Universidad de las Ciencias Informáticas

FACULTAD 6

Título: Reconocimiento de locutores para el Sistema de Gestión y Transmisión de Contenidos Audiovisuales.

Trabajo de Diploma para Optar por el Título de Ingeniero en Ciencias Informáticas

Autor: Mariemy Celia Pimienta Ortega

Tutor: Ing. Eduardo Cepero Utra

La Habana, 16 de junio de 2012

Año 54 de la Revolución

Frase

“El éxito es aprender a ir de fracaso en fracaso sin desesperarse.”

Winston Churchill

Político y hombre de estado británico. Primer Ministro del Reino Unido en dos períodos (1940-45 y 1951-55).

Declaración de Autoría

Declaramos ser autores de la presente tesis y reconocemos a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmo la presente a los ____ días del mes de _____ del año _____.

Mariemy Celia Pimienta Ortega

_____

Firma del Autor

Ing. Eduardo Cepero Utra

_____

Firma del Tutor

Datos de Contacto

Nombre y Apellidos: Eduardo Cepero Utra.

Sexo: M

Institución: Universidad de las Ciencias Informáticas.

Dirección de la institución: Carretera a San Antonio de los Baños, Km. 2 ½, Reparto: Torrens, Municipio: Boyeros, Provincia: La Habana.

Correo electrónico: eutra@uci.cu.

Teléfono del trabajo: -

Teléfono particular: -

Título de la especialidad de graduado: Ingeniero en Ciencias Informáticas.

Año de graduación: 2011.

Institución donde se graduó: Universidad de las Ciencias Informáticas.

Agradecimientos

A mis padres, por apoyarme y ayudarme a seguir adelante durante estos cinco años, por estar ahí cuando más los necesitaba, por estar en mis momentos de flaqueza: porque gracias a ellos soy la persona que soy hoy.

A mi hermano Frank, por guiarme y arrastrarme a este mundo tan lindo que es la Informática, por ser mi juez a la hora de tomar una decisión, por brindarme la felicidad de ser tía de una criaturita linda que está por venir.

A mis abuelos, Mima y Pipo, por darme la familia tan bella que tengo.

A mis tíos, por quererme y cuidarme no como una sobrina, sino, como una hija más.

A Amy y familia, por acogerme y permitirme entrar a sus vidas, por aconsejarme en los momentos difíciles y estar junto a mí en los momentos alegres.

A mi tutor Eduardo, por las veces que me devolvía el documento "lleno de sangre y envuelto en esparadrapo", por apoyarme y guiarme durante la mayor parte del proceso de la realización de la tesis.

A los profesores que conocí durante toda la carrera, por enseñarme lo que sé, por obligarme a ser más exigente conmigo misma.

A los amigos inseparables durante estos cinco años, a los que están y no están, a los nuevos amigos conocidos.

A las "Mimis" y a las "Pencs", por ser grandes amigas, por los momentos tan buenos que pasamos juntas, por las visitas al Mariel, por las cremitas de leche, por las yucas tan ricas que me comí, por ser tan ocurrentes y divertidas, las quiero mucho y no las voy a olvidar nunca.

A mis niños queridos: a mi hermanito Rojas, a Reinel, al Negro, a Lieter, a Fleitas, a Ulises, a Enrique; por ayudarme y soportarme durante todos estos años, por aumentar mi cultura general.

A todos gracias, y mil veces gracias…

Dedicatoria

A mi madre, por ser mi guía y luz.

A mi padre, por crearme la fortaleza y la confianza en mí misma.

A mi hermano, por brindarme sus consejos.

A mis abuelos, por quererme tanto.

A mi familia, por su apoyo incondicional.

A mis amigos, por brindarme ese tesoro que es la Amistad.

Resumen

El reconocimiento automático del locutor es una de las áreas de la biometría que más importancia ha adquirido en los últimos años. Pertenece a la rama de la Inteligencia Artificial y consiste en la identificación automática de las personas a través de su voz. El proceso de reconocimiento automático de locutores está ligado a las características fisiológicas y los hábitos lingüísticos de los mismos. El mismo conlleva a un procesado de audio que permite extraer las características del locutor y realizar una búsqueda que encuentre las posibles coincidencias mediante un proceso de reconocimiento de patrones. En el trabajo se presentan y analizan las principales técnicas de extracción y selección de rasgos acústicos, así como las de clasificación para el reconocimiento una persona a través de la voz. Para la extracción de características se utilizó el algoritmo MFCC (Mel Frequency Cepstral Coeficientes). La Cuantificación Vectorial (VQ) se utilizó para la clasificación de características y para el agrupamiento se utilizo el algoritmo LBG. Para la validación se creó un prototipo funcional, donde los resultados obtenidos se documentaron en el trabajo de investigación.

Palabras claves: Algoritmo, locutor, voz

Índice

Introducción 1
Capítulo 1. Fundamentación Teórica 5
1.1 Introducción. 5
1.2 Conceptos asociados al reconocimiento de locutores. 5
1.2.1 Reconocimiento Automático de Locutor (RAL) 5
1.2.2 Identificación Automática de Locutores 6
1.2.3 Verificación Automática de Locutores 6
1.3 Concepción del reconocimiento de locutores. 7
1.4 Situación Problemática. 8
1.5 Análisis de soluciones existentes 9
1.5.1 Verbio Speaker ID 9
1.5.2 IDENTIVOX 10
1.5.3 Métodos de extracción de rasgos acústicos del habla 11
1.5.4 Métodos de selección de rasgos 16
1.5.5 Métodos de clasificación 17
1.6 Conclusiones 24
Capítulo 2. Caracterización de la propuesta de solución 25
2.1 Introducción 25
2.2 Extracción de características del habla 25
2.2.1 Muestreo de la voz 26
2.2.2 Supresión de silencio 27
2.2.3 Ventanas Hamming 27
2.2.4 Transformada de Fourier 28
2.2.5 Coeficientes cepstrales en escala Mel (MFCC) 29
2.3 Algoritmo de reconocimiento 31
2.3.1 Cuantificación Vectorial (VQ) 32
2.3.2 Algoritmo de agrupamiento LBG 33
2.3.3 Distancia Euclidiana 35

Índice

2.4 Conclusiones 37

Capítulo 3. Validación de la propuesta de solución 38

3.1 Introducción 38

3.2 Validación de la propuesta de solución 38

3.3 Comparación de la propuesta de solución utilizando el algoritmo de clustering LBG y la propuesta de solución utilizando el algoritmo de clustering K-means. 41

3.4 Comparación de diferentes implementaciones de MFCC 42

3.5 Conclusiones 45

Conclusiones Generales 46

Recomendaciones 47

Bibliografía Referenciada 48

Bibliografía Consultada 51

Anexos 56

Anexo A 57

Anexo B 62

Glosario de Términos 67

Índice de Ecuaciones e Ilustraciones

Ecuación 1. Fórmula para medir la energía 27
Ecuación 2.Función para calcular las ventanas Hamming 28
Ecuación 3. Fórmula para calcular la Transformada Rápida de Fourier 28
Ecuación 4 Ecuación para transformar la frecuencia a frecuencias Mel 29
Ecuación 5 Cálculo de los MFCC 31
Ecuación 6 Cálculo de la distancia euclidiana entre dos puntos 35
Ecuación 7Cálculo del Porciento de Aciertos 41

Ilustración 1 Ventana Hamming 28
Ilustración 2 Representación de la escala Mel 30
Ilustración 3 Banco de filtros de frecuencias Mel 30
Ilustración 4 Vectores generados en la fase de entrenamiento antes de usar la Cuantificación Vectorial 32
Ilustración 5 Vectores de características representativas resultado después de usar la Cuantificación Vectorial 33
Ilustración 6 LBG para un caso de dos dimensiones 34
Ilustración 7 Flujo de procesos de la propuesta de solución para el reconocimiento de locutores. 36
Ilustración 8 Proceso de entrenamiento 39
Ilustración 9 Proceso de prueba 40

Índice de Tablas

Tabla 1. Ventajas y Desventajas de los métodos de clasificación.....22
Tabla 2 Porciento de aciertos para el algoritmo LBG 41
Tabla 3 Porciento de aciertos para el algoritmo LBG 42
Tabla 4 Porciento de acierto para el algoritmo K-means 42
Tabla 5 Implementación de MFCC con 12 filtros. 43
Tabla 6 Implementación de MFCC con 22 filtros. 43
Tabla 7 Implementación de MFCC con 32 filtros. 43
Tabla 8 Implementación de MFCC con 12 filtros. 44

Introducción

La comunicación es la acción y efecto de comunicar o comunicarse[1]. Desde épocas muy primitivas los hombres se comunicaban de forma muy rústica, por medio de señas y gemidos; esto provocaba malas consecuencias para el entendimiento entre los mismos, ya que les era imposible comprender en su totalidad lo que querían transmitir. Con el desarrollo del lenguaje se solucionaron los problemas de comprensión que existían entre las personas. Sin embargo, con el crecimiento de la población en las ciudades, el descubrimiento de los continentes y el proceso de emigración que trajo el mismo, se hizo necesario crear medios de comunicación de larga distancia para cubrir las necesidades de información y comunicación.

Uno de los primeros medios de comunicación que se creó fue la imprenta, esta revolucionó la comunicación, haciendo posible la reproducción de textos, que han trascendido a través de la historia hasta nuestros días, el único inconveniente que esto traía era el tiempo que se demoraba en llegar alguna noticia a las localidades lejanas, pero es ya con el surgimiento del telégrafo que las personas empiezan a comunicarse de una población a otra de una forma un poco más rápida. La radio ha sido un valioso medio de comunicación, por la rapidez de su difusión y por el alcance de su emisión. Pero es ya con la televisión, y más aún de Internet, que se da un gran paso en los medios de comunicación, al incorporarse la imagen en la transmisión de información.

El desarrollo de los medios de comunicación está estrechamente ligado al auge tecnológico que se ha venido dando en los últimos años. La comunidad científica está cada vez más interesada en crear dispositivos que faciliten la vida diaria de las personas, como por ejemplo, la creación de aplicaciones que utilicen reconocimiento de voz para poder interactuar con estas, y más reciente, la utilización de aplicaciones las cuales funcionan al reconocer a un locutor en específico.

Los sistemas que funcionan mediante el reconocimiento de locutores deben estar entrenados para responder a ciertas características particulares de los locutores, para así identificarlos con mayor precisión. Ejemplo de estos son los sistemas de seguridad y control, que necesitan una mayor exactitud a la hora de identificar a las personas, para evitar intrusos. Este método de identificación permitiría autorizar operaciones bancarias delicadas, realizar transacciones desde un cajero automático o permitir el acceso a transacciones o privilegios denegados. Además, los avances en el

1 Diccionario de la Real Academia de la Lengua Española

reconocimiento de locutor han sido de gran relevancia para campos como la investigación policial. Entre las principales ventajas que traen consigo estos tipos de sistemas podemos destacar su bajo coste de mantenimiento, el alto nivel de seguridad que ofrecen y la comodidad para el usuario.

En Cuba existen varios estudios sobre este tema, la mayoría de ellos aportados por el CENATAV[2], centro encaminado a investigaciones teóricas y aplicadas en el área del reconocimiento de patrones y la minería de datos, como parte de las técnicas biométricas, pero esto no evita la existencia de aplicaciones que utilicen este método de autenticación, principalmente por el poco desarrollo tecnológico con que cuenta Cuba. Este tipo de aplicaciones están dirigidas principalmente a áreas como la biometría.

La Universidad de Ciencias Informáticas (UCI) no se encuentra ajena al estudio de la evolución y desarrollo de estas técnicas biométricas; existen asignaturas que se imparten en el cuarto año de la carrera que poseen como material de estudio estas técnicas, se podría mencionar la asignatura de Seguridad Informática, en la cual se expone la utilidad del uso de la biometría para la seguridad de sistemas e instituciones, además en la asignatura de Inteligencia Artificial, se hace un esbozo sobre las redes neuronales, las cuales están ligadas al reconocimiento de locutores, ya que se pueden usar en la extracción de características de los hablantes, y además para clasificar dichas características obtenidas.

En el centro de desarrollo de Geoinformática y Señales Digitales (GEySED), en el departamento de Señales Digitales se desarrolla el Sistema de Gestión y Transmisión de Contenidos Audiovisuales (SIAV), encargado de automatizar el proceso de transmisión de medias, además de la gestión, almacenamiento y modificación de las mismas. Dentro de los subsistemas con que cuenta, el Subsistema Radial trabaja con todo lo relacionado con la gestión de las transmisiones radiales; el mismo no es capaz de detectar, por sí mismo, que locutor está transmitiendo en algún momento dado, es decir, de monitorizar locutores en sus transmisiones, por lo que surge la necesidad de agregarle una funcionalidad al Subsistema Radial, que automatice dicho proceso.

De lo planteado anteriormente surge como **problema a resolver** la incapacidad de monitorizar locutores en las transmisiones radiales del SIAV. Para dar solución al problema planteado se propone como **objetivo general** definir el flujo para el reconocimiento de locutores en el SIAV, teniendo como **objeto de estudio** las técnicas de reconocimiento de locutores. Se ha definido como **campo de**

2 CENATAV: Centro de Aplicaciones de Tecnologías de Avanzada.

acción los procedimientos y técnicas para el reconocimiento de locutores dentro del SIAV y se **defiende la idea** de que con la definición del flujo para el reconocimiento de locutores se garantizará el camino para la implementación de un componente que monitorice los locutores de una transmisión radial del SIAV.

Las tareas de investigación a cumplir son:

- Caracterizar los algoritmos de reconocimiento de locutores existentes a nivel mundial.
- Caracterizar las tendencias y tecnologías actuales más factibles para el desarrollo de la investigación.
- Descripción de la propuesta de solución.
- Validar la propuesta de algoritmo de detección de locutores seleccionado.

Para dar cumplimiento al grupo de tareas planteadas servirán de ayuda varios métodos científicos de investigación.

Métodos teóricos:

Analítico-Sintético: Mediante el uso de este método se puede realizar un análisis sobre los temas relacionados con el reconocimiento de locutores, basándose en las competencias. Dicho método propicia entender el tema, facilita el estudio del mismo, así como el arribo a nuevas conclusiones.

Histórico-Lógico: Este método permite realizar un estudio del estado del arte del objeto de estudio de la presente investigación, con el objetivo de conocer su evolución y desarrollo, además de conocer las tendencias actuales de los sistemas de reconocimiento de locutores a nivel mundial, y así dar respuesta al problema expuesto anteriormente.

Inductivo-Deductivo: Este método posibilita el desarrollo de un conocimiento más generalizado del reconocimiento de locutores, partiendo de aspectos específicos encontrados en documentos involucrados con la investigación del tema.

Métodos empíricos:

Análisis documental: Para determinar el marco teórico y contextual de la investigación así como las tendencias que manifiesta, se revisaron documentos, tesis de maestrías y doctorado, artículos, entre otros.

Introducción

Este trabajo está organizado de la siguiente manera:

Capítulo1. Fundamentación teórica: en este capítulo se presenta el estado del arte del objeto de estudio de la presente investigación, además de definir los conceptos teóricos asociados al tema que posibilitan un mejor entendimiento de la situación problemática planteada y el marco del problema planteado. Se exponen aplicaciones existentes que pueden ayudar a dar solución al problema del presente trabajo.

Capítulo2. Caracterización de la propuesta de solución: en este capítulo se expone la propuesta de solución para dar respuesta al problema científico de la presente investigación, además se describen con mayor profundidad los pasos para el reconocimiento de locutores a usar para la creación de la solución final, donde se analizarán las características con el fin de dar cumplimiento al objetivo general de la presente investigación.

Capítulo3.Validación de la solución: en este capítulo se realiza la validación de la propuesta de solución exponiendo los resultados.

Capítulo 1. Fundamentación Teórica

1.1 Introducción.

Para obtener una mayor comprensión del contenido de la presente investigación se deben tener en cuenta ciertos conceptos asociados al reconocimiento de locutores. Con el objetivo de dar cumplimiento a lo antes expuesto se ha redactado este capítulo, en donde se argumentan los conceptos y elementos que crean el asiento para la fundamentación teórica, los cuales guiarán hacia una posible solución del problema de investigación planteado.

Permitir conocer el significado de conceptos como reconocimiento automático de locutor, identificación automático de locutor y verificación automática de locutor, además la concepción del reconocimiento del locutor en sí, aumentará la noción de las ideas que se plantean en el documento. En este capítulo también se elabora el marco teórico de la investigación el cual permite señalar los antecedentes del problema, los conocimientos científicos acumulados producto de investigaciones anteriores y hacer un análisis crítico de la literatura existente para sustentar teóricamente el trabajo que se realiza.

Se examinan también algunas aplicaciones y se consideran posibles soluciones ya implementadas con la realización de un análisis detallado del estado del arte del reconocimiento de locutores a nivel mundial. Además, se explica con más profundidad aspectos relacionados con la situación problemática presentada.

1.2 Conceptos asociados al reconocimiento de locutores.

1.2.1 Reconocimiento Automático de Locutor (RAL)

El reconocimiento automático de locutores es el proceso de extracción de características relativas a la identidad de los individuos a partir de sus muestras de voz. Este proceso de reconocimiento tiene dos etapas básicas: el entrenamiento, que es la recolección de muestras de personas a identificar, y el reconocimiento, que es la comparación de las muestras tomadas con el locutor desconocido, tomando una decisión. Dentro del reconocimiento de locutores se reconocen dos tareas diferenciales: la Verificación Automática de Locutores (VAL) y la Identificación Automática de Locutores (IAL) (1)**.**

Modalidad biométrica que utiliza el habla de una persona, una característica influenciada tanto por la estructura física del tracto vocal del individuo como por las características de comportamiento del individuo, para fines de reconocimiento. Se divide en identificación y verificación de locutor (2).

Se define entonces como reconocimiento automático de locutores a la identificación automática de personas a través de la voz, teniendo dos etapas fundamentales: el entrenamiento y el reconocimiento Este proceso se divide en identificación y verificación de locutor.

1.2.2 Identificación Automática de Locutores

En la identificación de locutores, la muestra del locutor desconocido es comparada con las muestras de locutores conocidos que se tienen. En este modo se devuelve al locutor que más se acerque a una de las muestras, o se devuelve que no se ha encontrado a ningún locutor. Existen dos formas de operar en la identificación de locutores: sobre un conjunto cerrado, es decir, que el locutor desconocido es con certeza uno de los del grupo, y sobre conjunto abierto, es decir, que el locutor desconocido puede no estar en el grupo (3).

El objetivo de la identificación automática de locutor es, dada una señal de voz, determinar, dentro de un grupo de personas predeterminadas, la identidad de su "propietario". Hablamos de IAL de Grupo Cerrado si el locutor desconocido es con certeza uno de los del grupo, y de IAL de Grupo Abierto si el locutor puede ser alguien ajeno a ese grupo de personas. En el primer caso la respuesta del sistema será siempre una identidad, mientras que en el segundo existe la posibilidad de la respuesta "locutor rechazado" al no pertenecer al grupo de personas de referencia (4).

En la identificación de locutor el locutor no aporta información sobre su identidad y es el sistema el que determina quién es a partir de su voz dentro de un conjunto de posibles candidatos o, si se trata de identificación en conjunto abierto, si el locutor es conocido o no por el sistema (2).

Se define que la identificación automática de locutores consiste en encontrar la identidad de una o varias personas dentro de un grupo de personas determinadas.

1.2.3 Verificación Automática de Locutores

En la verificación de locutores, el locutor desconocido proclama una identidad, que es comparada con las muestras que se tiene. En este modo solo existen dos alternativas de decisión: aceptar o rechazar la identidad proclamada (3).

El objetivo es verificar si el locutor es quien asegura ser; la respuesta del sistema será binaria: identidad aceptada o rechazada (4).

Capítulo 1. Fundamentación Teórica

La tarea de los sistemas de verificación de locutor es determinar si el locutor es o no quién dice ser. Por último, el seguimiento y agrupamiento consiste en etiquetar qué locutor está hablando en un segmento de voz y cuándo se producen cambios de locutores (2).

Se define que la verificación automática de locutores consiste en aceptar o rechazar la identidad proclamada por el locutor.

1.3 Concepción del reconocimiento de locutores.

GEySED, uno de los dos centros de producción de la facultad 6 de la UCI, tiene como objetivo brindar soluciones informáticas a clientes que así lo deseen, ya sean de la propia universidad, nacionales e internacionales. Dentro de los proyectos pertenecientes a este centro se encuentra el SIAV, que realiza un producto con mismo nombre. Los objetivos que persigue son la automatización de procesos de transmisión de medias, además del almacenamiento y gestión de las mismas. Para ello cuenta con 11 subsistemas los cuales se mencionan a continuación:

Subsistema Web.

Subsistema Radial.

Subsistema de Monitoreo.

Subsistema de Programación.

Subsistema de la Transmisión y Administración de la Transmisión.

Subsistema de Seguridad.

Subsistema de Transferencia (Codificador).

Subsistema de Gestión de Medias.

Subsistema de Reporte.

Equipamiento.

Producción.

La razón de ser del Subsistema de Trasmisión Radial lo constituyen las transmisiones y señales digitales de audio.

Capítulo 1. Fundamentación Teórica

Las señales de audio que se generan en una cabina se envían al aire mediante un radioenlace a la planta de transmisiones, luego se introduce dicha señal en el transmisor que la procesa y amplifica para entregarla a la antena que se encarga de radiarla. En Cuba el proceso de transmisión comienza cuando la señal abandona la emisora y se le entrega a Radio Cuba y ETECSA para que la transporten y radien respectivamente. Durante el trayecto la señal se convierte en forma analógica, pues las consolas de transmisión en los estudios, los enlaces y transmisores responden a esa tecnología (5). Dentro del sistema SIAV, el Subsistema de Transmisión Radial es el que se encarga de realizar todo el proceso de transmisión de señales digitales de radio, mejorando así la calidad de la realización de las transmisiones radiales.

Para obtener un mejor control de los locutores dentro del Subsistema de Transmisión Radial, es necesaria la existencia de una aplicación que permita la monitorización de los locutores dentro de las emisoras. Cuando una emisora radial está al aire, la identificación del locutor se hace de forma manual, ya que no existe ningún sistema que lo haga posible de manera automática. Esta funcionalidad no solo es factible para el sistema, sino que pudiera ampliar sus límites dentro de las emisoras radiales cubanas o incluso comercializar este producto a nivel internacional.

1.4 Situación Problemática.

Los avances tecnológicos son un hecho consumado en la actualidad, y la sociedad se vuelve partícipe de todos los beneficios y ventajas que estos brindan en las principales esferas, ya sea automatizando un proceso o valiéndose de determinados equipos para agilizar o realizar acciones. Cuba se hace eco de esta vertiente en casi todos los sectores y renglones tanto económicos como sociales, informatizando la mayoría de las empresas e instituciones.

En algunas emisoras cubanas la gestión de los procesos asociados a la edición, producción, musicalización y transmisión radial se realizan de forma manual, haciendo muy engorrosa la realización de los mismos. En el Instituto Cubano de Radio y Televisión (ICRT) se ha introducido la digitalización en un porcentaje elevado, este proceso comienza a finales de la década de los 90, obedeciendo a un estudio prolongado en el cual se analizaron distintos programas para determinar los más adecuados para dicha institución. Esto conllevó a la compra de servidores y materiales informáticos para ayudar a la informatización de dichas emisoras (5). La UCI cuenta con una aplicación hecha por el proyecto SIAV utilizada por el ICRT, la cual va dirigida a la automatización de procesos que ocurren en las emisoras radiales, pero la misma no dispone de funciones que ayuden a la detección de locutores.

Capítulo 1. Fundamentación Teórica

El reconocimiento automático de personas a través de la voz se puede utilizar para gestionar la información relacionada con los locutores de una o varias emisoras. Esto permite tener un mayor control sobre, por ejemplo, la cantidad de veces que un locutor X ha intervenido en una transmisión. Además, se puede utilizar para mostrar información relacionada al locutor o a la emisora que el locutor representa. Con el reconocimiento automático de locutores se eliminan errores humanos que puedan ocurrir durante el proceso de identificación de locutores, como mostrar información incorrecta del locutor que está hablando en ese momento, o de la emisora que está transmitiendo.

La creación de esta funcionalidad en el Subsistema de Transmisión Radial dotará al sistema SIAV de la capacidad de detectar locutores en las transmisiones radiofónicas, lo cual ayudará a mejorar el proceso de gestión de locutores y tendrá como característica que será una solución propia del sistema.

1.5 Análisis de soluciones existentes

Según (6), el reconocimiento automático de locutores tiene un amplio rango de aplicaciones en diversas áreas, tales como controles de acceso a recintos físicos o virtuales, máquinas y contenidos, transacciones financieras y comerciales, investigación policial y resoluciones judiciales, por lo que las soluciones que existen se centran para este tipo de aplicaciones. El reconocimiento de locutores dentro de una transmisión radial no ha sido desarrollado, por lo que no existen soluciones existentes que tengan que ver con esta aplicación.

En (3) se exponen varios métodos que se utilizan en la creación de sistemas de reconocimiento de locutores para la identificación o verificación de los mismos. A continuación se podrán apreciar algunas soluciones creadas para el reconocimiento de locutores en algunas de las áreas mencionadas anteriormente, y los métodos más utilizados para el reconocimiento de locutores, así como sus características.

Algunas soluciones de software que se han desarrollado son:

1.5.1 Verbio Speaker ID

Verbio Speaker ID es un motor de verificación de locutor basado en tecnología biométrica de la voz e integrado con el sistema de reconocimiento del habla Verbio ASR[3] (7). Un sistema de seguridad autónomo, y también, complemento ideal de otros existentes incorporándoles el reconocimiento de la persona a través de la voz. Permite autenticar o verificar, tras un breve entrenamiento inicial, la

3 Verbio ASR: Es el motor de reconocimiento de voz de Verbio orientado a entornos telefónicos.

identidad del hablante. Regula el acceso de usuarios a los sistemas de información basándose en los parámetros biométricos únicos de la voz de cada persona, estableciendo un mecanismo para dotar de mayor seguridad a los sistemas de acceso convencionales en numerosos entornos, como la telefonía, multimedia, domótica[4], industrial, entre otros, mediante la huella vocal de los usuarios (8).

Verbio Speaker ID se utiliza para la seguridad y control de acceso de usuarios a sistemas de información, a través de la verificación o autenticación de los mismos. Es un software privativo, que para su utilización sería necesario todo el paquete de software de Verbio, privativos también. Para la identificación solo trabaja con las bases de datos de locutores que contiene Verbio ASR. Además, está enmarcado para un entorno telefónico o cualquier entorno que requiera o disponga de un sistema manos libres. Solo puede ser instalado en servidores con sistema operativo Windows (8).

Por lo antes expuesto, se decide no utilizar este software, ya que su uso solo se enmarca para la verificación de personas a través de la voz en entornos telefónicos, y no en la identificación de locutores en el entorno radial. Además, al ser privativo y solo poder utilizarse en plataformas privativas no se cumple con las normas de soberanía tecnológica por las que aboga el país.

1.5.2 IDENTIVOX

El software IDENTIVOX es un proyecto elaborado entre la Universidad Politécnica de Madrid y laGuardia Civil de Madrid. Esuna aplicación de Windows desarrollado con Microsoft Visual C++. Incorpora una biblioteca de clases, programada en ANSI C++ la cual pretende el desarrollo de aplicaciones para el reconocimiento automático de locutores y otras aplicaciones biométricas (9).

Basado en modelos de mezclas gaussianas en el reconocimiento del locutor independientedel texto, para uso forense, aplicando el modelo bayesiano para las conclusiones. Permite opciones de parametrización con coeficientes cepstrales en escala Mel (MFCC), además del uso opcional de los coeficientes de velocidad y aceleración para los efectos de la articulación conjunta. Es un software desarrollado en plataformas privativas y solo es compatible con el sistema operativo Windows.

Por lo antes expuesto, se decide no utilizar este software, ya que solo se enmarca para uso forense, y no en la identificación de locutores en el entorno radial. Además, al ser privativo y solo poder utilizarse en plataformas privativas no se cumple con las normas de soberanía tecnológica por las que aboga el país. El no utilizar este software no impide el poder usar alguna técnica para el reconocimiento de

4 Domótica: Aplicación de la informática a las tareas del hogar (65).

personas a través de la voz que este implícito en el mismo, como lo es la técnica de extracción de los MFCC, utilizada para la extracción de características del habla.

1.6 Métodos utilizados para el reconocimiento de personas por la voz

Existen métodos que se utilizan en la creación de aplicaciones para el reconocimiento de locutores. El estudio de estos es necesario, ya que ninguna de las soluciones existentes da solución al problema de la investigación. Estos se dividen en:

- Métodos para la extracción de rasgos acústicos del habla.
- Métodos de selección de rasgos.
- Métodos de clasificación.

A continuación se explican con mayor detalle.

1.6.1 Métodos de extracción de rasgos acústicos del habla

Los métodos de extracción de características se hacen necesarios por dos razones fundamentales: en primer lugar, para reducir el volumen de cálculo, y en segundo lugar, para evitar el problema conocido como la maldición de la dimensionalidad, que no es más que el número de vectores de entrenamiento necesario para una mejor caracterización del locutor crece exponencialmente con la dimensión del espacio (3)**.**

A continuación se realiza un análisis de los principales métodos para la extracción de características.

1.6.1.1 Banco de filtros

José Ramón Calvo, Rafael Fernádez, Gabiel Hernández en su obra Métodos de extracción, selección, y clasificación (3), se refieren a que los bancos de filtros son un término genérico que se refiere a los métodos que procesan una señal en múltiples bandas de frecuencia. Según (10) y (11) existen dos formas de utilizar los bancos de filtros en el reconocimiento del locutor: realizando la fusión a nivel de rasgo (fusión de entrada), o la fusión a nivel del clasificador (fusión de salida). En la primera se combinan los rasgos obtenidos en cada sub-banda en un solo vector de rasgos M-dimensional y se entrena un solo modelo de locutor. En la segunda, los rasgos de cada sub-banda son considerados independientemente y para cada sub-banda, se crea un modelo separado.

Este método tiene la ventaja sobre otras representaciones espectrales, que los rasgos tienen una interpretación física directa, por ejemplo, conociendo a priori el poder discriminativo de cada sub-

banda, se puede establecer un peso apropiado a cada una, o si alguna de las sub-bandas está contaminada por ruido, pueden usarse las no contaminadas (3).

José Ramón Calvo, Rafael Fernádez y Gabriel Hernández en su obra Métodos de extracción, selección, y clasificación (3) explican que la eficacia de este método depende significativamente de varios factores:

- La arquitectura del sistema: selección de las sub-bandas más críticas para el reconocimiento (incremento de peso en la decisión); la división optima de la banda total de frecuencias (número y tamaño de sub-bandas).
- La recombinación de la salida de cada reconocedor por sub-banda: nivel de recombinación, estrategias de recombinación, fusión de decisiones múltiples.
- Los rasgos a usar para el reconocimiento del locutor deben ser seleccionados: los rasgos que se usan para el reconocimiento del habla, por lo general no son recomendables para el reconocimiento del locutor. De igual modo, algunos rasgos apropiados para el reconocimiento usando todo el espectro[5], pueden ser ineficientes al aplicar un procesamiento multiespectral[6] (12).

1.6.1.2 Coeficiente de predicción linear

El tracto vocal puede considerarse como un tubo acústico conformado por cilindros de diferentes secciones transversales, sin perdidas ni ramificaciones, con una onda plana de sonido propagándose y reflejándose en toda su extensión desde la glotis hasta los labios, su efecto en la excitación glotal[7] es provocarle una serie de resonancias, es decir, el tracto vocal puede modelarse como un filtro todo-polo, siendo una buena aproximación para muchos sonidos del habla en condiciones acústicas favorables (3).

5 Espectro: Es la representación de las frecuencias que componen una señal de audio (63).

6 Multiespectral: Varios espectros dentro de una misma señal de audio.

7 Excitación glotal: La sucesión de aperturas y cierres de la glotis producen pulsos de presión del aire cuasi-periódicos emitidos por las cuerdas vocales. Debido a la naturaleza de relajación forzada de los movimientos de las cuerdas vocales, la onda glotal excita las cavidades de resonancia del tracto vocal que tiene un espectro muy rico en armónicos. Cuanto mayor es la intensidad de la voz, más corta es la duración en que está abierta la glotis (separadas las cuerdas vocales). Por el contrario, para los sonidos sordos las cuerdas vocales están abiertas y el flujo de aire continuo puede pasar libremente a través de ellas (3).

El método conocido como de coeficientes de predicción lineal (LPC) o de modelación autorregresiva propuesto en (13) ajusta los parámetros del filtro todo-polo al espectro del habla. Con un número suficiente de coeficientes, el método LPC puede considerarse una aproximación adecuada a la estructura espectral de muchos sonidos.

La idea del método LPC es que muestras adyacentes en la señal de la voz están altamente correlacionadas, y por tanto, el comportamiento de esta en un momento dado, puede predecirse de cierta cantidad de muestras anteriores.

Los LPC son altamente correlacionados[8] (10) y raramente se utilizan como rasgos por sí mismos, de los posibles coeficientes que se pueden obtener a partir del modelado de la voz como un proceso autorregresivo[9], los coeficientes cepstrales[10], conocidos como LPC-cepstrales o LPCC, han sido siempre utilizados por muchos sistemas de reconocimiento del locutor, al estar menos correlacionados (14). El método LPC proporciona una alternativa sencilla al método de derivar los coeficientes cepstrales de los LPC a partir de expresiones recursivas.

Se han aplicado otras transformaciones sobre los LPC que dan lugar a distintas familias de coeficientes menos correlacionados, susceptibles de ser utilizadas en tareas de reconocimiento del locutor y del habla, con mayor o menor acierto (15):

- Coeficientes de reflexión
- Coeficientes de razón logarítmica de áreas
- Coeficiente de reflexión ArcSin
- Coeficientes de pares de líneas espectrales

El modelo del cual se obtienen los coeficientes LPC presenta limitantes que resultan inconvenientes para su aplicación:

8 Correlación: Correspondencia o relación recíproca entre dos o más cosas o series de cosas (60).

9 Procesos autorregresivos: Deben su nombre a la regresión y son los primeros procesos estacionarios que se estudiaron. Un proceso autorregresivo es un proceso estocástico (61).Un proceso estocástico es aquel en el que se representan todos y cada uno de los pasos necesarios para realizar una actividad, además de las formas o maneras en que cada uno de los pasos puede ser llevado a efecto y sus respectivas probabilidades, dicho de otra manera, cualquier proceso en el que se involucren probabilidades es un proceso estocástico (62).

10 Coeficientes cepstrales: Se utilizan para caracterizar de forma más compacta un espectro (64).

- Los pulsos glotales no tienen una estructura espectral plana
- El tracto vocal no está compuesto solo por cilindros
- La cavidad nasal constituye un pasaje adicional
- Algunos sonidos se generan cerca de los labios como fricativos sordos [11]

Los LPC presentan problemas ante la degradación del habla producto del ruido y ante cambios en el canal, que los hacen poco útiles en ambientes reales (3)**.**

1.6.1.3 Rasgos cepstrales[12] obtenidos del espectro

La representación de la voz utilizando su espectro de potencia a corto término asume la cuasi – estacionariedad de la voz en segmentos de tiempo entre 20 y 30 ms, lo que permite obtener su comportamiento espectral en el segmento aplicando la transformada de Fourier. La obtención continua y solapada de los espectros a corto término de la voz da lugar al espectrograma. Los rasgos cepstrales pueden obtenerse también a partir del espectro de potencia a corto término de la señal del habla (Short-Time Fourier Transform, STFT). Dichos rasgos, representados en escalas logarítmicas y distorsionadas en frecuencia con escala Mel, son conocidos por Coeficientes Cepstrales en escala Mel (MFCC) (10) (15).

La representación logarítmica del espectro de potencia tiene las siguientes ventajas (3):

- Cuando la ganancia de la señal varia, la forma del espectro se preserva y solo de desplaza en amplitud
- La señal sonora puede modelarse como la convolución[13] de una excitación cuasi-periódica con el filtro variante en el tiempo que representa el tracto vocal, ambas componentes son fáciles de separar en el dominio de la potencia logarítmica, donde se suman.
- La distribución estadística del espectro en el dominio logarítmico tiene propiedades no presentes en el espectro de potencia lineal, que son convenientes en el reconocimiento del

11 Fricativos sordos: Sonido o fonema consonántico en cuya articulación los órganos que intervienen no obstruyen por completo el canal vocal, sino que el aire sale rosando entre ellos (65).

12 Rasgos cepstrales: Los rasgos cepstrales representan típicamente las propiedades de magnitud de espectro de una señal de habla, por lo que son ampliamente usadas en el procesamiento de voz (67).

13 Convolución: Es la respuesta a una excitación particular (66).

locutor y del habla.

El cepstrum de la voz se define como la transformada de Fourier inversa del logaritmo del espectro de potencia a corto término, y se utiliza para caracterizar tanto sonidos sonoros como sordos, con buenos resultados en la práctica tanto en el reconocimiento del locutor y del habla (3).

1.6.1.4 Rasgos dinámicos

La información dinámica de los rasgos espectrales o rasgos delta se estima calculando las derivadas de los rasgos en el tiempo y anexando dichas derivadas al vector de rasgos, elevando su dimensionalidad, lo que requiere un mayor volumen de datos para el entrenamiento de los clasificadores. Comúnmente se estiman también las derivadas en el tiempo de los rasgos delta, conocidas como rasgos delta delta y se anexan al vector de rasgos, creciendo aún más la dimensionalidad del espacio de rasgos. Los rasgos delta delta se pueden utilizar sobre cualquier rasgo espectral a corto término, especialmente los rasgos cepstrales y sus variantes (3).

1.6.1.5 Rasgos cepstrales delta desplazados

Este método es muy utilizado en reconocimiento de lenguaje. El uso de estos rasgos permite calcular un vector de rasgos pseudo-prosódico de una forma muy sencilla sin la necesidad de encontrar un modelo de la estructura prosódica de la señal de voz. Estos rasgos se determinan a partir de cualquiera de los rasgos acústicos (LPCC, MFCC, etc) (3).

1.6.1.6 Rasgos wavelet

En esquemas de extracción de rasgos diseñados con el propósito de reconocimiento del habla, los wavelets han tenido dos salidas: la primera es el uso de la transformada wavelet como un decorrelador[14] efectivo (16), y la segunda es la aplicación directa de la transformada wavelet a la señal de voz, tomando los coeficientes wavelets de mayor energía como rasgos, o utilizando las bandas de energía en lugar de las sub-bandas Mel (17). En el campo del reconocimiento del locutor, recientes avances se han observado con la utilización de la transformada wavelet para la obtención de rasgos del locutor. La teoría wavelet brinda una estructura muy flexible para obtener representaciones de señales con buenas resoluciones en ambos dominios: frecuencial y temporal. Permite también lidiar con problemas de ruido bien localizado en las frecuencias (3).

14 Decorrelador: Divide la señal en varias ramas.

1.6.2 Métodos de selección de rasgos

Una manera de elevar la efectividad de un sistema de reconocimiento de patrones es elevar la dimensionalidad de los rasgos, sin embargo esto puede causar problemas afectando la eficiencia del reconocedor, pues incrementa el costo computacional (18). Para solucionar dicho problema se requiere reducir adecuadamente la dimensionalidad del espacio de rasgos, obteniendo una compactación del mismo y elevando la eficiencia del reconocedor que trabajara con un conjunto menor de rasgos, menos redundantes, más relevantes y discriminativos (3).

Un método de selección de rasgos consiste en escoger el mejor sub-grupo de *k* rasgos de todos los *K* rasgos que representan al patrón, que el mismo permita mejor probabilidad de una clasificación correcta (3). A continuación se exponen diversos métodos de selección de rasgos.

1.6.2.1 Búsqueda exhaustiva

La búsqueda exhaustiva es el método más completo de búsqueda para selección de rasgos. La misma evalúa el error de clasificación con todas las posibles combinaciones de sub-grupos de k rasgos. La búsqueda exhaustiva tiene como inconveniente que requiere un enorme esfuerzo computacional, por lo que no se utiliza en la práctica (3).

1.6.2.2 Método de k – mejores

El método de k-mejores es el método más simple de los métodos de selección de rasgos. En este método el mejor sub-grupo de rasgos está compuesto por los k mejores rasgos, considerados independientes. Sin embargo, un sub-grupo de los k mejores rasgos no precisamente es el mejor sub-grupo de rasgos (3).

1.6.2.3 Selección hacia adelante

El método de selección hacia adelante comienza con el espacio de rasgos vacío y se van adicionando rasgos iterativamente, la prueba inicial se hace con cada rasgo individual, uno a uno, seleccionando los mejores k rasgos simples. Después se prueba con dos rasgos, incluyendo uno de los mejores seleccionado previamente y cada uno de los restantes K - k rasgos, el ciclo se repite hasta que se seleccionen los rasgos que se deseen (3).

1.6.2.4 Selección hacia atrás

Este método es una técnica de búsqueda paso a paso, llamada estrategia knock-out y comienza con el espacio total de K rasgos, todos los K sub-grupos de K-1 rasgos se usan para calcular el comportamiento y determinar el mejor sub-grupo de k-1 rasgos, el rasgo no usado queda fuera. El proceso se repite con k-1 sub-grupos de k-2 rasgos hasta que se llegue al sub-grupo de k rasgos que se desean (3).

1.6.2.5 Algoritmo l – r

Este algoritmo usa la selección hacia adelante y hacia atrás para obtener el mejor comportamiento del procedimiento de selección. Para cada iteración el algoritmo usa el procedimiento hacia delante para agregar l rasgos alsub-grupo y usa el procedimiento hacia atrás para eliminar los peores r rasgos del sub-grupo, hasta lograr los k rasgos más deseados. Existe una variante dinámica del algoritmo l – r conocida como Sequential Floating Forward Sequence (SFFS), que consiste en aplicar, después de cada paso hacia adelante, un número de pasos hacia atrás hasta que el sub-grupo resultante sea mejor que los anteriores, no habrá pasos hacia atrás si no se mejora el comportamiento (3).

1.6.2.6 Programación dinámica

La programación dinámica se utiliza para obtener el número óptimo de rasgos con mucho menos cálculos que la búsqueda exhaustiva. Esta es una técnica de optimización que cuando se utiliza en junto con alguna ecuación funcional para la selección de rasgos permite que dicha selección obtenga la máxima efectividad (3).

1.6.3 Métodos de clasificación

A partir de la extracción de los rasgos acústicos del habla y su respectiva selección, se debe encontrar un modelo que clasifique los rasgos efectivamente, además que sea lo suficientemente robusto ante las diferentes condiciones de variabilidad que pueda presentar la voz (3). A continuación se explican los principales métodos de clasificación de rasgos acústicos.

1.6.3.1 Distorsión dinámica en el tiempo

El algoritmo busca el mejor trayecto de alineamiento por medio de técnicas de programación dinámica entre la expresión de entrada y la referencia, realizando un mapeo lineal segmentado entre uno o ambos ejes de tiempo para alinear las dos señales; al concluir el mapeo, la distancia acumulada entre

las dos expresiones es la base de la puntuación. Si las expresiones son idénticas en el tiempo, el trayecto de alineamiento es una diagonal; si no son idénticas, las desviaciones de la diagonal representan las distancias requeridas a distorsionar.

El método puede aplicarse también para alinear las variaciones en el tiempo de rasgos correspondientes a la configuración dinámica de los articuladores y el tracto vocal, como la energía (15) y los coeficientes LPC y dinámicos (19).

Existen dos factores que afectan el funcionamiento del DTW en sistemas dependientes del texto: la detección del punto final y el establecimiento de restricciones al trayecto de alineamiento local. Debido a la simplicidad del sistema es fácilmente aplicable en tareas de control de acceso con password. Además el sistema es altamente dependiente de las expresiones de referencia, no permitiendo variabilidad en la señal de voz (3).

1.6.3.2 Métodos de cuantificación vectorial

En la cuantificación vectorial cada vector N-dimensional de entrada se representa por el más cercano codevector o centroide[15] de un pequeño grupo de vectores o codebook[16] altamente representativos de la distribución de vectores de entrada en el espacio N-dimensional. Este codebook se selecciona con los mejores representantes de los diferentes clúster o grupos en los que se hayan dividido los datos de entrada (3).

Para la aplicación de VQ al reconocimiento de locutores, se construye un codebook por cada locutor de la base, en el caso de la identificación del locutor se computa la distorsión de los vectores de rasgos de entrada con respecto a cada codebook y la menor distorsión nos indicara el locutor identificado; en el caso de la verificación del locutor, se computa la distorsión de los vectores de entrada con respecto al codebook del locutor que calma su identidad y se compara con un umbral, si la distorsión es menor que el umbral, se acepta como verificado el locutor (3).

Este método reduce sensiblemente la capacidad de almacenamiento en el cálculo del análisis espectral y reduce la complejidad computacional en el cálculo de distancias (3).

[15]Codevector o centroide: Vector representativo de una determinada región.

[16]Codebook: Conjunto finito de centroides.

1.6.3.3 Métodos discriminativos: redes neuronales

Las redes neuronales son modelos computacionales que emulan el comportamiento del celebro por medio de topologías que reflejan la interconexión de las células nerviosas (20). Una red neuronal consiste en un grupo de “neuronas” (nodos o celdas) que se distribuyen en capas y se interconectan por medio de caminos pesados. Cada neurona es un elemento procesador que entrega una salida simple a partir de múltiples entradas, dicha salida usualmente es controlada por una función de activación (3).

Las redes neuronales poseen una capa de entrada, una capa de salida y una o más capas ocultas, estando conectadas las salidas de una capa a las entradas de la próxima. Una red neuronal se entrena ajustando los pesos de los caminos de las uniones entre neuronas. El entrenamiento es supervisado si se conocen los patrones a aplicar a la capa de entrada y las clases correspondientes a obtener en la capa de salida, las neuronas correspondientes a las capas ocultas se entrenan de forma tal que cuando se cargue el patrón a la entrada, la salida de la clase a la cual corresponde el mismo se active (21).

1.6.3.4 Modelos ocultos de Markov

Los sistemas basados en Modelos ocultos de Markov son redes de estado que intentan modelar el mecanismo de producción del habla. Están integrados por un conjunto de estados (asimilables a las distintas posiciones en las que puede configurarse el tracto vocal durante una locución) que desembocan en un conjunto de posibles salidas (3).

Los modelos utilizados por los sistemas HMM para caracterizar la identidad de un locutor independiente del texto, son los denominados ergódicos, en los que no existe una ordenación correlativa[17] de las transiciones entre los distintos estados del modelo y, por lo tanto, resulta factible cualquier combinación de transición entre estados (3).

La principal ventaja de los sistemas de reconocimiento basados en HMM respecto a otros tipos ya referidos, la constituye su gran versatilidad, tanto en lo que se refiere a los procesos de entrenamiento como a ciertas características variables de la muestra: duración, contenido fonético o lingüístico, contexto, etc. Presenta una gran adaptabilidad a la variación de las condiciones de registro o del canal de transmisión (3).

[17] Continua, seguida, sucesiva, etc.

Este método presenta un alto costo computacional, y sus mejores resultados se encuentran en el reconocimiento del locutor dependiente de texto (3).

1.6.3.5 Modelos de mezclas gaussianas

Según (3) los Modelos de mezcal gaussianas modelan los distintos vectores de parámetros de una locución dada, realizando una suma ponderada (mezcla) de funciones de densidad de probabilidad gaussianas. Utilizando GMM podemos representar, con un alto grado de fidelidad, un amplio margen de distribuciones muestrales, tales como los diferentes coeficientes cepstrales que puede generar una locución concreta. Los GMM no precisarán, en la fase de entrenamiento, segmentar en estados ni entrenar la matriz de probabilidades de transiciones. En la etapa de reconocimiento no será necesario buscar la secuencia de estados de máxima verosimilitud, sino que bastará con acumular las probabilidades que asocia el modelo con cada uno de los vectores de entrada. Este método presenta un alto costo computacional implicando un tiempo considerable al crear los modelos.

Ventajas y desventajas de métodos de clasificación (3)

	Ventajas	Desventajas
Dynamic Time Warping (DWT)	Detectan y comparan tramos fonéticos de alta estabilidad (vocales abiertas, consonantes nasales) aplicando técnicas de correlación cruzada, coherencia, etc., para la medida de distancias. Los sistemas DTW han sido utilizados en algunas metodologías forenses como complemento a otros análisis clásicos.	Los principales inconvenientes de esos sistemas se relacionan con la enajenación de la información a nivel suprasegmental[18] y la necesidad de supervisión en las tareas de segmentación.

[18] Suprasegmental: Conocido como prosódica, es una característica del habla que afecta a un segmento más largo que el fonema, tales como el acento, la entonación, el ritmo, la duración y otros. El término suprasegmental implica la existencia de elementos que recaen sobre más de un segmento a la vez. Los suprasegmentales resultan de una utilización particular de recursos del aparato fonatorio (68).

Vector Quantization (VQ)	Reducción sensible de la capacidad de almacenamiento en el cálculo del análisis espectral y una reducción de la complejidad computacional en el cálculo de distancias (se puede usar cálculos tan simples como la distancia euclidiana o la de Mahalanobis).	Sus inconvenientes más significativos están relacionados con la distorsión espectral por el error de cuantificación (al representar cada vector por un representante).
Artificial Neural Network (ANN)	Las redes son robustas al ruido y permite tener en cuenta el contexto de la señal, pueden crearse redes que tengan un funcionamiento similar a las VQ, a las GMM, HMM y otros algoritmos en el reconocimiento del locutor.	Presentan como limitantes que la mayoría de las redes requieren almacenar todos los datos del entrenamiento durante la clasificación, requiriendo, en algunos casos, un volumen apreciable de memoria y poder de cálculo.

Hidden Markov Models (HMM)	Su gran versatilidad, tanto en lo que se refiere a los procesos de entrenamiento como a ciertas características variables de la muestra: duración, contenido fonético o lingüístico, contexto, etc. A todo ello, hemos de añadir su gran adaptabilidad a la variación de las condiciones de voz o del canal de transmisión y, lógicamente, su funcionalidad en condiciones dependientes de texto.	Alto costo computacional, y sus mejores resultados se encuentran en el reconocimiento del locutor dependiente del texto.
Modelos de mezclas de Gaussianas (GMM)	Las GMM pueden representar, con un alto grado de fidelidad, un amplio margen de distribuciones muestrales, como es el caso de los diferentes coeficientes cepstrales que pueden generar una locución.	Presentan un alto costo computacional implicando un tiempo considerable al crear los modelos

Tabla 1. Ventajas y Desventajas de los métodos de clasificación

Una evaluación comparativa de entre los diferentes métodos existentes no es una tarea simple debido a las diferentes configuraciones experimentales utilizadas por los grupos de investigadores. Esto trae como consecuencia de que los resultados reportados en la literatura no son directamente

comparables, debido a la gran variedad de bases de datos, opciones de parámetros para caracterizar la voz y la variedad de esquemas de post-procesamiento. Todo esto contribuye a la discordancia de los resultados reportados.

1.7 Conclusiones parciales

- Con el análisis de los conceptos asociados al dominio del problema se garantizó un mejor entendimiento del tema de la investigación.
- Según el estudio que se realizó sobre las soluciones existentes, se identificó que no existe ningún programa que se utilice con el fin de reconocer locutores dentro de una transmisión radial, sino que se especializan en otras áreas como la investigación forense y enla seguridad y control de sistemas.
- Para la creación de un algoritmo que reconozca locutores se deben extraer las características del hablante, realizar una selección de sus mejores características, y por último clasificar dichas características.

Capítulo 2. Caracterización de la propuesta de solución

2.1 Introducción

El presente capítulo tiene como objetivo presentar y explicar los pasos a seguir para la definición de un algoritmo como propuesta de solución para la creación de un sistema que sea capaz de reconocer locutores. La propuesta se apoya en una serie de investigaciones hechas sobre el tema, escogiendo las que arrojan mejores resultados, según los expertos y autores de dichas investigaciones.

Según (22), (23) y (24), los sistemas de reconocimiento de locutores tienen implícito dos fases: la fase de entrenamiento y la fase de prueba. En la fase de entrenamiento se le enseña al sistema a reconocer a los locutores, creando los modelos y estableciendo los umbrales de cada locutor que se encuentran en una base de datos, y en la fase de prueba se comparan los modelos de los locutores que pertenecen a la base de datos, con las muestras de voz del locutor desconocido, teniendo, como respuesta del sistema, el locutor identificado.

Para la creación de la propuesta de solución se identifican dos etapas significativas (22), (23), (24), (4) y (25), estas son: extraer las características del habla y utilizar el algoritmo de reconocimiento. Cada una de estas etapas cuenta con varios pasos para la realización de las mismas. Más adelante se explica con más detalles las características de estas etapas y de los procesos o pasos implícitos en ellas.

2.2 Extracción de características del habla

La extracción de características consiste en reducir la dimensión de un vector. Esta acción es necesaria para reducir el volumen de cálculo y para evitar que el número de entrenamiento de los vectores crezca exponencialmente junto con la dimensión del vector (25).

Lasse L. Molgaard, Kasper W. Jorgensen en Speaker Recognition (25), definen que para la extracción de características se deben tener en cuenta algunos criterios:

- Deben tener un alto poder discriminativo y ser altamente medibles.
- Deben ocurrir de forma natural y frecuentemente.

- Deben ser estables todo el tiempo.
- Fácil de medir.
- No ser susceptible de imitación.

En investigaciones realizadas por (22), (24) y (25) muestran que el objetivo principal de la extracción de características es simplificar el reconocimiento creando un resumen de la gran cantidad de datos de voz, sin que se pierda las propiedades acústicas que la define. Para lograr esto, los investigadores mencionados anteriormente, definen una serie de pasos, como el muestreo de la voz, la supresión de silencio, las ventanas Hamming y la Transformada de Fourier, que se deben cumplir, los cuales se explican a continuación.

2.2.1 Muestreo de la voz

En (22) se especifica que el primer paso en cada sistema de reconocimiento de locutores es extraer una señal digital que contiene la información contenida en la onda de presión que cada persona produce cuando habla. Para esto se utiliza un micrófono, para después convertir esta onda de presión en una señal analógica, y luego esta señal analógica se convierte en una señal digital mediante un convertidor analógico-digital.

La señal analógica resultante de la voz humana contiene la mayor parte de su energía entre 4 Hz y 4 KHz, por lo que se puede utilizar un filtro de paso bajo con una frecuencia de corte de 4 KHz a fin de reducir aliasing[19] en la digitalización de la señal. En (22) se propone que el ancho de banda del filtro y la tasa de muestreo se puede aumentar para obtener más información acerca de la señal de voz. En todos los casos, la frecuencia de muestreo deberá ser al menos el doble de la frecuencia más alta que se espera en nuestra señal.

Una vez que se tengan las muestras que se necesitan para cuantificar, se asigna un número digital para representar cada muestra. Cuantos más bits se utilizan para cuantificar esta cantidad, más precisa será la cuantificación de la señal. Es típico utilizar 8 ó 16 bits para cuantificar muestras de habla.

19 Aliasing: Es un efecto que produce una distorsión en una señal cuando se muestrea.

2.2.2 Supresión de silencio

Carlos Domínguez Sánchez en Speaker Recognition in a handheld computer (22), plantea que después de haber realizado el muestreo de la voz se procesa el habla, en la cual generalmente se divide el flujo de muestras en tramas o frames, donde cada frame es un grupo de N muestras. El primer frame contiene las primeras N muestras. Así, un solo frame puede contener muestras de M hasta M + N. Por lo tanto, se divide la señal de tal manera que no se superponen N - M muestras. Esta coincidencia permite procesar los frames de forma independiente.

Es posible medir la energía en cada frame aplicando la ecuación 1, donde N es el número de muestras por frame. El valor típico de N es 256 (22).

$$E_{frame} = \sum_{N} x[n]^2$$

Ecuación 1. Fórmula para medir la energía

Cuando la corriente de energías se analiza, si hay un número de frames consecutivos durante los cuales la energía es mayor que un umbral especificado, entonces el comienzo de un enunciado se ha encontrado. A la inversa, si hay un número de frames consecutivos en los que la energía es inferior al mismo, entonces no es necesario calcular más frames porque el enunciado ha terminado. Este proceso se repite con los siguientes frames con el fin de encontrar nuevos enunciados.

2.2.3 Ventanas Hamming

En, (22), (4), (28) y (29), se explica que el próximo paso del proceso es enventanar cada trama resultante después de realizar la supresión de silencio, para así minimizar las discontinuidades de la señal en el comienzo y fin en las mismas. En una función ventana existe un valor cero fuera de algún intervalo elegido. Cuando la señal o cualquier otra función son multiplicadas por una función ventana, el producto es también evaluado de cero fuera del intervalo (30). El ventaneado se hace para evitar problemas debidos al truncamiento de la señal. Según (25), la ventana más ampliamente utilizada en el procesamiento de la señal es la ventana de Hamming, su representación se muestra en la ilustración 2, y se describe por la ecuación 2, donde N es el número de muestras en cada frame.

Capítulo 2. Caracterización de la propuesta de solución

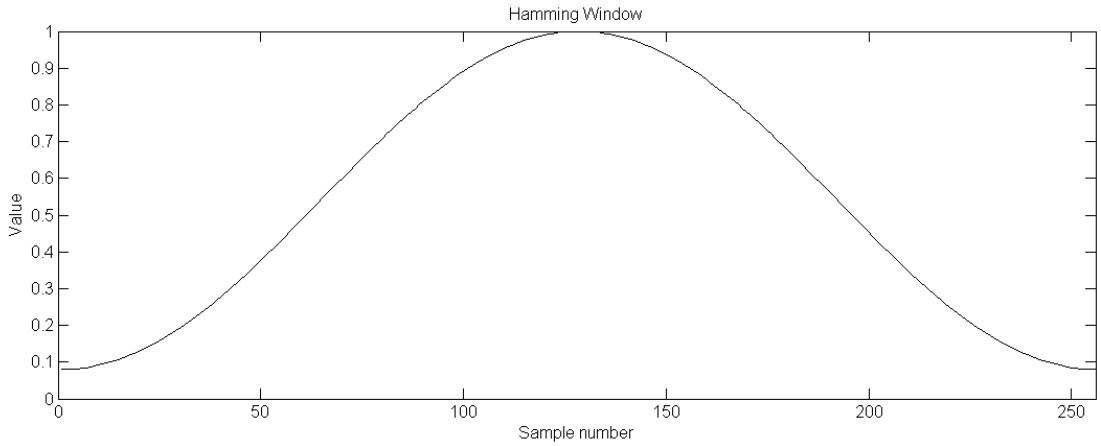

Ilustración 1 Ventana Hamming

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1$$

Ecuación 2.Función para calcular las ventanas Hamming

2.2.4 Transformada de Fourier

Según (22), explica que en este punto del proceso se tiene un número de tramas significativas, ya que se ha suprimido el silencio. Este proceso convierte cada tramade N muestras del dominio temporal al dominio de la frecuencia (4), (23). Una de las maneras más útiles para estudiar estas tramas es calcular su espectro, es decir, realizar una medición espectral. Esto se puede hacer fácilmente mediante el cálculo de la Transformada Rápida de Fourier (TRF) (22), (23), (25).

$$X_n = \sum_{k=0}^{N-1} X_k e^{-2\pi jk\frac{n}{N}}, n = 0,1,2,\dots,N-1$$

Ecuación 3. Fórmula para calcular la Transformada Rápida de Fourier

Donde N es el número de muestras, X_n son números complejos, j se utiliza para denotar una unidad imaginaria.

Con este algoritmo es factible obtener el mismo resultado que con una Transformada Discreta de Fourier (DFT) de un modo más rápido. Después de transformar cada frame al dominio de la frecuencia se obtiene un vector de valores.

2.2.5 Coeficientes cepstrales en escala Mel (MFCC)

Según (22), (23), (24) y (25), el elemento con más éxito utilizado para llevar a cabo el reconocimiento del locutor en los últimos años es el de coeficientes cepstrales en escala Mel. Este algoritmo consiste en aplicar un conjunto de filtros al espectro de una señal con el fin de medir la cantidad de energía que se encuentra en cada banda de frecuencia (canal). El resultado es una representación paramétrica (29) de expresión, al tiempo que reduce la cantidad de información que necesita para ser comparados entre las muestras de habla de un locutor dado y locutores grabados previamente.

Para calcular los MFCC primero se deben llevar las frecuencias a la escala de frecuencias Mel. Para llevar a cabo esta transformación se aplica la ecuación 4, siendo f la frecuencia a transformar (23), (4).

$$f_{mel} = 2595 log_{10}\left(1 + \frac{f}{700}\right)$$

Ecuación 4 Ecuación para transformar la frecuencia a frecuencias Mel

La escala Mel fue proyectado por Stevens, Volkmann y Newman en 1937.La escala Mel se basa principalmente en el estudio de observar el tono o frecuencia percibida por el ser humano. Es en general una aplicación lineal por debajo de 1000 Hz y logarítmicamente espaciada por encima (30). De este modo se da mayor importancia a la información contenida en lasbajas frecuencias en consonancia con el comportamiento del oído humano (4).

La figura siguiente muestra un ejemplo de frecuencia normal que se hace corresponder con la frecuencia de Mel.

Capítulo 2. Caracterización de la propuesta de solución

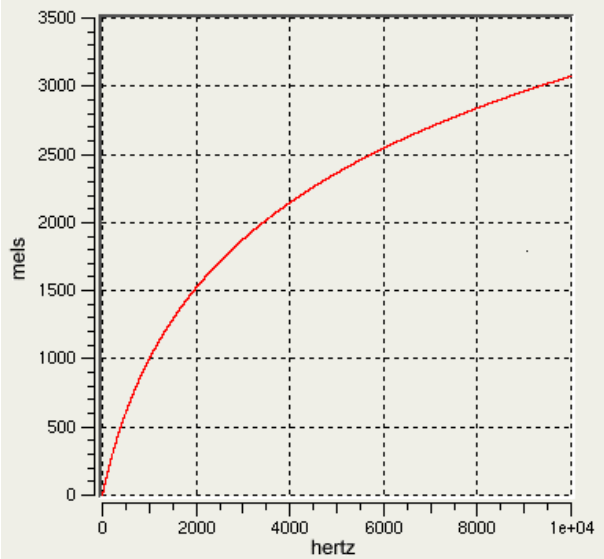

Ilustración 2 Representación de la escala Mel

La deformación de frecuencias Mel es más conveniente hacerla utilizando un banco de filtros con filtros centrados de acuerdo a las frecuencias Mel. La anchura de los filtros triangulares varía de acuerdo a la escala Mel, de modo que la energía total de registro es incluida en una banda crítica alrededor de la frecuencia central (22), (23), (4).

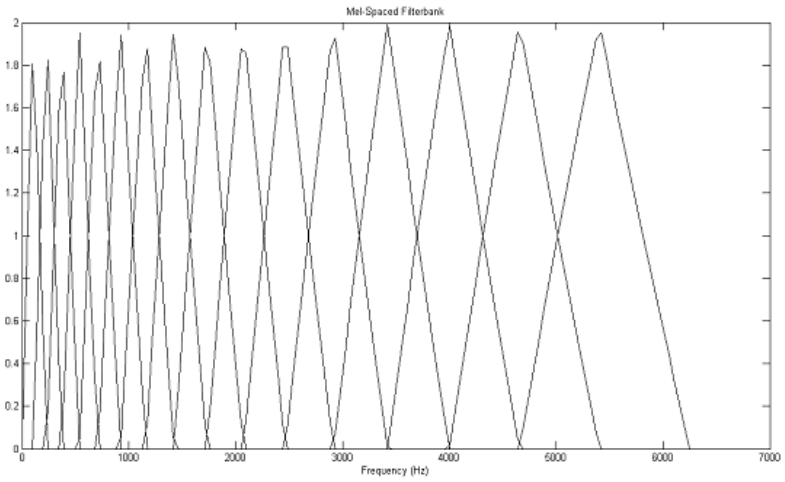

Ilustración 3 Banco de filtros de frecuencias Mel

Capítulo 2. Caracterización de la propuesta de solución

Se aplica al espectro de la señal el banco de filtros y luego se suma la salida para cada banda (4), dando lugar a los distintos coeficientes Mel, denotados por S_k con k = 1, 2,…., K, siendo K la cantidad de filtros (23), (4), (29). Finalmente se calcula el logaritmo se S_k, como los coeficientes Mel son números reales, se los puede convertir al dominio temporal mediante el uso de la Transformada Discreta del Coseno (29). El resultado es a lo que se le denomina MFCC. (29) expone que los MFCC se pueden calcular utilizando la ecuación

$$C_n = \sum_{k=1}^{K} (logS_k) \left[n \left(k - \frac{1}{2} \right) \frac{\pi}{K} \right] \; con \; n = 1, \dots, K-1$$

Ecuación 5 Cálculo de los MFCC

Donde cada elemento de la ecuación representa:

- k es la banda de frecuencias.
- n es el coeficiente MFCCen cuestión.
- $(logS_k)$ es el logaritmo de los coeficientes Mel.
- K es el número total de bandas o filtros.

2.3 Algoritmo de reconocimiento

En (30) y (25) se explica que un sistema de reconocimiento de locutor debe poder estimar las distribuciones de probabilidad de los vectores de características calculados. El almacenamiento de todos los vectores que se generan solo en el modo de entrenamiento resulta imposible, ya que estas distribuciones se definen en un espacio de alta dimensión. A menudo es más fácil empezar por cuantificar cada vector de característica para un número relativamente pequeño de los vectores de plantilla, con un proceso llamado Cuantificación Vectorial.

Para calcular la distorsión entre los vectores después de haber utilizado la Cuantificación Vectorial, se utilizan los algoritmos de agrupamiento, y por último se calcula la distancia entre el modelo de la voz del locutor desconocido y los modelos calculados previamente en la fase de entrenamiento. En los siguientes sub-epígrafes se explican con más detalles los procesos mencionados anteriormente.

2.3.1 Cuantificación Vectorial (VQ)

En (22), (23), (4), (26), (27) y (31), se refieren que existen muchos algoritmos que tratan de comprimir la información mediante el cálculo de centroides. Esto es: vectores en un espacio vectorial se agrupan en centroides y la etiqueta del centroide más cercano puede ser utilizado para representar el vector. Explican que lo ideal sería que sean pocos centroides pero bien separados (mientras más centroides se tenga, menor será la distorsión).

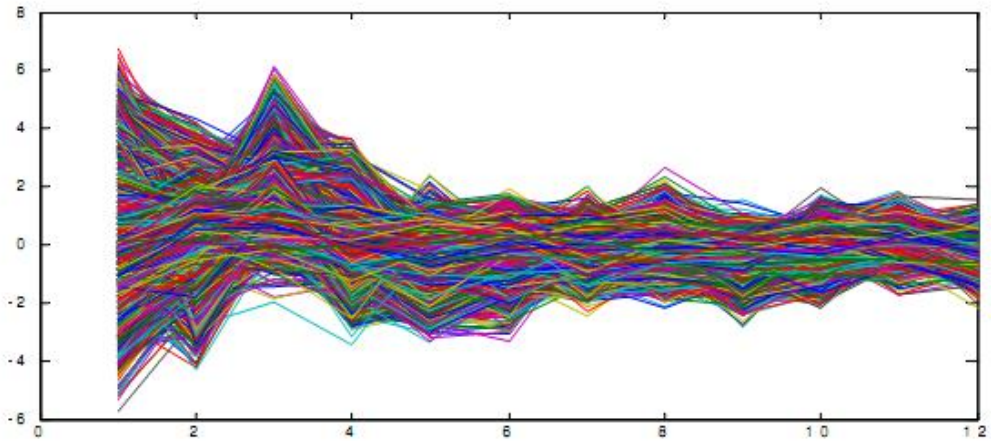

Ilustración 4 Vectores generados en la fase de entrenamiento antes de usar la Cuantificación Vectorial

La técnica de Cuantificación Vectorial consiste en extraer un pequeño número de características representativas de vectores como un medio eficaz de caracterizar las características de un locutor específico. Almacenar cada vector único que se genera a partir de la formación por medio de la Cuantificación Vectoriales imposible (30).

Capítulo 2. Caracterización de la propuesta de solución

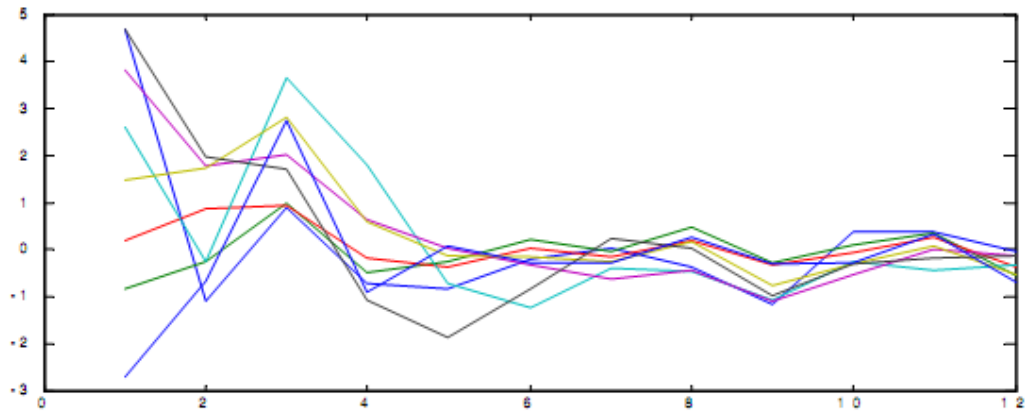

Ilustración 5 Vectores de características representativas resultado después de usar la Cuantificación Vectorial

Lasse L Mølgaard, Kasper W Jørgensen en Speaker Recognition (25), explican que utilizando estas funciones de datos de entrenamiento las características se agrupan para formar un libro de códigos para cada locutor. En la etapa de reconocimiento, los datos de la prueba del locutor se comparan con el libro de códigos de cada locutormidiendo la diferencia. Estas diferencias son usadas para tomar la decisión de reconocimiento.

2.3.2 Algoritmo de agrupamiento LBG

Después de la fase de entrenamiento, los vectores acústicos extraídos del hablante de entrada proporcionan un conjunto de vectores de entrenamiento. El paso siguiente que se hace es construir un codebook de hablantes específicos para este locutor usando los vectores de entrenamiento. En este punto de la investigación, los autores consultados se dividen en dos vertientes para realizar el agrupamiento, unos abogan por el uso del algoritmo K-means y otros por la utilización del algoritmo LBG. Para la propuesta de solución se optó por el algoritmo LBG.

Autores como (22), (23), (29), (31) y (32), exponen que existe un algoritmo bien conocido llamado LBG que se utiliza para agrupar un conjunto de vectores de entrenamiento L en un conjunto de vectores de un codebook M. El algoritmo es un algoritmo iterativo que, alternativamente, resuelve los criterios de optimización. El algoritmo requiere un codebook inicial. El codebook inicial se obtiene por el método de dividir. En este método, un codevector inicial se establece como el promedio de la secuencia de

entrenamiento. Este codevector se divide entonces en dos vectores. El algoritmo iterativo se ejecuta con estos dos vectores como el codebook inicial. Los dos últimos codevector se dividen en cuatro y el proceso se repite hasta que se obtenga en número de vectores deseados (29).

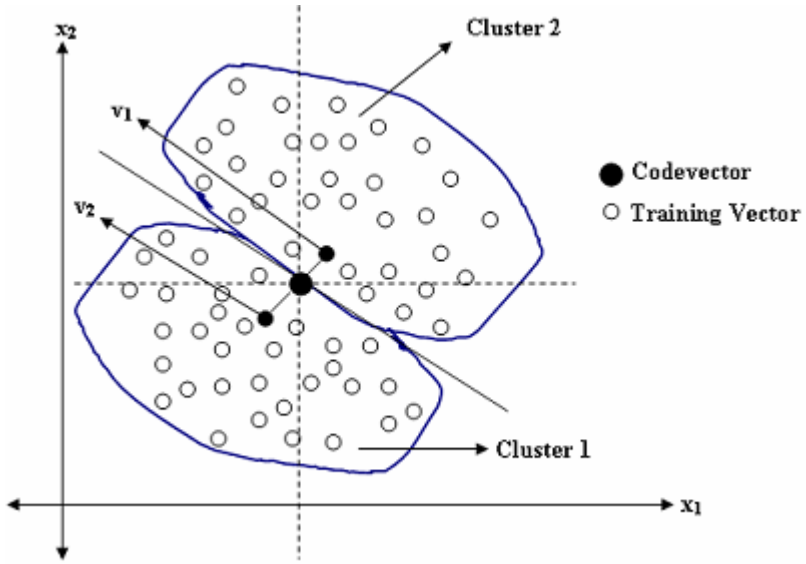

Ilustración 6 LBG para un caso de dos dimensiones

El algoritmo está implementado formalmente por el procedimiento recursivo siguiente:

1. Diseñar el codebook de un vector: este es el centroide de todo el conjunto de vectores de entrenamiento (no se requiere iteración aquí)
2. Duplicar el tamaño del codebook dividiendo cada y_n codebook acorde a la norma.

$$y_n^+ = y_n(1+\varepsilon)$$

$$y_n^- = y_n(1-\varepsilon)$$

   donde n varia de 1 al tamaño actual del codebook, y ε es un parámetro de división.
3. Búsqueda del vecino cercano: por cada vector de entrenamiento, se encuentra el codeword en el codebook actual que está más cerca, y asignar el vector a la correspondiente celda.
4. Actualizar el centroide: actualiza el codeword en cada celda utilizando el centroide de los vectores de entrenamiento asignados a esa celda.

5. Iteración 1: repetir los pasos 3 y 4 hasta que la distancia promedio caiga por debajo del umbral preestablecido.
6. Iteración 2: repetir los pasos 2, 3 y 4 hasta que el tamaño de codebook de M sea diseñado.

Intuitivamente, el algoritmo LBG diseña un codebook del vector M en cada etapa. El comienza primero con el diseño del codebook de un vector, después usa una técnica de división en los codeword para inicializar la búsqueda en el codebook de dos vectores, así sucesivamente hasta llegar al codebook de M vectores (23).

2.3.3 Distancia Euclidiana

En la fase de prueba se debe medir la distancia entre el modelo de la voz de un locutor desconocido y los modelos calculados previamente en la fase de entrenamiento.

Carlos Domínguez Sánchez en su tesis de maestria Speaker Recognition in a handheld computer (22), explica que partiendo de tener dos modelos de locutores A y B, un enfoque viable para medir la distancia entre A y B es la medición de la distancia euclidiana entre cada vector de características del locutor A y el mayor vector de características del locutor B, entonces se normaliza la distancia dividiendo por el número de vectores en el modelo A.

En (30) se define el cálculo de la distancia euclidiana entre dos puntos como:

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

Ecuación 6 Cálculo de la distancia euclidiana entre dos puntos

Donde P = (p_1, p_2,…, p_n) y Q = (q_1, q_2,…, q_n) son los puntos para medir la distancia.

Si la longitud del libro de códigos (codebook) es 1, entonces cada modelo es un vector que contiene la media de los coeficientes Mel de cada trama, y la distancia entre los modelos es la distancia elucídela entre los vectores. Por lo tanto ahora estamos en condiciones de reconocer el orador desconocido como uno de los oradores que se inscribieron durante la fase de entrenamiento (22).

Capítulo 2. Caracterización de la propuesta de solución

La figura 7 muestra el flujo final de la propuesta de solución descrita en este capítulo:

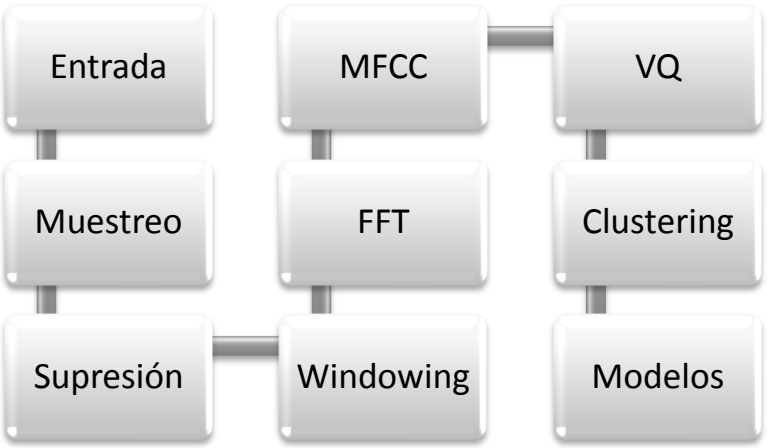

Ilustración 7 Flujo de procesos de la propuesta de solución para el reconocimiento de locutores.

Capítulo 2. Caracterización de la propuesta de solución

2.4 Conclusiones

- Con la caracterización de la propuesta de solución descrita en este capítulo se lograron definir los pasos para el reconocimiento de locutores.
- Los coeficientes cepstrales en escala Mel (MFCC) se utilizaron para la extracción de características, el algoritmo de Cuantificación Vectorial (VQ) como algoritmo de reconocimiento y el algoritmo LBG como algoritmo de agrupamiento.

Capítulo 3. Validación de la propuesta de solución

3.1 Introducción

El presente capítulo tiene como objetivo la validación del flujo o secuencia de pasos a seguir para la detección de locutores seleccionados. Para esto se utilizó el programa MATLAB que no es más que un entorno de computación y desarrollo de aplicaciones totalmente integrado, orientado para llevar a cabo proyectos en donde se encuentren implicados elevados cálculos matemáticos y la visualización gráfica de los mismos. MATLAB dispone también de un amplio abanico de programas de apoyo especializados, denominados Toolboxes. VOICEBOX es una herramienta para el procesamiento del habla consistente en rutinas en MATLAB. Las rutinas están disponibles como un archivo .zip y se hizo disponible bajo los términos de la Licencia Pública GNU.

3.2 Validación de la propuesta de solución

Para la validación de la propuesta se utilizaron 3 bases de datos, la primera con 8 sujetos, la segunda con 3 sujetos, y la tercera con 11 sujetos. Las bases de datos se dividen en dos carpetas, una para guardar los sujetos a entrenar y la otra para guardar los sujetos a comparar. Se utilizó el programa matemático MATLAB, ya que el mismo tiene incorporado funciones que se utilizan para el reconocimiento de locutores y que se utilizaron para la validación de la propuesta de solución, además de las creadas. Se utilizan dos funciones principales, *EntrenarDatos* y *ProbarDatos*, las cuales se utilizan para entrenar y probar los datos respectivamente, además de la función *mfcc,* la cual realiza la extracción de características, la función *vqlbg*, es la que permite la comparación de los vectores, la función *distancia*, calcula la distancia euclidiana entre los modelos entrenados y los modelos de prueba, y la función *melbf*, se utiliza para medir la amplitud de los bancos de filtros. La validación consta de dos etapas: la etapa de entrenamiento y la etapa de reconocimiento.

3.2.1 Etapa de entrenamiento

El primer paso es entrenar los sujetos que se encuentran en la carpeta “train” que contiene cada base de datos. Para esto se utiliza la función EntrenarDatos(directorio,cant), donde el parámetro *directorio* se refiere a la dirección en donde se encuentra la carpeta con los sujetos a entrenar, y *cant* la cantidad

de sujetos que están en el directorio. La respuesta a esto es el conjunto de vectores entrenados de cada sujeto, listos para comparar. Después de esta etapa de entrenamiento el sistema tiene conocimiento de las características de la voz de cada hablante. La ilustración 8 muestra el proceso de entrenar los datos a la base de datos de 8 sujetos haciendo uso de la herramienta MATLAB.

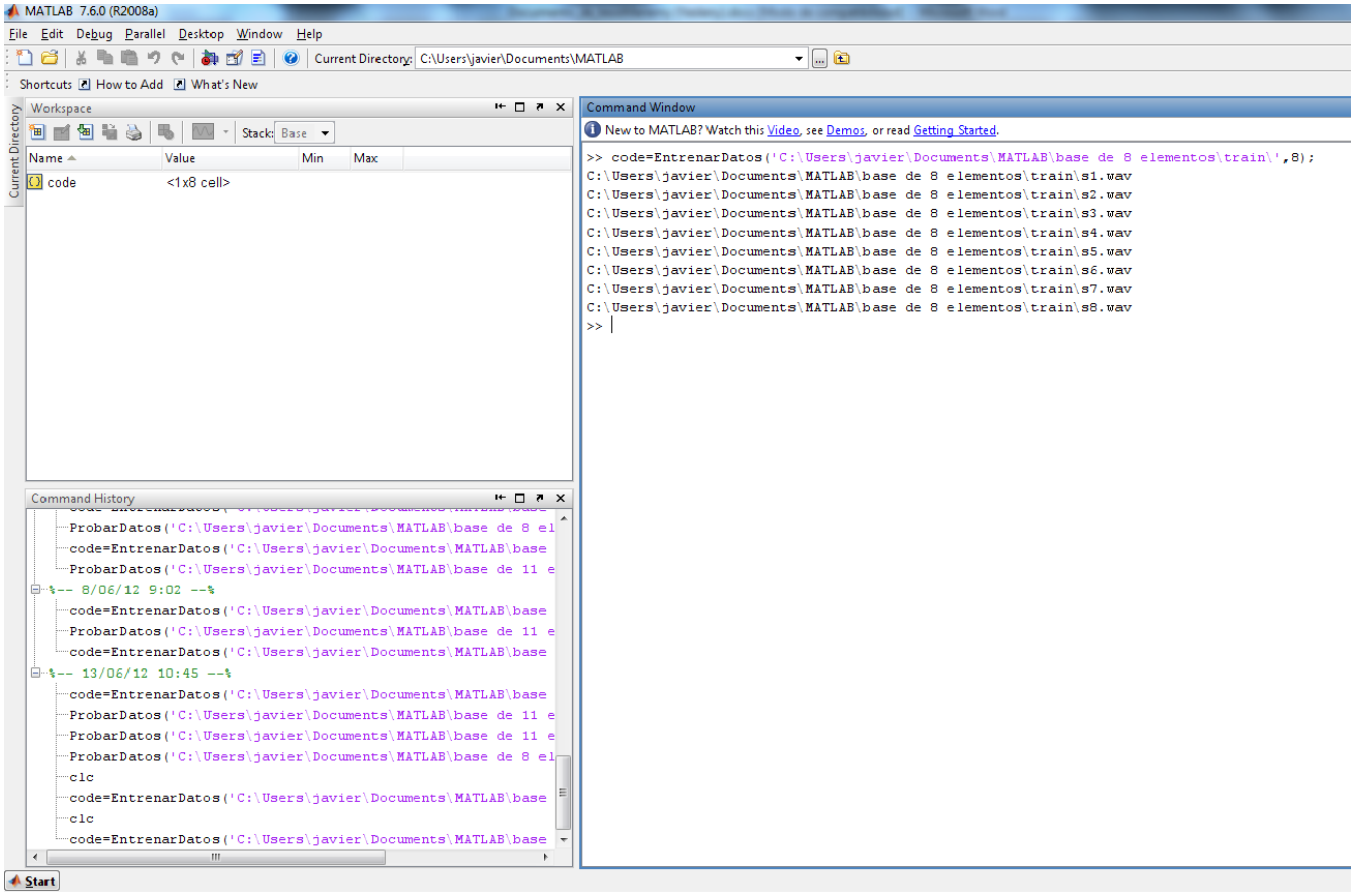

Ilustración 8 Proceso de entrenamiento

3.2.2 Etapa de reconocimiento

El próximo paso es probar que los sujetos entrenados se corresponden con los sujetos a comparar, es decir, de reconocer a los sujetos. Los sujetos a comparar se encuentran en una carpeta llamada “test” contenida en cada base de datos. Para esto se utiliza la función ProbarDatos(directorio, n, code), donde *directorio* es la dirección que contiene los archivos de prueba, *n* la cantidad de sujetos que están en el directorio, y *code* son los vectores anteriormente entrenados. La respuesta son los

locutores identificados. La ilustración 9 muestra el proceso de prueba haciendo uso de la herramienta MATLAB.

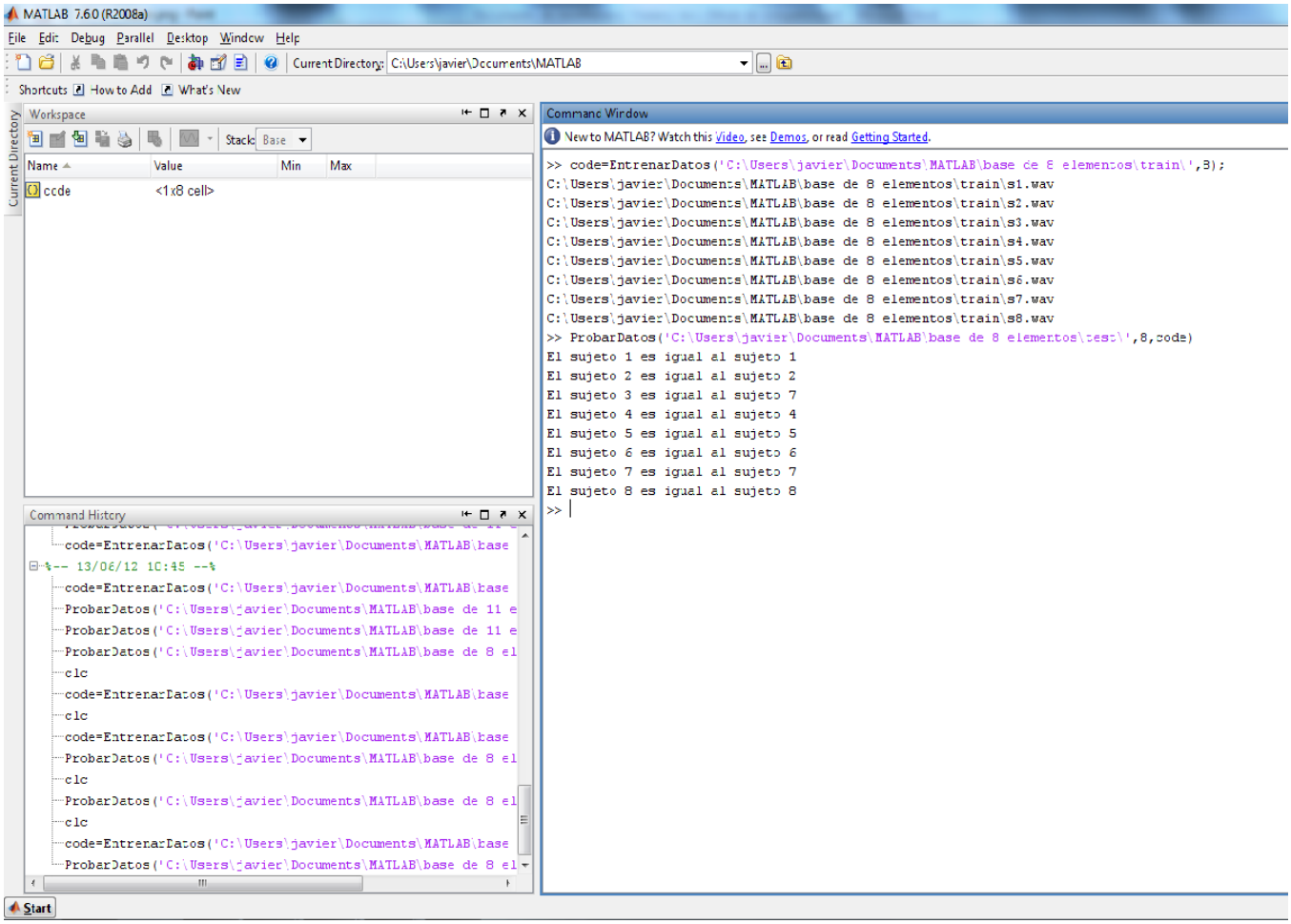

Ilustración 9 Proceso de prueba

Como se puede observar en la ilustración 9 el sistema reconoce a 7 de los 8 sujetos, por lo que el porciento de aciertos para el mismo es de un 87.5%. Se probaron también con las bases de datos de 3 y de 11 sujetos, arrojando los siguientes resultados:

- Para la base de datos con 3 sujetos el sistema reconoció a los 3 sujetos
- Para la base de datos con 11 sujetos el sistema reconoció a 10 sujetos.

La tabla 2 muestra el por ciento de acierto del sistema con la utilización de las 3 bases de datos:

Capítulo 3. Validación de la propuesta de solución

Cantidad de sujetos	Total de aciertos	Porciento de aciertos
3	3	100%
8	7	87.5%
11	10	91.0%

Tabla 2Porciento de aciertos para el algoritmo LBG

Estos resultados permiten observar que, mientras menos sujetos se tengan la identificación tendrá un mayor porciento de éxito. Además, aunque para la base de datos con 11 sujetos el sistema reconoce a 10 sujetos, el porciento de aciertos sigue siendo elevado en comparación con la base de datos de 8 sujetos.

3.3 Comparación de la propuesta de solución utilizando el algoritmo de clustering LBG y la propuesta de solución utilizando el algoritmo de clustering K-means.

Según (24) y (29), el algoritmo K-means es una forma de agrupar los vectores de entrenamiento para obtener vectores de características. En este algoritmo se agrupan los vectores basados en atributos en k particiones. Se utiliza el medio k de los datos generados para agrupar los vectores. El proceso del algoritmo K-means comienza con la partición del conjunto inicial en K grupos o clases. Después se calcula el punto medio, o centroide, de cada conjunto. Se construye una nueva partición mediante la asociación de cada punto con el más cercano centroide. Luego se vuelven a calcular los centroides de los nuevos grupos, y el algoritmo se repite hasta alcanzar un determinado umbral o número de iteraciones.

Para la realización de esta comparación se usaron 3 bases de datos, la primera con 8 sujetos, la segunda con 3 sujetos, y la tercera con 11 sujetos. La comparación se hizo mediante el cálculo del porciento de aciertos como se muestra en la figura:

$$PA = \frac{\#\ aciertos}{total} \times 100\%$$

Ecuación 7Cálculo del Porciento de Aciertos

Primero se realizó la prueba de la propuesta de solución utilizando el algoritmo de clustering LBG arrojando los siguientes resultados:

Cantidad de sujetos	Total de aciertos	Porciento de aciertos
3	3	100%
8	7	87.5%
11	10	91.0%

Tabla 3Porciento de aciertos para el algoritmo LBG

Las pruebas de la propuesta de solución utilizando el algoritmo de clustering K-means arrojaron los siguientes resultados:

Cantidad de sujetos	Total de aciertos	Porciento de aciertos
3	2	66,6%
8	2	25.0%
11	3	27,3%

Tabla 4Porciento de acierto para el algoritmo K-means

Como se puede observar la utilización del algoritmo LBG como algoritmo para el clustering en la propuesta de solución arroja mejores resultados que la utilización del algoritmo K-means para el clustering. Esta prueba ayuda a reafirmar la decisión tomada en el epígrafe 2.3.2, de utilizar el algoritmo LBG como algoritmo de clustering.

3.4 Comparación de diferentes implementaciones de MFCC

Según (28)el rendimiento de los coeficientes de frecuencia Mel-Cepstrum (MFCC) puede ser afectado por el número de filtros. En este trabajo, se realizan varios experimentos de comparación para medir el rendimiento del reconocimiento de locutor variando el número de filtros de MFCC a 12, 22, 32 y 42. El reconocedor alcanza el rendimiento máximo en el número de filtro K = 32.

Hablante	No. de intentos	Falsa Aceptación	Falso Rechazo
S1	4	0	0
S2	4	0	1
S3	4	0	2
S4	4	0	0

S5	4	0	2
Total	20	0	5

Tabla 5Implementación de MFCC con 12 filtros.

Hablante	No. de intentos	Falsa Aceptación	Falso Rechazo
S1	4	0	0
S2	4	0	2
S3	4	0	2
S4	4	0	0
S5	4	0	3
Total	20	0	7

Tabla 6 Implementación de MFCC con 22 filtros.

Hablante	No. de intentos	Falsa Aceptación	Falso Rechazo
S1	4	0	0
S2	4	0	0
S3	4	0	1
S4	4	0	0
S5	4	0	2
Total	20	0	3

Tabla 7 Implementación de MFCC con 32 filtros.

Hablante	No. de intentos	Falsa Aceptación	Falso Rechazo
S1	4	0	0
S2	4	0	0
S3	4	0	2
S4	4	0	1

S5	4	0	1
Total	20	0	4

Tabla 8 Implementación de MFCC con 12 filtros.

Capítulo 3. Validación de la propuesta de solución

3.5 Conclusiones parciales

- Según los resultados que se muestran en la tabla 2 el algoritmo LBG, para 3 sujetos tiene un 100% de aciertos, para 8 sujetos, un 87,5% de aciertos y para 11 sujetos un 100% de aciertos. Sin embargo, según como se muestra en la tabla 3, para el algoritmo K-means, con 3 sujetos muestra un 66,6% de aciertos, para 8 sujetos, un 25% de aciertos, y para 11 sujetos un 27, 3% de aciertos.
- Para la implementación del MFCC se recomienda la utilización de 32 filtros, ya que fueron los que mejores resultados proyectaron, además esa cantidad permite aumentar la precisión y no requiere de mucho más tiempo.
- La utilización de la Cuantificación Vectorial es una manera simple y eficiente de realizar la identificación de locutores.

Conclusiones Generales

- Según el estudio que se realizó sobre las soluciones existentes, se identificó que no existe ningún programa que se utilice con el fin de reconocer locutores dentro de una transmisión radial, sino que se especializan en otras áreas como la investigación forense y en la seguridad y control de sistemas.
- Para la creación de un algoritmo que reconozca locutores se deben extraer las características del hablante, realizar una selección de sus mejores características, y por último clasificar dichas características.
- Los coeficientes cepstrales en escala Mel (MFCC) se utilizaron para la extracción de características, el algoritmo de Cuantificación Vectorial (VQ) como algoritmo de reconocimiento, y el algoritmo LBG como algoritmo de agrupamiento.

Recomendaciones

Una vez vencidos los objetivos de esta investigación, y teniendo en cuenta las experiencias obtenidas a lo largo de su desarrollo se recomienda:

- Aplicar la propuesta de solución a una base de datos con un mayor número de sujetos para arribar a conclusiones más precisas.
- Mejorar la propuesta de solución en términos de ruido para aumentar la precisión en malas condiciones.
- Diseñar una herramienta que permita monitorizar a los locutores en las transmisiones radiales.

Bibliografía Referenciada

1. Facultad de Ciencias Exactas, Ingeniería y Agrimensura. *Facultad de Ciencias Exactas, Ingeniería y Agrimensura.* [En línea] 2010. [Citado el: 12 de Noviembre de 2011.] www.fceia.unr.edu.ar/prodivoz/speaker_verification.pdf.

2. **Elizalde, Cristina Esteve.** *RECONOCIMIENTO DE LOCUTOR DEPENDIENTE DE TEXTO MEDIANTE ADAPTACIÓN DE MODELOS OCULTOS DE MARKOV FONÉTICOS.* Madrid : s.n., 2007.

3. *Métodos de extracción, selección, y clasificación.* **Calvo, José Ramón, Fernádez, Rafael y Hernández, Gabiel.** 2142, La Habana : s.n., 2008. ISSN 2072-6287.

4. **Crhistyan Czech, Fabian Miodownik, Alexis Ravaschio.** *Reconocimiento de locutores a partir de archivos en formato MP3.* 2005.

5. radiocubana. *radiocubana.* [En línea] [Citado el: 16 de Noviembre de 2011.] http://www.radiocubana.cu/index.php/articulos-especializados-sobre-la-radio/50-nuevas-tecnologias/1471-radio-cubana-para-oirte-mejor.

6. **Cabeceran, Mireia Farrús i.** *FUSING PROSODIC AND ACOUSTIC INFORMATION FOR SPEAKER RECOGNITION.* Barcelona : s.n., 2008.

7. Verbio Speaker Technologies. *Verbio Speaker Technologies.* [En línea] Speaker Tecchnologies S.L. [Citado el: 13 de Junio de 2012.] http://www.verbio.com/webverbio3/index.php/es/tecnologia/verbio-asr.html.

8. mobile2010. *mobile2010.* [En línea] [Citado el: 23 de Noviembre de 2011.] http://mobile2010.b2bmatchmaking.com/index.php?page=cat_tech&action=detail&params[id]=161.

9. it.kth. *it.kth.* [En línea] [Citado el: 20 de Enero de 2012.] http://web.it.kth.se/~maguire/DEGREE-PROJECT-REPORTS/101117-Carlos_Dominguez-with-cover.pdf.

10. **Kinnunen, T.** *Spectral Features for Automatic Text-Independent Speaker Recognition. .* Finland : Department of Computer Science, University of Joensuu., 2003.

11. **Ramachandran, R., Farrell, K., Ramachandran R. y Mammone R.** *Speaker recognition – general classifier approaches and data fusion methods.* 2002.

12. **Higgins, J. E., Damper, R. I. y Harris, C. J.** *A Multi-Spectral Data Fusion Approach to Speaker Recognition.* 1998.

13. **Atal, B.S. y Schroeder, M.R.** *Predictive Coding of Speech Signals. Report of the 6th Int. Congress on Acoustics.* Tokyo : s.n., 1968.

Bibliografía Referenciada

14. **Atal, B.S.** *Automatic recognition of speakers from their voices.* 1976.

15. **CAMPBELL, JOSEPH P.** *Speaker Recognition: A Tutorial.* USA : s.n., 1997. 0018-9219(97)06947-8.

16. **Tufekci, Z. y Gowdy, J.N.** *Feature extraction using discrete wavelet transform for speech recognition.* USA, : Proc. of the IEEE SoutheastCon, 2000.

17. *High resolution speech feature parameterization for monophone based stressed speech recognition.* **Sarikaya, R. y Hansen, H.L.** 7, USA : IEEE Signal Proc. Letters,, 2000., Vol. 7.

18. **Duda, R. O., Hart, P. E. y Stork, D. G.** *Pattern Classification.* s.l. : John Wiley and Sons, Inc., 2001.

19. **Furui, S.** *An overview of Speaker Recognition Technology.* 1994.

20. **Morgan, D.B. y Scofield, C.L.** *Neural Networks and Speech Processing.* s.l. : Kluwer Academic Publishers, 1991.

21. **Llerena, Y.** *State of the art in Speaker Recognition.* Ciego de Ávila : s.n., 2006.

22. **SÁNCHEZ, CARLOS DOMÍNGUEZ.** *Speaker Recognition in a handheld computer.* Stockholm : s.n., 2010. TRITA-ICT-EX-2010:285.

23. *Design Of An Automatic Speaker Recognition System Using MFCC, Vector Quantization And LBG Algorithm.* **Kumar, P. Mallikarjuna Rao and Ch.Srinivasa.** 8, 2011, Vol. 3. ISSN : 0975-3397.

24. **Feng, Ling.** *Speaker Recognition.* Lyngby, Denmark : s.n., 2004. ISSN 1601-233X.

25. **Lasse L Mølgaard, Kasper W Jørgensen.** *Speaker Recognition.* 2005.

26. **DARSHAN MANDALIA, PRAVIN GARETA.** *Speaker Recognition Using MFCC and Vector Quantization Model.* Ahmedabad : s.n., 2011.

27. **Srinivasan, A.** *Speaker Identification and Verification using Vector Quantization and Mel Frequency Cepstral Coefficients.* India : s.n., 2012. ISSN: 2040-7467.

28. **Tiwari, Vibha.** *MFCC and its applications in speaker recognition.* INDIA : s.n., 2010. ISSN 0975-8364.

29. **Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman.** *SPEAKER IDENTIFICATION USING MEL FREQUENCY CEPSTRAL COEFFICIENTS .* Dhaka, Bangladesh : s.n., 2004. ISBN 984-32-1804-4.

30. speaker-recognition.googlecode. *speaker-recognition.googlecode.* [En línea] [Citado el: 15 de Enero de 2012.] http://speaker-recognition.googlecode.com/files/Finally_version1.pdf.

31. **Sriram Bhattaru, Amit Kumar.** *Text Independent Speaker Recognition Using MFCC Technique an Vector Quantization using LBG Algorithm.* Rourkela : s.n.

Bibliografía Referenciada

32. *Vector Quantization Approach for Speaker Recognition using MFCC and Inverted MFCC.* **Rajan, Satyanand Singh and E.G.** 1, Hyderabad : s.n., 2011, Vol. 17. 0975 – 8887.

33. Real Academia Española. *Real Academia Española.* [En línea] [Citado el: 22 de Noviembre de 2011.] http://buscon.rae.es/draeI/SrvltConsulta?TIPO_BUS=3&LEMA=patr%C3%B3n.

34. Real Academia Española. *Real Academia Española.* [En línea] [Citado el: 20 de Noviembre de 2011.] http://buscon.rae.es/draeI/SrvltConsulta?TIPO_BUS=3&LEMA=monitorizar.

35. Real Academia Española. *Real Academia Española.* [En línea] [Citado el: 21 de Noviembre de 2011.] http://buscon.rae.es/draeI/SrvltConsulta?TIPO_BUS=3&LEMA=locutor.

60. Real Academia Española. *Real Academia Española.* [En línea] [Citado el: 11 de Junio de 2012.] http://buscon.rae.es/draeI/SrvltConsulta?TIPO_BUS=3&LEMA=correlaci%C3%B3n.

61. Departamento de Estadística. *Departamento de Estadística.* [En línea] [Citado el: 11 de Junio de 2012.] http://halweb.uc3m.es/esp/Personal/personas/imolina/MiDocencia/SeriesTemporales/Tema3SeriesEstud.pdf.

62. Instituto Tecnológico de Chihuahua. *Instituto Tecnológico de Chihuahua.* [En línea] [Citado el: 10 de Junio de 2012.] http://www.itch.edu.mx/academic/industrial/sabaticorita/_private/07Procesos%20estocasticos.htm.

63. Estudio Marhea. *Estudio Marhea.* [En línea] 2004. [Citado el: 10 de Junio de 2012.] http://www.estudiomarhea.net/.

64. **Fuentes, Luis Javier Rodríguez.** *Estudio comparativo de varias representaciones paramétricas para el reconocimiento automático del habla.* Bilbao : s.n., 1994. DEE-I/2/94.

65. *Diccionario Manual de la Lengua Española.* s.l. : Larousse Editorial, S.L., 2007.

66. **R., Jorge A. Olivera.** *Convolución: Un proceso natural en los sistemas lineales e invariantes en el tiempo.* [Documento] Monterrey : s.n.

67. **Oneisys Núñez Cuadra, José Ramón Calvo de Lara.** *Métodos de extracción de rasgos para la identificación del idioma: estado del arte.* Ciudad de La Habana : s.n., 2010. ISSN 2072-6287.

68. CENTRO ESPECIALIZADO EN LENGUAJE Y APRENDIZAJE. *CENTRO ESPECIALIZADO EN LENGUAJE Y APRENDIZAJE.* [En línea] [Citado el: 14 de Junio de 2012.] http://www.nataliacalderon.com/suprasegmental-g-81.xhtml.

Bibliografía Consultada

1. Facultad de Ciencias Exactas, Ingeniería y Agrimensura. *Facultad de Ciencias Exactas, Ingeniería y Agrimensura.* [En línea] 2010. [Citado el: 12 de Noviembre de 2011.] www.fceia.unr.edu.ar/prodivoz/speaker_verification.pdf.

2. **Elizalde, Cristina Esteve.** *RECONOCIMIENTO DE LOCUTOR DEPENDIENTE DE TEXTO MEDIANTE ADAPTACIÓN DE MODELOS OCULTOS DE MARKOV FONÉTICOS.* Madrid : s.n., 2007.

3. *Métodos de extracción, selección, y clasificación.* **Calvo, José Ramón, Fernádez, Rafael y Hernández, Gabiel.** 2142, La Habana : s.n., 2008. ISSN 2072-6287.

4. **Crhistyan Czech, Fabian Miodownik, Alexis Ravaschio.** *Reconocimiento de locutores a partir de archivos en formato MP3.* 2005.

5. radiocubana. *radiocubana.* [En línea] [Citado el: 16 de Noviembre de 2011.] http://www.radiocubana.cu/index.php/articulos-especializados-sobre-la-radio/50-nuevas-tecnologias/1471-radio-cubana-para-oirte-mejor.

6. **Cabeceran, Mireia Farrús i.** *FUSING PROSODIC AND ACOUSTIC INFORMATION FOR SPEAKER RECOGNITION.* Barcelona : s.n., 2008.

7. Verbio Speaker Technologies. *Verbio Speaker Technologies.* [En línea] Speaker Tecchnologies S.L. [Citado el: 13 de Junio de 2012.] http://www.verbio.com/webverbio3/index.php/es/tecnologia/verbio-asr.html.

8. mobile2010. *mobile2010.* [En línea] [Citado el: 23 de Noviembre de 2011.] http://mobile2010.b2bmatchmaking.com/index.php?page=cat_tech&action=detail&params[id]=161.

9. it.kth. *it.kth.* [En línea] [Citado el: 20 de Enero de 2012.] http://web.it.kth.se/~maguire/DEGREE-PROJECT-REPORTS/101117-Carlos_Dominguez-with-cover.pdf.

10. **Kinnunen, T.** *Spectral Features for Automatic Text-Independent Speaker Recognition. .* Finland : Department of Computer Science, University of Joensuu., 2003.

11. **Ramachandran, R., Farrell, K., Ramachandran R. y Mammone R.** *Speaker recognition – general classifier approaches and data fusion methods.* 2002.

12. **Higgins, J. E., Damper, R. I. y Harris, C. J.** *A Multi-Spectral Data Fusion Approach to Speaker Recognition.* 1998.

13. **Atal, B.S. y Schroeder, M.R.** *Predictive Coding of Speech Signals. Report of the 6th Int. Congress on Acoustics.* Tokyo : s.n., 1968.

14. **Atal, B.S.** *Automatic recognition of speakers from their voices.* 1976.

15. **CAMPBELL, JOSEPH P.** *Speaker Recognition: A Tutorial.* USA : s.n., 1997. 0018-9219(97)06947-8.

Bibliografía Consultada

16. **Tufekci, Z. y Gowdy, J.N.** *Feature extraction using discrete wavelet transform for speech recognition.* USA, : Proc. of the IEEE SoutheastCon, 2000.

17. *High resolution speech feature parameterization for monophone based stressed speech recognition.* **Sarikaya, R. y Hansen, H.L.** 7, USA : IEEE Signal Proc. Letters,, 2000., Vol. 7.

18. **Duda, R. O., Hart, P. E. y Stork, D. G.** *Pattern Classification.* s.l. : John Wiley and Sons, Inc., 2001.

19. **Furui, S.** *An overview of Speaker Recognition Technology.* 1994.

20. **Morgan, D.B. y Scofield, C.L.** *Neural Networks and Speech Processing.* s.l. : Kluwer Academic Publishers, 1991.

21. **Llerena, Y.** *State of the art in Speaker Recognition.* Ciego de Ávila : s.n., 2006.

22. **SÁNCHEZ, CARLOS DOMÍNGUEZ.** *Speaker Recognition in a handheld computer.* Stockholm : s.n., 2010. TRITA-ICT-EX-2010:285.

23. *Design Of An Automatic Speaker Recognition System Using MFCC, Vector Quantization And LBG Algorithm.* **Kumar, P. Mallikarjuna Rao and Ch.Srinivasa.** 8, 2011, Vol. 3. ISSN : 0975-3397.

24. **Feng, Ling.** *Speaker Recognition.* Lyngby, Denmark : s.n., 2004. ISSN 1601-233X.

25. **Lasse L Mølgaard, Kasper W Jørgensen.** *Speaker Recognition.* 2005.

26. **DARSHAN MANDALIA, PRAVIN GARETA.** *Speaker Recognition Using MFCC and Vector Quantization Model.* Ahmedabad : s.n., 2011.

27. **Srinivasan, A.** *Speaker Identification and Verification using Vector Quantization and Mel Frequency Cepstral Coefficients.* India : s.n., 2012. ISSN: 2040-7467.

28. **Tiwari, Vibha.** *MFCC and its applications in speaker recognition.* INDIA : s.n., 2010. ISSN 0975-8364.

29. **Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman.** *SPEAKER IDENTIFICATION USING MEL FREQUENCY CEPSTRAL COEFFICIENTS .* Dhaka, Bangladesh : s.n., 2004. ISBN 984-32-1804-4.

30. speaker-recognition.googlecode. *speaker-recognition.googlecode.* [En línea] [Citado el: 15 de Enero de 2012.] http://speaker-recognition.googlecode.com/files/Finally_version1.pdf.

31. **Sriram Bhattaru, Amit Kumar.** *Text Independent Speaker Recognition Using MFCC Technique an Vector Quantization using LBG Algorithm.* Rourkela : s.n.

32. *Vector Quantization Approach for Speaker Recognition using MFCC and Inverted MFCC.* **Rajan, Satyanand Singh and E.G.** 1, Hyderabad : s.n., 2011, Vol. 17. 0975 – 8887.

33. Real Academia Española. *Real Academia Española.* [En línea] [Citado el: 22 de Noviembre de 2011.] http://buscon.rae.es/draeI/SrvltConsulta?TIPO_BUS=3&LEMA=patr%C3%B3n.

Bibliografía Consultada

34. Real Academia Española. *Real Academia Española.* [En línea] [Citado el: 20 de Noviembre de 2011.] http://buscon.rae.es/draeI/SrvltConsulta?TIPO_BUS=3&LEMA=monitorizar.

35. Real Academia Española. *Real Academia Española.* [En línea] [Citado el: 21 de Noviembre de 2011.] http://buscon.rae.es/draeI/SrvltConsulta?TIPO_BUS=3&LEMA=locutor.

36. Interactive and Coperative Technologies Lab. [En línea] http://ict.udlap.mx/people/ingrid/Clases/IS412/index.html.

37. forensicscience. *forensicscience.* [En línea] [Citado el: 23 de Noviembre de 2011.] www.forensicscience.pl/pfs/47_gonzales.pdf.

38. cslu.ogi. *.cslu.ogi.* [En línea] [Citado el: 24 de Noviembre de 2011.] http://www.cslu.ogi.edu/HLTsurvey/ch1node9.html.

39. **Maider Zamalloa, Germán Bordel, Luis Javier Rodriguez, Mikel Peñagarikano, Juan Pedro Uribe.** *SELECCIÓN Y PESADO DE PARÁMETROS ACÚSTICOS MEDIANTE ALGORITMOS GENÉTICOS PARA EL RECONOCIMIENTO DEL LOCUTOR.* Pais Vasco : s.n., 2006.

40. **NEIRA, ELKIN RAMON GARAVITO.** *MODELO DE IDENTIFICACIÓN DE LOCUTOR EN ENTORNOS GSM, APLICACIÓN EN COLOMBIA. .* Bogotá, D.C. : s.n., 2010.

41. **E. Martínez Torrico, J. González Rodríguez, J. Ortega García.** *Reconocimiento de Locutores con GMMs y HMMs basado en Número de Identificación Personal (PIN).* Madrid : s.n.

42. **Rodríguez, José Luis Oropeza.** *Algoritmos y Métodos para el Reconocimiento de Voz en Español Mediante Sílabas .* México D.F. : s.n., 2006.

43. **Beigi, Homayoon.** *Fundamentals of Speaker Recognition.* NY : Springer New York Dordrecht Heidelberg London, 2011. ISBN 978-0-387-77591-3.

44. **Liu, Ying y Pearlman, William A.** *MULTISTAGE LATTICE VECTOR QUANTIZATION FOR HYPERSPECTRAL IMAGE COMPRESSION.* NY : s.n.

45. **Bryan L. Pellom, John H.L. Hansen.** *An Efficient Scoring Algorithm for Gaussian Mixture Model based Speaker Identification.* North Carolina : s.n., 1997. SPL.SA.1.8.

46. *A Novel Full-Search Vector Quantization Algorithm Based on the Law of Cosines.* **Mielikainen, Jarno.** 6, 2002, Vol. IX. 10.1109/LSP.2002.800507.

47. *Methods to Accelerate a Competitive Learning Algorithm Applied to VQ Codebook Design.* **JR, E.L. BISPO, y otros.** 3, Brasil : s.n., 2010, Vol. XI. 193-203.

48. **LABRADOR, ENRIQUE PRIETO.** *ESTUDIO COMPARATIVO DE PARÁMETROS ESPECTRALES PARA CLASIFICACIÓN DE AUDIO.* MADRID : s.n., 2008.

49. **UREÑA, RUBÉN SOLERA.** *MÁQUINAS DE VECTORES SOPORTE PARA RECONOCIMIENTO ROBUSTO DE HABLA.* MADRID : s.n., 2011.

Bibliografía Consultada

50. **Fernández, Laura Docóo.** *Aportaciones a la Mejora de los Sistemas de Reconocimiento.* 2001.

51. cnx.org. *cnx.org.* [En línea] http://cnx.org/content/m14201/1.3/.

52. msdn.microsoft. *msdn.microsoft.* [En línea] [Citado el: 15 de Marzo de 2012.] http://msdn.microsoft.com/en-us/vstudio/cc482921.aspx.

53. **Pericas, Francisco Javier Hernando.** *TECNICAS DE PROCESADO Y REPRESENTACION DE LA SEÑAL DE VOZ PARA EL RECONOCIMIENTO DEL HABLA EN AMBIENTES RUIDOSOS .* Barcelona : s.n., 1993.

54. my.fit. *my.fit.* [En línea] [Citado el: 13 de Abril de 2012.] http://my.fit.edu/~vkepuska/ece5526/TIMIT_Corpus/MATLAB/voicebox/.

55. *Reconocimiento del locutor dependiente del texto con modelos acusticos del habla.* **Alfonso, Ivis Rodés y Lara, José Ramón Calvo de.** 2142, Ciudad de La Habana : s.n., 2009. ISSN 2072-6287.

56. *Métodos de compensacion de ruido en reconocimiento de locutores.* **González, Dayana Ribas y Lara, José R. Calvo de.** 2142, Ciudad de La Habana : s.n., 2010. ISSN 2072-6287.

57. *New Clustering Algorithm for Vector Quantization using Rotation of Error Vector .* **Sarode, H. B. Kekre and Tanuja K.** 3, Mumbai : s.n., 2010, Vol. 7. ISSN 1947-5500.

58. *A Fast Search Algorithm for Vector Quantization Using Mean Pyramids of Codewords .* **Chen, Chang-Hsing Lee and Ling-Hwei.** 2/3/4, 1995, Vol. 43. 0090-6778/95$.

59. *Speaker Verification Using MFCC and Support Vector Machine .* **Luo, Shi-Huang Chen and Yu-Ren.** Hong Kong : s.n., 2009, Vol. 1. ISBN: 978-988-17012-2-0.

60. Real Academia Española. *Real Academia Española.* [En línea] [Citado el: 11 de Junio de 2012.] http://buscon.rae.es/draeI/SrvltConsulta?TIPO_BUS=3&LEMA=correlaci%C3%B3n.

61. Departamento de Estadística. *Departamento de Estadística.* [En línea] [Citado el: 11 de Junio de 2012.] http://halweb.uc3m.es/esp/Personal/personas/imolina/MiDocencia/SeriesTemporales/Tema3SeriesEstud.pdf.

62. Instituto Tecnológico de Chihuahua. *Instituto Tecnológico de Chihuahua.* [En línea] [Citado el: 10 de Junio de 2012.] http://www.itch.edu.mx/academic/industrial/sabaticorita/_private/07Procesos%20estocasticos.htm.

63. Estudio Marhea. *Estudio Marhea.* [En línea] 2004. [Citado el: 10 de Junio de 2012.] http://www.estudiomarhea.net/.

64. **Fuentes, Luis Javier Rodríguez.** *Estudio comparativo de varias representaciones paramétricas para el reconocimiento automático del habla.* Bilbao : s.n., 1994. DEE-I/2/94.

65. *Diccionario Manual de la Lengua Española.* s.l. : Larousse Editorial, S.L., 2007.

Bibliografía Consultada

66. **R., Jorge A. Olivera.** *Convolución: Un proceso natural en los sistemas lineales e invariantes en el tiempo.* [Documento] Monterrey : s.n.

67. **Oneisys Núñez Cuadra, José Ramón Calvo de Lara.** *Métodos de extracción de rasgos para la identificación del idioma: estado del arte.* Ciudad de La Habana : s.n., 2010. ISSN 2072-6287.

68. CENTRO ESPECIALIZADO EN LENGUAJE Y APRENDIZAJE. *CENTRO ESPECIALIZADO EN LENGUAJE Y APRENDIZAJE.* [En línea] [Citado el: 14 de Junio de 2012.] http://www.nataliacalderon.com/suprasegmental-g-81.xhtml.

Anexos

Los Anexos que a continuación se presentan brindan un conjunto de elementos para una futura implementación y comprensión del funcionamiento básico del reconocimiento de locutores. Se estructura de la siguiente manera:

Anexo A: Se presentan las funciones vinculadas al reconocimiento.

Anexo B: Se presentan las funciones de apoyo que brinda MATLAB

Anexo A

Función **“EntrenarDatos.m”**

Se utiliza para entrenar los datos que se encuentran en el directorio que se pasa por parámetro

```
functioncode = EntrenarDatos(directorio,cant)


k = 16;            %numero de centroides

%Fase de entrenamiento
for i = 1:cant
    file = sprintf('%ss%d.wav', directorio, i);
disp(file);

    [s, fs] = wavread(file);

signal = mfcc(s, fs);         %Calcular mfcc
code{i} = vqlbg(signal, k);   %Entrenar el codebook
end

end
```

Función **“ProbarDatos.m”**

Se utiliza para probar los datos que se encuentran en el directorio que se pasa por parámetro

```
functionProbarDatos(testdir, n, code)


for k=1:n
  file = sprintf('%ss%d.wav', testdir, k);
  [s, fs] = wavread(file);

  v = mfcc(s, fs);

distmin = inf;
  k1=0;
```

```
for l=1:length(code)
d = distancia(v, code{l});
dist = sum(min(d,[],2)) / size(d,1);

if (dist<distmin)
distmin = dist;
k1=l;
end
end
msg = sprintf('El hablante %d compara con el hablante %d', k, k1);
disp(msg);
end

end
```

Función**"distancia.m"**

Se utiliza para calcular la distancia entre los codeword de cada codebook

```
function d = distancia(x, y)

[M, N]= size(x);
[M2, P] = size(y);

if(M ~= M2)
    error('Las dimensiones de las matrices no coinciden')
end

d = zeros(N, P);
if(N < P)
    copies = zeros(1,P);
for n=1:N
d(n,:) = sum((x(:, n+copies) - y) .^2, 1);
end
else
    copies = zeros(1,N);
for p=1:P
        d(:,p) = sum((x - y(:, p+copies)) .^2, 1)';
end
```

```
end

d = d.^0.5;
end
```

Función**“mfcc.m”**

Se utiliza para calcular los MFCC, para la extracción de características.

```
function signal = mfcc(s, fs)
m = 100;
n = 256;
l = length(s);

nbFrame = floor((l - n) / m) + 1;

for i=1:n
for j=1:nbFrame
        M(i, j)= s(((j - 1) * m) + i);
end
end

h = hamming(n);
M2 = diag(h) * M;

for i=1:nbFrame
    frame(:, i) = fft(M2(:, i));
end

m = Melbf(20,n,fs);
n2 = 1 + floor(n / 2);
z = m * abs(frame(1:n2, :)).^2;

signal = dct(log(z));
end
```

Función **“Melbf.m”**

Se utiliza para convertir a frecuencias MEL.

```
function m = Melbf (nf, l, fs)
f0 = 700/fs;
fn2 = floor(l / 2);

lr = log(1 + 0.5/f0) / (nf + 1);

bl = l * (f0 * (exp([0 1 nf nf+1] * lr) - 1));

b1 = floor(bl(1)) + 1;
b2 = ceil(bl(2));
b3 = floor(bl(3));
b4 = min(fn2, ceil(bl(4))) - 1;

pf = log(1 + (b1:b4)/l/f0) / lr;
fp = floor(pf);
pm = pf - fp;

r = [fp(b2:b4) 1+fp(1:b3)];
c = [b2:b4 1:b3] + 1;
v = 2 * [1-pm(b2:b4) pm(1:b3)];

m = sparse(r, c, v, nf, 1+fn2);
end
```

Función **“vqlbg.m”**

Se utiliza para realizar la Cuantificación Vectorial, usando el algoritmo de agrupamiento LBG.

```
function v = vqlbg(d, k)
e = .01;
v = mean(d, 2);
dpr = 10000;

for i=1:log2(k)
```

```
v = [v*(1+e), v*(1-e)];

while (1 == 1)
        z = distancia(d, v);
        [m, ind] = min(z, [], 2);
t = 0;
for j=1:2^i
            v(:, j) = mean(d(:, find(ind == j)), 2);
            x = distancia(d(:, find(ind == j)), v(:, j));
for q=1:length(x)
                t = t+x(q);
end
end
if (((dpr - t) / t) < e)
break;
else
dpr = t;
end
end
end

end
```

Anexo B

Función**“wavread.m”**

Y=WAVREAD(FILE): Lee un archivo WAVE especificado por el archivo de cadena y devuelve los datos de la muestra en Y. Se adjunta el ".wav" si no se da la extensión.

```
function [sig, sf, bits]=wavread(wavefile, siz)

ifnargin< 1
  help wavread;
return;
end

ifisempty(findstr(wavefile,'.'))
wavefile=[wavefile,'.wav'];
end

fid=fopen(wavefile,'r','ieee-le');
if fid == -1
  error('Can''t open .WAV file for input!');
end;

% read riff chunk
header=fread(fid,4,'uchar');
header=fread(fid,1,'ulong');
header=fread(fid,4,'uchar');

% read format sub-chunk
header=fread(fid,4,'uchar');
header=fread(fid,1,'ulong');

format(1)=fread(fid,1,'ushort');                % Format
format(2)=fread(fid,1,'ushort');                % Channel
format(3)=fread(fid,1,'ulong');                 % Samples per second
header=fread(fid,1,'ulong');
```

```
block=fread(fid,1,'ushort');
format(4)=fread(fid,1,'ushort');                % Bits per sample

% read data sub-chunck
header=fread(fid,4,'uchar');
nbyteforsamples=fread(fid,1,'ulong');

nsamples=nbyteforsamples/block;
if(nargin< 2)
siz = [1 nsamples];
end

sf = format(3);
bits = format(4);
if(strcmp(siz, 'size'))
  sig = [nsamples, format(2)];
fclose(fid);
return;
end

st = siz(1);
et = siz(2);

if (format(4)+format(2) == 9)
fseek(fid, (st-1), 0);
  [sig, cnt] = fread(fid, [1, et-st+1],'uchar');
  sig = (sig-128)/128;
end
if (format(4)+format(2) == 10)
fseek(fid, (st-1)*2, 0);
  [sig, cnt] = fread(fid, [2,et-st+1], 'uchar');
  sig = (sig-128)/128;
end

if (format(4)+format(2) == 17)
fseek(fid, (st-1)*2, 0);
```

```
  [sig, cnt] = fread(fid, [1, et-st+1], 'short');
  sig = sig/32768;
end

if (format(4)+format(2) == 18)
fseek(fid, (st-1)*4, 0);
  [sig, cnt] = fread(fid,[2,et-st+1],'short');
  sig = sig/32768;
end
fclose(fid);

sig = sig';
```

Función **"haming.m"**

HAMMING(N): Devuelve la ventana simétrica de Hammingde N-punto en un vector columna.

```
function c = hamming (m)

if (nargin~=1)
print_usage ();
endif

if (~(isscalar (m) && (m == round (m)) && (m > 0)))
    error ('hamming: m has to be an integer > 0');
end

if (m == 1)
    c = 1;
else
    m = m - 1;
    c = 0.54 - 0.46 * cos (2 * pi * (0:m)' / m);
end

end
end
```

Funcion “dct”

Y = DCT(X): Devuelve la transformada discreta del coseno de X. El vector Y es del mismo tamaño que X y contiene los coeficientes de la transformada discreta del coseno.

```
function b=dct(a,n)

ifnargin == 0,
    error(generatemsgid('Nargchk'),'Not enough input arguments.');
end

ifisempty(a)
   b = [];
return
end

% If input is a vector, make it a column:
do_trans = (size(a,1) == 1);
ifdo_trans, a = a(:); end

ifnargin==1,
  n = size(a,1);
end
m = size(a,2);

% Pad or truncate input if necessary
if size(a,1)<n,
aa = zeros(n,m);
aa(1:size(a,1),:) = a;
else
aa = a(1:n,:);
end

% Compute weights to multiply DFT coefficients
ww = (exp(-i*(0:n-1)*pi/(2*n))/sqrt(2*n)).';
```

```
ww(1) = ww(1) / sqrt(2);

if rem(n,2)==1 || ~isreal(a), % odd case
% Form intermediate even-symmetric matrix
y = zeros(2*n,m);
y(1:n,:) = aa;
  y(n+1:2*n,:) = flipud(aa);

% Compute the FFT and keep the appropriate portion:
yy = fft(y);
yy = yy(1:n,:);

else% even case
% Re-order the elements of the columns of x
y = [ aa(1:2:n,:); aa(n:-2:2,:) ];
yy = fft(y);
ww = 2*ww;  % Double the weights for even-length case
end

% Multiply FFT by weights:
b = ww(:,ones(1,m)) .* yy;

ifisreal(a), b = real(b); end
ifdo_trans, b = b.'; end
end
```

Glosario de Términos

Identificador de voz: Aplicación que determina de entre un número de muestras a quien corresponde la muestra anónima aportada.

Locutor: Persona que habla ante el micrófono, en las estaciones de radiotelefonía, para dar avisos, noticias, programas, etc. (35)

Monitorizar: Observar mediante aparatos especiales el curso de uno o varios parámetros fisiológicos o de otra naturaleza para detectar posibles anomalías. (34)

Patrón: Modelo que sirve de muestra para sacar otra cosa igual. (33)

Patrón de voz: Modelo de voz que sirve para comparar con varias muestras de voz.

Secuencia de voz: Serie o sucesión de sonidos vocales que guardan estrecha relación entre sí.

Técnicas biométricas: Las técnicas biométricas se utilizan para medir características corporales o de comportamiento de las personas con el objeto de establecer una identidad.