UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN INGENIERÍA
ELÉCTRICA– INSTRUMENTACIÓN

INVESTIGACIÓN Y DESARROLLO DEL MÉTODO DE RECONOCIMIENTO DE VOZ USANDO CLASIFICADORES NEURONALES

TESIS
QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN INGENIERÍA

PRESENTA:
García Fragoso Nestor Abdy

TUTOR PRINCIPAL
Dra. Baydyk Mykolaivna Tetyana CCADET UNAM

CIUDAD DE MÉXICO, Junio 2017

Agradecimientos

La vida es la que se encarga de hacerme madurar.

A lo largo de mi vida me ha tocado convivir con mucha gente. Familia, amigos conocidos, las experiencias que he vivido cerca de ellos me han servido para aprender y guiarme en mi vida.

Le doy gracias especialmente a mi mamá por que ha sido el pilar de mi vida y sin ella no hubiera podido lograr cada una de mis metas. A mi abuela, mis tías y primos, porque siempre que necesite apoyo sé que puedo contar con su ayuda.

Mis amigos han sido una parte importante de mi vida y crecimiento, puedo decir que me han guiado y ayudado a tomar buenas decisiones. Siempre que necesito consejo. En estos últimos años me he encontrado con nuevos amigos que se volvieron mis mejores amigos; y descubrí que los viejos amigos siguen siendo grandes compañeros.

A mis profesores que me ponen retos para seguir aprendiendo, me ayudaron a superar mis límites y a descubrir que nada es imposible.

Gracias a l universidad Nacional Autónoma de México por haberme permitido continuar con mis estudios y al grupo de investigadores del CCADET, que creyeron en mí y me dieron la oportunidad de continuar con mi crecimiento académico en su programa de posgrado.

Y muchas gracias a mis tutores, que confiaron en mí. La Dra. Tetyana Baydyk, y el Dr- Ernst Kussul, que fueron muy generosos y pacientes. Ya que sin su ayuda este trabajo no habría sido posible.

Y finalmente, pero no menos importante, a mis mascotas que fueron la inspiración necesaria para continuar trabajando.

Este trabajo forma parte del proyecto PAPIIT IT 100817.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

CAPÍTULO 1

INTRODUCCIÓN

1.1 Descripción general

El proyecto de la tesis consiste en emplear técnicas de inteligencia artificial para el reconocimiento de hablantes. De las tareas de reconocimiento de hablantes, se pueden diferenciar dos tipos: la primera es la identificación del hablante, responde a la pregunta, ¿Quién habla?; la segunda tarea consiste en verificar al hablante, responde a la pregunta ¿Eres quién dices ser? El tema principal de la tesis es desarrollar una metodología que permita identificar al hablante, esto se hace comparando una determinada señal de voz con un conjunto de registros de voz para identificar a cuál hablante pertenece.

Para cumplir con el trabajo se han propuesto dos tareas principales, la primera es construir una base de datos de voz de diferentes hablantes, y la segunda tarea es implementar un clasificador que permita identificar a los hablantes de la base datos.

Se espera obtener como resultados una base de voz para su uso en trabajos posteriores y el sistema de reconocimiento de voz basado en redes neuronales.

El presente trabajo forma parte del proyecto PAPIIT IT 100817 a cargo del laboratorio de Computación Neuronal del CCADET y de la Dra. Tetyana Baydyk y al Dr. Ernst Kussul.

1.2 Objetivos

Desarrollar un clasificador neuronal artificial, adaptado del clasificador LIRA grayscale, que permite hacer el reconocimiento de voz para un conjunto determinado de hablantes; el sistema emplea espectrogramas de las señales de voz como entrada del sistema, extrae las características e identifica al hablante.

Diseñar y crear una base de datos de voz de diferentes hablantes. La base de datos no está limitada a la adquisición de señales de voz de diferentes personas; además debe cumplir con diferentes requisitos de adquisición y conversión para poder usarlos en el sistema, a continuación, sus características:

- Todas las señales de voz deben cumplir con las mismas condiciones de adquisición.
- Todas deben estar tener la misma codificación. Esto es, archivos wav.
- Todas las grabaciones deben tener un contenido fonético representativo del lenguaje.
- Las frases deben ser las mismas y ser dichas por cada hablante.
- El ancho de banda de las grabaciones también debe ser estar unificado.

- Se emplearán los mismos filtros y la misma cantidad de filtros para obtener los espectrogramas por cada señal de voz.

Objetivos secundarios son los siguientes:

- Investigar diferentes metodologías empleadas en el reconocimiento de voz.
- Desarrollar y emplear filtros resonantes obtención de características frecuenciales de las señales de voz. En esta etapa se hace la conversión de una señal de audio a una imagen o espectrograma.
- Crear diferentes programas de apoyo para la administración, manipulación y pre procesamiento de los datos.

1.3 Problema a resolver y alcance

El principal problema es establecer un nuevo método de reconocimiento de hablantes mediante su espectrograma de voz y el uso de clasificadores neuronales.

El presente trabajo tiene planeado cubrir las siguientes tareas, que además incluyen la adquisición de datos, transformación de datos, extracción de características, procesamiento y obtención de resultados.

1. Investigar diferentes metodologías empleadas en el reconocimiento de voz.
2. Adquisición de muestras de voz, de diferentes hablantes, para la creación de una base de datos de voz humana.
3. Desarrollar y emplear filtros resonantes de audio para la extracción de características de las señales de voz. En esta etapa se hace la conversión de una señal de audio a una imagen o espectrograma.
4. Crear diferentes programas de apoyo para la administración, manipulación y pre procesamiento de los datos.
5. Desarrollar un clasificador neuronal que pueda extraer las características de voz, por cada hablante y clasificarlos.
6. Evaluar la eficiencia del clasificador de acuerdo a sus resultados.

1.4 Metodología

El diagrama de la Fig. 1.1, muestra el procedimiento empleado para la realización del trabajo.

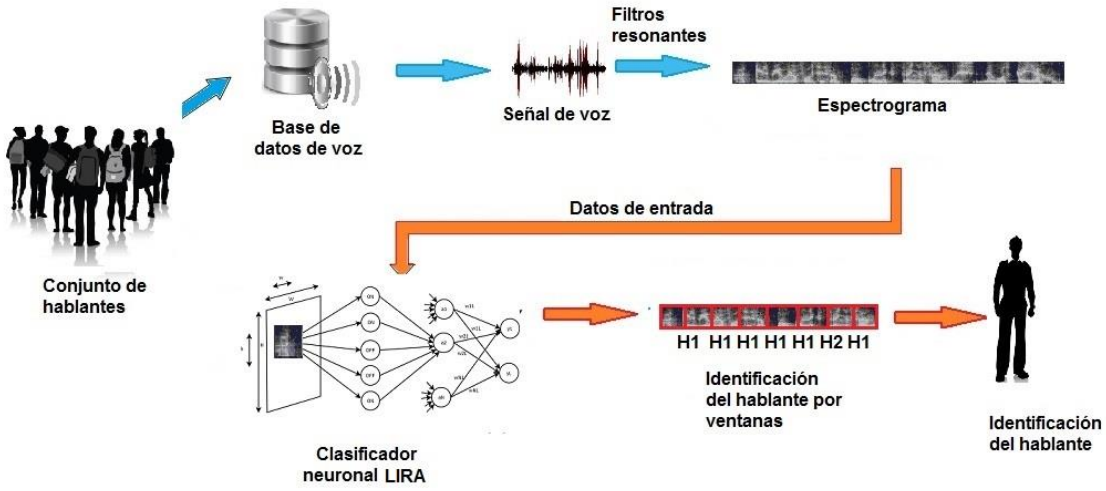


Fig. 1.1 Metodología empleada para el reconocimiento de voz.

En primer lugar, de un conjunto de personas se realizan registros de su voz. La obtención de muestras de la voz se hace con grabaciones de audio digital, todas tomadas obtenidas bajo las mismas condiciones de grabación, esto es, misma codificación, un ancho de banda unificado, condiciones de un ambiente silencioso para evitar otras fuentes de sonido. De esta manera se estructura la base de datos de la voz. Además a cada hablante se le asigna una clave de identificación única, y se registra su sexo.

Para las grabaciones se usaron 15 frases distintas, dichas cada una por cada hablante; estas frases emplean todo el conjunto fonético del lenguaje español e inglés. El objetivo de usar este tipo de declaraciones es el de tener la mayor información posible de los cambios fonéticos de la voz por cada persona.

Una vez obtenidas las señales de audio digital de voces por cada hablante, se diseñaron 64 filtros resonantes para la extracción de frecuencias naturales de la voz humana. Y usando las frecuencias seleccionadas por los filtros se construye el espectrograma de voz.

El espectrograma de cada frase, es transformado en una imagen de 64 x 1000 pixeles. La cual servirá como entrada para el clasificador neuronal LIRA.

El clasificador neuronal LIRA grayscale, siglas de Limited Receptive Area Grayscale, es una red neuronal especializada en la clasificación de imágenes [11].

El sistema esta implementado en códigos de Matlab, el cual por ser un algoritmo de carácter numérico matricial es una herramienta muy potente para la solución de este tipo de tareas.

El clasificador neuronal consta de 4 capas: una capa de entrada (S), una capa intermedia (I), una capa de asociación (A) y una capa de salida (R). La capa de entrada recibe pixeles de una ventana seleccionada al azar de la imagen original, los valores de los pixeles pueden están dentro del rango [0 255]. La capa S se conecta a la capa A mediante la capa I. por medio de un proceso de activación de neuronas.

La capa I está formada por neuronas ON y neuronas OFF, cada una de estas neuronas tiene asociado un umbral de excitación que determinara la activación de una neurona de la capa A. Para el caso de las neuronas ON, el valor del pixel debe ser mayor al umbral de excitación para activar la neurona; para el caso de las neuronas OFF el valor de pixel debe ser menor para excitar la neurona. Las neuronas de la capa A, solo se activarán si todas las neuronas ON y las neuronas OFF de la capa I de una ventana están activadas.

Finalmente, todas las neuronas de la capa A están conectadas con todas las neuronas de la capa R, y tienen asociadas un peso. Para cada neurona de la capa R, que también corresponde con una clase, se suman los pesos de las neuronas activas de la capa A, y se selecciona la de mayor valor para seleccionar la clase ganadora.

Antes de probar el clasificador, este debe ser entrenado. Para el proceso de entrenamiento se usa la regla de Hebb.

Las especificaciones del clasificador neuronal LIRA implementado son:

- Neuronas de la capa A: 64000.
- Tamaño de las ventanas: 20 x 20 pixeles.
- Neuronas ON: 3.
- Neuronas OFF: 2.
- Tamaño de los espectrogramas: 64 x 1000 pixeles.
- Cada espectrograma se segmenta en imágenes de 64 x 100 pixeles con un solapamiento del 50%.
- 15 espectrogramas por hablante. 12 espectrogramas para entrenamiento y 3 para el reconocimiento.

- ⇨ Además se cuenta con el apoyo y equipos del laboratorio de Computación neuronal del CCADET a cargo del Dr. Ernst Kussul.

1.5 Descripción del contenido

El presente trabajo está dividido en 8 capítulos, en los cuales se describen los principios, conceptos y métodos empleados para la resolución del problema de reconocimiento del hablante. A continuación, se describe de forma resumida el contenido de cada uno de los capítulos.

En el capítulo 2, se describen los antecedentes de la computación hasta su empleo como herramienta para el reconocimiento de la voz. Se definen conceptos relacionados con la identificación de los hablantes y de las metodologías existentes empleadas para realizar esta tarea.

En el capítulo 3, se definen conceptos básicos de acústica para el tratamiento de señales. Se especifica el procedimiento de generación de la voz en el ser humano y sus características que permiten identificar a su portador. Finalmente se menciona el procedimiento para convertir una señal analógica a digital.

En el capítulo 4, se definen conceptos de bases de datos. Se hace la descripción detallada de la creación de una base de datos digital de registros de voz, se menciona el proceso de obtención de los registros de voz de diferentes hablantes, las frases empleadas. Además, en este capítulo se describe el proceso de filtración para la obtención de los espectrogramas usados como discriminantes en la tarea de reconocimiento del hablante por su voz. Finalmente se hace una revisión completa como está conformada la base de datos, incluyendo los registros en formato de audio digital y su correspondiente espectrograma, así como información básica de cada hablante.

En el capítulo 5 se describe la red neuronal empelada para la tarea de reconocimiento de hablantes. En esta descripción se especifican las diferentes capas de la red neuronal y la función de cada una de ellas. La forma en que se obtienen las características de cada espectrograma, los cálculos empleados, el método de entrenamiento y la forma de obtención de los resultados.

En el capítulo 6 se detallan las diferentes pruebas realizadas durante la investigación y los resultados obtenidos.

Y finalmente en el último capítulo se señalan las conclusiones obtenidas,

CAPÍTULO 2

ANTECEDENTES

En este capítulo se hace una breve reseña sobre el reconocimiento de patrones en voz. Iniciando desde luego con el uso del cálculo computacional hasta llegar a la inteligencia artificial como herramienta para la simulación de la resolución de tareas empleando las computadoras u otros dispositivos electrónicos.

De manera más específica se menciona la evolución de las redes neuronales como algoritmo para solucionar problemas empleando computadoras simulando como lo haría un cerebro humano. Y finalmente se describen algunos trabajos sobre reconocimiento de la voz humana empleando diferentes enfoques.

2.1 Historia del cálculo computacional

En las antiguas civilizaciones del periodo clásico surgen dos detonadores que son el punto clave para el desarrollo de la computación actual: la sistematización del razonamiento y el desarrollo de métodos de cálculo [1].

El razonamiento sistematizado tiene su origen en la antigua Grecia durante el periodo clásico (600-300 AC), los filósofos dieron origen a las matemáticas formales. Principalmente tres filósofos que aportaron conocimiento favorable para las matemáticas formales; Platón (427-347 AC) presentó la abstracción de las ideas, Aristóteles (384-332 AC) desarrolló el razonamiento deductivo y sistematizado, y Euclides (325-265 AC) fue quien desplegó el método axiomático diferenciando entre principios (definiciones, axiomas y postulados) y teoremas [2].

El término algoritmo fue empleado por primera vez en el año 825, por los árabes. La palabra se deriva del nombre del autor de un famoso libro persa, Abu Ja'far Mohammed ibn Musa al-Khowarizmi [3].

El primer mecanismo de almacenamiento de datos en forma de tarjetas perforadas fue creado en 1801, por Francia, Joseph Marie Jacquard. Él diseño un telar, con la singularidad de que este podría leer patrones de tarjetas perforadas y copiarlas a las telas, así sólo debía elegir la tarjeta y la máquina se encargaba tejer el telar. Estas tarjetas perforadas son consideradas como el primer mecanismo de almacenamiento de datos.

El padre de la computación fue Charles Babbage, fundador de la Royal Astronomical Society de Inglaterra. Él diseño una máquina de propósito general sin que estuviera restringida a una tarea específica. Su colega Ada Byron fue la primera en desarrollar programas para esta máquina y por lo tanto es la primera programadora de la historia [2].

Georges Boole (1854) tuvo una contribución destacada en su libro "Una investigación sobre las leyes de la Verdad", en este trabajo establece:

> El proceso del razonamiento mediante una representación simbólica. Para ello utilizó variables que solo podían adoptar dos valores “1” (verdadero) y “0” (falso) …; simbolizó el operador lógico “OR” con el símbolo “±” y el símbolo “AND” con el ”*”; y realizó un estudio en profundidad del algebra de las expresiones que sólo contenían este tipo de variables.

Durante la Segunda Guerra Mundial, en Londres crearon la primera computadora diseñada para decodificar los mensajes de radio cifrados de los alemanes, este proyecto estuvo a cargo de Alan Turing y la computadora se llamaba Colossus.

En 1939 surge la ENIAC (Electronic Numerical Integrator and Calculator) proyecto organizado por Estados Unidos dirigido por John Atanasoff y Clifford Berry. Su sucesor la EDVAC (Electronic Discrete Variable Automatic Calculator) incorporó métodos matemáticos de John Von Neumann que permitía corregir algunas deficiencias de la ENIAC, y hoy en día es conocida como la arquitectura Von Neuman.

2.2 Sistemas de reconocimiento

Un patrón es un objeto que ya ha sido clasificado y puede usarse para clasificar otros objetos similares. El reconocimiento de patrones por lo tanto es clasificar los objetos similares de acuerdo a sus características.

De manera más general: un patrón es un conjunto de características que describen a una entidad u objeto a través de un conjunto de características que toman la forma de variables [4]

El proceso de clasificar y comparar las características de un objeto para obtener un patrón es relativamente sencillo para un ser humano. Cualquiera que sea el tipo de información, ya sean imágenes, texto, sonidos, olores o personas; el proceso cognitivo sigue siendo el mismo, todo comienza desde la percepción sensorial.

Desde un punto de vista biológico el proceso inicia con percepción de un estímulo sensorial del ambiente, después de eso el estímulo debe repetirse varias veces para que se pueda convertir en un patrón; al percibir un estímulo nuevo en un contexto diferente este se convertirá en un detonador de la memoria que permitirá asociar el nuevo estímulo con los anteriores y finalmente permitirá establecer una correspondencia entre ambas sensaciones.

El reconocimiento de patrones mediante el uso de computadoras es muy diferente a la parte biológica; sin embargo, con el uso de sensores y convertidores

analógico-digitales una computadora es capaz de observar el ambiente continuamente, cuantificar las características físicas de

Un sistema de reconocimiento consta generalmente de 3 etapas [4]:

1. Adquisición de datos
2. Extracción de características o parametrización
3. Clasificación

2.3 Reconocimiento de voz

El proceso de tratamiento de la voz, comprende un amplio rango de funciones y tareas en aplicaciones. La Fig. 2.1, describe algunas aplicaciones en las que interviene la voz, y establece el lugar que tiene el reconocimiento del hablante en este proceso. El proceso es una adaptación de [5].

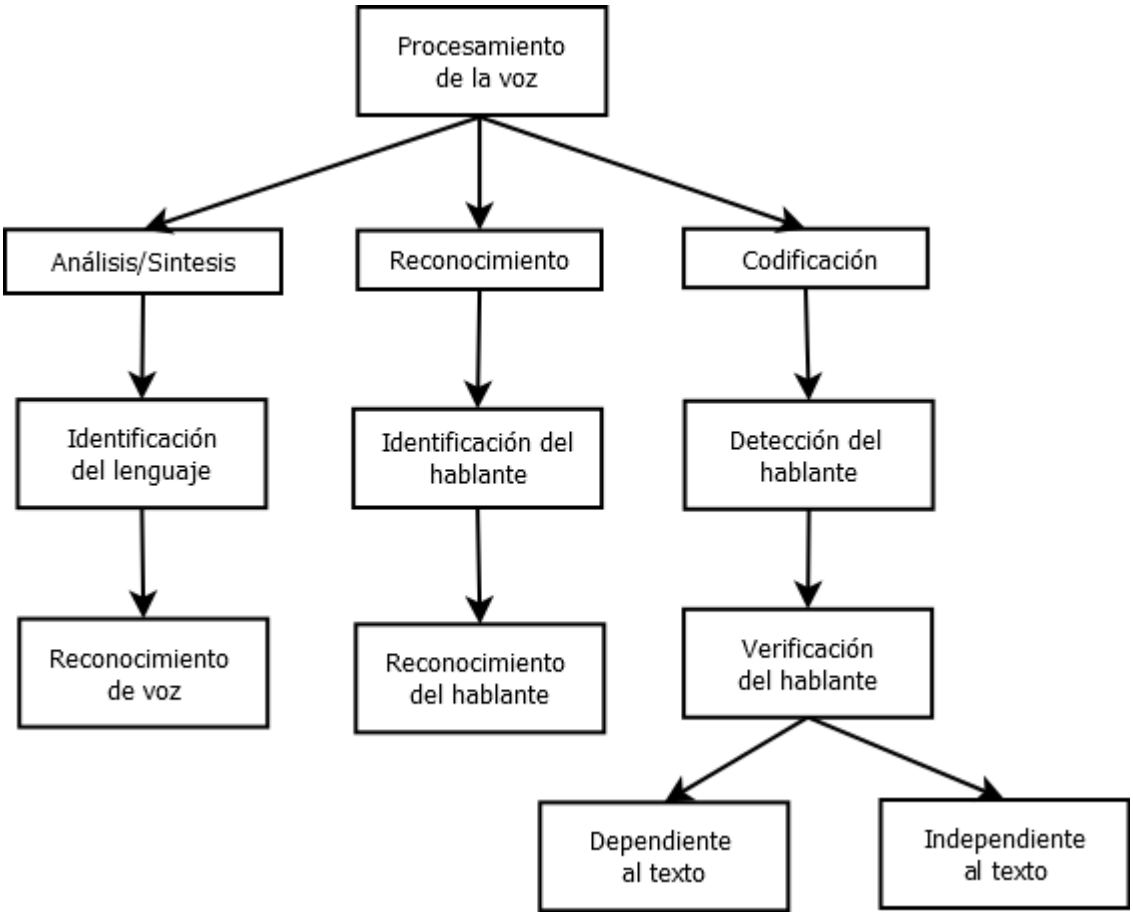


Fig. 2.1 Procesamiento de la voz

La voz es una señal compleja, es el resultado de varias transformaciones ocurridas por diferentes niveles: semántica, lingüística, articulatoria y acústica. Las

diferencias en estas transformaciones aparecen como diferencias en las propiedades acústicas de la señal de voz. Las diferencias relacionadas con el hablante son el resultado de la combinación anatómica inherente al tracto vocal y a los hábitos aprendidos del lenguaje. En el reconocimiento del hablante, todas estas diferencias pueden emplearse para discriminar hablantes [5].

La tarea de reconocimiento de hablantes es un proceso muy simple para los humanos, y la dificultad es poder emular este procedimiento en una computadora. Desde hace casi un siglo, se ha estudiado la capacidad de los humanos para poder reconocer voces [6]. Esta tarea se volvió aún más ambiciosa tras el desarrollo de las computadoras digitales y el procesamiento paralelo, el objetivo del reconocimiento del hablante cambio a un reconocimiento automático del hablante, de manera más similar a como lo hacen las personas [7].

Los sistemas de reconocimiento automático de hablantes (ARS, Automatic Speaker Recognition son usados para verificar e identificar a una persona, Fig. 2.2; estos permiten un control automatizado de servicios de voz, por ejemplo transacciones bancarias o el control de información confidencial.

Los ARS, están divididos en dos clases dependiendo su función: Identificación Automática del hablante (ASI, Automatic Speaker Identificaction) y Verificación Automática del hablante (ASV, Automatic Speaker Verification). Los sistemas ASI responden a la pregunta ¿quién eres?, mientras que los sistemas ASV responden a la pregunta ¿eres quién dices ser? [5].

Fig. 2.2. Tareas en el reconocimiento de hablantes. Identificación y verificación.

La verificación requiere distinguir la voz de un hablante de una base de datos de un grupo de voces desconocidas en el sistema. Los hablantes conocidos que reclaman su identidad a través de su voz son conocidos como “reclamantes”, el resto de los hablantes conocidos o desconocidos por el sistema son llamados

“impostores”. Esto genera dos tipos de errores durante la verificación: falsos aceptantes, el sistema toma a un impostor como un reclamante; y el segundo, falsos rechazos, el sistema toma como impostor a un reclamante. De esta manera la verificación es la parte más importante del reconocimiento de hablantes

Los modernos sistemas ASI y ASV consisten en 6 componentes clave: filtrado y convertidor A/D, removedor de silencios, interfaz de procesamiento, empate de patrones, decisión lógica y matriculado.

La sección de filtrado y el convertidor A/D son los responsables de capturar la voz del usuario, en la Fig. 2.3 se muestra el esquema de este proceso. El silencio es removido de la voz y esta se convierte en una serie representativa de características espectrales que realzan las propiedades específicas presentes en la voz. Usando estas características, la sección de empate, las relaciona con los modelos almacenados y calcula una probabilidad de distorsión para cada modelo. Una vez obtenidos estos resultados el sistema toma una decisión para validar o verificar al hablante, matriculándolo con un identificador. Sin embargo estos sistemas primero deben ser entrenados [5].

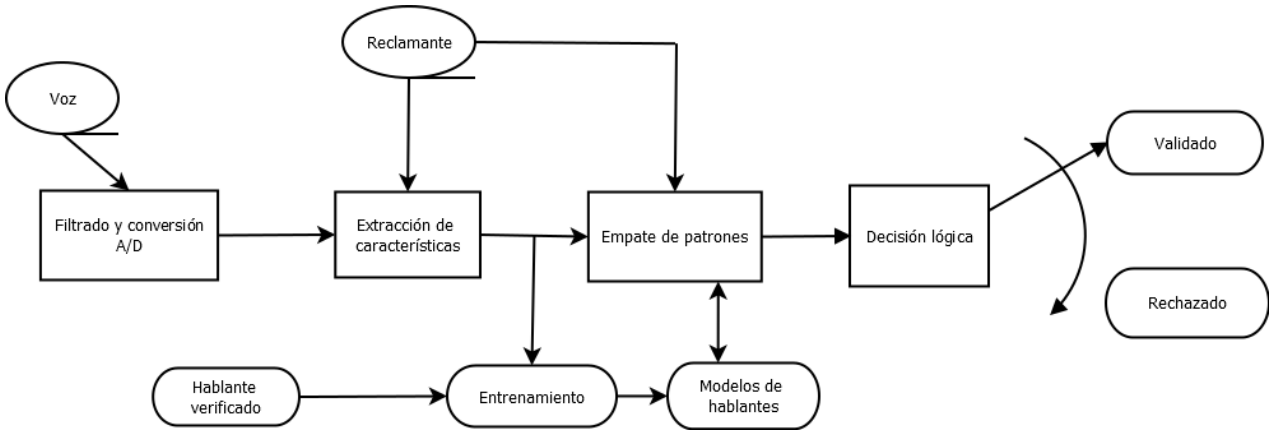


Fig. 2.3. Sistema de verificación.

CAPÍTULO 3

DEL AUDIO ANALÓGICO AL AUDIO DIGITAL

Todos los algoritmos están conformados por entradas, que son procesadas por un mecanismo principal, y salidas que son los resultados del proceso. El tipo de entradas determinan la complejidad del sistema, y son la parte clave del algoritmo. Las entradas pueden ser desde datos simples hasta, salidas de otros subprocesos, archivos o señales.

En este capítulo se intenta presentar las características que definen al audio, las características de la voz humana y el mecanismo que hace posible la identificación de entre personas. Se hace una reseña del proceso de transformación de una señal de audio analógica a una señal digital. Y finalmente, en la última parte se muestran las diferentes herramientas que existen para analizar las señales de audio digital.

Un archivo de audio digital es una estructura de datos que permite almacenar información acústica del mundo exterior o bien generada por un sintetizador de sonidos. La cantidad y calidad de la información contenida en estos archivos depende de los estándares de codificación usadas para su creación, estos son la frecuencia de muestreo, nivel de cuantización, número de canales, filtros, etc.

Los estándares de codificación de archivos de audio actuales buscan dos cosas fundamentales; por un lado reducir al mínimo el tamaño de la señal digital y por otro conservar la mayor calidad posible. Los algoritmos de compresión más sencillos reducen muy poco con poco procesamiento. Los más complejos ofrecen una reducción mayor pero el procesamiento también es mucho más elaborado.

En cuanto a calidad, existen muchos parámetros para especificar el nivel de calidad de una señal sonora digital de una señal. Sin embargo la relación más aceptada para establecer este nivel consiste en una relación entre la señal de audio original y la señal de audio analógica, la exactitud de estas pruebas solo considera la sensibilidad humana. Para refinar estas pruebas, otro método consiste en determinar la relación de señal a ruido donde la diferencia es más exacta que para señales idénticas para el oído humano.

3.1 Principios de acústica

La acústica es la ciencia que estudia las ondas sonoras en su generación, propagación y recepción, así como las circunstancias que producen estos tres factores cuando se crea una perturbación acústica [8].

Sin embargo, Sergi Jorga hace una descripción más acertada del sonido [9]. De este estudio podemos asumir al sonido como una compleja interacción de un objeto vibrante, un medio transmisor, el oído y el cerebro. La Fig. 3.1, muestra la relación que existe entre estas partes y la generación del sonido.

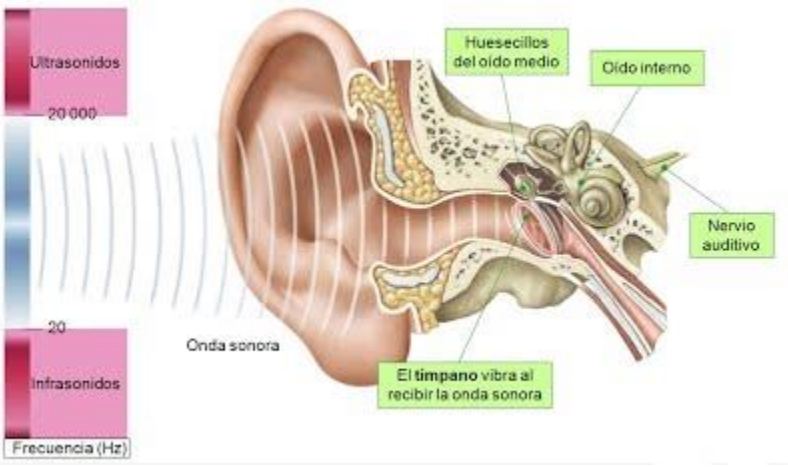


Fig. 3.1. Relación del sonido con el odio humano.

En la parte izquierda aparece el rango de frecuencias audibles y a la derecha las partes del oído que son estimuladas por las perturbaciones del aire, hasta excitar el nervio auditivo conectado al cerebro. Para que un objeto sea audible por el ser humano, la oscilación de este objeto debe ser aproximadamente entre 20 y 20,000 veces por segundo. Cuando un objeto oscila desplaza las partículas del medio que lo rodea, comprimiendo y descomprimiendo periódicamente las moléculas que lo integran, de esta manera se producen pequeños cambios de presión en forma periódica. Como las moléculas van desplazando a las contiguas, la variación periódica de la presión se propaga originando las ondas sonoras. Al entrar al oído estas ondas de presión son recibidas por el tímpano y transmitidas al cerebro por el nervio auditivo, el cual lo interpreta como sonido.

Las unidades de presión, no son suficientes para medir las vibraciones de las partículas del aire, los esfuerzos de compresión y descompresión generados por las oscilaciones del objeto son demasiado pequeñas como para poder usarse como referencia. Para ello se recurre al decibelio del nivel de presión sonora (dB SPL), el cual toma como referencia el menor nivel de presión sonora que el oído humano puede detectar. El menor sonido audible para el ser humano es de 0 dB SPL. La Tabla 3.1 muestra algunas características importantes que relacionan los niveles de audición con el oído humano

Tabla 3.1 Limites del sonido audible

	Sonidos Agudos	**Sonidos graves**
Longitud de onda	2[cm]	17 [m]
Frecuencia	0.05[ms]	50[ms]
Periodo	20[kHz]	20[Hz]

3.1.1 Características de propagación del sonido.

Durante la propagación del sonido se observan tres fenómenos acústicos que pueden distorsionar la señal original [10]. Estos son:

- La reflexión: se produce cuando una onda choca contra una superficie de otro medio, esto provoca que la onda original se divida en dos señales una de reflexión y otra de transmisión dividendo entre ambas la energía de la señal original. La proporción entre la señal de reflexión y la señal de transmisión depende del ángulo de inclinación entre los medios y del medio.
- La absorción: es producida por la fricción de la señal con el medio de transmisión, el roce constante libera energía en forma de calor que es sustraída de la señal. La cantidad de la pérdida de energía depende de la frecuencia de la señal siendo mayor para altas frecuencias que para las bajas frecuencias.
- La difracción: se produce cuando la trayectoria de la señal es interrumpida por diferentes obstáculos que se encuentran en el medio, la presencia de los obstáculos disminuye la intensidad de los sonidos que pasan a través de ellos.

3.1.2 Sonidos periódicos

Los sonidos rara vez son producidos por una única perturbación [11]; los sonidos analógicos son producidos en su mayoría por múltiples perturbaciones periódicas. Estas perturbaciones pueden dividirse en ciclos, donde cada ciclo contiene información entre dos perturbaciones sucesivas.

En la Fig. 3.2 se muestran las diferentes componentes necesarias para definir una onda periódica. Son 4 componentes, Ec (3.1):

$$f(t) = A\sin(2\pi f t + \varphi) = A\sin\left(\frac{2\pi t}{T} + \varphi\right) \qquad (3.1)$$

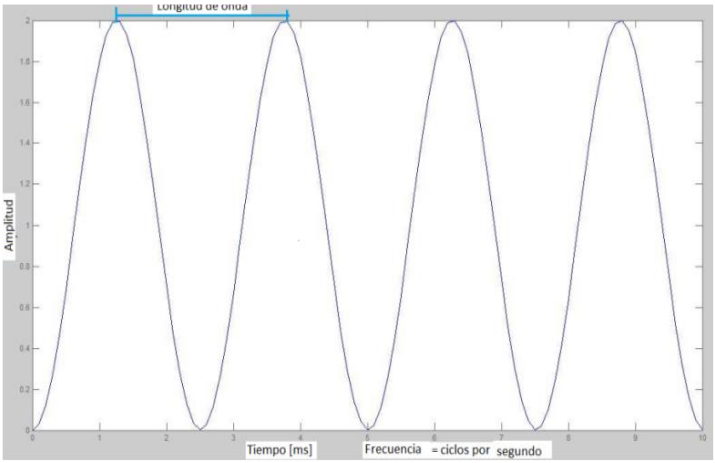


Fig. 3.2 Ejemplo de una señal periódica.

La longitud de onda λ: es la distancia que existe entre dos oscilaciones. Se mide en unidades de metros o pies.

La frecuencia *f,* Ec. (1) es el número de oscilaciones que ocurren en un segundo. Su unidad de medida es el Hz, Ec. 3.2.

$$1Hz = 1\ ciclo = \frac{1}{s} \qquad (3.2)$$

El periodo *T,* Ec. (3.3), es el tiempo transcurrido entre una oscilación y otra, se mide en segundos.

$$T = \frac{1}{f} \qquad (3.3)$$

La amplitud A, es el punto máximo que alcanza una oscilación. También se denomina valor pico.

La fase indica la posición de la partícula que oscila en el momento de empezar a contar el tiempo *t,* es decir $t = 0[s]$.

3.1.3 Espectro

Cualquier sonido periódico puede representarse como una serie de frecuencias características, estas frecuencias características son el resultado de la descomposición de una señal.

La información sobre las frecuencias que contiene un determinado sonido y sus respectivas amplitudes constituyen al espectro. El espectro se obtiene calculando la energía que aporta a cada frecuencia al sonido total [10].

La representación de un espectro de un sonido se puede hacer de dos formas, la primera es mediante tabla que relacione el número del armónico con su amplitud; y la segunda es mediante una gráfica que relaciona el intervalo de tiempo con el rango de frecuencias presentes en el sonido, esta grafica es llamada espectrograma y constituye una herramienta fundamental en el análisis acústico.

3.2 Psicoacústica, análisis analógico del audio

El cerebro es capaz de reconocer la voz de diferentes personas y asociarlas a un sujeto en específico. El estudio sobre las propiedades acústicas que permiten la identificación de la voz de una persona a otra, o de cualquier sonido en general lo hace la psicoacústica.

La psicoacústica, es la ciencia encargada del estudio de la percepción subjetiva de las cualidades del sonido. Éstas son intensidad, tono y timbre. Dichas cualidades o características, están a su vez determinadas por los propios parámetros del sonido, sobretodo la amplitud y la frecuencia [12].

3.2.1 Altura

La altura es la cualidad que nos permite distinguir un sonido agudo, de un sonido grave, esto mediante la frecuencia de vibración del objeto [13].

Esta cualidad es puramente subjetiva, sin embargo la propiedad física que relaciona la altura con la percepción es la frecuencia de vibración. La condición necesaria para poder percibir una altura es que la frecuencia debe ser aproximadamente periódica.

La forma de representar las alturas es mediante el uso de octavas; la altura está determinada por la frecuencia, y mediante una relación de frecuencias es posible representar los cambios en la altura, esta relación de frecuencias es lo que se conoce como octava.

3.2.2 Timbre

Esta cualidad es la más importante para la identificación de hablantes, y que por ella nos es posible reconocer a la persona a la cual pertenece. La concibe en conceptos musicales como la cualidad que nos permite distinguir un instrumento musical de otro, o una voz de otra [13]. Esta cualidad depende principalmente de los armónicos del sonido.

El timbre tiene una repercusión más importante que la altura. Dos o más voces pueden tener la misma altura, pero no las oímos de la misma manera, debido a que en los sonidos la frecuencia más grave es la que determina el periodo y la

altura. La frecuencia más grave se conoce como frecuencia fundamental, y las posteriores frecuencias son sus armónicos, que corresponden a sonidos agudos. La suma de los armónicos determina el timbre de la voz, y es lo que la identifica de las demás.

El primer matemático que emprendió el estudio de sobre los armónicos, fue Jean Baptiste Fourier, el descubrió que una señal por compleja que sea puede descomponerse en una suma algebraica de señales sinusoidales armónicas, obtenidas de una original.

Los factores que influyen en la estructura armónica son los números, magnitud y fluctuación de los armónicos, junto con la presencia o ausencia de armónicos superiores; al ancho de banda de la señal y la energía aportada a la misma por los armónicos en relación con la energía total.

Existen dos enfoques para el análisis del timbre, [11]:

> El primer enfoque estudia los sonidos aislados, y se propone identificar todos los elementos que los distinguen de otros sonidos, intervienen dos elementos: el espectro y las envolventes. Hay una envolvente primaria, que es la que determina la forma en que varía en el tiempo la amplitud general, y una serie de envolventes secundarias, que corresponden a las variaciones temporales relativas de los armónicos o de los parciales.

El segundo enfoque clasifica los sonidos según la fuente, [11]:

> Busca las características comunes a todos los sonidos de una voz, y las que los distinguen de los sonidos de otras voces. El elemento fundamental de este análisis es la existencia de resonancias en los componentes accesorios al mecanismo propiamente dicho de producción del sonido, resonancias que filtran el sonido, favoreciendo determinadas frecuencias más que otras.

3.2.3 Intensidad

La intensidad nos permite distinguir un sonido fuerte de un sonido suave y depende de la amplitud de las vibraciones del cuerpo [13]. El orden de la intensidad de un sonido es muy variable y se trata de un crecimiento logarítmico en las amplitudes de la señal. La intensidad se mide en decibelios.

3.3 La voz humana

La voz humana forma parte del sistema de comunicación que utiliza señales acústicas emitidas y recibidas por los seres humanos. La principal utilidad de los sistemas de comunicación es que sirven para transmitir información.

Todo sistema de comunicación se forma de diferentes componentes: emisor, receptor, mensaje, código, canal y contexto [14]. Para el caso de los sistemas de comunicación por voz, el emisor es la persona que expresa el mensaje y, más concretamente su sistema fonatorio que traduce el mensaje a una sucesión de sonidos. El receptor es el sistema auditivo que percibe los sonidos emitidos por el transmisor y los transforma en impulsos nerviosos que son interpretados por el cerebro. El mensaje es la declaración que se comunica. El código es el lenguaje hablado. El canal es el medio en el que se propaga la señal, principalmente el aire, aunque también puede ser un medio de transmisión electrónico, como en el caso de teléfonos cableados. El contexto los constituyen los factores que alteren la señal transmitida y el motivo de la conversación.

3.3.1 Anatomía del aparato fonatorio y del sistema auditivo

La voz humana se produce voluntariamente por medio del aparato fonatorio. Este aparato está formado por los pulmones como fuente de energía en forma de flujo de aire, la laringe, que contiene las cuerdas vocales, la faringe, las cavidades oral y nasal y una serie de elementos articulatorios: los labios, los dientes, el alveolo, el paladar y la lengua, Fig. 3.3.

La cuerdas vocales son dos membranas dentro de la laringe que vibran al pasar el flujo de aire de los pulmones, la frecuencia de vibración de las cuerdas vocales juntos con los movimientos de los elementos articulatorios son los que generan los diferentes sonidos emitidos por una persona.

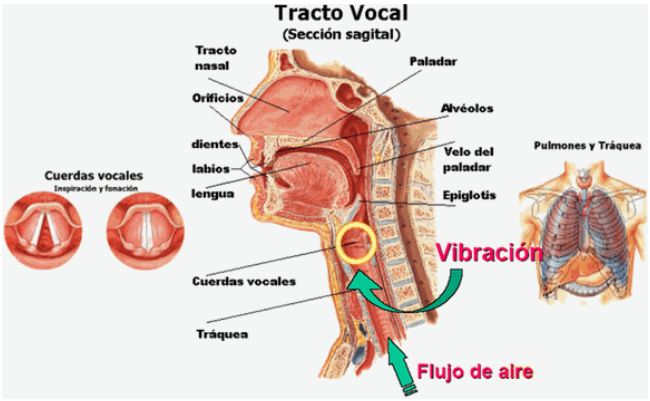


Fig. 3.3. Partes del aparato fonatorio [15].

El sistema auditivo es el sistema que nos permite percibir los oídos. La Fig. 3.4 representa las componentes fisiológicas del oído humano.

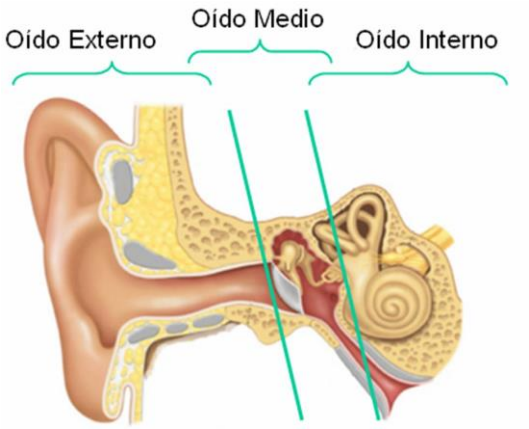


Fig. 3.4 Fisiología del oído humano [16].

Su estructura se compone de tres partes principales [14]:

1. Oído externo: Diseñado estructuralmente para recoger las ondas sonoras y dirigirlas al interior durante el proceso de audición.
2. Oído medio: Transforma la energía acústica en energía mecánica transmitiéndola y amplificándola hasta el oído interno.
3. Oído interno: Se realiza la transformación de la energía mecánica, producida por las ondas sonoras, en energía eléctrica.

3.4 Análisis del audio digital

El audio digital se genera mediante proceso de transformación de audio analógico a un conjunto de información binaria, mediante técnicas de conversión analógica-digital.

El audio es analógico en su origen y, por lo tanto, los sistemas digitales deben transformar este carácter analógico, mediante los procesos de muestreo y codificación. Las muestras de una señal de audio deben cumplir el teorema de muestreo, el teorema establece una condición: la señal debe ser de banda limitada. Además el muestreo debe realizarse con cierta precaución, ya que puede producir un aliasing. El error de cuantificación se puede minimizar aplicando técnicas de dithering [8].

El sistema de conversión completo contiene dos conversores, un conversor analógico digital a la entrada y un conversor digital-analógico a la salida. El proceso comienza cuando la señal ingresa al convertidor de entrada, durante el proceso la señal es muestreada, cuantificada y codificada, otros procesos más

complejos utilizan filtros para eliminar frecuencias indeseables, o eliminar ruido. La Fig. 3.5 muestra un diagrama con las etapas del proceso A/D:

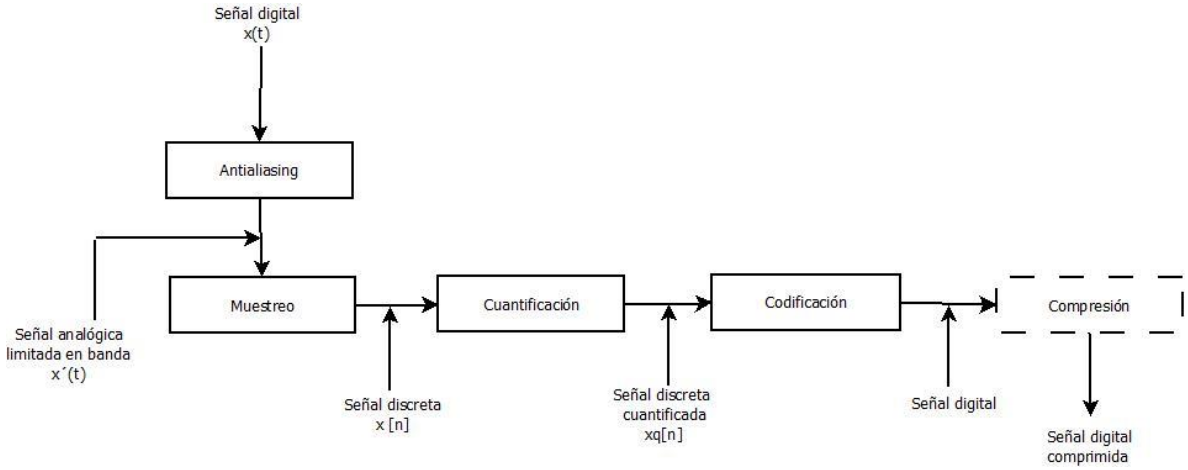


Fig. 3.5 Proceso de conversión de una señal analógica a una señal digital.

El muestreo toma pequeñas porciones de datos de la señal original cada periodo de tiempo determinado por el ciclo de trabajo del sistema, la cuantificación discretiza las muestras, y la codificación asigna valores binarios a cada valor discreto. Las tasas de muestreo y el número de bits por muestra determinan la calidad del audio. El proceso inverso recupera parte de la señal original.

3.4.1 Teorema del muestreo

El teorema de muestreo establece que una señal continúa limitada en banda puede ser remplazada por una secuencia discreta de muestras sin pérdida de información, y describe cómo se pueden reconstruir la señal original a partir de esas muestras [8].

> "El teorema especifica que la frecuencia de muestreo debe ser al menos el doble de la frecuencia máxima original. Las señales de audio con frecuencias entre 0 y la frecuencia de Nyquist (S/2) Hz pueden especificarse exactamente con S muestras por segundo" (Nyquist-Shannon, 1949).

El algoritmo para señales de audio analógicas, es el siguiente:

1. Primero las señales se pasan por un filtro pasa-bajas, para que la respuesta en frecuencia quede limitada en banda y no exceda la frecuencia de Nyquist, además el filtro también elimina aquellas frecuencias que no son percibidas por el oído. La Fig. 3.6 muestra una señal preparada para iniciar el proceso de conversión, esta señal ya se encuentra con un ancho de banda limitado.

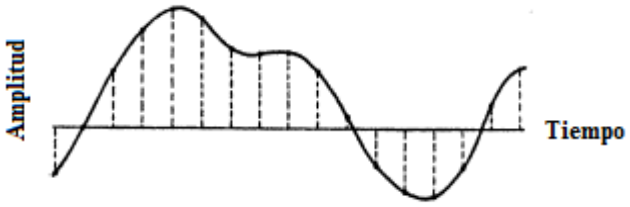


Fig. 3.6 Señal analógica filtrada.

2. La señal es muestreada para obtener valores en amplitud, estos valores representan impulsos de la señal. La señal muestreada contiene la misma información que la señal a la salida del filtro. El proceso de muestreo de hace con la convolución de la señal original con una función tren de impulso. El periodo de la función impulso determinara la separación entre datos. En la Fig. 3.7 se observan amplitudes de la onda tras el muestreo.

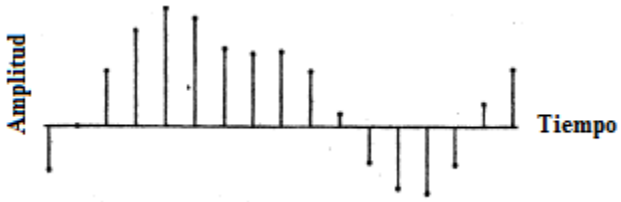


Fig. 3.7 Señal analógica muestreada.

3. Posteriormente, cada muestra es cuantificada y codificada. Con este procedimiento es posible reconstruir la señal a su forma original, sin ningún tipo de pérdida de información. La Fig. 3.8 da una visión simple pero coherente del proceso de reconstrucción.

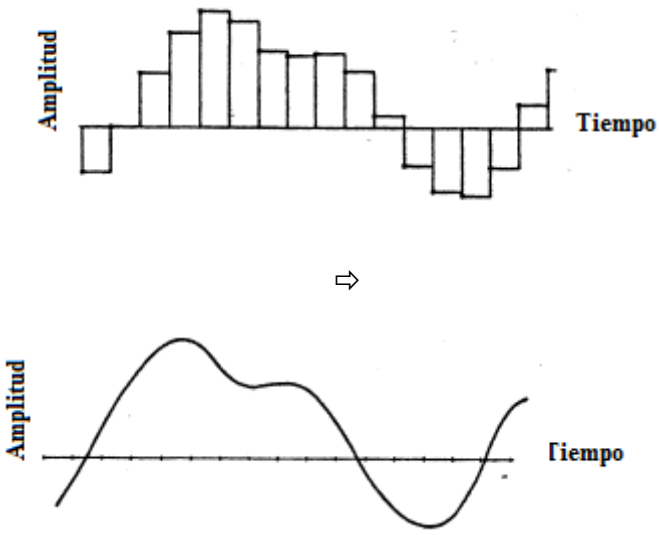


Fig. 3.8 Reconstrucción de una señal analógica de una señal digital.

Una señal muy variable requiere una tasa de muestreo mucho más alta esto requiere un proceso de cuantificación más intenso, la selección de una frecuencia de muestreo es muy importante porque determina el ancho de banda del sistema.

El teorema de muestreo especifica cómo debe muestrearse una señal para asegurar un determinado ancho de banda. La frecuencia de muestreo debe ser por los menos el doble de la frecuencia máxima de la señal de audio para tener un muestreo sin pérdida de información [8]. El uso de un filtro asegura que las frecuencias superiores sean eliminadas para evitar al aliasing. Y para asegurar la recuperación de la señal original también es necesario colocar otro filtro paso-bajas para eliminar la distorsión de armónicos totales.

3.4.2 Aliasing

El aliasing es un fenómeno anómalo que aparece en el proceso de muestreo. El aliasing puede crear componentes falsas en una señal, estas componentes aparecen dentro del ancho de banda de la señal y son indetectables. Durante el proceso de muestreo sin pérdidas, la condición que se debe cumplir es tener un ancho de banda limitado. El aliasing aparece cuando se viola el teorema de muestreo. La frecuencia más alta de la señal de audio debe ser igual o inferior a la frecuencia de Nyquist [8].

Cuando las frecuencias alcanzan la frecuencia de Nyquist, se crean dos muestras por ciclo. Con frecuencias más altas, el proceso continúa creando muestras pero con información falsa de la señal. Los componentes de aliasing no solo ocurren alrededor de la frecuencia de muestreo sino también en múltiplos de la misma.

La forma de resolver este problema es empleando un filtro paso-bajas, el filtro debe proporcionar una atenuación a partir de la frecuencia de Nyquist para asegurar que el espectro de la señal no tenga frecuencias superiores. Sí el filtro no es capaz de corregir el aliasing, una vez se muestree la señal no habrá forma de eliminar las componentes erróneas en la señal digital.

3.4.3 Cuantificación

El muestreo representa los instantes de medida, y la cuantificación representa los valores de medida, en audio representa la amplitud de la señal en los instantes de muestreo [8] Una señal analógica puede representarse mediante una serie de pulsos, la amplitud de cada pulso indica el valor numérico de la señal en ese preciso instante.

La precisión del valor de la cuantificación está determinada por la resolución del sistema. La resolución del sistema depende de la cantidad de bits utilizados para representar la amplitud.

En una cuantificación uniforme, la amplitud de la señal se transforma en un determinado número de niveles de cuantificación, todos de igual tamaño, para poder precisar lo más posible el valor original. Los sistemas de audio de alta calidad utilizan al menos 65 000 niveles de cuantificación.

La cuantificación el proceso en el cual a cada muestra se le asigna un valor de un conjunto finito de niveles [17]. Teóricamente el muestreo es un proceso que no produce pérdidas de información, por el contrario, la cuantificación, no importa cuál sea la escala ni cuántos sean los niveles de cuantificación, o el código utilizado siempre existe un error.

3.4.4 Dither

El dither es una técnica empleada en audio digital para eliminar la distorsión producida por el proceso de cuantificación. Su misión es transformar la distorsión en ruido blanco [17]. El dither es un ruido de bajo nivel incorrelado con la señal de ruido, se añade a la señal de audio antes de ser muestreada. Cuando se añade el dither la amplitud de la señal cuantificada se balancea en torno a los niveles de cuantificación.

El dither no enmascara el error de cuantificación; más bien permite al sistema de digitalización codificar amplitudes inferiores al bit menos significativo [8].

3.4.5 Codificación

La codificación sigue un estándar, se ha establecido la numeración binaria como método de asignación de valores a las muestras, principalmente por el hecho de que los niveles de cuantificación siempre son potencias de 2; (2^n).

3.4.6 Compresión

Estrictamente, la compresión de la señal resultante no es un proceso de la conversión analógica-digital, debido a que la señal digital ya fue obtenida en la codificación; sin embargo es prudente mencionar este proceso porque muchas de las aplicaciones que requieren una conversión analógica – digital o viceversa han sometido a las señales a un proceso de compresión de información. Para las señales digitales de audio, el método más difundido en la actualidad es el mp3.

Los sistemas de compresión analizan el espectro de la señal de audio, obteniendo una distribución en tonos y bandas para después, aplicando métodos psicoacústicos, eliminan el contenido que el oído no puede percibir [17]

En resumen, el muestreo y la cuantificación son los dos procesos fundamentales de la digitalización. El muestreo determina el ancho de banda de la señal y por lo tanto la respuesta en frecuencia. La cuantificación determina el rango dinámico del sistema, que puede ser medido como la relación señal/ruido.

CAPÍTULO 4

ORGANIZACIÓN Y ADMINISTRACIÓN DE LA BASE DE DATOS

La Base de Datos es un mecanismo de organización de información muy sofisticado. Ésta permite almacenar cantidades masivas de datos, permite recuperar datos completamente o que cumplan criterios de búsqueda, facilita la distribución de los datos balanceadamente para su análisis, estudio o procesamiento, y permite la toma de decisiones a partir de los datos seleccionados.

Los beneficios de las bases son muchos, si se saben aprovechar correctamente, todo inicia desde un minucioso análisis para el diseño de los datos. Saber que datos y que tipos de datos de datos se van almacenar suele tener ventajas en la distribución de la información, evitar sobrecarga de datos, administrar correctamente el espacio de disco y obtener resultados correctos de las consultas.

En este apartado se intenta, además de obtener como resultado una base de datos estable, desarrollar un correcto diseño de la estructura interna, gestionar correctamente los recursos a mi disposición para obtener beneficios extras como mejorar el tiempo de respuesta de consultas, adaptar correctamente los espacios de memoria y crear inserciones perfectamente organizadas.

Para lograr estos objetivos se cuenta con herramientas de gestión de bases de datos, de diseño y los propios lenguajes de manipulación y definición de datos. Si bien la base de datos global no constituye un reto para su creación, por el hecho de manejar pocas tablas, algunas de ellas incluso con muy pocos registros; sí se requiere un especial cuidado en aquellas cuyos registros son de mucha importancia para la etapa de análisis de datos. Y por tanto el mayor tiempo de procesamiento lo ocuparan estos métodos de análisis y los tiempos de consultas deben ser lo más cortos posibles.

4.1 Conceptos teóricos

Las Bases de datos son sistemas de información organizada en archivos digitales que permiten la administración, almacenamiento, organización y recuperación de datos para minimizar la tarea de lecturas de información distribuida en diferentes medios. La principal ventaja al contar con una base de datos es la facilidad con que los datos son recuperables, dejando como tarea principal la forma en que serán utilizados

Los beneficios secundarios de la base de datos son, que al tener un análisis previo sobre la información que se desea almacenar en esta estructura, es evitar redundancia de información, permite tener un acceso controlado al insertar nuevos elementos, permite seleccionar la información con la que se va a trabajar, minimiza los tiempos de búsqueda de algún registro en específico y optimiza la

lectura de los datos que el sistema o programa principal va a utilizar como entradas.

En este capítulo se aborda la planeación de una base de datos sencilla que permita suplir funciones al programa de reconocimiento de hablantes, y así ejecutarse de forma más óptima. Los puntos que se cubrirán en el desarrollo de la base de datos son los siguientes:

- ⇨ Planeación de la base datos.
- ⇨ Análisis del tipo de información que se almacenará.
- ⇨ Selección del tipo información almacenada.
- ⇨ Creación de registros por hablantes y asociación son sus señales se voz y espectrogramas.

Una vez hecho una correcta planeación de la base de datos, la tarea de la creación se vuelve más sencilla. Y solo queda como prioridad los sistemas que se encargará de utilizar esta información, una consecuencia directa es que los tiempos son más cortos y se evita la sobrecarga de memoria al acceder solo al registro solicitado por el programa principal.

Las tareas que se debe realizar son: selección del sistema manejador de bases de datos, creación de la base de datos, control de acceso a usuarios, creación y asignación de privilegios, carga de datos, creación de procedimientos almacenados, ejecución de programas de inserción de datos, monitoreo de usuarios, rendimientos de consultas, crecimiento de estadísticas y optimización.

4.1.1 Sistemas de bases de datos

Date define a los sistemas de bases de datos como un sistema computarizado para guardar registros [18]. El sistema debe ser capaz de adaptarse de acuerdo a la tarea para el cual fue diseñado:

- ✓ Agregar nuevos archivos vacíos a la base de datos.
- ✓ Insertar datos dentro de los archivos existentes.
- ✓ Recuperar datos de los archivos existentes.
- ✓ Modificar datos en archivos existentes.
- ✓ Eliminar datos en archivos existentes.
- ✓ Eliminar archivos existentes de la base de datos.

La Fig. 4.1, muestra los elementos con los que se compone una base de datos. Estos elementos deben sincronizarse correctamente para una mayor eficiencia en las consultas.

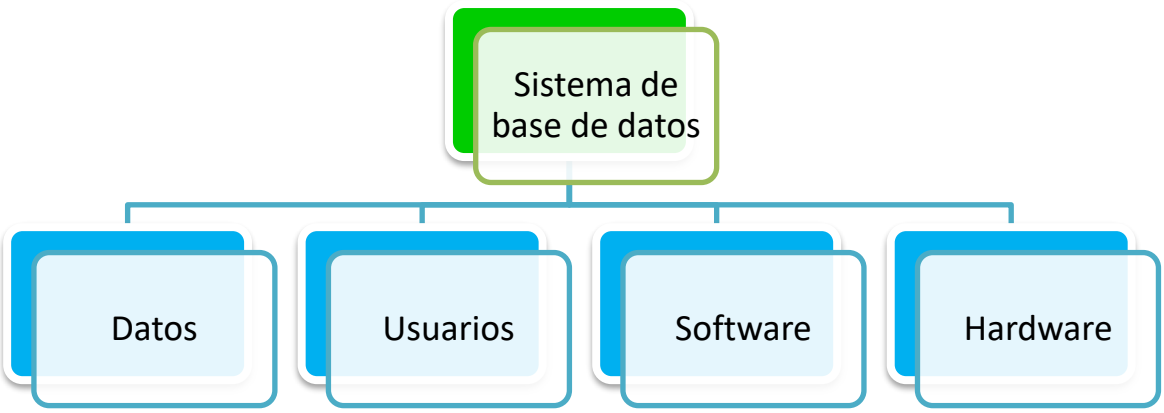


Fig.4.1. Elementos que conforman un sistema de base de datos.

Los datos son los registros almacenados. Están integrados por cada tipo de información, referencia u archivo asociado al registro. La palabra datos se deriva del vocablo latín para *da*; por lo tanto, los datos son hechos dados, a partir de los cuales es posible inferir hechos adicionales. Un hecho dado corresponde a su vez a lo que en la lógica se denomina proposición verdadera [18].

El hardware, está compuesto por los sistemas físicos que permiten el almacenamiento, presentación y componentes de e/s que permiten la inserción, modificación o acceso a los datos. El componente más importante de hardware, lo componen los discos donde se almacenan los datos, y es prioridad del administrador, siempre tener copias de los registros.

La capa de software conocida como el administrador de base de datos o el servidor de base de datos, es la encargada de procesar todas las solicitudes procedentes de los usuarios.

Los usuarios se dividen en tres clases distintas, Fig. 4.2:

1. Programadores, son los responsables de escribir programas de aplicación de bases de datos en algún lenguaje de programación.
2. Usuarios finales, son aquellos que interactúan con el sistema de bases de datos a través de aplicaciones en línea, o bien puede usar una interfaz proporcionada como parte integral del software del sistema de bases de datos.

3. El administrador de base de datos, encargado de gestionar los recursos de la base de datos.

Fig. 4.2. Tipos de usuarios de una base datos. Recuperado de [20], el 11 de octubre de 2016.

4.1.2 Ventajas de las bases de datos

La ventaja de utilizar una base datos son múltiples, pero las más importantes son las descritas en [19], de forma simplificada son las siguientes:

a) Los datos pueden compartirse. Es posible cumplir con los requerimientos de aplicaciones nuevas sin tener que agregar información adicional a la base de datos.
b) Es posible reducir la redundancia. No es necesario crear duplicados de datos para distintas aplicaciones.
c) Es posible evitar inconsistencia de datos. Existe un mecanismo conocido como propagación de actualizaciones que permite reflejar las modificaciones entre aplicaciones de manera automática.
d) Es posible brindar manejo de transacciones. Una transacción es una unidad lógica, que por lo general comprende varias operaciones de la base de datos.
e) Es posible mantener la integridad. Permite asegurar que los datos de la base de datos sean correctos.
f) Es posible hacer cumplir la seguridad.
g) Es posible equilibrar los requerimientos en conflicto.

4.2 Bases de datos de voz

La base de datos para el reconocimiento de hablantes consiste en un conjunto de archivos de audio en formato WAV, su correspondiente espectrograma en imagen BMP y un identificador del hablante al que pertenece.

4.2.1 Diseño de la base de datos

La base de datos ha sido realizada siguiendo el proceso de Fig. 4.3:

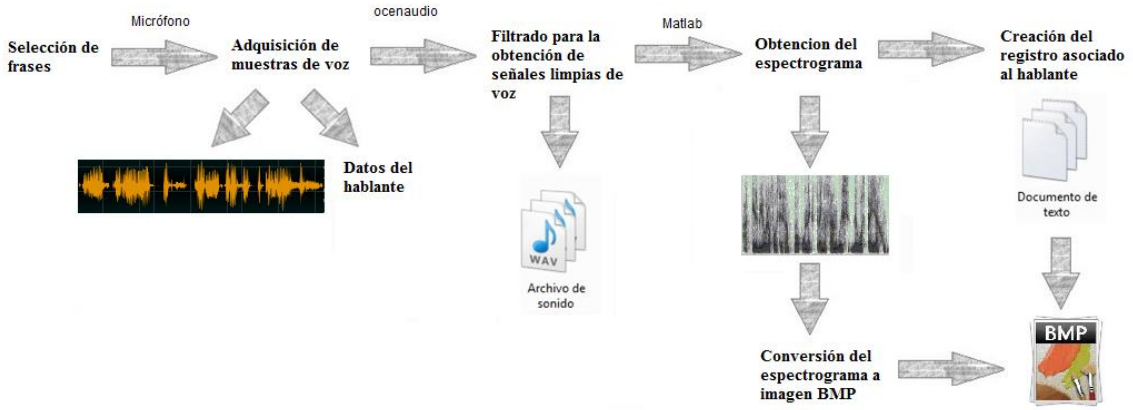


Fig. 4.3. Proceso de creación y transformación de muestras de voz de diferentes hablantes para obtener sus espectrogramas.

4.2.2 Selección de las frases

Las frases utilizadas constituyen una lista de 15 pangramas. Un pangrama (del griego: παν γραμμα, «todas las letras») o frase holoalfabética es un texto que usa todas las letras posibles del alfabeto de un idioma. En acústica los pangramas sirven para captar el sonido de todas las letras del alfabeto.

Los pangramas son frases diseñadas que utilizan todas las letras del alfabeto al menos una vez en su contenido. De esta forma se puede asegurar que todos los sonidos fonéticos que pueden ser producidos por la voz serán generados en una sola declaración.

Se tomó en cuenta el idioma, esto con el objetivo de generalizar la identificación de un hablante independientemente del idioma en el que hable:

El conjunto de frases está formado por 10 declaraciones en español y 5 en inglés:

1. Whisky bueno: excitad mi frágil pequeña vejez.
2. El viejo señor Gómez pedía queso, kiwi y habas, pero le ha tocado un saxofón.
3. Aquel biógrafo se zampo un extraño sándwich de vodka y ajo.
4. El veloz murciélago hindú comía feliz cardillo y kiwi. La cigüeña tocaba el saxofón detrás del palenque de paja.
5. Hoy bajo su valor la wulfenita, extraño molibdato que se cotiza por kilogramo.
6. El pingüino Wenceslao hizo kilómetros bajo exhaustiva lluvia y frío, añoraba a su querido cachorro
7. Tengo un libro de papiroflexia sobre las hazañas y aventuras de Don Quijote de la Mancha en Kuwait.

8. Le gustaba cenar un exquisito sándwich de jamon con zumo de piña y vodka fría.
9. Queda gazpacho, fibra, látex, jamón, kiwi y viña.
10. Manchaba una y otra la pequeña hoja de fax con kiwi y grasa.

En inglés
1. Sphinx of black quartz, judge my vow.
2. The five boxing wizards jump quickly.
3. Pack my box with five dozen liquor jugs.
4. A quick brown fox jumps over the lazy dog.
5. Sexy zebras just prowl and vie for quick hot matings.

4.2.3 Adquisición y tratamiento de muestras de voz

Para el proceso de obtención de las muestras de voz se usaron a 20 hablantes. Cada hablante leyó las 15 frases, y se obtuvo una grabación por cada una. Esto en una sola sesión por hablante. Las grabaciones se hicieron usando un micrófono convencional.

Las consideraciones durante la toma de muestras fueron las siguientes:

- Grabaciones de todas las frases por hablante en una sola sesión.
- Las grabaciones se hicieron en un cuarto aislado de ruido y sin reverberación.
- El formato de audio empleado para guardar las muestras fue WAV.
- El ancho de banda limitado por cada grabación fue de 200Hz a 3500Hz.
- Frecuencia de muestreo: 16000Hz.
- Resolución: 16 bits.

La información obtenida de los hablantes se describe en la tabla 4.1:

Tabla 4.1. Datos de los hablantes obtenidos

Información	Descripción
Id_hablante	Código numérico que permite identificar al hablante.
Id_delaracion	Código numérico que permite identificar la grabación.
Genero	Hombre o mujer
Fecha_obtencion	Fecha de la sesión cuando se obtuvieron las muestras.

Para este trabajo se usó un total de 20 hablantes, de los cuales 10 son hombres y 10 mujeres. Cada uno hizo una grabación por frase, dando un total de 300 grabaciones y 300 espectrogramas.

4.2.4 Filtro resonante para el ventaneo espectrograma

El diseño del espectrograma se hizo a partir de un modelo creado a partir de filtros resonantes. El modelo inicial del filtro parte de una función senoidal igualada con la suma de la misma senoidal defasada 3/pi hacia la izquierda y a hacia la derecha, Ec. (4.1).

$$\sin\left(t+\frac{3}{\pi}\right)+\sin\left(t-\frac{3}{\pi}\right)=\sin(t) \qquad (4.1)$$

Despejamos el seno defasado a la izquierda y generalizamos la ecuación para cualquier función $y(t)$ evaluada entre 0 y 1. Podemos simplificar la igualdad matemática y obtener una representación, como la Ec. 4.2.

$$y(t+1)=y(t)-y(t-1) \qquad (4.2)$$

Donde $t=1,2,3$ ...., donde todos los números de esta secuencia serán parte de un seno con un periodo igual a 2π. Fig. 4.4.

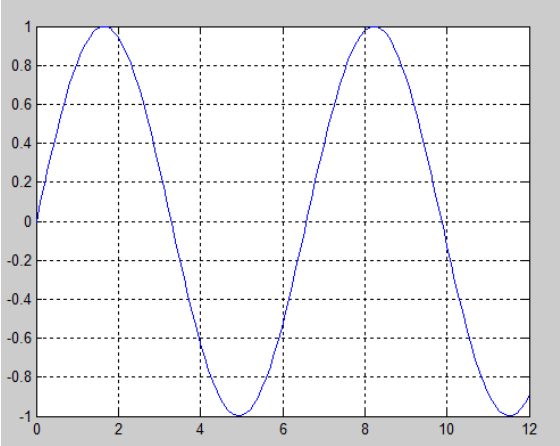


Fig. 4.4. Función $y(t)$. El periodo de esta función es igual a 2π.

Para controlar la agudeza del filtro multiplicando la función por un factor de calidad λ, y además sumamos la función de entrada a la derecha de la expresión, Ec. (4.3):

$$y(t+1)=\big(y(t)-y(t-1)\big)(1-\lambda)+x(t) \qquad (4.3)$$

El modelo analógico de la Ec. (4.3), le aplicamos la transformada Z, Ec. (4.4) para obtener un modelo digital de la función de transferencia, Ec.(4.5)

$$Yz = (Y - YZ^{-1})(1 - \lambda) + X \qquad (4.4)$$

$$H = \frac{Y}{X} = \frac{1}{z - (1 - z^{-1})(1 - \lambda)} \qquad (4.5)$$

La ventana trabajara con 128 coeficientes, para obtener estos valores evaluamos la función se transferencia con $z = e^{\frac{\pi n}{128}}$, para n=0,1, 2,…,127.

Su representación espectral es similar a un filtro gaussiano resonante Fig. 4.5. El parámetro de calidad afecta directamente la agudeza de la gaussiana. En la figura 3 se puede ver la diferencia del espectro para una $\lambda = 0$, para $\lambda = 0.5$ y para $\lambda = 0.99$.

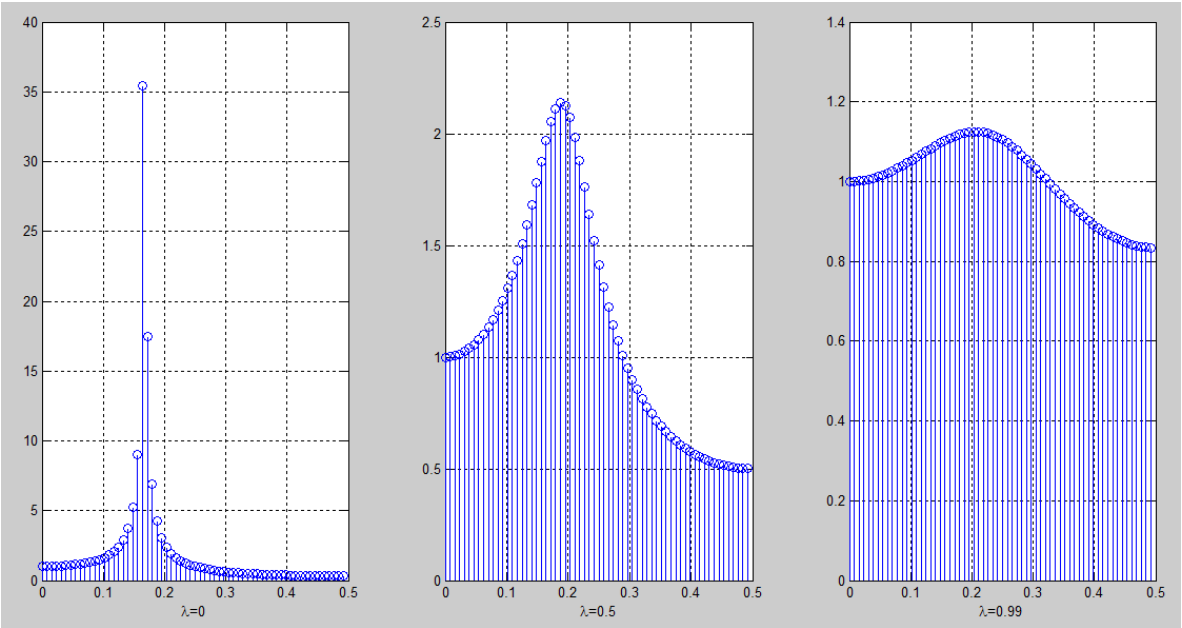


Fig. 4.5. Espectro del filtro para diferentes valores de λ.

Muchas características de la voz están escondidas en altas frecuencias, sin embargo estas frecuencias presentan amplitudes pequeñas. Y son de importancia para el reconocimiento de la persona, ya que son peculiaridades principales para cada persona. Para representar altas frecuencias se utiliza $\lambda \approx 1$. Para representar frecuencias altas son necesarios más periodos de tiempo. Los periodos corresponden al desarrollo de resonancia y depende del parámetro de calidad.

4.2.5 El espectrograma

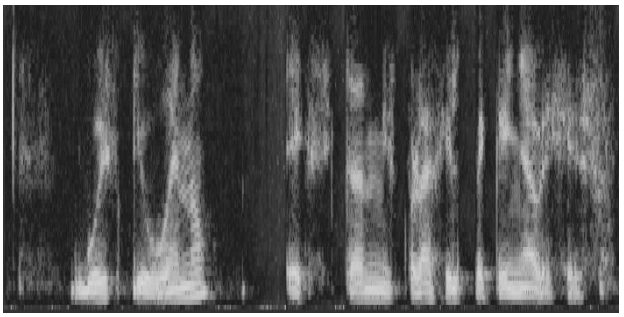
El proceso para la creación del espectrograma fue el siguiente:

En el dominio temporal se empleó una ventana rectangular de 128 coeficientes para recorrer la grabación. A estos elementos seleccionados se les aplica la transformada rápida de Fourier para obtener su contenido espectral y además se le aplica el filtro resonante de con un factor de calidad de $\lambda = 0.99$.

Para el recorrido completo de la grabación se empleó un solapamiento del 95% en la ventana temporal. Y las imágenes obtenidas de todas las grabaciones se normalizaron para tener una dimensión de 1000x 100 pixeles

Finalmente, la imagen obtenida fue almacenada en un archivo BMP es escala de grises. La imagen de la Fig. 4.5 es el espectrograma en escala de grises, en formato BMP. Además se muestra una comparación de espectrogramas entre un hombre y una mujer, para una frase dicha en español, Fig. 4.6 y otra frase dicha en ingles Fig. 4.7. En la Fig. 4.8 se muestran los espectrogramas de todas las frases dichas por un hablante.

a)



b)

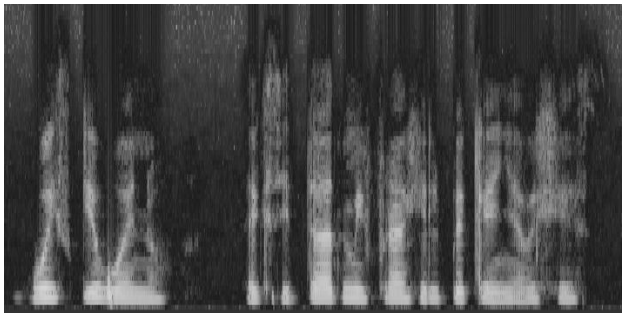
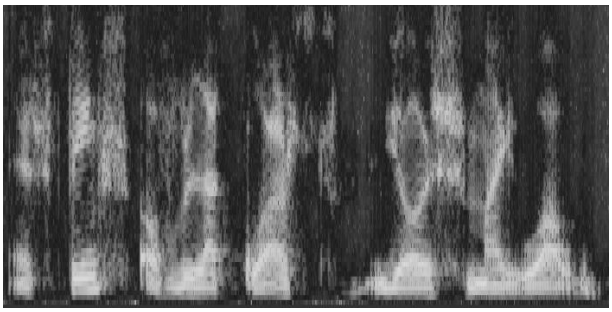


Fig. 4.6. Espectrogramas en escala de grises y formato BMP, de a) hombre y b) mujer para una frase en español.

a)



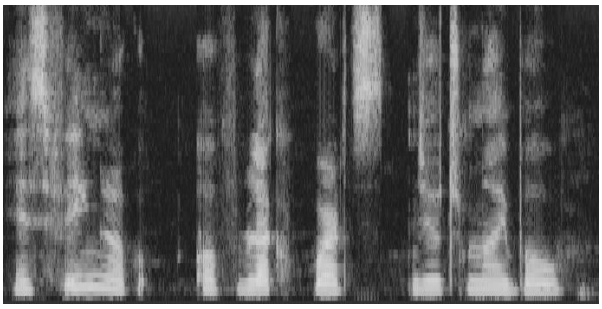
b) 

Fig. 4.7. Espectrograma hombre y mujer para una frase en inglés.

El nombrado del espectrograma incluye al identificador del hablante y el identificador de la declaración. Estos a su vez están relacionados con la información del hablante en un archivo de datos que vincula la información del hablante con sus espectrogramas.

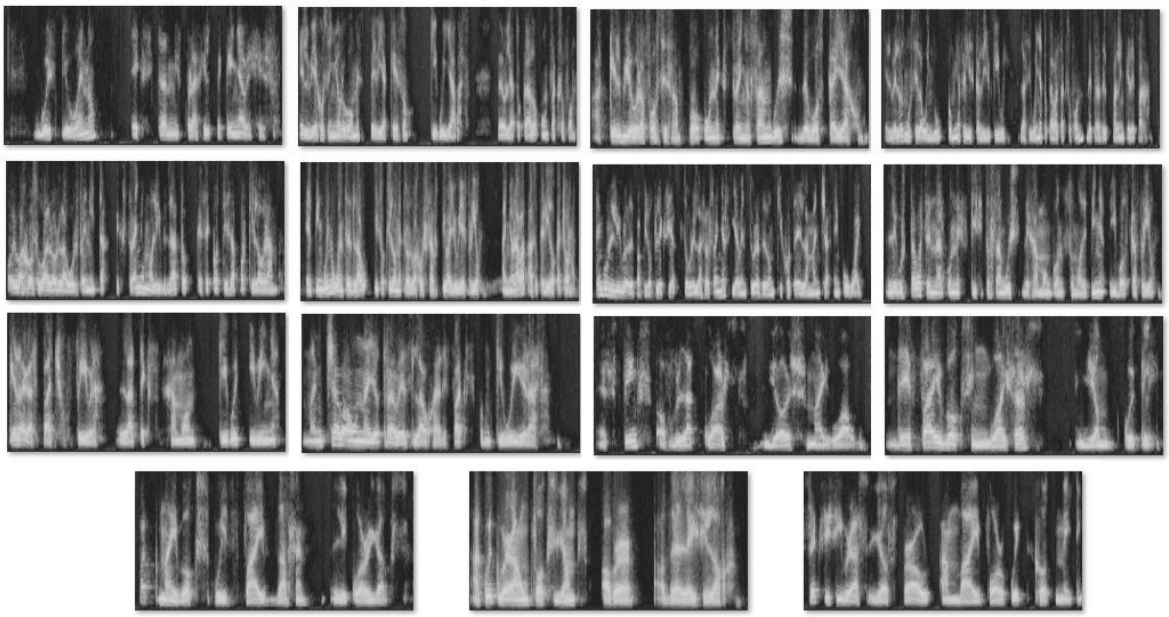


Fig. 4.8. Espectros de las 15 frases dichas por un hablante.

CAPÍTULO 5

CLASIFICADOR NEURONAL LIRA

Las redes neuronales – junto con los algoritmos evolutivos y la lógica borrosa, forman parte de una rama de la ciencia informática conocida como inteligencia artificial [23].

Formalmente las redes neuronales son grafos dirigidos, con las siguientes propiedades [21]:

1) A cada nodo i se le asocia una variable de estado x_i.
2) A cada conexión de los nodos i y j se le asocia un peso $w_{ij} \in R$.
3) A cada nodo i se le asocia un umbral θ_i.
4) Para nodo i se define una función $f_i(x_j, w_{ij}, \theta_i)$ que depende de los estados de los nodos, de los pesos de sus conexiones, del umbral del nodo. Esta función proporciona el nuevo estado del nodo.

En términos generales, en las redes neuronales, los nodos son las neuronas y las conexiones son las sinapsis.

En este capítulo se presenta a detalle el diseño y la implementación del clasificador neuronal LIRA, se hacen un desglose de los puntos que se deben cubrir como la información necesaria para su funcionamiento, los nuevos datos generados durante el proceso y finalmente la salida. Se desarrolla la trascripción de un modelo de caja negra del sistema propuesto, primero a diagramas de flujo y posteriormente a código ejecutable.

En cuanto a la red neuronal se explica la arquitectura de la red, la distribución y el funcionamiento de cada capa, las conexiones entre ellas y los métodos de excitación de las diferentes neuronas propias de cada capa.

Para explicar el método de excitación de la neurona primero de hace una analogía de la neurona artificial con la neurona biológica, y analizan sus diferencias y semejanzas.

5.1 Modelo biológico y modelo artificial

El sistema nervioso biológico está formado por células de funcionalidad especifica llamadas neuronas. Estas células tienes la capacidad de comunicarse mediante micro impulsos eléctricos entre sí. En conjunto, el objetivo principal es dar al organismo información sensorial de su entorno. La Fig. 5.1 contiene una imagen sencilla de un neurona bilógica.

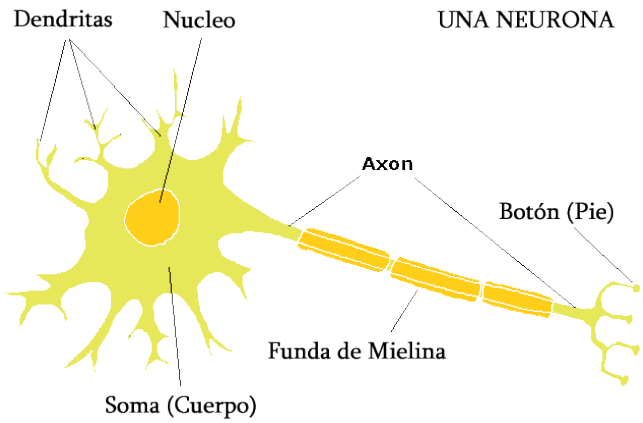


Fig. 5.1. Neurona biológica.

La información de una neurona en general a través de las prolongaciones llamadas axones, que terminan en uniones especializadas denominadas sinapsis. La sinapsis puede localizarse en las prolongaciones neuronales llamadas dendritas o en el cuerpo celular neuronal, el denominado soma [22].
Las neuronas pueden variar de tamaño considerablemente. Pueden poseer numerosas dendritas que aumentan muchas veces el área receptora de la neurona. El axón de la neurona puede ser bastante corto o de más de un metro de longitud.
Las neuronas pueden dividirse en tres grandes grupos: receptoras, intermedias o de salida. Las neuronas receptoras extraen información externa y la envían al cerebro. Transforman las señales captadas por lo sentidos en impulsos eléctricos que viajan a través de sus axones. Las neuronas de salida transmiten las señales a los órganos, músculo, etc.
La manera de simular el comportamiento de una neurona biológica es mediante un modelo de entradas y salidas, el diagrama que muestra la relación entre las partes de la neurona artificial y la bilógica está en la Fig. 5.2. El conjunto de entradas forma un vector, cada elemento de entrada se multiplica por su peso correspondiente a la conexión y se obtiene su suma. La suma de las entradas por sus pesos sirve para obtener el nivel de excitación de la neurona, el cual es empleado en la función de excitación. El resultado de la función es comparado con un umbral, de tal manera que, si supera el umbral, la neurona tendrá una respuesta positiva; sino se supera el umbral la neurona tendrá una respuesta negativa.

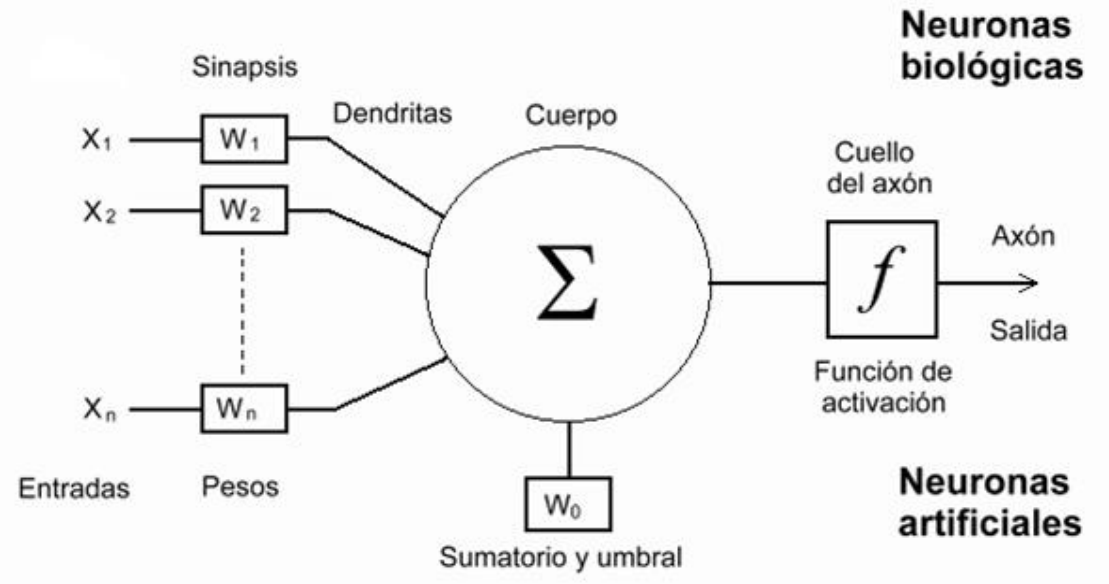


Fig. 5.2. Neurona artificial.

5.2 Estructura de un sistema neuronal artificial

Las redes neuronales artificiales imitan la estructura del sistema nervioso, con el objetivo de construir sistemas de procesamiento paralelos, distribuidos y adaptativos, que pueden considerarse de cierto modo inteligentes [21].

Esto no es posible, el cerebro y el sistema de una computadora son muy diferentes. La ejecución de un programa en una computadora con un CPU, es de manera secuencial, en comparación con el trabajo paralelo que realiza el sistema nervioso en un organismo por muy simple. A pesar de que las neuronas biológicas son mucho más lentas y simples que una CPU, el problema para emular su comportamiento subyace en la cantidad de información que captan y que son capaces de identificar de forma muy sencilla, tales como el reconocimiento del habla, visión computacional, identificación de objetos, etc.

Para resolver estos problemas se ha recurrido a tratar de imitar el funcionamiento del cerebro para resolverlos, de forma conveniente se utilizan sistemas que copien la estructura de las redes neuronales biológicas con el fin de obtener un funcionamiento similar.

Los elementos básicos del sistema nervioso biológico son las neuronas, que se agrupan en conjuntos de millones de ellas organizadas en capas, estos sistemas a su vez forman un sistema con una función específica. Un conjunto de los subsistemas da lugar a un sistema global, el sistema nervioso. En la creación de un sistema neuronal artificial se puede establecer una estructura jerárquica similar,

Fig. 5.3. El elemento esencial de partida es la neurona artificial, que se organizará en capas, varias capas constituyen una red neuronal; y por último, una red neuronal junto con las interfaces de entrada y salida, más los módulos de apoyo constituyen el sistema global del proceso [21].

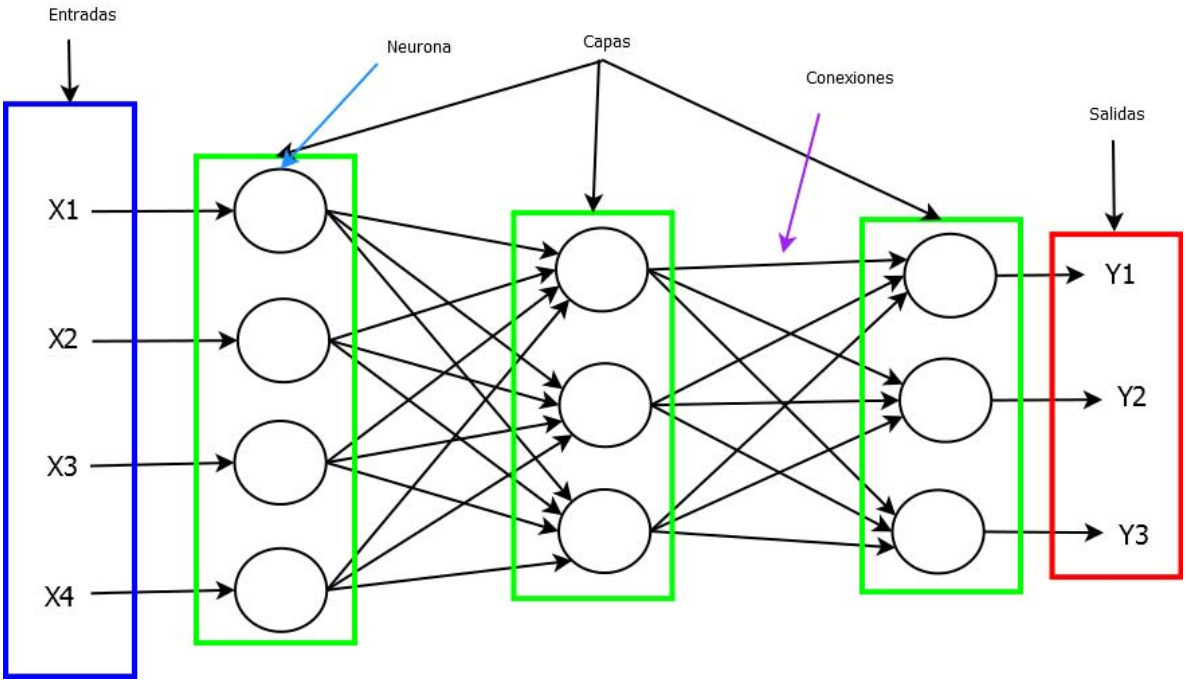


Fig. 5.3. Estructura de una red neuronal.

5.3 Arquitectura del clasificador neuronal LIRA Grayscale

El clasificador neuronal LIRA Grayscale, es una red neuronal diseñada para la clasificación de imágenes. Su nombre proviene de las siglas en inglés, LImited Receptive Area Grayscale, y hace referencia a imagen en escala de grises, que será usada como entrada en la estructura de la red neuronal.

En trabajos previos, el clasificador LIRA, fue utilizado exitosamente en la tarea de reconocimiento y clasificación de microtornillos fabricados por micromáquinas, este clasificador fue implementado en lenguaje C++, y compilado usando C++ Builder [23].

En este contexto, el clasificador neuronal LIRA fue adaptado para trabajar con espectrogramas de audio, específicamente de voz humana, e implementado en scripts de Matlab.

La Fig. 5.4, es un diagrama de la arquitectura del clasificador neuronal LIRA, en él se muestran las cuatro capas que lo componen, siendo: la capa de entrada (S), la capa intermedia (I), la capa asociativa (A) y la capa de salida (R).

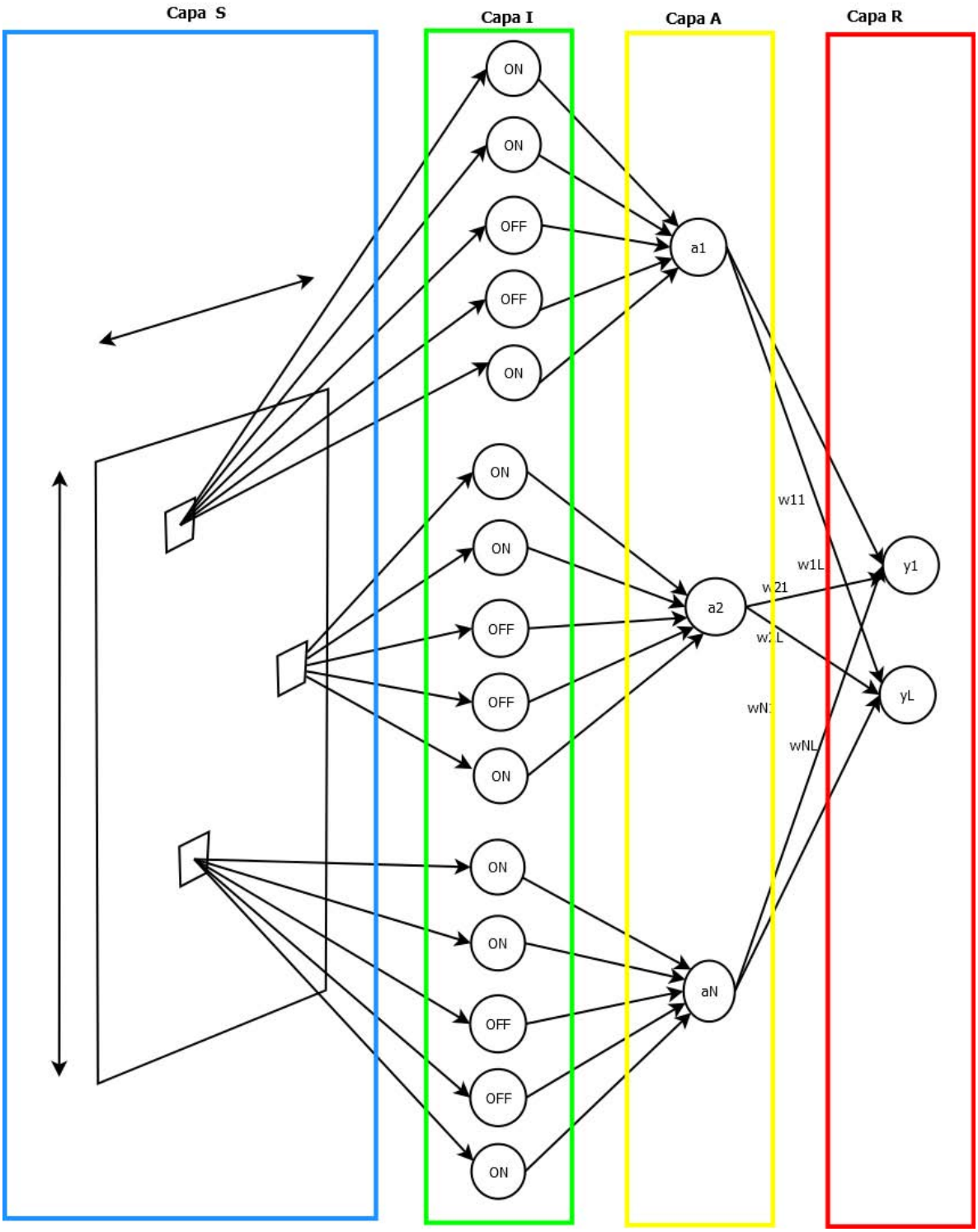


Fig. 5.4 Arquitectura del clasificador LIRA Grayscale.

La arquitectura del clasificador es usada para determinar un hablante por cada imagen. Los espectrogramas tienen una longitud de 1000 pixeles, y por cada uno de ellos se selecciona de manera progresiva una región del espectro de 100

pixeles de longitud por la altura del espectro, iniciando desde el pixel 1 para la primera región; el avance de las regiones se hace usando un porcentaje de solapamiento sol entre regiones. Cada una de estas regiones constituye una imagen que será usada en el clasificador LIRA.

5.3.1. Capa de entrada

El número de neuronas de la capa S o de entrada corresponden a una ventana seleccionada aleatoriamente en una región del espectrograma.

Las imágenes son archivos BMP en escala de grises, por lo que el valor que puede tomar una neurona está dentro del rango [0,255]. El procedimiento para obtener estos valores es el siguiente: se selecciona un número N de ventanas con un tamaño de 20 x 20 pixeles de manera aleatoria en la región del espectro, por cada una de estas ventanas se seleccionan además cinco pixeles de los cuales se toma su valor numérico, que serán las entradas de las neuronas ON, OFF, Fig. 5.4. La selección de las N ventanas y de los pixeles solo se realiza una vez durante todo el proceso, estas ventanas y pixeles serán los mismos para cada zona de la imagen analizada. Por las dimensiones de la imagen y el número de neuronas, se puede tener una misma ventana asociada a diferentes neuronas, pero la selección de los 5 pixeles será diferente. El objetivo de seleccionar primero una ventana y posteriormente cinco puntos de esta, es para delimitar el esparcimiento de los puntos a un área muy reducida.

Como se trata de una imagen, para la selección de las ventanas solo basta con seleccionar una coordenada de la imagen estudiada y almacenarla en un registro, esta coordenada corresponde a la esquina superior izquierda de la ventana, Fig. 5.5. Y para la selección de los pixeles se obtienen 5 coordenadas, pero dentro del rango [0,19]. El registro de estos puntos queda almacenado en un archivo, que sirve como patrón para todas las imágenes. Para obtener el valor del pixel, a la coordenada de la ventana se le suma la coordenada del pixel, y se extrae su valor.

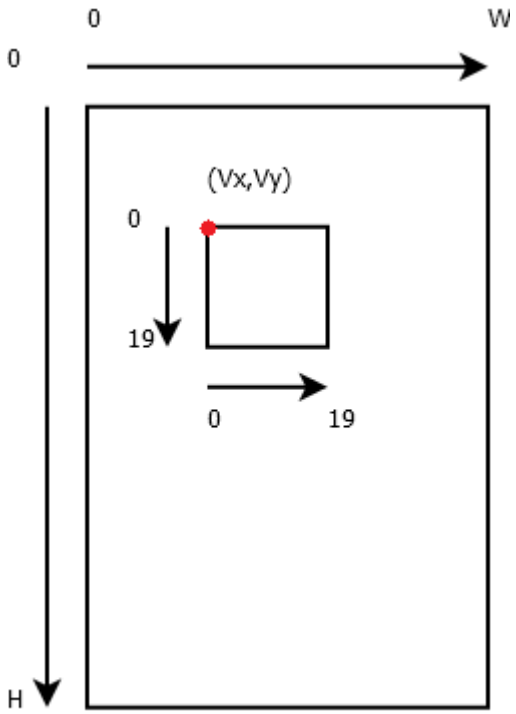


Fig. 5.5 Selección de la ventana.

El número total de neuronas de la capa S es igual a N.

5.3.2. Capa intermedia

La capa I o intermedia, es una capa de procesamiento. Esta capa está formada por dos tipos de neuronas: neuronas ON y neuronas OFF. El total de neuronas ON y OFF es igual al número de neuronas de la capa de entrada. Además, cada neurona tiene asociado un umbral de excitación T.

Las neuronas de la capa de S se conectan directamente a las neuronas de la capa I, en relación uno a uno, de manera sistemática las neuronas de la capa S se dividen en puntos positivos y en puntos negativos. Los puntos positivos están conectados con las neuronas ON y los puntos negativos están conectados con las neuronas OFF.

El valor del umbral de excitación T es seleccionado de forma aleatoria del rango de valores [0,255]. Este valor sirve para activar su neurona asociada de la capa intermedia. Para el caso de las neuronas ON, esta se activará en caso de que el valor de la neurona conectada de la capa S sea mayor o igual al umbral de dicha neurona. Para las neuronas OFF, estas se activarán cuando el valor de la neurona conectada de la capa S sea menor o igual al umbral de dicha neurona.

Cuando una neurona de la capa intermedia este activa, tendrá como salida un valor de “1”, de otro modo su valor de salida será 0.

El número de neuronas de la capa I es $N\times5$.

5.3.3. Capa asociativa

Las neuronas de la capa A están vinculadas a cada grupo de 5 neuronas provenientes de su ventana correspondiente de la capa de entrada. Las salidas de cada grupo de neuronas de la capa Intermedia de una ventana están conectadas directamente con una neurona de la capa A, en esta neurona se realiza una operación AND con los valores de salida provenientes de las neuronas de la capa intermedia.

La salida de una neurona de la capa asociativa será uno si todas sus entradas tienen un valor de uno, en caso de que cualquiera esta sea un cero, la salida de la neurona será cero.

El número de neuronas de la capa A es N.

5.3.4. Capa de salida

Las neuronas de la capa R o de salida, son las diferentes clases. Cada una de estas neuronas esta conectadas a todas las neuronas de la capa asociativa. Y por cada conexión tienen asociado un peso w_{kj}. La forma de obtener a la neurona ganadora es empleando una suma ponderada de la ecuación (5.1). La neurona ganadora será aquella cuyo valor sea el mayor de todas las salidas.

$$y_i = \sum_{k=1}^{N} w_{ik} a_k \qquad (5.1)$$

Antes de poder emplear el clasificador con éxito, este debe ser entrenado modificando los pesos de las conexiones de la neurona asociativa y de salida. Para ello se emplea la regla de Hebb.

5.4 Proceso de entrenamiento

La red neuronal emplea un proceso de entrenamiento supervisado en el que se aplica un método de selección del ganador. Este método consiste en aplicar una regla para la selección del ganador con una pequeña modificación: sea y_g la salida de la neurona ganadora y y_c la salida de la neurona competidora.

Los pesos de todas las conexiones entre las neuronas de las capas A y R deben estar inicializados en 1. En la primera etapa del entrenamiento primero se debe codificar la imagen presentada en la capa S, la codificación se hace en las capas I y A. En esta etapa se pretende conocer el número y posición de neuronas que se activan para dicha imagen.

Una vez codificada la imagen, la siguiente etapa consiste en calcular la suma ponderada, Ec. (5.1), de la capa R, empleando las neuronas activas de la capa A y los pesos correspondientes a las conexiones con cada neurona de la capa R.

Una vez calculadas las salidas de cada neurona de la capa R, la salida de clase esperada es decrementada por un factor de excitación adicional T_e. Usando la ecuación (5.2).

$$y_r = y_r(1 - T_e) \qquad (5.2)$$

Y se selecciona la neurona con la máxima excitación y_g como ganadora, g. Una vez obtenida esta neurona se procede a comparar ambas neuronas, la neurona esperada r y la neurona ganadora$g$. Si ambas son la misma, no se hace modifica ningún peso. Por el contrario, si son diferentes los pesos de las conexiones de la neurona r son incrementados en una unidad, mientras que los pesos de las conexiones de la neurona g son decrementados en dos unidades, teniendo como límite inferior el 0, asegurando que no existan pesos negativos.

El proceso termina cuando se tiene un porcentaje de errores aceptable durante el entrenamiento o cuando ya no se tiene ningún error. Un error es una modificación en la matriz de pesos.

El proceso de entrenamiento se convierte en un ciclo multiple, este proceso debe efectuarse para cada región del espectro, en todo el espectro, por cada hablante y por cada frase dicha por el hablante, delas imágenes seleccionadas para esta tarea.

5.5 Construcción del clasificador

La Fig. 5.6 contiene al diagrama de flujo general con el cual se implementó el clasificador. Este diagrama divide al sistema completo en 4 subprocesos: el primero de ellos consiste en la creación de archivos auxiliares iniciales que contienen información necesaria para los demás subprocesos; el siguiente subproceso es el de la codificación, en él se extraen las características de las imágenes y se almacenan en otro archivo auxiliar. El proceso de entrenamiento es iterativo y utiliza los códigos obtenidos del proceso anterior para modificar la matriz de pesos, además en este proceso se hace un conteo del número de errores por

ciclo el cual es utilizado para evaluar el rendimiento del clasificador. El último subproceso es el de reconocimiento, en el cual se evalúa el clasificador con códigos de imágenes que no participaron en el entrenamiento.

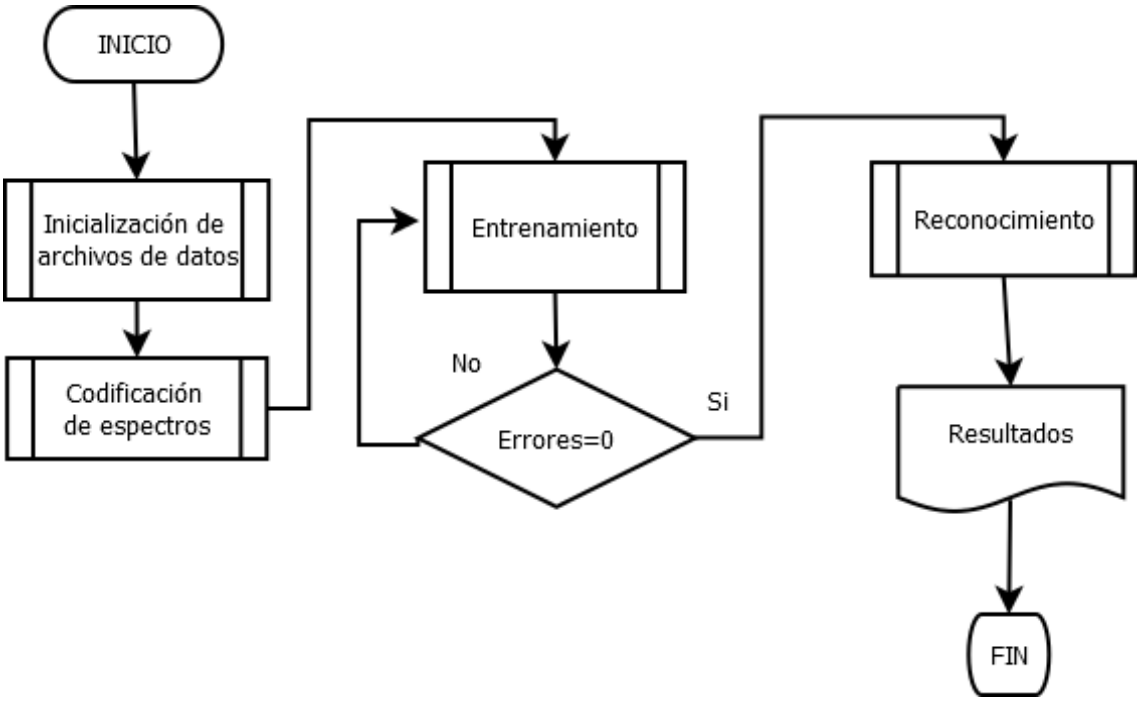


Fig. 5.6 Diagrama de flujo del clasificador LIRA Grayscale

Los parámetros usados para la implementación son:

- ✓ Número de neuronas asociativas, N: 64000.
- ✓ Número de neuronas positivas por ventana, ON: 3.
- ✓ Número de neuronas negativas por ventana, OFF: 2.
- ✓ Tamaño de la imagen: $100 \times h$ pixeles, donde h es la altura del espectro.
- ✓ Tamaño de las ventanas: 20×20 pixeles.
- ✓ Número de espectros por hablante:15.
- ✓ Clases de salida:13.
- ✓ Relación de espectros para entrenamiento-reconocimiento:80%-20%.

5.5.1 Archivos auxiliares

La capa de entrada y la capa intermedia requieren de información necesaria para poder funcionar.

Para la capa de entrada se necesitan seleccionar de manera aleatoria N ventanas de la imagen. Para esto solo se necesita especificar las coordenadas de un pixel,

además de cada ventana se deben elegir 5 pixeles, 3 que servirán como entrada para las neuronas ON y 2 pixeles para las neuronas OFF.

Para la capa intermedia se necesitan elegir de manera aleatoria los umbrales asociados a cada neurona ON y OFF.

Como ya se había especificado, N es el número de neuronas de la capa asociativa.

Los archivos iniciales son 5:

- Ventanas: Contiene las coordenadas (x,y) de la esquina inferior izquierda de la ventana en la imagen seleccionado de manera aleatoria de la imagen. La selección se hace N veces.
- P_positivos: Contiene las coordenadas de los puntos positivos, elegidos dentro de cada ventana. Contiene dos matrices de tamaño $3\times N$. Una matriz contiene posiciones en x y la otra contiene las posiciones en y. Seleccionadas de manera aleatoria dentro del rango [1,20].
- P_negativos: Contiene las coordenadas de los puntos negativos, elegidos dentro de cada ventana. Contiene dos matrices de tamaño $2\times N$. Una matriz contiene posiciones en x y la otra contiene las posiciones en y. Seleccionadas de manera aleatoria dentro del rango [1,20].
- Umbrales: Contiene una matriz de tamaño $5\times N$. Cada fila contiene los umbrales de las neuronas de la capa intermedia, seleccionados dentro del rango [0,255]. El orden de los umbrales para la disposición de las neuronas es: ON, OFF, ON, OFF, ON.
- Pesos: Contiene una matriz inicializada en 1's de tamaño $13\times N$.

5.5.2 Codificación

La Fig. 5.7 contiene el algoritmo diseñado para obtener las neuronas de la capa asociativa que fueron activadas durante el proceso. La salida del clasificador es un archivo más que contiene el identificador del hablante, el número total de neuronas que fueron activadas durante la codificación y, finalmente los índices de estas neuronas activas, denotadas por A en el diagrama de flujo, que corresponden a las neuronas de la capa asociativa con salida "1". La codificación tuvo aproximadamente el 20% del tiempo total de procesamiento.

El proceso de la codificación es el siguiente:

1) Se indica el tamaño de solapamiento en tre imágenes (sol), las dimensiones de la imagen del espectrograma ([EsW, EsH]), las dimensiones de la

imagen extraída del espectrograma ([ImW, imH]) y el tamaño de la ventana ([winV, winH]), esta última es de 20 x 20.

2) Se cargan a memoria los archivos con las variables auxiliares de la red neuronal, estos son los umbrales de excitación, los puntos negativos y positivos de la capa I, y la posición de las ventanas.
3) Se carga la imagen del espectro a memoria y se inicializa en 0 una variable (m) que servirá para recorrer el espectrograma.
4) Se selecciona la región del espectrograma determinada por el solapamiento y por m.
5) Se inicializa un índice (i) en 1, que nos servirá para identificar la neurona de la capa A que haya sido activada.
6) Para la primera ventana determinada por su posición, se leen los valores de los puntos positivos (pp) y negativos (pn), también determinados por su posición. Estas coordenadas ya fueron leídas de los archivos.
7) Se aplica una operación AND, con los resultados de las comparaciones de los valores negativos y positivos con sus respectivos umbrales. Para los puntos negativos la comparación es “mayor que” y para los valores positivos la comparación es “menor que”.
8) Si el resultado de la operación AND es 1, se incrementa el índice (i) y se almacena en una variable llamada A.
9) Si el resultado es 0, solo se incrementa el valor del índice (i).
10) El proceso se vuelve iterativo hasta completar el número total de neuronas (N).
11) Se le agrega un encabezado a la variable A, que incluye el identificador del hablante y el número total de neuronas asociativas activadas. Se crea una nueva fila a la variable, para la siguiente imagen.
12) El índice m se incrementa en 1. Y el proceso se repite hasta recorrer todo el espectrograma.
13) Se guardan en un archivo los códigos del espectrograma denotados por A.

*Este proceso solo es para un espectrograma de un hablante, se debe repetir para todos los espectrogramas de todos los hablantes. Es decir, son 15 espectrogramas por 13 hablantes.

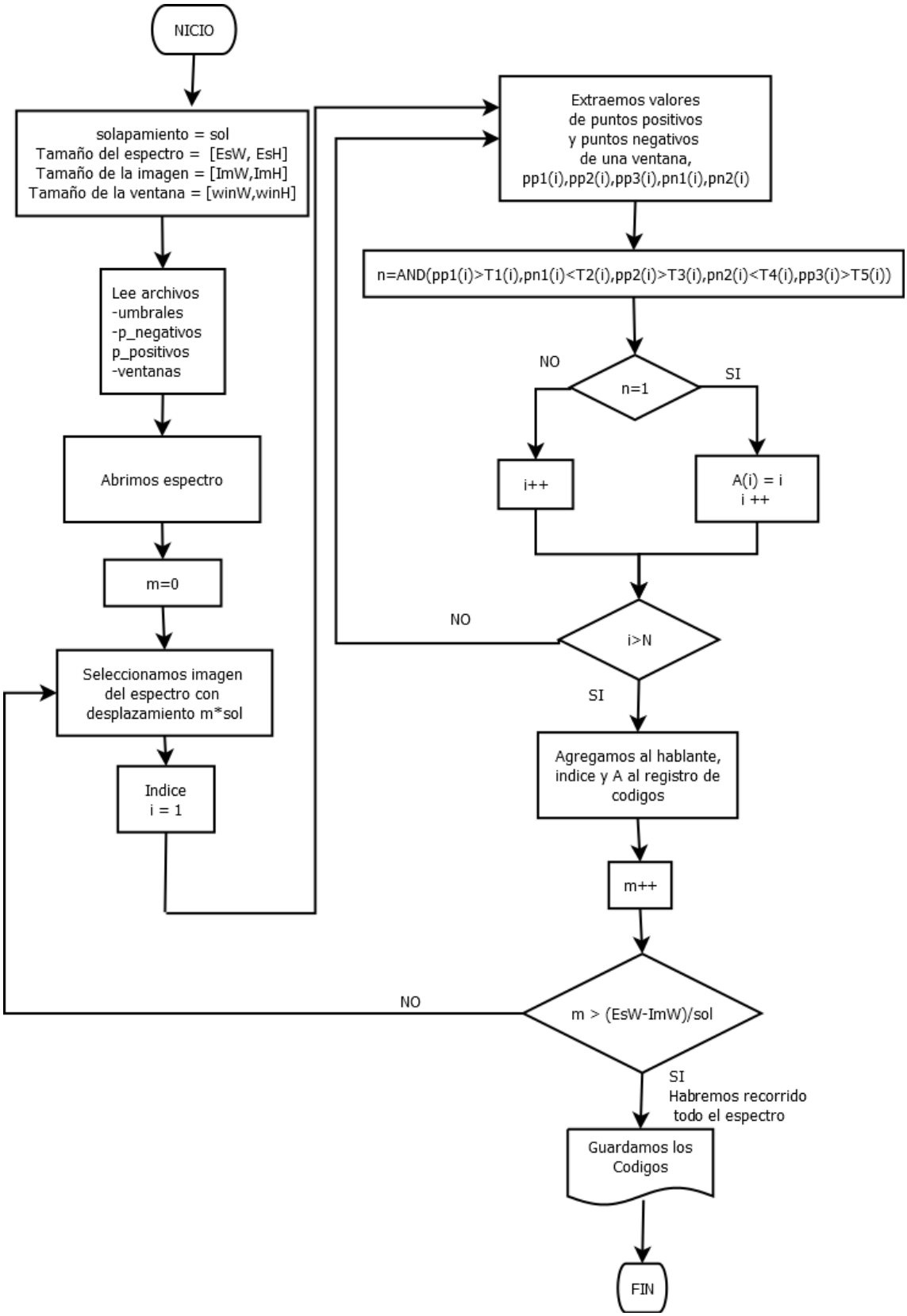


Fig. 5.7 Diagrama de flujo para la codificación de imágenes.

5.5.3 Entrenamiento

El método de entrenamiento es un proceso iterativo, en el cual se adaptan los pesos de las conexiones entre la capa de salida y la capa asociativa, para asegurar que la imagen mostrada pertenece a su clase real. Se usa el procedimiento de aprendizaje con la regla de Hebb para modificar los pesos. La Fig. 5.8 muestra el diagrama de flujo usado para el entrenamiento. El proceso utiliza los códigos obtenidos en la etapa anterior.

El seudocódigo, para el entrenamiento de un espectrograma es el siguiente:

1) Leemos los archivos, de códigos y de pesos. Códigos contiene una matriz (A) con las neuronas que fueron activadas en cada espectro y su hablante, y pesos contiene una matriz (w) de dimensión (13 x N), que es el número de hablantes totales y N, el número de neuronas totales.
2) Se inicializa el número de errores (error) en 0.
3) Se inicializa un índice auxiliar (i) que servirá para recorrer las filas de la matriz (A). Se asigna un valor de excitación adicional de 0.3 a Te.
4) Se lee el código contenido en la fila (i) de la matriz A, y su respectivo hablante (r).
5) El código leído contiene índices de las neuronas activadas. Se extraen los pesos correspondientes a estas neuronas de la matriz (w), de todos los hablantes. Y la matriz resultante se denota como (wa).
6) Se obtiene la suma (y) por hablante de (wa).
7) A la suma del hablante real (y(r)) se le aplica el factor de excitación adicional.
8) Se selecciona un ganador (g). El ganador es el hablante cuya suma sea la mayor.
9) Si el ganador (g), es el mismo que el hablante leído del código (r). Se verifica que la suma no sea 0. Si la suma es 0, a todos los pesos (wa) del hablante (r) se le suma 1, se incrementa el error en 1..
10) Si el ganador (g), es diferente a (r). Se le suma 1 a todos los pesos (wa) de (r), y se les restan 2 a los pesos (wa) de los demás hablantes, y se incrementa en el error en 1.
11) El índice se incrementa en 1, y se los pesos de (w) se actualizan con los pesos de (wa).
12) El proceso se repite hasta recorrer la matriz de códigos (A).
13) El error se almacena en un archivo.

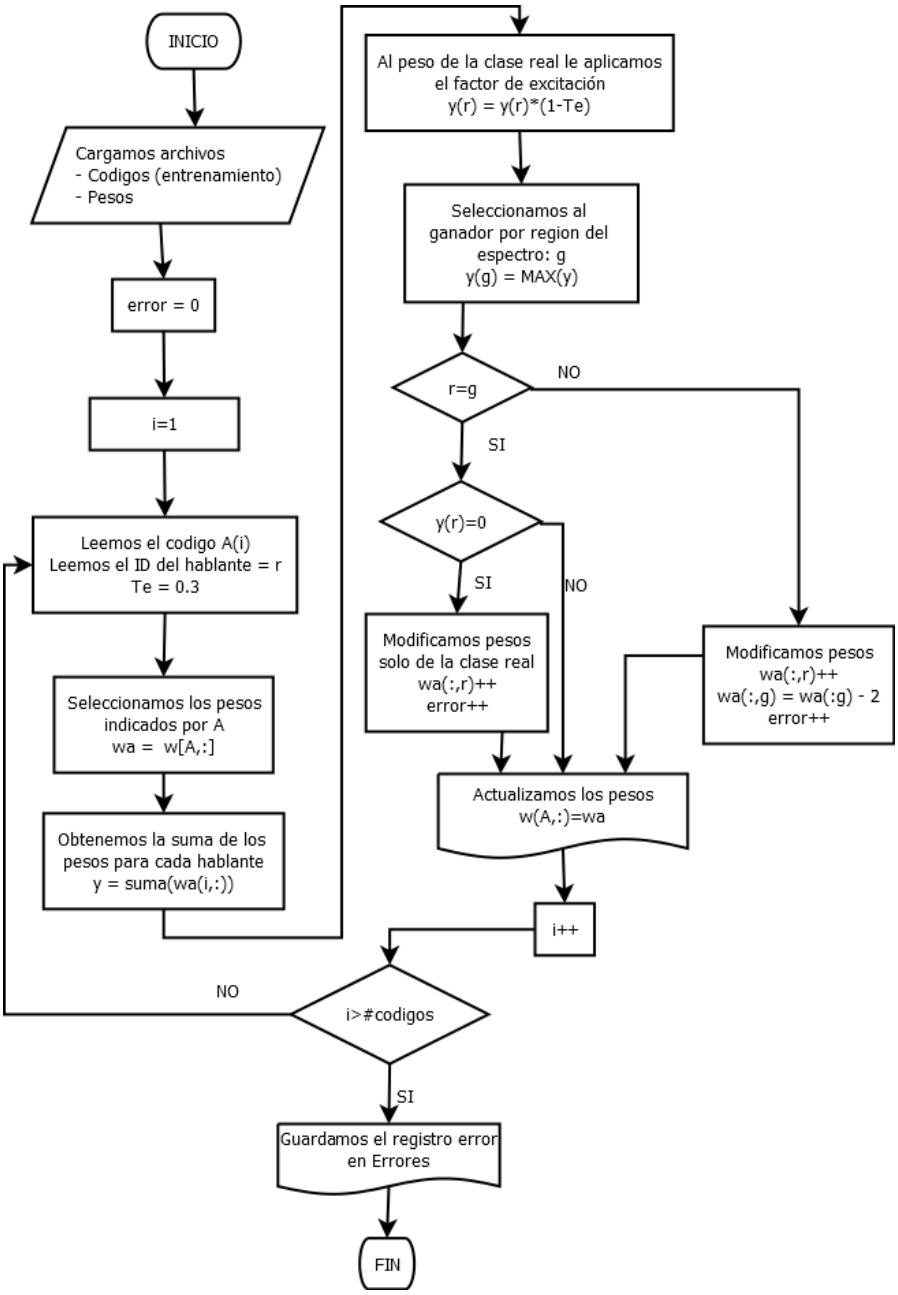


Fig. 5.8 Diagrama de flujo para el entrenamiento del clasificador.

5.5.4 Reconocimiento

El último proceso es el de reconocimiento, y es el más rápido. En el solo se presenta los códigos de un espectro, y el clasificador determina a que hablante pertenece. Se puede incorporar un ciclo para determinar de una sola ejecución la clasificación de todos los espectros. El porcentaje de error está determinado por los espectros clasificados que no correspondan con su clase original. La Fig. 5.9 muestra el diagrama de flujo para el reconocimiento de un espectro.

El pseudocodigo es el siguiente:

1) Se leen los códigos correspondientes al reconocimiento de un espectro (A) y la matriz de peso (w).
2) Se inicializa un índice que servirá para recorrer la matriz (A).
3) Se lee el código determinado por (i) de la matriz (A), y su correspondiente hablante (R).
4) Se extraen los pesos determinados por A(i) de la matriz (w), de todos los hablantes. Y la matriz resultante se denota como (wa).
5) Se obtiene la suma (y) por hablante de (wa).
6) Se selecciona el ganador (g) cuya suma sea la mayor. Este ganador se almacena en un vector (GG). Y finalmente se incrementa el índice (i).
7) El proceso se repite hasta leer todos los códigos de (A).
8) Del vector (GG) se obtiene un único ganador (G) del espectrograma mediante la moda de este vector.
9) Se compara el ganador (G) con el leído de los códigos (R).
10) Si son iguales se marca como acierto. Y se guardan.
11) Si son diferentes se marca como error. Y se guardan.

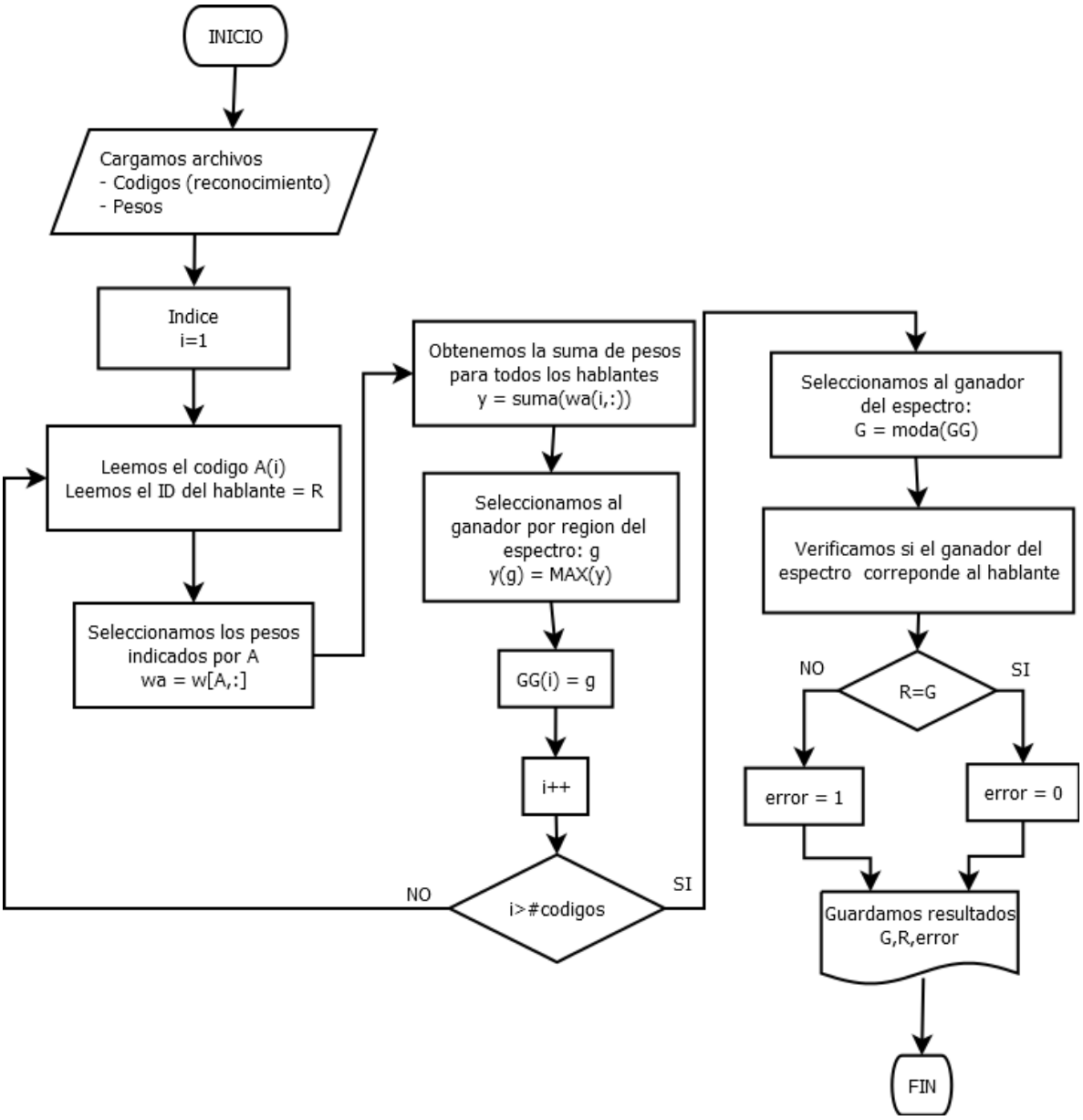


Fig. 5.9 Diagrama de flujo para el reconocimiento de un espectro.

Adicionalmente, además del código escrito para el clasificador, se creó código para obtener el reconocimiento de todos los espectros destinados para este fin dentro de un ciclo, y otro para visualizar la evolución de los errores durante el entrenamiento usan una gráfica que muestre el número de errores contra el numero de ciclos de entrenamiento.

CAPÍTULO 6

PRUEBAS Y RESULTADOS

6.1 Resultados

De manera general este trabajo el resultado de la aplicación de metodologías de diferentes disciplinas. La Física proporciono los conceptos necesarios para tratar el sonido de la voz en forma analógica, la Fisiología sirvió para comprender como se genera la voz en el tracto vocal, la computación y el tratamiento de señales sirvieron para realizar una conversión de datos y extraer información de importancia de la voz; finalmente, la Inteligencia Artificial proporciono los métodos adecuados para el reconocimiento de la señal y su debida clasificación.

Se realizaron 4 diferentes experimentos diferentes, de los cuales 3 de los experimentos se realizaron una vez y el que obtuvo mejores resultados se hizo 4 veces, variando los espectros de entrenamiento y de reconocimiento.

A continuación, los detalles por cada experimento.

6.1.1 Experimento 1

La tabla 6.1 contiene los parámetros usados en el experimento 1.

Tabla 6.1. Parámetros del clasificador neuronal.

Numero De capas de entrada	1
Número de capas intermedias	2
Numero de capas de salida	1
Número de neuronas asociativas, N	64000
Número de neuronas positivas por ventana, ON	3
Número de neuronas negativas por ventana, OFF	2
Tamaño de las ventanas	20×20 pixeles.
Número de espectros por hablante	15
Clases de salida	13
Solapamiento	50%
Altura de la imagen	64 pixeles
Relación de espectros para entrenamiento-reconocimiento	80%-20%.

Sin embargo para este experimento no fue posible obtener un entrenamiento exitoso del clasificador neuronal, por lo que se modificaron algunos parámetros de la Tabla 6.1. Debido a esto, no se realizó la etapa de reconocimiento

6.1.2 Experimento 2

El experimento 2 fue el que obtuvo mejores resultados en la etapa de reconocimiento, en este se obtuvo porcentaje de errores igual a 0% durante el entrenamiento en todas las pruebas realizadas. Pudiendo aplicar las matrices de pesos obtenidas en las etapas de reconocimiento.

Los parámetros usados para la implementación son los mismos de la Tabla 6.1, Con la siguiente modificación:

a) Solapamiento: 50%
b) Altura de la imagen: 500 pixeles.

Para este experimento por ser el que mejores resultados tuvo en rendimiento, reconocimiento y entrenamiento se dividió en 5 diferentes pruebas. En cada prueba se seleccionaron los espectros de entrenamiento y de reconocimiento de la manera mostrada en la tabla 6.2 se muestra el número de los espectros por cada hablante usados en el reconocimiento, los restantes fueron usados para entrenamiento.

Tabla 6.2 Espectros empleados en cada prueba para reconocimiento.

Prueba	ID de Espectros
2a	1,2,11
2b	3,4,12
2c	5,6,13
2d	7,8,14
2e	9,10,15

En la Fig. 6.1, Fig. 6.2, Fig. 6.3, Fig. 6.4 y Fig. 6.5, se muestran las curvas de error para cada una de las pruebas, en todas ellas se puede observar que el máximo error es aproximadamente del 18%, y que todas convergen a 0%. El entrenamiento termina cuando se obtienen cero errores durante un ciclo, las curvas de error solo muestran hasta el último ciclo de entrenamiento. En las curvas de errores acumulados, no se aprecia un porcentaje de 0, porque la curva representa en promedio de errores durante todo el entrenamiento.

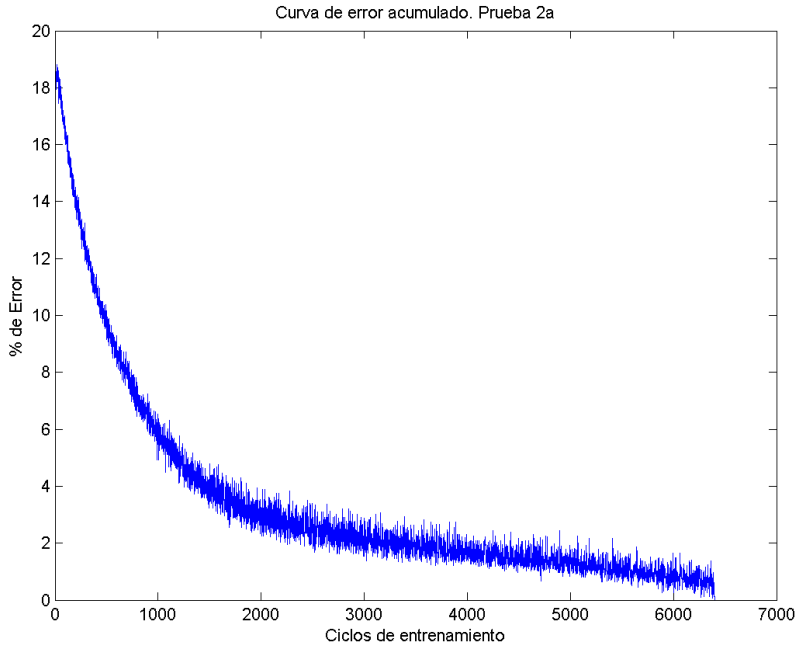


Fig. 6.1. Curva de errores durante el entrenamiento, para la prueba 2a.

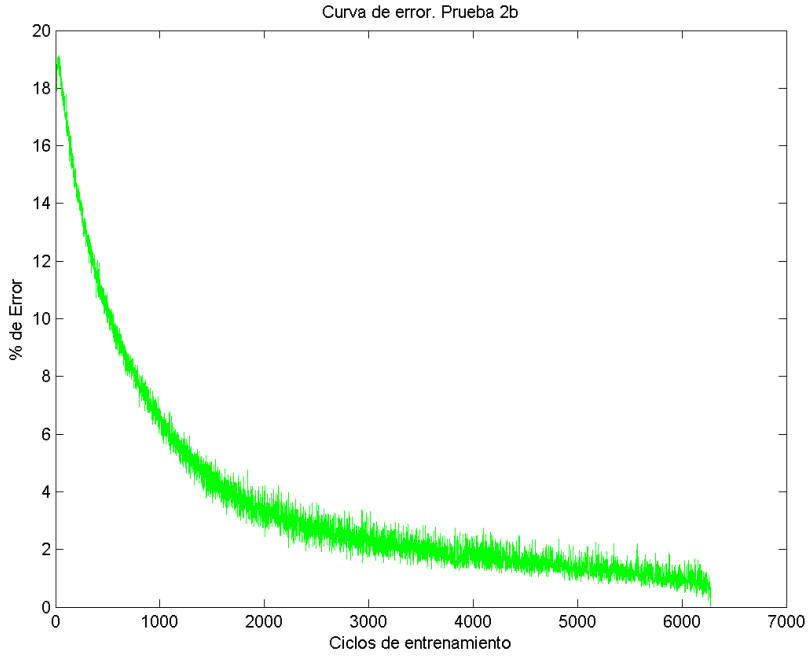


Fig. 6.2. Curva de errores durante el entrenamiento, para la prueba 2b.

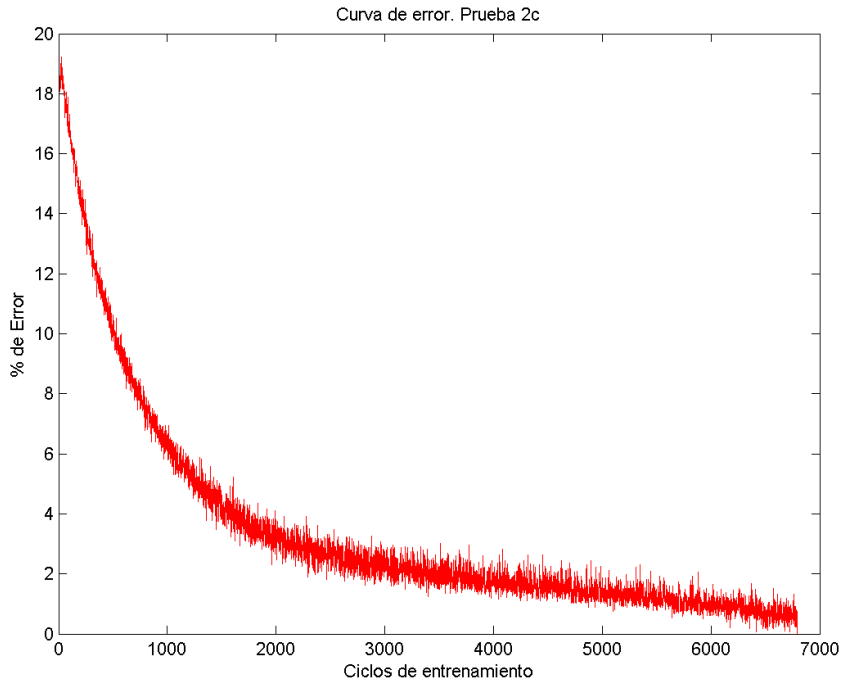


Fig. 6.3. Curva de errores durante el entrenamiento, para la prueba 2c.

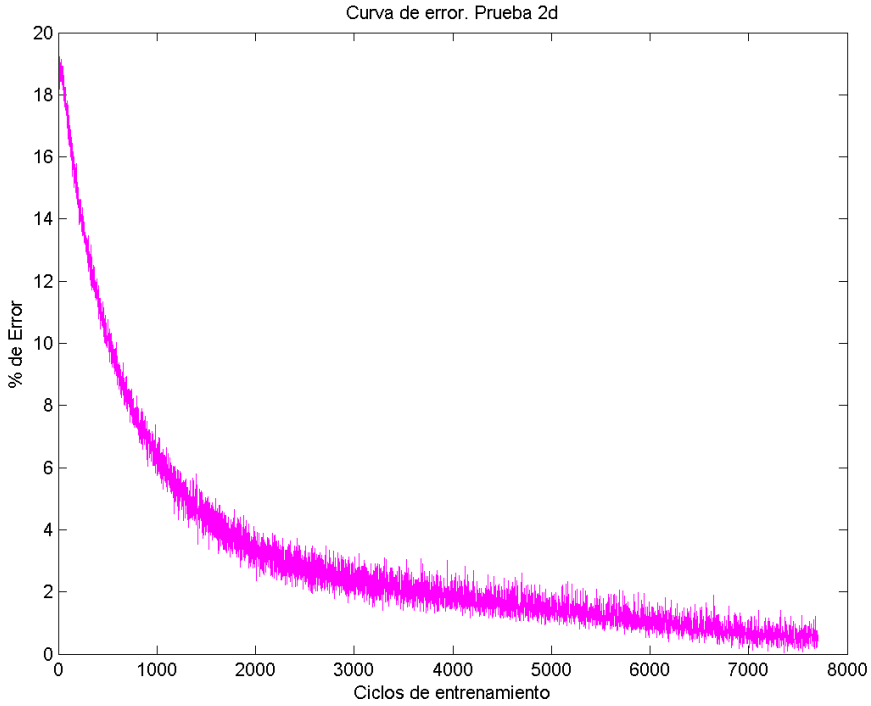


Fig. 6.4. Curva de errores durante el entrenamiento, para la prueba 2d.

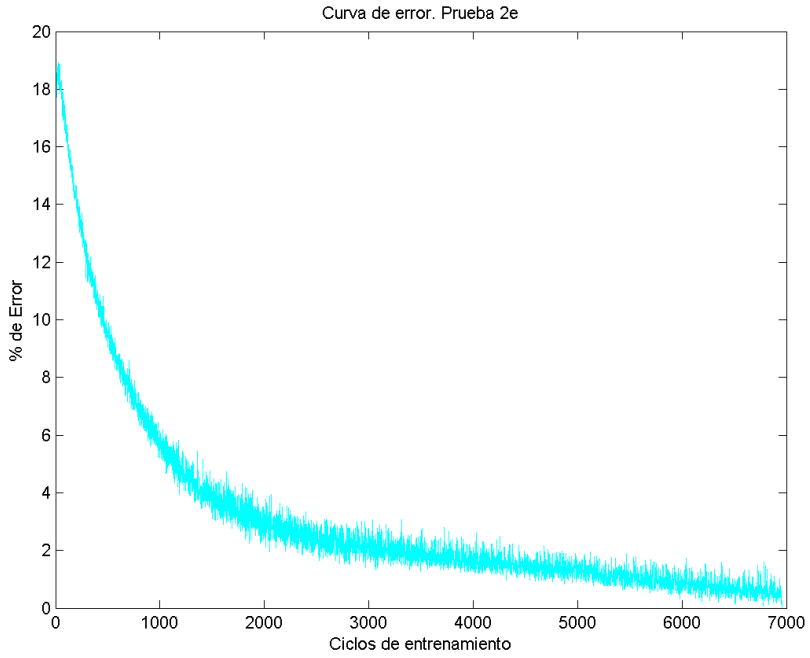


Fig. 6.5. Curva de errores durante el entrenamiento, para la prueba 2e.

Las curvas suavizadas se muestran en las Fig. 6.6, Fig. 6.7, Fig. 6.8, Fig. 6.9, Fig. 6.10. En ellas se muestran el porcentaje de error acumulado en los ciclos de entrenamiento.

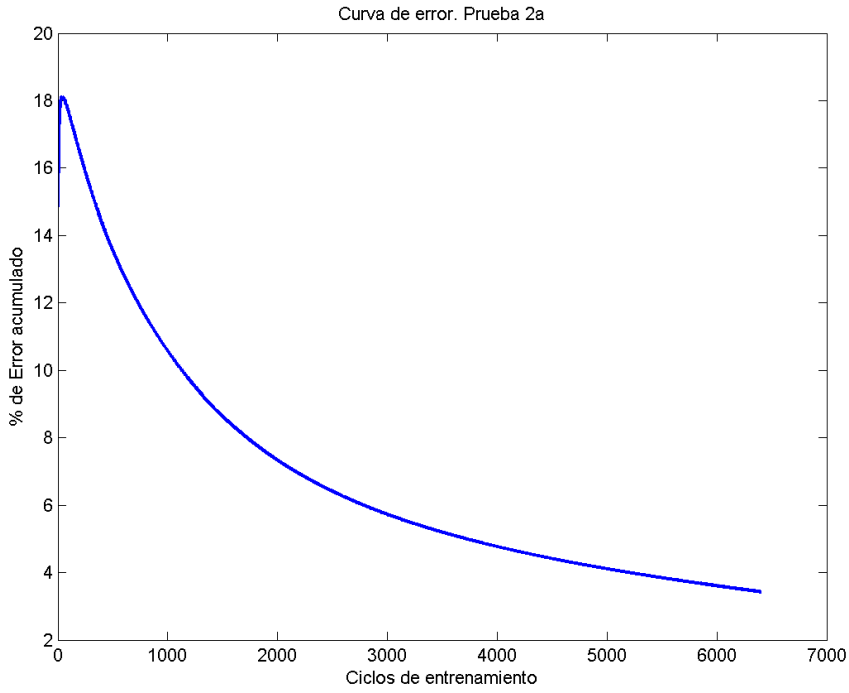


Fig. 6.6. Curva de errores acumulado durante el entrenamiento, para la prueba 2a.

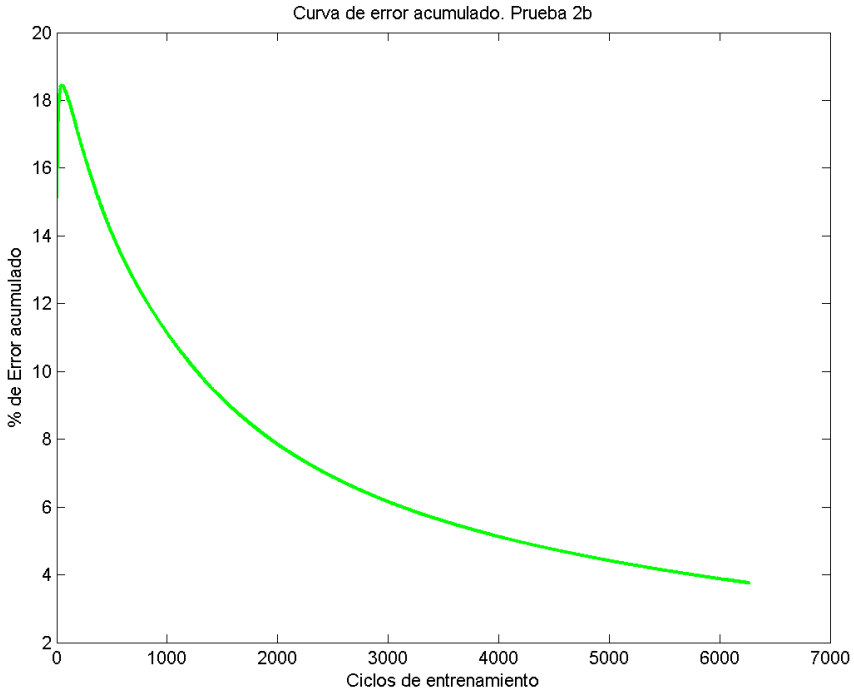


Fig. 6.7. Curva de errores acumulado durante el entrenamiento, para la prueba 2b.

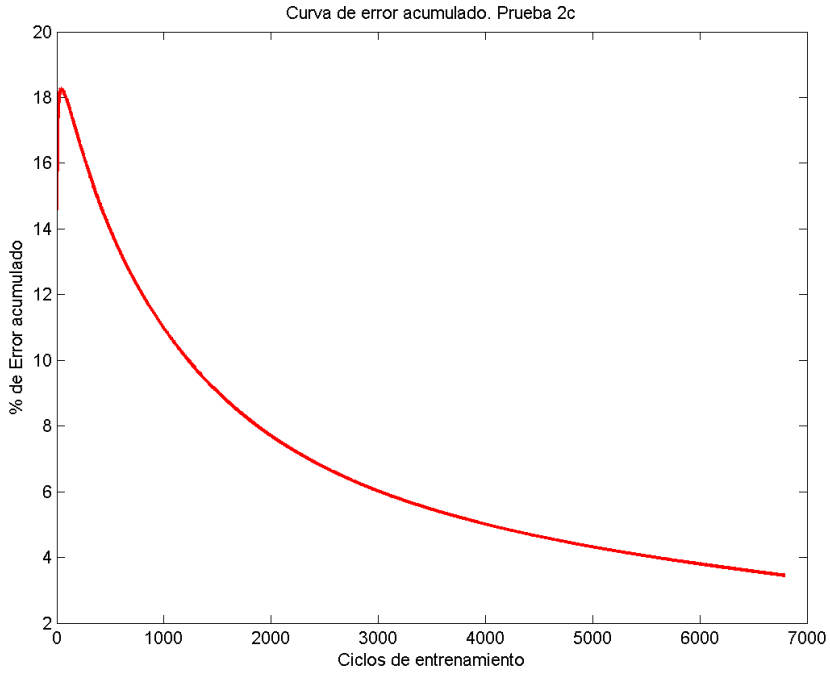


Fig. 6.8. Curva de errores acumulado durante el entrenamiento, para la prueba 2c.

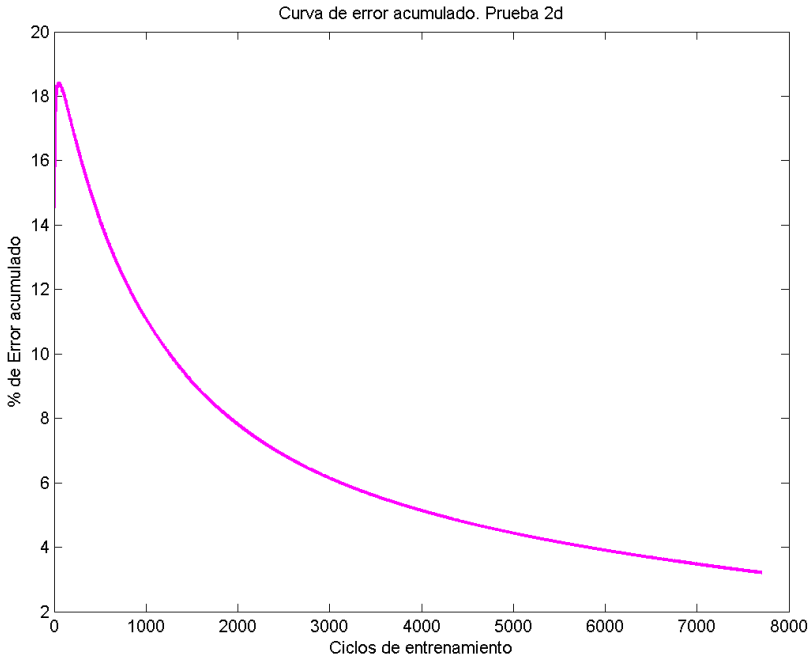


Fig. 6.9. Curva de errores acumulado durante el entrenamiento, para la prueba 2d.

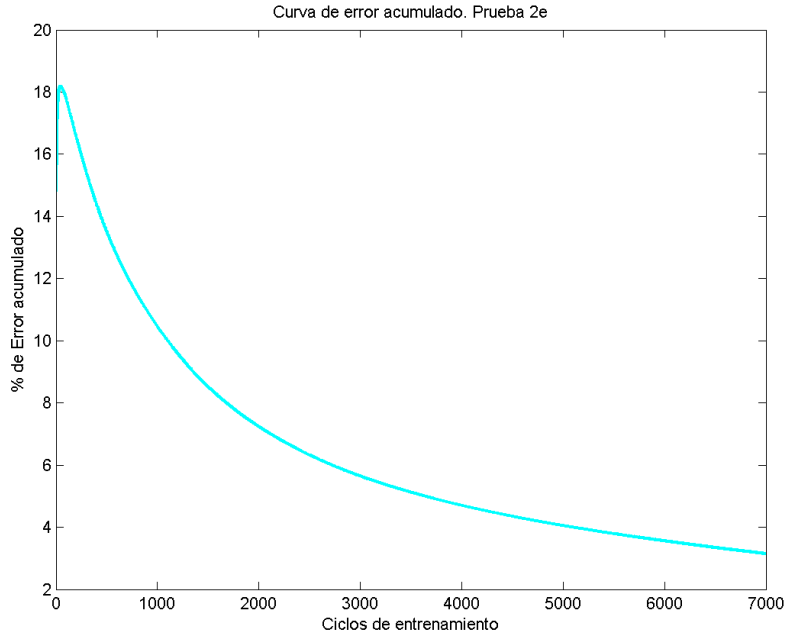


Fig. 6.10. Curva de errores acumulado durante el entrenamiento, para la prueba 2e.

En cada una de estas pruebas se realizó la etapa de reconocimiento. Los resultados obtenidos se encuentran el a tabla 6.3. La selección de la prueba 2b y 2d fueron las que obtuvieron y mejor porcentaje de reconocimiento, teniendo 4 errores por 39 espectrogramas evaluadas.

Para calcular el porcentaje de error, se hizo una comparación de la clasificación de los hablantes obtenidos por el sistema de reconocimiento contra los esperados. Cuando los valores del identificador obtenido contra el real no empatan, se considera un error. Para las pruebas del experimento 2 se obtuvieron tablas de confusión, Tablas 6.3.

Tabla 6.3 Tabla de confusión para las pruebas del experimento 2.

Tabla de confusion del experimento 2a

Hablantes reales \ Hablantes obtenido	1	2	3	4	5	6	7	8	9	10	11	12	13
1		0	0	0	0	0	0	0	0	0	0	0	0
2	0		0	0	0	0	0	0	0	0	0	0	0
3	0	0		0	0	0	0	0	0	0	0	0	0
4	0	0	0		0	0	0	0	0	0	0	0	0
5	0	0	0	0		0	0	0	0	0	0	0	0
6	0	0	0	0	0		2	0	0	0	0	0	0
7	0	0	0	0	0	1		0	0	0	0	0	0
8	0	0	0	0	0	0	0		0	0	0	0	0
9	0	0	0	0	0	0	0	0		0	0	0	0
10	0	0	0	0	0	0	0	0	0		0	0	0
11	0	0	0	0	0	0	0	0	1	0		0	1
12	0	0	0	0	0	0	0	0	0	0	0		0
13	0	0	0	0	0	0	0	0	0	0	0	0	

Tabla de confusion del experimento 2b

Hablantes reales \ Hablantes obtenido	1	2	3	4	5	6	7	8	9	10	11	12	13
1		0	0	0	0	0	0	0	0	0	0	0	0
2	0		0	0	0	0	0	0	0	0	0	0	0
3	0	0		0	0	0	0	0	0	0	0	0	0
4	0	0	0		0	0	0	0	0	0	0	0	0
5	0	0	0	0		0	0	0	0	0	0	0	0
6	0	0	0	0	0		0	0	0	1	0	0	0
7	0	0	0	0	0	1		0	0	0	0	0	0
8	0	0	0	0	0	0	0		0	0	0	0	0
9	0	0	0	0	0	0	0	0		0	0	0	0
10	0	0	0	0	0	1	0	0	0		0	0	0
11	0	0	0	0	0	0	0	0	0	0		0	0
12	0	0	0	0	0	0	0	0	0	0	0		0
13	0	0	0	0	0	0	0	0	0	0	0	0	

Tabla de confusion del experimento 2c

Hablantes reales \ Hablantes obtenido	1	2	3	4	5	6	7	8	9	10	11	12	13
1		0	0	0	0	0	0	0	0	0	0	0	0
2	0		0	0	0	0	0	0	0	0	0	0	0
3	0	0		0	0	0	0	0	0	0	0	0	0
4	0	0	0		0	0	0	0	0	0	0	0	0
5	0	0	0	0		0	0	0	0	0	0	0	0
6	0	0	0	0	0		1	0	0	0	0	0	0
7	0	0	0	0	0	1		0	0	1	0	0	0
8	0	0	0	0	0	0	0		0	0	0	0	0
9	0	0	0	0	0	0	0	0		0	0	0	1
10	0	0	0	0	0	0	0	0	0		0	0	0
11	0	0	0	0	0	0	0	0	0	0		0	0
12	0	0	0	0	0	0	0	0	0	0	0		0
13	0	0	0	0	0	0	0	0	0	0	0	0	

Tabla de confusion del experimento 2d

Hablantes reales \ Hablantes obtenido	1	2	3	4	5	6	7	8	9	10	11	12	13
1		0	0	0	0	0	0	0	0	0	0	0	0
2	0		0	0	0	0	0	0	0	0	0	0	0
3	0	0		0	0	0	0	0	0	0	0	0	0
4	0	0	0		0	0	0	0	0	0	0	0	0
5	0	0	0	0		0	0	0	0	0	0	0	0
6	0	0	0	0	0		1	0	0	1	0	0	0
7	0	0	0	0	0	0		0	0	0	0	0	0
8	0	0	0	0	0	0	0		0	0	0	0	0
9	0	0	0	0	0	0	0	0		0	0	0	0
10	0	0	0	0	0	0	0	0	0		0	0	0
11	0	0	0	0	0	0	0	0	0	0		0	1
12	0	0	0	0	0	0	0	0	0	0	0		0
13	0	0	0	0	0	0	0	0	0	0	0	0	

Tabla de confusion del experimento 2e

Hablantes reales \ Hablantes obtenido	1	2	3	4	5	6	7	8	9	10	11	12	13
1		0	0	0	0	0	0	0	0	0	0	0	0
2	0		0	0	0	0	0	0	2	0	0	0	0
3	0	0		0	0	0	0	0	0	0	0	0	0
4	0	0	1		0	0	0	0	0	0	0	0	0
5	0	0	0	0		0	0	0	0	0	0	0	0
6	0	0	0	0	0		0	0	0	0	0	0	0
7	0	0	0	0	0	0		0	0	1	0	0	0
8	0	0	0	0	0	0	0		0	0	0	0	0
9	0	0	0	0	0	0	0	0		0	0	0	1
10	0	0	0	0	0	0	0	0	0		0	0	0
11	0	0	0	0	0	0	0	0	1	0		0	0
12	0	0	0	0	0	0	0	0	0	0	0		0
13	0	0	0	0	0	0	0	0	0	0	0	0	

El porcentaje calculado se obtiene sumando el total de confusiones entre el número total de frases de reconocimiento (39), y estos se resumen en la Tabla 6.4.

Tabla 6.4. Resultados de reconocimiento para el experimento 2.

Prueba	% de error
2a	12.82
2b	7.69
2c	10.25
2d	7.69
2e	17.95

Los errores obtenidos fueron para diferentes hablantes, se descarta un error por voz parecida.

6.1.3 Experimento 3

Para mejorar la eficiencia del clasificador se aumentó el solapamiento de imágenes, con el objetivo de tener más información durante el entrenamiento, sin embargo, con esta modificación se disminuyó el rendimiento del clasificador, porque al tener una mayor numero de imágenes para procesar el tiempo durante la codificación, el entrenamiento y reconocimiento aumento de manera lineal al número de imágenes procesadas.

La Fig. 6.11, muestra la curva de error durante el entrenamiento, en ella se puede observar que no fue posible obtener 0 errores inclusive aumentando el número de ciclos. El mayor porcentaje fue de, se tomaron los mismos parámetros del experimento anterior, cambiando solo el solapamiento entre imágenes. La Fig. 6.12, es la curva de error acumulado para todos los ciclos realizados.

Los parámetros usados para la implementación son los mismos de la Tabla 6.1, Con la siguiente modificación:

a) Solapamiento: 75%
b) Altura de la imagen: 64 pixeles.

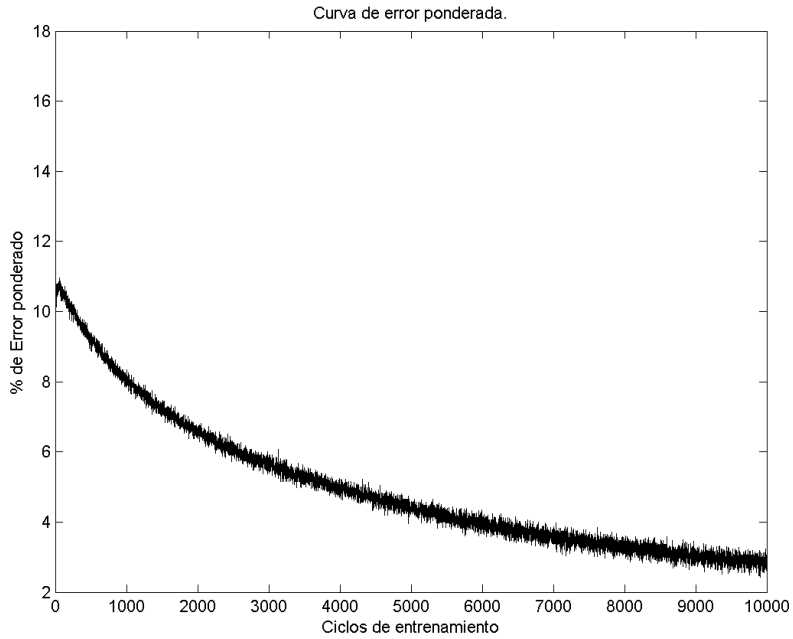


Fig. 6.11. Curva de errores durante el entrenamiento, para el experimento 3.

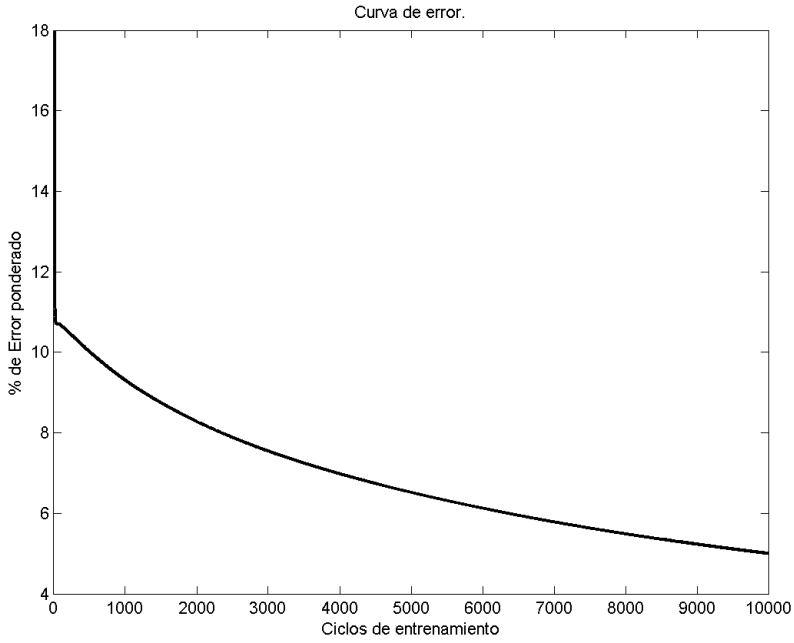


Fig. 6.12. Curva de errores acumulado durante el entrenamiento, para el experimento 3.

A pesar de que no se obtuvo un número de errores iguales a cero de realizó la etapa de reconocimiento. La cual no dio mejores resultados a los anteriores.

Se obtuvo un porcentaje de reconocimiento del 28.20%.

6.1.4 Experimento 4

Para este experimento se usaron espectro de 500 pixeles de alto, solapamiento del 75% y los demás parámetros de la Tabla 6.1, fue el último experimento realizado. Los resultados obtenidos no fueron mejores que el anterior, pero con ello se descartó la idea de que se podía mejorar la eficacia del clasificador aumentando el solapamiento. La curva de error durante el entrenamiento y la curva de error acumulada se muestran en la Fig. 6.13 y 6.14, respectivamente.

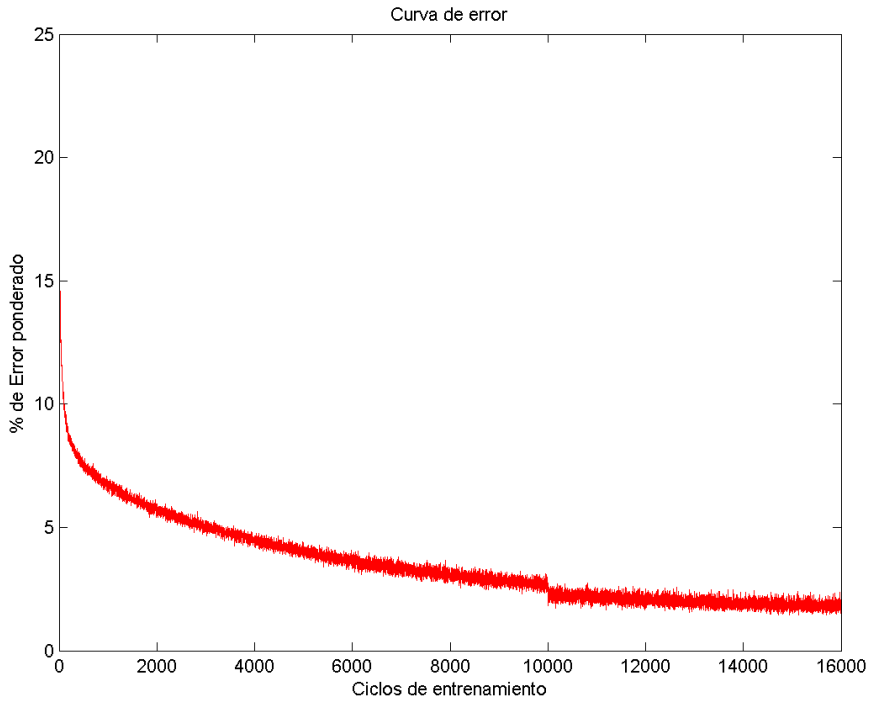


Fig. 6.13. Curva de errores durante el entrenamiento, para el experimento 4.

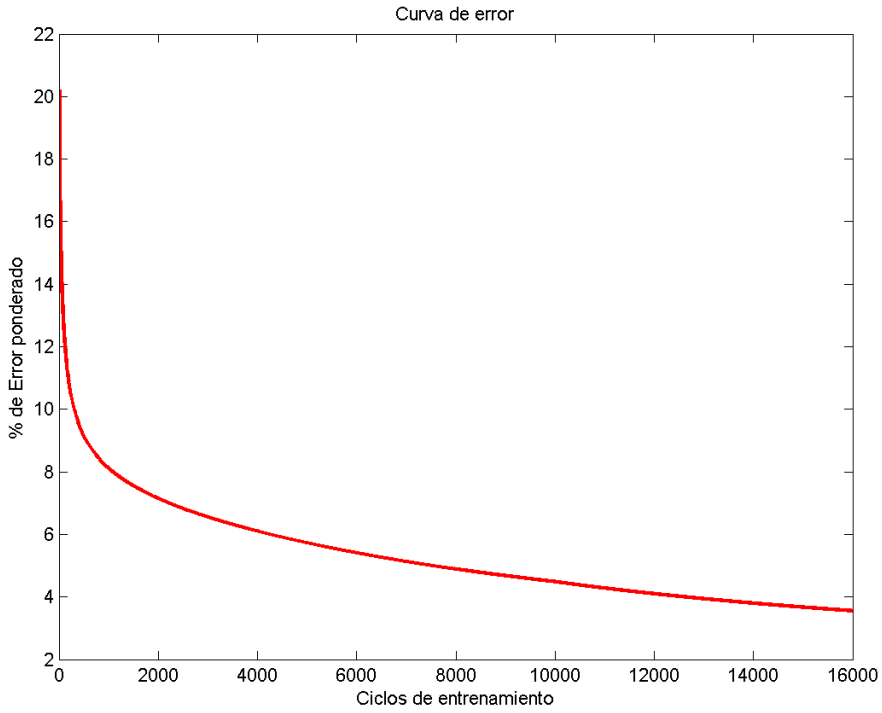


Fig. 6.14. Curva de errores acumulado durante el entrenamiento, para el experimento 4.

El porcentaje de reconocimiento fue mejor, pero no superó el obtenido en las pruebas del experimento 2. El reconocimiento fue del 25.64%.

6.1.5 Resumen de resultados

La tabla 6.5 contienen los resultados obtenidos para cada experimento, y las tablas de confusión para el experimento 3 y el 4 en la tabla 6.6 y 6.7 respectivamente:

Tabla 6.5 Resultados de experimentos con el clasificador.

Experimento	Altura	Solapamiento	% De error
1	ImE=64 pixeles	50%	100%
2	ImE=500 pixeles	50%	7.69%
3	ImE=64 pixeles	25%	28.20%
4	ImE=500 pixeles	25%	25.64%

En la tabla se observa que el mejor resultado se obtuvo con una parametrización del experimento 2.

Tabla 6.6 Tabla de confusion del experimento 3

Hablantes reales \ Hablantes obtenido	1	2	3	4	5	6	7	8	9	10	11	12	13
1		0	0	0	0	0	0	0	0	1	0	0	0
2	0		0	0	0	0	0	0	1	0	0	0	0
3	0	0		0	0	0	0	0	0	0	1	0	0
4	0	0	0		0	0	0	0	0	0	0	0	0
5	0	0	0	0		0	0	0	0	0	0	0	0
6	0	0	0	0	0		0	0	0	3	0	0	0
7	0	0	0	0	0	0		0	0	3	0	0	0
8	0	0	0	0	0	0	0		0	0	0	0	0
9	0	0	0	0	0	0	0	0		0	0	0	1
10	0	0	0	0	0	0	0	0	0		0	0	0
11	0	0	0	0	0	0	0	0	0	0		0	0
12	0	0	0	0	0	0	0	0	0	0	0		0
13	0	0	0	0	0	0	0	0	0	0	0	0	

Tabla 6.7 Tabla de confusion del experimento 4

Hablantes reales \ Hablantes obtenido	1	2	3	4	5	6	7	8	9	10	11	12	13
1		0	0	0	0	0	0	0	0	0	0	0	0
2	0		0	0	0	0	0	0	0	0	0	0	0
3	0	0		0	0	0	0	0	0	0	1	0	0
4	0	0	1		0	0	0	0	0	0	0	0	0
5	0	0	0	0		0	0	0	0	0	0	0	0
6	0	0	0	0	0		0	0	0	3	0	0	0
7	0	0	0	0	0	0		0	0	3	0	0	0
8	0	0	0	0	0	0	0		0	0	0	0	0
9	0	0	0	0	0	0	0	0		0	0	0	1
10	0	0	0	0	0	0	0	0	0		0	0	0
11	0	0	0	0	0	0	0	0	0	0		0	0
12	0	0	0	0	0	0	0	0	0	0	0		1
13	0	0	0	0	0	0	0	0	0	0	0	0	

Además, se realizó un experimento extra, el cual se hacía con las mismas frases usadas durante el entrenamiento pero dichas 3 meses después de las usadas para el entrenamiento. En este caso solo se emplearon dos hablantes y 12 frases por hablante, se obtuvo un reconocimiento del 100% empleando la matriz de pesos del experimento 2.

6.2 Interfaz para el reconocimiento.

Con el fin de crear una aplicación utilizable se creó una interfaz de usuario. La Fig. 6.17 contiene el recuadro de la interfaz de usuario. Consta de las siguientes partes:

- Un listbox para seleccionar el espectro que se analizará.
- Un button para iniciar el proceso de reconocimiento.
- Una imagen en la que se muestra el espectro que se está evaluando.
- Una ventana del espectro que se reconocerá de manera individual.
- Una etiqueta que muestra al hablante de la ventana individual
- Una etiqueta que muestra al hablante obtenido de la moda de los hablantes reconocidos en cada ventana. Este es el resultado final y contiene al hablante identificado.

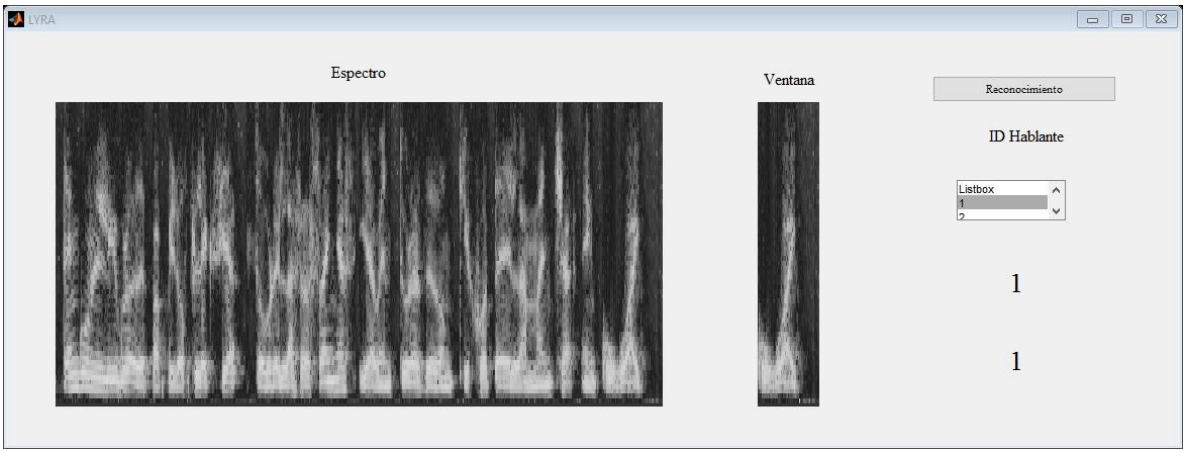


Fig. 6.15. Interfaz de usuario para el reconocimiento

Esta interfaz utiliza la matriz de pesos obtenida por el entrenamiento y los códigos de los espectros que evaluados.

6.3 Interpretación de resultados

Los resultados obtenidos fueron aceptables debido al tiempo dedicado a probar distintos parámetros de configuración de la red neuronal. Sin embargo, es necesario aclarar la metodología empleada para lograr el progreso en resultados, y con lo cual queda abierta la puerta para continuar con la experimentación.

Se empleó la metodología de los algoritmos genéticos. Con las siguientes consideraciones

- Los parámetros de la Tabla 6.1 empleados para combinar y sus conjuntos de valores posibles fueron la altura del espectrograma (64 pixeles; 500 pixeles), el solapamiento (25%, 50%), y el conjunto de muestras para entrenamiento y reconocimiento, definidos en la Tabla 6.2.

- Solo se consideraron dos iteraciones, definiendo 4 configuraciones distintas cada una.
- En la primera iteración se combinaron la altura del espectrograma y el solapamiento. De estas se eligió la mejor, y en la segunda iteración se combinó con los distintos conjuntos de muestras de entrenamiento y reconocimiento.
- No hubo mutación en los parámetros.

De esta manera fue el avance en la investigación. Aunque los resultados no fueron los esperados, sirvieron para buscar nuevas formas de mejorar al sistema.

Las características que pueden ser mejoradas son las siguientes:

- Agregar una clase para hablantes no registrados, este problema fue considerado posterior a la implementación y pruebas, y no pudo ser corregido a tiempo.
- No normalizar los espectrogramas en el tiempo. Al crear los espectrogramas, estos tenían una longitud distinta en el eje de tiempo, pero al convertirlos a imagen su longitud fue reducida a 1000 pixeles para todos, comprimiendo la información. Esto se hizo para mejorar el control en la codificación de las características, pero es posible omitir la normalización.
- Aumentar el tamaño de la base datos de voz con más frases y con más hablantes, tomadas en distintos tiempos.
- Mejorar la calidad del espectrograma, para tener una mejor resolución de los cambios de la voz.
- Para el proceso de experimentación, además de los parámetros investigados modificar el número de neuronas totales del clasificador, el número de neuronas ON y OFF, investigara para más valores de solapamiento, modificar el tamaño de la ventana. Que son parámetros importantes para identificar el patrón de voz de un individuo.

CAPÍTULO 7

CONCLUSIONES

Para resolver la tarea de reconocimiento del hablante se preparó una base de voces. Del proceso de selección de las frases dichas por cada hablante fue indispensable usar aquellas que contuvieran información fonética, para distinguir a los diferentes hablantes. Se grabaron a 13 personas, cada persona pronunció 15 frases, de las cuales 10 fueron en español y 5 en inglés. Las personas que participaron fueron hombres y mujeres, de entre 20 y 40 años. Las frases obtenidas se guardaron en el formato estándar WAV, este formato es usado en sistemas de tratamiento de señales de voz. A todas las muestras de voz fueron filtradas para obtener sus correspondientes espectrogramas, estos espectrogramas fueron usados como imágenes de entrada para el clasificador neuronal. Posteriormente los espectrogramas obtenidos fueron transformados a imágenes de formato BMP. Este formato nos asegura que no haya perdida de información por compresión.

Para distinguir a los hablantes se empleó el clasificador neuronal LIRA. Este clasificador fue programado en Matlab y fue adaptado para la tarea de clasificación de hablantes. Cada imagen de espectrograma fue escaneada con diferentes tamaños de ventanas. Para el desarrollo del código, lo más importante fue diseñar las subrutinas de forma que no consumieran mucho tiempo y memoria del procesador. Para eso se aislaron en subprocesos las diferentes etapas del clasificador, de esta manera se separó en tres partes: codificación, entrenamiento y reconocimiento.

Con el código fuente separado, también fue posible medir el tiempo de procesador consumido en cada etapa: la codificación requiere un 20%, el entrenamiento un 78% y el reconocimiento un 2%.

Se obtuvieron buenos resultados en los procesos de entrenamiento. Después de los 700 ciclos se obtuvieron 0 errores. Y en el proceso de reconocimiento se obtuvo como mejor resultado, un 7.69% de errores de identificación.

La elección del ventaneo fue lo más complicado, segmentos muy cortos aumentan considerablemente el tiempo de codificación y de entrenamiento, y dan peores resultados durante el reconocimiento.

Finalmente, para el reconocimiento, la identificación del hablante no tuvo errores, demostrando que el algoritmo es adaptable para las tareas de reconocimiento.

A pesar de estar concluido el trabajo, aun se puede mejorar el método. Algunas de las mejoras que se pueden realizar son refinar los espectrogramas, mejorar la calidad de la grabación de los hablantes, realizar diferentes grabaciones en diferentes momentos o con alguna alteración, como la telefónica. Mejoras en la programación son optimización de código con cómputo paralelo, implementar un algoritmo genético para obtener la configuración óptima de los parámetros para el clasificador.

A pesar de que se emplearon pocos sujetos de prueba como hablantes es posible agregar nuevos elementos, solo debe ejecutarse nuevamente el proceso de entrenamiento con estos elementos, y el proceso de reconocimiento no se ve alterado. Y el tiempo se

incrementa de forma lineal al aumentar el número de hablantes. El reconocimiento puede ser probado para frases distintas a las empleadas en los experimentos, solo codificando el nuevo espectro.

8. REFERENCIAS

[1] Güimi. Historia de la computación, Recuperado el 1 Abril de2013 de guimi.net.

[2] García F. N. *Análisis e interpretación de patrones musicales.* México. UNAM. 2014.

[3] Knuth D. E. *The Art of Computer Programming: Volume 1. Fundamental Algorithms.* (2ª Ed.) Massachusetts: Addison-Wesley Publishing Company. 2002.

[4]Devijver P. R. y Kittler J. (1982), *Pattern Recognition: A Statistical Approach.* New Jersey: Prentice-Hall. 1982.

[5] Campbell J. P., *Speaker recognition: a tutorial*, Proc. IEEE, vol. 85, no. 9, pp. 1437-1462, 1997.

[7] Reynolds, D. A., A *Gaussian mixture modeling approach to text independent speaker identification*, Ph.D. Thesis, Georgia Institute of Technology, 1992.

[8] Pohlmann K., *Principles of digital audio.* (5ª Ed.) New York: McGraw-Hill, 2002.

[9] Serdi J. P., *Audio digital y MIDI***.** España: Anaya, 1997.

[10] Instituto Politécnico de Madrid. Escuela Universitaria Técnica de Telecomunicación, Manual Técnico de sonido**.** Recuperado el 08 Julio del 2013 de http://www.diac.upm.es/escuela, 2000.

[11] Miyara F. *Acústica y Sistemas de sonido*, Argentina: UNR, 1999.

[12] Arribas J. *Psicoacústica.* Recuperado el 12 de Julio de 2013 de http://www.lpi.tel.uva.es/~nacho/docencia/ing_ond_1.htm,2006.

[13] Torres J. L. *Educación Musical.* México: Porrúa, 1972.

[14] Fuentes, J. L., *Gramática moderna de la lengua española.* Bibliográfica Internacional. Madrid, España, 1988

[15] Fuente de internet: http://3.bp.blogspot.com/-wulpZ35yAc0/UyO7ItegW9I/AAAAAAAAAeo/k8wsjEE0wDc/s1600/fonador.gif. Recuperado el 19 de octubre de 2016.

[16] Fuente de internet http://elcuerpohumano.webcindario.com/imagenes/El%20oido.jpg. Recuperado el 06 de octubre de 2016.

[17] Martínez F. J. *Tutorial Web de Técnicas de Digitalización de Audio para la asignatura Tratamiento Digital de Audio*. España: Universidad Carlos III de Madrid, Departamento de Teoría de la Señal y Comunicaciones. Tesis de licenciatura en Ingeniería Técnica en Telecomunicaciones, 2009.

[18] Date C. J. *Introducción a los sistemas de bases de datos.* 7ª Ed. México: Pearson Educación, 2001.

[19] García F. N. *Análisis e interpretación de patrones musicales.* México. UNAM, 2004.

[20] Tipos de usuarios de bases de datos. http://2.bp.blogspot.com/-T6Pta6xkFEQ/TV70mLJ-QNI/AAAAAAAAAIY/u075W8siBmo/s1600/SIMPLE.gif. Recuperado el 11 de octubre de 2016.

[21] Del Brio B. M., Sanz. A.; Redes neuronales y sistemas borrosos. Alfaomega 3ª Ed. México: 2007.

[22] Hansen J. T.; *Netter, Cuaderno de anatomía.* Elsevier Masson, España, 2014.

[23] Baydyk T., Kussul E. *Redes neuronales, vision computacional y micromecánica.* Itaca, México, 2009.