UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE INGENIERÍA MECÁNICA

DESARROLLO E IMPLEMENTACIÓN DE UN SISTEMA DE RECONOCIMIENTO AUTOMÁTICO DEL HABLANTE MEDIANTE REDES PERCEPTRÓN MULTICAPA, PARA MEJORAR LA TASA DE RECONOCIMIENTO

INFORME DE SUFICIENCIA

PARA OPTAR EL TÍTULO PROFESIONAL DE: INGENIERO MECATRÓNICO

ALCIDES GUILLERMO JOO AGUAYO

PROMOCIÓN 2010 - I

LIMA - PERÚ

2013

Dedicado a *mi madre Julia,*
y a mis hermanos David y Verónica.

Agradecimientos

A mis profesores Julio Casquero e Iván Calle por su apoyo para el desarrollo del presente trabajo, asesoría y motivación constante. Al Ing. Fredy Sotelo Valer por su certera asesoría para el presente trabajo.

A mis compañeros de trabajo y grupo del curso de titulación por sus comentarios valiosos, en especial a Felipe Gómez por su aporte tan crítico y apegado al carácter de investigación de este trabajo.

A mis compañeros de trabajo por apoyarme durante las horas que estuve ausente para poder realizar mi investigación.

A las personas que colaboraron conmigo permitiéndome grabar sus voces para poder realizar los ensayos y pruebas de reconocimiento de voz: Diana, Juan, Mario, Carlos y Smith, sin sus voces no se habría podido realizar ningún testeo de reconocimiento, por su tiempo y apoyo mis agradecimientos.

A Dios y al universo por haber conspirado para mantenerme firme y no decaer a pesar las adversidades que se presentaron durante el desarrollo de la investigación que necesito de esfuerzo y dedicación y que comprende la aplicación de los conocimientos adquiridos para la carrera como Ingeniero Mecatrónico.

TABLA DE CONTENIDOS

LISTA DE FIGURAS

PRÓLOGO

Una de las características importantes que identifica a una persona es su voz y nosotros los seres humanos usamos la voz para identificar a las personas así como el género de cada una de ellas. Este hecho se puede emplear para el diseño de sistemas de reconocimiento automático los que podrían reemplazar a los métodos tradicionales como el uso de tarjetas de identificación, claves de acceso, etc. A pesar de la existencia de sistemas comerciales de reconocimiento basados en la voz, el desarrollo de un sistema alternativo no solo nos permite apreciar la aplicación del procesamiento de señales e inteligencia artificial en el reconocimiento del hablante sino que nos permite el desarrollo de tecnología nacional aplicada y orientada a la solución de problemas reales.

El presente trabajo consiste en el diseño e implementación de un sistema automático de reconocimiento basado en la voz usando una red perceptrón multicapa. Para tal fin, la presente tesis se ha organizado en cinco capítulos, los que están expuestos en el orden siguiente:

En el capítulo 1, se mencionan los antecedentes del informe de sustentación de tesis, mencionando los trabajos previos y las referencias de aplicación existentes los cuales de alguna manera inspiraron y guiaron la presente investigación. Se

plantean los objetivos a los que se desean llegar en el presente informe de suficiencia. Se explica la justificación y alcances de este trabajo para concluir este capítulo listando los recursos empleados para desarrollar la investigación.

En el capítulo 2, se describe el sistema de reconocimiento, que partes tiene y en que consiste este sistema. Así mismo se describe el cómo es que el sistema realiza el reconocimiento de voz.

En el capítulo 3, se realiza la identificación del problema, dando detalles de cuáles son los límites dentro de los cuales se desea enmarcar esta investigación. Una vez detallado el problema se expone la hipótesis de trabajo.

En el capítulo 4, se exponen los fundamentos teóricos necesarios, de los cuales se estudiara en primer lugar las principales características de la voz humana. Se hará un breve estudio fisiológico de la voz, analizando su proceso de formación y sus principales características. Luego se hará un estudio del pitch y de las frecuencias formantes, ya que son las características que son únicas en cada persona y por tal basándonos en estas medidas podemos distinguir a una persona de otra.

Se presentaran un conjunto de algoritmos del área de procesamiento de señales para la etapa de pre procesamiento y para la etapa de extracción de características. De esta manera, será posible representar de manera compacta y única la señal de voz de cada persona, y esta nueva representación será usada más adelante

como las entradas del sistema de reconocimiento tanto para la etapa de entrenamiento como para la etapa de prueba y funcionamiento.

Luego de estudiar la voz y el tratamiento de señales de voz, se presenta una reseña de la red perceptrón multicapa analizando las propiedades que la hacen adecuadas para la tarea de clasificación y reconocimiento. Se mostrara el algoritmo de optimización usado para la tarea de optimización de la red. Se discutirá la función de error adecuada para la tarea de clasificación y se mostrara el algoritmo de back-propagation como un método computacionalmente eficiente para el cálculo del gradiente de la función de error.

Se concluye el capítulo de fundamentos teóricos dando una breve explicación de los sistemas de reconocimiento, según las ciencias computacionales como parte del enfoque que permitirá comprender el panorama de esta investigación.

En el capítulo 5, se explica la estructura del sistema de reconocimiento, como se realiza la adquisición de señales de voz, la obtención de la señal pre-procesada, la extracción de características de las señales de voz, la obtención del sistema de reconocimiento y la comprobación de la calidad del sistema de reconocimiento de patrones de voz usando redes perceptrón multicapa.

Para la comprobación de la calidad se muestran las pruebas y resultados del sistema de reconocimiento. En este capítulo se muestra los patrones característicos de cada persona, y se hace un análisis de la red neuronal para encontrar los

parámetros que resulten en una mayor tasa de reconocimiento. Finalmente para comprobar se hace un análisis del funcionamiento del sistema en tiempo real con la finalidad de verificar el correcto funcionamiento del sistema.

Para concluir el informe, se presentan las Conclusiones y Recomendaciones encontradas en el desarrollo del presente trabajo y se adjuntan los anexos pertinentes para complementar la investigación.

Este informe es útil porque nos permitió no solo optar por el título de ingeniero sino también porque al momento de realizarlo se aplicaron varios de los tópicos para el área de ciencias computacionales permitiendo realizar un sistema sencillo pero eficaz para realizar el reconocimiento de voz, este trabajo permitirá a las nueva generaciones tener un referencia para seguir ampliando las investigaciones sobre este tema tan interesante.

CAPÍTULO 1

INTRODUCCIÓN

En los últimos años se viene dando un auge de los sistemas denominados inteligentes, debido a la aplicación de la inteligencia artificial, en este ámbito tenemos a los sistemas de reconocimiento de patrones como parte de un enfoque que se ocupa de los procesos sobre ingeniería, computación y matemáticas relacionados con objetos físicos o abstractos, cuyo propósito de extraer información que nos permita establecer propiedades entre conjuntos de dichos objetos para poder relacionarlos, clasificarlos y ordenarlos de alguna manera.

Más adelante se verá que los estudios, indican que la voz obedece a ser representada mediante una serie de valores característicos siendo objeto de ser asignada a través de patrones o conjuntos característico; estos patrones pueden ser entrenados en una red neuronal y aplicando el reconocimiento de patrones al conjunto de entrenamiento podemos realizar una clasificación o un reconocimiento de voz del hablante. Finalmente implementaremos un sistema de reconocimiento del hablante que nos permita probar tal hecho, de modo tal que se utilizan las diversas herramientas del procesamiento digital de señales para el tratamiento de la voz así

como diversas herramientas de la inteligencia artificial tales como las redes neuronales multicapa y el reconocimiento de patrones.

1.1. Antecedentes

Se tiene la necesidad de desarrollar sistemas que permitan reconocer la voz del hablante utilizando nuevas formas que permitan mejorar la capacidad de reconocimiento, esto se traduce en aumentar la efectividad de reconocimiento, así mismo que se desarrollen de una forma rápida y sencilla y que signifiquen una aplicación directa de los fundamentos adquiridos durante la etapa de pregrado. Se requiere un sistema que sea capaz de reconocer a un sujeto por la palabra que emite, que este basado en una red perceptrón multicapa y que si atas de reconocimiento sea mayor al 70%. Para el presente informe es importante tener en cuenta los siguientes antecedentes.

1.1.1. Antecedente 1

a) Apellidos y nombres: Luna Ortega Carlos Alejandro, Martínez Romo Julio Cesar, Mora González Miguel.
b) Lugar o Institución donde se hizo la investigación: Universidad Politécnica de Aguas calientes, Instituto Tecnológico de Aguascalientes, Universidad de Guadalajara México. Año 2006

c) Título: Reconocimiento de Voz con Redes Neuronales, DTW[1] y Modelos Ocultos de Markov.[2]

d) Propósito: Realizar un el diseño de un "reconocedor de voz" usando redes neuronales artificiales (ANN), alineamiento dinámico del tiempo (DTW) y modelos ocultos de Markov (HMM) así como la realización de un algoritmo de reconocimiento.

e) Sustento teórico: Esta investigación sustenta que debido al auge de los sistemas de reconocimiento de voz dada la creciente necesidad de tener sistemas que puedan controlar de manera no física diversos sistemas de seguridad, sistemas para personas discapacitadas, sistemas para almacenamiento de información, y otras aplicaciones. Las dificultades para la extracción de características están que por lo general las personas no repiten dos veces lo que dicen, se tiene el efecto del estado de ánimo o el estado de salud del hablante, el tiempo etc., para dar solución a estos problemas se buscó desarrollar un algoritmo que usa como algoritmos base algunos que realizan un mayor desempeño entre ellos destacan las ANN para el aprendizaje, el DTW para el referenciado en el tiempo y los HMM para tener de manera completa el algoritmo y dar un modelo de mayoría.

f) Referencia a los resultados: Los resultados muestran un sistema de reconocimiento capaz de reconocer a un individuo y palabras a partir de números del 0 al 9 para poder marcar números y hacer llamadas, el reconocimiento se realiza en un 97% sobre un solo individuo.

---

[1] DTW: Dynamic Time Warping: Alineamiento dinámico del tiempo
[2] http://redalyc.uaemex.mx/pdf/944/94403203.pdf

g) Relación con el trabajo actual: Este trabajo orientado a ayudar a personas discapacitadas, sirvió de guía para aplicar el reconocimiento de patrones y la aplicación de redes neuronales en el reconocimiento de voz.

1.1.2. Antecedente 2

a) Apellidos y nombres: Merlo G, Fernández V.

b) Lugar o Institución donde se hizo la investigación: Laboratorio De sistemas inteligentes, Universidad de Buenos Aires Argentina. Año 1997

c) Título: Reconocimiento de la voz Mediante una red neuronal Kohonen[3]

Propósito: Elaborar un sistema de reconocimiento de voz utilizando una red neuronal Konohen.

d) Sustento teórico: El reconocimiento de voz mediante diversas técnicas tales como cadenas ocultas de Markov y ANN es tema de investigación constante, el sistema de reconocimiento debe realizar especial énfasis en el tratamiento digital de la señal y preparando una codificación para una correcta entrada a la red neuronal.

e) Referencia a los resultados Los resultados indican una taza de reconocimiento del hablante de un 65% en las cifras del 0-9 y del 85% para las vocales.

f) Relación con el trabajo actual, el trabajo permitió dar referencias del cuidado que se debe tener al tratar la señal para obtener un buen conjunto de entrenamiento de la red, el hecho de tener una buena muestra depende mucho de las condiciones de grabación y toma de voces. También se relaciona con el

[3] http://www.itba.edu.ar/archivos/secciones/c11-reconocimientodevozconkohonen-cacic97.pdf

hecho que se aplica otro método en el reconocimiento de voz aunque con resultados diferentes de lo esperado, pero ello se debe a que en ese año los sistemas de adquisición de datos y grabación multimedia no eran del todo accesibles y fáciles de manejar. Este trabajo ahonda en el efecto de trabajar con las señales, digitalizadas y tratadas y luego utilizar el conjunto patrón como elemento de reconocimiento.

1.2.Objetivo Principal

Desarrollar e implementar un sistema de reconocimiento del hablante mediante redes perceptrón multicapa, para mejorar la tasa de reconocimiento.

1.3. Objetivos secundarios

Pre-procesamiento

Implementar algoritmos de pre-procesamiento para señales de voz con el fin de minimizar el ruido y eliminar señales no deseadas que no correspondan a señales de voz.

Extracción de características

Investigar e implementar algoritmos de extracción de características que permitan hallar las características únicas de la voz de cada persona. Estas características son

las que más adelante usará la red neuronal como entradas, a la hora del entrenamiento y luego para reconocer a una persona determinada.

Red Neuronal

Implementar la red perceptrón multicapa en la computadora para la tarea de clasificación. Este programa, contendrá las funciones de propagación hacia adelante, propagación hacia atrás, y las funciones de optimización que harán el entrenamiento del sistema.

Entrenamiento de la red

Entrenar la red perceptrón usando una base de datos con las voces de cinco personas para la tarea de reconocimiento de las mismas, verificando el desempeño del sistema, el tiempo de entrenamiento y el error de validación.

Sistema de reconocimiento.

Implementar el sistema de reconocimiento en tiempo real usando una interface de usuario de tal manera que sea amigable y fácil de usar.

1.4. Justificación

Esta investigación justifica su realización por qué es necesario profundizar en la investigación de los diversos sistemas de reconocimiento de patrones, como en este caso particular lo es los patrones de voz, así mismo buscar la implementación de los mismos a bajo costo.

Este informe está justificado porque aplica un método para realizar la identificación de patrones de voz, así como el reconocimiento del hablante a través de palabras mediante los elementos descriptores propios de la voz humana.

En la actualidad es muy importante, dentro del campo de la ingeniería Mecatrónica, profundizar la investigación de los sistemas inteligentes cuya principal característica es la de tomar decisiones que permitan automatizar las tareas. Los sistemas inteligentes estudiados por la Inteligencia Artificial y aplicados en la actualidad por las ciencias computacionales[4], tienen como uno de sus pilares el estudio y aplicación de las redes neuronales, las cuales permiten tomar decisiones a partir de conjuntos de entrada.

1.5. Alcances de la investigación

Dentro de los alcances del presente informe se quiere desarrollar e implementar un sistema de reconocimiento de voz que sea capaz de reconocer el género del hablante y al hablante mismo, así como la palabra que el hablante emitió.

Este sistema utilizará una computadora con el software MATLAB, un micrófono multimedia para computadora y la grabación de voces de varios individuos.

[4] Las ciencias de la computación o ciencias computacionales son aquellas que abarcan las bases teóricas de la información y la computación, asa como su aplicación en sistemas computacionales.

El reconocimiento de voz estará orientado al reconocimiento de una palabra específica.

La aplicación del método estará orientada a minimizar el error y a mejorar la tasa de error durante el reconocimiento.

La idea de este trabajo es la de poder implementar en un futuroen una tarjeta pequeña capaz de realizar el proceso indicado, a través de la creación de una interfaz portátil del programa; pero eso es objeto de otro trabajo, esta investigación se centrara; es decir este trabajo concluye con implementar un programa elaborado en MATLAB que permita reconocer patrones de voz dentro de un conjunto de entrenamiento de cinco personas de un grupo de cuatro palabras y con una efectividad de reconocimiento de hasta 80%.

1.6. Recursos empleados

Para la investigación presente se utilizaron los siguientes recursos:

a) Micrófono multimedia para PC

El micrófono (ver Figura 1.1.) es un elemento capaz de captar ondas sonoras convirtiendo la potencia acústica en eléctrica de similares características ondulatorias.

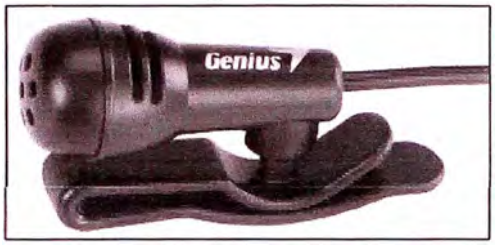

Figura 1.1. Micrófono multimedia para computadora.

b) Computadora o laptop portátil

Sistema Operativo: Windows XP, Vista, 7

Plataforma: x86 – x64 (32 bits ó 64 bits)

Disco Duro: 4 GB (espacio necesario para la instalación)

Memoria RAM: 512MB

Capacidad para soportar MATLAB R2010a

c) Software científico de uso académico

MATLAB en su versión: 7.10.0.499 (R2010a)- 32 bit.

Sistema Operativo: Windows XP, Vista, 7

Plataforma: x32 – x64

Disco Duro: 6 GB (espacio necesario para la instalación)

Memoria RAM: 1 GB

d) Funciones y bibliotecas de MATLAB:

Las herramientas adicionales para el desarrollo de la investigación se encuentran en el uso de las funciones propias de MATLAB para el manejo de datos, en este caso de audio. Ver Figura 1.2.

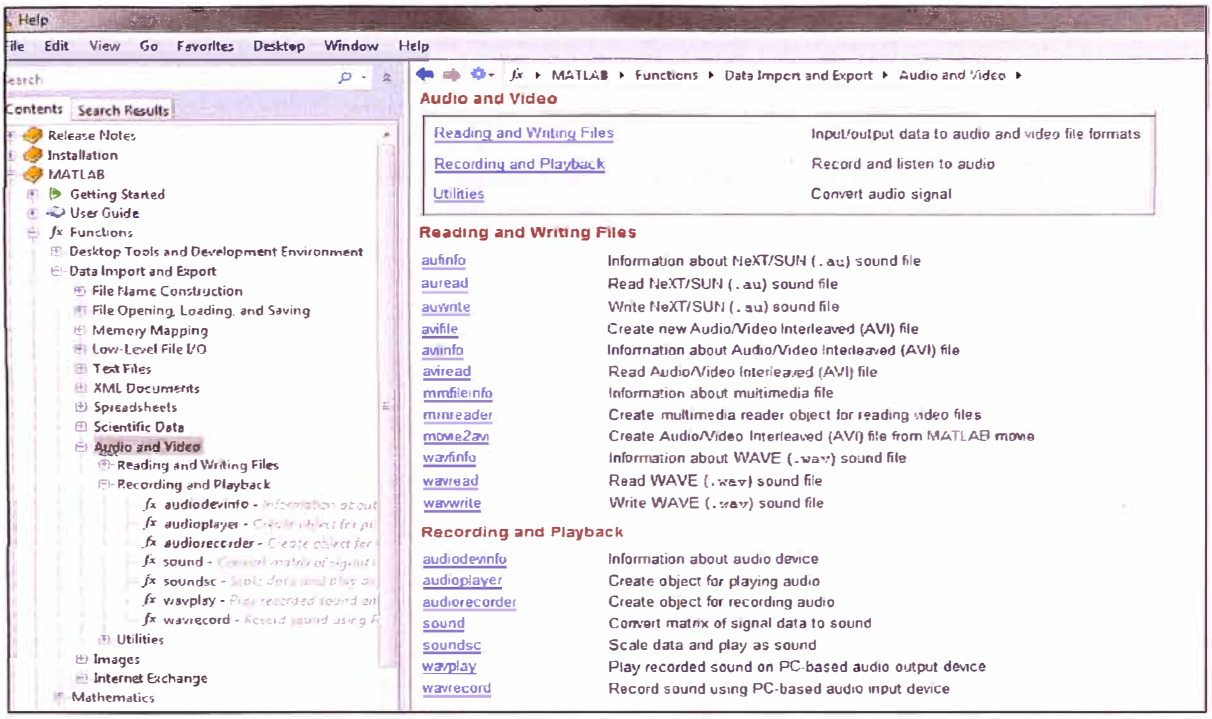

Figura 1.2. Pantalla de ayuda de uso de manejo de datos de audio de MATLAB.

Dentro de las herramientas adicionales se encuentran los programas Final_01 (Apéndice B) para la grabación de las voces. Las condiciones de grabación de que se tomaron para esta ocasión fueron las de minimizar el ruido exterior, mantener una cierta distancia entre el micrófono y el hablante y que el hablante por su puesto se encuentre en buen estado de salud y ánimo.

e) Interfaz y/o tarjeta de desarrollo

La interfaz para la adquisición de los datos de audio se realizó a través de una tarjeta de sonido multimedia, con las siguientes características mínimas:

Resolución de audio de 12-bit

Conversión analógica a digital de 16 bits de entradas analógicas a velocidades de muestreo de 96 kHz.

Conversión digital a analógica de 16 bits de fuentes digitales a 96 kHz a salida de altavoces 2.1 analógica.

CAPÍTULO 2

DESCRIPCIÓN DEL SISTEMA DE RECONOCIMIENTO DE VOZ

2.1. Descripción del sistema de reconocimiento de voz

Para el desarrollo de este sistema, primero se toma una muestra de voz en forma analógica con un micrófono multimedia solicitando a un determinado sujeto que diga una palabra específica, a la vez se graba y digitaliza a través de la computadora usando MATLAB, esta grabación se realiza a través de un programa para poder automatizar la actividad de toma de muestras lo denominaremos final_01.m y se encuentra en el apéndice. La señal grabada se debe acondicionar previamente para obtener lo que se necesita y eliminar lo que se necesita. Luego se procesa la señal acondicionada para extraer las características propias de cada voz. Cada una de estas palabras fue preparada, luego de tomada la muestra, es decir se acondiciono la señal de voz, y luego se extrajeron las características que son propias de cada voz (LPC y Pitch), después se utilizaron estas características como un conjunto patrón para poder entrenar a la red neuronal.

Se grabaron las palabras: hola, exceso, reconocer, y conexión con un grupo de cinco personas de modo tal que cada persona repite cada una de estas palabras treinta veces, es decir seiscientas muestras de palabras grabadas en total. Al repetir constantemente se podrán obtener patrones de voz por cada palabra y por cada sujeto, de ese modo podremos a través de las características previamente extraídas realizar el entrenamiento de la red neuronal. Una vez que la red neuronal esta entrenada, el sistema podrá realizar el reconocimiento de voz de cualquiera de las palabras probadas indicando el género del hablante, su nombre y la palabra que emitió. Ver figura 2.1.

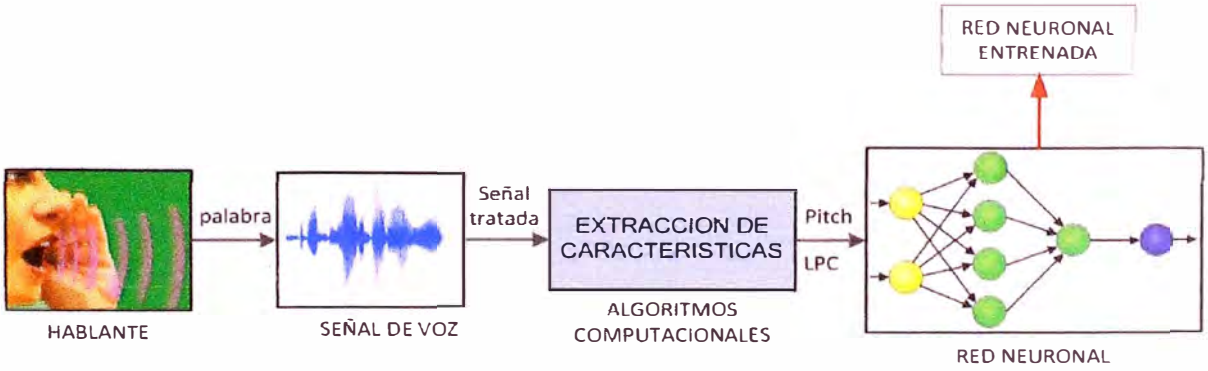

Figura 2.1. Sistema de Reconocimiento de Voz.

Para la grabación de las muestras se establecieron los parámetros y características que según el programa final_01.m se pueden verificar: estos se realizan a través del comando audiorecorder y sus parámetros son: frecuencia de muestreo $F_s = 16000$ (influye en la calidad del sonido, y como no vamos a reproducir se ha tomado este valor como referencia por recomendación de la ayuda de MATLAB), número de bits $N_{bits} = 16$ (ya que el sistema operativo es Windows), y el

canal de grabación será mono canal, puesto que solo nos interesa el sonido de la voz. Como lo que se grabara serán únicamente palabras el tiempo promedio para decir una palabra puede tomarse como 2 segundos, luego a la hora de implementar el programa de registro automático de muestras de voz se utilizó un tiempo de 2.2 segundos para poder dar tiempo al hablante de tomar aire y estar alerta a la siguiente muestra. En el programa se contempló la cantidad de muestras que se desea tomar, el nombre del archivo y la palabra que el hablante emitió, así mismo se determinó que el tamaño de la matriz característica debería ser 12 que es el número de coeficientes LPC (ver fundamento teórico, capitulo 3) por el tamaño de muestras.

Para el almacenamiento de la señal de voz una vez grabada con el comando "audiorecorder" procedemos a almacenar la señal con el comando "wavwrite" el cual almacena el archivo de voz grabado en formato "wav" para que este pueda ser reproducido, comparado y utilizado después. Nosotros por razones de automatización de tareas hicimos que en el programa final_01.m se grabe la muestra, se elimine parte de la señal que es ruido y graficamos la señal de voz tomada durante el muestreo para verificar que la señal que acabamos de grabar tenga la data apropiada para el estudio. Por ejemplo vamos a realizar utilizando el programa final_01.m la grabación de 5 muestras de voz con el hablante: Guillermo y con la palabra "ejemplo uno", durante la ejecución del programa este pedirá automáticamente si se desea tomar una nueva voz o se desea seguir cada muestra ira graficando la señal de voz y al final del muestreo se muestra una pantalla como en la fig.2.2.

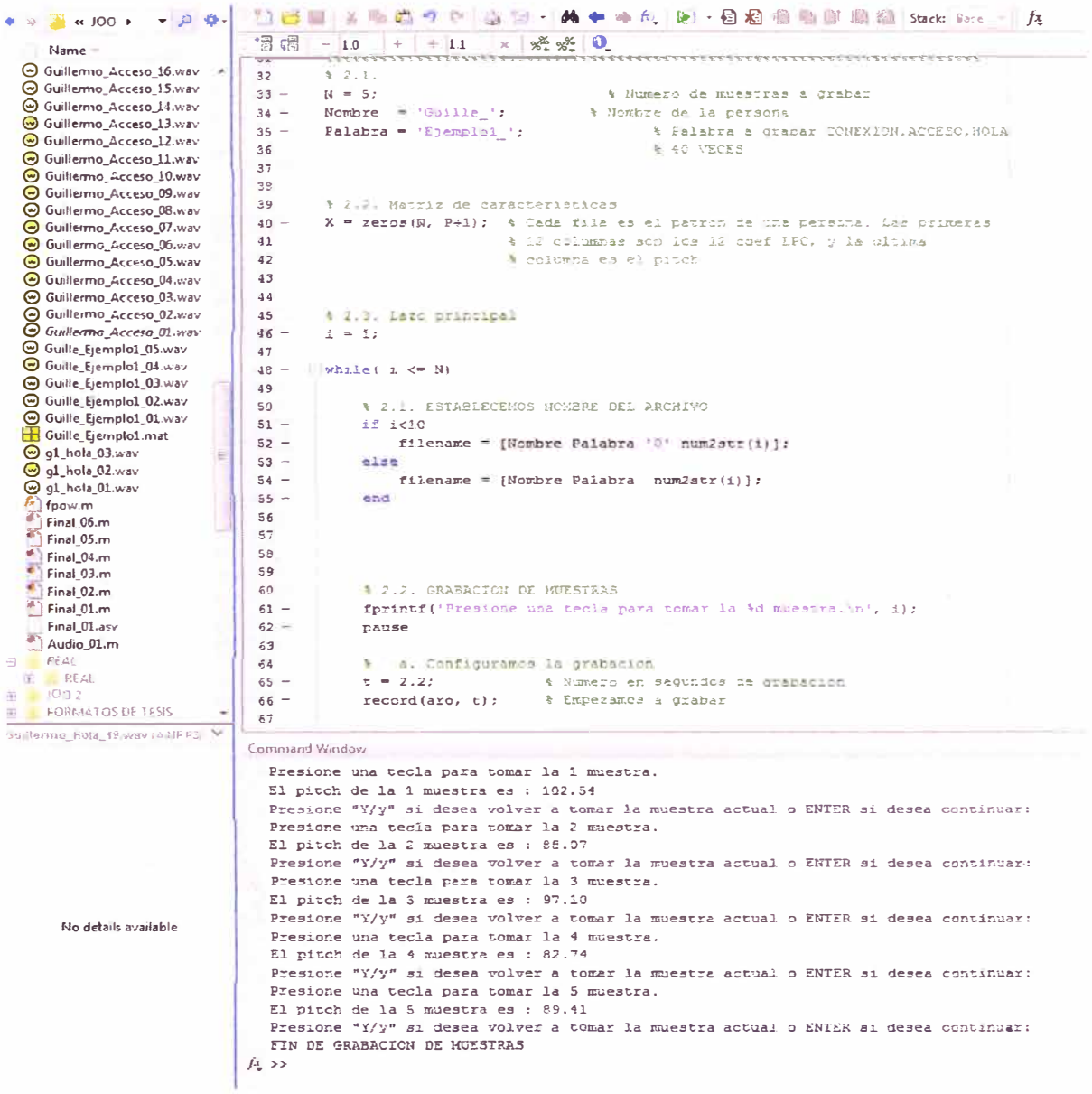

Figura 2.2. Grabación y almacenamiento de muestras de voz

2.2. Proceso de reconocimiento de voz

El proceso de reconocimiento de voz se realiza una vez que la red ha sido entrenada, y en caso el sujeto sea del grupo de personas que está en el conjunto de entrenamiento y este emite una palabra del grupo de palabras seleccionadas, el

sistema responderá indicando el nombre del sujeto y la palabra que emitió. En caso el sujeto una palabra que no pertenece al conjunto de palabras que se emplearon para el experimento y por lo tanto se pertenece al conjunto patrón el sistema indicara que no se reconoce al hablante. Ver figura 2.3.

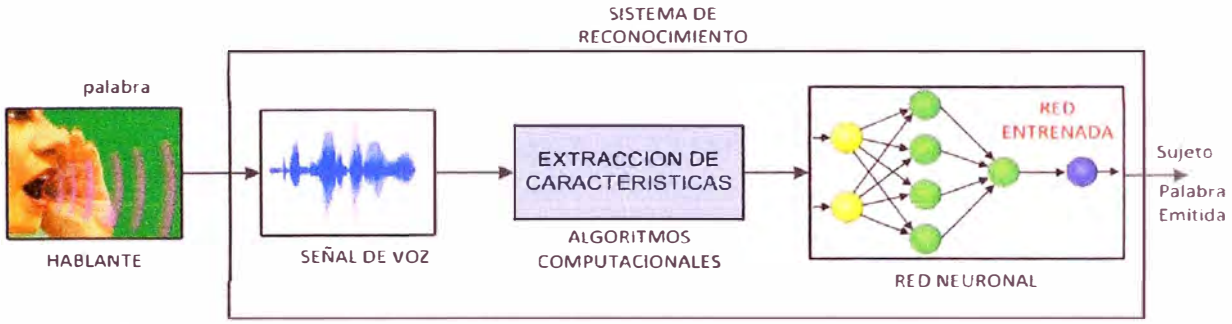

Figura 2.3. Proceso de reconocimiento de voz.

Para el proceso de muestreo, grabación y reconocimiento de voz se utilizó el programa Final_01.m que tiene como diagrama de flujo la figura 2.4

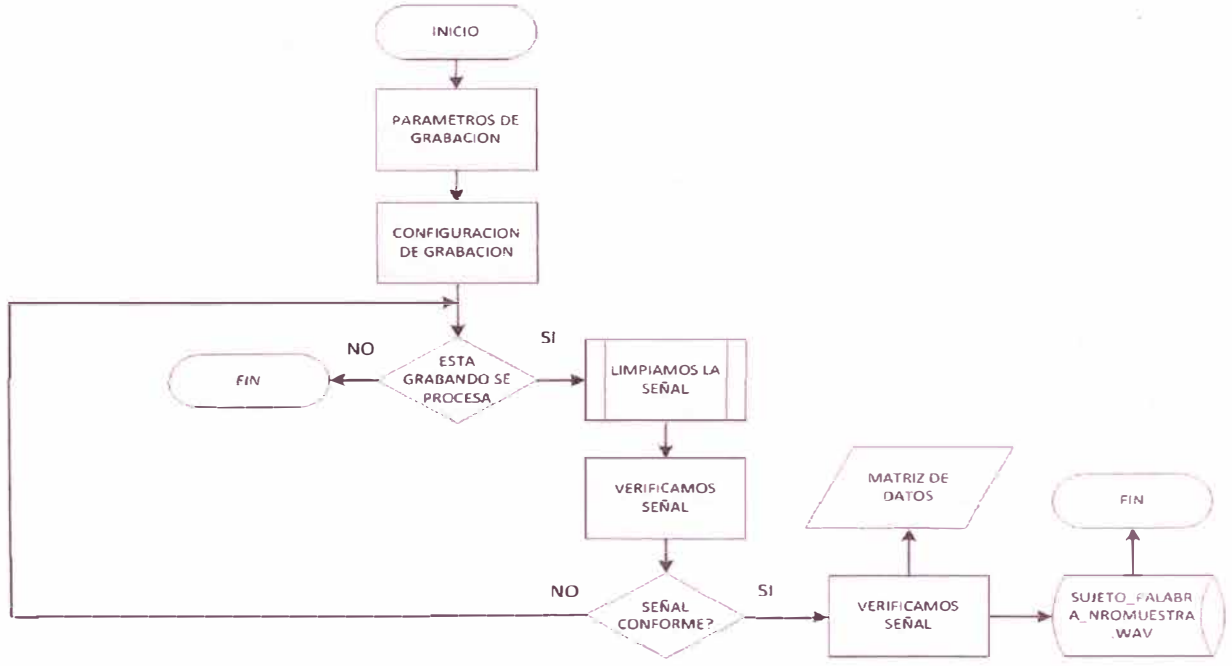

Figura 2.4. Diagrama de flujo grabación y registro de muestras de voz.

CAPITULO 3

IDENTIFICACIÓN DEL PROBLEMA Y DETERMINACIÓN DE LA HIPÓTESIS DE TRABAJO

3.1. Identificación del Problema

En los últimos años hemos presenciado una creciente demanda por sistemas que permitan identificar automáticamente la identidad de una persona y que puedan ser utilizados en aplicaciones como monitoreo, control de acceso, seguridad. reconocimiento de palabras etc., esto debido a que los sistemas más comunes de identificación, tales como el uso de una tarjeta o el uso de alguna clave, son fáciles de burlar y/o falsificar. Por este motivo, la mayoría de los sistemas desarrollados hasta el momento hacen uso de las características únicas de cada persona tales como la huella dactilar, el iris, el ADN, etc., con el fin de lograr tales propósitos, teniendo cada uno de estos métodos ventajas y desventajas, así como su costo al momento de implementar.

Una de las características que identifican a una persona es la voz, los seres humanos usamos la voz para identificar a diferentes personas así como el género de

las mismas podemos utilizar este hecho. Podemos utilizar este hecho para el diseño de sistemas de reconocimiento automático los que pueden reemplazar a los métodos tradicionales de identificación de personas como el uso de tarjetas de identificación, claves de acceso, etc. En el presente trabajo se ha escogido trabajar con sistemas de reconocimiento basado en la voz debido a que estas señales se pueden trabajar de manera sencilla, económica y de manera no intrusiva, usando sensores muy sencillos, y debido a que son particularmente adecuados en las siguientes aplicaciones:

Autentificación de transacciones. (Identificación Biométrica)

Por ejemplo en la prevención del fraude de compras en las compras por teléfono usando tarjetas de crédito. En este caso se el sistema verifica la identidad de la persona que hace las compras a través del teléfono y/o celular. Normalmente es equipado por gobiernos, fuerzas de seguridad, laboratorios forenses, seguridad para empresas corporativas, banca online, telefonía, gestión de call centers. (Ver Figura 3.1.)

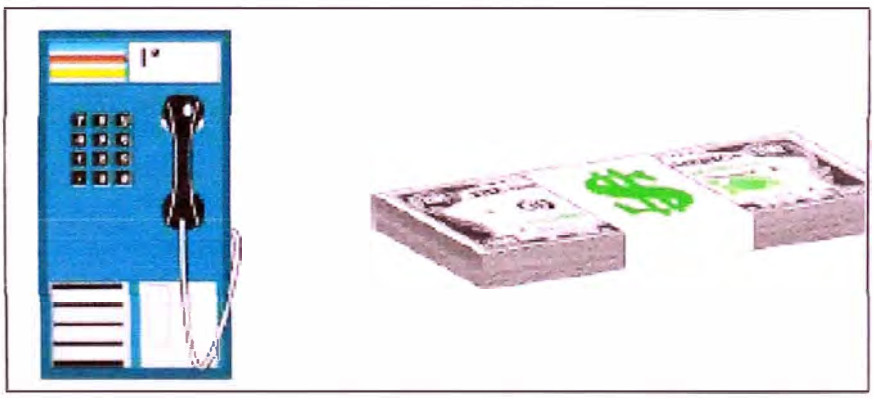

Figura 3.1. Autentificación de transacciones.

- **Control de acceso.**

Donde se requiere que solo personas seleccionadas (ingenieros especializados, gerente, etc.) tengan acceso a determinadas instalaciones, o a ciertos computadores y redes de datos. En estos sistemas no es suficiente el uso de claves de acceso, ya que estos pueden ser fácilmente burlados, por lo que se requiere sistemas de autentificación más sofisticados, robustos y difíciles de ser burlados, acceso por voz en sistemas computarizados.

Figura 3.2. Control de acceso

- **Monitoreo**

Donde se requiere el monitoreo por ejemplo de determinadas personas en ambientes como la prisión o cuando se desea vigilar a personas con arresto domiciliario. En estos casos, se debe detectar la presencia de la persona y/o cuando se desea registrar los comentarios o conversaciones de la misma.

Figura 3.3. Monitoreo de personas.

Ayuda a discapacitados.

El cual permite dar órdenes a otros sistemas para que los discapacitados puedan tener una herramienta que les permita realizar su vida de modo normal y sin limitaciones.

Figura 3.4. Silla de ruedas controlada por comandos de voz.

A pesar de que ya existen sistemas comerciales de reconocimiento basado en la voz, el desarrollo de un sistema alternativo no solo permite apreciar la aplicación del procesamiento de señales e inteligencia artificial en la tarea de reconocimiento del hablante sino que permite el desarrollo de tecnología nacional aplicada a la solución de problemas reales.

Como ya se indicó en los antecedentes en el capítulo uno, la tasa de los sistemas de reconocimiento que usan otras topologías neuronales esta entre 60 y 85% y se realiza utilizando reconocimiento de vocales.

Ante estas consideraciones nos planteamos la siguiente pregunta::

"¿Es posible diseñar e implementar un sistema de reconocimiento del hablante mediante redes perceptrón multicapa, que pueda reconocer al sujeto y la palabra que emitió a través de los elementos descriptores de la voz, mejorando la tasa de reconocimiento desde 60% hasta 80% utilizando palabras?"

3.2. Hipótesis de trabajo.

El presente trabajo de investigación esta ordenado y dividido de acuerdo al siguiente diagrama de medios fines (Ver Figura 3.5.).

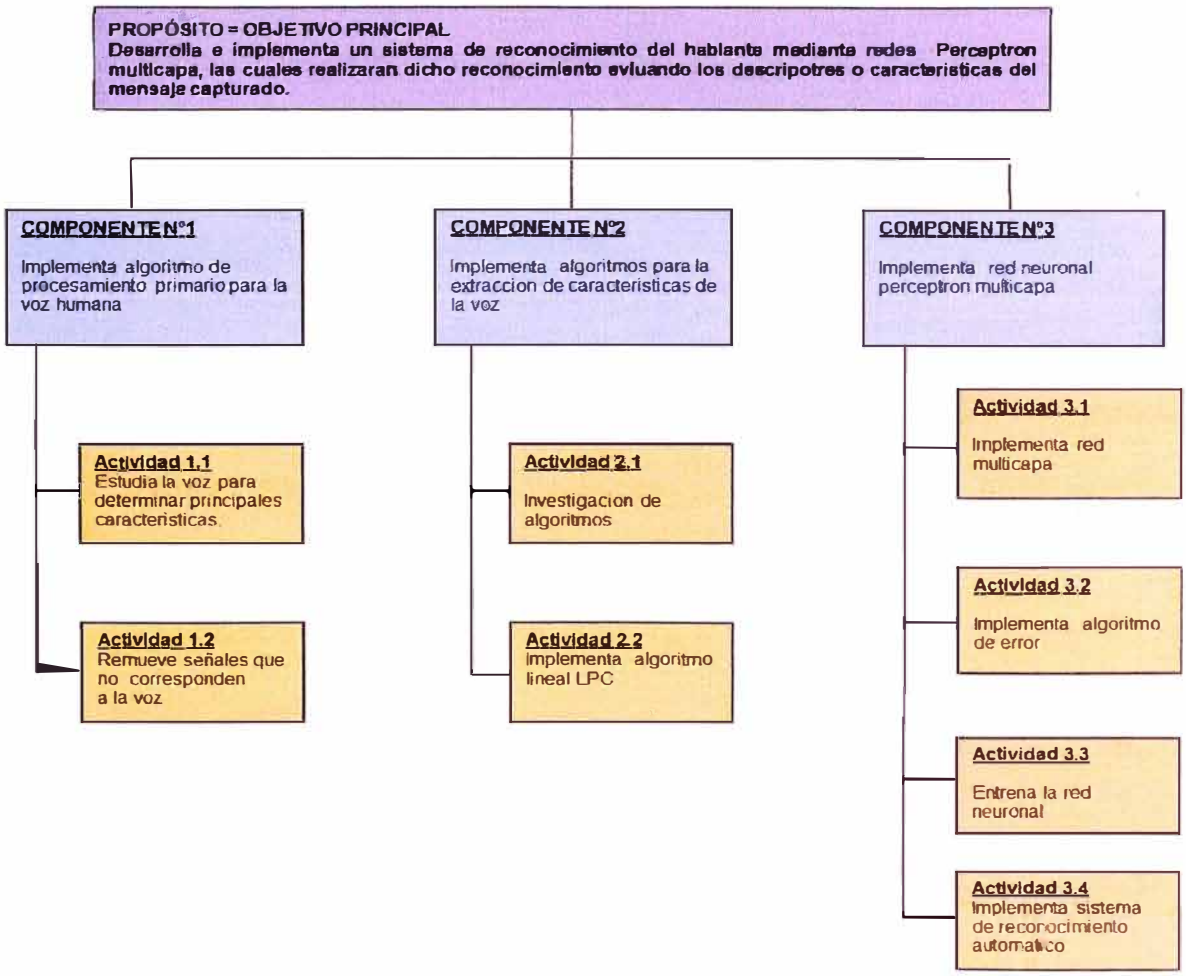

Figura 3.5. Diagrama medios fines

Para lo cual se tuvieron las siguientes consideraciones: Componente1, se consideró que el sistema va a reconocer al hablante a través de sus elementos descriptores y que el sistema va a reconocer la palabra que emite el hablante en cada caso, para lo cual es necesario primero conocer y determinar los elementos descriptores. Se removerán las señales que no correspondan a la voz, esto con el fin de quedarnos solo con los elementos que nos interesan para el estudio. Es posible que al obtener elementos descriptores de la voz, estos nos puedan servir para caracterizarla.

En el componente 2, se implementa un algoritmo de reconocimiento basado en los elementos descriptores y se obtendrán los patrones de reconocimiento. Es posible obtener patrones para luego aplicar una clasificación a través de alguna herramienta de decisión.

En el componente 3, se considera realizar dicha clasificación implementando la red, usando redes perceptrón multicapa y se desarrollara íntegramente en MATLAB. Es posible implementar el algoritmo de reconocimiento que permita clasificar y reconocer al hablante a través de los elementos descriptores mejorando la tasa de reconocimiento hasta 80%.

CAPITULO 4

FUNDAMENTO TEÓRICO

4.1. Mapa teórico.

Es conveniente en este capítulo presentar el siguiente mapa teórico (ver Figura 4.1.) el cual relaciona y presenta los fundamentos necesarios que se tomaron en cuenta para el desarrollo del presente informe.

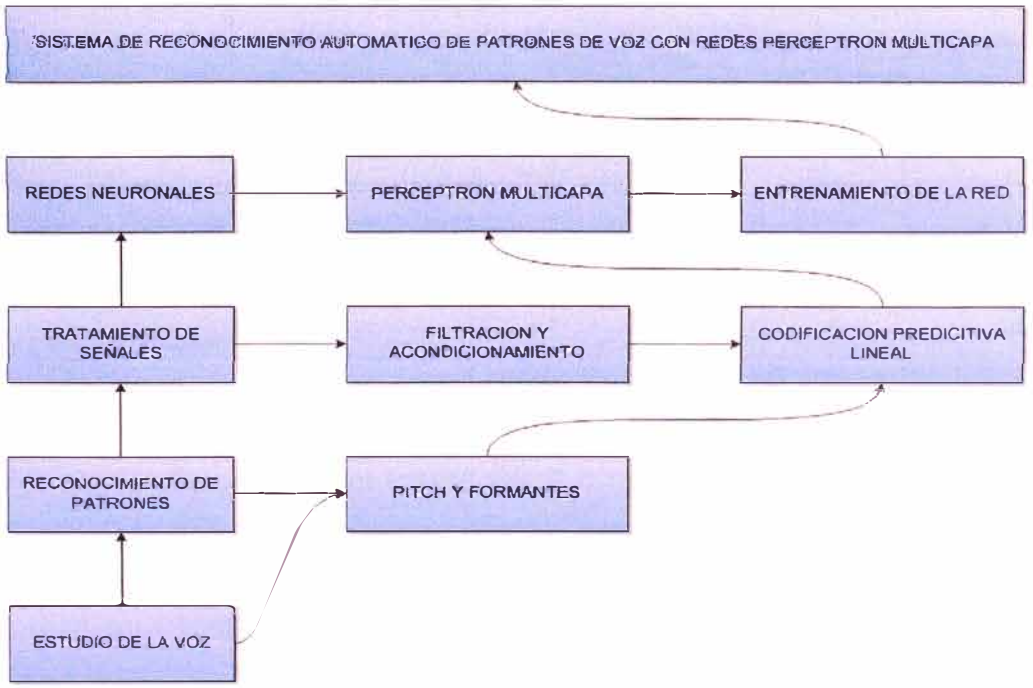

Figura 4.1. Mapa teórico.

4.2. La voz humana y sus características.

La función principal asociada a la transmisión de voz es la transmisión de un mensaje; sin embargo, junto con este mensaje, el receptor recibe una cantidad adicional de información acerca del género, identidad, estado emocional, salud, la fuerza de pronunciación, la entonación, etc., de la persona que emite el mensaje. La fuente de esta información reside en las características fisiológicas y de comportamiento de cada persona [1].

En esta parte se estudiaran las principales características de la voz humana. Se hará un breve estudio fisiológico de la voz, analizando el proceso de formación de la voz y sus principales características. Luego se hará estudio del Pitch y de las frecuencias formantes, que son las características únicas en cada persona y por tal basándonos en estas medidas podemos distinguir a una persona de otra. Como se verá más adelante, la mayoría de algoritmos de extracción de características, que permiten identificar a una persona, se basan en aproximaciones de estas características. La voz y su funcionamiento son tema de constante estudio de la fonología y el lenguaje, vamos a presentar los aspectos relevantes para la tesis.

4.2.1. Estudio fisiológico del tracto vocal

Las características fisiológicas del tracto vocal se muestran en la Figura 4.2. la forma del tracto vocal, determinado por la posición de la lengua, mandíbula, labios y dientes, crean un conjunto de resonancias acústicas en respuesta a los flujos

periódicos de aire generados por el pulmón cuando se quiere producir una señal de voz. Debido a que las características fisiológicas son únicas en cada persona, el conjunto de resonancias es único para cada persona.

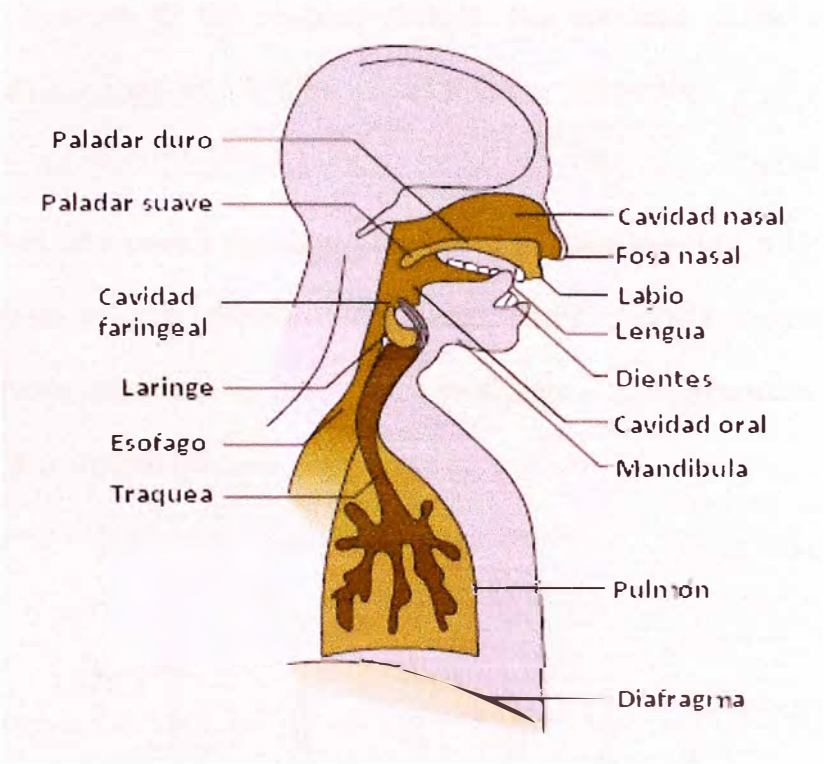

Figura 4.2. Fisiología del tracto vocal.

Cuando se emiten sonidos por la boca, estos pasan a través de dos sistemas antes de que el sonido tome su forma final. El primer sistema, conocido como tracto laringeal, es el generador del Pitch o la frecuencia fundamental de la vibración del aire; y el siguiente, conocido como tracto supra-laringeal, es el sistema que modula los armónicos del Pitch creados por el primer sistema dando la forma final a la señal de voz.

a) El Pitch

El Pitch es la característica más distintiva entre los hombres y mujeres. El Pitch de una persona tiene su origen en las cuerdas vocales y está definido como la frecuencia de vibración de las cuerdas vocales. Por ejemplo, si las cuerdas vocales vibran 300 veces por segundo, se dice que el Pitch es de 300Hz.

Cuando el aire pasa a través de las cuerdas vocales vibra a la frecuencia del Pitch, pero además se crean diversos armónicos. Estos armónicos ocurren en enteros múltiplos del Pitch, tal como se muestra en la figura 4.3., y decrecen en amplitud a razón de 12dB por octavo [1].

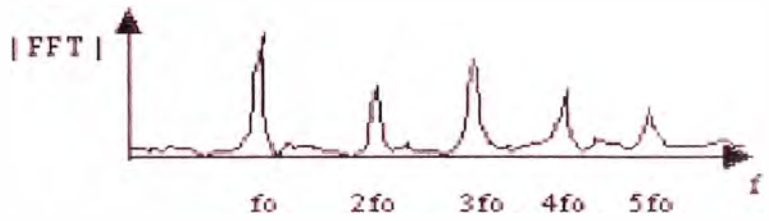

Figura 4.3. El Pitch y sus armónicos.

La razón por la que el Pitch difiere entre hombres y mujeres es el volumen, tamaño y la tensión de la cavidad de la faringe. En la niñez el Pitch tiene un valor aproximado de 250Hz, y la longitud de las cuerdas vocales es aproximadamente 10.4mm. Luego de la pubertad y con el desarrollo corporal el cuerpo humano varia sus dimensiones por lo que el Pitch cambia de valor. Las cuerdas vocales en los

varones se incrementa a una longitud entre 15-25mm y el de las mujeres entre 13-15mm. Estos incrementos en tamaño tienen correlación con la disminución del Pitch. En los varones, el Pitch promedio cae entre 60 a 120Hz, y entre las mujeres el Pitch cae entre 120 a 200Hz.

b) Las Frecuencias Formantes

Cuando el aire sale del sistema generador del Pitch, este ingresa al segundo sistema, llamado tracto supralaringeal, donde el aire empieza a reverberar a determinadas frecuencias determinados por el diámetro y la longitud de las cavidades supra-laringeales. Estas reverberaciones son llamadas "resonancias" o "frecuencias formantes [2]."

Las frecuencias formantes son frecuencias resonantes del tracto vocal que aparecen en el espectro de voz como picos claramente apreciables tal como se muestra en la figura 4.4, donde se aprecia claramente la presencia de tres picos en el diagrama de frecuencia de una señal corta de voz.

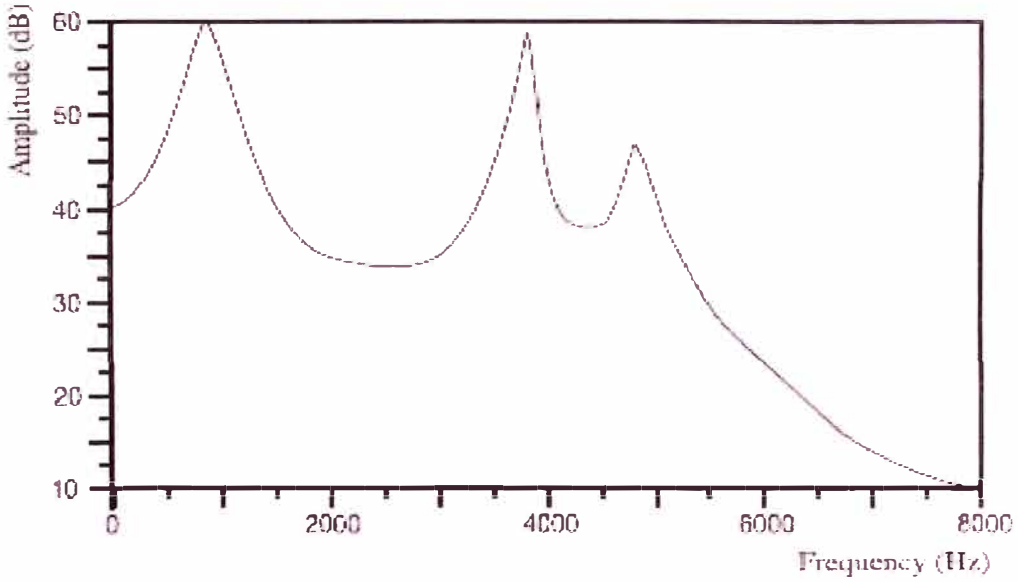

Figura 4.4. Espectro de 20ms de una señal de voz.

En la producción de voz, las frecuencias formantes cambian debido a la posición de la lengua, mandíbula, y otras características del tracto vocal. Cada tipo de frecuencia formante está relacionado con la vocal que se desea pronunciar, de donde se reconocen dos cosas:

- Cada frecuencia formante tiene un correspondiente ancho de banda.

- Cada frecuencia formante está en un intervalo bien definido dentro del espectro de la voz.

Las frecuencias formantes para cada vocal son bastante similares entre los seres humanos dado que estos deben ser reconocibles como un sonido en particular.

4.2.2. Producción de la voz como un sistema lineal.

El tracto vocal de una persona crea diferentes frecuencias formantes para cada vocal. Se puede ver a este sistema como un filtro variable donde la entrada está dada por el Pitch y sus armónicos generados por las cuerdas vocales, mientras que las salidas del filtro (que es el sonido que se percibe desde la boca) es la ganancia de los armónicos que caen en las frecuencias formantes de la vocal que se desea pronunciar.

La teoría acústica de la producción de voz asume que el proceso de producción de voz es un sistema lineal, que consiste en un filtro y una fuente [1]. Este modelo, que se muestra en la figura 4.5, captura el proceso de producción de voz descrito anteriormente.

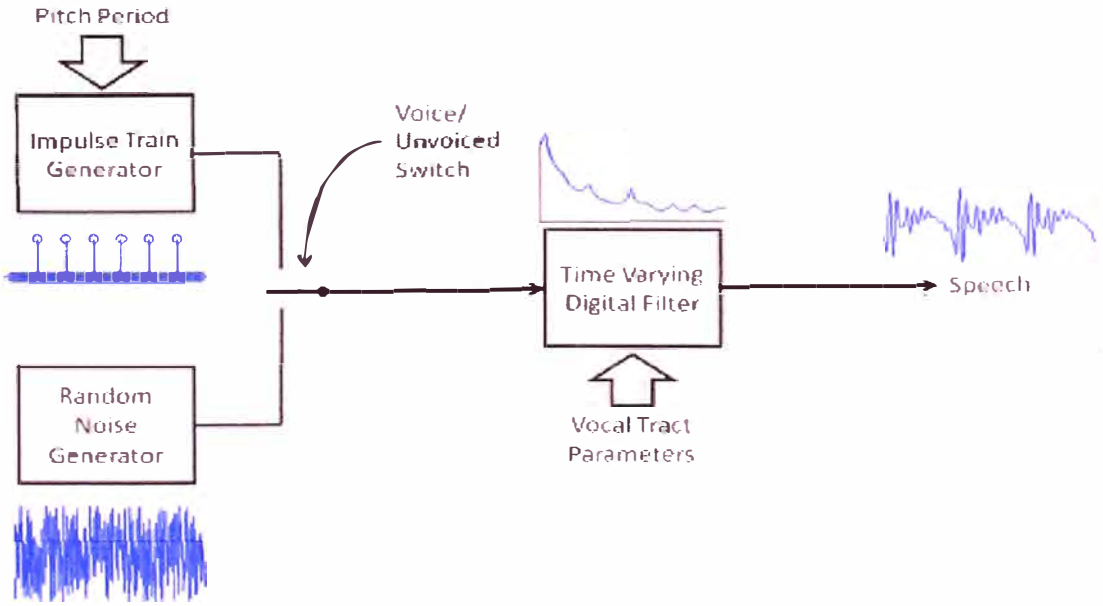

Figura 4.5. Modelo de un sistema de generación de voz.

La fuente. Que consiste de un generador de pulsos a cierta frecuencia fundamental para señales de voz, o de ruido blanco para señales de no voz. La fuente imita al sistema de generación del Pitch.

Filtro Tracteal. Que filtra la señal de la fuente, modificando las resonancias de la señal de pulsos. Este parte imita el proceso de generación de las frecuencias formantes.

4.2.3. Comparación de características

Para el reconocimiento de personas, se debe identificar las características más útiles de tal manera que el sistema de reconocimiento tenga las siguientes características [1]:

No puedan ser imitadas y/o controladas por el hablante.

Inalterables por los estados de ánimo y/o de salud de cada hablante.

Independiente del ruido causado por el proceso de grabación.

El Pitch es bueno a la hora de distinguir el género porque el Pitch de las mujeres es generalmente más grande que el Pitch de los hombres. Sin embargo, puede ser fácilmente alterado por el hablante, lo que haría que el sistema de reconocimiento produzca resultados erróneos. Otro problema con el Pitch es que puede ser afectado por el estado de ánimo y de salud.

Como el Pitch, las frecuencias formantes también pueden usarse para distinguir el género de un hablante ya que estas ocurren a más altas frecuencias en una mujer que en un hombre. Por ejemplo, la primera y segunda frecuencia formante de la letra "u" en una mujer están en 370Hz y 950Hz y las de un hombre están en 300Hz y 870Hz. Por otro lado, la estructura de cada frecuencia formante varía entre cada persona, es decir es única para cada persona, por lo que esta característica puede ser usada como base para un sistema de reconocimiento automático de personas.

Como se verá más adelante, el uso conjunto del Pitch y las frecuencias formantes, se pueden usar como buenos indicadores para identificar la identidad de una persona siempre y cuando la persona hable de manera normal sin tratar de modificar su voz.

4.3. Herramientas de procesamiento de señales de voz

La primera etapa de los sistemas de reconocimiento se denomina comúnmente "extracción de características" que consiste en aplicar a la señal "cruda" o señal inicial de entrada (por ejemplo, la matriz de pixeles de una imagen o el vector que representa la voz) uno o varios algoritmos que se encargan de extraer una serie de características que representan únicamente a la señal "cruda" de entrada. Además, también es común la etapa de pre procesamiento, que se encarga de aplicar a la señal de entrada un conjunto de transformaciones con el fin de atenuar y/o eliminar las señales no deseadas, es decir retirar las partes con las que no vamos a

obtener información útil, de esta manera simplificar el trabajo que se va a realizar más adelante.

En este capítulo se va a presentar un conjunto de algoritmos del área de procesamiento de señales para la etapa de pre procesamiento y para la etapa de extracción de características. De esta manera, va a ser posible representar de manera compacta y única la señal de voz de cada persona, y esta nueva representación será usada más adelante como las entradas del sistema de reconocimiento tanto para la etapa de entrenamiento como para la etapa de prueba y funcionamiento, vamos a considerar lo siguiente:

4.3.1. Pre-procesamiento de la voz

Como en la mayoría de aplicaciones, cualquier señal de voz que se tenga siempre esta corrompida por una cantidad finita de ruido. Esto se debe a imperfecciones a la hora de grabar la señal de voz, a la pérdida de información debida a la digitalización de la data, etc. En general es ruido indeseado que siempre va al medio acústico, a las condiciones de la fonación, mediante el uso de un sistema de grabación por ejemplo con micrófono y una pc.

Por otro lado, a la hora de tomar muestras de voz existen periodos de tiempo en que no se tiene ninguna señal. Esto aparece por ejemplo al inicio de una grabación de voz, ya que siempre el hablante no empieza a hablar al mismo tiempo que se inicia una grabación, y al final de una grabación.

4.3.2. Remoción de las señales que no corresponde a la voz

La primera tarea es remover las señales que no corresponda a la voz de una grabación. Para esto, es conveniente usar un algoritmo de detección de energía, en donde la energía de una señal se voz está dada por:

$$E = \frac{1}{N}\sum_{n=0}^{N-1}|x(n)|^2 \qquad (4.1)$$

Donde N, corresponde a la longitud de la señal x(n). Entonces, hallando la potencia de la señal de voz mostrado en la figura 4.6., donde la potencia se evalúa cada 20ms, se tiene el grafico de potencia que se muestra en la figura 4.7.

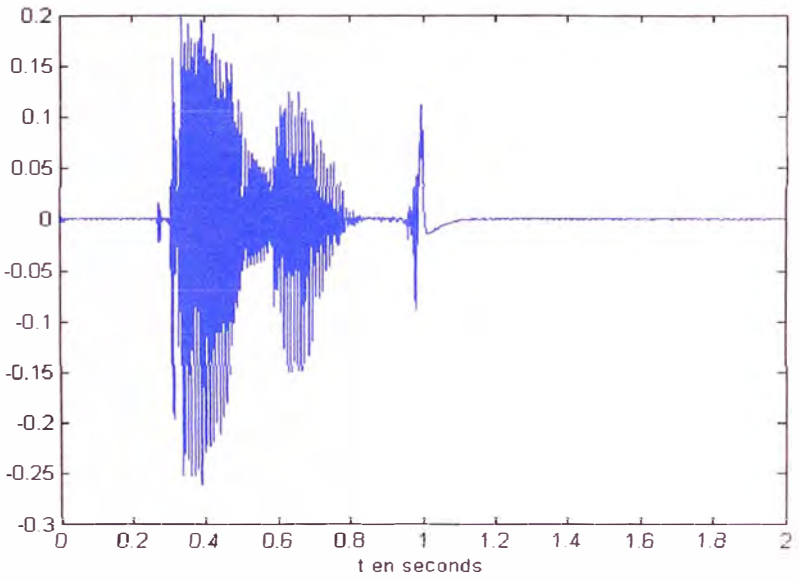

Figura 4.6. Señal de voz.

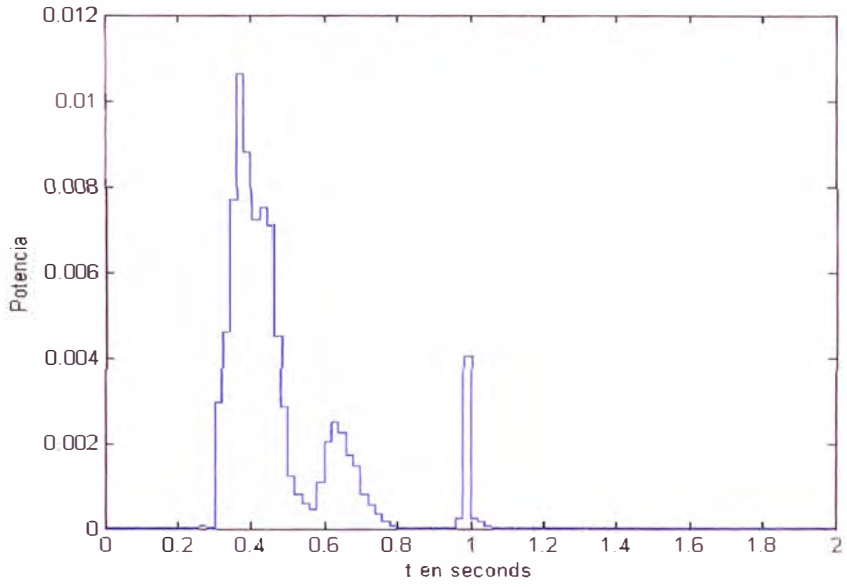

Figura 4.7. Potencia de la señal de voz.

Como se puede apreciar, las partes que corresponde a señales de voz se caracterizan por tener una gran potencia mientras que el resto tiene un valor pequeño de potencia. Esta observación, nos permite notar que con el fin de eliminar las señales que no corresponde a las de voz, se debe eliminar aquellos segmentos que tengan poca potencia.

Entonces luego se aplicar el proceso de eliminación de señales que no corresponde a la voz a la señal mostrada en la Figura 4.7., se tiene la señal que se muestra en la figura 4.8., donde se ve claramente como esta nueva señal no contiene señales que tienen potencia muy pequeña. Es decir se ha logrado eliminar aquellos intervalos donde solo se registra el ruido del ambiente, dejando solo el sonido de la señal de voz.

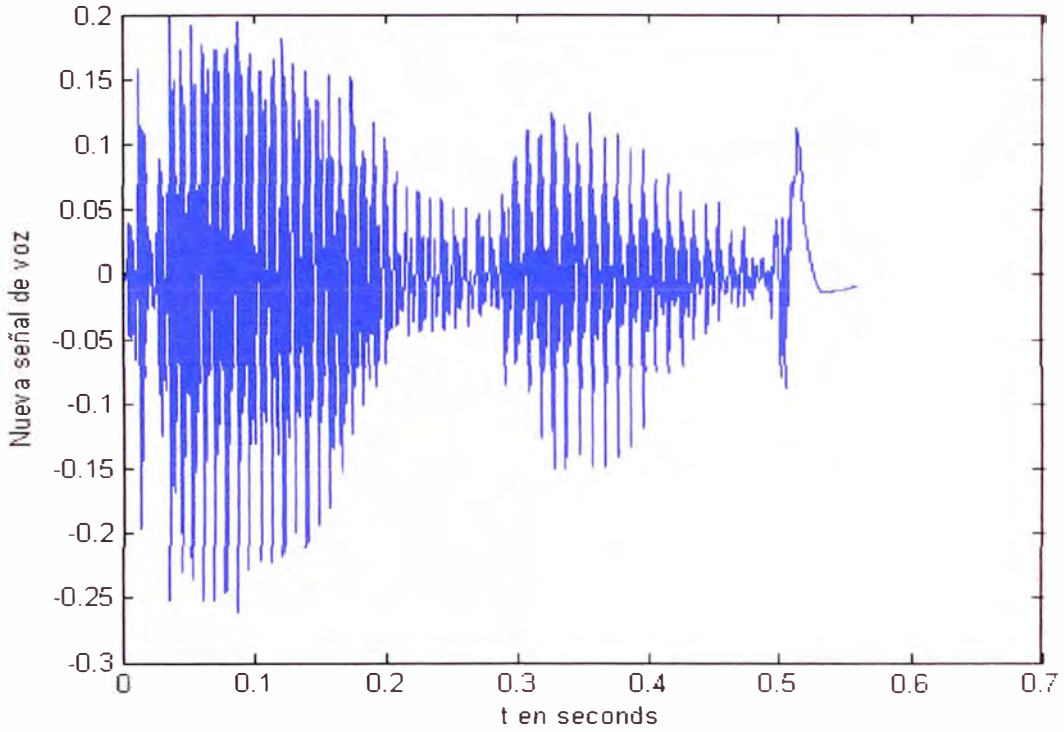

Figura 4.8. Señal de voz cortada.

4.3.3. Filtrado de señales de voz

Como se mencionó anteriormente, la gran mayoría de señales siempre están afectadas por ruido, los cuales tienen componentes de altas frecuencias. Así, con el fin de mejorar la proporción señal/ruido, es conveniente filtrar las señales [2]. Entonces para este fin se puede aplicar el filtrado usando una ventana. En la figura 4.9 se muestran las ventanas más comunes usadas en el procesamiento de señales.

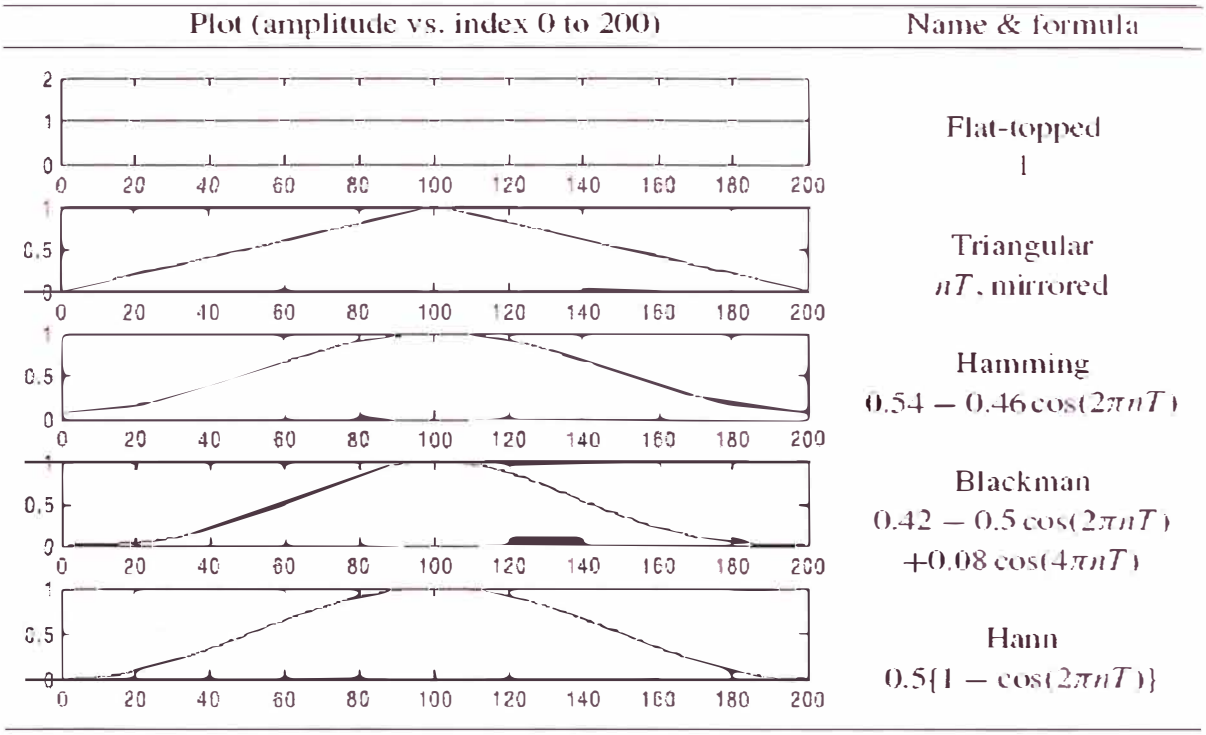

Figura 4.9. Ventanas más comunes.

La ventana con más uso es la ventana de Hamming [2], por lo que se usará esta ventana en el filtrado de nuestras señales de voz. Usualmente el filtrado no se hace sobre toda la señal sino sobre intervalos de aproximadamente 30ms. En las siguientes figuras se muestra el resultado del filtrado de un segmento de una señal de voz.

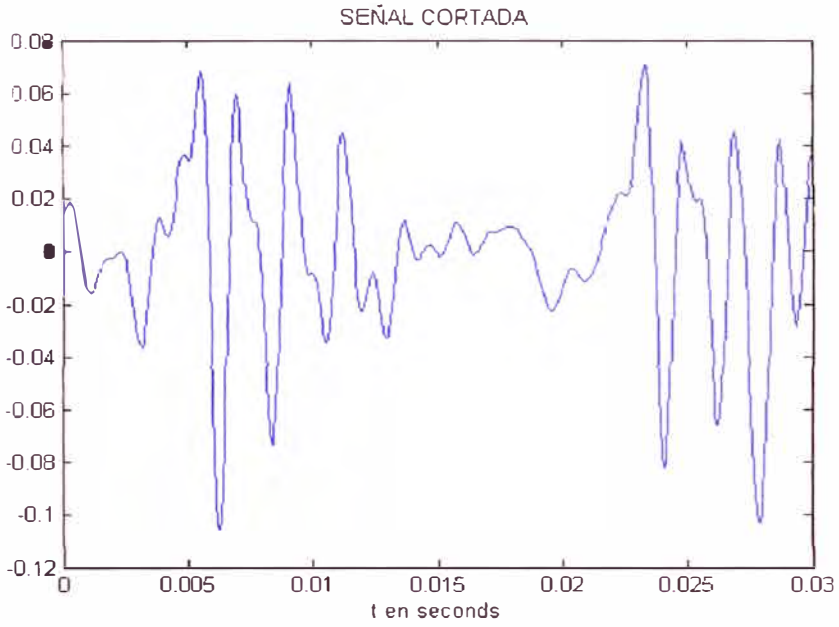

Figura 4.10. Segmento de una señal de voz.

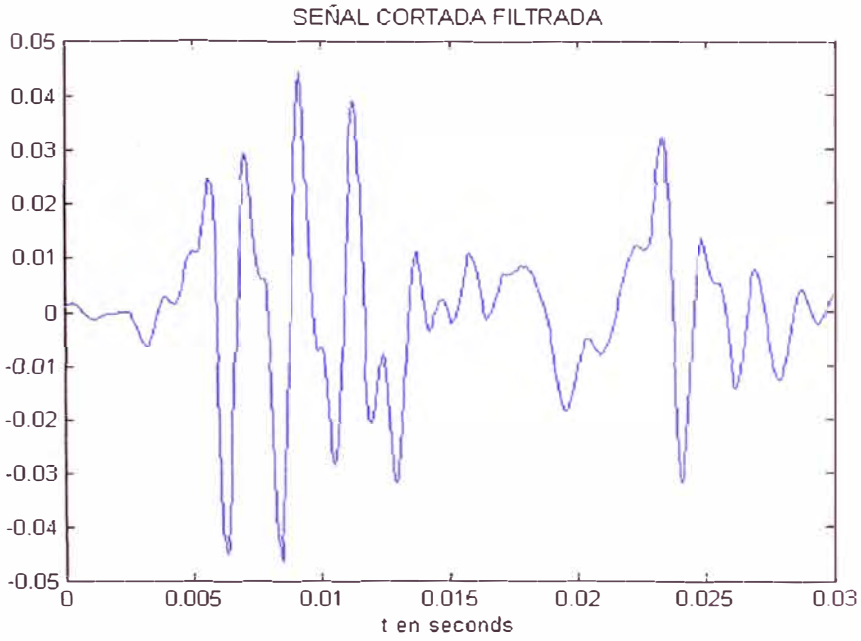

Figura 4.11. Segmento filtrado de una señal de voz.

4.4. Extracción de características

Luego de la etapa de pre procesamiento sigue la etapa de extracción de características que consiste en aplicar uno o más algoritmos que se encargar de representar únicamente la señal de entrada. Tal como se mencionó anteriormente, una persona puede ser caracterizada por su Pitch y sus frecuencias formantes. En esta sección se describirá el algoritmo de Codificación Predictiva Lineal (Linear Predictive Coding) como un método para extraer estimaciones de las frecuencias formantes, y se describirá un algoritmo para la detección del Pitch.

4.4.1. Codificación predictiva lineal LPC

La predicción lineal ha sido por varias décadas el ingrediente principal en la tecnología de la comunicación y la codificación de voz [1]. En el área del procesamiento digital de la voz el algoritmo LPC ofrece un método poderoso y bastante simple que provee un estimado para aproximar el espectro a través de una serie de coeficientes.

Este algoritmo se basa en la observación que la voz producida por el sistema muscular permanece estacionaria por aproximadamente 30ms. Esta característica implica que las 240 muestras de una señal de voz grabada a 8KHz (que corresponde a una duración de 30ms), son similares, y pueden ser parametrizadas por un conjunto más pequeño de valores. Estos valores son conocidos como los coeficientes de

predicción lineal. Estos coeficientes son polinomios generadores de un filtro digital que, cuando son estimulados por una señal de entrada, recrea las características de la señal original. Aunque esta señal recreada no es idéntica a la señal original, su respuesta en frecuencia es idéntica a la original.

Un predictor lineal de orden P está representado por P coeficientes $a[0], a[1], ..., a[P-1]$, y está dado por:

$$y[n] = x[n] + \sum_{p=0}^{P-1} a(p)y(n-p) \tag{4.2}$$

Donde, x[n] es el vector de audio de entrada, y $y(n)$ es el vector de salida de audio, que tiene las características vocales codificadas en los coeficientes en "a".

Para hallar los valores de los coeficientes, primero se asume que, dado una señal de voz pseudo estacionaria, la siguiente muestra en el instante "n" puede ser representada como una combinación lineal de las "P" muestras pasadas. Esta combinación se muestra en la siguiente ecuación:

$$x[n] = a_1 x[n-1] + a_2 x[n-2] + ... + a_P x[n-P] \tag{4.3}$$

El error entre la muestra predicha y la muestra real muestra la habilidad del sistema de hacer predicciones, y como tal buscamos minimizar el siguiente error:

$$e[n] = x[n] - x'[n] \tag{4.4}$$

En realidad lo que se hace es minimizar el error cuadrático medio (RMS) sobre todas las "n" muestras:

$$E = \sum_n e^2[n] = \sum_n \left\{ x[n] - \sum_{k=1}^{P} a_k x[n-k] \right\} \tag{4.5}$$

Con el fin de determinar los coeficientes LPC que minimizan la ecuación de error (4.5) se debe hallar la derivada de tal ecuación e igualarla a cero:

$$\frac{\partial E}{\partial a_j} = -2\sum_n x[n-j]\left\{ x[n] - \sum_{k=1}^{P} a_k x[n-k] \right\} = 0 \tag{4.6}$$

De donde se tiene la siguiente ecuación, que nos permite hallar los coeficientes "a":

$$\sum_{k=1}^{P} a_k \sum_n x[n-j]x[n-k] = \sum_n x[n]x[n-j], \qquad j = 1,\ldots,P \tag{4.7}$$

Para resolver el conjunto de ecuaciones (4.7) existe una gran cantidad de métodos. Los métodos más comunes son los de la covarianza y el de auto correlación. El primero divide la voz en una serie de segmentos y minimiza el error sobre cada segmento de N muestras. El método de auto correlación asume que la señal es estacionaria con energía finita, con un rango de sumatoria infinita (lo cual se

cumple si aplicamos una ventana a la señal). Nosotros optaremos por el método de auto correlación por ser más preciso.

Entonces, si asumimos la sumatoria infinita se tiene:

$$\sum_{n=-\infty}^{\infty} x[n-j]x[n-k] \equiv \sum_{n=-\infty}^{\infty} x[n-j+1]x[n-k+1] \tag{4.8}$$

De donde se tiene:

$$\sum_{n=-\infty}^{\infty} x[n-j+1]x[n-k+1] \equiv \sum_{n=-\infty}^{\infty} x[n]x[n+j-k] \tag{4.9}$$

Usando estos resultados, la ecuación (4.7) se puede reformular como:

$$\sum_{k=1}^{P} a_k \sum_{n=-\infty}^{\infty} x[n]x[n+j-k] = \sum_{n=-\infty}^{\infty} x[n]x[n-j] \tag{4.10}$$

Recordando que la función de auto correlación está dada por:

$$R[k] = \sum_{n=-\infty}^{\infty} x[n]x[n+k] \tag{4.11}$$

La ecuación (4.10) puede ser representada por un conjunto de P ecuaciones lineales de la siguiente forma:

$$\begin{bmatrix} R(0) & R(1) & \cdots & R(P-1) \\ R(1) & R(0) & \cdots & R(P-2) \\ \vdots & \vdots & \ddots & \vdots \\ R(P-1) & R(P-2) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(P) \end{bmatrix} \tag{4.12}$$

La técnica estándar más utilizada para resolver el sistema anterior está dada por el algoritmo Durbin - Levinson [3] y será la usada en nuestro sistema.

Con el fin de apreciar este algoritmo, en la siguiente figura se muestran los coeficientes LPC hallados al analizar 9 muestras de voz de una mujer.

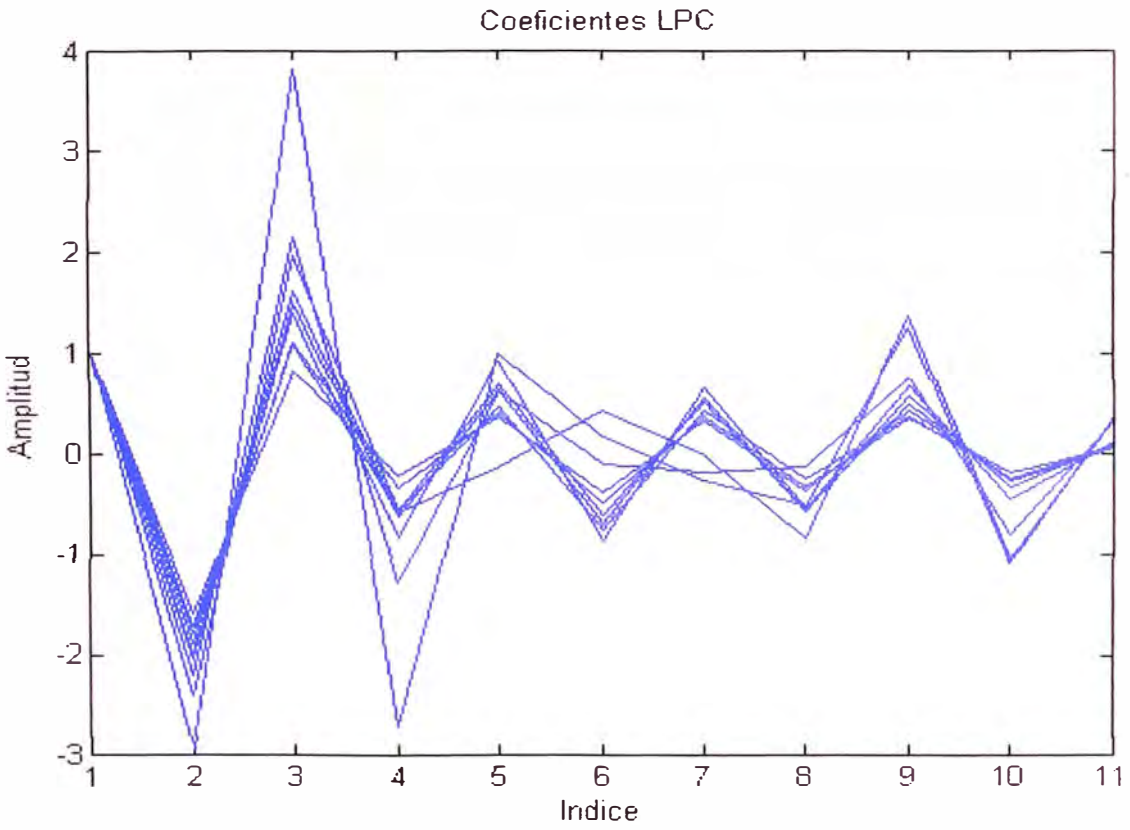

Figura 4.12. Coeficientes LPC de una mujer.

Ahora, en la siguiente figura se muestran los coeficientes LPC hallados al analizar 9 muestras de voz de un hombre.

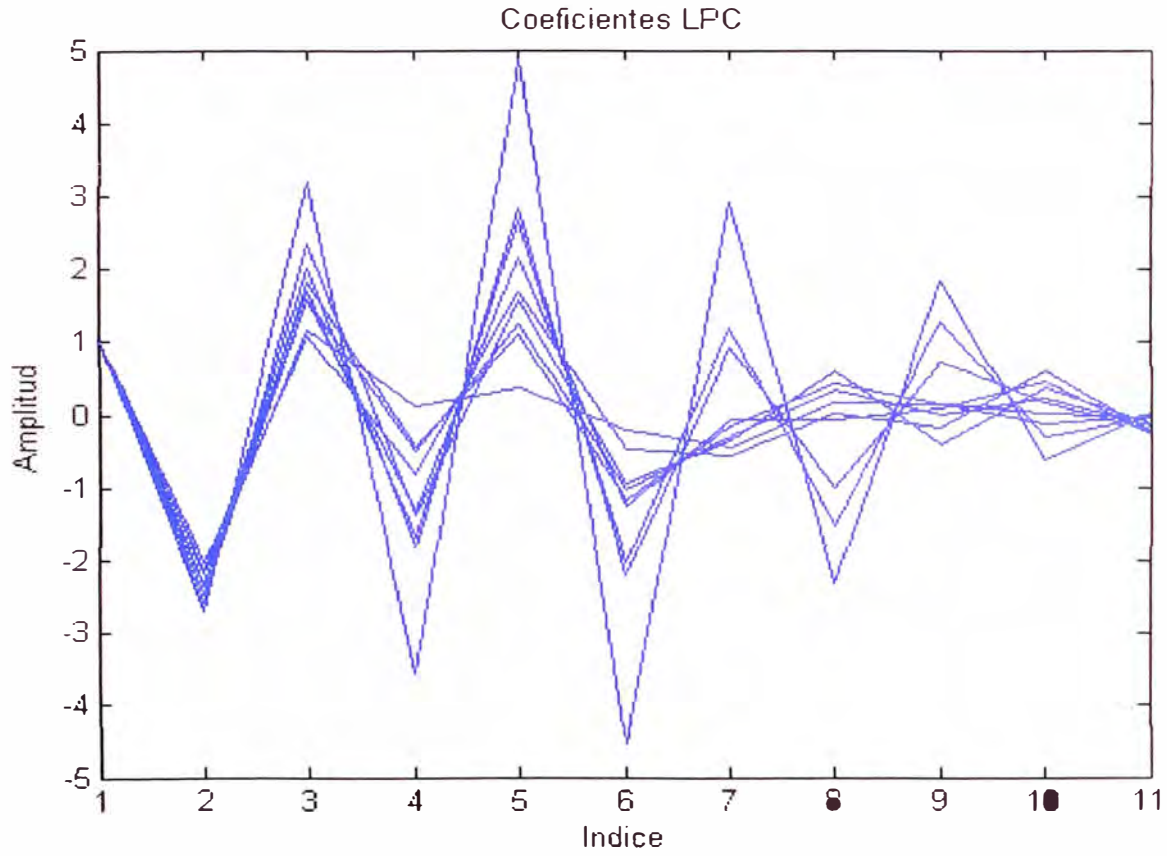

Figura 4.13. Coeficientes LPC de un varón.

Como se observa, los picos de los coeficientes en cada persona tienden a estar en las mismas ubicaciones, y además estos presentan diferentes valores para la señal de voz de la mujer y hombre analizados. Entonces se tiene que los coeficientes LPC proveen en método de extraer las características únicas de voz de cada persona, por lo que basándonos en esta medida será posible clasificar o identificar a una persona.

4.4.2. Algoritmo de detección del Pitch

En un sistema práctico de determinación del hablante es sumamente beneficioso determinar si la persona es un varón o una mujer. Como se mencionó anteriormente el "Pitch" provee una medida para realizar tal distinción. En esta sección se va a describir el algoritmo Harmonic Product Spectrum para tal fin.

El método HPS (Harmonic Product Spectrum) [4] utiliza la transformación en frecuencia (Transformada de Fourier) sobre segmentos cortos de tiempo ajustados por una ventana para poder hallar el Pitch. Este algoritmo está basado en la observación que la voz esta esencialmente compuesto de la frecuencia fundamental f_0 "el Pitch" y una serie de armónicos que ocurren en enteros múltiplos del Pitch (ver Fig. 4.14).

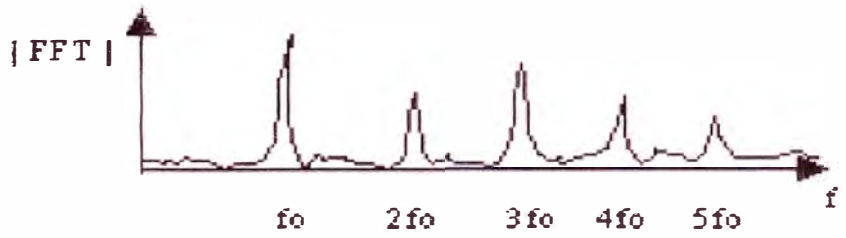

Figura 4.14. Diagrama en frecuencia de una señal de voz.

Entonces si desmembramos esta señal por 2, 3,4,..., y multiplicamos la señal original por estas señales separadas, el producto debe ser una señal con un único pico ubicado en la frecuencia fundamental (ver Fig. 4.15).

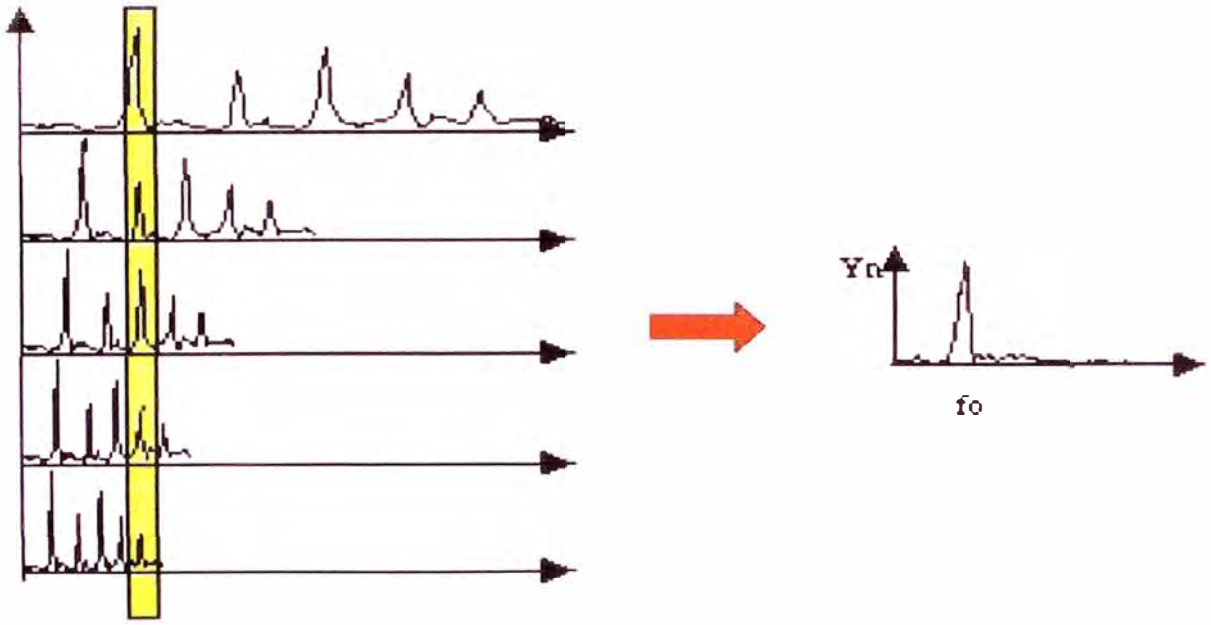

Figura 4.15. Algoritmo HPS.

En la práctica, dado que la una señal de voz, puede contener una variedad de elementos, es mejor hallar el Pitch para segmentos de 30ms y presentar el promedio como el valor final del valor del Pitch de cierta señal de audio [2].

Con el fin de probar este algoritmo y determinar la precisión de nuestro algoritmo de detección de Pitch se muestra en la figura 4.16 el Pitch promedio hallado al implementar el algoritmo anterior y aplicar a 9 señales de voz de una mujer y de un hombre. Como se observa en la figura, claramente el Pitch de una mujer es mayor al Pitch de un hombre, además los valores obtenidos están dentro del rango establecido anteriormente. En general un valor de 150Hz, es la línea divisoria entre el Pitch de una varón y de una mujer.

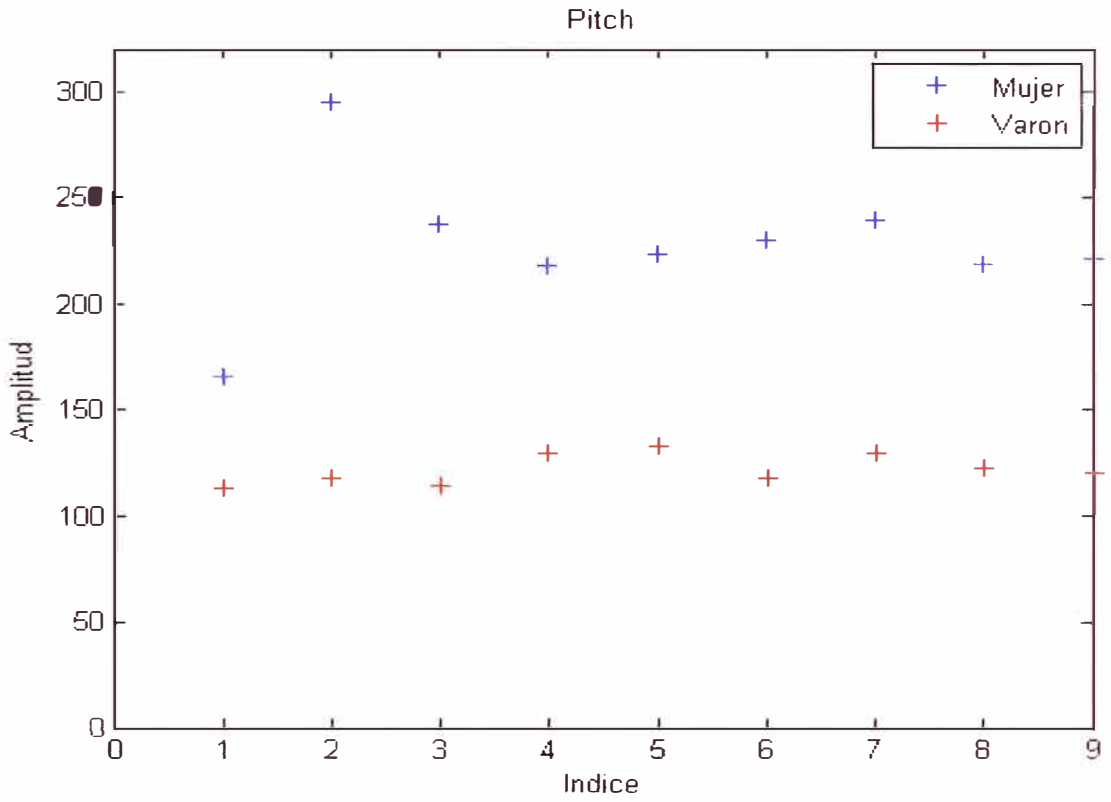

Figura 4.16. Pitch promedio de una mujer y un varón.

Como se verá más adelante, será el Pitch junto con los coeficientes LPC de cada persona las entradas de la red neuronal Perceptrón multicapa. Es decir la red neuronal aprenderá a clasificar a cada persona en base a estas medidas.

4.5. Las redes neuronales y la red perceptrón

4.5.1. Introducción

Las redes neuronales artificiales son modelos estadísticos capaces de desarrollar la tarea de reconocimiento de patrones. Estos modelos tienen sus orígenes

como representaciones de los procesamientos de información que se dan en los sistemas biológicos. Sin embargo, desde la perspectiva del reconocimiento de patrones, la imposición de criterios biológicos no es muy deseable en el desarrollo de sistemas óptimos de reconocimiento. Por lo que en este capítulo se mostraran las redes neuronales desde un enfoque estadístico.

En este capítulo se mostraran los principales conceptos de las redes neuronales que son relevantes en la tarea de clasificación de patrones. Se verán los conceptos de propagación hacia adelante, propagación hacia atrás del error, análisis de la función de error, gradiente de la función de error y los algoritmos de optimización. Todos estos conceptos son importantes a la hora de implementar la red neuronal para la tarea de reconocimiento de patrones.

4.5.2. La red Perceptrón multicapa

Desde una perspectiva matemática, la red Perceptrón multicapa se puede ver como función no lineal [5] con una serie de parámetros adaptivos donde la data fluye hacia adelante, es decir desde las entradas hacia las salidas y no hay algún tipo de retro alimentación. La red más común, y la que se va a usar en el siguiente trabajo tiene la estructura que se muestra en la figura 4.17. Como se observa esta red tiene tres capas, una capa de entrada, una capa de unidades ocultas, y la capa de salida. Donde las dimensiones de las entradas y de las salidas están dadas por la aplicación que se le dé a la red.

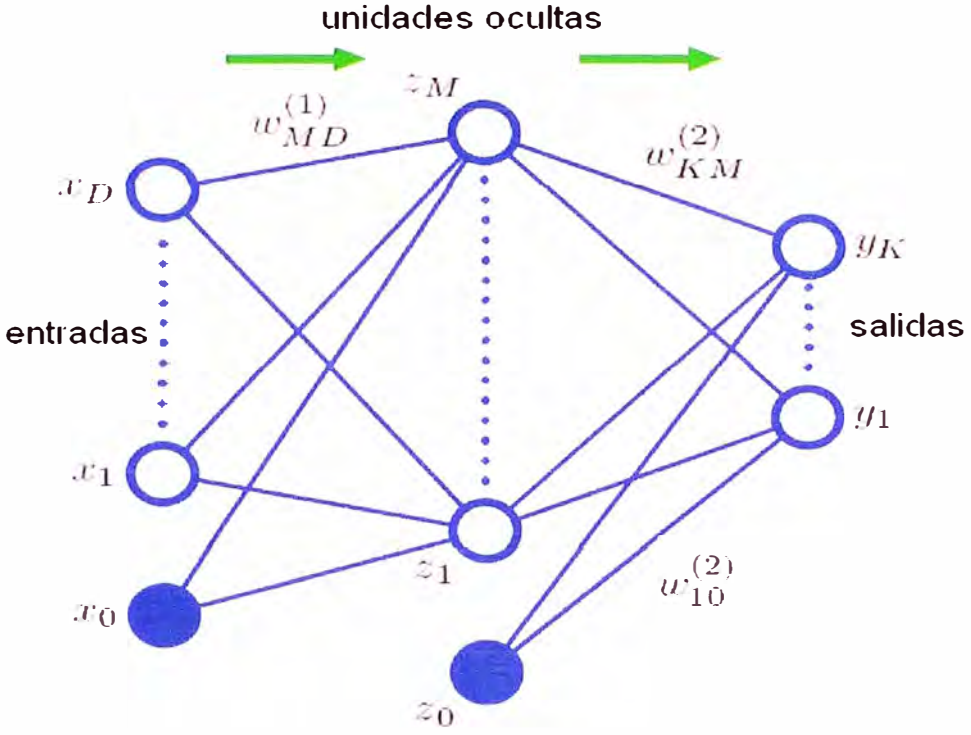

Figura 4.17. Estructura de una red Perceptrón multicapa.

En esta red las entradas $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ a la red se transforman de acuerdo a las siguientes ecuaciones:

$$a_j = \sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \qquad (4.13)$$

$$z_j = h(a_j) \qquad (4.14)$$

Donde $j = 1, \ldots, M$, representa el número de unidades ocultas. Los parámetros $w_{ji}^{(1)}$ y $w_{j0}^{(1)}$ representan los pesos y los bías de la unidad oculta j, y representan la primera capa de parámetros. Las cantidades a_j se conocen como las activaciones de las unidades ocultas las cuales se transforman usando una función de

activación $h(a_j)$, que típicamente se escoge como la función sigmoidal $h(\cdot) = \tanh(\cdot)$, dando lugar a las cantidades z_j que son las salidas de cada unidad oculta.

Las salidas de las unidades ocultas se transforman nuevamente de acuerdo a las ecuaciones:

$$a_k = \sum_{j=1}^{M} w_{kj}^{(2)} \cdot z_j + w_{k0}^{(2)} \qquad (4.15)$$

$$y_k = \sigma(a_k) \qquad (4.16)$$

La función de activación de las salidas se escoge de acuerdo al tipo de aplicación que se le quiere dar a la red. En los problemas de regresión (donde los valores de las salidas son continuos) la función de activación es la identidad:

$$y_k = a_k \qquad (4.17)$$

En los poblemos de clasificación binaria la función de activación es la función Logistic Sigmoid [5]:

$$y_k = \sigma(a_k) = \frac{1}{1+\exp(-a_k)} \qquad (4.18)$$

Y en los problemas de clasificación donde hay varias clases la función de activación es la función Softmax [5]:

$$y_k = \sigma(a_k) = \frac{\exp(a_k)}{\sum_{q=1}^{K} \exp(-a_q)} \qquad (4.19)$$

En el caso de nuestro sistema de reconocimiento se usara las activaciones de salida tipo Softmax debido a que nuestra red tiene varias salidas y en cada caso, es decir en cada patrón de entrada la red solo debe activar una salida mientras que las demás deben tener una salida pequeña. Por ejemplo, la primera salida tendrá un valor grande de salida cada vez que el usuario "Guillermo" ingrese su voz, mientras que la segunda salida tendrá un valor mayor cada vez que el usuario "Marcela" ingrese al sistema.

4.5.3. Función de error

Tal como se muestra en [5], el entrenamiento de la red es equivalente a la minimización de su función correspondiente de error. Esta función de error es dependiente de los parámetros del sistema (los pesos y los bías de las diferentes capas) y del conjunto de entrenamiento (las entradas con sus salidas deseadas). Cabe añadir que esta función es del tipo suave, es decir que tiene un gradiente, el cual puede ser hallado usando el método de back - propagation.

Entonces una vez que se ha encontrado los parámetros que minimizan esta

función de error se dice que el sistema ha sido entrenado, o lo que es lo mismo que es capaz de predecir correctamente las salidas para cualquier entrada.

a) Conjunto de entrenamiento

El conjunto de entrenamiento de la red está dado por el conjunto $D = \{\mathbf{X}, \mathbf{T}\}$, donde $\mathbf{X} = \{x_n\}$ es el conjunto de entradas, $\mathbf{T} = \{t_n\}$ es el conjunto de salidas deseadas, y $n = 1, \ldots, N$ representa el número de patrones de entrenamiento.

En nuestro caso las entradas $\mathbf{X} = \{x_n\}$ estarán dadas por los coeficientes LPC y el pitch que son entregados por el módulo de extracción de características (ver fig. 4.18), y las salidas deseadas $\mathbf{T} = \{t_n\}$ son la codificación que se le da a cada persona.

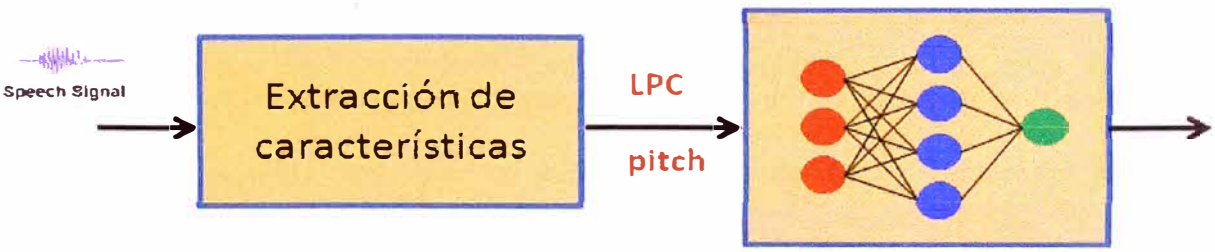

Figura 4.18. Esquema del módulo de reconocimiento.

b) Función de error en la clasificación

En el caso del problema de clasificación binaria con una sola salida, las salidas del conjunto de entrenamiento t se etiquetan tal que $t = 1$ denota pertenencia a la clase 1 y $t = 0$ denota pertenencia a la clase 2. Por otro lado se tiene que la salida

$y(\mathrm{x}, \mathrm{w})$ de la red dada en (4.6) satisface $0 \leq y(\mathrm{x}, \mathrm{w}) \leq 1$, lo que hace posible darle la siguiente interpretación probabilística:

- $p(C_1 \mid \mathrm{x}) = y(\mathrm{x}, \mathrm{w})$: Representa la probabilidad de que la entrada pertenezca a la clase 1.

- $p(C_2 \mid \mathrm{x}) = 1 - y(\mathrm{x}, \mathrm{w})$: Representa la probabilidad de que la entrada pertenezca a la clase 2.

De donde es posible describir las salidas por una distribución de Bernoulli, que es una distribución para variables binarias, de la forma:

$$p(t \mid \mathrm{x}, \mathrm{w}) = y(\mathrm{x}, \mathrm{w})^t \left\{1 - y(\mathrm{x}, \mathrm{w})\right\}^{1-t} \qquad (4.20)$$

En donde se demuestra que la maximización de su función de probabilidad, que busca hallar aquella red que es más probable de haber generado la data de entrenamiento, corresponde a la minimización de la siguiente función de error:

$$E(\mathrm{w}) = -\sum_{n=1}^{N} \left\{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\right\} \qquad (4.21)$$

En el caso de las redes con varias salidas, cada salida tiene una etiqueta o target dado por $\{0,1\}$ y solo una de ellas se activa cada vez. Entonces, en este tipo de redes se tiene que la función de probabilidad del sistema está dado por:

$$p(t \mid \mathrm{x}, \mathrm{w}) = \prod_{k=1}^{K} y_k(\mathrm{x}, \mathrm{w})^{t_k} \left\{1 - y_k(\mathrm{x}, \mathrm{w})\right\}^{1-t_k} \qquad (4.22)$$

Tomando el logaritmo negativo de la función de probabilidad se tiene la siguiente función de error, que es la que se usara en el entrenamiento de nuestro sistema de reconocimiento.

$$E(\mathrm{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn} \ln y_k(\mathrm{x}_n, \mathrm{w}) \qquad (4.23)$$

c) Gradiente de la función de error

Tal como se verá más adelante, la mayoría de algoritmos que hacen el entrenamiento de la red hacen uso del gradiente de la función de error. De hecho el entrenamiento se puede dividir en dos etapas. La primera que busca evaluar el gradiente del error en un punto dado, y la segunda que busca ajustar los parámetros tomando en cuenta estas derivadas. En esta sección se describirá el algoritmo de propagación hacia atrás o back-propagation para hallar eficientemente las derivadas de la función de error.

El algoritmo de back-propagation tal como se muestra en [5], de manera general consta de las siguientes etapas:

1. Aplicar un vector de entrada x_n a la red y propagarlo a través de la red usando las ecuaciones (4.1), (4.2), (4.3) y (4.4). Esto con el fin de hallar las activaciones de las diversas unidades.

2. Evaluar los parámetros deltas δ_k de las unidades de la capa de salida usando las ecuación:

$$\delta_k = y_k - t_k \qquad (4.24)$$

3. Propagar hacia atrás los δ_k para hallar los δ_j que corresponden a los deltas de las unidades escondidas, usando la ecuación:

$$\delta_j = h'(a_j)\sum_k w_{kj}\delta_k \qquad (4.25)$$

Gráficamente, el proceso que realiza la ecuación (4.19) está representada por la siguiente figura, donde se aprecia como los deltas de las salidas se propagan hacia atrás.

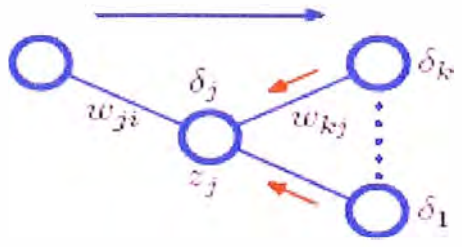

Figura 4.19. Propagación hacia atrás.

4. Finalmente hallamos las derivadas usando las ecuación:

$$\frac{\partial E}{\partial w_{ji}} = \delta_j z_i \qquad (4.26)$$

A continuación se mostrar el algoritmo back-propagation que se implementara en este trabajo el cual considera funciones de activación de las unidades escondidas tipo logistic sigmoid:

$$h(a) = \tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}} \qquad (4.27)$$

El cual tiene la propiedad:

$$h'(a) = 1 - h(a)^2 \qquad (4.28)$$

Entonces para cada patrón en el conjunto de entrenamiento, primero evaluamos la propagación hacia adelante usando:

$$a_j = \sum_{i=0}^{D} w_{ji}^{(1)} x_i \qquad (4.29)$$

$$z_j = \tanh(a_j) \qquad (4.30)$$

$$a_k = \sum_{j=0}^{M} w_{kj}^{(2)} \cdot z_j \qquad (4.31)$$

$$y_k = \sigma(a_k) \qquad (4.32)$$

Luego, evaluamos los deltas de las salidas:

$$\delta_k = y_k - t_k \qquad (4.33)$$

Entonces, propagamos hacia atrás los deltas de las salidas para hallar los deltas de las unidades escondidas:

$$\delta_j = (1 - z_j^2) \sum_{k=1}^{K} w_{kj} \delta_k \qquad (4.34)$$

Finalmente las derivadas del error con respecto a los pesos de la primera y segunda capa están dados por:

$$\frac{\partial E}{\partial w_{ji}^{(1)}} = \delta_j x_i \qquad \frac{\partial E}{\partial w_{kj}^{(2)}} = \delta_k z_j \qquad (4.35)$$

4.5.4. Algoritmos de entrenamiento

Una vez obtenido la función de error y el gradiente de esta función con respecto a los parámetros, lo que sigue es hallar los parámetros que minimizan la función de error.

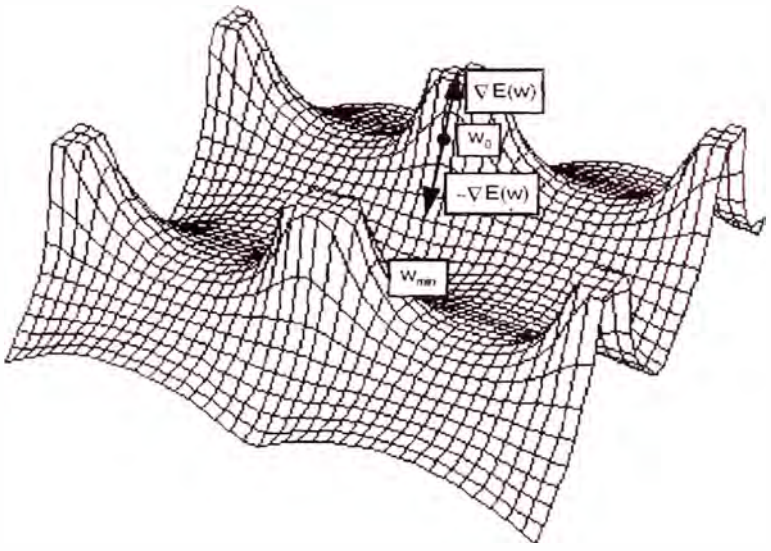

Figura 4.20. Típica función de error de una red Perceptrón multicapa.

Debido a la complejidad de la función de error la localización del punto de error mínimo no se puede hacer a través de una formula cerrada sino que se realiza a través de métodos iterativos que requieren la elección de un punto inicial y donde se hace una búsqueda en el espacio de error en una sucesión de pasos de la forma:

$$w^{(\tau+1)} = w^{(\tau)} + \Delta w^{(\tau)} \qquad (4.36)$$

Donde τ representa el paso de la iteración y $\Delta w^{(\tau)}$ representa el incremento del vector de parámetros en el paso τ. Diferentes elecciones de este incremento representan diferentes algoritmos de optimización tales como el método de disminución del gradiente, gradiente conjugado, etc.

Típicamente, estos algoritmos de optimización hacen uso del gradiente del error para determinar el vector de incremento de parámetros $\Delta w^{(\tau)}$. Este gradiente se calcula de manera óptima haciendo uso de la técnica de retro-propagación del error.

4.5.5. Método de quasi-newton

Tal como se muestra en [5], el algoritmo de quasi-newton es uno de los algoritmos mejor acondicionados para hallar el mínimo de una función, lo que significa que se puede usar para hallar el mínimo de una función de error.

Este algoritmo hace la búsqueda iterativa en el espacio de parámetros, buscando la minimización del error, usando la fórmula:

$$\mathrm{w}_{j+1} = \mathrm{w}_j + \alpha_j \mathrm{G}_j \mathrm{g}_j \qquad (4.37)$$

Dónde:

α_j : Longitud del paso que se hace en cada búsqueda

G_j : Aproximación de la inversa de la matriz Hesiana de la función a minimizar.

g_j : Gradiente de la función a minimizar

Como se observa en (4.35) este algoritmo hace uso explícito del gradiente de la función de error. En el caso de la red neuronal a usar, este gradiente se calcula usando el algoritmo back-propagation descrito en la sección anterior.

La matriz G se halla de manera iterativa usando la fórmula de Broyden-Fletcher - Goldfarb - Shannon (BFGS):

$$G_{j+1} = G_j + \frac{\mathrm{pp}^T}{\mathrm{p}^T \mathrm{v}} - \frac{(G_j \mathrm{v})\mathrm{v}^T G_j}{\mathrm{v}^T G \mathrm{v}} + (\mathrm{v}^T G_j \mathrm{v})\mathbf{u}\mathbf{u}^T \qquad (4.38)$$

Dónde:

$$\mathrm{p} = \mathrm{w}_{j+1} - \mathrm{w}_j \qquad (4.39)$$

$$\mathrm{v} = \mathrm{g}_{j+1} - \mathrm{g}_j \qquad (4.40)$$

$$\mathbf{u} = \frac{\mathbf{p}}{\mathbf{p}^T \mathbf{v}} - \frac{G_j \mathbf{v}}{\mathbf{v} G_j \mathbf{v}} \qquad (4.41)$$

El valor inicial de la matriz G es la matriz identidad. Se demuestra que cada paso de este algoritmo minimiza el error. Lo que es más este método es mucho más óptimo que el método tradicional de disminución del gradiente.

4.5.6. Regularización de la red

El número de entradas y salidas de la red está determinada por la dimensión de los elementos del conjunto de entrenamiento, sin embargo aún queda el problema de determinar el número de unidades escondidas "M" a usar. Este problema se llama como selección del modelo y es de importancia capital en el reconocimiento de patrones.

Si se usa un modelo muy simple (valor de M pequeño) el modelo no es capaz de reproducir de manera de adecuada la data (under-fitting), si se usa un modelo muy complejo (valor de M grande) el modelo reproduce de manera exacta el conjunto de entrenamiento, sin embargo, exhibe un comportamiento muy oscilante y no predice correctamente las salidas dada nuevas entradas (over-fitting), en general los modelos de complejidad intermedia son los que predicen mejor las salidas para nuevos valores de entradas (ver Fig. 4.21).

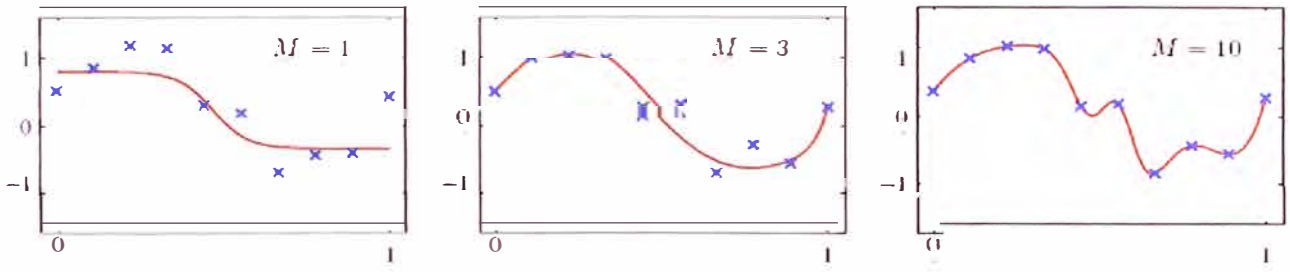

Figura 4.21. Problema de selección del modelo.

Una solución a este problema se hace mediante la adición de términos de regularización a la función de error. En particular el regulador más simple es el cuadrático, dando lugar a un error regularizado de la forma:

$$\tilde{E}(\mathbf{w}) = E(\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \qquad (4.42)$$

Entonces el problema de complejidad de modelo se traslada a encontrar el valor adecuado del valor λ.

4.6. Reconocimiento de patrones

Mucha de la información que se maneja en la vida real se presenta en la forma de patrones complejos: caras, textos escritos, enfermedades, música, flores, piezas industriales, etc. En el contexto de este trabajo se nombró reconocimiento de patrones como reconocimiento de patrones de voz.

Aunque la aplicabilidad de las diversas técnicas resulta, a priori, muy amplia, no hay un método que sea solución óptima. Diversas razones hacen que los sistemas de reconocimiento de formas sean muy específicos del problema a resolver:

La naturaleza de los patrones: caracteres escritos, símbolos, dibujos, imágenes biomédicas, objetos tridimensionales, firmas, huellas dactilares, espectrogramas, imágenes de teledetección, cromosomas, etc.

Los requerimientos del sistema, especialmente en tiempo de respuesta hace que algunos métodos de reconocimiento, aun siendo superiores en éxito no sean aplicables en la práctica.

Factores económicos: un sistema equipado con diferentes sensores y equipos de procesamiento muy potentes pueden dar resultados muy satisfactorios pero no pueden ser costeados por los usuarios.

A grandes rasgos podemos dividir los diferentes esquemas de reconocimiento de patrones como se muestra a continuación (Carnell, 1998):

4.6.1. Sistema de detección y clasificación de patrones

La Clasificación de patrones es el acto de asignar una etiqueta de clase a un objeto, un proceso físico o un evento. La asignación se basa siempre en las mediciones que se obtienen de ese objeto o proceso o evento, las mediciones estarán disponibles a partir de un sistema sensorial.

CAPITULO 5

SISTEMA DE RECONOCIMIENTO DE PATRONES DE VOZ MEDIANTE RED PERCEPTRÓN MULTICAPA

5.1. Estructura del sistema de reconocimiento

Hasta este punto se ha discutido los principales componentes del sistema de reconocimiento de hablante. Se ha discutido la etapa de pre-procesamiento y extracción de características en donde se elimina las señalas no deseadas y se extraen una serie de características que distinguen únicamente a la señal de voz de entrada; y la etapa de reconocimiento la cual se implementa usando una red Perceptrón multicapa.

En este capítulo, mostraremos en detalle la arquitectura final del sistema de reconocimiento del hablante. Primero mostraremos como se conectan el sistema de adquisición de voz con el sistema de extracción de características, y luego mostraremos como se da la conexión entre el sistema de extracción de características

y el sistema de reconocimiento, estableciendo de esta manera el número de entradas de la red neuronal.

Tal como se muestra en la Figura 5.1., el sistema de reconocimiento consta básicamente de dos etapas bien definidas: La etapa de extracción de características en donde primero se eliminan las señales no deseadas para luego extraer una serie de características que describen la señal de voz; y la etapa de reconocimiento que toma las características halladas como entradas y aprende la relación entre las entradas con la persona que las ha generado [6], [7], [8], [9].

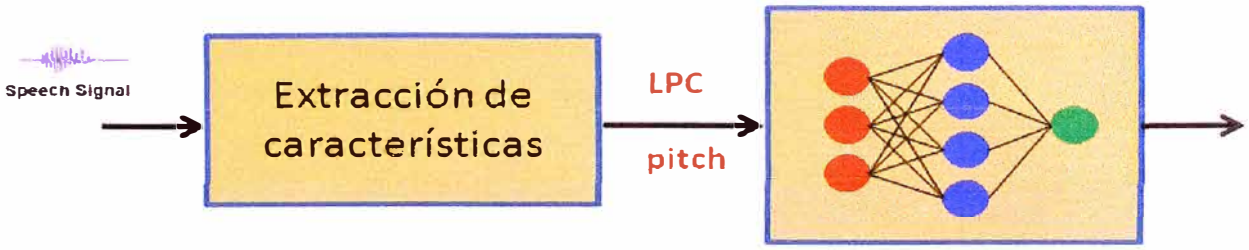

Figura 5.1. Esquema del sistema de reconocimiento.

Además se debe tomar en cuenta que en el sistema existe una etapa previa que es la etapa de toma de señales de voz, la cual se hace a través de micrófonos y software de adquisición de datos; una etapa de pre procesamiento de señales donde se eliminan las señales no deseadas; y una etapa posterior que se encarga de realizar una acción determinada en base a los resultados obtenidos, por ejemplo puede decidir si permite o no el acceso de una persona a un lugar determinado o si le permite el

acceso a cierta computadora. En las siguientes secciones se describirán en detalle cada una de estas etapas, tanto con el fin de tener una mayor compresión del sistema así como de facilitar su implementación.

5.2. Adquisición de señales de voz

La primera etapa consiste en adquirir la voz de una persona y hacerla disponible al sistema de reconocimiento. En el caso de nuestro sistema, la voz se adquiere por computadora a través de un micrófono, el cual dependiendo de la computadora puede estar ya integrado o puede ser conectado fácilmente a la PC (ver Fig. 5.2).

Figura 5.2. Etapa de adquisición de voz.

En el caso de la lectura de datos de la voz, el software MATLAB provee una serie de funciones que permite la lectura y escritura de señales de voz tanto en formato " *.wav" como en otros formatos. En nuestro caso, usaremos el formato

".wav" no solo por su facilidad y gran aceptación sino por la facilidad con la que se puede especificar parámetros tales como: la frecuencia de muestreo, número de canales, y el número de bits a usar.

Entonces, luego de conectar el micrófono y usar las funciones de lectura de audio lo que se tiene es un vector "x" que consiste en números en formato "double" que especifican los valores de la señal en cada instante de tiempo. Graficando este vector se tiene un resultado parecido al que se muestra en la Fig. 5.3.

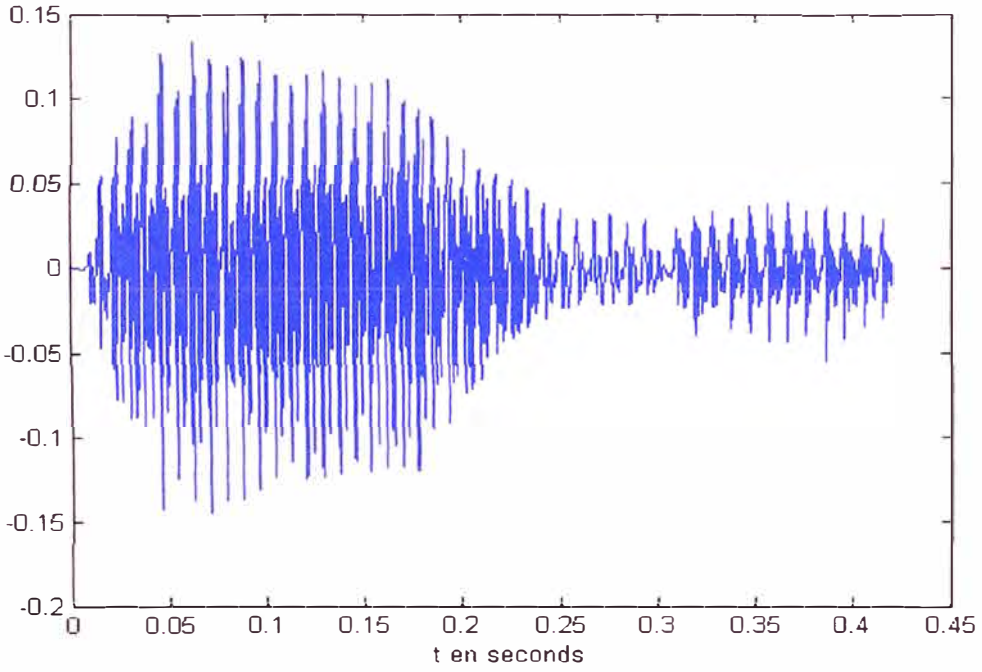

Figura 5.3. Señal de voz.

5.3. Obtención de la señal pre - procesada

La siguiente etapa consiste en eliminar las señales no deseadas. Tal como se mencionó en el capítulo 4, lo que hace esta etapa es tomar la señal de voz de entrada y eliminar el ruido y las señales que no corresponde a voz, y produce como salida una nueva señal de voz "limpia" (ver Fig. 5.4).

Figura 5.4. Etapa de pre procesamiento.

Por ejemplo para la señal de voz mostrada en la Fig. 5.5., la señal que se obtiene es la que se muestra en la Fig. 5.6.

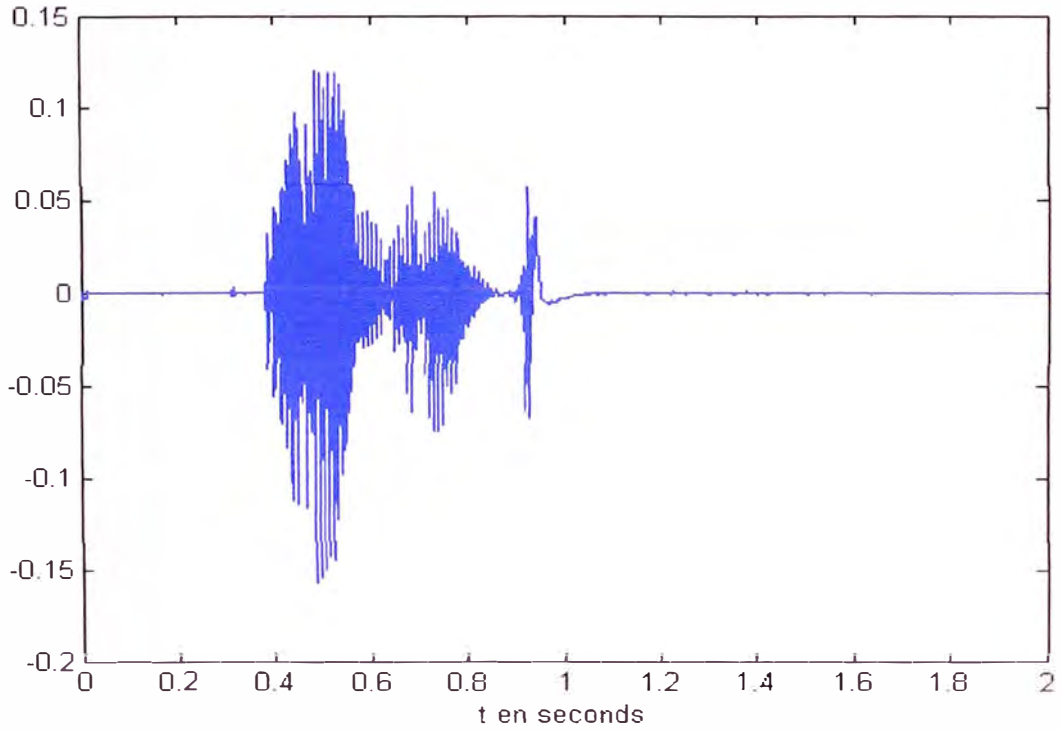

Figura 5.5. Señal de voz.

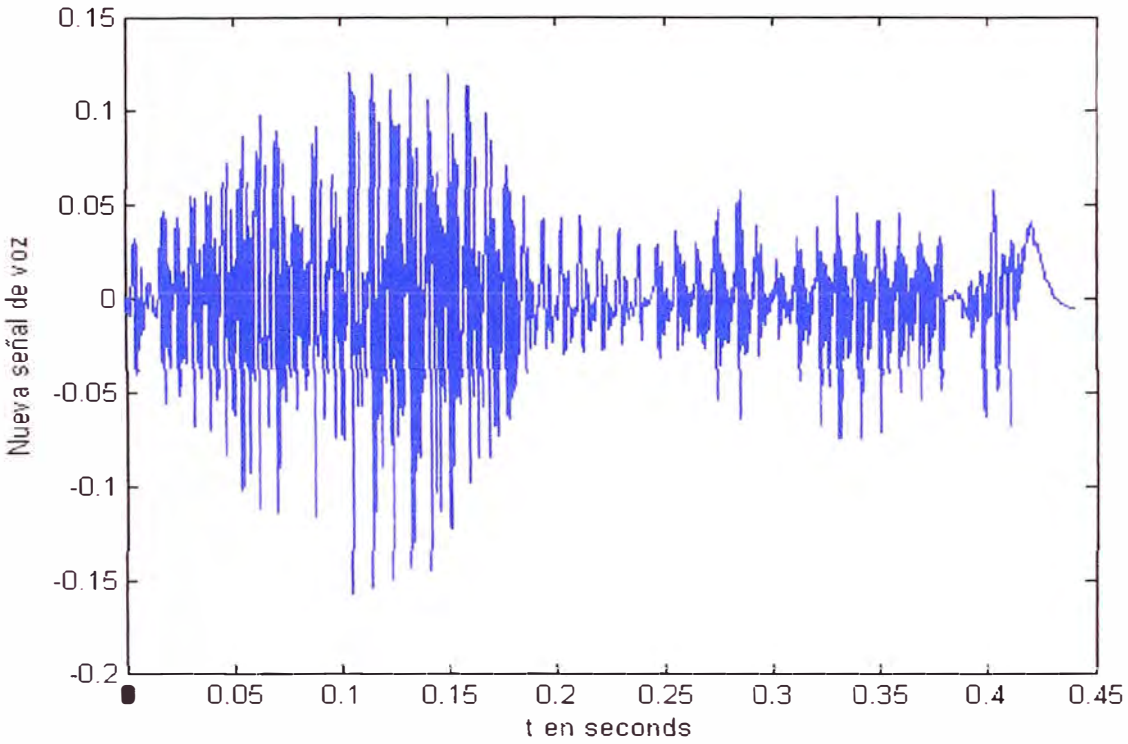

Figura 5.6. Señal de voz limpiada.

5.4. Extracción de características de las señales de voz

La etapa de extracción de características consiste en tomar la señal de voz "limpia" y extraer una serie de características únicas que distinguen a cada persona (ver Fig. 5.7).

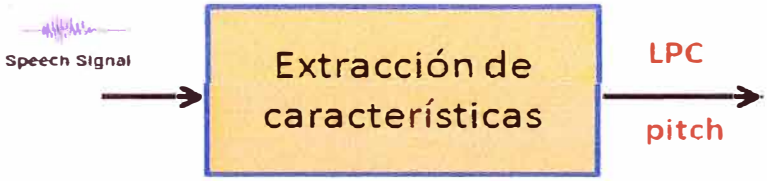

Figura 5.7. Etapa de extracción de características.

Como se mencionó en los capítulos previos, se han escogido los coeficientes LPC y el pitch como estos indicadores ya que estos proveen una serie de medidas únicas que permiten representa a cada persona.

Entonces, luego de aplicar estos algoritmos a la señal de voz lo que se tiene es un vector "x" de características (ver. Fig. 5.8) que consta de los coeficientes LPC y el valor del pitch de la señal de voz aplicada.

$$\mathbf{x} = \begin{pmatrix} a_1 \\ \vdots \\ a_p \\ p \end{pmatrix}_{13x1}$$

LPC

pitch

Figura 5.8. Vector de características.

5.5. Obtención del sistema de reconocimiento

En la etapa de reconocimiento (ver. Fig. 5.9) se toman el vector de características "x" que contiene los coeficientes LPC y el pitch y la red neuronal se encarga de aprender la identidad de la persona que género este vector.

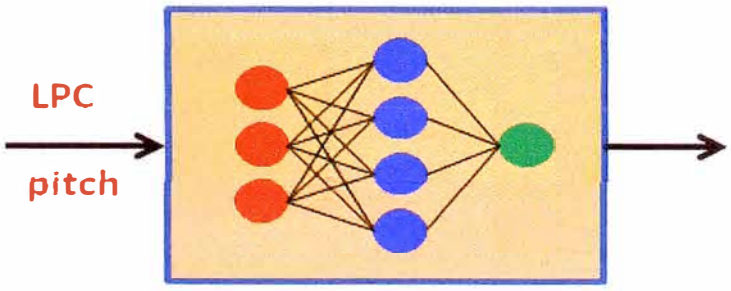

Figura 5.9. Etapa de reconocimiento.

Para el entrenamiento de la red se toman una cantidad determinada de muestras de voz para cada persona, toda esta data constituirá nuestro conjunto de entrenamiento. Además se tomaran unas muestras adicionales de cada persona, con el fin de tener nuestro conjunto de prueba que nos servirá para medir el desempeño de la red.

Debido a la dimensionalidad de las entradas será necesario construir una base de datos con una cantidad suficiente de muestras de voz, esto debido a la gran cantidad de pesos y bías que tiene la red neuronal. Además, con el fin de incrementar el desempeño de la red, se deberá usar un parámetro adecuado de regularización.

5.6. Comprobación de la calidad del sistema de reconocimiento

Una vez que el sistema ha sido entrenado, el sistema ser capaz de reconocer la identidad de las personas que están en su base de datos de entrenamiento. El proceso que sigue la voz en esta etapa es el que se muestra en la figura 5.10, donde los parámetros de la red neuronal ya están bien definidos.

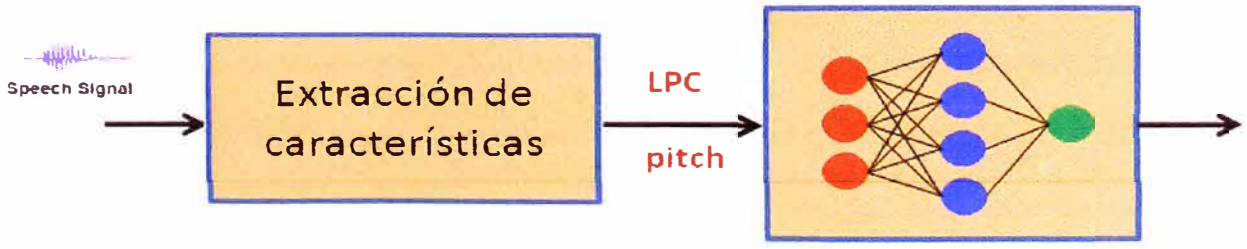

Figura 5.10. Etapa de funcionamiento.

Las salidas de la red son capaces de indicar la identidad de la persona que emitió la voz de entrada y en base a esta salida, se pueden desarrollar cualquiera de las aplicaciones que se muestran en el Capítulo 2.

Una vez establecido la estructura del sistema de reconocimiento queda evaluar el desempeño de nuestro sistema. Para tal fin se tiene que establecer el conjunto de entrenamiento que consiste en las señales de voz de un número determinado de personas. Una vez obtenido este conjunto, se procede con el entrenamiento del sistema, para finalmente verificar su capacidad de reconocimiento.

5.6.1. Consideraciones de ensayo

En este capítulo se harán las pruebas y resultados del sistema de reconocimiento en las señales de voz de un número determinado de personas.

En primer lugar se obtendrá la base de datos, estos es las señales de voz de un número de personas, luego se mostraran los vectores de características de cada persona.

Entonces se hará el entrenamiento del sistema, es decir el entrenamiento de la red neuronal, para finalmente se evaluar los resultados obtenidos, es decir la capacidad de reconocimiento del sistema en las personas que componen la base de datos.

5.6.2. Tipo de prueba

Para realizar la prueba se probaron una a una cada una de las partes del programa, para luego entrenar la red y obtener el conjunto de entrenamiento. Se tomaron muestras de voz por cada palabra para obtener patrones de voz en cada palabra dando un total de 30 muestras por cada palabra de un total de 4.

5.6.3. Variables independientes y dependientes

a) Variables independientes: El número de personas, que son 5.
b) Las variables dependientes son: Conjunto de entrenamiento, que depende de la cantidad de personas, y el número de unidades escondidas.

Conjunto de entrenamiento

El conjunto de entrenamiento de nuestro sistema consiste de las voces de 5 personas, tres hombres y dos mujeres. Para cada persona se han registrado 30 muestras de las siguientes palabras "Hola", "Acceso", y "Conexión", lo cual significa que para cada persona se tienen 90 muestras, de donde 60 muestras se usaran para el entrenamiento de la red y 30 muestras se usaran para la etapa de prueba. Entonces en total se tienen 360 patrones de entrenamiento y 120 patrones de prueba. Como se observa, la data de entrenamiento es en tamaño mayor que la data de prueba, esto es porque se necesita abundante data para entrenar correctamente todos los pesos de la red.

Como se observa en la data de entrenamiento, se ha usado un número igual de varones y mujeres, esto es con el fin de ver si la red tiene problemas a la hora de distinguir entre hombres y mujeres. Por otro lado, se debe tener en cuenta que se va a entrenar la red en el reconocimiento basado en ciertas palabras, esto significa que nuestro sistema es del tipo "dependiente del texto".

Número de unidades escondidas: Selección adecuada

Tal como se muestra en [5], el número de unidades escondidas se escoge de tal manera que se maximice el rendimiento de la red. En nuestro caso, se escoge aquel número que maximice la tasa de reconocimiento. En la Fig. 5.3. se muestra el resultado de la tasa de reconocimiento en la palabra "conexión" usando diferente número de unidades escondidas, como se observa se tiene el mayor rendimiento cuando el número de unidades escondidas es 12, por lo que este valor es el que se usa en la arquitectura final del sistema. Ver figura 5.11.

De igual manera, se hizo un análisis similar con las demás palabras evaluadas, donde se tuvo que el número de unidades escondidas de 12 resulta en un mayor rendimiento. Cabe añadir que estas evaluaciones se hicieron en el conjunto de prueba y no en el conjunto de entrenamiento, esto con el fin de evaluar el correcto aprendizaje de la red neuronal.

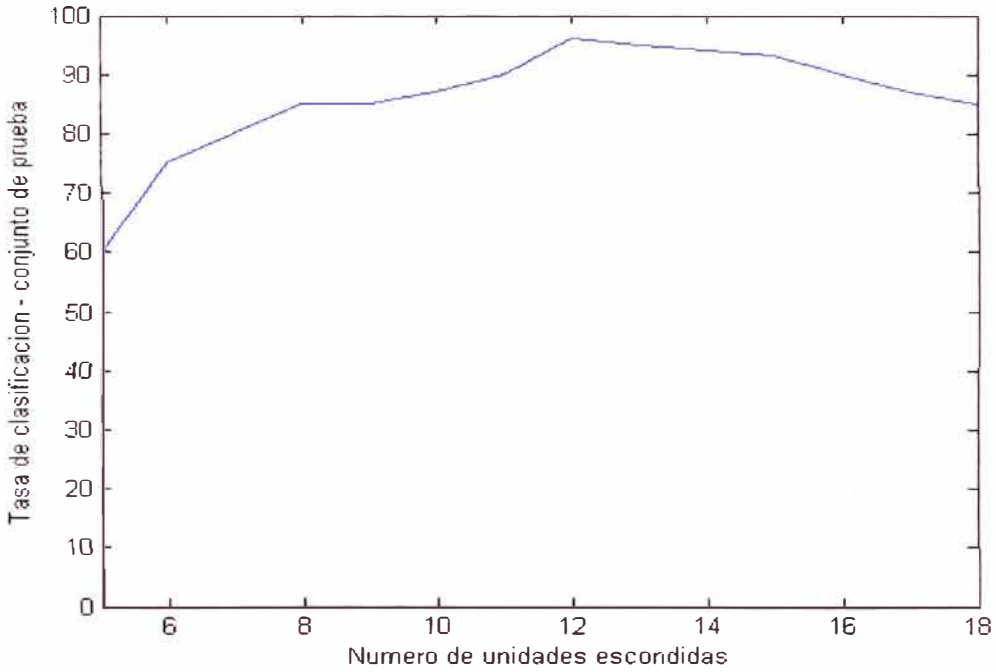

Figura 5.11. Selección del número de unidades escondidas

5.6.4. Extracción de características

Como se observó en los capítulos anteriores, se va a aplicar un algoritmo de extracción de características a las muestras de voz. En esta sección se va a hacer una inspección visual de las características de cada persona, es decir se va a visualizar los coeficientes LPC y el valor del pitch, para de esta manera apreciar los patrones de cada persona y como estos son diferentes para cada persona.

a. Características de la primera persona

En la figura 5.12., se muestran los coeficientes de la primera persona "Arturo" usando la palabra hola, tal como se puede apreciar, el patrón de los coeficientes LPC

está bien definido alrededor de ciertos valores, lo que permite usarlos en la etapa de clasificación. En la figura 5.13, se muestra el pitch hallado, donde se observa que los valores hallados son menores de 150, lo cual es válido.

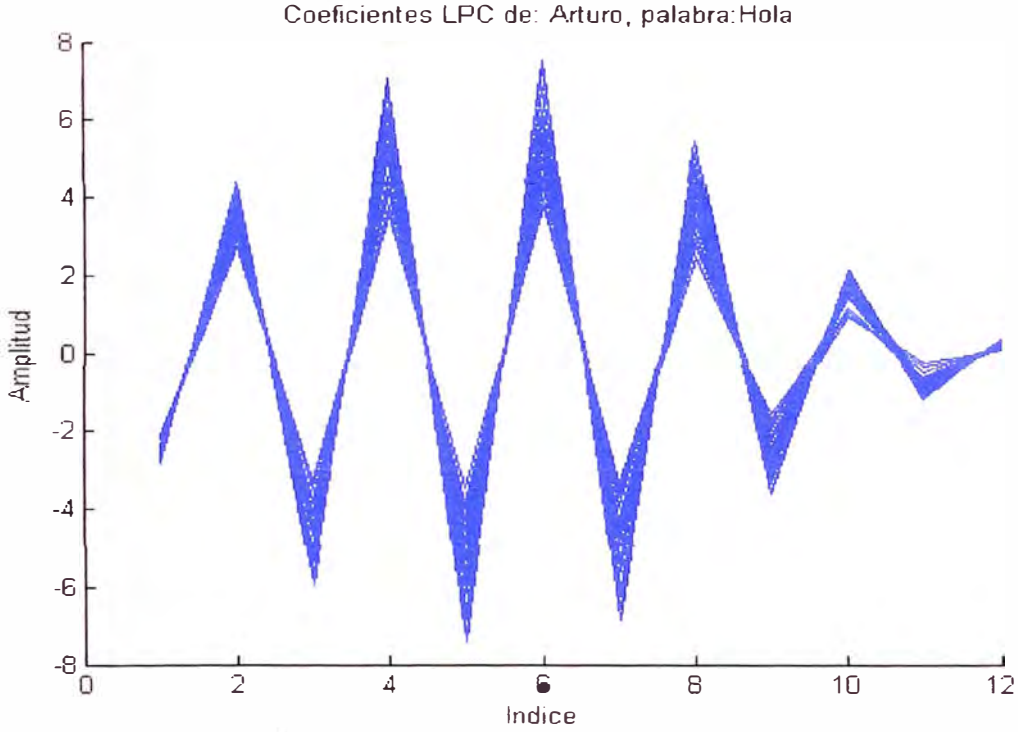

Figura 5.12. Coeficientes LPC de la 1era persona, palabra "Hola".

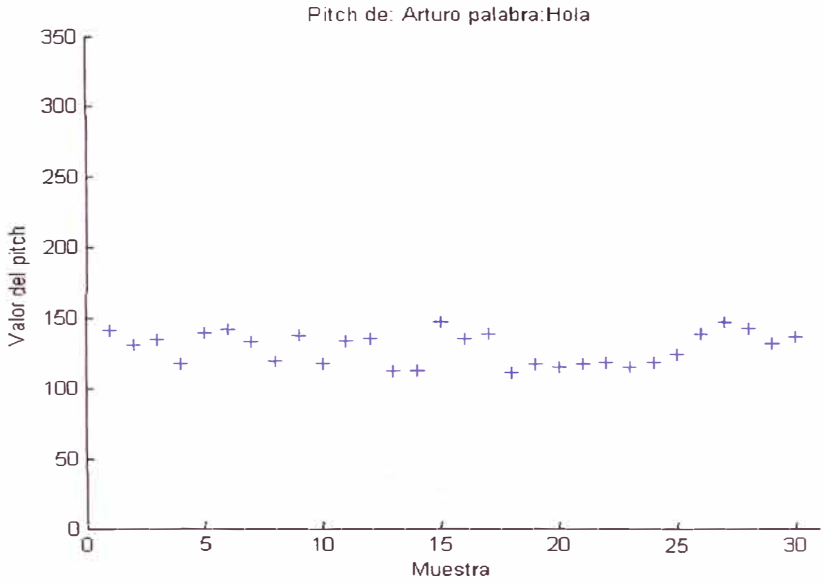

Figura 5.13. *Pitch* de la 1era persona, palabra "Hola".

En las figura 5.14, y 5.15, se muestran los coeficientes y el pitch de la primera persona "Arturo" usando la palabra "Acceso", verificándose las conclusiones anteriores.

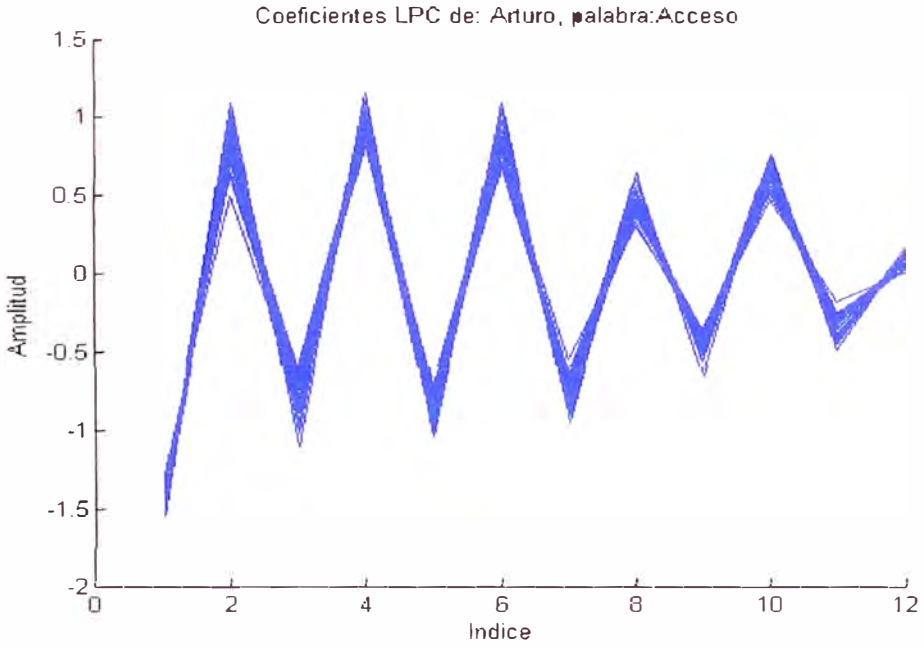

Figura 5.14. Coeficientes LPC de la 1era persona, palabra "Acceso".

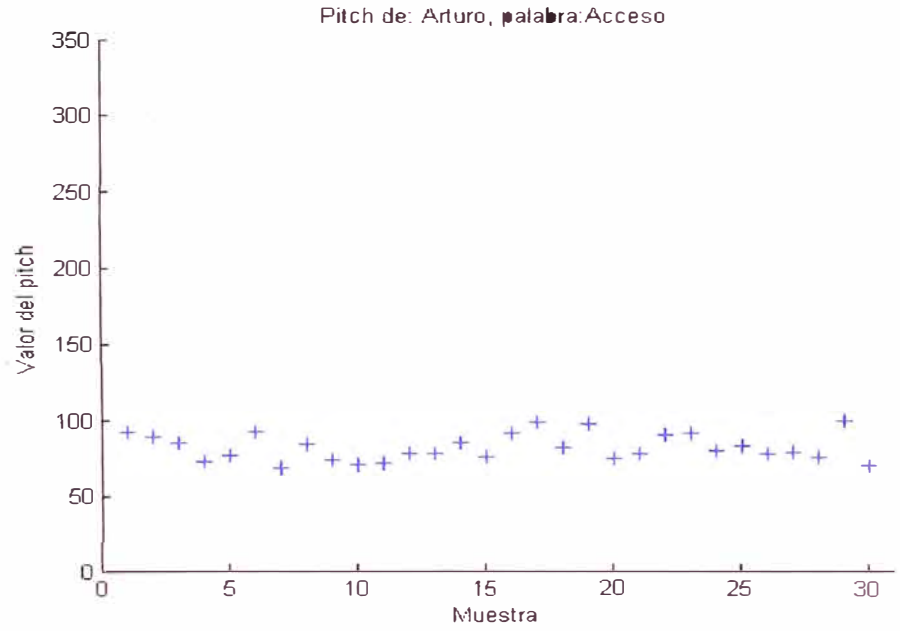

Figura 5.15. Pitch de la 1era persona, palabra "Acceso".

En las figura 5.16., y 5.17., se muestran los coeficientes y el pitch de la primera persona "Arturo" usando la palabra "Conexión", verificándose las conclusiones anteriores.

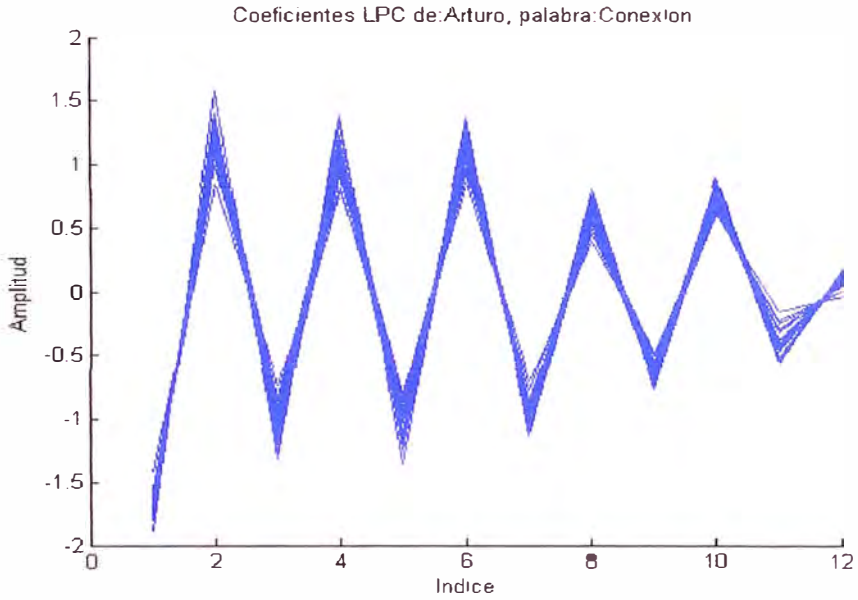

Figura 5.16. Coeficientes LPC de la 1era persona, palabra "Conexión".

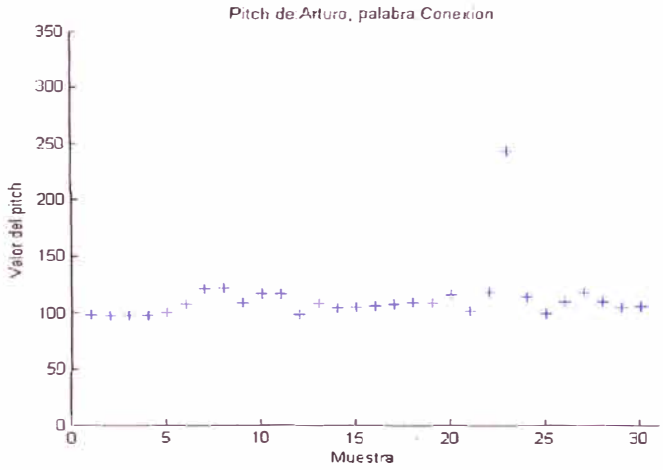

Figura 5.17. Pitch de la 1era persona, palabra "Conexión".

b. Características de la segunda persona

Ahora, en las siguientes figuras 5.18, 5.19, y 5.20, se muestran los coeficientes LPC de la 2da persona, donde se observa que se repite nuevamente el patrón de características de la misma persona.

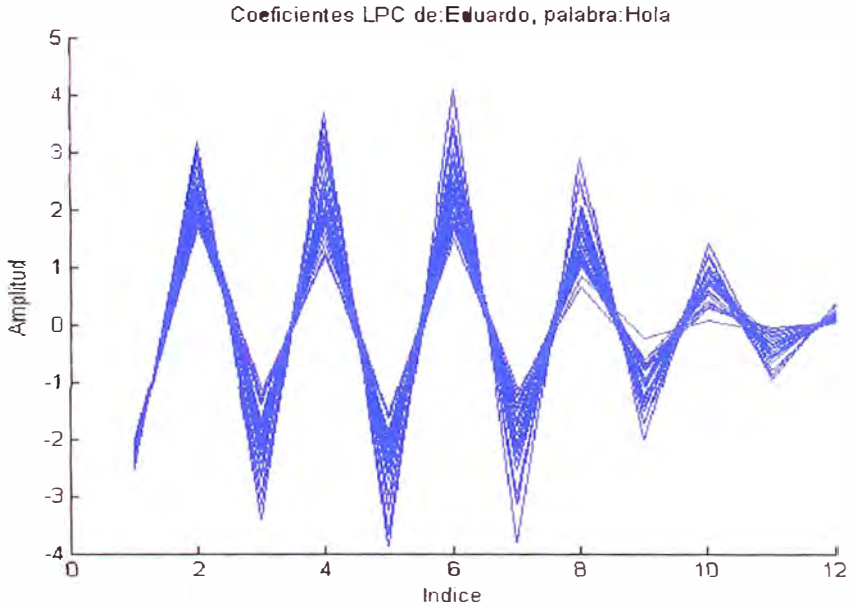

Figura 5.18. Coeficientes LPC de la 2da persona, palabra "Hola"

Es importante notar que estos patrones se repiten, lo cual confirma que los descriptores seleccionados son adecuados, y que el algoritmo LPC luego del número de repeticiones o muestras grabadas evidencia un patrón, el cual nos permite entrenar a la red y obtener el conjunto de entrenamiento.

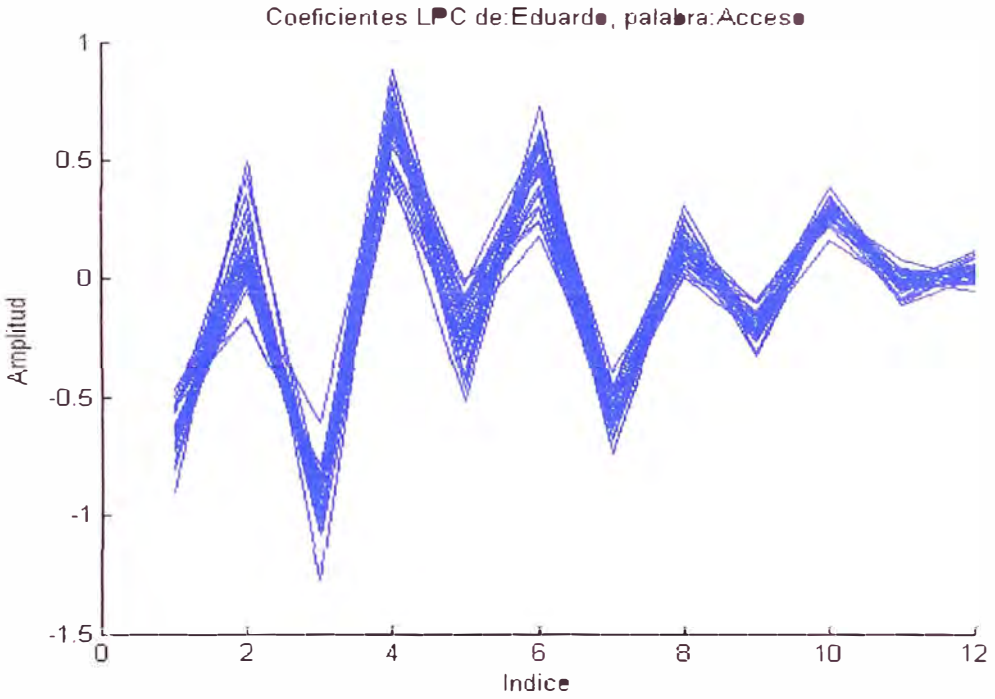

Figura 5.19. Coeficientes LPC de la 2da persona, palabra "Acceso"

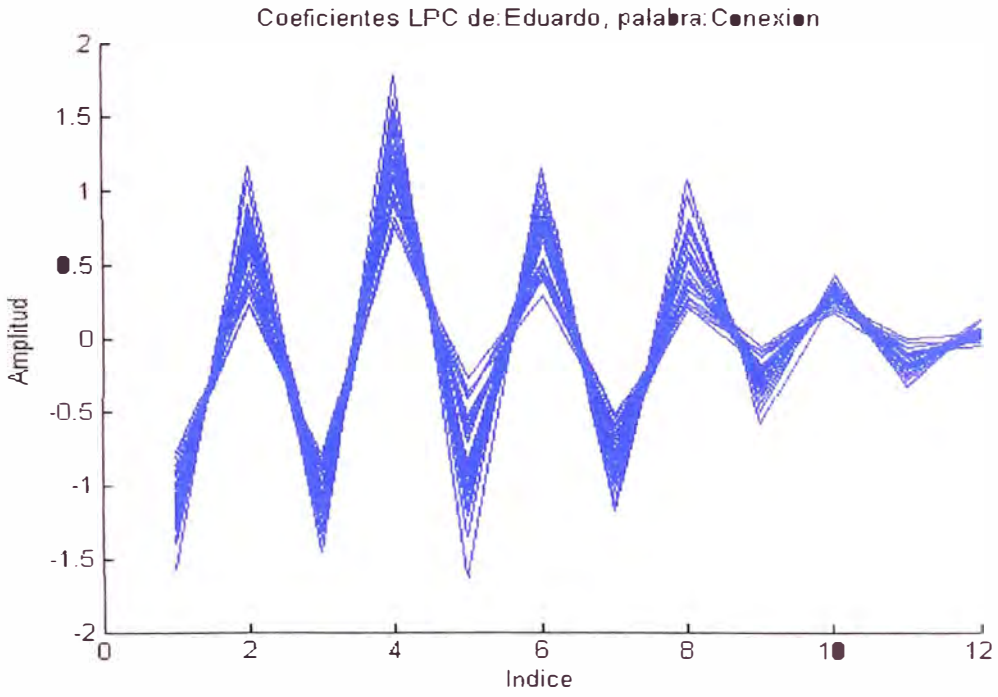

Figura 5.20. Coeficientes LPC de la 2da persona, palabra "Conexión"

c. Características de la tercera persona

En las figuras 5.21 y 5.22 se muestran los coeficientes LPC y el pitch de la 3era persona que es una mujer, donde se puede ver nuevamente que los valores LPC siguen un patrón característico, y además se cumple que el pitch es superior a 150 tal como se esperaba. Cabe observar en este punto que si bien los coeficientes parecen tener la misma forma que alguno de los coeficientes anteriores, se debe considerar la amplitud para apreciar las diferencias entre los coeficientes LPC de diferentes personas

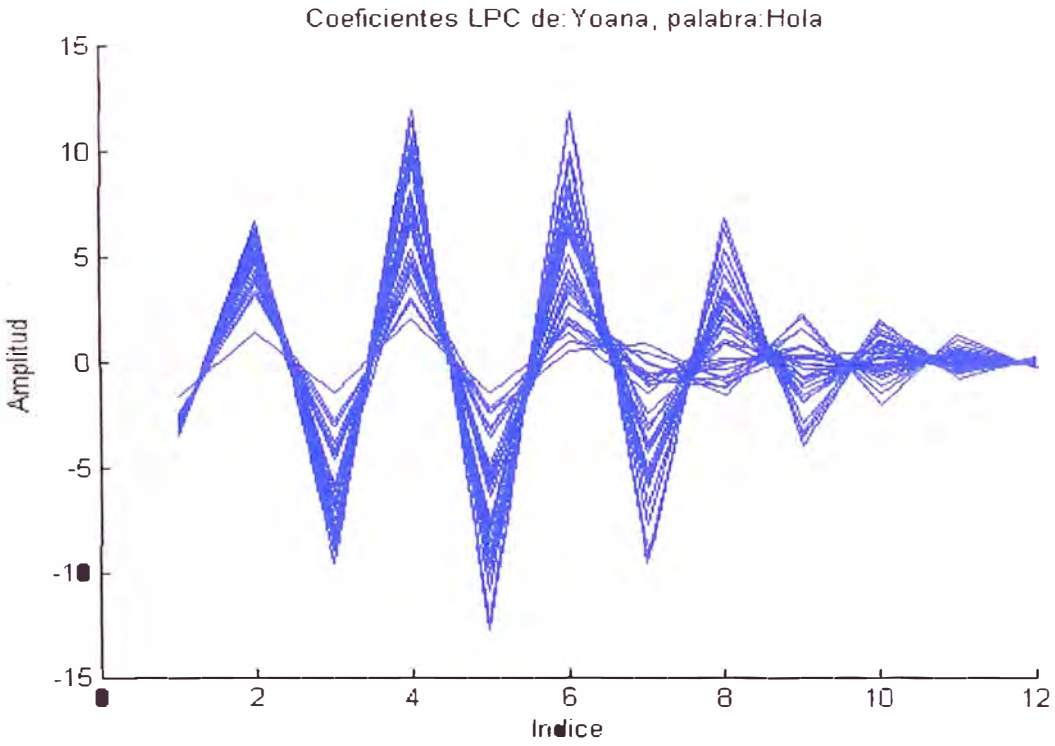

Figura 5.21. Coeficientes LPC de la 3ra persona, palabra "Hola".

Según la figura 5.22, se confirma el género del hablante al menos durante el 96.67% del muestreo.

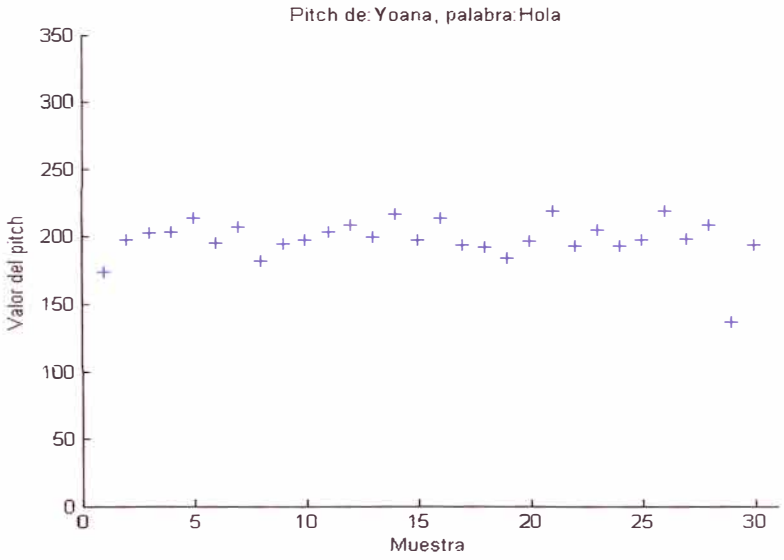

Figura 5.22. Pitch de la 3era persona, palabra "Hola".

Seguidamente, en las figuras 5.23 y 5.24 se muestran los coeficientes LPC y el pitch usando la segunda palabra "Acceso".

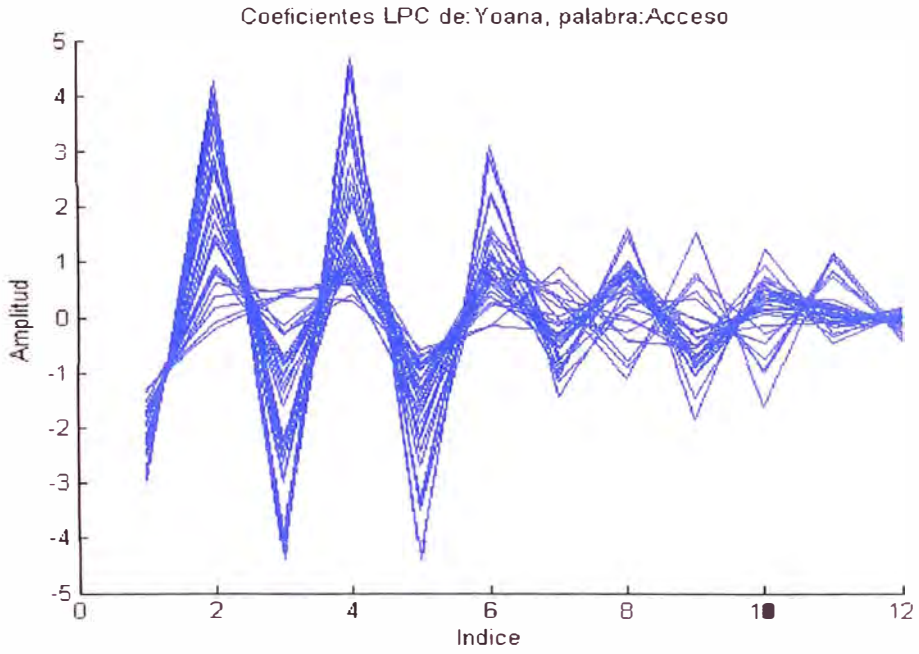

Figura 5.23. Coeficientes LPC de la 3ra persona, palabra "Acceso".

Figura 5.15.

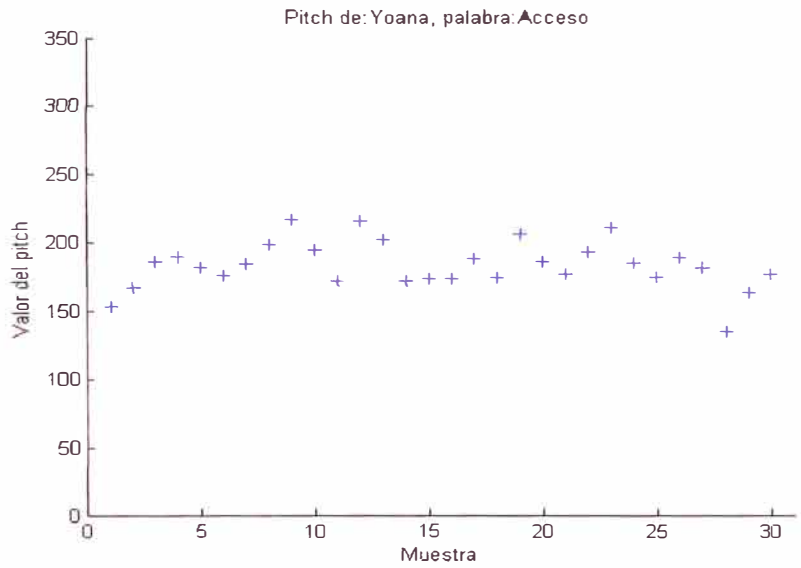

Figura 5.24. Pitch de la 3era persona, palabra "Acceso".

Seguidamente, en las figuras 5.25 y 5.26 se muestran los coeficientes LPC y el pitch usando la tercera palabra "Acceso".

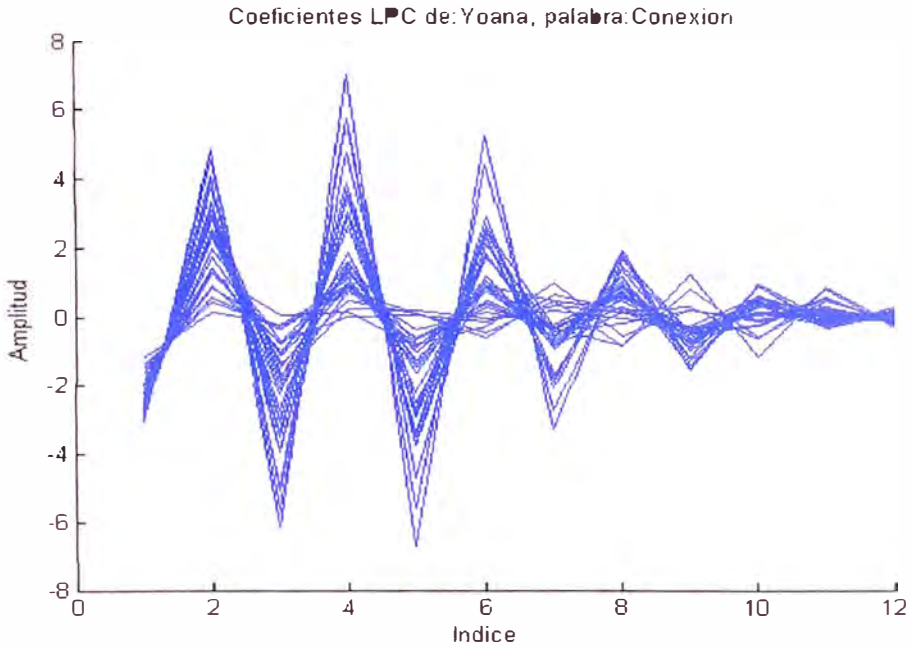

Figura 5.25.Coeficientes LPC de la 3ra persona, palabra "Conexión".

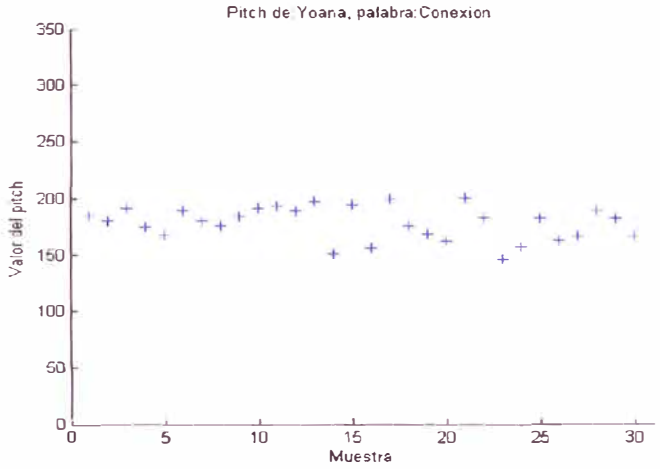

Figura 5.26. Pitch de la 3era persona, palabra "Conexión".

d. Características de la cuarta persona

En las figuras 5.27 y 5.28 se muestran los coeficientes LPC de la 4ta persona que también es una mujer.

Aquí también vemos que el patrón que le corresponde es único y es repetitivo para las diferentes muestras de voz.

Durante este muestreo se tiene más aleatoriedad en cuanto al valor del Pitch, dando un 87% de efectividad en el reconocimiento, esto debido al estado de animo de la persona a la hora de tomar las muestras.

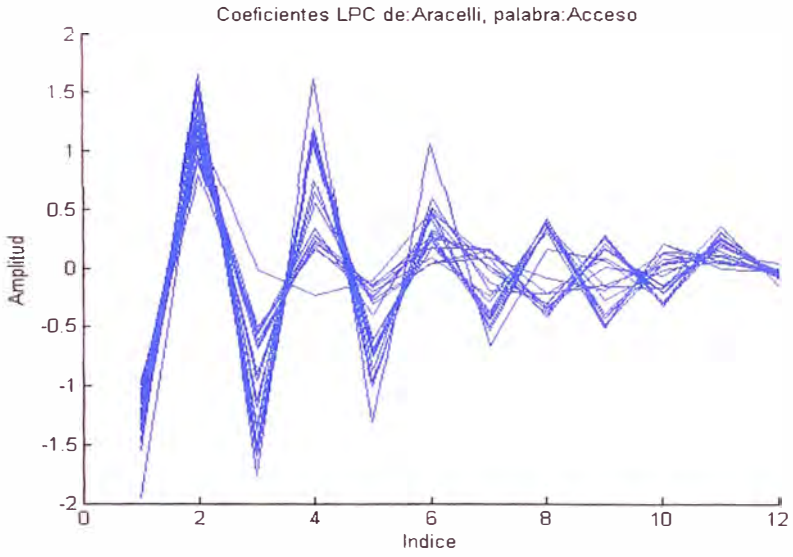

Figura 5.27. Coeficientes LPC de la 4ta persona, palabra "Acceso"

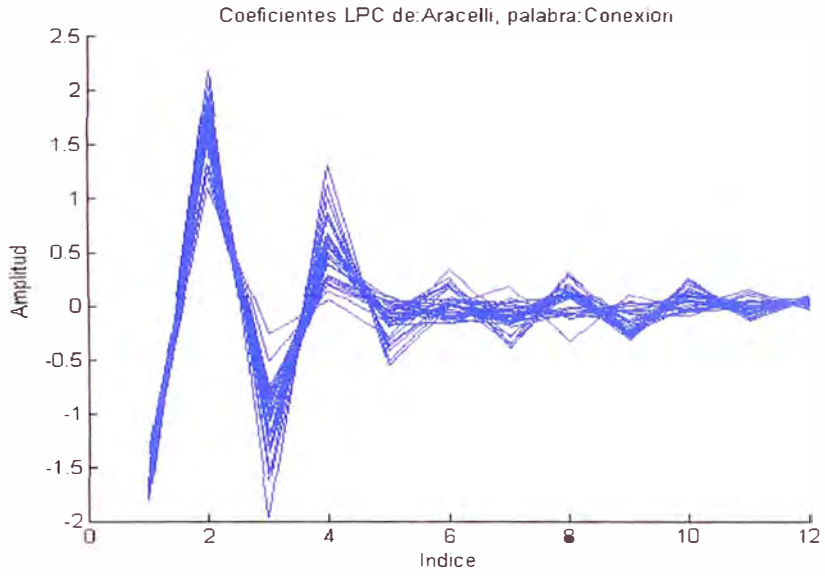

Figura 5.28. Coeficientes LPC de la 4ta persona, palabra "Conexión".

En la figura 5.29 se muestra el pitch de la 4ta persona, y como puede verse el pitch corresponde a valores mayores de 150 debido a que es una mujer.

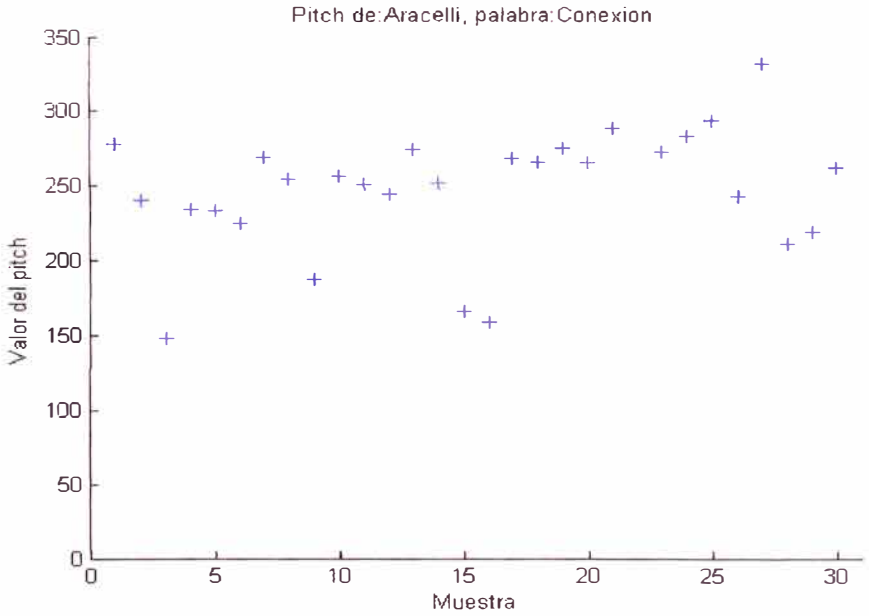

Figura 5.29. Pitch de la 4ta persona, palabra "Conexión".

e. Características de la quinta persona

Finalmente, en las figuras 5.30 y 5.31 se muestran los coeficientes LPC y el pitch de la 5ta persona en una palabra determinada.

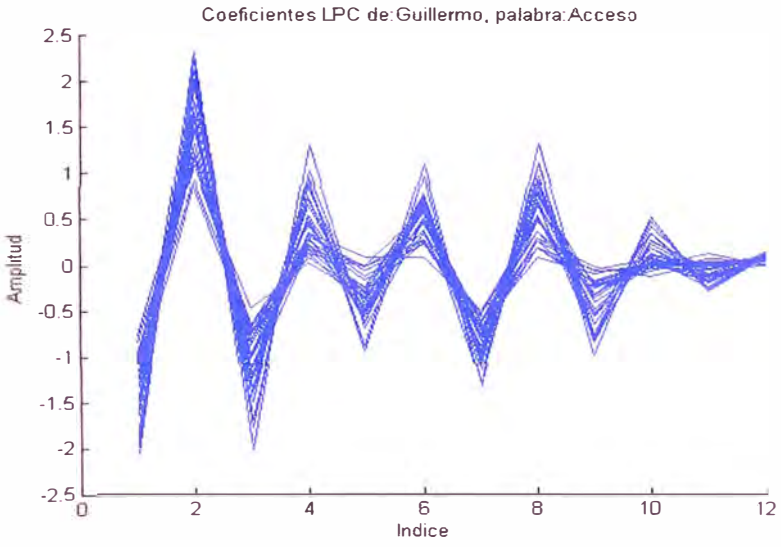

Figura 5.30. Coeficientes LPC de la 5ta persona, palabra "Acceso".

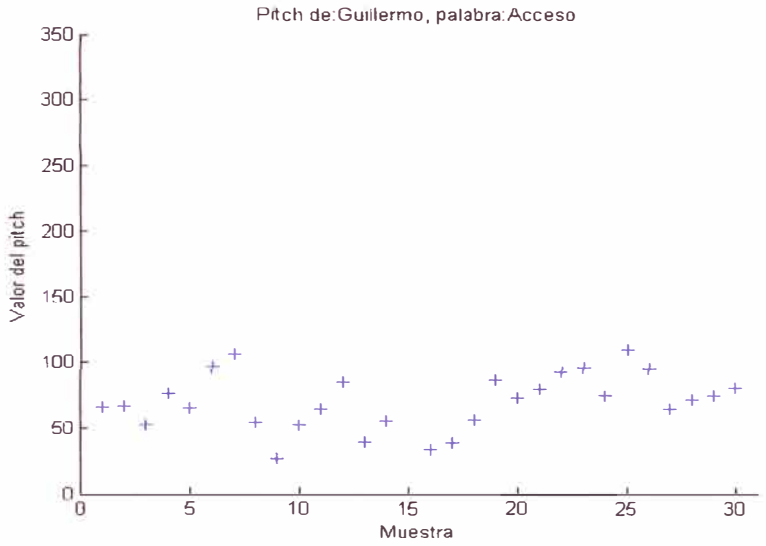

Figura 5.31. Pitch de la 5ta persona, palabra "Acceso".

5.6.5. Comprobación de la calidad del SRP[5] de voz del hablante

Entrenamiento de la red neuronal

Una vez definido el conjunto de entrenamiento queda entrenar la red neuronal. La arquitectura de la red neuronal ya fue definida en el capítulo anterior, y solo queda definir el número de unidades escondidas. Con el fin de no tener perdida de información se ha escogido un número de unidades escondidas mayor al número de entradas [5], específicamente se ha usado 15 unidades escondidas, y se va a manejar como parámetro variable el coeficiente de regularización. Como se vio en el capítulo 4 permite hallar el modelo más óptimo.

Cabe recordar que para el entrenamiento de la red se ha usado el algoritmo de optimización "quasi-newton" para la minimización de la función de error junto con

[5] Sistema de reconocimiento de patrones

el algoritmo de "back-propagation" para el cálculo del gradiente del error. Además en el entrenamiento se han usado un número máximo de 200 iteraciones.

a. Resultados usando la palabra "Conexión"

En las figuras 5.32 se muestran los resultados del entrenamiento de la red neuronal usando la palabra "conexión". En estas graficas el eje X muestra los valores que predicen la red, es decir la identidad de la persona según la red; y los valores en el eje Y muestra los valores reales de la identidad de la persona. Como se aprecia, la tasa de clasificación en el conjunto de entrenamiento es de 100%.

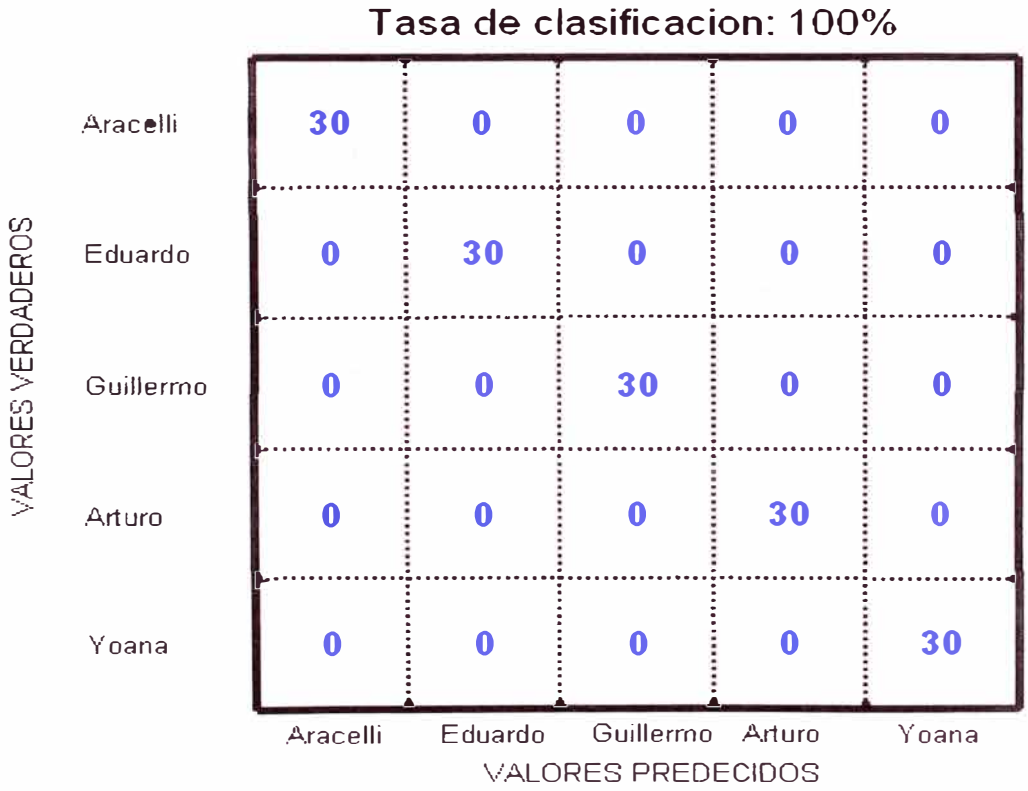

Figura 5.32. Resultados conjunto de entrenamiento - palabra "Conexión".

En figura 5.33 se muestran los resultados en el conjunto de prueba. Como se observa la tasa de reconocimiento es de 96%.

Además se puede observar que la tasa de reconocimiento entre varones es perfecto, mientras que la tasa de reconocimiento entre mujeres hay dos muestras que tienen error, ya que estas corresponden a "Yoana" pero la red las predice como "Aracelli". Sin embargo, la tasa de reconocimiento es bastante alta en el conjunto de prueba, lo que significa que el sistema ha aprendido a reconocer a una persona basándose en las medidas LPC de su voz.

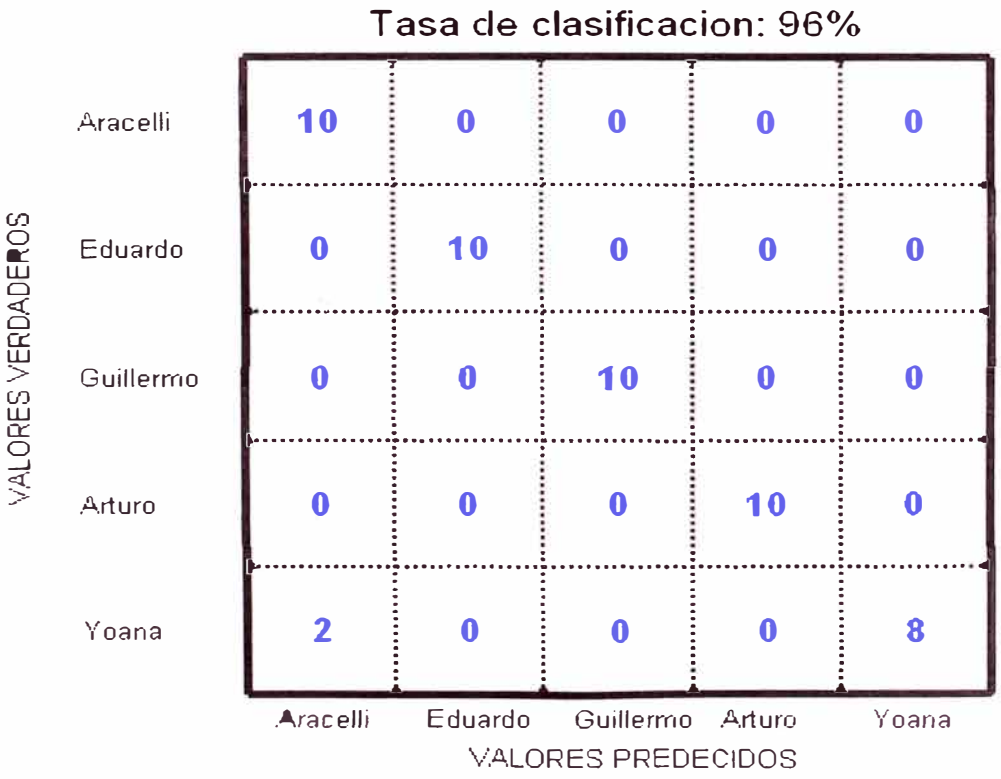

Figura 5.33. Resultados conjunto de prueba - palabra "Conexión".

b. Resultados usando la palabra "Acceso"

En la figura 5.34 se muestra los resultados del conjunto de entrenamiento usando la palabra "Acceso". Como se observa la tasa de clasificación es de 100%.

En la figura 5.35 se muestra los resultados del conjunto de prueba usando donde se observa que la tasa de clasificación es de 86%, presentándose un mayor problema de clasificación entre mujeres, pero también

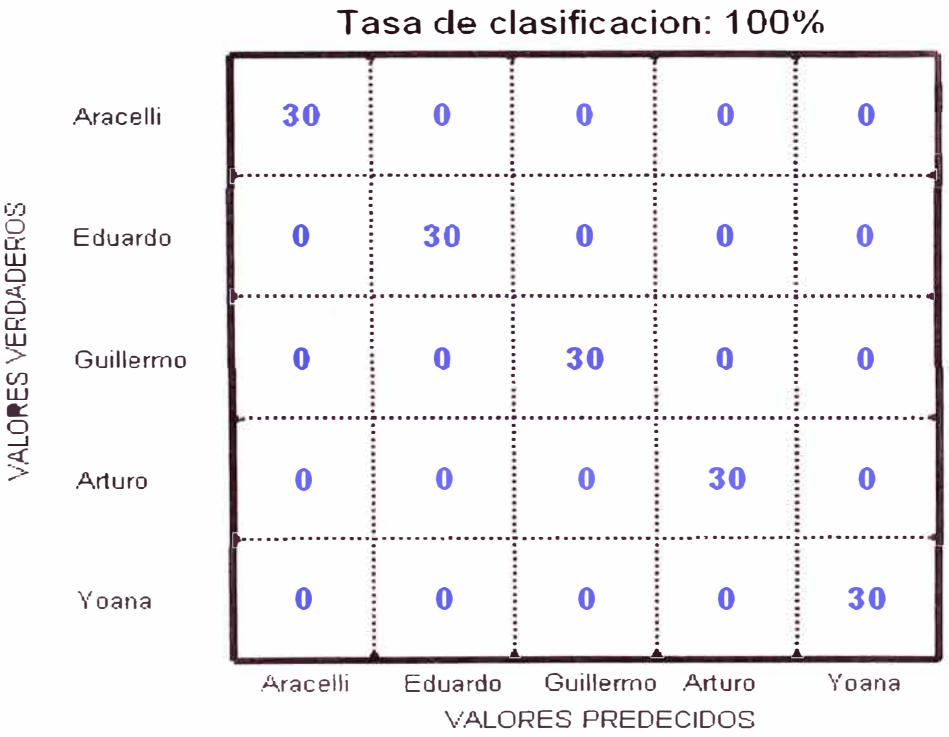

Figura 5.34.Resultados conjunto de entrenamiento - palabra "Acceso".

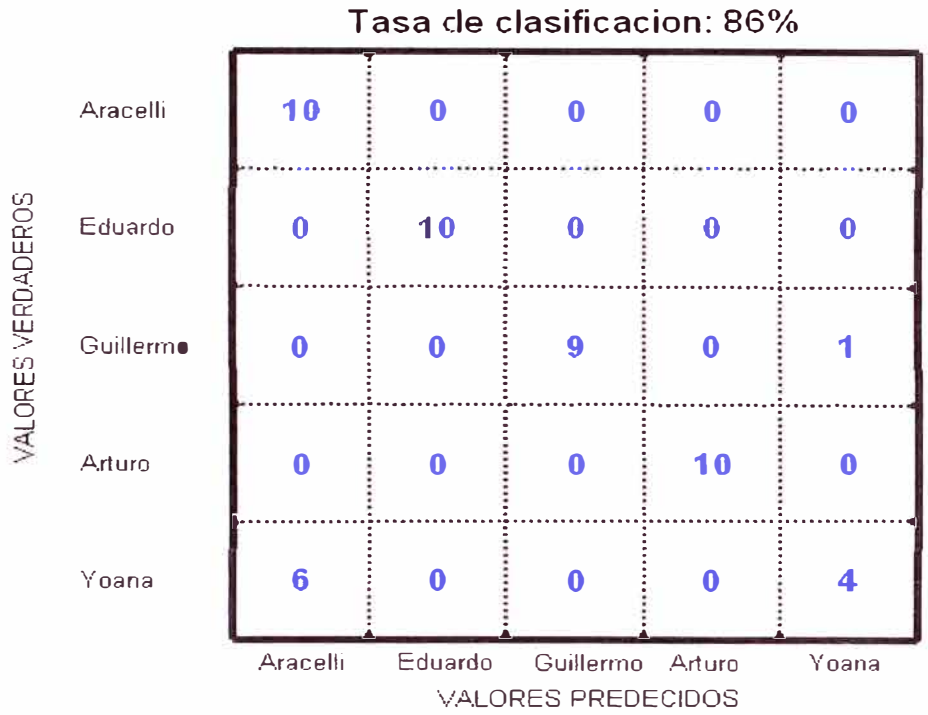

Figura 5.35. Resultados conjunto de prueba - palabra "Acceso".

c. Resultados usando la palabra "hola"

En la figura 5.36 y 5.37 se muestran los resultados de reconocimiento tanto en la etapa de entrenamiento como de prueba. Como se observa, en este caso la tasa de reconocimiento en prueba ha disminuido considerablemente, aun cuando la tasa de reconocimiento en el entrenamiento sigue siendo de 100%. Este resultado nos indica claramente que la palabra "hola" no es muy adecuada para el desarrollo de sistemas de reconocimiento, ya que no permite una buena tasa de clasificación.

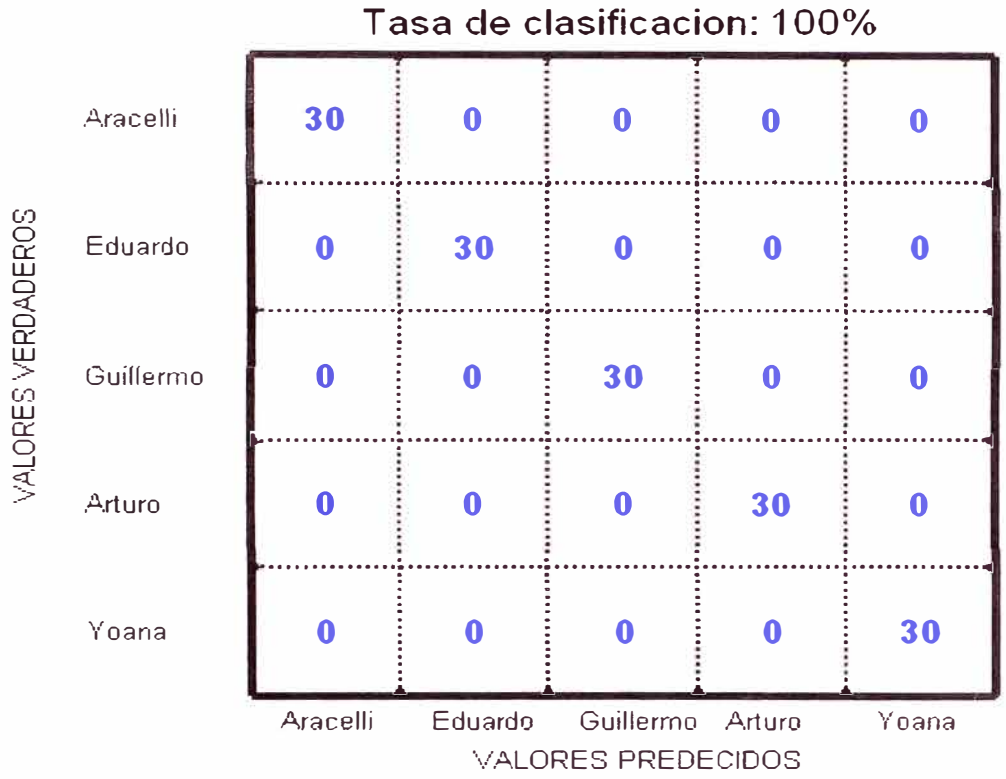

Figura 5.36. Resultados conjunto de entrenamiento - palabra "Hola".

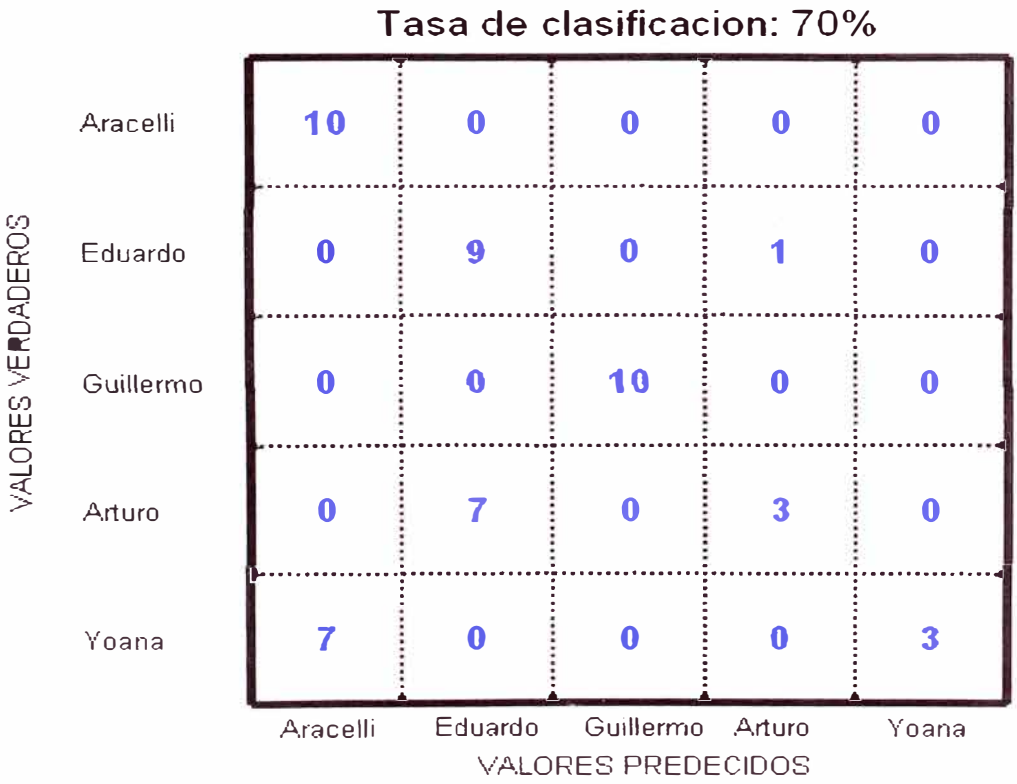

Figura 5.37. Resultados conjunto de prueba - palabra "Hola".

d. Resumen de los resultados:

A continuación se muestra una tabla resumen (Ver tabla 5.1) con los resultados de las pruebas obtenidas indicándose la cantidad de aciertos durante la predicción del sistema.

Tabla 5.1. Resumen de resultados.

PERSONA / PALABRA	PRIMERA ARTURO	SEGUNDA EDUARDO	TERCERA YOANA	CUARTA ARACELLI	QUINTA GUILLERMO	TASA DE CLASIFICACIÓN
HOLA	3	9	3	10	10	70%
ACCESO	10	10	4	10	9	86%
CONEXIÓN	10	10	8	10	10	96%
					PROMEDIO	84%

La mayor cantidad de desaciertos se da en la voz de la tercera persona, llegando el sistema a "confundir "con la cuarta persona también del mismo género. Las tasas de clasificación mejoran respecto de los antecedentes propuestos mostrando claramente un aumento de la tasa de clasificación, así como una forma alternativa de realizar el reconocimiento de los patrones de voz mediante redes perceptrón multicapa.

5.6.6. Sistema final de reconocimiento

Una vez encontrado los parámetros de la red que dan una mayor tasa de reconocimiento, la tarea de reconocimiento simplemente consiste en almacenar en memoria la red neuronal, para más adelante usarla en cualquier momento simplemente cargando los valores de los pesos y usando la propagación hacia delante

de la red neuronal. Tal como se observó en los experimentos anteriores, el uso de la palabra conexión permite una mayor tasa de clasificación por lo que será la palabra a usar en el sistema de reconocimiento en tiempo real que se muestra en la figura 5.38.

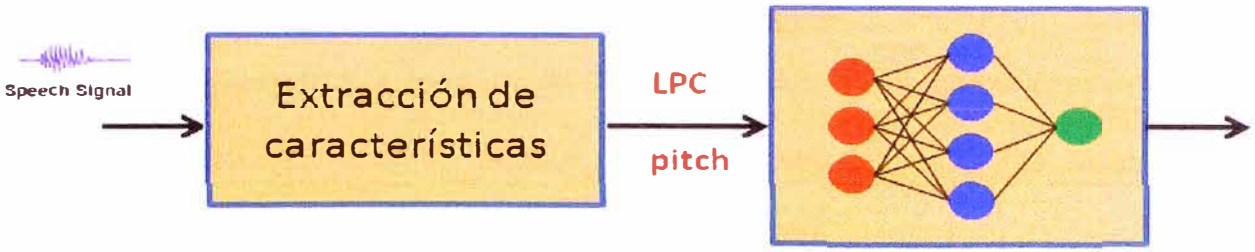

Figura 5.38. Sistema final de reconocimiento.

En la figura 5.38 se muestra la interface de usuario desarrollada para el reconocimiento de personas. Como se puede apreciar el sistema pide una señal de voz, y automáticamente muestra la identidad de la persona. Al final el programa responde con un mensaje:

SISTEMA DE RECONOCIMIENTO DEL HABLANTE USANDO REDES NEURONALES

Presione una tecla para tomar su muestra de voz.

La identidad del usuario es: Guillermo

Conclusiones y recomendaciones

Conclusiones

1. Se implementó un sistema de reconocimiento de personas basado en la voz. En particular usando la palabra "Conexión" y usando las voces de cinco personas, tres hombres y dos mujeres, se obtuvo una tasa de reconocimiento de 100% en el entrenamiento y de 96% en el conjunto de prueba. También se experimentó con las palabras "Acceso" y "Hola" pero se obtuvieron menores tasas de reconocimiento, usando las herramientas del procesamiento de señales para la extracción de características de las señales de voz y las redes neuronales para el aprendizaje del reconocimiento de patrones.

2. Se desarrolló algoritmos para el filtrado y mejoramiento de las señales. Antes de la etapa de extracción de características es muy importante el desarrollo de algoritmos para eliminar las señales que no son de voz, además se requiere hacer un filtrado para eliminar el ruido de alta frecuencia.

3. Se implementó un algoritmo que permite determinar el pitch promedio de cada muestra tomada. .Los coeficientes LPC aplicados a las señales de voz permiten extraer características únicas de las señales de voz de cada persona. Estos coeficientes son usados más adelante como entradas de la red neuronal. De igual manera, el pitch es una característica que nos permite distinguir

entre hombres y mujeres, y esta características ayuda al sistema en la tarea de reconocimiento de personas.

4. Se usó La red Perceptrón multicapa exitosamente en la tarea de reconocimiento de personas basado en la voz. Usando el algoritmo de back-propagation se puede hallar entrenar rápidamente el sistema, en especial permite el cálculo eficiente del gradiente de la función de error, el cual se requiere para el posterior entrenamiento de la red.

5. Los resultados del entrenamiento usando las palabras "Hola", "Acceso" y "Conexión" son del 100% en todos los casos; sin embargo, la tasa de reconocimiento en la prueba solo es aceptable en el caso de la palabra "Conexión", donde se tiene una tasa de 96%, ya que en las otras dos palabras la tasa de reconocimiento es muy baja.

6. Es posible implementar sistemas de reconocimiento del hablante, aplicando reconocimiento de patrones, que procesando la voz y a través de sus elementos descriptores realicen el reconocimiento del hablante, mejorando la tasa de reconocimiento que estaba en 60% y que puede llegar hasta 90%, superando los objetivos iniciales planteados en la hipótesis del presente informe.

Recomendaciones:

1. Queda como tarea evaluar el desempeño del sistema si lo entrenamos con más muestras de voz. Como se menciona en la literatura, si el conjunto de entrenamiento es mayor, la tasa de reconocimiento tiende a incrementarse. La desventaja de esto es el esfuerzo en tiempo y dinero de la toma de más datos.

2. Otro punto a evaluar a futuro es si usando otro tipo de características de voz, se puede incrementar la tasa de reconocimiento del sistema. Si se obtienen mejores características se garantiza que el desempeño del sistema se puede incrementar de manera notable.

3. El sistema de reconocimiento en tiempo real se implementó exitosamente, y queda como tarea el desarrollo de interfaces más amigables al usuario, tales como interfaces GUI. Además, se propone como tarea el desarrollo del sistema de reconocimiento en el lenguaje C o en un entorno Python.

4. Las ciencias computacionales, en la actualidad vienen desarrollando nuevas herramientas para el procesamiento y sistemas de toma de decisiones inteligentes, siendo las redes bayesianas una alternativa interesante para analizar en el campo de reconocimiento de patrones.

Bibliografía

1. J. Benesty, M. Mohan Sondi. Yiteng Huang. *Springer Handbook of Speech Processing*. Springer-Verlag Berlin Heidelberg (2008).
2. Ian MacLoughlin. *Applied Speech and Audio Processing*. Cambridge University Press (2009).
3. Makhoul, J. *Linear prediction: A tutorial review*. Proceedings of the IEEE 64, 4.
4. G. Middelton, *Pitch Detection Algorithms*. http://cnx.rice.edu/content/m11714
5. Christopher. M. Bishop. Neural Networks for Pattern Recognition. Clarendom Press Oxford (1995)
6. R.V Pawar, P. P: Kajave, S. N. Mali. *Speaker Identification using Neural Networks*. World Academy of Science, Engineering and Technology 2005.
7. M. Shaban Al-Ani, T. Sultan, K. Aljebory. *Speaker Identification: A Hybrid Approach Using Neural Networks and Wavelet Transform*. Journal of Computer Science 2007.
8. D. A. Reynolds, L. P. Heck. *Automatic Speaker Recognition, Recent Progress, Current Applications, and Future Trends*. MIT Lincon Laboratory.
9. B. J. Love, J. Vining, X. Sun, *Automatic Speaker Recognition Using Neural Networks*. Electrical and Computer Engineering Department Texas University.
10. B. Yegnanarayana, K. Sharat Reddy and S. P. Kishore. Source and System Features for Speaker Recognition Using AANN Models.

Apéndice

Código Fuente

MATLAB DEL SISTEMA DE RECONOCIMIENTO AUTOMÁTICO

```
% Final_01 : Escritura automatica de archivos de audio
% Author   : Guillermo Joo
% Descripcion: Este archivo nos permite escribir automaticamente
archivos
%              de audio.
clc; clear all, close all;


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%
%
% 1. ESTABLECEMOS PARAMETROS DE GRABACION
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%
Fs = 16000;          % Frecuencia de muestreo
Nbits = 16;          % Numero de bits
ch = 1;              % Numero d canales de entrada del microfono
                     % 1-Mono(default), 2-stereo

% Creacion de un objeto de grabacion de audio
aro = audiorecorder(Fs, Nbits, ch);



%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%
%
% 2. CONFIGURACION DE LA GRABACION
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%
N = 3;                          % Numero de muestras a grabar
Nombre  = 'Alex_';           % Nombre de la persona
Palabra = 'Conexion_';              % Palabra a grabar - hola,
acceso, conexion, entrada

 for i = 1:N
```

```
    % a. Establecemos el nombre de archivo
    if i<10
        filename = [Nombre Palabra '0' num2str(i)];       % Name of
the file
    else
        filename = [Nombre Palabra  num2str(i)];
    end


    % b. Empezamos a grabar cuando se presiona una tecla
    fprintf('Presione una tecla para tomar la %d muestra.\n', i);
    pause

    t = 2.0;              % Numero en segundos de grabacion
    record(aro, t);       % Grabamos

    %   Esperamos hasta que la grabacion termine
    while(isrecording(aro))
        continue;
    end



    % 3. Guardamos la grabacion con el nombre especificado
    speech = getaudiodata(aro, 'double');
    wavwrite(speech, Fs, Nbits, filename);

end


fprintf('FIN DE GRABACION DE MUESTRAS\n');
```

```
% Final_02 : ELIMINACION DE RUIDO Y GRABACION DE LA SEÑAL
% Author    : Guillermo Joo
% Descripcion: Este archivo nos permite eliminar señales que no son
%              de voz automaticamente para luego guardarlo en
archivos
%              de audio.
clc; clear all, close all;
```

```
%
% 1. PARAMETROS INICIALES
%
```

```
% 1.1. Numero de archivos a
limpiar N = 20;
```

```
% 1.2. Archivos a limpiar
Nombre  = 'Veronica_';            % Nombre de la persona
Palabra = 'Hola_';                % Palabra Hola, acceso, conexion,
entrada
```

```
% 1.3. Parametros de limpieza
t = 0.02;                   % (seg.) Longitud de cada
pmin = 1e-4;                segmento % Potencia minima
```

```
%
% 2. LECTURA Y LIMPIEZA DE ARCHIVOS
%
```

```
for i = 1:N

    % a. Establecemos el nombre de archivo de audio a
    limpiar if i<10
        filename = [Nombre Palabra '0' num2str(i)];
    else
        filename = [Nombre Palabra  num2str(i)];
    end


    % b. Leemos el archivo de audio
```

```
    [y, Fs, Nbits] = wavread(filename);



    % c. Limpiamos el ruido
    [signal signal_cut] = noise_removal(y, Fs, t, pmin);



    % d. Guardamos la grabacion con el nombre especificado
    clean_filename = [ filename '_clean'];       % Nombre del archivo
    wavwrite(signal_cut, Fs, Nbits, clean_filename);


    % e. Mensajes
    fprintf('Se ha limpiado el %d archivo de audio y se ha
guardado', i)
    fprintf('el archivo limpio.\n')
    fprintf('Presione una tecla para continar.\n');
    pause

end
```

```
% SISTEMA DE RECONOCIMIENTO DE PERSONAS EN TIEMPO REAL USANDO REDES
% NEURONALES PERCEPTRON MULTICAPA
clc; clear all; close all

% CARGAMOS LA RED NEURONAL ENTRENADA
load red.mat


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%
% 1. OBTENEMOS EL VECTOR DE AUDIO
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%
% 1.1. Establecemos parametros de grabacion
Fs = 16000;            % Frecuencia de muestreo
Nbits = 16;            % Numero de bits
ch = 1;                % Numero d canales de entrada del microfono
                       % 1-Mono(default), 2-stereo


% 1.2. Creacion de un objeto de grabacion de audio
aro = audiorecorder(Fs, Nbits, ch);


% 1.3. Empezamos a grabar cuando se presiona una tecla
disp('SISTEMA DE RECONOCIMIENTO DEL HABLANTE USANDO REDES
NEURONALES')
disp('--------------------------------------------------------------
-')
disp(' ')
fprintf('Presione una tecla para tomar su muestra de voz.\n');
pause

t = 2.0;               % Numero en segundos de grabacion
record(aro, t);        % Grabamos
%   Esperamos hasta que la grabacion termine
while(isrecording(aro))
    continue;
end


% 1.4. Obtenemos el vector de datos
y = getaudiodata(aro, 'double');




%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%
% 2. PROCESAMOS LA SEÑAL
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%
% 2.1. Limpiamos el ruido
%
```

```
%   - Parametros de limpieza
t = 0.02;                    % (seg.) Longitud de cada segmento
pmin = 1e-4;                 % Potencia minima

%   - Llamamos a la funcion
[~, signal_cut] = noise_removal(y, Fs, t, pmin);



% 2.2. Hallamos los coeficientes LPC de la señal limpia
P = 12;
a = lpc(signal_cut, P);
if(0)
    plot(a, '-b')
    hold on
end


% 2.3. Hallamos el pitch promedio de la señal limpia
t = 0.03;
pitch = pitch_promedio(signal_cut, t, Fs);


% 2.4. Guardamos el patron de caracteristicas
X = a(2:end);

% Consideramos el pitch
if(net.nin == 13)
    X = [a(2:end)  normal(pitch)];
end




%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%
% 3. EVALUAMOS LA SALIDA DE LA RED NEURONAL
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%
% 3.1. Propagamos la data por la red entrenada
YY = mlpfwd(net, X);

% 3.2. Hallar la identidad de la persona
[~, index] = max(YY);
switch index
    case 1
        disp('La identidad del usuario es: Aracelli')
    case 2
        disp('La identidad del usuario es: Arturo')
    case 3
        disp('La identidad del usuario es: Diana')
    case 4
        disp('La identidad del usuario es: Guillermo')
    case 5
```

```
        disp('La identidad del usuario es: Marcela')
    case 6
        disp('La identidad del usuario es: Menphis')
end
```

```
function pitch = hps(x,Fs)
% HPS Calculo del pitch de una señal de voz
%
% Inputs:
%     x: Señal de audio
%   - Fs: Frecuencia de muestreo
%
% Outputs:
%   - pitch: Valor del pitch


% 1. PARAMETROS INICIALES
fx = log(abs(fft(x, 2048)));
len = length(fx);


% 2. COMPRESION ARMONICA
fx1 = fx(1:2:len);
fx2 = fx(1:3:len);
fx3 = fx(1:4:len);
fx4 = fx(1:5:len);


% 3. HALLAMOS EL INDICE DE MAYOR PICO
len4 = length(fx4);
Px = 2*(fx(1:len4) + fx1(1:len4) + fx2(1:len4) + fx3(1:len4) +
fx4(1:len4));
[~,I]=max(exp(Px(1:len4)));


% 4. HALLAMOS LA FRECUENCIA FUNDAMENTAL (PITCH)
F = 0:Fs/(len):Fs;
pitch = F(I);


end
```

```
function [a i] = lpc_promedio(x, t, Fs, P)


% Calculamos el numero de puntos correspondiente al intervalo "t"
N = Fs*t;

% Maximo valor de numero de puntos
Nt = length(x);

% Array para guardar los coeficientes a
a = zeros(1, P+1);

i = 0;
m = 1;
n = N;

while(n < Nt)

    % Obtener los segmentos
    seg = x(m:n);
    seg = seg.*hamming(N);

    % Hallamos los coeficientes
    a = a + lpc(seg, P);

    % Actualizar "i" y los valores del segmento
    i = i+1;
    m = i*N + 1;
    n = (i+1)*N;
end

%y1_seg = y1(1:N);
%y2_seg = y1(N+1:2*N);
%y3_seg = y1(2*N+1:3*N);


a = 1/i*a;



end
```