UNIVERSIDAD NACIONAL DE INGENIERIA
FACULTAD DE INGENIERIA MECANICA

"DISEÑO DE UN SISTEMA IDENTIFICADOR DE PERSONAS MEDIANTE EL RECONOCIMIENTO DE VOZ USANDO REDES NEURONALES"

TESIS

PARA OPTAR EL TITULO PROFESIONAL DE

INGENIERO MECATRONICO

JOSUÉ VALENTIN USCATA BARRIENTOS

PROMOCION 2006-I

LIMA-PERU

2009

DEDICATORIA:

Dedico esta obra a mis padres Teofanes Máximo Uscata Crisante y Máxima Barrientos Carvajal; a mi hermano mayor Edgar Jorge Uscata Barrientos y al menor Marcos Máximo Uscata Barrientos

TABLA DE CONTENIDO

CAPITULO IV.

CAPITULO V.

CAPITULO VIII

CAPITULO VIII

PROLOGO

La presente tesis está conformada por tres unidades conteniendo ocho capítulos en las que se desarrolla este trabajo.

La primera unidad Formulación del Problema se revisa las distintas técnicas de identificación de personas así como también se presenta una de estas técnicas que es el reconocimiento de voz y se usa en conjunto con la herramienta de las redes neuronales para resolver el problema de identificación de personas, la unidad está compuesta por los capítulos: introducción y generalidades, así como también, el de formulación del problema. La segunda unidad Marco Teórico donde se desarrolla la teoría necesaria para el desarrollo de este trabajo, está compuesta por los capítulos: procesamiento de señales, redes neuronales artificiales, la voz humana, etapa de adquisición y extracción de las características de la señal de voz y etapa de entrenamiento de la red de Kohonen. La tercera unidad Desarrollo de la Solución integra las actividades que se desarrollan para solucionar el problema planteado y los resultados de las pruebas realizadas, esta unidad esta compuesta por: identificación de personas mediante el reconocimiento de voz usando redes neuronales, así como también, pruebas y resultados.

En el primer capítulo, se hace la introducción explicando el inicio de este trabajo enlazando la teoría recibida en el pregrado con la investigación y los beneficios que se consiguen con esta combinación.

En el segundo capítulo, se plantea el problema al que se le propone una solución con este trabajo como también las especificaciones técnicas de este. También se detallan los objetivos generales y específicos de la tesis.

En el tercer capítulo, se hace un repaso a la teoría revisada para el desarrollo de este trabajo definiendo las señales y sistemas e identificando a las señales aleatorias. Se hace una revisión de los temas de análisis de frecuencia y la transformada de Fourier para realizar estos análisis en las señales aleatorias. La voz es una señal aleatoria para ello se

tocan temas de la estimación espectral de potencia y los estimadores de Bartlett y el mejoramiento de este que es el estimador de Welch.

En el cuarto capítulo, se hace una revisión teórica a las redes neuronales presentando la clasificación y tipos de redes neuronales artificiales, se hizo énfasis en la teoría de la red de Kohonen mostrando su arquitectura, algoritmo de aprendizaje, la etapa de funcionamiento y la etapa de aprendizaje.

En el quinto capítulo, se hace una revisión de la voz humana viendo los conceptos preliminares de comunicación, lenguaje, la fonología y la fonética. También se hace una revisión breve del aparato fonatorio, se muestra la clasificación de los sonidos de voz concluyendo con un análisis espectral de la voz humana del cual se tomo como iniciativa y base de este trabajo.

En el sexto capítulo, se revisa la etapa de adquisición y extracción de las características de la señal de voz, empezando por las características de la tarjeta de sonido que es el hardware encargado de la adquisición de la señal. Se hace una revisión de la naturaleza del sonido digital y uno de los formatos de almacenamiento. En este capítulo también se presenta la forma de adquisición y filtrado de señal para este trabajo así como también la extracción de las características frecuenciales de voz.

En el séptimo capítulo, se revisa el entrenamiento para la red neuronal de kohonen.

En el octavo capítulo, se presentan las pruebas y resultados al momento de realizar la adquisición, tratamiento digital y extracción de las características frecuenciales de la señal de voz, para luego realizar la prueba de validación.

Para realizar este trabajo recibí el apoyo de mi asesor de tesis el Prof: Ing. Msc. Ricardo Raul Rodriguez Bustinza, el Prof: Ing. Mario Borja Borja, el Prof: Dr. Jorge del Carpio Salinas y de mi compañero de estudios el Ing. Msc. David Ronald Achanccaray Diaz.

UNIDAD I

CAPITULO I

INTRODUCCION

La universidad nos dan las herramientas teóricas de las que podemos hacer uso para resolver problemas que se nos presentan en el día a día, como ejemplo podemos hacer mención de la vez que hicimos la medición de la elongación de un resorte al aplicarle una fuerza, la teoría nos dice que el resorte tiene un comportamiento lineal entonces al obtener el cociente entre la fuerza y la elongación obtenemos la constante de elasticidad de dicho resorte [1], con esta constante podemos escalar la elongación en kilogramos con marcas sobre una lamina de metal y con esto obtener el dinamómetro que en el campo agrícola los agricultores la llaman romanilla y es muy usado para hacer la medición de la masa del producto obtenido de la tierra. Otro ejemplo de mayor nivel seria el uso de la Transformada de Fourier formula matemática que nos sirve para obtener las frecuencias que componen una señal [2], esta señal es leída por su respectivo sensor, dependiendo del tipo de señal, para luego ser muestreada y posteriormente filtrada para reducirle, por ejemplo, el ruido; esta señal leída, muestreada y filtrada es recibida por un procesador pasándose a cambiar de nombre de señal a dato y estos datos estarían listos para aplicarle el algoritmo de Fourier y así mostrarnos de una forma muy aproximada a la real las frecuencias que componen la señal adquirida; en las centrales de telecomunicaciones este equipo es usado integrando

una antena, filtros, pantalla y un DSP (Procesador Digital de Señales) para mostrar las frecuencias que se encuentran viajando por el medio ambiente.

El propósito de esta tesis es que haciendo uso de la investigación y de la teoría, el procesamiento de señales digitales en conjunto con las redes neuronales, resolver el problema de reconocer a las personas por medio de su voz y así poder identificarlas. Con esto se puede demostrar que la teoría que se da en el pregrado de la mano con la investigación conforma una poderosa herramienta. Podemos decir que la investigación permite avanzar, estar atento de las últimas tecnologías, estar actualizados permanentemente, hacer pruebas tomando como referencia la teoría para evitar los errores, dar un valor agregado.

Para incentivar esto puedo citar como ejemplo lo que le paso a nuestro colega Arquímedes que al presentársele un problema y estando muy concentrado en esté se dio cuenta de lo que en el futuro se conocería como **El Principio de Arquímedes** que además de salvarle la vida le dio fama [3].

Este trabajo es un diseño que puede tener muchos derivados de lo que me permite poder decir sobre los alcances y limitaciones del estudio y es que solo estarían acotados por la automotivación y la imaginación de cada uno de los colegas.

CAPITULO II

GENERALIDADES Y FORMULACION DEL PROBLEMA

2.1 Planteamiento Del Problema

Hoy en día hay sistemas que requieren de la identificación de las personas para poder acceder a los mismos como por ejemplo un ambiente que solo el personal autorizado o capacitado puede entrar, los equipos tales como controladores o administradores de datos que solo algunas personas tienen la autorización de reconfigurar o acceder a este equipo.

Otro ejemplo seria el registro en forma personal de la hora de ingreso y salida de las trabajadores para poder contabilizar las horas trabajadas durante la semana, la manera clásica es a través de palabras claves o password que no presentan mucha seguridad por que pueden ser copiados fácilmente esto sucede en el caso de la restricción a los ambientes, en el caso del registro del trabajador este le puede dar a un amigo la palabra clave y hacer el registro por el amigo.

La pregunta sería como poder identificar a las personas que están autorizadas para una determinada tarea o que los trabajadores hagan el registro de forma personal?.

Una manera seria acceder al ambiente o hacer el registro mediante el reconocimiento de *voz* de la persona.

Otra seria a través de la huella de los dedos reconociendo las formas de las crestas papilares y los surcos interpapilares que no cambian sus características a lo largo de toda la vida

Otra forma de acceso seria por reconocimiento de iris obteniendo imágenes y extrayendo patrones de esta para compararla con patrones extraídas anteriormente como base.

El reconocimiento facial también consiste en tomar patrones del rostro en diferentes expresiones faciales de la persona autorizada para luego compararla y permitir el acceso.

Al observar los sensores que se necesitarían para cada método como por ejemplo en el de reconocimiento de voz se necesita de un micrófono, para el de huellas digitales se necesitaría una pantalla lectora de huellas digitales, para el caso de reconocimiento de iris se podría usar una cámara fotográfica, para el caso de reconocimiento facial se usaría también una cámara o una video cámara, en conclusión de todos estos casos el menos costoso seria un micrófono comercial.

Si nos fijamos en los recursos de hardware, para el procesamiento de cada método se necesitaría de un procesador o un microcontrolador para los diferentes métodos

En el procesamiento de imagen del iris, huellas digitales y faciales se usa el procesamiento de imágenes los cuales requieren de gran memoria y además de una gran velocidad de procesamiento.

En comparación con el procesamiento de voz que necesita menos memoria y en cuanto a la velocidad de procesamiento podría ser menor al que se usa para imágenes para obtener una respuesta en un tiempo aceptable, es por estas razones que en este trabajo se usara el reconocimiento de voz, para esto hay varios métodos haciendo uso de la teoría de procesamiento de voz [4].

Si se escoge el procesamiento de voz para esta tarea, ahora que parte de la voz se usara?.

Para responder esta pregunta debemos mencionar que para realizar este trabajo se observaron 2 formar de analizar la voz.

La primera seria la que hace el análisis de la voz observando su espectrograma es decir la variación de las frecuencias conforme transcurre el tiempo, cuando hablamos emitimos sonidos que contiene un determinado espectro de frecuencias en instante de tiempo y que ha medida que va pasando el tiempo este, el espectro de frecuencia, varia. Este análisis requiere de una gran cantidad de data y es por esa razón que no la usamos como método de análisis.

Si se mantiene constante el espectrograma a lo largo del tiempo que podría ser pronunciando una vocal entonces esto sería como analizar el espectrograma en un instante de tiempo, entonces solo seria cuestión de obtener su espectro de frecuencias. Para realizar esta tarea, en el caso las vocales, se podrían ver a la vocal como una señal aleatoria y aplicarle los métodos para las señales aleatorias y así obtener el espectro de la señal, estos espectros pasarían a ser los patrones de entrenamiento de la red para que después estos patrones entrenados identifiquen al usuario comparando su patrón de frecuencia con alguno que esta entrenado, en este trabajo se usara esta teoría de procesamiento de voz como análisis de los datos de entrada y luego se usaran las redes neuronales para la síntesis de estos datos y poder identificar a las personas por medio del reconocimiento de voz [4].

¿Será posible usando la teoría de procesamiento de voz en conjunto con el de redes neuronales diseñar un sistema identificador de personas por el reconocimiento de su voz?

2.2 Especificaciones Técnicas

Para realizar la grabación se requiere de una tarjeta de sonido con una frecuencia de muestreo de 44,1 Khz y 16 bits de resolución como mínimo, la mayoría de las tarjetas de sonido actuales ya cuentan con estos requerimientos mínimos [5].

Los instrumentos con el cual se hace la adquisición de voz serian un micrófono para PC con un conector jack que va directo a la tarjeta de sonido y según el estándar PC99 de Microsoft la entrada esta seria la de color rosado, este micrófono seria genérico, es decir los que se encuentran en el mercado.

El sistema diseñado está compuesto por 2 aplicativos, el primero es para registrar la voz del grupo de personas a identificar y que solo se realizara una sola vez dependiendo si se quiere cambiar del grupo de personas y la segunda parte es la que se encargara de identificar a la persona pidiéndole su voz.

Para un perfecto funcionamiento del sistema es necesario ***recolectar*** la voz de las personas, que serán reconocidas, en forma clara, con una intensidad normal y libre de ruido publico, esto se necesita para un buen entrenamiento así como también en el momento del reconocimiento. Según las pruebas se hace un mejor reconocimiento cuando en las etapas de registro de voz y la de identificación no hay variación en cuanto a la intensidad de voz, la ubicación del micrófono respecto a la persona, la posición de la persona, la pronunciación normal de la persona, es decir que no esté sometido a una influencia que afecte la emisión normal de la voz.

El sistema diseñado soporta un número ilimitado de usuarios ya que se le proporciona 2 neuronas para cada persona, el inconveniente es en el tiempo de procesamiento al momento de reconocer (una centésima de segundo por persona) y mucho mas tiempo se necesitaría para el entrenamiento (0.5 segundo para 2 personas). Estos tiempos dependen de la velocidad del CPU, en este caso la velocidad del procesador usado es 1.8 GHz.

2.3 Objetivos

2.3.1 Objetivos General

- Diseñar un sistema que se encargue de identificar a las personas por medio del reconocimiento de su voz usando redes neuronales artificiales.

2.3.2 Objetivo Específico

- Capturar la voz a frecuencias de muestreo de 16Khz en un formato sin compresión de la data para no perder la calidad.
- Encontrar patrones de la voz que distingan a cada persona.
- Implementar un modulo de aprendizaje de los patrones de voz de cada persona.
- Implementar un modulo que identifique a la persona por reconocimiento de los patrones de su voz.

2.4 Metodología

- Investigar y revisar las diferentes herramientas matemáticas en cuanto a procesamiento y reconocimiento de voz se refieren.
- Seleccionar las herramientas apropiadas que servirán para conseguir los patrones de voz de las personas.
- Proponer la solución del problema usando las herramientas seleccionadas para realizar el aprendizaje y la identificación de la voz.
- Implementar un prototipo para probar la propuesta de la solución al problema planteado desarrollándolo en el software MATLAB 6.5.

UNIDAD II

CAPITULO III

PROCESAMIENTO DE SEÑALES

3.1 Señales y Sistemas

Una señal la podemos definir como una cantidad física que varía a lo largo del tiempo. Matemáticamente podemos definir las señales a través de una variable independiente que es el tiempo. Por ejemplo:

$$s(t) = \cos(2\pi 100 * t) - 4sen(2\pi 750 * t)$$

Otro ejemplo de señal podría ser la señal natural del electrocardiograma (ECG). Esta señal proporciona información, estado del corazón del paciente, al doctor. De forma similar un electroencefalograma (EEG) proporciona información sobre la actividad cerebral.

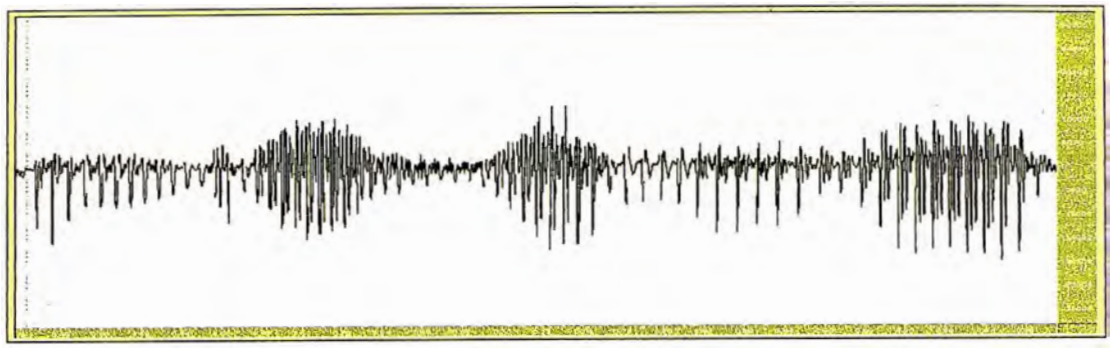


Figura 3.1: La figura muestra la señal de voz vista a lo largo del tiempo

Las señales de voz como se muestra en la Figura 3.1, los electrocardiogramas y los electroencefalogramas son ejemplos de señales que llevan información y que varían como funciones de una única variable independiente, el tiempo. Una imagen constituye un ejemplo de señal que varia con 2 variables independientes. Las dos variables independientes en este caso son las coordenadas espaciales. Estos son unos pocos ejemplos del inagotable número de señales naturales que se pueden encontrar en la práctica.

Asociados a las señales naturales se encuentran los medios con los que se generan. Por ejemplo, las señales de voz se generan al forzar el paso del aire a través de las cuerdas vocales.

Las imágenes se obtienen exponiendo película fotográfica ante un paisaje u objeto. Por lo tanto, la forma en la que se generan las señales se encuentran asociada con un sistema que responde ante un estimulo o fuerza.

En una señal de voz el sistema esta constituido por las cuerdas vocales y el tracto vocal, también llamado cavidad vocal. El estimulo en combinación con el sistema se llama fuente de señal. Por lo tanto, tenemos fuentes de voz, de imágenes y de otros tipos de señales.

Un sistema se puede definir también como un dispositivo físico que realiza una operación sobre una señal. Por ejemplo, un filtro que se usa para reducir el ruido y las

interferencias que corrompen la señal conteniendo la información deseada se denomina sistema. En este caso, el filtro realiza algunas operaciones sobre la señal, cuyo efecto es reducir (filtrar) el ruido y la interferencia presentes en la señal deseada [4].

3.2 Análisis Frecuencial de Señales y Sistemas

El objetivo básico de desarrollar herramientas de análisis frecuencias es proporcionar una representación matemática para las componentes frecuenciales contenidas en una cierta señal.

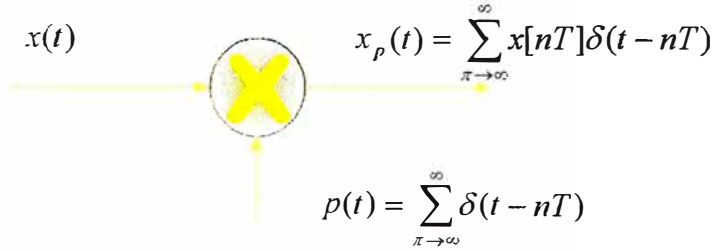
Como en la física, el termino espectro se emplea al referirse al contenido en frecuencia de una señal. El proceso de obtención del espectro de frecuencias de una señal dada, usando las herramientas matemáticas básicas, se conoce como análisis frecuencial o espectral. A su vez, el proceso de determinación del espectro de una señal en la práctica, basado en mediciones reales de la señal, se denomina estimación espectral. Esta distinción es muy importante. En un problema practico, la señal que esta siendo analizada no conduce a una descripción matemática exacta. La señal suele ser portadora de cierta información que intentamos extraer. Si esta información que deseamos extraer se puede obtener directa o indirectamente a partir del contenido espectral de la señal, hacemos estimación espectral sobre la señal que porta la información y así obtenemos una estimación del espectro de la señal. De hecho, podemos ver la estimación espectral como un tipo de análisis espectral realizado sobre señales obtenidas de fuentes físicas como por ejemplo las señales de voz, EEG, ECG, etc. Los instrumentos de software o programas empleados para obtener estimaciones espectrales de tales señales se conocen como analizadores espectrales.

Las herramientas usadas para el análisis es la trasformada discreta de fourier es por eso a continuación describiremos la teoría referente a esta herramienta [4].

3.3 La Transformada Discreta De Fourier (DFT)

Tal como pasa en el caso continuo, la serie de Fourier discreta es aplicable solamente a señales periódicas. Para señales aperiódicas aplicamos la transformada de Fourier discreta, que podemos deducir a partir de la transformada de Fourier continua [4].

Sea el siguiente sistema, que muestrea la señal continua *x(t)*:



$$x_p(\omega) = \int_{-\infty}^{\infty} x_p(t).e^{-j\omega t}\,dt = \int_{-\infty}^{\infty}\sum_{n=-\infty}^{\infty} x[nT].\delta(t-nT).e^{-j\omega t}\,dt =$$

$$= \int_{-\infty}^{\infty}\sum_{n=-\infty}^{\infty} x[nT].e^{-j\omega nt}.\delta(t-nT)dt = .\sum_{n=-\infty}^{\infty} x[nT].e^{-j\omega nt}\int_{-\infty}^{\infty}\delta(t-nT)dt =$$

$$= \sum_{n=-\infty}^{\infty} x[nT].e^{-j\omega nt}$$

Pasamos de tiempo continuo a tiempo discreto con un cambio de variable $\Omega = \omega T$, y obtenemos la transformada discreta de Fourier de $x(n)$:

$$x_p(\Omega) = \sum_{n=-\infty}^{\infty} x[n].e^{-j\Omega n}$$

3.4 Estimación Espectral de Potencia

Muchos de los fenómenos que ocurren en la naturaleza se caracterizan mejor estadísticamente en términos de promedios. Por ejemplo, fenómenos metereológicos como fluctuaciones en la temperatura y presión del aire se caracterizan mejor estadísticamente como procesos aleatorios. Los ruidos térmicos de voltajes generados en resistencias y equipos electrónicos son otros ejemplos de señales físicas que se modelan bien como procesos aleatorios.

Debido a los cambios en el tiempo de estas señales debemos hacer el análisis estadístico de dichas señales que trabaja con características promediadas de estas. La autocorrelación de un proceso aleatorio es el promedio estadístico apropiado que se usan para este tipo de señales al obtener sus características en el dominio de la frecuencia, al aplicarle la transformada de Fourier a esta función se obtiene la densidad espectral de potencia.

Hay varios métodos para estimar el espectro de potencia y a continuación mencionaremos algunos de los cuales se hizo una elección para realizar este trabajo.

3.5 Bartlett: Promediado De Periodogramas

El método de Bartlett (1948) es un estimador consistente del espectro de potencia que realiza un promediado del periodograma. La mejora respecto al periodograma reside en la reducción de varianza.

Supongamos *K* realizaciones incorreladas de un proceso aleatorio $x(n)$. Cada realización tiene una longitud *L*. El periodograma de cada realización $x_i(n)$ es:

$$\hat{S}_{per}^{(i)}(e^{jw}) = \frac{1}{L}\left|\sum_{n=0}^{L-1} x_i(n)e^{-jnw}\right|^2 \qquad ;i = 1,2,...,K$$

El promedio de estos periodogramas define el estimador de Bartlett:

$$\hat{S}_B(e^{jw}) = \frac{1}{K}\sum_{i=1}^{K} \hat{S}_{per}^{(i)}(e^{jw})$$

El valor medio de $S_x(e^{jw})$:

$$E\left\{\hat{S}_B(e^{jw})\right\} = E\left\{\hat{S}_{per}^{(i)}(e^{jw})\right\} = \frac{1}{2\pi}\hat{S}_B(e^{jw}) * W_B(e^{jw})$$

Donde $W_B(e^{jw})$ es la transformada de Fourier de la ventana de Bartlett, $W_B(k)$, que se extiende desde $-L$ hasta L. Este estimador es asintóticamente no sesgado. Con esta suposición de realizaciones incorreladas, obtenemos una varianza de $S_x(e^{jw})$:

$$Var\left\{\hat{S}_B(e^{jw})\right\} = \frac{1}{K^2}\sum\left[Var\left\{\hat{S}_{per}^{(i)}(e^{jw})\right\}\right] = \frac{1}{K}Var\left\{\hat{S}_{per}^{(i)}(e^{jw})\right\} \approx \frac{1}{K}\hat{S}_x^2(e^{jw})$$

K veces inferior a la del periodograma y tiende a cero cuando K tiende al infinito. En consecuencia, $S_x(e^{jw})$ resulta ser un estimador consistente del espectro de potencia si K y L pueden tender a infinito (es decir, si son lo suficientemente altos).

En este método hemos supuesto realizaciones incorreladas de un proceso, pero esta situación, en la práctica, es difícil de conseguir. Generalmente se dispone de una sola realización de un proceso $x(n)$ de longitud N, y Bartlett propone dividir este proceso en K secuencias no solapadas de longitud L, donde $N = KL$:

$$x(n) = x(n + iL) \qquad \begin{cases} n=0,1,...,L-1 \\ i=0,1,...,K-1 \end{cases}$$

El estimador de Bartlett con este supuesto es:

$$\hat{S}_B(e^{jw}) = \frac{1}{N}\sum_{i=0}^{K-1}\left|\sum_{n=0}^{L-1} x(n+i\cdot L)\cdot e^{-jw}\right|^2$$

El estimador de Bartlett es asintóticamente no sesgado:

$$E\left\{\hat{S}_B(e^{jw})\right\} = \frac{1}{2\pi} S_x(e^{jw}) * W_B(e^{jw})$$

Los periodogramas utilizados para el promediado tienen longitud *L*, por ello, la resolución de este método es:

$$\mathrm{Re}\,s[\hat{S}_B(e^{jw})] = 0.89\frac{2\pi}{L} = 0.89\frac{2\pi}{N}$$

Esta expresión es K veces mayor que en el periodograma.

Como las secuencias $x_i(n)$ generalmente están correladas, (a no ser que $x(n)$ sea ruido blanco), la reducción en la varianza no es tan grande como hemos visto anteriormente. De todas formas, la varianza va a ser inversamente proporcional a K, y asumiendo que las secuencias de datos están aproximadamente incorreladas, para valores elevados de N, la varianza resulta ser, aproximadamente:

$$Var\{\hat{S}_B(e^{jw})\} \approx \frac{1}{K} Var\left\{\hat{S}_{per}^{(i)}(e^{jw})\right\} \approx \frac{1}{K}\hat{S}_x^2(e^{jw})$$

Si permitimos que K y *L* tiendan a infinito cuando ***N*** tiende a infinito, el estimador de Bartlett será un estimador consistente del espectro de potencia. Para un valor fijo de N, el método de Bartlett ofrece un compromiso entre resolución y varianza a través de los valores de K y L. Es decir, podremos reducir la varianza a costa de una pérdida de resolución espectral, y viceversa, pues al ganar en resolución veremos incrementada la varianza [6].

3.6 Welch: Promediado De Periodogramas Modificados

Welch, en 1967, propuso dos modificaciones al método de Bartlett. En primer lugar, permitió el solapamiento de segmentos de datos. Este efecto se conoce como overlap. Suponiendo que entre dos secuencias sucesivas existe un desplazamiento de D puntos y que cada secuencia consta de L puntos de longitud, la secuencia $i-esima$ viene determinada por la expresión:

$$x_i(n) = x(n + i \cdot D) \qquad ; n = 0,1,\ldots, L-1$$

El solapamiento entre dos secuencias consecutivas $x_i(n)$ y $x_{i+1}(n)$ es de $L-D$ puntos, y si las K secuencias cubren una longitud de N puntos, entonces:

$$N = L + D \cdot (K-1)$$

Supongamos que no existe solape entre las secuencias $(D = L)$; tendremos $K = N / L$ secciones de longitud L, como en el método de Bartlett. Si permitimos que las secuencias posean un solape del 50% $(D = L/2)$, el número de secciones K de longitud L es:

$$K = 2 \cdot N / L - 1$$

De esta manera, se mantiene la resolución (la longitud de la secuencia no varía) del método de Bartlett, pero al doblar el número de periodogramas modificados que van a promediarse, se reduce la varianza.

Con un 50% de solape entre las secuencias, también podemos formar *K* secuencias de longitud $2L$, donde K es:

$$K = N / L - 1$$

Así, mejoramos la resolución manteniendo la misma varianza que en el método de Bartlett.

Con el overlap o solapamiento es posible incrementar el número y/o la longitud de las secuencias que van a ser promediadas, logrando de esta forma una reducción en la varianza, siempre con un compromiso en la resolución del método de estimación espectral.

La segunda propuesta consiste en enventanar cada secuencia $x_i(n)$ con una ventana general $w(n)$ (no sólo con la ventana rectangular), antes de calcular el periodograma. De esta manera se obtiene un periodograma modificado por cada secuencia enventanada:

$$\hat{S}_M^{(i)}(e^{jw}) = \frac{1}{LU}\left|\sum_{n=0}^{L-1} w(n) \cdot x_i \cdot e^{-jw}\right|^2$$

El estimador de Welch es el promedio de los periodogramas modificados:

$$\hat{S}_W(e^{jw}) = \frac{1}{K}\sum_{n=0}^{K-1}\hat{S}_M^{(i)}(e^{jw})$$

Y su expresión general es:

$$\hat{S}_W(e^{jw}) = \frac{1}{KLU}\sum_{i=0}^{K-1}\left|\sum_{n=0}^{L-1} w(n)x(n+i\cdot D)\cdot e^{-jw}\right|^2$$

Donde *U*:

$$U = \frac{1}{L}\sum_{n=0}^{L-1}\left|w(n)\right|^2$$

Vamos a examinar las prestaciones del método de Welch.

Obtenemos la expresión del valor medio:

$$E\left\{\hat{S}_W(e^{jw})\right\} = E\left\{\hat{S}_M(e^{jw})\right\} = \frac{1}{2\pi LU}S_x(e^{jw}) * \left|W(e^{jw})\right|^2$$

Donde $W(e^{jw})$ es la transformada de Fourier de la ventana $\mathrm{w(n)}$ de L puntos, utilizada para formar los periodogramas modificados. Observamos que el método de Welch es asintóticamente no sesgado. La resolución depende de la ventana utilizada, y se define para el ancho de banda a 3 dB de dicha ventana. La varianza es difícil de calcular, pues el overlap no permite realizar la suposición de operar con secuencias incorreladas. De todas formas, para una ventana de Bartlett, con un solapamiento del 50%, se demuestra que la varianza es, aproximadamente:

$$Var\{\hat{S}_W(e^{jw})\} \approx \frac{9}{8K}\hat{S}_x^2(e^{jw})$$

Comparando con el método de Bartlett,

$$Var\{\hat{S}_B(e^{jw})\} \approx \frac{1}{K}Var\left\{\hat{S}_{per}^{(i)}(e^{jw})\right\} \approx \frac{1}{K}\hat{S}_x^2(e^{jw})$$

Observamos que, para un número dado de secciones *K*, la varianza con el método de Welch es mayor que con el método de Bartlett en un factor 9/8. No obstante, para unos valores fijos de datos, *N*, y resolución (longitud de la secuencia *L*), con un 50% de solapamiento, se obtiene el doble de secciones para promediar en el método de Welch. Si expresamos la varianza en términos de *L* y *N*, con ese 50% de overlap, tenemos:

$$Var\left\{\hat{S}_w(e^{jw})\right\} = \frac{9}{16}\frac{L}{N}S_x^2(e^{jw})$$

Y como N/L es el número de secciones utilizadas en el método de Bartlett, obtenemos la siguiente relación:

$$Var\left\{\hat{S}_w(e^{jw})\right\} \approx \frac{9}{16}Var\left\{\hat{S}_B(e^{jw})\right\}$$

Resumiendo, es posible incrementar el número de secuencias a promediar para una cantidad fija de datos, incrementando el overlap, pero esto supone una mayor carga computacional, así como un aumento en la correlación de las secuencias $x_i(n)$, por lo que las prestaciones disminuyen al incrementar K para un valor dado de N. Solapes típicos son 50% y 75% [6].

CAPITULO IV

REDES NEURONALES ARTIFICALES

4.1 Fundamentos de las Redes Neuronales Artificiales

Es difícil pensar de aquellas poderosas computadoras que pueden resolver extensos algoritmos con variables en coma flotante en tan solo pocos segundos pero no pueden entender el significado de las figuras o distinguir entre objetos de diferentes clases. Los sistemas de computación secuencial, dan buenos resultados en la solución de problemas de los diferentes campos de la ingeniería pero aun así están incapacitados de entender el mundo.

Esta dificultad de los sistemas de computo que trabajan en forma secuencial, como los desarrollados por Von Neuman, hizo que la mayoría de investigadores concentren su atención en desarrollar nuevos sistemas de tratamiento de la información, que permitan resolver los problemas cotidianos, tal como los hace el cerebro humano; este órgano biológico cuenta con varias características deseables para cualquier sistema de procesamiento digital, tales como:

- Es robusto y tolerante a fallas, diariamente mueren neuronas sin afectar su desempeño.

- Es flexible, se ajusta a nuevos ambientes por medio de un proceso de aprendizaje, no hay que programarlo.
- Puede manejar información difusa, con ruido o inconsistencia,
- Es altamente paralelo.
- Es pequeño, compacto y consume poca energía.

El cerebro humano constituye una computadora muy notable, es capaz de interpretar información imprecisa suministrada por los sentidos a un ritmo increíblemente veloz. Logra discernir un susurro en una sala ruidosa, un rostro en un callejón mal iluminado y leer entre líneas un discurso; lo más impresionante de todo, es que el cerebro aprende sin instrucciones explicitas de ninguna clase y crea las representaciones internas en un lugar llamado por nosotros como mente en el momento en que pensamos.

Impresionados por la eficiencia del cerebro humano, varios investigadores desarrollaron teorías sobre Redes Neuronales Artificiales (RNA) que tratan de copiar el funcionamiento de las redes neuronales biológicas, estas RNA se usaron para aprender estrategias de resolución con ejemplos típicos tomados como patrones; estos sistemas no requerían de una programación para ejecutar si no que aprendían de la experiencia.

Estas RNA han tenido mucho éxito en las siguientes aplicaciones tales como el filtrado señales, procesamiento de voz, planeamiento, predicción, control optimizado y lo que nos interesa a nosotros s la clasificación de patrones.

Las computadoras son de tipo serial, es decir, ejecutan sus comandos en forma secuencial por la arquitectura que tienen en comparación con las RNA que son de tipo paralelo por que ingresan los datos de forma paralela estas son procesadas y se obtienen las salidas según la cantidad de neuronas que se dispongan y la memoria estaría situada en las conexiones entre neuronas y se las llaman pesos.

Las RNA son una teoría que aun esta en proceso de desarrollo y prometen mucho ya que estos sistemas están basados en el comportamiento de las neuronales biológicas y el comportamiento de estas aun siguen desconocidas pero mientras pasa el tiempo se irán descubriendo mas cosas sobre estas y se tendrá mas base para generar nuevos conceptos.

4.1.1 Introducción a la Neurona Biológica

Se cree que el sistema nervioso contiene alrededor de cien mil millones de neuronas. Y se presentan en muchas formas algunas de forma muy peculiar con un cuerpo celular o soma de 10 a 80 micras de longitud continuando un extenso árbol de ramificaciones llamado árbol dendrítico debido a que sus ramas se las llama dendritas y su tronco es llamado axón de cuya longitud varia desde las 100 micras hasta el metro en caso de las neuronas motoras.

Desde el punto de vista funcional, las neuronas tienen tres componentes principales, las dendritas, el cuerpo de la célula o soma y el axón. Las dendritas, son el árbol receptor de la red, son como fibras nerviosas que cargan de señales eléctricas el cuerpo de la célula. El cuerpo de célula, realiza la suma de esas señales de entrada. El axón es una fibra larga que lleva la señal desde el cuerpo de la célula hacia otras neuronas. El contacto entre el axón de una célula y una dendrita de otra célula es llamado sinapsis, la longitud de la sinapsis es determinada por la complejidad del proceso químico que estabiliza la función de la red neuronal. Un esquema simplificado de la conexión de dos neuronas biológicas se observa en la Figura 4.1

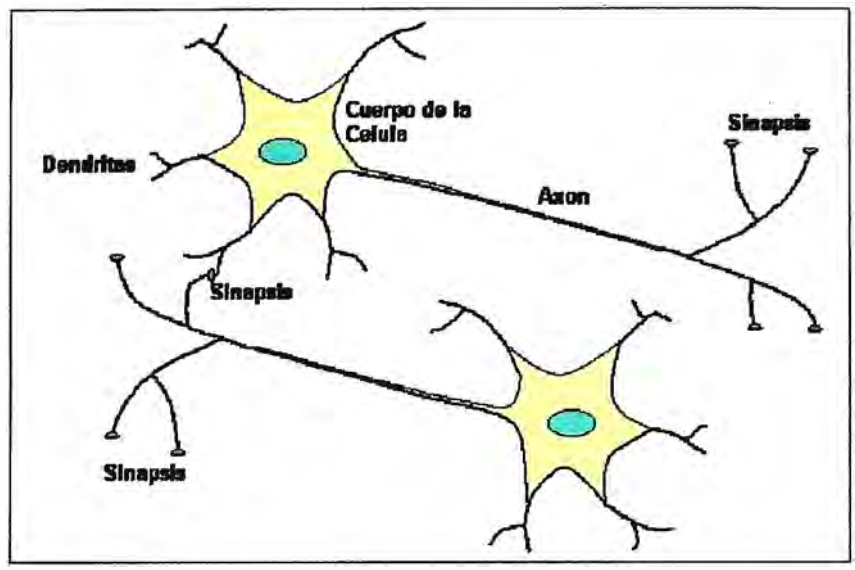


Figura 4.1: Neuronas Biológicas

Algunas de las estructuras neuronales son determinadas en el nacimiento, otra parte es desarrollada a través del aprendizaje, proceso en que nuevas conexiones neuronales son realizadas y otras se pierden por completo. El desarrollo neurológico se hace critico durante los primeros años de vida, por ejemplo esta demostrado que si un cachorro de gato, se le impide usar uno de sus ojos durante un periodo corto de tiempo, este nunca desarrollara una visión normal en ese ojo.

Las estructuras neuronales continúan cambiado durante toda la vida, estos cambios consisten en un refuerzo o debilitamiento de las uniones sinápticas, por ejemplo se cree que nuevas memorias son formadas por las modificaciones de esta intensidad entre sinapsis, así el proceso de recordar el rostro de un nuevo amigo, consiste en alterar varias sinapsis.

Como consecuencia de los nuevos estudios de la base neuronal de los sistemas mnémicos (relacionados con la memoria), se creía que el almacenamiento de la memoria asociativa, requería de un circuito neuronal muy complejo. Entre quienes comenzaron a oponerse a este enfoque se hallaba Donald O. Hebb, profesor de la universidad de Milner, Hebb sugirió que el aprendizaje asociativo podría ser producido por un mecanismo celular sencillo y propuso que las asociaciones

podrían formarse por una actividad neuronal coincidente: "Cuando un axón de la célula A excita la célula B, se produce algún proceso de desarrollo o cambio metabólico en una ambas células, de suerte que la eficacia de A, como célula excitadora de B, se intensifica". Según la regla Hebbiana de aprendizaje, el que coincida la actividad de las neuronas presinápticas (suministran el impulso de entrada) con la de las postsinápticas (reciben el impulso) es muy importante para que se refuerce la conexión entre ellas, este mecanismo es llamado pre-postasociativo, del cual puede observarse un ejemplo de la Figura 4:2.

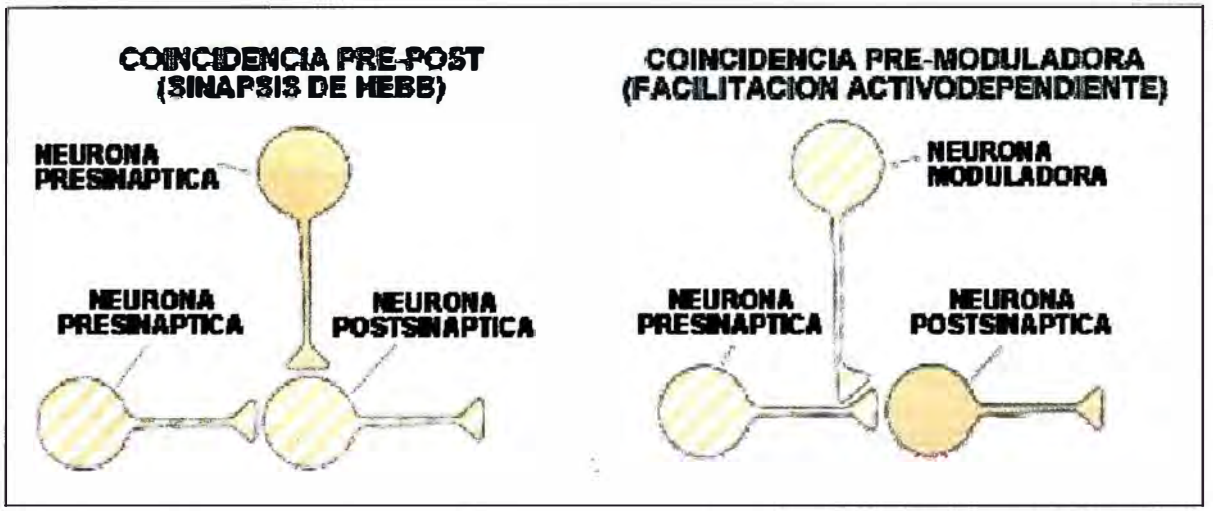


Figura 4.2: Cambios asociativos de la fuerza sinápticas durante el aprendizaje.

Todas las neuronas conducen la información de forma similar, esta viaja a lo largo de axones en breves impulsos eléctricos, denominados potenciales de acción; los potenciales de acción que alcanzan una amplitud máxima de unos 100 mV y duran 1ms, son resultado del desplazamiento a través de la membrana celular de iones de sodio dotados de carga positiva, que pasan desde el fluido extracelular hasta el citoplasma intracelular, la concentración extracelular de sodio supera enormemente la concentración intracelular.

La membrana en reposo mantiene un gradiente de potencial eléctrico de -70mv, el signo negativo se debe a que el citoplasma intracelular está cargado negativamente con respecto al exterior; los iones de sodio no atraviesan con facilidad la membrana en reposos, los estímulos físicos o químicos que reducen el gradiente de potencial, o que despolaricen la membrana, aumentan su permeabilidad al sodio y el flujo de este ión hacia el exterior acentúa la despolarización de la membrana, con lo que la permeabilidad al sodio y el flujo ión hacia el exterior acentúa la despolarización de la membrana , con lo que la permeabilidad al sodio se incrementa más aún. Alcanzando un potencial critico denominado "umbral", la realimentación positiva produce un efecto regenerativo que obliga al potencial de membrana a cambiar de signo. Es decir, el interior de la célula se toma positivo con respecto al exterior, al cabo de 1 ms, la permeabilidad del sodio decae y el potencial de membrana retorna a -70mv, su valor de reposo. Tras cada explosión de actividad iónica, el mecanismo de permeabilidad del sodio se mantiene refractario durante algunos milisegundos; la tasa de generación de potenciales de acción queda así limitada a unos 200 impulsos por segundo, o menos.

Aunque los axones puedan parecer hilos conductores aislados, no conducen los impulsos eléctricos de igual forma, como hilos eléctricos no serian muy valiosos, pues su resistencia lo largo del eje es demasiado baja; la carga positiva inyectada en el axón durante el potencial de acción queda disipada uno o dos milímetros mas adelante, para que la señal recorra varios centímetros es preciso regenerar el frecuentemente el potencial de acción a lo largo del camino, la necesidad de reforzar repetidamente esta corriente eléctrica limita a unos 100 metros por segundo la velocidad máxima de viaje de los impulsos, tal velocidad es inferior a la millonésima de la velocidad de una señal eléctrica por un hilo de cobre.

Los potenciales acción, son señales de baja frecuencia conducidas de forma muy lenta, éstos no pueden saltar de una célula a otra, la comunicación entre neuronas viene siempre mediada por transmisores químicos que son liberados en la sinapsis.

Un ejemplo de comunicación entre neuronas y del proceso químico de la liberación de neurotransmisores se ilustra en la Figura 4:3.

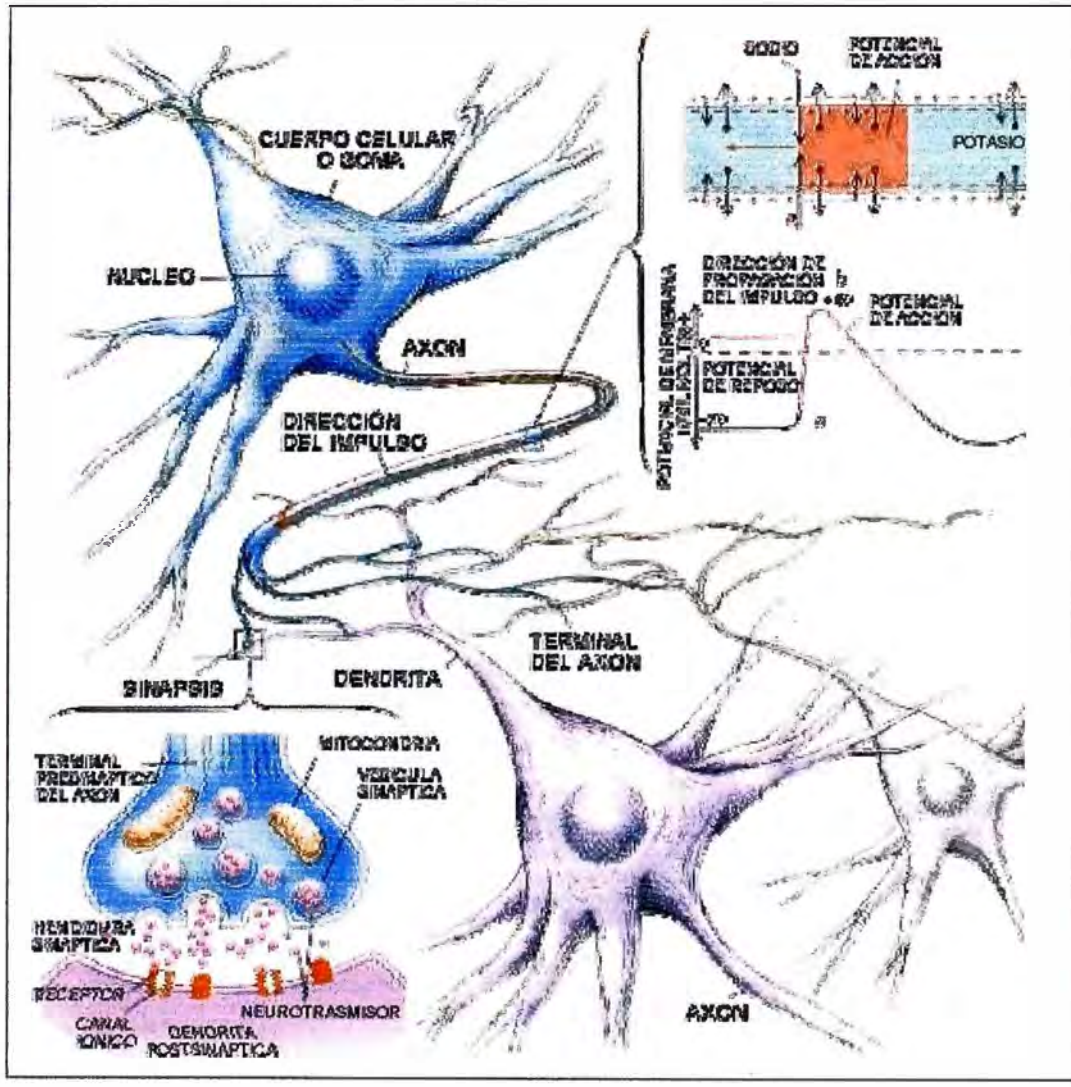


Figura 4.3: Comunicación entre neuronas.

Cuando el potencial de acción llega al Terminal de un axón son liberados transmisores alojados en diminutas vesículas, que después son vertidos en una

hendidura de unos 20 nanómetros de anchura que separa la membrana presináptica; durante al apogeo del potencial de acción, penetran iones de calcio en el terminal nervioso, su movimiento constituye la señal determinante de la exocitosis sincronizada, esto es la liberación coordinada de molécula neurotransmisoras. En cuanto son liberados, los neurotransmisores se enlazan con receptores postsinapticos, instando el cambio de la permeabilidad de la membrana.

Cuando el desplazamiento de carga hace que la membrana se aproxime al umbral de generación de potenciales de acción, se produce un efecto excitador y cuando la membrana resulta estabilizada en la vecindad el valor de reposo se produce un efecto inhibidor. Cada sinapsis produce un efecto, para determinar la intensidad (frecuencia de los potenciales de acción) de la respuesta cada neurona ha de integrar continuamente hasta 1000 señales sinápticas, que se suman en el soma o cuerpo de la célula.

En algunas neuronas los impulsos se inician en la unión entre el axón y el soma, y luego se transmiten a lo largo del axón a otras células. Las sinapsis pueden ser excitatorias o inhibitorias según el neurotransmisor que se libere, cada neurona recibe de 10.000 a 100.000 sinapsis y su axón realiza una cantidad similar de sinapsis.

La sinapsis se clasifican según su posición en la superficie de la neurona receptora en tres tipos: axo-dendriticas, axo-axónicas. Los fenómenos que ocurren en la sinapsis son de naturaleza química, pero tienen efectos eléctricos laterales que se pueden medir.

En la Figura 4.4 se visualiza el proceso químico de una sinapsis y los diferentes elementos que hacen parte del proceso tanto en la neurona presináptica, como en la postsinaptica, como en la postsináptica.

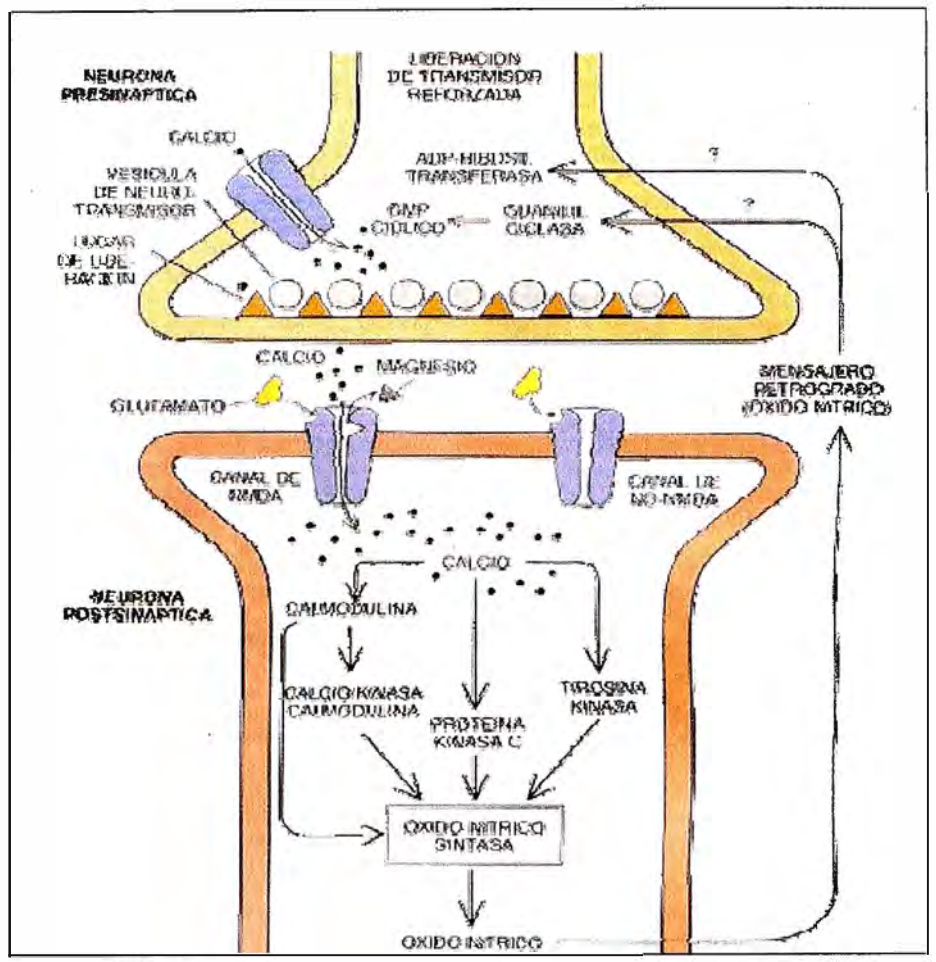


Figura 4.4: Proceso químico de una sinápsis.

Las RNA no alcanzan la complejidad del cerebro, sin embrago hay dos aspectos similares entre redes biológicas y artificiales, primero los bloques de construcción de ambas redes son sencillos elementos computacionales (aunque las RNA son mucho más simples que las biológicas) altamente interconectados; segundo, las conexiones entre neuronas determinan la función de la red.

4.1.2 Características De Una Red Neuronal Artificial

Ahora podemos señales a la neurona como un sistema que sus entradas para recibir los datos, tiene un centro de procesamiento dependiendo del tipo de trabajo que se desee que realice y su respectiva salida.

En diferentes tratados sobre RNA nombran a la neurona artificial de diferentes formas como nodo, neuronodo, celda, unidad o elemento de procesamiento (PE); en la siguiente figura se puede observar la similitud con la neurona biológica.

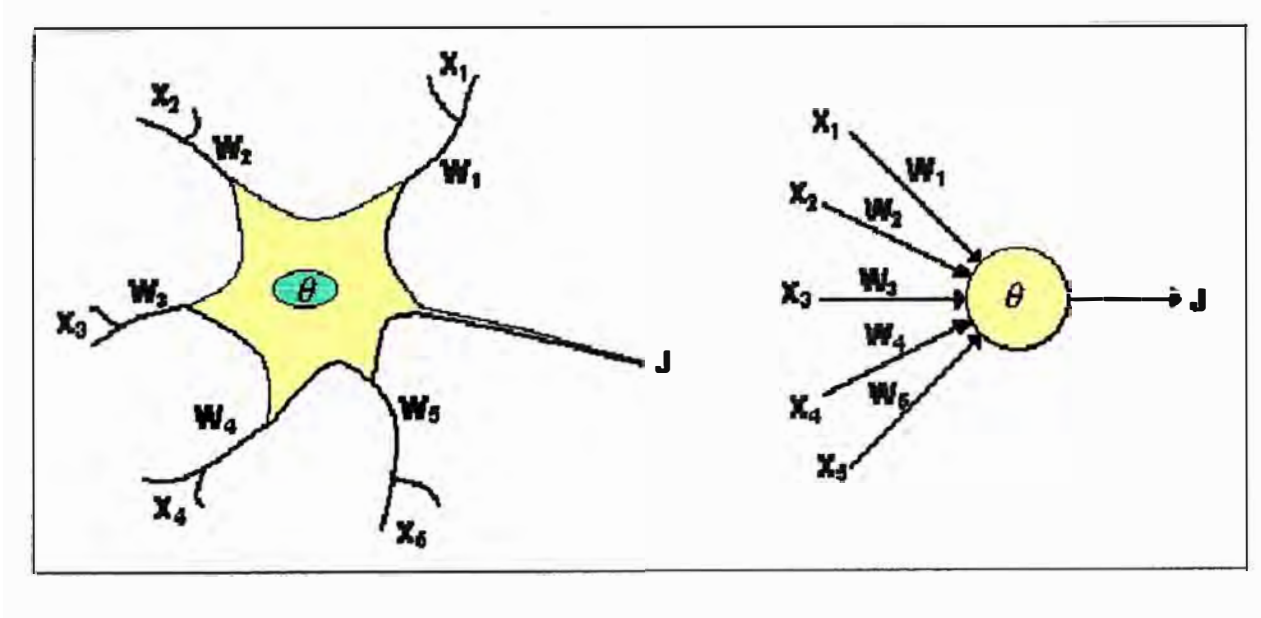


Figura 4.5: De la neurona biológica a al neurona artificial.

Como se vio en el proceso biológico y ayudándonos de la Figura 4.5, de las neuronas biológicas se pueden observar las siguientes analogías con las artificiales:

- Las entradas X_i representan las señales que provienen de otras neuronas y que son capturadas por las dendritas.
- Los pesos W_i, son la intensidad de la sinapsis que conecta dos neuronas; tanto *Xi* como W_i son valores reales.
- θ es la función umbral que la neurona debe sobrepasar para activarse; este proceso ocurre biológicamente en el cuerpo de la célula.

Las entradas de la neurona artificial son variables que pueden ser continuas o discretas a diferencia de las biológicas que solo son pulsos discretos. Cada señal de entrada pasa a través de una ganancia o peso, llamado peso sináptico o fortaleza de la conexión cuya función es similar a la función sináptica de la neurona biológicas. Las ganancias pueden ser excitatorias o inhibitorias, el nodo acumula todas las señales de entrada operándolas con su peso obteniendo una salida neta o total y las hace pasar por una función umbral y dependiendo del tipo de esta se obtiene la salida. De lo anterior podemos anotar:

$$neta_i = \sum_{i=1}^{n} W_i X_i = \overset{\rho\rho}{XY}$$

Para un mejor entendimiento del proceso se muestra la Figura 4.6.

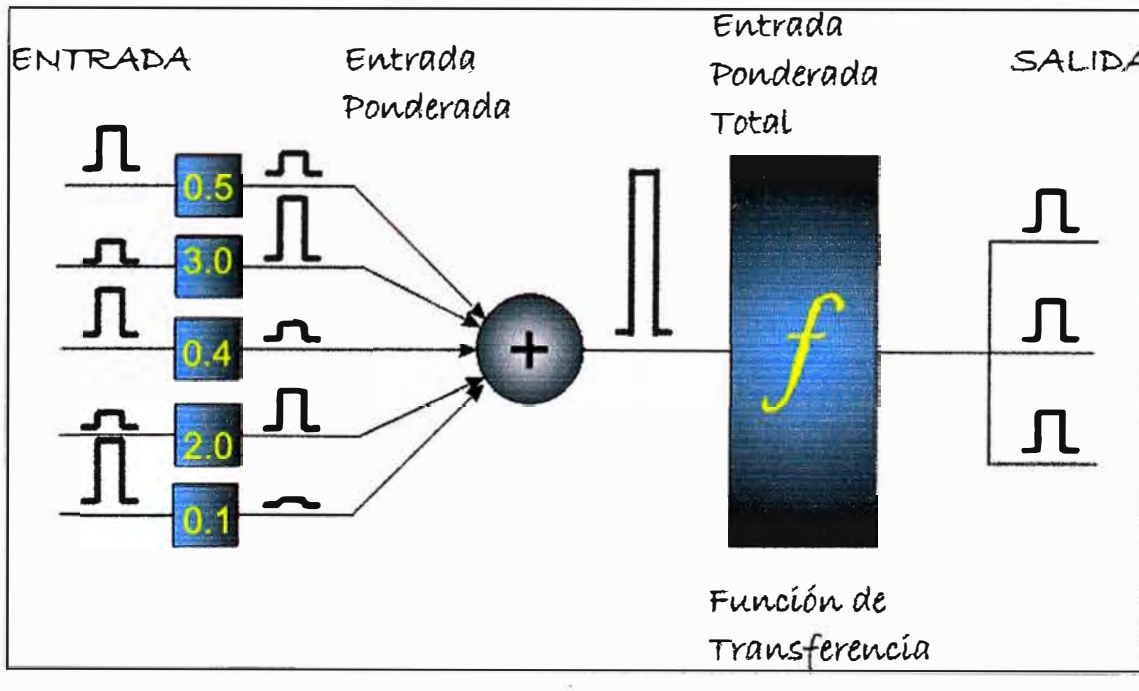


Figura 4.6: Proceso de una red neuronal.

Dependiendo del tipo de función que se le asigne a f se obtiene la activación y la salida seria.

$$x_i = f_i(neta_i)$$

Donde f_i representa la función de activación para una sola neurona correspondiente a la función escogida para operar la entrada total o neta y obtener la salida especifica de la neurona.

4.2 Clasificación de las Redes Neuronales

Como en todo sistema hay una variedad de RNA para realizar los diferentes tipos de procesos que existen variando en su arquitectura en cuanto a la forma de entrada y salida de la RNA, en la función umbral o de transferencia; es por ellos podemos hacer una clasificación de ellas presentándola en la Figura 4.7.

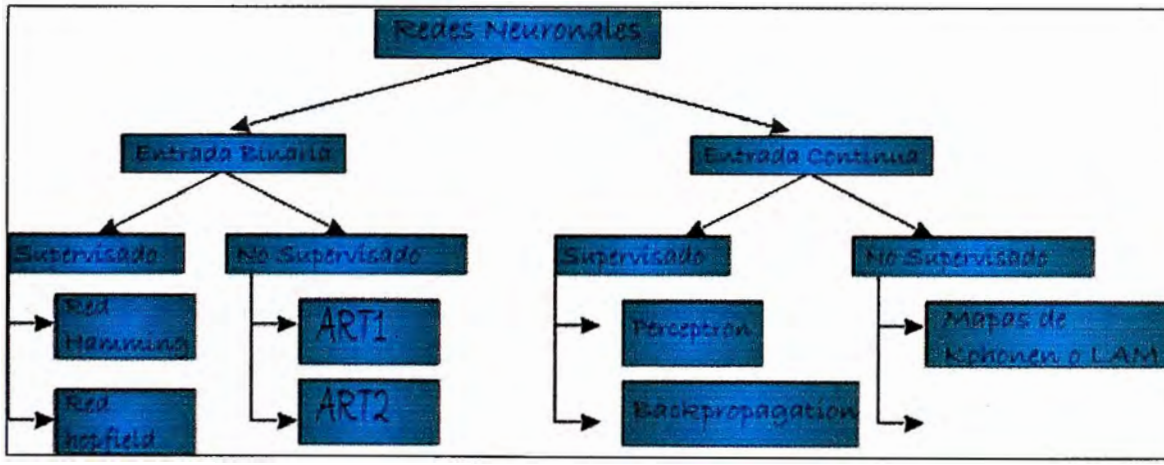


Figura 4.7: Proceso de una red neuronal.

Como se menciono anteriormente las señales de entrada son continuas o discretas que se pasan la forma binaria para el procesamiento; y así podemos dividirlos por su forma de aprendizaje que serian lo supervisados y no supervisados.

4.3 Redes Neuronales No Supervisadas

Son aquellos que no necesitan de un "maestro" que les enseñe o indique si se esta operando correcta o incorrectamente pues no dispone de ninguna salida objetivo hacia el cual le red neuronal deba tender. Así, durante el proceso de aprendizaje debe descubrir por si misma rasgos comunes, regularidades, correlaciones o categorías en los datos de entrada, e incorporarlos a su estructura interna de conexiones (pesos). Se dice en este caso que las neuronas deben autoorganizarse en función de los estímulos (señales o datos) procedentes del exterior. Para obtener resultados de calidad la red requiere un cierto nivel de redundancia en las entradas procedentes de espacio sensorial, hablando técnicamente de un grupo grande de patrones de aprendizaje.

Los procesamientos de este tipo de redes pueden ser: análisis de similitud entre patrones, análisis de comportamientos principales, clasificación de patrones, codificación, etc.

4.4 Redes Neuronales Competitivas.

En los últimos diecisiete años, las redes neuronales artificiales (RNA) han emergido como una potente herramienta para el modelado estadístico orientada principalmente al reconocimiento de patrones, tanto en la vertiente de clasificación como de predicción.

Las RNA poseen una serie de características admirables, tales como la habilidad para procesar datos con ruido o incompletos, la alta tolerancia a fallos que permite a la red operar satisfactoriamente con neuronas o conexiones dañadas y la capacidad de responder en tiempo real debido a su paralelismo inherente.

4.5 Los mapas autoorganizados de Kohonen

En 1982 Teuvo Kohonen presentó un modelo de red denominado mapas autoorganizados o SOM (Self-Organizing Maps), basado en ciertas evidencias descubiertas a nivel cerebral y con un gran potencial de aplicabilidad práctica. Este tipo de red se caracteriza por poseer un aprendizaje no supervisado competitivo.

Ahora veremos en qué consiste este tipo de aprendizaje. Durante el proceso de aprendizaje la red autoorganizada debe descubrir por sí misma rasgos comunes, regularidades, correlaciones o categorías en los datos de entrada, e incorporarlos a su estructura interna de conexiones. Se dice, por tanto, que las neuronas deben autoorganizarse en función de los estímulos (datos) procedentes del exterior.

Dentro del aprendizaje no supervisado existe un grupo de modelos de red caracterizados por poseer un aprendizaje competitivo. En el aprendizaje competitivo las neuronas compiten unas con otras con el fin de llevar a cabo una tarea dada. Con este tipo de aprendizaje, se pretende que cuando se presente a la red un patrón de entrada, sólo una de las neuronas de salida (o un grupo de vecinas) se active. Por tanto, las neuronas compiten por activarse, quedando finalmente una como neurona vencedora anulándose las perdedoras, que son forzadas a sus valores de respuesta mínimos.

4.5.1 Arquitectura

Un modelo SOM está compuesto por dos capas de neuronas. La capa de entrada (formada por N neuronas, una por cada variable de entrada) se encarga de recibir y transmitir a la capa de salida la información procedente del exterior. La capa de salida (formada por M neuronas) es la encargada de procesar la información y formar el mapa de rasgos. Normalmente, las neuronas de la capa de salida se organizan en forma de mapa bidimensional como se muestra en la Figura 4.8, aunque a veces también se utilizan capas de una sola dimensión (cadena lineal de neuronas) o de tres dimensiones (paralelepípedo).

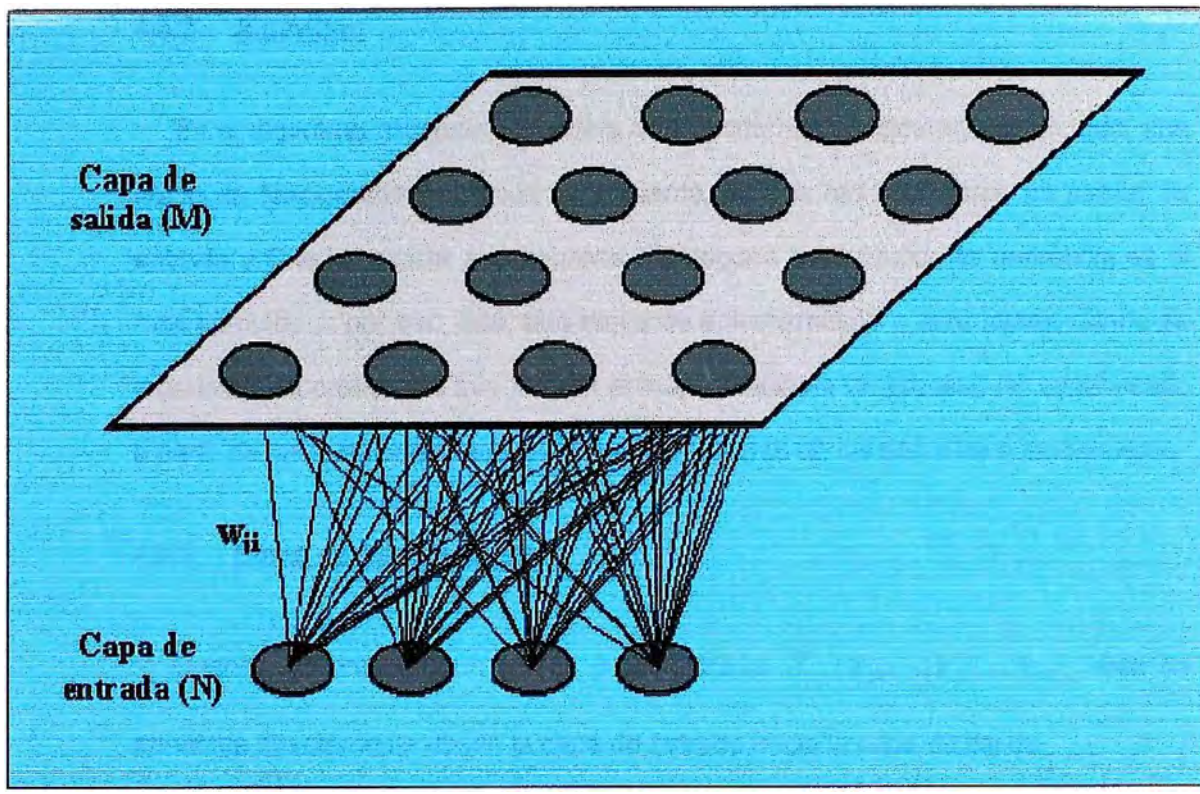


Figura 4.8: Arquitectura del SOM.

Las conexiones entre las dos capas que forman la red son siempre hacia delante, es decir, la información se propaga desde la capa de entrada hacia la capa de salida. Cada neurona de entrada i está conectada con cada una de las neuronas de salida j mediante un peso w_{ji}. De esta forma, las neuronas de salida tienen asociado un vector de pesos W_j llamado vector de referencia (o *codebook*), debido a que constituye el vector prototipo (o promedio) de la categoría representada por la neurona de salida j. Entre las neuronas de la capa de salida, puede decirse que existen conexiones laterales de excitación e inhibición implícitas, pues aunque no estén conectadas, cada una de estas neuronas va a tener cierta influencia sobre sus vecinas. Esto se consigue a través de un proceso de competición entre las neuronas y de la aplicación de una función denominada de vecindad como veremos más adelante.

4.5.2 Algoritmo

En el algoritmo asociado al modelo SOM podemos considerar, por un lado, una etapa de funcionamiento donde se presenta, ante la red entrenada, un patrón de entrada y éste se asocia a la neurona o categoría cuyo vector de referencia es el más parecido y, por otro lado, una etapa de entrenamiento o aprendizaje donde se organizan las categorías que forman el mapa mediante un proceso no supervisado a partir de las relaciones descubiertas en el conjunto de los datos de entrenamiento.

4.5.3 Etapa de funcionamiento

Cuando se presenta un patrón *p* de entrada $X_p : x_{p1},...,x_{pi},...,x_{pN}$, éste se transmite directamente desde la capa de entrada hacia la capa de salida.

En esta capa, cada neurona calcula la similitud entre el vector de entrada X_p y su propio vector de pesos W_j o vector de referencia según una cierta medida de distancia o criterio de similitud establecido. A continuación, simulando un proceso competitivo, se declara vencedora la neurona cuyo vector de pesos es el más similar al de entrada.

La siguiente expresión matemática representa cuál de las M neuronas se activará al presentar el patrón de entrada X_p:

$$y_{pj} = \begin{cases} 1 .....\min \left\| X_p - W_j \right\| \\ 0 ....resto \end{cases}$$

Donde y_{pj} representa la salida o el grado de activación de las neuronas de salida en función del resultado de la competición (1 = neurona vencedora, 0 = neurona no vencedora), $\left\| X_p - W_j \right\|$ representa una medida de similitud entre el

vector o patrón de entrada $X_p : x_{p1},...,x_{pi},...,x_{pN}$, y el vector de pesos $X_p : w_{p1},...,w_{pi},...,w_{pN}$, de las conexiones entre cada una de las neuronas de entrada y la neurona de salida j.

En esta etapa de funcionamiento, lo que se pretende es encontrar el vector de referencia más parecido al vector de entrada para averiguar qué neurona es la vencedora y, sobre todo, en virtud de las interacciones excitatorias e inhibitorias que existen entre las neuronas, para averiguar en qué zona del espacio bidimensional de salida se encuentra tal neurona. Por tanto, lo que hace la red SOM es realizar una tarea de clasificación, ya que la neurona de salida activada ante una entrada representa la clase a la que pertenece dicha información de entrada. Además, como ante otra entrada parecida se activa la misma neurona de salida, u otra cercana a la anterior, debido a la semejanza entre las clases, se garantiza que las neuronas topológicamente próximas sean sensibles a entradas físicamente similares.

Por este motivo, la red es especialmente útil para establecer relaciones, desconocidas previamente, entre conjuntos de datos.

4.5.4 Etapa de aprendizaje

Se debe advertir, en primer lugar, que no existe un algoritmo de aprendizaje totalmente estándar para la red SOM. Sin embargo, se trata de un procedimiento bastante robusto ya que el resultado final es en gran medida independiente de los detalles de su realización concreta. En consecuencia, trataremos de exponer el algoritmo más habitual asociado a este modelo.

El algoritmo de aprendizaje trata de establecer, mediante la presentación de un conjunto de patrones de entrenamiento, las diferentes categorías (una por neurona de salida) que servirán durante la etapa de funcionamiento para realizar clasificaciones de nuevos patrones de entrada.

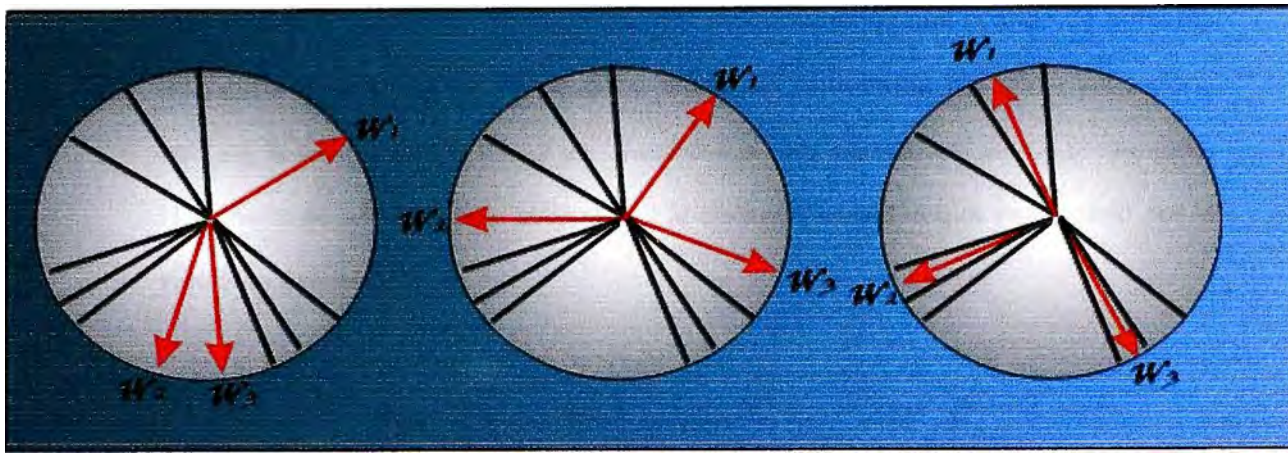
De forma simplificada, el proceso de aprendizaje se desarrolla de la siguiente manera. Una vez presentado y procesado un vector de entrada, se establece a partir de una medida de similitud, la neurona vencedora, esto es, la neurona de salida cuyo vector de pesos es el más parecido respecto al vector de entrada. A continuación, el vector de pesos asociado a la neurona vencedora se modifica de manera que se parezca un poco más al vector de entrada. De este modo, ante el mismo patrón de entrada, dicha neurona responderá en el futuro todavía con más intensidad. El proceso se repite para un conjunto de patrones de entrada los cuales son presentados repetidamente a la red, de forma que al final los diferentes vectores de pesos sintonizan con uno o varios patrones de entrada y, por tanto, con dominios específicos del espacio de entrada. Si dicho espacio está dividido en grupos, cada neurona se especializará en uno de ellos, y la operación esencial de la red se podrá interpretar como un análisis de clusters.

La siguiente interpretación geométrica del proceso de aprendizaje puede resultar interesante para comprender la operación de la red SOM. El efecto de la regla de aprendizaje no es otro que acercar de forma iterativa el vector de pesos de la neurona de mayor actividad (ganadora) al vector de entrada. Así, en cada iteración el vector de pesos de la neurona vencedora rota hacia el de entrada, y se aproxima a él en una cantidad que depende del tamaño de una tasa de aprendizaje.

En la Figura 4.9 se muestra cómo opera la regla de aprendizaje para el caso de varios patrones pertenecientes a un espacio de entrada de dos dimensiones, representados en la figura por los vectores de color negro. Supongamos que los vectores del espacio de entrada se agrupan en tres clusters, y supongamos que el número de neuronas de la red es también tres. Al principio del entrenamiento los vectores de pesos de las tres neuronas (representados por vectores de color rojo) son aleatorios y se distribuyen por la circunferencia. Conforme avanza el aprendizaje, éstos se van acercando progresivamente a las muestras procedentes

del espacio de entrada, para quedar finalmente estabilizados como centroides de los tres clusters.

Figura 4.9: proceso de aprendizaje en dos dimensiones.



Al finalizar el aprendizaje, el vector de referencia de cada neurona corresponderá con el vector de entrada que consigue activar la neurona correspondiente.

En el caso de existir más patrones de entrenamiento que neuronas de salida, como en el ejemplo expuesto, más de un patrón deberá asociarse con la misma neurona, es decir, pertenecerán a la misma clase. En tal caso, los pesos que componen el vector de referencia se obtienen como un promedio (centroide) de dichos patrones.

Además de este esquema de aprendizaje competitivo, el modelo SOM aporta una importante novedad, pues incorpora relaciones entre las neuronas próximas en el mapa.

Para ello, introduce una función denominada zona de vecindad que define un entorno alrededor de la neurona ganadora actual (vecindad); su efecto es que durante el aprendizaje se actualizan tanto los pesos de la vencedora como los de las neuronas pertenecientes a su vecindad. De esta manera, en el modelo SOM se

logra que neuronas próximas sintonicen con patrones similares, quedando de esta manera reflejada sobre el mapa una cierta imagen del orden topológico presente en el espacio de entrada.

Una vez entendida la forma general de aprendizaje del modelo SOM, vamos a expresar este proceso de forma matemática. Recordemos que cuando se presenta un patrón de entrenamiento, se debe identificar la neurona de salida vencedora, esto es, la neurona cuyo vector de pesos sea el más parecido al patrón presentado. Un criterio de similitud muy utilizado es la distancia euclídea que viene dado por la siguiente expresión:

$$\min\left\|X_p - W_j\right\| = \min \sum_{j=1}^{N} (x_{pi} - W_{ji})^2$$

De acuerdo con este criterio, dos vectores serán más similares cuanto menor sea su distancia.

Una medida de similitud alternativa más simple que la euclídea, es la correlación o producto escalar:

$$\min\left\|X_p - W_j\right\| = \max \sum_{j=1}^{N} x_{pi} \cdot W_{ji}$$

Según la cual, dos vectores serán más similares cuanto mayor sea su correlación.

Identificada la neurona vencedora mediante el criterio de similitud, podemos pasar a modificar su vector de pesos asociado y el de sus neuronas vecinas, según la regla de aprendizaje:

$$W_{k+1} = W_k + \mu_k \cdot (X_k - W_k)$$

Donde k hace referencia al número de ciclos o iteraciones, esto es, el número de veces que ha sido presentado y procesado todo el juego de patrones de entrenamiento llamados épocas, μ_k es la tasa de aprendizaje que, con un valor inicial entre 0 y 1, decrece con el número de iteraciones (n) del proceso de aprendizaje. La variación de los pesos se hará en la zona de vecindad alrededor de la neurona vencedora j^* en la que se encuentran las neuronas cuyos pesos son actualizados.

Tradicionalmente el ajuste de los pesos se realiza después de presentar cada vez un patrón de entrenamiento, como se muestra en la regla de aprendizaje expuesta.

Con cada iteración el rate de aprendizaje disminuiría de la siguiente manera:

$$\mu_k = \mu_0 \cdot \left(1 - \frac{k}{It}\right)$$

Donde It es el número de iteraciones que se harán.

Cuando un mapa autoorganizado es diseñado por primera vez, se deben asignar valores a los pesos a partir de los cuales comenzar la etapa de entrenamiento. En general, no existe discusión en este punto y los pesos se inicializan con pequeños valores aleatorios, por ejemplo, entre -1 y 1 ó entre 0 y 1, aunque también se pueden inicializar con valores nulos o a partir de una selección aleatoria de patrones de entrenamiento [7].

CAPITULO V

LA VOZ HUMANA

5.1 CONCEPTOS PRELIMINARES

5.1.1 Comunicación y lenguaje

Los sistemas de comunicación transportan información. Nos proponemos estudiar un sistema de comunicación específico, el de la comunicación a través de señales de voz, es decir señales acústicas tradicionalmente emitidas y recibidas por seres humanos en forma oral.

Históricamente, desde la Antigua Grecia se han realizado intentos por generar voces artificiales. En muchos casos eran simplemente juegos de tuberías conectadas a un locutor humano, en otros auténticos ingenios acústicos capaces de producir sonoridades vocálicas. El desarrollo de la telefonía a principios del siglo XX motivó intensas investigaciones sobre las propiedades de la voz y la audición con el fin de mejorar la calidad de la comunicación telefónica. El proceso continuó y hoy en día las tecnologías existentes permiten, por ejemplo, disponer de sistemas de comunicación oral hombre-máquina.

En todo sistema de comunicación hay varios componentes: emisor, receptor, mensaje, código, canal y contexto. Es necesario conocer algunos aspectos de cada uno de ellos para poder integrar sistemas que funcionen de manera eficaz y eficiente. En nuestro caso el emisor es el conjunto integrado por el cerebro que "piensa" el mensaje y el aparato fonatorio que lo "traduce" a una emisión acústica. El receptor es el aparato auditivo que recibe la onda sonora y la transforma en impulsos nerviosos que luego son interpretados por el cerebro. El mensaje es la idea a comunicar. El código es el lenguaje hablado. La combinación del mensaje y el código constituyen la señal. El canal puede ser el medio en el cual se propaga la onda sonora (en general el aire) o un medio de transmisión electrónico que constituye en sí mismo otro subsistema de comunicación cuyas propiedades son bien conocidas y que se aproxima en muchos casos (aunque no siempre) a la idealidad. El contexto puede tener un sinnúmero de componentes, que van desde factores puramente subjetivos o psicológicos, como el interés, la atención, la motivación hasta factores físicos tales como respuesta en frecuencia, interferencias, distorsiones, ruido, etc.

5.1.2 Algunos Conceptos Sobre Lenguaje

La lengua es un sistema de signos lingüísticos que permiten la comunicación en una comunidad. Es un sistema pues cada uno de sus elementos tiene entidad propia y entidad relativa a su posición o relación con los otros elementos. Es un código de signos. Tiene carácter social, ya que es común a una sociedad.

El habla es el acto de seleccionar los signos de entre los disponibles y organizarlos a través de ciertas reglas. Materializa el código. Es individual, vale decir que cambia de un individuo a otro. Los signos pueden corresponder al lenguaje escrito o al oral.

El lenguaje es un sistema articulado ya que los sonidos y otros componentes se integran entre sí. Está formado por signos lingüísticos, nombre que recibe la señal en el lenguaje.

El lenguaje tiene modalidades regionales llamadas dialectos.

Un signo es algo que reemplaza a otra cosa para comunicarla en un mensaje.

Los signos lingüísticos se clasifican en dos tipos: significado y significante. El significado es el concepto mental, idea o contenido a comunicar. El significante es la imagen, ya sea gráfica o acústica que se le asigna.

La relación entre significado y significante es arbitraria o convencional, aunque no necesariamente discrecional: involucra acuerdos tácitos, explícitos o normativos en una comunidad lingüística.

En el lenguaje escrito, el significante es la grafía escrita, formada por combinaciones de letras, en tanto que en el lenguaje hablado es su realización acústica mediante la palabra hablada.

Las palabras son los elementos libres mínimos del lenguaje. La sintaxis es el conjunto de reglas para la coordinación de las palabras en frases u oraciones. En su versión escrita las palabras están formadas por letras o grafemas, es decir unidades gráficas mínimas, y, en el caso oral, por fonemas.

Los fonemas son la unidad fónica ideal mínima del lenguaje. Se materializan a través de los sonidos, pero de una manera no unívoca. Las variantes de los fonemas se denominan alófonos.

Los monemas son unidades mínimas con significado, que puede ser gramatical, dando origen a los morfemas, o léxico, representado por los lexemas. Los morfemas tienen relación con la gramática, o la forma de organizar o dar estructura a las

categorías básicas del lenguaje (género, número, tiempo o persona de los verbos, etc.), mientras que los lexemas se refieren a significados externos al lenguaje mismo.

Las palabras constan de al menos un monema, siendo las más comunes bimonemáticas, que incluyen un lexema y un morfema. En la tabla siguiente se dan dos ejemplos en los que se identifican los componentes de la palabra.

Tabla 5.1: Identificación de los componentes de la palabra.

Palabra	**Monemas**		**Grafemas**	**Fonemas**
	Lexema	Morfema		
Gato	Gat	o	G, a, t, o	/g/, /a/, /t/, /o/
Amaban	Ama	ban	A, m, a, b, a, n	/a/, /m/, /a/, /b/, /a/, /n/

5.1.3 Fonología y fonética

La Fonología estudia los fonemas, es decir el modelo fónico convencional e ideal del lenguaje. La Fonética, en tanto, se refiere a los sonidos en el habla, incluyendo su producción acústica y los procesos físicos y fisiológicos de emisión y articulación involucrados.

Así, la Fonología es el estudio de los sonidos de la lengua en cuanto a su carácter simbólico o de representación mental. Procede detectando regularidades o recurrencias en los sonidos del lenguaje hablado y sus combinaciones, y haciendo abstracción de las pequeñas diferencias debidas a la individualidad de cada hablante y de características suprasegmentales como la entonación, el acento (tónico, es decir por aumento de la intensidad y agógico, por aumento de la

duración), etc. Cada uno de los sonidos abstractos así identificados es un fonema. Uno de los objetivos de la fonología es acotar al máximo la cantidad de fonemas requeridos para representar cada idioma de una manera suficientemente precisa.

La Fonética estudia experimentalmente los mecanismos de producción y percepción de los sonidos utilizados en el habla a través del análisis acústico, articulatorio y perceptivo. Se ocupa, por consiguiente, de las realizaciones de los fonemas.

5.2 La voz humana

5.2.1 Breve explicación de la anatomía del aparato fonatorio

La voz humana se produce voluntariamente por medio del aparato fonatorio. Éste está formado por los pulmones como fuente de energía en la forma de un flujo de aire, la laringe, que contiene las cuerdas vocales, la faringe, las cavidades oral (o bucal) y nasal y una serie de elementos articulatorios: los labios, los dientes, el alvéolo, el paladar, el velo del paladar y la lengua tal como se puede ver en la Figura 5.1. Las cuerdas vocales son, en realidad, dos membranas dentro de la laringe orientadas de adelante hacia atrás como se muestra en la Figura 5.2. Por adelante se unen en el cartílago tiroides (que puede palparse sobre el cuello, inmediatamente por debajo de la unión con la cabeza; en los varones suele apreciarse como una protuberancia conocida como manzana de Adán). Por detrás, cada una está sujeta a uno de los dos cartílagos aritenoides, los cuales pueden separarse voluntariamente por medio de músculos. La abertura entre ambas cuerdas se denomina glotis.

Cuando las cuerdas vocales se encuentran separadas, la glotis adopta una forma triangular. El aire pasa libremente y prácticamente no se produce sonido. Es el caso de la respiración.

Cuando la glotis comienza a cerrarse, el aire que la atraviesa proveniente de los pulmones experimenta una turbulencia, emitiéndose un ruido de origen aerodinámico conocido como aspiración (aunque en realidad acompaña a una espiración o exhalación). Esto sucede en los sonidos denominados "aspirados" (como la h inglesa). Al cerrarse más, las cuerdas vocales comienzan a vibrar a modo de lengüetas, produciéndose un sonido tonal, es decir periódico. La frecuencia de este sonido depende de varios factores, entre otros del tamaño y la masa de las cuerdas vocales, de la tensión que se les aplique y de la velocidad del flujo del aire proveniente de los pulmones. A mayor tamaño, menor frecuencia de vibración, lo cual explica por qué en los varones, cuya glotis es en promedio mayor que la de las mujeres, la voz es en general más grave. A mayor tensión la frecuencia aumenta, siendo los sonidos más agudos. Así, para lograr emitir sonidos en el registro extremo de la voz es necesario un mayor esfuerzo vocal. También aumenta la frecuencia (a igualdad de las otras condiciones) al crecer la velocidad del flujo de aire, razón por la cual al aumentar la intensidad de emisión se tiende a elevar espontáneamente el tono de voz.

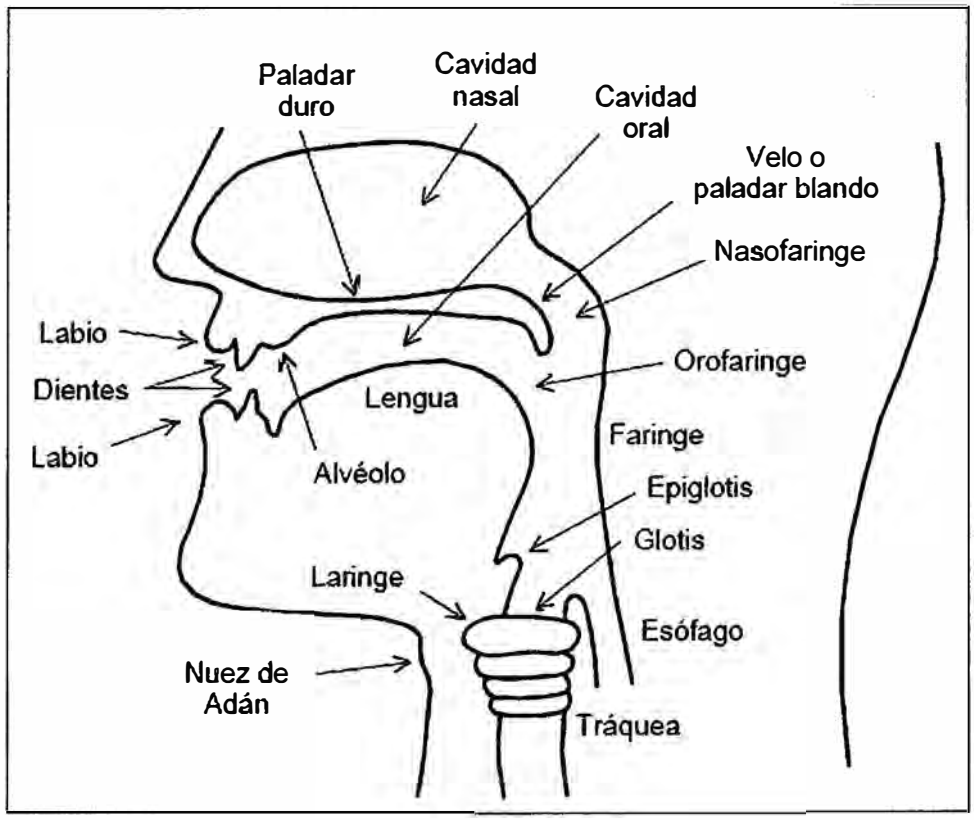


Figura 5.1: Corte esquemático del aparato fonatorio humano.

Finalmente, es posible obturar la glotis completamente. En ese caso no se produce sonido. Sobre la glotis se encuentra la epiglotis, un cartílago en la faringe que permite tapar la glotis durante la deglución para evitar que el alimento ingerido se introduzca en el tracto respiratorio. Durante la respiración y la fonación (emisión de sonido) la epiglotis está separada de la glotis permitiendo la circulación del flujo de aire. Durante la deglución, en cambio, la laringe ejecuta un movimiento ascendente de modo que la glotis apoya sobre la epiglotis.

La porción que incluye las cavidades faríngea, oral y nasal junto con los elementos articulatorios se denomina genéricamente cavidad supraglótica, en tanto que los espacios por debajo de la laringe, es decir la tráquea, los bronquios y los pulmones, se denominan cavidades infraglóticas.

Varios de los elementos de la cavidad supraglótica se controlan a voluntad, permitiendo modificar dentro de márgenes muy amplios los sonidos producidos por las cuerdas vocales o agregar partes distintivas a los mismos, e inclusive producir sonidos propios. Todo esto se efectúa por dos mecanismos principales: el filtrado y la articulación.

El filtrado actúa modificando el espectro del sonido. Tiene lugar en las cuatro cavidades supraglóticas principales: la faringe, la cavidad nasal, la cavidad oral y la cavidad labial. Las mismas constituyen resonadores acústicos que enfatizan determinadas bandas frecuenciales del espectro generado por las cuerdas vocales, conduciendo al concepto de formantes, es decir una serie de picos de resonancia ubicados en frecuencias o bandas de frecuencia que, según veremos, son bastante específicas para cada tipo de sonido.

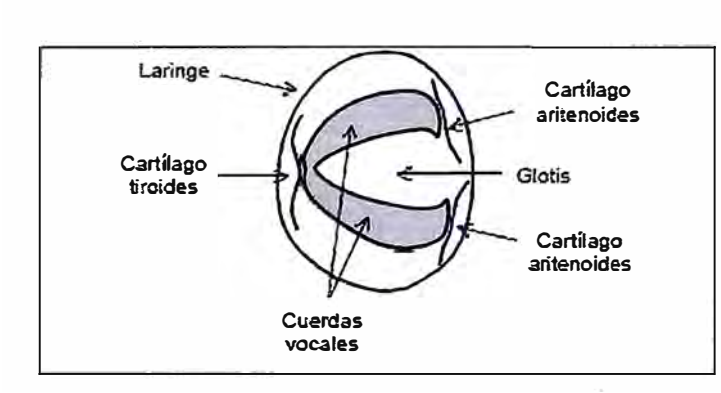


Figura 5.2: Corte esquemático de la laringe según un plano horizontal.

La articulación es una modificación principalmente a nivel temporal de los sonidos, y está directamente relacionada con la emisión de los mismos y con los fenómenos transitorios que los acompañan. Está caracterizada por el lugar del tracto vocal en que tiene lugar, por los elementos que intervienen y por el modo en que se produce, factores que dan origen a una clasificación fonética de los sonidos que veremos luego.

5.2.2 Clasificación de los sonidos de la voz

Los sonidos emitidos por el aparato fonatorio pueden clasificarse de acuerdo con diversos criterios que tienen en cuenta los diferentes aspectos del fenómeno de emisión.

Estos criterios son:

a) Según su carácter vocálico o consonántico.

b) Según su oralidad o nasalidad

c) Según su carácter tonal (sonoro) o no tonal (sordo)

d) Según el lugar de articulación

e) Según el modo de articulación

f) Según la posición de los órganos articulatorios

g) Según la duración

5.2.2.1 Vocales y consonantes

Desde un punto de vista mecanoacústico, las vocales son los sonidos emitidos por la sola vibración de las cuerdas vocales sin ningún obstáculo o constricción entre la laringe y las aberturas oral y nasal. Dicha vibración se genera por el principio del oscilador de relajación, donde interviene una fuente de energía constante en la forma de un flujo de aire proveniente de los pulmones. Son siempre sonidos de carácter tonal (cuasiperiódicos), y por consiguiente de espectro discreto. Las consonantes, por el contrario, se emiten interponiendo algún obstáculo formado por los elementos articulatorios. Los sonidos correspondientes a las consonantes pueden ser tonales o no dependiendo de si las cuerdas vocales están vibrando o no. Funcionalmente, en el castellano las vocales pueden constituir palabras completas, no así las consonantes.

5.2.2.2 Oralidad y nasalidad

Los fonemas en los que el aire pasa por la cavidad nasal se denominan nasales, en tanto que aquéllos en los que sale por la boca se denominan orales. La diferencia principal está en el tipo de resonador principal por encima de la laringe (cavidad nasal y oral, respectivamente). En castellano son nasales sólo las consonantes "m", "n", "ñ"

5.2.2.3 Tonalidad

Los fonemas en los que participa la vibración de las cuerdas vocales se denominan tonales o también sonoros. La tonalidad lleva implícito un espectro cuasi periódico. Como se puntualizó anteriormente, todas las vocales son tonales, pero existen varias consonantes que también lo son: "b", "d", "m", etc. Aquellos fonemas producidos sin vibraciones glotales se denominan sordos. Varios de ellos son el resultado de la turbulencia causada por el aire pasando a gran velocidad por un espacio reducido, como las consonantes "s", "z", "j", "f".

5.2.2.4 Lugar y modo de articulación (consonantes)

La articulación es el proceso mediante el cual alguna parte del aparato fonatorio interpone un obstáculo para la circulación del flujo de aire. Las características de la articulación permitirán clasificar las consonantes. Los órganos articulatorios son los labios, los dientes, las diferentes partes del paladar (alvéolo, paladar duro, paladar blando o velo), la lengua y la glotis. Salvo la glotis, que puede articular por sí misma, el resto de los órganos articula por oposición con otro. Según el lugar o punto de articulación se tienen fonemas:

- Bilabiales: oposición de ambos labios.
- Labiodentales: oposición de los dientes superiores con el labio inferior.

- Linguodentales: oposición de la punta de la lengua con los dientes superiores.
- Alveolares: oposición de la punta de la lengua con la región alveolar.
- Palatales: oposición de la lengua con el paladar duro.
- Velares: oposición de la parte posterior de la lengua con el paladar blando.
- Glotales: articulación en la propia glotis.

A su vez, para cada punto de articulación ésta puede efectuarse de diferentes modos, dando lugar a fonemas:

- Oclusivos: la salida del aire se cierra momentáneamente por completo.
- Fricativos: el aire sale atravesando un espacio estrecho.
- Africados: oclusión seguida por fricación.
- Laterales: la lengua obstruye el centro de la boca y el aire sale por los lados.
- Vibrantes: la lengua vibra cerrando el paso del aire intermitentemente.
- Aproximantes: La obstrucción muy estrecha que no llega a producir turbulencia.

Los fonemas oclusivos (correspondientes a las consonantes "b", "c", "k", "d", "g", "p", "t") también se denominan a veces explosivos, debido a la liberación repentina de la presión presente inmediatamente antes de su emisión. Pueden ser sordos o sonoros, al igual que los fricativos ("b" intervocálica, "f", "j", "h" aspirada, "s", "y", "z"). Sólo existe un fonema africado en castellano, correspondiente a la "ch". Los laterales ("l", "ll") a veces se denominan líquidos, y son siempre sonoros. Los dos fonemas vibrantes del castellano (consonantes "r", "rr") difieren en que en uno de ellos ("r") se ejecuta una sola vibración y es intervocálico, mientras que en el otro

("rr") es una sucesión de dos o tres vibraciones de la lengua. Finalmente, los fonemas aproximantes (la "i" y la "u" cerradas que aparecen en algunos diptongos) son a veces denominados semivocales, pues en realidad suenan como vocales. Pero exhiben una diferencia muy importante: son de corta duración y no son prolongables.

En la Tabla 5.2 se indican las consonantes clasificadas según el lugar y el modo de articulación, la sonoridad y la oro-nasalidad. En algunos casos una misma consonante aparece en dos categorías diferentes, correspondiente a las diferencias observadas.

Tabla 5.2: Clasificación de las consonantes de la lengua castellana según el lugar y el modo de articulación y la sonoridad.

Lugar de articulación	Modo de articulación								
	Oral								Nasal
	Oclusiva		Fricativa		Africada	Lateral	Vibrante	Aproximante	
	Sorda	Sonora	Sorda	Sonora	Sorda	Sonora	Sonora	Sonora	Sonora
Bilabial	p	b, v		b, v				w	m
Labiodental			f						
Linguodental			z						
Alveolar	t	d	s	y	ch	l	r, rr		n
Palatal				(y)	(ch)	ll		i	ñ
Velar	k	g	j						
Glotal			h						

5.2.2.5 Posición de los órganos articulatorios (vocales)

En el caso de las vocales, la articulación consiste en la modificación de la acción filtrante de los diversos resonadores, lo cual depende de las posiciones de la lengua (tanto en elevación como en profundidad o avance), de la mandíbula inferior, de los labios y del paladar blando. Estos órganos influyen sobre los formantes, permitiendo su control.

Podemos clasificar las vocales según la posición de la lengua como se muestra en la Tabla 5.3.

Tabla 5.3: Clasificación de las consonantes de la lengua castellana según el lugar y el modo de articulación y la sonoridad.

Posición vertical	Tipo de vocal	Posición horizontal (avance)		
		Anterior	Central	Posterior
Alta	Cerrada	i		u
Media	Media	e		o
Baja	Abierta		a	

Otra cualidad controlable es la labialización, es decir el hecho de que se haga participar activamente los labios. Las vocales labializadas, también definidas como redondeadas, son las que redondean los labios hacia adelante, incrementando la longitud efectiva del tracto vocal. La única vocal labializada en el castellano es la

En otros idiomas, como el francés, el portugués, el catalán y el polaco, así como en lenguas no europeas como el guaraní o el hindu, existe también el matiz de oralidad o nasalidad. En las vocales orales el velo (paladar blando) sube, obturando la nasofaringe, lo cual impide que el aire fluya parcialmente por la cavidad nasal. En las vocales nasalizadas (u oronasales) el velo baja, liberando el paso del aire a través de la nasofaringe. Se incorpora así la resonancia nasal.

5.2.2.6 Duración

La duración de los sonidos, especialmente de las vocales, no tiene importancia a nivel semántico en el castellano, pero sí en el plano expresivo, por medio del énfasis o acentuación a través de la duración. En inglés, en cambio, la duración de una vocal puede cambiar completamente el significado de la palabra que la contiene [8].

5.3 Análisis Espectral De La Voz Humana

Tomando de referencia a la teoría se sabe que la mejor forma de tratar a las vocales pronunciadas es como variable aleatoria [4]. Además, el análisis en la frecuencia de esas señales es obteniendo la estimación de su espectro de frecuencias, el análisis en la frecuencia de una señal aleatoria se realiza al aplicarle la trasformada de Fourier a la función de autocorrelación, esto produce a lo que se llama densidad espectral de potencia (power spectral density PSD). El método usado para este trabajo es la estimación de Welch que es una modificación del método de Bartlett [6].

Usando partes de código del sistema diseñado se puede obtener la PSD de cada una de las vocales como por ejemplo se muestra en la figura 5.3, que son las PSD de las vocales pronunciadas por el autor de este trabajo.

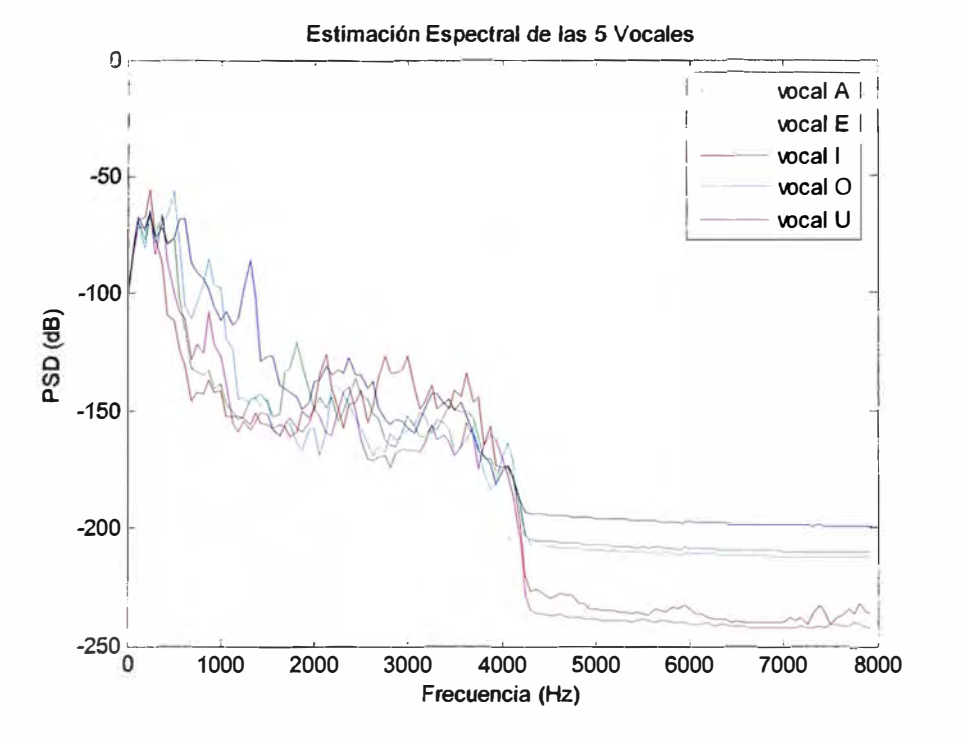


Figura 5.3: La figura muestra la estimación espectral de las 5 vocales para una misma persona.

Ahora pasaremos a observar como se presenta la PSD al pronunciar la vocal 'A' en diferentes intensidades de voz.

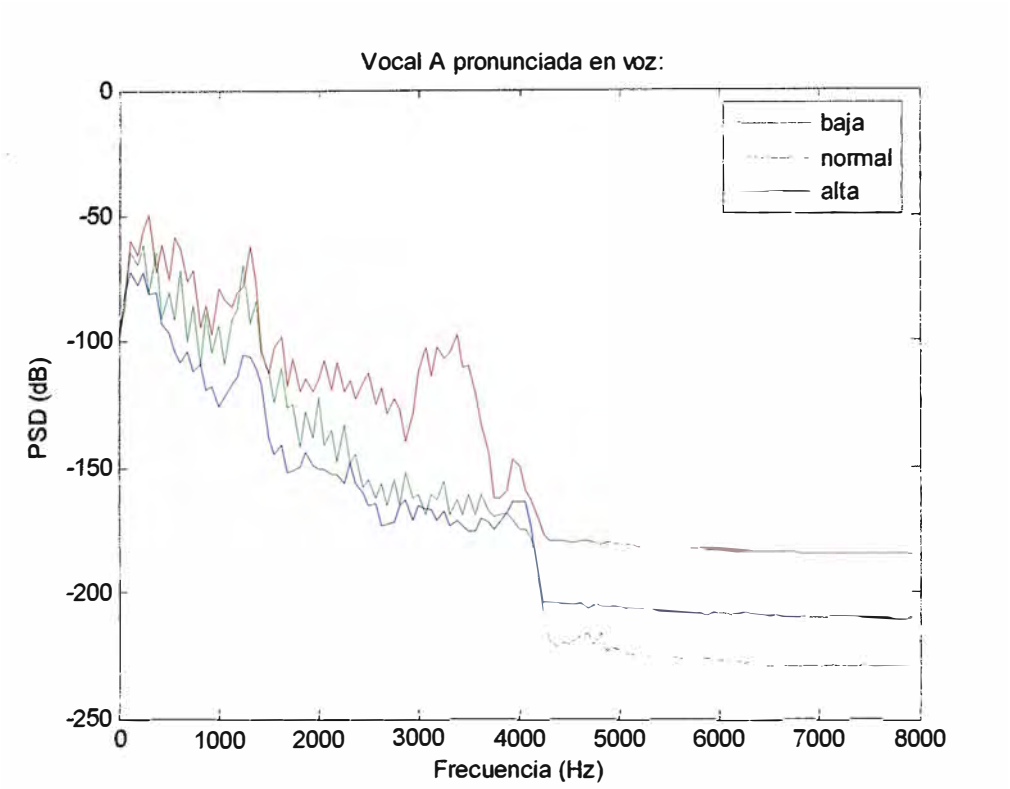


Figura 5.4: La figura muestra la PSD de la vocal A pronunciadas a diferentes intensidades de voz

Como se observa en la Figura 5.4 la intensidad de voz es reflejada en el área bajo la curva de la PSD, esta prueba puede simular el estado de animo de las personas y podemos tener un indicador del estado de animo de las personas al calcular el área bajo la curva.

También podemos observar que la formante, pico formado en el espectro de una determinada frecuencia, de la vocal A de la Figura 5.3 que se encuentra entre 1000 y 2000 Hz. se mantiene en la Figura 5.4, de esto podemos deducir que cada vocal tiene un PSD que se mantiene constante a lo largo del tiempo como se puede observar en la Figura 5.5 al pronunciar la 'E' en diferentes tiempos pero manteniendo la misma intensidad de voz.

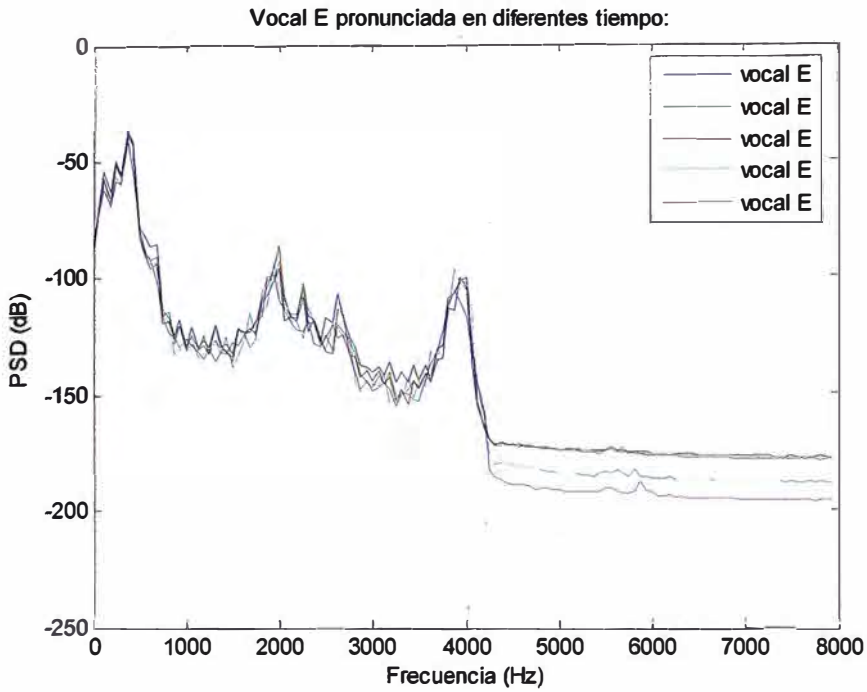


Figura 5.5: La figura muestra la PSD de la vocal E pronunciada en diferentes tiempos.

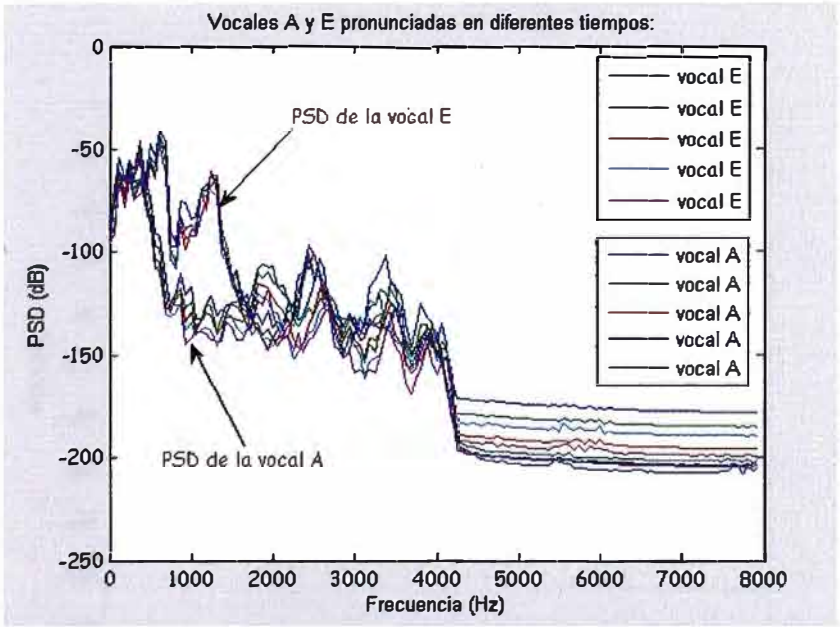


Figura 5.6: La figura muestra la PSD de las vocales A y E

En la Figura 5.6 se observa un grupo de 5 PSD para las vocales A y E y notamos que las que son de la misma vocal tienden al misma grafica y pero los de diferentes vocales difieren

a lo largo de la frecuencia, esto podemos usarlo como un reconocedor de vocales ya podemos comparar graficas de una señal de entrada con graficas patrones para de ahí deducir a que vocal le pertenece la grafica de la señal entrada.

Lo anterior es para un mismo usuario o persona, pero lo que se quiere es extender este análisis, es por ello ahora se analizara los espectro de 4 personas, estas son miembros de una misma familia, se aclara esto porque también se quiere aprovechar análisis en cuanto al parentesco con la voz ya que las voces entre hermanos o entre padre e hijo guardan alguna similitud.

La idea es entender por que son iguales las voces entre familiares, debe de existir alguna razón del por qué?. Esto se podría saber después de realizar un análisis en la frecuencia, los miembros de la familia son: Máximo (padre de 59 años), Máxima (madre de 51 años), Josué (hijo mayor de 23 años) y Marcos (hijo menor de 9 años). En las siguientes figuras se observaran las PSD de los miembros de la familia pronunciando la misma vocal.

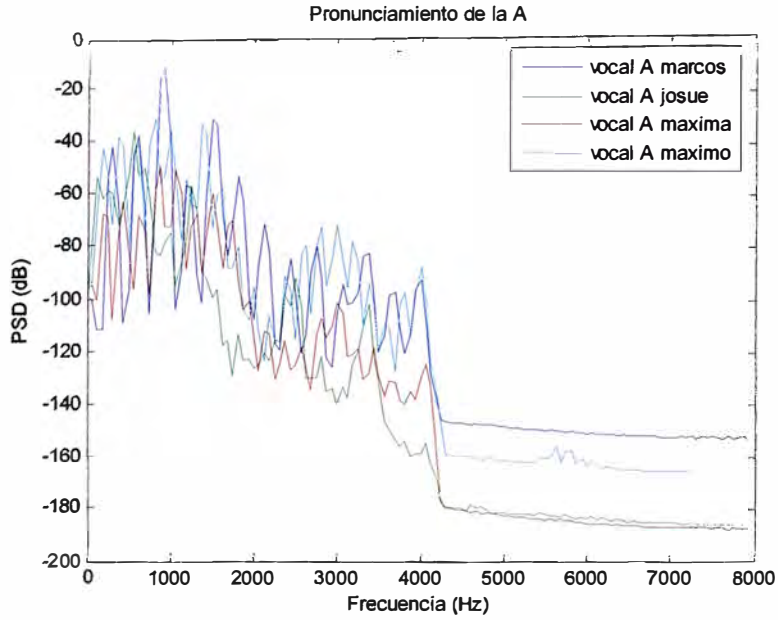


Figura 5.7: La figura muestra la PSD de la vocal A pronunciada por diferentes miembros de una familia

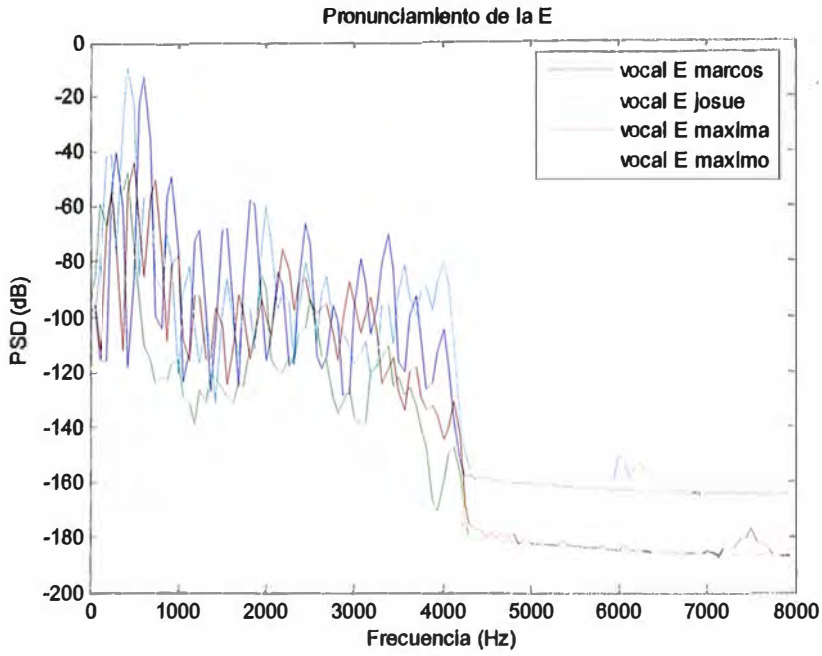


Figura 5.8: La figura muestra la PSD de la vocal E pronunciada por diferentes miembros de una familia

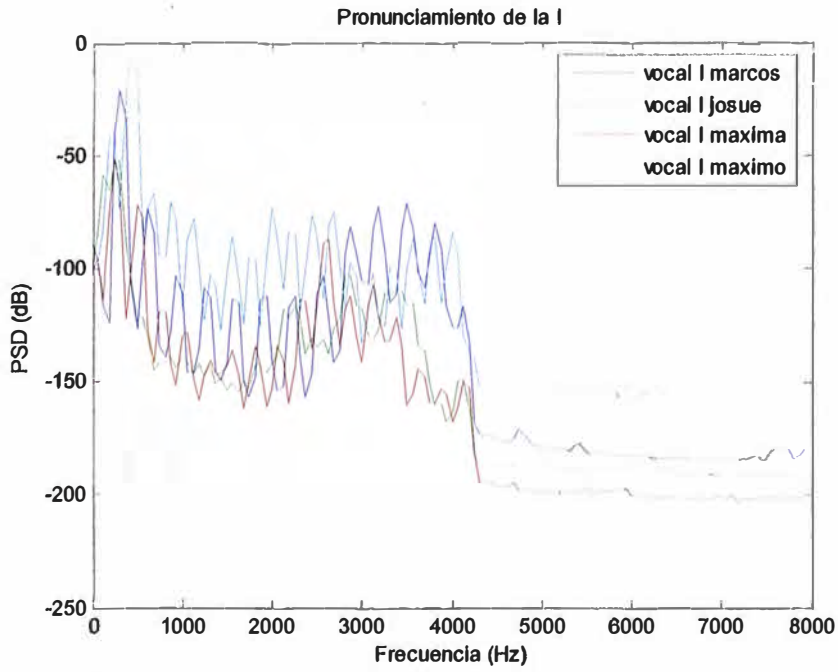


Figura 5.9: La figura muestra la PSD de la vocal I pronunciada por diferentes miembros de una familia

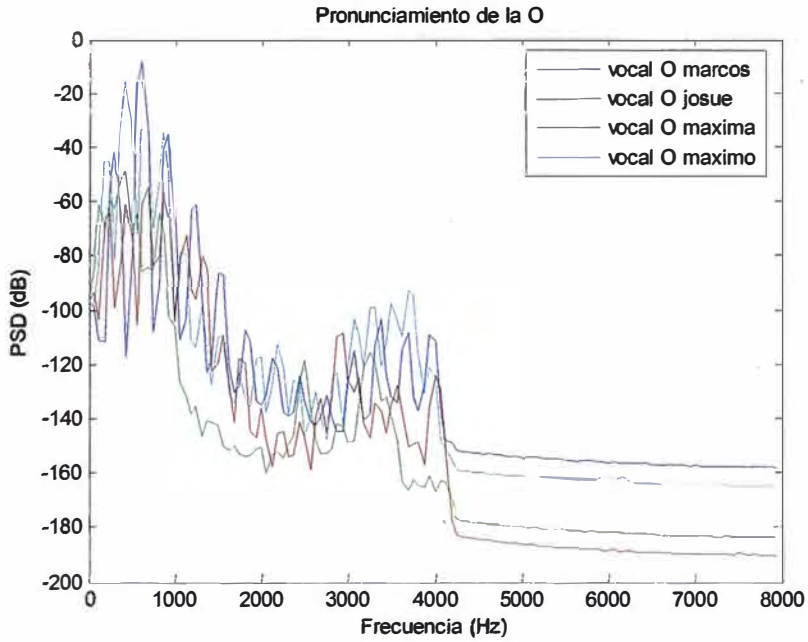


Figura 5.10: La figura muestra la PSD de la vocal O pronunciada por diferentes miembros de una familia

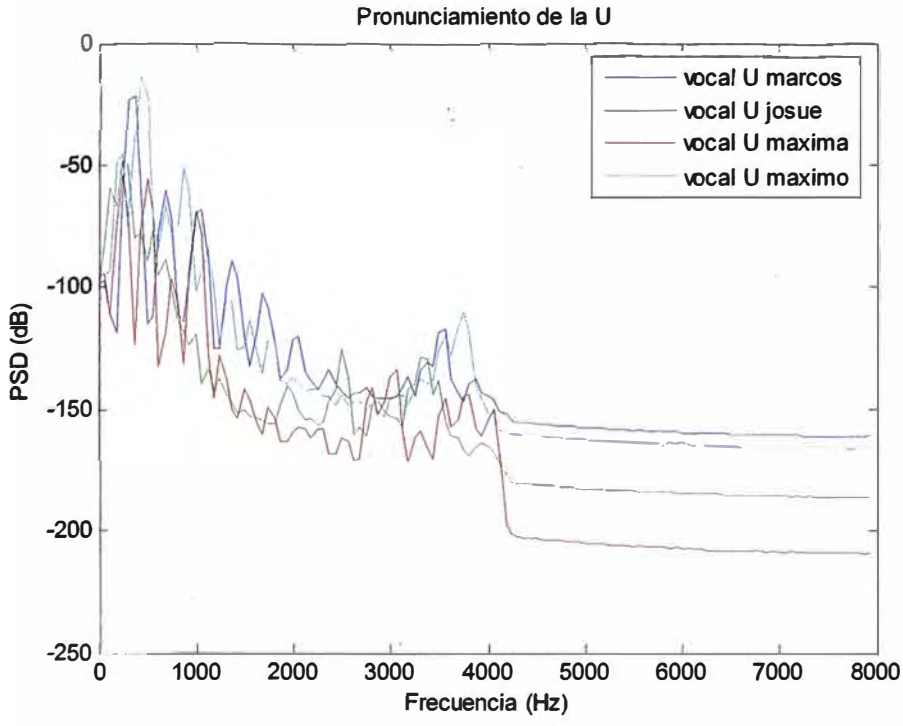


Figura 5.11: La figura muestra la PSD de la vocal U pronunciada por diferentes miembros de una familia

Observando las pruebas anteriores se puede ver la tendencia a una curva para cada cuadro, de lo que podemos concluir que cada vocal presenta formantes que la caracterizan o diferencian entre ellas, esta diferencia la podemos usar para hacer el reconocimiento de vocales pronunciada por cualquier usuario o persona, pero para este trabajo seremos mas estrictos, es decir, que no aceptaremos el parecido entre las PSD de la misma vocal pronunciada si no mas bien aprovecharemos las diferencia para hacer la identificación de la persona ya que el PSD de la vocal de una persona no varia en el tiempo, bajo ciertas condiciones, pero si se diferencia de las demás personas al pronunciar la misma vocal como se ha demostrado en las anteriores figuras.

Para el trabajo de identificación de personas por medio del reconocimiento de voz necesitaremos, por cada persona, un grupo de PSD de la misma vocal

El sistema diseñado hace uso de estas PSD que pasaran a llamarse patrones y se obtendrá un patrón promedio o patrón centroide, que le pertenecerá a una persona; entendiéndose por patrón promedio a los pesos resultantes de la red neuronal. Cada persona registrada en el sistema dejara su patrón promedio que la identificara y las diferenciara de las demás personas. Este conjunto de patrones promedio formaran el sistema de reconocimiento de voz.

En el momento de hacer la identificación de la persona el sistema le pedirá que diga la vocal con la cual se registro tomándolo como dato este, el sistema, empezara a comparar con los demás patrones promedio encontrando al de mayor semejanza y con esto se identificara a la persona dueña de la voz, este patrón promedio encontrado seria para nosotros el PSD promedio de la vocal pronunciada en el registro de la persona. En caso de no encontrar ninguna que se parezca se tratara de una persona que no ha sido registrada.

Para que el sistema de una mayor seguridad a la identificación se propone que se registren dos vocales diferentes, que pasara a ser patrón promedio de la vocal, para cada persona; en este caso se tendrán dos patrones promedio para cada persona y al momento

de hacer la comparación las condiciones serian de que los dos patrones promedio que tengan el mayor parecido al par de patrones registrados y que coincidan con los dos patrones de la persona corresponderá a la persona identificada.

Ahora necesitamos de una herramienta que nos permita hacer dos procesos; el primer proceso consiste en obtener el patrón promedio de los patrones de muestra que se obtienen al momento de hacer el registro de la persona y el segundo proceso sería la de realizar las comparaciones para hallar el patrón promedio mas parecido para cada vocal y asegurarse de que le pertenezcan a la misma persona. Para esta labor que mejor herramienta que la que ofrecen las redes neuronales con un tipo de red llamada la Red de Kohonen, esta red nos permite hacer las dos operaciones que mencionamos líneas arriba, la red de Kohonen nos permite, por medio de una clasificación modificada de patrones, obtener el patrón promedio, a esto se le llama el aprendizaje de la red neuronal. Se necesitara una neurona por cada vocal y cada persona necesitara 2 vocales, entonces por persona necesitaremos dos neuronas que contengan el patrón promedio que se obtendrá por medio del aprendizaje de la red, estos componentes del patrón promedio serán guardados en los pesos de la red neuronal, todo esto realiza el primer proceso. Para el segundo proceso, al momento de identificar a una persona la red neuronal empezara a hacer las comparaciones activándose las neuronas que tengan un vector de pesos con mayor semejanza a los patrones de entrada, en caso de que estas neuronas activadas no le pertenezcan a la misma persona se asumirá de que la persona no esta registrado y de que por ejemplo no se le permita el pase al siguiente nivel.

Con esto se puede decir que se está frente a un sistema identificador de personas por medio del reconocimiento de la voz usando redes neuronales que en los siguientes capítulos de detallara su diseño.

CAPITULO VI

ETAPA DE ADQUISICION Y EXTRACCION DE LAS CARACTERISTICAS DE LA SEÑAL DE VOZ

6.1 Tarjeta de sonido

Una tarjeta de sonido o placa de sonido es una tarjeta de expansión para computadoras que permite la entrada y salida de audio bajo el control de un programa informático.

6.1.1 Características generales

Una tarjeta de sonido típica incorpora un chip de sonido que por lo general tiene el convertidor Digital-Análogo el cual cumple con la importante función de "traducir" formas de ondas grabadas o generadas digitalmente en una señal analógica y viceversa. Esta señal es enviada a un conector (para audífonos) en donde se puede conectar cualquier otro dispositivo como un amplificador, un altavoz, etc.

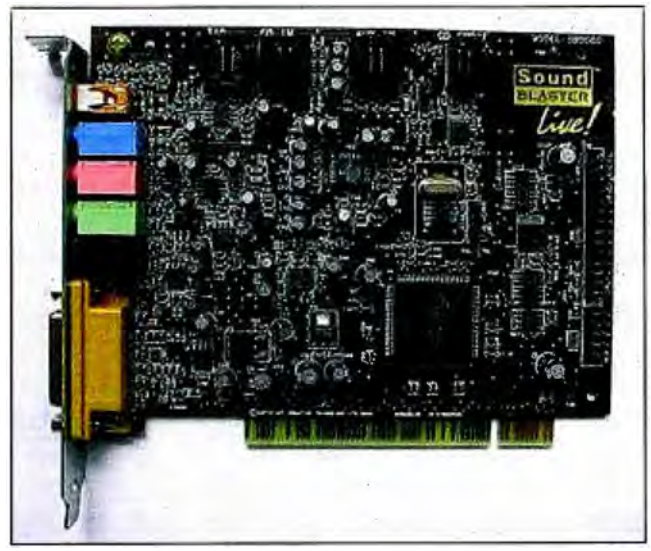


Figura 6.1: Tarjeta de sonido comercial con el estándar PC99.

Los diseños más avanzados tienen más de un chip de sonido, y tienen la capacidad de separar entre los sonidos sintetizados (usualmente para la generación de música y efectos especiales en tiempo real utilizando poca cantidad de información y tiempo del microprocesador y quizá compatibilidad MIDI) y los sonidos digitales para la reproducción.

Esto último se logra con DACs (por sus siglas en inglés Digital-Analog-Conversor o Convertidor-Digital-Analógico), que tienen la capacidad de reproducir múltiples muestras digitales a diferentes tonos e incluso aplicarles efectos en tiempo real como el filtrado o distorsión. Algunas veces, la reproducción digital de multi-canales puede ser usado para sintetizar música si es combinado con un banco de instrumentos que por lo general es una pequeña cantidad de memoria ROM o flash con datos sobre el sonido de distintos instrumentos musicales. Otra forma de sintetizar música en las PC's es por medio de los "códecs de audio" los cuales son aplicativos diseñados para esta función pero consumen mucho tiempo de microprocesador.

La mayoría de las tarjetas de sonido también tienen un conector de entrada o "Line In" por el cual puede entrar cualquier tipo de señal de audio proveniente de otro

dispositivo como micrófonos, casseteras entre otros y luego así la tarjeta de sonido puede digitalizar estas ondas y guardarlas en el disco duro del computador.

Otro conector externo que tiene una tarjeta de sonido típica es el conector para micrófono. Este conector está diseñado para recibir una señal proveniente de dispositivos con menor voltaje al utilizado en el conector de entrada "Line-In" [9].

6.1.2 Conexiones

Casi todas las tarjetas de sonido se han adaptado al estándar PC99 de Microsoft que consiste en asignarle un color a cada conector externo, de este modo:

Tabla 6.1: Entradas y salidas de la tarjeta de sonido comercial con los colores del estándar PC99.

Color	Función
Rosa	Entrada analógica para micrófono.
Azul	Entrada analógica "Line-In".
Verde	Salida analógica para la señal estéreo principal (altavoces frontales).
Negro	Salida analógica para altavoces traseros.
Naranja	Salida Digital SPDIF (en ocasiones como salida análoga para altavoces centrales).

6.2 Fundamentos del sonido digital

6.2.1 Naturaleza Del Sonido

Diferentes escritos definen al sonido de diferentes maneras como:

> *Desde el punto de visto físico se define al sonido como cualquier interrupción del estado de reposo del aire que provoca vibraciones en el tímpano, las*

> *cuales son interpretadas como sonido por el sistema nervioso central.*
>
> *Desde el punto de vista psicológico, estrictamente hablando, es la interpretación cortical de las vibraciones percibidas a través del sentido auditivo (Brosnahan & Malmberg)[10].*

El sonido es la materia por excelencia del lenguaje porque es la única materia que se transmite desde el sistema emisor al sistema receptor y, por tanto, es la materia misma de la comunicación entre los dos sistemas.

La audición en los seres humanos, ocurre siempre que una vibración tenga una frecuencia comprendida entre unos 15 a 20.000 hercios, hertzios o hertz, y su intensidad sea la suficiente para llegar al oído interno. Cuando las vibraciones pasan estos márgenes se habla de ultrasonidos y no son perceptibles al ser humano.

Las características del sonido se pueden medir y para ello se usa las unidades de hertz (Hz) que miden la frecuencia de un sonido o sea cuantas veces vibra en un segundo, y los decibeles (Db) que mide la intensidad (amplitud) de una onda.

El oído y un micrófono incorporado a la tarjeta de sonido tienen un funcionamiento similar. Ambos transforman las vibraciones del aire en una señal eléctrica que puede ser comprendida y almacenada por sus respectivos cerebros. Esta señal puede ser guardada, manipulada o reproducida por los medios electrónicos adecuados.

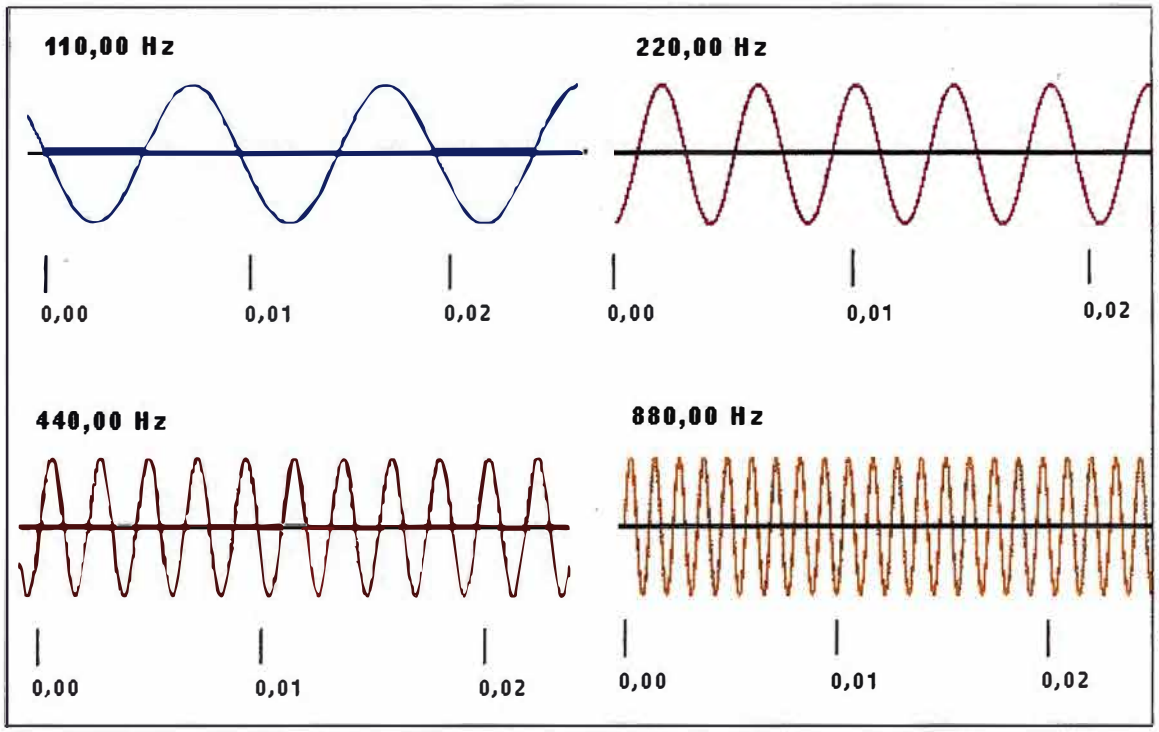


Figura 6.2: Muestra de cuatro sonidos con distintas frecuencias

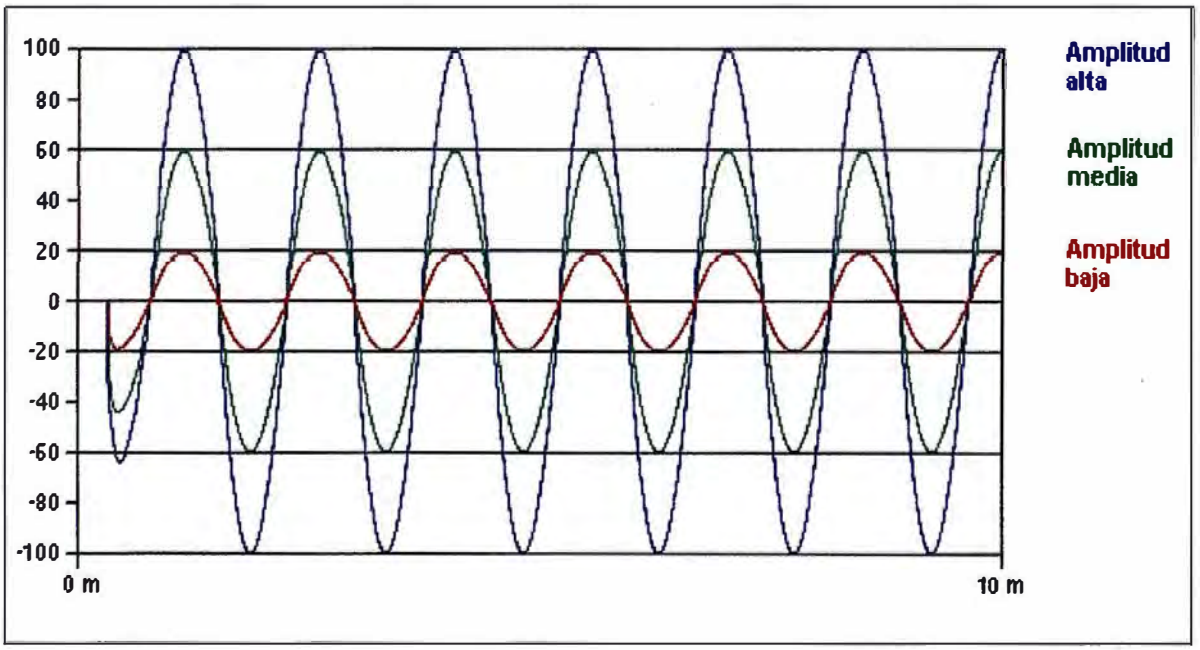


Figura 6.3: Muestra de un mismo sonido a tres distintas intensidades.

6.2.2 Computador Y Sonido

La palabra digital nos indica la presencia de procesos numéricos para concretar un hecho (imagen, sonido, etc.), los sistemas de audio digital tienen circuitos eléctricos para guardar el registro de la música, en pocas palabras lo que hacen estos circuitos es grabar una larga cadena de números (digitalización o muestreo) con un dispositivo llamado conversor análogo digital (ADC), que se encarga de monitorear la evolución de la onda y asignarle a cada momento un valor numérico, luego ese valor numérico es decodificado por un conversor llamado digital-análogo (DAC).

La calidad del sonido depende de la frecuencia del muestreo y a la resolución. Frecuencias de muestreo o de sample, se refiere al numero de mediciones que se hacen por segundo. Cuanto mayor sea el numero de muestras mejor es la calidad del sonido, por ejemplo si la velocidad de muestreo es de una cada un segundo las variaciones del sonido que se produzcan en el intermedio no serán registradas. Según estudios [4], la frecuencia de muestreo debe ser el doble del sonido mas alto que se pueda escuchar, como el oído humano puede escuchar aproximadamente hasta los 20.000 Hz, la frecuencia optima de muestreo será de 44,1 Khz. (44.100 hercios), esta es la frecuencia que se usa en los CD de música.

Como los instrumentos o las voces humanas no pasan la frecuencia de los 10 Khz., con una frecuencia de muestreo de 32 Khz. se obtiene muy buenos resultado. Al disminuir significativamente la frecuencia de muestreo el sonido se vuelve opaco o poco nítido pues se pierden las frecuencias agudas.

La resolución, el término hace referencia a la exactitud de las medidas de frecuencia. Se mide en bit, si la resolución es de 8 bit tenemos 256 niveles posibles (2^8=256). Si se amplía a 16 bit el rango se extiende a 65.535 (2^{16}=65536). Como

referencia se puede decir que un disco compacto se graba a 44,1 Khz. y a una resolución de 16 bits.

Si se desea digitalizar 3 minutos de música a un muestreo de 44,1 Khz. y almacenando por cada muestra dos bytes (16 bits) se obtiene lo siguiente:

$$3\min x60\frac{seg}{\min}x44100\frac{muestras}{seg}x2\frac{bytes}{muestra}=15876000bytes$$

El cálculo nos indica que para almacenar una canción de tres minutos con calidad profesional se necesitaran 16 MB aproximadamente. Es aquí donde surge el problema, pues aunque los discos duros u otros medios de almacenamiento han crecido mucho, tener varias canciones significaría ocupar gran parte del disco del computador, Para solucionar este problema se han desarrollado formatos de archivo que permiten realizar grabaciones de sonido con muy buena calidad usando un método de compresión, el problema es que el sonido no puede ser editado para ser modificado.

6.2.3 Formato WAV

El formato WAV, (Waveform Audio File) es un formato de archivo originario de Microsoft Windows 3.1. Es el formato para almacenar sonidos mas utilizado por los usuarios de Windows, la flexibilidad de este formato lo hace muy usado para el tratamiento del sonido pues puede ser compreso y grabado en distintas calidades y tamaños. Las frecuencias de muestreo van desde los 11025, 22050, 44100 Hz. Aunque los archivos ***wav*** pueden tener un excelente sonido comparable a la del CD (16 bites y 44,1 Khz. estéreo) el tamaño necesario para esa calidad es demasiado grande (especialmente para los usuarios de Internet) una canción convertida a ***wav*** puede ocupar fácilmente entre 20 y 30 Mb. La opción mas pequeña es grabar a 4 bits y los Khz lo mas bajo posible, el problema es la baja calidad del sonido, los

ruidos, la estática, incluso cortes en el sonido, por esta razón casi siempre se usa para muestras de sonido. La ventaja mas grande es la de su compatibilidad para convertirse en varios formatos por medio del software adecuado, un ejemplo de ello es pasar de ***wav*** a Mp3 [12].

6.3 Adquisición De La Señal De Voz

En esta primera parte nos encargamos de adquirir las señales de voz, primero conectaremos un micrófono comercial a la tarjeta de sonido de la computadora tal como muestra la Figura 6.4.

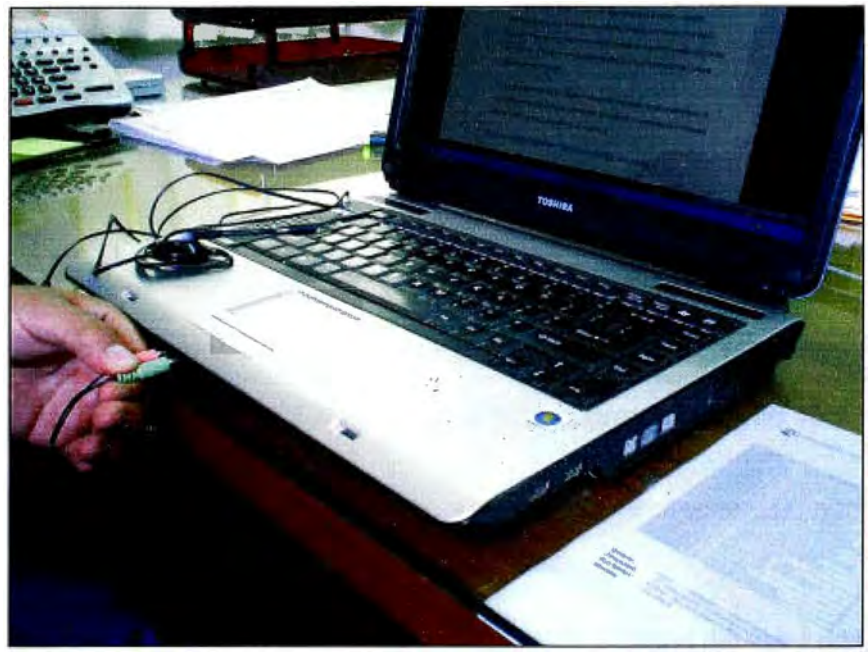


Figura 6.4: Conexión del micrófono a la tarjeta de sonido de la computadora

Se ejecuta el aplicativo ***registrar_voz.m***, el sistema pedirá que se pronuncie la palabra dos vocales separadamente, la primera vocal es la vocal "I" y luego la vocal "A", esto se hace para cada usuario, grabándolas en formato de audio que no realice compresión como es el formato "wav", para la configuración de la grabación tenemos que tener en cuenta lo siguiente, según la teoría, la voz cuenta con una gama de frecuencias importantes de hasta

de 8 Khz. Con el criterio del teorema del muestreo se toma una frecuencia de muestreo de $Fm = 16Khz.$ con esta frecuencia de muestreo se tomaran 7680 muestras de la señal de voz con 16 bits para cada muestra.

$x_i(n)$.....muestra de la señal de voz

El proceso de recolectar 7680 muestras se realizara 10 veces, por cada vocal y por cada persona, estos 10 datos vendría a ser la cantidad de patrones que se usara en la parte del aprendizaje de la red. Si lo guardamos en memoria en forma de matrices se vera que cada persona tendrá 2 matrices, una por cada vocal, y en cada matriz hay 10 columnas que estarán compuestas por las 7680 muestras.

6.4 Filtrado Digital De La Señal de Voz

Basándonos en aplicaciones de telefonía solo tomaremos un ancho de hasta 4 Khz. es por esto hacemos pasar la señal por un filtro elíptico con una frecuencia de corte $Fc = 4Khz.$ El filtro usado es un pasabanda del tipo elíptico de orden 8 con un rizado de pasabanda en decibelios (Rp) de 0.01 y una rizado de supresión (Rs) de 40 y la frecuencia de corte normalizada de $Fn = \frac{Fc}{\frac{Fm}{2}}$, los valores de la frecuencia normalizada se encuentran entre los valores de 0 y 1 donde 1 viene a ser la mitad del valor de la frecuencia de muestreo.

Con todos estos valores obtenemos el filtro $G(z)$ [11], consiguiendo con este filtro estrechar la zona de transición entre bandas, así como atenuar los efectos del ruido de altas frecuencias.

$$G(z) = \frac{0.0802z^8 + 0.255z^7 + 0.567z^6 + 0.85z^5 + 0.982z^4 + 0.851z^3 + 0.567z^2 + 0.255z + 0.0802}{z^8 - 0.183z^7 + 2.27z^6 - 0.449z^5 + 1.74z^4 - 0.332z^3 + 0.485z^2 - 0.072z + 0.0296}$$

Aun así con este recorte se puede distinguir la voz de las personas por que en esta zona están concentradas el tono y en el timbre de la voz.

6.5 Extracción De Las Características Frecuenciales De La Señal de Voz

Se sabe que las vocales están clasificadas como señales aleatorias y el análisis en frecuencia es distinto al que se le hace a las señales deterministas es por esto que se le aplica la teoría de periodograma, es decir, se le aplica la trasformada de fourier discreta a la función de autocorrelación produciendo la densidad espectral de potencia, que es un equivalente a el análisis en la frecuencia de una señal determinista. Como contamos con una gran cantidad de datos podemos usar el método de Welch que hace como una especie de promedio de periodogramas, es decir, a las 7680 muestras las dividimos en 30 segmentos, esto hace que cada segmento contenga 256 muestras, de los cuales a cada segmento se le calcula su periodograma enventanada con la ventana de Hamming para luego aplicarle el método de Welch haciendo un promedio de ellas, con esto obtenemos su densidad espectral de la señal (PSD) que cuenta con 256 puntos o muestras. El PSD es simétrico y periódico es por esto solo se toma la mitad de puntos del primer periodo.

La expresión que usamos es la siguiente:

$$\hat{S}_W(e^{jw}) = \frac{1}{KLU}\sum_{i=0}^{K-1}\left|\sum_{n=0}^{L-1} w(n)x(n+i\cdot D)\cdot e^{-jw}\right|^2$$

Donde *U*:

$$U = \frac{1}{L}\sum_{n=0}^{L-1}\left|w(n)\right|^2$$

L: Numero de puntos del segmento

K: La cantidad de segmentos

Todo este procedimiento lo realiza la función ***y_Y.m*** donde se encuentra como subfunción la función de ***welch.m***. En la Figura 6.5 muestra gráficamente el proceso detallado anteriormente.

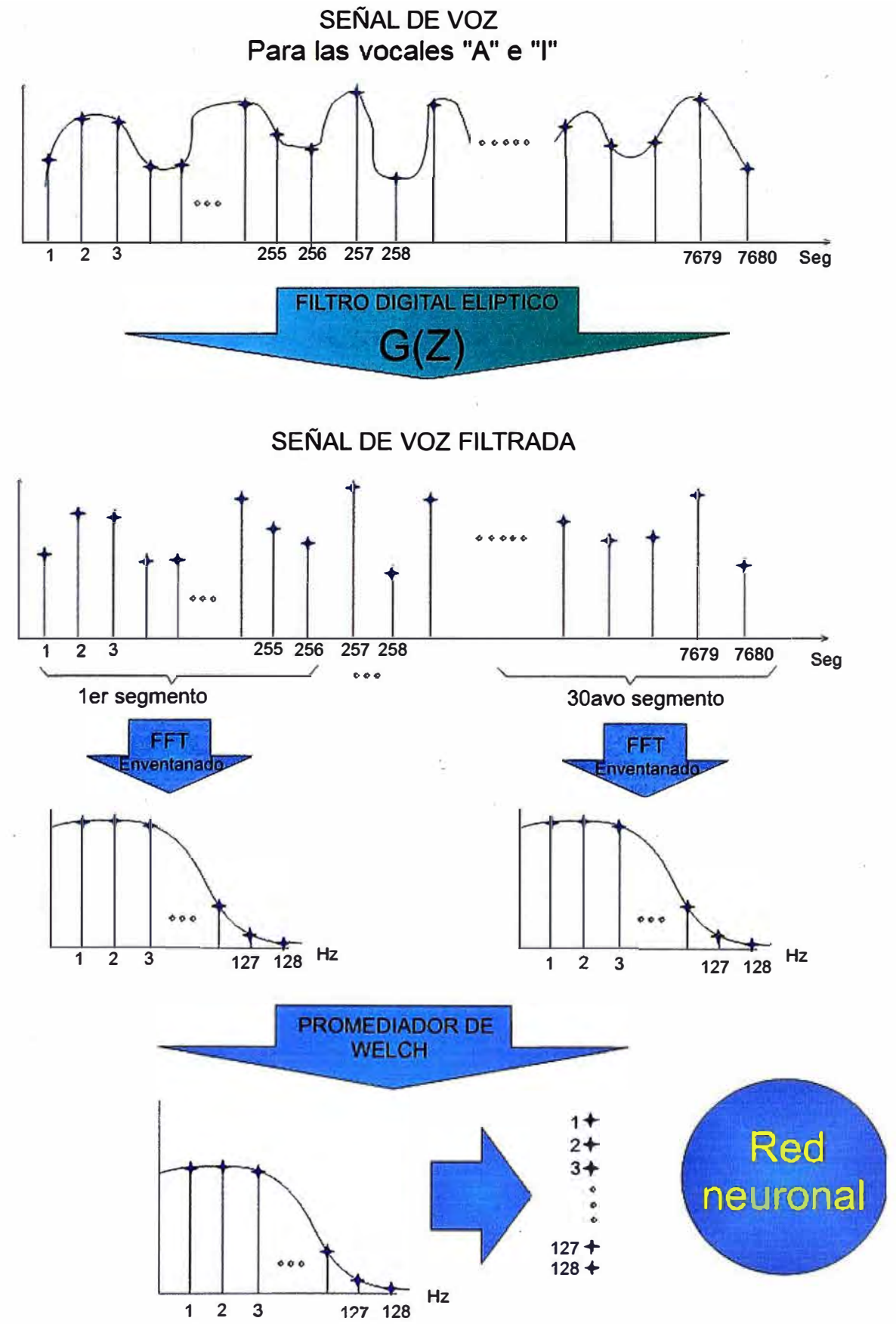


Figura 6.5: La figura muestra las etapas de procesamiento de voz para obtener la densidad espectral que pasan a ser los patrones de entrenamiento de la red neuronal

Todo este procedimiento descrito anteriormente se hace para una sola silaba del cual solo nos interesa el tiempo de pronunciamiento de la vocal pronunciada, entonces cada vocal pronunciada va tener su espectro de frecuencias muestreadas en 256 puntos de los cuales solo se tomaran 128 muestras.

En conclusión de cada vocal se obtendrá su espectro que consta de 128 puntos o muestras, estos 2 espectros, de las vocales "I" y "A", son los que pertenecen al usuario, no puede haber dos usuarios con los mismos espectros de vocal iguales, ahora este grupo de muestras pertenecientes al espectro lo llamaremos patrones de entrenamiento.

Estos patrones serán guardados en memoria en forma de matriz esto quiere decir que cada persona tendrá 2 matrices, una por cada vocal, y cada matriz estará compuesta por 10 columnas de 128 muestras cada matriz, estas muestras son el muestreo de la PSD de la señal.

CAPITULO VII

ETAPA DE ENTRENAMIENTO DE LA RED DE KOHONEN

7.1 Entrenamiento De La Red Neuronal

En la data tomada del PSD de cada señal se presentan ligeras variaciones en cada patrón perteneciente a la misma vocal pronunciada del mismo usuario y necesitamos de una operación que nos consiga una especie de patron promedio de este grupo

Esto se hace tomando los 10 patrones, de la misma vocal pronunciada y del mismo usuario, del cual obtenemos un patrón promedio que se parece a todo ese grupo de patrones, ahora para hacer esto se trabajara con la red neuronal competitiva de Kohonen. La cualidad de esta red es que permite realizar la operación que se menciono haciendo el entrenamiento de la red. De este entrenamiento se obtendrá el patrón resultante al que se le llamara peso de la neurona. Como se estableció que para realizar una mejor identificación es necesario usar dos vocales esto quiere decir que se necesitara de 2 neuronas por cada usuario, para el entrenamiento usaremos los 10 patrones por neurona, el entrenamiento consiste en hacer parecer al peso de la neurona a todos los patrones pertenecientes a una misma vocal del mismo usuario, esto lo conseguimos a través del siguiente ***algoritmo de aprendizaje***.

$$W_{k+1} = W_k + \mu_k . (X_k - W_k)$$

La forma del entrenamiento se explica en la parte teórica, pero aquí podemos decir que se empezaría con un peso inicial que vendría a ser el primer patrón obtenido, luego este empezaría a compararse con cada uno de los patrones y el criterio de comparación seria la distancia euclídea.

$$\min \left\| X_p - W_j \right\| = \min \sum_{j=1}^{N} (x_{pi} - W_{ji})^2$$

La neurona que presenta menor distancia es la que se activa o la ganadora y el algoritmo de aprendizaje se encarga que el vector de pesos apunte a la dirección de dicho patrón con mas parecido, es decir, tiende a parecerse en modulo y dirección, para que no exista problema con el criterio de la distancia euclídea es por eso que se normalizan los patrones es decir que su modulo sea siempre la unidad.

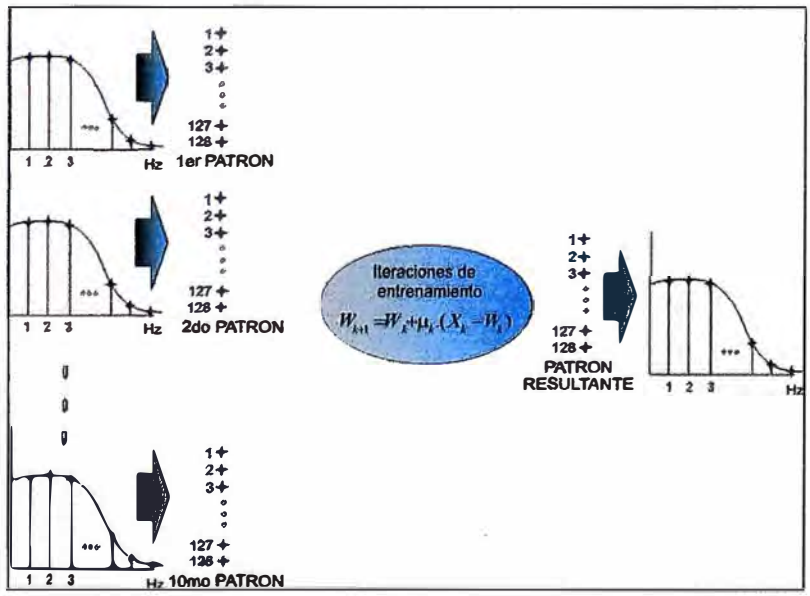


Figura 7.1: La figura muestra la etapa de entrenamiento de la red neuronal usando los patrones adquiridos anteriormente

Con esto obtendremos 2 espectros que le pertenecen a la persona, al decir la vocal '***A'*** y la vocal '***O'***, estos espectros están compuestos por 128 muestras que quedaran guardadas en los pesos de su respectiva neurona situadas en la capa de entrada, entonces la red neuronal estará compuesta por la capa de entrada y la capa de salida estará compuesta por las dos neuronas por persona. Toda esta operación la realiza la función ***entrenamiento.m***

El modelo de la red neuronal se muestra en la Figura 7.2.

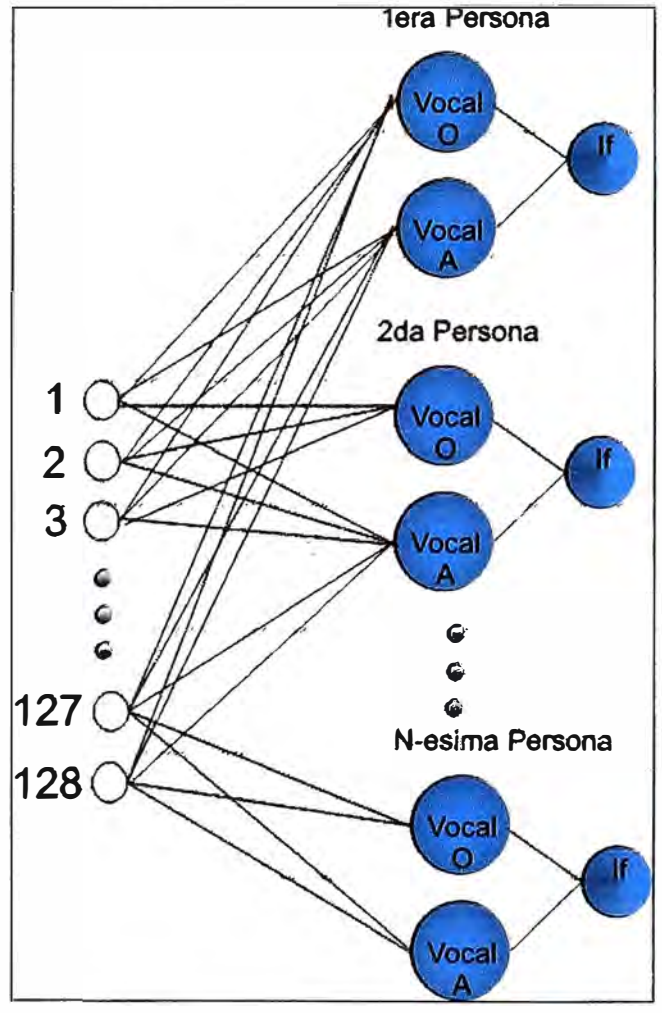


Figura 7.2: La figura muestra el modelo de la red neuronal a usar en este trabajo

Esta seria el modelo de la red neuronal entrenada que estaría lista para la identificación de las personas esta red competitiva consta de dos neuronas por persona y se conectan para detectar que las dos neuronas se han activado y así identificar a las personas.

UNIDAD III

CAPITULO VIII

IDENTIFICACION DE PERSONAS MEDIANTE EL RECONOCIMIENTO DE VOZ USANDO REDES NEURONALES

El identificador de personas consta de dos procesos, el primer proceso de capacitación que se encarga del aprendizaje de las voces de las personas a identificar y el segundo proceso de funcionamiento que es la identificación de la personas por el reconocimiento de voz. En la Figura 8.1 muestra el diagrama de bloques general.

Figura 8.1: Diagrama de General de la Identificación de personas.

Las pruebas de identificación de personas para la presente propuesta se realizo en un prototipo desarrollado en MATLAB 6.5.

8.1 Proceso de Capacitación.

Este primer proceso está relacionada con la función ***registrar_voz.m*** (VER APENDICE E.1), en esta función se adquiere la señal de voz, luego se realiza un filtrado digital para extraer las características frecuenciales, todo se realiza con el fin de obtener los patrones de entrenamiento y que por medio de este entrenamiento se obtendrá un vector de pesos resultante pertenecientes a una neurona. Para una identificación más segura se registrara las características frecuenciales en dos neuronas para cada persona.

8.1.1 Diagrama de Blogues

En la Figura 8.2 se muestra el diagrama de bloques del primer proceso que se encarga de establecer el identificador definido por la red neuronal, esta red neuronal es compuesta por vectores de pesos resultante provenientes del entrenamiento con los patrones obtenidos en la parte de extracción de características frecuenciales, la extracción de estas características frecuenciales se realiza sobre las voces ingresadas de cada persona que han sido adquiridas y luego tratadas.

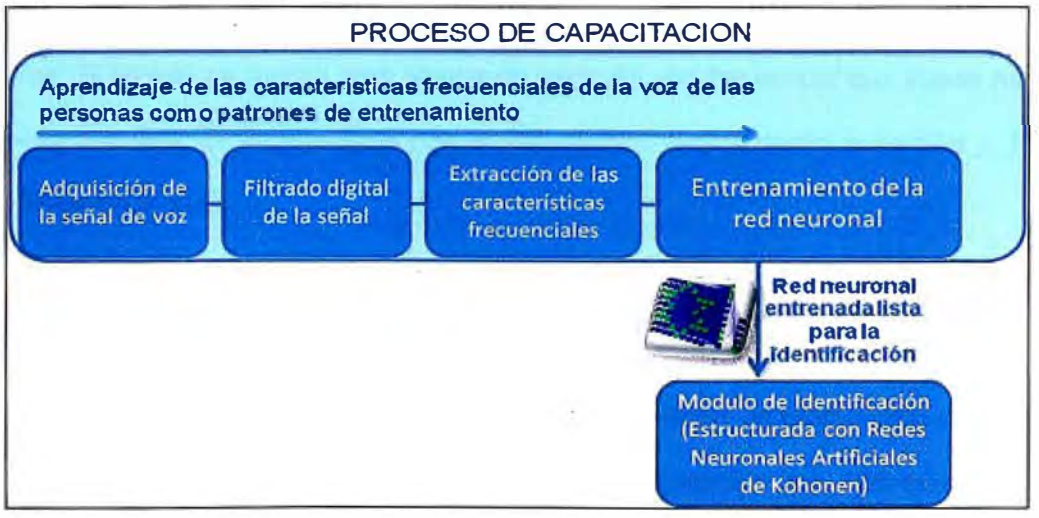


Figura 8.2: Diagrama de bloques de la primera etapa del Identificador de personas.

8.1.2 Adquisición de la Señal de Voz

En esta fase se adquiere la señal de voz, conectando un micrófono a la tarjeta de sonido, se corre la función ***registrar_voz.m*** (VER APENDICE E.1) esta función solicitará la cantidad de usuarios a registrar luego de esto pedirá que cada persona a registrar pronuncie 2 vocales, el sistema pide las vocales I y A. No necesariamente tienen que ser dichas vocales, el sistema es capaz de recibir 2 vocales cualesquiera de las 5 que existen siempre y cuando al momento de hacer la identificación se solicitaran las mismas vocales pronunciadas y en el mismo orden. Cada vocal pronunciada será muestreada hasta obtener 7680 puntos o muestras con una frecuencia de muestreo de 16Khz, esta operación se realiza para obtener un patrón de entrenamiento en la etapa de extracción de características frecuenciales para cada persona. En la etapa de entrenamiento se usaran 10 patrones por lo que la acción ejecutada anteriormente se realizara 10 veces por cada persona. De esta tarea se encarga la función ***escuchando.m*** (VER APENDICE E.3). Dentro de ***escuchando.m*** se usa el comando wavrecord que captura la voz y la guarda en formato WAV (VER 6.2.3 Y 6.3).

8.1.3 Filtrado Digital de la Señal

El filtrado digital se realiza sobre las señales de las vocales adquiridas con un filtro pasa bajos de tipo elíptico (VER 6.4). Este filtrado se realiza con cada una de las 7680 muestras. El filtrado se realiza para limpiar de ruido de alta frecuencia que pueda haber ingresado a la data. El filtrado digital con un filtro elíptico realiza dentro la función ***y_Y. m*** (VER APENDICE E.4).

8.1.4 Extracción de las Características Frecuenciales

En esta fase con la data filtrada se procede para realizar un promediado de Welch. Para esto de cada 7680 puntos se dividen en 30 segmentos con 256 puntos cada segmento, a cada segmento se le realiza su transformada de Fourier tomando solo la mitad de puntos usados esto debido a la simetría. A esta transformada de Fourier la

llamaremos espectro de frecuencias de la señal, entonces a esto 30 espectro obtenidos se procede a realizar el algoritmo de Welch (VER APENDICE E.5) para obtener un promedio de estos 30 espectros, a este promedio se le llamará PSD (Power Spectral Density – Densidad Espectral de Potencia). La razón de usar Welch es debido a que las vocales son señales consideradas no deterministicas o aleatorias, el espectro de frecuencia de esta señal no es constante, el tiempo siempre presenta diferencias, es por ello se requiere del algoritmo de Welch para obtener un espectro representante de los 30 tomados. Esta tarea es realizada por la función ***welch.m***.

8.1.5 Entrenamiento de la Red Neuronal Artificial

Con lo anterior se obtuvo los 10 patrones de entrenamiento para cada vocal, como anteriormente se dijo sobre la seguridad en la identificación se usaran 2 neuronas para cada persona. Como cada persona dice 2 vocales se trabajaran con las primeras vocales pronunciadas por cada persona, entonces cada persona tiene 10 patrones para entrenar su respectiva neurona asignada a su primera vocal. La red neuronal artificial usada es la de Kohonen (VER 4.5.1). Se usa el algoritmo de aprendizaje competitivo (VER 4.5.2 Y APENDICE E.6). En primer lugar se usa el primer patrón como el vector de pesos inicial para cada neurona. Todos los pesos son agrupados en un solo bloque luego se coge el primero y se compara con cada vector de pesos de su respectiva neurona midiendo la similitud con la distancia de Euclides entonces aquella neurona con mayor parecido al patrón ingresado será la ganadora y será la que realice el entrenamiento. De lo anterior podemos decir que como los patrones de una misma persona tiene semejanza entonces al ingresar estos patrones y se las compara con la neurona a la que fue inicializada con un vector de pesos perteneciente al mismo grupo dicha neurona siempre resultara ganadora, por lo que se realizara el entrenamiento con todos los patrones de ese grupo produciendo así mediante el entrenamiento un vector de pesos resultante en cada neurona, el conjunto de estas neuronas definirán a la red que se usara como identificador de personas. Se realiza el mismo procedimiento para definir la neurona de

la otra vocal correspondiente a la misma persona. Los pesos que definen a la red resultado del entrenamiento son guardados en archivo de datos ***.mat*** con el comando ***save*** de MATLAB. Esta tarea es realizada por la función ***entrenamiento.m***.

8.2 Etapa de la Identificación de Personas

En esta etapa se realiza lo mismo que se hizo en la primera etapa de adquirir la señal luego tratarla y de ahí obtener sus características frecuenciales pero en este caso ya no serán 10 si no solo una con la que se hará la identificación haciendo una comparación con cada vector de pesos de las neuronas. Así el que tenga mayor similitud será la persona identificada y para asegurar fiabilidad de la identificación se recurre a la segunda neurona ya que tiene que coincidir con la misma persona identificada, siendo este el caso se habrá identificado a la persona exitosamente.

8.2.1 Diagrama de Bloques

En la Figura 8.3 se muestra el diagrama de bloques de la segunda etapa ya con la rede definida como el identificador de personas.

Figura 8.3: Diagrama de bloques de la primera etapa del Identificador de personas.

8.2.2 Adquisición de la Señal de Voz, Filtrado Digital y extracción de las Características Frecuenciales

En esta parte se adquiere la señal de las 2 vocales pronunciadas de la persona a identificar en formato wav (VER 6.2.3 Y 6.3), se cogen los 7680 puntos de cada vocal para limpiarla de ruido con el filtro elíptico. Luego se divide en 30 segmentos para realizar la transformada de Fourier sobre cada segmento y con el algoritmo de Welch obtener el periodograma de estos 30 segmentos. Con esto se obtiene la característica frecuenciales, es decir, la densidad espectral de potencia de la señal de voz. Como son 2 vocales serian 2 espectros de frecuencia que pasaría tomarse como vectores de entrada a la red neuronal para ser evaluada (VER APENDICE E.2).

8.2.3 Modulo Identificador de Personas

Se carga al modulo identificador con la estructura de la red neuronal artificial definida por los vectores de pesos resultantes obtenidos y guardados en el proceso de capacitación. Luego los 2 vectores que vienen a ser las características frecuenciales de la voz de la persona son ingresados al modulo. Luego en el modulo se realiza la comparación con cada par de vectores pertenecientes al par de neuronas definidos para cada persona, cuando coinciden en similitud con las 2 neuronas pertenecientes a una misma persona, en ese caso la persona ha sido identificada. Para verificar la similitud entre los vectores de entrada y los vectores de pesos usamos la distancia de Euclides. Todos estos pasos están comprendidos en la función ***acceso.m*** (VER APENDICE E.2).

CAPITULO IX

PRUEBAS Y RESULTADOS

En este capítulo se procederá a realizar las pruebas del sistema empezando por adquirir la base de datos que contendrá características en frecuencia presentes en el par de vocales pronunciadas por cada persona, como se vio en el Capitulo VI se realizaran los paso de adquisición de señal para luego realizar el tratamiento digital de la señal de voz para proceder a la extracción de sus características frecuenciales finalizando con la comprobación del sistema en la parte de validación de la red neuronal.

9.1 Adquisición de la Señal.

En esta etapa procedemos a correr el programa ***registrar_voz.m***, el sistema pedirá la cantidad de personas a registrar como se muestra en la figura 9.1.

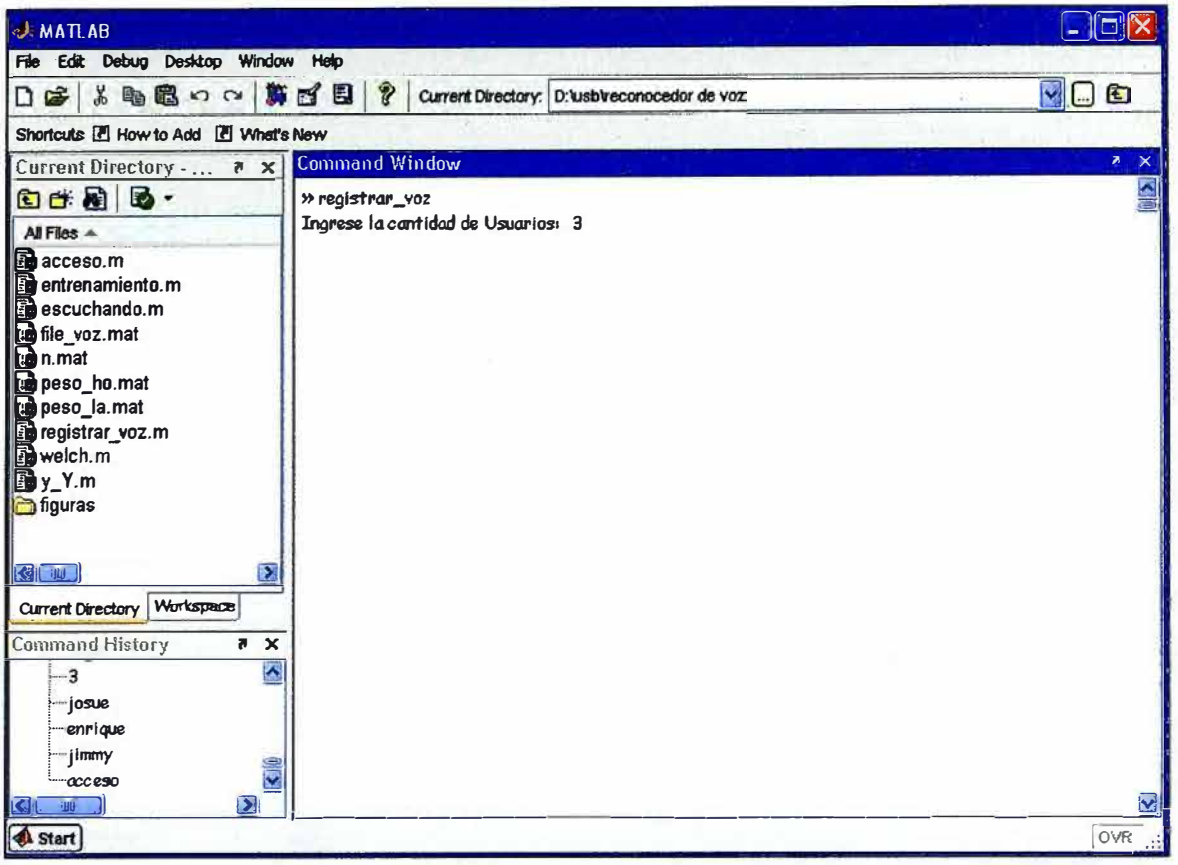


Figura 9.1: El sistema pide la cantidad personas a registrar

Se hará la prueba con 3 personas de nombres Josue, Enrique y Jimmy, el sistema pedirá que digan un par de vocales, los mismos para cada persona y en el mismo orden, en este caso se pide que mencionen las vocales "I" y "A". Esta tarea la realiza la función ***escuchando.m***.

Primero se ingresara la voz de Josue el sistema le pedirá que pronuncie y mantenga la vocal "I" por menos de medio minuto, luego le pedirá que pronuncie y mantenga la vocal "A" por el mismo tiempo. Después de ingresar las dos vocales el sistema pide el nombre de la persona que se está registrando para en un futuro identificar a la persona por medio de su voz.

Se realiza el mismo procedimiento de registro de voz para Enrique y Jimmy para completar el registro de las 3 personas como se observa en la figura 9.2.

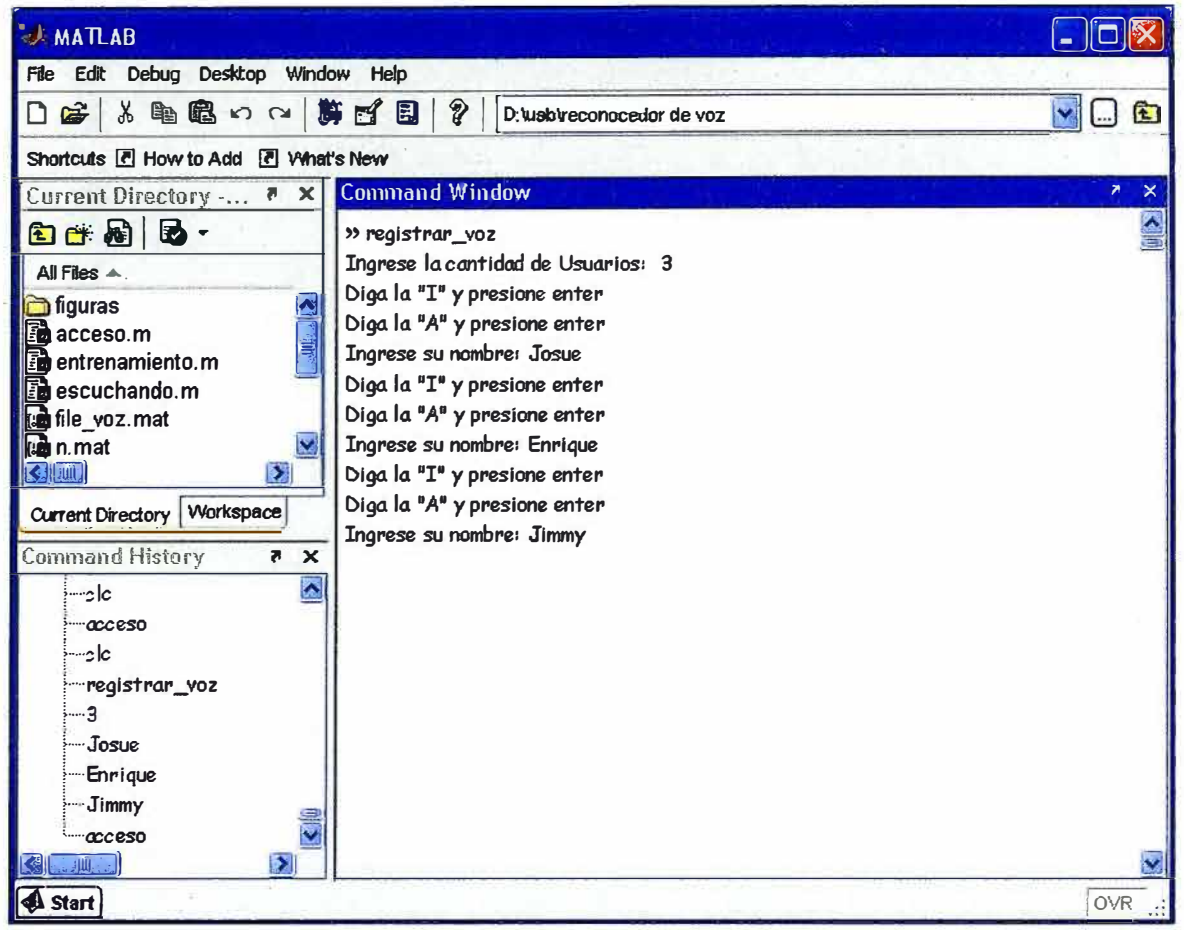


Figura 9.2: El sistema pide la cantidad personas a registrar

9.2 Filtrado Digital de la Señal de voz y Extracción De La Característica Frecuencial De La Entrada.

Cuando se termina de registrar la cantidad de personas indicadas al sistema, en este caso es 3, el sistema inmediatamente procede a realizar el tratamiento digital de la señal de voz registrada en vocales. En esta parte se procede a realizar el análisis en la frecuencia de la señal de voz usando la función ***y_Y.m*** haciendo un filtrado y usando la función ***welch.m*** para llevar la señal al análisis en la frecuencia obteniendo las Densidades Espectrales de Potencia (Power Spectral Density PSD) de cada señal. Se recomienda el uso del método de Welch para analizar en la frecuencia a las señales aleatorias como es el caso de la voz. En las siguientes figuras se muestra el análisis en la frecuencia (PSD) de las 10 muestras pertenecientes a cada vocal.

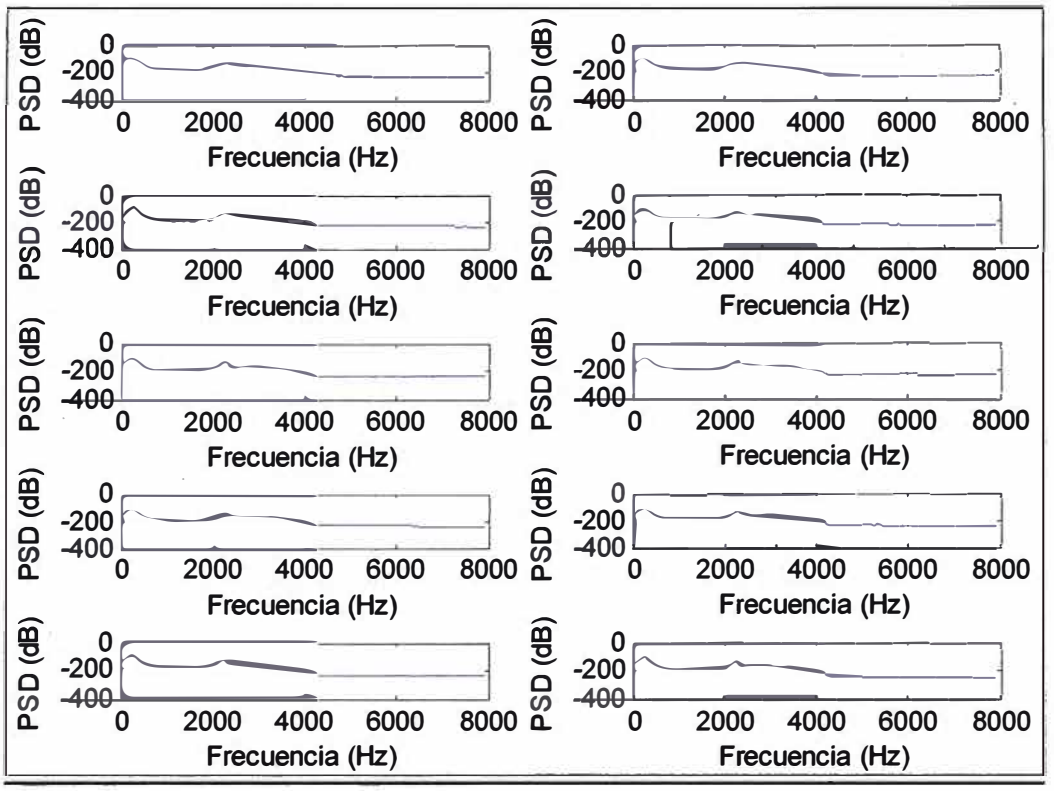


Figura 9.3: PSD de las 10 muestras de la vocal "I" pertenecientes a Josue

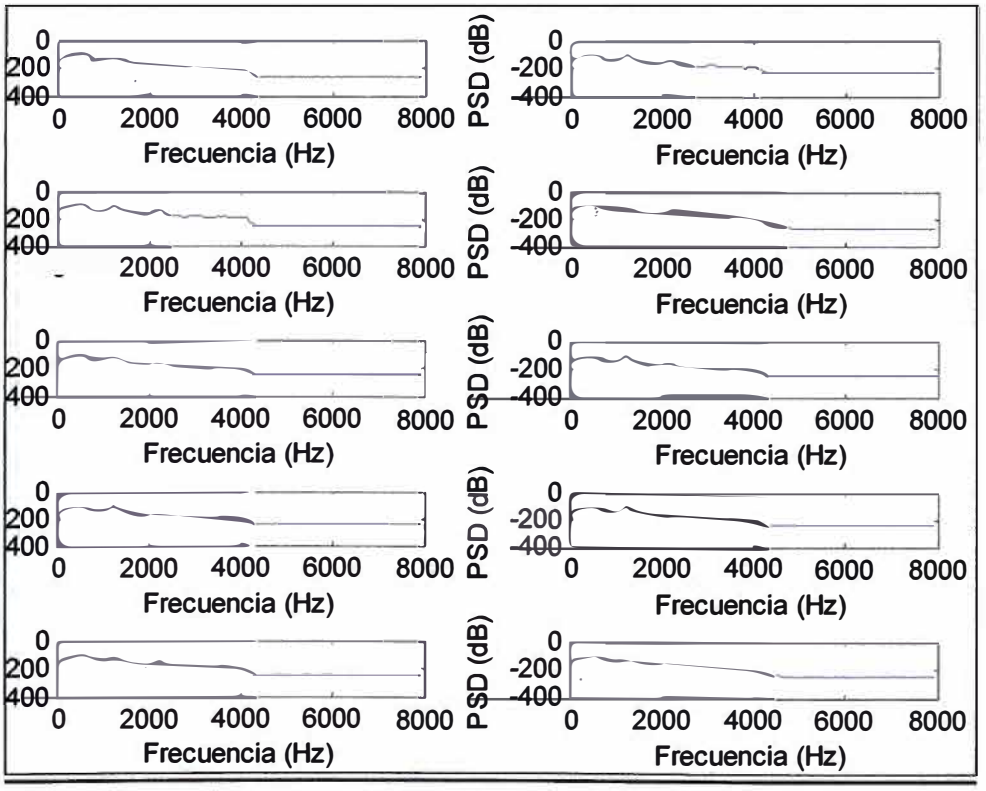


Figura 9.4: PSD de las 10 muestras de la vocal "A" pertenecientes a Josue

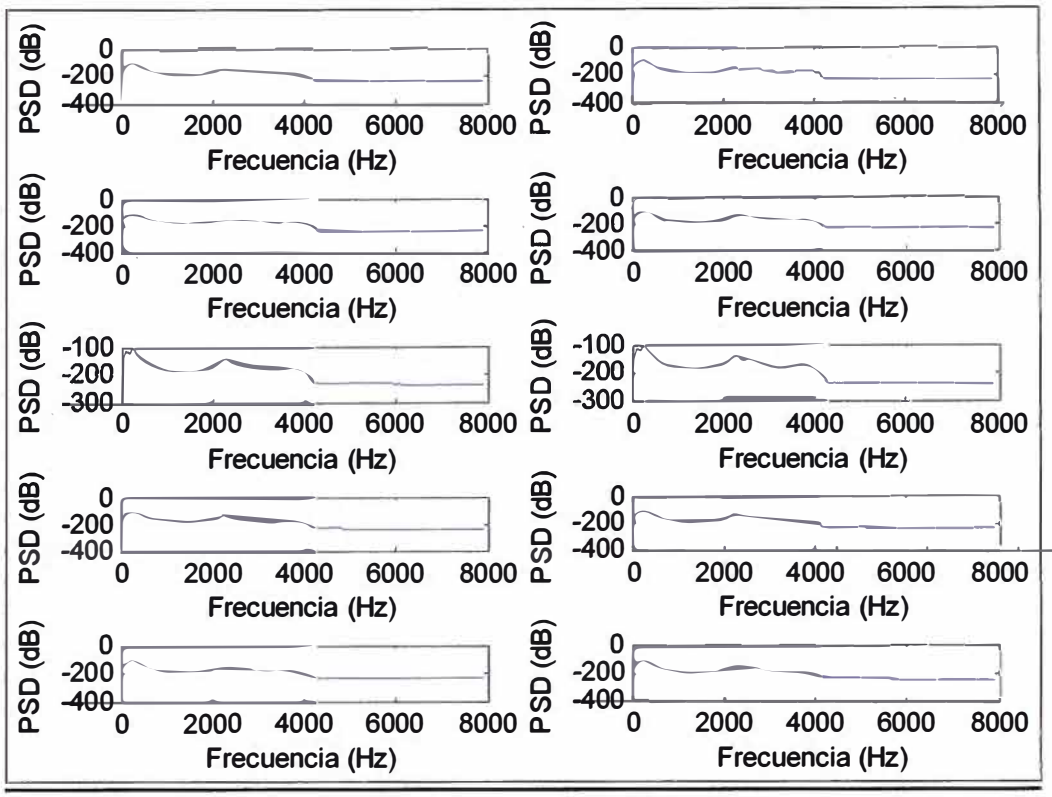


Figura 9.5: PSD de las 10 muestras de la vocal "I" pertenecientes a Enrique

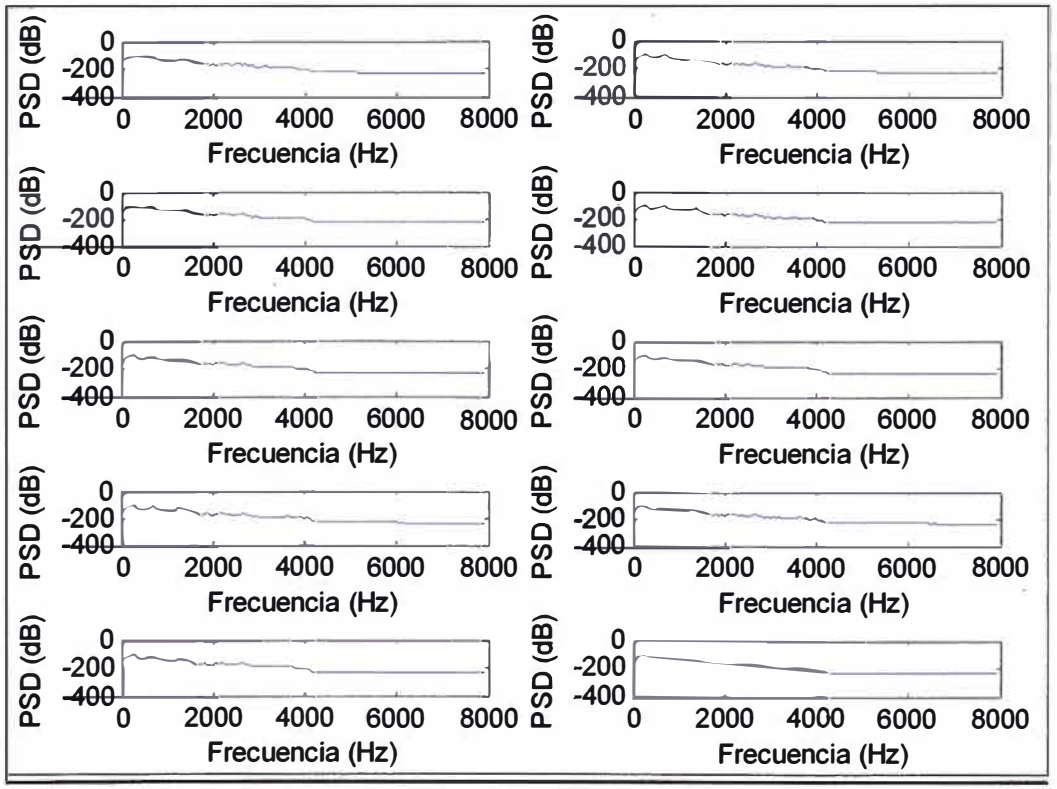


Figura 9.6: PSD de las 10 muestras de la vocal "A" pertenecientes a Enrique

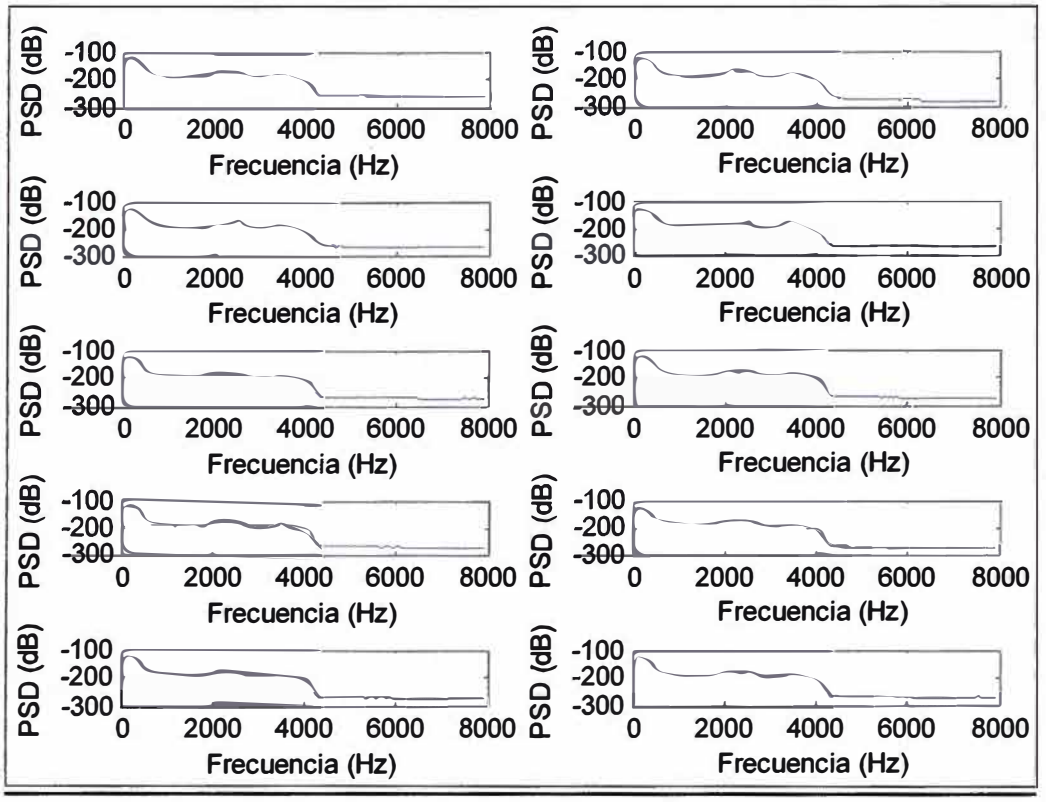


Figura 9.7: PSD de las 10 muestras de la vocal "I" pertenecientes a Jimmy

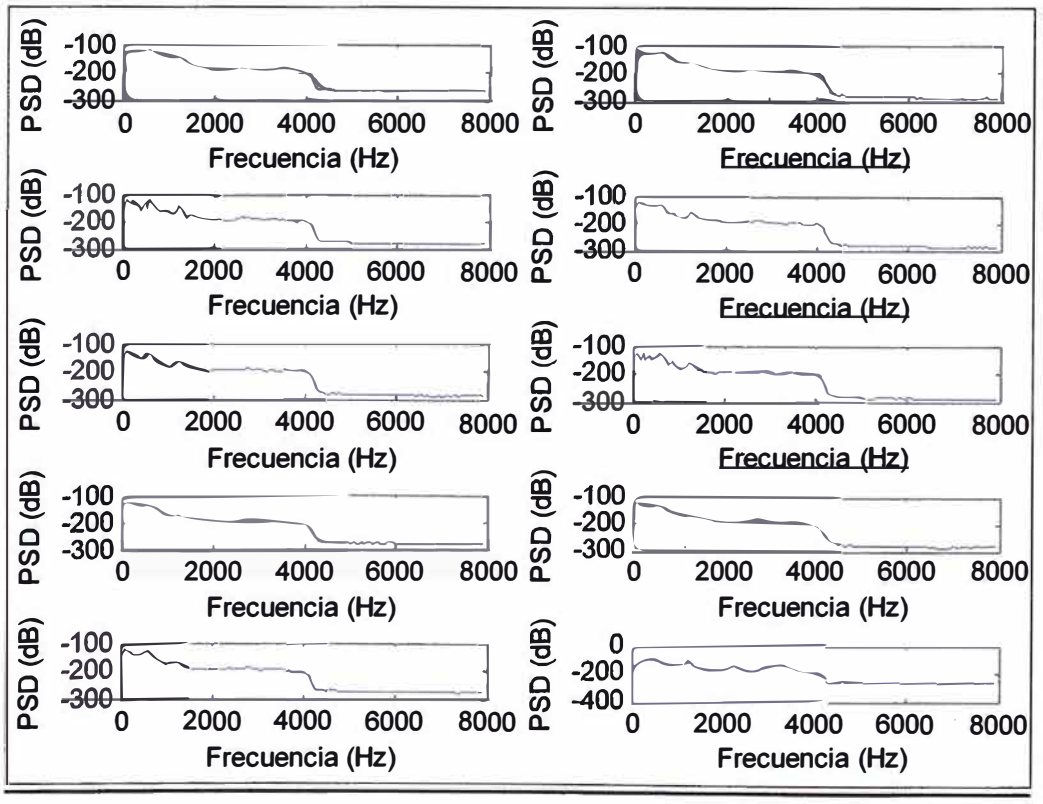


Figura 9.8: PSD de las 10 muestras de la vocal "A" pertenecientes a Jimmy

9.3 Entrenamiento de la Red Neuronal

En esta etapa se procede al entrenamiento de la red neuronal, a cada persona se le asigna dos neurona una para cada vocal pronunciada, cada neurona tiene 128 pesos, cada uno de estos pesos contiene un valor del muestreo del PSD de cada vocal. En la etapa de adquisición de datos el sistema adquirió 10 pares de PSD para cada persona esto hace un total de 30 pares PSD que pasarían a llamarse parámetros de entrenamiento de la red. El sistema hace un recorrido de estos 30 pares de parámetros y los hace ingresar a los tres pares de neuronas activando el par correspondiente a la persona y asi realizar el entrenamiento. La cantidad de épocas que realiza el sistema para entrenar al sistema es de 100 epocas. Figura 9.10 muestra que par de neuronas se está activando conforme recorren los 30 pares de parámetros, en el mejor de los casos se deben activar los 10 primeros con la primera neurona, los diez segundo con la segunda neurona y los 10 terceros con la tercera neurona, esto se observa en Figura 9.9 demostrando que el entrenamiento es un éxito.

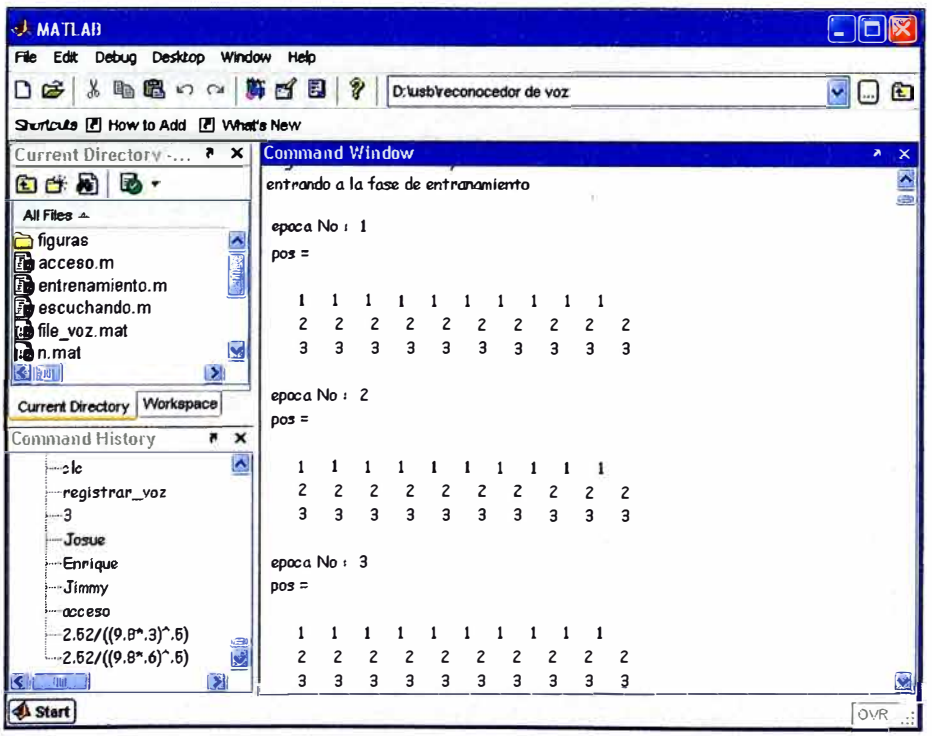


Figura 9.9: Etapa de Entrenamiento de la Red Neuronal en sus Primeras Épocas.

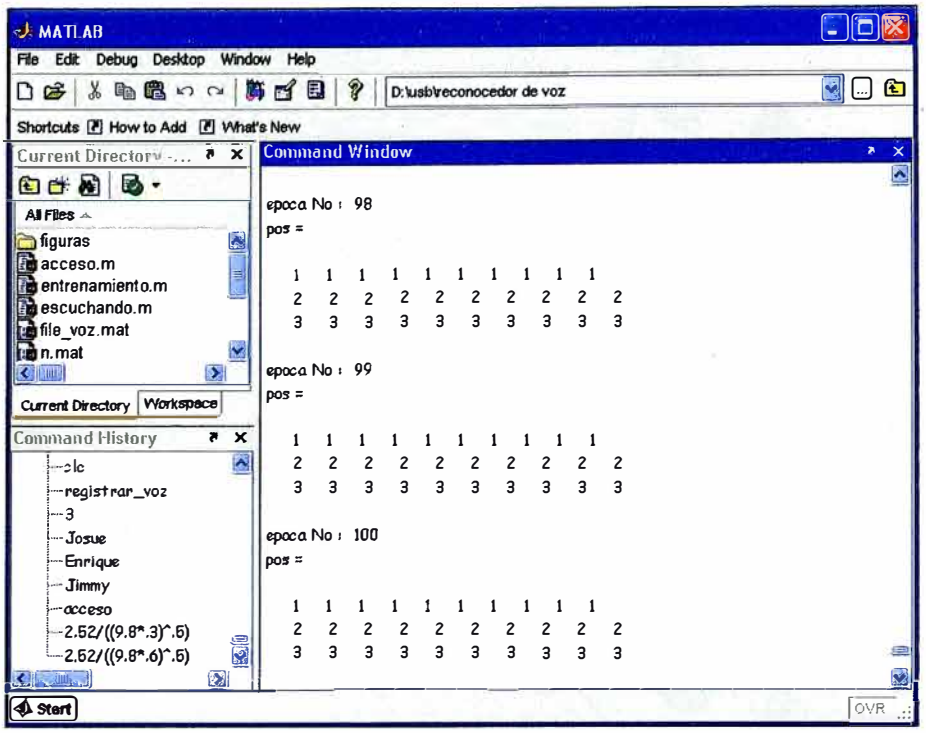


Figura 9.10: Etapa de Entrenamiento de la Red Neuronal en sus Primeras Épocas.

9.4 Validación de la Red Neuronal Competitiva.

Obtenido las características frecuenciales de cada vocal de la persona que a pedido la identificación se procede ha buscar la neurona que contiene las características frecuenciales de su voz de las vocales 'A' y 'I', primero empieza buscando entre todas las neuronas que contienen la letra 'A' y con el criterio de la distancia euclídea se encontrara al que tienes menos distancia y la neurona se activara , luego de esto lo mismo sucederá con la letra 'I' y la neurona activada tendrá que coincidir con la neurona activada correspondiente a la letra 'A' de la misma persona, si cumple con esta condición el sistema diseñado a identificado a la persona, en caso de que no se cumpla con esta condición entramos al caso de que la persona no ha sido registrada anteriormente. Esta lógica se encuentra esquematizada de forma general en Figura 9.11

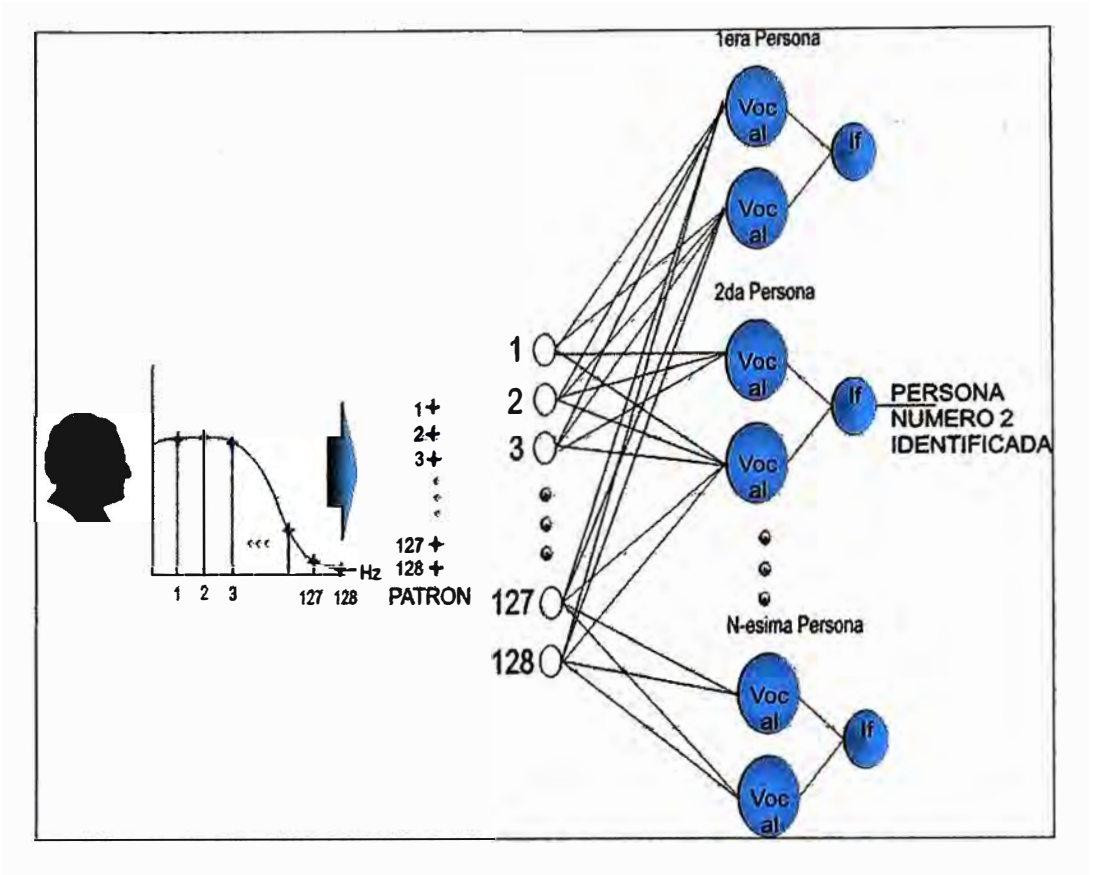


Figura 9.11: Esquema de la identificación de la persona usando el sistema identificador de personas.

La operación descrita anteriormente es realizada por la función ***acceso.m***. El orden para identificar a las persona será primero Enrique luego le seguirá Jimmy culminando con Josue. Como se observa en Figura 9.12. Adicionalmente el sistema permite adicionarle un código de registro dependiendo del orden en que se realizo su registro, esto también se muestra en la figura.

En la Tabla 9.1 muestra una tabla en la que se muestra las activaciones de la red al momento de ingresar un usuario para ser identificado, la activación se da en la neurona con menor distancia euclidiana. En la tabla se sombrea de color verde la neurona activada con menor distancia euclidiana.

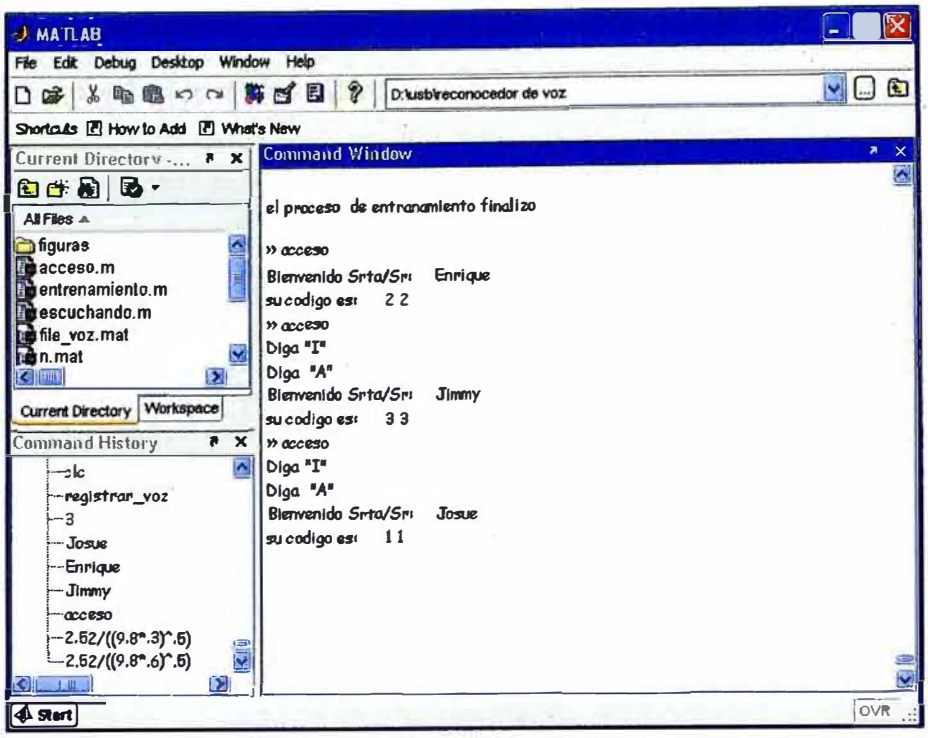


Figura 9.12: Identificación de la persona en MATLAB.

			RED ENTRENADA					
			Jimy		Josue		Enrique	
			Vocal I	Vocal A	Vocal I	Vocal A	Vocal I	Vocal A
USUARIOS	Enrique	Vocal I	0.0916		0.086		0.0843	menor
		Vocal A		0.1114		0.1004	menor	0.094
	Jimy	Vocal I	0.0815	menor	0.0861		0.1096	
		Vocal A	menor	0.0625		0.0636		0.0669
	Josue	Vocal I	0.1026		0.0651	menor	0.1005	
		Vocal A		0.0764	menor	0.0514		0.0614

Tabla 9.1: Prueba de la red entrenada por medio de la distancia euclidiana.

Los tiempos empleados para el procesamiento de la señal, el aprendizaje y la identificación dependen del procesador que realiza la operación así como también del sistema operativo que distribuye la capacidad del procesador a proceso ejecutados en paralelo al sistema

identificador. En la Tabla 9.2 muestra las pruebas realizadas para hallar estos tiempos variando la cantidad de personas registradas en el sistema identificador.

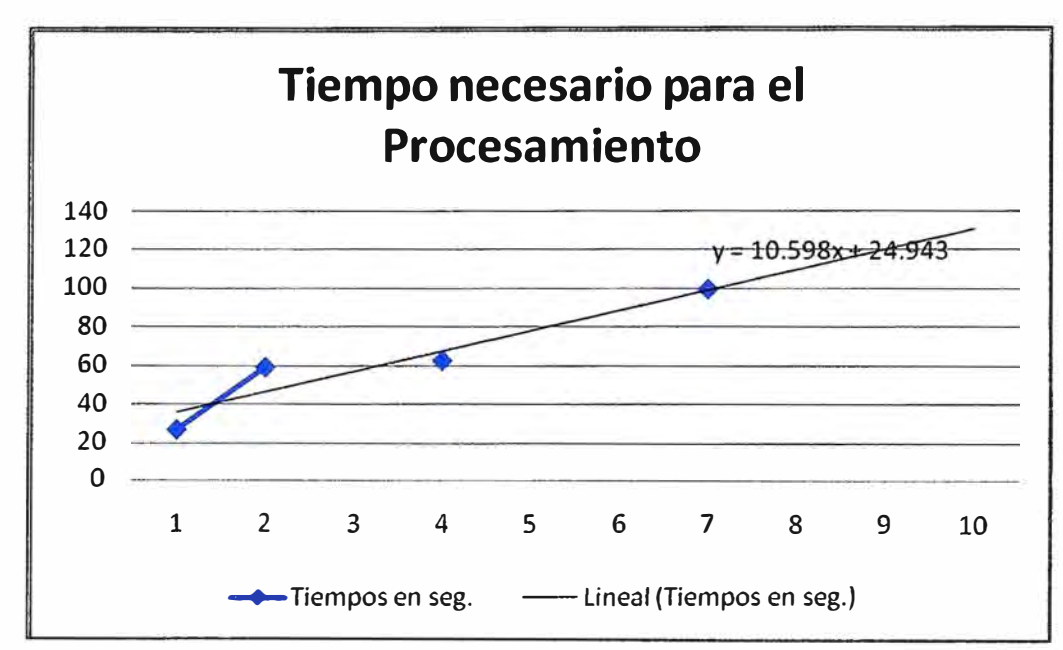
Número de Personas Registradas	1 persona	2 personas		3 personas			4 personas			
Orden de Identificación	1ra. Persona	1ra. persona	2da. persona	1ra. persona	2da. persona	3ra. persona	1ra. persona	2da. persona	3ra. persona	4ta. Persona
Tiempo de Procesamiento en segundos	27.051	59.1668		62.6025			99.3239			
Tiempo de Entrenamiento en segundos	0.265	0.4769		1.1076			1.0769			
Tiempo de Identificación en segundos	6.7269	10.2702	12.1829	7.6582	7.5585	5.5252	8.3809	8.3908	12.8308	10.1402

Tabla 9.2: Prueba de los tiempos de procesamiento, entrenamiento e identificación.

Como se menciono anteriormente los tiempos varían en cada prueba debido al procesador usado y los programas que se estén ejecutando al mismo tiempo y que son administrados por el sistema operativo. En este caso la primera prueba se realizo con el registro de una sola persona requiriendo para el procesamiento de la señal de voz que comprende la adquisición de la señal de voz y la extracción de características frecuenciales de la señal de voz 27 segundos aproximadamente, para el entrenamiento fue necesario un tiempo de0.2 segundos aproximadamente y para la identificación se necesito de 6.7 segundos aproximadamente. Para el caso de tres personas registradas se necesito de 59 segundos aproximadamente para el procesamiento de la señal, para el entrenamiento se necesito de 0.4 segundos aproximadamente para las 3 personas registradas y para la identificación la primera persona registrada se necesito de 7 segundos aproximadamente, para la segunda persona registrada se necesito de 7.5 segundos aproximadamente y para la tercera persona se necesito de 5.5 segundos aproximadamente.

Según los resultados se observa que a mayor cantidad de personas se precisa de mayor tiempo, esto se observa en cada uno de los tres procesos en el que se analiza el tiempo que emplean para ejecutarse.

Para una cantidad de personas registradas el tiempo de identificación es similar para cualquier persona registrada.



Grafica 9.1: Tendencia en el tiempo necesario para el procesamiento de la señal según la cantidad de personas registradas.

Como se observa en la Grafica 9.1, se puede decir que para las pruebas realizadas variando el número de personas registradas los tiempos necesarios para el procesamiento obedecen a la tendencia lineal mostrada en dicha grafica, es decir, que por cada persona se necesitan de 10.6 segundos aproximadamente con un tiempo fijo de 24.9 segundo aproximadamente independiente de la cantidad de personas que se registren.

CONCLUSIONES

- La señal de voz esta compuesta por el tono y timbre que son componentes en frecuencia de la señal de voz y es con estas componentes que se distinguen y diferencian las voces de diferentes personas.
- La voz presenta un rango de frecuencias de hasta 23 Khz pero de los cuales solo hasta los 4 Khz se encuentran las principales o donde se conservan el timbre y el tono de voz.
- Las personas de menor edad tienen su espectro de frecuencias mas concentrado en la zona de altas frecuencias y la de mayor edad la tienen concentrado en la zonas de baja frecuencias, es por esto se diferencian a las personas con voz aguda o voz grave, la voz aguda es aquélla que esta compuesta de altas frecuencias y la voz grave es la que esta compuesta de bajas frecuencias.
- Generalmente las mujeres y los niños tiene la voz aguda mientras que los hombres tienen la voz grave, esto se debe a sus características físicas del tracto vocal.
- Las personas que tienen la misma voz son aquellas que tienen la zona del tono y el timbre iguales en magnitud al observar su densidad espectral.
- Lo que permiten las redes neuronales es diferenciar los espectros de frecuencia de voz de cada persona luego de haber sido entrenada la red.
- Conforme a las pruebas se observa una eficacia del 100% y conforme se incremente la cantidad de personas su eficiencia se verá afectada debido al tiempo que se emplea en el proceso de identificación pero la necesidad de mayor tiempo empleado es debido a la mayor robustez del sistema diseñado.

RECOMENDACIONES

- Para una mejor identificación es mejor tener mas pesos de referencia para que la característica de la envolvente de el espectro de la frecuencia sea mas fiel y así diferenciar de las demás envolventes de cada persona, pero esto se debe hacer con cuidado ya que se necesitaría de una mayor cantidad de memoria y además un mayor tiempo de procesamiento.
- Las variantes en la programación pueden darse para un razonamiento difuso con lo cual la discriminación seria aun mejor.
- A mayor cantidad de usuarios lleva a tomar en cuenta la primera recomendación.
- Idealmente este sistema podría soportar un grupo ilimitado de usuarios pero a mayor cantidad de usuarios mayor es la probabilidad de que exista mas de una persona con el mismo tono y timbre de voz, es por esto que la cantidad de usuarios estaría limitado a las características del grupo de usuarios en cuanto a la diferencia de sus voces.
- El sistema diseñado requiere de dos vocales cualesquiera de las 5 que existen y esas vocales usadas serán requeridas a la hora de la identificación.

REFERENCIAS

[1] Landau y Lifschitz: "Teoría de la Elasticidad", Reverté, 1969. ISBN 84-291-4080-8.

[2] Bergland, G. D. "A Guided Tour of the Fast Fourier Transform." IEEE Spectrum 6, 41-52, July 1969.

[3] http://es.wikipedia.org/wiki/Arqu%C3%ADmedes.

[4] J.G. Proakis, D.G. Manolakis. Digital Signal Processing: Principles, algorithms and applications". Prentice-Hall, Inc. 1996.

[5] http://www.conozcasuhardware.com/quees/tsonido1.htm.

[6] http://physionet.cps.unizar.es/~eduardo/docencia/tds/librohtml/welch1.htm.

[7] http://ohm.utp.edu.co/neuronales/.

[8] http://www.eie.fceia.unr.edu.ar/~acustica/biblio/fonatori.pdf.

[9] http://es.wikipedia.org/wiki/Tarjeta_de_sonido.

[10] http://ebooks.unibuc.ro/filologie/spaniola/1.htm.

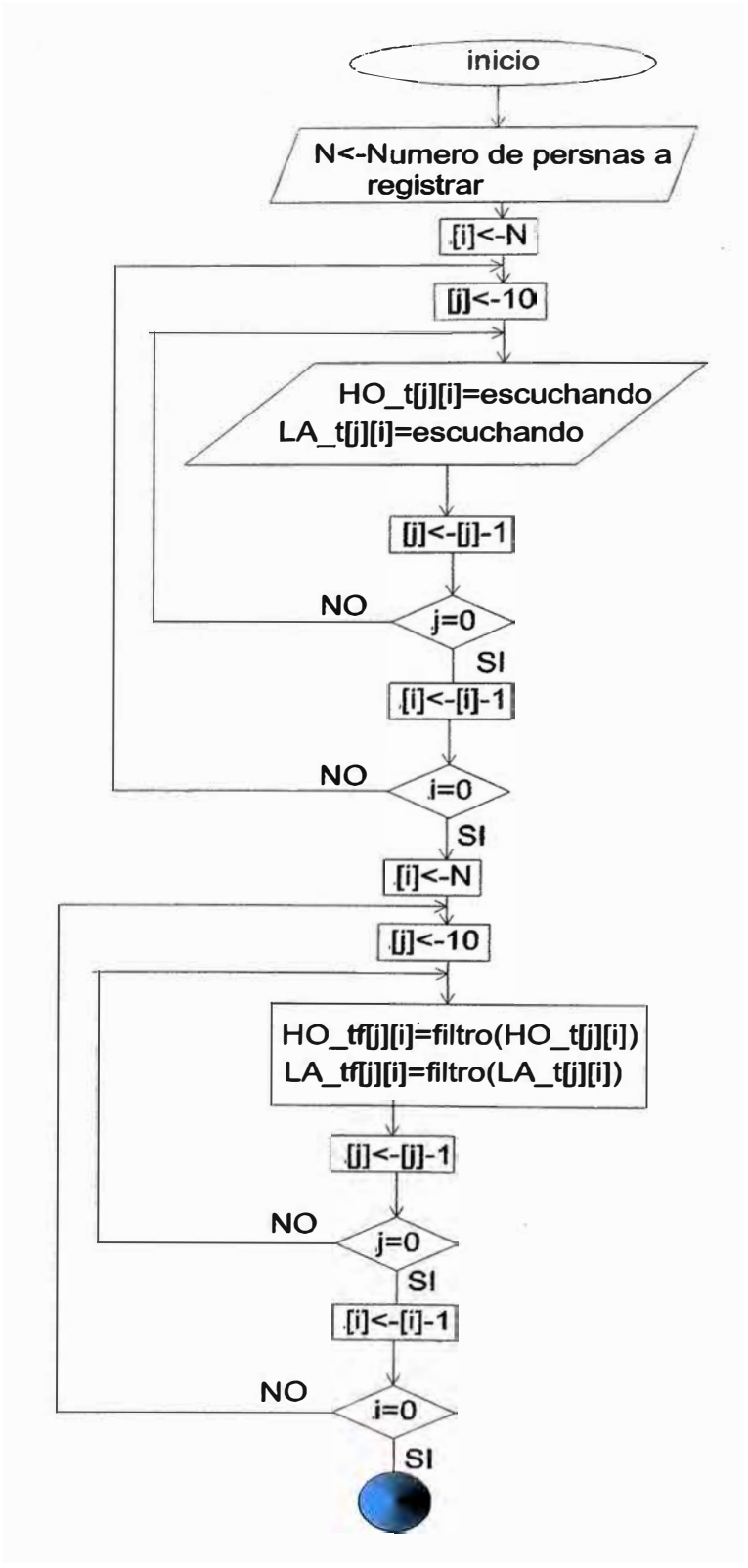
[11] http://www.tecnun.com/asignaturas/tratamiento%20digital/tema8.pdf.

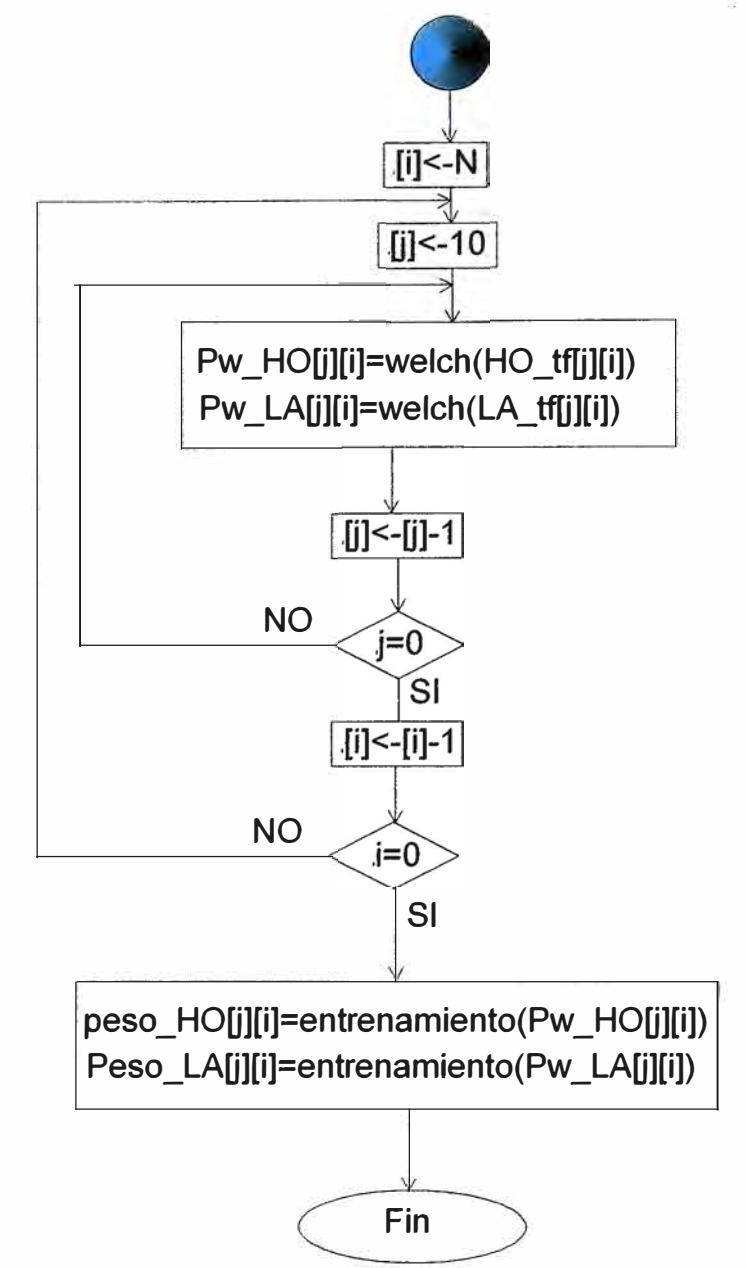
[12] http://www.microsoft.com/whdc/device/audio/multichaud.mspx

APÉNDICE

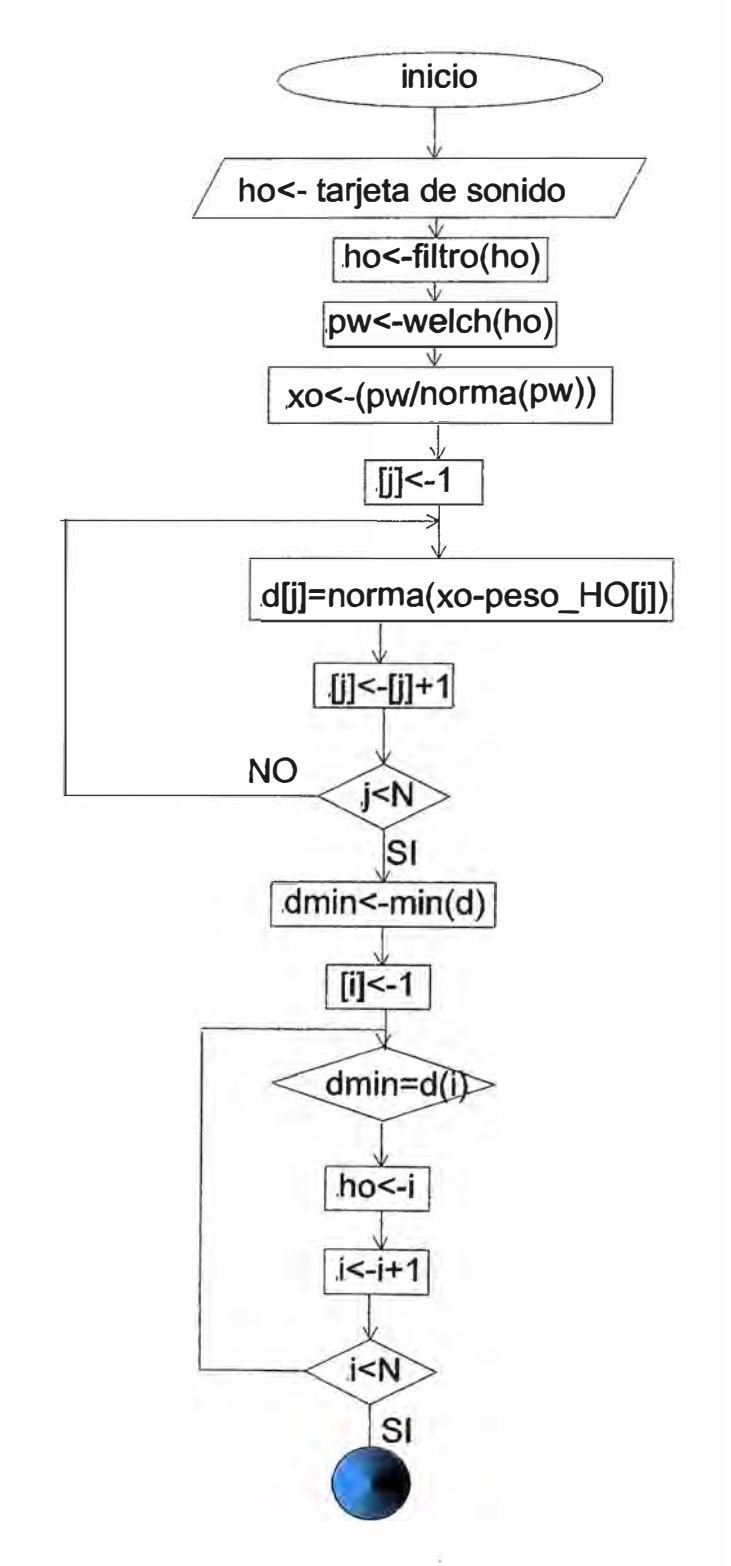
A. DIAGRAMAS DE FLUJO

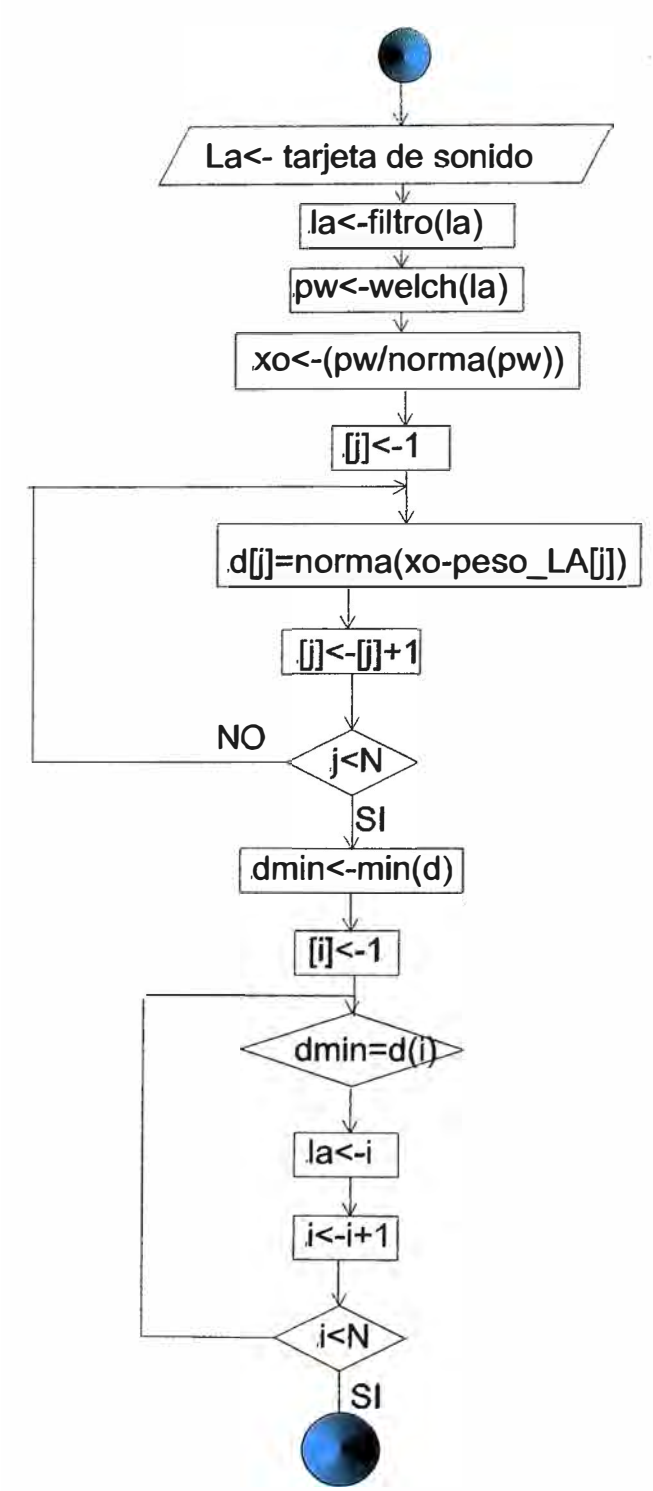
A.1 PARTE DE CAPACITACIÓN




[i]<-N
[j]<-10
Pw_HO[j][i]=welch(HO_tf[j][i])
Pw_LA[j][i]=welch(LA_tf[j][i])
[j]<-[j]-1
NO
j=0
SI
[i]<-[i]-1
NO
i=0
SI
peso_HO[j][i]=entrenamiento(Pw_HO[j][i])
Peso_LA[j][i]=entrenamiento(Pw_LA[j][i])
Fin

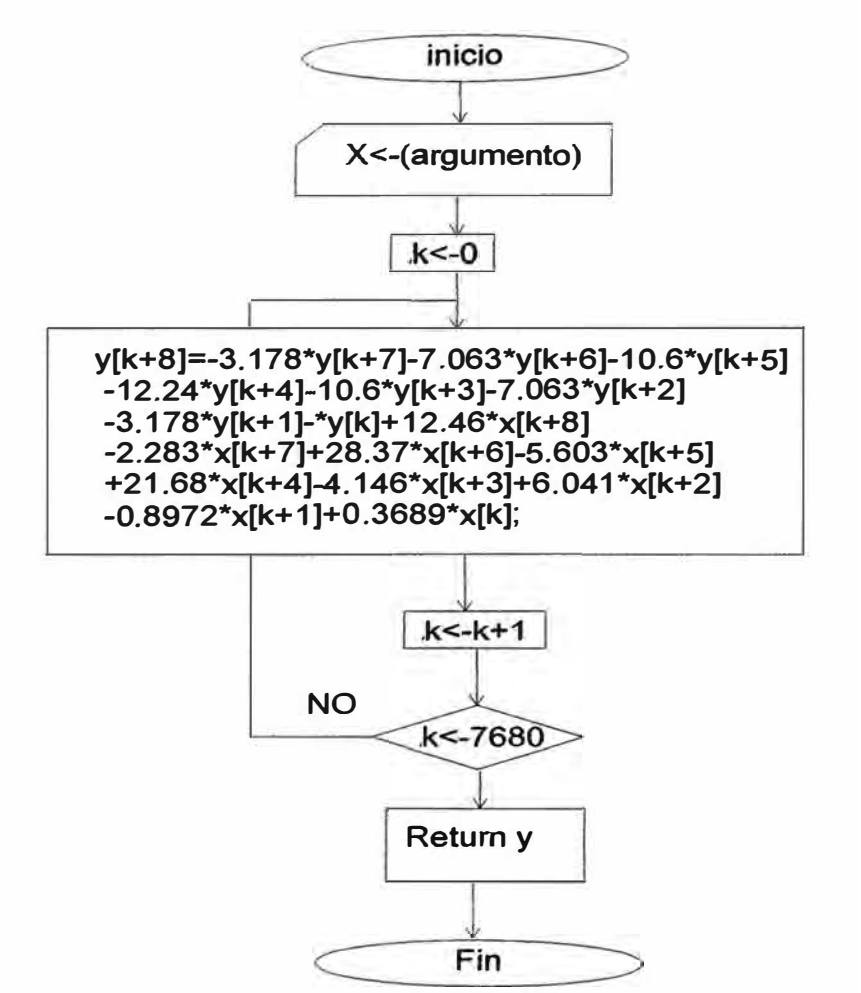
A.2 PARTE DE FUNCIONAMIENTO



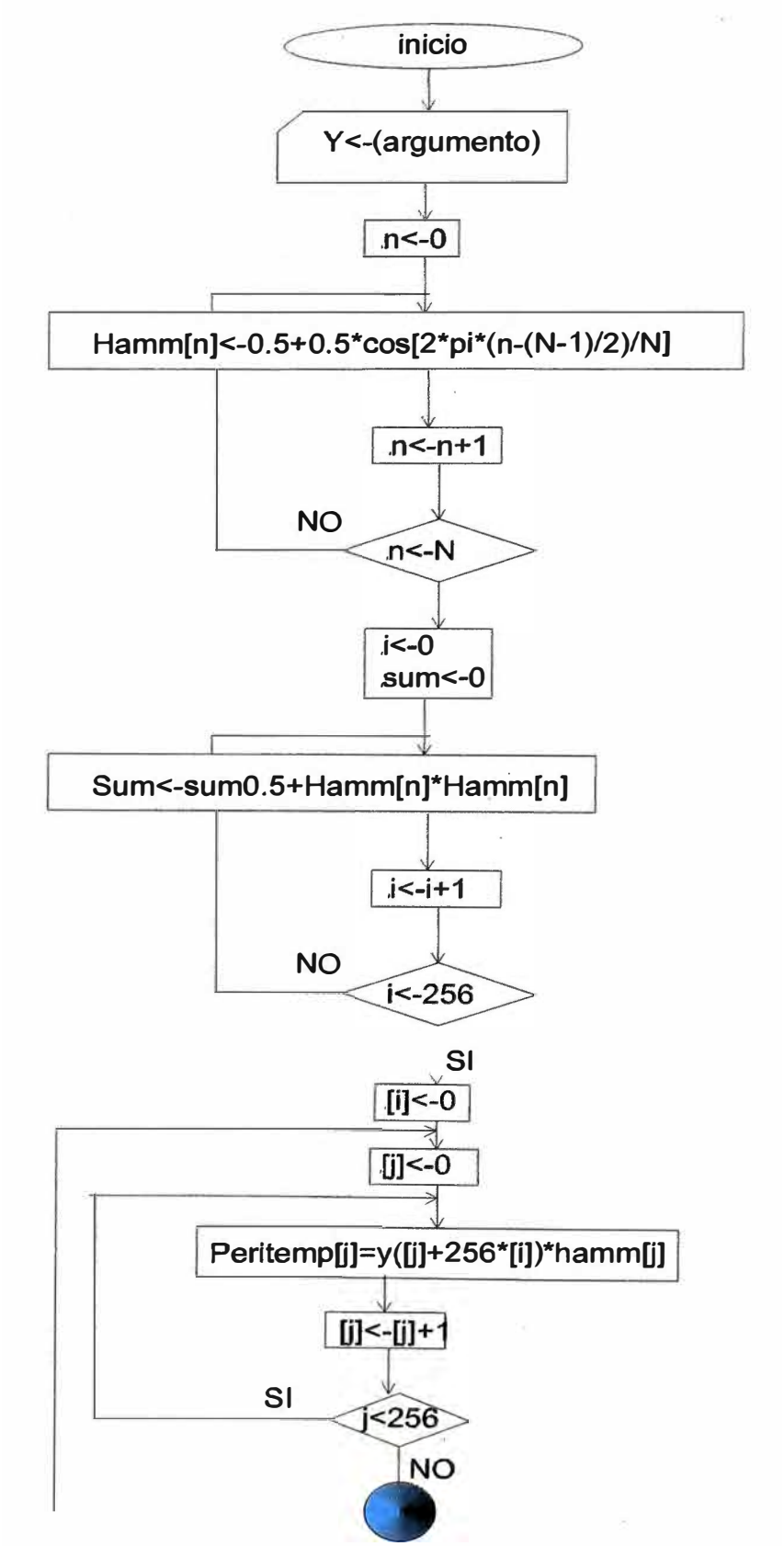

La<- tarjeta de sonido
la<-filtro(la)
pw<-welch(la)
xo<-(pw/norma(pw))
[j]<-1
d[j]=norma(xo-peso_LA[j])
[j]<-[j]+1
NO
j<N
SI
dmin<-min(d)
[i]<-1
dmin=d(i)
la<-i
i<-i+1
i<N
SI

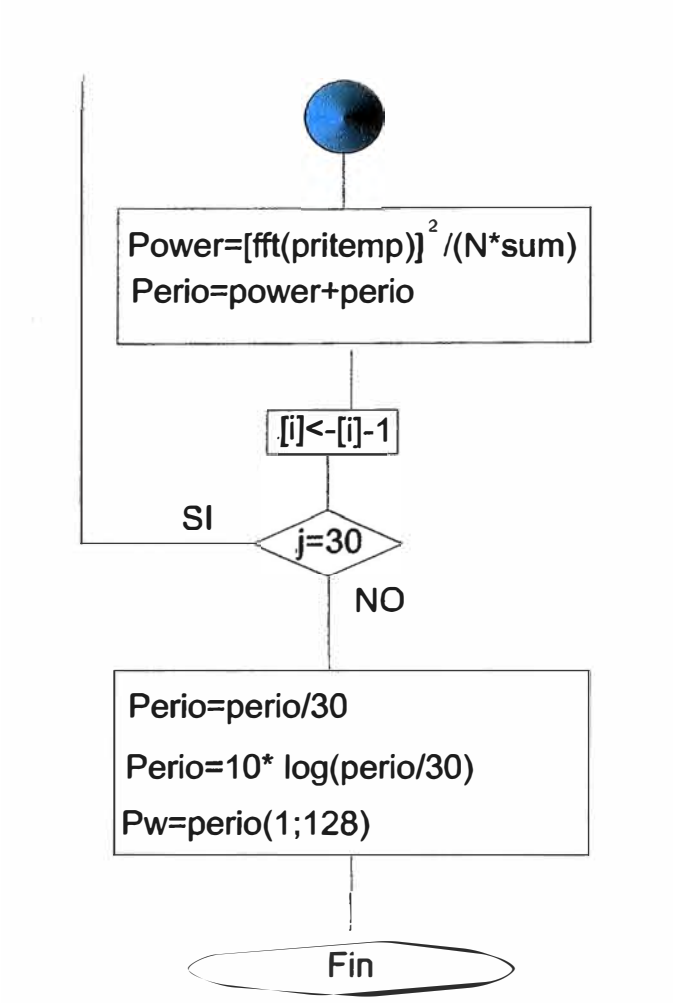
A continuación se realizan los siguientes diagramas de flujo usados para el diseño, estos son ***escuchando***, ***filtro***, ***welch*** y ***entrenamiento,*** además estos usan otros subdiagramas como ***fft*** y ***norma***. ***escuchando*** es un diagrama de flujo que hace una conexión con la tarjeta de sonido.

A.3 DIAGRAMA DE FLUJO DEL SUBPROGRAMA FILTRO

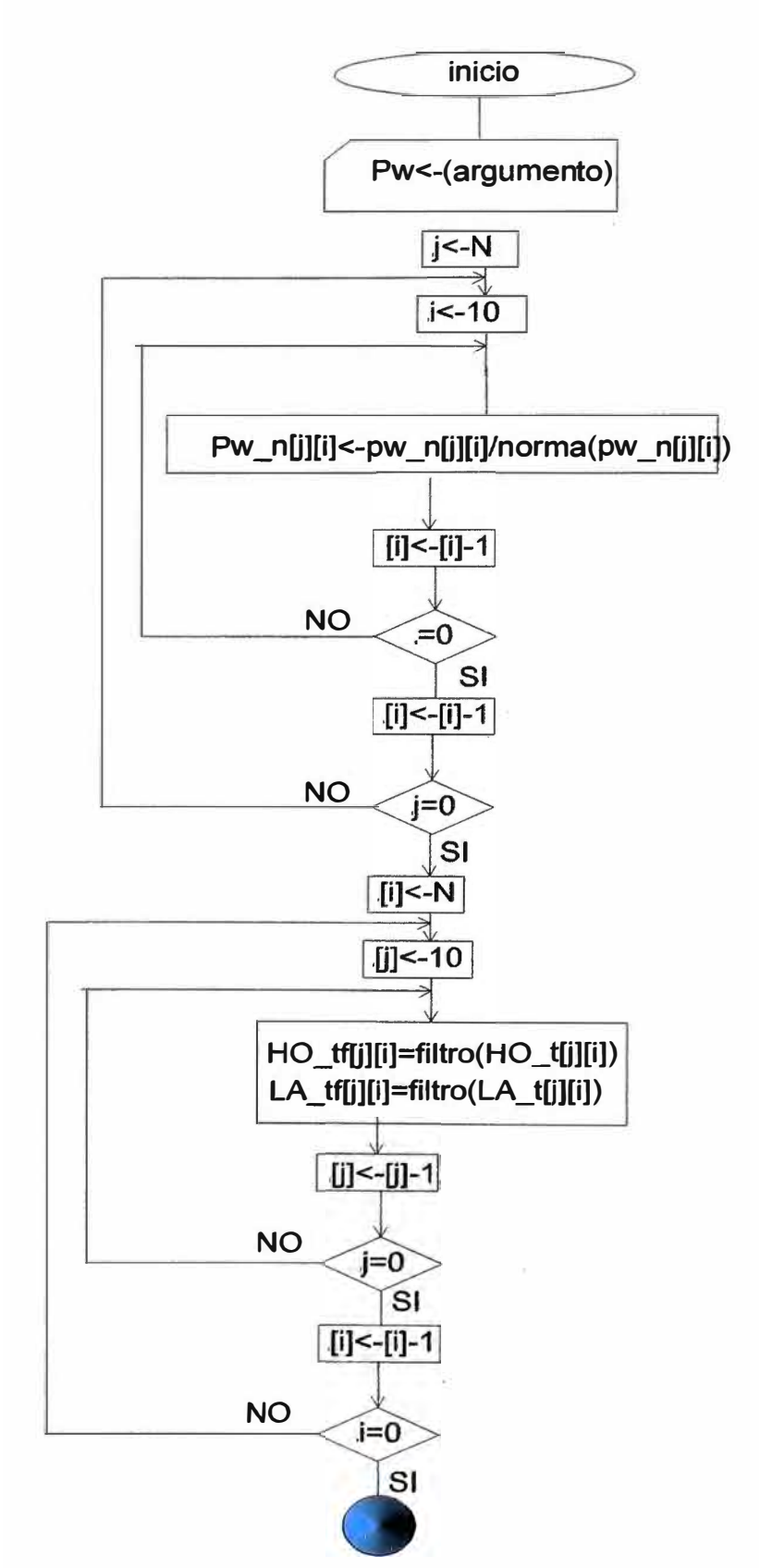


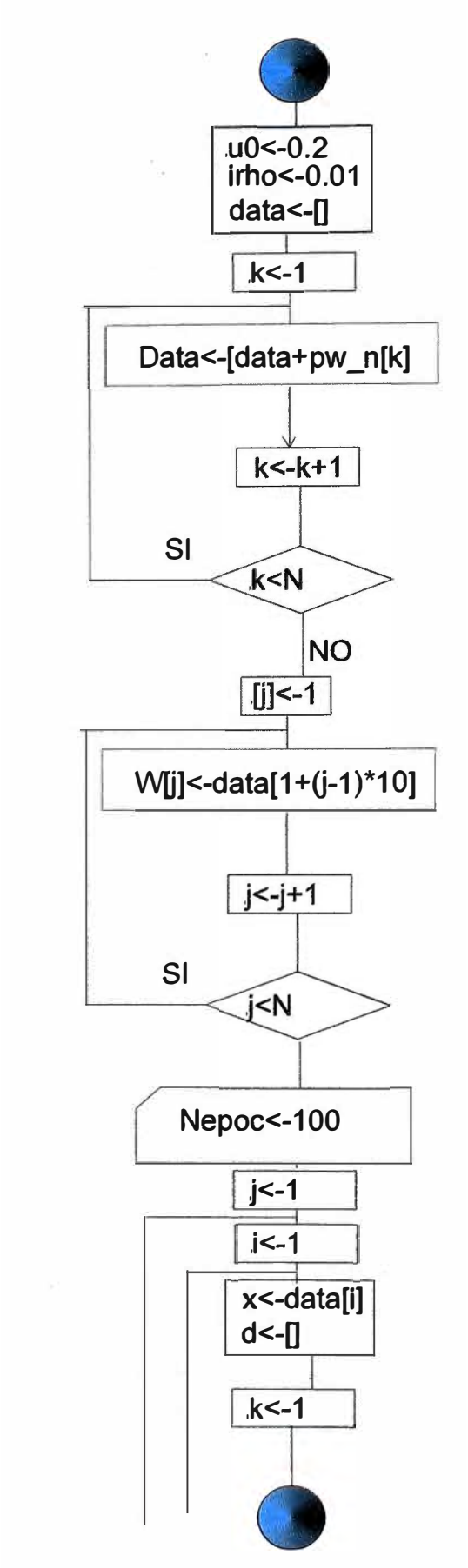
A.4 DIAGRAMA DE FLUJO DEL SUBPROGRAMA WELCH

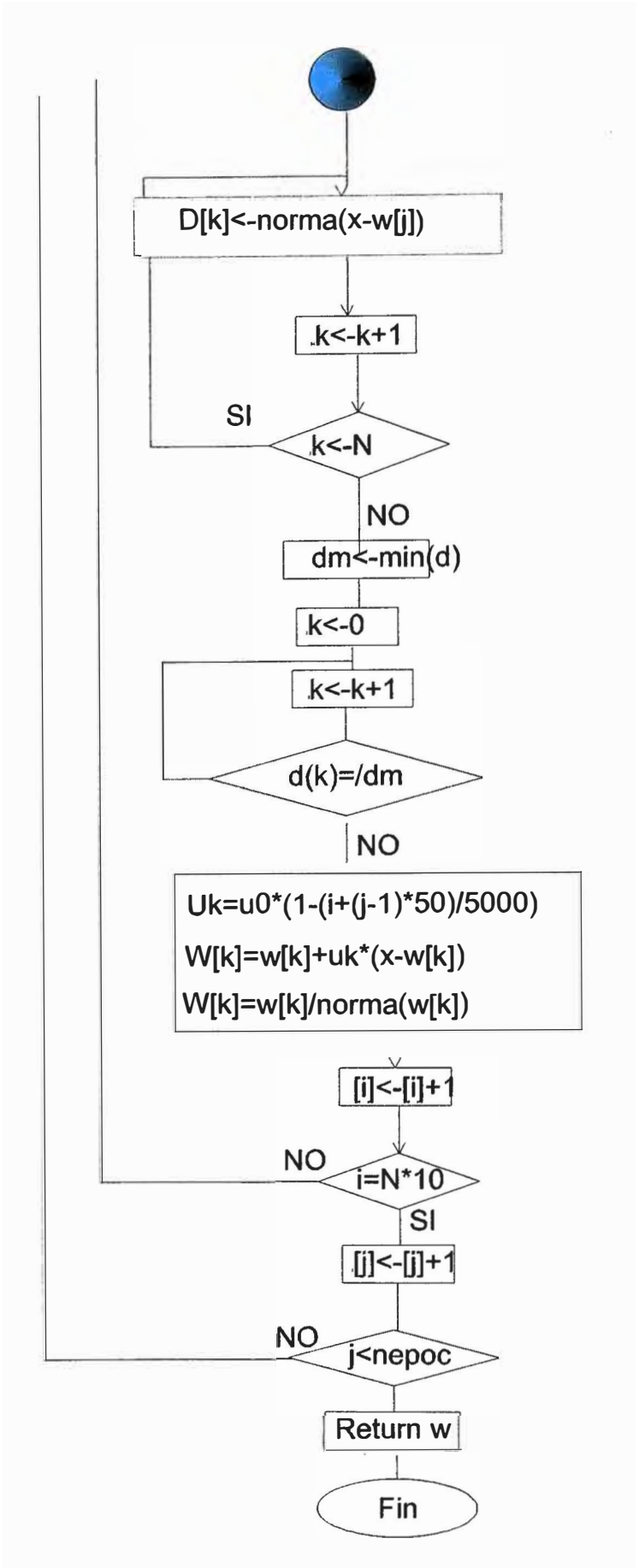



Power=[fft(pritemp)]^2 /(N*sum)
Perio=power+perio
[i]<-[i]-1
SI
j=30
NO
Perio=perio/30
Perio=10* log(perio/30)
Pw=perio(1;128)
Fin

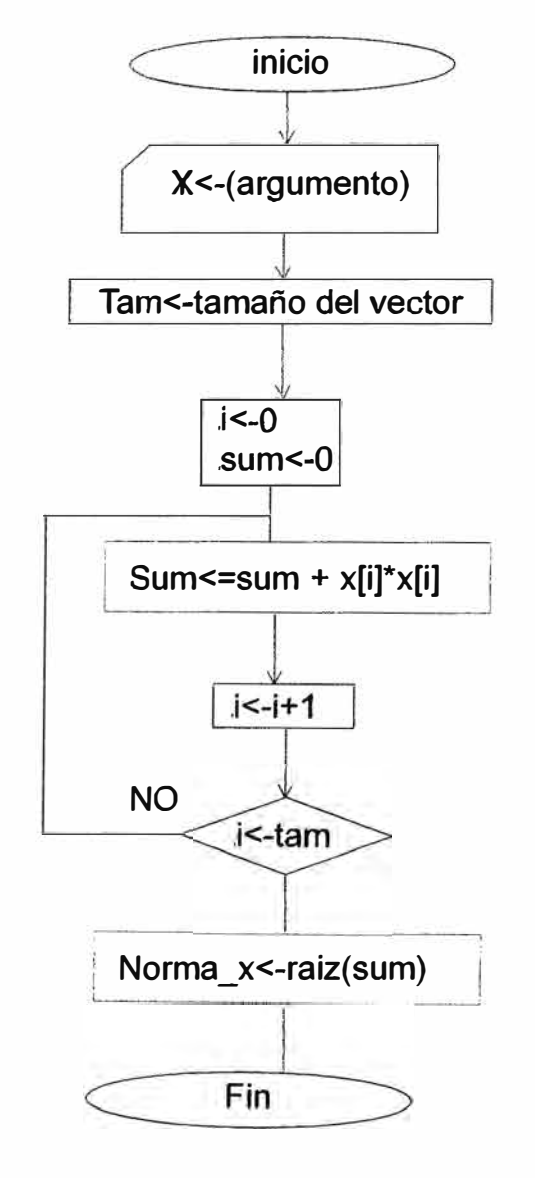
A.5 DIAGRAMA DE FLUJO DEL SUBPROGRAMA ENTRENAMIENTO



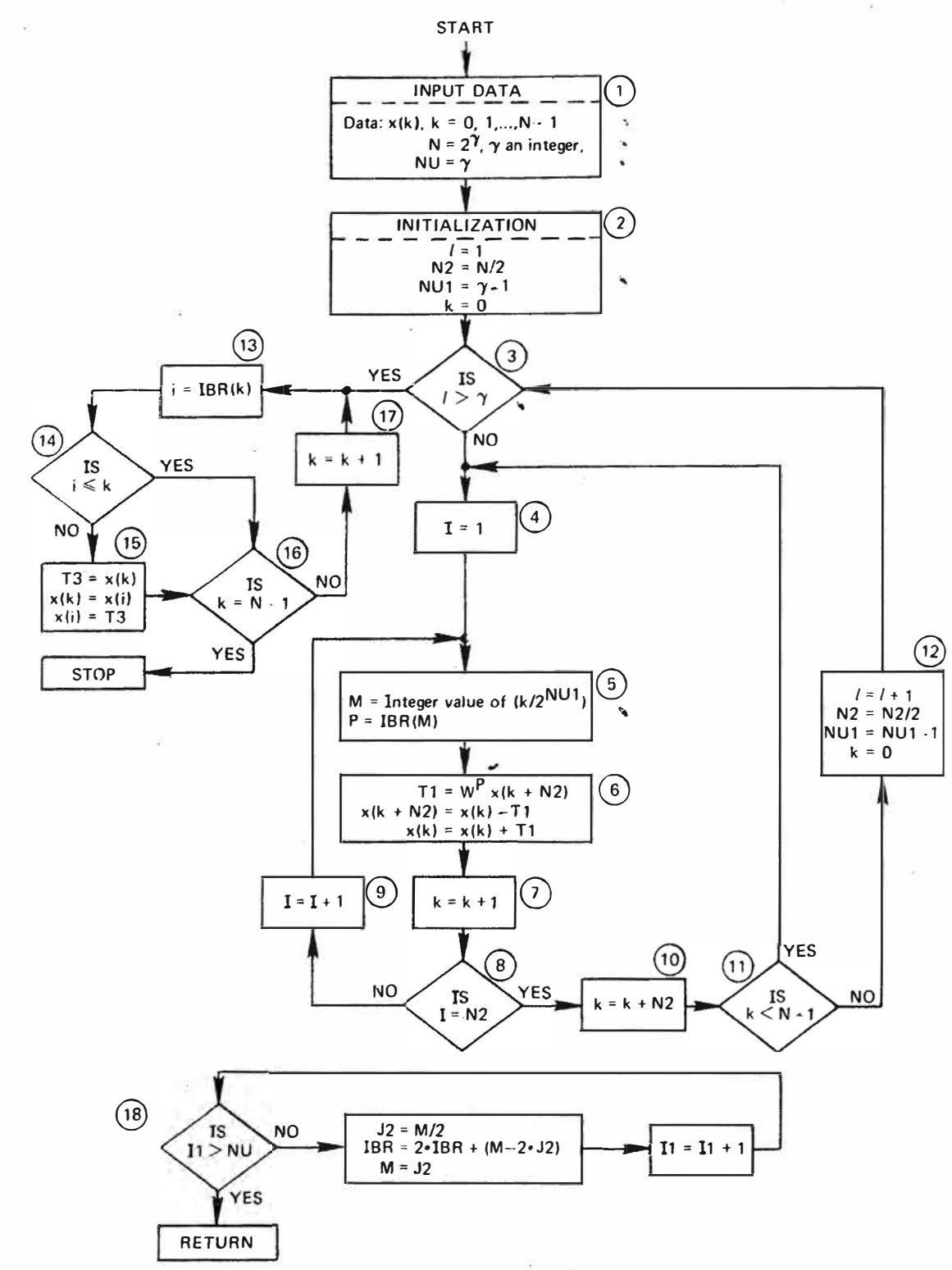

u0<-0.2
irho<-0.01
data<-[]
k<-1
Data<-[data+pw_n[k]
k<-k+1
SI
k<N
NO
j<-1
W[j]<-data[1+(j-1)*10]
j<-j+1
SI
j<N
Nepoc<-100
j<-1
i<-1
x<-data[i]
d<-[]
k<-1


D[k]<-norma(x-w[j])
k<-k+1
SI
k<-N
NO
dm<-min(d)
k<-0
k<-k+1
d(k)=/dm
NO
Uk=u0*(1-(i+(j-1)*50)/5000)
W[k]=w[k]+uk*(x-w[k])
W[k]=w[k]/norma(w[k])
[i]<-[i]+1
NO
i=N*10
SI
[j]<-[j]+1
NO
j<nepoc
Return w
Fin

A.6 DIAGRAMA DE FLUJO DEL SUBPROGRAMA DE NORMA



A.7 DIAGRAMA DE FLUJO DEL SUBPROGRAMA FFT



E FUNCIONES DESARROLLADAS EM MATLAB

E.1: PROGRAMA PARA EL APRENDIZAJE DE LAS VOCES: *registrar_voz.m*

```
% INGRESO DE DATA DE USUARIOS
voz = struct('HO_t',{[]},'LA_t',{[]},'nombre',{''},'HO_w',{[]},'LA_w',{[]});
N=input('Ingrese la cantidad de Usuarios:  ');
for z=1:N
    fprintf('Diga la "I" y presione enter \n')
   voz.HO_t{z}=escuchando(N);
    fprintf('Diga la "A" y presione enter \n')
  voz.LA_t{z}=escuchando(N);
   voz.nombre{z}=input('Ingrese su nombre: ','s');
end
for z=1:N
   figure(2*z-1)
voz.HO_w{z}=y_Y(voz.HO_t{z});
   figure(2*z)
voz.LA_w{z}=y_Y(voz.LA_t{z});
end
HO_w=voz.HO_w;
LA_w=voz.LA_w;
save('n','N')
save('file_voz','voz')
%ENTRENAMIENTO
fprintf('entrando a la fase de entranamiento ')
entrenamiento;
fprintf('el proceso  de entranamiento finalizo')
```

E.2: PROGRAMA QUE SE ENCARGA DE LA IDENTIFICACION: *acceso.m*

```
function acceso
 load('peso_ho')
load('peso_la')
load('n')
load('file_voz')
Fs=16000;
fprintf('Diga "I"\n')
pause;
y=wavrecord(7680,Fs);
[b,a] = ellip(8,0.01,40,4000/8000);
yf=filter(b,a,y);
pw=welch(yf,7680,256,Fs);
xo=(pw(1:128)/norm(pw(1:128)))';
for i=1:N
d(i)=norm(xo-w_ho(:,i))
end
dm=min(d);
for j=1:N
   if dm==d(j)
      ho=j;
   end
end
fprintf('Diga  "A"\n')
pause;
y=wavrecord(7680,Fs);
[b,a] = ellip(8,0.01,40,4000/8000);
yf=filter(b,a,y);
pw=welch(yf,7680,256,Fs);
xa=(pw(1:128)/norm(pw(1:128)))';
for i=1:N
d(i)=norm(xa-w_la(:,i))
end
dm=min(d);
for j=1:N
   if dm==d(j)
      la=j;
   end
end
if la==ho
   fprintf('Bienvenido Srta/Sr:    %s  \n', voz.nombre{la})
   fprintf('su codigo es:     %g %g \n', ho,la)
else
   fprintf('No autorizado   \n')
end
```

E.3: PROGRAMA QUE SE ENCARGA DE CAPTURAR LA SEÑAL DE VOZ: *escuchando.m*

```
function [yt]=escuchando(N)
pause;
Fs=16000;
for i=1:10
yt(:,i)=wavrecord(7680,Fs);
end
```

E.4: FUNCION PARA EXTRAER LAS CARACTERISTICAS FRECUENCIALES DE LA VOZ: *y_Y.m*

```
function [Y_w]=y_Y(y_t)
Fs=16000;
pw=zeros(256,10);
[b,a] = ellip(8,0.01,40,4000/8000);
for i=1:10
yf=filter(b,a,y_t(:,i));
subplot(5,2,i)
pw(:,i)=WELCH(yf,7680,256,Fs)';
maxpw(i)=max(abs(pw(1:128,i)));
Y_w(:,i)=pw(1:128,i)/maxpw(i);
end
```

E.5: FUNCION PARA CALCULAR EL PROMEDIO DE PERIODOGRAMAS: welch.m

```
function [perio]=welch(y,M,N,Fs)
perio=0;
sum=0;
hamm=hamming(N);
for i=1:N
   sum=sum+hamm(i)*hamm(i);
end
sum=sum/N;
peritemp=0;
for i=0:((M/N)-1)
   for j=1:N
      peritemp(j)=y(j+N*i)*hamm(j);
   end
   power=(abs(fft(peritemp)).^2)/(N*sum);
   perio=power+perio;
end
perio=perio*N/M;
f=0:1/N*Fs:Fs-1;
perio=10*log(perio);
plot(f(1:N/2),perio(1:N/2))
xlabel('Frecuencia (Hz)')
ylabel('PSD (dB)')
```

E.6: PROGRAMA USADO PARA REALIZAR EL APRENDIZAJE COMPETITIVO DE LA RED NEURONAL ARTIFICIAL: *entrenamiento.m*

```
%red kohonen competitiva
%entrenamiento para reconocer vocales
function  entrenamiento
load('file_voz')
load('n')
ho=voz.HO_w;
la=voz.LA_w;
% entrenamiento para primera vocal
for j=1:N
   for i=1:10
      ho_n{j}(:,i)=ho{j}(:,i)/norm(ho{j}(:,i));
   end
end
%Entrenamiento de la red con
%aprendizaje competitivo Kohonen
u0=0.2;
rho=0.01;
data=[];
for k=1:N
data=[data  ho_n{k}];
end
 w=[];
for i=1:N
w(:,i)=data(:,1+(i-1)*10);
end

nepoc=100;
pos=zeros(N,10);
for j=1:nepoc
p=1;
q=0;
   for i=1:N*10
%          i=i+1
      x=data(:,i);
      d=[];
      for k=1:N
         d(k)=norm(x-w(:,k));
      end

      dm=min(d);
      k=1;
      while d(k)~=dm
```

```
            k=k+1;
        end
        q=q+1;
        if q>10
           q=i-(p)*10;
           p=p+1;
        end

        pos(p,q)=k;
%            <-----------------
        uk=u0*(1-(i+(j-1)*50)/5000);
        w(:,k)=w(:,k)+uk*(x-w(:,k));
        w(:,k)=w(:,k)/norm(w(:,k));
    end
    fprintf('epoca No :  %g  ', j)
    pos
end
w_ho=w;
clear w
% entrenamiento para la segunda vocal
for j=1:N
    for i=1:10
        la_n{j}(:,i)=la{j}(:,i)/norm(la{j}(:,i));
    end
end
%Entrenamiento de la red con
%aprendizaje compeitivo Kohonen
u0=0.2;
rho=0.01;
data=[];
for k=1:N
data=[data  la_n{k}];
end

w=[];
for i=1:N
w(:,i)=data(:,1+(i-1)*10);
end
nepoc=100;
pos=[];
for j=1:nepoc
p=1;
q=0;
    for i=1:N*10
        x=data(:,i);
        d=[];
```

```
        for k=1:N
           d(k)=norm(x-w(:,k));
        end
        dm=min(d);
        k=1;
            while d(k)~=dm
                k=k+1;
            end
        q=q+1;
        if q>10
           q=i-(p)*10;
           p=p+1;
        end

         pos(p,q)=k;
%            <-----------------

        uk=u0*(1-(i+(j-1)*50)/5000);
              w(:,k)=w(:,k)+uk*(x-w(:,k));
              w(:,k)=w(:,k)/norm(w(:,k));
    end
    fprintf('epoca No :  %g  ', j)
     pos
end
w_la=w;
save('peso_ho','w_ho')
save('peso_la','w_la')
```