NumEvents
WTA vowel channel
Out VowelNeuronSET: /o/ Natural Speech
vowel channel
time
x 10^6
Out VowelNeuronSET: /o/ Natural Speech
NumEvents
vowel channel
Out VowelWTAsSET: /o/ Natural Speech
WTA vowel channel
x 10^6
Out VowelWTAsSET: /o/ Natural Speech
NumEvents
WTA vowel channel

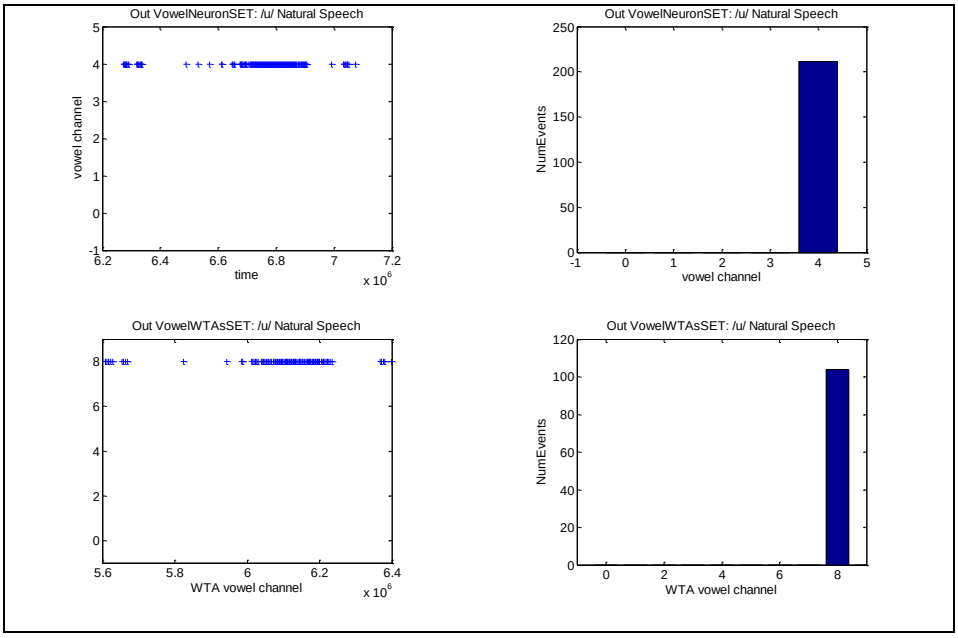


Figura 92. Salida de los bloques *VowelNeuronSET* y *VowelWTAsSET*. En estas figuras se representa el número de eventos emitidos como respuesta a los fonemas /a/, /e/, /i/, /o/ y /u/.

7.4.3. Bloque delayNeuronChain

Los fonemas vocálicos reconocidos, cuyos valores pueden ser 0 para el fonema /a/, 2 para el fonema /e/, 4 para /i/, 6 para /o/ y 8 para el fonema /u/, se retrasan un tiempo en el módulo *delayNeuronChain*. El fonema retrasado se distingue porque su dirección asociada es una dirección impar. La asignación de esta dirección impar es realizada por el módulo *AERMapper*. En este caso, las direcciones de los eventos retrasados serán 1 para el fonema /a/, 3 para el fonema /e/, 5 para /i/, 7 para /o/ y 9 para el fonema /u/. De esta forma a la salida de esta primera etapa tendremos una secuencia de eventos *AER* correspondiente al fonema vocálico identificado, seguido con un cierto retraso de otra secuencia de eventos *AER* del mismo fonema identificado. Solo que las direcciones de los eventos de la primera secuencia son

direcciones pares y las direcciones de la segunda secuencia son impares permitiendo así distinguir entre un evento emitido en su tiempo y un evento retrasado, Figura 93. Esta notación permitirá a la siguiente etapa del sistema, *words recognition*, reconocer una palabra, ya que va a ser capaz de identificar dos fonemas que aunque se han pronunciado uno detrás de otro, gracias a este bloque, *delayNeuronChain*, sus eventos *AER* asociados van a coincidir en el tiempo.

El bloque *delayNeuronChain* es una cadena de neuronas de retraso, *delayNeuron*. Esta estructura modular va a permitir configurar el tiempo de retraso de un evento, simplemente añadiendo a la cadena un número determinado de neurona de retrasos. En nuestro sistema, el bloque *delayNeuronChain* está formada por una única *delayNeuron*.

7.4.4. Módulo AERMapper

El módulo *AERMapper* incrementa en 1 las direcciones de los eventos *AER* que le llegan a través de su puerto *AER* de entrada, y los emite con esta nueva dirección a través de su puerto *AER* de salida.

7.4.5. Módulo AERSplitter

El módulo *AERSplitter* recibe una secuencia de eventos *AER* a través de su puerto *AER* de entrada y emite esa misma secuencia a través de sus dos puerto *AER* de salida. Es el encargado de replicar cada evento *AER* que recibe como entrada y enviar una copia a la salida de esta etapa y otra copia a la cadena de retraso, *DelayNeuronChain*.

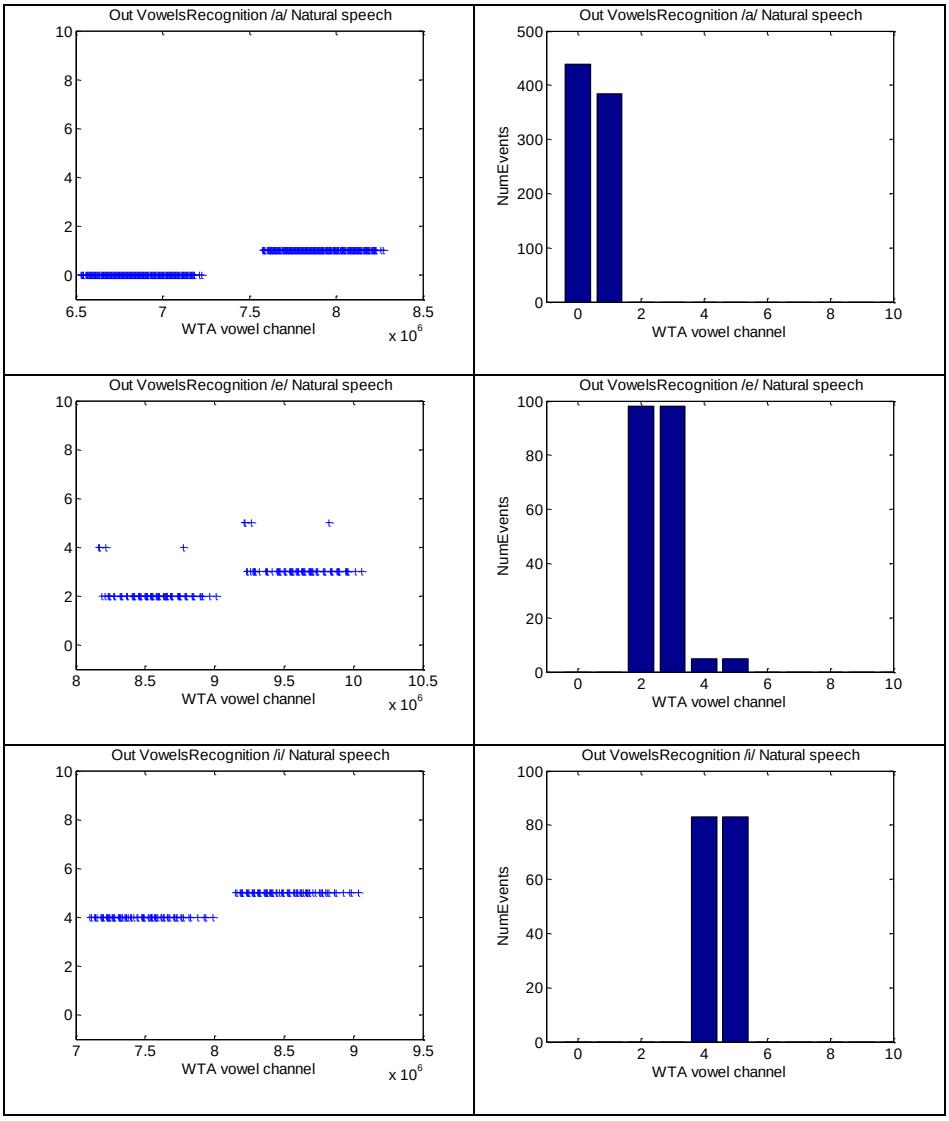
En la siguiente Tabla 20 se resume los recursos de la *FPGA Cyclone IV E* utilizados para la implementación del bloque de reconocimiento de fonemas, formado por el bloque *VowelNeuronSet*, el bloque *VowelWTAsSet*, un módulo

AERSplitter, un módulo *AERMapper,* una neurona *delayNeuron* y un arbitrador de eventos *AER*, Figura 87.

Tabla 20. Recursos usados en la *FPGA Cyclone IV E* para la implementación del bloque *Vowels Recognition*.

	Elementos lógicos		**Memoria (bits)**
	Registros	***LUTs***	
VowelNeuronSet (5 *RNeuron* + arbitrador *AER*)	616/ 114480 (< 1 %)	226/ 114480 (< 1 %)	1680/3981312 (< 1 %)
VowelWTAsSet (5 *WTANeuron* + arbitrador *AER*)	529/ 114480 (< 1 %)	327/ 114480 (< 1 %)	1680/3981312 (< 1 %)
AERMapper	7/ 114480 (< 1 %)	2/ 114480 (< 1 %)	
AERSplitter	5/ 114480 (< 1 %)	1/ 114480 (< 1 %)	
delayNeuron	3148/ 114480 (3 %)	2788/ 114480 (2 %)	
Arbitrador *AER*	11/ 114480 (< 1 %)	2/ 114480 (< 1 %)	
Total	4316/ 114480 (4 %)	3346/ 114480 (3 %)	3360/3981312 (< 1 %)

En la siguiente Figura 93 se ilustra la salida del sistema de reconocimiento de fonemas, *vowels recognition*. Cada fila se corresponde con un fonema vocálico. La primera columna muestra la secuencia de eventos AER emitidos en esta etapa. En ella se aprecia la distinción entre los eventos con direcciones pares y los eventos con direcciones impares (eventos retrasados). De nuevo, hay que destacar que en el proceso de identificación del fonema /e/ aparece algún evento con una dirección incorrecta. Esto se debe a que el bloque *VowelWTAsSET* implementa una red competitiva blanda.


Out VowelsRecognition /a/ Natural speech
WTA vowel channel
x 10^6
Out VowelsRecognition /a/ Natural speech
NumEvents
WTA vowel channel
Out VowelsRecognition /e/ Natural speech
WTA vowel channel
x 10^6
Out VowelsRecognition /e/ Natural speech
NumEvents
WTA vowel channel
Out VowelsRecognition /i/ Natural speech
WTA vowel channel
x 10^6
Out VowelsRecognition /i/ Natural speech
NumEvents
WTA vowel channel

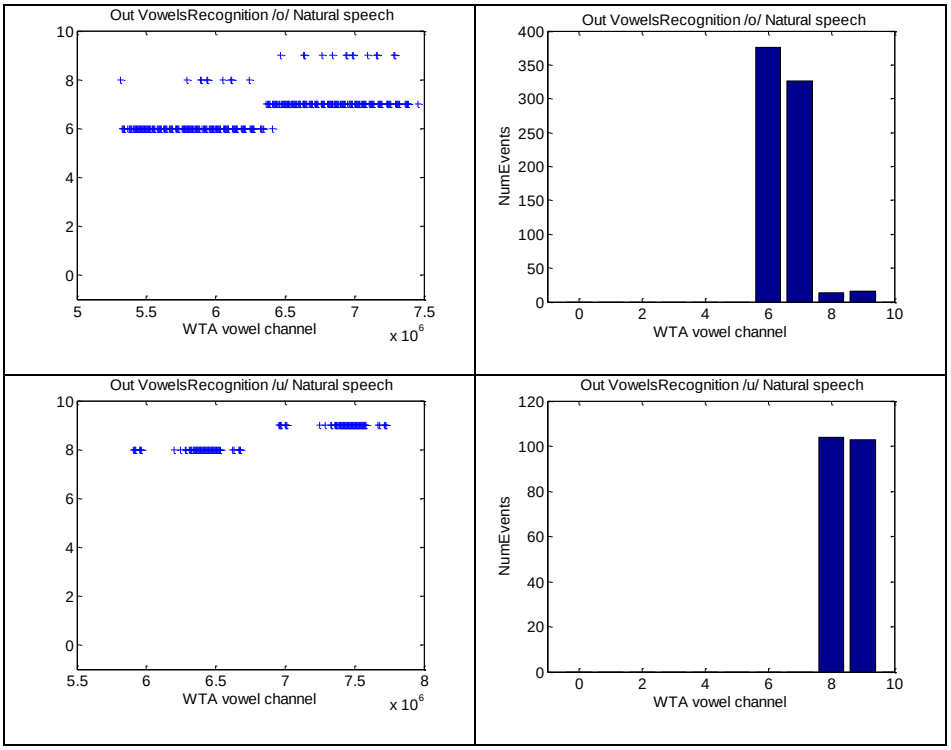


Figura 93. Salida de *VowelsRecognition*. En estas figuras se representa el número de eventos emitidos como respuesta a los fonemas /a/, /e/, /i/, /o/ y /u/.

7.5. Sistema de reconocimiento de palabras

En esta segunda etapa, *words recognition,* se va a identificar una colección de palabras a partir de la información de los fonemas vocálicos reconocidos en la etapa anterior, *vowels recognition*. En la Figura 94 se muestra el diagrama de bloques del sistema de reconocimiento de palabras que está formado por los bloques *WordNeuronSet* y *WordWTAsSet*.

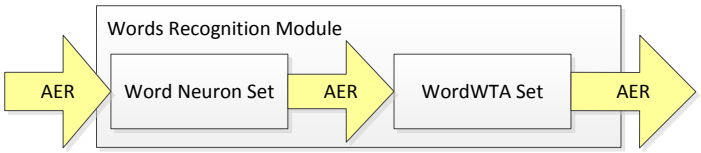


Figura 94. Módulo *WordsRecognition.*

7.5.1. Bloque WordNeuronSet

En este sistema se ha considerado que una palabra es una secuencia de dos fonemas vocálicos. De modo que, este sistema reconocerá la palabra RIMA, por ejemplo, cuando identifique la secuencia "ia". Se ha descrito como en la etapa anterior, *vowelsRecognition*, cada vez que se reconoce un fonema vocálico se emite un evento *AER* asociado a dicho fonema y con un cierto retraso se vuelve a emitir el mismo evento *AER*. De manera que a la entrada del módulo *WordsRecognition*, para la palabra RIMA tendremos la siguiente secuencia de eventos *AER*: llegan eventos asociados al fonema /i/ (dirección 4); después llegan eventos asociados al fonema /i/(retrasado) (dirección 5) y eventos asociados al fonema /a/ (dirección 0); y después llegan los eventos asociados al fonema /a/ (retrasado) (dirección 1). Es precisamente en el momento en que le llegan los eventos asociados a /i/(retrasado) y eventos asociados a /a/, direcciones 5 y 0, cuando nuestro sistema es capaz de reconocer la secuencia "ia" y por tanto la palabra RIMA. Esta coincidencia en el tiempo se observa en la siguiente Figura 95.

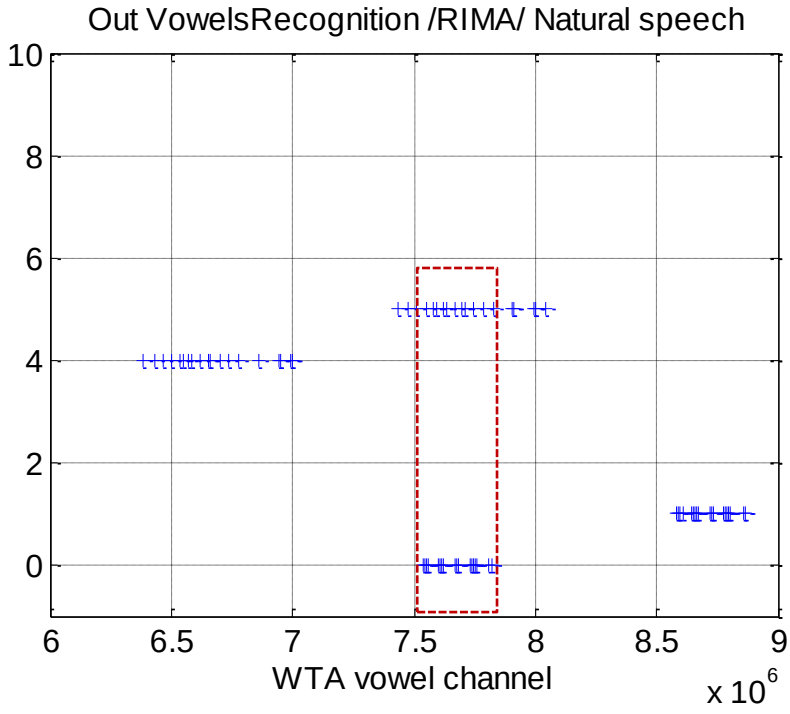


Figura 95. Salida de la etapa *VowelsRecognition* tras pronunciarse la palabra RIMA (secuencia de fonemas /i/, /a/). El recuadro rojo marca el periodo de tiempo en el que se produce el emparejamiento de los eventos.

Interfaz del bloque *WordNeuronSET*

El bloque *WordNeuronSET* tiene dos puertos de comunicación *AER*, uno de entrada y otro de salida. Este bloque analiza la secuencia de eventos *AER* que recibe a través del puerto de entrada y emite a través del puerto de salida una secuencia de eventos *AER* correspondiente a la palabra reconocida.

Descripción funcional del bloque *WordNeuronSET*

La secuencia de eventos *AER* procedente de la etapa anterior, *words recognition*, que contienen la información de los fonemas vocálicos identificados, llega en paralelo a la entrada de un conjunto de *RNeuron*, una por cada palabra que se quiere reconocer, Figura 96.

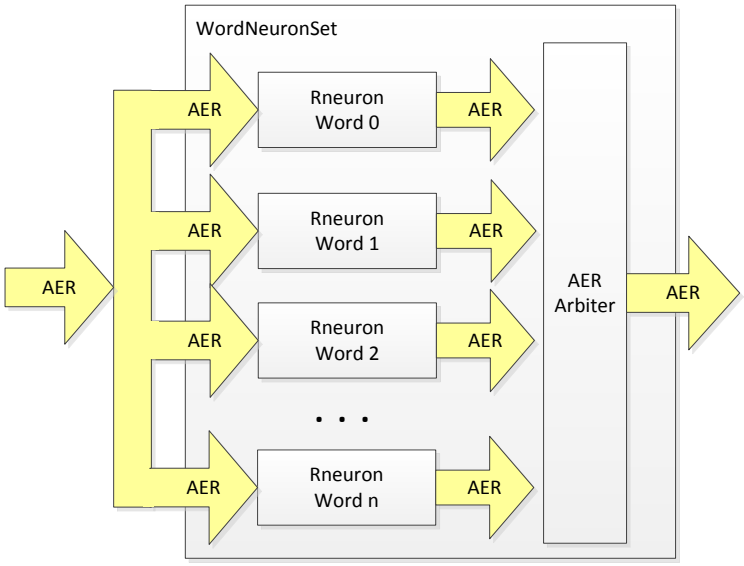


Figura 96. Bloque WordNeuronSet

El funcionamiento de este bloque es equivalente al bloque *VowelNeuronSet*. En lugar de identificar para cada vocal las dos bandas de la cóclea digital pulsante asociadas a las dos frecuencias formantes, en este bloque se identifican los dos fonemas vocálicos que constituyen una palabra, el primero de ellos retrasado. Las direcciones asociadas a estos fonemas vocálicos identificados permiten definir el valor de patrón o máscara de cada *RNeuron* (Tabla 21). Para configurar el valor de umbral de disparo de las entradas de cada *RNeuron* se ha tenido en cuenta el número de eventos que llegan a esta etapa del sistema. Se ha observado que en general la segunda sílaba de una palabra produce un número menor de eventos, aunque esto va

depender del tipo de consonante que le acompaña. Y en cuanto al tiempo de reinicio, todas tienen el mismo valor, 5 ms.

Tabla 21. Patrón para identificar las 11 palabras definidas en el sistema.

	Patrón	**Direcciones asociadas a los fonemas vocálicos**
RNeuron AE	"0000000110"	1 (a_{delay}), 2(e)
RNeuron AO	"0001000010"	1 (a_{delay}), 6(o)
RNeuron EA	"0000001001"	3 (e_{delay}), 0(a)
RNeuron EO	"0001001000"	3 (e_{delay}), 6(o)
RNeuron IA	"0000100001"	5 (i_{delay}), 0(a)
RNeuron IE	"0000100100"	5 (i_{delay}), 2(e)
RNeuron IO	"0001100000"	5 (i_{delay}), 6(o)
RNeuron OA	"0010000001"	7 (o_{delay}), 0(a)
RNeuron UA	"1000000001"	9 (u_{delay}), 0(a)
RNeuron UE	"1000000100"	9 (u_{delay}), 2(e)
RNeuron UI	"1000010000"	9 (u_{delay}), 4(i)

La salida de este bloque *WordNeuronSet* es una secuencia de eventos *AER* correspondiente a la palabra reconocida. Sus valores pertenecen al rango de 0 a 10 (0 para la secuencia "AE", 1 para "AO", 2 para "EA", 3 para "EO", 4 para "IA", 5 para "IE", 6 para "IO", 7 para "OA", 8 para "UA", 9 para "UE" y 10 para "UI"). En la siguiente Figura 97 se muestra un ejemplo de salida de esta etapa después de pronunciarse la palabra RIMA. En la primera fila, figura de la izquierda, se puede observar la secuencia de eventos que le llega al bloque *WordsNeuronSet*, que han sido generados en la etapa anterior. En la figura de la izquierda se puede observar cómo existen eventos de la dirección 5 (fonema /i/ retrasada) y de la dirección 0 (fonema /a/) que coinciden en el tiempo. En la segunda fila, se muestra que el sistema ha reconocido la secuencia "IA". Ha detectado 10 coincidencias de la secuencia IA y por tanto ha generado 10 eventos con la dirección 4.

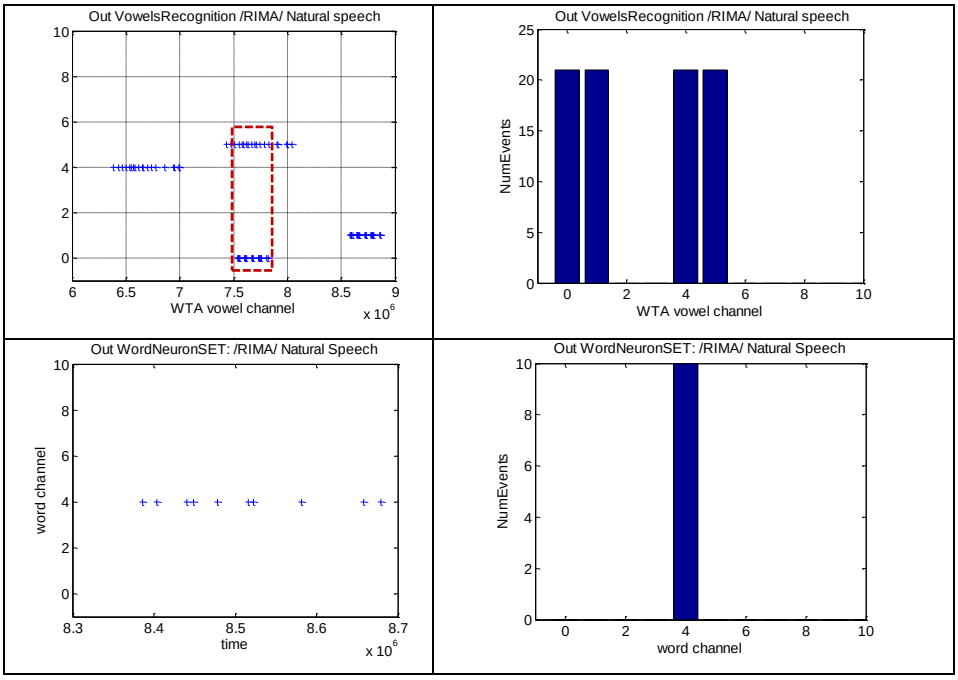


Figura 97. Salida del bloque *WordNeuronSet* tras pronunciar la palabra RIMA (secuencia de fonemas "IA" que se corresponde con la dirección 4).

En el ejemplo anterior, la palabra ha sido identificada de forma unívoca; sin embargo, puede ocurrir que en el proceso de identificación de una palabra aparezcan eventos de distintas direcciones. Aunque hay que resaltar que ocurre en pocos casos y que, al igual que ocurría con la identificación de los fonemas vocálicos, la diferencia entre el número de eventos con la dirección correcta y el número de eventos con otras direcciones es significativa y por lo tanto el sistema es capaz de reconocer la palabra correcta.

7.5.2. Bloque WordWTAsSET

En el caso del reconocimiento de palabras también se ha usado un conjunto de neuronas ganadoras *WTANeuron*, una por cada palabra que se quiere identificar, para una mayor exactitud en el proceso de identificación.

Interfaz del bloque *WordWTAsSET*

El bloque *WordWTAsSET* tiene dos puertos de comunicación *AER*, uno de entrada y otro de salida. Por el puerto de entrada recibe una secuencia de eventos *AER*, correspondiente a las palabras identificadas en el bloque anterior, *WordNeuronSET*. Esta información es analizada para elegir la palabra ganadora, información que será enviada a través del puerto de salida como una secuencia de eventos *AER*.

Descripción funcional del bloque *WordWTAsSET*

El bloque *WordWTAsSET,* Figura 98, está formado por un conjunto de neuronas *WTANeuron*, una por cada palabra a identificar.

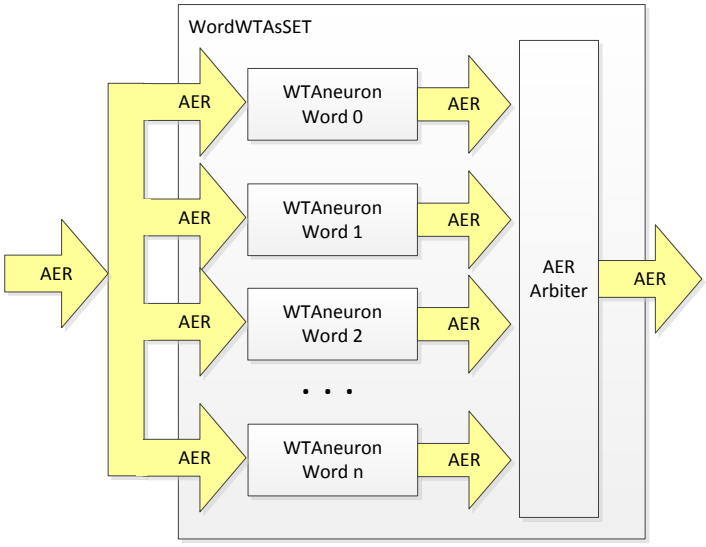


Figura 98. Bloque WordWTASet.

Cada *WTANeuron* va a elegir una palabra como la ganadora coincidiendo con el valor de su parámetro identificador o máscara. Todas tienen el mismo valor de reinicio, 2,5 ms (valor calculado empíricamente). El valor de los pesos excitadores, Ve, e inhibidores, Vi, de las entradas y el valor del umbral de potencial de cada una de ellas se ha elegido después de un proceso de testeo del sistema. En la siguiente Tabla 22 se muestran los valores elegidos para un proceso de reconocimiento óptimo.

Tabla 22. Parámetros de configuración de las *WTANeuron* del bloque *WordsWTAsSET*.

	AE	**AO, OA, UI**	**EA,EO,IE,UE**	**IA**	**IO**	**UA**
Ve	4	4	4	4	4	4
Vi	2	8	1	4	5	4
Umbral de potencial	2	15	1	4	1	1

En la siguiente Tabla 23 se resume los recursos de la *FPGA Cyclone IV E* utilizados para la implementación de la etapa de reconocimiento. En la primera fila se presentan los resultados correspondientes a la etapa de reconocimiento de fonemas, obtenidos de la Tabla 20; y en la segunda y tercer fila los correspondientes al bloque de reconocimiento de palabras, formado por el bloque *WordNeuronSet*, y el bloque *WordWTAsSet*, Figura 94.

Tabla 23. Recursos usados en la *FPGA Cyclone IV E* para la implementación de la etapa de reconocimiento.

	Elementos lógicos		**Memoria (bits)**
	Registros	***LUTs***	
Vowels recognition	4316/ 114480 (4 %)	3346/ 114480 (3 %)	3360/3981312 (< 1 %)
WordNeuronSet (11 *RNeuron* + arbitrador *AER*)	1341/ 114480 (1 %)	464/ 114480 (< 1 %)	3696/3981312 (< 1 %)

WordWTAsSet (11 *WTANeuron* + arbitrador *AER*)	1154/ 114480 (1 %)	751/ 114480 (< 1 %)	3696/3981312 (< 1 %)
Total	6811/ 114480 (6 %)	4561/ 114480 (4 %)	10752/3981312 (< 1 %)

A continuación, se presenta en la Tabla 24 un resumen de los recursos de la *FPGA Cyclone IV E* utilizados para la implementación de todo el sistema. Hay que tener en cuenta que junto a la cóclea digital pulsante formada por 21 bandas (Figura 75), y el sistema de reconocimiento (Figura 82), el sistema total también incluye un módulo para obtener la señal sonora a partir del *codec* de audio WM8731/WM8731L (Microelectronics, 2009), de la placa de desarrollo EP4CE115F29C7. Esta interfaz ha requerido el uso de un PLL[56], para generar una correcta señal de reloj.

Tabla 24. Resumen de recursos usados en la *FPGA Cyclone IV E* para la implementación de todo el sistema.

Registros	9786/ 114480 (9 %)
LUTs	23521/ 114480 (21 %)
Memoria (bits)	10752/3981312 (< 1 %)
Multiplicador 9-bits	210/532 (39 %)
PLL	1/4 (25 %)

7.6. Visión global del sistema

En la siguiente Figura 99 se resume todo el proceso llevado a cabo por el sistema descrito en este trabajo, desde la cóclea digital pulsante hasta la identificación de la

[56] *PLL*, siglas en inglés *Phase-Locked Loop*. Dispositivo electrónico en el que la frecuencia y la fase son realimentados. Son usados en las FPGAs para el manejo y adaptación de los recursos de reloj.

palabra. Cada fila de la figura se corresponde con la salida de una etapa del sistema: la 1ª fila, la cóclea digital pulsante; la 2ª fila, el sistema de reconocimiento de fonemas vocálicos, *VowelsRecognition*; y la 3ª fila, el sistema de reconocimiento de palabras, *WordsRecognition*. En cada fila se muestra la información en forma de cocleograma, secuencia de eventos *AER* a lo largo del tiempo, y en forma de histograma, número de eventos *AER* emitidos por cada banda o canal.

A partir de la señal de audio, que en este caso se corresponde con la pronunciación de la palabra "LETRA", nuestro sistema es capaz de convertir los valores *PCM* de dicha señal en una secuencia de eventos *AER*. Este proceso se lleva a cabo en la cóclea digital pulsante, formada por 21 bandas (par de filtro paso banda y generador de pulso), Figura 75. Cada una de las bandas responderá ante el estímulo de la señal de audio de un modo más activo, si las componentes en frecuencia de dicha señal de audio coinciden con la frecuencia central del filtro paso banda.

En la primera columna de la primera fila, se muestra la salida de la cóclea, en ella se distingue claramente los dos golpes de voz correspondientes a las dos sílabas de la palabra LETRA. En la primera parte se observa como las bandas que emiten un mayor número de eventos (bandas 1 y 9), son las bandas cuyos filtros tienen su frecuencia central próxima a la frecuencia de las formantes F_1 y F_2 del fonema vocálico /e/ (ver Tabla 17). En la segunda parte, las bandas que más emiten (bandas 3 y 7), tienen su frecuencia central próxima a la frecuencia de las formantes F_1 y F_2 del fonema vocálico /a/. En el cocleograma se puede observar una división en el segundo golpe de voz correspondiente al sonido /r/ seguido del sonido /a/.

Es interesante resaltar que se va a emitir un mayor número de eventos para la primera sílaba, porque su pronunciación es más fuerte que en el caso de la segunda sílaba, por ser una palabra llana. También hay que destacar que las bandas que disparan ante un determinado fonema vocálico pueden sufrir alteraciones según el contexto de la vocal, es decir, según la consonante que le acompañe.

En la 2ª fila, 3ª y 4ª fila, se muestra la salida de la etapa de reconocimiento de fonemas vocálicos, *vowels recognition*. Esta etapa recibe la secuencia de eventos *AER* procedentes de la cóclea digital pulsante. Estos eventos *AER* de entrada, tienen una dirección perteneciente al rango 0-20, correspondiente a las 21 bandas de la cóclea. En esta etapa de reconocimiento de fonemas vocálicos se distinguen tres fases. En primer lugar, la secuencia de eventos *AER*, con direcciones en el rango 0-20, llega en paralelo a un conjunto de 5 *RNeuron*, encargadas de la identificación de los 5 fonemas vocálicos de la lengua española. El proceso de reconocimiento realizado por estas neuronas consiste básicamente en la detección de un determinado patrón o máscara. El patrón de cada *RNeuron* se define a partir del valor de las bandas más activas de la cóclea para el fonema vocálico que se quiere reconocer, Tabla 18. A la salida de este conjunto de *RNeuron* se tiene una secuencia de eventos *AER*, correspondiente al fonema vocálico identificado. Los eventos *AER* emitidos, en este caso, tienen direcciones en el rango de 0-4 (0 para el fonema /a/, 1 para /e/, 2 para /i/, 3 para /o/ y 4 para /u/). La identificación de un fonema no es unívoco, es decir pueden aparecer en la salida de este bloque eventos *AER* cuyas direcciones no se correspondan con el fonema identificado. Se muestra en la segunda fila que coincidiendo en el tiempo aparecen eventos del fonema /a/ (dirección 0), junto a eventos de otros fonemas.

Por tanto, la salida de este bloque llega a un segundo bloque formado por 5 *WTANeuron* (3ª fila), que se comportan como una red neuronal competitiva, de manera que a la salida de este bloque se emitirá una secuencia de eventos *AER* cuyas direcciones identifican al fonema vocálico reconocido como el fonema vocálico ganador.

A continuación (4ª fila), esta secuencia de eventos *AER* pasa por una cadena de retraso, *delayNeuronChain*, encargada de retrasar dicha secuencia de eventos. De manera que a la salida de esta etapa, aparece una secuencia de eventos *AER* correspondiente al fonema vocálico identificado, seguido con un cierto retraso de

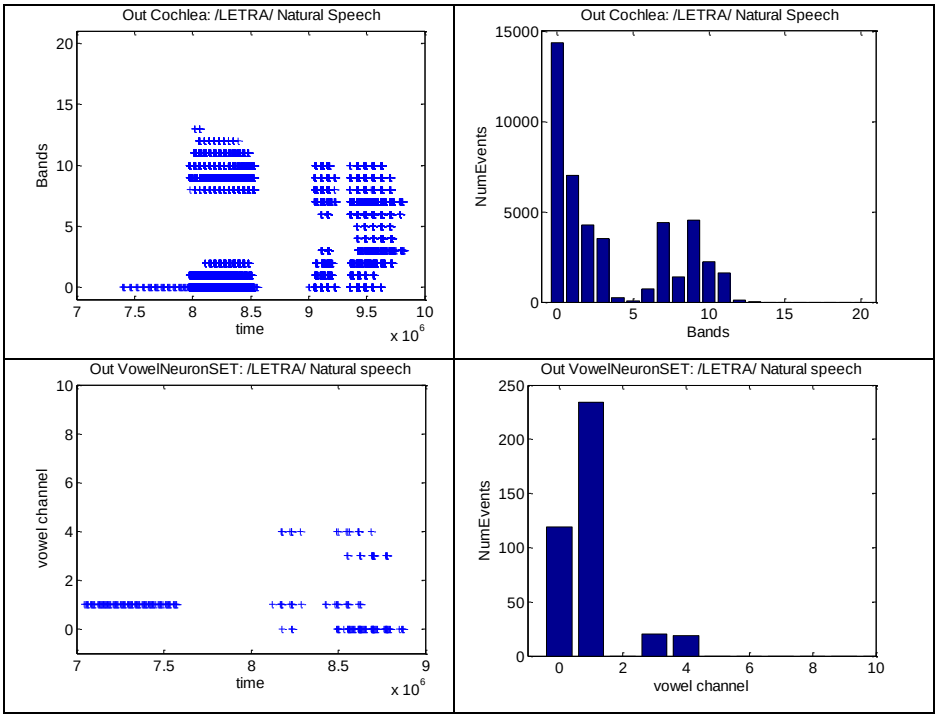
otra secuencia de eventos *AER* del mismo fonema identificado. Solo que las direcciones de los eventos de la primera secuencia son direcciones pares y las direcciones de la segunda secuencia son impares permitiendo así distinguir entre un evento emitido en su tiempo y un evento retrasado. Se ha implementado un módulo *AERMapper* encargado de realizar este cambio de valor de dirección para los eventos retrasados.

En la figura de la cuarta fila y primera columna se observa como el sistema ha reconocido los fonemas vocálicos /e/ (eventos *AER* con dirección 2) y /a/ (eventos *AER* con dirección 0). La secuencia de eventos retrasados tienen las direcciones 3 y 1 respectivamente. Existe una coincidencia en el tiempo de eventos con direcciones 3 y 0 correspondientes a eventos retrasados del fonema /e_{delay}/ y eventos en su tiempo del fonema /a/. Esta es la entrada que recibe la siguiente etapa del sistema, el reconocimiento de palabras, *words recognition*.

En la etapa de reconocimiento de palabras también se distinguen dos fases. En una primera fase la secuencia de eventos *AER*, que contienen la información de los fonemas vocálicos identificados, llega en paralelo a un conjunto de *RNeuron,* que constituyen el bloque *WordNeuronSET* (5ª fila). Cada *RNeuron* de este bloque es capaz de identificar una determinada palabra, usando un algoritmo de detección de un determinado patrón o máscara. En este caso, el patrón de cada *RNeuron* se define a partir del valor de las direcciones asociadas a los fonemas vocálicos que constituyen la palabra a reconocer, Tabla 21. A la salida de este conjunto de *RNeuron* se tiene una secuencia de eventos *AER*, correspondiente a la palabra identificada. Experimentalmente se ha observado, que en el reconocimiento de palabras, la secuencia de eventos *AER* que contienen la información de la palabra identificada no suele contener eventos AER con direcciones incorrectas. Pero a pesar de todo, a esta etapa también se le ha añadido un bloque de neuronas *WTANeuron*, el bloque *WordWTAsSET*, para garantizar que en la salida del sistema sólo aparecerán eventos

AER cuya dirección se corresponde a la palabra que es identificada y que es la ganadora.

En la figura de la quinta fila y primera columna se observa como el sistema ha reconocido una palabra que está formada por la secuencia de los fonemas vocálicos /e/ y /a/ (eventos *AER* con dirección 2), como es el caso de la palabra LETRA. El sistema ha sido capaz de reconocer la coincidencia en el tiempo de los eventos *AER* asociados al fonema vocálico /e_{delay}/ y /a/.



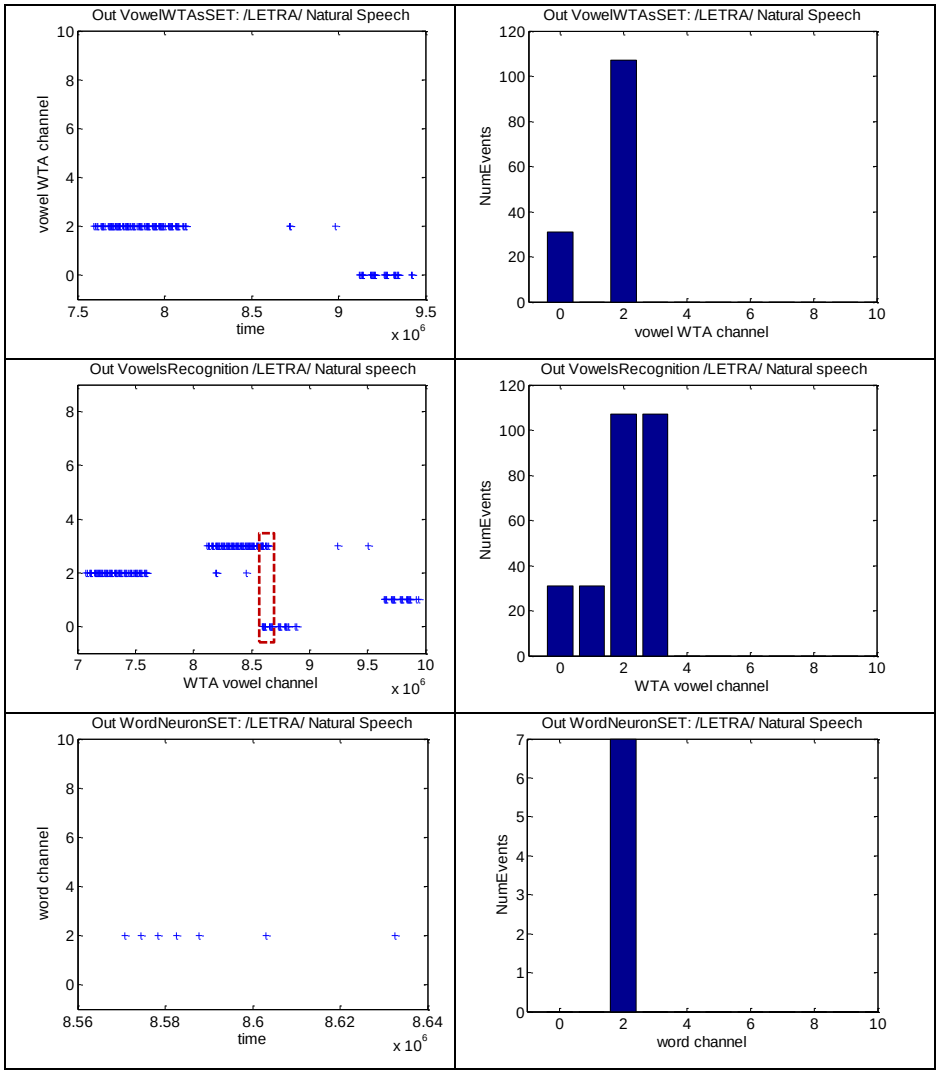


Figura 99. Ejemplo de las salidas de las etapas del sistema.

7.7. Escalabilidad del sistema

El sistema propuesto en este trabajo es capaz de reconocer un conjunto de 11 palabras bisílabas de la lengua española, a partir de la identificación de los fonemas vocálicos que la componen. Sin embargo, gracias al diseño modular presentado, basado fundamentalmente en tres elementos básicos: la neurona de reconocimiento *RNeuron*, la neurona ganadora *WTANeuron* y la neurona de retraso *delayNeuron*; es posible el desarrollo de otros sistemas de reconocimiento combinando estos mismos elementos.

A continuación se describen diferentes posibilidades de configuración del sistema.

7.7.1. Reconocimiento de un conjunto de fonemas

Se ha descrito como el bloque de reconocimiento de fonemas vocálicos, *VowelNeuronSET*, es capaz de identificar cada uno de los 5 fonemas vocálicos de la lengua española. Este bloque se puede generalizar para permitir el reconocimiento de cualquier conjunto de fonemas, ya sean fonemas vocálicos o consonánticos, o incluso fonemas de cualquier otra lengua.

Para ello, será necesario definir el conjunto de neuronas de reconocimiento, formado por tantas *RNeuron* como fonemas se quiera identificar, (Figura 89). Por ejemplo, sea num_{fonema} el número de fonemas que se quieren reconocer. Cada *RNeuron* tendrá su propio identificador, que no es más que un valor perteneciente al rango de 0 a $(num_{fonema} - 1)$, que se usará como dirección del evento *AER* asociado al fonema que representa. Para configurar los parámetros de cada *RNeuron*, la máscara o patrón y los valores de umbral de disparo de las entradas, se hará un estudio de la respuesta de la cóclea digital pulsante para determinar cuáles son las bandas activas de la cóclea.

Además, también se ampliará y configurará el bloque de neuronas ganadoras, *WTANeuron*; tantas como *RNeuron* (Figura 91).

Con esta arquitectura, a la salida de la etapa de reconocimiento de fonemas se tendrá una secuencia de eventos AER, cuyas direcciones tendrán un valor perteneciente al rango de valores de 0 a $(num_{fonema} - 1)$, asociados a cada uno de los fonemas identificados en esta etapa.

Hasta ahora, se ha descrito como cada fonema tiene asociado una única neurona de reconocimiento. Sin embargo, el diseño del sistema también permite que un fonema en concreto pueda ser identificado por más de una neurona de reconocimiento, *RNeuron*. En este caso, bastaría con asignar el mismo identificador (dirección *AER*) a cada una de las *RNeuron* ligadas al mismo fonema. Esta posibilidad permite que el sistema pueda reconocer un fonema pronunciado por diferentes hablantes, por ejemplo un hombre y una mujer, o hablantes con distintas edades, etc., ya que, como se ha expuesto en el capítulo 3, existen diferentes factores que pueden hacer variar la frecuencia de las formantes de un determinado fonema. Por lo tanto, cada *RNeuron* tendrá definida una máscara diferente de acuerdo a las bandas activas de la cóclea para cada pronunciación del mismo fonema.

7.7.2. Reconocimiento de un conjunto de palabras

Existe la posibilidad de ampliar el número de palabras que el sistema puede reconocer. En este caso, el bloque *WordNeuronSET* estaría formado por tantas neuronas de reconocimiento, *RNeuron,* como palabras se quiera reconocer (Figura 96). Si consideramos que el sistema va a identificar $num_{palabra}$, cada *RNeuron* tendrá asignado un identificador cuyo valor va a pertenecer al rango de 0 a $(num_{palabra} - 1)$. Para configurar los parámetros de cada *RNeuron*, concretamente su máscara o patrón y sus valores de umbral de disparo de las entradas, se hará un estudio de la salida de la anterior etapa, etapa de reconocimiento de fonemas, para conocer el número de

eventos *AER* que se emiten para cada dirección *AER,* las cuales representan a los distintos fonemas que se pueden identificar en la etapa anterior, y que servirán para el reconocimiento de una palabra en concreto. Las direcciones que aparecen en la secuencia de eventos *AER* de entrada, con valores pertenecientes al rango 0 a $(num_{fonema} - 1)$, tras la pronunciación de una palabra, servirán para definir la máscara de la *RNeuron* de esa palabra; y el número de eventos por cada dirección *AER* servirá para definir el valor del umbral de disparo de las diferentes entradas. En esta etapa, también hay que ampliar el bloque de neuronas ganadoras, *WTANeuron*, tantas como palabras a identificar (Figura 98).

7.7.3. Reconocimiento de palabras con más de una sílaba

En este trabajo se ha presentado una arquitectura del sistema que permite el reconocimiento de palabras bisílabas. Esto ha sido posible, gracias a la incorporación de un elemento de retraso, *delayNeuron*, que ha permitido hacer coincidir en el tiempo una secuencia de eventos *AER* retrasados (asociados a la primera sílaba de la palabra) con otra secuencia de eventos AER emitidos en su tiempo (asociados a la segunda sílaba de la palabra) (Figura 95).

Si se quiere ampliar esta arquitectura para que el sistema reconozca palabras de más de dos sílabas será necesario añadir nuevas cadenas de retraso; si la palabra tiene tres sílabas, el sistema estará formado por dos cadenas de retraso, Figura 100; si la palabra tiene cuatro sílabas, el sistema tendrá tres cadenas de retraso, y así sucesivamente. Dependiendo de la posición de la sílaba en la palabra, la cadena de retraso estará formada por un número distinto de neuronas de retraso, *delayNeuron*. Es decir, si es la última sílaba de la palabra, no se retrasará, en su camino no existe ninguna cadena de retraso; si es la penúltima sílaba, se añade una cadena de retraso formada por una única neurona *delayNeuron*; si es la antepenúltima sílaba, se añade una cadena de retraso formada por dos neuronas *delayNeuron*; y así sucesivamente.

Obviamente, será necesario cambiar el módulo *AERSplitter* para que replique la secuencia de eventos *AER* en las diferentes cadenas de retraso. Su número de puertos *AER* de salida va a depender del número de cadenas de retraso del sistema. Si tenemos una cadena de retraso, el módulo *AERSplitter* tendrá dos puertos de salida; si tenemos dos cadenas de retraso, tendrá 3 puertos de salida, etc.

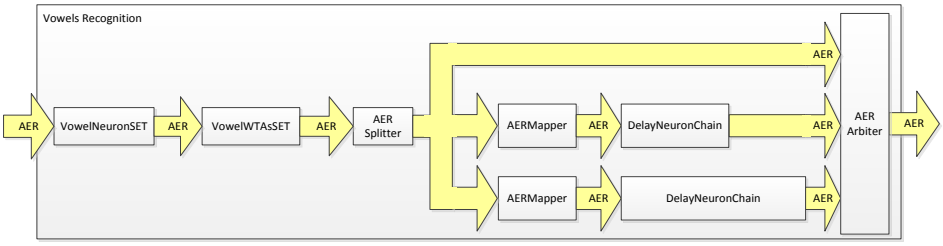


Figura 100. Ejemplo de arquitectura del sistema para el reconocimiento de palabras de tres sílabas.

7.7.4. Reconocimiento de frases

Una vez que el sistema es capaz de identificar cualquier palabra, el sistema también se puede configurar para que reconozca frases, teniendo en cuenta que una frase no es más que una secuencia de palabras.

Para ello, se tiene que añadir un nuevo nivel de reconocimiento que será el encargado de reconocer frases. Es decir, a la salida del bloque de reconocimiento de palabras, *words recognition*, se conectará un nuevo bloque que a partir de la secuencia de eventos AER, que contiene la información de las distintas palabras identificadas, será capaz de identificar la frase pronunciada por el hablante.

Este nuevo bloque estará compuesto por un bloque de cadenas de retraso, con neuronas *delayNeuron*, un bloque de neuronas de reconocimiento, *RNeuron*, y un bloque de neuronas ganadoras, *WTANeuron*, Figura 101.

El primer bloque de cadenas de retraso se configurará dependiendo del número de palabras que tienen las frases a identificar. Si la frase tiene dos palabras, el bloque tendrá 1 cadena de retraso; si tiene tres palabras, tendrá 2 cadenas de retraso, y así sucesivamente. La idea de este bloque es hacer coincidir en el tiempo los eventos *AER* que identifican a las distintas palabras que forman la frase, para que en el siguiente bloque de reconocimiento de frase se pueda aplicar el mismo mecanismo de detección de patrones usado para el reconocimiento de palabras.

El segundo bloque de reconocimiento de frase estará formado por un conjunto de neuronas *RNeuron*; tantas *RNeuron* como frases se quieran identificar. La configuración de los parámetros de estas *RNeuron* va a depender de la salida de la etapa anterior, *words recognition*. En la máscara de cada *RNeuron* se marcará las distintas direcciones de las palabras que forman la frase que representa esta neurona. La salida de este bloque será una secuencia de eventos AER, donde las direcciones de estos eventos representan las distintas frases que el sistema puede reconocer.

Al igual que en las anteriores etapas, también se añadirá un tercer bloque formado por neuronas ganadoras, *WTANeuron*, tantas como frases a reconocer; con la finalidad de que en la salida del sistema se presente la frase identificada que a su vez es la frase ganadora.

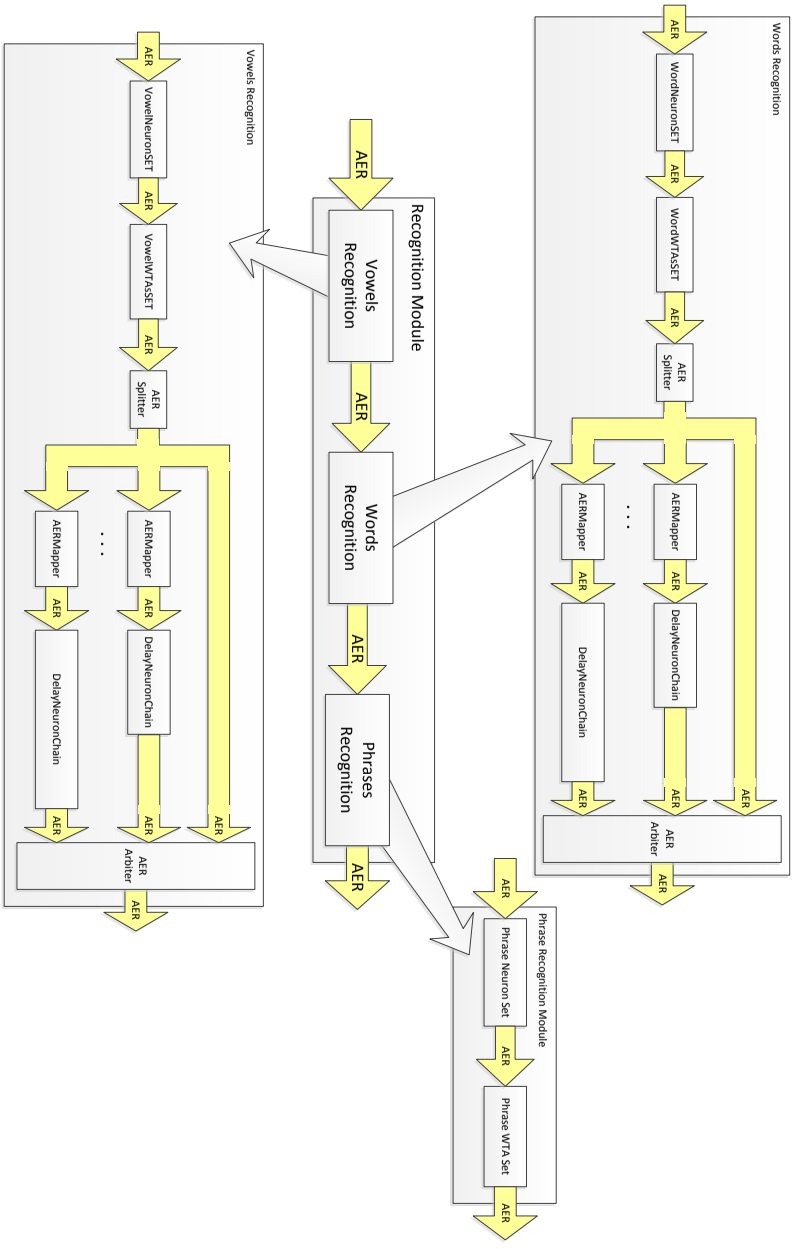


Figura 101. Ejemplo de arquitectura del sistema para el reconocimiento de frases.

Capítulo 8

Experimentos

Una vez explicado el sistema propuesto, se presentan una serie de experimentos que van a permitir comprobar las capacidades y viabilidad de la implementación del mismo.

8.1. Entorno de experimentación

La base de datos o *corpus*, utilizado en la evaluación del sistema propuesto, contiene 10 palabras de la lengua española: *cero, dique, fase, goma, letra, musa, nube, pino, poda* y *rima*. Cada palabra ha sido pronunciada y grabada, por cada uno de los siete hablantes que han participado en el experimento, obteniendo una base de datos de 10x7 archivos en formato *wav*. Junto a las palabras, también se grabaron las 5 vocales de la lengua española, añadiendo a la base de datos 5x7 archivos más. Estos últimos archivos, correspondientes a las vocales, han sido utilizados en la etapa de

configuración del sistema, para obtener los valores de las formantes F1 y F2 que caracterizan a cada fonema vocálico (Tabla 17), así como los umbrales de las *RNeuron* y los parámetros de las *WTANeuron* (Tabla 19).

Las grabaciones se realizaron en un laboratorio con ambiente semi-controlado; se ha usado el software matemático *Matlab* (ver en el anexo los scripts *SoundVowels* y *SoundWords*); una frecuencia de muestreo de 11025 Hz, resolución de 16 bits y adquisición monoaural.

Los siete hablantes que han intervenido en las grabaciones tienen el español como lengua materna y una edad comprendida entre 20 y 50 años. Todos varones.

El hardware usado (Figura 102) en la implementación y pruebas del sistema es la placa de desarrollo EP4CE115F29C7, que dispone de una *FPGA* de la familia *Cyclone IV E* (Altera, 2010), descrito en el capítulo 6. Como interfaz con el PC, se ha usado la placa *USBAERmini2*, descrita en el capítulo 2. Esta placa permite obtener los eventos *AER* de salida y enviarlos al PC por medio de una conexión *USB* para su posterior análisis.

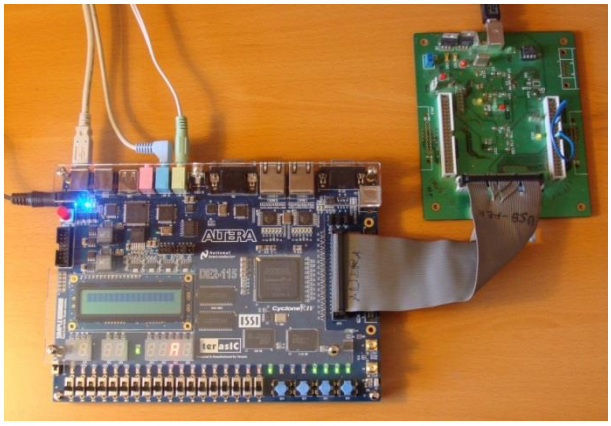


Figura 102. Hardware usado en la implementación y pruebas del sistema, placa de desarrollo EP4CE115F29C7 y placa *USBAERmini2,* respectivamente.

8.2. Experimentos y resultados

A continuación se muestran los resultados de dos grupos de experimentos. En el primero de ellos, el sistema está configurado para la identificación de los 5 fonemas vocálicos de la lengua española (ver script *Exp1*, en el capítulo de anexos). En el segundo, el sistema fue configurado para el reconocimiento de las 10 palabras enumeradas con anterioridad. Esta discriminación se hace, como ya se ha explicado en el capítulo 7, basándose únicamente en la presencia de los fonemas vocálicos (ver script *Exp2*, en el capítulo de anexos).

8.2.1. Experimento con vocales

En este experimento se ha probado la etapa de reconocimiento de fonemas del sistema, *vowels recognition*. Con estas pruebas se demuestra tanto la capacidad del sistema para la identificación de los fonemas vocálicos como la eficacia de las neuronas ganadoras, *WTANeuron* en el proceso de identificación.

Cada grabación de una vocal se ha repetido 10 veces y se ha calculado la tasa de acierto tanto a la salida del bloque de reconocimiento de fonemas vocálicos, *VowelNeuronSET*, como a la salida del bloque de neuronas ganadoras, *VowelWTAsSET*. En las siguientes tablas se muestran los resultados de identificación de cada vocal, para uno de los hablantes. En cada tabla, se muestra el número total de eventos *AER* emitidos tras la pronunciación de la vocal, así como el número de eventos AER para cada una de las direcciones asociadas a una vocal. Hay que resaltar la tasa de aciertos del fonema vocálico /e/, donde se consigue, con el procedimiento *winner-take-all*, una mejora del 41% al 99%; y en el fonema vocálico /o/, del 79% al 98%.

Este proceso descrito para este hablante, cuyos resultados se muestran desde la Tabla 25 hasta la Tabla 34, se ha repetido para cada uno de los siete hablantes que han intervenido en el experimento. En la Tabla 35 se resume la tasa de acierto de la etapa de reconocimiento de fonemas vocálicos del sistema, para todos los hablantes.

En la Tabla 25 se muestran los resultados obtenidos para la identificación del fonema vocálico /a/ pronunciada por el hablante 1. Se observa que el mayor número de eventos que se obtienen corresponde al fonema vocálico /a/, lográndose una tasa de acierto de reconocimiento del 96% en las 10 iteraciones. Una vez aplicado el procedimiento *winner-take-all* se obtienen los resultados que se muestran en la Tabla 26; aunque el número de eventos para el fonema vocálico /a/ es menor, comparado con los resultados mostrados en la Tabla 25, debido a los mecanismos de excitación e inhibición de las neuronas ganadoras, *WTANeuron,* se consigue una tasa de acierto del 100%. Se han eliminado todos los eventos *AER* correspondientes al fonema vocálico /u/.

Tabla 25. Resultados del reconocimiento del fonema /a/ pronunciada por el hablante 1.

Número de Iteración	**/a/**	**/e/**	**/i/**	**/o/**	**/u/**	**Total de eventos AER**	**Tasa de acierto (%)**
1	617	0	8	0	15	640	96,00
2	615	0	8	0	15	638	96,00
3	617	0	8	0	15	640	96,00
4	615	0	8	0	15	638	96,00
5	617	0	8	0	15	640	96,00
6	615	0	8	0	14	637	96,00
7	614	0	8	0	15	637	96,00
8	618	0	8	0	15	641	96,00
9	615	0	8	0	15	638	96,00
10	618	0	8	0	15	641	96,00

Tabla 26. Resultados del reconocimiento del fonema /a/, después de la fase de *winner-take-all,* pronunciada por el hablante 1.

Número de Iteración	**/a/**	**/e/**	**/i/**	**/o/**	**/u/**	**Total de eventos AER**	**Tasa de acierto (%)**
1	460	0	0	0	0	460	100,00
2	458	0	0	0	0	458	100,00
3	464	0	0	0	0	464	100,00
4	453	0	0	0	0	453	100,00
5	452	0	0	0	0	452	100,00
6	459	0	0	0	0	459	100,00
7	455	0	0	0	0	455	100,00
8	462	0	0	0	0	462	100,00
9	458	0	0	0	0	458	100,00
10	457	0	0	0	0	457	100,00

En la Tabla 27 se muestran los resultados obtenidos para la identificación del fonema vocálico /e/ pronunciada por el hablante 1. La tasa de acierto en este caso es inferior al 50%, debido a la aparición de eventos *AER* correspondiente al fonema vocálico /i/. El sistema, así configurado, tiene dificultad para discriminar entre el fonema vocálico /e/ e /i/ correspondientes al hablante 1. Sin embargo, al aplicarse el procedimiento *winner-take-all* se obtiene los resultados que se muestran en la Tabla 28, una tasa de acierto del 99% en todos los casos. El sistema ha conseguido anular eventos cuyas direcciones no se correspondían con el de la vocal /e/.

Tabla 27. Resultados del reconocimiento del fonema /e/ pronunciada por el hablante 1.

Número de Iteración	/a/	/e/	/i/	/o/	/u/	Total de eventos AER	Tasa de acierto (%)
1	0	164	216	0	0	380	43,00
2	0	161	222	0	0	383	42,00
3	0	163	222	0	0	385	42,00
4	0	164	221	0	0	385	42,00
5	0	161	218	0	0	379	42,00
6	0	163	221	0	0	384	42,00
7	0	162	219	0	0	381	42,00
8	0	164	220	0	0	384	42,00
9	0	162	219	0	0	381	42,00
10	0	160	222	0	0	382	41,00

Tabla 28. Resultados del reconocimiento del fonema /e/, después de la fase de *winner-take-all,* pronunciada por el hablante 1.

Número de Iteración	/a/	/e/	/i/	/o/	/u/	Total de eventos AER	Tasa de acierto (%)
1	0	99	1	0	0	100	99,0
2	0	95	1	0	0	96	99,0
3	0	95	1	0	0	96	99,0
4	0	100	1	0	0	101	99,0
5	0	97	1	0	0	98	99,0
6	0	96	1	0	0	97	99,0
7	0	97	1	0	0	98	99,0
8	0	99	1	0	0	100	99,0
9	0	97	1	0	0	98	99,0
10	0	97	1	0	0	98	99,0

En el proceso de identificación del fonema vocálico /i/, el sistema logra una tasa de acierto del 100%, como se observa en la Tabla 29 y la Tabla 30. De estos resultados solo hay que resaltar la disminución de eventos *AER* a la salida del bloque

de neuronas *WTANeuron*, debido a los mecanismos de excitación e inhibición que las caracterizan.

Tabla 29. Resultados del reconocimiento del fonema /i/ pronunciada por el hablante 1.

Número de Iteración	**/a/**	**/e/**	**/i/**	**/o/**	**/u/**	**Total de eventos AER**	**Tasa de acierto (%)**
1	0	0	234	0	0	234	100,00
2	0	0	232	0	0	232	100,00
3	0	0	231	0	0	231	100,00
4	0	0	228	0	0	228	100,00
5	0	0	230	0	0	230	100,00
6	0	0	228	0	0	228	100,00
7	0	0	232	0	0	232	100,00
8	0	0	230	0	0	230	100,00
9	0	0	230	0	0	230	100,00
10	0	0	233	0	0	233	100,00

Tabla 30. Resultados del reconocimiento del fonema /i/, después de la fase de *winner-take-all,* pronunciada por el hablante 1.

Número de Iteración	**/a/**	**/e/**	**/i/**	**/o/**	**/u/**	**Total de eventos AER**	**Tasa de acierto (%)**
1	0	0	83	0	0	83	100,00
2	0	0	83	0	0	83	100,00
3	0	0	82	0	0	82	100,00
4	0	0	83	0	0	83	100,00
5	0	0	82	0	0	82	100,00
6	0	0	83	0	0	83	100,00
7	0	0	83	0	0	83	100,00
8	0	0	83	0	0	83	100,00
9	0	0	83	0	0	83	100,00
10	0	0	83	0	0	83	100,00

Tabla 31. Resultados del reconocimiento del fonema /o/ pronunciada por el hablante 1

Número de Iteración	/a/	/e/	/i/	/o/	/u/	Total de eventos AER	Tasa de acierto (%)
1	0	0	0	321	90	411	78,00
2	0	0	0	622	155	777	80,00
3	0	0	0	624	153	777	80,00
4	0	0	0	622	155	777	80,00
5	0	0	0	626	156	782	80,00
6	0	0	0	620	155	775	80,00
7	0	0	0	624	157	781	79,00
8	0	0	0	621	155	776	80,00
9	0	0	0	624	154	778	80,00
10	0	0	0	624	158	782	79,00

Tabla 32. Resultados del reconocimiento del fonema /o/, después de la fase de *winner-take-all,* pronunciada por el hablante 1.

Número de Iteración	/a/	/e/	/i/	/o/	/u/	Total de eventos AER	Tasa de acierto (%)
1	0	0	0	440	9	449	98,0
2	0	0	0	442	9	451	98,0
3	0	0	0	438	9	447	98,0
4	0	0	0	442	9	451	98,0
5	0	0	0	436	8	444	98,2
6	0	0	0	442	9	451	98,0
7	0	0	0	443	9	452	98,0
8	0	0	0	435	9	444	98,0
9	0	0	0	441	9	450	98,0
10	0	0	0	439	9	448	98,0

En la Tabla 31 se muestra la tasa de acierto para el reconocimiento de la vocal /o/ pronunciada por el hablante 1; siendo en las 10 iteraciones superior al 78%. En ella, se observa que junto a eventos *AER* asociados al fonema vocálico /o/ aparecen

eventos *AER* correspondientes al fonema vocálico /u/, aunque en una menor cantidad. Gracias al bloque de neuronas ganadoras se consiguen reducir estos eventos, obteniéndose una tasa de acierto del 98%, como se observa en la Tabla 32.

Tabla 33. Resultados del reconocimiento del fonema /u/ pronunciada por el hablante 1.

Número de Iteración	**/a/**	**/e/**	**/i/**	**/o/**	**/u/**	**Total de eventos AER**	**Tasa de acierto (%)**
1	0	0	0	0	74	74	100,00
2	0	0	0	0	75	75	100,00
3	0	0	0	0	76	76	100,00
4	0	0	0	0	76	76	100,00
5	0	0	0	0	76	76	100,00
6	0	0	0	0	76	76	100,00
7	0	0	0	0	76	76	100,00
8	0	0	0	0	75	75	100,00
9	0	0	0	0	74	74	100,00
10	0	0	0	0	77	77	100,00

Tabla 34. Resultados del reconocimiento del fonema /u/, después de la fase *winner-take-all,* pronunciada por el hablante 1.

Número de Iteración	**/a/**	**/e/**	**/i/**	**/o/**	**/u/**	**Total de eventos AER**	**Tasa de acierto (%)**
1	0	0	0	0	38	38	100,00
2	0	0	0	0	38	38	100,00
3	0	0	0	0	39	39	100,00
4	0	0	0	0	39	39	100,00
5	0	0	0	0	38	38	100,00
6	0	0	0	0	38	38	100,00
7	0	0	0	0	38	38	100,00
8	0	0	0	0	37	37	100,00
9	0	0	0	0	38	38	100,00
10	0	0	0	0	38	38	100,00

En el proceso de identificación del fonema vocálico /u/, al igual que para el fonema vocálico /i/, el sistema logra una tasa de acierto del 100%, como se observa en la Tabla 33 y la Tabla 34. De nuevo solo hay que resaltar la disminución de eventos *AER* a la salida del bloque de neuronas *WTANeuron*, debido a los mecanismos de excitación e inhibición.

De los resultados obtenidos se puede deducir que el sistema implementado es capaz de identificar los 5 fonemas vocálicos de la lengua española para el hablante 1, a partir del cual se han configurado todos los parámetros del mismo. Hay que resaltar, de nuevo, la eficiencia del procedimiento *winner-take-all*, con el que se ha conseguido mejorar la tasa de aciertos en los casos de los fonemas vocálicos /e/ y /o/. También, hay que recordar que este bloque de *WTANeuron* ha sido configurado para formar una red de neuronas que implementan una competición blanda, como se describe en el capítulo 7. Es por esto, por lo que después de la etapa de *winner-take-all* aparecen eventos *AER* que no se corresponden con la vocal identificada. En este sentido, sería interesante estudiar en un futuro los resultados al cambiar la implementación por una competición dura, donde una vez que una neurona dispara, no se va a permitir que ninguna otra emita un pulso, porque van a recibir un pulso inhibidor de la neurona ganadora.

Este proceso descrito para el hablante 1 se ha repetido para cada uno de los siete hablantes que han intervenido en el experimento. En la Tabla 35 se muestra la tasa de acierto promedio obtenida por el sistema, después de analizar para cada uno de los hablantes las 10 iteraciones por cada vocal grabada.

Tabla 35. Resultados (tasa de acierto %) del reconocimiento de vocales, después de la fase de *winner-take-all,* para todos los hablantes del experimento.

Número de hablante	**/a/**	**/e/**	**/i/**	**/o/**	**/u/**
1	100	99.50	100	98	100
2	100	95.70	100	96.70	100
3	100	0	0	86.70	100
4	91.10	96.60	100	98.30	100
5	98.90	86	100	97.70	95.50
6	100	65.10	100	89	100
7	100	100	100	88.90	100

Se observa que la tasa de acierto es superior al 65% en todos los casos excepto para el hablante 3, para el cual el sistema no es capaz de discriminar entre los fonemas vocálicos /e/ e /i/. Estos resultados demuestran que la sintonización del sistema es válida para la gran mayoría de los hablantes analizados. Para construir un sistema de reconocimiento de voz totalmente independiente del hablante se propone una nueva línea de trabajo futuro en la que se trate de implementar un sistema capaz de adaptarse a las características de la voz de una hablante cualquiera.

En la Tabla 36 se resume el porcentaje de acierto y de fallo del sistema obtenido a partir del análisis de los resultados de este experimento. Para cada señal de estímulo correspondiente a un fonema vocálico, se detalla la tasa de acierto promedio para todos los hablantes así como la tasa de fallos. Para el fonema vocálico /a/ el sistema es capaz de identificarlo en el 98,57% de los casos, independientemente del hablante y falla el 1,43% de las veces, identificando el fonema vocálico /u/ en lugar de la vocal /a/; para el fonema vocálico /e/, el sistema lo identifica con una tasa de acierto del 77,56% y falla el 1,17% porque reconoce una /i/ o bien se equivoca el 21,17% porque identifica una /u/ en lugar de una /e/; para el fonema vocálico /i/ el sistema acierta con un 100%; para el fonema vocálico /o/, el sistema lo reconoce en un 93,61% de los casos y falla el 6,37% de la veces, porque lo confunde con el

fonema /u/ y el 0.01% porque identifica la vocal /a/ en su lugar; y por último, para el fonema vocálico /u/ el sistema es capaz de identificarlo con una tasa de acierto del 99,36%, sólo en un 0,64% fallará al identificar una /o/ en lugar de la vocal /u/. En este experimento se ha obtenido una tasa de acierto global del 93,82%.

Tabla 36. Tasa de acierto (%) y tasa de fallo (%) de la etapa de reconocimiento de vocales del sistema.

	/a/	**/e/**	**/i/**	**/o/**	**/u/**
Estímulo /a/	98,57	0,00	0,00	0,00	1,43
Estímulo /e/	0,00	77,56	1,17	0,00	21,27
Estímulo /i/	0,00	0,00	100,00	0,00	0,00
Estímulo /o/	0,01	0,00	0,00	93,61	6,37
Estímulo /u/	0,00	0,00	0,00	0,64	99,36
Tasa de acierto global	**93,82**				
Tasa de fallo global	**6,18**				

Esta misma información se presenta gráficamente en la siguiente Figura 103. En las cinco primeras gráficas se muestra el porcentaje de acierto y de fallo del sistema, para cada uno de los fonemas vocálicos usados como estímulos del sistema. En la última grafica se representa la tasa de acierto global del sistema del 93,82% frente a la tasa de fallos del 6,18%.

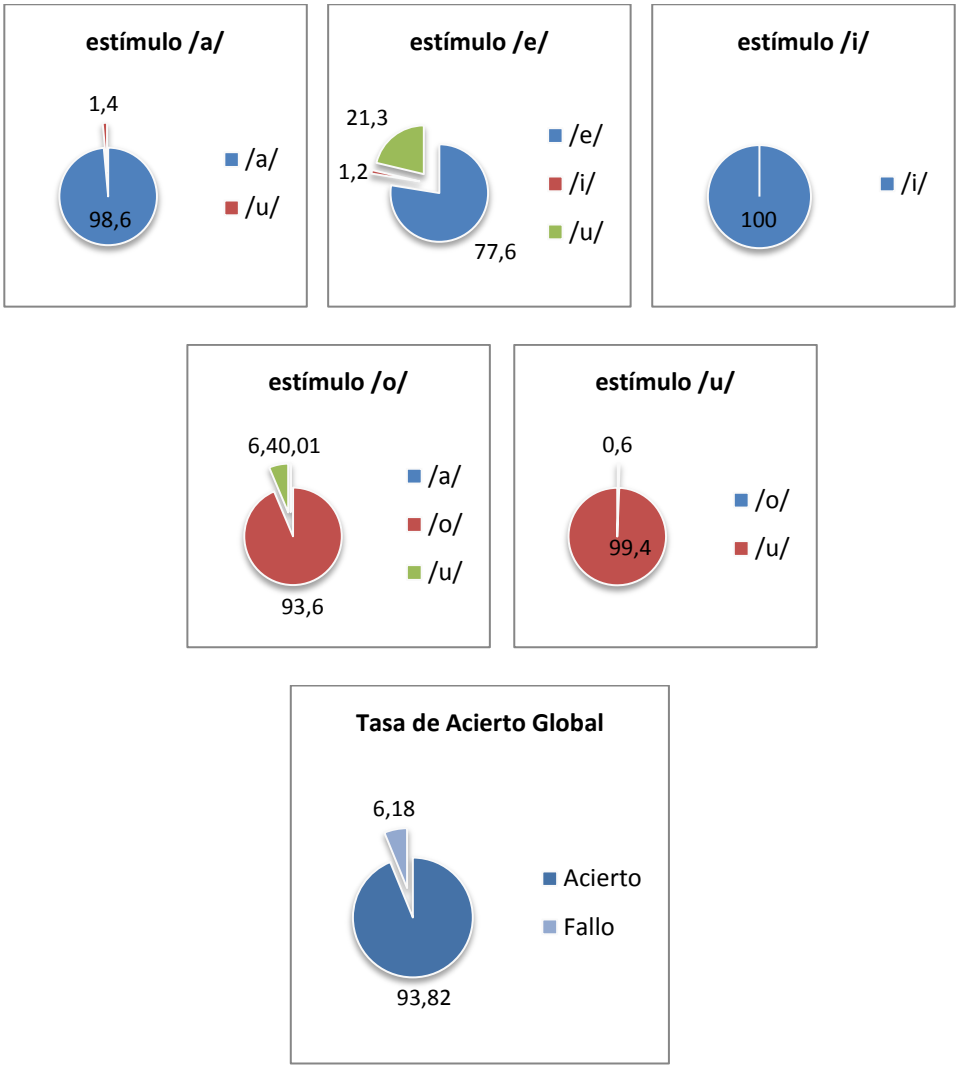


Figura 103. Tasa de acierto global del sistema.

8.2.2. Experimento con palabras

En este experimento se ha probado la etapa de reconocimiento de palabras del sistema, *words recognition*. Con estas pruebas se demuestra la capacidad del sistema para

el reconocimiento de un conjunto reducido de 10 palabras bisílabas de la lengua española.

Al igual que el experimento de las vocales, cada grabación de una palabra se ha repetido 10 veces y se ha calculado la tasa de acierto del sistema. En la Tabla 37 se presenta el resumen de los resultados obtenidos para los siete hablantes que han intervenido en los experimentos.

Tabla 37. Resultados (tasa de acierto %) del reconocimiento de palabras para todos los hablantes del experimento.

	Hablante 1	**Hablante 2**	**Hablante 3**	**Hablante 4**	**Hablante 5**	**Hablante 6**	**Hablante 7**
CERO	100	55.7	0	100	0	100	0
DIQUE	100	100	0	78.3	100	100	100
FASE	89.6	14.9	0	72.3	0	73.7	64.0
GOMA	88.1	98.8	93.3	93.5	73.3	63.3	79.1
LETRA	82.3	100	0	0	55.8	100	0
MUSA	56.7	45.6	0	20	5.2	0.6	5.8
NUBE	100	100	0	97.3	100	100	100
PINO	75.0	97.5	0	0	95	0	0
PODA	89.0	91.6	65	86.8	21.4	60.3	90.4
RIMA	71.5	73.8	0	87.5	12.3	0	2

Para analizar estos resultados hay que considerar que el proceso de reconocimiento de palabras se ha basado en la identificación previa de los fonemas vocálicos que integran cada palabra. Es decir, se ha partido de la información que nos proporcionan las formantes F1 y F2 de cada fonema vocálico. Por lo tanto, no se ha tenido en cuenta la posible variación de estas frecuencias debido fundamentalmente al contexto de las consonantes, las cuales acompañan a las vocales.

En los resultados se refleja esta influencia del contexto consonántico, explicado en el capítulo 3; existe una mayor dificultad de reconocimiento en aquellas palabras que contienen una consonante oclusiva sorda (/p/), la cual va a provocar una mayor presencia de frecuencias bajas alterando así el valor esperado de las frecuencias F1 y sobre todo F2 de la vocal considerada. De igual forma, se ven afectada las palabras con una consonante nasal sonora (/n/) y vibrante sonora (/r/). Sin embargo, aquellas palabras que contienen una consonante fricativa sorda (/s/) van a mostrar una mayor presencia de frecuencias altas, alterando por tanto los resultados. La influencia de los fonemas consonánticos y su inclusión en el sistema conforman una nueva línea de trabajo futuro; ésta tendrá el objetivo de obtener un sistema capaz de identificar palabras a partir de la identificación de todos los fonemas de una lengua, y no solo de los fonemas vocálicos.

También, hay que destacar que en las primeras pruebas de este experimento se obtuvo un bajo porcentaje de acierto del hablante 5, debido a que su velocidad de habla es diferente respecto de los otros hablantes del experimento. Esta diferencia provocó que los eventos correspondientes a la segunda sílaba se generaran antes del tiempo esperado. En este caso, este factor de velocidad impidió el solape en el tiempo de los eventos retrasados asociados a la primera sílaba y de los eventos asociados a la segunda sílaba, y por tanto la palabra no pudo ser identificada. Para este hablante, se ha cambiado el tiempo de retraso configurado de la neurona *delayNeuron*, para adaptarlo a su velocidad, obteniendo así unos mejores resultados, que son los que se muestran en la Tabla 37. Debido a la variabilidad que se produce en el habla, sobre todo por la diversidad de hablantes, se plantea como una nueva línea de trabajo futuro la adaptación automática del sistema a la velocidad de cada hablante.

Por último, resaltar que los mejores resultados se encuentran para el hablante 1, a partir del cual han sido configurados todos los elementos del sistema. Para este hablante, el sistema es capaz de reconocer todas las palabras de la prueba con una

tasa superior al 56% en todos los casos. En este punto, también se puede considerar la idea de ampliar el sistema, como una nueva línea de trabajo, con el fin de que los diferentes parámetros de configuración del sistema se cambien automáticamente en función de las características de los diferentes hablantes que usen el sistema.

Capítulo 9

Conclusiones y trabajos futuros

9.1. Resumen de aportaciones

Las aportaciones que se enumeran en este apartado, resumen los logros más relevantes que se han alcanzado en esta tesis, los cuales cumplen los objetivos definidos en este trabajo. Con esta tesis se han aportado nuevas razones para demostrar que es posible la construcción de un sistema complejo de reconocimiento del habla basado en el paradigma neuromórfico, usando exclusivamente el esquema de comunicación neuromórfico AER. Y además, se ha demostrado la viabilidad de implementar estos sistemas neuromórficos sobre una plataforma hardware como son las FPGAs, prescindiendo del uso del computador convencional. A continuación se enumeran las aportaciones de este trabajo:

- Se ha descrito e implementado una nueva cóclea artificial neuromórfica basada en filtros digitales y en generadores de pulsos, la cual es capaz de separar las

diferentes componentes frecuenciales de una señal de audio compleja. Esta cóclea convierte la señal sonora en pulsos, que se van a transmitir usando el protocolo de comunicación neuromórfico AER.

- Se ha descrito e implementado una nueva neurona, llamada *RNeuron*, capaz de reconocer patrones, a partir de una información representada en pulsos. Esta celda ha sido usada para reconocer los distintos fonemas vocálicos, así como la identificación de cada una de las palabras.

- Se ha descrito e implementado una nueva neurona, llamada *WTANeuron*, que modela la función excitadora-inhibidora, presente en las redes de neuronas biológicas.

- Se ha descrito e implementado una nueva neurona, llamada *delayNeuron*, que permite modelar cadenas de retrasos de pulsos para conseguir la simultaneidad en el tiempo de sucesos temporalmente separados, que facilitará la identificación de patrones temporales, como son secuencias de fonemas que forman palabras; secuencias de palabras que forman frases, etc.

- Basado en las tres neuronas anteriores, se ha descrito un mecanismo de generalización que permite la construcción de un sistema capaz de identificar un conjunto arbitrario de fonemas, palabras y frases. En este punto hay que destacar que, en las diferentes etapas del sistema, cada una de las neuronas *RNeuron* y *WTANeuron* trabajan de forma individual a partir de la misma información, ofreciendo de este modo un procesamiento paralelo.

- Toda la información usada por los diferentes elementos del sistema está codificada en frecuencia de pulsos y ha sido transmitida usando el protocolo de comunicación neuromórfico AER.

- La implementación hardware de todos los elementos que integran el sistema se ha realizado usando exclusivamente un dispositivo digital, una FPGA; no se ha

usado un computador convencional. Esta plataforma digital garantiza la construcción de un sistema de bajo coste, bajo consumo y capaz de realizar un procesado paralelo y en tiempo real.

- Los resultados obtenidos en las pruebas y experimentos con datos reales, demuestran la viabilidad del sistema de reconocimiento del habla propuesto, el cual está basado en el paradigma neuromórfico.

9.2. Conclusiones

Con el trabajo desarrollado en esta tesis, se ha puesto de manifiesto el gran potencial que tienen los sistemas neuromórficos para el procesado de la información sensorial. Se ha iniciado el camino en este nuevo enfoque neuromórfico implementado en sistemas digitales (FPGAs) sobre el proceso de reconocimiento del habla, y en general sobre el procesado de audio. Con los diferentes elementos presentados en este trabajo, se ha construido una base que abre la posibilidad de construir nuevos sistemas neuromórficos complejos de procesado de audio, con el objetivo de imitar la estructura y el funcionamiento del sistema auditivo humano y, de este modo, investigar y aprender más sobre el procesado que se realiza en el cerebro del sonido; principal objetivo de la ingeniería neuromórfica.

En el siguiente apartado se hace un guiño a las futuras investigaciones que, a partir de los conocimientos de ingeniería e inspiradas en la biología, contribuyan al desarrollo de nuevos sistemas neuromórficos para el procesado de la señal de audio en particular, o de cualquier información sensorial en general.

9.3. Trabajos futuros

- Se ha expuesto en este documento, que la frecuencia fundamental de la señal de voz, llamada F0, es una característica que permite discriminar a un hablante. Se propone como trabajo futuro introducir esta característica en el sistema para identificar a un hablante.

- Se ha considerado que el primer y segundo formante de la señal de voz son las claves acústicas definitivas para la descripción y clasificación de las vocales. Sin embargo, también se ha descrito que cada tracto bocal origina modelos espectrales y distribuciones formánticas diferentes, originando que un mismo fonema vocálico puede tener distintas representaciones acústicas debido a la diversidad de los hablantes. Por lo tanto, se plantea ampliar el sistema propuesto añadiendo nuevas *RNeuron* para la identificación del mismo fonema. Esto añade al sistema información redundante que mejorará la eficiencia del proceso de identificación de fonemas, al permitir un reconocimiento independiente del hablante.

- Para construir un sistema de reconocimiento de voz totalmente independiente del hablante se propone una nueva línea de trabajo futuro en la que se trate de implementar un sistema capaz de adaptarse a las características de la voz de un hablante cualquiera.

- Estudiar en un futuro los resultados del sistema al cambiar la implementación de la neurona *WTANeuron* por una competición dura, donde una vez que una neurona dispara, no se va a permitir que ninguna otra emita un pulso, porque van a recibir un pulso inhibidor de la neurona ganadora.

- Incluir el mecanismo de aprendizaje en las neuronas *RNeuron* y *WTANeuron*. Este aprendizaje modificaría el peso de los patrones permitiendo que las neuronas aprendan fonemas, palabras o frases.

- Dado que el contexto del discurso modifica la velocidad del mismo, un desarrollo futuro sería que la neurona *delayNeuron* adaptará automáticamente el tiempo de retraso a la velocidad del discurso.

- La influencia de los fonemas consonánticos y su inclusión en el sistema conforman una nueva línea de trabajo futuro; ésta tendrá el objetivo de obtener un sistema capaz de identificar palabras a partir de la identificación de todos los fonemas de una lengua, y no solo de los fonemas vocálicos. Por tanto, se propone ampliar el sistema para que identifique todos los fonemas de la lengua española o de cualquier otro idioma.

- Dado que el sistema auditivo biológico se basa en la información que le aportan dos cócleas, una línea de trabajo futuro sería incluir esta información redundante, aumentando así la efectividad del procesamiento del sistema.

- Teniendo en cuenta la información de dos cócleas es posible realizar un procesamiento estéreo para la localización de la fuente de un sonido.

Anexos

Scripts de Matlab

10.1. Script MaxBand

Se realiza un barrido en frecuencias en el rango 200 Hz – 12 kHz. Para ello, se emite un conjunto de señales seno; la frecuencia de cada señal seno coincide con la frecuencia central de cada banda de la cóclea. Se obtiene el número máximo de eventos *AER* emitidos por cada una de las 21 bandas de la cóclea artificial pulsante, para cada una de las señales seno. Se representa en una figura el valor de la banda que más eventos emite para cada una de las señales seno.

```
%inicio parámetros de la figura
bandas = [0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20];
frec = {'300' '455' '570' '700' '845' '1000' '1175' '1375' '1600'
'1860' '2160' '2510' '2925' '3425' '4050' '4850' '5850' '7050'
'8600' '10750'  };
figure;
title('Histogram');
xlabel('Bandas');
ylabel('Num Events')
hold on;
```

```
%fin parametros de la figura


%defino las frecuencias de las señales seno
%en función de las frecuencias de las bandas de la cóclea
f = [200 400 510 630 770 920 1080 1270 1480 1720 2000 2320 2700
3150 3700 4400 5300 6400 7700 9500 12000];

f1 = [ f(2)-f(1) f(3)-f(2) f(4)-f(3) f(5)-f(4) f(6)-f(5) f(7)-
f(6) f(8)-f(7) f(9)-f(8) f(10)-f(9) f(11)-f(10) f(12)-f(11)
f(13)-f(12) f(14)-f(13) f(15)-f(14) f(16)-f(15) f(17)-f(16)
f(18)-f(17) f(19)-f(18)  f(20)-f(19) f(21)-f(20) 1000];
%el ultimo elemento 1000 para que salga la frecuencia 12500

f1 = f + f1/2;

Suma = zeros(19,21);

%genero las señales seno y obtengo la salida del sistema con el
script NumEvents.
for i=1:1:21
       f1(i)
       sound(sin(2*pi*f1(i)*(0:1/48000:2)),48000,16);
       NumEvents
       pause(1); %1 segundo

       [sumaEv maxSuma maxBanda] = sumaAER(inaddr);

       Suma(i,:) = sumaEv; %para representar en la figura

       filename=sprintf('PCoc21_1ch_%d.mat',  f1(i));
       filename = [here '\EXP1\' filename ]
       save (filename, 'inaddr', 'ints', 'sumaEv', 'maxSuma',
       'maxBanda');%, 'totalDiscreteValues','f1', 'absFFT', 'f');
end

plot( bandas, Suma)
legend (frec);
```

10.2. Script AllBandAllF

Se realiza un barrido en frecuencias en el rango 200 Hz – 12 kHz. Para ello, se emite un conjunto de señales seno; la frecuencia de cada señal seno coincide con la frecuencia central de cada banda de la cóclea. Se obtiene el número máximo de

eventos *AER* emitidos por cada una de las 21 bandas de la cóclea artificial pulsante, para cada una de las señales seno, al igual que en el script *MaxBand.m*. En este caso, la representación de la figura es distinta, en función de la frecuencia de las señales seno se representan las bandas activas.

```
bandas = {'0' '1' '2' '3' '4' '5' '6' '7' '8' '9' '10' '11' '12'
'13' '14' '15' '16' '17' '18' '19' '20' };

f = [200 400 510 630 770 920 1080 1270 1480 1720 2000 2320 2700
3150 3700 4400 5300 6400 7700 9500 12000];
frec = [ f(2)-f(1) f(3)-f(2) f(4)-f(3) f(5)-f(4) f(6)-f(5) f(7)-
f(6) f(8)-f(7) f(9)-f(8) f(10)-f(9) f(11)-f(10) f(12)-f(11)
f(13)-f(12) f(14)-f(13) f(15)-f(14) f(16)-f(15) f(17)-f(16)
f(18)-f(17) f(19)-f(18)  f(20)-f(19) f(21)-f(20) 1000];

frec = f + frec/2;
maxfrec = numel(frec);

for i=1:1:maxfrec
        sound(sin(2*pi*frec(i)*(0:1/48000:2)),48000,16);
        NumEvents;
        pause(1); %1 segundo
        [sumaEv maxSuma maxBanda] = sumaAER(inaddr);
    for banda=1:1:21
        EvBanda(i,banda) = sumaEv(banda)
    end
end

figure;
xlabel('Freq(Hz)');
ylabel('Num Events')
hold on;

plot( frec,EvBanda, '-o')
legend (bandas);
```

10.3. Script NumEvents

Lee una secuencia de eventos *AER* capturados por la placa *USBAERmini2*. Cada evento *AER* se caracteriza por una dirección y un instante de tiempo.

```
inaddr=[];
ints=[];
sumaEv = [];

%Configuro la placa
monitortime=3;
usb0 = factory.getInterface(0);

if (isempty(usb0))

       usb0=net.sf.jaer.hardwareinterface.usb.cypressfx2.CypressFX
       2MonitorSequencer.instance.getFirstAvailableInterface;
end

if ~usb0.isOpen()
        usb0.open
end

usb0.setOperationMode(1)

%se inicia la captura de eventos AER
usb0.setEventAcquisitionEnabled(true)
tic
while toc<monitortime

inpacket=usb0.acquireAvailableEventsFromDriver.getPrunedCopy()
    ints=[ints; inpacket.getTimestamps()];
    inaddr=[inaddr; inpacket.getAddresses()]

end

inpacket=usb0.stopMonitoringSequencing();
inaddr=[inaddr; inpacket.getAddresses()];
ints=[ints; inpacket.getTimestamps()];
```

10.4. Script *sumaAER*

Dada una secuencia de diferentes valores, obtiene el número de veces que se repite cada valor. A partir de esta información muestra el valor que se repite más veces y su número de ocurrencias. Esta función es usada para conocer la dirección *AER* que se emite un mayor número de veces.

```
function [suma, maxi, maxbanda]=sumaAER(mat)

lm=sort(mat);
lm=lm';

numel(lm);
suma = zeros(1,21);
Meventos = [1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1];


j=1;
for i=0:1:numel(suma)-1
    fin = 0;
    while j<=numel(lm) & fin==0
        if lm(j)==(i)
           suma(i+1)=suma(i+1)+1;
           j=j+1;
        else
            fin = 1;
            i+1;
            suma(i+1)=suma(i+1)/Meventos(i+1);
        end
    end
end

[maxi maxbanda]=max(suma);
maxi
maxbanda =maxbanda-1

end
```

10.5. Script SoundVowels

Función encargada de la grabación de cada una de las vocales, así como de distintas secuencias de vocales. A esta función se le pasa como parámetro un valor entero, que se usará como identificador del hablante. Se genera un fichero *wav* para cada grabación. Se ha usado la frecuencia de muestreo de 11025 Hz.

```
function SoundVowels(i)
       Fs = 11025
       here=fileparts(mfilename('fullpath'));
```

```
        'Diga la vocal a (pulse enter)'
        pause
        'inicio'
        ya = wavrecord(2*Fs, Fs, 'int16');
        'fin'
        wavplay(ya,Fs);
        filename=sprintf('%d_a.wav',  i);
        filename = [here '\EXPVowels\' filename ]
        wavwrite(ya, Fs, filename);


        'Diga la vocal e (pulse enter)'
        pause
        'inicio'
        ye = wavrecord(2*Fs, Fs, 'int16');
        'fin'

        wavplay(ye,Fs);
        filename=sprintf('%d_e.wav',  i);
        filename = [here '\EXPVowels\' filename ]
        wavwrite(ye, Fs, filename);

        %......
        El código que sigue es similar. Se repite para cada vocal
        que se quiera grabar.
        %......

end
```

10.6. Script SoundWords

Función encargada de la grabación de cada una de las palabras usadas en los experimentos. A esta función se le pasa como parámetro un valor entero, que se usará como identificador del hablante. Se genera un fichero *wav* para cada grabación. Se ha usado la frecuencia de muestreo de 11025 Hz.

```
function SoundWords(i)
        Fs = 11025
        here=fileparts(mfilename('fullpath'));

        'Diga la palabra RIMA (pulse enter)'
        pause
        'inicio'
```

```
        y1 = wavrecord(3*Fs, Fs, 'int16');
        'fin'

        wavplay(y1,Fs);
        filename=sprintf('%d_RIMA.wav',  i);
        filename = [here '\EXPWords\' filename ]
        wavwrite(y1, Fs, filename);

        'Diga la palabra LETRA (pulse enter)'
        pause
        'inicio'
        y2 = wavrecord(3*Fs, Fs, 'int16');
        'fin'

        wavplay(y2,Fs);
        filename=sprintf('%d_LETRA.wav',  i);
        filename = [here '\EXPWords\' filename ]
        wavwrite(y2, Fs, filename);
        %......
        El código que sigue es similar. Se repite para cada palabra
        que se quiera grabar.
        %......

end
```

10.7. Script *Exp1*

Este script ha sido usado para analizar la respuesta de cada una de las etapas del sistema a partir del estímulo de una vocal. Para cada una de las grabaciones de una vocal se obtiene el número total de eventos emitidos por cada banda/neurona del sistema. Este procedimiento se ha repetido 10 veces para cada vocal; y para cada uno de los hablantes que han participado en los experimentos. El siguiente trozo de código se repite para cada uno de los hablantes.

```
vocales = {'a', 'e' ,'i' ,'o', 'u'};


for j=1:1:5  %5 vocales
  for i=1:1:10 %10 iteraciones por cada vocal
        filename = strcat('1_',vocales(j), '.wav') %hablante 1

        [y, Fs] = wavread(char(filename));
```

```
        wavplay(y, Fs, 'async');

        NumEvents;
        pause(1); %1 segundo
        [sumaEv maxSuma maxBanda] = sumaAER(inaddr);

         for ivocal=1:2:10
       indice =i+(j-1)*10;
             EvVocal1(indice,ivocal) = sumaEv(ivocal);
         end
    end
end

%......
El código que sigue es similar. Se repite para cada hablante
%......
```

10.8. Script *Exp2*

Este script ha sido usado para analizar la respuesta de cada una de las etapas del sistema a partir del estímulo de una palabra. Para cada una de las grabaciones de una palabra se obtiene el número total de eventos emitidos por cada banda/neurona del sistema. Este procedimiento se ha repetido 10 veces para cada palabra; y para cada uno de los hablantes que han participado en los experimentos. El siguiente trozo de código se repite para cada uno de los hablantes.

```
word = {'CERO', 'DIQUE', 'FASE', 'GOMA', 'LETRA', 'MUSA', 'NUBE',
'PINO', 'PODA', 'RIMA' };

for j=1:1:10 %10 palabras
       for i=1:1:10 %10 iteraciones por cada palabra

        filename = strcat('1_',word(j), '.wav') %hablante 1

        [y, Fs] = wavread(char(filename));
        wavplay(y, Fs, 'async');

        NumEvents;
        pause(2); %1 segundo
        [sumaEv maxSuma maxBanda] = sumaAER(inaddr);
```

```
          for iword=1:1:11
             indice =i+(j-1)*10;
             EvWord11(indice,iword) = sumaEv(iword);
          end
    end
end

%......
El código que sigue es similar. Se repite para cada hablante
%......
```

Bibliografía y referencias

0'Shaughnessy, D. (1987). *Speech Communication: Human and Machine*. Massachusetts: Addison Wesley Pub. Comp.

Abdelatty Ali, A. M., der Spiegel, J. V., & Mueller, P. (2002). Robust Auditory-Based Speech Processing Using the Average Localized Synchrony Detection. *IEEE Transactions on Speech and Audio Processing*, *10*(5), 279-292.

Altera. (2010). Cyclone IV Device Datasheet, *3*, 1-44. Retrieved from http://www.altera.com/literature/hb/cyclone-iv/cyiv-53001.pdf

Alvarez-López, M. A. (2004). *Reconocimiento de voz sobre diccionarios reducidos usando modelos ocultos de Markov*. Universidad Nacional de Colombia.

Andreou, A. G., Meitzler, R. C., Strohbehn, K., & Boahen, K. A. (1995). Analog VLSI neuromorphic image acquisition and pre-processing systems. *Neural Networks*, *8*(7-8), 1323-1347. doi:10.1016/0893-6080(95)00098-4

Argyrakis, P., Hamilton, A., Webb, B., Zhang, Y., Gonos, T., & Cheung, R. (2007). Fabrication and characterization of a wind sensor for integration with neuron circuit. *Microelectronic Engineering*, *84*, 1749-1753.

Avis, C., Shihab, S., Giacomo, I., & Tim, H. (2001). Report on Telluride 2001.

Baum, L. E. (1972). An inequality and associated maximation techniques in statistical estimation for probabilistic functions of Markov processes. *Inequalitiies* (Academic P., pp. 1-8).

Becanovic, V., Hosseiny, R., & Indiveri, G. (2004). Object tracking using multiple neuromorphic vision sensors. In D. Nardi, M. Riedmiller, C. Sammut, & J. Santos-Victor (Eds.), *RoboCup 2004: Robot Soccer World Cup VIII* (pp. 426-433). Berlin: Springer.

Bengio, Y. (1993). A connectionist approach to speech recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, *7*(4), 647-667.

Berner, R. (2006). *Highspeed USB2.0 AER Interfaces*.

Boahen, K. a. (2000). Point-to-point connectivity between neuromorphic chips using address events. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, *47*(5), 416-434. doi:10.1109/82.842110

Borzone de Manrique, M. I. (1979). On the recognition of isolated Spanish vowels. In H. Hollien & P. Hollien (Eds.), *Current Issues in the Phonetic Sciences* (pp. 677-681). Amsterdam.

Bothe, H., Persson, M., Biel, L., & Rosenholm, M. (1999). Multivariate sensor fusion by a neural network model. *Fusion'99 Proceedings* (pp. 1094-1011).

Bourland, H. (1995). Towards increasing speech recognition error rates. *Eurospeech*, 883-894.

Bourland, H., & Morgan, N. (1993). Continuous speech recognition by connectionist statistical methods. *IEEE transactions on neural networks*, *4*(6), 893-909.

Bourland, H., & Morgan, N. (1994). *Connectionist speech recognition - A hybrid approach*. (Kluwer Academic, Ed.).

Békésy, G. von. (1960). *Experiments in Hearing*. New York: McGraw-Hill Inc.

CappoCacciaWorkshop. (2012). The 2012 CapoCaccia Cognitive Neuromorphic Engineering Workshop. Retrieved from http://capocaccia.ethz.ch/capo/wiki/2012

Casacuberta, F., & Vidal, E. (1987). *Reconocimiento Automático del Habla*. (Marcombo, Ed.) (1st ed.).

Caviar Project. (2006). Convolution Address-Event-Representation (AER) Vision Architecture for Real-Time. Retrieved from http://www2.imse-cnm.csic.es/caviar/

Chan, V. Y., Jin, C. T., & Schaik, A. V. (2010). Adaptive sound localization with a silicon cochlea pair. *Frontiers in neuroscience*, 196. doi:10.3389/fnins.2010.00196

Chan, V., Jin, C. T., & v. Schaik, A. (2012). Neuromorphic audio-visual sensor fusion on a sound-localizing robot. *Frontiers in neuroscience*, *6*(February), 21. doi:10.3389/fnins.2012.00021

Chistovich, L. A., Sheikin, R. L., & Lublinskaja, V. V. (1979). Centres of gravity and spectral peaks as the determinants of vowel quality. In L. y Öhman (Ed.),

Frontiers of Speech Communication Research (pp. 143-157). Londres: Academic Press.

Clarke, C., Qiang, L., Peremans, H., & Hernandez, A. (2004). FPGA implementation of a neuromimetic cochlea for a bionic bat head. In J. Becker (Ed.), *Field-programmable logic and applications* (pp. 1073-1075). Berlin.

Cole, R. A. (1997). *Survey of the State of the Art in Human Language Technology*. Cambridge University Press.

Dau, T. (1996). *Modeling auditory processing of amplitude modulation*. Universidad de Oldenburg.

Davis, K. H., Biddulph, R., & Balashek, S. (1952). Automatic Recognition of Spoken Digits. *The Journal of the Acoustic Society of America*, *24*(6), 637-642.

Delbruck, T., & Lichtsteiner, P. (2007). Fast sensory motor control based on event-based hybrid neuromorphic-procedural system. *IEEE International Symposium on Circuits and Systems*, (ISCAS), 845-848. Ieee. doi:10.1109/ISCAS.2007.378038

Delbruck, T., & Mead, C. A. (1996). *Analog VLSI phototransduction by continuous-time, adaptive, logarithmic photoreceptor circuits*. Pasadena, California.

Deller, J. R., Proakis, J. G., & Hansen, J. H. (1993). *Discrete-time processing of speech signals*. Macmillan publishing company New York.

Domínguez-Morales, M., Jimenez-Fernandez, A., Paz, R., López-Torres, M. R., Cerezuela-Escudero, E., Linares-Barranco, A., Jimenez-Moreno, G., et al. (2011). An Approach to Distance Estimation with Stereo Vision Using

Address-Event-Representation. In B.-L. Lu, L. Zhang, & J. Kwok (Eds.), *Neural Information Processing* (Vol. 7062, pp. 190-198). Springer Berlin Heidelberg. doi:10.1007/978-3-642-24955-6_23

Dugast, C., Devillers, L., & Aubert., X. (1994). Combining TDNN and HMM in a hybrid system for improved continuousspeech recognition. *IEEE Transactions on Speech and Audio Processing*, *2*(1), 217-223.

Duggirala, V., Studebaker, G. A., Pavlovic, C., & Sherbecoe, R. L. (1988). Frequency Importance Functions for a Feature Recognition Test Material. *The journal of the Acoustical Society of America*, *83*(6), 2372-2382.

Dundur, R., Latte, M. V., Kulkarni, S. Y., & Venkatesha, M. K. (2008). Digital Filter for Cochlear Implant Implemented on a Field-Programmable Gate Array. *PWASET*, *33*, 468-472.

Ewe, C. T., Cheung, P. Y. K., & Constantinides, G. A. (2004). Dual Fixed-Point: An Efficient Alternative to Floating-Point Computation. In S. V. (Eds.). J. Becker, M. Platzner (Ed.), *Field Programmable Logic and Application* (pp. 200-208). Berlin, Heidelberg: Springer-Verlag.

Fastl, H., & Zwicker, E. (2007). *Phycho-acoustics. Facts and models.* (S.-V. Berlín & H. G. & C. K, Eds.)*Springer Series in Information Sciences* (third edit.).

Fern, V. (1997). Antecedentes y desarrollo de los sistemas actuales de reconocimiento, *1*, 321-346.

Fernández Planas, A. M. (1993). Estudio del campo de dispersión de las vocales castellanas. *Estudios de Fonética Experimental, V*, 129-145.

Flanagan, J. L. (1972). *Speech Analysis, Synthesis and Perception.* (K. und K. in Einzeldarstellungen, Ed.) (2nd ed.).

Fletcher, H. (1940). Auditory patterns. *Reviews of modern physics*, *12*, 47-65.

Forney, G. (1973). The viterbi algorithm. *Proceedings of the IEEE*, *61*(3), 268-278.

Gambin, I., Grech, I., Casha, O., Gatt, E., & Micallef, J. (2010). Digital cochlea model implementation using Xilinx XC3S500E Spartan-3E FPGA. *International Conference on Electronics, Circuits and Systems (ICECS)* (pp. 946-949).

Ghitza, O. (1994). Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, *2*(1), 115-132.

Ghulam, M., Horikawa, J., & Nitta, T. (2006). A pitch-synchronous peak-amplitude based feature extraction method for noise robust ASR. *ICASSP*, 505-508.

Gil Loyzaga, P. (2005). *Estructura y función de la corteza auditiva. Bases de la vía auditiva ascendente.* (E. Salesa, E. Perelló, & A. Bonavida, Eds.) (pp. 23-38). Barcelona: Masson.

Glasberg, B., & Moore, B. (1990). *Derivation of auditory filter shapes from notched noise data.* Hearing Res.

Glover, M., Hamilton, A., & Smith, L. S. (2002). Analogue VLSI leaky integrated-and-fire neurons and their use in a sound analysis system. *Analog Integrated Circuits and Signal Processing*, *30*(2), 91-100.

Gold, B., & Morgan, N. (2000). *Speech and Audio Signal Processing*. John Wiley and sons.

Graham, D. W. (2006). *A biologically inspired front end for audio signal processing using programmable analog circuitry*. Georgia Institute of Technology.

Greenberg, S. (1996). Auditory processing of speech. In Lass (Ed.), *Principles of Experimental Phonetics* (pp. 362-407). Sant Louis: Mosby.

Gómez-Rodríguez, F., Linares-Barranco, A., Miró, L., Liu, S.-C., v. Schaik, A., Etienne-Cummings, R., & Lewis, M. A. (2007). AER Auditory Filtering and CPG for Robot Control. *2007 IEEE International Symposium on Circuits and Systems* (pp. 1201-1204). Ieee. doi:10.1109/ISCAS.2007.378268

Gómez-Rodríguez, F., Paz, R., Miro, L., & Linares-Barranco, A. (2005). Two Hardware Implementations of the Exhaustive Synthetic AER Generation Method, 534-540.

Halberstom, B., & Raphael, L. J. (2004). Vowel normalization: the role of fundamental frequency and upper formants. *Journal of Phonetics2*, *32*(3), 423-434.

Handel, S. (1993). *Listening. An Introduction to the Perception of Auditory Events*. Cambridge, Massachusets: The MIT Press.

Harrison, R. R., & Koch, C. (1998). An analog VLSI model of the fly elementary motion detector. *Advances in Neural Information Processing Systems*, *10*, 880-886.

Helmoltz, H. F. von. (1874). *Théorie Physiologique de la Musique*. (M. G., Ed.). Paris.

Hermansky, H., Morgan, N., Bayya, A., & Kohn, P. (1992). Rasta-plp speech analysis technique. *IEEE*, 121-124.

Higgins, C. M., & Koch, C. (1999). Multi-Chip Neuromorphic Motion Processing. In D. Wills & S. DeWeerth (Eds.), *Proc. 20th Anniversary Conference on Advanced Research in VLSI* (pp. 309-323). Los Alamitos: IEEE Computer Society.

Hubbard, A. (1993). A travelling-wave amplifier model of the cochlea. *Science*, *259*, 68-71.

Häfliger, P. (2007). CAVIAR Hardware Interface Standards , Version 2.01.

IEEE. (2008). VHDL. Retrieved from http://www.vhdl.org/

INE. (2012). The Institute of Neuromorphic Engineering. Retrieved from www.ine-web.org

Indiveri, G. (1998). Analogue VLSI model of locust DCMD neuron response for computation of object aproach. *Neuromorphic Systems: engineering silicon from neurobiology*, 47-60.

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, *64*(4), 532-556.

Jimenez-Fernandez, Angel, Lujan-Martinez, C., Paz-Viecente, R., Jimenez, G., & Civit, A. (2009). From Vision Sensor to Actuators , Spike Based Robot Control through Address-Event-Representation. *Lecture Note in Computer Science*, *2*, 797-804.

Johnson, K. (1997). *Acoustic and Auditory Phonetics* (2nd ed.). Londres.

Juang, B., & Rabiner, L. R. (2006). Speech Recognition, Automatic: History. In K. Brown (Ed.), *Encyclopedia of Language & Lingistics (2ª edition)* (pp. 806-819). doi:http://dx.doi.org/10.1016/B0-08-044854-2/00906-8

Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (2000). *Principles of Neural Science* (4th ed.). New York: McGraw-Hill.

Kates, J. (1991). A Time-Domain Digital Cochlear Model. *IEEE Signal Processing Magazine*, *39*(12), 2573-2592.

Kates, J. (1993). Accurate Tuning Curves in a Cochlear Model. *IEEE, Speech and Audio Proc.*, *1*(4), 453-462.

Katsiamis, A., Drakakis, E., & Lyon, R. F. (2007). Practical Gammatone-Like Filters for Auditory Processing. *URASIP Journal on Audio, Speech, and Music Processing*, *2007*.

Kewley-Port, D. (1995). Thresholds for formant-frequency discrimination of vowels in consonantal context. *The Journal of the Acoustic Society of America*, *97*(5), 3139-3146.

Kim, D.-S., Lee, S.-Y., & Kil, R. M. (1999). Auditory Processing of Speech Signals for Robust Speech Recognition in Real-World Noisy Environments. *Audio*, *7*(1), 55-69.

Koickal, T. J., Hamilton, A., Tan, S. L., Covington, J. A., Gardner, J. W., & Pearce, T. C. (2005). Analog VLSI circuit implementation. of an adaptive neuromorphic olfaction chip. *In IEEE International Symposium on Circuits and Systems (ISCAS)*, *54*, 60-73.

Landercy, A., & Renard, R. (1977). *Éléments de phonétique*. Mons - Bruxelles: Centre International de Phonétique Appliquée - Didier.

Lazzaro, J. (1991). Biologically-based Auditory Signal Processing in Analog VLSI. *Signals, Systems and Computers*, *2*, 790-794. doi:10.1109/ACSSC.1991.186555

Lazzaro, J., & Mead, C. A. (1989a). Circuit models of sensory transduction in the cochlea. *In Analog VLSI implementations of neural networks*, 85-101.

Lazzaro, J., & Mead, C. A. (1989b). A silicon model of auditory localization. *Neural computation*, *1*, 47-57.

Lazzaro, J., & Mead, C. A. (1989c). Silicon modelling of pitch perception. *Proceedings of the National Academy of Sciences of the United States of America*, *86*, 9597-9601.

Lazzaro, J., & Wawrzynek, J. (1997). Speech recognition experiments with silicon auditory models. *Analog Integrated Circuits and Signal Processing*, *13*, 37-51.

Leong, M. P., & Jin, C. T. (2003). An FPGA-Based Electronic Cochlea. *EURASIP Journal on Applied Signal Processing*, 629-638.

Li, C., Delbruck, T., & Liu, S.-C. (2012). Real-Time Speaker Identification using the AEREAR2 Event- Based Silicon Cochlea. *IEEE International Symposium on Circuits and Systems*, 1159-1162.

Lichtsteiner, Patrick, Posch, C., & Delbruck, T. (2008). A 128x128 120 dB 15 us Latency Asynchronous Temporal Contrast Vision Sensor. *IEEE Journal of Solid-State Circuits*, *43*(2), 566-576. doi:10.1109/JSSC.2007.914337

Lieberman, P., & Blumstein, S. E. (1988). *Speech physiology, speech perception and acoustic phonetics*. Cambridge: Cambridge University Press.

Lim, S. C., Temple, A. R., & Jones, S. (1997). VHDL-based design of biologically inspired detection system. *IEEE international Conference on Neural Network* (pp. 922-927). Houston, USA.

Linares-Barranco, A., Gomez-Rodriguez, F., Jimenez-Fernandez, A., Delbruck, T., & Lichtensteiner, P. (2007). Using FPGA for visuo-motor control with a silicon retina and a humanoid robot. *2007 IEEE International Symposium on Circuits and Systems* (pp. 1192-1195). Ieee. doi:10.1109/ISCAS.2007.378265

Lippmann, R. P. (1989). Pattern Clasification using neural networks. *IEEE Communications Magazine*, *27*(11), 47-54.

Lippmann., R. P. (1989). Review of neural networks for speech recognition. *Neural Computation*, *1*(1), 1-38.

Liu, S.-C., & Delbruck, T. (2010). Neuromorphic sensory systems. *Current opinion in neurobiology*, *20*(3), 288-95. Elsevier Ltd. doi:10.1016/j.conb.2010.03.007

Liu, S.-C., v. Schaik, A., Minch, B. A., & Delbruck, T. (2010). Event-Based 64-Channel Binaural Silicon Cochlea with Q Enhancement Mechanisms. *IEEE Int Symp Circuits Syst*, 2027-2030.

Liu, W., Andreou, A. G., & Goldstein, M. H. (1993a). Analog cochlear model for multiresolution speech analysis. *Advances in Neural Information Processing Systems*, *5*, 666-673.

Liu, W., Andreou, A. G., & Goldstein, M. H. (1993b). Voiced speech representation by an analog silicon model of the auditory periphery. *IEEE Trans. Neural Networks*, *3*(3), 477-487.

Lyon, R. F. (1982). A computational model of filtering, detection, and compression in the cochlea. *International Conference on Acoustics, Speech and Signal Processing*, 1282-1285.

Lyon, R. F. (1996). The All-Pole Gammatone Filter and Auditory Models. *Apple Computer, Inc.*

Lyon, R. F., & Mead, C. A. (1988). An analog electronic cochlea. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *36*(7), 1119-1134. doi:10.1109/29.1639

López Bascuas, L. E. (1997). La percepción del habla: problemas y restricciones computacionales. *Anuario de Psicología*, 3-19.

Mahowald, M. (1992). *VLSI Analogs of Neuronal Visual Processing : A Synthesis of Form and Function. Technology*. California Institute of Technology.

Mahowald, M. (1993). The Address-Event Representation Communcation Protocol. AER 0.02.

Mahowald, M. (1994). *An analog VLSI system for stereoscopic vision* (1st ed.). Boston: Klumber Academic Publisher.

Makhoul, J. (1973). Spectral Analysis of Speech by Linear Prediction. *Ieee Transactions On Audio and Electroacoustics*, (3), 140-148.

Marrero Aguiar, V., & Martín, Y. (2001). Discriminación auditivas de los rasgos distintivos acústicos en palabras aisladas: oídos normales y patológicos. In J.

Díaz García (Ed.), *Actas del II Congreso de Fonética Experimental* (pp. 258-266). Sevilla.

Marrero Aguiar, V., Santos, A., & Cárdenas, M. R. (1993). Feature Discrimination And Pure Tone Audiometry. In R. Aulanko & Y. Korpijaakko-Huuhka (Eds.), *Proceedings of the Third Congress Of The International Clinical Phonetics And Linguistics Association* (pp. 121-128). Helsinki.

Martínez Celdrán, E. (1995). En torno a las vocales del español: análisis y reconocimiento. *Estudios de Fonética Experimental*, *7*, 195-218.

MathWorks. (2012). Filter Design HDL Coder. Retrieved from http://www.mathworks.es/products/filterhdl/

McEwan, A., & v. Schaik, A. (2003). An analogue VLSI implementation of the Meddis inner hair cell model. *EURASIP Journal on Advances in Signal Processing*, *7*, 639-648.

Mead, C. A. (1990). Neuromorphic electronic systems. *Proceedings of the IEEE*, *78*(10), 1629-1636. doi:10.1109/5.58356

Mead, C. A., & Mahowald, M. (1988). A silicon model of early visual processing. *Neural Networks*, *1*(1), 91-97. doi:10.1016/0893-6080(88)90024-X

Meddis, R. (1986). Simulations of mechanical to neural transduction in the auditory receptor. *Journal of the Acoustical Society of America*, *79*(3), 702-711.

Meddis, R. (1988). Simulation of auditory-neural transduction: Futher studies. *Journal of the Acoustical Society of America*, *83*, 1059-1063.

Meddis, R., Hewitt, M., & Shackleton, T. (1990). Implementation details of a computational model of the inner hair-cell/auditory-auditory nerve synapse. *Journal of the Acoustical Society of America*, *87*, 1813-1816.

Mermelstein, P. (1978). Difference limens for formant frequencies on steady-state and consonant-bound formants. *Journa/ of the Acoustic Society of America*, *63*, 572-580.

Microelectronics, W. (2009). WM8731/WM8731L datasheet. *Audio*, (April).

Mishra, A., & Hubbard, A. E. (2002). A Cochlear Filter Implemented With a Field-Programmable Gate Array, *49*(1), 54-60.

Morgado Estévez, A. (2003). *Analisis y modelado de sistemas pulsantes bioinspirados basados en buses de altas prestaciones: Bus AER*. Universidad de Cádiz.

Mugliette, C., Grech, I., Casha, O., Gatt, E., & Micallef, J. (2011). FPGA active digital cochlea model. *International Conference on Electronics, Circuits and Systems (ICECS)* (pp. 699-702).

Neto, J., Almeida, L., Hochberg, M., Martins, C., Nunes, L., Renals, S., & Robinson, T. (1995). Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system. *Eurospeech* (pp. 2171-2174).

Oster, M., & Liu, S.-C. (2005). Spiking Inputs to a Winner-take-all Network. *NIPS*.

Pamies, A., & Fernández Planas, A. M. (2006). Sobre la percepción de la duración vocálica en español. In J. Durán (Ed.), *Actas del V Congreso Andaluz de Lingüística general* (pp. 501-512). Granada.

Patterson, R., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., & Allerhand, M. (1992). Complex Sounds and Auditory Images. In Y. Cazals, L. Demany, & K. Horner (Eds.), *Auditory Phisiology and Perception* (pp. 429-446). Pergamon Press.

Pearson, M., Nibouche, M., Gilhespy, I., Gurney, K., Melhuish, C., Mitchison, B., & Pipe, A. G. (2006). A hardware based implementation of a tactile sensory system for neuromorphic signal processing applications. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, *4*.

Peterson, L. C., & Bogert, B. P. (1950). A dynamical theory of the cochlea. *Journal of the Acousticcal Society of America*, *22*, 369-381.

Quilis, A., & Esgueva, M. (1983). Realización de los fonemas vocálicos españoles en posición fonética normal. In M. Esgueva & M. Cantarero (Eds.), *Estudios de fonética I* (pp. 137-252). Madrid: Consejo Superior de Investigaciones Científicas.

Quilis, A., & Fernández, J. A. (1985). *Curso de Fonética y Fonología Españolas, para estudiantes Angloamericanos* (11th ed.). Collectanea Phonetica, Consejo Superior de Investigaciones Científicas, Instituto de Filología, Madrid.

Rabiner, L. R., & Juang, B. H. (1986). An introduction to hidden Markov models. *ASSP Magazine, IEEE Publication*, *3*(1), 4-16.

Rabiner, L. R., & Juang, B. H. (1993). *Fundamentals of speech recognition*. Prentice Hall PTR.

Ravishankar, M. K. (1996). *Efficient algorithms for speech recognition*. Carnegie Mellon University.

Renals, S., McKelvie, D., & McInnes, F. (1991). A comparative study of continuous speech recognition using neural networks and hidden Markov models. *ICASSP* (pp. 369-372).

Robinson, T. (1994). An application of recurrent nets to phone probability estimation. *IEEE Transactions on Neural Networks*, *5*(2), 298-305.

Romero, J. (1989). Campos de dispersión auditivos de las vocales del castellano. Percepción de las vocales. *Estudios de Fonética Experimental, III*, 181-206.

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, *65*(6), 386-408.

Roy, D. (2006). Design and developmental metrics of a "skin-like" multi-input quasicompliant robotic gripper sensor using tactile matrix. *Journal of Intelligent and Robotic Systems*, *46*(4), 305-337.

Ruggero, M. A., Rich, N. C., & Robles, L. (1990). Middle-ear response in the chinchilla and its relationship to mechanics at the base of the cochlea. *Journal of the Acoustical Society of America*, *87*(4), 1612-1629.

Sakoe, H., Isotani, R., Yoshida, K., Iso, K., & Watanabe, T. (1989). Speaker-independent word recognition using dynamic programming neural networks. *ICASSP* (pp. 29-32).

Schauer, C., & Paschke, P. (1999). A spike-based model of binaural sound localization. *International Journal Neural System*, *9*(5), 447-452.

Seneff, S. (1986). A computational model for the peripheral auditory system: Application to speech recognition research. *ICASSP*, 1983-1986.

Seneff, S. (1988). A joint Synchrony/Mean-Rate Model of Auditory Speech Processing. *Journal of Phonetics*, *16*, 55-76.

Serrano-Gotarredona, R., Oster, M., Lichtsteiner, P., Linares-Barranco, A., Paz-Vicente, R., Gomez-Rodriguez, F., Camunas-Mesa, L., et al. (2009). CAVIAR: a 45k neuron, 5M synapse, 12G connects/s AER hardware sensory-processing- learning-actuating system for high-speed visual object recognition and tracking. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, *20*(9), 1417-38. doi:10.1109/TNN.2009.2023653

Serrano-Gotarredona, R., Oster, M., Lichtsteiner, P., Linares-Barranco, A., Paz-Vicente, R., Gomez-Rodriguez, F., Riis, H. K., et al. (2005). AER Building Blocks for Multi-Layer Multi-Chip Neuromorphic Vision Systems. *Advances in Neural Information Processing Systems*, (1).

Singer, E., & Lippmann., R. P. (1992). A speech recognizer using radial basis function neural networks in an HMM framework. *ICASSP* (pp. 629-632).

Sivilotti, M. (1991). *Wiring considerations in Analog VLSI Systems, with application to Field-Programmable Networks*. Cal. Inst of Tech.

Slaney, M. (1993). *An efficient implementation of the Petterson-Holdsworth auditory filter bank*.

Stevens, S., Volkman, J., & Newman, E. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, *8*, 185-190.

TellurideWorkshop. (2012). Telluride Neuromorphic Cognition Engineering Workshop 2012. Retrieved from http://ine-web.org/telluride-conference-2012/telluride-2012/index.html

Thibodeau, G. (1998). *Estructura y función del cuerpo humano* (pp. 180-183). Harcourt Brace.

Traunmüller, H. (1987). Phase vowels. In Schouten (Ed.), *The Psychophysics of Speech Perception* (pp. 377-384). Dordrecht: Martinus Nijhoff Pub.

Varela Rincón, J., & Loaiza Pulgarín, J. (2008). *Reconocimiento de palabras aisladas mediante redes neuronales sobre FPGA. Annals of Physics*. MIT Press.

Vasarhelyi, G., Adam, M., Vazsonyi, E., Kis, A., Barsony, I., & Ducso, C. (2006). Characterization of an integrable single-crystalline 3-d tactile sensor. *IEEE Sensors journal*, *6*(4), 928-934.

Velasco-Medina, J., & Hernan-Meza Escobar, J. (2007). Arquitecturas Hardware para una Cóclea Electrónica usando FPGAs. In Uninorte (Ed.), *XII Simposio de Tratamiento de Señales, Imágenes y Visión Artificial*. Colombia.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. (1989). Phoneme recognition using time delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *37*(3), 328-339.

Watts, L. (1993). *Cochlear Mechanics : Analysis and Analog VLSI*. California Institute of Technology.

Watts, L., Kerns, D. A., Lyon, R. F., & Mead, C. A. (1992). Improved Implementation of the Silicon Cochlea. *IEEE Journal of Solid-State Circuits*, *27*(5).

Wong, C. K., & Leong, P. H. W. (2006). An FPGA-based electronic cochlea with dual fixed-point arithmetic. *Proc. International Conference on Field Programmable Logic and Applications (FPL)* (pp. 205-210).

Wong, W. K., Neoh, T. M., Loo, C. H., & Ong, P. C. (2008). Bayesian fusion of auditory and visual spatial cues during fixation and saccade in humanoid robot. *15th International Conference on Advances in Neuro-Information Processing* (pp. 1103-1109).

Xing, X. (2000). *Characterization and redesign of an electronic cochlea chip*. Boston University.

Yang, Z., Murray, A. F., Woergoetter, F., Cameron, K. L., & Boonsobhak, V. (2006). A neuromorphic depth-from-motion vision model with stdp adaptation. *IEEE transactions on neural networks*, *17(2)*, 482-495.

Yu, T., Schwartz, A., Harris, J., Slaney, M., & Liu, S.-C. (2009). Periodicity detection and localization using spike timing from the AER EAR. *IEEE International Symposium on Circuits and Systems*, 109-112. doi:10.1109/ISCAS.2009.5117697

Zwicker, E. (1961). Subdivision of the Audible Frequency Range into Critical Bands. *Journal of the Acoustical Society of America*, *33*(2), 248.

v. Schaik, A., & Liu, S.-C. (2005). AER EAR : A Matched Silicon Cochlea Pair with Address Event Representation Interface. *IEEE International Symposium on Circuits and Systems*, 4213-4216.

v. Schaik, A., & Vittoz, E. (1997). A silicon model of amplitude modulation detection in the auditory brainstem. *Advances in NIPS* (pp. 741-747). MIT Press.

v. Schaik, A., Fragnière, E., & Vittoz, E. (1995). Improved Silicon Cochlea using Compatible Lateral Bipolar Transistors. *NIPS*, (November), 28-30.

v. Schaik, A., Fragnière, E., & Vittoz, E. (1996). An analogue electronic model of ventral cochlear nucleus neurons. *In 5th International Conference on Microelectronics for Neural Networks and Fuzzy Systems (MicroNeuro '96)*.