Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Ciudad de México

Presentado por:

Alicia Gabriela Ortiz Acosta

Víctor Hugo Arellano Rodríguez

Extracción de parámetros y reconocimiento de voz esofágica utilizando transformada wavelet y redes neuronales

Presentado el día: 16 de Noviembre del 2007

Asesor:

Dr. Alfredo Mantilla Caeiros

Profesor:

M. en C. Edgar Omar López Caudana

Sinodales:

Dr. Jorge Brieva Rico

Dr. Ricardo Fernández del Busto y Ezeta

Índice

Introducción **1**

Objetivos **5**

Justificación **6**

Capitulo 1: Antecedentes y marco teórico **7**

1.1 Anatomía del aparato fonador y auditivo

1.1.1 Anatomía del aparato fonador

1.1.2 Órgano principal de generación de voz

1.1.3 Fonética Acústica

1.1.4 Anatomía del sistema auditivo humano

1.1.5 Oído externo, oído medio, oído interno

1.1.6 Bandas críticas

1.2 Transformada Wavelet

1.2.1 Análisis en resoluciones múltiples (MRA)

1.2.2 Transformada Wavelet

1.2.3 Transformada Wavelet discreta

1.2.4 Ejemplo de funciones Wavelet utilizadas en procesamiento de voz

1.3 Redes Neuronales

1.3.1 Introducción

1.3.2 Topologías de redes neuronales

1.3.3 Algoritmo de propagación hacia atrás

Capitulo 2: Sistema de análisis de voz y extracción de parámetros.....25

2.1 Esquema general

2.2 Sistema de análisis de voz

2.2.1 Acondicionamiento

2.2.2 Segmentación

2.2.3 Detección de segmentos vocalizados

2.3 Extracción de parámetros utilizando wavelet

2.3.1 Construcción de un Wavelet madre a partir de un modelo del oído interno

2.3.2 Muestreo del plano escala - traslación

Capitulo 3: Sistema de análisis y extracción de parámetros..... 32

3.1 Algoritmo de la red neuronal

3.1.1 El perceptrón

3.1.2 Algoritmo de aprendizaje

3.1.3 Vector de parámetros, clasificación y filtrado

3.2 Modificaciones realizadas

Capitulo 4: Resultados..... 39

4.1 Comparación con otras funciones Wavelet

4.1.2 Comparación de vocales según método

4.2 Resultados del sistema

4.2.1 Etapa de acondicionamiento

4.2.2 Etapa de segmentación y detección de segmentos vocalizados

4.2.3 Etapa de extracción de características

4.2.4 Etapa de reconocimiento (Red neuronal)

4.3 Resultados ingresando una señal de voz específica

4.4 Resultados de la etapa de reconocimiento en al personalizar la red neuronal

Conclusiones..... 54

Anexos..... 57

Anexo A: Tabla comparativa y gráfica de resultados

Anexo B: Adaptaciones del código en Matlab

Anexo C: Pruebas de las modificaciones a los parámetros del sistema

Bibliografía.....72

Introducción

La comunicación es un fenómeno indispensable para la relación grupal de los seres vivos por medio del cual obtenemos información acerca de nuestro entorno y otros estilos de vida. Cada ser humano es capaz de compartir dicha información haciendo partícipes a otros de la misma forma, su evolución.

La comunicación humana se da entre dos personas y no necesariamente debe ser verbal, se puede dar también de forma gestual, o escrita, aunque la forma más común de comunicación se da de forma oral. Para que el mecanismo de comunicación funcione adecuadamente se debe tener un emisor que tenga la capacidad de enviar un mensaje a un receptor, es decir, una persona que hable y otra que escuche.

Para producir el habla, el emisor hace uso de su aparato respiratorio y de otros componentes tales como centros nerviosos que se encargan del habla, centros de control respiratorio situados en la corteza cerebral, estructuras de articulación, resonancia dentro de la boca y las cavidades nasales.

La laringe es un elemento físico muy importante para la producción del habla. Sin embargo cada ser humano esta expuesto a contraer infecciones o enfermedades provocadas ya sea por la edad, o por exponernos a entornos químicos a los cuales no estamos acostumbrados. Tal es el caso del cáncer que se presenta como una de las causas de muerte con mayor índice en el mundo. Esta enfermedad, se presenta como un crecimiento irregular de células las cuales van destruyendo tejido y poco a poco extendiéndose a diferentes partes del cuerpo. El cáncer se presenta principalmente en órganos vitales del cuerpo humano, tal como la laringe.

El cáncer de laringe es el segundo cáncer en incidencia del tracto digestivo superior. Cada año en el mundo se diagnostican 136 000 casos de este tipo de cáncer y su porcentaje de supervivencia a 5 años se sabe es del 68%. La primera laringectomía con éxito se llevó en el año de 1873 y desde ese día, estudios de biomédica se han fijado como objetivo ofrecerle una calidad de vida favorable a dichos pacientes.

El cáncer de laringe es una enfermedad cuyos factores de riesgo se deben a alguna afección médica ya existente, edad avanzada (comúnmente entre los 50 y 70 años de edad), obesidad, desnutrición o tabaquismo. Sin embargo este padecimiento tiene muchas formas de curación; desde radiaciones o tratamiento láser, quimioterapia, etc. En el caso en que estos métodos no sean efectivos, se debe proceder a una laringectomía. La laringectomía es un procedimiento quirúrgico que consiste en extirpar la laringe al paciente desde la base de la lengua hasta la tráquea, incluyendo la musculatura endolaríngea con fascia cervical superficial y el hueso hioides, seguido de la creación de una nueva abertura llamada estoma, para que el paciente pueda respirar.

Una de las discapacidades más grandes que trae consigo este procedimiento es la pérdida de la voz debido a que al separar el ducto de aire que permite respirar por medio de la boca, nariz y esófago se impide utilizar el aire proveniente de los pulmones. Así también se pierden las cuerdas vocales y la capacidad de hablar por medio de ellas.

Hoy en día se tienen 3 opciones muy recomendables para restituir la capacidad de comunicación oral en un paciente laringectomizado: voz con prótesis fonatoria, la laringe electrónica y voz esofágica.

El primer método consiste en colocar una válvula en la estoma para permitir que el aire proveniente los pacientes y es difícil de dominar.

El siguiente método hace uso de una laringe electrónica la cual consiste en un dispositivo electrónico que produce voz de forma electromecánica. Dicho dispositivo se coloca en el tracto vocal sustituyendo a la laringe natural y genera una excitación. Dicha excitación consiste en una transducción de las vibraciones mecánicas en eléctricas y se crea la voz electrónica. La desventaja de este método es que la voz producida por el dispositivo tiende a ser monótono y artificial, lo que permita muchas veces no ser entendida al cien por ciento.

El método de voz esofágica tiene su principio fundamental y así mismo la producción de voz en la inyección, succión y deglución de aire desde la cavidad oral hacia el segmento faringoesofágico. Es decir, el paciente es enseñado a tomar aire en su boca y forzarlo hacia el esófago cerrando con la lengua contra

el techo del paladar. Cuando el aire es expulsado por medio de eructos, produce la vibración de las paredes del esófago y de la faringe, produciendo un sonido de tono que es la voz de los laringectomizados. Entonces el paciente articula este sonido grave con la lengua, dientes y paladar, como lo hacían al hablar normalmente.

La ventaja de este método es que no utiliza ningún tipo de aparato o alguna cirugía posterior a la laringectomía, es totalmente natural y se adquiere con entrenamiento. Y como desventaja se tiene que la rehabilitación puede ser larga dependiendo del paciente, no se sabe a ciencia cierta cuanto dura el período de entrenamiento para hablar con voz esofágica.

Por otra parte, la voz esofágica presenta características que pueden dificultar su comprensión. Esto se debe a que las cuerdas vocales no intervienen y como resultado se tiene que algunos fonemas no presentan la claridad debida para su entendimiento, tal es el caso de los fonemas vocálicos.

Para poder resolver dicha problemática se han desarrollado diferentes sistemas de adquisición y reconocimiento de voz esofágica en los cuales se identifican distintas regiones donde la voz es menos inteligible y son reemplazadas por voz sintetizada electrónicamente.

Se necesita obtener las características de cada segmento de voz y clasificar los fonemas obtenidos, de la misma forma trabajar en un algoritmo que permita la síntesis de la voz.

Los fonemas vocálicos son los más difíciles de pronunciar para un laringectomizado debido a que son los mas afectados por la perdida de las cuerdas vocales, y por lo mismo el sistema analizado en este trabajo limita su alcance al reconocimiento de las 5 vocales del español.

De la misma manera se involucraron en el sistema métodos que nos pueden dar mucha información sobre las características de la voz esofágica, tal es el caso de la transformada Wavelet que nos dará información sobre los segmentos vocalizados en tiempo y frecuencia de una señal de voz que sea introducida al sistema.

Al haberse diseñado y utilizado anteriormente estos sistemas, en este proyecto se ha trabajado en el entendimiento del modelo del oído humano en

el cual se basan los algoritmos aquí programados, de la misma manera se han hecho algunas modificaciones al código de cada programa para disminuir su complejidad y sobre todo obtener resultados gráficos que logren darnos mayor información que los que dan los algoritmos originales.

Siguiendo los objetivos planteados para el desarrollo de este proyecto se hicieron comparaciones con otras funciones wavelet para llevar a cabo la validación de dicho algoritmos.

A continuación se presentan todos los fundamentos teóricos necesarios para el desarrollo de este proyecto, al igual que cada modificación realizada y el resultado de la misma. Para finalizar se muestran los resultados obtenidos para cada una de las etapas del sistema de reconocimiento de voz esofágica.

Objetivo General:

1. Validación de algoritmos de extracción de parámetros y reconocimiento de voz esofágica.

Objetivos Específicos:

1. Extracción de parámetros usando una función Wavelet basada en el modelo del oído humano.
2. Comparación de la función Wavelet del oído humano con otras funciones Wavelet.
3. Validación de un algoritmo de reconocimiento basado en redes neuronales.

Justificación

La capacidad de poder comunicarse con una persona por medio del habla es prácticamente indispensable para un ser humano. Al perder la capacidad de hablar, la calidad de vida del ser humano en cuestión se torna mas complicada, sin dejar atrás que se distorsiona también la vida de los seres que la rodean.

Al haberse sometido a una laringectomía, el paciente pierde esa capacidad de hablar, al igual que tiene imposibilidad de llevar aire proveniente de los pulmones hacia la boca.

Se han propuesto distintos métodos para lograr superar dicha problemática, sin embargo la calidad de voz que se tiene en los pacientes no es la más adecuada. La mayoría de estos métodos se basan en la implantación de un dispositivo en la laringe del paciente; un método enteramente natural es de voz esofágica, en cual solo se debe entrenar al paciente para que hable fluyendo aire desde el esófago.

El objetivo de este trabajo consiste en contribuir al desarrollo de un sistema que permita mejorar las características acústicas de una señal de voz esofágica. Permitiendo así, ayudar a que la dicción de cada paciente al cual le fue extirpada la laringe sea de mayor calidad y por consiguiente regresarle una calidad de vida a la cual estaba acostumbrado.

1.1 Anatomía del aparato fonador y auditivo

1.1.1 Anatomía del aparato fonador

A través de la voz, el ser humano es capaz de producir sonidos con diferentes frecuencias. Sin embargo, la información verbal mas certera que podemos obtener se encuentra en un rango de frecuencias de los 500 Hz a los 2.5 KHz.
La voz tiene 3 propiedades fundamentales, estas son:

I) Tono: se refiere al número de veces por segundo que las cuerdas vocales se unen durante la fonación.

II) Intensidad de la voz: Depende de que tan juntas se encuentren las cuerdas vocales entre si, de la cantidad de presión de aire por debajo de la laringe, la frecuencia fundamental de la voz y la resonancia producida en el tracto vocal.

III) Timbre: Esta propiedad nos permite diferenciar entre un sonido y otro; esta determinada principalmente por el contenido armónico y las dinámicas características del sonido.

Ahora bien, se puede explicar de manera sencilla como se lleva a cabo la producción y emisión de sonidos verbales. Estos se deben a la acción o funcionamiento secuenciado, sincronizado y automático de una corriente de aire, un vibrador sonoro, un resonador y articuladores.

Estos cuatro elementos generan los sonidos del habla en el siguiente orden:

I) Los pulmones suministran el aire que atraviesan los bronquios, la tráquea y sincronizan las cuerdas vocales ubicadas en la laringe.

II) El aire sufre una modificación en la caja de resonancia de la nariz, la boca y garganta, en la que se amplifica y se forma el timbre de la voz.

III) Los órganos articuladores van finalmente a modelar esa columna sonora transformándola en fonemas, sílabas y palabras.

Los pasos anteriores, muestran los principales componentes del aparato fonador vinculados con la producción de la voz (ver Figura 1.1.1). Los pulmones simulan una fuente de energía acústica y la corriente de aire que se desplaza por la tráquea es modulada en las cuerdas vocales que vibran haciendo de oscilador.

Los sonidos sordos, o no vocalizados, se producen cuando se cierran y abren abruptamente las cavidades laríngea, bucal y nasal. La configuración del tracto vocal es también muy variable, ya que también son parte de él las articulaciones, la mandíbula, la lengua, los labios, y el velo del paladar. Este último, realiza la función de válvula que controla la comunicación entre el tracto bucal y el nasal.

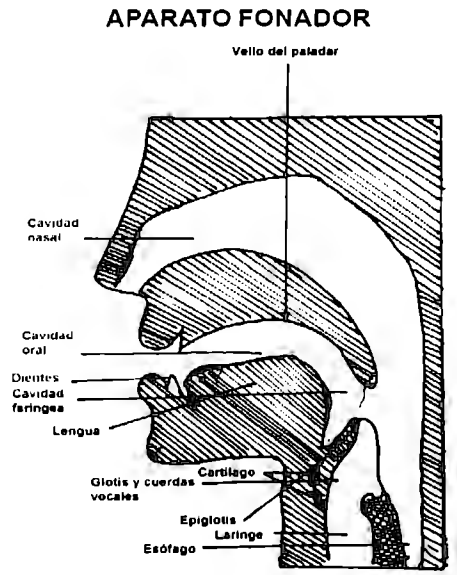


Figura 1.1.1 Aparato fonador humano

1.1.2 Órgano principal de generación de la voz

El órgano principal de la producción de la voz es la laringe, que es también un conducto para el paso del aire. Sus caras laterales están parcialmente cubiertas por la tiroides, que es un cartílago que sobresale de la garganta y tiene la forma de un libro al revés. Detrás de este cartílago se encuentran las cuerdas vocales.

Estas cuerdas vocales estas constituidas por dos repliegues superiores que son las cuerdas falsas o bandas ventriculares y dos repliegues inferiores que son las verdaderas cuerdas vocales.

Los repliegues inferiores son los que producen las primeras características del sonido:

I) Si dichas cuerdas se aproximan y producen vibración se origina un sonido vocalizado, de lo contrario será un sonido no vocalizado.
II) La vibración provoca una onda o tono y unos armónicos que al filtrarlos producen el timbre del sonido.

III) Al pasar el aire hacia las cuerdas vocales con mayor o menor energía se produce la intensidad de voz.

IV) La duración se produce por un impulso psicomotriz a través del nervio recurrente hacia el diafragma.

Estos son los mecanismos fisiológicos que dan lugar a la producción de la voz, mismos que se encuentran regulados y controlados por el sistema nervioso central.

1.1.3 Fonética acústica

Los órganos que intervienen en la articulación del sonido son móviles o fijos. Son móviles los labios, la mandíbula, la lengua y las cuerdas vocales, que reciben el nombre de órganos articulatorios. Son fijos los dientes, los alvéolos, el paladar duro y el paladar blando.
Los sonidos se producen cuando se ponen en contacto dos órganos articulatorios, también cuando se ponen en contacto un órgano fijo y otro articulatorio.

El modo de articulación se determina por la disposición de los órganos móviles en la cavidad bucal y cómo impiden o dejan el libre paso del aire, esta acción puede llevarse a cabo de diversas formas:

I. La interrupción instantánea y completa del paso del aire para las implosivas.
II. Dejar abierto el paso nasal pero interrumpido el oral para las nasales.
III. Producir un contacto con la lengua pero dejar libre el paso del aire a uno y otro lado para las laterales.
IV. Producir una leve interrupción primero y dejar el paso libre después para las africadas.
V. Permitir el paso del aire por un paso estrecho por el que el aire pasa rozando para las fricativas.
VI. Permitir el paso libre del aire por el centro de la lengua sin fricción alguna para las vocales.

Se emiten diferentes clases de vocales según varíe la posición de la lengua, tanto a partir de su eje vertical (alta, media y baja), como a partir de su eje horizontal (anterior, central y posterior). Por ejemplo, en español son vocales altas las vocales [i] y la [u]. Son vocales medias la [e] y la [o] y es vocal baja la [a]. Así, la lengua va de abajo arriba para pronunciar las dos vocales seguidas de la palabra "aire", pero desciende a una posición media para pronunciar su última vocal. Son vocales

anteriores del español la [i] y la [e], las vocales posteriores son la [o] y la [u], la [a] es la vocal central. La lengua se mueve de atrás hacia adelante para emitir las vocales de la palabra "totales".

1.1.4 Anatomía del sistema auditivo

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

I) Captación y procesamiento mecánico de las ondas sonoras.
II) Conversión de la señal acústica (mecánica) en impulsos nerviosos, y transmisión de dichos impulsos hasta los centros sensoriales del cerebro.
III) Procesamiento neural 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, mientras que la etapa de procesamiento neural, en la cual se producen las diversas sensaciones auditivas, se encuentra ubicada en el cerebro. Así pues, se pueden distinguir dos regiones o partes del sistema auditivo: la región periférica, en la cual los estímulos sonoros conservan su carácter original de ondas mecánicas hasta el momento de su conversión en señales electroquímicas y la región central, en la cual se transforman dichas señales en sensaciones e intervienen procesos cognitivos, mediante los cuales se asigna un contexto y un significado a los sonidos.

1.1.5 Oído externo, medio e interno

Oído externo:
La única parte visible del oído es el pabellón auditivo o aurícula que, debido a su especial forma helicoidal, es la primera parte del oído en reaccionar ante el sonido. La aurícula funciona como una especie de embudo que ayuda a dirigir el sonido hacia el interior del oído. Sin la presencia de este embudo las ondas sonoras tomarían una ruta directa hacia el conducto auditivo, esto haría que el proceso de audición fuera difícil e ineficaz ya que gran parte del sonido se perdería y sería más difícil escuchar y comprender los sonidos.

El conducto auditivo, además de proteger el tímpano, actúa como un audífono natural que amplifica automáticamente los sonidos bajos y menos penetrantes de la

voz humana. De este modo, el oído compensa parte de la debilidad de la voz humana y hace más fácil oír y comprender una conversación normal.

El tímpano el cual señala el inicio del oído medio, es extremadamente sensible. Cuenta con 3 capas, la primera es un recubrimiento de piel similar al que tiene el canal auditivo. La segunda es una membrana elástica gracias a la cual el tímpano es capaz de convertir los cambios de presión presentes en el oído externo en vibraciones mecánicas que se transportan al oído medio. Y la tercera consiste en una estructura mucosa consistente con las paredes del oído medio.

Oído medio:
Las vibraciones se transmiten al interior por medio de tres huesos: martillo, yunque y estribo. La ventana oval es una membrana que recubre la entrada a la cóclea en el oído interno. Cuando el tímpano vibra, las ondas sonoras pasan por el martillo y el yunque hacia el estribo y posteriormente hacia la ventana oval.
Cuando las ondas sonoras se transmiten desde el tímpano a la ventana oval, el oído medio funciona como un transformador acústico, amplificando las ondas sonoras antes de que lleguen al oído interno.
La trompa de Eustaquio se encuentra también en el oído medio e iguala la presión del aire a ambos lados del tímpano, garantizando que la presión no se acumule en el oído.

Oído interno:
El oído interno es una intrincada zona de tubos y conductos, conocido como laberinto. En él, se encuentra la cóclea, donde las ondas sonoras se transforman en impulsos eléctricos que se envían al cerebro. El cerebro traduce esos impulsos en sonidos que podemos reconocer y entender.

1.1.6 Bandas críticas

Si el oído es estimulado con un tono puro, las distintas regiones de la membrana basal responderán en diferentes posiciones. Será posible encontrar una sección donde la membrana sufra un desplazamiento máximo; sin embargo, las regiones cercanas a ésta también se verán afectadas por el estímulo.
Si ahora se tiene una señal audible que contiene dos o más frecuencias, la respuesta de la membrana basal será la superposición o suma de los efectos de cada uno de sus componentes. En este caso, cada uno de los componentes

espectrales afectará una posición en particular de la membrana basal y sus alrededores. Si se encuentran demasiado cerca, será imposible distinguirlos.
Por lo anterior, se dice que el oído no es capaz de distinguir las componentes espectrales presentes en un estímulo. Para poder distinguir cada una de esas frecuencias y por lo tanto obtener una mayor comprensión sobre lo que escuchamos es necesario tener una aproximación que nos dé a conocer la resolución aproximada del oído.
El ancho de banda crítico es la mínima diferencia necesaria en la frecuencia de dos tonos para que se puedan distinguir como tonos independientes.

1.2 Transformada Wavelet

1.2.1 Análisis en resoluciones múltiples (MRA)

El análisis en resoluciones múltiples es una técnica que permite analizar señales en múltiples bandas de frecuencia.
El objetivo del análisis en resoluciones múltiples es expandir una señal en una base de funciones cuyas propiedades tiempo-frecuencia se adapten a la estructura de la señal. Un tipo de análisis en resoluciones múltiples, que ha sido utilizado exitosamente en varias aplicaciones es la transformada Wavelet en donde la resolución espectral aumenta a medida que disminuye la frecuencia durante la descomposición, o dicho de otra forma, la resolución temporal aumenta conforme se incrementa la frecuencia de las componentes a identificar en la señal. Al permitir este tipo de variación en las dispersiones temporales y espectrales es posible separar las componentes que tienen una mayor energía en la señal.
Para identificar y extraer las características más representativas de una señal, se inicia el proceso con una secuencia de valores muestreados de una variable física. Estos datos pueden ser el valor promedio de la señal durante un cierto periodo de muestreo. Después el flujo de datos se divide en pequeños segmentos y cada uno de estos es aproximado como el valor promedio de las muestras que lo forman.

1.2.2 Transformada Wavelet

La transformada Wavelet muestra una representación en tiempo - frecuencia, es decir, es capaz de dar información en tiempo y frecuencia simultáneamente, la ventaja de esta representación es que una componente espectral en algún instante de tiempo puede darnos información particularmente interesante para cada estudio.

La transformada Wavelet es eficiente para el análisis de señales no estacionarias y de rápida transitoriedad. Esta transformada, provee análisis de resoluciones múltiples con ventanas dilatadas. El análisis de las frecuencias de mayor rango se realiza usando ventanas angostas y el análisis de las frecuencias de menor rango se hace utilizando ventanas anchas.

La transformada Wavelet de una función $f(t)$ es la descomposición de $f(t)$ en un conjunto de funciones base $\psi_{\sigma,\tau}(\tau)$. La transformada Wavelet se define como:

$$\gamma(\tau,s)=\left\langle f,\psi_{\tau,s}\right\rangle=\int_{-\infty}^{\infty} f(t)\psi_{\tau,s}(t)dt \qquad (1.2.1)$$

En la ecuacion anterior, $\left\langle f,\Psi_{\tau,s}\right\rangle$ se conoce como el producto escalar de f(t) con $\Psi_{\tau,s}$ y se calcula realizando la integral que se muestra.

La transformada wavelet, es definida a través de la familia de funciones ($\psi_{\sigma,\tau}$), a las que se les llaman funciones wavelet hijas, estas son generadas a partir de una función wavelet madre, mediante la traslación y escala.

Entonces, las funciones wavelet hijas son generadas a partir de la traslación y cambio de escala de una misma función wavelet $\psi(t)$, llamada "wavelet madre" y se define como:

$$\psi_{\tau,s}(t)=\frac{1}{\sqrt{s}}\cdot\psi\left(\frac{t-\tau}{s}\right) \qquad (1.2.2)$$

Donde s es el factor de escala y τ es el factor de traslación.

Las funciones wavelet hijas $\psi_{\sigma,\tau}(t)$ generadas de la misma función Wavelet madre $\psi(t)$ tienen diferente escala s y ubicación τ, pero tienen todas la misma forma. Se utilizan siempre factores de escala $s > 0$. Así, cambiando el valor de s se cubren rangos diferentes de frecuencias. Valores grandes de parámetros corresponden a frecuencias de menor rango, o una escala grande de $\psi_{\sigma,\tau(\tau)}$. Valores pequeños de s corresponden a frecuencias de mayor rango o una escala muy pequeña de $\psi_{\sigma,\tau}(t)$.

1.2.3 Transformada Wavelet discreta

Si se desea utilizar la transformada Wavelet como una herramienta para el procesamiento de señales, es necesario definir cierto margen de precisión en los cálculos, es decir, un muestreo en el plano escala – traslación. Al llevarse a cabo este proceso, se necesita conservar íntegramente la información contenida en la transformada Wavelet.

Para este muestreo, la escala debe ser discretizada a espacios que sigan un comportamiento geométrico. Se elige para esto, potencias enteras de 2, que se define como:

$$s_j = 2^j$$

La traslación, entonces es definida como:

$$\tau_{jk} = k \cdot s_j$$

Donde j y k son números enteros.

Con esto, la Wavelet madre cambia a una forma discreta y se define como:

$$\psi_k^j(t) = 2^{\frac{-j}{2}} \cdot \psi\left(2^{-j} \cdot t - k\right) \tag{1.2.3}$$

La transformada Wavelet realiza una división de cualquier señal de energía finita en varias proyecciones de la misma sobre espacios definidos por las Wavelet hijas.
El muestreo en tiempo es pequeño para el análisis utilizando Wavelet de pequeña escala, mientras que es grande para el análisis con Wavelet de gran escala. La posibilidad de variar el factor de escala *s* permite usar Wavelet de escala muy pequeña para concentrar el análisis en singularidades de la señal. Cuando sólo los detalles de la señal son de interés, unos pocos niveles de descomposición son necesarios. Por lo tanto el análisis Wavelet provee una forma más eficiente de representar señales transitorias.

1.2.4 Ejemplo de funciones Wavelet utilizadas en procesamiento de voz

Existen diferentes Wavelet que ya son utilizadas de forma constante y que tienen definiciones establecidas. Sin embargo, la elección de un tipo de Wavelet depende de la aplicación específica que se le vaya a dar. Actualmente existen muchas aplicaciones en las que las Wavelet actúan de manera directa, una de esas aplicaciones es en procesamiento de voz.

A continuación se presentan algunas de las funciones Wavelet mas utilizadas dentro del procesamiento de voz y con las cuales se realizó una comparación con la función Wavelet basada en el funcionamiento del oído humano.

Haar:

Esta es la Wavelet más simple y antigua, se describe con la siguiente función:

$$h(x) = \begin{cases} 1: 0 \leq x < \frac{1}{2}. \\ -1: \frac{1}{2} \leq x < 1. \\ 0: \quad \text{otro valor.} \end{cases} \tag{1.2.4}$$

Su gráfica se muestra en la Figura 1.2.4.1, donde se puede observar que es una Wavelet sencilla y tiene una forma cuadrada lo cual no es lo más óptimo para el procesamiento de voz.

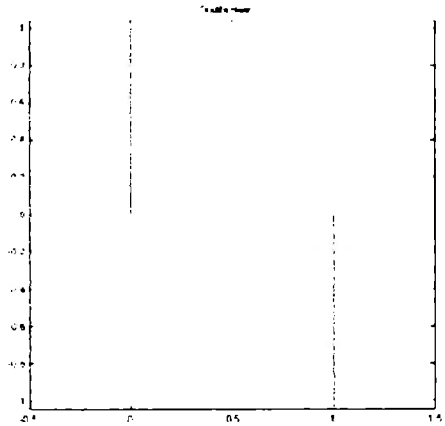


Figura 1.2.4.1 Wavelet de Haar

Mexican hat:

El nombre de esta Wavelet proviene de la forma que describe su gráfica que está dada por:

$$\psi(x) = \left(\frac{2}{\sqrt{3}}\pi^{-1/4}\right)(1-x^2)e^{-x^2/2} \tag{1.2.5}$$

Esta función Wavelet es simétrica como se observa en la Figura 1.2.4.2, por lo que le permite examinar a las señales de un modo simétrico, esta función Wavelet se utiliza en procesamiento de voz por su forma Gaussiana.

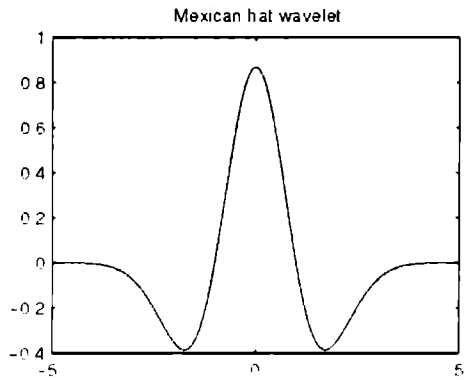


Figura 1.2.4.2 Wavelet de Mexican hat

Morlet:

La expresión para definir esta Wavelet es la siguiente:

$$\psi(x) = e^{-x^2/2}\cos(5x) \tag{1.2.6}$$

La Wavelet de Morlés tiene una forma simétrica como muestra la Figura 1.2.4.3 y tiene una forma similar a la de Mexican hat por lo que también es utilizada en procesamiento de voz.

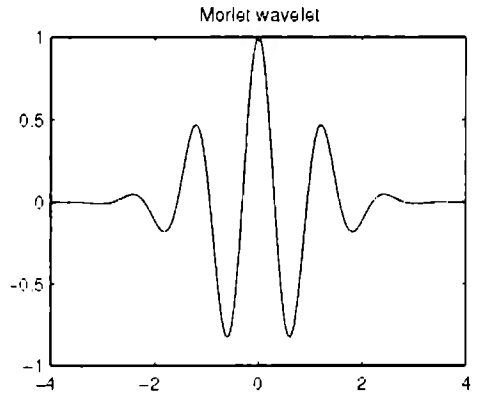


Figura 1.2.4.3 Wavelet de Morlet

Meyer:

Esta Wavelet tiene una función de la siguiente forma:

$$\hat{\psi}(\omega) = (2\pi)^{-1/2} e^{i\omega/2} \sin\left(\frac{\pi}{2} v\left(\frac{3}{2\pi}|\omega| - 1\right)\right) \qquad if \qquad \frac{2\pi}{3} \le |\omega| \le \frac{4\pi}{3}$$

$$\hat{\psi}(\omega) = (2\pi)^{-1/2} e^{i\omega/2} \cos\left(\frac{\pi}{2} v\left(\frac{3}{4\pi}|\omega| - 1\right)\right) \qquad if \qquad \frac{4\pi}{3} \le |\omega| \le \frac{8\pi}{3} \tag{1.2.7}$$

La Wavelet de Meyer tiene una grafica como la mostrada en la figura 1.2.4.4, tiene una forma similar a la Gaussiana lo que hace que el procesamiento de voz se pueda realizar.

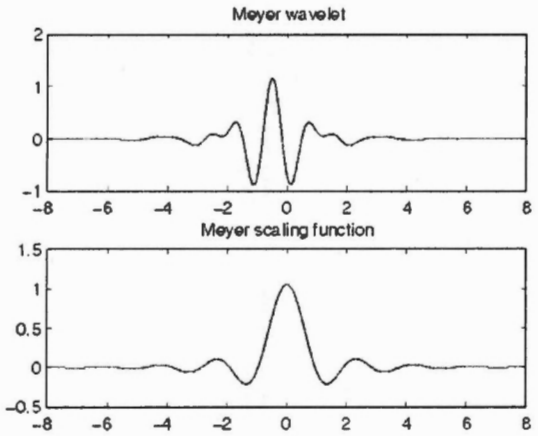


Figura 1.2.4.4 Wavelet de Meyer

1.3 Redes neuronales

1.3.1 Introducción

Durante años, el ser humano ha mostrado gran interés en el tema de reproducir la habilidad cognoscitiva por medio de recursos artificiales, es decir, a través del tiempo el ser humano ha pretendido imitar el funcionamiento del cerebro para distintos temas de investigación, a dichas aplicaciones se les conoce como "Inteligencia Artificial". Uno de los múltiples métodos para llevar a cabo la reproducción de la habilidad cognoscitiva son las llamadas redes neuronales.
Con el uso de redes neuronales se busca la solución de problemas complejos, no como una secuencia de pasos, sino como la evolución de unos sistemas de computación inspirados en el cerebro humano y dotados por tanto de cierta "inteligencia", estos sistemas son la combinación de elementos muy simples interconectados que procesan información y consiguen la resolución de problemas relacionados con reconocimiento de patrones, predicción y control entre otras aplicaciones.

Una red neuronal es un modelo de interconexión de las neuronas que intentan reproducir el comportamiento del cerebro. Es decir, una red neuronal es un procesador de información recurrente y distribuido que tiene una propensión natural para organizar conocimientos experimentales. Su similitud con el cerebro se encuentra en que su conocimiento es adquirido por un proceso de aprendizaje

que se guarda en las interconexiones de las neuronas, este proceso es mejor conocido como peso sináptico.

Las redes neuronales son sistemas no lineales que pueden ser fácilmente adaptables ya que a través del tiempo sus pesos sinápticos van cambiando adaptándose así a los cambios que vaya sufriendo la red en el transcurso de aprendizaje. Sus componentes principales son:

I) Unidad de proceso: La neurona artificial

Comúnmente las neuronas están agrupadas en 3 capas que son la de entrada, la capa de salida y las capas ocultas.

En la figura 1.3.1.1 se puede observar el modelo no lineal de una neurona:

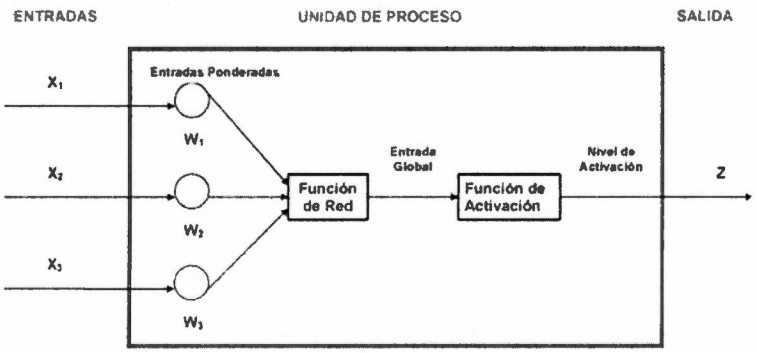


Figura 1.3.1.1 Modelo no lineal de una neurona

Donde x_k son las señales de entrada, w_k son los pesos de la neurona k y la función de red es un sumador y Z es la salida de la red.

II) Capas:

Conjunto de neuronas cuyas entradas provienen de la misma fuente y cuyas salidas se dirigen al mismo destino.

III) Función de activación:

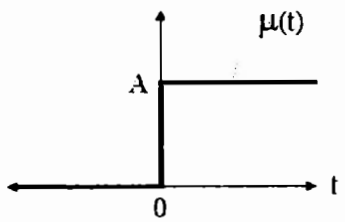
Es la última etapa en una neurona y su labor es la de entregar una salida acotada en términos de las entradas y salidas.

Existen cinco funciones de transferencia típicas que determinan distintos tipos de neuronas:

- *Función Escalón:* La salida puede tomar solo valores de "0" o "1". La función de umbral se define:

$$u(t) = \begin{cases} 0, t < 0 \\ 1, t > 0 \end{cases} \qquad (1.3.1)$$

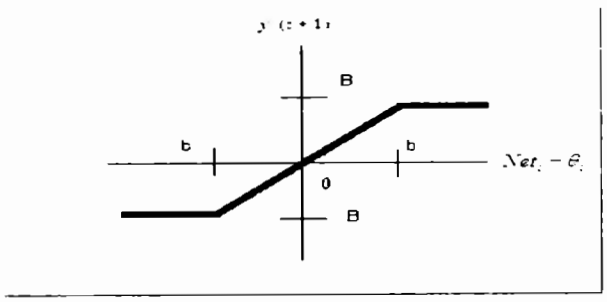
Su gráfica:



- *Función Lineal a tramos:* Esta función lineal esta limitada por dos intervalos definidos de la siguiente manera:

$$y_i(t+1) = \begin{cases} b & si\left[Net_i <= b + \theta_i\right] \\ Net_i - \theta_i si & b + \theta_i < Net_i < B + \theta_i \\ B & si\left[Net_i >= B\right] \end{cases} \qquad (1.3.2)$$

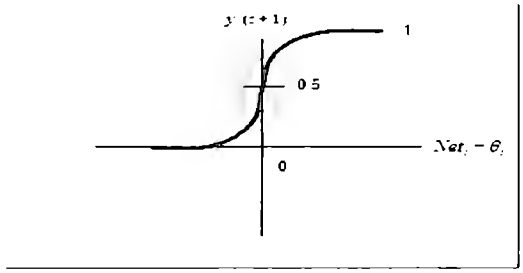
Su gráfica tiene la siguiente forma:



- *Función Sigmoidal:* Esta función es de gran ventaja para el uso de redes neuronales debido a que es una función creciente con cambios suaves. Se puede variar su pendiente ajustándola a las características de alguna red neuronal en particular. Se define de la siguiente manera:

$$y(t+1) = \frac{1}{1+e^{-(Net_i - \theta_i)}} \qquad (1.3.3)$$

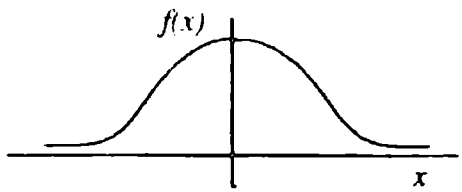
Su gráfica se muestra de la siguiente forma:



- *Función Gaussiana:* Esta función solo está delimitada para el rango de "0" a "1" y presenta cambios suaves. Está definida de la siguiente forma:

$$y = Ae^{-Bx^2} \tag{1.3.4}$$

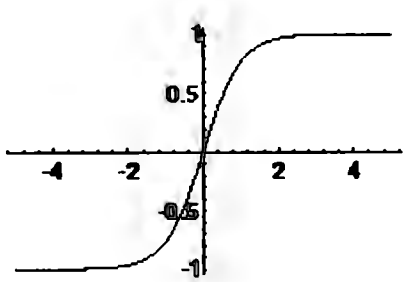
Su gráfica tiene la siguiente forma:



- *Función tangente hiperbólica:* Esta función presenta cambios suaves y es creciente al igual que la función sigmoidal. Se define de la siguiente manera:

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{1.3.5}$$

Su grafica tiene la siguiente forma:



1.3.2 Topologías de redes neuronales

La arquitectura de una red neuronal, se refiere a la forma en la cual se encuentran interconectadas las neuronas en sus diferentes capas. Dicha interconexión es una parte muy importante para desempeño de la red.
Existen 4 clases de redes neuronales según su arquitectura:

Redes neuronales mono capa

Este tipo de topología sólo cuenta con una capa de neuronas de entrada y una de salida. En las redes mono capa se establecen conexiones laterales entre las neuronas que van en un solo sentido. En la capa de entrada no se realiza ningún tipo de procesamiento, por lo que la información solo es transferida a la salida, quien será la que procesará finalmente la información.

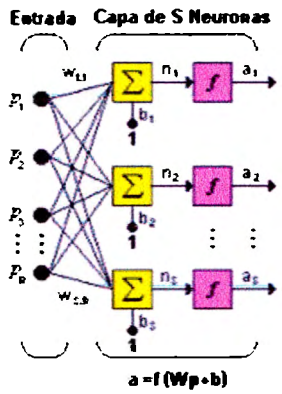


Figura 1.3.2.1 Red Neuronal Mono capa

Redes neuronales multicapa

Las redes multicapa disponen de conjuntos de neuronas jerarquizadas en distintas capas, con al menos una capa de entrada y otra de salida. Eventualmente una o varias capas ocultas. Normalmente todas las neuronas de una capa reciben señales de otra capa anterior y envían señales a la capa posterior. A estas conexiones se las conoce como conexiones hacia delante. El procesamiento de la información se lleva a cabo en las capas ocultas, las cuales reciben la información proveniente de la capa de entrada, a su vez, la capa de salida recibe la información procesada y finalmente genera una respuesta total para el sistema perteneciente al patrón de activación dado por los nodos que conforman la capa de entrada.

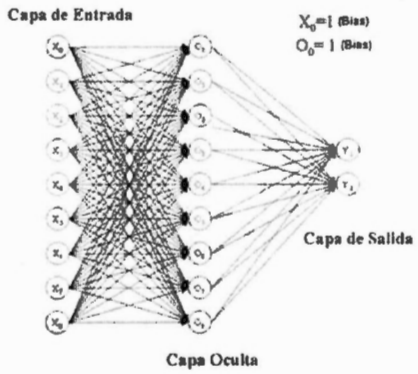


Figura 1.3.2.2 Red Neuronal multicapa

Redes recurrentes

Esta topología de red tiene la estructura de la red multicapa, la característica distintiva es que cuenta con un lazo de retroalimentación, es decir, la capa de salida alimenta también a la capa de entrada lo cual tiene un impacto en el funcionamiento de la red.

Estructuras matriciales

Las estructuras matriciales están formadas por una capa de entrada, seguida de un arreglo de una o mas dimensiones. Cada neurona de la capa de salida está organizada en filas y columnas para llevar a cabo el procesamiento de la información. No tiene lazos de retroalimentación.

1.3.3 Algoritmo de propagación hacia atrás

Uno de los algoritmos mas utilizados al entrenar una red neuronal es el algoritmo de propagación hacia atrás. Para llevar a cabo este entrenamiento es necesario ajustar los pesos y umbrales, de tal manera que el error entre la salida deseada y el de la salida actual, sea mínimo. Para lograr esto, necesitamos obtener la derivada de los pesos (EW). En otras palabras, se debe calcular como es el cambio del error según va variando el valor de los pesos.

El algoritmo, calcula cada error obteniendo primero que tan rápido cambia el error conforme a la salida de una neurona en la capa de salida (EA). El EA, es la diferencia entre la salida actual y la deseada.

Primero, calcula el peso total de la neurona X_j usando la siguiente expresión:

$$X_j = \sum_i y_i W_{ij} \tag{1.3.7}$$

Donde y_i son las señales de entrada de la capa anterior y W_{ij} es el peso de la conexión entre la neurona *i* y *j*.

Utilizando la señal de activación, las salidas se determinan y la red neuronal calcula el error E definido por la siguiente expresión:

$$E = \frac{1}{2}\sum_i (y_i - d_i)^2 \tag{1.3.8}$$

Donde y_j es la salida de la neurona *i* y d_j es la salida de la neurona. Esta expresión aplica para cada una de las diferentes capas.

Los cuatro pasos a seguir por el algoritmo son:

I) Calcula que tan rápido cambia el error con respecto a la salida de una neurona en la capa de salida.

$$EA_j = \frac{\partial E}{\partial y_j} = y_j - d_j \tag{1.3.9}$$

II) Calcula la tasa de cambio del error con respecto a la entrada de un nodo en la capa de salida.

$$EI_j = \frac{\partial E}{\partial x_j} = \frac{\partial E}{\partial y_j} x \frac{\partial y_j}{\partial x_j} = EA_j y_j (1 - y_j) \tag{1.3.10}$$

III) Se determina como va cambiando el error con respecto al cambio del peso entre el enlace de esa neurona y la anterior.

$$EW_{ij} = \frac{\partial E}{\partial W_{ij}} = \frac{\partial E}{\partial x_j} x \frac{\partial x_j}{\partial W_{ij}} = EI_j y_i \tag{1.3.11}$$

IV) Para las capas intermedias, se calcula como cambia el error con respecto a la salida de la capa previa.

$$EA_i = \frac{\partial E}{\partial y_i} = \sum_j \frac{\partial E}{\partial x_j} x \frac{\partial x_j}{\partial y_i} = \sum_j EI_j W_{ij} \tag{1.3.12}$$

Finalmente se calcula la corrección del peso para cada uno de los nodos:

$$\Delta W_{ij} = -\eta \frac{\partial E}{\partial W_{ij}} = -\eta EW_{ij} \tag{1.3.13}$$

Y la corrección para los umbrales:

$$\Delta \theta_j = -\eta \frac{\partial E}{\partial \theta_j} = -\eta EA_j \theta_j (1 - \theta_j) \tag{1.3.14}$$

Donde θ_j es el umbral y η es el factor de aprendizaje.

En este capítulo se hablo sobre el funcionamiento y las características del aparato auditivo y del aparato fonador. También se hablo sobre la parte teórica de la transformada Wavelet y las redes neuronales, las cuales forman parte fundamental del proyecto. En este capítulo se mostró de forma detallada el marco teórico, es decir, todos los antecedentes teóricos que se necesitan saber para el entendimiento del presente proyecto.

En el siguiente capítulo se expondrá la parte más importante del proyecto, que consiste en el sistema de análisis y la extracción de parámetros por medio de una Wavelet basada en el modelo del oído humano.

2.1 Esquema general

El sistema de análisis de voz utilizado para el desarrollo de este proyecto tiene como objetivo principal identificar los segmentos vocálicos de una señal de voz esofágica. Los conceptos presentados en el capítulo anterior son la base del sistema de reconocimiento de voz propuesto para lograr la identificación y clasificación correcta de los fonemas vocálicos presentes en una señal de voz.

Las características presentes en la voz de un ser humano tienden a variar según el locutor, la intensidad o incluso el idioma. Por lo anterior, se dificulta mucho el generalizar un sistema de reconocimiento de voz tomando en cuenta cada una de esas características.

El sistema de reconocimiento de voz utilizado a continuación es una herramienta programada en el ambiente de Matlab cuya finalidad es acondicionar la señal de voz para posteriormente extraer y resaltar sus características más significativas y así lograr el reconocimiento correcto de las 5 vocales del idioma español según corresponda el segmento de voz analizado.

En la figura 2.1.1 se muestra el diagrama de bloques del sistema de reconocimiento de voz, el cual se explicara con mayor detalle a continuación.

Figura 2.1.1 Diagrama a bloques del sistema de reconocimiento de voz

2.2 Sistema de análisis de voz

2.2.1 Acondicionamiento

Esta etapa consiste en filtrar una señal de audio para reducir el ruido, posteriormente la señal se normaliza y se amplifican sus componentes de mayor importancia para el reconocimiento de la voz. El objetivo principal de esta etapa

consiste en modificar las características espectrales de dicha señal de audio y optimizarlas para obtener resultados más claros en las etapas posteriores.

Para la implementación en Matlab de esta etapa, el filtrado se llevó a cabo por medio de un filtro FIR pasa altas para eliminar las componentes frecuenciales provenientes del ruido, así mismo, utilizamos otro filtro FIR pasa bajas con el mismo objetivo. Con ambos filtros, se pretende entregar a la salida una señal limpia de ruido.

Para el reconocimiento de la voz es necesario obtener las componentes de la señal donde haya más energía, por lo anterior, se debe resaltar cada una de las componentes de la señal. Esta parte se logró al adherir al sistema un filtro pasa altas llamado filtro de pre-énfasis. Debido a la pendiente de 20dB por década que tiene, este filtro amplifica las componentes de alta frecuencia y cumple con 2 funciones importantes: equilibra una atenuación de similar magnitud presente en las secciones vocalizadas de la voz e imita la sensibilidad adicional del oído humano a sonidos de frecuencias altas.

Con base en trabajos previamente hechos, se seleccionó un filtro de pre-énfasis con respuesta al impulso de la siguiente manera:

$$h = \begin{bmatrix} 1 & -0.4 \end{bmatrix}$$

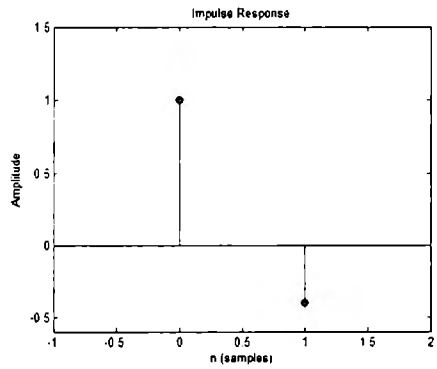


Figura 2.2.1 Respuesta al impulso

Para finalizar esta etapa, se normalizó la señal ya filtrada con el fin de garantizar que a la salida del pre-procesamiento, la señal tenga la misma energía. Esto es de gran relevancia ya que más adelante la energía y la potencia nos ayudarán a tomar decisiones importantes sobre el contenido fonético de la misma.

2.2.2 Segmentación

En esta etapa necesitamos dividir las señales de mayor y menor energía, es decir, dividirlas en segmentos vocalizados y no vocalizados. Sabemos que al hablar, nuestras cuerdas vocales vibran al presentar un comportamiento vocalizado. Es por

eso que para la etapa de segmentación dividimos las señal en diferentes tramas que posteriormente clasificaremos en vocalizadas y no vocalizadas. Este proceso evitará que más adelante se pierda tiempo en calcular segmentos de la señal que no nos ayudarán para el reconocimiento de voz adecuadamente.

La identificación de los segmentos vocalizados se lleva trama a trama, por lo tanto, es muy importante que la duración de cada trama solo contenga un fonema con información relevante para identificar su contenido. Con el objetivo de no tener errores por esta segmentación, cada trama se multiplica por una ventana de cambios suaves. Esta ventana debe tener una duración ligeramente mayor en tiempo que cada trama para evitar pérdidas de información entre una secuencia y otra.

Al programar esta etapa en Matlab, se consideró una segmentación de tramas de 15ms empleando una ventana de 30ms. Para el suavizado de la señal se utilizó una ventana de Hamming que presenta la siguiente ecuación:

$$W[k+1] = 0.54 - 0.46\cos\left(2\pi\frac{k}{(n-1)}\right), k = 0,1,2,...,n-1$$

Después de pasar la señal acondicionada por la ventana de Hamming tendremos como resultado una señal con secuencias de menor longitud, que corresponden a la información de entrada en distintos intervalos de tiempo.

2.2.3 Detección de segmentos vocalizados

En esta sección se debe identificar por medio del clasificador los segmentos de la señal que son vocalizados. Como se mencionó anteriormente, un segmento de la señal de voz se considera vocalizado si la generación de flujo de aire es alterado por la vibración de las cuerdas vocales del locutor. La diferencia de los segmentos vocalizados y no vocalizados se encuentra en que los segmentos vocalizados presentan un movimiento periódico y a su vez, transportan una mayor energía con respecto de los segmentos no vocalizados. Para poder clasificar estos segmentos, se manipularon ambas características.

Primero, se calcula la potencia promedio en cada segmento con la siguiente fórmula:

$$Pprom = \frac{1}{n}\cdot\sum_{k=1}^{n}\left|x[k]^2\right|$$

Donde x representa el segmento de la señal a analizar y n el numero de muestras que lo forman.

La energía transportada por la señal de voz puede variar dependiendo de cada persona, o por diferentes factores tales como el tono al hablar, la presencia de ruido, etc. Esto tiene como consecuencia la dificultad de definir un intervalo fijo para la potencia promedio y por lo tanto, hace mas complicada la clasificación de segmentos vocalizados. Es por esta razón que el algoritmo para la detección de secuencias vocalizadas compara la potencia promedio del segmento con la que posee la señal en la cercanía del mismo. El intervalo en la segmentación es de 200ms, este intervalo se definió de manera empírica como la duración promedio de una sílaba al hablar de manera normal. Posteriormente se tuvo que fijar un límite superior y otro inferior de la potencia promedio, la cual será comparada con la del segmento de la señal. Esto se hizo con el fin de evitar que segmentos que no son vocalizados pero que tienen un comportamiento similar sean clasificados como vocalizados por ejemplo, intervalos de silencio, o algún diptongo.

A continuación se determina si el segmento es vocalizado o no vocalizado por medio de un teorema propuesto por Greenwood, el cual consiste en fijar dos umbrales y hacer la división entre la potencia promedio del segmento y la señal. Si el cociente es menor al umbral inferior, la trama se considera no vocalizada, en cambio si este cociente es mayor al umbral superior, los datos se consideran como vocalizados. Si el cociente llegase a estar entre ambos umbrales, este criterio no es suficiente para determinar la clasificación y es entonces cuando aplicamos el criterio de cruces por cero, para aprovechar la periodicidad del segmento vocalizado.

El algoritmo de cruces por cero consiste en determinar dos umbrales con valores propuestos por Mark Greenwood y Andrew Kinghom. Dichos valores son de 1200 y 5000 cruces por cero en cada segundo. Se compara el número de cruces por cero de la señal con el intervalo. Si el número se encuentra dentro del intervalo, entonces el segmento se considera como vocalizado, de lo contrario es marcado como no vocalizado. Sin embargo este intervalo de 1200 a 5000, en el código implementado en Matlab es variable, debido a que el valor de cruces por cero se modificará con el cambio de locutor, con la edad, género.

Al finalizar esta etapa, tendremos a la salida una cadena de valores lógicos que indican el contenido de la señal, es decir si es vocalizado o no. Este programa en Matlab se llama isVoiced.

La siguiente parte de esta etapa consiste en la extracción de las características de los segmentos, cabe mencionar que es de las partes más importantes ya que debemos obtener características de calidad para que no afecte el desempeño del proyecto.

2.3 Extracción de parámetros utilizando Wavelet

La extracción de características de una señal de voz es la etapa de mayor importancia para el desarrollo de este proyecto. Los resultados obtenidos en este bloque son muy significativos para lograr los objetivos del proyecto. Por lo tanto es importante saber analizar los resultados de la extracción de parámetros y recordar que este diseño se hizo en base a la fisiología del sistema auditivo humano.

2.3.1 Construcción de un Wavelet madre a partir de un modelo del oído interno

En el sistema auditivo humano, la cóclea es el órgano encargado de transformar una vibración mecánica en un impulso nervioso que posteriormente es interpretado como un sonido por el cerebro. Dentro de la cóclea, la membrana basal junto con los vellos exteriores realiza una descomposición espectral de las señales mecánicas. Sin embargo esta descomposición no posee la misma resolución para todas las frecuencias. Es decir, entre mayor sea la frecuencia característica para una sección de la membrana basal, mayor será el rango de frecuencias para las cuales ésta responderá con un impulso en el nervio auditivo. Este fenómeno nos permite observar lo que ocurre dentro de nuestro oído como el análisis de resoluciones múltiples de una señal y por lo tanto se puede emplear la transformada wavelet para obtener una extracción de características similar a la que lleva a cabo el oído humano.

Con lo anterior se pretende definir un wavelet madre tomando en cuenta el modelo del oído propuesto por Zhang [18].

El modelo de Zhang propone que la dinámica de una sección de la membrana basal que presenta una frecuencia característica *fc*, es similar a la de un filtro *gamma-tone* sintonizado en esa misma frecuencia. Este filtro debe su nombre a su representación temporal, la cual consiste en el producto de una distribución gama por un tono.

La distribución gama indica la probabilidad de que hayan ocurrido cierto número de eventos que presentan una distribución de Poisson desde un tiempo inicial. Puede escribirse:

$$P[\alpha,\theta](x)=\frac{x^{\alpha-1}\cdot e^{\frac{-x}{\theta}}}{\Gamma(\alpha)\cdot\theta^{\alpha}}\rightarrow x>0$$

(2.3.1)

La función gamma se multiplica por un tono. Éste es elegido de tal manera que su frecuencia sea igual a la tasa con la que ocurren los eventos Poisson que dan origen a la distribución gama.

$$\Psi_{\theta}^{\alpha}(t) = \frac{1}{(\alpha-1)!\,\theta^{\alpha}} \cdot t^{\alpha-1} \cdot e^{-\frac{t}{\theta}} \cdot \cos(2\pi \cdot \frac{1}{\theta} \cdot t) \rightarrow t > 0 \tag{2.3.2}$$

Donde α es la forma y θ es la escala. Para esta ecuación en particular α = 3 ya que es el orden del filtro utilizado y θ es el inverso de la frecuencia.

La transformada de Fourier de la función $\Psi^{\alpha}(\tau)$ puede escribirse:

$$\Psi^{\alpha}(\omega) = \frac{1}{2} \cdot (\alpha-1)! \left[\frac{1}{\left[1 + i \cdot (\omega - 2\pi)\right]^{\alpha}} + \frac{1}{\left[1 + i \cdot (\omega + 2\pi)\right]^{\alpha}} \right] \tag{2.3.3}$$

$\Psi^{\alpha}(t)$ puede considerarse una wavelet madre para cada valor positivo de su parámetro debido a que la expresión anterior toma un valor real para todos los valores positivos de ω.

2.3.2 Muestreo del plano escala - traslación

La discretización de las variables anteriores se lleva a cabo con la finalidad de llevarlas a un sistema digital para procesamiento de señales. Es por eso que se necesita muestrear el eje de la escala y el eje de traslación, por lo tanto los valores que se tengan para dicho escalamiento no pueden tomar cualquier valor real. El objetivo principal de esto es que al obtener la transformada wavelet ya discretizada pueda ser calculada por una computadora e incorporar el principio de las bandas críticas en la descomposición.

La extracción de características deseada, hace uso de una descomposición de la señal en bandas que asemejan la respuesta en frecuencia de la cóclea. Debido a eso, se necesita un escalamiento en ambos ejes: escala y traslación. Esta descomposición muestra de manera directa las componentes para cada una de las bandas definidas por la escala de Bark.

La escala de Bark consiste en un mapeo logarítmico en el cual se pretende que la resolución espectral del oído humano sea de un Bark para una frecuencia característica. Debido a que es un fenómeno enteramente biológico, no se tiene una fórmula matemática que relacione la frecuencia en Hz con la frecuencia en Barks. Sin embargo se propone una interpolación hecha por Schroeder [19] que las relaciona:

$$Z = 7\ln\left(\frac{f}{650} + \sqrt{\left(\frac{f}{650}\right)^2 + 1}\right) \tag{2.3.4}$$

Esta escala de Bark, presenta 24 frecuencias críticas. Dichas bandas se definen mediante incrementos de 1 Bark a lo largo de toda la banda audible.

El muestreo de la escala apropiado para la transformada wavelet continua se hace con base a la interpolación mostrada, de tal forma que al descomponer la señal se pueda tener de manera concreta la información sobre cada banda del oído humano.

$$s_j = \frac{1}{325} \cdot \frac{e^{\frac{j}{7}}}{e^{\frac{2j}{7}} - 1} \quad j = 1, 2, \ldots$$

El muestreo propuesto para el eje de la escala cumple con el teorema de Littlewood-Paley [20], de modo que no existe pérdida de información durante la discretización propuesta para la transformada wavelet continua.

En este capítulo hemos presentado la primera parte del sistema, como es que se lleva a cabo el filtrado de la señal y las diferentes etapas para llegar a la extracción de características. En el siguiente capítulo se explicará como se lleva a cabo la clasificación por medio de una red neuronal, que es la ultima etapa del sistema, se explicara como funciona esta red y los resultados que se obtuvieron.

3.1 Algoritmo de la red neuronal

3.1.1 El perceptrón

La arquitectura de la red neuronal utilizada para desarrollar los objetivos de este proyecto estuvo basada en el perceptrón. Esta arquitectura es la forma más simple de red neuronal que se conoce y su objetivo es representar las propiedades más básicas de un sistema inteligente al cual se pueda adaptar.

El perceptrón es una red de alimentación directa, esto significa que la información fluye desde la capa de entrada en dirección hacia la capa de salida.

El Perceptrón es un clasificador el cual asigna a un vector de N valores un valor binario, usando una transformación no lineal. Así cada vector pertenece a una de las particiones que crea el perceptrón.

Está formado de una sola neurona con pesos y umbral ajustables, a su vez, tiene un tipo de aprendizaje supervisado, es decir, necesita conocer los valores esperados para cada una de las entradas.

La salida de la red se calcula como la suma de todas las señales de entrada multiplicadas por su peso y finalmente limitada por una función de activación.

3.1.2 Algoritmo de aprendizaje

Se lleva a cabo siguiendo la regla Delta en la cual se establece que el valor de los pesos debe ser ajustado por la diferencia entre la salida actual y la deseada.

$$\Delta w_j = \eta(d_j - y_j)x_j \qquad (3.1.2.1)$$

Donde Δw_j es el cambio en el peso, η es el factor de aprendizaje, d_j es la salida deseada, y_j la salida actual y x_j la señal de entrada.

Si la salida es correcta, nada se cambia, en caso contrario, los pesos que conectan las entradas que dan las salidas erróneas, son modificados de forma en que se reduzca el error cometido.

3.1.3 Vector de parámetros, clasificación y filtrado

En el sistema propuesto la arquitectura de la red neuronal es utilizada al finalizar la etapa de creación del vector de parámetros, la cual se encarga de enviar al clasificador la información proveniente de la extracción de características.

La creación del vector de parámetros se lleva a cabo descartando inicialmente todos los segmentos de la señal que fueron detectados como no vocalizados para evitar cálculos innecesarios.

En esta etapa, la señal vuelve a segmentarse debido a que la información obtenida de la extracción de características no se encuentra dividida, esto se hace por una ventana rectangular siguiendo un proceso similar al de segmentación. A continuación, se extrae la característica más significativa de cada una de las bandas provenientes de la extracción de características.

Además de la información obtenida en la descomposición por transformada Wavelet, es necesario obtener también la energía total de la señal previa a la descomposición. La razón de lo anterior consiste en que la energía total en una señal física se conserva siempre, por lo cual, junto con la normalización de la primera etapa permitirá al clasificador realizar una discriminación entre los posibles fonemas vocálicos contenidos en un segmento.

El vector de parámetros consta de las siguientes componentes:

- Energía total del segmento
- Energía contenida en cada una de las bandas utilizadas por la transformada wavelet.
- Cambio de energía de la señal con respecto al segmento anterior.
- Cambio en la energía contenida en cada banda con respecto a la trama previa.

La clasificación de cada vector de parámetros insertado a la de la red se lleva a cabo a la salida de la red neuronal, es decir, por cada segmento vocalizado se construye un vector de parámetros el cual se suministra a la red neuronal que posteriormente es clasificado según la clase de vocal correspondiente.

La red neuronal artificial clasifica su salida en cinco clases representando a cada vocal. La arquitectura propuesta para cumplir con este objetivo es una red neuronal artificial de varias capas ante-alimentadas, entrenada con el algoritmo de propagación hacia atrás.

La red neuronal consta de una capa de entrada cuyo numero de nodos es igual a la longitud del vector de parámetros y su función es solamente pasar la información a la siguiente capa; consta también de dos capas ocultas interconectadas con un número de nodos ajustable que se encargan de procesar la información y utilizan como función de activación la tangente hiperbólica.

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \qquad (3.1.3.1)$$

Por último, se tiene una capa de salida con 5 nodos que representan a cada una de las vocales y cuya salida tiene un comportamiento lineal, es decir, la salida es igual a la entrada.

Cada neurona utilizada determina su entrada como la suma de la señal multiplicada por un desplazamiento el cual es calculado, al igual que los pesos en cada neurona, esto basado en el algoritmo de propagación hacia atrás.

El entrenamiento de la red se lleva a cabo con la inserción de archivos de audio para obtener un conjunto de datos característicos. Al inicio del entrenamiento se crean vectores de parámetros para cada archivo de entrenamiento, de la misma forma se crean vectores de parámetros deseados para la salida de la red según el contenido del archivo.

Previo al entrenamiento, se normalizan las componentes de entrada y salida de cada vector obtenido. Para llevar a cabo dicha normalización, a la entrada y a la salida de la red se resta a cada elemento del vector la media y posteriormente se divide entre la desviación estándar presente en dicha componente, es decir:

$$\overset{)}{V}_{i} = \frac{V_{i} - \mu[V_{i}]}{\sigma[V_{i}]} \qquad (3.1.3.2)$$

Este proceso de normalización evita que componentes muy significativas o de magnitud muy grande del vector impacten considerablemente el desempeño de la red.

Por otro lado, cuando la red ya ha sido entrenada y se utiliza para la clasificación de algún vector de parámetros, éste es normalizado de la misma manera que a la entrada del entrenamiento. Sin embargo, a la salida se realiza la operación inversa con los resultados obtenidos. Es decir, los resultados obtenidos se multiplican por la desviación estándar de las salidas deseadas y al resultado se le suma la media calculada durante el entrenamiento.

Como parámetros de entrenamiento para la red neuronal y siguiendo las características del algoritmo de propagación se estableció el valor de factor de aprendizaje inicial, el cual es incrementado en un porcentaje anteriormente establecido con cada iteración que disminuye el error cuadrático medio de la red y es reducido en otro porcentaje después de cada iteración que incrementa en más de un 4% el error cuadrático medio. El entrenamiento finaliza cuando el error cuadrático medio logra converger a un valor o llega a un número máximo de iteraciones.

Inicialmente se establecieron los valores de factor de aprendizaje en 0.1, el incremento del factor de aprendizaje cuando reduce el error cuadrático medio en 5% y el porcentaje de reducción para el error cuando aumenta el error más del 4% en 30%.

Cada uno de esos valores fue modificado al momento de llevar a cabo la validación del algoritmo de entrenamiento en el código de Matlab train e initThresholds con el objetivo de optimizar y entregar resultados más consistentes en esta fase.

A la salida del clasificador tendremos un conjunto de cinco componentes por cada vector de parámetros ingresado a la red, el cual representa a que vocal pertenece dicho segmento vocalizado. La red determina a que vocal corresponde cada segmento vocalizado al comparar el valor más alto con un umbral de activación y el segundo valor más alto con un umbral de diferenciación, ambos umbrales pueden ser ajustados en la red. Si el valor mayor se encuentra por encima del umbral de activación y el segundo valor más grande se encuentra por debajo del valor de diferenciación, la vocal clasificada será la correspondiente a la clase enviada por la red neuronal. De lo contrario, no se puede clasificar y envía "_" indicando ausencia de componentes vocálicas.

Finalmente llegamos a la etapa de filtrado, la cual tiene como objetivo entregar una cadena de caracteres que representen cada uno de los fonemas vocálicos dentro del archivo de audio suministrado a la red.

Para descartar errores que se hayan cometido en la clasificación a la cadena obtenida se le aplica un filtro modal, donde inicialmente se eliminan los símbolos "_" que aparezcan donde hay un gran numero de vocales. Se ubica una ventana sobre cada símbolo y si la mitad de la cadena o más de ella está formada por vocales, el símbolo se elimina.

A continuación, divide la cadena de vocales en conjunto, los cuales son filtrados también por una ventana que recorre carácter a carácter y asigna a cada posición la moda de los elementos dentro de la ventana. Finalmente, se concatenan los grupos de vocales resultantes con el símbolo "_" entre ellos.

Se genera la cadena de caracteres la cual representa el reconocimiento de las vocales contenidas en la palabra suministrada a la red con un archivo de audio.

A continuación, presentaremos las modificaciones hechas al sistema para mejorar su funcionamiento en la parte de entrenamiento y en la clasificación.

3.2 Modificaciones realizadas

A continuación se mostraran las modificaciones que se realizaron para hacer más eficiente la red neuronal y por lo tanto mejorar el reconocimiento y la clasificación de las vocales dentro de la misma red neuronal.

En el entrenamiento de la red se modificaron algunos de los parámetros que venían previamente establecidos en el código. Para obtener mejores resultados se utilizó el método de prueba y error, donde se fueron cambiando los parámetros de entrenamiento y clasificación de la red hasta obtener un resultado favorable. Se realizaron muchas pruebas con las que se obtuvieron distintos resultados, en esta sección solo se muestran los resultados más significativos o más importantes que se lograron dentro de la red neuronal y el clasificador.

Dentro de la red neuronal, la Figura 3.2.1 muestra los parámetros originales y los parámetros modificados, donde se obtuvo un decremento en el error cuadrático medio después del entrenamiento la red, reduciéndolo hasta un 1.14%. Con lo cual podemos

decir que la red neuronal se está entrenando correctamente, con base en que el error dentro de ese entrenamiento es muy bajo.

Parámetros	Originales	Modificados
Factor de aprendizaje	0.01	0.001
Incremento si reduce el error	1.05	1.10
Decremento si aumenta el error	.7	.8
Iteraciones	5000	5000
Error Cuadrático medio deseado	0.1	0.001
Error Inicial	153.295	153.518
Error después de entrenar	7.98856	1.14549
Archivos de entrenamiento	26 por vocal	26 por vocal

Figura 3.2.1 Tabla de parámetros originales y modificados durante el entrenamiento de la red neuronal

En lo referente a la clasificación de los segmentos vocalizados después de pasar por la red neuronal, también se obtuvieron resultados favorables al variar algunos de los parámetros que determinan el funcionamiento del clasificador.

Estos parámetros y sus modificaciones se muestran en la Figura 3.2.2. En esta tabla se muestran los parámetros que venían establecidos originalmente en el código y los parámetros modificados. Con los parámetros modificados se mejoró la clasificación de los segmentos vocalizados en un 6%, lo que aumentó por consecuencia el porcentaje de reconocimiento del método utilizado en este proyecto, que es el de la Wavelet basado en el modelo del oído humano.

Los umbrales de activación y de diferenciación se modificaron de acuerdo a varias pruebas realizadas y se intentó establecerlos de tal forma que la clasificación fuera la mejor y correcta para todas las vocales en la capa de salida.

Parámetros	Originales	Modificados
Umbral de activación	0.5	0.53
Umbral de diferenciación	0.33	0.47
Nodos por capa		
1a. Capa oculta	54	50
2a. Capa oculta	10	9
Capa de salida	5	5

Tabla 3.2.2 Tabla de parámetros originales y modificados durante la clasificación de los segmentos vocalizados.

En la parte de filtrado modal, no se encontró que hubiera alguna falla. Sin embargo, se modificaron algunas partes del código para observar si mejoraba en algún sentido el reconocimiento, los resultados obtenidos no arrojaron ningún tipo mejora, por lo tanto no se modificó esta sección ya que el filtrado modal esta trabajando correctamente y no presenta errores; por lo tanto se quedó como estaba programado originalmente.

En este capítulo se mostró el tipo y el funcionamiento de la red neuronal utilizada en este proyecto. También se dieron a conocer algunos de los resultados obtenidos al modificar ciertos parámetros de la red neuronal.

La red neuronal es muy importante dentro del proyecto ya que en base a su buen funcionamiento, mejorará el reconocimiento y la clasificación de los segmentos vocalizados.

Con este capítulo se finaliza el sistema del proyecto, dando paso a la siguiente sección donde se expondrán los resultados obtenidos.

4.1 Comparación con otras funciones Wavelet

La importancia de este apartado se centra en la demostración de la efectividad de la función Wavelet basada en el modelo del oído humano en comparación con diferentes funciones Wavelet propuestas para aplicaciones similares.

En este capítulo encontraremos diferentes comparaciones donde se incluyen el reconocimiento en general, estadísticamente según las vocales clasificadas correctamente, las gráficas en 3 dimensiones de diferentes funciones en las cuales podremos obtener más información sobre el contenido energético y espectral de la señal de voz.

Para este estudio de comparación se utilizaron 7 diferentes funciones Wavelet, estas son: Haar, Daubuchies 4, Daubuchies 10, Mexican Hat, Morlet, Meyer y la Wavelet basada en el modelo del oído humano o Ear.
Habiendo programado con Matlab la extracción de características para cada Wavelet y modificando en el programa *signalParams.m* las funciones necesarias para llamar mandar a cada método y llevar a cabo el reconocimiento con las características provenientes de cada función, se procedió a hacer pruebas con cada una de las funciones Wavelet.

Las pruebas se llevaron a cabo suministrando al sistema 107 diferentes archivos de voz, 60 de voz normal y 47 de voz esofágica. Se documentaron en la Tabla 4.1.1 cada una de las cadenas de caracteres arrojadas por la red neuronal. En seguida se prosiguió a contar las vocales reconocidas correctamente y se le asignaba un porcentaje según el número de vocales reconocidas correctamente en cada palabra. Al finalizar dicho conteo sacamos los promedios según método y vocal, tanto para voz esofágica y voz normal. A continuación se presentan los resultados de los diferentes métodos y los archivos de audio que se le añadieron a la red después de ser debidamente entrenada.

VOZ NORMAL

alabra	ear	haar	daub4	daub 10	mexican hat	morlet	meyer
eja	_a_e_a_	_e_a_e_	_a_e_e_	_a_e_a_	_a_e_a_	_a_e_a_	_a_e_a_
rigo	_a_i_o_	_o_i_u_	_a_i_u_	_a_i_u_	_a_i_u_	_a_i_o	_a_i_u_
icto	_a_i_o_	_i_i_u_	_a_i_o_	_a_i_u_	_a_i_u_	_a_i_o_	_a_i_u_
iou	_a_e_i_o_u_	_a_a_a_a_a_	a_e_i_o_u_	_a_e_i_o_u_	_a_e_i_o_u_	_a_e_i_o_u_	_a_e_i_o_u_
ilar	_ai_a_	_e_a_	_ae_a_	_ai_a_	_ai_a_	_ai_a_	_ai_a_
la	_a_a_	_e_e_	_a_a_	_a_a_	_a_a_	_a_a_	_a_a_
bido	_e_i_o_	_a_i_o_	_e_i_u_	_e_i_u_	_e_i_u_	_e_i_o_	_e_i_u_
sar	_e_a_	_a_i_	_o_a_	_o_a_	_o_a_	_e_a_	_o_a_
po	_u_o_	_u_i_	_u_u_	_u_u_	_u_u_	_u_o_	_u_o_
da	_u_a_	_u_o_	_u_o_	_u_a_	_u_a_	_u_a_	_u_a_
co	_o_o_	_o_o_	_o_o_	_o_o_	_o_o_	_o_o_	_o_o_
narillo	_a_i_o_	_u_u_u_	_i_ui_i_	_a_e_o_	_a_a_o_	_a_e_o_	_a_e_o_
scado	_e_a_o_	_u_u_u_	_a_i_u_	_e_a_o_	_e_a_o_	_e_a_o_	_e_a_o_

VOZ ESOFÁGICA

alabra	ear	haar	daub4	daub 10	mexican hat	morlet	meyer
beja	_ae_a_	_ao_a_	_ae_a_	_ae_a_	_ae_a_	_ae_a_	_ae_a_
brazar	_a_a_a_	_a_a_a_	_a_a_a_	_a_a_a_	_a_a_a_	_a_a_a_	_a_a_a_
dicto	_a_i_o_	_a_o_a_	_a_e_a_	_a_i_o_	_a_i_o_	_a_i_o_	_a_i_o_
eiou	_a_e_i_o_u_	_a_a_i_a_o_	_a_i_i_i_i_	_a_e_i_o_u_	_a_e_i_o_u_	_a_e_i_o_u_	_a_e_i_o_u_
hilar	_a_a_	_ao_o_	_ao_a_	_a_a_	_a_a_	_a_a_	_a_a_
ache	_a_e_	_a_a_	_a_ei_	_a_uia_	_a_oi_	_a_e_	_a_e_
ebido	_e_i_o_	_o_o_o_	_i_i_o_	_e_i_o_	_e_i_o_	_e_i_o_	_e_i_o_
ruma	_u_a_	_o_a_	_e_a_	_u_a_	_u_a_	_u_a_	_u_a_
asa	_a_o_	_i_i_	_i_i_	_a_o_	_a_o_	_a_o_	_a_a_
uda	_u_a_	_a_a_	_a_a_	_u_o_	_u_o_	_u_o_	_u_a_
oco	_o_o_	_o_a_	_o_a_	_o_o_	_o_o_	_o_o_	_o_o_
bogado	_a_o_a_o_	_a_a_o_a_	_a_o_a_a_	_a_o_a_o_	_a_o_a_o_	_a_o_a_o	_a_o_a_o

Tabla 4.1.1 Resultados por método para cada función Wavelet

En la Figura 4.1.2 observamos la gráfica de resultados generales obtenidos para cada Wavelet. Se muestra el promedio de los porcentajes calculados para cada palabra según el reconocimiento correcto.

El resultado de dicha gráfica muestra que, como se había previsto, el porcentaje del método Ear fue el que mejor porcentaje de reconocimiento obtuvo de las vocales con un 74.07%, en segundo lugar se observa la función Morlet al tener un 68% de efectividad. Las demás funciones se acercan, sin embargo los resultados no se vuelven tan relevantes por la mínima diferencia entre ellos. La función Wavelet que si se encuentra muy por debajo de la media es la de Haar, sin dejar atrás a la función Daubuchies 4. Esto se debe a que la señal para estas dos últimas es una señal cuadrada y para las otras funciones se tiene una forma de onda gaussiana.
En la Figura 4.1.3 se nota una mejoría significativa para los demás métodos, esto es en más de un 10% a excepción del método "Ear" que solo incrementa en un 6%, sin embargo su porcentaje sigue estando por encima de los demás métodos.

Figura 4.1.2 Porcentajes de reconocimiento por método para voz normal

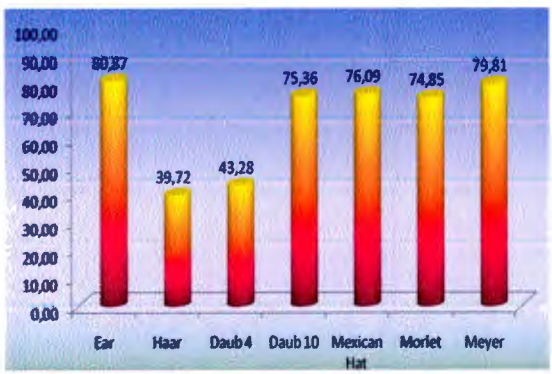


Figura 4.1.3 Porcentajes de reconocimiento por método para voz esofágica

La Figura 4.1.4 muestra el promedio general obtenido del resultado de los porcentajes de cada achivo de audio sin importar si es de voz normal o de voz esofágica. Con esta gráfica se logra demostrar que el método Ear es más efectivo que los otros a pesar de su cercanía estadística.

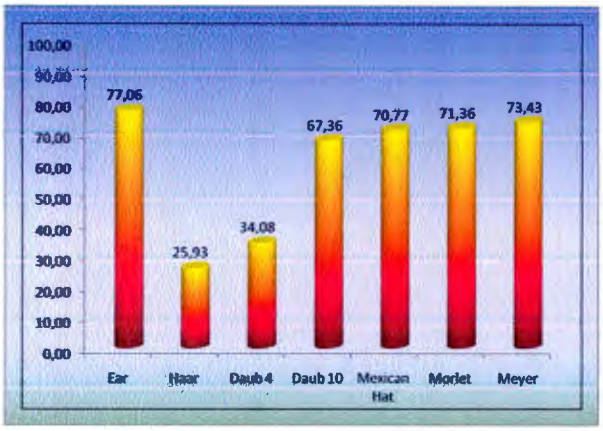


Figura 4.1.4 Porcentajes de reconocimiento generales por método

4.1.2 Comparación de vocales según método

Para un tipo de investigación estadístico de este tipo no es suficiente el comparar la función Wavelet basada en el modelo del oído humano con otras funciones Wavelet para demostrar su efectividad. En esta etapa fuimos un poco más lejos al sacar estadísticas sobre cual de las vocales en específico eran las mejor reconocidas por el sistema propuesto.

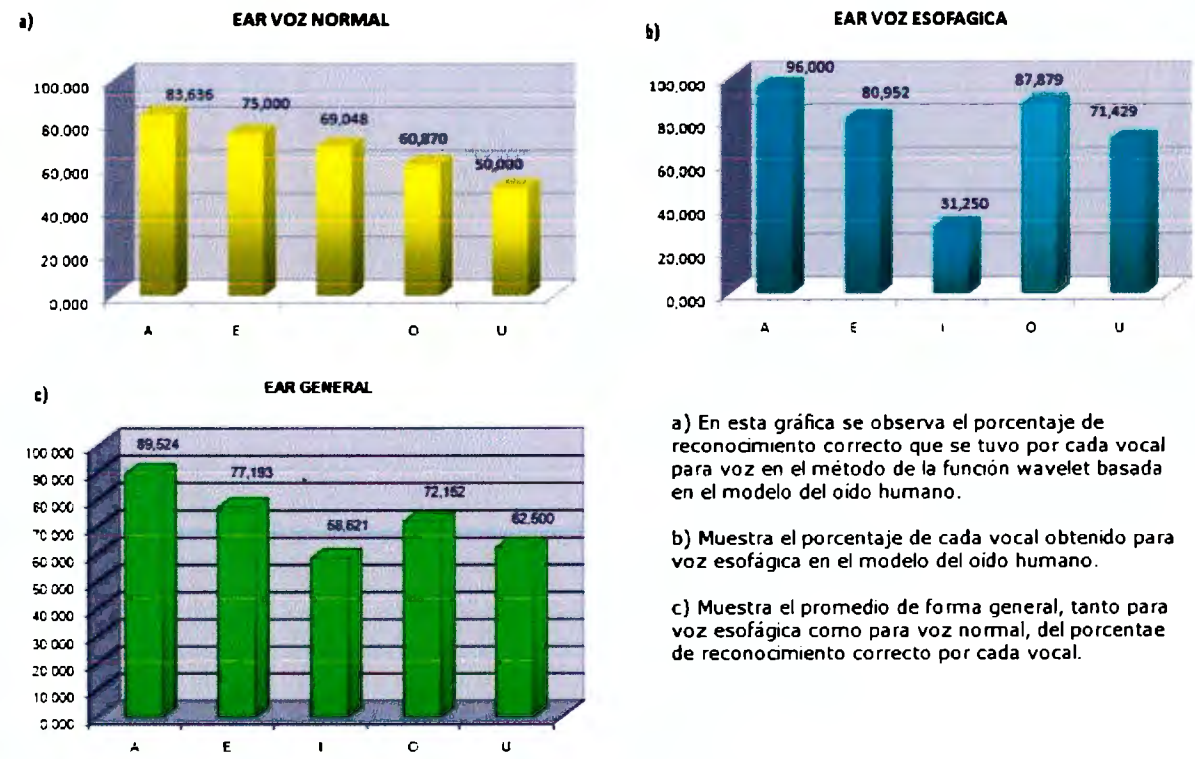
Esta parte del estudio estadístico se realizó contando inicialmente cuántas vocales de cada clase el sistema debía clasificar después de haber insertado los 60 archivos de voz normal y los 47 de voz esofágica, es decir, saber cuántas "a" había en total de las 107 palabras, "e" de cada archivo y así sucesivamente. Al término de este conteo, se obtuvieron 105 letras "a", 57 letras "e", 58 letras "i", 79 letras "o" y finalmente 48 letras "u".

Como siguiente paso, se contaron las vocales reconocidas correctamente y las vocales reconocidas erróneamente o que simplemente no reconoció en cada uno de los métodos analizados. Lo anterior se realizó, teniendo como principal objetivo obtener el porcentaje de las vocales reconocidas correctamente, este proceso se hizo para cada clase de vocal tanto en voz normal como para voz esofágica según el número de vocales. En voz normal, se tenían 55 letras "a", 36 letras "e", 42 letras "i", 46 letras "o" y finalmente 20 letras "u", para voz esofágica se tenían 50 letras "a", 21 letras "e", 16 letras "i", 36 letras "o" y finalmente 28 letras "u".

Finalmente se sacó el promedio general de cada vocal reconocida correctamente por método y cuál de esas vocales era la mejor reconocida según la función Wavelet que se estaba analizando.

En la Figura 4.1.2.1 se muestran los resultados obtenidos para el método Ear. En la gráfica a) se ve que la vocal "a" es la mejor reconocida por este método para voz

normal con un 83.63%, estando por debajo la vocal "e" con un 75%, la "i" con el 69.04%, la "o" con 60.87% y finalmente la "u" con el 50%. En la gráfica b) semuestran los resultados para los archivos de voz esofágica, estando por encima de todas la letra "a" y el porcentaje de reconocimiento más bajo es el de la letra "i". Finalmente, en la gráfica c) de la misma figura tenemos el promedio general de ambos tipos de voz. Los resultados obtenidos muestran también que la vocal mejor reconocida en este método es la letra a con un 89.52% y la más baja es la letra "i" con el 58.62%.



a) En esta gráfica se observa el porcentaje de reconocimiento correcto que se tuvo por cada vocal para voz en el método de la función wavelet basada en el modelo del oído humano.

b) Muestra el porcentaje de cada vocal obtenido para voz esofágica en el modelo del oído humano.

c) Muestra el promedio de forma general, tanto para voz esofágica como para voz normal, del porcentae de reconocimiento correcto por cada vocal.

Figura 4.1.2.1 Porcentajes de reconocimiento por vocal para método Ear

Este mismo análisis se realizó para cada uno de los métodos comparados teniendo como resultados que la vocal mejor reconocida por el sistema es la "a" ya que cada uno de ellos obtuvo un porcentaje, aunque distinto, mayor que las demás vocales.
El resto de las gráficas obtenidas para cada uno de los métodos puede verse también dentro del Anexo A.

4.1.3 Comparación de cada método según vocal

Otro punto importante para este apartado se centra en los resultados que se obtuvieron llevando a cabo otro estudio estadístico comparando las Wavelets. Este se basó en calcular el porcentaje mejor reonocido para cada vocal según el método; es decir, que método reconoció mejor la letra "a", la letra "e", etc.

Tomando en cuenta la información que ya se tenía sobre cuántas vocales había en total y cuántas reconoció correctamente, procedimos a tabular cada método y su porcentaje de reconocimiento correcto en la Tabla 4.1.3.1.

	VOZ NORMAL				
MÉTODO	A	E	I	O	U
EAR	83,636	75,000	69,048	60,870	50,000
HAAR	16,364	8,333	7,143	26,087	75,000
DAUB 4	20,000	22,222	52,381	8,696	35,000
DAUB 10	80,000	75,000	57,143	36,957	25,000
MEX HAT	87,273	77,778	64,286	47,826	55,000
MORLET	85,455	69,444	50,000	65,217	55,000
MEYER	89,091	63,889	57,143	54,348	55,000

VOZ ESOFÁGICA					
MÉTODO	A	E	I	O	U
EAR	96,000	80,952	31,250	87,879	71,429
HAAR	76,000	19,048	18,750	42,424	50,000
DAUB 4	80,000	28,571	12,500	30,303	46,429
DAUB 10	92,000	71,429	31,250	93,939	67,857
MEX HAT	92,000	76,190	31,250	96,970	67,857
MORLET	94,000	71,429	31,250	87,879	71,429
MEYER	98,000	80,952	25,000	90,909	71,429

Tabla 4.1.3.1 Porcentaje de reconocimiento por vocal para voz esofágica y voz normal

En las siguientes gráficas, Figura 4.1.3.2 se muestran los resultados obtenidos para los archivos de voz esofágica.

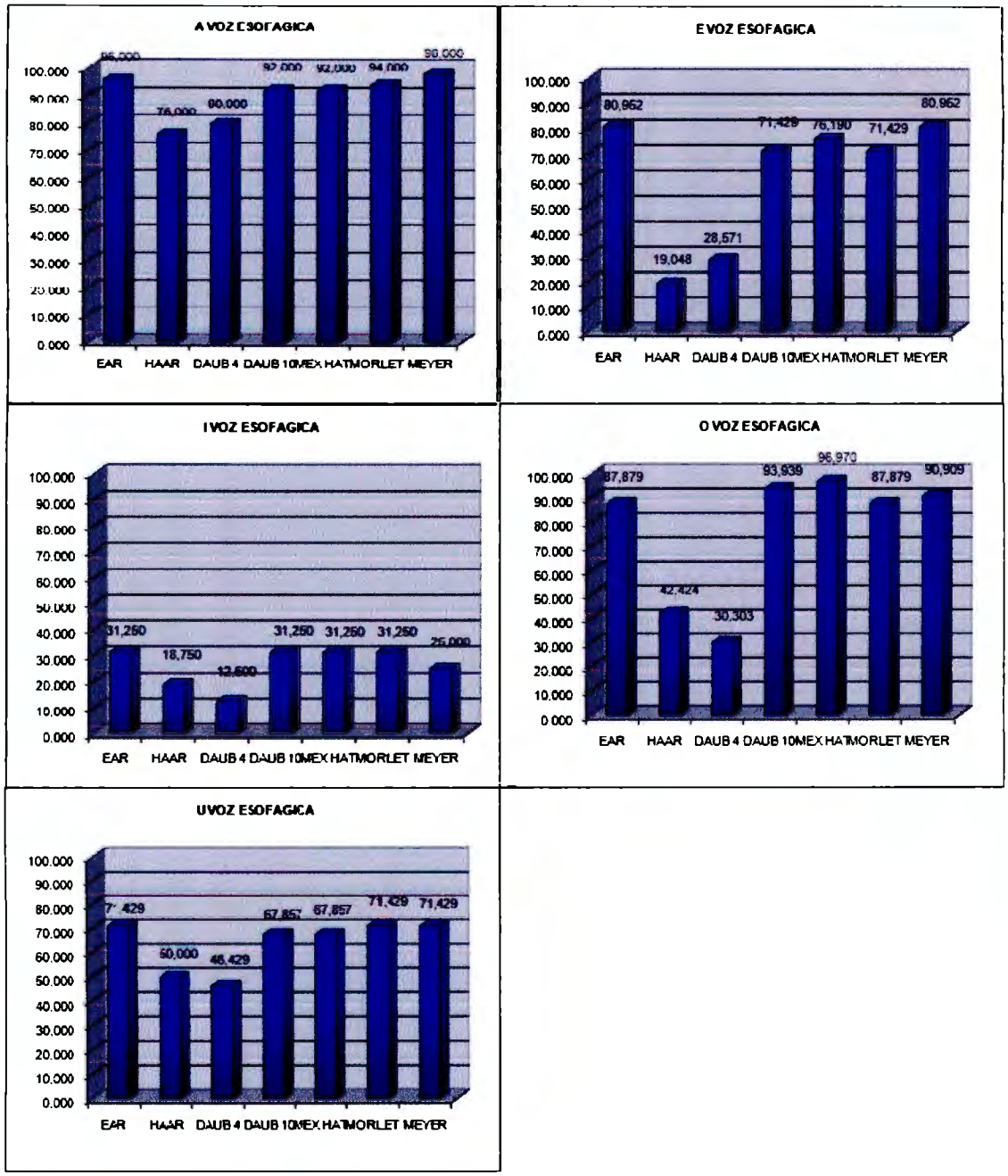


Figura 4.1.3.2 Porcentaje de reconocimiento correcto para las vocales de voz esofágica por método

La figura anterior muestra que tan efectiva es la función Wavelet basada en el modelo del oído humano con respecto a las demás funciones Wavelet.

En la gráfica para la letra "a" notamos que el porcentaje de reconocimiento de cada una de las funciones Wavelet es muy alto y aceptable, aún así, la función Ear esta por encima de las demás Wavelets a excepción de la Wavelet de Meyer cuyo

reconocimiento esta en un 98%. En cambio, en la letra "e" ambas funciones mencionadas anteriormente se encuentran también por encima de todas las demás con un 80.9%. La vocal con más problemas de reconocimiento se encontró en la letra "i", la cual no rebasó el 32% en ninguna de las funciones. La letra "o" también tiene un porcentaje de reconocimiento muy grande, oscilando entre valores porcentuales tales como 87% y 96% en todas las funciones Wavelet a excepción de la Wavelet de Haar y de Dabuchies4 que están por debajo del 4%. Finalmente se tiene la gráfica de la letra "u" la cual no tiene mucha variación porcentual entre el total de las funciones Wavelet.

En esta sección, podemos concluir que los resultados obtenidos con este estudio estadístico nos muestran que la función Wavelet basada en el modelo del oído humano tiene un porcentaje de reconocimiento mayor a cualquiera de las Wavelet con las cuales se llevó a cabo su comparación.

Con este estudio estadístico, se pudo demostrar que la función Wavelet del método Ear, al estar enfocada a trabajar como lo hace el oído humano, funciona de manera eficiente y mejor que las otras funciones Wavelets. Esto se debe principalmente a que se tiene una aplicación específica, a diferencia de las demás funciones Wavelet que han sido propuestas para aplicaciones similares.

4.2 Resultados del sistema

En esta sección del capítulo, se mostrarán los resultados obtenidos en el sistema, es decir, en la etapa de acondicionamiento, segmentación, detección de segmentos vocalizados, extracción de características y reconocimiento. Los resultados obtenidos se muestran para cada etapa por separado. Dentro de estos resultados se incluyen gráficas que se obtuvieron en las diferentes etapas del sistema.

Estos resultados muestran el proceso que se llevó a cabo para llegar a la etapa de extracción de características y reconocimiento de voz, las cuales son en conjunto, el punto más importante del presente proyecto.

4.2.1 Etapa de acondicionamiento

En esta etapa se implementaron tres diferentes filtros para limpiar la señal de voz. Los filtros implementados son filtros tipo FIR y cumplen con funciones específicas.

Las gráficas que se presentan a continuación, pertenecen a los tres filtros en el dominio del tiempo y en el de la frecuencia.

En la Figura 4.2.1.1 se muestra un filtro FIR pasa-altas con frecuencia de corte de 100Hz. Este filtro tiene la finalidad de eliminar componentes de ruido de baja frecuencia, específicamente de 60Hz.

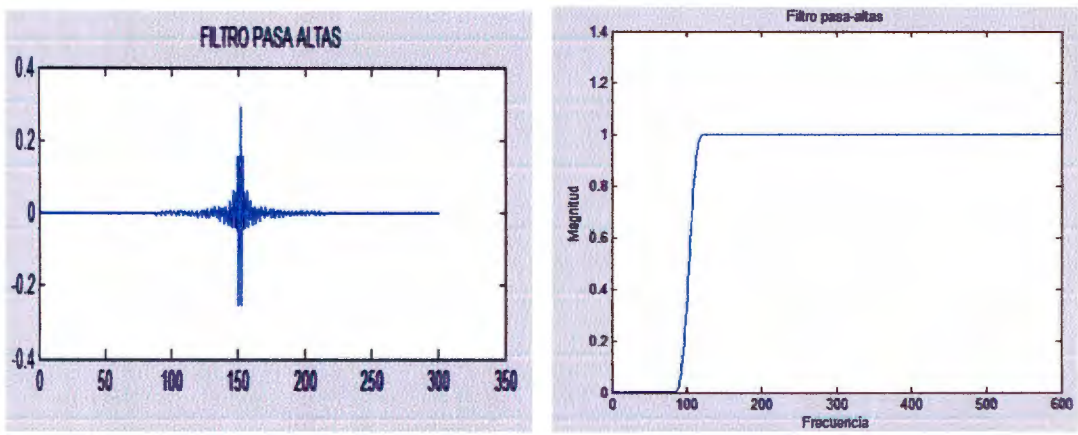


Figura 4.2.1.1 Filtro pasa-altas

En la Figura 4.2.1.2 se muestra un filtro FIR pasa-bajas con frecuencia de corte de 900Hz. Se eligió esta frecuencia de corte ya que las frecuencias mayores a esta, en voz normal se consideran no inteligibles.

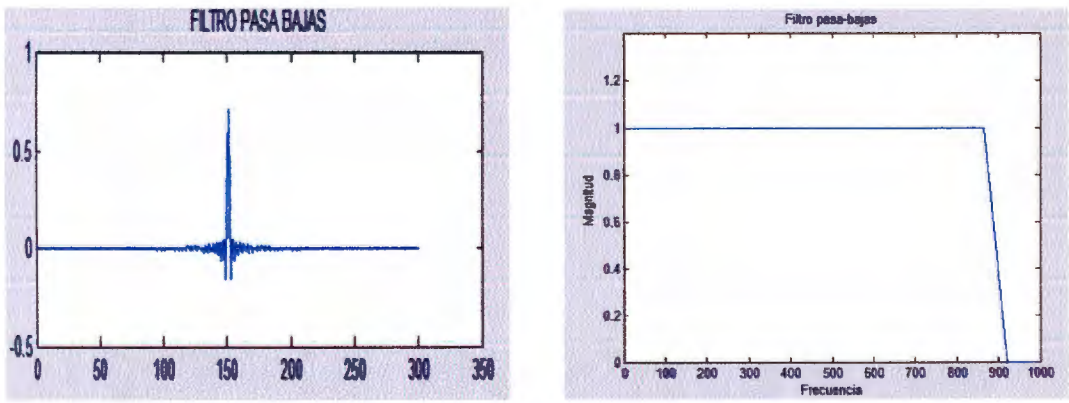


Figura 4.2.1.2 Filtro pasa-bajas

La Figura 4.2.1.3 muestra el filtro de Pre-énfasis, el cual es un filtro FIR pasa-altas. Este filtro cumple con la función de enfatizar las componentes de alta frecuencia dentro del rango que los filtros anteriores han definido, esto debido a que por naturaleza las componentes vocálicas tienden a atenuarse.

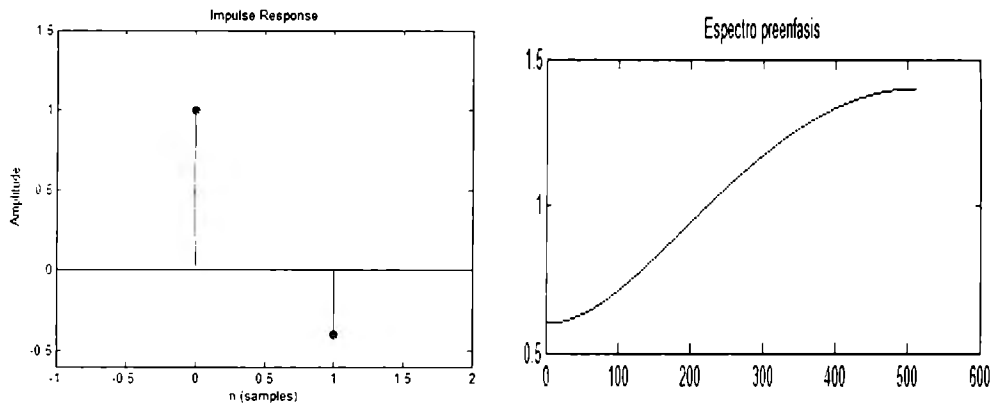


Figura 4.2.1.3 Filtro de pre-enfasis

4.2.2 Etapa de segmentación y detección de segmentos vocalizados

En estas dos etapas del sistema, se validó su correcto funcionamiento. Durante la etapa de segmentación se verificó que la señal fuera segmentada correctamente en los intervalos establecidos y que la ventana utilizada, en este caso de 30 ms, también realizara su función que es la de resaltar y colocar dentro de cada segmento la caracteristica más significativa en el centro del segmento.
En la etapa de detección de segmentos vocalizados se validó que los dos métodos utilizados para determinar si el segmento es vocalizado o no vocalizado, en este caso la potencia promedio y los cruces por cero, cumplen con su función y funcionan correctamente. Estos dos métodos cuentan con parámetros variables para adaptarlos en caso de no obtener los resultados esperados.

4.2.3 Etapa de extracción de caracteristicas

Durante esta etapa también se validó su funcionamiento por medio de varias pruebas e ir siguiendo paso a paso el código y su función. Dentro de esta etapa, se lograron obtener gráficas representativas de la función Wavelet. Esto se obtuvo modificando el código y haciendo algunas adaptaciones al mismo. Estas adaptaciones y modificaciones pueden observarse en el Anexo B, donde se presenta el codigo en MatLab.

4.2.4 Etapa de reconocimiento (Red neuronal)

En esta sección se realizaron muchas pruebas para mejorar el funcionamiento de la red neuronal en general. Se cambiaron varios de los parámetros, utilizando el método de prueba y error, hasta alcanzar los resultados deseados. Los resultados

más significativos que se obtuvieron dentro de la red neuronal, pueden observarse en el capítulo 3 en la sección 3.2 específicamente.
Todas las pruebas realizadas y los parámetros modificados, se pueden observar en el Anexo C.

4.3 Resultados ingresando una señal de voz especifica

A continuación se presentarán los resultados de todo el sistema, para una palabra especifica, esto con el objetivo de mostrar que el sistema funciona de manera eficiente y que se lograron alcanzar los objetivos específicos de este proyecto.
Se realizaron varias pruebas con varias palabras para demostrar lo anterior, sin embargo por razones de longitud y contenido repetitivo, en esta sección solo se presentará una sola palabra tanto de voz normal como de voz esofágica.
Para ver los resultados de las otras palabras, referirse al Anexo D situado al final de este documento.
La palabra que fue utilizada para el análisis, fue "a e i o u", esto con la finalidad de que se observen claramente todas las vocales, los valores de cada una de éstas y así observar que el sistema funciona correctamente para cada una de ellas.
La figura 4.3.1 muestra las señales originales que fueron cargadas, es la misma señal para voz normal y voz esofágica, la señal es la palabra "a e i o u". Estas señales aún no han pasado por ninguna etapa del sistema.

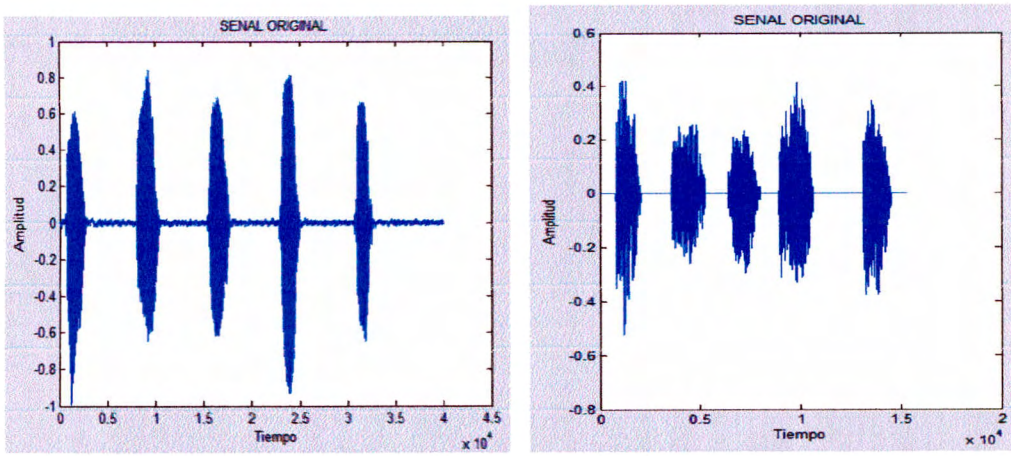


Figura 4.3.1 Señales de voz originales. A la izquierda señal de voz normal y a la derecha señal de voz esofagica

La Figura 4.3.2 muestra la señal después de la etapa de acondicionamiento, donde se eliminaron componentes de alta y baja frecuencia con los filtros antes

mencionados. Estas gráficas se muestran antes de la normalización, la cual consistirá en elevar la energía de la señal para igualarla a la de la entrada y mejorar el análisis posterior.

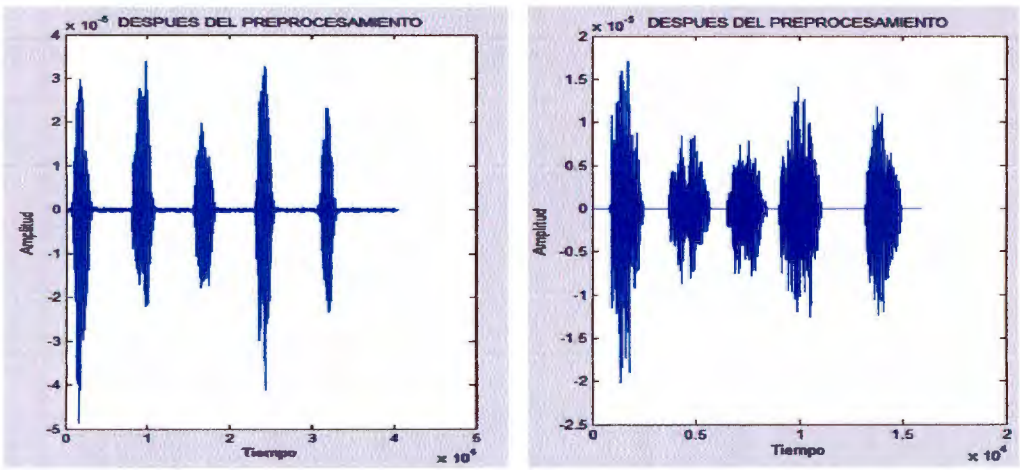


Figura 4.3.2 Señal de voz despues del acondicionamiento

En la Figura 4.3.3 se puede observar la potencia de la señal, de cada una de las vocales en específico. Con esto validamos que los segmentos vocalizados fueron selecciondos correctamente por medio de los dos criterios explicados anteriormente.

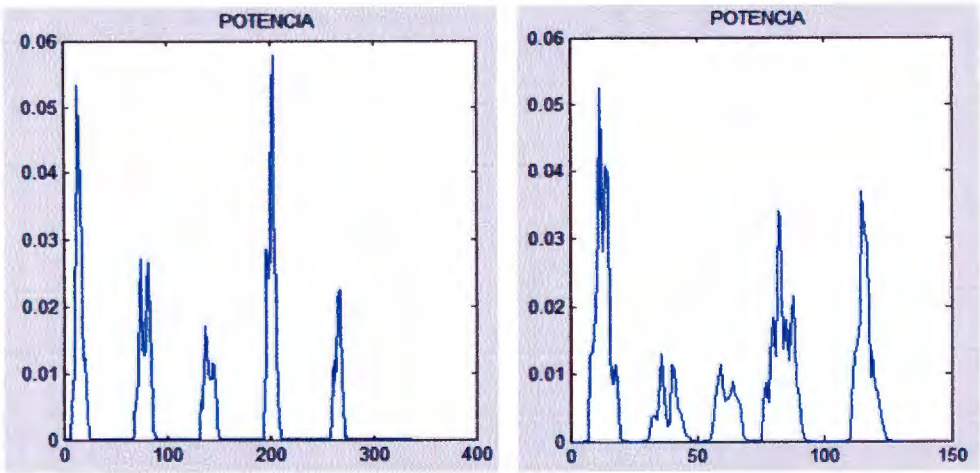


Figura 4.3.3 Potencia de los segmentos vocalizados

La Figura 4.3.4 muestra la extracción de características de la transformada Wavelet basada en el modelo del oído humano. Se puede observar que las características extraídas para este tipo de palabra corresponden a las vocales, lo que nos indica que el sistema trabaja correctamente y por lo tanto, la Wavelet propuesta funciona

correctamente. Estas gráficas se lograron obtener en 3D y están representadas en amplitud, escala y traslación, que son los parámetros establecidos de una Wavelet.

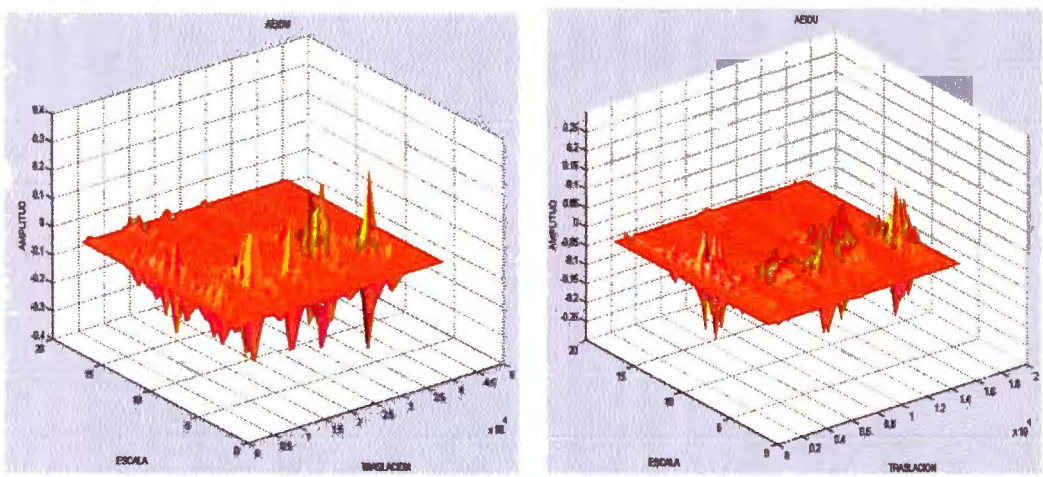


Figura 4.3.4 Extracción de características por medio de transformada Wavelet

4.4 Resultados de la etapa de reconocimiento al personalizar la red neuronal

En esta sección presentaremos los resultados obtenidos al personalizar la red neuronal. Esto es, entrenar la red neuronal con archivos de un sólo locutor y posteriormente clasificar de manera satisfactoria las palabras introducidas al sistema. Los resultados que se presentan a continuación son de un locutor de voz normal.

El locutor y las palabras que se grabaron para este análisis, fueron de un hombre de 21 años, de voz normal.

En la Tabla 4.4.1 se presenta la comparación del entrenamiento de la red con el training data que se tenía anteriormente y con el training data del locutor especificado en esta sección.

La red se entrenó con 25 archivos de cada vocal, es decir, 125 archivos en total. Se puede observar que el porcentaje de las vocales reconocidas, para este grupo específico de palabras, aumentó de un 52% a un 55% es decir el reconocimiento por vocal se elevó en un 3%.

Palabra	*Antes*	*Personalizado*
Bebida	_a_e_a_	_e_e_a_
Aire	_a_e_	_a_e_
Bocina	_oi_a_	_u_i_a_
Carro	_a_	_a_
Chile	_e_	_e_
Droga	_o_a_	_o_a_
Fuego	_a_o_	_a_o_
Formula	_a_a_	_a_a_
Hoja	_a_a_	_o_a_
Hola	_a_a_	_o_a_
León	_ea_	_ea_
Libro	_ea_	_ea_
Luz	_o_i_	_ao_
Música	_i_a_	_ai_a_
Paz	_a_	_a_
Radio	_a_e_	_a_e_
Sexo	_a_a_	_a_e_
Técnica	_e_i_a_	_e_a_a_
Reconocimiento	**52%**	**55%**

Tabla 4.4.1 Comparación antes y después de personalizar el training data

Posteriormente para determinar si mejoraba el reconocimiento de alguna manera, se agregaron 25 archivos más por vocal al entrenamiento de la red neuronal, en total se agregaron 250 archivos de entrenamiento, 50 por cada vocal.

En la Tabla 4.4.2 se muestran los resultados obtenidos de la comparación entre el entrenamiento anterior y el personalizado. Donde claramente se puede observar que el reconocimiento mejoró considerablemente en un 12% con respecto al reconocimiento anterior. Este mejoramiento en el reconocimiento, se lo atribuimos al número de archivos de entrenamiento que se introdujeron.

Palabra	*Antes*	*Personalizado*
Bebida	_a_e_a_	_e_a_
Aire	_a_e_	_a_e_e_
Bocina	_oi_a_	_e_e_i_a_
Carro	_a_	_a_
Chile	_e_	_e_e_
Droga	_o_a_	_o_a_
Fuego	_a_o_	_o_o_
Formula	_a_a_	_o_o_a_
Hoja	_a_a_	_o_a_
Hola	_a_a_	_o_a_
León	_ea_	_eo_
Libro	_ea_	_eao_
Luz	_o_i_	_o_i_
Música	_i_a_	_a_e_a_
Paz	_a_	_a_
Radio	_a_e_	_a_a_
Sexo	_a_a_	_e_a_
Técnica	_e_i_a_	_e_i_a_
Reconocimiento	**52%**	**64%**

Tabla 4.4.2 Comparación antes y después de introducir más archivos de entrenamiento

En este capítulo mostramos los resultados obtenidos a lo largo del desarrollo de este proyecto. Todos los resultados fueron favorables ya que se lograron los objetivos establecidos. El proyecto se basó principalmente en la validación y en la obtención de resultados específicos, los cuales están especificados y respaldados en esta sección del presente documento.

Con este capítulo se finaliza este documento, dando pie a avances posteriores para el mejoramiento del proyecto.

Conclusiones

Este proyecto propone la contribución a un sistema que ayude a mejorar la calidad de la voz a las personas que se les ha extirpado la laringe y hablan ahora por medio del método de voz esofágica. Para el desarrollo de este proyecto se conjuntaron conceptos sobre fonética, acústica, anatomía del oído humano, análisis espectral de una señal de audio y la descomposición que lleva a cabo la membrana basal. Otros conceptos fundamentales para el desarrollo de este trabajo fue la herramienta matemática que nos proporciona la transformada Wavelet, la arquitectura de una red neuronal y su funcionamiento para llegar al reconocimiento de voz.

Este algoritmo de reconocimiento de voz esofágica está enfocado directamente a clasificar por medio de la red neuronal las cinco vocales del español. Cada uno de los códigos presentados en este proyecto y simulados en Matlab fue validado mediante pruebas para obtener información valiosa para, más adelante, poder enfocarnos en mejorar el desempeño del sistema y que el reconocimiento de cada una de las vocales fuera aún más efectivo.

En la etapa de acondicionamiento de la señal se diseñaron filtros tipo FIR pasa altas, pasa bajas y de pre-énfasis cuya frecuencia de corte están en 100Hz, 900Hz y 300Hz respectivamente. Al llevar a cabo el filtrado de la señal se logró quitar el ruido de frecuencias de ruido de 60Hz, las que estuvieran por debajo de los 100Hz y por encima de los 900Hz. Este rango de frecuencias es porque la voz no es inteligible arriba de 1KHz y por debajo de 100Hz. Con esto tenemos como resultado una señal más limpia cuyos componentes vocálicos muestran mayor energía y por lo tanto se pueden explotar sus características para lograr llevarlas a un sistema de clasificación. Finalmente con la normalización de la señal se logra mantener la cantidad de energía contenida en la señal original.

Más adelante, se llevó al sistema a detectar cuáles eran sus segmentos vocalizados y no vocalizados con el fin de eliminar los segmentos en donde no haya una vocal involucrada. Al tener en esta etapa resultados satisfactorios en ese sentido, se prosigue con la descomposición de la señal para obtener los coeficientes wavelet y con ello, la extracción de parámetros.

La transformada wavelet nos proporciona información útil de la señal tanto en traslación, escala y amplitud. En esta etapa se obtuvo la gráfica en tres dimensiones de la transformada wavelet de la señal ya acondicionada. Esta herramienta nos fue de gran ayuda visualmente para el análisis de una señal, dicha gráfica nos muestra como

es que a menor escala, tenemos mayor frecuencia. Los coeficientes wavelet obtenidos y graficados muestran el comportamiento de la señal y lo más importante, permiten visualizar fácilmente las vocales que se tiene en una palabra o en una frase al analizar la amplitud de energía que se tiene en cada una.

Llegando a la etapa de clasificación de segmentos vocalizados se hicieron muchas pruebas, para empezar se logró reducir el error cuadrático medio al término del entrenamiento de la red neuronal, es decir, con los parámetros originales del sistema de entrenamiento se tenía un error de 7.98% hasta un 1.14% en dicho resultado influyó que lo parámetros de entrenamiento utilizados fueron más estrictos que los establecidos. Sin embargo, el haber reducido el error cuadrático medio en el entrenamiento, no quiere decir que se va a reconocer el 98.86%.

Seguido del entrenamiento, se llevaron a cabo las pruebas ya de reconocimiento de voz y clasificación de segmentos vocalizados por medio de la red neuronal. En el funcionamiento original de la red teníamos un porcentaje de reconocimiento correcto para voz esofágica de 80.87% y para voz normal un 77.06%, después de llevar a cabo modificaciones al sistema logramos incrementar ambos porcentajes en un 6%, llegando a un porcentaje de reconocimiento de 86% para voz esofágica y del 83% para voz normal.

Posteriormente se llevó a cabo la personalización de la red, es decir, entrenar la red con archivos de voz de un locutor para reconocer a ese mismo locutor. En esta parte introdujimos a la red diferentes palabras como se mencionó en el capítulo de resultados, una serie de diferentes palabras fueron grabadas al igual que archivos de entrenamiento, los resultados de esto fue que para ese conjunto de palabras se tenía un porcentaje de reconocimiento del 52% con los archivos de entrenamiento que ya se tenían, al personalizar la red ese porcentaje de reconocimiento se incrementó en un 3% para llegar a un 55%. Ese incremento demostró que si se personaliza la red a cada locutor, ésta mejora su desempeño. Sin embargo al añadir aún mas archivos de entrenamiento a la red neuronal para el mismo locutor, el porcentaje de reconocimiento incrementó considerablemente en un 12%.

La parte anterior demuestra que es posible personalizar la red neuronal y tener resultados de clasificación más satisfactorios que si la red se generaliza a cualquier locutor.

A continuación se analizaron los resultados obtenidos en las pruebas realizadas para validar la función wavelet con otras funciones.

El estudio estadístico hecho para demostrar que la función basada el modelo del oído humano está por encima de otras funciones wavelet, cuya aplicación es similar a la desarrollada en este proyecto, fue todo un éxito al mostrarnos que los porcentajes de reconocimiento correcto siempre tuvo un valor mayor a las demás como se mostró en el capítulo de resultados.

Este resultado no fue ninguna sorpresa, ya que esta función wavelet fue desarrollada para que trabajara tal y como lo hace el oído humano, a diferencia de las demás wavelets ya eran funciones propuestas para análisis de aplicaciones semejantes, pero ninguna de ellas específicamente para el reconocimiento de voz esofágica.

Trabajo a futuro

Ya que el trabajo de validación se ha realizado para este sistema, deduciendo que la red de entrenamiento incrementa su buena funcionalidad al personalizarse, se debe obtener una gran cantidad de archivos de entrenamiento según la persona cuya voz deba ser reconocida. Estos archivos de entrenamiento deberán ser los más limpios posibles de ruido y grabados en condiciones óptimas para evitar errores en la clasificación provenientes de las grabaciones.

A largo plazo se tiene en mente llevar estos algoritmos a un DSP para no dejar este sistema en una simulación solamente y demostrar su funcionamiento físicamente desarrollada en hardware.

Finalmente, se mencionó anteriormente que la arquitectura de red utilizada es el perceptrón, esta arquitectura es la más básica de una red neuronal por lo que una prueba contundente para el sistema sería la de variar esa arquitectura para ver si el reconocimiento de la misma mejora.

Anexo A

Tabla de resultados obtenidos de cada una de las pruebas hechas suministrando al sistema los 107 archivos de audio.

VOZ NORMAL

Palabra	ear	haar	daub4	daub 10	mexican hat	morlet	meyer
	voz normal				voz normal		
H23							
abeja	_a_e_a_	_e_a_e_	_a_e_e_	_a_e_a_	_a_e_a_	_a_e_a_	_a_e_a_
abrigo	_a_i_o_	_o_i_u_	_a_i_u_	_a_i_u_	_a_i_u_	_a_i_o	_a_i_u_
adicto	_a_i_o_	_i_i_u_	_a_i_o_	_a_i_u_	_a_i_u_	_a_i_o_	_a_i_u_
aniou	_a_e_i_o_u_	_a_a_a_a_a_	_a_e_i_o_u_	_a_e_i_o_u_	_a_e_i_o_u_	_a_e_i_o_u_	_a_e_i_o_u_
ahilar	_ai_a_	_e_a_	_ae_a_	_ai_a_	_ai_a_	_ai_a_	_ai_a_
bala	_a_a_	_e_e_	_a_a_	_a_a_	_a_a_	_a_a_	_a_a_
bebido	_e_i_o_	_a_i_o_	_e_i_u_	_e_i_u_	_e_i_u_	_e_i_o_	_e_i_u_
beso	_e_a_	_a_i_	_o_a_	_o_a_	_o_a_	_e_a_	_o_a_
cupo	_u_o_	_u_i_	_u_u_	_u_u_	_u_u_	_u_o_	_u_o_
duda	_u_a_	_u_o_	_u_o_	_u_a_	_u_a_	_u_a_	_u_a_
foco	_o_o_	_o_o_	_o_o_	_o_o_	_o_o_	_o_o_	_o_o_
agua	_a_o_	_u_u_	_u_i_	_a_o_	_a_a_	_a_o_	_a_o_
amarillo	_a_i_o_	_u_u_u_	_i_ui_i_	_a_e_o_	_a_a_o_	_a_e_o_	_a_e_o_
ángel	_a_e_	_u_u_	_i_i_	_a_e_	_a_e_	_a_e_	_a_e_
armonía	_a_o_i_	_u_u_u_	_u_u_u_	_a_o_i_	_a_o_i_	_a_o_i_	_a_o_i_
atmosfera	_a_a_e_	_u_u_u_	_a_i_iuiui_	_a_a_e_	_a_a_ea_	_a_a_ea_	_a_a_ea_
balón	_ao_	_u_	_iui_	_ao_	_ao_	_ao_	_ao_
cangrejo	_a_e_a_	_u_u_u_	_i_u_u_	_a_e_a_	_a_e_a_	_a_e_o_	_a_e_o_
cantidad	_a_e_a_	_u_u_u_	_u_u_u_	_i_e_a_	_a_i_a_	_a_e_a_	_a_e_a_
carretera	_a_ea_	_u_u_	_i_u_	_a_ea_	_a_ea_	_a_ea_	_a_ea_
contador	_o_a_o_	_u_e_u_	_u_u_u_	_o_a_o_	_o_a_o_	_o_a_o_	_o_a_o_
cuadro	_a_o_	_u_u_	_i_i_	_a_a_	_a_a_	_a_o_	_a_o_
cubeta	_o_e_a_	_u_u_u_	_i_i_e_	_o_e_a_	_u_e_a_	_a_e_a_	_o_e_a_
disciplina	_e_i_e_	_u_u_u_	_u_u_u_	_e_i_e_	_e_i_e_	_e_i_e_	_e_i_e_
difícil	_e_i_i_	_u_u_u_	_i_i_i_	_e_e_i_	_e_e_i_	_e_e_i_	_e_e_i_
dinero	_i_eo_	_u_u_	_i_ui_	_i_eo_	_i_eo_	_i_eo_	_i_eo_
distorsión	_e_o_ia_	_u_u_u_	_i_i_i_	_e_o_ia_	_e_o_ia_	_e_o_ia_	_e_o_ia_
economía	_e_o_i_	_u_u_u_	_u_i_iu_	_e_o_i_	_e_o_i_	_e_o_i_	_e_o_i_
adiós	_a_a_	_u_u_	_i_i_	_a_ae_	_a_a_	_a_o_	_a_o_
electrodo	_i_a_o_	_u_u_u_	_i_i_i_	_i_a_o_	_i_a_o_	_i_a_o_	_i_a_o_
espionaje	_e_a_a_e_	_u_u_u_u_	_u_u_u_u_	_e_e_i_e_	_e_e_a_e_	_e_e_i_e_	_e_e_a_e_
flor	_ao_	_u_	_iu_	_ao_	_ao_	_ao_	_ao_
fuego	_oa_o_	_u_u_	_u_u_	_a_o_	_oao_o_	_o_o_	_a_o_
guerra	_e_a_	_u_u_	_i_i_	_e_a_	_e_a_	_e_a_	_e_a_
guitarra	_i_a_a_	_u_u_u_	_i_i_u_	_i_a_a_	_i_a_a_	_i_a_a_	_i_a_a_
herramienta	_ea_i_a_	_u_u_u_	_u_u_u_	_ei_i_a_	_ea_i_a_	_ea_i_a_	_ea_i_a_
hola	_a_a_	_u_u_	_i_i_	_a_a_	_a_a_	_a_a_	_a_a_
holocausto	_a_a_o_	_u_u_u_	_i_i_i_	_aoa_a_a_	_a_a_o_	_aoa_a_o_	_aoa_a_o_
indio	_i_io_	_u_u_	_i_ui_	_i_iaeo_	_i_ieo_	_i_ieo_	_i_ieo_
información	_i_o_a_e_	_u_u_u_u_	_u_u_u_u_	_i_o_a_e_	_i_o_a_a_	_e_o_a_a_	_i_o_a_a_

Anexo A

ingeniero	_i_e_ie_	_u_u_u_	_u_u_u_	_i_e_ie_	_i_e_ieo_	_i_e_ieo_	_i_e_ieo_
jamaica	_a_a_a_	_u_u_u_	_u_i_u_	_a_a_a_	_a_a_a_	_a_a_a_	_a_a_a_
judío	_o_io_	_u_u_	_i_i_	_o_io_	_o_io_	_o_ieo_	_o_io_
llaves	_a_	_u_	_iuiu_	_a_	_a_	_a_	_a_
máscara	_a_a_	_u_u_	_u_u_	_a_a_	_a_a_	_a_a_	_a_a_
mujer	_o_e_	_u_u_	_u_u_	_o_e_	_o_e_	_o_e_	_o_e_
navidad	_a_a_	_u_u_	_a_u_	_a_a_	_a_a_	_a_a_	_a_a_
nubes	_o_e_e_	_e_u_u_	_u_i_a_	_o_e_e_	_o_e_e_	_o_e_e_	_o_e_e_
perfume	_e_oe_	_u_u_	_i_ui_	_e_oe_	_e_oe_	_e_eo__	_e_eo__
pescado	_e_a_o_	_u_u_u_	_a_i_u_	_e_a_o_	_e_a_o_	_e_a_o_	_e_a_o_
química	_i_a_	_u_u_	_i_e_	_i_a_	_i_a_	_i_e_	_i_a_
recompensa	_e_o_e_a_	_u_u_u_u_	_u_i_u_i_	_e_o_e_a_	_e_o_e_a_	_e_o_e_a_	_e_o_e_a_
reproductor	_a_a_a_	_u_u_u_	_i_i_u_	_a_a_a_	_a_a_a_	_a_a_a_	_a_a_a_
revolución	_e_o_u_a_	_u_u_u_u_	_u_u_u_u_	_e_o_o_a_	_e_o_u_a_	_e_o_u_a_	_e_o_u_a_
sonido	_i_i_o_a_	_u_u_u_u_	_i_a_u_i_	_i_i_o_a_	_i_i_o_a_	_i_i_o_a_	_io_i_o_e_
uniforme	_u_i_oae_	_u_u_u_	_u_u_u_	_u_i_oae_	_u_i_oae_	_u_i_oae_	_u_i_oae_
universidad	_o_a_a_	_u_u_u_	_u_u_u_	_o_e_a_	_e_a_a_	_o_a_a_	_o_a_a_
vino	_i_o_	_u_u_	_i_i_	_i_o_	_i_o_	_i_o_	_i_o_
violeta	_ieie_a_	_u_u_	_u_u_	_ieie_a_	_ieie_a_	_ie_i_	_ieie_a_
violín	_ei_	_u_	_i_	_ei_	_ei_	_ei_	_ei_

VOZ ESOFÁGICA

Palabra	ear	haar	daub4	daub 10	mexican hat	morlet	meyer
H30							
abeja	_ae_a_	_ao_a_	_ae_a_	_ae_a_	_ae_a_	_ae_a_	_ae_a_
abrazar	_a_a_a_	_a_a_a_	_a_a_a_	_a_a_a_	_a_a_a_	_a_a_a_	_a_a_a_
adicto	_a_i_o_	_a_o_a_	_a_e_a_	_a_i_o_	_a_i_o_	_a_i_o_	_a_i_o_
aeiou	_a_e_i_o_u_	_a_a_i_a_o_	_a_i_i_i_i_	_a_e_i_o_u_	_a_e_i_o_u_	_a_e_i_o_u_	_a_e_i_o_u_
ahilar	_a_a_	_ao_o_	_ao_a_	_a_a_	_a_a_	_a_a_	_a_a_
bache	_a_e_	_a_a_	_a_ei_	_a_uia_	_a_oi_	_a_e_	_a_e_
bebido	_e_i_o_	_o_o_o_	_i_i_o_	_e_i_o_	_e_i_o_	_e_i_o_	_e_i_o_
bruma	_u_a_	_o_a_	_e_a_	_u_a_	_u_a_	_u_a_	_u_a_
casa	_a_o_	_i_i_	_i_i_	_a_o_	_a_o_	_a_o_	_a_a_
duda	_u_a_	_a_a_	_a_a_	_u_o_	_u_o_	_u_o_	_u_a_
foco	_o_o_	_o_a_	_o_a_	_o_o_	_o_o_	_o_o_	_o_o_
abogado	_a_o_a_o_	_a_a_o_a_	_a_o_a_a_	_a_o_a_o_	_a_o_a_o_	_a_o_a_o	_a_o_a_o
abolir	_a_o_oe_	_a_o_oao_	_a_o_e_	_a_o_e_	_a_o_eo_	_a_o_oe_	_a_o_e_
abono	_a_o_a_	_a_o_a_	_a_a_a_	_a_o_a_	_a_o_a_	_a_o_a_	_a_o_a_
absoluto	_a_o_o_o_	_a_a_a_a_	_a_o_a_a_	_a_o_o_o_	_a_o_o_o_	_a_o_o_o_	_a_o_o_o_
acetato	_a_o_a_o_	_a_a_a_a_	_a_o_a_a_	_a_o_a_o_	_a_o_a_o_	_a_o_a_o_	_a_o_a_o_
acidez	_a_e_o_	_a_a_a_	_a_a_o_	_a_e_o_	_a_e_o_	_a_e_o_	_a_e_o_
abusar	_a_o_a_	_a_a_a_	_a_a_a_	_a_e_a_	_a_o_a_	_a_o_a_	_a_e_a_
acierto	_a_a_e_o_	_a_a_a_a_	_a_a_a_a_	_a_a_e_o_	_a_a_e_o_	_a_a_e_o_	_a_a_e_o_
acreditar	_a_o_o_a_	_a_a_a_a_	_a_u_a_a_	_a_o_o_a_	_a_o_o_a_	_a_o_o_a_	_a_o_o_a_
adaptar	_a_a_a_	_a_o_a_	_a_a_a_	_a_a_a_	_a_a_a_	_a_a_a_	_a_a_a_
adepto	_a_e_o_	_a_a_a_	_a_e_a_	_a_e_o_	_a_e_o_	_a_e_o_	_a_e_o_

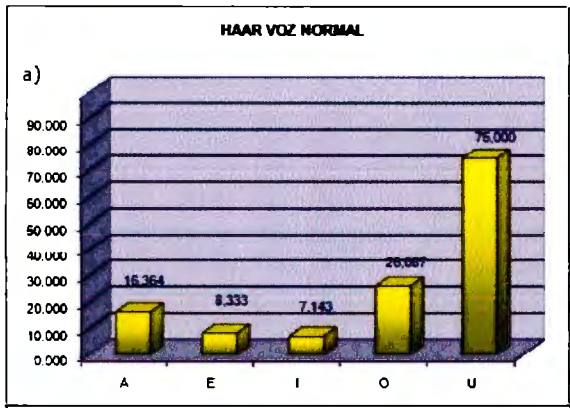
Anexo A

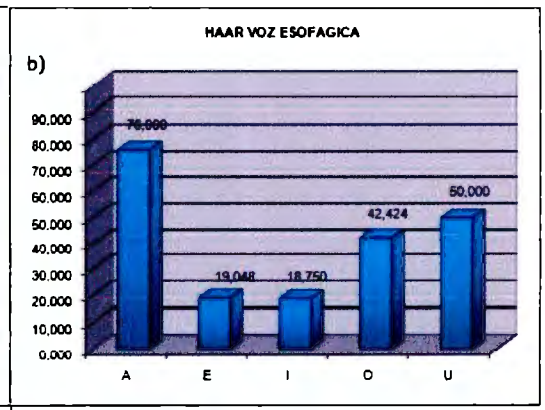
acustica	_a_o_o_o_a_	_a_u_a_a_a_	_a_o_u_a_a_	_a_o_o_o_a_	_a_o_o_o_a_	_a_o_o_o_a_	_a_o_o_o_a_
actitud	_a_o_o_	_a_a_o_	_a_a_a_	_a_o_o_	_a_o_o_	_a_o_o_	_a_o_o_
adios	_a_	_a_	_a_	_a_	_a_	_a_	_a_
aditivo	_a_o_e_o_	_a_a_a_a_	_a_a_a_a_	_a_o_e_o_	_a_o_e_o_	_a_e_e_o_	_a_o_e_o_
adobo	_a_o_o_	_a_o_o_	_a_o_o_	_a_o_o_	_a_o_o_	_a_o_o_	_a_o_o_
adolecer	_a_o_e_e_	_a_a_a_a_	_a_a_a_a_	_a_o_e_e_	_a_o_e_e_	_a_o_e_e_	_a_o_e_e_
boton	_o_o_	_o_o_	_o_o_	_o_o_	_o_o_	_o_o_	_o_o_
bruja	_u_a_	_u_o_	_u_a_	_u_a_	_u_a_	_u_a_	_u_a_
buche	_u_e_	_u_a_	_u_e_	_u_e_	_u_e_	_u_oe_	_u_e_
buro	_u_o_	_u_o_	_u_o_	_u_o_	_u_o_	_u_o_	_u_o_
consultorio	_o_o_o_	_i_i_i_	_i_i_i_	_o_o_o_	_o_o_o_	_o_o_o_	_o_o_o_
cubeta	_o_e_a_	_o_o_a_	_o_o_a_	_o_e_a_	_o_e_a_	_o_e_a_	_o_e_a_
cubo	_o_u_o_	_o_o_o_	_o_u_e_	_o_u_o_	_o_u_o_	_o_u_o_	_o_u_o_
dilema	_i_i_a_	_a_o_a_	_a_o_a_	_e_e_a_	_e_e_a_	_o_i_a_	_e_e_a_
dolar	_o_a_	_i_i_	_i_i_	_o_a_	_o_a_	_o_a_	_o_a_
dulce	_u_o_	_o_a_	_o_a_	_u_o_	_u_o_	_u_o_	_u_o_
duro	_u_o_	_e_a_	_o_a_	_u_o_	_u_o_	_u_o_	_u_o_
dureza	_o_e_a_	_o_a_o_	_o_o_a_	_o_e_a_	_o_e_a_	_o_e_a_	_o_e_a_
faro	_ao_	_a_	_i_	_ao_	_ao_	_a_	_ao_
lodo	_o_o_	_i_i_	_i_i_	_o_o_	_o_o_	_o_o_	_o_o_
lugar	_a_o_	_i_i_	_i_i_	_a_o_	_a_o_	_a_o_	_a_o_
lupa	_u_a_	_i_i_	_i_i_	_u_a_	_u_a_	_u_a_	_u_a_
parcela	_a_e_a_	_i_i_i_	_i_i_i_	_a_e_a_	_a_e_a_	_a_e_a_	_a_e_a_
pared	_a_e_	_i_i_	_i_i_	_a_e_	_a_e_	_a_e_	_a_e_
vidente	_i_e_e_	_i_i_i_	_i_i_i_	_i_e_o_	_i_e_e_	_i_e_e_	_i_e_e_

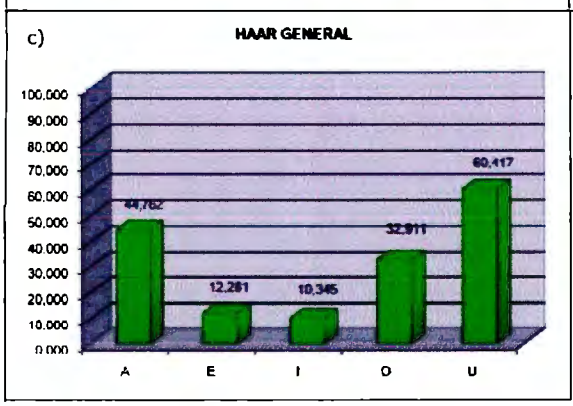
A continuación se presentan los resultados obtenidos para cada función wavelet correspondientes al estudio estadístico de cada vocal por método.

A) *WAVELET DE HAAR*

Anexo A







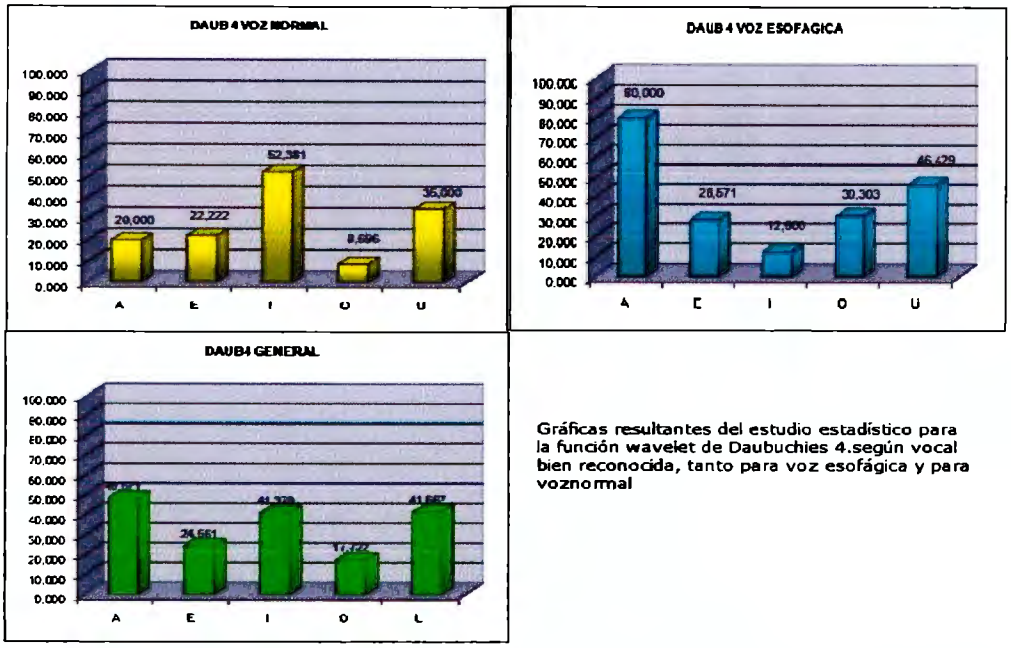
a) Muestra el promedio de las vocales reconocidas correctamente con la funcion wavelet de Haar para voz normal.

b)Muestra el porcentaje de vocales reconocidas correctamente para voz esofágica con la wavelet de Haar.

c)Promedio general del reconocimiento para la waveet de Haar por vocal.

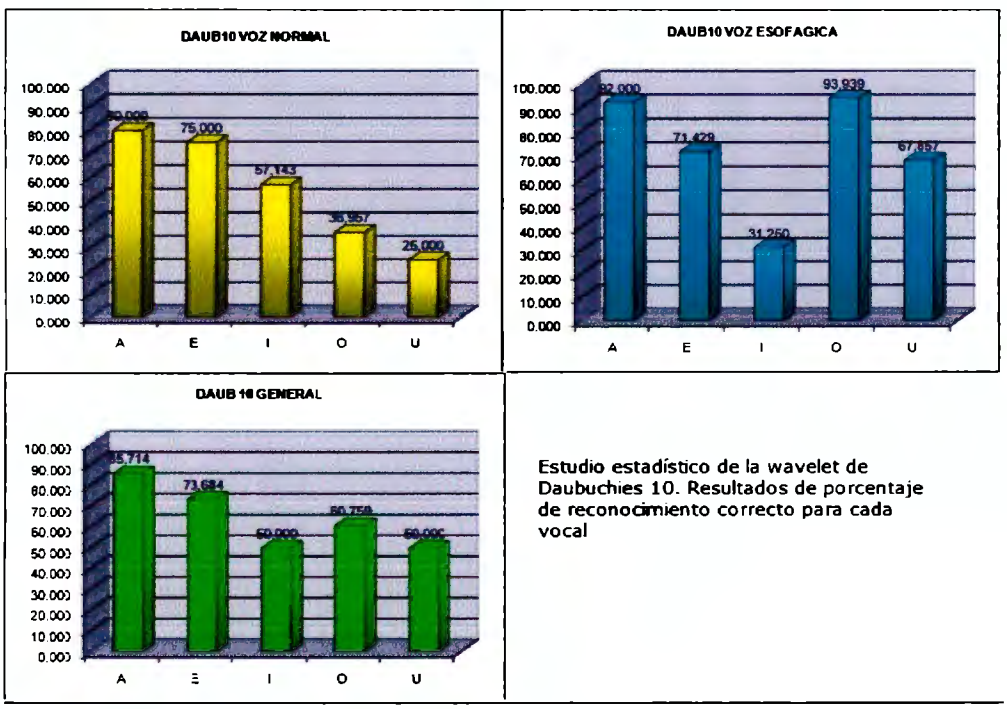
FUNCIÓN WAVELET DABUCHIES4

Anexo A



Gráficas resultantes del estudio estadístico para la función wavelet de Daubuchies 4.según vocal bien reconocida, tanto para voz esofágica y para voznormal

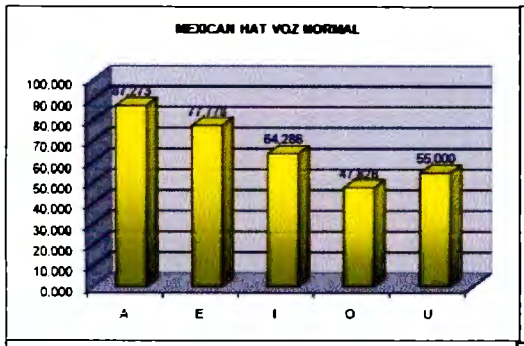
FUNCIÓN WAVELET DABUCHIES10

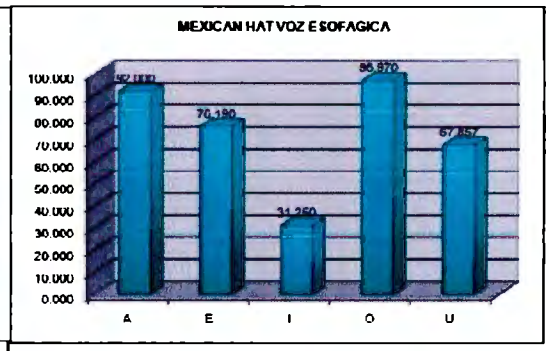


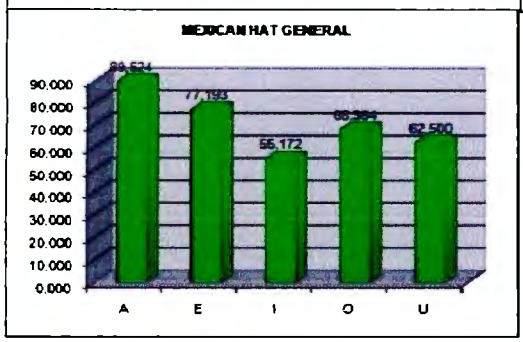
Estudio estadístico de la wavelet de Daubuchies 10. Resultados de porcentaje de reconocimiento correcto para cada vocal

FUNCIÓN WAVELET MEXICAN HAT

Anexo A

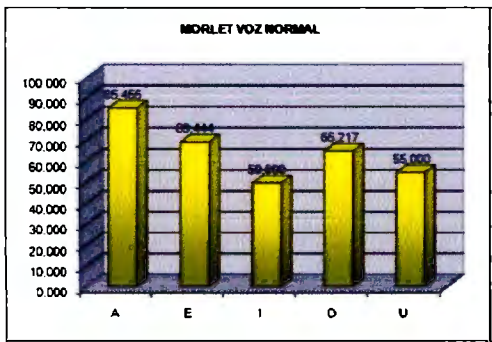


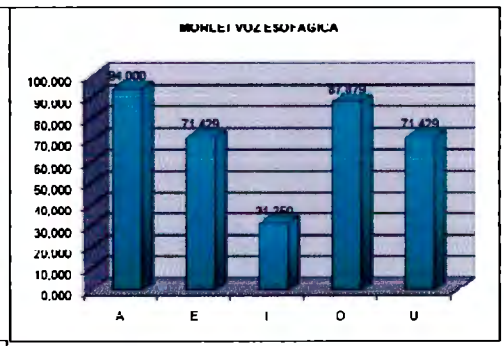


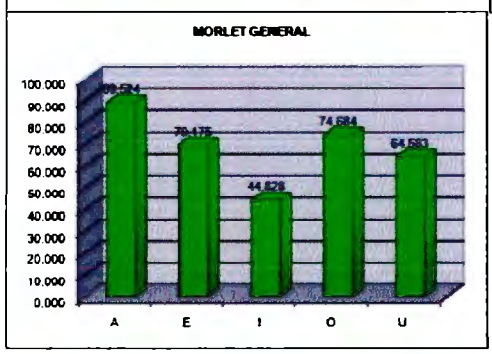


Resultados del estudio estadístico realizado para cada vocal en cada método.
Podemos ver que la voal "a" es la mejor reconocida en esta función.

FUNCIÓN WAVELET MORLET



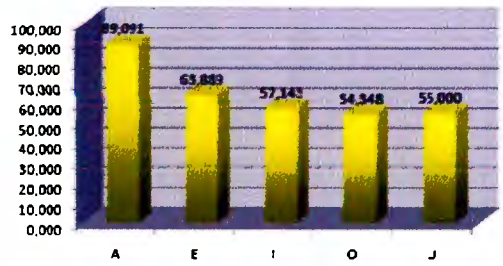


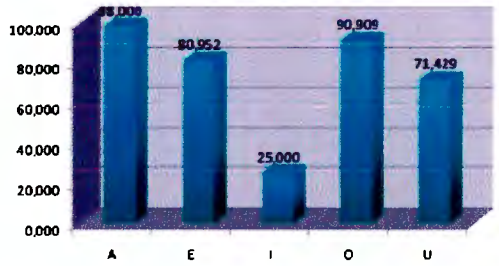


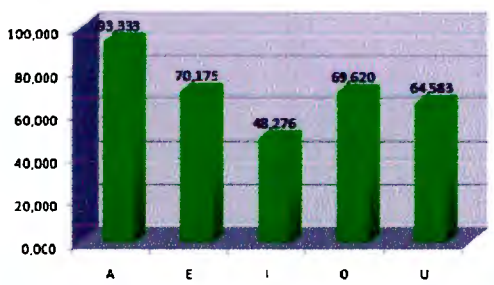
Gráfica de resultados para el estudio estadístico de las vocales tanto en voz esofágica y voz norma.
Finalmente se encuentran los resultados generales para esta función wavelet.

FUNCIÓN WAVELET MEYER

Anexo A







Resultados estadísticos para la función wavelet de Meyer.
a) Voz normal

b) Voz esofágica

c) Premedio general.

Se presenta el código en Matlab con el cual se logro hacer la extracción de características y mostrarlas en 3D. Las graficas que salen producto de este código, están representadas en amplitud, escala y traslación.

Plot Wavelet 3D

```
function plotWavelet(data,fs,method,n,maxScales);
   %data=load('data/ui/data.txt');
   data=wavread('audio/H23/abeja.wav');
   method=load('data/method.txt');
   n=17;
   fs=8000;
   data=data(:);
   waveletDecomposition=[];
 if(strcmp(method,'filters'))
      for(i=1:n)
         if(exist(['data\filters\B' num2str(i) '.txt'],'file'))
            B=load(['data\filters\B' num2str(i) '.txt']);
            signalWindow=conv(data,B)';

waveletDecomposition=[waveletDecomposition;signalWindow(1:length(data))];
         end;
      end;
   else
      if(nargin<5)
         maxScales=1;
      end;
      for(Z=1:n)
         fc=325*(exp(2*Z/7)-1)./exp(Z/7);
         fprintf('%d) ',Z);
         save data/fc.txt fc   -ascii;
         fc2(Z)=fc;

waveletDecomposition=[waveletDecomposition;earDWT2(data,fs,fc,3,0.03,0.001,1,
[],maxScales)'];
         save data/waveletDecomposition.txt waveletDecomposition -ascii;

       end;
   end;
   if(length(waveletDecomposition(:,1))<n)
      waveletDecomposition=[waveletDecomposition;zeros(n-
length(waveletDecomposition(:,1)),length(waveletDecomposition(1,:)))];
   end;
         save data/fc2.txt fc2   -ascii;


      coef=load('data/waveletDecomposition.txt');
     % subplot(1,1,1),plot(coef),title('wave')

Z=coef
mesh(Z)
colormap hot
xlabel('TRASLACION')
ylabel('ESCALA')
zlabel('AMPLITUD')
```

```
%title('CUPO')
axis auto
```

A continuación se presentan las modificaciones que se hicieron ala red neuronal para lograr mejorar su funcionamiento. Se muestran los parámetros modificados y los resultados obtenidos para esos parámetros para el fin antes descrito. Solo se aplicaron a un cierto número de palabras para observar el cambio.

Prueba 1:

```
%Error cuadratico medio deseado
  errorMin=0.1;
  %Factor de aprendizaje
  learningRatio=0.05
  %Coeficiente de incremento para el factor de aprendizaje en caso de
  %reducir el error
  learningRatioI=1.20;
  %Coeficiente de decremento para el factor de aprendizaje en caso de
  %aumentar el error mas alla de lo tolerado
  learningRatioD=0.7;
  %Peso dado al cambio anterior en los pesos y desplazamientos para
  %calcular el cambio siguiente en los mismos parametros (Solo en caso de
  %que el error no aumente más de lo tolerado)
  delayRatioI=0.95;
  %Maximo incremento en el error tolerado
  errorRatio=1.10;
```

Palabra	Ear nuevo	Prueba 1 antes
abeja	_a_a_	_a_e_a_
abrigo	_a_i_	_a_i_o_
adicto	_a_i_o_	_a_i_o_
ahilar	_ai_a_	_ai_a_
bala	_a_	_a_a_
bebido	_e_i_o_	_e_i_o_
besar	_o_a_i_	_e_a_
cupo	_u_o_	_u_o_
duda	_u_a_	_u_a_
foco	_o_o_	_o_o_

Prueba 2:

```
%Error cuadratico medio deseado
  errorMin=0.3;
  %Factor de aprendizaje
  learningRatio=0.08
  %Coeficiente de incremento para el factor de aprendizaje en caso de
  %reducir el error
  learningRatioI=1.22;
  %Coeficiente de decremento para el factor de aprendizaje en caso de
  %aumentar el error mas allá de lo tolerado
```

learningRatioD=0.7;
%Peso dado al cambio anterior en los pesos y desplazamientos para
%calcular el cambio siguiente en los mismos parametros (Solo en caso de
%que el error no aumente más de lo tolerado)
delayRatioI=1;
%Maximo incremento en el error tolerado
errorRatio=1.10;

No se entreno la red de forma correcta por lo que no se obtuvieron resultados.

Prueba 3:

2000 iteraciones
%Error cuadratico medio deseado
errorMin=0.001;
%Factor de aprendizaje
learningRatio=0.001
%Coeficiente de incremento para el factor de aprendizaje en caso de
%reducir el error
learningRatioI=1.10;
%Coeficiente de decremento para el factor de aprendizaje en caso de
%aumentar el error mas alla de lo tolerado
learningRatioD=0.8;
%Peso dado al cambio anterior en los pesos y desplazamientos para
%calcular el cambio siguiente en los mismos parámetros (Solo en caso de
%que el error no aumente mas de lo tolerado)
delayRatioI=.95;
%Maximo incremento en el error tolerado
errorRatio=1.05;

Error inicial: 153.433, Error después del entrenamiento: 4.11476

Palabra	Ear nuevo	Prueba 3 antes
adios	_a_a_	_a_a_
agua	_a_o_	_a_o_
amarillo	_a_e_o_	_a_e_o_
ángel	_a_e_	_a_e_
armonía	_a_o_u_u_a_	_a_o_i_
atmósfera	_a_a_ea_	_a_a_e_
balón	_a_	_ao_
cangrejo	_a_e_a_	_a_e_a_
cantidad	_a_e_a_	_a_e_a_
carretera	_a_ea_	_a_ea_
contador	_o_a_o_	_o_a_o_
dificil	_i_i_i_e_	_e_i_i_
fuego	_a_o_	_oa_o_
indio	_i_io_	_i_ieo_
uniforme	_u_i_o_e_	_u_i_oae_

Prueba 4:

8000 iteraciones
%Error cuadratico medio deseado
errorMin=0.001;
%Factor de aprendizaje
learningRatio=0.001
%Coeficiente de incremento para el factor de aprendizaje en caso de
%reducir el error
learningRatioI=1.10;
%Coeficiente de decremento para el factor de aprendizaje en caso de
%aumentar el error mas alla de lo tolerado
learningRatioD=0.8;
%Peso dado al cambio anterior en los pesos y desplazamientos para
%calcular el cambio siguiente en los mismos parámetros (Solo en caso de
%que el error no aumente más de lo tolerado)
delayRatioI=.95;
%Maximo incremento en el error tolerado
errorRatio=1.05;

Error inicial: 153.295, Error después del entrenamiento: 1.36841

Palabra	Ear nuevo	Prueba 4 antes
adios	_a_e_	_a_a_
agua	_a_o_	_a_o_
amarillo	_a_ae_o_	_a_e_o_
ángel	_a_e_	_a_e_
armonía	_a_o_ia_	_a_o_i_
atmósfera	_a_a_a_	_a_a_e_
balón	_a_	_ao_
cangrejo	_a_e_o_	_a_e_a_
cantidad	_a_e_a_	_a_e_a_
carretera	_a_ea_	_a_ea_
contador	_o_a_o_	_o_a_o_
difícil	_i_i_i_e_e_	_e_i_i_
fuego	_a_o_	_oa_o_
indio	_i_ao_	_i_ieo_
uniforme	_o_i_oae_	_u_i_oae_

Prueba 5:

Sin filtro
4000 Iteraciones
%Error cuadratico medio deseado

errorMin=0.001;
%Factor de aprendizaje
learningRatio=0.1
%Coeficiente de incremento para el factor de aprendizaje en caso de
%reducir el error
learningRatioI=1.05;
%Coeficiente de decremento para el factor de aprendizaje en caso de
%aumentar el error mas alla de lo tolerado
learningRatioD=0.8;
%Peso dado al cambio anterior en los pesos y desplazamientos para
%calcular el cambio siguiente en los mismos parámetros (Solo en caso de
%que el error no aumente mas de lo tolerado)
delayRatioI=.95;
%Maximo incremento en el error tolerado
errorRatio=1.05;

Error inicial: 153.518, Error después del entrenamiento: 5.79501

Palabra	Ear nuevo	Prueba 5 antes
adios	_a_ueo_	_a_a_
agua	_a_o_	_a_o_
amarillo	_a_aei_i_eao_	_a_e_o_
ángel	_a_e_e_	_a_e_
armonia	_oa_o_ia_	_a_o_i_
atmósfera	_a_o_aea_	_a_a_e_
balón	_ao_	_ao_
cangrejo	_a_e_e_e_o_	_a_e_a_
cantidad	_ea_e_a_	_a_e_a_
carretera	_ae_ea_	_a_ea_
contador	_o_a_o_	_o_a_o_
dificil	_ie_e_ei_e_	_e_i_i_
fuego	_aoaea_o_	_oa_o_
indio	_i_eieao_	_i_ieo_
uniforme	_ui_oa_ae_	_u_i_oae_

Prueba 6:

Sin Filtro
%Error cuadratico medio deseado
errorMin=0.01;
%Factor de aprendizaje
learningRatio=0.05
%Coeficiente de incremento para el factor de aprendizaje en caso de
%reducir el error
learningRatioI=1.08;
%Coeficiente de decremento para el factor de aprendizaje en caso de
%aumentar el error mas allá de lo tolerado
learningRatioD=0.7;
%Peso dado al cambio anterior en los pesos y desplazamientos para
%calcular el cambio siguiente en los mismos parámetros (Solo en caso de

%que el error no aumente más de lo tolerado)
delayRatioI=.98;
%Maximo incremento en el error tolerado
errorRatio=1.04;

Error inicial: 153.433, Error después del entrenamiento: 3.77693

Palabra	Ear nuevo	Prueba 6 antes
adios	_ao_eao_	_a_a_
agua	_a_o_	_a_o_
amarillo	_a_aei_ieao_	_a_e_o_
ángel	_a_e_	_a_e_
armonía	_a_o_eu_ua_	_a_o_i_
atmósfera	_a_oa_aeoea_	_a_a_e_
balón	_aoa_	_ao_
cangrejo	_a_e_o_o_	_a_e_a_
cantidad	_ea_e_a_	_a_e_a_
carretera	_ae_ea_	_a_ea_
contador	_o_a_o_	_o_a_o_
difícil	_ie_ei_euie_e_	_e_i_i_
fuego	_aoa_o_	_oa_o_
indio	_iui_uieao_	_i_ieo_
uniforme	_ui_oaeoe_	_u_i_oae_

Prueba 7:

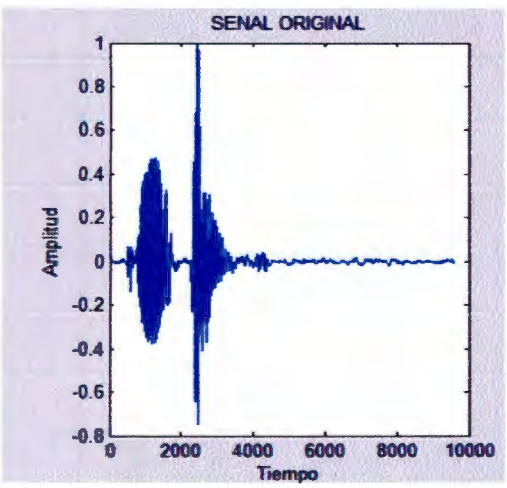
Con filtro
%Error cuadratico medio deseado
errorMin=0.1;
%Factor de aprendizaje
learningRatio=0.05
%Coeficiente de incremento para el factor de aprendizaje en caso de
%reducir el error
learningRatioI=1.08;
%Coeficiente de decremento para el factor de aprendizaje en caso de
%aumentar el error mas allá de lo tolerado
learningRatioD=0.8;
%Peso dado al cambio anterior en los pesos y desplazamientos para
%calcular el cambio siguiente en los mismos parámetros (Solo en caso de
%que el error no aumente más de lo tolerado)
delayRatioI=.95;
%Maximo incremento en el error tolerado
errorRatio=1.04;

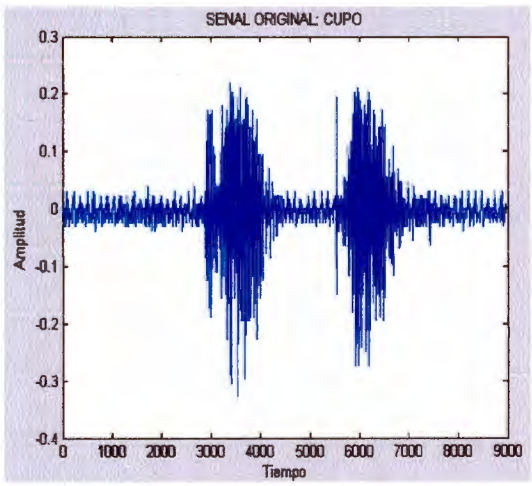
Error inicial: 153.295, Error despues del entrenamiento: 7.98856

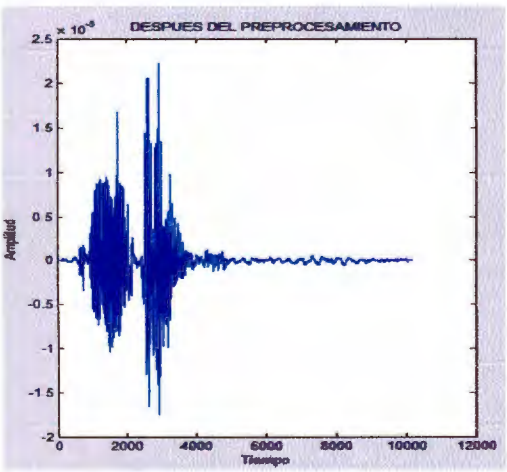
Palabra	Ear nuevo	Prueba 7 antes
adios	_a_	_a_a_
agua	_a_o_	_a_o_
amarillo	_a_o_	_a_e_o_
ángel	_a_e_	_a_e_
armonía	a_o_i_	_a_o_i_
atmósfera	_a_a_a_	_a_a_e_
balón	_a_	_ao_
cangrejo	_a_e_o_	_a_e_a_
cantidad	_e_a_e_e_a_	_a_e_a_
carretera	_e_a_e_e_a_	_a_ea_
contador	_o_o_	_o_a_o_
dificil	_i_i_	_e_i_i_
fuego	_a_	_oa_o_
indio	_i_	_i_ieo_
uniforme	_i_e_	_u_i_oae_

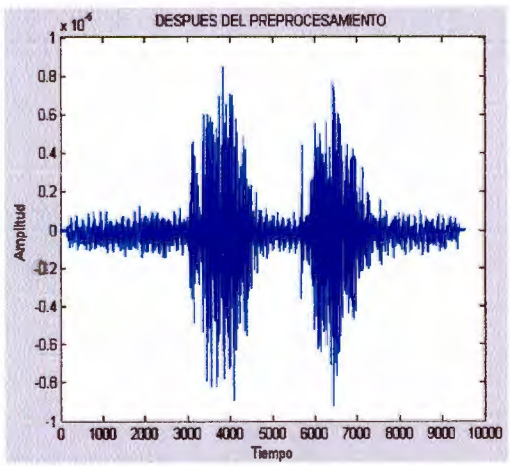
En este anexo se presentan los resultados obtenidos al ingresar diversos archivos de audio al sistema. Las graficas obtenidas y mostradas a continuación, están separadas por palabra y están divididas en dos columnas, del lado izquierdo para voz normal y del lado derecho para voz esofágica.

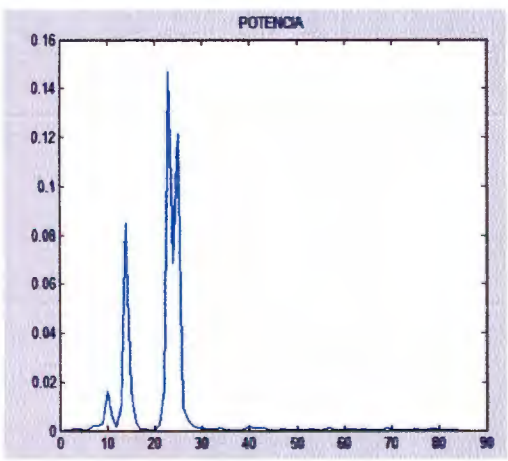
Palabra: CUPO

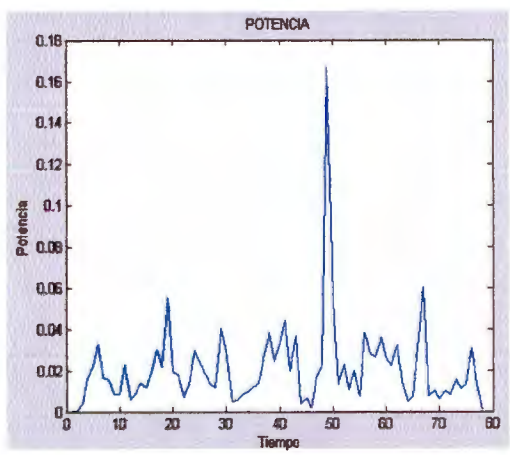


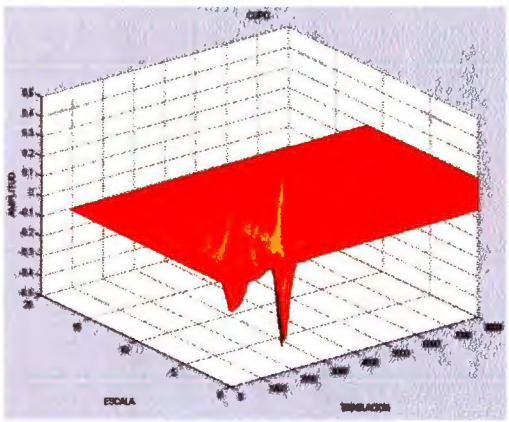


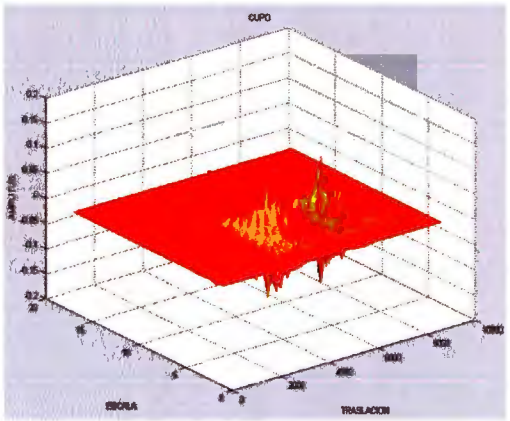




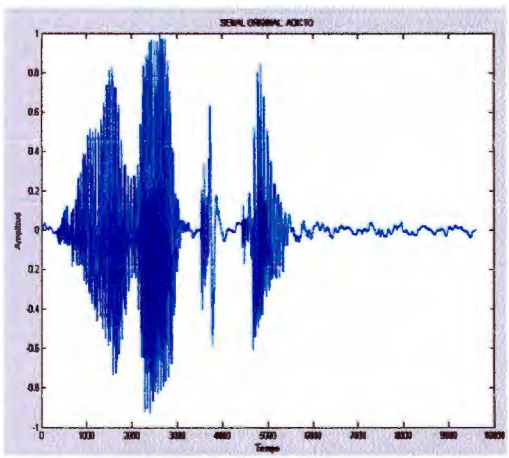


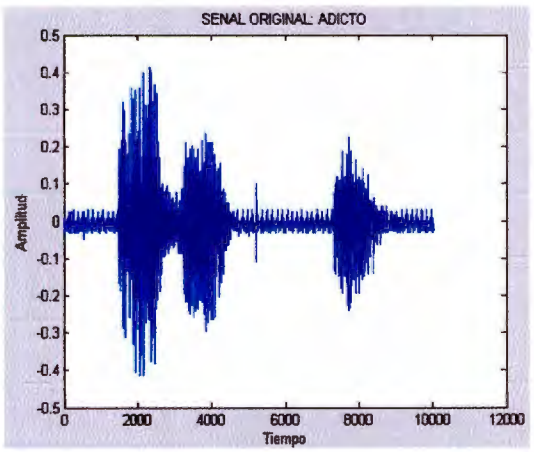


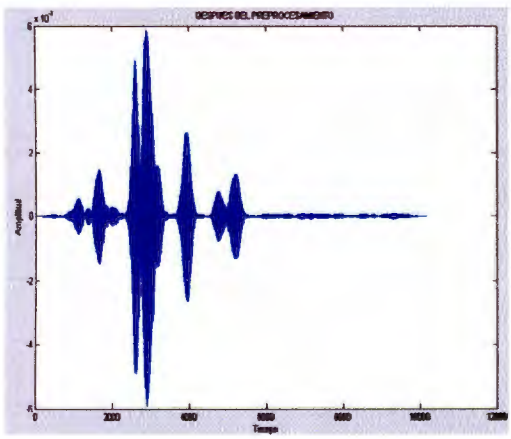


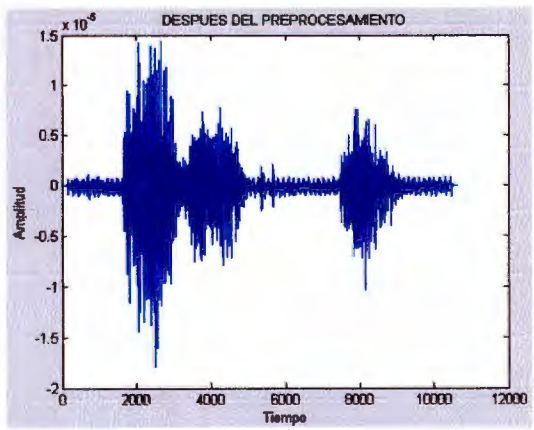


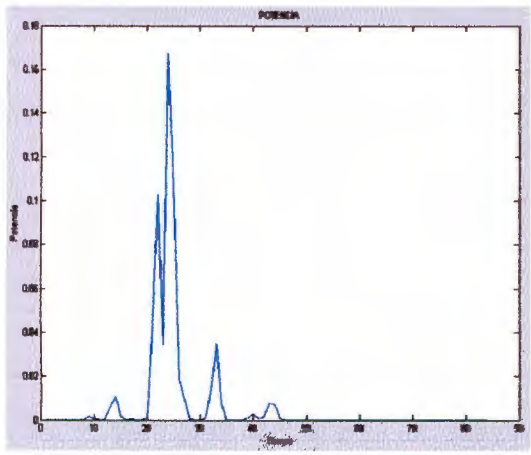
Palabra: ADICTO

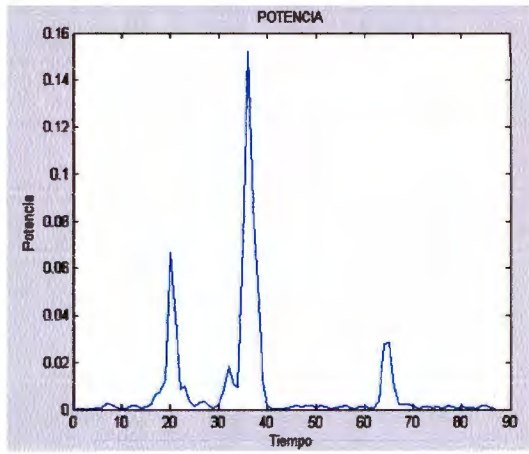


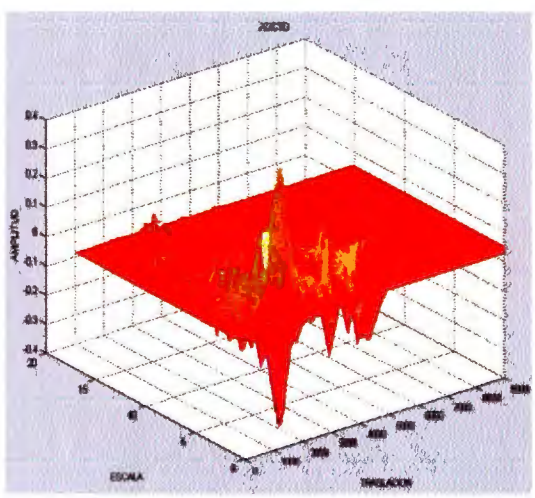


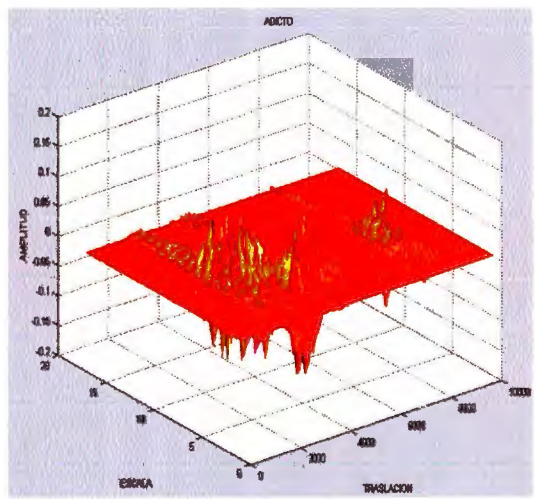




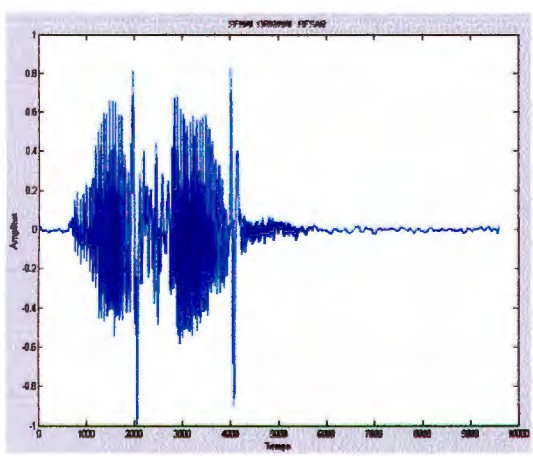


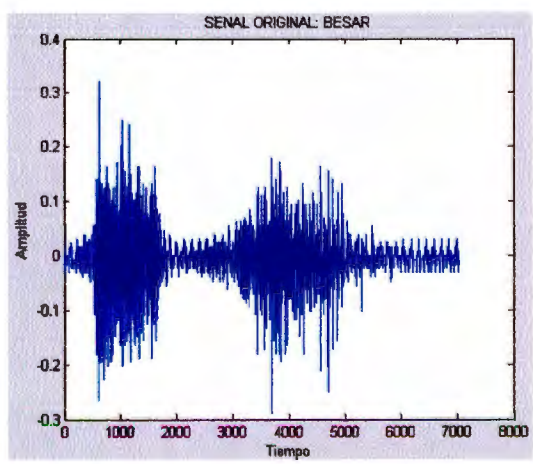




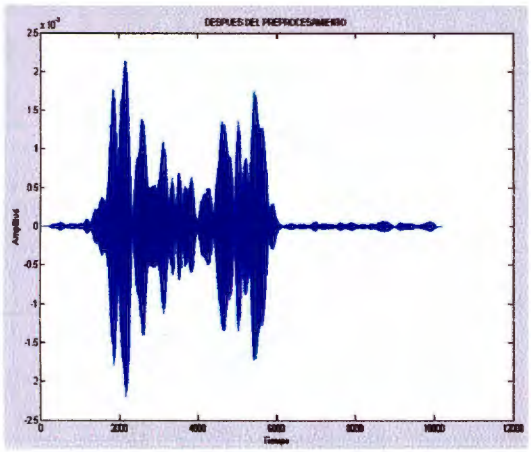


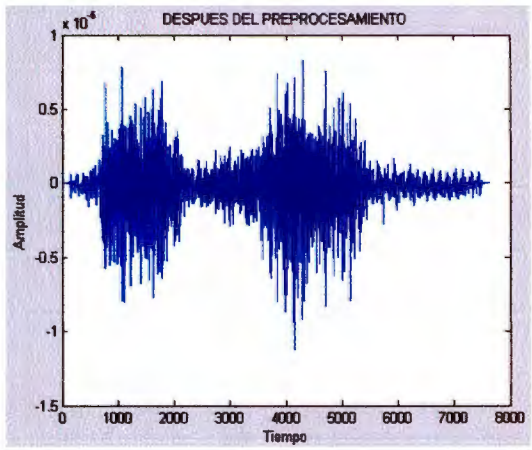
Palabra: BESAR

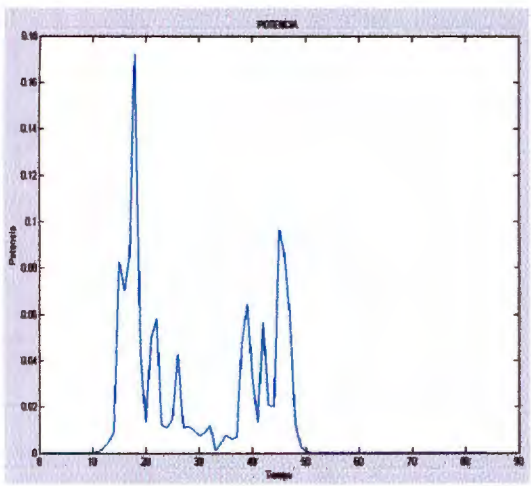


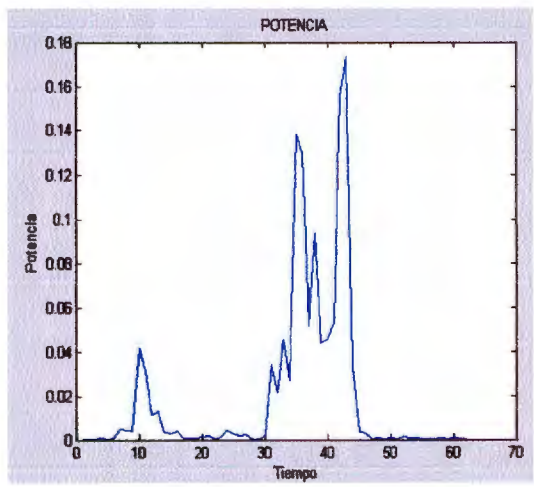


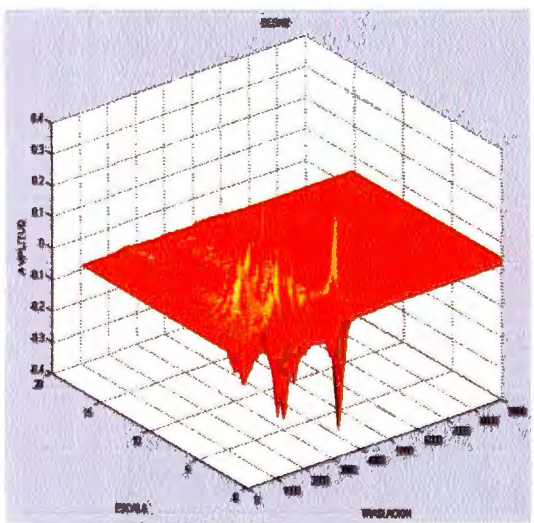
Anexo D

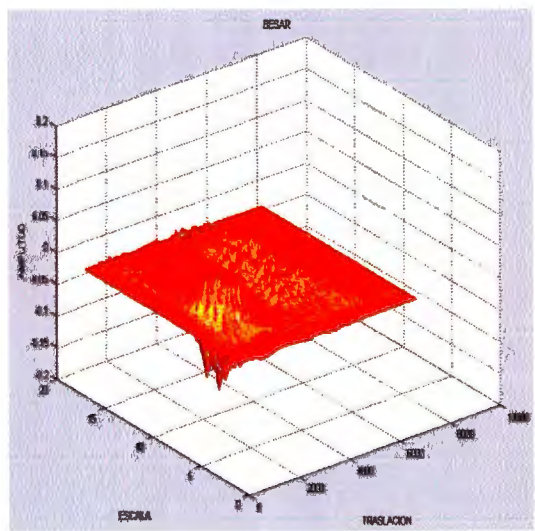

DESPUES DEL PREPROCESAMIENTO
Amplitud
Tiempo


x 10^-5
DESPUES DEL PREPROCESAMIENTO
Amplitud
Tiempo

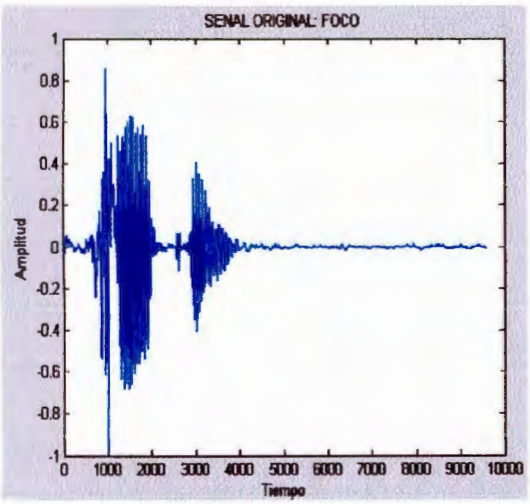

POTENCIA
Potencia
Tiempo

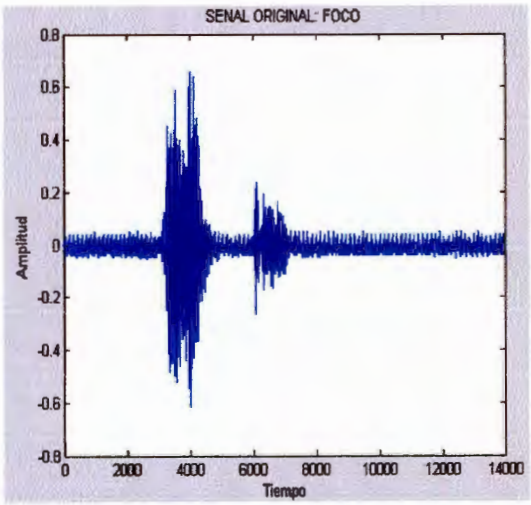

POTENCIA
Potencia
Tiempo

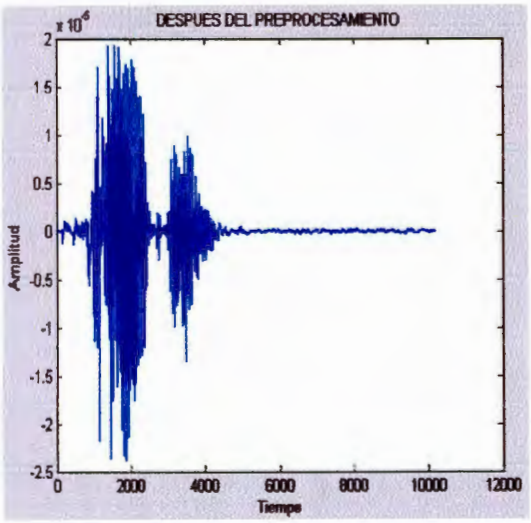

ESCALA
TRASLACION

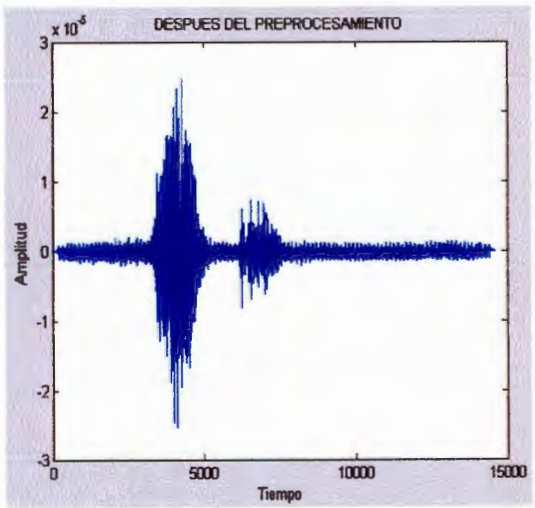

ESCALA
TRASLACION

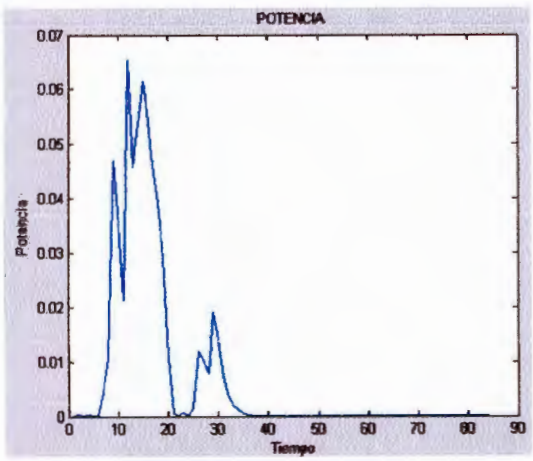
Palabra: FOCO

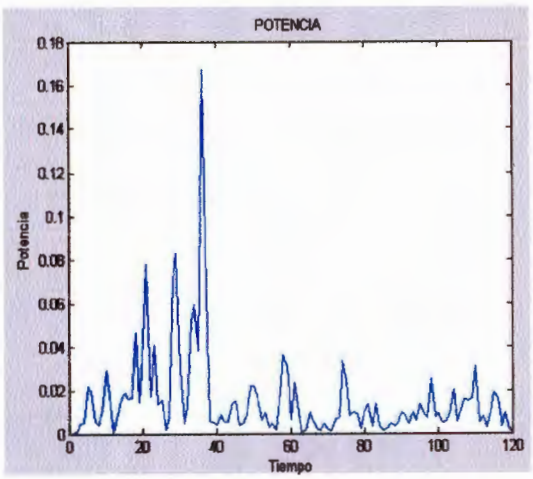


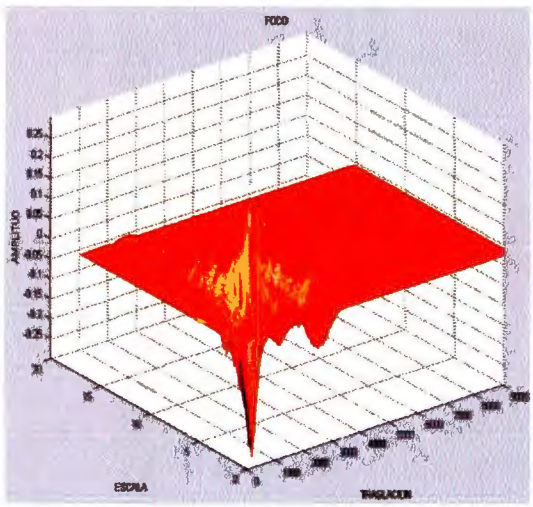


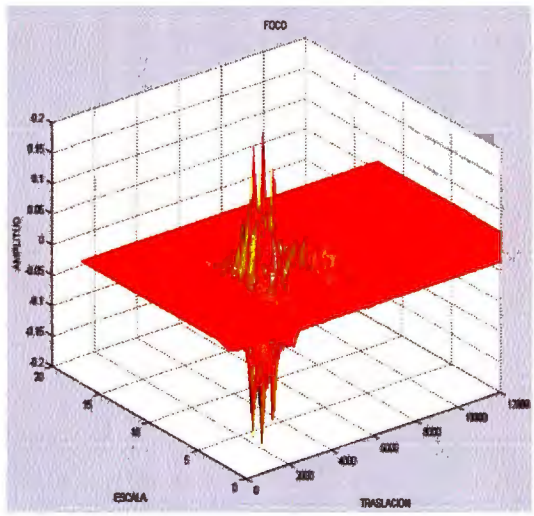




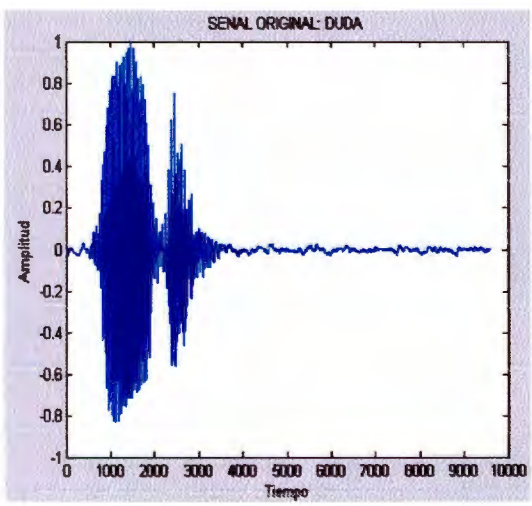


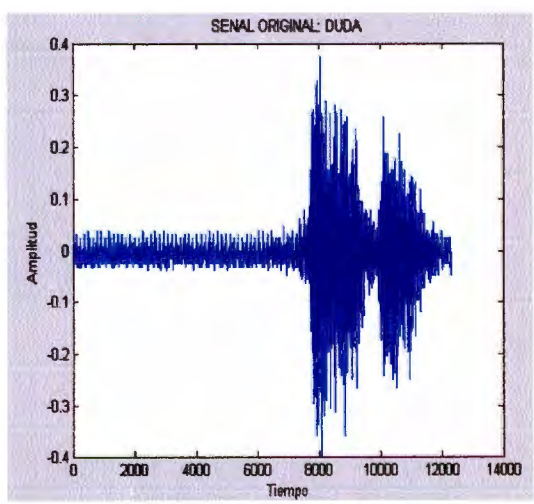


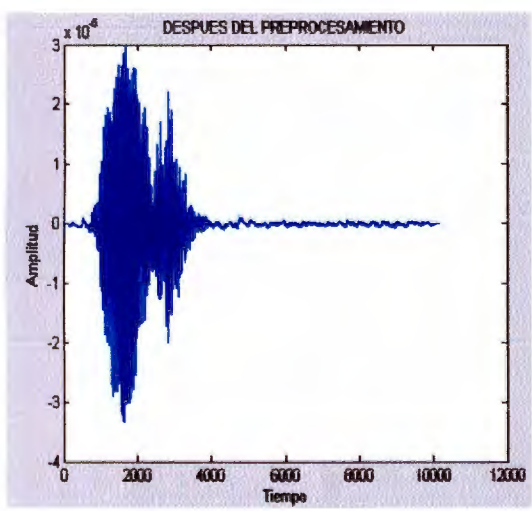


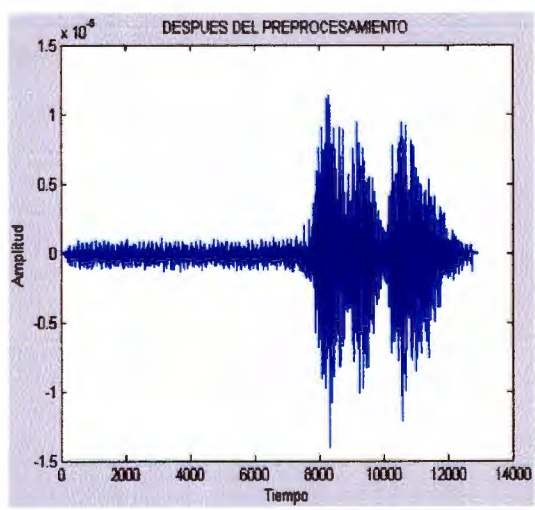


Palabra: DUDA

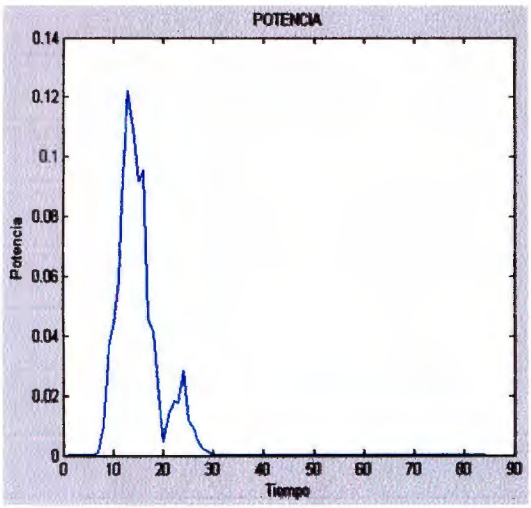


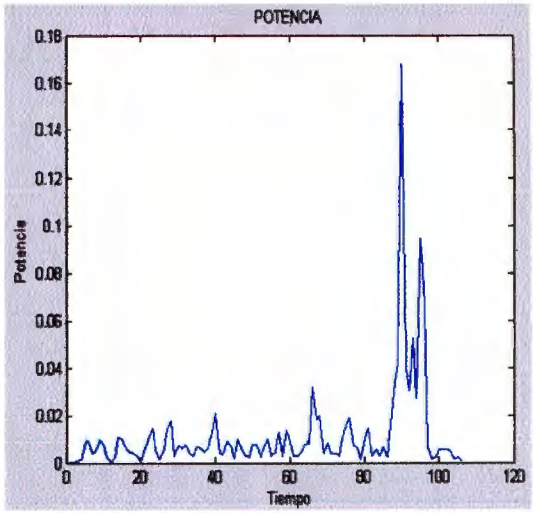


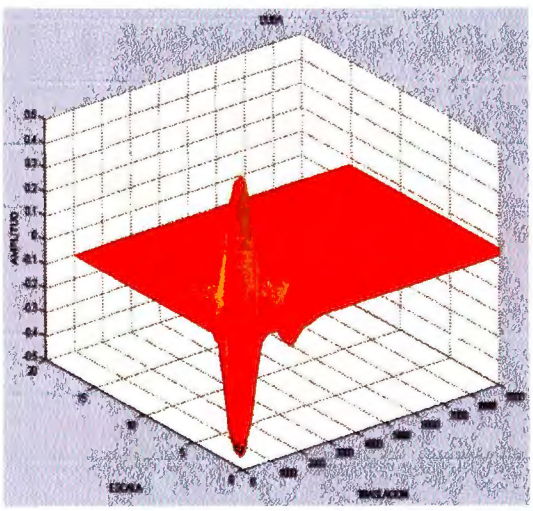


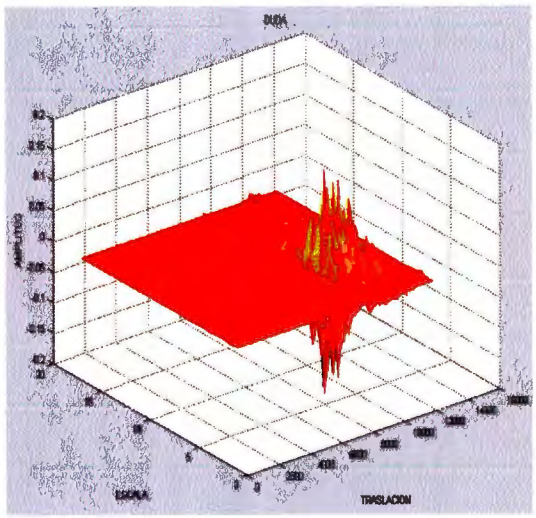


Anexo D


POTENCIA
Potencia
Tiempo
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0 10 20 30 40 50 60 70 80 90


POTENCIA
Potencia
Tiempo
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0 20 40 60 80 100 120


ESCALA
TRASLACION


ESCALA
TRASLACION

Bibliografía

[1] Treviño López, Jorge Alberto
Reconocimiento de voz esofágica empleando redes neuronales artificiales/ Jorge Alberto
Treviño López, Patricia Isabel Ortal Vite; asesor Alfredo Mantilla Caeiros.

[2] Reconocimiento de voz y fonética acústica / Bernal, Jesús; Bobadilla, Jesús y Gómez, pedro/ 356 págs.

[3] http://users.rowan.edu/~polikar/wavelets/wttutorial.html

[4] http://acta.otorrinolaringol.esp.medynet.com/textocompleto/actao.pdf s. J. Pérez ruíz.

[5] El oído humano, ¿genera sonido?. Ciencia y desarrollo. Mayo-junio, no. 122: pp 52-59.

[6] http://healthlibrary.epnet.com/getcontent.aspx?token=8482e079-8512-47c2-960c-a403c77a5e4c&chunkiid=103913

[7] http://sisbib.unmsm.edu.pe/bibvirtualdata/libros/linguistica/leng_nino/pdf/explor_producc.pdf

[8] http://www.ee.ryerson.ca/~jsantarc/html/theory.html

[9] Brown, D. & Rothery, p. 1993. Models in biology. Ed. Springer- Verlag.

[10] Alonso, G. & Becerril, j.l. 1993. Introducción a la inteligencia artificial. Ed. Multimedia ediciones s.a. Barcelona.

[11] Tutorial de redes neuronales de la universidad politécnica de Madrid:
http://www.gc.ssr.upm.es/inves/neural/ann2/anntutorial.html

[12] Fundamental papers in wavelet theory / edited by Christopher Heil and David F. Walnut. Autor: Heil, Christopher, 1960.
Princeton, N.J. ; Woodstock : Princeton University Press, c2006.

[13] Introduction to time-frequency and wavelet transforms / Shie Qian
Autor: Qian, Shie, 1949
Upper Saddle River, N.J. : Prentice Hall PTR, c2002

[14] Adaptive filters : theory and applications / B. Farhang-Boroujeny
Autor: Farhang-Boroujeny, B
Chichester ; New York : Wiley, c1999

[15] Redes neuronales : conceptos fundamentales y aplicaciones / Dr. Edgar Nelson Sánchez Camperos, M.C. Alma Yolanda Alanís García
Autor: Sánchez Camperos, Edgar Nelson
Madrid : Pearson Prentice Hall, c2006.

[16] Controls tutorials for Matlab and simulink.
http://www.library.cmu.edu/ctms/

[17] Wavelet Toolbox 4 \ Matlab
https://tagteamdbserver.mathworks.com/ttserverroot/Download/39028_8797v04_wavelet_web.pdf

[18] Xuedong Zhang, et.al., "A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and supression", Acoustical Society of America, 2001.

[19] Schroeder MR, et.al., "Suving: Automatic silence/unvoiced/voiced classification of speech", Department of computer Science, University of Sheffield. (http://plaza.ufl.edu/dongsoo7/speech1.pdf)

[20] Teorema de Littlewood-Paley
http://www.springerlink.com/content/kjn147l3118751k6/