Universidad Autónoma de Querétaro.

Maestría en Instrumentación y Control Automático

Tesis del proyecto:

Sistema de reconocimiento de voz usando perceptrón multicapa y Coeficientes Cepstrales de Mel.

Asesor: Dr. Edgar Alejandro Rivas Araiza.

Alumno: Juan Guillermo García Guajardo.

Santiago de Querétaro, Qro. Mayo 2011

Portada Interna de Tesis

Universidad Autónoma de Querétaro
Facultad de Ingeniería
Maestría en Ciencias en Instrumentación y Control Automático

SISTEMA DE RECONOCIMIENTO DE VOZ USANDO PERCEPTRÓN MULTICAPA Y COEFICIENTES CEPSTRALES DE MEL

TESIS

Que como parte de los requisitos para obtener el grado de Maestro en Ciencias en Instrumentación y Control Automático.

Presenta:
Juan Guillermo García Guajardo

Dirigido por:
Dr. Edgar Alejandro Rivas Araiza

SINODALES

Dr. Edgar Alejandro Rivas Araiza Presidente	Firma
Dr. Manuel Toledano Ayala Secretario	Firma
M. en C. Alfonso Noriega Ponce Vocal	Firma
M. en C. Aurora Femat Díaz Suplente	Firma
M. en C. Guillermo Ronquillo Lomelí Suplente	Firma
Dr. Gilberto Herrera Ruiz Nombre y Firma Director de la Facultad	Dr. Luis Gerardo Hernández Sandoval Nombre y Firma Director de Investigación y Posgrado

Centro Universitario
Querétaro, Qro.
Mayo 2011
México

Resumen

El avance de la tecnología ha traído un innumerable desarrollo de algoritmos y sistemas encargados de facilitar el trabajo cotidiano del ser humano, tal es el caso del reconocimiento de voz que en nuestros días es algo que continuamente se está introduciendo en dispositivos como celulares y computadoras por la accesibilidad remota y fácil interacción con sistemas complejos. Sin embargo, el reconocimiento automático de voz es un problema que aún no se encuentra completamente resuelto debido a la variabilidad de la voz. En este trabajo se presenta una metodología para poder reconocer comandos aislados mono-locutor usando la parametrización de la señal de voz a través de la técnica de los Coeficientes Cepstrales de Mel, los cuales se inspiran en la percepción del oído humano ante los sonidos de su entorno y trabajan en el plano frecuencial que nos proporciona la Transformada Rápida de Fourier, estos parámetros sirven como entrada a una Red Neuronal Artificial para poder comparar patrones de distintas palabras. Obteniendo resultados de acierto cercanos al 90% en un grupo de veinte distintos comandos. A la salida de esta Red Neuronal se adecuó un algoritmo encargado de enviar la palabra correcta por el protocolo de comunicación RS-232 hacia un módulo GSM con la finalidad de tener conectividad con el entorno, permitiendo enviar mensajes de texto a cualquier otro dispositivo móvil de forma automática.

(**Palabras clave:** Coeficientes Cepstrales de Mel, Redes Neuronales Artificiales, Perceptrón Multicapa, Transformada Rápida de Fourier, Parametrización, RS-232, GMS.)

SUMMARY

Nowadays a lot of systems and algorithms have been developing to easier daily work, improving always the Human Machine Interaction (HMI), in the case of Voice Recognizers (VR), which are implemented continuously on cellular phones and computers for the remote access in the presence of complex systems. However, the automatic speech recognizing is not entirely solved because the voice signal changes from each person, language and geographical region. This work shows a methodology of command recognizer mono-locutor, that uses Mel Frequency Cepstral Coefficients (MFCC) parameterization, which is inspired in human perception hear, for this set upon we must be in a frequency plane that in this case Fast Fourier Transform (FFT) was employed. After applying MFCC algorithm we introduce the vector generated as an input of Artificial Neural Network (ANN), previously trained, to compare and activate the respective command neuron. Resulting in 90% of twenty different commands probed were successfully recognized. Over the ANN output an algorithm was implemented to send the correct word thought Global System for Mobile communications (GSM) module using RS-232 communication protocol with the purpose of add connectivity and allow automatic way to send text messages to another's mobile devices.

(**Keywords:** Mel Frequency Cepstral Coefficients, Artificial Neural Networks, Neuron, Fast Fourier Transform, RS-232, Global System for Mobile Communications.)

A mis padres, hermanos y sobrinos

Que siempre me dieron un motivo y estuvieron presentes desde la distancia en esta lucha interminable que tengo de superarme.

A mi novia

Que me apoyo desde el primer día y fue mi soporte más cercano durante estos años de aprendizaje.

Agradecimientos

Quiero agradecer especialmente a mi director de tesis el Dr. Edgar Alejandro Rivas Araiza, persona que en realidad admiro por el empeño y dedicación que tiene hacia cualquier actividad que realiza. Quién me enseño con el ejemplo de que el trabajo no tiene horario, agenda o fatiga cuando en realidad te gusta lo que estas haciendo.

Índice

Lista de figuras

Lista de Tablas

Capítulo 1:

Introducción.

En este capítulo se da el panorama en el que se encuentra el proyecto, así como el planteamiento teórico y su respectiva justificación, se presenta la metodología a seguir y se plantean distintas metas trazadas para desarrollar el proyecto.

En el capítulo 2 se revisan los trabajos que anteceden a esta tesis y se dividen en dos partes principales, una en donde se describen a manera de breve historia los primeros sistemas de reconocimiento de voz, y la otra parte en donde se presentan varios artículos actuales en donde manejan las técnicas modernas del reconocimiento de voz.

El capítulo 3 tiene la finalidad de explicar la fundamentación teórica necesaria para desarrollar los algoritmos, fortaleciendo este trabajo de investigación, con bases matemáticas.

En el capítulo 4 se muestra la experimentación práctica, mostrando los métodos y el proceso que se necesitó para desarrollar el trabajo.

Finalmente en el capítulo 5 se exponen e interpretan los resultados obtenidos, también se presentan las conclusiones y propuestas de trabajo a futuro.

1.1 Marco global del proyecto

Para comenzar, es necesario introducir a la terminología empleada en la tesis desde sus orígenes pero de manera explícita y sencilla.

El fenómeno físico que permite este objeto de estudio es el sonido, que es el efecto de la propagación de una onda longitudinal a través de un medio elástico, la onda es generada por el movimiento vibratorio de un cuerpo (Guitart, 2001).

Las ondas tienen propiedades que pueden darnos información acerca del cuerpo de origen como su amplitud (que es la variación máxima del desplazamiento) y la frecuencia (que es

una medida que indica el número de veces en las que se repite un suceso en cierto tiempo), ésta tiene como unidad internacional de medida el Hercio (Hz).

Los seres humanos percibimos el sonido por medio del aparato auditivo (captando sólo ondas sonoras que tengan una frecuencia entre 20 y 20,000 Hz) que recibe la información, la conduce, la modifica adaptándola a cierto modelo de captación humana y la amplifica para después entregarla a los conductos neuronales.

Emitimos sonidos por la boca con ayuda del aparato fonador, que está conformado por varios órganos del cuerpo, entre los cuales están los pulmones, la cavidad nasal y bucal, la faringe y laringe, (Encarta, 2006), con éste aparato, el aire que inspiramos se puede distorsionar para provocar un sonido deseado. Cabe mencionar que la voz, como espectro de frecuencia, es útil solo en los primeros 10 Khz, ya que los sonidos que podemos emitir como parte de una palabra se genera en este rango, pero un mensaje puede ser comprendido con tan solo 8 Khz ya que la mayor parte del contenido frecuencial se encuentra dentro este rango. El conjunto de varios sonidos puede formar parte de un lenguaje (que se define como "un conjunto de signos estructurados que dan a entender una cosa"), en el caso del lenguaje oral, los signos, son las señales sonoras recibidas. (Soriano, 2004).

La comunicación oral entre humanos se da gracias a la mezcla entre hablar y escuchar por parte de dos o más individuos, pero para comprender lo que los demás intentan expresar, es necesario asociar los sonidos recibidos con otros que se han escuchado anteriormente y que se les han dado algún significado. Esto provoca una reacción cerebral que nos ayuda a descifrar el mensaje y así poder responder. (Fonseca 2003).

Ya con estos conceptos definidos podemos mencionar que los sistemas digitales encargados de reconocer el habla humana, en su primera etapa captan señales sonoras por medio de un micrófono (Serajul, 2008), que es un transductor electroacústico que convierte las ondas sonoras en variaciones eléctricas.

Pero para que dichas máquinas puedan hacer uso de ésta información, es necesario convertir la variación de la señal con respecto al tiempo a lenguaje binario, que se compone de unos o ceros. Esto se logra mediante un convertidor Analógico–Digital, mejor conocido como ADC de sus siglas en inglés *Analog to Digital Converter*. La función de éste dispositivo es tomar una muestra cada cierto tiempo definido y representar el valor que está a la entrada con un valor finito (discreto) de amplitud que puede ser cualquiera de los 2^n valores posibles, donde *n* es el número de bits de salida. (Donald, 2005).

Se necesita segmentar la señal de voz para eliminar datos que no pertenecen a la representación de una palabra, esto es, elegir cierto valor de umbral que deseche muestras que no cumplan con una amplitud específica. (Kotti, 2007) (Yang, 2008).

1.2 Justificación.

Desde el comienzo de la era digital la humanidad ha buscado la forma de poder manipular las máquinas de una manera simple y rápida, esto ha llevado a realizar muchos y variados lenguajes tanto de programación como de interpretación de archivos, pero sin duda una de las mejores formas de comunicación es el lenguaje oral ya que se emplea en la vida cotidiana y tiene la facilidad de no tocar en lo más mínimo ningún dispositivo.

Hablar siempre ha sido de lo más natural para la mayoría de nosotros, crecemos y aprendemos un sinfín de palabras durante toda la vida, esto nos ayuda a entendernos y a su vez expresar pensamientos, sentimientos, ordenar que se realice una tarea, preguntar por algo que nos interesa, etc. Poder llevar esta herramienta al plano digital es sin duda un gran avance de la tecnología. Ya que une al ser humano con las máquinas modernas, aunque no se cuente con conocimiento alguno acerca del funcionamiento o programación del sistema. (Stiefelhagen, 2004).

Es el inicio de una nueva etapa, la que aún solo se encuentra en libros o películas visionarias, en las que la vida cotidiana se mezcla de tal forma que se toma como normal que un robot o computadora haga las tareas que se le piden directamente o sea capaz de entablar una conversación rica en información.

México necesita hoy formar parte de esta evolución con recursos humanos capaces de desarrollar sistemas de buen nivel para así ser autosuficientes y no tener que importar tecnología que muchas veces es costosa y celosa en cuanto a datos de diseño o manejo técnico.

No podemos quedarnos con los brazos cruzados viendo al futuro solo en los ojos de naciones vecinas, tenemos que adaptarnos a éste presente que cambia continuamente ya que sólo con trabajo y esfuerzo la prosperidad llega.

Es un gran reto, sin duda, adentrase en el área de la investigación de cualquier índole, pero gracias al internet ésta tarea se facilita enormemente poniendo a nuestro alcance artículos o publicaciones de científicos reconocidos que décadas atrás sólo eran para unos cuantos, así se encuentran las bases teóricas para hacer un buen modelo, con la ventaja de compartir o comparar trabajo.

El tema del procesamiento de digital de señales se ha ido perfilando al lado de algoritmos matemáticos que respaldan la información con bases teóricas bien definidas. Esto mejora los procesos a veces un tanto empíricos de solucionar un problema.

Éste proyecto pretende ser parte de la línea de investigación enfocada a edificios inteligentes, que se trabaja en esta Universidad, y que se enfoca a la aplicación de algoritmos que nos ayuden a facilitar el manejo de aparatos electrónicos o electromotrices dentro de edificios mediante un ente principal que controle y regule los procesos deseados.

El término inteligente se debe a que las máquinas digitales se han transformado de tal forma que un sistema puede ser capaz de distinguir su entorno simulando la forma en que un ser humano percibe su alrededor, pueden tomar decisiones "propias" conforme a las bases con las que fueron programadas y sobre todo se pretende lograr la autosuficiencia para aprender nuevas cosas por si solas.

Es bien conocido que existen programas en el mercado similares al que se desea elaborar, en cuanto al hecho de reconocer palabras. La ventaja de éste, es la posibilidad de trabajar con un sistema especializado que cuente con un diseño accesible para la comunidad en general y así hacer que de manera gradual se introduzcan en nuestra vida diaria este tipo de dispositivos.

El poder interactuar con un sistema mediante la voz como instrumento principal nos da accesibilidad remota, libre del uso de manos y ayuda a manipular dispositivos electrónicos de una forma simple, así el ser humano puede manejar su entorno al mencionar una o varias palabras de comando conocidas.

1.3 Descripción del problema.

El problema principal gira en torno a cómo diseñar un sistema autómata capaz de reconocer palabras específicas de comando que nos ayuden a manipular elementos de un edificio, como lo son puertas, ventanas, iluminación, etc.

Se necesita hacer que la máquina comprenda lo que se le dice por medio de la voz, lograr que el sonido se convierta en variaciones eléctricas y a su vez en números binarios que se procesen como un conjunto de señales que formen una palabra en concreto y que se identifique como tal.

Esto conlleva a pequeñas fracturas en los algoritmos, como los errores que aparecen al trabajar bajo diferentes ambientes o condiciones de entorno, que generan ruido ambiental distinto, para ello se necesita implementar un sistema de segmentación automática. El tiempo de reconocimiento máximo y mínimo con el que se puede procesar una palabra sin que ésta se fragmente o se mezcle con alguna otra palabra. (Adjoudj, 2005) (Chang, 2009).

En cuanto a un sistema mono locutor, las variaciones que se presentan al estar expuesto a distintas emociones, al estrés, o enfermedades, perturban la señal de entrada hasta el punto en que para el sistema la palabra mencionada es irreconocible.

Cuando se experimenta con diferentes personas, surgen diferencias de acento, tonalidad de voz, rapidez, que son los más comunes.

Se deben tener presentes los errores comunes de programación, como el manejo incorrecto de tipos de dato, errores de aproximación, fallas en espacios reservados de memoria.

No se pretende solucionar cada uno de los errores que han surgido durante varias décadas al construir sistemas de reconocimiento, sino de tener un sistema capaz de identificar palabras para realizar una tarea en específico, que sea eficiente y que cuente con una estructura abierta flexible, que facilite la implementación y mejoramiento en un futuro.

1.4 Planteamiento teórico

Hipótesis: *"Es posible desarrollar un sistema capaz de reconocer un número finito de palabras usando la parametrización de Mel en unión con las Redes Neuronales Artificiales".*

1.5 Objetivos

Desarrollar un sistema capaz de reconocer confiablemente al menos veinte comandos de voz para comunicarse con otros dispositivos o personas a través de un enlace inalámbrico de radiofrecuencia (GSM) integrado en el mismo sistema.

Otro objetivo importante es programar mediante un lenguaje estándar, de tal manera que el código puede ser retomado y adaptado para distintas aplicaciones.

1.5.1 Objetivos Específicos:

1) Desarrollar un software que permita evaluar de manera paramétrica los siguientes algoritmos:

a) Transformada Rápida de Fourier.
b) Coeficientes Cepstrales de Mel.
c) Red neuronal perceptron multicapa
d) Calculo de cruces por cero y energía de señal

2) Creación de base de datos de prueba

3) Experimentar para hallar la estructura de red neuronal mas adecuada para reconocimiento.

4) Evaluar la segmentación automática de voz mediante RNA y energía de la señal.

5) Adaptar modulo GSM. Para Poder dar conectividad inalámbrica con otros dispositivos dentro del mismo edificio.

6) Integración, validación y corrección del sistema para obtener tasa de reconocimiento del 80% o superior.

1.6 Metas

Estas se van a ir incrementando conforme la investigación avance y con cada meta básica cumplida surgirá una nueva que mejore el resultado final.

1. Primero se necesita desarrollar el software de los algoritmos necesarios para reconocer un número razonable de palabras comando.

2. Hacer simulaciones bajo distintas condiciones de grabación y mejorar la tasa de reconocimiento.

3. Conectar y adaptar el sistema a un módulo GSM que nos ayudará a comunicarnos con el exterior.

4. Generar base de datos con distintos individuos.

5. Implementar todas las partes antes señaladas para hacer una aplicación conjunta del sistema.

1.7 Resultados Esperados, Posibles Aplicaciones y Uso del Proyecto

Se espera tener un sistema que sea capaz de reconocer por lo menos 15 palabras relacionadas con edificios inteligentes con una probabilidad de error del 3-10%, que pueda trabajar bajo condiciones de ruido y con un tiempo de respuesta aceptable.

El sistema se podrá adaptar a un sin fin de procesos, tales como control de cargas a distancia que ayudaría principalmente a personas que han perdido alguna extremidad a manejar los elementos de un edificio con tan sólo mencionar palabras verbalmente.

Otra posible aplicación es el poder entablar un enlace con otra persona por medio de algún tipo de tecnología de comunicación como lo es el GSM (Global System for Mobile Communications), mandar mensajes previamente grabados en una plantilla, entablar una conversación o controlar dispositivos desde distancias lejanas.

Como trabajo a futuro, se puede implementar un sistema de identificación de voz que mezclado con el reconocedor de palabras puedan dar acceso selectivo a un área en particular, que funjan como medios de seguridad.

Otra posible aplicación futura del reconocimiento de voz puede ser un sistema "guiador electrónico" para personas invidentes, en el cual por medio de voz se menciona una dirección de destino deseada y éste deberá ir guiando mediante señales auditivas la ruta más eficiente a seguir, esto con el fin de que la persona pueda caminar con seguridad en una ciudad desconocida sin temor a perderse.

Capítulo 2:

Estado del Arte.

En este capítulo revisaremos los trabajos relacionados con el propuesto en esta tesis, al final se hará una comparación en base a los resultados obtenidos en los artículos investigados.

Sin duda el reconocimiento automático del habla (*ASR, Automatic Speech Recognition*), ha evolucionado notablemente desde su aparición hace ya más de 50 años y durante ese transcurso han surgido un sin fin de prototipos que aplican técnicas probabilísticas y matemáticas. En primera instancia se verán algunos sistemas y también aportaciones que han ido moldeando esta rama de la tecnología.

Después se describirán brevemente las técnicas modernas usadas en cada artículo, mientras que las empleadas en esta tesis se verán a detalle en el capítulo 3 de este documento. Por ello es necesario dividir el reconocimiento de voz en la actualidad en dos partes principales, la extracción de características y la identificación.

2.1 Evolución del Reconocimiento de voz

En 1952 Davis, Biddulph y Balashek de los Laboratorios Bell desarrollaron un sistema para reconocer dígitos aislados (del 0 al 9) enfocado a un sólo locutor, obteniendo un 98% de efectividad. Cabe mencionar que este sistema se desarrolló análogamente y fue diseñado para un solo locutor. (Davis, 1952).

Durante la década de los 60´s, se comenzó a trabajar con vocabularios pequeños, dependientes del locutor y con palabras de flujo discreto, que es la forma donde se remarcan las pausas entre palabras y frases (Juang, 2004).

Suzuki y Nakata en 1961 de los Laboratorios de Investigación de Radio en Tokio, reconocieron las vocales en japonés mediante un analizador de espectro de 26 canales. (Suzuki, 1961).

Sakai y Doshita en 1962 usaron el primer segmentador de voz para el análisis y reconocimiento de vocales y consonantes, empleando la técnica de cruce por cero y las bandas de frecuencia de energía. La eficiencia fue del 90% para las vocales y 70% para las consonantes (Sakai, 1962).

En 1963 Nagata de los laboratorios *NEC* en Japón, construyó el hardware de un banco de filtros con 8 espectros de bandas con el cual se reconocían 10 dígitos aislados en japonés (Nagata, 1963).

Durante esta década los investigadores que trabajaban con procesamiento de voz comprendieron la complejidad del desarrollo de una verdadera aplicación, es por ello que comenzaron a trabajar con vocabularios pequeños, dependientes de locutor y con palabras de flujo discreto, que es la forma donde se remarcan las pausas entre palabras y frases. (Masanobu 2007).

El sistema de Fry y Denes de la Universidad de Inglaterra fue capaz de reconocer 4 vocales y 9 consonantes, esto gracias a la incorporación de información estadística que les permitió reconocer secuencias de *fonemas en inglés.* (Fry, 1959)

En los 70´s se desarrolló el primer sistema de reconocimiento de voz comercial, el llamado *"VIP-100 System"*, desarrollado por la compañía Threshold Technology, Inc. Uno de los principales usos de éste fue la organización de paquetes en instalaciones de la compañía FedEx.

Fue en este periodo que la ARPA (Advanced Research Projects Agency) del departamento de defensa de los Estados Unidos de América funda el programa de *Speech Understanding Research* (SUR), con una duración de 5 años en los cuales no se cumplieron las metas establecidas, pero se dejaron buenas aportaciones en el tema, como las reglas sofisticadas de fonética. (Juang, 2004).

También se introdujeron las técnicas del *warping*, el modelado probabilístico y el algoritmo de retropropagación al reconocimiento de voz. (Sakoe, 1978).

El sistema desarrollado por Raj Reddy de la Universidad de Carnegie Mellon tuvo el nombre de *"Harpy"*, en el cual se podía usar un vocabulario de 1,011 palabras con un nivel de reconocimiento razonable. Éste se introdujo en el campo del reconocimiento continuo basado en la grabación dinámica de fonemas. (Lowerre, 1990)

El *"DRAGON system"* de Jim Baker que apareció en 1980, cambió el enfoque basado en reconocimiento de patrones para usar métodos de modelo probabilístico como los Modelos Ocultos de Markov.

Assit, una empresa de telecomunicaciones desarrolló el sistema llamado "Sicare Light" para el control de puertas, ventanas, etc. Enfocado principalmente para personas que presentan alguna falta de sus extremidades.

IBM unió esfuerzos con Fred Jelinek para crear *"Voiceactivated typewriter" (VAT)* que tenía como función principal convertir una sentencia dictada en una secuencia de letras y palabras que podían ser mostradas en una pantalla o impresas en papel. El sistema, llamado *Tangora*, incluyó la técnica del modelo de lenguaje el cual establece reglas de estadística gramática o sintáctica. (Jelinek, 1975).

En esta década la idea de Red Neuronal Artificial fue introducida al campo del reconocimiento de voz satisfactoriamente gracias al procesado paralelo distribuido, una forma particular de este procesado fue el perceptrón multicapa. (Lippmann, 1990).

2.2 Introducción a las técnicas modernas de reconocimiento de voz

A partir de los 90´s, las herramientas empleadas se combinan para formar sistemas más eficientes y los algoritmos se van mejorando. También los costos de desarrollo disminuyen y los grandes vocabularios comienzan a ser algo normal. Las aplicaciones independientes de locutor y el flujo continuo (sin pausas significantes entre dictado) comenzaron a ser comunes. Los sistemas actuales hacen énfasis en que se puede dividir la tarea de reconocer la voz mediante 2 principales etapas principales: La extracción de características y la Identificación de la palabra.

2.2.1 Métodos de Extracción de Características

El objetivo principal de la extracción de características es transformar la señal de entrada en otra que contenga menos parámetros pero que sea representativa de la señal original, conteniendo así su forma en esencia y su relación entre elementos de la señal, ya sea en un plano temporal o de frecuencia. Esto ayuda notablemente al desempeño del software, ahorrando así espacio de memoria y tiempo de respuesta del sistema. A continuación se muestra un mapa conceptual, que contiene los principales algoritmos empleados en la actualidad en cuanto a extracción de características:

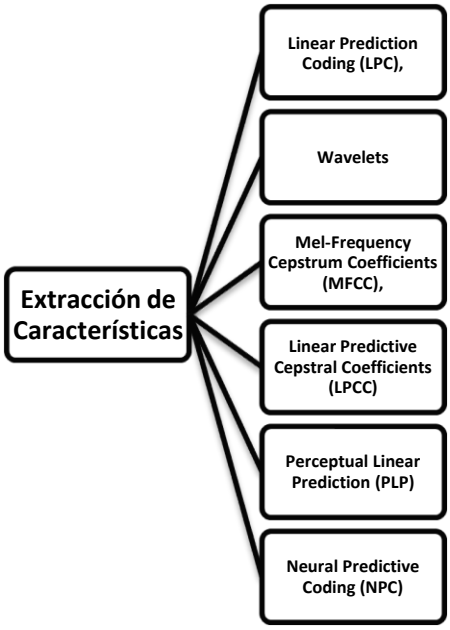


Figura 2.1: Técnicas de extracción de características actuales.

2.2.1.1 Predicción Lineal

La Predicción Lineal (LP) en el reconocimiento de voz, consiste en modelar el tracto vocal como un filtro digital constituido únicamente por polos (respuesta infinita al impulso o IIR), permitiendo así calcular la próxima muestra como una suma ponderada de las muestras pasadas. Este filtro de predicción se traduce en la función de transferencia de la ecuación:

$$H(z) = \frac{G}{1 - \sum_{i=1}^{P} a_i \bullet z^{-i}}$$

Donde *G* es la ganancia del filtro que depende de la naturaleza de la señal (sonora o no sonora). Entonces, dada la señal *s*(*n*), el problema consistirá en determinar los coeficientes de predicción y la ganancia.

Entonces, serán los coeficientes de predicción los que se usarán como parámetros de reconocimiento de palabras. Se han realizado varias modificaciones a este algoritmo para hacer más eficiente el reconocimiento de voz.

2.2.1.2 Transformada Wavelet

El propósito de la Transformada Wavelet (TW) es la descomposición de una señal *x(t)* en una combinación lineal de versiones dilatadas y desplazadas de la función madre $\Psi(t)$, lo cual se denota a través de*:*

$$X(\tau, a) = \frac{1}{\sqrt{a}} \int x(t) \Psi^{\bullet}_{\tau,a}(t) dt$$

$$\Psi^{\bullet}_{\tau,a}(t) = \Psi^{\bullet}\left(\frac{t-\tau}{a}\right)$$

Donde τ corresponde al desplazamiento de la Wavelet madre y *a* es la respectiva escala. Entre los conjuntos de *Wavelets* más usados están la Haar, Morlet, Daubechies y Coifman. Sin embargo, para el reconocimiento de voz se han empleado típicamente la Morlet y la Daubechies.

2.2.2 Métodos de Identificación

Esta parte de reconocimiento de voz, compara una señal de entrada con el conocimiento que tiene de otras señales previamente analizadas, teniendo así un clasificador ó identificador de señales, el cual es capaz de mostrar la similitud que existe entre dicha entrada y cada una de las señales con las que cuenta el sistema.

Los métodos actuales se ven reflejados en el siguiente mapa conceptual:

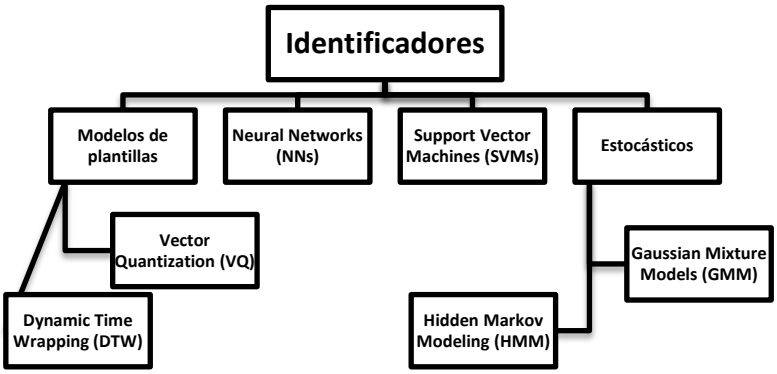


Figura 2.2: Técnicas de identificación actuales.

2.2.2.1 Máquinas de Vectores de Soporte

Las SVM inicialmente fueron desarrolladas por Vapnik y su grupo de colaboradores en los Laboratorios Bell AT&T y presentadas como una novedosa técnica para clasificación de patrones.

Como entrenamiento de clasificadores se emplean Funciones de Base Radial o Funciones Polinomiales, entre otras. Una de las principales ideas detrás de esta técnica es la de separar las clases por medio de una superficie que maximice el margen entre ellas, a diferencia de las técnicas usadas para el entrenamiento de las ANN, las cuales buscan una superficie que separa las clases con el menor número de errores de entrenamiento.

2.2.2.2 Modelos de Mezclas de Gaussianas

Un GMM está compuesto, básicamente, de una superposición de *M funciones de densidad de* probabilidad (*fdp) gaussianas,* donde cada *fdp está* ponderada por un coeficiente de peso.

Por cada clase se estiman los parámetros de los GMM que incluyen los coeficientes de ponderación y las medias y matrices de covarianza de cada fdp gaussiana.

2.2.2.3 Cuantificación Vectorial aplicada

La idea básica de la *Cuantificación Vectorial (VQ*) es la de sustituir un cierto vector de parámetros, obtenido del análisis de un cierto segmento de señal, por un vector similar, llamado *vector código* perteneciente a un diccionario finito y prefijado de vectores. Cada vector código tiene asociado un cierto índice que se convierte en la salida del cuantificador.

2.2.2.4 Modelos ocultos de Márkov

Un modelo oculto de Márkov o HMM es un modelo estadístico en el que se asume que el sistema a modelar es un proceso de Márkov de parámetros desconocidos. El objetivo es determinar los parámetros desconocidos (u *ocultos*, de ahí el nombre) de dicha cadena a partir de los parámetros observables. Los parámetros extraídos se pueden emplear para

llevar a cabo sucesivos análisis, por ejemplo en aplicaciones de reconocimiento de patrones.

2.2.2.5 Alineamiento Temporal Dinámico

Algoritmo para medir la similitud entre dos secuencias que pueden variar en el tiempo o la velocidad.

Las semejanzas en los patrones de voz serían detectadas cuando una persona hable lento, rápido, o incluso si hay aceleraciones y deceleraciones en el transcurso. Este algoritmo nos permite encontrar la coincidencia óptima entre dos secuencias dadas.

2.3 Trabajos Investigados

En esta sección haré un resumen de algunos de los artículos investigados que se enfocan tanto al reconocimiento de voz como al reconocimiento de persona, aplicando las técnicas actuales, como las que se vieron en los cuadros sinópticos anteriores.

Algunos de estos trabajos hacen uso de bases de datos reconocidas internacionalmente como lo son: NIST (*National Institute of Standards and Technology*), TIMIT (Texas Instruments and Massachusetts Institute of Technology) y CMU (Carnegie Mellon University).

La mayoría de los trabajos hacen uso de la comparación entre varios métodos de reconocimiento respecto a un identificador o viceversa.

2.3.1 Reconocimiento de comandos de voz usando la transformada wavelet y máquinas de vectores de soporte

En este artículo se comparan dos algoritmos de identificación de voz, uno es el clasificador mediante redes neuronales artificiales y el otro mediante una máquina de vectores de soporte, en donde la extracción de características es realizada mediante paquetes Wavelet.

Los modelos basados en la transformada discreta wavelet (DWT) usan los paquetes wavelet (PW). Estos esquemas, han sido combinados con clasificadores como las redes neuronales artificiales RNA y análisis de discriminantes lineales, obteniéndose en aplicaciones para el reconocimiento de comandos un porcentaje de acierto del 85% para la base datos NIST y 90% para la base de datos TI64.

Para el caso de las máquinas de vectores de soporte, han sido utilizadas para el reconocimiento de vocales con un rendimiento del 71.72% y 85.13%, ambas usando la base de datos TIMIT.

En cuanto el reconocimiento de fonemas con un rendimiento del 77.6% y al reconocimiento de palabras se han presentado resultados con un rendimiento del 88.4%, con la base de datos OGI alphadigit.

En el experimento llevado a cabo por Marín y su grupo de trabajo, en el cual se construyó una base de datos de 113 muestras de voz de diferentes personas sin importar género, los cuales pronunciaron los dígitos del uno al cinco en el idioma español. Del total de muestras recolectadas, 90% fueron empleadas para entrenar al sistema, mientras que el 10% restante fue para validar.

La segmentación empleada combina la técnica de detección de inicio por medio de un umbral dado y el método de cruces por cero en bloques de 5 milisegundos. Ya detectada la señal, se comienza a dividir en bloques de 64 ms sin traslape.

Se emplearon paquetes wavelet de 24 nodos de descomposición, obteniendo 24 características cada 64 ms, se implementan varias alteraciones en base a este esquema.

Los Parámetros de los sistemas de identificación fueron:

- RNA del tipo Perceptrón multicapa. Cada red cuenta con 5 salidas y fueron entrenadas con el algoritmo *backpropagation con gradiente conjugado* escalado.
- SVM Tiene dos funciones núcleo (kernel):

    a) Tipo RBF (*Radial Basis Functions)*

    b) Funciones polinomiales de orden diferente (poly-n).

Los resultados fueron satisfactorios para la wavelet PW5+cepstro, con db4 y 8, arrojando valores de reconocimiento hasta del 96% para el identificador tipo red neuronal y para las máquinas de vectores con RBF (Marín, 2006).

2.3.2 Artificial neural network & mel-frequency cepstrum coefficients-based speaker recognition

En este artículo se habla de cómo se emplearon las dos técnicas principales en esta tesis, como lo son los coeficientes en escala de Mel y las redes neuronales artificiales, pero en este caso se enfocan a la identificación de locutor y no al reconocimiento de voz como tal.

Se empleó una red neuronal con el algoritmo de "backpropagation", se segmenta la señal en tramas entre los 5 y los 100mseg, en el cual se caracteriza la señal con 160 valores, estos sirven a su vez como entrada a la red neuronal, la cual tiene una capa oculta de 150 neuronas.

La red fue entrenada tanto en condiciones ideales como en condiciones de ruido inducido. También se hace uso del algoritmo LGB ¨Diseño de Cuantización de vector" como medio de comparación contra la red neuronal.

La base de datos empleada fue propia, la cual incluye 294 señales de voz distintas de 142 sujetos diferentes. También se usaron las bases de datos ASR y CMU. El mejor algoritmo de identificación de la base ASR fue el LGB con una taza de reconocimiento del 80.01%, para el caso de la base de dados propia, el mismo algoritmo se desempeñó mejor con un 85.74%, pero en el caso de los datos contenidos en la CMU fue muy notable el desempeño, ya que con redes neuronales se alcanzó un 90.66%, mientras que el algoritmo LGB un 69.33% (Adjoudj, 2005).

2.3.3 Selección de características usando HMM para la identificación de patologías de voz

Este artículo es interesante ya que compara los métodos de extracción de características de análisis de componentes principales y análisis discriminante lineal. Los cuales tuvieron un rendimiento del 76.25% y 91.45% respectivamente, las muestras de patología fueron sobre de labio y paladar hendido, ya con las características obtenidas se emplean los Modelos Ocultos de Markov, para la identificación de patología.

Tomaron una primera base de datos de 160 muestras de la vocal sostenida, /a/, pronunciada por 80 niños con labio leporino y paladar hendido. Y otra base de datos conformada por 320 muestras de la misma vocal pero por 160 pacientes con voz normal y 160 con pacientes que presenten una de las patologías antes mencionadas.

En la etapa de identificación se contrastaron los HMM con los modelos de mezclas gaussianas GMM.

Los resultados obtenidos fueron que la extracción de análisis discrimínate es un buen método empleando los HHM como identificador, ya que de otra forma los resultados no fueron lo suficientemente buenos (Álvarez, 2004).

2.3.4 Implementación de un reconocedor de palabras aisladas dependiente del locutor

La codificación se lleva a cabo mediante las técnicas de Predicción Lineal y Cepstrum real, mientras que la etapa de clasificación se realiza mediante el alineamiento temporal dinámico (DTW), que permite independencia del intervalo de tiempo de cada muestra de voz.

Implementaron la etapa de reconocimiento de palabras usando el algoritmo de alineamiento temporal dinámico (DTW), el cual es capaz de discriminar entre palabras con duración temporal independientes de la señal.

Para efectos del trabajo, la adquisición se desarrolló con una frecuencia de muestreo *fs* = 11025 Hz, una cuantificación de 16 bits y calidad de sonido *mono estéreo*.

Antes de entrar a la etapa de extracción de características, la señal de voz se segmenta a intervalos de 20 a 30 ms, tiempo durante el cual la señal se considera casi estacionaria.

El sistema de reconocimiento fue sometido a prueba para un conjunto de 10 palabras. Se escogió a modo de ejemplo un vocabulario compuesto por los 10 dígitos. Los experimentos se llevaron a cabo pronunciando 20 veces cada una de las palabras. Cabe destacar que las pruebas se realizaron por un solo locutor y en condiciones de ausencia de ruido de fondo.

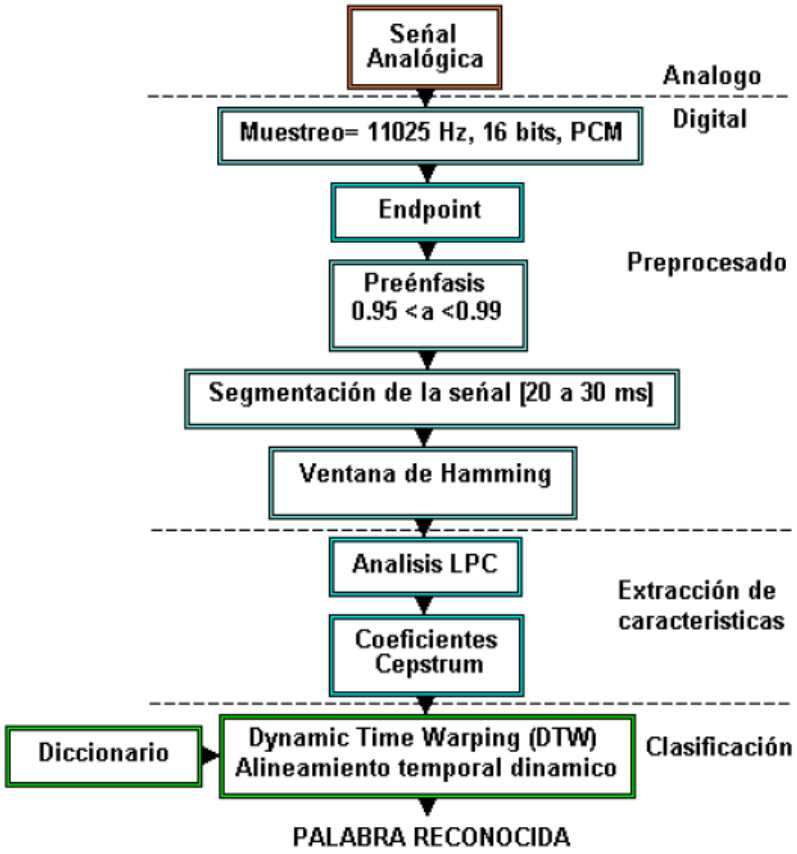


Figura 2.3: Procedimiento del reconocedor de palabras aisladas.

Los resultados obtenidos demuestran que el uso de estas técnicas permite obtener un 85% de clasificación correcta (San Martín, 2004).

2.3.5 An automatic speaker recognition system

Este artículo usa los MFCC para la extracción de características, el resultado de ésta caracterización se introduce a un proceso de Vector de Cuantización para su identificación.

Se hace la comparación con tres diferentes tipos de fuente emisora de sonido: 1) música, 2) Parlante en inglés y 3) Parlante en bengalí. Se logra un rango de reconocimiento es cercano al 90%.

Se eligieron los VQ por no tener un desarrollo computacional complejo, al final se compara con las redes neuronales artificiales y los modelos ocultos de Markov, haciendo solamente uso de los datos obtenidos por referencias (Chakraborty, 2008).

El universo de muestras fue de 70 archivos. El diagrama del proceso es el siguiente:

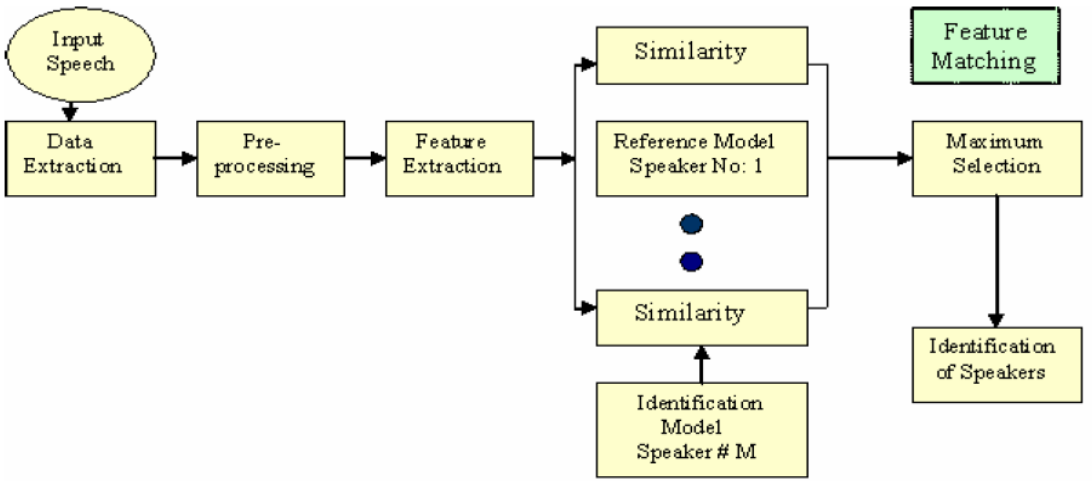


Figura 2.4: Diagrama de bloques del sistema de reconocimiento de persona.

2.3.6 Speaker identification based on the frame linear predictive coding spectrum technique

En este artículo se hace una variación de extracción de características mediante codificación predictiva lineal llamada frame linear predictive coding spectrum, (FLPCS), la cual añade información como lo es cepstrum y lo aplica a tramas de tiempo.

En la etapa de identificación se emplea la red neuronal de regresión general (general regression neural network, GRNN), que es una variante de las redes neuronales artificiales. Por otro lado, se emplean los modelos de Mezclas Gaussianas GMM. Estos dos métodos nos dan un tiempo de entrenamiento y de identificación reducido.

La ventaja del modo texto-dependiente es que al usarse no se necesitan grandes sentencias, en comparación con el modo texto-independiente, en el cual las sentencias cortas pueden incrementar la velocidad de clasificación.

La frecuencia de muestreo empleada es de 16KHz para una gama sonidos con frecuencias de voz de entre 0 y 8KHz.

La base de datos comprende 50 locutores, 25 hombres y 25 mujeres, cada uno de los cuales repitió 50 veces una de 5 sentencias diferentes.

Las sentencias de entrada fueron divididas en ventanas tipo haming de 20ms, traslapadas cada 12.5ms.

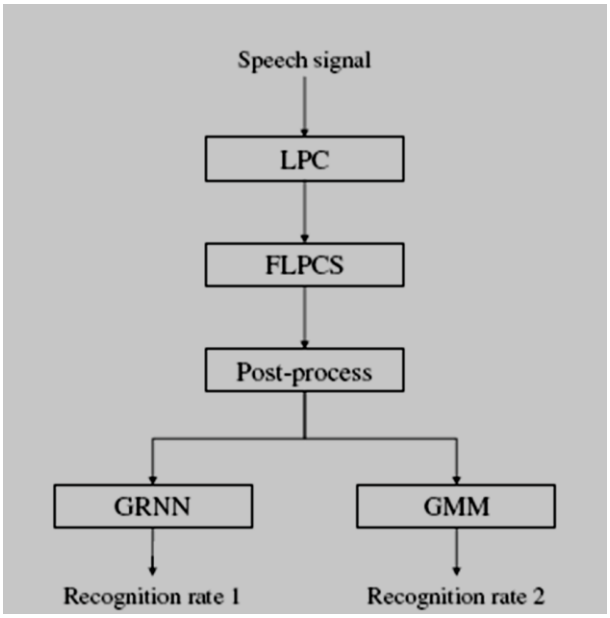


Figura 2.5: Diagrama de flujo del análisis del habla.

Los resultados obtenidos fueron favorables para el identificador GMM, quien tuvo un rango de reconocimiento desde el 85% para un orden de LPC de 13, hasta un 99.2% para un LPC de orden 50. Estos resultados opacan los obtenidos por la GRNN que obtuvo un valor mínimo de reconocimiento del 77.2% y un máximo de 91.6%. Se tomaron los valores más altos y bajos de todas las sentencias mencionadas.

Concluyendo que para la extracción por medio de FLPCS, el método de GMM es más veloz y más eficiente, en cuanto a reconocimiento, que el GRNN. (Jian-Da, 2009)

2.3.7 Classification of audio signals using SVM and RBFNN

En este artículo se propusieron algoritmos para la clasificación de 6 diferentes tipos de audio: música, noticias, deportes, comerciales, caricaturas y películas.

La señal se caracteriza mediante coeficientes Cepstrales predictivos lineales y coeficientes Cepstrales en escala de Mel. Para clasificar se hace uso de las máquinas de vectores de soporte contrastado con el método extendido RBFNN de las redes neuronales.

Usaron 100 comerciales en diferentes lenguajes, 100 archivos musicales, 100 caricaturas, 100 cortos de película, 100 cortos de deportes y 100 fragmentos de noticias en inglés y Tamil. La duración de estos varía entre 1 y 10 segundos.

La frecuencia de muestreo es de 8 Khz a 16 bits, usando tramas de 20ms con 10ms de traslape.

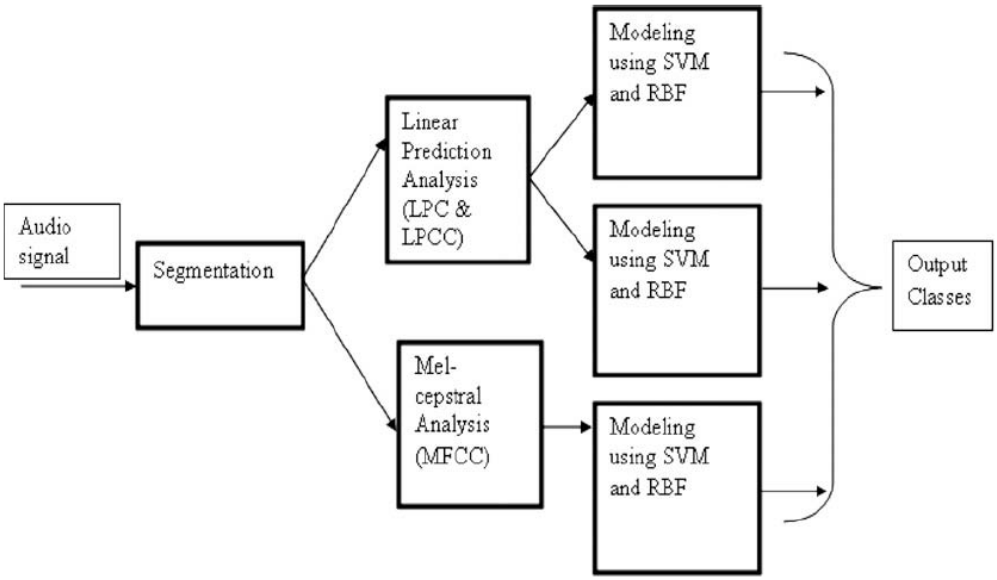


Figura 2.6: Diagrama de bloques del clasificador de audio.

Los resultados experimentales muestran que la tasa de reconocimiento de los vectores de soporte es del 92% y del algoritmo mejorado de redes neuronales es del 93%, teniendo un mejor desempeño con la extracción de características mediante coeficientes Cepstrales de Mel (Dhanalakshmi, 2009).

2.3.8 Aplicación de RNA y HMM a la verificación automática de locutor

El enfoque de este trabajo fue el reconocer al locutor para fines de seguridad. Menciona que se ha evolucionado de usar simples operaciones manuales, hasta emplear el uso de la biometría que incluye huellas dactilares, iris, rostro, etc. En este caso se detectó una contraseña de 3 dígitos del cero al nueve, aunado a la identificación de persona, para reforzar la seguridad.

Usa el modo de "text-prompted" o dependiente de texto, se requirió un identificador de dígitos basado en Modelos Ocultos de Markov, y en el caso del verificador de locutor una red neuronal.

Después de ser detectada la señal se segmenta en periodos de 23ms con una ventana tipo haming para extraer 16 coeficientes Cepstrales de Mel.

La red es del tipo perceptrón multicapa con 32 capas ocultas, y 16 neuronas de entrada y una de salida. La función empleada fue la de tangente hiperbólica sigmoidea. Se entrenó con el método de "back propagation".

La cadena de Markov empleada cuenta con un diccionario de 128 vectores creado con el algoritmo de *Linde-Buzo-Gray* (LGB), la cual contiene 5 estados y topología izquierda-derecha; cada estado tiene un vector de probabilidad discreta asociado con 128 símbolos.

Su sistema en forma de bloque se muestra en la siguiente figura:

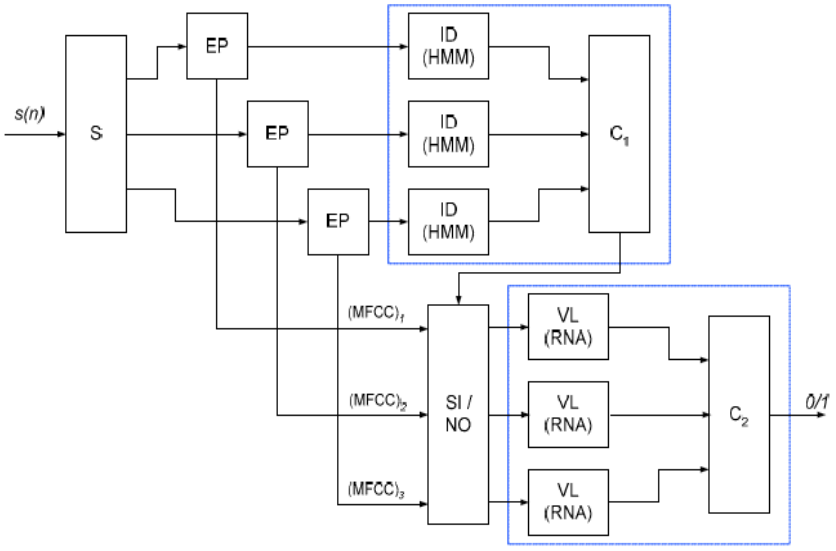


Figura 2.7: Flujo de la señal en el sistema de verificación de locutor.

Dónde:

(S)	Segmentación
(EP)	Extracción de parámetros
(ID)	Identificación de dígitos
(VL)	Verificación de locutor
C1 y C2	Comparadores

El sistema como tal alcanzó un rendimiento cercano al 95%. La base de datos fue de 1800 pronunciaciones de 20 locutores diferentes (Alegre, 2007).

2.4 Evolución de las Telecomunicaciones.

La tecnología de telefonía móvil representa un aspecto importante en nuestra vida cotidiana tanto en negocios como en asuntos personales, el acceso instantáneo a personas donde quiera que ellos estén es ahora parte de nuestra cultura de comunicación, de hecho para ciertas personas se genera una dependencia tal que si su dispositivo falla se sientes ansiosos e incomunicados.

Después de la invención del telégrafo un número considerable de personas trabajó en transmitir sonido sobre cables, en 1857 el Italiano-Americano Antonio Meucci desarrollo el primer sistema de teléfono primitivo, pero a falta de financiamiento no se pudo dar la difusión adecuada y por esa razón tradicionalmente el reconocimiento como inventor del teléfono se le da a Alejandro Graham Bell, quien concibió esta idea el verano de 1874 pero la primera frase transmitida por este medio se da en 1876 con la frase "Mr Watson, come here, I want you", ya que Bell había derramado algo de acido sobre su ropa y necesitaba asistencia.

Cuando el sistema del teléfono se estableció, el siguiente mayor desarrollo en esta área fue la tecnología Wireless (sin cables) ó radio, James Clerk Maxwell fue el primero en deducir matemáticamente la existencia de ondas electromagnéticas. Y el italiano Marconi exploró y aplicó las ondas hercianas para usos de comunicación, demostrando en 1901 comunicaciones a grandes distancias transmitiendo una señal a través del atlántico.

El siguiente gran paso fue la invención del transistor con el cual se permitieron circuitos electrónicos más pequeños y portables, haciendo posible los primeros "Walkie-takies" en EUA por Motorola en 1940 los cuales aunque pesados (16Kg) permitían la comunicación entre militares en continuo movimiento. Lo que impulso a las grandes compañías de comunicaciones como Bell a obtener el primer celular comercial en mayo de 1978 en Bahrain, esta primera fase tuvo 2 células y 250 suscriptores.

Rápidamente otras empresas como Advanced Mobile Service (AMPS) empezaron la comercialización en Chicago. Pero la venta masiva se da hasta 1983 en EUA siendo el "Nordic Mobile Telephone "(NMT) el primer sistema en venta. Teniendo una banda de frecuencias de 450MHz.

Naturalmente la primera telefonía celular se da de forma análoga, pero la iniciativa "Groupe Speciale Mobile" que después cambiaría su nombre a "Global System for Mobile Communications" de sus siglas GSM estuvieron entusiasmado por trabajar con sistemas digitales y estándares. Con el GSM establecido fue posible enviar mensajes de texto SMS (short Message Service). Dando pie a la nueva era de comunicaciones, donde para 2004

más de 45 billones de mensajes fueron enviados cada mes y para febrero del mismo año un billón de suscriptores GSM conectados.

Alrededor de 1990 la industria celular fue todo un éxito y analistas descubrieron que la gente deseaba usar mas allá sus equipos especialmente para los servicio de envío-recepción de datos al ver el significante crecimiento del uso del internet. El primer paso fue el conocido como "General Packet Radio System"(GPRS), que junto con el "Enhanced Data rates for Global Evolution" (EDGE) fueron nombrados la segunda generación de sistemas de este tipo, que tiene como objetivo el dar soporte a la actividad de datos.

Pero al tener un ancho de banda limitado surgen los sistemas 3G en Europa conocidos como Universal Mobile Telecommunications System (UMTS), usando un ancho de banda CDMA (W-CDMA), esto nos ayuda a tener sistemas con mayor resolución en el significado de datos de voz y video, permitiendo así las video llamadas y un sin fin de aplicaciones y navegación en tiempo real.

2.4.1 Breve Descripción de la Red GSM.

El Sistema Global para las comunicaciones móviles (GSM) es un sistema estándar de comunicación inalámbrica. Por medio de esta red es posible el intercambio de información, principalmente de equipos móviles. Con ello es posible enviar o recibir tanto Voz, Datos y mensajería SMS.

La arquitectura del sistema GSM ha demostrado ser muy exitosa, muchos de los nombres y las ideas han sido adoptados por otros sistemas. Sus conceptos básicos también se han incorporado en los nuevos sistemas como UMTS / W-CDMA 3G.

Los elementos principales del sistema son la Estación Transceiver Base (Base Transceiver Station, BTS), la Estación de Control Base (Base Station Controller, BSC), el Centro de conmutación móvil (Mobile Switching Centre, MSC) y las áreas de registro y autenticación (Figura 2.8). Estos incluyen el Registro de Ubicación Inicial (Home Location Register, HLR), el Registro de Ubicación Visitante (Visitor Location Register, VLR), el Registro de identificación del Equipo (Equipment Identity Register, EIR) y el Centro de autenticación (Authentication Center, AUC).

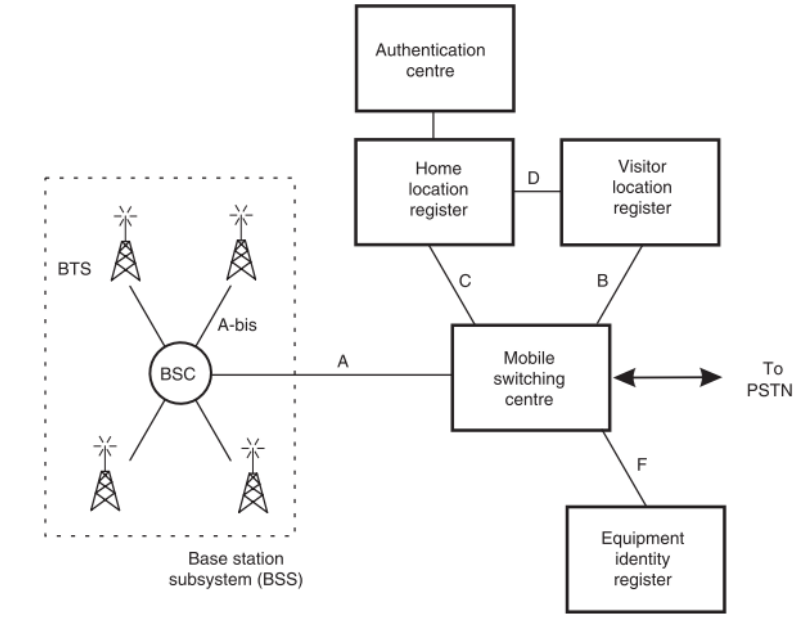


Figure 6.1 The GSM network configuration.

Figura 2.8: Red GSM.

El BTS es el área principal de comunicación con los móviles. El BTS transmite y recibe las señales del manejo de protocolos de interfaz. El BTS se vincula a un BSC, que controla un pequeño grupo de BTS. Estos dos están vinculados mediante una interfaz conocida como la interfaz A-bis. Como todas las otras interfaces entre los elementos de la red GSM, están rigurosamente definidas, para permitir el uso a equipos de diferentes fabricantes.

El BSC gestiona una o más BTS. Maneja la puesta en marcha del canal de radio, salto de frecuencia control y traspasos. Se vincula con el MSC o el centro de conmutación móvil a través de una interfaz denominada A. Otra tarea que normalmente maneja el BSC es convertir los datos de voz de 13 kbps utilizadas en enlaces de radio estándar a 64 kbps utilizadas por la PSTN.

El núcleo de la red es el sub-sistema MSC, o centro de conmutación móvil, que actúa como un nodo de conmutación de la PSTN. Además de esto, las interfaces con AuC para proporcionar autenticación y permitiendo a los usuarios entrar en la red. También interactúa con el HLR y VLR para proporcionar información de localización para la red, por ello las llamadas pueden ser dirigidas a la BTS correcta, incluidas las que posiblemente tengan que enviarse a los móviles que están con roaming. Además de esto las coordenadas de entrega.

Hay varios identificadores diferentes que se incorporan en el GSM estándar para proporcionar flexibilidad sin dejar de mantener el nivel de seguridad requerido. Un número

de estos identificadores se almacenan en una tarjeta, conocida como “Subscriber Identity Module” (tarjeta SIM). Esta pequeña tarjeta de memoria se inserta en un móvil para proporcionar información sobre el suscriptor. Las tarjetas SIM también contienen otra información, incluida la agenda telefónica, esto permite al usuario cambiar de equipo móvil sin perder su registro de teléfono y contactos. Otros identificadores se almacenan en el equipo móvil propio, o en la red.

2.5 Conclusiones

Para la mayoría de los casos el identificador del tipo Red Neuronal Artificial expone resultados mayores al 90%, pero mucho depende de las condiciones de creación de la red para poder elevar el porcentaje de acierto.

También es cierto que para vocabularios más grandes, los identificadores van perdiendo tasa de reconocimiento. Esto complica la tarea de tener un diccionario completo para reconocer.

Las condiciones de validación de los resultados son necesarias ya que en algunos artículos muestran tazas de reconocimiento demasiado elevadas, probablemente porque el universo de muestras de pruebas fue muy corto, como en el caso del primer artículo mencionado, en el cual solo se valida con un 10%.

Sin duda, la extracción de características mediante el cepstrum es de lo más usado, ya sea en forma de coeficientes de Mel o algoritmos modificados de wavelets o predicciones lineales.

Capítulo 3

Metodología.

En este capítulo se relatará el proceso necesario para desarrollar el sistema de reconocimiento automático de voz, se dará a conocer el material empleado y las técnicas que nos permiten llevar a cabo el objetivo de la tesis. El proceso de desarrollo consta de varios módulos validados a través de simulaciones como lo son: la segmentación automática de voz, la Transformada Rápida de Fourier, la obtención de los Coeficientes Cepstrales de Mel y la Red Neuronal Artificial, que son los pilares del sistema. Varios de estos algoritmos se dividen en sub rutinas.

3.1 Recursos Materiales y Humanos

Las simulaciones fueron realizadas en una computadora portátil con 4Gb de memoria RAM, trabajando a 2.1 GHz y con una tarjeta de sonido Conexant High Definition Audio Cx20549 @ nVIDIA nForce 430. El micrófono empleado tiene las siguientes características: apto para frecuencias entre 50Hz y 44.1KHz. La programación de las simulaciones se realizó con el uso de MatLab R2008a con un enfoque de programación en lenguaje c para la compatibilidad de un posible cambio de plataforma o dispositivo programable.

3.2 Descripción del Proceso

La investigación de los métodos y algoritmos toma una parte importante en el proceso ya que hay que tener bien definido lo que se va a hacer, lo que se puede modificar y las herramientas que quedan en un segundo plano y que se pueden emplear en proyectos alternos o comparativos de funcionamiento.

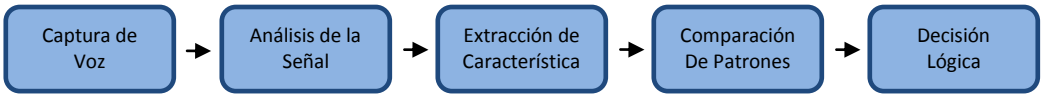


Figura 3.1.- Diagrama General de un Sistema de Reconocimiento de Palabras.

La mayor parte de los sistemas de reconocimiento de reconocimiento siguen la estructura del diagrama de bloques 3.1, donde la captura de voz siempre se hace a través de un micrófono, aunque existen técnicas híbridas como la mostrada por (Chen, 2009) en donde se hace uso de una video cámara para leer los labios y complementar el reconocimiento de la voz.

En el análisis de la señal se pueden encontrar filtros que mejoren la calidad o que la adapten a cierto modelo requerido, ésta manipulación va desde modificar la amplitud, el tiempo o desfasamiento, por ejemplo, estas transformaciones sirven para moldear la señal y adaptarla al siguiente paso de procesamiento.

La extracción de características es necesaria para reducir el número de datos que entran a los algoritmos de comparación con el fin de reducir el tiempo de procesamiento, pero se debe de cumplir el principio de representar la esencia de los datos de entrada.

La comparación de Patrones se encarga de identificar entre un universo finito de patrones previamente conocidos por el sistema, cual se parece más a la que hay a la entrada para así poder identificarlo.

La decisión lógica es el fin para el cual es diseñado el sistema, este puede ser para controlar cargas, para tomar notas o manipular algún aparato, las formas de uso más empleadas en la actualidad. Aunque existen un sin número de posibles aplicaciones enfocadas primordialmente a mejorar la interacción del ser humano con las máquinas complejas.

La metodología llevada a cabo en esta tesis muestra la estructura típica de un sistema de reconocimiento de palabras como se observa en la figura 3.2.

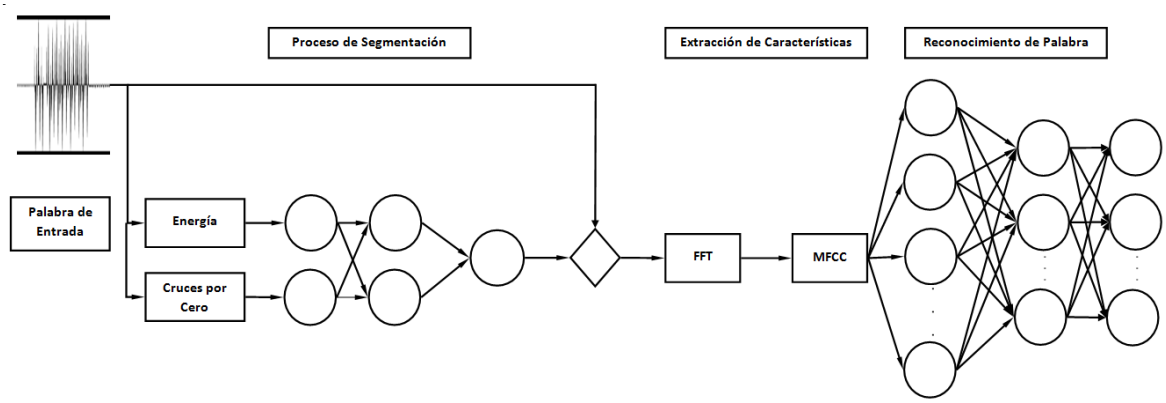


Figura 3.2.- Diagrama a bloques del proceso.

La captura de la señal de voz se hace a través de un micrófono. Se realiza un procedimiento de segmentación automática de voz, el cual nos permite eliminar silencios ó sonidos extraños al sistema para así poder procesar solo el espectro que representa una palabra, para ello se hace uso de los cálculos de energía y los cruces por cero como entradas a una red neuronal artificial encargada de identificar si existe un espectro de palabra. La extracción de características se realiza en el plano frecuencial y se adapta a un modelo que asimila a como el ser humano percibe los sonidos. Para identificar la palabra de entrada se hace uso de una red neuronal artificial que tiene como entrada el vector que caracteriza una palabra de entrada y como salida una neurona por palabra que se activa al momento de detectar similitud con los datos de entrenamientos.

3.3 Captura de Voz

El mecanismo encargado de realizar la percepción es de la voz es el transductor electroacústico, nosotros lo conocemos cotidianamente como micrófono, el cual convierte las ondas sonoras, como las emitidas por el aparato fonador humano, en señales eléctricas que pueden discretizarse a través de un sistema de adquisición, que tiene dentro un convertidor analógico digital para su representación y manejo en lenguaje máquina.

El micrófono sin duda desempeña un factor importante ya que dependiendo de su configuración y diseño puede atenuar el ruido externo ó amplificar la señal de voz de entrada, mientras más alta sea la fidelidad o relación señal a ruido (Signal to noise ratio, SNR), el precio se eleva notablemente pero se eliminan interferencias que puedan afectar la respuesta del sistema.

Se requiere un sistema de adquisición de datos para poder manipular la señal en el marco digital, la tarjeta de sonido de la computadora toma este papel y por ser un elemento integrado desde hace muchos años en las computadoras se tiene la posibilidad de poder configurar la frecuencia de muestreo y el tamaño de la palabra (número de bits por dato).

Se necesita generar un banco de pruebas para poder almacenar las distintas palabras determinado número de veces, es necesario al trabajar con redes neuronales artificiales ya que estas se tienen que entrenar previamente.

3.3.1 Bancos de Pruebas

En primera instancia se creó una base de datos con la cual trabajar, esto es grabar ciertas veces cada palabra deseada para así poder tener una referencia con la cual se pueda entrenar la red neuronal. En la primer etapa de viabilidad del proyecto se hicieron muestras de adquisición empleando la grabadora de sonidos de Windows Xp con una frecuencia de muestreo de 8khz, esto debido a que las frecuencias mínimas para descifrar un mensaje de voz humana es de 4KHz y para cumplir con el teorema de muestreo de Nyquist se tiene que tomar una frecuencia de por lo menos el doble de la frecuencia máxima de entrada. El formato digital es de 7 bits más signo. El micrófono empleado fue de marca Steren, bajo las siguientes características: apto para frecuencias entre 50Hz y 16KHz, con una sensibilidad de -60dB – 3dB. Éste micrófono no tiene una ganancia señal/ruido eficiente, por lo cual se pudo observar la robustez de la RNA.

Con este banco de prueba se entrena y simula la red neuronal artificial y se obtiene su respuesta.

Figura 3.3.- Grabadora de Sonidos de Windows.

Se crearon 2 distintos bancos de pruebas para observar el comportamiento de la red neuronal ante una palabra parametrizada mediante los Coeficientes Cepstrales de Mel.

Banco 1:

{‘gato’, ’carro’, ’hola’}
3 Distintas palabras pronunciadas
8 Grabaciones por Palabra para entrenar la RNA
8 Grabaciones por Palabra para comprobar la RNA

Banco 2:

{‘abre’, ’apaga’, ’cierra’, ’enciende’, ’lámpara’, ’puerta’ y ’ventana’}
7 Distintas palabras pronunciadas
10 Grabaciones para fines de entrenamiento de la RNA
10 Grabaciones para fines de comprobar la respuesta de la RNA.

Pero al hacer uso de la grabadora de sonidos de Windows, la adquisición de señales de voz resultaba tediosa y era poco flexible en ciertos aspectos de grabación. Por eso se creó una interfaz gráfica de usuario en MatLab que permite la creación de bases de datos (BD) tan flexible como las funciones nos lo permiten y con la facilidad de guardar en un solo archivo toda la información necesaria.

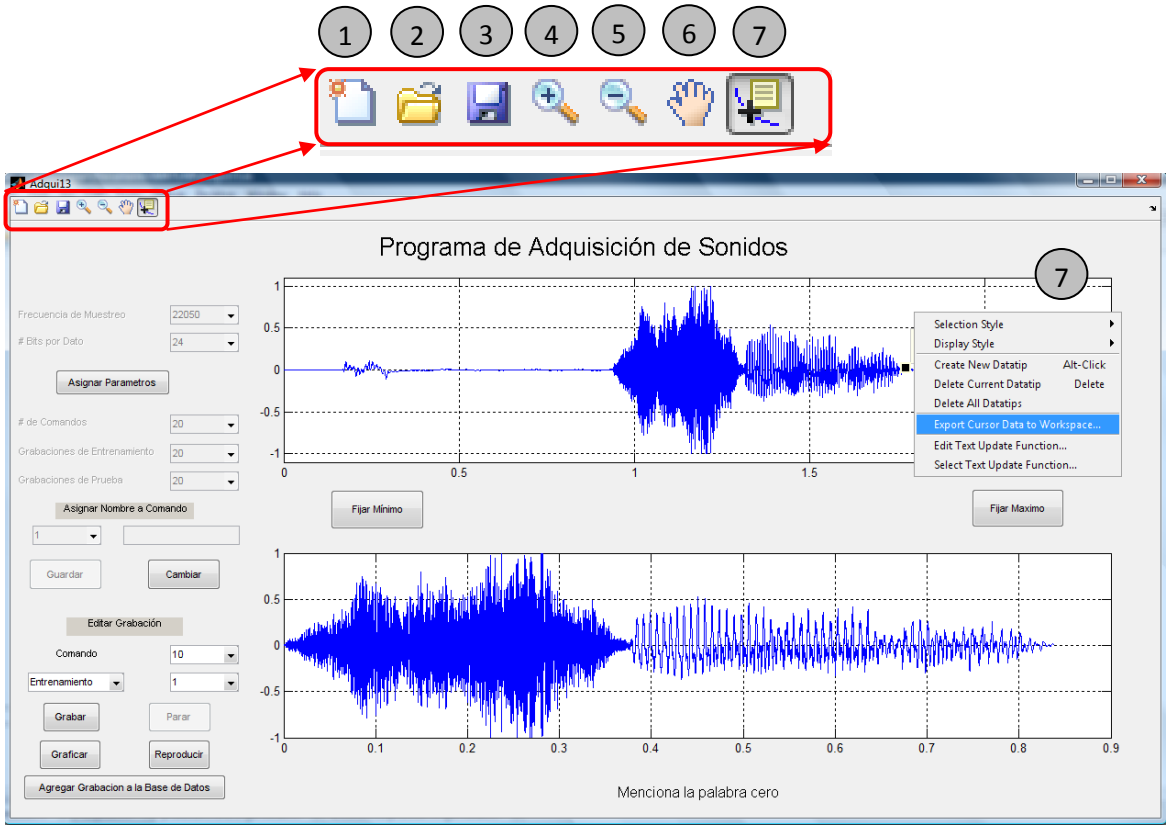
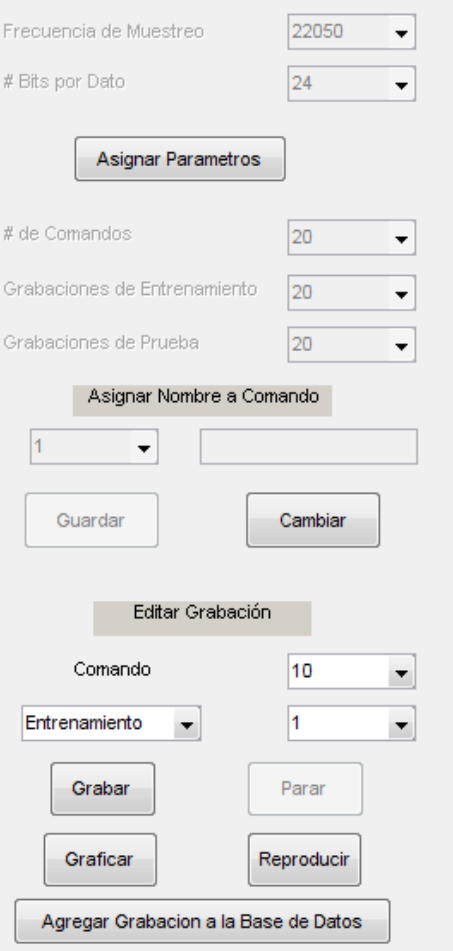


Figura 3.4.- Interfaz Gráfica de Usuario para la Adquisición de Sonidos.

En la figura 3.1 se puede apreciar la interfaz gráfica de usuario en el menú que sobresale se puede notar:

1. Crear nueva base de datos, a través de este botón podemos crear una nueva base de datos.
2. Abrir base de datos, es muy útil ya que en bases de datos muy extensas es muy difícil llevar a cabo todas las grabaciones por eso se puede dividir en varias sesiones de toma de datos sin pérdida de información.
3. Guardar Archivo, nos permite ir guardando modificaciones a la base de datos.
4. Acercar, nos ayuda a analizar de cerca una señal grabada.
5. Alejar, por medio de esta se puede regresar a la vista de la gráfica original después de haber acercado.
6. Mover, ayuda desplazarse en los ejes x y
7. Dato del Cursor, Nos da la posición en x y de un punto fijado.
8. Exportar dato al espacio de trabajo, con ayuda de este podemos exportar el dato para poder manipularlo o también si después de exportar el dato presionamos "fijar mínimo" o "fijar máximo" podemos segmentar manualmente la palabra.

La barra que está situada a la derecha de la interfaz gráfica es de configuración y selección:



Selecciona entre las Frecuencias de Grabación Disponibles.

Selecciona el Tamaño de dato.

Fija el número de comandos que tiene la BD.

Fija el número de Grabaciones de entrenamiento

Fija el número de Grabaciones de prueba

Se puede Asignar un nombre a cada comando por entrenar.

Esta parte es de navegación de la BD podemos seleccionar el comando, la grabación y separar los datos de entrenamiento y prueba. Para grabar la palabra solamente se ubica el lugar de destino, se presiona el botón "Grabar", se menciona la palabra dirigiéndose al micrófono y se presiona el botón "Parar" para finalizar La Grabación. Posteriormente se puede Graficar Agregar a la BD o Reproducir.

Figura 3.5.- Configuración y navegación de la GUI.

3.4 Segmentación Automática de Palabras

El segmentar la voz es una práctica habitual en cualquiera de sus modalidades ya sea sementando palabras, sílabas o fonemas esto nos ayuda a separar únicamente la información de interés a analizar.

El reconocimiento automático de voz es un problema que aún no se encuentra completamente resuelto debido a la variabilidad de la voz, como ejemplo al mencionar dos veces la misma palabra los espectros de la señal no son los mismos difieren tanto en amplitud como en relación temporal. Los algoritmos de reconocimiento automático de palabras aisladas requieren una segmentación eficaz para obtener buenos resultados.

Uno de los métodos actuales más empleados para segmentar es el llamado algoritmo de Viterbi el cual detecta una secuencia de salida en tiempo discreto usando la probabilidad de máquinas de estado finitas, el algoritmo está fuertemente ligado a la técnica de los Modelos Ocultos de Markov y es aplicado a la segmentación de fonemas (Forney, 2005; Toledano, 2003).

El objetivo de la segmentación es el de aislar sólo el contenido de información de entrada que pertenezca exclusivamente a la señal de voz, así en un sistema de tiempo continuo se puede descartar el ruido ambiental y los momentos de silencio disminuyendo la memoria requerida y también la complejidad del cómputo. Cuanto más preciso sea el proceso de segmentación, mejor será la futura caracterización y por consiguiente aumenta la probabilidad de acierto del conjunto de encargado de reconocer la voz (Ching, 1999). Gráficamente se puede apreciar mediante la siguiente figura.

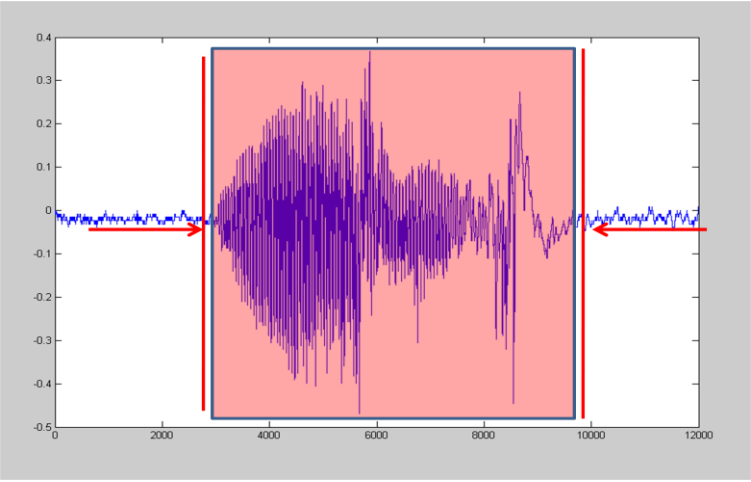


Figura 3.6.- Segmentación de Voz.

La importancia de poder hacer este procedimiento automáticamente radica en la flexibilidad de desarrollo, entrenamiento y la aplicación del sistema. A continuación se describirán los algoritmos empleados para la segmentación seguido de la descripción del procedimiento.

3.4.1 Energía y Detección de Cruces por Cero

Varios métodos han surgido para resolver el problema de la segmentación, uno de los cuales es el de umbralización que fija un valor de amplitud de señal para determinar si comienza o termina una palabra, sin embargo carece de una fundamentación de análisis de la señal, lo cual hace este método muy susceptible al ruido ambiental y tiene un mal desempeño en condiciones ajenas a las de entrenamiento. Por ello se eligieron técnicas que involucran el determinar el comportamiento de la voz.

Los términos de bandas de Energía y cruces por cero se introdujeron al área del reconocimiento de voz en 1961 por Sakai y Toshiyuki, que emplearon estas técnicas para hacer el primer segmentador de voz con una eficiencia del 90% para vocales y 70% para las consonantes, este análisis fue aplicado a una máquina de escribir activada por voz (Sakai, 1962).

La voz cuenta con 2 diferentes tipos de sonido, los sonoros (voiced) y los sordos (unvoiced) (Priyabrata, 2009), los primeros son producidos al forzar el aire a través de la glotis con la tensión de las cuerdas vocales ajustadas de tal forma que puedan vibrar en una oscilación relajada, mientras que los sonidos sordos son generados por la formación de una constricción en el mismo punto en el tracto vocal y forzando el aire a través de la constricción a una velocidad lo suficientemente alta para producir turbulencia. Por estas definiciones se sabe que los sonidos sonoros tienen una mayor energía y menor taza de cruces por cero comparados con los sonidos sordos (Gyucheol, 2001).

En una señal continua, la Energía total E en el intervalo de tiempo t1 a t2 está definida como:

$$E = \int_{t_1}^{t_2} |x(t)|^2 \, dt \qquad (1)$$

Para el caso de las señales discretas donde N es el número de muestras de la señal, la energía se define por:

$$E = \sum_{m=0}^{N-1} x(m)^2 \tag{2}$$

Los cruces por cero indican el número de veces que una señal continua toma el valor de cero. Para las señales discretas, un cruce por cero ocurre cuando dos muestras consecutivas difieren de signo, o bien una muestra toma el valor de cero.

$$z = \sum_{m=0}^{N-1} \left| sign\left[x(m)\right] - sign\left[x(m-1)\right] \right| \tag{3}$$

Donde *sign* es la función signo.

$$sign[x[n]] = \begin{cases} 1 & x[n] \geq 0 \\ -1 & x[n] < 0 \end{cases} \tag{4}$$

La técnica propuesta para lograr esta acometida es la de las redes neuronales, que tiene como entrada la energía y los cruces por cero de determinado número de muestras, este número debe de ser menor que el de la ventana empleada para aplicar el análisis de la señal, y su salida es lo más cercana al valor de '1' si es una posible palabra y cercana al '0' si es una señal de silencio relativo.

3.4.2 Redes Neuronales Artificiales

Las Redes Neuronales Artificiales (RNA) son sistemas de procesamiento de la información, cuya estructura y funcionamiento están inspirados en las redes neuronales biológicas (Metin, 2007). Consisten en un gran número de elementos simples de procesamiento llamados nodos o neuronas que están organizados en capas. Cada neurona está conectada con otras neuronas mediante enlaces de comunicación, cada uno de los cuales tiene un peso asociado. En los pesos se encuentra el conocimiento que tiene la RNA acerca de un determinado problema. La conexión de las neuronas colabora para producir un estímulo de salida.

El empleo de las RNA puede orientarse en dos direcciones, como modelos para el estudio del sistema nervioso y los fenómenos cognitivos, o bien como herramientas para la resolución de problemas prácticos como la clasificación de patrones y la predicción de funciones.

Las RNA han sido aplicadas de forma satisfactoria en la predicción de diversos problemas en diferentes áreas de conocimiento como: biología, medicina, economía, ingeniería, psicología, etc. Obteniendo excelentes resultados respecto a los modelos derivados de la estadística clásica. El paralelismo de cálculo, la memoria distribuida y la adaptabilidad al entorno, han convertido a las RNA en potentes instrumentos con capacidad para aprender relaciones entre variables sin necesidad de imponer presupuestos o restricciones de partida en los datos. Deben ser entrenadas, para posteriormente procesar automáticamente la respuesta que deseamos. Tienen varias funciones como:

- La función de propagación o ponderación que se encarga de transformar las diferentes entradas que provienen de la sinapsis en el potencial de la neurona. Normalmente se usa como función de propagación la suma ponderada de las entradas multiplicadas por los pesos. En esta función se interpreta como un regulador de las señales que se emiten entre neuronas al ponderar las salidas que entran a la neurona.

- La función de activación que combina el potencial postsináptico que nos proporciona la función de propagación, con el estado actual de la neurona, para conseguir el estado futuro de activación de la neurona (Yu Hen Hu, 2002). Sin embargo, es muy común que las redes neuronales no tomen su propio estado como un parámetro y que por tanto no se considere. Ésta función es normalmente creciente y monótona, y las funciones más comunes son: lineal, escalón, hiperbólicas o tangenciales.

- La Función de salida convierte el estado de la neurona en la salida hacia la siguiente neurona que se transmite por las sinapsis. Usualmente no se considera y se toma la identidad, esto es, de manera que la salida es el propio estado de activación de la neurona. Existen algunas redes que transforman su estado de activación en una salida binaria y para eso usan la función escalón antes mostrada como salida. (Weifeng 2008).

3.4.2.1 El Perceptrón

Está constituido por un conjunto de sensores de entrada que reciben los patrones de entrada a reconocer o clasificar y una neurona de salida que se ocupa de clasificar a los patrones de entrada en dos clases, según que la salida de la misma sea 1 (activada) o 0 (desactivada) (Michael, 2007).

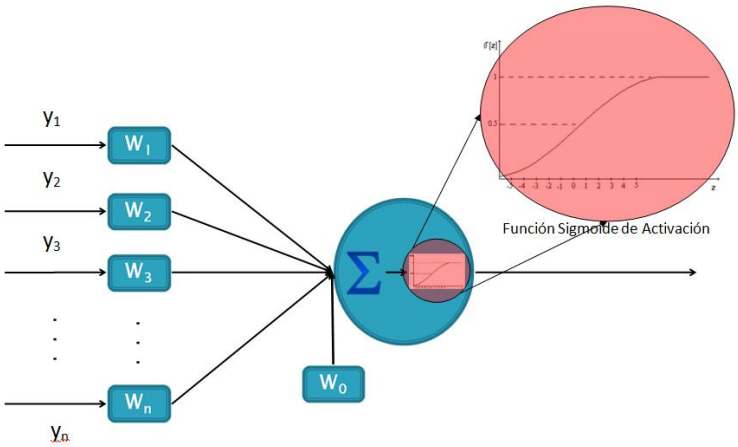


Figura 3.7.- Perceptrón Simple.

Un perceptrón se refiere a una neurona artificial y también a la unidad básica de inferencia en forma de discriminador lineal. El perceptrón simple es una red que consta de dos capas de neuronas. Esta red admite valores binarios o bipolares como entrada para los sensores y los valores de su salida están en el mismo rango que los de entrada. La función de la primera capa es hacer de sensor, por ella entran las señales a la red. La segunda capa realiza todo el procesamiento. La manera de interconectar ambas capas es que cada neurona de la primera capa esté unida con todas las de la segunda capa.

Pero éste tipo de perceptrón tiene una serie de limitaciones, entre las cuales destaca el no poder clasificar conjuntos linealmente dependientes. Por ello surge el perceptrón multicapa, que retoma la definición del perceptrón simple pero se le añaden una serie de capas intermedias que convierten las funciones linealmente dependientes en independientes, gracias a la transformación de la capa oculta (Palmer, 2001).

3.4.3 Metodología de Segmentación

El procedimiento empleado para la segmentación fue:

1) Crear un banco de información compuesto de varios vectores de datos correspondientes a digitalizaciones de palabras como: “abre”, “cierra”, “ventana”, “puerta”, etc.

   Esta base de datos nos sirve para tener la información necesaria para entrenar la RNA. Y consta de 6 distintas palabras grabadas en 10 distintas ocasiones.

2) Dividir cada grabación en ventanas de 12.5 ms, equivalentes a 100 muestras a una frecuencia de muestre fs=8Khz, usualmente se proponen con tamaño de 10-30ms. (A. S. Kolokolov, 2003).

   Se realiza con el fin de obtener las características de la voz para un número específico de datos, el necesario para poder hacer el posterior análisis y el justo para que no sea demasiado tiempo de procesamiento.

3) Obtener la energía y cruces por cero de cada ventana.

   El cálculo se realizó con ayuda de la Ec. (2) y (3), de donde se obtienen las características de la señal que nos permiten definir cuando una trama de datos pertenece a algún segmento de alguna palabra. Se requiere normalizar estos valores para adecuarse a la entrada de la RNA.

4) Segmentar manualmente las grabaciones.

   Esto es, estimar visualmente el inicio y fin de cada palabra y crear una señal que represente como '1' cuando exista la palabra y como cero cuando no exista, como lo es en el inicio y fin de cada grabación; posteriormente se obtiene la media de cada 100 muestras, los resultados de esta operación se agregan a un vector.

5) Se forma una matriz general con los vectores de energía y cruces por cero, y un vector con la salida deseada, estos forman las dos entradas y la salida deseada con la cual se entrena la RNA.

Como se observa en la siguiente figura, la estructura de la red neuronal comprende 6 entradas y una salida. Donde el valor de la salida depende de los valores de cruces por cero y energía pasados y presentes.

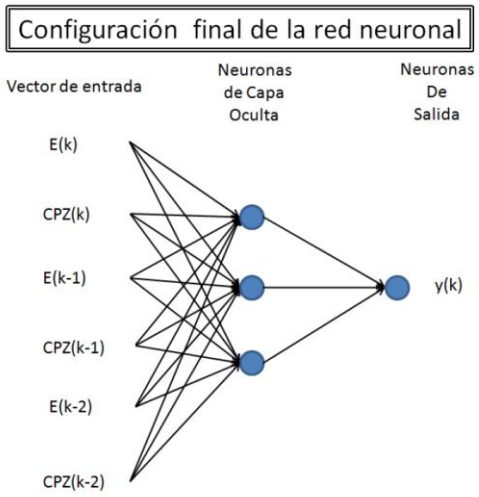


Figura 3.8.- Configuración de la RNA de la segmentación.

3.5 Extracción de Características

En esta etapa del reconocimiento de voz propuesto se realiza la transformación del extenso número de datos de entrada pertenecientes a una palabra en un reducido vector de datos que representa la esencia de la palabra original, para fines de un manejo reducido de datos, accesible y que implica un tiempo de procesamiento menor.

3.5.1 Transformada Rápida de Fourier

Como el origen de la voz son las vibraciones realizadas por nuestro aparato fonador, el tratamiento de esta señal puede llevarse al plano frecuencial para su análisis, por ello se hace uso de la herramienta matemática de la Transformada Rápida de Fourier que es una variación de la Transformada Discreta de Fourier, la cual descompone una señal de entrada en componentes seno y coseno de distintas frecuencias. (Proakis, 1996). Para tener una respuesta más rápida del sistema es necesario tomar segmentos en el tiempo que nos permitan realizar operaciones previas, antes de que toda la palabra sea pronunciada, para así no tener que hacer todo el cálculo hasta el final ya que esto llevaría un tiempo computó mayor y una salida más tardía, también el cálculo de la Trasformada Rápida de Fourier es más sencilla.

La FFT (de sus siglas en inglés Fourier Fast Transform), toma *n* muestras discretas y nos arroja un vector de números complejos, los cuales tienen una amplitud y una fase, se toma la amplitud, pero solo la primera mitad se toma en cuenta ya que la segunda es un espejo de la primera, como se puede observar en la siguiente figura.

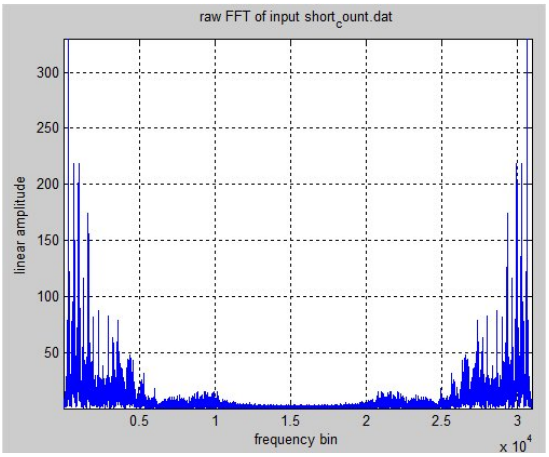
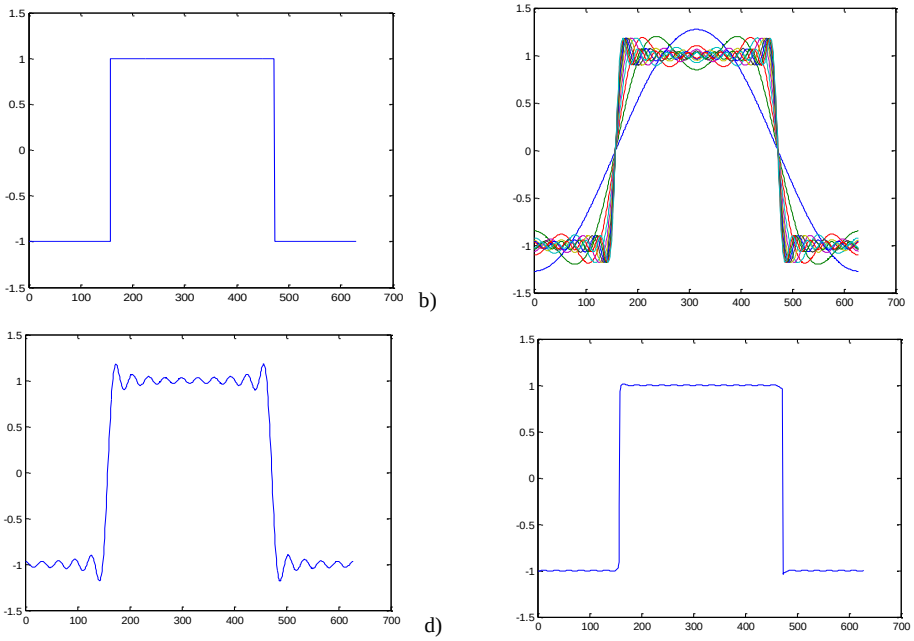


Figura 3.9.- Ejemplo de la Gráfica de la Amplitud de Transformada Rápida de Fourier.

Para ejemplificar el funcionamiento de este procedimiento matemático, se presentan las siguientes imágenes donde 3.3 a) es de una señal cuadrada y 3.3 b) son todas las señales coseno que la forman en conjunto, se puede apreciar distintas frecuencias de oscilación.

a) b)



c) d)

Figura 3.10.- Ejemplo de Transformada Rápida de Fourier.

Dependiendo del número de componentes que se tengan mejor será la reconstrucción de la señal como se muestra en las figuras 3.3 c) y d), la imagen reconstruida se va pareciendo más a la original.

El tamaño de la ventana es fijo y para fines prácticos se toman valores en potencia de 2 para facilitar el cálculo de la FFT. El objetivo es tomar el espectro generado por el pitch (la frecuencia fundamental a nivel de percepción de la voz y cuyo valor medio habitual está entre 80 y 1100 Hz). Por eso se emplean ventanas entre 128 y 256 milisegundos a una frecuencia de muestreo de 8KHz.

La Transformada Discreta de Fourier arroja como resultado dos vectores uno con las amplitudes de los componentes coseno (llamado parte real) y otro con las amplitudes de los componentes seno (llamado parte imaginaria). A continuación se muestran las respectivas ecuaciones:

$$\mathrm{Re}\,X[k] = \sum_{i=0}^{N-1} x[i]\cos\left(\frac{2\pi ki}{N}\right) \tag{5}$$

$$\mathrm{Im}\,X[k] = -\sum_{i=0}^{N-1} x[i]\sin\left(\frac{2\pi ki}{N}\right) \tag{6}$$

En donde *N* es el número total de muestras, *i* es el índice de la muestra de entrada y *k* es el índice de la frecuencia analizada. (Franco, 2009).

Así tendremos arreglos de números rectangulares, que se pueden expresar en su forma polar, que nos da la amplitud y fase de los componentes seno y coseno.

Pero con el fin de poder agilizar este proceso se modificó este algoritmo para reducir el tiempo de cómputo, en la siguiente figura se describe la operación para obtener la Transformada Rápida de Fourier, podemos observar que el que sea un número de muestras en potencia de dos hace que se pueda hacer operaciones en pareja.

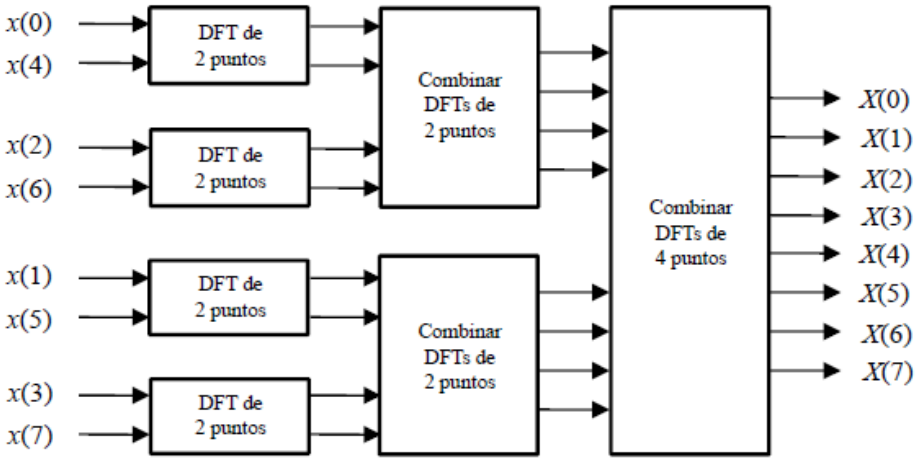


Figura 3.11.- Cálculo de la FFT para 8 muestras.

Donde cada bloque "DFT" es una operación llamada mariposa debido a su apariencia y éste es el elemento básico de la FFT.

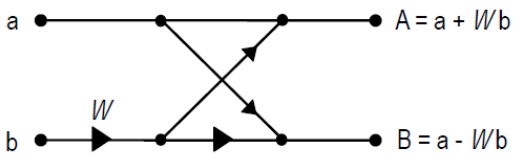


Figura 3.12.- Diagrama de la Operación mariposa para la FFT.

Donde W es igual a $e^{-2\pi fi/N}$

3.5.2 Coeficientes Cepstrales de Mel

Debido a que se debe trabajar con un número de muestras mínimo de 8000 por segundo, es necesario reducir el tamaño para poder manejar esta cantidad de información en un tiempo mínimo, es por eso que se tiene que generar un vector reducido y representativo de la entrada. (Chakraborty, 2008). Esto se logra a través de los distintos métodos explicados en el capítulo anterior, en nuestro caso mediante el método de los Coeficientes Cepstrales de Mel (también conocido como Mel Frecuency Cepstral Coeficients, MFCC) el cual interpreta la señal en una escala que simula la percepción del oído humano ante los sonidos. (Alvarado, 2008).

Para obtener los coeficientes, es necesario trabajar en la denominada Escala de Mel, que es una escala logarítmica basada en la percepción humana del pitch. Es por eso que en la figura 3.6 se puede observar que la respuesta mayor se tiene en los primeros 1000Hz, para después atenuarse.

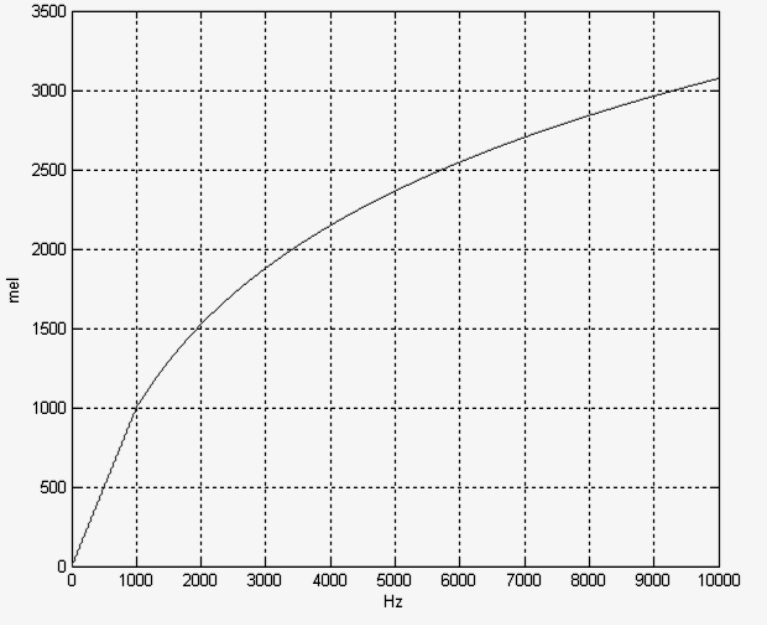


Figura 3.13.- Escala de Mel.

Después de haber obtenido la FFT, relacionamos los dos planos de frecuencia mediante las siguientes 2 fórmulas:

$$\hat{f}_{mel}^{-1} = f_{in} = 700 \bullet \left[\exp\left(\frac{\hat{f}_{mel}}{1127} \right) - 1 \right] \qquad (7)$$

$$\hat{f}_{mel}^{-1} = 1127 \bullet \ln\left[1127\left(1+\frac{f_{in}}{700}\right)\right] \tag{8}$$

La primera corresponde a pasar de frecuencia Mel a frecuencia en Hz y la segunda lo contrario.

Las ecuaciones necesarias para elaborar el banco de filtros necesario para obtener los coeficientes son las siguientes:

$$f_{bi} = \left(\frac{N}{F_s}\right)\hat{f}_{mel}^{-1}\left(\hat{f}_{mel}(f_{low}) + i\frac{\hat{f}_{mel}(f_{high}) - \hat{f}_{mel}(f_{low})}{M+1}\right) \tag{9}$$

$$H_i(k) = \begin{cases} 0 & for \quad k < f_{b_{i-1}} \\ \dfrac{(k - f_{b_{i-1}})}{(f_{b_i} - f_{b_{i-1}})} & for \quad f_{b_{i-1}} \le k \le f_{b_i} \\ \dfrac{(f_{b_{i+1}} - k)}{(f_{b_{i+1}} - f_{b_i})} & for \quad f_{b_i} \le k \le f_{b_{i+1}} \\ 0 & for \quad k > f_{b_{i+1}} \end{cases}, i = 1, 2, ..., M \tag{10}$$

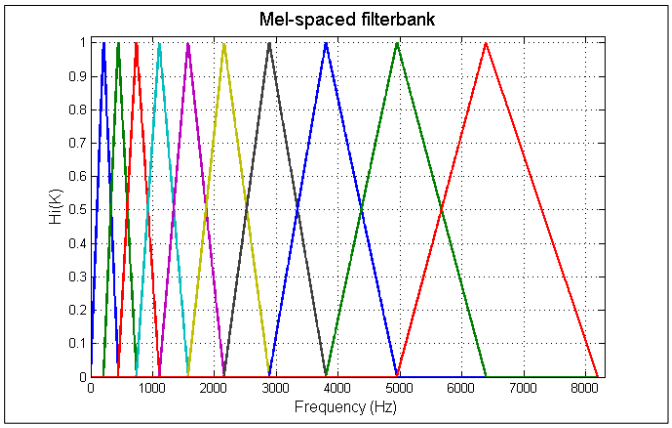


Figura 3.14.- Ejemplo de banco de filtros generado.

En donde *N* es número de muestras, F_{low} es la frecuencia inferior del banco de filtros, F_{high} frecuencia superior, *M* es el número de Filtros, *H* es la Matriz de filtros, *i* es el índice de filtro (del 1-M), y *k* es el número de muestra.

Ahora con el banco de filtros se puede obtener el cepstrum de la señal (proveniente de spectrum, espectro), el cual se calcula mediante:

$$X_i = \log_{10}\left(\sum_{k=0}^{N-1} |X(k)| \bullet H_i(k)\right) \tag{11}$$

De donde |X(k)| es la magnitud de la FFT.

Hasta este punto la reducción del vector de entrada es notable y contenida en el vector de energía X_i.

El último paso es obtener los Coeficientes Cepstrales de Mel de la siguiente forma:

$$C_j = X_i \bullet \cos\left(j \bullet \left(\frac{i-1}{2} \right) \bullet \frac{\pi}{M} \right) \tag{12}$$

En la siguiente figura se puede observar un ejemplo gráfico del cómo se reducen los datos de entrada, en el lado derecho tenemos una palabra de duración 2 segundos grabada a 8Khz, al hacer la transformación mediante los MFCC se observa que se pasa de 16000 datos originales a tan solo 120.

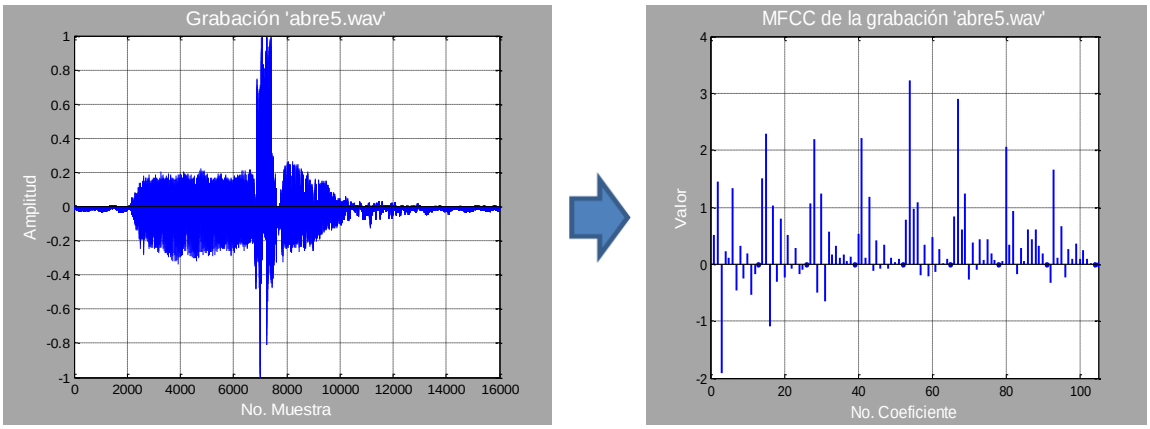


Figura 3.15.- Ejemplo de Aplicar los Coeficientes Cepstrales de Mel.

3.6 Reconocimiento de Palabras

Como se presentó en la sección 3.4.2 las redes neuronales son útiles en problemas donde se encuentra variabilidad en la señal de entrada, ideal para describir no linealidades como es el caso de la voz. La estructura de la RNA propuesta es la que se muestra en la figura 3.16 donde la entrada es un vector de Coeficientes Cepstrales de Mel pertenecientes a una palabra, se tiene una capa oculta y una neurona de salida por cada palabra entrenada.

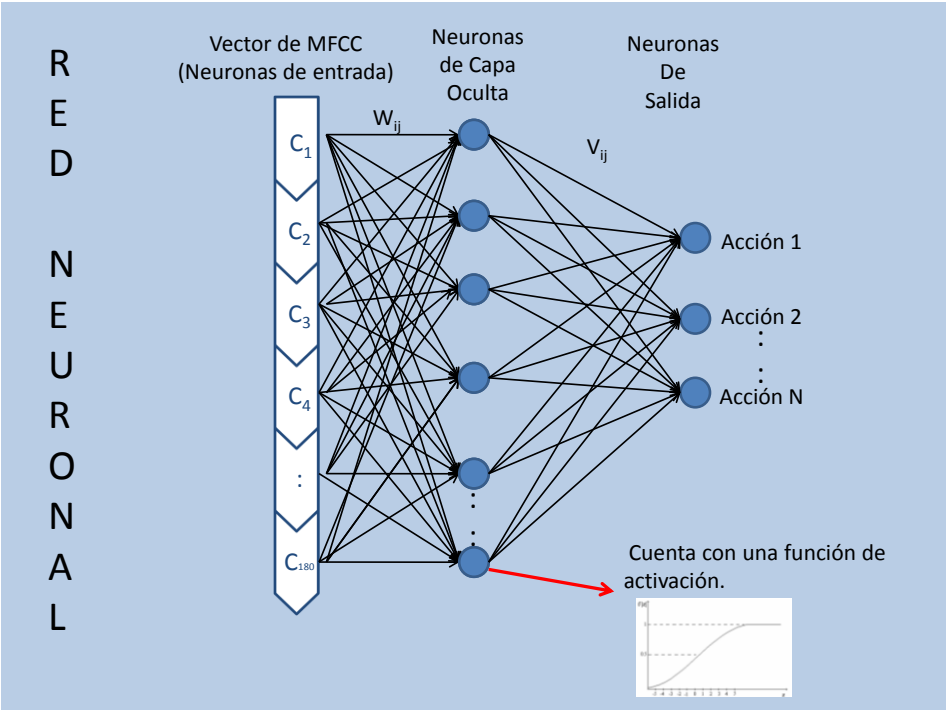


Figura 3.16.- Configuración de la RNA Encargada de Reconocer las Palabras.

En esta sección se optó por emplear las Redes Neuronales Artificiales (RNA) por su robustez al momento de clasificar conjuntos, esto nos permite una mayor tolerancia ante el ruido exterior para así poder tener resultados más satisfactorios.

En la práctica los MFCC se deben normalizar en base al mayor y al menor registrado, para que la entrada a la red neuronal trabaje en el rango de valores de cero y uno.

3.7 Operaciones de la Red GSM

3.7.1 Inicialización

Cuando un teléfono móvil GSM se enciende por primera vez, se tarda unos segundos antes de que una llamada pueda realizarse. No sólo el móvil tiene que inicializarse, también tiene que conectarse a la red de modo que está listo para hacer y recibir llamadas. En la primera etapa de vinculación con la red se verifican las frecuencias disponibles para una BTS. El móvil explorará todas las frecuencias disponibles, teniendo en cuenta las señales que se pueden recibir. Asimismo, observa las señales que contienen un canal de control y toma la información de broadcast, incluyendo la identificación del sistema, la

información de control de acceso y la potencia inicial de transmisión. Las ubicaciones de los canales de mensajería y paginación también se envían. Un identificador único para la estación base es enviado, junto con un identificador de red. El móvil elige la frecuencia faro más fuerte y entonces usando un canal de acceso aleatorio, envía un mensaje de registro. La red responde, verificando la autenticación y almacenando la ubicación del móvil en el HLR y VLR, según corresponda.

3.7.2 Iniciación de llamadas

Cuando se realiza una llamada desde un móvil, el móvil tiene que comunicarse con el BTS para que la llamada pueda ser configurada y se realice la conexión. Se envía una ráfaga de acceso usando el acceso aleatorio de canal. Este mensaje contiene un número de 5 bits que temporalmente identifica el móvil a la red. Al menos 3 bits dentro del identificador de mensaje, el tipo de mensaje que se solicita – la respuesta, el origen de la llamada, o incluso una reconexión de una temporalmente llamada perdida.

Existe la posibilidad de que otro móvil envíe una solicitud de acceso aleatorio al mismo tiempo, y si no se recibe el reconocimiento, se espera cierta cantidad de tiempo aleatorio para que no colapse con el mismo móvil otra vez, y luego volverá a enviar su solicitud de acceso.

Cuando la BTS recibe correctamente la solicitud de acceso, se devuelve el mismo número al azar y dirige el móvil a un canal de radio específico y ranura de tiempo. Normalmente la red va a solicitar autenticación y, si tiene éxito, la llamada puede comenzar.

3.7.3 Recepción de llamadas

Un proceso similar al usado en iniciar una llamada. Obviamente, hay algunas diferencias, porque el móvil tiene que estar en alerta constante de buscar una llamada entrante. Cuando se solicita hacer una llamada a un móvil específico, la red verifica la última ubicación registrada con la ubicación de registro. La red envía un mensaje una página en el canal de paginación de un grupo de BTS en la región donde el móvil se encontraba anteriormente.

Cuando el móvil recibe el mensaje de paginación, responde en su canal de control. Entonces la red asigna un canal de radio y una ranura de tiempo a la cual se envía el móvil. Y se realiza la llamada hasta que uno de los extremos la termina.

3.7.4 Finalizar Llamada

Una llamada puede ser finalizada por cualquiera de los extremos de la conexión. Cuando esto se detecta, una serie de mensajes son intercambiados entre la red y el móvil. Varios mensajes se envían para garantizar que el vínculo no se ha roto accidentalmente y que es verdaderamente una terminación real de llamada.
Una vez finalizada la llamada, la red libera el canal de circulación y el móvil vuelve a su estado de reposo monitoreo del canal de paginación. También se comprueba que la BTS que se estaba usando todavía ofrece la señal óptima. De esta manera se asegura que si en un futuro se recibe una llamada, sea la mejor posición para recibirla.

3.7.5 Traspasos

Uno de los elementos clave de un teléfono móvil es que es capaz de moverse y aún permanecer conectado. Esto significa que cuando el móvil se mueve fuera del rango de una BTS (es decir, fuera de una célula) y entra en la siguiente, debe ser posible transferir la llamada sin ninguna interrupción perceptible para el usuario.

Se necesita una cantidad considerable de tiempo de red para asegurar que este proceso, conocido como traspaso, suceda correctamente, ya que es un elemento esencial del control de la red. Al ocurrir cualquier problema rápidamente se pasa a otra red. El término para esto es "Churn".

En GSM que utiliza técnicas TDMA la emisora transmite sólo por una de ocho ranuras, del mismo modo, el receptor recibe sólo una ranura de ocho. Como resultado, la sección de RF del móvil podría estar inactiva durante seis ranuras de un total de ocho. Este no es el caso, ya que durante las ranuras que no se está comunicando con el BTS se escanean los canales de radio en busca de balizas frecuencias que pueden ser más fuertes o más apropiadas. Además de esto, cuando el móvil se comunica con una BTS particular una de las respuestas que hace es enviar una lista de canales de radio de las frecuencias faro de BTS vecinas. Analiza esto e informa de la calidad del vínculo con la BTS. De esta manera el teléfono móvil ayuda a la decisión de traspaso, y como un resultado de esto se conoce como Mobile-Assisted HandOver (MAHO).
La red sabe la calidad del vínculo entre móvil y la BTS, así como la fuerza de la BTS local percibida por el móvil. También se conoce la disponibilidad de los canales en las células cercanas. Como resultado se tiene toda la información que se necesita para poder tomar una decisión de cambiar de una BTS a otra.

Si la red decide que es necesario que el móvil cambie, le asigna un nuevo canal y un tiempo de ranura. Se informa a la BTS y al móvil del cambio. El móvil entonces sintoniza durante el período de no transmisión o recepción, es decir, en un período de inactividad.

3.7.6 Salto en Frecuencia

Una de las facilidades que ofrece el GSM estándar es permitir un modo de operación de frecuencia llamado salto. Esta es una forma efectiva de operación de espectro ensanchado. Esencialmente, cuando una señal utiliza saltos de frecuencia se mueve de una frecuencia a otra, permaneciendo en una frecuencia determinada durante un corto período de tiempo, suficiente para enviar una ráfaga de datos. Entonces se mueve a otra frecuencia, donde se envía otra ráfaga de datos. Para operar en este modo, el transmisor y el receptor deben seguir el patrón de salto al mismo tiempo.

Hay una serie de ventajas del salto en frecuencia. Es una técnica utilizada por militares para evitar interferencias de señales hostiles, y también evita el espionaje porque la señal no se puede recibir fácilmente a menos que se conozca el patrón de salto. Para las aplicaciones en telecomunicaciones móviles, se utiliza sobre todo porque disminuye el nivel de interferencia. Mediante el uso de saltos de frecuencia, si un canal se bloquea, tendrá sólo un efecto transitorio. Por otra parte, como el canal es realmente compartido por varios móviles saltando de un canal a otro, se reduce el nivel de interferencia. Esto tiene ventajas significantes para los planificadores de red, que a menudo tienen que diseñar sistemas para situaciones en el peor de los casos. Como resultado de esto, el nivel de reutilización de frecuencias se puede mejorar, lo que crea ventajas operacionales y financieras.

Otra ventaja es que reducen los efectos de desvanecimiento selectivo. Una señal que llega a un móvil o BTS será la suma de las señales que llegan del transmisor a través de varios caminos como resultado de reflexiones. Como las longitudes de los caminos serán diferentes, a veces se producirá una señal más grande, mientras que otras veces se tienden a anular. Como esto es algo depende de la frecuencia en uso, moviendo a un canal diferente se mejorará la situación. De nuevo, el problema tiende a ser reducido.

El salto de frecuencia es relativamente fácil de implementar en las redes GSM debido que la sección RF del móvil se está moviendo entre las distintas frecuencias para transmitir y recibir, el cambio de frecuencia se realiza durante los períodos muertos cuando ni el emisor ni el receptor están activos.

Con el fin de coordinar la transmisión y recepción de los canales y asegurarse de que el BTS y el móvil estén en sincronía, el algoritmo de salto se emite en el canal de control.

3.7.7 Servicio de Mensajes Cortos

El servicio de mensajes cortos (SMS) es una característica que es ampliamente utilizada en los teléfonos móviles GSM. Proporciona la capacidad de enviar y recibir mensajes de texto desde y hacia teléfonos móviles. El texto puede comprender palabras o números, o una combinación alfanumérica. SMS se creó como parte de la Fase 1 GSM estándar, y su uso ha crecido más allá de todas las expectativas. El primer mensaje corto se cree que se han enviado en diciembre de 1992 desde un ordenador personal (PC) a un teléfono móvil en la red GSM de Vodafone en el Reino Unido. Cada mensaje corto puede tener hasta 160 caracteres de longitud cuando se usan alfabetos latinos, y 70 caracteres de longitud cuando alfabetos no latinos (como el árabe y el chino).
SMS es lo que se denomina como "almacenamiento y reenvío de servicio". Los mensajes del remitente son enviados a un centro SMS y luego hacia el destinatario. Cada red de telefonía móvil que soporta SMS a uno o más centros de mensajería para manejar y gestionar los mensajes cortos. Una vez emitido una confirmación de mensaje se entrega, y esto puede ser seleccionado para aparecer en el teléfono del usuario.
Los mensajes cortos se pueden enviar y recibir al mismo tiempo con la voz de GSM, los datos y llamadas de fax. Esto es posible porque la voz, datos y llamadas de fax se toman desde el canal de tráfico, mientras que el SMS utiliza la ruta de señalización.

3.7.7.1 Comandos AT.

Los comandos AT son instrucciones codificadas que conforman el lenguaje de comunicación entre un usuario y un terminal módem y son de carácter genérico en su mayoría, ya que un mismo comando funciona en modelos de distintas marcas, haciendo que un programa basado en comandos AT sea inmensamente robusto y compatible con la mayor parte de los dispositivos disponibles en el mercado.

La gran parte de los módems disponibles reconocen los comandos AT más utilizados. Por lo mismo, la tecnología GSM ha adaptado el uso de estos comandos, teniendo comandos específicos que pueden ser encontrados en documentación especializada sobre el módulo GSM. Dependiendo del módulo usado, es la implementación que se le da a los comandos y no depende del medio de comunicación, que puede ser serial, infrarrojo o Bluetooth.

Los comandos AT, poseen en su mayoría un prefijo dado por ‘AT’. Cada acción que se desee viene precedida por este prefijo.

Algunos comandos, llevan al final un signo de interrogación (?). Esto quiere decir que se está pidiendo información. Mientras que un signo igual (=) quiere decir que se está configurando un parámetro, donde luego del signo igual se ingresa el valor o valores de los parámetros separados por coma que se desean ajustar. La expresión igual-interrogación (=?), se usa para obtener todo el rango de valores posibles que se pueden configurar.

Capítulo 4

Experimentos y Resultados.

Este capítulo está dedicado a mostrar los distintos experimentos realizados y al análisis de los resultados que se obtuvieron. Primero verán las pruebas y resultados de la viabilidad del proyecto, después los resultados de aplicar la segmentación automática de voz, pasando por aplicaciones diseñadas para la manipulación de la información y finalmente las pruebas definitivas para encontrar una configuración óptima del sistema.

El Procedimiento que se dio para seleccionar el número de comandos propuestos, fue incremental y consta de 3 faces distintas.

1.- Identificación de 3 palabras diferentes: Fué el primer acercamiento que se tuvo con todos los algoritmos integrados con fines prácticos de confirmación de los resultados positivos esperados.

2.- Identificación de 8 palabras diferentes: Durante esta fase se esperaba que al incrementar el número de comandos, la RNA se comportara de manera similar, manteniendo su capacidad de identificar palabras correctamente.

3.- Identificación de 20 palabras diferentes: En está fase final se propuso un universo finito de 20 palabras que nos dan las combinaciones necesarias para el control e interacción con edificios inteligentes.

A continuación se mostrarán los resultados de cada una de estas faces con su respectivo análisis. Las gráficas referentes a la respuesta neuronal muestran con un color diferente a la respuesta de cada una de las neuronas de salida, asi mismo en el eje de las abscisas se hace referencia la base de datos de prueba, en donde cada punto es un archivo de prueba en cuestión y éstos se agrupan en conjuntos n archivos, donde n es el número de veces que se prueba la palabra, pertenecientes a una misma palabra.

El primer acercamiento al reconocimiento de palabras se realizó mediante una simulación de prueba para 3 palabras distintas: {‘gato’, ’carro’, ’hola’}. En esta primera prueba se configuró con los parámetros más destacados en la literatura del tema. Como el número de coeficientes cepstrales de mel, el tamaño de la ventana y la frecuencia de muestreo.

FASE 1:

La prueba preliminar se realizó con sólo 3 distintas palabras {‘gato’, ’carro’, ’hola’}, Grabadas a 8Khz, Con un tamaño de ventada de 1024 muestras pertenecientes a 128ms y con 15 MFCC por ventana. Se usaron 10 grabaciones de cada palabra para el entrenamiento y 8 distintas para su validación.

En la siguiente gráfica se observa el resultado de este primer acercamiento, se fijó un objetivo de 0.65 a la salida de cada neurona como medio de aceptación de la palabra. Los resultados fueron favorables, ya que de las 24 grabaciones, 24 acertaron con el inconveniente de un falso positivo en la grabación 21 perteneciente a la palabra ‘gato’, pero que sin embargo se identificó como ‘carro’ también.

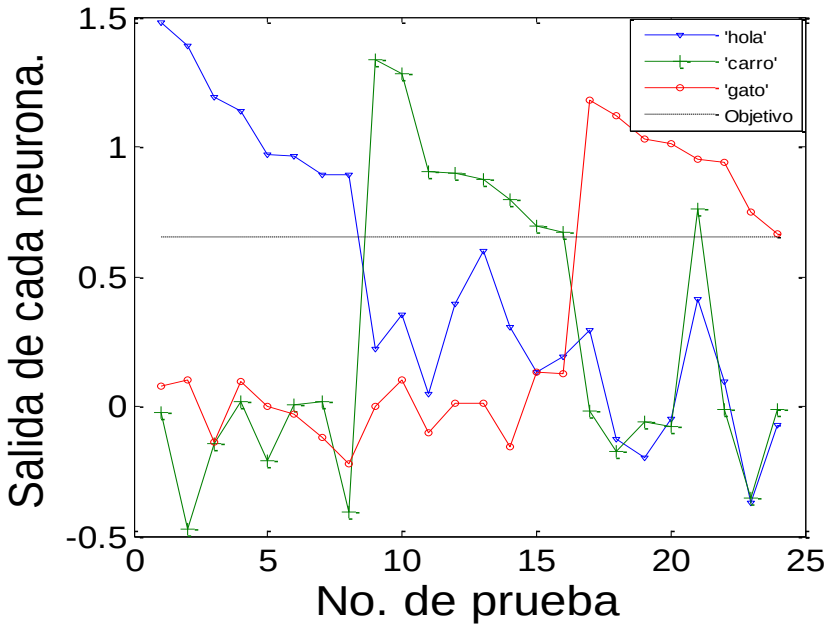


Figura 4.1.- Respuesta a la red neuronal, Fase 1.

Después de la primera fase, se amplía el número a 7 distintas palabras que son: {‘abre’, ’apaga’, ’cierra’, ’enciende’, ’lámpara’, ’puerta’ y ’ventana’}, referentes al control de cargas en edificios inteligentes, para los cuales se tomaron 10 distintas grabaciones con variaciones en la pronunciación. A este set de entrenamiento se le variaron distintos parámetros para mejorar la respuesta. Se emplearon 7 neuronas en la capa de salida. Cada una de las cuales se entrenó de tal forma para que arrojaran un valor cercano a ‘1’ cuando se introdujera una entrada similar a la palabra correspondiente, y un valor cercano a ‘0’ cuando entrada no pertenezca a la palabra. La siguiente imagen muestra un boceto de cómo está conformada la RNA.

Para cada red se formó una matriz de entrada de Nx10x7, donde N es un vector del tamaño de la entrada a la RNA y se expresa como el tiempo máximo de grabación multiplicado por el número de MFCC por ventana y dividido entre el tamaño de la ventana, si el espacio temporal de alguna de las grabaciones no era el suficiente para ocupar N MFCC se rellenaba con ceros lo restante.

En ambas pruebas la velocidad de grabación fue de 8Khz, el modo mono aural, y con un tamaño de dato de 7 bits más uno de signo. Las grabaciones fueron realizadas en ambiente de oficina.

A continuación se mostraran los distintos casos de entrenamiento para esta fase.

FASE 2, CASO I:

Para este experimentó se implementó la ventana de 128 ms, equivalente a 1024 muestras, se definió un banco de filtros de 13, dando como resultado 169 datos máximos de entrada a la RNA, el resultado se refleja en la siguiente gráfica:

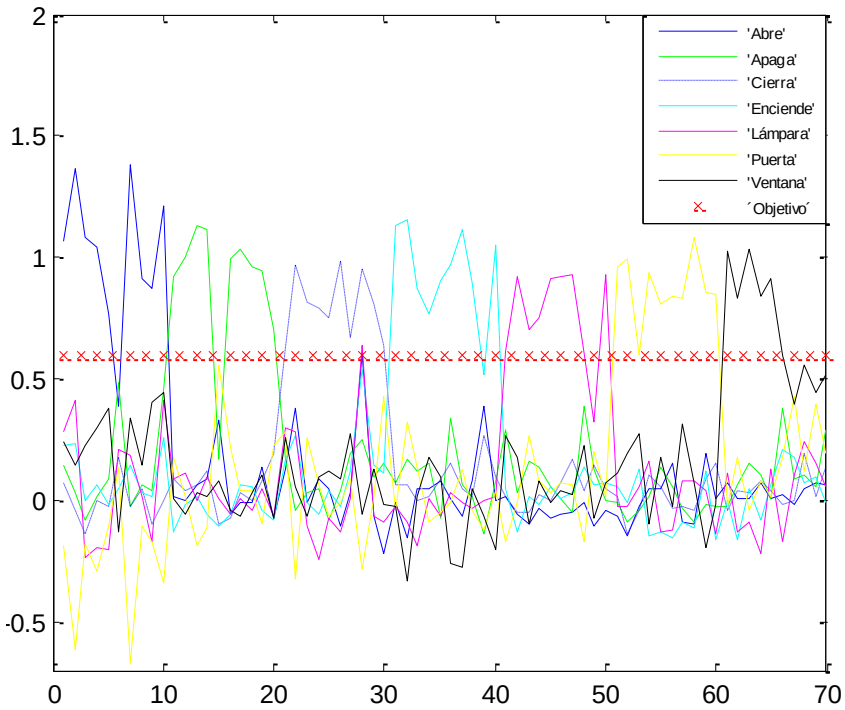


Figura 4.2.- Respuesta a la red neuronal, Fase 2, Caso I.

La siguiente tabla nos muestra la respuesta de la red neuronal al introducir las palabras de prueba, se fijó un umbral de 0.6 para declarar acertada la salida de la red neuronal, la eficacia de este sistema es del 80%.

		"abre"	"apaga"	"cierra"	"enciende"	"lámpara"	"puerta"	"ventana"
No. de grabaciones	Entrenamiento	10	10	10	10	10	10	10
	Comprobación	10	10	10	10	10	10	10
	Acertadas	9	9	8	8	7	10	5
	Falso Positivo	0	0	1	0	0	0	0

Tabla 4.1 Resultados de la Fase 2, Caso I.

FASE 2, CASO II:

Se propuso cambiar el tamaño de la ventana a 1 segundo para tomar las frecuencias de 0-4KHz, pero los resultados no fueron favorables ya que la taza de reconocimiento decremento a menos del 70%, añadiendo un número significante de resultados falsos positivos, por lo cual se desechó este tamaño de ventana para trabajar con la red neuronal. Al modificar el tamaño de ventana, se modifica el número de coeficientes por ventana, para poder cumplir con el número de entradas de la red neuronal, por ello el número de coeficientes por ventana para este caso es de 26.

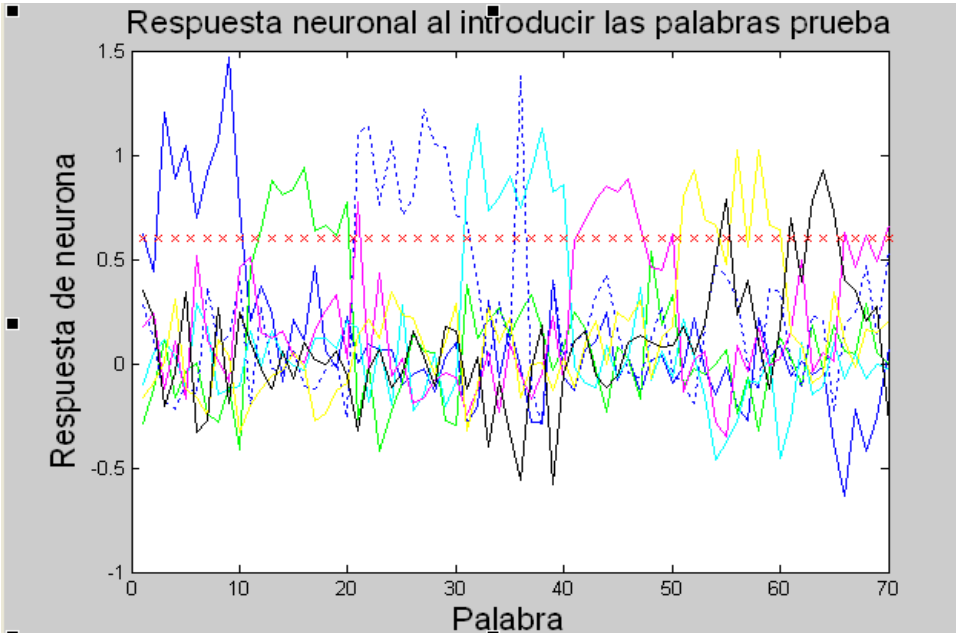


Figura 4.3.- Respuesta a la red neuronal, Fase 2, Caso II.

En la siguiente tabla se aprecian los resultados a detalle, siendo estos los menos aptos hasta el momento para implementar en la tarjeta.

		"abre"	"apaga"	"cierra"	"enciende"	"lámpara"	"puerta"	"ventana"
No. de grabaciones	Entrenamiento	10	10	10	10	10	10	10
	Comprobación	10	10	10	10	10	10	10
	Acertadas	7	7	10	8	6	8	5
	Falso Positivo	0	0	1	1	0	1	2

Tabla 4.2 Resultados de la Fase 2, Caso II.

FASE 2, CASO III:

Para este caso, se retomó el valor de la ventana de 128ms, pero esta vez cambiado el número de MFCC de 13 a 15, incrementando la taza de acierto aproximadamente al 90%.

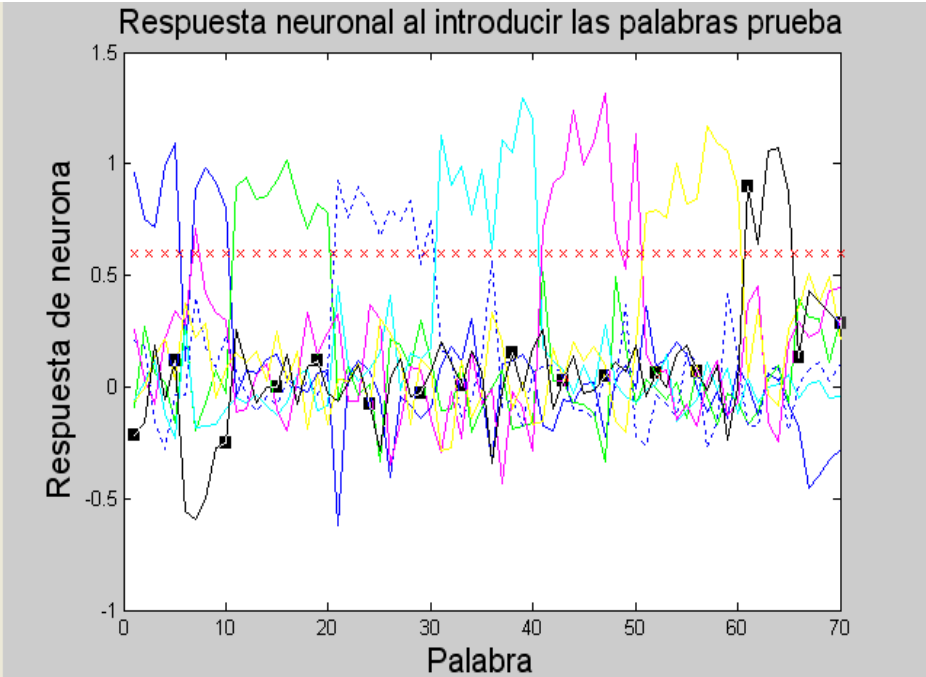


Figura 4.4.- Respuesta a la red neuronal, Fase 2, Caso III.

Una ventaja de emplear ventanas de duración corta es el simplificar el cálculo de la transformada de Fourier y del algoritmo de reducción de características.

		"abre"	"apaga"	"cierra"	"enciende"	"lámpara"	"puerta"	"ventana"
No. de grabaciones	Entrenamiento	10	10	10	10	10	10	10
	Comprobación	10	10	10	10	10	10	10
	Acertadas	9	9	9	10	9	10	6
	Falso Positivo	1	0	0	0	0	0	0

Tabla 4.3 Resultados de la Fase 2, Caso III.

FASE 2, CASO IV:

En este último experimento, se cambió el tamaño de la ventana a 256 mseg teniendo un desempeño del 92% de acierto solo con un falso positivo, este es el mejor resultado que se tiene hasta el momento. Sin embargo el comportamiento a la respuesta de la última palabra de prueba sigue un patrón en todos los casos de no reconocer las últimas 4 grabaciones pertenecientes a la palabra {‘ventana’}.

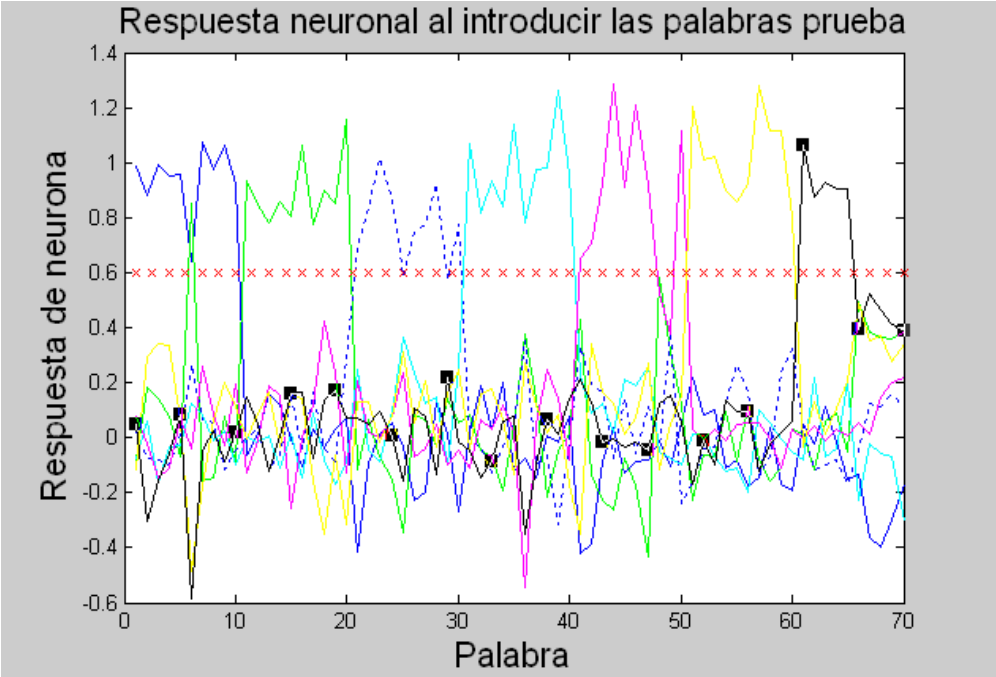


Figura 4.5.- Respuesta a la red neuronal, Fase 2, Caso IV.

		“abre"	“apaga"	“cierra"	“enciende”	“lámpara”	“puerta”	“ventana”
No. de grabaciones	Entrenamiento	10	10	10	10	10	10	10
	Comprobación	10	10	10	10	10	10	10
	Acertadas	10	10	9	10	9	10	6
	Falso Positivo	1	0	0	0	0	0	0

Tabla 4.4 Resultados de la Fase 2, Caso IV.

Cada una de estas grabaciones fue segmentada manualmente mediante estimación visual para identificar los límites de la palabra. Solo se empleó una capa para agilizar los resultados, se experimentó con más capas intermedias, sin embargo los resultados no eran lo suficientemente significativos como para sacrificar tiempo de procesado. No existe una regla general para determinar el número de neuronas que necesita una capa intermedia, algunos textos dicen que la mitad de neuronas de la capa de entrada bastan para una buena identificación, mientras otros dicen que con la raíz cuadrada del número de entradas es suficiente, se tomó el valor intermedio entre estos dos criterios.

Las neuronas de capa oculta y las de capa de salida tienen como función de activación la sigmoide.

Las primeras pruebas se realizaron con el uso del kit de herramientas de desarrollo de redes neuronales artificiales que provee MatLab. En la figura 4.6 se observa el entorno gráfico al momento de entrenamiento, cabe destacar la importancia de los parámetros denominados Época y Error Objetivo, el primero sirve para fijar el número máximo de ciclos de ajuste de pesos, mientras que el segundo fija un error mínimo a lograr, el entrenamiento termina cuando alguno de estos 2 parámetros cumple su acometida.

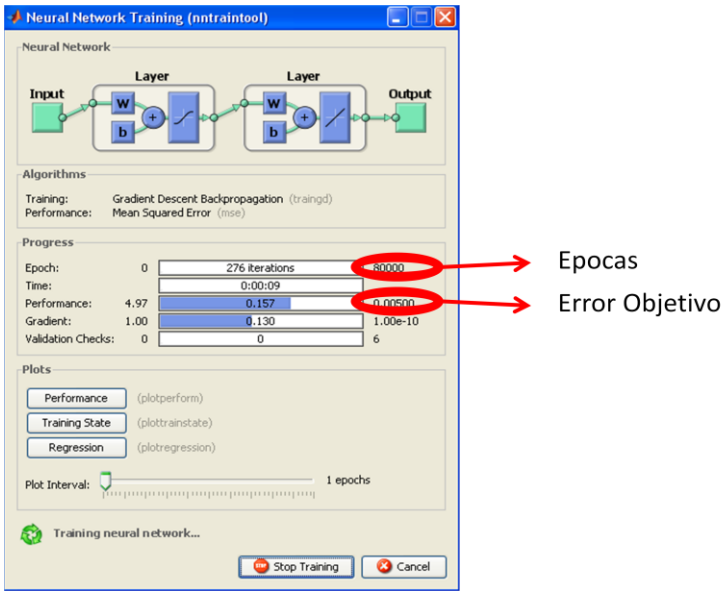


Figura 4.6.- Entorno gráfico de RNAs MatLab.

Se observa que para el caso IV la respuesta fue satisfactoria ya que la mayoría de las palabras probadas fue acertada, esto anima a poder incrementar número de palabras para una mejor interacción con dispositivos electrónicos. Pero para poder emplearla en un sistema real es necesario adecuar la señal de entrada a una segmentación automática de palabra que nos permita segregar solo el contenido de interés.

FASE 3:

En esta fase final se entrenó la RNA de tal forma en la que pudiera identificar veinte distintas palabras de comando, los resultados son alentadores y cubren alrededor del 90% de las palabras acertadas correctamente, las gráficas 4.7 y 4.8, también en tabla 4.4.

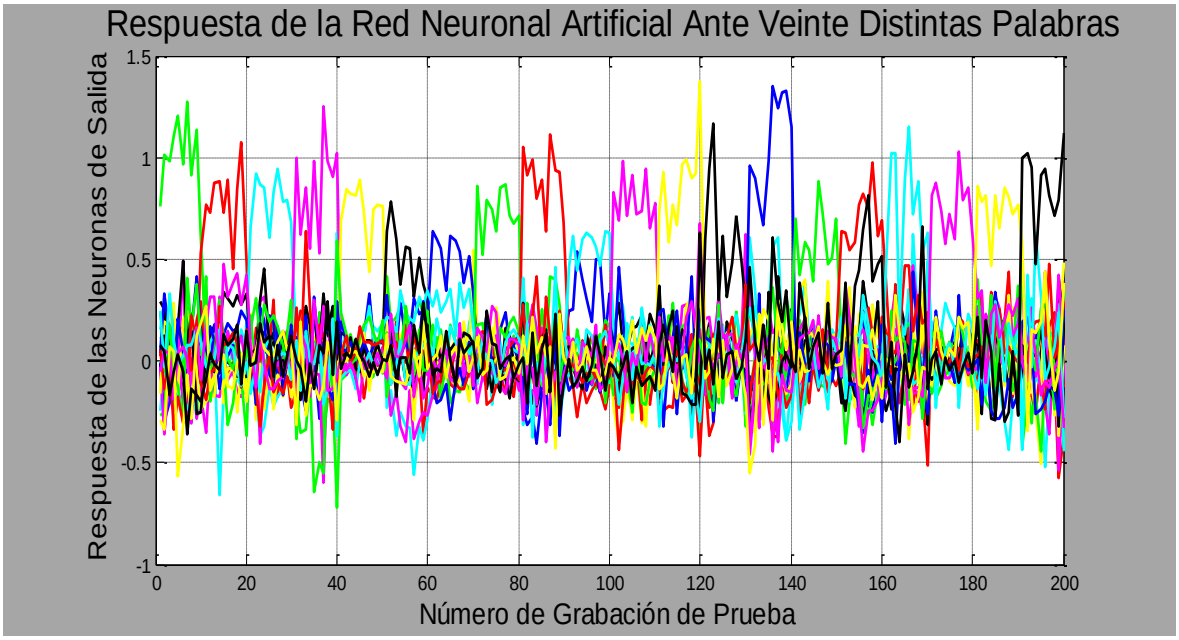


Figura 4.7.- Respuesta a la red neuronal, Fase 3.

En la gráfica 4.7 se puede observar la respuesta de la RNA, notándose que la mayoría de las respuestas positivas sobrepasan un umbral de 0.5, mientras que solo unas cuantas respuestas falsas positivas sobrepasan el 0.4 de valor de salida. En la tabla 4.5 se explone el desglose de estos resultados en donde la mayoría de las palabras se identifican correctamente, para el caso en donde existía un falso positivo el algoritmo de salida evalúa las respuestas de las neuronas que sobrepasan este umbral y selecciónа la de mayor magnitud.

		"Uno"	"Dos"	"Tres"	"Cuatro"	"Cinco"	"Seis"	"Siete"	"Ocho"	"Nueve"	"Cero"	"Acciona"	"Aparató"		"Cancelar"		"desactivar"	"llamada"	"mensaje"	"puerta"	"ventana"	Total
No. de grabaciones	Entrenamiento	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	200
	Comprobación	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	200
	Acertadas	10	9	10	9	9	7	8	10	10	8	10	10	8	9	8	8	9	9	10	9	180
	Falso Positivo	0	0	0	1	0	0	0	0	0	2	0	0	1	2	0	1	1	0	0	0	8

Tabla 4.5 Resultados de la fase 3.

Para interpretar la gráfica 4.8 se toma en cuenta que una respuesta correcta es aquella en la que el módulo del número de grabación % 20, es igual al valor de la ordenada, esto és de una forma más digerible que por cada 10 archivos de muestra se incrementa en 1 el valor de neuronas con máxima exitación. Notándose que solo diez respuestas no corresponden a su debido valor de respuesta.

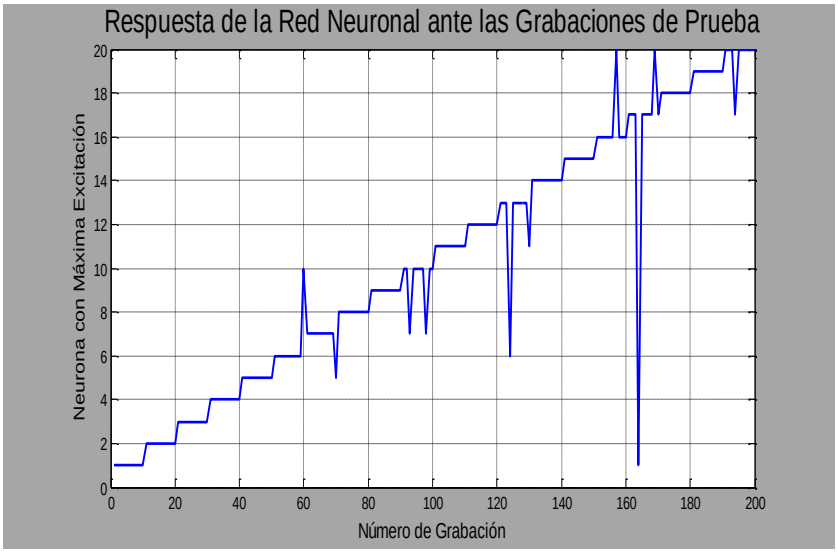


Figura 4.8.- Disperción de Respuesta Neuronal de Salida.

Pero no fue nada simple el encontrar la configuración óptima en la que se logre identificar el mayor número de respuestas correctas, para ello y por que no se tiene una manera precisa y definida de estructura neuronal, se optó por hacer pruebas exhaustivas de variación de número de parámetros con algoritmos automáticos que nos permitan extraer los parámetros más sobresalientes y aplicarlos.

A continuación se mostrarán los resultados obtenidos al aplicar esta metodología empírica para encontrar la configuración de RNA adecuada para este proyecto en específico.

4.1 Configuración óptima de la red neuronal artificial

No existe una fórmula de manejo de redes neuronales artificiales que nos indique el número de neuronas entrada, capaz intermedias, neuronas de capa oculta ó función de activación usar, solo en base a la experimentación se puede llegar a un resultado satisfactorio que se adecue a la respuesta deseada.

En la búsqueda de la configuración de red neuronal que mejor desempeño tenga y que se adapte más al tipo de datos de entrada se hicieron pruebas en donde se dejan fijos todos los parámetros con excepción del que se desea analizar. La frecuencia de muestreo empleada fue para todos los casos de 22050Hz, el tamaño de dato es de 22 bits, el número de épocas objetivo para entrenar la red neuronal artificial es de 20000

Variando el número de palabras

Parámetros	
Filtros Empleados	13
Tamaño de Ventana	8192
Entradas a la RNA	65
Neuronas de capa Oculta	130
Grabaciones de Entrenamiento	10

Tabla 4.6 Parámetros de entrenamiento fijos, ante variación de número de palabras.

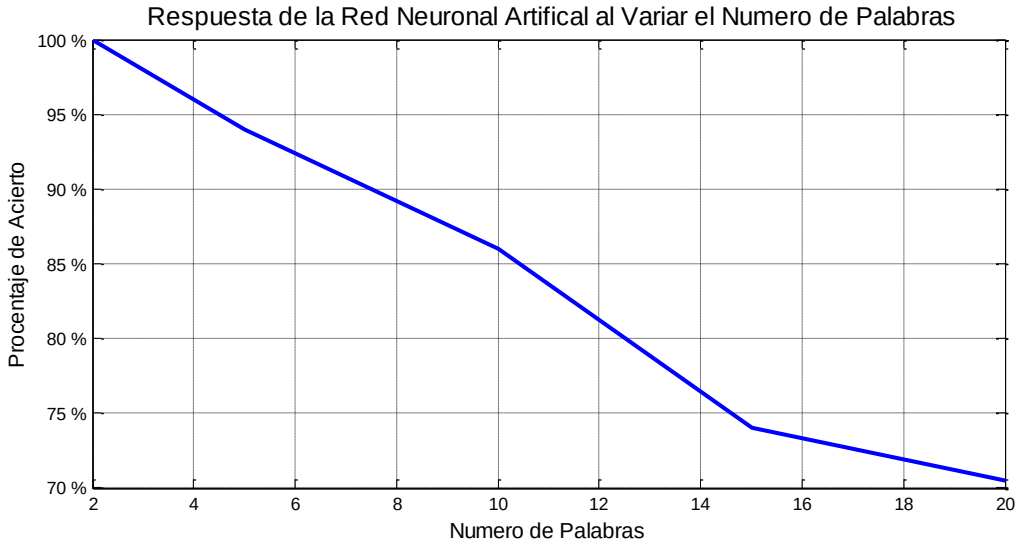


Figura 4.9.- Variación de Número de Palabras

En la gráfica se puede observar que conforme aumenta el número de palabras que debe identificar la red neuronal artificial, el porcentaje de acierto decrementa, Esto impide que se tengan grandes vocabularios dentro de una red neuronal artificial única, se pueden hacer arreglos de varias redes neuronales para poder mejorar esta relación de número de palabras.

Variando el número de grabaciones por palabras para entrenar.

Este punto es crítico ya que de él depende el tiempo en el que el usuario puede hacer el entrenamiento del sistema, no debe de extenderse demasiado para que sea rápida, lo menos tediosa posible y al mismo tiempo eficiente.

Parámetros	
Grabaciones de Entrenamiento	13
Tamaño de Ventana	8192
Entradas a la RNA	52-65
Neuronas de capa Oculta	204-260
Diferentes Palabras	20

Tabla 4.7 Parámetros de entrenamiento fijos, ante variación de número de grabaciones.

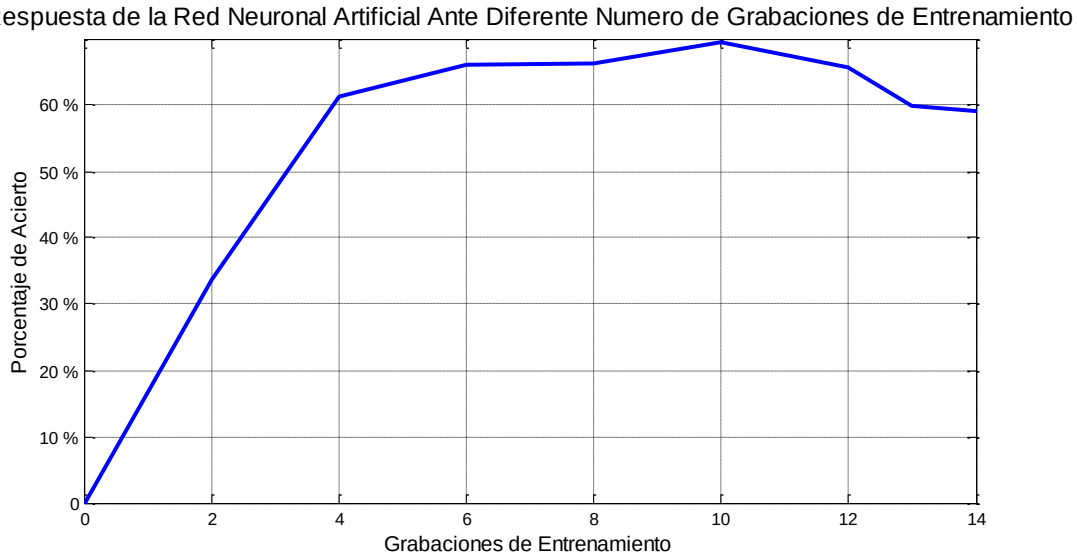


Figura 4.10.- Variación de número de grabaciones de entrenamiento.

Como la figura nos muestra, el mejor número para tener un buen desempeño es de alrededor de las 10 grabaciones por palabra. Muy pocas grabaciones no alcanzan un resultado satisfactorio y después de las diez grabaciones no el porcentaje de acierto decae para después estabilizarse. Esto también es indicio de un sobre entrenamiento, en donde se satura de información a la red neuronal artificial

Variando el número de filtros.

Variar el número de Coeficientes Cepstrales de Mel nos ayuda a identificar donde se obtiene el número necesario de elementos por ventana analizada. Esto influye directamente con el número de entradas de la RNA.

Parámetros	
Grabaciones de Entrenamiento	10
Tamaño de Ventana	8192
Entradas a la RNA	25-200
Neuronas de capa Oculta	50-400
Diferentes Palabras	20

Tabla 4.8 Parámetros de entrenamiento fijos, ante variación de número de filtros.

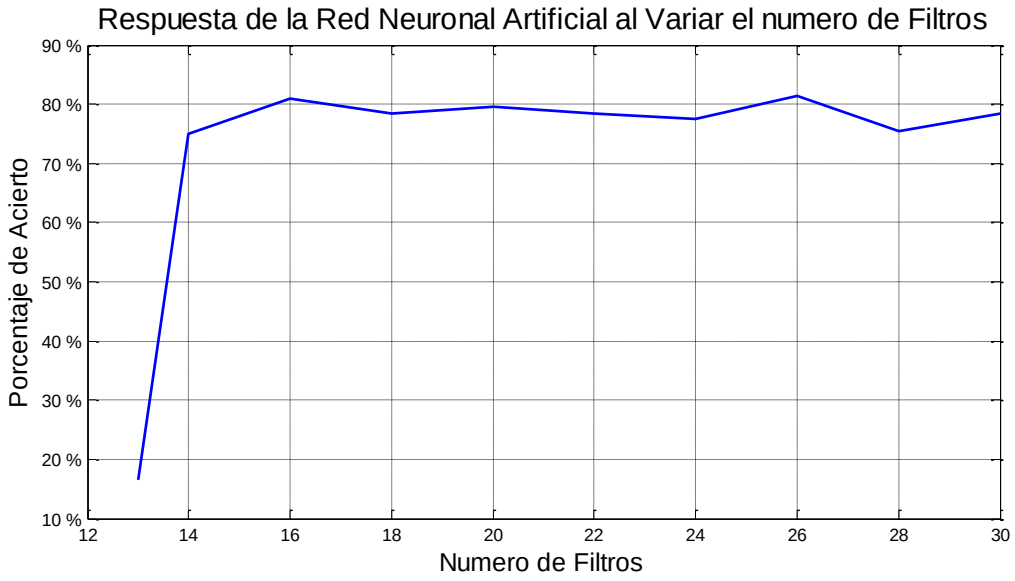


Figura 4.11.- Variación de Número de Filtros.

Los resultados obtenidos son fácilmente leídos ya que si son muy pocos los elementos que describen una palabra, el banco de filtros tomará grandes cantidades de frecuencia en solo unos cuantos filtros y confundirá las entradas, pero al aumentar el número de filtros demasiado se corre el peligro de que el espectro de frecuencias importante como lo es el primer kilohercio quede en sólo en los primeros filtros, desperdiciando y haciendo muy parecido el espectro de las palabras en los filtros restantes.

Variando el número de neuronas en la capa oculta.

Los textos no mencionan cual es la relación entre neuronas de entrada y capa oculta, por eso se realizó esta prueba con distintas escalas de multiplicación de las neuronas de capa oculta que van desde 0.5 a 16 veces el número de neuronas de la capa de entrada.

Parámetros	
Grabaciones de Entrenamiento	10
Tamaño de Ventana	8192
Entradas a la RNA	25-200
Neuronas de capa Oculta	50-400
Diferentes Palabras	20

Tabla 4.9 Parámetros de entrenamiento fijos, ante variación de número de neuronas en la capa oculta.

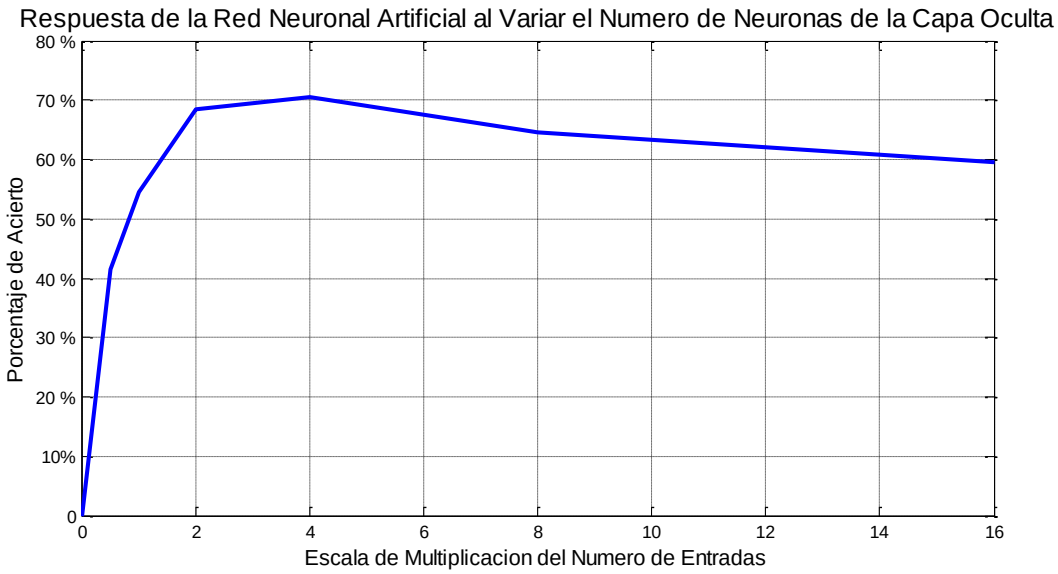


Figura 4.12.- Variación de Número de Neuronas de la Capa Oculta.

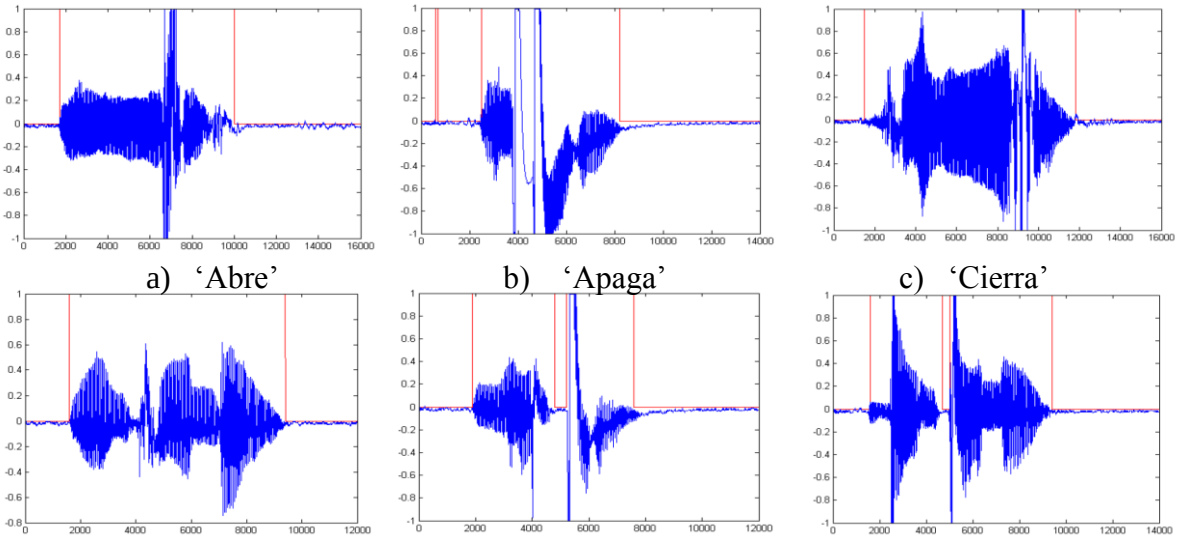
4.2 Resultados de la Segmentación Automática de Voz

Una eficiente segmentación automática de voz nos ayuda a eliminar datos innecesarios, a reducir el procesamiento, disminuir el tiempo de ejecución, aumentar la caracterización de la señal y por supuesto como herramienta para realizar el reconocimiento automático de voz sin la necesidad de depender de la apreciación humana. Sin embargo este proceso es muchas veces dificl de realizar de una manera precisa, en este trabajo en particular el problema más recurrente fue siempre el de la separación que hay entre sílabas. Pero en general el procedimiento adoptado obtiene buenos resultados.

Para validar la efectividad del sistema se probó la red neuronal bajo vectores de entrada de palabras distintas a las del conjunto de grabación, observándose la salida de la red neuronal para cada una. Se espera que mientras se apliquen datos de entrada pertenecientes a una palabra, la salida de la red este próxima al valor uno. En el instante en que los datos dejan de presentar la palabra, la salida de la red exhibirá un valor cercano a cero.

La cantidad de datos de prueba corresponde a 70 grabaciones, la validación se realizó de comparando la salida de la red neuronal ante una entrada distinta a la de entrenamiento. Se obtuvieron aciertos en segmentación del 85%.

A continuación se muestran algunos de los resultados más relevantes de este experimento, cada una de las siguientes imágenes representa los datos de una palabra distinta, el espectro de azul es el perteneciente a la señal de voz, mientras que el rojo es la respuesta del algoritmo de reconocimiento automático de voz.



a) 'Abre' b) 'Apaga' c) 'Cierra'

d) 'Enciende' e) 'Puerta' f) 'Ventana'

Figura 4.13.- Respuesta de la red neuronal ante distintas palabras de entrada.

Se puede observar que los falsos negativos (salida inactiva cuando el resultado es correcto) o falsos positivos (salida activa con estimulo incorrecto) se presentan en un intervalo muy corto, por lo cual se implementó un filtro de media posterior que ayuda a mejorar la respuesta de la RNA, este filtro consta del valor promedio de las cinco entradas previas y cinco posteriores a la muestra analizada, mejorando notablemente el resultado de la segmentación.

En la siguiente grafica se muestra en color azul una palabra de entrada y en rojo la segmentación al aplicar la RNA y el filtro de media, mostrándose la efectividad del algoritmo propuesto.

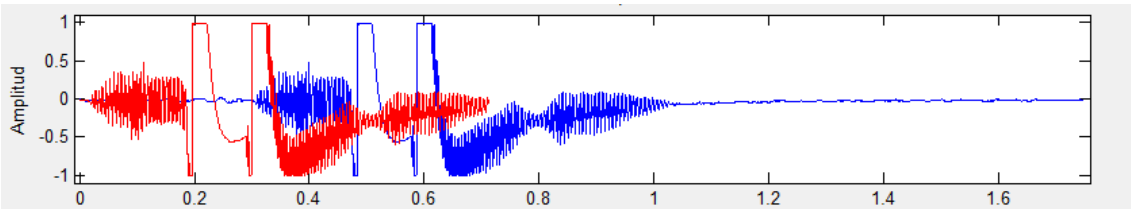


Figura 4.14.- Segmentación de la palabra ‘apaga’.

El proceso en forma gráfica se muestra en la siguiente secuencia. Al inicio se divide en ventanas de 12.5 mseg, cada una de las cuales promedia y se normalizan la energía (línea roja) y los cruces por cero (línea azul), estos valores se introducen a la RNA la cual bajo cierto umbral nos da una respuesta que puede ser ‘1’ o ‘0’ (línea verde) si pertenece al conjunto de palabras entrenadas, después se aplica el filtro de media a las cinco muestras anteriores y posteriores (línea negra).

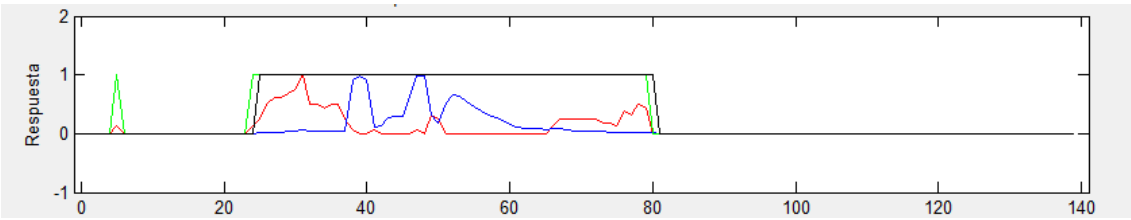


Figura 4.15.- Respuesta de la red neuronal ante distintas palabras de entrada.

Ya con el problema de la segmentación automática resulto el siguiente paso trazado fue el de crear una interfaz que permita hacer pruebas en tiempo pseudoreal.

4.3 Interfaz Gráfica de Usuario del Laboratorio de Pruebas

Con el fin de automatizar y metodizar en módulos el proceso se diseñó una interfaz gráfica en MatLab, esté programa facilita la iteración y pruebas de variables de diseño del sistema de reconocimiento automático de comandos así como también permite una simulación en tiempo pseudoreal. Se puede simular el comportamiento del comparador de patrones de archivos previamente grabados distintos a los simulados y grabaciones realizadas directamente desde la misma interfaz.

A continuación se muestra un pequeño tutorial sobre el funcionamiento y manejo de dicha interfaz gráfica.

La interfaz funciona de 2 distintas maneras, la primera es abriendo un archivo guardado en el disco duro, el paso 1 es presionar el icono de la carpeta, ubicado en la parte superior izquierda de la ventana principal

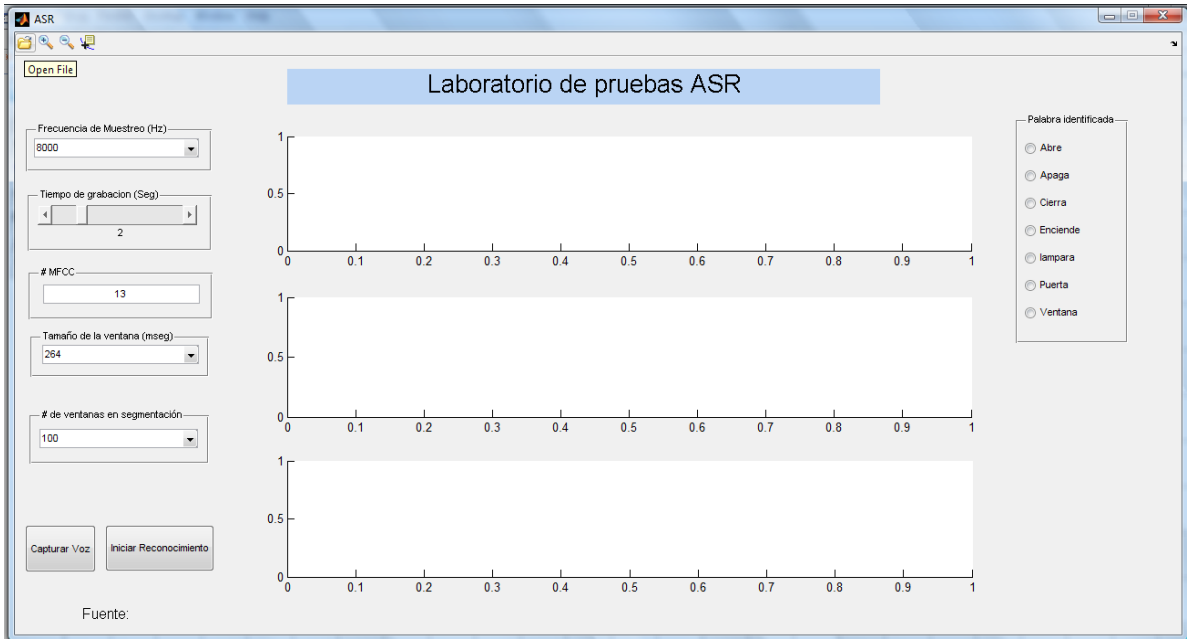


Figura 4.16.- Interfaz Gráfica de Usuario para hacer pruebas en tiempo pseudoreal.

Acto seguido se abre una ventana en la cual buscamos el archivo, teniendo la posibilidad de navegar en forma de explorador:

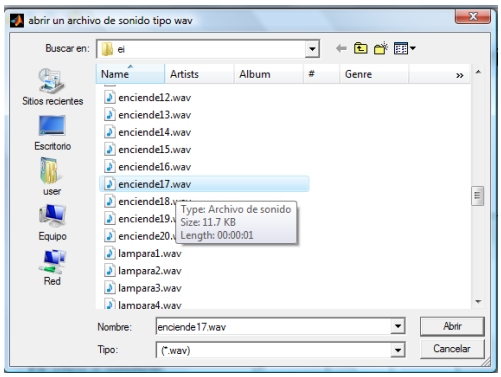


Figura 4.17.- Selección de archivos .wav previamente grabados en la pc.

Después de seleccionar el archivo con doble click, o presionando el botón abrir, se prosigue a seleccionar las condiciones de simulación, es muy importante que éstas coincidan con las de la red neuronal a simular, las condiciones por default son las que aparecen al principio, 13 Coeficientes Cepstrales de Mel, Ventana de 264 milisegundos y ventana de segmentación de 100 muestras.

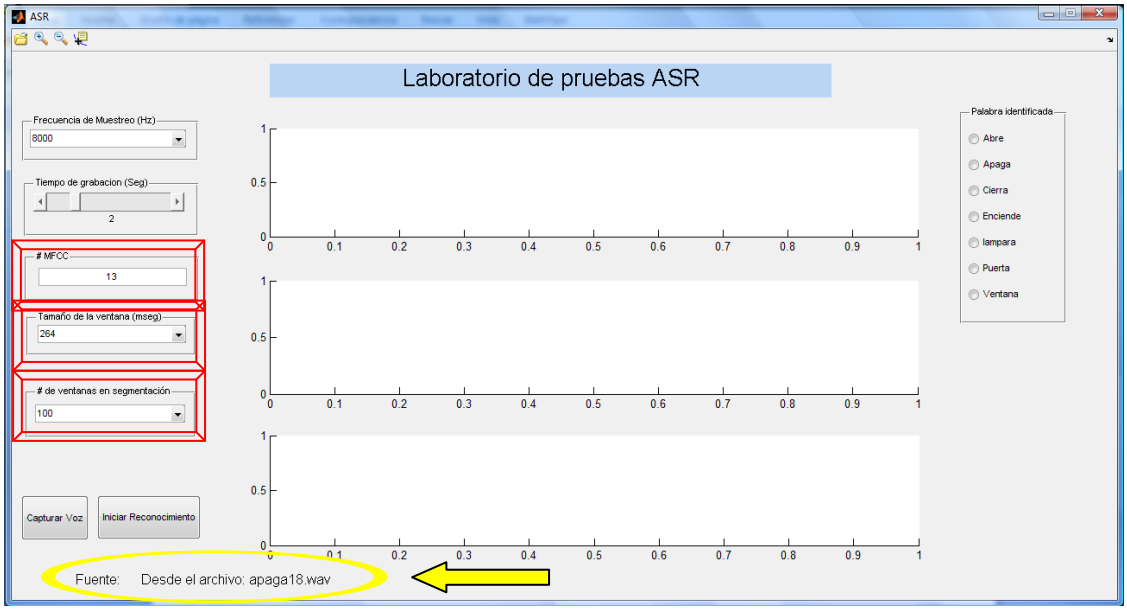


Figura 4.18.- Configuración de los parámetros para la simulación.

Nótese que en la parte inferior izquierda aparece la fuente de la señal de voz, para este caso es desde un archivo tipo 'wav' predefinido. El siguiente paso es iniciar el reconocimiento con el botón del mismo nombre, dando como resultado algo similar a lo que a continuación se presenta.

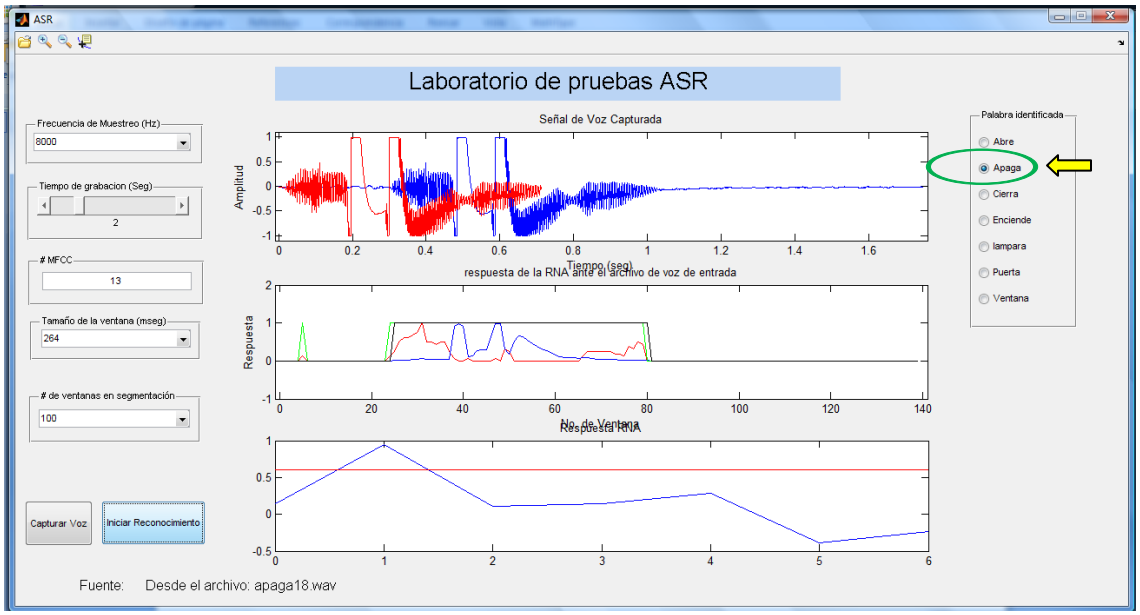


Figura 4.19.- Prueba al introducir un archivo de voz al laboratorio de pruebas.

La primer Gráfica muestra con azul la señal de voz original adquirida y con rojo la señal de voz segmentada, con la que se trabajará para su reconocimiento, se puede observar una reducción de más de la mitad del número total de muestras, y que la señal resultante es verdaderamente la palabra en cuestión. En la segunda gráfica se observa con el color rojo los cruces por cero de la señal, con azul la energía, con verde la respuesta inmediata de la RNA encargada de segmentar la voz y con negro el resultado final del algoritmo, para esta última señal se aplicó un filtro de media mezclando las 5 muestras anteriores y posteriores a la ventana analizada. En la última gráfica se ve la respuesta de cada neurona ante la señal de entrada analizada, están ordenadas alfabéticamente, como complemento se agrega el arreglo de botones tipo radial que están en la parte izquierda de la ventana principal, Solo el botón que esté relacionada con la palabra reconocida se activará.

En el segundo método de simulación, mediante la grabación directa de la palabra, primero se tiene que seleccionar los parámetros de grabación de frecuencia y tiempo de muestreo, cuidando de igual forma los parámetros de la red neuronal que se probará. Después de seleccionar dichos parámetros, se prosigue a presionar el botón de capturar voz, después se presiona iniciar reconocimiento y los resultados serán similares a los mostrados con el método de simulación anterior.

4.4 Pruebas de Robustes

Con la finalidad de experimentar el sistema ante distintas condiciones para verificar su funcionalidad se realizaron varias propuestas que se verán en esta sección. El enfoque de estas pruebas esta dirigido más a la interación con el usuario final que el proceso de validación del algoritmo en sí, pero si pueden ser parte importante a considerar en trabajos futuros.

4.4.1 Palabras Similares

Primeramente se analizará la respuesta de una RNA de 5 neuronas de salida, pero con el caso en específico en el que las palabras tienen una pronunciación fonética similar, para el cual se tomaron como muestras cinco palabras en donde la única variante es la primera letra. Los resultados muestran un 72 % de las muestras identificadas correctamente.

Los resultados se aprecián gráficamente en la figura 4.14, los cuales no fueron del todo satisfactorios, ya que sólo 36 de las 50 grabaciones de prueba se identificaron correctamente. Este resultado se debe sin lugar a duda a que la mayoría de las ventanas finales contienen un espectro muy similar, por lo tanto una suma ponderada de diferencias muy estrecha. Entonces los primeros coeficientes de la primera ventana son los responsables de diferenciar entre palabras, pero no tienen el suficiente peso para modificar la salida de la RNA de una manera satisfactoria.

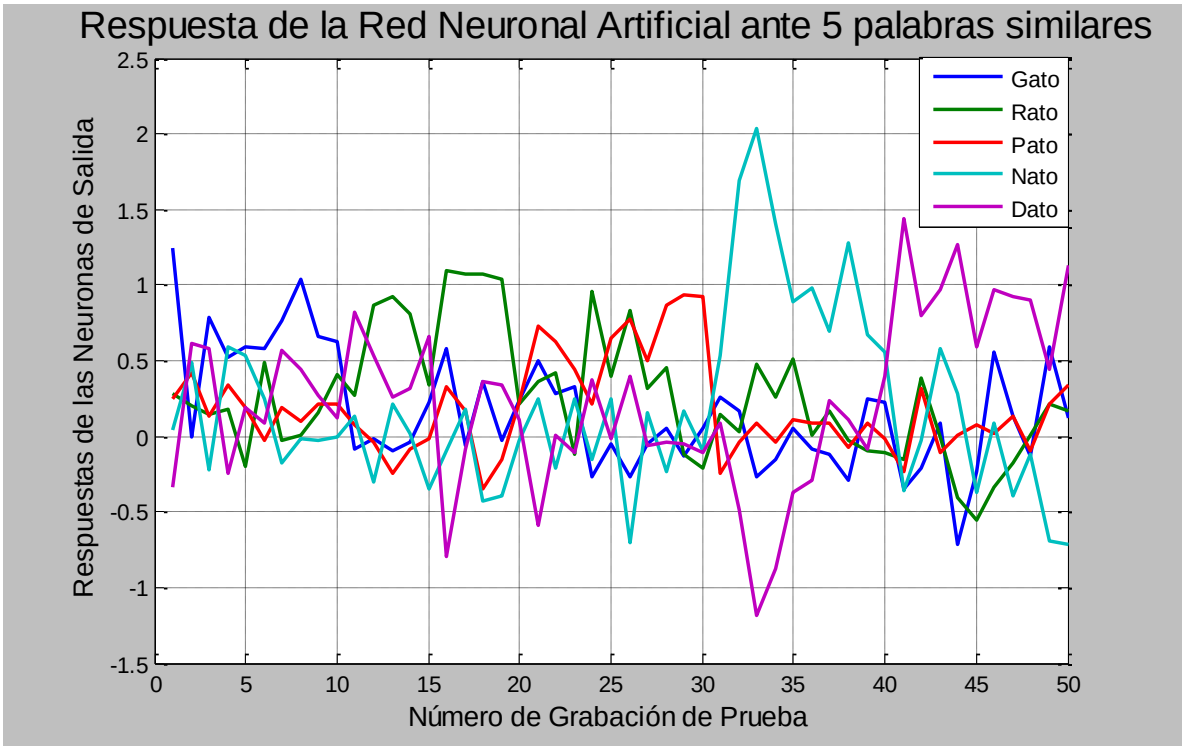


Figura 4.20.- Prueba robustez, ante 5 palabras de pronunciación similar.

4.4.2 Reducción de Procesamiento por Desctriptor Estadístico

Uno de los problemas más grandes que se presentaron durante el desarrollo del sistema, fué el gran tiempo invertido en el entrenamiento. Principalmente en la Fase 3, ya que al manejar una base de datos tan basta en información, entrenar la RNA se hacia una tarea tediosa y cansada, que aunque no se hacia de forma manual, sí privaba de desarrollar otras actividades.

Fue entonces que el maestro Guillermo Ronquillo Lomelí, hizó una observación de caso de prueba interesante, la cual disminuye dicho tiempo al realizar un análisis estadístico previo al entremaniento. El método más simple y que se incluye en este trabajo es el de realizar un promedio de todos los coeficientes cepstrales de mel pertenecientes a una misma palabra para poder tomar así sólo un vector que represente a todas las grabaciones de entrenamiento por cada palabra y reducir la matriz de entrenamiento considerablemente.

Los resultados no fueron los esperados, pero tampoco fueron del todo excluyentes, de hecho esta herramienta aunque no se implementó en el trabajo no se puede dar por descartada, ya que se puede utilizar otro método estadístico o ampliar el número de archivos de prueba que para este caso fueron 10 por cada palabra. La siguiente gráfica muestra la respuesta de la red neuronal al emplear este método.

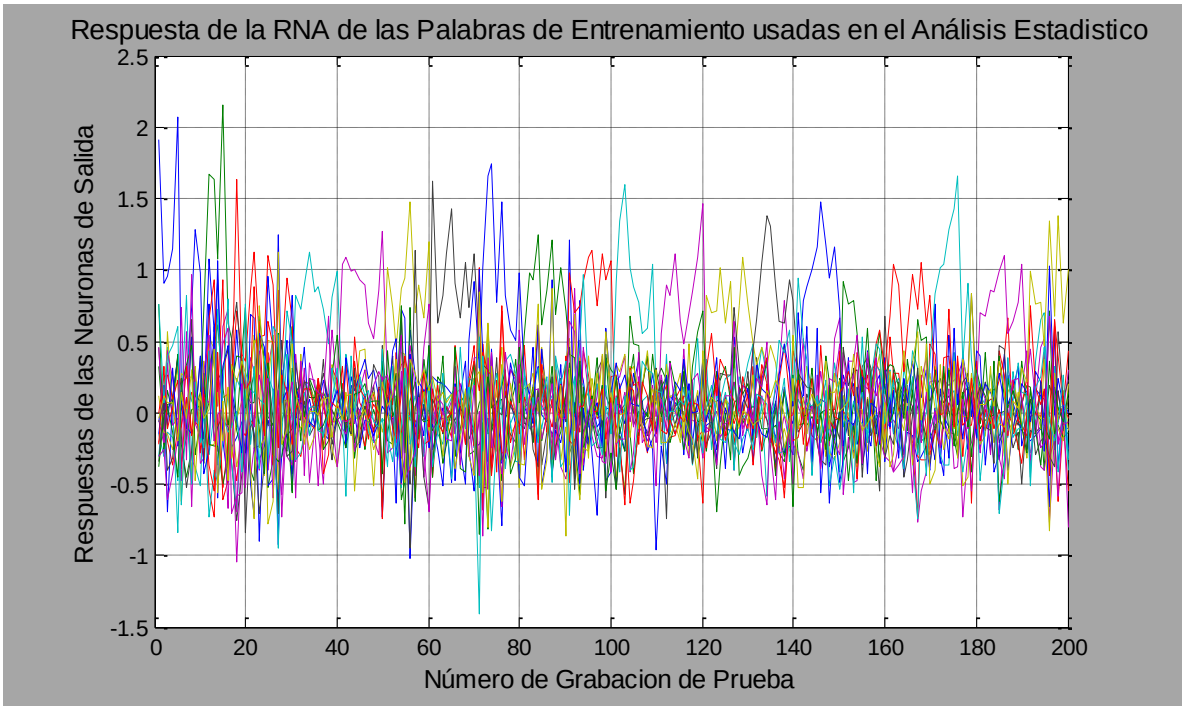


Figura 4.21.- Respuesta de la RNA ante las muestras de entrenamiento estadístico.

Se aprecia que para la gran mayoría de las grabaciones empleadas para obtener el vector promedio, se identifica correctamente la palabra. Mientras que en la gráfica 4.16 referente a las grabaciones de prueba, el resultado no es tan alentador.

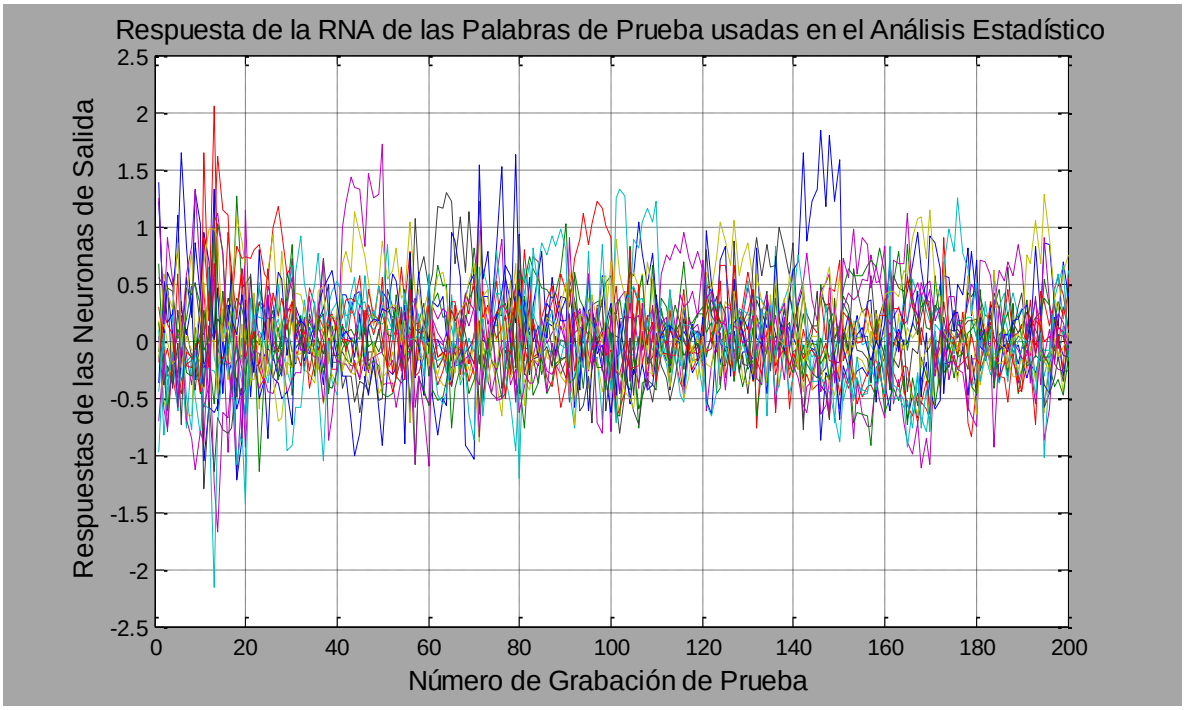


Figura 4.22.- Respuesta de la RNA ante las muestras de prueba estadístico.

4.4.3 Prueba ante Personas Diferentes

Con el objetivo de observar el comportamiento del sistema en distintas personas se requirió de la ayuda de individuos de prueba, hombre y mujer, los cuales amablemente se prestaron a realizar las base de datos personal empleando la interfaz gráfica de adquisición de voz, grabando archivos tanto de entrenamiento como de prueba. Resultando que no importa el género ni la persona siempre se llega a resultados exitosos de reconocimiento de comandos siempre y cuando se tenga una buena base de datos para entrenamiento, esto es, que se mencionen de una manera natural, clara y completa las palabras de comando.

Los resultados varían de un 80 – 90% de palabras identidificadas correctamente, se aplicaron los mismos parámetros de entrenamiento de la RNA que se obtuvieron en la caracterización descrita a principios de este capítulo, es son: 30 MFCC, el doble de neuronas en la capa oculta que las de entrada (90-120, dependiendo de la velicidad de pronunciación del individuo en custión), 20 comandos pronunciados, 10 grabaciones de entrenamiento y 10 de prueba.

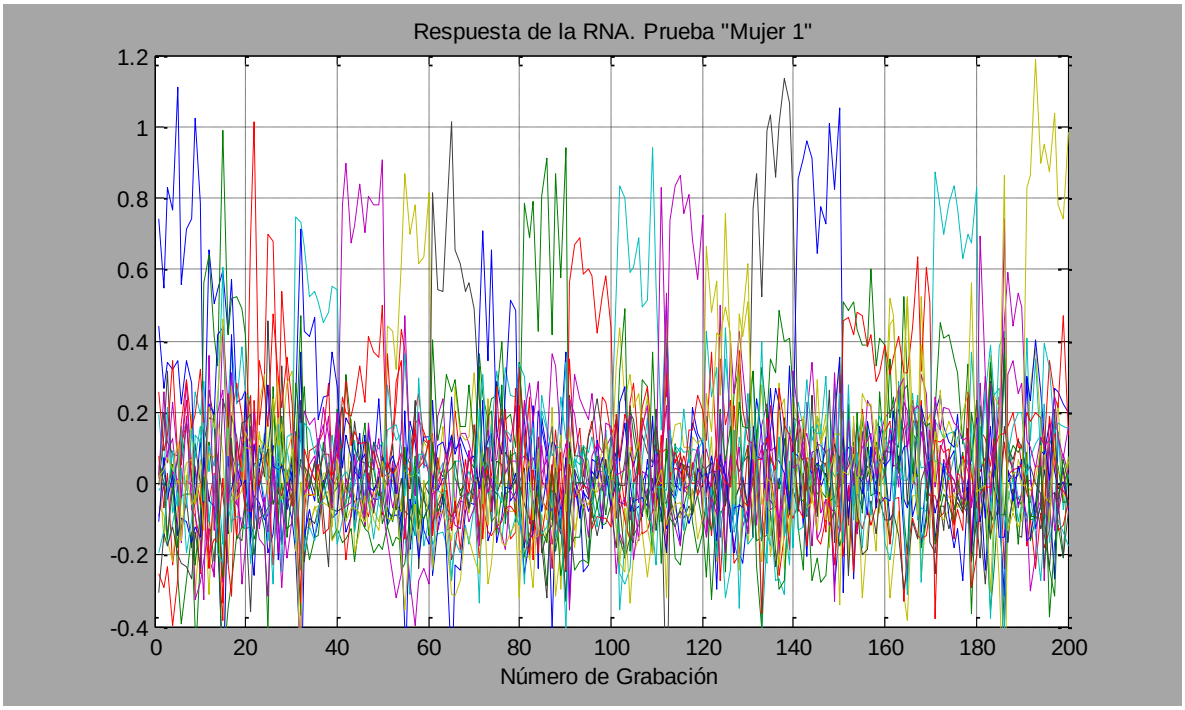


Figura 4.23.- Pruebas de robustez, con distintas personas. Respuesta RNA ante individuo 1.

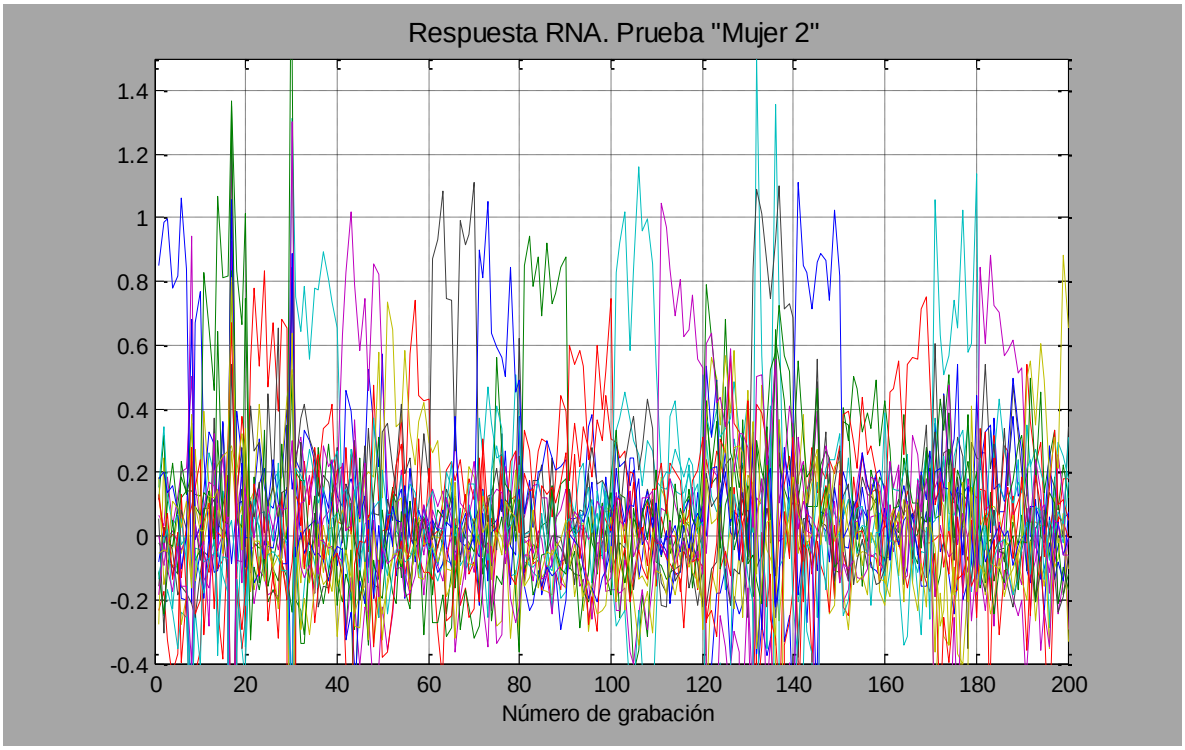


Figura 4.24.- Pruebas de robustez, con distintas personas. Respuesta RNA ante individuo 2.

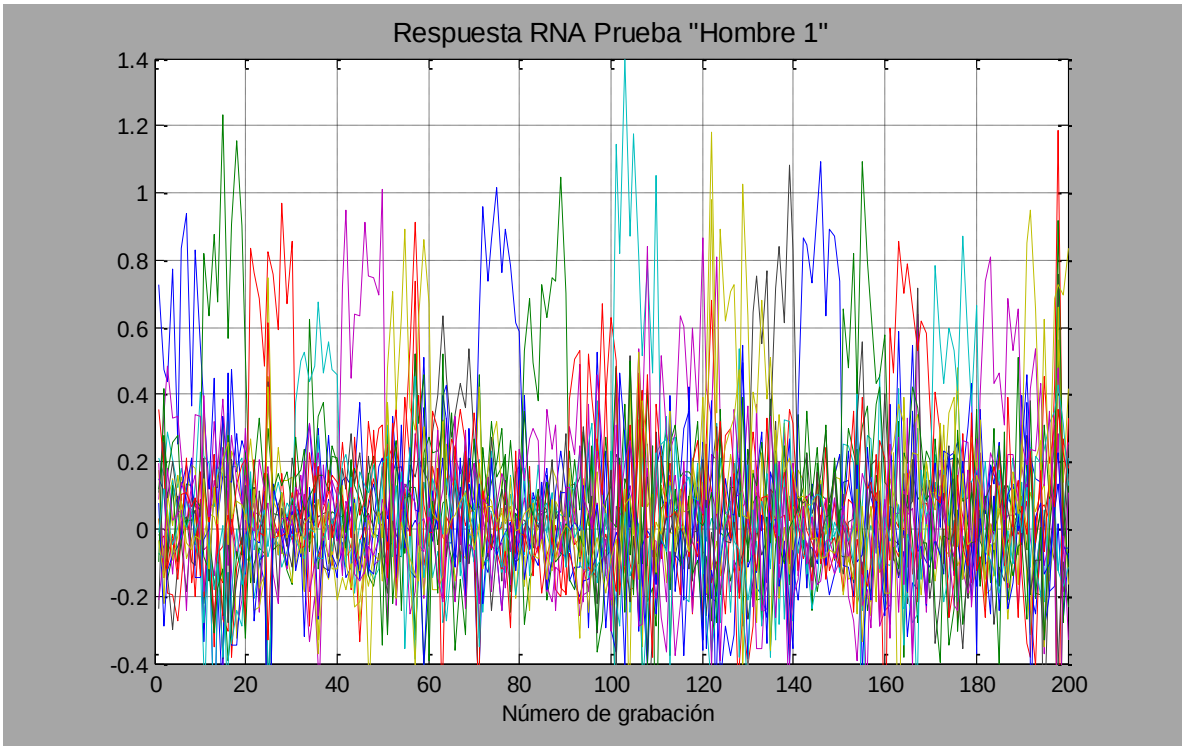


Figura 4.25.- Pruebas de robustez, con distintas personas. Respuesta RNA ante individuo 3.

4.5 Acoplamiento del Sistema.

En la figura 4.15 se muestra un sistema de reconocimiento de voz aplicado a la comunicación con el exterior, pensado para personas que tengan problemas de movilidad, como lo puede ser la paraplejia, el usuario final solo tiene que mencionar una palabra de comando y el sistema enviará la palabra escuchada a un módulo GSM a través de una interfaz tipo serial RS-232, el módulo codificará el mensaje y el destinatario puede ver la palabra o el mensaje generado por esa palabra en la pantalla de su teléfono celular.

También está pensado para hacer lo contrario enviar un mensaje a una casa u oficina en donde desde un teléfono celular se mande un comando a un módulo GSM huésped que a su vez active un sistema de activación de cargas.

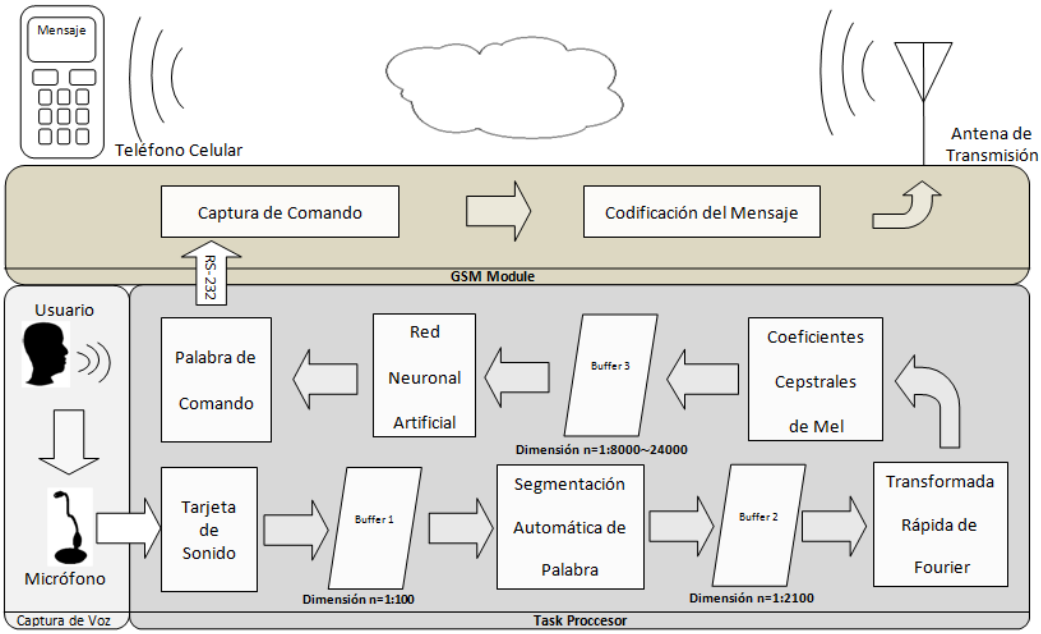


Figura 4.26.- Sistema de Reconocimiento de palabras a distancia con enlace GSM.

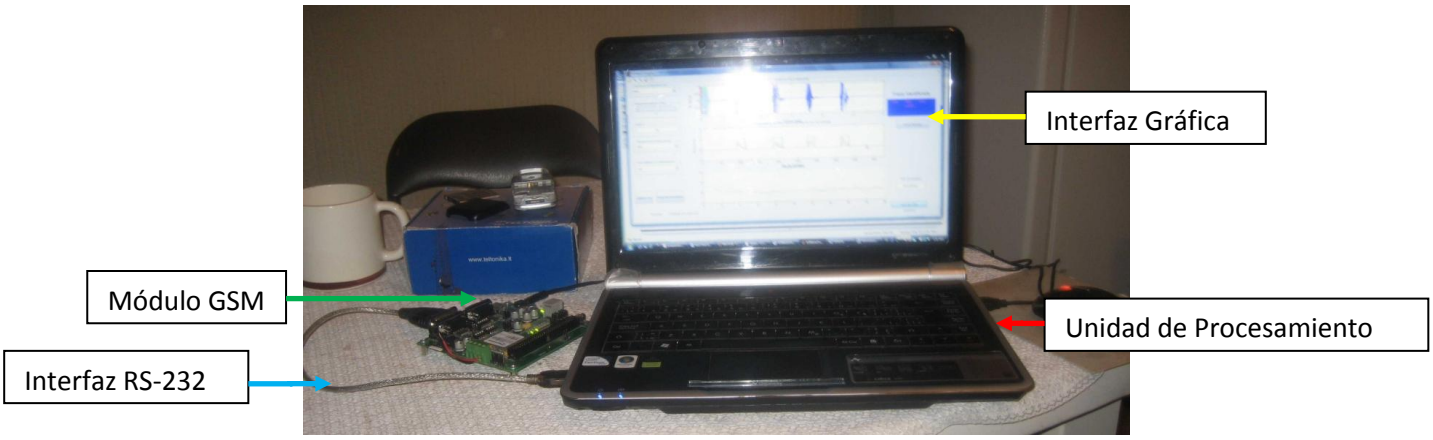


Figura 4.27.- Sistema Físico de Reconocimiento de palabras a distancia con enlace GSM.

4.6 Funcionamiento del Sistema.

A continuación de describirá paso por paso, como es que el usuario interactúa con el sistema. En primera instancia y antes de poder hacer el reconocimiento automático de palabras, es necesario crear la base de datos propia del usuario para así poder entrenar la red neuronal artificial, Esto se hace como previamente se describió en el capítulo anterior.

Teniendo una red neuronal artificial entrenada se prosigue a abrir la interfaz gráfica de usuario principal, dentro de esta se necesita configurar el tiempo de muestreo empleado, el número de Coeficientes Cepstrales de Mel empleados, el tamaño de las ventas tanto de segmentación como de filtrado, es muy importante que se configuren estrictamente los mismos parámetros con los cuales fue entrenada la RNA para que se tenga un funcionamiento real y eficiente.

El tiempo de grabación variable es un parámetro útil al momento de elegir entre frases de comando cortas o largas, después de definir este tiempo, es necesario accionar el botón de "Capturar Voz" (véase fig 4.20 (2)) y cuando aparezca la leyenda 'Grabando' debajo de dicho botón se prosigue a dictar la frase que se desea identificar.

Cuando se cuente con una grabación válida (esto es, que todas las palabras de comando mencionadas pertenezcan a la base de datos con la cual se entrenó la RNA) se presiona el botón de "Iniciar Reconocimiento"(fig 4.20 (3)) automáticamente se hace todo el procesamiento digital de voz y externan los resultados del procesamiento en las gráficas(fig 4.20 (4)), y también se muestra la frase reconocida en forma de texto (fig 4.20 (5)) .

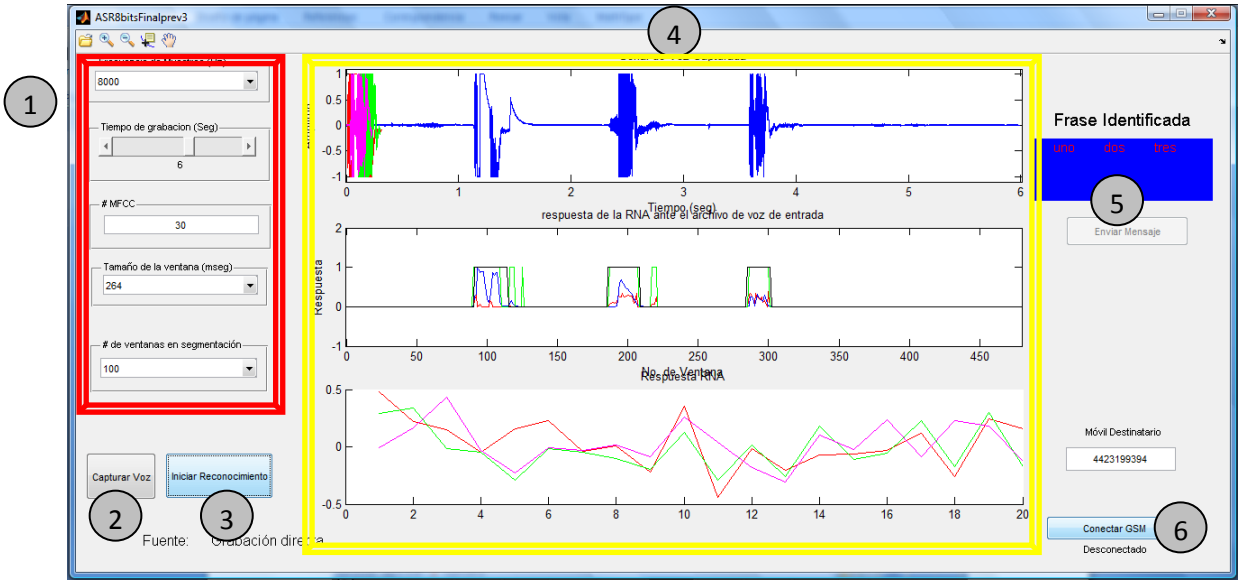


Figura 4.28.- Interfaz Gráfica: Detección de Palabras

Con la frase reconocida se prosigue a conectar el sistema a la red GSM para esto primero se tiene que activar la interfaz de comunicaciones serial RS-232 y verificar el funcionamiento del módulo GSM. Esto lo hace el usuario de forma manual con el botón de "Conectar GSM" (fig 4.20 (5)), si se llegara a detectar un error con cualquiera de las 2 comunicaciones, se mostrará en el cuadro de texto de estado, ubicado en la parte inferior del botón en juego (fig 4.21 (1)). Si todo resulta satisfactorio este cuadro de texto mostrará la leyenda "conectado" y el nombre del botón cambiará a "Desconectar" (fig 4.21 (2)), para poder hacer la desconexión manual, este proceso se puede hacer de forma automática, se muestra de forma manual para fines de práctica.

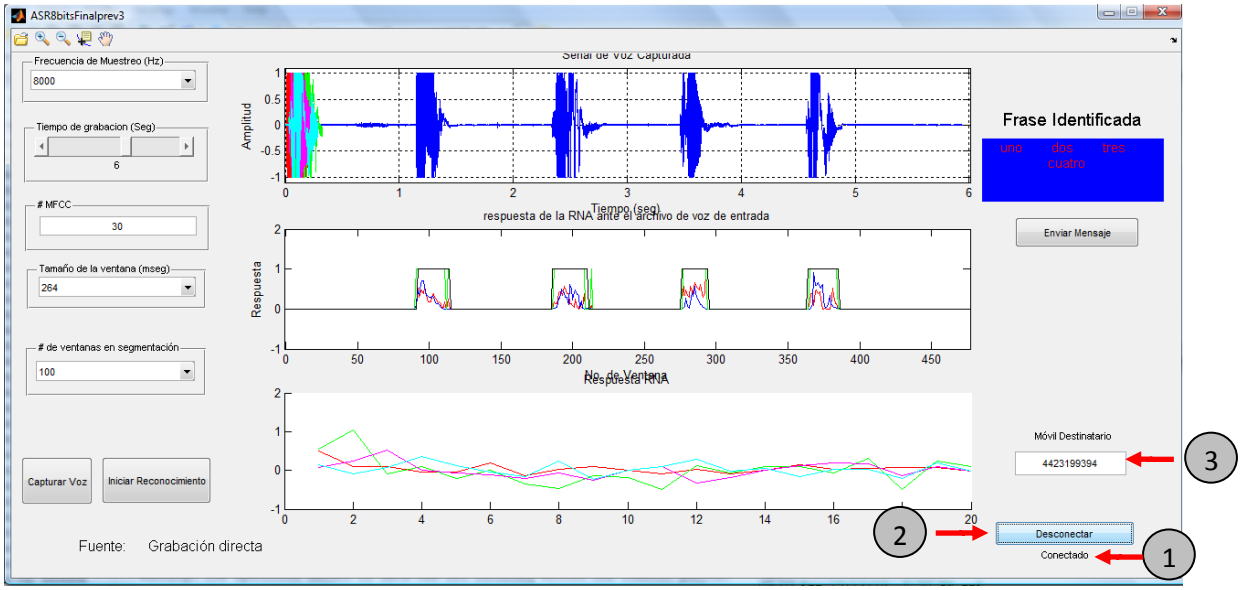


Figura 4.29.- Interfaz Gráfica: Establecer Conexión GSM.

Algo importante que se debe tener antes de mandar el mensaje reconocido es el número de teléfono móvil del destinatario (fig 4.21 (3)). Esté debe de ser un número de 10 dígitos, sin contener caracteres, si alguna de estas dos condiciones falla al momento de escribir el número, aparecerá el siguiente mensaje de error.

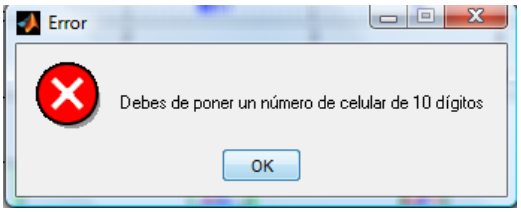


Figura 4.30.- Interfaz Gráfica: Error de Número de Destinatario.

Cuando se tiene un mensaje descifrado, un número de teléfono móvil válido y una conexión con el módulo GSM establecida, el paso final es enviar el mensaje de comando al destinatario. Si todo es correcto aparecerá el siguiente mensaje en la pantalla.

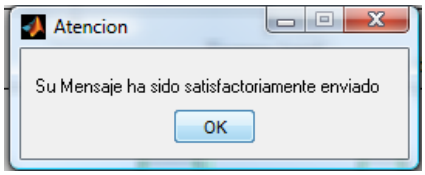


Figura 4.31.- Interfaz Gráfica: Mensaje Satisfactorio.

Muchos de estos pasos se pueden omitir al momento de emplear el sistema como producto, el único evento que no se puede omitir es el de dar la orden de inicio de grabación y claro está el mencionar la frase de comando. Pero hacer pasos intermedios nos permiten tener un mejor manejo de todo el entorno que envuelve al sistema.

Capítulo 5

Conclusiones.

En este último capítulo se describirá la configuración más óptima que conlleva al mejor reconocimiento para el sistema diseñado, las condiciones para que estos resultados sean consistentes y por último se menciona posible trabajo a futuro para realizar en esta rama.

En el estado del arte investigado es claro que el método de Modelos Ocultos de Markov es el más empleado actualmente en distintos dispositivos, por tal motivo se han desarrollado bastantes herramientas para su diseño y pruebas, pero aún con el porcentaje de acierto aceptable que pueden llegar a presentar, muchos de estos sistemas no son universales y su desempeño depende de gran parte de la articulación de las palabras, es así que es muy difícil reconocer el habla nativa de ciertas regiones en donde aunque se habla el mismo idioma, las palabras se pronuncian de una forma distinta.

Los resultados de emplear la extracción de características mediante los coeficientes de Mel y clasificar a través de las Redes Neuronales Artificiales muestran un proceso alterno pero eficaz de poder analizar y reconocer palabras completas como comandos.

La salida depende indudablemente de las condiciones de grabación como el ruido ambiental diferente al de entrenamiento y la sensibilidad del micrófono, es por eso que proceso de entrenamiento debe contener conjuntos grabados bajo distintas condiciones para que la red neuronal tenga una versatilidad apta al momento de clasificar la entrada.

En cuanto a la segmentación automática de voz se observó que para el inicio y fin de palabra el diseño es capaz de responder satisfactoriamente. Pero cuando existen palabras con sílabas separadas por demasiado silencio, se segmenta como si fueran 2 distintas palabras dando resultados falsos negativos, los falsos positivos se dan cuando un sonido ajeno al entrenamiento se captura, pero al aplicar el filtro de media posterior a la RNA, éste desperfecto se reduce al mínimo. Llegando a obtener resultados cercanos al 95%, una buena segmentación.

Se puede concluir que se cumple con el objetivo de la hipótesis, ya que cada uno de nuestros elementos teóricos aporta consistencia y efectividad al resultado final que es el de identificar un número finito de comandos.

Por una parte la segmentación automática ayuda indudablemente al sistema a desechar toda aquella señal que no pertenezca a una pronunciación de voz, se pudiera comparar su

funcionamiento con el de un filtro de energía y cruces por cero, permitiendo solo combinaciones válidas. El tener solo dos elementos de entrada por ventana también es una ventaja, y aunque se pueden agregar más definiciones que caractericen a una señal de voz como por ejemplo la varianza de la señal o algunos otros elementos probabilísticos, al hacer esto sin duda el nivel de detección de voz neta se incrementaría, aunque el tiempo de procesado tendría un aumento considerable.

El emplear el algoritmo de redes neuronales artificiales para segmentar la palabra es poco visto en el estado del arte ya que por el hecho de que se trabaja más con los Modelos Ocultos de Markov para el reconocimiento, se emplean otros métodos enfocados también a la probabilidad de eventos. El usar una red neuronal explota una de sus cualidades como lo es la robustez que da mayor seguridad de segregar conjuntos extraños a los de entrenamiento.

El resultado de emplear esta segmentación es bueno aunque se puede optimizar al encontrar una configuración de red que se amolde al tipo de señal, esto es sin duda una tarea difícil ya que no se puede probar el resultado de una red, hasta tenerla diseñada y para poder obtener la mejor sería necesario implementar una gran variedad de estas antes de dar con la adecuada.

Por otro lado el uso del método de los Coeficientes Cepstrales de Mel también sirve como un filtro que adapta la señal de entrada a como el ser humano la percibe, esto marca una segunda parte eliminación de sonidos que el humano no asimila y que nos innecesarios en el proceso de reconocimiento. Otra ventaja de utilizar este método es la reducción de la señal original a unos cuantos coeficientes que sin duda es precisa para un tiempo de procesado menor al momento de clasificar el conjunto total.

En la última etapa, la Red Neuronal Artificial clasifica la palabra segmentada en una de varias posibilidades con las que fue entrenada, dejando inactivas las salidas al momento que entra una señal desconocida. El valor del umbral se fija en base a la observación de resultados previos, pero en las gráficas mostradas se puede observar que para señales extrañas el nivel de las salidas no sobrepasa el 50% en su mayoría.

El objetivo de emplear una red neuronal artificial en la segmentación y otra en la etapa clasificación fue un para probar como es el comportamiento de éstas en cascada, que aunque tienen un elemento intermedio que no pertenece a este género, si da una idea de que se pueden utilizar redes interconectadas como en el cerebro humano para resolver un problema exponiendo buenos resultados como conjunto. El motivo de no utilizar una RNA única que tenga como entrada la señal de voz y como salida las posibles palabras entrenadas, es a causa del exceso en los datos de entrada que son mínimamente 8000 por segundo, esto implicaría demasiadas multiplicaciones y un tiempo de procesado que sobrepasaría nuestras necesidad de hacerlo en tiempo real.

Se pudo constatar que el arreglo de la red neuronal artificial influye demasiado en la efectividad del sistema, pero con las pruebas realizadas se logró una buena taza de reconocimiento.

El código del software diseñado queda a disposición de los interesados para hacer de su labor de investigación cómoda y sencilla.

5.1 Trabajo futuro

En esta sección se describe el objetivo al que se enfoca este trabajo, así como ideas y propuestas de futuras aplicaciones o complementos que se pueden llevar a cabo con estas u otra técnicas.

Para hacer sistemas de alta calidad es necesario trabajar con sistemas en tiempo real en donde se aprovechen al máximo cada ciclo de reloj y se realice el mayor número de cálculos entre toma de muestras, esto nos da un sistema de rápida respuesta.

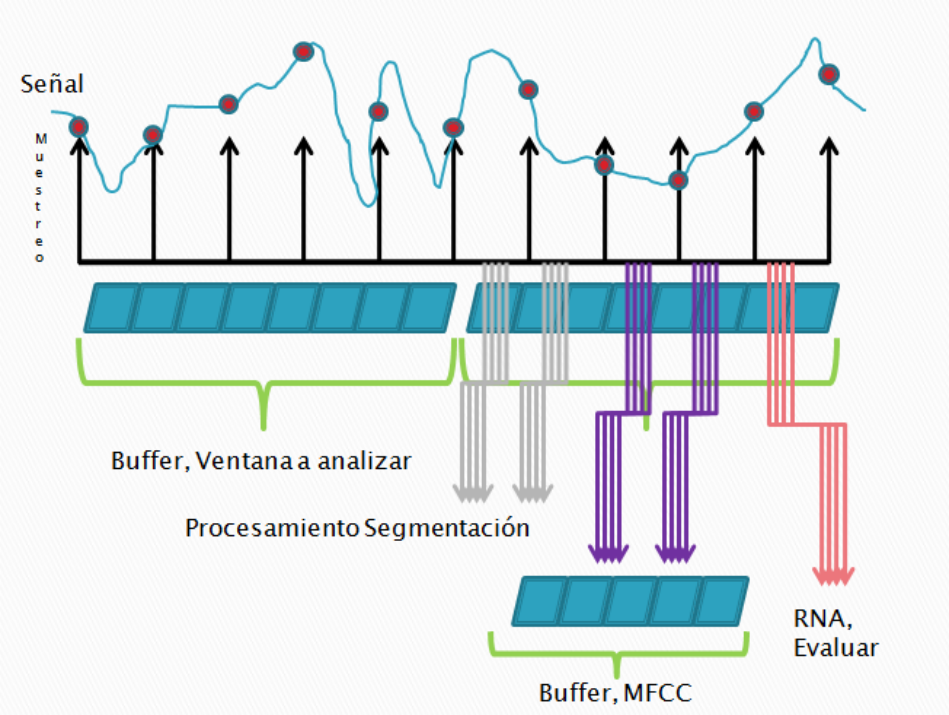


Figura 5.1.- Adquisición en tiempo real.

Una investigación alterna para emplear los métodos de esta tesis es el reconocer la voz mediante fonemas, sílabas ó letras. Para este caso un buen reconocimiento se vería reflejado en la identificación de palabras continuas sin necesidad de tener un universo de palabras en la base de datos, solo algunos sonidos conjuntamente con reglas y modelos de pronunciación, así el sistema será capaz de descifrar cualquier palabra pronunciada y no solo un conjunto previamente entrenado.

También una buena práctica es probar con un método de extracción de características distinto como los vistos en el capítulo dos, en los que también se intenta describir la voz humana a través de distintas herramientas matemáticas.

Bibliografía

Adjoudj Réda, Boukelif Aoued, "*Artificial Neural Network & Mel-Frequency Cepstrum Coefficients-Based Speaker Recognition".* SETIT 2005. 3rd International Conference: Sciences of Electronic. Technologies of Information and Telecommunications, March 27-31, 2005 – TUNISIA

Alegre F. L. "Aplicación de RNA y HMM a la Verificación Automática de Locutor". IEEE Latin America Transactions, Vol. 5, No. 5, September 2007, Pp 329-337.

Alvarado Valderrama Jorge Edilberto. Reconocimiento de Palabras Aisladas utilizando MFCC y Dynamic Time Warping. Escuela Académico Profesional de Informática Universidad Nacional de Trujillo, Perú. 2008. jorgealvarado@seccperu.org

Álvarez Mauricio; Germán Castellanos. "Selección De Características Usando HMM Para La Identificación De Patologías De Voz". Revista Scientia Et Technica Año X, No X, Mes 200x. Utp. Issn 0122-1701 1. Pp.1-4. Malvarez@Ohm.Utp.Edu.Co, Gcastell@Ieee.Org.2004.

A. S. Kolokolov. "Preprocessing and Segmentation of the Speech Signal in the Frequency Domain for Speech Recognition". Automation and remote control, Volume 64, Number 6. June, 2003, pp. 985-994.

Benesty Jacob, M. Mohan Sondhi, Yiteng Huang. "Springer Handbook of Speech Processing", Springer-Verlag Berlin Heidelberg 2008

Biblioteca de Consulta Microsoft® Encarta® 2006. Microsoft Corporation. Reservados todos los derechos.

Castellanos Moisés. Federico, Jesús. "Reconocimiento de voz con redes neuronales". ACM Transactions on Computational Logic. Vol. 5. No. N. 2007.

Clemente Eduardo. Vargas, Alcira. Olivier, Alejandra. Kirschning, Ingrid. Cervantes, Ofelia. Entrenamiento y evaluación de reconocedores de voz de propósito general basado en redes neuronales "feed – forward" y modelos ocultos de Markov. TLATOA – CENTIA. Vol. 15. 1999.

Chang-Wen Hsu. Lin-Shan Lee. Higher Order Cepstral Moment Normalization for Improved Robust Speech Recognition. IEEE Transactions On Audio, Speech, And Language Processing, vol. 17, No. 2, February 2009.

Chen Tsuhan. Audiovisual Speech Processing "Lip Reading and Lip Synchronization". IEEE Signal Processing Magazine, January 2001. pp 9-21.

Chakraborty. P. F. Ahmed. Md. Monirul Kabir. Md. Shahjahan1. Kazuyuki Murase. An Automatic Speaker Recognition System. University of Fukui, Bunkyo, Fukui, Japan. Springer-Verlag Berlin Heidelberg 2008. pp.517-526. jahan@eee.kuet.ac.b. murase@synapse.his.fukui-u.ac.jp.

Ching-Tang Hsieh, Mu-Chun Su, Eugene Lai and Chih-Hsu Hsu . "A Segmentation Method for Continuous Speech Utilizing Hybrid Neuro-Fuzzy Network". Department of Electrical Engineering, Tamkang University. Taipei Hsien, Taiwan 251, R.O.C. Journal of Information Science and Engineering 15, pp. 615-628. 1999.

Davis K. H., R. Biddulph, and S. Balashek, Automatic Recognition of Spoken Digits, J. Acoust. Soc. Am., Vol 24, No. 6, pp. 627-642, 1952.

Dhanalakshmi P; S. Palanivel, V. Ramalingam. "Classification of audio signals using SVM and RBFNN", Expert Systems with Applications, 36 (2009) pp. 6069–6075. ELSEVIER.

Donald Christiansen. Charles Alexander. Standard Handbook of Electronic Engineering. McGraw-Hill 5th Ed. 2005.

Esparza Arellano María Elena. J. Benito Avalos Briseño. Reconocimiento de Voz. Instituto Tecnológico de Aguascalientes. jbenitomx@yahoo.com.mx

Faudez Zanuy, Marcos. “Tratamiento digital de voz e imagen y aplicación a la multimedia”. Marcombo. Boixareu editors .pp.97. 2000

Franco Gasca Luis Alfonso. Apuntes de Procesamiento Digital de Señales, Verano de la Maestría en Instrumentación y Control 2009. Universidad Autónoma de Querétaro.

Fonseca Yerena Socorro. Comunicación Oral. Fundamentos y Práctica estratégica. Prentice Hall, 2da ed. 2003.

Forney G. David, Jr. “The Viterbi Algorithm: A Personal History”. Signal Processing Magazine, IEEE, Vol 23. July 2006.MIT Cambridge, MA 02139 USA forneyd@comcast.net. pp. 120-142

Fry D. B.; P. Denes, The Design and Operation of the Mechanical Speech Recognizer at University College London, J. British Inst. Radio Engr., Vol. 19, No. 4, pp. 211-229, 1959.

Gómez Rojas, Germán Alonso. Henao López, Juan Carlos. Salazar Isaza, Harold. Entrenamiento de una red neuronal artificial usando el algoritmo simulated annealing. Scientia et technica. Vol. X. No. 24. 2004.

Guitart Jorge M. Sonido y Sentido. Teoría práctica de la pronunciación del español. Georgetown Studies in Spanish Lingustics series. Washington, D.C. 2001.

Gyucheol Jang. Sooyoung Woo. Chang D. Yoo. “Voice Segmentation Algorithm”. Proceedings of ICSP 2001. Agust 22-24, 2001 Daejon, Korea. Korea Advanced Institute os Science and Technology 373-1 Kusong-dong. Yousong-u. Taejon 305-701.Korea.

Jelinek F; L. R. Bahl, and R. L. Mercer, Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech, IEEE Trans. On Information Theory, Vol. IT-21, pp. 250-256, 1975

Jian-Da Wu, Bing-Fu Lin. “Speaker identification based on the frame linear predictive coding spectrum technique”. Expert Systems with Applications 36 (2009) pp. 8056–8063. ELSEVIER.

Jordi Adell. Antonio Bonafonte “Towards Phone Segmentation for Concatenatenative Speech Synthesis”.5th ISCA Speech Syntesis Workshop – Pittsburg, 2004.

Juang B.H.; Lawrence R. Rabiner. Automatic Speech Recognition – A Brief History of the Technology Development. Georgia Institute of Technology, Atlanta Rutgers University and the University of California, Santa Barbara. 2004.

Kaschel C., Héctor. Watkins, Francisco. Sanjuán U. Enrique. Comprensión de voz mediante técnicas digitales para el procesamiento de señales y aplicación de formatos de comprensión de imágenes. Revista Facultad de Ingeniería. Universidad De Tarapacá. Vol. 13. No. 3. 2005.

Kotti Margarita. Vassiliki Moschou. Constantine Kotropoulos. Speaker segmentation and clustering. Artificial Intelligence and Information Analysis Lab, Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece, Elsevier 2007.

Lippmann R. P. Review of Neural Networks for Speech Recognition, Readings in Speech Recognition, A. Waibel and K. F. Lee, Editors, Morgan Kaufmann Publishers, pp. 374-392, 1990

Lowerre B. The HARPY Speech Understanding System, Trends in Speech Recognition, W. Lea, Editor, Speech Science Publications, 1986, reprinted in Readings in Speech Recognition, A. Waibel and K. F. Lee, Editors, pp. 576-586, Morgan Kaufmann Publishers, 1990.

Marín Jorge I.; Pablo A. Muñoz; Francisco J. Ibargüen, Reconocimiento De Comandos De Voz Usando La Transformada Wavelet Y Máquinas De Vectores De Soporte. Revista Scientia Et Technica Año Xii, No 31, Agosto De 2006 Utp. Issn 0122-1701, Pp 35-40. Jorgemarin@Uniquindio.Edu.Co, Pabloandresm@Yahoo.Com, Fjibarg@Yahoo.Com, Grupos Gama Y Gdsproc Ceifi, Facultad De Ingeniería Universidad Del Quindío.

Masanobu Nakamura. Koji Iwano. Sadaoki Furui. Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan. Elsevier 2007.

Metin Akay. Handbook of Neural Engineering. IEEE Press Editorial Board. 2007

Michael Gerber. Tobias Kaufmann. Beat Pfister. Perceptron-Based Class Verification. Speech Processing Group Computer Engineering and Networks Laboratory ETH Zurich, Switzerland. M. Chetouani et al. (Eds.): NOLISP 2007,

Nagata K.; Y. Kato, and S. Chiba, Spoken Digit Recognizer for Japanese Language, NEC Res. Develop., No. 6, 1963.

Oropeza Rodríguez, José Luis. Suárez Guerra, Sergio. Algoritmos y método para el reconocimiento de voz en español mediante sílabas. Computación y sistemas. Vol. 9. No. 3. 2006.

Palmer, A. Montaño, J.J. Jiménez, R. Tutorial sobre Redes Neuronales Artificiales: El Perceptrón Multicapa. Área de Metodología de las Ciencias del Comportamiento. Facultad de Psicología. Universitat de les Illes Balears. Revista Electrónica De Psicología. Vol. 5, _o. 2, Julio 2001.www.Psicologia.com .alfonso.palmer@uib.es

Pardo, José D. Castro,. José A. Iseda, Gilberto P. Torres M. César. Mattos, Lorenzo. Reconocimiento automático del habla utilizando la transformada de Fourier y redes neuronales. Revista colombiana de física. Vol. 38. No. 4. 2006.

Press William H. Saul A. Teukolsky. William T. Vetterling. Brian P. Flannery.Numerical Recipes in C. The Art of Scientific Computing. Cambridge University Press 2da ed. 1997.

Priyabrata Sinha. "Speech Processing in Embedded Systems", Springer Science Business Media. pp.41. 2009.

Proakis John G. Dimitris G. Manolakis. Digital Signal Processing. Principles, Algorithms, and Applications. Prentice-Hall. 3ra ed. 1996.

Rabiner Lawrence. Biing-Hwang Juang, "Fundamentals of Speech Recognition", Prentice Hall. 1993

Sakai J.; S. Doshita, The Phonetic Typewriter, Information Processing 1962, Proc. IFIP Congress, Munich, 1962.

Sakoe H.; S. Chiba, Dynamic Programming Algorithm Quantization for Spoken Word Recognition, IEEE Trans. Acoustics, Speech and Signal Proc., Vol. ASSP-26, No. 1, pp. 43-49, Feb. 1978.

San Martín S. César. Carrillo A., Roberto. Implementación de un reconocedor de palabras aisladas dependiente del locutor. Revista Facultad de Ingeniería. U.T.A. Vol. 12. No. 1. 2004.

Serajul Haque. Roberto Togneri. Anthony Zaknich. Perceptual features for automatic speech recognition in noisy environments. School of Electrical, Electronic and Computer Engineering, University of Western Australia, Crawley, Australia. Elsevier. 2008.

Soriano Mas Carles, Gemma Guillazo Blanch, Diego Antonio Redolar Ripoll, Meritxell Torras García, Anna Vale Martínez, Fundamentos de Neurociencia, ed. UOC, 2004

Stiefelhagen, R.; C. Fuegen, P. Gieselmann, H. Holzapfel, K. Nickel, A.Waibel. Natural Human-Robot Interaction using Speech, Gaze and Gestures, IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004.

Suzuki J.; K. Nakata, Recognition of Japanese Vowels—Preliminary to the Recognition of Speech, J. Radio Res. Lab, Vol. 37, No. 8, pp. 193-212, 1961.

Todor Ganchev. Nikos Fakotakis. George Kokkinakis. Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task. Wire Communications Laboratory. University of Patras, Greece. 2004. tganchev@wcl.ee.upatras.gr

Toledano Torre Doroteo, Luis A. Hernández Gómez, Member, IEEE, and Luis Villarrubia Grande. "Automatic Phonetic Segmentation". IEEE Transactions on Speech and Audio Processing, vol. 11, no. 6, november 2003.

Viver, Javier. PHILIPS: Intelligent speech interpretation – la tecnología inteligente de reconocimiento de voz. Procesamiento del lenguaje natural. No. 35. 2005.

Weifeng Li. Kenichi Kumatani. John Dines. Mathew Magimai-Doss. Herv´e Bourlard. A Neural Network Based Regression Approach for Recognizing Simultaneous Speech. IDIAP Research Institute, Martigny, Switzerland. Springer-Verlag Berlin Heidelberg. 2008.

www.es.wikipedia.org.

Yang Shao. Soundararajan Srinivasan. Zhaozhang Jin. DeLiang Wang. A computational auditory scene analysis system for speech segregation and robust speech recognition. The Ohio State University, Columbus. USA. Elsevier. 2008.

Yu Hen Hu, Jenq-Neng Hwang. "Handbook of Neural Network Signal Procesing". CRC Press LLC.2002.