Universidad de las Ciencias Informáticas

Facultad 6

TRABAJO DE DIPLOMA PARA OPTAR POR EL TÍTULO DE INGENIERO EN CIENCIAS INFORMÁTICAS

Título: Técnica automática para la detección de palabras claves sobre audio

Autor: Yanier Rodríguez López.

Tutor: Ing. Adnan Fuentes Díaz.

Curso: 2012-2013

Declaración de Autoría

Declaro que soy el único autor de este trabajo y autorizo a la Universidad de las Ciencias Informáticas a hacer uso del mismo en su beneficio.

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

Datos de Contacto

DATOS DE CONTACTO

TUTOR: Ing. Adnan Fuentes Díaz **E-mail:** afuentesd@uci.cu.

- Ingeniero en Ciencias Informáticas, Universidad de Ciencias Informáticas, 2011.
- Vinculado al Proyecto Productivo SGPM.
- Rol de arquitecto y desarrollador dentro del proyecto productivo.

Resumen

Hoy en día la creciente disponibilidad y constante intercambio de información contenida en archivos audiovisuales genera el uso de mejores sistemas de recuperación de información. El presente trabajo tiene como objetivo proponer una técnica para el procesamiento de señales de audio con el propósito de recuperación de la información, utilizando la detección de palabras clave.

Los sistemas para la detección de palabras clave procesan la señal de voz por medio de modelos del lenguaje (palabras, fonemas, sílabas) utilizando para esto una serie de descriptores que extraen las características particulares de la misma. Entre estos descriptores el más usado para describir las particularidades de la voz son los Coeficientes Cepstrales en escala Mel o *Mel Frecuency Cepstral Coefficients* (MFCC por sus siglas en inglés). Además se utilizó algoritmos y descriptores para la detección de extremos de la señal de voz. Esto es una fase muy útil en la cual se obtiene de una señal los segmentos que aportan información significativa.

El modelado de las características fundamentales de la voz se realizó por medio de Modelos Ocultos de Markov o *Hidden Markov Models* (HMM por sus siglas en inglés) ya que constituyen un potente método estadístico para la modelación acústica del habla. Se confeccionó un modelo acústico basado en palabras.

La herramienta de ensayo utilizada fue *Matrix Laboratory* (MATLAB por sus siglas en inglés) para la implementación de los procesos de tratamiento de la señal de audio.

Se utilizaron medidas de eficiencia conocidas como *Recall, Precision*, *Accuracy* y la Tasa de Verdaderos Negativos o *True Negative Rate* (TNR por sus siglas en inglés) para el análisis de los resultados. Entre los resultados obtenidos se encuentran la implementación del algoritmo COPER para la detección de extremos de la señal de audio así como la utilización de la Energía y la Tasa de Cruces por Cero (*Zero Crossing Rate* ZCR por sus siglas en inglés) para la detección de extremos.

Palabras claves: recuperación de información, acústica, palabra clave.

Índice de contenido

ÍNDICE

INTRODUCCIÓN 1

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA SOBRE LA DETECCIÓN DE PALABRAS CLAVE EN AUDIO 7

1.1 Conceptos asociados a la investigación 7

1.1.1 Técnica 7

1.1.2 Frecuencia 7

1.1.3 Espectrograma 8

1.1.4 Acústica 8

1.1.5 Palabras clave 9

1.1.6 Recuperación de información 9

1.2 Situación problemática 10

1.3 Objeto de Estudio 11

1.3.1 Características acústicas de la voz 11

1.3.1.1 Fono y Fonema 11

1.3.1.2 Modelos fonéticos 12

1.3.1.3 Formantes 13

1.3.1.4 Aparato fonador 14

1.3.2 Reconocimiento automático del habla 14

1.3.3 Detección de palabras clave en archivos de audio 16

1.3.3.1 Detección de palabras clave basado en reconocedores de habla continua de gran vocabulario 17

1.3.3.2 Detección de palabras clave basado en modelos de relleno 18

1.3.3.3 Detección de palabras clave basado en reconocedores de sub-unidades de palabra 20

Índice de contenido

1.3.4 Procesamiento de la señal de voz 21
1.3.4.1 Ventaneo de la señal 22
1.3.5 Descriptores de la señal de habla 23
1.3.5.1 Volumen 24
1.3.5.2 Energía en Tiempo Corto (STE) 24
1.3.5.3 Tasa de Cruces por Cero (*Zero Crossing Rate*) 25
1.3.5.4 Autocorrelación 26
1.3.5.5 Coeficientes de Predicción lineal (LPC) 27
1.3.5.6 Coeficientes Cepstrum 28
1.3.5.7 Coeficientes Cepstrales en Escala Mel (MFCC) 29
1.4 Análisis de soluciones existentes 32
1.5 Técnicas para la reducción de ruido en audio 33
1.5.1 Métodos para disminuir el ruido en banda 34
1.6 Conclusiones Parciales 36
CAPÍTULO 2: PRESENTACIÓN DE LA SOLUCIÓN PROPUESTA 38
2.1 Clasificadores usados para el modelado de las señales de habla 38
2.1.1 Distorsión Dinámica del Tiempo (DTW) 39
2.1.2 Modelos Ocultos de Markov (HMM) 40
2.1.3 Cuantización Vectorial (VQ) 42
2.1.3.1 Algoritmo K-means 44
2.1.3.2 Vecino más cercano (k-Nearest Neighbor) 45
2.1.4 Redes Neuronales (ANN) 46
2.1.5 Modelos de Mixturas Gausianas (GMM) 47
2.2 Descripción de la solución propuesta 49

Índice de contenido

2.2.1 Detección de extremos de la señal 51
2.2.1.1 Método basado en Energía y ZCR 52
2.2.1.2 Método basado en el algoritmo COPER 54
2.2.2 Clasificación basada en HMM 54
2.2.2.1 Entrenamiento de un HMM 55
2.2.2.2 Reconocimiento mediante HMM 56
2.3 Esquema de utilización de la técnica propuesta 57
2.4 Conclusiones parciales 58
CAPÍTULO 3: ANÁLISIS DE LOS RESULTADOS OBTENIDOS 59
3.1 Herramientas de ensayo 59
3.2 Medidas de eficiencia 59
3.3 Pruebas y experimentos 61
3.3.1 Experimentos en la detección de extremos de la señal de voz 62
3.3.2 Experimentos de detección de palabras clave sobre audio 63
3.3.2.1 Experimentos con diferentes cantidades de coeficientes MFCC 65
3.3.2.1.1 Experimentos con diferentes cantidades de coeficientes MFCC sin ΔMFCC-ΔΔMFCC 65
3.3.2.1.2 Experimentos con diferentes cantidades de coeficientes MFCC con ΔMFCC 65
3.3.2.1.3 Experimentos con diferentes cantidades de coeficientes MFCC utilizando ΔMFCC-ΔΔMFCC 66
3.3.2.1.4 Análisis de los experimentos con diferentes cantidades de coeficientes MFCC 66
3.3.2.2 Experimentos con coeficientes LPC, LPCC y MFCC 67
3.4 Conclusiones parciales 68
CONCLUSIONES GENERALES 69
RECOMENDACIONES 69

Índice de contenido

REFERENCIAS BIBLIOGRÁFICAS 71

ANEXOS 81

ANEXO 1 TABLAS 81

ANEXO 2 FIGURAS 83

ANEXO 3 FUNCIONES 85

INTRODUCCIÓN

Desde los inicios de la humanidad la comunicación oral es la forma de intercambio entre personas más usada constituyendo una gran fuente de intercambio. A causa del incesante desarrollo de las tecnologías de información y comunicaciones, hoy día una elevada cantidad de materiales audiovisuales son almacenados digitalmente en audio. El procesamiento de información constituye una necesidad para la realización de actividades tanto económicas como sociales, esto unido al desarrollo de archivos de audio disponibles por los ordenadores en constante intercambio, ha creado una necesidad urgente de recuperar inteligentemente las frases habladas, grabadas en estos archivos de audio. Aunque se tenga una gran base de datos de archivos de audio, la única información que se puede obtener de forma directa de los mismos son el título o alguna descripción previa sobre el contenido del mismo. Los problemas aparecen cuando se quiere tener conocimiento sobre un contenido de forma específica.

Actualmente el desarrollo de aplicaciones y técnicas destinadas a imitar tanto a la producción de voz humana como sistemas para su identificación tienen un amplio uso en esferas tecnológicas y sociales; esto ha sido posible gracias al desarrollo de investigaciones científicas destinadas a encontrar formas de mejorar el desempeño de estos sistemas en la identificación de la mejor secuencia de fragmentos de palabras. Estos sistemas son capaces de gestionar la señal de voz emitida por un determinado individuo, sometiendo la misma a un proceso de digitalización para la obtención de sus características, las cuales permiten enmarcar su comportamiento y posibilita la identificación de lenguajes, idiomas y términos, los cuales pueden ser usados en tecnologías de recuperación de la información. La detección de características de la voz posibilita la localización de un término específico, como por ejemplo la búsqueda de palabras en archivos de audios amplios, que pueden ser utilizadas en tecnologías de recuperación de la información. El proceso de búsqueda de palabras en audio puede ser dividido en tres categorías: la búsqueda de palabras aisladas, donde un usuario o locutor pronuncia una sola palabra o comando y puede ser utilizado para la autenticación o personalización de ordenadores, el reconocimiento de palabras en habla continua con un vocabulario muy amplio que es capaz de reconocer cualquier palabra. Estos sistemas necesitan una gran base de datos de modelos de palabras y la búsqueda de palabras clave, la cual es muy utilizada para localizar fragmentos de palabras así como frases en archivos previamente grabados evitando la reproducción entera de los mismos.

El reconocimiento automático del habla posibilita la entrada de comandos o instrucciones a través del habla natural a sistemas informatizados. Los sistemas de reconocimiento automático del habla son muy útiles para interactuar con aplicaciones o sistemas informáticos como por ejemplo la telefonía móvil o acceso a instalaciones y ordenadores, reduciendo el tiempo necesario de introducción de datos por ejemplo autenticarse en un ordenador personal mediante la voz o la realización de un dictado.

El desarrollo de sistemas de este tipo gracias al desarrollo de la inteligencia artificial constituye un gran avance en el desarrollo de esferas tecnológicas como: protección y seguridad, en la informática, robótica; así como el disfrute y bienestar social mediante la personalización de objetos personales como una computadora o hasta un hogar, además de sus usos en la telefonía móvil.

Las técnicas de localización de palabras o frases en archivos de audio son usadas en tratamientos de edición de señales de audio para la realización de sustituciones, borrados o inserciones de términos a la señal. También posibilita realizar la indexación de archivos de audio por concepto de palabras clave, lo cual es muy utilizado para recuperar información específica contenida en los mismos sin la necesidad de reproducir grabaciones completas (García, 2008).

Estos sistemas han estado en constante avance gracias al desarrollo de algoritmos matemáticos de reconocimiento de patrones y herramientas de predicción estadística. En (Luengo, y otros, 2009) se presenta un sistema de detección automática de vocales en la voz, mediante un identificador basado en Modelos Ocultos de Markov (*Hidden Markov Model* HMM por sus siglas en inglés) de grupos fonéticos. En (Ericson, 2009) se analiza el reconocimiento de habla continua mediante fonemas. En (Lokesh, y otros, 2012) se estudian características del habla robustas para mejorar la efectividad en el reconocimiento del habla.

En Cuba existen centros de investigación en los que se ha logrado publicaciones científicas en el campo del reconocimiento automático del habla. Ejemplos de estos centros de investigación científica se encuentran el Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV) y la Asociación Cubana de Reconocimiento de Patrones.

Entre las publicaciones científicas desarrolladas en CENATAV se encuentran (CENATAV, 2012):

- Métodos de extracción, selección y clasificación de rasgos acústicos para el reconocimiento del

locutor.

- Reconocimiento del locutor dependiente del texto con modelos acústicos del habla.
- Autenticación biométrica por el habla del usuario en las redes de info-comunicaciones.
- Métodos de reconocimiento automático del idioma.
- Topological Representation of Speech for Speaker Recognition.
- Métodos de extracción de rasgos para la identificación del idioma: estado del arte.

Todas estas investigaciones científicas han sido certificadas y algunas han sido publicadas fuera del ámbito nacional.

La Universidad de las Ciencias Informáticas (UCI) cuenta con una infraestructura productiva organizada en centros de producción, entre estos se encuentra el centro Geoinformática y Señales Digitales (GEySED) perteneciente a la Facultad 6. Dentro de GEYSED se encuentra el departamento Señales Digitales que tiene dentro de sus líneas de investigación el procesamiento y utilización de señales digitales, éstas están interrelacionadas con un inmenso conjunto de disciplinas de gran alcance social, entre las que se encuentran las telecomunicaciones, los avances en la exploración del espacio, la medicina y la arqueología, televisión digital, los sistemas de información y el entretenimiento audiovisual. A medida que los sistemas de comunicación se van convirtiendo en una necesidad tanto de la vida social como en esferas científicas, la importancia de un procesamiento de señales sofisticados se hace cada vez más relevante. Dentro del departamento Señales Digitales se manejan grandes flujos de datos de materiales audiovisuales y se encargan de actividades como: su publicación, captura, vigilancia, procesamiento en tiempo real y la recuperación de los mismos para préstamo o consulta de un usuario o sistema. La recuperación de materiales audiovisuales en el departamento Señales Digitales se lleva a cabo realizando búsquedas por filtros sobre metadatos que son descritos manualmente por un documentalista. La mayoría de estos archivos recuperados poseen audio y no se aprovechan al máximo las posibilidades de extracción de metadatos que el mismo contiene. Esto trae consigo que las búsquedas no sean lo suficientemente exactas, logrando resultados lejanos de lo que el usuario recuperador espera. Además estos metadatos son afectados cuando su registro se realiza de forma manual conduciendo a

problemas como:

- Ambigüedad de las palabras claves a registrar.
- Muchas palabras definidas como claves.
- Dilatación de almacenamiento del material para que sea recuperado por palabras claves.
- Sujeto a apreciación del personal encargado de documentar el material.
- El documentalista puede agotarse por el propio proceso de descripción de materiales arrastrando con su agotamiento una transcripción imprecisa del material.

La necesidad de lograr resultados satisfactorios en la recuperación de materiales audiovisuales fundamenta el **problema a resolver** como: El registro de palabras claves en archivos de audio dentro de los productos del departamento Señales Digitales resulta impreciso cuando este proceso se ejecuta manualmente.

La investigación se enfoca en el **objeto de estudio:** Técnicas para el procesamiento digital de señales de audio. Delimitando como **campo de acción:** El procesamiento de señales digitales de audio mediante técnicas enfocadas a la detección de palabras claves en archivos de audio en el departamento Señales Digitales.

Para darle solución al problema planteado se traza el **objetivo general** de: Proponer una técnica automática para la detección de palabras clave en audio para los productos del departamento Señales Digitales.

Teniendo en cuenta el problema de esta investigación así como el objetivo que se propone la misma, se tiene como **idea a defender:** La aplicación en el departamento Señales Digitales de una técnica automática para la detección de palabras clave en audio, permitirá mejores resultados en el proceso de recuperación de información de materiales audiovisuales.

Para la realización del objetivo propuesto se darán cumplimiento a las siguientes **tareas de la investigación:**

1- Caracterizar las técnicas de reconocimiento de habla.

2- Identificar técnicas para la reducción de ruido en audio.

3- Seleccionar técnica para el reconocimiento de palabras clave en audio.

4- Validar técnica para la detección de palabras clave en audio.

5- Proponer un esquema en que se satisfaga las necesidades del departamento Señales Digitales para la detección de palabras clave sobre audio.

Para la realización de la investigación se utilizan los siguientes métodos científicos:

Métodos Teóricos:

- **Histórico-Lógico:** Se utilizará para el estudio de los antecedentes sobre las técnicas de reconocimiento del habla y detección de palabras claves en audios.

- **Analítico-Sintético:** Para analizar la documentación y la bibliografía existente sobre reconocimiento del habla y la detección de palabras claves sobre audio.

Métodos Empíricos:

- **Matemáticos-Estadísticos:** Fueron utilizados para el procesamiento de la información obtenida a través de técnicas del nivel empírico, el uso de las propiedades de la matemática funcional y probabilística, para la creación de los modelos del habla y el entrenamiento. Se utilizaron las posibilidades de las aplicaciones informáticas para procesar la información cuantitativa, realización de pruebas y demostraciones.

La investigación estará estructurada en tres capítulos:

En el capítulo I se abordan temas sobre las técnicas y procedimientos para la detección de palabras claves en audio, así como un estudio para identificar las técnicas para la reducción de ruido en audio.

En el capítulo II se seleccionará una técnica para la detección de palabras claves sobre archivos de audio así como la propuesta de un esquema para la utilización de la técnica seleccionada en el departamento Señales Digitales.

En el capítulo III se expone la validación de la técnica para la detección de palabras claves sobre audio propuesta.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA SOBRE LA DETECCIÓN DE PALABRAS CLAVE EN AUDIO

En la actualidad el estudio de técnicas y procedimientos enfocados al procesamiento del habla ha posibilitado el desarrollo de sistemas capaces tanto para el reconocimiento de la misma como para agilizar procesos de búsqueda o recuperación de información. En este primer capítulo se realiza un estudio de las técnicas existentes en el ámbito de la detección de palabras clave sobre archivos de audio, así como técnicas para la reducción de ruido en audio. Además se explican conceptos relacionados con el dominio de la investigación para una mejor comprensión del mismo.

1.1 Conceptos asociados a la investigación

1.1.1 Técnica

Según (Definición, 2008) “la palabra técnica proviene de *téchne*, un vocablo de raíz griega que se ha traducido al español como “arte” o “ciencia”. Esta noción sirve para describir a un tipo de acciones regidas por normas o un cierto protocolo que tiene el propósito de arribar a un resultado específico, tanto a nivel científico como tecnológico, artístico o de cualquier otro campo”. En otras palabras, una técnica es un conjunto de procedimientos reglamentados y pautas que se utiliza como medio para llegar a un cierto fin.

Es considerada también como un conjunto de procedimientos o recursos que se usan en una ciencia o en una actividad determinada. Habilidad que tiene una persona para hacer uso de estos procedimientos o recursos. Aplicación práctica de los métodos y de los conocimientos relativos a las diversas ciencias (Farlex Inc, 2010).

Es el seguimiento de un conjunto de procedimientos y operaciones en una actividad científica, deportiva o laboral para el cumplimiento de un determinado objetivo.

1.1.2 Frecuencia

Es la medida de rapidez de la oscilación de una onda por unidad de tiempo. Indica la cantidad de veces que un ciclo de la señal se repite en un segundo (Topon, y otros, 2008).

La frecuencia es el número de repeticiones de un fenómeno por unidad de tiempo. La frecuencia de patrones ondulatorios como el sonido, las ondas electromagnéticas (como la radio o la luz), las señales eléctricas, u otras ondas, indica el número de ciclos de la onda repetitiva por segundo. La unidad de frecuencia del Sistema Internacional es el hercio o *hertz* (Hz), llamado así en honor al físico alemán *Heinrich Rudolf Hertz*. 1 Hz representa un ciclo (u onda) por segundo. Es inversamente proporcional a la longitud de onda (la distancia entre dos crestas). Es igual a la velocidad dividida por la longitud de onda (GreenFacts, 2008).

En el tratamiento de señales de audio la frecuencia constituye el número de muestras de señal que aparecen en un segundo de transmisiones.

1.1.3 Espectrograma

Es una visualización del espectro conformado por la representación de las energías (eje y) para cada componente de frecuencia (eje x) para un segmento de onda de corta duración de 2 a 3 ciclos de frecuencia.

El espectrograma (también llamado sonograma) consiste en la representación gráfica del espectro de frecuencias de la emisión sonora. Puede revelar rasgos como altas frecuencias o modulaciones de amplitud que no pueden apreciarse incluso aunque estén dentro de los límites de frecuencia del oído humano. Normalmente, representa el tiempo sobre el eje horizontal, la frecuencia sobre el eje vertical y la amplitud de las señales mediante una escala de grises o de colores. Otra representación que suele utilizarse es la que sitúa la frecuencia en el eje horizontal y la amplitud en el eje vertical. Este modelo no muestra información sobre el tiempo, por lo que la interpretación solamente resulta útil para análisis de puntos específicos de la secuencia (Hernández, 2010).

Un espectrograma es una forma de representar gráficamente en el tiempo la gama de frecuencias de una señal sonora.

1.1.4 Acústica

La acústica, etimológicamente procede del griego *akustike* (Ciencia del sonido), es un término que define a la vez una ciencia y una tecnología. Como ciencia estudia las propiedades de las vibraciones de las

partículas de un medio susceptible de engendrar sonidos, infrasonidos o ultrasonidos. Como tecnología se orienta a la producción, la transmisión y del análisis de los efectos del sonido (López, 2012).

La acústica es una rama de la física interdisciplinaria que estudia el sonido, infrasonido y ultrasonido, es decir ondas mecánicas que se propagan a través de la materia (tanto sólida como líquida o gaseosa, no se propagan en el vacío) por medio de modelos físicos y matemáticos. A efectos prácticos, la acústica estudia la producción, transmisión, almacenamiento, percepción o reproducción del sonido (Rochaix, y otros, 2010).

Estudia los fenómenos relacionados a la producción del sonido, su recepción y sus variaciones vibratorias.

1.1.5 Palabras clave

Según (Pergaminovirtual, 2012) una palabra clave es un conjunto de caracteres que puede utilizarse para buscar una información en un buscador.

Según (Babylon, 2012) una palabra clave es cualquier búsqueda, palabra que brinde información significativa que pueda ser usada como criterio de búsqueda. Además (Oxford University Press, 2013) plantea que una palabra clave o *keyword* es un palabra o frase que representa alguna información y que puede ser usada en sistemas de recuperación de información para indicar un contenido.

1.1.6 Recuperación de información

Según (Oliván, y otros, 2007) La recuperación de Información es el proceso de comunicación entre los usuarios y los sistemas, que mediante un término se logra recuperar un conjunto de elementos, que son evaluados por el interesado en la búsqueda, para lograr satisfacer su necesidad de información.

La recuperación de la Información, tiene por objetivo: identificar, localizar, seleccionar y acceder a los recursos de información útiles al usuario (Molina, 2011). El proceso de recuperación se lleva a cabo mediante consultas a la base de datos donde se almacena la información estructurada, mediante un lenguaje de interrogación adecuado. Es necesario tener en cuenta los elementos clave que permiten hacer la búsqueda, determinando un mayor grado de pertinencia y precisión, como son: los índices, palabras clave, tesauros y los fenómenos que se pueden dar en el proceso como son el ruido y silencio documental (Sayán, 2011).

Puede definirse como un proceso donde se accede a una información previamente almacenada, mediante herramientas informáticas que permiten establecer ecuaciones de búsqueda específicas. Dicha información ha debido de ser estructura previamente a su almacenamiento.

1.2 Situación problemática

Hoy día el desarrollo creciente de las tecnologías de información y las comunicaciones facilita el intercambio de flujos de datos entre personas, empresas y comunidades científicas. Esto ha favorecido áreas del desarrollo tecnológico en campos como la informática, salud, robótica, entre otros. Un ejemplo de esto lo constituye el aumento de las investigaciones y aplicaciones destinadas al reconocimiento de voz. Dentro del mundo del reconocimiento de voz cobra mucha importancia la detección de palabras claves sobre archivos de audio con el principal objetivo de la extracción y recuperación de información. Esta importancia radica en el simple hecho de la creciente cantidad de información en lenguaje oral que es guardada en formato digital en la actualidad. Es por esto que son necesarios cada vez más sistemas de acceso a archivos de audio que proporcionen posibilidades de búsquedas de usuarios sobre grandes bases de datos de audios directamente sobre estos archivos y no sobre algún resumen o descripción del contenido de los mismos realizada por algún documentalista.

Actualmente los proyectos productivos que forman parte del departamento Señales Digitales manejan gran cantidad de información audiovisual, a la cual se le aplican diferentes procesos para facilitar su gestión, como por ejemplo: su publicación, captura, catalogación, vigilancia, procesamiento en tiempo real y recuperación de los mismos.

Los procesos de recuperación de materiales audiovisuales dentro del departamento Señales Digitales son el resultado de la descripción manual de metadatos de información. Este proceso manual puede provocar que las búsquedas realizadas no sean del todo la esperada por los usuarios de sistemas de recuperación audiovisual. La descripción de metadatos de información es realizada por personas encargadas de documentar los materiales audiovisuales, esto puede traer consigo factores que afectan a los metadatos asociados a cada material, entre estos factores se encuentra: agotamiento del documentalista por largos procesos de descripción de materiales arrastrando un mal registro de la palabra, frase o tema que se analiza; además la realización de la descripción de estos materiales está sujeta a la interpretación del documentalista.

La recuperación de materiales audiovisuales resulta deficiente cuando dichos materiales no están bien descritos o si se hace necesario reproducir un determinado material y analizarlo totalmente. Estos factores pueden causar exceso en consumo de recursos importantes como, altos costos computacionales, tiempo para realizar otras búsquedas, además de causar cansancio en la persona que está realizando la documentación del audiovisual, provocando la posibilidad de que cometa errores en los resultados finales y no se cumplan las expectativas que el usuario o sistema recuperador espera.

Ante la presente situación se determinó que dichos inconvenientes podían ser solucionados definiendo un flujo para la detección de palabras claves en el departamento Señales Digitales, empleando para ello técnicas automáticas.

1.3 Objeto de Estudio

Las técnicas para el tratamiento de señales del habla están en constante evolución en gran medida gracias al desarrollo de modelos capaces de procesar las características acústicas del habla. Estas técnicas posibilitan el desarrollo de procedimientos enfocados a la detección de palabras clave en archivos de audio.

1.3.1 Características acústicas de la voz

La señal de voz es la composición de una secuencia de sonidos y silencios. Esta señal está compuesta por palabras que a su vez están compuestas por fonemas, estos constituyen la unidad básica del habla.

El habla puede verse como una secuencia de unidades básicas, de sonidos o fonemas. Los fonemas son unidades teóricas, definidas para estudiar el nivel fonético-fonológico de una lengua humana. Son unidades lingüísticas abstractas y no pueden observarse directamente en la señal de voz. Un mismo fonema se aplica a muchos sonidos ligeramente diferentes llamados realizaciones del fonema o alófonos.

1.3.1.1 Fono y Fonema

Un sonido o fono se caracteriza por una serie de rasgos fonéticos y articulatorios. Es cualquiera de las posibles realizaciones acústicas de un fonema (Rojo, 2011).

Un fonema es cada uno de los sonidos que distinguen significados en una lengua, es decir que poseen valor distintivo. Es una unidad fónica mínima e invisible capaz de diferenciar significados (Sintagmas, 2010), (Bermejo, 2008).

En el castellano se definen 24 fonemas entre consonánticos y vocálicos. Los fonemas son unidades básicas utilizadas para el estudio del nivel fónico-fonológico de la lengua. Son aquellos sonidos del habla que nos permiten distinguir las palabras en una lengua. Los mismos poseen características como (Bermejo, 2008), (Rojo, 2011), (Gila, 2012):

- Diferenciadores: cada fonema se delimita dentro del sistema por las cualidades que se distinguen de los demás.
- Son indivisibles, no se pueden descomponer en unidades menores.
- Son abstractos no son sonidos, sino modelos de sonido.

Acústicamente un fonema es un sonido que se distingue por un patrón característico. Por otro lado un alófono es una de las diferentes pronunciaciones que puede llegar a tener un fonema. Los fonemas se dividen de acuerdo a su manera de articulación en dos grupos fundamentales, estos son fonemas vocálicos y fonemas consonánticos (García, 2008).

1.3.1.2 Modelos fonéticos

Los fonemas representan un nivel superior de fragmentación de palabras. La modificación de un fonema puede cambiarle el sentido a la misma. Hay que distinguirlo de alófono, que es cada una de las pronunciaciones reales del modelo ideal que representa el fonema. En función del grado de resolución que se quiera que presenten las unidades fonéticas se pueden obtener modelos más o menos específicos. La manera en que se hace la división en unidades fonéticas depende del contexto donde el alófono se localice:

- Monofonemas: son unidades totalmente libres de contexto. Un monofonema tiene en cuenta todas las posibles realizaciones de un fonema independientemente de sus vecinos.

- Bifonemas: son unidades que dependen solo de uno de sus contextos, ya sea este el derecho (bifonema derecho) o el izquierdo (bifonema izquierdo).

- Trifonemas: son unidades que dependen de ambos contextos a la vez.

- Trifonemas generalizados: Dado que el número de unidades va aumentando al ir considerando más detenidamente la posición de los fonemas en su contexto, puede llegar a ser tan elevado que su entrenamiento no fuera posible. Surge el trifonema generalizado como un primer nivel de compartición en el que varios trifonemas cercanos se agrupan para reducir el número de modelos y para que el entrenamiento de estos sea mejor.

De esta forma, un reconocedor necesita disponer de un conjunto de unidades que permita construir cualquier palabra o frase a partir de su concatenación. Los fonemas representan este conjunto completo y reducido de unidades a partir de las cuales se puede generar cualquier palabra. Estas unidades pueden modelarse con mayor o menor resolución en función del contexto a considerar.

La creación de modelos acústicos, para su posterior uso en la detección de palabras claves, se realiza en dos fases (Bermejo, 2008), (Rojo, 2011), (Gila, 2012):

- Extracción de las características del sonido del habla: Por semejanza con el funcionamiento del sistema humano, la extracción de esas características, que llamaremos parámetros, se realiza en el dominio de la frecuencia. Es la asignación de los parámetros extraídos a las representaciones discretas de nuestro diseño (fonemas, trifonemas, sílabas, palabras) correspondientes, con el objetivo de crear un modelo para cada una de las representaciones discretas que las identifique.

- Entrenamiento y reconocimiento de modelos para cada fonema a identificar: A partir de la extracción de parámetros se construirán una serie de modelos estadísticos con los cuales se identificarán, con cierta probabilidad, fonemas en otras locuciones. Por ello, se mide la distancia entre el conjunto de parámetros que constituye el modelo y los parámetros de la pronunciación a reconocer.

1.3.1.3 Formantes

Son frecuencias que entran en resonancia en las cavidades nasales y orales, saliendo hacia el exterior como la información más importante del habla. Un formante es el pico de intensidad en el espectro de un

sonido; se trata de concentración de energía que se obtiene de una determinada frecuencia. Los formantes son los elementos que sirven para distinguir componentes del habla humana, sobre todo las vocales y los sonidos sonantes (Rincón, y otros, 2008).

1.3.1.4 Aparato fonador

Se define como aparato fonador al andamiaje biológico del ser humano controlado por el sistema nervioso central en correlación con un conjunto de peculiaridades morfológicas y mecanismos fisiológicos, los cuales influyen de manera decisiva sobra las características del habla. La emisión de un fonema exige la realización de determinadas maniobras neuromusculares, así como la generación de corriente de aire que debe ser modulada a diferentes niveles del aparato fonador (Otero, 2010).

1.3.2 Reconocimiento automático del habla

El reconocimiento automático del habla (RAH) o reconocimiento automático de voz es una rama de la Inteligencia Artificial que tiene como objetivo permitir la comunicación hablada entre seres humanos y sistemas de cómputo (Romero, 2012). El problema que se plantea en un sistema (RAH) es el de hacer cooperar un conjunto de informaciones que provienen de diversas fuentes de conocimiento (acústica, fonética, fonológica, léxica, sintáctica, semántica y pragmática), en presencia de ambigüedades, incertidumbres y errores inevitables para llegar a obtener una interpretación aceptable del mensaje acústico recibido (Otero, 2010).

Estos sistemas son capaces de gestionar la señal de voz emitida por un individuo. Dicha señal ha sido pasada por un proceso de digitalización para obtener elementos de medición (muestras), las cuales permiten denotar su comportamiento e implementar procesos de tratamiento de la señal, enfocados al reconocimiento.

En un sistema de reconocimiento automático del habla SRAH la señal de voz se ve inmersa en dos bloques importantes: entrenamiento y reconocimiento. Dicho entrenamiento es una de las etapas más críticas dentro de estos sistemas y gran parte del éxito de un sistema de reconocimiento de voz recae en esta etapa (Rodríguez, 2006).

El objetivo de un SRAH (Sistema de Reconocimiento Automático del Habla) es extraer la información lingüística contenida en la señal de voz. Para ello determina la secuencia de palabras más verosímil dada la señal de voz de entrada. Dicho criterio puede expresarse matemáticamente de la siguiente manera: $\boldsymbol{W = arg\ max_i\{P(W_i|O)\}}$**.** Siendo $\boldsymbol{W_i}$ la hipótesis de $\boldsymbol{i}$ sobre la secuencia de palabras y $\boldsymbol{O}$ la secuencia de observaciones obtenida a partir de la señal de voz. Esta secuencia de observaciones no será más que la secuencia de vectores de parámetros que representan la señal de voz y se obtienen para cada trama ($\boldsymbol{O = \{O_1, O_2, \dots, O_T\}}$). Donde $\boldsymbol{O_T}$ denota el vector de parametrización obtenido para el instante $\boldsymbol{t}$, y $\boldsymbol{T}$ indica el número de vectores de parámetros que extraemos de la señal observada (Vicente, 2007), (Becerra, y otros, 2009).

Los SRAH, son la implementación algorítmica de un análisis detallado de las características del habla, y que dependerá del lenguaje usado y de otros conceptos de extracción de características de la señal de voz. Un SRAH puede contemplar: reconocimiento de palabras aisladas, identificación de palabras claves en habla continua, reconocimiento de palabras conectadas así como en un discurso continuo. En estos sistemas una vez extraídas las características acústicas y del tracto vocal de la señal de voz, estas pasan a ser procesadas por modelos o clasificadores en los cuales se lleva a cabo el proceso de entrenamiento del sistema para así pasar a la fase final de reconocimiento. Después se realiza una búsqueda para encontrar la secuencia de segmentos con mayor probabilidad de ser reconocidos (García, 2008). Ver figura 1.

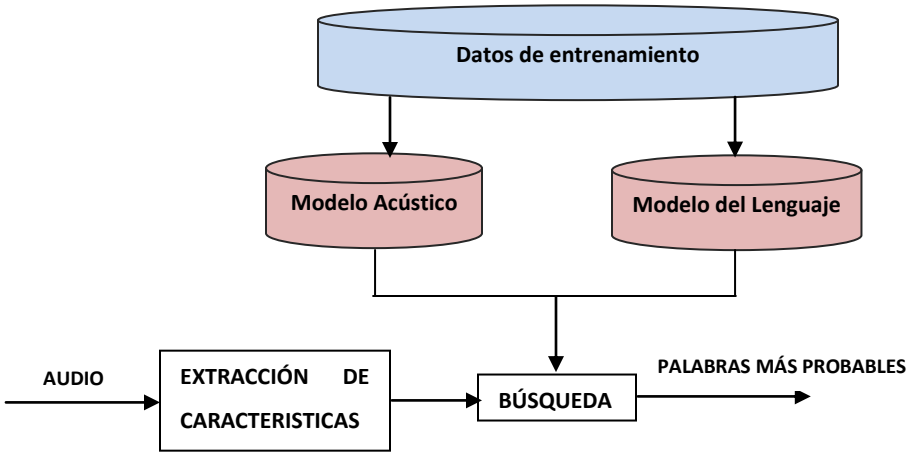


Figura. 1 Arquitectura de un Sistema de Reconocimiento Automático del Habla

1.3.3 Detección de palabras clave en archivos de audio

Dentro de las aplicaciones del tratamiento de señales de voz y sistemas de reconocimiento automático del habla cobra mucha importancia la detección de palabras clave con el objetivo de la extracción de información de los segmentos de audio a reconocer. Esta importancia se reafirma en el hecho de la existencia de una gran cantidad de información en formato de audio.

Uno de los métodos más usados para la detección de palabras clave es la búsqueda basada en los sistemas de reconocimiento de habla continua de gran vocabulario o *Large Vocabulary Continuos Speech Recognition* (LVCSR por sus siglas en inglés) (Gila, 2012). Este tipo de sistemas se basan en un diccionario que suele contener unas 65.000 palabras y tienen la mejor tasa de acierto siempre y cuando las palabras que se quieran buscar estén dentro del vocabulario (Noguerales, y otros, 2010), (Gila, 2012).

El problema de este tipo de sistemas viene cuando la palabra que se desea reconocer es una de las llamadas palabras fuera de vocabulario o *Out-Of-Vocabulary* (OOV por sus siglas en inglés), que son normalmente nombres propios, acrónimos o extranjerismos (García, 2008), (Parada, y otros, 2011), (Gila, 2012). Según (Logan, y otros, 2004), (Noguerales, y otros, 2010) el 12 % de las palabras que forman parte de una consulta típica de un usuario son OOV. A ello se le suma el hecho de que los idiomas van creciendo con lo que aparecen nuevas palabras que no están en el vocabulario a no ser que éste se actualice. El problema de las palabras OOV siempre existirá independientemente del número de palabras contempladas en el vocabulario de los LVCSR (Noguerales, y otros, 2010).

Estos problemas ocasionados por las palabras OOV han creado la necesidad de utilizar sistemas que realicen un reconocimiento más eficiente de las consultas realizadas por usuarios o que al menos sean un buen complemento de los sistemas basados en LVCSR. Es por esto que aparecen otros dos grandes métodos de reconocimiento de palabras clave, estos son: los sistemas basados en modelos de relleno y los sistemas basados en reconocedores de voz de sub-unidades de palabras (Gila, 2012).

Los sistemas basados en modelos de relleno presentan una tasa mejor de aciertos que los basados en sub-unidades de palabra, por detrás de los sistemas LVCSR (los cuales son los que mejores resultados obtienen), pero presentan el problema de que en caso que se quiera cambiar la palabra a buscar hay que reprocesar todo el audio (García, 2008), (Noguerales, y otros, 2010), (Gila, 2012).

Los sistemas basados en reconocedores de sub-unidades de palabra no necesitan reprocesar el audio cuando el vocabulario de la aplicación cambia, al utilizar la decodificación acústica con un vocabulario compuesto por fonemas, grafemas, sílabas, cuyo número siempre se mantiene constante en un idioma. Sin embargo este tipo de sistemas al no considerar ningún modelo de palabra durante el proceso de reconocimiento es el que peor resultado obtiene (Noguerales, y otros, 2010), (Gila, 2012).

Los sistemas de detección de palabras clave basados en Modelos Ocultos de Markov (*Hidden Markov Model* HMM por sus siglas en inglés) requieren de un entrenamiento inicial que se realiza en un segmento de habla grabado, durante la fase de entrenamiento un conjunto de unidades acústicas son adquiridas y usadas para modelar el habla (García, 2008).

1.3.3.1 Detección de palabras clave basado en reconocedores de habla continua de gran vocabulario

El reconocimiento de habla continua de gran vocabulario se caracteriza por transcribir el habla pronunciada de forma natural, cubriendo un vocabulario extenso.

Según (Noguerales, y otros, 2010) las mayores preocupaciones para emplear sistemas LVCSR para la recuperación de información de voz lo constituye su tasa de error, que resultaba especialmente elevada si las condiciones del audio y la voz sobre las que trabaja el reconocedor no coinciden con las condiciones de entrenamiento del mismo. Este problema de robustez de los reconocedores de voz frente a variaciones del tipo de audio o del tipo de habla es una de las principales limitaciones para el uso de sistemas LVCSR en la recuperación de información y se hace prácticamente imprescindible que el sistema de reconocimiento de voz se adapte al entorno donde se va a usar mediante un laborioso proceso manual que exige la mayor cantidad de datos lo más adaptables posibles a dicho entorno. No obstante los sistemas LVCSR proporcionan información útil para la recuperación de información, haciendo prácticamente imprescindible su uso (Noguerales, y otros, 2010).

Los sistemas de detección de palabras clave basados en reconocedores de habla continua de gran vocabulario (LVCSR) utilizan un diccionario completo de la lengua concreta en la que se está buscando términos a partir del cual se hace un reconocimiento cuya unidad más pequeña es la palabra. Después de ese reconocimiento y con él la obtención de todas las palabras presentes en el audio, se hace uso de un

detector de palabras clave con el que obtenemos la localización de los términos que estamos buscando. En el siguiente diagrama (ver figura 2) se muestran los bloques de este proceso (Gila, 2012).

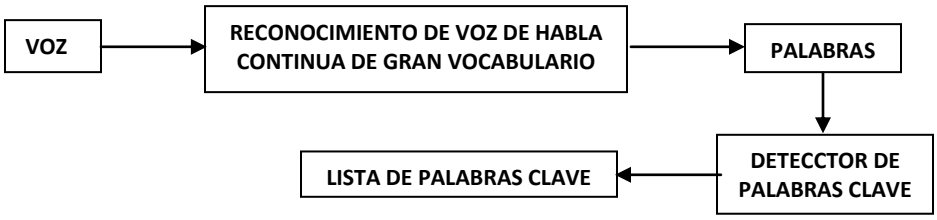


Figura. 2 Diagrama de bloques de un sistema de detección de palabras clave basado en LCVSR

La señal de voz es procesada a través de un reconocedor de habla continua de gran vocabulario y las palabras que forman parte de la consulta del usuario son detectadas por el detector de palabras clave a partir de la salida del reconocedor de voz (Noguerales, y otros, 2010).

El gran problema de estos sistemas es la imposibilidad que tienen de encontrar términos fuera de vocabulario (OOV), ya que en el diccionario que usan nunca están presentes nombres propios, acrónimos o extranjerismos. A ello se le suma que estos sistemas tienen un importante problema de robustez ya que su tasa de error es elevada si las condiciones del audio y la voz sobre las que trabaja el reconocedor no coinciden con las de entrenamiento (Gila, 2012).

En este tipo de sistemas las búsquedas de palabras clave se realizan sobre la cadena de palabras resultante la transcripción del sistema LVCSR (García, 2008). En (Noguerales, y otros, 2010) plantea que los sistemas de detección de palabras clave basados en reconocedores de habla continua de gran vocabulario (LVCSR) no eran del todo eficientes debido a la existencia de palabras como por ejemplo nombres propios, acrónicos, extranjerismos que se encuentran fuera del vocabulario (OOV).

En (Gila, 2012), (Noguerales, y otros, 2010) se expresa que estos sistemas son los más usados para la detección de palabras clave, además que aportan los mejores resultados. Y para mejorar el problema de las palabras OOV es necesario periódicamente una actualización de la base de datos del sistema.

1.3.3.2 Detección de palabras clave basado en modelos de relleno

Estos sistemas a diferencia de los LVCSR, contienen en su diccionario únicamente las palabras clave que se desean extraer del audio junto con unos modelos especiales de relleno. El proceso de decodificación

acústica obtiene la secuencia más probable de palabras que aparecen en el audio. Es por esto que hay que tener en cuenta no sólo las palabras clave en el proceso, sino también cualquier tipo de palabra que pueda aparecer en el audio. Esto se consigue con los modelos de relleno que intentan absorber las palabras que aparecen en el audio que no se corresponde con ninguna de las palabras clave que se desea buscar.

A la salida del sistema reconocedor se obtiene el conjunto de términos buscados así como los modelos de relleno que se sitúan en las zonas del audio donde no existen palabras clave. Después de esto utilizando unas medidas de confianza (diseñadas para descartar las hipótesis propuestas con una puntuación muy baja) se obtiene la localización de los términos buscados (García, 2008), (Noguerales, y otros, 2010), (Gila, 2012). Ver diagrama de la figura 3.

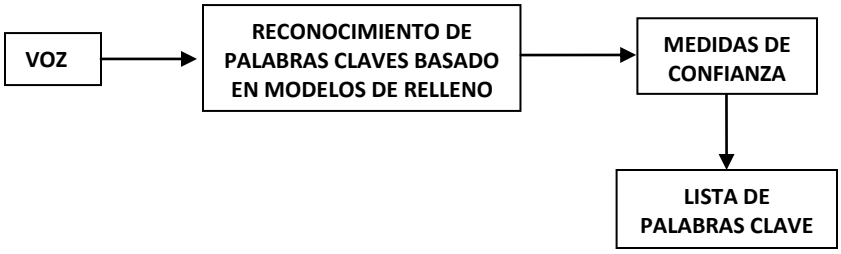


Figura. 3 Diagrama de bloques de un sistema de detección de palabras clave basado en modelos de relleno

Existen distintas formas de entrenar tanto los modelos de relleno como las palabras clave a buscar. El entrenamiento de los modelos puede realizarse mediante las unidades de palabra (fonemas, sílabas, palabras), esta etapa es de vital importancia. A lo largo del tiempo se han propuesto distintas formas de hacerlo.

En (Cuayahuitl, y otros, 2004) compararon el uso de fonemas, palabras y sílabas como modelos de relleno observaron para un sistema para el idioma español, que el modelo de sílaba era el que mejores resultados obtenía al ser más robusto que el de fonema y a diferencia del modelo de palabra es capaz de cubrir toda la variación del idioma.

También se han estudiado varias formas de realizar las medidas de confianza de las hipótesis obtenidas. Por ejemplo en (Cuayahuitl, y otros, 2004), (Xin, y otros, 2001), (Tejedor, y otros, 2008) y (Szoke, y otros,

2005) determinan las medidas de confianza descartando las hipótesis cuya puntuación sea menor que un umbral.

Estas medidas de confianza logran aumentar el rendimiento del sistema de detección de palabras clave basados en modelos de relleno, en términos de aciertos y falsas alarmas.

1.3.3.3 Detección de palabras clave basado en reconocedores de sub-unidades de palabra

Este tipo de sistemas son muy usados en mecanismos de extracción de información en contenidos de audio, aunque no presenta mejores resultados que los métodos basados en modelos de relleno ni basados en el reconocimiento de habla continua de gran vocabulario. (NIST, 2006), (Noguerales, y otros, 2010).

Este método tiene como ventaja que las búsquedas realizadas son muy rápidas. Es por eso que estos sistemas de detección de palabras clave basados en reconocedores de sub-unidades de palabra siempre están formados por una primera fase *offline* en la que se realiza el reconocimiento de voz basado en sub-unidades de palabra (habitualmente fonemas, pero también grafemas[1], sílabas, grafones[2]) y una segunda fase *online* en la que se hace uso de un detector de palabras clave que se encarga de buscar las palabras en la secuencia reconocida en la primera fase. Además, se suelen realizar unas medidas de confianza para descartar falsas hipótesis (Noguerales, y otros, 2010), (Gila, 2012). Ver figura 4.

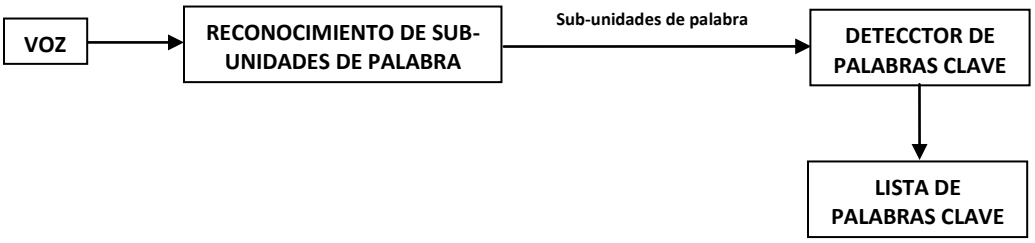


Figura. 4 Diagrama de bloques de un sistema de detección de palabras clave basado en modelos de sub-unidades de palabra

La utilidad de los sistemas de detección de palabras clave basados en reconocedores de voz de sub-unidades de palabra, radica en el hecho de que los fonemas, sílabas, grafemas en definitiva, las sub-

[1] *Unidad lingüística mínima de la lengua escrita. Son las letras, números y símbolos lingüísticos.*
[2] *Son fragmentos de palabra que se entrenan a partir de la secuencia de fonemas y grafemas de las palabras.*

unidades de palabra se mantienen inalterables en un idioma. De esta forma, estos sistemas se convierten en idóneos para la localización de palabras fuera de vocabulario (OOV).

A continuación se muestran algunas propuestas para el uso de este tipo de sistemas. Por ejemplo en (Amir, y otros, 2001) propusieron un sistema en el que se obtenía una secuencia de fonemas que correspondía a la secuencia de palabras más probable reconocida por un reconocedor de palabras de gran vocabulario. De esta secuencia de fonemas se obtienen las hipótesis tras la búsqueda de palabras clave. En (Thambiratnam, y otros, 2007) y (Szoke, y otros, 2005) desarrollaron sistemas basados en un reconocedor de fonemas en los que para cada instante de tiempo existía más de una hipótesis del fonema encontrado. Y (Larson, y otros, 2007) propusieron un reconocedor silábico con el que configuraban una secuencia similar al de los sistemas basados en reconocedor de fonemas.

1.3.4 Procesamiento de la señal de voz

Uno de los problemas a la hora de realizar la detección de palabras clave sobre archivos de audio y a la vez constituye una necesidad es la selección de un conjunto de características propias de una señal de voz, o sea aquellas que capturen mejor su estructura temporal y espectral. La extracción de estas características sirve para poder discriminar dentro de una señal si es silencio o no.

En el procesamiento de la señal de voz se extraen las características que utilizará posteriormente el reconocedor, en el proceso de extracción se divide la señal en una colección de segmentos. Posteriormente se obtiene una representación de características acústicas más distintivas para cada segmento. Con estas características obtenidas, se construye un conjunto de vectores que constituyen la entrada al siguiente módulo. Una manera de representar el sonido es graficándolo en forma de onda (*waveform*). En el eje horizontal representa el tiempo y el vertical la amplitud. Una limitación es el hecho de que no describe explícitamente el contenido de la señal en términos de propiedades (García, 2008).

Una representación más adecuada para el análisis de la señal de voz son los espectrogramas. Un espectrograma es una representación de la señal de voz de acuerdo a las variaciones de la energía con respecto al tiempo. Ver figura 5.

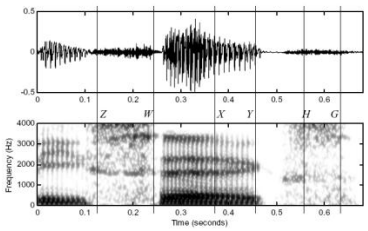


Figura. 5 Señal de voz y su correspondiente espectrograma de ancho de banda

1.3.4.1 Ventaneo de la señal

Al analizar la señal de voz hay que tener en cuenta que la misma varía constantemente en el dominio del tiempo, es por eso que resulta más eficaz aplicar el análisis a pequeñas porciones de la misma, ya que el principal objetivo es observar la variación de los diferentes parámetros calculados, esto es posible segmentando la señal mediante de ventanas.

Las ventanas según (López, 2009), (Rodríguez, 2006) son funciones matemáticas usadas en el análisis de señales para evitar discontinuidades al principio y al final de cada uno de los bloques analizados. Así pues una ventana se utiliza cuando es necesario acotar de forma voluntaria el tamaño de la señal con que se va a trabajar. Entre las clasificaciones de las ventanas sobresalen las ventanas cuadráticas y las ventanas Hamming. Del mismo modo, para evitar discontinuidades, la ventana no se desplaza cada vez todo el tamaño de su longitud, si no que se desplaza la mitad de su longitud para así irse superponiendo y lograr una ponderación más adecuada del valor que se está calculando, ver figura 6.

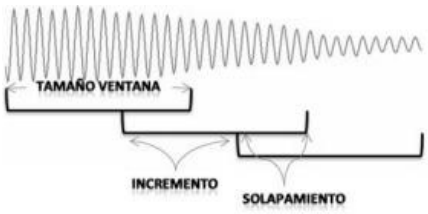


Figura. 6 Proceso de ventaneo de una señal

La duración promedio de una ventana para el tratamiento de señales de habla según (Rincón, y otros, 2008), (García, 2008), (Facundo, y otros, 2008), (Ramírez, 2008) es de 20-40 ms (milisegundos)[3] tiempo en que la señal de habla se considera prácticamente invariable. En (Martínez, y otros, 2008) proponen el uso de ventas de 10-45 ms y (Rocamora, y otros, 2007) se utilizan ventanas de 25ms. Según (Mezcua, 2005) la mayoría de los sistema de tratamiento de voz realizan el análisis en tramas de voz de 10 a 30 ms.

Según (Martínez, y otros, 2008), (López, 2009) la ventana rectangular es la de menor complejidad si solo se quisiera acotar la señal de audio pero la utilización de esta ventana modifica demasiado la señal en el dominio de la frecuencia, es por eso que se utilizan otro tipo de ventanas que modifican en gran medida en el espacio temporal pero que dejan prácticamente la señal invariable en el dominio de frecuencia, tal es el caso de la ventana Hamming.

La ventana Hamming suaviza los valores de los extremos evitando así discontinuidades bruscas entre segmentos consecutivos para ello no se desplaza en todo su tamaño sino que se solapa o superpone la mitad de su dimensión y en (Ramírez, 2008), (López, 2009) viene representada por la siguiente fórmula, ver figura 7, en la que ***N*** indica la longitud de la ventana en muestras.

La ventana Hamming es una función coseno[4] que busca acercar a cero o cerca de cero los extremos de los bloques y así reducir el salto (Cruz, 2010). Por tanto, su objetivo será evitar que las características de los extremos varíen la interpretación de lo que ocurre en la parte más significativa de las muestras seleccionadas.

$$\boldsymbol{w(n) = \begin{cases} 0.54 - 0.46\, cos\left(\dfrac{2\pi n}{N-1}\right), 0 \leq n \leq N-1 \\ \\ 0, \qquad caso\ contrario \end{cases}}$$

Figura. 7 Fórmula de la ventana Hamming

1.3.5 Descriptores de la señal de habla

[3] *Los milisegundos son una unidad de tiempo que corresponde a la milésima fracción de un segundo.*
[4] *La función coseno asocia a cada número real, x, el valor del coseno del ángulo cuya medida en radianes es x.*

A la hora de realizar el análisis de una señal de habla es necesario utilizar descriptores que posibiliten caracterizar el comportamiento de la misma. La utilización de estas características es muy útil en los procesos del análisis de la señal con respecto al tiempo o su comportamiento en el dominio de la frecuencia. Además permiten el desarrollo de procesos enfocados a la detección de palabras o términos dentro de una señal de habla.

1.3.5.1 Volumen

La distribución del volumen de una señal de audio revela la variación temporal de la magnitud de la señal, la cual es importante para la clasificación en segmentos de silencio o sonoros. Para calcular el valor del volumen en un *frame* se puede utilizar la media cuadrática o *Root Mean Square* (RMS, por sus siglas en inglés) de la magnitud de la señal. Específicamente el volumen del *frame* n-ésimo es calculado por la fórmula de la figura 8 (García, 2010).

$$V(n)\sqrt{\frac{1}{N}\sum_{i=0}^{N-1} S^2{}_n(i)}$$

Figura. 8 Cálculo del Volumen

Donde $S_n(i)$ es la muestra i-ésima en el *frame* n-ésimo y N es la longitud del *frame*. El volumen de una señal de audio depende de la ganancia dada en el sistema de grabación y digitalización.

1.3.5.2 Energía en Tiempo Corto (STE)

La Energía en Corto Tiempo o *Short Time Energy* (STE por sus siglas en inglés) no es un estado físico real, es un número escalar que se le asigna a un estado de un sistema físico, es un valor arbitrario a cada una de las muestras que se van a analizar. En el caso de las señales de audio, se suele asemejar a la intensidad sonora de la misma, siendo señales con mucha energía aquellas que tienen un sonido muy alto. Se calcula a partir del valor al cuadrado de la señal acotada a un tipo de ventana en concreto, ver figura 9. Siendo $x(m)$ la señal de audio y $w(m)$ es la ventana con la que se va a hacer el acotado de la señal que tiene longitud N se define energía en corto tiempo como (López, 2009):

$$\boldsymbol{STE = 1/N \sum_{m=1}^{N} [x(m)w(n-m)]^2}$$

Figura. 9 Fórmula de Energía en Corto Tiempo

La energía es escogida como característica teniendo en cuenta que los fonemas sordos contienen más energía que los segmentos de silencio. La señal de voz es limitada en banda, aproximadamente entre 20 Hz y 20 Khz (región de audición para un humano normal). Sin embargo la mayor parte de la misma se concentra por debajo de 2 Khz (Rueda, y otros, 2006).

En (Al-Shoshan, 2006), (López, 2009) se plantea que la STE de una señal de habla brinda una representación de la variación de amplitud de la misma. Además expresa que las señales de música no contienen tantos espacios de silencio como las de habla por lo que la STE es mayor en las señales de música que en las de habla. Esto es muy útil para procesos de discriminación de segmentos de audio.

1.3.5.3 Tasa de Cruces por Cero (*Zero Crossing Rate*)

Se denomina cruce por cero al hecho de que muestras consecutivas tengan diferentes signos algebraicos, es decir, el cruce por cero se considera cuando la señal a analizar pasa de positivo a negativo o viceversa. Por tanto se considera Tasa de Cruces por Cero a la cantidad de veces que una señal cambia de signo en una cantidad de tiempo determinada. En (Subramanian, 2004), (López, 2009), (Liang, y otros, 2010) es objeto de análisis como descriptor para la clasificación de segmentos de audio en silencio, ruido, habla o música.

Su cálculo se rige por la siguiente ecuación, ver figura 10 (Cortizo, y otros, 2009).

$$\boldsymbol{Z_s(m) = 1/N \sum_{n=m-N+1}^{m} \frac{|sign\{s(n)\} - sign\{s(n-1)\}|}{2}}$$

Figura. 10 Fórmula de la Tasa de Cruces por Cero

Donde $\boldsymbol{s(n)}$ representa la señal en el instante $\boldsymbol{n}$, $\boldsymbol{N}$ es la longitud de la señal a procesar y $\boldsymbol{sign}$ es la función signo y toma los valores de la figura 11.

$$\boldsymbol{sign}\{\boldsymbol{s}(\boldsymbol{n})\} = \begin{Bmatrix} +\mathbf{1}, \boldsymbol{s}(\boldsymbol{n}) \ \geq \mathbf{0} \\ -\mathbf{1}, \boldsymbol{s}(\boldsymbol{n}) < 0 \end{Bmatrix}$$

Figura. 11 Fórmula de la función signo

Se puede utilizar para determinar el carácter sonoro o no sonoro de un determinado segmento de audio, los segmentos sonoros tienen una menor tasa de cruces por cero que los tramos sin sonido.

Una tasa de ZCR elevada indica que el segmento de voz tiene un contenido espectral en alta frecuencia importante, mientras que una tasa baja implica que casi toda la señal está en baja frecuencia. Esta distinción del espectro permite obtener otra forma de separar los segmentos de voz sonoros de los sordos. Un segmento sonoro posee un espectro centrado en baja frecuencia y uno sordo tiene una componente en alta frecuencia superior (Bleda, y otros, 2010).

Por lo general las señales de voz presentan una alternación de sonidos con voz y sin voz, esto no ocurre en las señales de música. Esto provoca que la variación de ZCR sea mucho mayor en las señales que contienen habla que las de música (Al-Shoshan, 2006), (López, 2009), (García, 2010), (Liang, y otros, 2010).

1.3.5.4 Autocorrelación

La autocorrelación es una medida del parecido de una señal consigo misma y se emplea para determinar el tono (*pitch*) de la voz. El *pitch*, en tratamiento de voz, es la frecuencia fundamental a la cual vibran las cuerdas vocales, esta vibración produce una repetición en el patrón temporal de la señal y esta repetición se ve reflejada en forma de un máximo local en la función de autocorrelación. En ausencia de vibración de las cuerdas vocales, esto es, cuando el segmento es sordo, la autocorrelación no indica nada (Bleda, y otros, 2010).

Su cálculo en el dominio temporal se realiza por medio de la siguiente expresión, ver figura 12 (Martinez, 2010).

$$\boldsymbol{R}_{\mathbf{m}}[\boldsymbol{k}] = \sum_{\substack{\boldsymbol{n}=\mathbf{0} \\ \boldsymbol{k}=\mathbf{0},\mathbf{1},\mathbf{2},\ldots,\boldsymbol{p}.}}^{\boldsymbol{N}-\mathbf{1}} \{\boldsymbol{w}[\boldsymbol{m}-\boldsymbol{n}]\boldsymbol{x}[\boldsymbol{n}]\}\{\boldsymbol{w}[\boldsymbol{m}-(\boldsymbol{n}+\boldsymbol{k})]\boldsymbol{x}[\boldsymbol{n}+\boldsymbol{k}]\}$$

Figura. 12 Fórmula de la función de Autocorrelación

Donde $\boldsymbol{x}$ representa a la señal de voz, $\boldsymbol{w}$ representa al perfil de la ventana de análisis, $\boldsymbol{N}$ es el tamaño de la ventana y $\boldsymbol{p}$ es el número de elementos a calcular de la autocorrelación.

1.3.5.5 Coeficientes de Predicción lineal (LPC)

Los Coeficientes de Predicción Lineal o *Linear Prediction Coefficients* (LPC por sus siglas en inglés) son muy utilizados en la extracción de características del tracto vocal, proporcionan un sistema de análisis–síntesis para las señales de voz. Teniendo una muestra de voz considerando a ser estacionaria en una ventana o tramas estas pueden ser sintetizadas y modeladas por una combinación lineal. Los LPC extraen la envolvente espectral de la señal de voz utilizando una pequeña cantidad de recursos computacionales y obteniendo una buena resolución en frecuencia y sobre todo en el tiempo.

Se obtienen a partir del análisis de la señal de voz empleando para ello la autocorrelación. Mediante los coeficientes LPC se obtiene un filtro, denominado comúnmente filtro LPC o filtro del tracto vocal, cuya respuesta en frecuencia es aproximadamente igual a la del tracto vocal de la persona que ha producido el sonido, es decir, es un filtro capaz de proporcionar el mismo timbre de voz al sonido que pasa a través del mismo (Bleda, y otros, 2010).

Según (Borrero, y otros, 2011), (Alejandro, y otros, 2006), (Rojo, 2011) los LPC están basados en la, producción del habla. Se utiliza debido a que proporciona un modelo adecuado de la señal de voz; sus parámetros se ajustan a las características del tracto vocal; representa la envolvente espectral de la señal de forma comprimida; los parámetros obtenidos mediante predicción lineal muestran un espectro suavizado que proporciona la información más representativa de la voz y es un método preciso, adecuado para cálculo, tanto por su sencillez como su rapidez de ejecución.

La idea de un modelo LPC es que muestras adyacentes en la señal de voz están altamente correlacionadas, y por lo tanto, el comportamiento de esta en un momento dado, puede predecirse de

cierta cantidad de muestras anteriores realizando para esto la siguiente aproximación, ver figura 13 (Calvo, y otros, 2008).

$$\boldsymbol{s[n] = \sum_{k=1}^{p} a[k]s[n-k]}$$

Figura. 13 Aproximación de la señal mediante LPC

Donde $\boldsymbol{p}$ es el orden de la predicción. Luego, el objetivo del método LPC consiste en determinar los coeficientes de predicción $\{\boldsymbol{a[k]}|\boldsymbol{k = 1, \dots, p}|\}$ tal que el error de predicción promedio o residuo sea mínimo.

Otra alternativa para el uso de los coeficientes de predicción LPC son los coeficientes LPC-ceptrales que son el resultado de aplicarle el análisis cepstral (cepstrum) a salida del filtro LPC. Los coeficientes ceptrales tienen una gran importancia, ya que aprovecha los beneficios de la predicción lineal, eliminando ciertos aspectos de ruido y perturbación incluidos dentro de las tramas de datos. Estos coeficientes también pueden ser calculados mediante un proceso iterativo cuya fuente del mismo se basa en los coeficientes LPC (Carmona, 2006), (Calvo, y otros, 2008), (Fook, y otros, 2010), (Venkateswarlu, y otros, 2012), (Bhattacharje, 2013).

Estos coeficientes LPCC son muy usados en sistemas de reconocimiento automático del habla y de locutores (Carmona, 2006), (Calvo, y otros, 2008), (Venkateswarlu, y otros, 2012);

1.3.5.6 Coeficientes Cepstrum

El Cepstrum es una modificación no lineal del espectro que permite determinar, entre otras cosas, el tono de la voz y la respuesta en frecuencia del tracto vocal. Como en el caso de la autocorrelación, el pitch aparecerá como un máximo local dentro del cepstrum, y el tracto vocal estará representado por los primeros coeficientes del cepstrum. El cálculo del cepstrum es un poco más elaborado que el de los parámetros anteriores y debe realizarse tal y como indica el diagrama de bloques: ver figura 14 (Bleda, y otros, 2010).

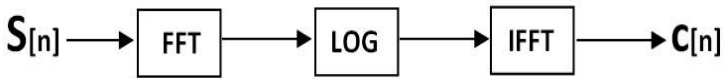


Figura. 14 Cálculo del Cepstrum

Una vez obtenido el espectro del segmento de voz **S** mediante la Transformada Rápida de Fourier (*Fourier Fast Transform* FFT pos sus siglas en inglés), se calcula el logaritmo del espectro (cepstrum complejo) o del módulo del espectro (cepstrum real) para luego regresar al dominio temporal mediante la Transformada Inversa de Fourier (*Inverse Fourier Fast Transform* IFFT por sus siglas en inglés). Cuando el segmento de voz tratado es sonoro, aparece en el cepstrum un máximo cuya separación del origen (en número de muestras) es equivalente al período del pitch de la voz, ver figura 40 anexo 2 (Bleda, y otros, 2010).

1.3.5.7 Coeficientes Cepstrales en Escala Mel (MFCC)

Los coeficientes Cepstrales en escala Mel (MFCC) son características espectrales basadas en la percepción auditiva humana, las cuales son extensamente usadas para el procesamiento de audio y reconocimiento del habla. En (Vicente, 2007), (Calvo, y otros, 2008), (Kua, y otros, 2010), (Saravia, 2011), (Bermejo, 2008) se expresa que los MFCC son la parametrización más popular de la señal de habla en los sistemas de reconocimiento automático del habla, así como sistemas para el procesamiento de señales de voz y en (Martínez, y otros, 2008) se considera que los coeficientes MFCC ofrecen mejores resultados para el tratamiento de señales de voz que los Coeficientes de Predicción Lineal (LPC).

En (Rojo, 2011), (Saravia, 2011) se considera que entre las muchas técnicas de parametrización del habla, la más empleada es *Mel Frecuency Cepstral Coefficients* o MFCC. MFCC es la técnica de parametrización del habla más utilizada en los sistemas automáticos de reconocimiento de voz, principalmente porque se adapta bien a las hipótesis utilizadas para estimar las distribuciones de estado en HMMs y, también, debido a la robustez de ruido superior que ofrece sobre otras técnicas alternativas de extracción de características, como, por ejemplo, LPCC y LPC.

Para obtener más información, como por ejemplo la de coarticulación de fonemas, es necesario introducir datos de la velocidad y aceleración de los parámetros. Así surgen los MFCC-Delta (o ΔMFCC) y los MFCC- Delta-Delta (o $\Delta\Delta$MFCC), que representan la evolución temporal de los fonemas en su transición a

otros fonemas. Los ΔMFCC se calculan como la variación de los coeficientes MFCC con respecto a un instante de tiempo. Por ello, son denominados coeficientes de velocidad (ya que dan los cambios por tiempo) o de primera derivada. Los coeficientes ΔΔMFCC representan la variación de los coeficientes de velocidad, por lo que son llamados coeficientes de aceleración (Bermejo, 2008), (Rojo, 2011), (Saravia, 2011).

Los MFCC representan la envolvente espectral de la señal de voz, obteniendo así de ellos importantes características del habla. En concreto, el primer coeficiente, C_0, indica la energía de la señal. Y el segundo coeficiente, C_1, tiene una razonable interpretación como indicador del balance global de energía entre bajas y altas frecuencias (Bermejo, 2008), (O'Shaughnessy, 2008).

Están basados en la descomposición de la señal de voz. Esta señal es dividida en ventanas o *frames*, a cada ventana se le aplica la Transformada Rápida de Fourier (Zuazua, y otros, 2005), (Rojo, 2011), (Rincón, y otros, 2008), (Ramírez, 2008). Este resultado es transformado con ayuda de filtros (*filter-bank*[5]) que usan la escala Mel en función de la frecuencia (Ong, y otros, 2004), (Rojo, 2011), (Rincón, y otros, 2008) para luego aplicar el logaritmo y la Transformada Discreta del Coseno (DCT) (Rojo, 2011), y obtener los coeficientes Cepstrales en escala Mel (MFCC) como parámetros acústicos de la señal de voz (Iqbal, y otros, 2011).

Los pasos básicos para realizar un análisis acústico haciendo uso de los MFCC se muestran en el diagrama de la figura 16 (Rincón, y otros, 2008):

- Espectro de Potencia: Este se obtiene de cada uno de los bloques ventaneados usando la Transformada Rápida de Fourier.

- Banco de Filtros en la escala Mel: Es un banco de filtros triangulares que se encargan del cálculo del espectro promedio alrededor de cada frecuencia central con anchos de banda que aumentan de manera progresiva.

[5] *Los Filter-bank en escala Mel son un banco de filtros triangulares que se encargan del cálculo del espectro promedio alrededor de cada frecuencia central con anchos de banda que aumentan de forma progresiva.*

- Calculo de la Energía log: Este cálculo se realiza a la salida de cada filtro, para tratar de simular la relación no lineal existente entre la intensidad emitida y la intensidad percibida cuya naturaleza es netamente logarítmica.

- Coeficientes Cepstrum: Los coeficientes cepstrum derivados de la escala Mel se obtienen a partir de la Derivada Discreta del Coseno de las salidas de los filtros.

La escala Mel son escalas utilizadas para el tratamiento de audio, se utiliza debido a que la sonoridad, altura y timbre de una señal acústica son dependientes entre sí, la modificación de una modifica las otras. Las bandas Mel son un conjunto de filtros de banda crítica que modelan el comportamiento del sistema auditivo humano, para esto se basa en la escala de frecuencia Mel que es lineal en frecuencias por debajo de 1000 Hz y logarítmica en frecuencias por encima de 1000 Hz (Peeters, 2004), ver figura 15. Estos coeficientes MFCC son más robustos en presencia de ruido que los LPC.

$$\boldsymbol{M} = \begin{cases} \boldsymbol{M} = \boldsymbol{f}, \qquad \boldsymbol{f} < 1000\boldsymbol{Hz} \\ \boldsymbol{M} = \boldsymbol{fc}\left(\boldsymbol{1} + \boldsymbol{log_{10}}\left(\frac{\boldsymbol{f}}{\boldsymbol{fc}}\right)\right), \boldsymbol{f} > 1000\boldsymbol{Hz} \end{cases}$$

Figura. 15 Conversión de Hz a escala Mel

Donde $\boldsymbol{M}$ es la frecuencia expresada en Mel y $\boldsymbol{f}$ en Hz , y $\boldsymbol{fc}$ = 1000 Hz.

En (Sung, y otros, 2009), (Adrian, y otros, 2010), (Rojo, 2011) y (Rúa, y otros, 2009) se utilizan 12 coeficientes MFCC para el reconocimiento del habla mediante el modelado fonético, además se incluyen el log-Energy y los delta MFCC y delta-delta MFCC conformando un vector de 39 coeficientes. Donde la utilización de las variaciones delta y delta-delta son los que mejor representan las características del habla. También en (Calvo, y otros, 2008) y (Caravaca, 2011) se plantea que es muy común utilizar 12 coeficientes cepstrales MFCC así como 12 rasgos delta y 12 delta-delta. En (Bermejo, 2008) se utilizan 13 coeficientes MFCC así como sus delta y delta-delta, conformando un vector de 39 coeficientes para la creación de los modelos acústicos.

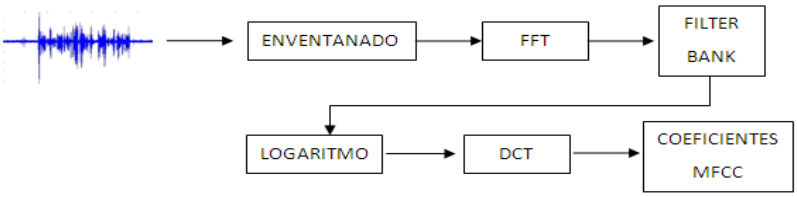


Figura. 16 Proceso de extracción de los coeficientes MFCC

1.4 Análisis de soluciones existentes

Existen varios sistemas de reconocimiento automático del habla, algunos de los cuales están ya siendo comercializados. Se pueden destacar los denominados *Dragon Dictate* (*Dragon Systems*) y *Naturally Speaking* (IBM). La mayoría de estos sistemas están basados en Modelos Ocultos de Markov (*Hidden Model Markov HMM* por sus siglas en inglés) o su versión híbrida con redes neuronales y llegan a un reconocimiento del orden del 95%, en discurso continuo, para un único hablante, con un micrófono de buena calidad y en un ambiente de bajo ruido.

Dragon Naturally Speaking. Es un software que a través de él puedes crear documentos de Microsoft Office utilizando la voz y redactar todo tipo de cartas, informes y mensajes de correo electrónico. Esta herramienta fue creada por *Nuance Communications Inc*. dentro de la gama de herramientas *Dragon Naturally Speaking Solutions* (Ecured, 2011).

Enfocado puramente al sector profesional (en gran medida por su elevado precio), *Dragon Dictate* permite transcribir de forma precisa archivos de audio grabados en casi cualquier formato existente ya que soporta archivos .wav, .m4a, .m4v, .mp4 y .aif, siendo este último formato el que utilizan los programas como *Dragon Dictation* para iPhone y iPad. Entre sus características, se encuentra la precisión de la aplicación y una forma de acometer las correcciones mediante comandos de voz (Hipertextual SL, 2012).

Verbio WordSpotting es el sistema de reconocimiento de *Verbio* orientado a la búsqueda de palabras claves en fragmentos de audio, con múltiples finalidades, tales como indexación de contenidos, análisis de llamadas recibidas en *call center*. Basado en el sistema de reconocimiento de voz independiente de locutor *Verbio ASR*, permiten la búsqueda de texto en fragmentos de audio con velocidades de proceso muy superiores al tiempo real. La API de *Verbio WordSpotting* desarrollada por *Verbio* permite identificar palabras aisladas, conjuntos de palabras e incluso parámetros concretos dentro de ficheros de audio o

muestras de sonido obtenidas en tiempo real. El resultado se vuelca en un fichero que contiene, junto a cada palabra detectada, el instante temporal de la aparición, así como su nivel de confianza o score (Verbio Tecnologies . S. L, 2011).

ReconVox es un producto de altas prestaciones de reconocimiento del habla. Gracias a su capacidad de reconocer tanto órdenes aisladas como habla continua para cualquier locutor sin necesidad de entrenamiento previo, permite desde controlar dispositivos electrónicos mediante la voz hasta acceder a servicios automáticos manejados mediante sentencias complejas a través de cualquier teléfono. Está disponible tanto en forma de DLL (*Dynamic Link Library*) para plataformas *Windows* como en forma de librería dinámica para plataformas *UNIX/Linux* (DTEC, 2009).

La mayoría de las aplicaciones existentes actualmente para el reconocimiento del habla están enfocadas mayormente a la identificación del locutor con dependencia o independencia del texto, para la realización de dictados sobre editores de textos o la redacción de mensajes de correos electrónicos, transcripción de textos, autenticación mediante la voz o personalización de ordenadores. Las principales dificultades para la obtención de estas aplicaciones anteriormente analizadas son en su mayoría herramientas privativas, además la implementación de sus tecnologías para servidores está sobre plataformas privativas, así como incompatibilidades con sistemas de libre acceso.

1.5 Técnicas para la reducción de ruido en audio

Es muy común que en un sistema de reconocimiento automático del habla se empleen técnicas para la reducción de ruido en archivos de audio, la presencia de ruido podría provocar variaciones en las características esenciales de la voz trayendo como consecuencias bajos índices en los resultados de pruebas de estos sistemas de reconocimiento.

Por ruido se entiende toda componente de tensión o intensidad indeseada que se superpone con la componente de señal que se procesa o que interfiere con el proceso de medida. El ruido de un sistema se puede clasificar de la siguiente manera:

Ruido interno o inherente: que corresponden al que se genera en los dispositivos electrónicos como consecuencia de su naturaleza física (ruido térmico, ruido por cuantización de las cargas, ruido de semiconductor, etc.). El ruido inherente es de naturaleza aleatoria.

Ruido externo o interferencias: al que se genera en un punto del sistema como consecuencia de acoplamiento eléctrico o magnético con otro punto del propio sistema, o con otros sistemas naturales (tormentas, etc.) o construidos por el hombre (motores, equipos, etc.). El ruido de interferencia puede ser periódico, intermitente, o aleatorio. Normalmente se reduce, minimizando el acoplo eléctrico o electromagnético, bien a través de blindajes, o bien, con la reorientación adecuada de los diferentes componentes y conexiones.

Según sea la naturaleza del ruido que exista, se deben utilizar diferentes criterios. En el caso de que se trate de:

Ruido blanco, es aquel que posee una densidad espectral de tensión de ruido o de intensidad de ruido constante.

Ruido rosa o ruido 1/F: es aquel que posee una densidad espectral de potencia de ruido que es inversamente proporcional a la frecuencia.

Ruido aditivo: señal aleatoria estacionaria que se supone independiente de la señal principal y que por tanto se suma en potencia a ésta.

Los primeros pasos están dados en el filtrado común de la señal contaminada, no obstante se hizo cada vez más necesario la reducción del ruido que aún queda compartiendo el ancho de banda de la señal útil (Moyano, 2005).

Una señal recibida, contaminada con Ruido Aditivo, por ejemplo con Ruido Blanco, es afectada en su banda de frecuencia y fuera de esta por las componentes de ruido. Los filtros de banda analógica y digital separan intervalos de frecuencia, así lo hace un filtro pasa banda, permite el paso de una banda de frecuencia y el resto la rechaza. Este tipo de filtro diferencia la señal con respecto al ruido fuera de su banda útil y es capaz de separarla. Es decir, se basa en el contenido de frecuencia diferente que hay entre el ruido fuera de banda y la señal útil.

1.5.1 Métodos para disminuir el ruido en banda

El proceso de disminuir la influencia del ruido también puede ser mediante la descripción estadística del modelo de ruido y de la señal a utilizar, estas tienen una característica en común: es su estadística de segundo orden.

- Filtrado Adaptativo *Wiener*: Se basa en separar ambas señales por su diferencia en cuanto a la correlación cruzada de ambas. Si este cálculo es igual a cero entonces se está en presencia de una condición idónea para emplear el método.

El principio de funcionamiento del filtrado adaptativo *Wiener* se explica a partir de las cinco magnitudes siguientes: ver figura 17.

1. $\boldsymbol{u(0), u(1), u(2)}, \ldots$: entrada del filtro.
2. $\boldsymbol{w_{0,} w_1, w_2, \ldots, w_N}$: coeficientes del filtro.
3. $\boldsymbol{d(n)}$: denominada señal de referencia, es la respuesta deseada del filtro.
4. $\boldsymbol{y(n)}$: respuesta del filtro.
5. $\boldsymbol{e(n)}$: es la diferencia entre la salida del filtro y la señal deseada.

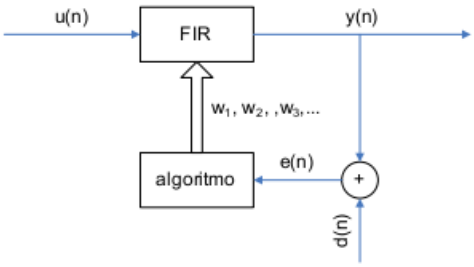


Figura. 17 Esquema del Filtro Adaptativo

La idea del filtrado es hacer $\boldsymbol{d(n)}$ lo más parecido posible a $\boldsymbol{y(n)}$**.** En este sentido es de esperar que con un buen diseño, calculando correctamente los coeficientes del filtro por parte del bloque algoritmo en la figura 17, para cada instante, $\boldsymbol{y(n)}$ se aproxime a $\boldsymbol{d(n)}$, de modo que la medida de similitud entre las señales estará dada por comparar dos valores: $\boldsymbol{y(n)}$ y $\boldsymbol{d(n)}$**,** (Takuya Yoshioka, 2008).

- Sustracción Espectral: Un enfoque intuitivo para la supresión de ruido es substraer una estimación del mismo al habla contaminada. Las técnicas de sustracción espectral realizan esta tarea sobre la representación de la señal ruidosa en el dominio de Fourier. Apoyándose en la mayor importancia de la magnitud en la percepción del habla, el enfoque consiste en estimar la magnitud del espectro de la señal limpia, y asignarle la fase de la señal contaminada. En cada tramo de análisis, la estimación de la magnitud del habla limpia se obtiene restando una estimación del espectro de magnitud del ruido al espectro de magnitud de la señal ruidosa (Tomassi, y otros, 2005).

Esta técnica intenta estimar la señal limpia a partir del espectro de la señal ruidosa, sustrayéndole una versión estimada del ruido que es generalmente actualizada durante los intervalos de ausencia de voz.

- Estimadores de *Ephraim* y *Malah*: En lugar de plantear un problema de optimización directamente sobre la señal limpia y la estimada, los estimadores de *Ephraim-Malah* minimizan el error cuadrático medio sobre la magnitud o sobre el logaritmo de la magnitud de los coeficientes en la representación de Fourier del habla limpia. El enfoque supone que estos coeficientes, al igual que los del ruido, son variables aleatorias estadísticamente independientes cuyas componentes reales e imaginarias pueden modelarse mediante una distribución gaussiana con media cero (Tomassi, y otros, 2005).

Estos estimadores tratan minimizan el error cuadrático medio para la amplitud o para el logaritmo de la amplitud del espectro de la señal limpia, suponiendo que tanto el ruido como la señal de interés son procesos aleatorios independientes.

Existan algunas variaciones de LPC para combatir el ruido entre ellas se encuentra: OSALPC (*One Sidec Autocorrelation Linear Predictive Coding* por sus siglas en ingles). En esta técnica se realiza Predicción Lineal en el dominio de la autocorrelación en lugar de sobre la señal misma y su uso en reconocimiento de habla en ambientes ruidosos es atractivo debido a sus prestaciones con respecto a la técnica LPC convencional en condiciones severas de ruido blanco aditivo y su simplicidad computacional (Javier Hernando, 1997).

1.6 Conclusiones Parciales

Se analizaron algunas soluciones existentes y entre los principales factores para la obtención de estas aplicaciones son en su mayoría herramientas privativas. Analizando los métodos de detección de palabras

clave existentes donde los mejores resultados se obtienen utilizando sistemas de reconocimiento de habla continua basados en vocabularios de gran tamaño. Después de analizar una serie de descriptores se tiene que los mejores descriptores de una señal de voz son los coeficientes LPC, LPC-Cepstrales y MFCC que caracterizan el tracto vocal y son muy utilizados en sistemas de reconocimiento automático del habla y de detección de palabras clave sobre audio. Se analizaron descriptores como energía y ZCR los cuales son muy útiles para lograr la detección de extremos de una señal además de discriminar entre espacios de silencio o sonido.

CAPÍTULO 2: PRESENTACIÓN DE LA SOLUCIÓN PROPUESTA

En el presente capítulo se realiza una descripción sobre los distintos clasificadores utilizados en el modelado de las características acústicas del habla para su reconocimiento automático y el desarrollo de sistemas de tratamiento de señales de voz, así como se presentan las técnicas y descriptores utilizados para la segmentación y extracción de características de la señal de voz para su posterior clasificación.

2.1 Clasificadores usados para el modelado de las señales de habla

A partir de la obtención de los rasgos acústicos de la voz, previamente extraídos y seleccionados se hace necesario encontrar un modelo que clasifique los mismos y sea lo suficientemente robusto ante la variabilidad del habla.

Estos métodos de clasificación surgieron a causa del desarrollo de las investigaciones sobre reconocimiento del habla. En la década del 70 predominaba la clasificación con el uso de plantillas de palabras. Ya para la década de los 80s el proceso de clasificación se hacía mediante la Distorsión Dinámica en el Tiempo (DTW *Dynamic Time Warping*) y la Cuantización Vectorial (VQ) además se introdujo la aplicación de Redes Neuronales para el reconocimiento del habla. En los 90s se comenzó el desarrollo de clasificadores basados en enfoques estadísticos con los Modelos Ocultos de Markov (HMM *Hidden Model Markov*) y los Modelos de Mezclas Gausianas (GMM) (Calvo, y otros, 2008). Ver figura 41 anexo 2.

Estos modelos de clasificación trabajan a partir de un conjunto de modelos captados automáticamente en fase de entrenamiento, la pronunciación es decodificada. Las dos fases que componen a la mayoría de estos modelos, son: la fase de entrenamiento, donde se generan los modelos de referencia, empleando un conjunto de bases de datos de voces grabadas con la suficiente variabilidad; y la fase de reconocimiento, donde se realiza una comparación entre las pronunciaciones y las referencias obtenidas, eligiendo la secuencia de palabras cuya distancia a los modelos de referencia sea menor (Calvo, y otros, 2008), (García, 2008).

La señal de habla es muy variable en el tiempo, además está compuesta por espacios de sonido y silencio. La transcripción del habla utilizando las características acústicas de la misma, se apoya en el

número finito y diferenciado de unidades fonéticas que contiene el lenguaje hablado. Los fonemas se estructuran en palabras, y éstas en frases que modelan las ideas (García, 2008).

A continuación se analizan algunos de los clasificadores más utilizados.

2.1.1 Distorsión Dinámica del Tiempo (DTW)

Dinamic Time Warping o Distorsión Dinámica del Tiempo se basa en la alineación temporal de las características de una pronunciación respecto a un patrón de referencia antes de calcular una puntuación general. La referencia se adapta estirándose o comprimiéndose en el dominio temporal, así los vectores de características se alinean con el patrón antes de hacer la medida de distancia (García, 2008). Es muy utilizado en sistemas que se basan en el reconocimiento del locutor dependiente del texto (Calvo, y otros, 2008), (Moreno, y otros, 2010). En (Martín, y otros, 2004) fue usado para la implementación de un reconocedor de palabras aisladas dependiente del locutor y se concluye que hasta cierto punto brinda resultados aceptables pero se presenta un problema dado que requiere más procesamiento que las técnicas empleadas en sistemas independientes del locutor, para este caso en particular tuvo una efectividad del 85%. Los modelos DTW comparan el vector de entrada con un conjunto de rasgos acústicos ordenados en el tiempo que constituyen las expresiones a reconocer.

Estas expresiones son almacenadas en las plantillas durante la fase de entrenamiento. En el reconocimiento se alinea de manera óptima la secuencia de datos de entrada con el modelo de referencia almacenado con anterioridad. Una vez concluida la comparación, 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 (Moreno, y otros, 2010).

Entre las ventajas que posee este modelo es que usa algoritmos muy simples y con un grado de computación reducida, es por eso que se utiliza generalmente en aplicaciones sencillas que obtienen muy buenos resultados. Como es el caso de (Borrero, y otros, 2011) donde fue usado para el reconocimiento de palabras aisladas para un mini-robot usando para el experimento 4 palabras y voces de 25 hombres y 25 mujeres. En esta tarea sólo alcanzó una efectividad de 84.45 %.

El principal inconveniente es la dificultad de generalizar, ya que requiere un alto procesado en tareas muy complejas con grandes vocabularios (García, 2008). Una desventaja del mismo es que es altamente dependiente de las expresiones de referencia, imposibilitando la variabilidad en la señal de voz (Moreno, y otros, 2010).

2.1.2 Modelos Ocultos de Markov (HMM)

Los Modelos Ocultos de Markov (*Hiden Markov Model* HMM por sus siglas en inglés) según (Sung, y otros, 2009), (Deng, y otros, 2007), (Vicente, 2007), (Rojo, 2011), (Ceballos, y otros, 2010), (Alfonso, y otros, 2009), (García, 2008), (Bermejo, 2008) se han convertido en uno de los métodos estadísticos más potentes para el modelado acústico de las señales de habla. Sus principios han sido utilizados con éxito en reconocimiento de voz automático, análisis de tonos y formantes, mejora de expresión, síntesis de voz, comprensión de lenguaje hablado (García, 2008), (Bermejo, 2008), (Otero, 2010). Un HMM es una máquina de estados finita, en la que las observaciones son una función probabilística del estado, es decir, el modelo es un proceso doblemente estocástico (Martínez, y otros, 2008), formado por un proceso estocástico oculto no observable directamente, que corresponde a las transiciones entre estados y un proceso estocástico observable cuya salida es la secuencia de vectores espectrales (García, 2008), (Bermejo, 2008), (Gila, 2012).

Los modelos HMM surgieron de la necesidad de utilizar técnicas de modelado estadístico que fueran capaces de afrontar el problema de la variabilidad de la voz, que crece de manera importante cuando la complejidad y el tamaño del vocabulario aumentan. Ésta es la diferencia fundamental con los reconocedores de voz basados en DTW, que utilizan otras pronunciaciones deterministas como patrones de referencia. En (García, 2008) se expresa que los modelos HMM presentan mayor capacidad de modelado de las características acústicas del habla que los DTW. Estos modelos se utilizan en dos niveles. Por un lado, para modelar palabras mediante la utilización de modelos acústicos. Por otro, para modelar el lenguaje, definiendo una red de reconocimiento en la que se imponen las restricciones gramaticales, a partir de las cuales se realiza la decodificación de la secuencia de palabras más probable utilizando la información del nivel acústico. En cualquier caso, la utilización de los HMM en el reconocimiento de voz se aplica teniendo en cuenta dos suposiciones (Tera, 2002):

- La señal de voz puede ser separada en pequeños segmentos (estados) en los que la señal puede considerarse prácticamente estacionaria. Se asumen que las transiciones entre segmentos contiguos son instantáneas.

- La probabilidad de observación de que un vector de características se genere sólo depende del segmento actual y no de símbolos anteriores o posteriores.

Cada HMM viene dimensionado por tres ejes de referencia: la matriz de probabilidades de transición entre estados, el vector de probabilidades del estado inicial y la distribución de probabilidades de salida u observación. Respecto a esta última, existe la posibilidad de trabajar con HMM de observación discreta o *Discrete Hidden Model Markov* (DHMM, por sus siglas en inglés) o continua llamados *Continuos Hidden Model Markov* (CHMM, por sus siglas en inglés) (Rojo, 2011). En general, los CHMM modelan con mayor fidelidad y producen tasas más altas de identificación que los no continuos, aunque, en condiciones de trabajo en tiempo real, los discretos parecen resultar más rápidos (Moreno, y otros, 2010), (Rojo, 2011).

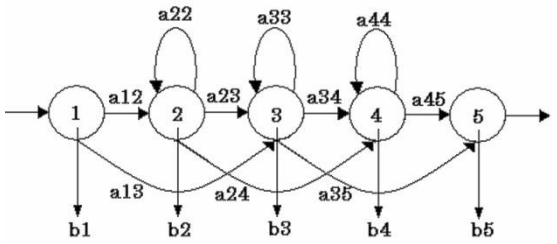


Figura. 18 Estados de un HMM

Una gran ventaja de los sistemas de reconocimiento que utilizan HMM respecto a los demás es su gran versatilidad, tanto en los procesos de entrenamiento como en algunas características variables de la muestra: duración y contexto (Acevedo, 2008), (Moreno, y otros, 2010). Los HMM son ampliamente aplicados en sistemas de reconocimiento de voz, ya que con su uso, se puede realizar la agrupación en clases de las palabras o fonemas independientemente por diferente HMM, siendo discriminada cada clase fácilmente en la etapa de reconocimiento (Acevedo, 2008).

Un modelo HMM está compuesto por los siguientes elementos (Bermejo, 2008), (García, 2008), (Saravia, 2011), (Rojo, 2011):

- ***N:*** Número de estados del modelo. Los estados se denotan como: $\boldsymbol{S} = \{\boldsymbol{S_1}, \boldsymbol{S_2}, \dots, \boldsymbol{S_N}\}$**.** Donde $\boldsymbol{q_i(t)}$ indica la posibilidad de que un vector en el tiempo $\boldsymbol{t}$ esté en el estado $\boldsymbol{S_i}$.

- ***A:*** Matriz de transición de probabilidades de estados. Cada uno de los elementos se define como: $\boldsymbol{a_{ij}} = \boldsymbol{p}\big(\boldsymbol{q_j}\big)(\boldsymbol{t}+\mathbf{1})|\boldsymbol{q_i}(\boldsymbol{t}))$.

- ***B:*** Matriz de probabilidades de generación de vectores o distribuciones. Cada uno de sus elementos se define como la probabilidad de que un determinado vector de características en un determinado tiempo haya sido generado por la distribución correspondiente: $\boldsymbol{b_j}(\boldsymbol{y}(\boldsymbol{t})) = \boldsymbol{p}\big(\boldsymbol{y}(\boldsymbol{t})|\boldsymbol{q_i}(\boldsymbol{t})\big)$.

- **π:** Vector con las distribuciones de estados inicial: $\boldsymbol{\pi_i} = \boldsymbol{p}\big(\boldsymbol{q_i}(\mathbf{1})\big)$.

Así cualquier modelo HMM puede ser definido de la siguiente manera: $\boldsymbol{\lambda} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$**.**

Dada la definición anterior para un HMM, existen tres problemas básicos a resolver para que el modelo pueda ser usado en aplicaciones reales (X. Huang, 2001), (Bermejo, 2008):

- Problema de Evaluación: Dada una secuencia de observaciones y un modelo, se busca cómo calcular la probabilidad de que la secuencia observada haya sido producida por dicho modelo.

- Problema de Decodificación: Dada una secuencia de observaciones y un modelo, se busca cómo elegir una secuencia de estados que sea óptima en algún sentido.

- Problema de Aprendizaje: Dada una secuencia de observaciones de entrenamiento, se busca cómo obtener los parámetros del modelo de forma óptima.

Los modelos HMM como parte de los modelos probabilísticos de reconocimiento basados patrones son en la actualidad en aplicaciones de reconocimiento del habla o sistemas para el procesamiento de voz, así como mejorar su rendimiento formando sistemas híbridos de reconocimiento.

2.1.3 Cuantización Vectorial (VQ)

En la cuantización vectorial cada vector N-dimensional de entrada (un punto en el o espacio N-dimensional) se representa por el más cercano *codevector* o centroide de una pequeño grupo de vectores

o *codebook* 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 *clusters* o grupos en los que se hayan dividido los datos de entrada (Calvo, y otros, 2008). La cuantificación vectorial consiste en representar las características de los fonemas como un espacio vectorial de dimensión el numero de parámetros, de forma que al fonema a reconocer se le asigna el vector cuya distancia a él sea mínima. Por tanto, los fonemas quedarán representados por unos vectores determinados centroides de forma que todos los puntos que caigan en una zona determinada se asignarán a dicho vector (Gila, 2012).

Una parte importante en cualquier tipo de procesamiento de voz es la optimización de los algoritmos en cuanto a velocidad y almacenamiento, entonces, la cuantificación de vectores trae consigo la idea de clasificar un conjunto de vectores, luego de lo cual se buscarán los mejores representantes para reducir el tamaño de la información a manejar. La forma de medir la fidelidad de un cuantificador es determinar el error que éste produce al reemplazar los datos de entrada que recibe por los vectores representantes, dicho parámetro es llamado error por distorsión. La finalidad de un cuantificador es obtener un conjunto de vectores representativos llamado *codebook*, que presente el menor error por distorsión, por ejemplo para cuantificar los vectores de observación. Entre las principales ventajas se encuentra (Martínez, y otros, 2008):

- Reduce el almacenamiento de la información de análisis.
- Se reduce el cálculo para determinar distancias entre vectores espectrales. La representación del VQ se limita a una tabla que contiene las distancias entre pares de vectores del *codebook*.
- Representación discreta de las señales de voz. Asociando una característica fonética con cada vector del *codebook*, el proceso de elección del vector que mejor lo representa es equivalente a asignar una característica fonética a cada segmento de voz.

Principales desventajas:

- Distorsión en la representación del vector. Hay un número finito de vectores en el *codebook*, el proceso de elección del mejor representante es equivalente a cuantificar el vector y conduce a un cierto nivel de error de cuantificación.

- El almacenamiento requerido para los vectores del *codebook* no es pequeña.

Para construir un cuantificador de vectores se necesita los siguientes componentes: Ver figura 19.

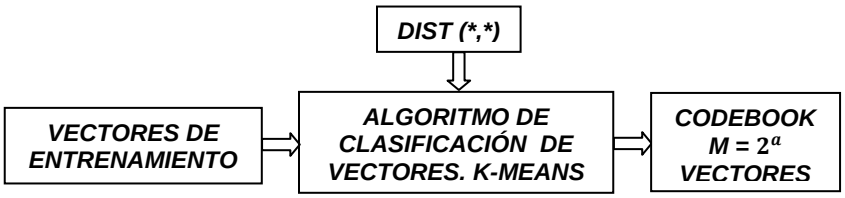


Figura. 19 Componentes de un VQ

- Un gran número de vectores de observación, $\{\boldsymbol{V_1}, \boldsymbol{V_2}, \dots, \boldsymbol{V_N}\}$ que conforman el grupo de entrenamiento. El grupo de entrenamiento se usa para crear el grupo de vectores del *codebook* óptimo que representa la variabilidad espectral observada en el grupo de entrenamiento. Se determina el tamaño del *codebook* como $\boldsymbol{M = 2^a}$, siendo a el número de bits necesarios para codificar ***M*** palabras de código, por lo tanto se necesitan ***n>M*** vectores para que sea eficaz.

- Una medición de distancia entre cada par de vectores espectrales de observación para agrupar el conjunto de vectores de entrenamiento como así también para asociar o clasificar vectores arbitrarios a cada entrada del *codebook*.

- Un procedimiento de clasificación para ubicar y calcular los centroides. Sobre la base del particionamiento que clasifica el grupo de ***n*** vectores en ***M*** *clusters* o sectores, primero elegimos el número ***M***, observaciones del *codebook*, para luego proceder a la clasificación.

Finalmente, luego del proceso de clasificación (entrenamiento) queda como resultado del mismo un libro de códigos o *codebook* (Martínez, y otros, 2008).

2.1.3.1 Algoritmo K-means

Por largo tiempo K-Means ha sido considerado el estándar dentro del *clustering* y permanece con gran fuerza en este campo. Lo primero que necesita es que sea especificado con anticipación el número de *cluster* deseado: este es el parámetro k. Entonces k valores son elegidos aleatoriamente y asignados como centros de *clusters*. Seguidamente todos los vectores son asignados al centro de *cluster* más cercano de acuerdo a una medida de distancia. Después los centroides o medios de los centro de *clusters*

son calculados. Estos centroides son tomados para ser los nuevos centros para sus respectivos *clusters*. Finalmente se repite el proceso completo con los nuevos centros.

Esto ocurrirá iterativamente hasta que los centros se estabilicen o estos no cambien mucho. Este método es simple, efectivo y fácil de probar, busca en su funcionamiento minimizar una función objetivo dada por el cuadrado total de la distancia desde cada documento de cada *cluster* a cada centro de *cluster* (Rodríguez, 2011).

El algoritmo está compuesto por los siguientes pasos:

1. Elegir el número de *clusters*, k.
2. Generar aleatoriamente k *clusters* y determinar el centro de los *clusters*, o directamente tomar aleatoriamente k objetos como los centros del *clusters*.
3. Asignar cada objeto al centro de *cluster* más cercano.
4. Recalcular los nuevos centros de *clusters*.
5. Repetir el paso 3 y 4 hasta que el criterio de convergencia sea alcanzado (usualmente que las asignaciones no cambien entre una iteración y otra o los centros no cambien mucho)

2.1.3.2 Vecino más cercano (k-Nearest Neighbor)

En el algoritmo del vecino más cercano o *k-Nearest Neighbor* (KNN por sus siglas en inglés) la función de proximidad puede decidir la clasificación de un nuevo ejemplo atendiendo a la clasificación del ejemplo o de la mayoría de los *k* ejemplos más cercanos. Admite también funciones de proximidad que consideren el peso o coste de los atributos que intervienen, lo que permite, entre otras cosas, eliminar los atributos irrelevantes (Martineaux, y otros, 2008). Una función de proximidad clásica entre dos instancias $\boldsymbol{X_i}$ y $\boldsymbol{X_j}$, si suponemos que un ejemplo viene representado por una $\boldsymbol{n}$ tupla de la forma $(\boldsymbol{a_1}(\boldsymbol{x}), \boldsymbol{a_2}(\boldsymbol{x}), \dots, \boldsymbol{a_n}(\boldsymbol{x}))$ en la que $\boldsymbol{a_r}(\boldsymbol{x})$ es el valor de la instancia para el atributo $\boldsymbol{a_r}$, es la distancia euclídea, que se muestra en la ecuación de la figura 20 (Fook, y otros, 2010), (Graña, y otros, 2012).

$$D(X_i, X_j) = \sqrt{\sum_{l=1}^{n} (X_{il} - X_{jl})^2}$$

Figura. 20 Ecuación de la distancia Euclidiana

2.1.4 Redes Neuronales (ANN)

Las redes neuronales (ANN por sus siglas en inglés) son modelos computacionales que emulan el comportamiento del cerebro humano por medio de tipologías que reflejan la interconexión de las células nerviosas (Moreno, y otros, 2010).

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.

En sentido general 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. Ver figura 21. Una ANN “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 a las clases correspondientes a obtener en la capa de salida; las neuronas correspondientes a las capas ocultas se entrenan de forma tal que se cargue el patrón a la entrada, la salida de la clase a la cual corresponde el mismo, se active (Llerena, 2006).

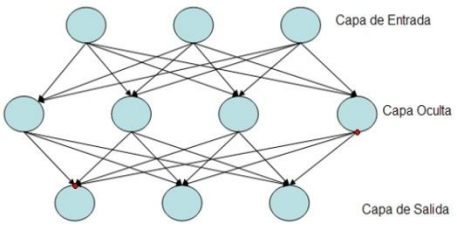


Figura. 21 Ejemplo de una red neuronal artificial

Las ANN son estructuras con capacidad de clasificación y discriminación, intrínsecamente no lineales, que les permite aprender una determinada tarea a partir depares observación-objetivo, sin necesidad de suposiciones sobre el modelo subyacente (García, 2008).

La capacidad de aprendizaje, estructura de cálculo distribuido-paralelo, tolerancia a fallos, adaptabilidad y plasticidad, son características que posee una RNA(Red Neuronal Artificial), a través de las cuales, puede llegar a resolver problemas que necesitarán gran cantidad de tiempo en ordenadores convencionales, además de poseer gran velocidad y robustez, ya que el conocimiento adquirido se encuentra repartido por toda la red, de forma que si se daña una parte, se continúan generando cierto número de respuestas correctas. Dentro de las principales tareas de una RNA se encuentra la de clasificar datos, es decir, debe identificar si una entrada pertenece a una clase o a otra.

Las principales ventajas de este enfoque son: la sencillez con que esta red de neuronas interconectadas puede entrenarse a partir de un conjunto de datos de referencia de forma automática, proceso denominado aprendizaje, y la posibilidad de procesado en paralelo, por lo que es necesario equipos de gran potencia y rapidez. Su principal inconveniente es la dificultad de controlar si este entrenamiento se está produciendo de forma adecuada o no, al ser un método no supervisado.

2.1.5 Modelos de Mixturas Gausianas (GMM)

Los Modelos de Mixturas Gausianas o *Gaussian Mixture Model* (GMM por sus siglas en inglés) es un técnica muy empleada en sistemas de reconocimiento independientes del texto y de locutores (Saravia, 2011). Estos modelan los distintos vectores de parámetros de una locución dada, realizando una suma ponderada (mezcla) de funciones de densidad de probabilidad gausianas. Pueden entenderse como un sistema en el que se aglutinan las virtudes de aquellos otros basados en técnicas VQ y los denominados clasificadores gausianos uni-modales. También pudieran apreciarse como HMM de un sólo estado. A diferencia de los HMM ergódicos de varios estados, los GMM no precisarán en la fase de entrenamiento segmentar en estados ni entrenar la matriz de probabilidades de transiciones. Además, en la etapa de reconocimiento, no será necesario buscar la secuencia de estados de máxima verosimilitud (algoritmo de *Viterbi*), sino que bastará con acumular las probabilidades que asocia el modelo con cada uno de los vectores de entrada (Padilla, 2012).

El modelo de densidades de mezclas gausianas representa una suma pesada de M (número de mezclas) componentes de densidad descritas de la siguiente manera: $\boldsymbol{p}\left(\frac{x}{\lambda}\right) = \sum_{i=1}^{M} \boldsymbol{p_i b_i}(\boldsymbol{x})$, (Moreno, y otros, 2010).

Donde:

- *x* es un vector D-dimensional y la matriz de rasgos: $\boldsymbol{X} = \left(\boldsymbol{x_1}, \boldsymbol{x_{2,\ldots,}} \boldsymbol{x_T}\right) = \begin{pmatrix} \boldsymbol{c_{1,1}} \boldsymbol{c_{1,2}} & \cdots & \boldsymbol{c_{1,T}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{c_{D,1}} \boldsymbol{c_{D,2}} & \ldots & \boldsymbol{c_{D,T}} \end{pmatrix}$

Es una sucesión de variables aleatorias indexadas por una variable discreta, el tiempo *(t=1,...., t).* Cada una de las variables aleatorias del proceso tiene su propia función de distribución de probabilidad y asumiremos que son independientes.

- $\boldsymbol{b_i}(\boldsymbol{x})$: son las componentes de densidad, con $\boldsymbol{i}$ = 1,2,…, *M*. Cada componente es una función gausiana de la forma: Ver figura 22.

$$\boldsymbol{b_i}(\boldsymbol{x}) = \frac{\mathbf{1}}{(\mathbf{2}\boldsymbol{\pi})^{\boldsymbol{D}/\mathbf{2}} |\sum_{\boldsymbol{i}}|^{\mathbf{1}/\mathbf{2}}} \boldsymbol{exp}\{-\frac{\mathbf{1}}{\mathbf{2}}(\boldsymbol{x} - \mu_{\boldsymbol{i}})' \sum_{\boldsymbol{i}}^{-\mathbf{1}} (\boldsymbol{x} - \mu_{\boldsymbol{i}}),\}$$

Figura. 22 Ecuación de la función gausiana

Donde $\mu_{\boldsymbol{i}}$ y $\sum_{\boldsymbol{i}}$ son el vector de medias y la matriz de covarianzas correspondientes a la i-ésima mezcla y se obtienen de la matriz de rasgos $\boldsymbol{X}$.

- $\boldsymbol{p_i}$ son los pesos de las mezclas $\boldsymbol{i}$= 1,2, …, $\boldsymbol{M}$ y satisfacen la condición $\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{M}} \boldsymbol{p_i} = \mathbf{1}$.

- $\boldsymbol{\lambda} = \{\boldsymbol{p_i}, \mu_{\boldsymbol{i}}, \sum_{\boldsymbol{i}}\}$ con $\boldsymbol{i}$ = 1,…,$\boldsymbol{M}$, representa el modelo de mezclas gausianas donde se encuentran los parámetros que caracterizan la señal de voz.

En la tabla 8 del anexo 2 se muestra un resumen de cada clasificador (Calvo, y otros, 2008).

El DTW llegó a ser la metodología más madura y de consecución de los mejores resultados en reconocimiento de palabras antes de la aparición de los modelos de Markov, a partir de este momento, éstos tomaron el protagonismo, dada su mayor flexibilidad, robustez y su capacidad de abordar problemas más complejos (García, 2008).

2.2 Descripción de la solución propuesta

Constituye una necesidad para el departamento Señales Digitales la recuperación directa de archivos de voz de manera automática, evitando dilataciones y posibles errores en el proceso, logrando una mayor rapidez y efectividad en su recuperación. Esto posibilitaría la reducción en la tardanza de respuesta sobre peticiones de préstamo de materiales audiovisuales, así como una mejor transparencia en el proceso.

La detección de palabras clave en archivos de audio consta de un conjunto de fases dentro de las cuales se realizan diferentes actividades de tratamiento de la señal. Entre los procedimientos propuestos para llevar a cabo la detección de palabas clave se encuentra el procesamiento de la señal de audio que se está analizando mediante su segmentación para evitar la pérdida de tramas de información y facilitar el proceso de análisis de la misma para una posterior extracción de características esenciales. En (Tejada, y otros, 2003), (Roj, y otros, 2005) , (Zade, y otros, 2008), (Luengo, y otros, 2009), (Gurlekian, y otros, 2010) se utiliza la fase de la detección de extremos de la señal a analizar como fase inicial del procesamiento de la señal de voz antes de pasar a la fase de extracción de características. Estas características son empleadas en la fase de entrenamiento quedando todo listo para su clasificación. El funcionamiento de todas estas etapas para la detección de palabras clave están estrechamente relacionadas entre sí, donde un mal resultado de una podría provocar bajos resultados finales. A continuación se muestran las diferentes etapas de la solución propuesta para la detección de palabras clave sobre archivos de audio, ver figura 23.

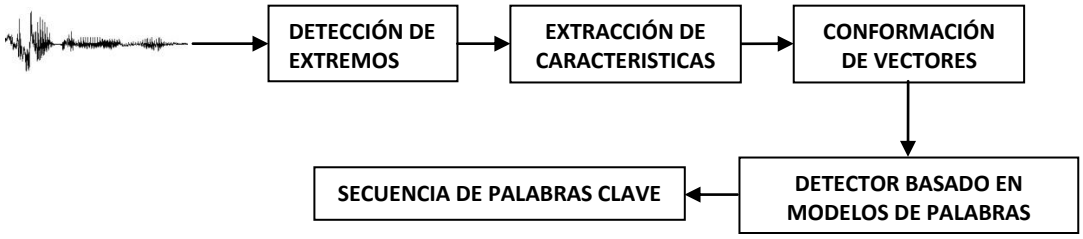
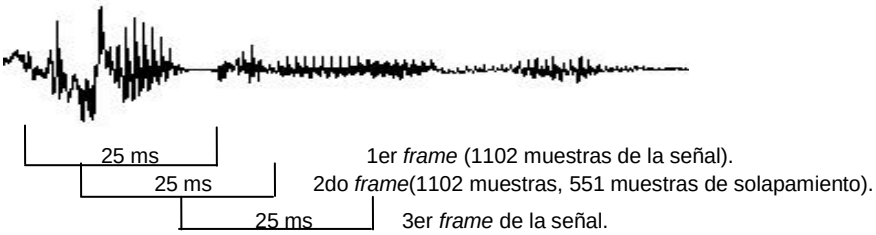


Figura. 23 Flujo de la solución propuesta

El esquema anterior involucra una serie de técnicas estudiadas en el capítulo I. El proceso de segmentación de la señal a través de ventanas se realiza con ventanas de 25 ms (milisegundos) que es un tiempo en el que se considera a la señal de voz prácticamente invariable (García, 2010). Para la

elección de la ventana a utilizar se tendrá en cuenta que la ventana hamming es muy útil para el análisis de en el dominio de la frecuencia.

Además se trabajarán con archivos de audio con formato *wav* (*waveform audio format*) el mismo es un formato de audio sin compresión, este tipo de formato posee una calidad digital excelente y se utilizarán archivos de audio con una frecuencia de 44.1 KHz. Entre más alta la frecuencia de una señal se contará con una mayor cantidad de elementos de la misma, aportando fidelidad a la extracción de parámetros de la señal. O sea la frecuencia indica la cantidad de muestras de la señal en un segundo, por lo tanto entre mayor sea la frecuencia se tendrá una mayor cantidad de muestras de la señal favoreciendo el proceso de extracción de las características acústicas de la voz. El tamaño de las ventanas dependerá de la frecuencia con que se está procesando la señal de voz. Para un tamaño de 25 ms y una frecuencia de 44.1 KHz, el tamaño de la ventana en cuanto a número de muestras será de 1102 muestras para la frecuencia establecida, con un solapamiento equivalente a 551 muestras. Ver figura 24.



1 Ventaneo de la señal a 44.1 KHz de frecuencia en ventanas de 25 ms

Una vez terminado este proceso de ventaneo se realiza la detección de extremos, para eliminar segmentos de silencio. Esto es útil para reducir la cantidad de información a procesar. Como resultado de estos procesos se obtiene la señal de habla a procesar.

Esta señal de habla limpia resultante se le realiza el proceso de extracción de características; utilizando para ello una serie de descriptores previamente analizados, en este caso se utilizarán los coeficientes MFCC ya que aparece en la bibliografía consultada como la característica más significativa de las señales de voz. De este descriptor se tomarán los primeros doce coeficientes así como 12 delta MFCC y 12 delta-delta MFCC para cada *frame*, esta cantidad de coeficientes es muy utilizada en los trabajos consultados (G.Saráchaga, 2006). Como resultado se obtiene un vector de longitud igual al número de *frames* obtenidos en el proceso ventaneo de la señal procesada.

Una vez conformado los vectores de características de la señal que se está procesando pasan a ser sometidos al modelado de los mismos, para la determinación de probabilidades de aparición en los modelos del repositorio, estos modelos estarán conformados por modelos de palabras. Se realizará la búsqueda de las palabras que están dentro del vocabulario a nivel de palabra. Dando como fase final una secuencia de modelos acústicos de palabras que representan la transcripción de la señal procesada.

2.2.1 Detección de extremos de la señal

La detección de extremos de la señal que se está procesando, es una etapa muy importante y necesaria ya que en esta se puede reducir la cantidad de datos a procesar contribuyendo al ahorro de recursos computacionales. Cuando se detecta los extremos de una señal se eliminan fragmentos de silencio quedando solo las muestras que contienen información significativa.

En (Schwarz, 2008) la detección de actividad de voz es muy importante en su aplicación a técnicas de reconocimiento basadas en fonemas. El reconocimiento de fonemas es usado para la transcripción del habla a cadenas de fonemas.

En (Rueda, y otros, 2006) y (Kola, y otros, 2011) se expresa que la detección de extremos o actividad de voz es un paso muy importante en el procesamiento de señales de habla, en específico para los sistemas de reconocimiento automático del habla.

En (Caravaca, 2011) el proceso de detección de extremos, es el encargado de segmentar y etiquetar un audio en fragmentos clasificados como silencio o no, indicando en cuales de estos fragmentos de la señal se ha detectado un segmento sonoro. Esta herramienta es de gran utilidad en el procesado de señales de voz, como puede ser codificación, reconocimiento y transmisión discontinua.

En (Miltre, y otros, 2012) se expresa que en el ámbito del tratamiento de señales de habla identificar el silencio así como el principio y fin de una frase hablada ofrece como ventaja una mayor claridad de la información. Puede ser usado además en el tratamiento y limpieza de audio.

De forma generalizada, podemos decir que el algoritmo de detección de extremos básico puede descomponerse en dos partes (Caravaca, 2011):

1. Cálculo de umbrales y toma de medidas y/o extracción de características.

2. Aplicación de la regla de decisión en función de los umbrales.

Como parte de la investigación se propone la detección de extremos por medio de dos métodos , el primer método es basado en el análisis de dos descriptores caracterizados en el capítulo I, estos son la energía y la Tasa de Cruces por Cero. Y el segundo método es basado en el algoritmo COPER.

2.2.1.1 Método basado en Energía y ZCR

Los descriptores de Energía y ZCR brindan una valiosa información de la señal, los segmentos de señal que sean sonoros tienen una mayor energía que los segmentos sordos, y por el contrario los segmentos sonoros tienen una ZCR menor que los no sonoros.

En (Lu, y otros, 2000) se utiliza la ZCR para la segmentación del audio en sonido o silencio, producto de sus resultados experimentales obtuvieron el valor de 0.04 como umbral para el ZCR. Otros autores como (Rueda, y otros, 2006) y (Tong, y otros, 2006) utilizan la energía ya que los fonemas sordos contienen mayor energía que los segmentos de silencio. En (Reyes, y otros, 2002) se expresa que la detección de extremos basadas en energía y ZCR y reflejan que estos descriptores fallan considerablemente cuando esta se encuentra en un ambiente ruidoso además se plantea que los niveles de umbral se obtienen experimentalmente analizando el contenido de la energía y el cruce por cero que poseen las palabras pronunciadas así como el ruido de fondo; proponen el procedimiento de la figura 42 en el anexo 2.

En (Mohanty, y otros, 2010) propone un método basado en energía y ZCR para la detección de extremos de la señal , donde se calculan sus valores y son comparados con un umbral, los valores superiores son etiquetados como segmentos de sonido. Este trabajo define ZCR como un importante parámetro para la detección de extremos enfatizando que los segmentos sonoros contienen una menor ZCR que los no sonoros. En este caso se obtuvo el resultado del 70% manejando archivos de 1 a 5 segundos de duración con 12 niveles de ruido de fondo.

En (Facundo, y otros, 2008) se procede al cálculo de la energía luego se determina un umbral de decisión (el obtenido en este caso fue de 0.02) que se obtuvo a partir del análisis de las pruebas experimentales, a partir de este umbral se etiquetan como segmentos sonoros los que tengan un valor de energía mayor que dicho umbral. En (Riso, 2009) se utiliza la energía para la eliminación de los segmentos de silencio prolongados en secciones de la señal.

En (Zade, y otros, 2008) utilizan un bloque para la detección de extremos de la señal para luego pasar a la extracción de las características principales de la señal de voz, para esto utiliza los descriptores de energía y ZCR. También se expresa que una incorrecta detección de extremos puede reducir el rendimiento de un reconocedor de voz.

En (Zanuy, 2000) también usa tasa de cruces por cero y sumas de muestras por trama este método fue concebido para su aplicación al reconocimiento de palabras con un segundo por separado, su porcentaje de efectividad es de 96% en condiciones ideales, en ambientes de ruido su efectividad es de 56%.

En (Peeters, 2004) se utiliza la ZCR como descriptor para la clasificación de señales de audio, además se plantea que es la medida del cambio de signo de una señal donde los segmentos de habla tienden a tener un valor más pequeño a diferencia a los segmentos de ruido que tienden a tener un valor alto. También se usa la energía para la detección de extremos de la señal, determinan un umbral del 20% del valor máximo de la energía de la señal a lo largo del tiempo.

En (Bachu, y otros, 2008) utilizan ZCR y la energía, la ZCR es menor para los segmentos de voz que para los segmentos de silencio en una señal de voz, y la energía en mayor para los segmentos de habla y menor para los de silencio. Y proponen un diagrama para la utilización de estos dos descriptores, como se muestra en la figura 42 del anexo 2.

En el siguiente esquema se muestra la utilización de estos descriptores propuesta para la detección de extremos, ver figura 25.

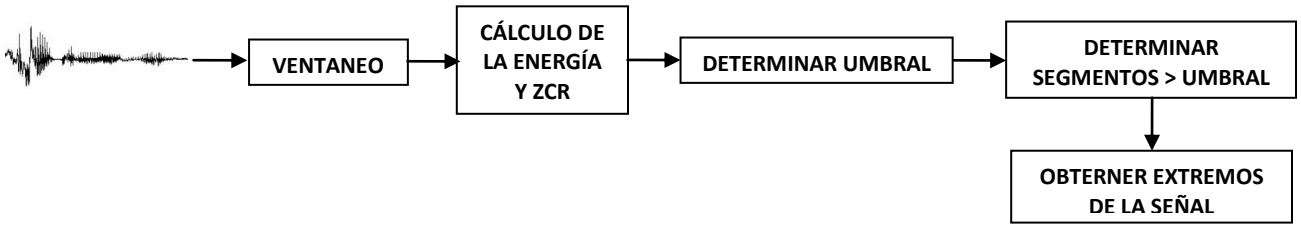


Figura. 24 Detección de extremos de la señal utilizando Volumen y ZCR

Después de realizado el ventaneo de la señal para cada *frame* resultante se calcula su energía y tasa de cruces por cero, a partir de estos vectores se determina un umbral en el que cada *frame* que contenga valores de energía y ZCR superiores a dicho umbral será etiquetado como un segmento sonoro. Para

finalizar el proceso se desechan las partes de la señal que sean detectadas como silencio, así como las partes sonoras que sean tan pequeñas que no representen información significativa a analizar.

2.2.1.2 Método basado en el algoritmo COPER

Uno de los factores que pueden provocar bajos resultados en la detección de extremos basada en ZCR es el ruido de fondo, si este posee una alta frecuencia se dificulta la delimitación pudiendo tomar como segmentos de silencio partes de abundante ruido. El algoritmo COPER propuesto por (Reyes, y otros, 2002) relaciona los cambios de signos de la señal y proporciona información sobre los cambios de energía de la misma. A continuación se muestra la ecuación que rige este algoritmo, ver figura 26.

$$\boldsymbol{COPER} = \sum_{\boldsymbol{m=0}}^{\boldsymbol{N}} |\boldsymbol{x}[\boldsymbol{m}]|\boldsymbol{x}[\boldsymbol{m}]| - \boldsymbol{x}[\boldsymbol{m}-\mathbf{1}]|\boldsymbol{x}[\boldsymbol{m}-\mathbf{1}]||$$

Figura. 25 Fórmula del algoritmo COPER

Donde $\boldsymbol{x}$ es la señal y $\boldsymbol{N}$ es el tamaño de la ventana. Este algoritmo sigue en detalle los cambios temporales que se llevan a cabo durante una pronunciación y proporciona suficiente información para ser utilizado en la detección de extremos. En (Reyes, y otros, 2002) se ha determinado que el algoritmo COPER proporciona una respuesta aceptable a señales con ruido de 15 dB[6] proporcionando mejoras notables con respecto a los algoritmo que utilizan energía y ZCR. A continuación se muestra el esquema de su utilización, ver figura 27.

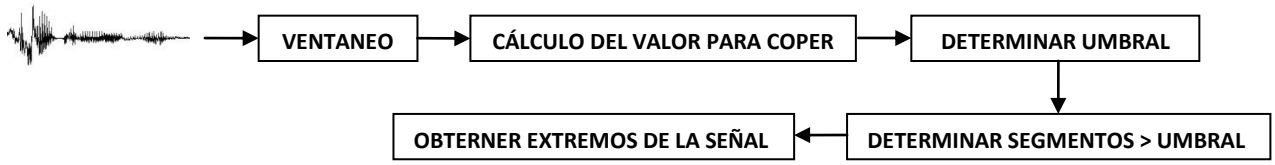


Figura. 26 Detección de extremos mediante algoritmo COPER

2.2.2 Clasificación basada en HMM

[6] *Un decibelio es una unidad que sirve para la comparación de niveles de potencia o de tensión en acústica. Resultan adecuados para valorar la percepción de los sonidos por un oyente.*

Los modelos HMM son modelos generativos que modelan un proceso como una secuencia de estados que se conectan por transiciones, cada estado tiene asociado una salida de una observación con su correspondiente distribución de probabilidades. Cada transición está asociada a una probabilidad que la caracteriza. Cada estado depende sólo del estado previo (Bermejo, 2008). Recibe el nombre de estados ocultos ya que no pueden ser tratados de forma directa sino a través de sus observaciones, en el caso del procesamiento del habla estas observaciones corresponden a los vectores confeccionados por la extracción de descriptores de la señal. La clasificación basada en patrones mediante modelos HMM cuenta con tres etapas principales, estas son (García, 2008), (Rojo, 2011), (Saravia, 2011): inicialización, entrenamiento y clasificación. En la etapa de inicialización se establecen parámetros como la cantidad de estados ocultos y observables del modelo $\boldsymbol{N}$, la matriz de probabilidades de transición $\boldsymbol{A}$ y la matriz de probabilidades de distribuciones $\boldsymbol{B}$. Inicialmente estos parámetros son generados aleatoriamente y forman el modelo $\boldsymbol{\lambda} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$.

En aplicaciones para el tratamiento de señales de voz se utilizan mayormente HMM con topología de izquierda a derecha o sea la secuencia de transición entre estados están ordenados de izquierda a derecha. Ya que la señal de voz es continua siempre hacia adelante se escogen esta tipología de modelos HMM. Ejemplo de un modelo tipo izquierda-derecha se muestra en la figura 28.

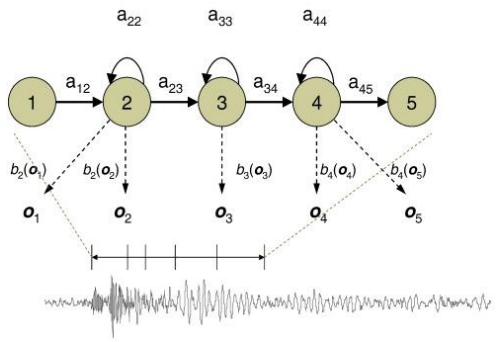


Figura. 27 Ejemplo de HMM tipología de izquierda-derecha

2.2.2.1 Entrenamiento de un HMM

Se plantea cómo deben ajustarse los parámetros del modelo inicial $\boldsymbol{\lambda} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$ para tratar de maximizar la probabilidad de la secuencia de observación $\boldsymbol{O}$ dado el modelo $\boldsymbol{P(O/\lambda)}$. El mayor inconveniente que

presenta esto es que no existe un método analítico conocido que maximice el conjunto de parámetros a partir de los datos de entrenamiento (Moreno, y otros, 2010).

Para lograr esto utiliza algoritmos para realizar el cálculo o estimación de los parámetros que definen el modelo. Se trata básicamente de un proceso iterativo que maximiza en forma local la probabilidad de que una secuencia de observación haya sido generada por un modelo particular $P(O/\lambda)$ y que garantiza en cierta forma la convergencia del proceso a partir de un modelo inicial aleatorio (Martínez, y otros, 2008).

Uno de los métodos más conocidos para la realización de esta tarea es la técnica de Maximización de la Estimación (EM por sus siglas en inglés) (Bilmes, 2000), (Londoño, y otros, 2007), (Rojo, 2011). El algoritmo EM es un algoritmo iterativo que la idea básica es estimar, a partir de un modelo inicial λ, a un nuevo modelo $\bar{\lambda}$ que cumpla con $p(X|\bar{\lambda}) \geq p(X|\lambda)$. En la próxima iteración el nuevo modelo será el inicial y el proceso se repetirá hasta que converja, $p(X|\bar{\lambda}) \geq p(X|\lambda) \leq e$, con $e > 0$ fijo o cuando sobrepase la cantidad de iteraciones deseadas (Calvo, y otros, 2008). Una especialización del algoritmo de Máxima Estimación es el algoritmo *Baum-Welch* el cual se explica en (Moreno, y otros, 2010), (Welch, 2003), (Bermejo, 2008), (Rojo, 2011). Entonces dada una secuencia muestral se busca mediante el método de maximización de la estimación obtener el modelo HMM que tenga más probabilidades de generar la secuencia indicada.

2.2.2.2 Reconocimiento mediante HMM

La detección de palabras claves mediante modelos HMM sigue el flujo que muestra el siguiente diagrama de la figura 29.

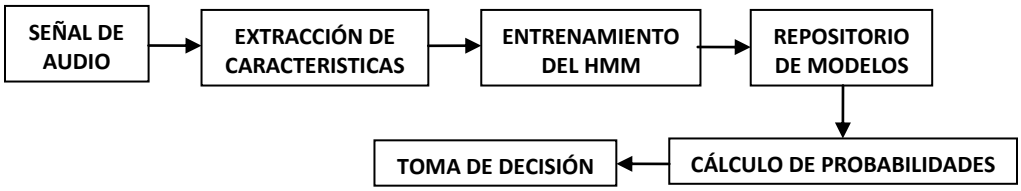


Figura. 28 Reconocimiento mediante HMM

En el proceso de entrenamiento para cada palabra clave se obtiene un modelo óptimo que las caracteriza, con estos modelos se construye un repositorio de modelos que servirán como referencia para cuando se introduzca una nueva palabra a detectar evaluarla contra los modelos del repositorio para determinar el

que más probablemente haya podido generarla. La toma de decisión se realiza comparando las probabilidades obtenidas obteniendo la máxima. De esta forma se asocia el modelo a la palabra a reconocer y se obtiene del repositorio la palabra asociada al modelo de mayor probabilidad.

En esta fase se presenta el problema de encontrar la secuencia de estados óptima para un conjunto de observaciones. Para resolver esto se utiliza el algoritmo *Viterbi* que básicamente trata de maximizar la probabilidad $P(O\ |q,\ \lambda)$. Para encontrar la mejor secuencia de estados $\boldsymbol{Q} = \{\boldsymbol{q_1}, \boldsymbol{q_2}, \dots, \boldsymbol{q_T}\}$ para una secuencia de observación dada $\boldsymbol{O} = \{\boldsymbol{o_1}, \boldsymbol{o_2}, \dots, \boldsymbol{o_T}\}$ se define la variable (ver figura 30):

$$\boldsymbol{\delta_i(i)} = \max_{q_1, q_2, \dots, q_t - 1} \boldsymbol{P}[\boldsymbol{q_1}, \boldsymbol{q_2}, \dots, \boldsymbol{q_t} = \boldsymbol{i}, \boldsymbol{o_1}, \boldsymbol{o_2}, \dots, \boldsymbol{o_t} | \boldsymbol{\lambda}]$$

Figura. 29 Maximización de la probabilidad de estados

que representa la secuencia de estados con mayor probabilidad en el instante $\boldsymbol{t}$ que acaba en el estado $\boldsymbol{o_t}$ generando las $\boldsymbol{t}$ primeras observaciones. El funcionamiento del algoritmo *Viterbi* se explica en (Bermejo, 2008), (Moreno, y otros, 2010), (Rojo, 2011).

2.3 Esquema de utilización de la técnica propuesta

Actualmente los procesos de recuperación de información en materiales audiovisuales dentro de los productos del departamento de Señales Digitales, son el resultado de procedimientos de búsqueda sobre metadatos que son descritos manualmente. Estos productos del departamento procesan grandes flujos de datos y son responsables de tareas como: la publicación, captura, vigilancia y procesamiento en tiempo real además de la recuperación de materiales audiovisuales. El Sistema Gestor de Procesos de Media (SGPM) posibilita la integración de sistemas mediante una plataforma de indexación y codificación de funcionalidades para la realización de procesos sobre materiales audiovisuales. Esta aplicación está concebida para que se integre con otros sistemas de forma automática a partir de la configuración de servicios, basados en estándares y actualización de los estados de procesos.

La técnica propuesta para la detección de palabras clave puede ser utilizada mediante su integración al SGPM, el mismo se encargaría de realizar el procesamiento de la señal de voz. Una vez realizada la integración al SGPM su utilización se refleja en el esquema siguiente, ver figura 31.

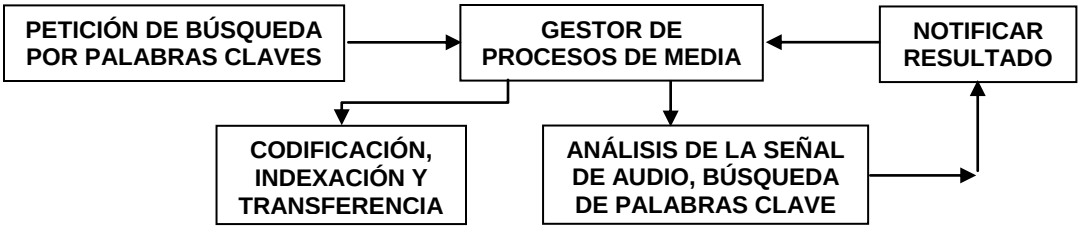


Figura. 30 Esquema de utilización de la técnica propuesta

El SGPM mediante estándares de comunicación recibiría peticiones de búsquedas sobre archivos de audio por concepto de palabras claves, cada petición trae asociado el objeto de búsqueda, el SGPM determina que componente es el que está asociado a esta petición. La petición es procesada en el modulo para el análisis de la señal de audio, el cual es el encargado de realizar la búsqueda por concepto de palabras clave. Como resultado del procesamiento de la petición original el componente de procesamiento encargado de la realización de búsqueda de palabras claves sobre audios notifica al gestor un resultado. De ahí el SGPM brinda una acción de respuesta a la petición mediante un componente encargado de esta tarea.

2.4 Conclusiones parciales

En este capítulo se realizó una caracterización de los principales modelos para la clasificación de vectores de características. Se seleccionaron los Modelos Ocultos de Markov (HMM) como método para el modelado de las características espectrales del audio ya que presenta flexibilidad en el modelado de las rápidas variaciones del habla. Se determinó que se emplearán modelos acústicos a nivel de palabra. Se utilizarán los coeficientes MFCC como descriptor de las características del habla, el cual es uno de los descriptores más utilizados en el procesamiento de señales de voz. Se utilizará el algoritmo COPER para la detección de extremos de la señal el cual consume menos recursos computacionales y el más robusto ante el ruido que los sistemas basados en energía y ZCR.

CAPÍTULO 3: ANÁLISIS DE LOS RESULTADOS OBTENIDOS

3.1 Herramientas de ensayo

Para realizar las pruebas prácticas del procesamiento del audio implementación del proceso de detección de palabras clave se utilizó el software MATLAB sobre Sistema Operativo GNU/Linux Ubuntu 12.04. El hardware utilizado fue una computadora de escritorio con procesador Intel Core Duo a 2.20 GHz y 1 GB de memoria RAM.

Matrix Laboratory (MATLAB por sus siglas en inglés). Es un software para realizar cálculos numéricos con vectores y matrices. Como caso particular se puede también trabajar con números escalares, tanto reales como complejos, con cadenas de caracteres y con otras estructuras de información más complejas. Una de las capacidades más atractivas es la de realizar una amplia variedad de gráficos en dos y tres dimensiones. MATLAB tiene también un lenguaje de programación propio.

Dentro de la instalación del MATLAB se cuenta con una serie de herramientas adicionales (toolbox) para el trabajo en diferentes campos y aplicaciones matemáticas con una gran variedad de funciones implementadas que optimizan mucho el trabajo para el cual se esté usando el software. La mayoría de las funciones están escritas en el lenguaje abierto MATLAB, brinda la posibilidad de inspeccionar los algoritmos, modificar el código fuente y crear sus propias funciones personalizadas.

Por otra parte se utilizó el toolbox *VOICEBOX* creado por *Mike Brookes* bajo licencia de código abierto GPL. Es un toolbox para el procesamiento de señales de habla. De este se utilizó la función melcepst para el cálculo de los coeficientes MFCC e incluye ΔMFCC (delta MFCC) y $\Delta\Delta$MFCC (delta delta MFCC) que corresponden a la derivada y a la derivada de la derivada de los coeficientes MFCC extraídos.

3.2 Medidas de eficiencia

Las medidas de eficiencia son una herramienta estadística que posibilitan resumir la información que se muestrea para tener conocimiento generales de la misma. Para esto es común el uso de términos como: falsos positivos y falsos negativos los cuales son usados para la descripción de posibles errores en los procesos de decisión estadísticos.

Cuando se realiza un proceso de clasificación o detección de palabras sobre audios se produce un falso positivo una vez que se detecta o reconoce alguna palabra cuando no se encuentra en el segmento analizado y un falso negativo cuando no realiza la detección correcta de la palabra existiendo la misma en el segmento de audio que se analiza. Ambas clasificaciones permiten calcular la eficacia del sistema.

Análogamente se definen los términos verdadero positivo y verdadero negativo para describir las decisiones correctas de cualquier proceso. Un verdadero positivo se produce cuando se detecta una palabra clave correctamente. Por otra parte, cuando no se detecta ninguna palabra clave sobre el segmento que se analiza y se confirma que no lo hay se está en presencia de un verdadero negativo. Ver tabla 1.

CONDICIÓN DE DETECCIÓN DE PALABRAS CLAVES.		**DETECCIÓN CORRECTA.**	**DETECCIÓN INCORRECTA.**
RESULTADO OBTENIDO	**POSITIVO**	**Palabra clave + Positivo = Verdadero Positivo**	**No Palabra clave + Positivo = Falso Positivo**
	NEGATIVO	**Palabra clave + Negativo = Falso Negativo**	**No Palabra clave + Negativo = Verdadero Negativo**

Tabla 1 Tabla ilustrativa de los términos para describir los resultados de clasificación

Para la medición de la eficiencia se tomarán cada uno de los segmentos de palabras analizados. Siendo así las mediciones no se realizan como tal al proceso de clasificación, el cual se basa precisamente en los segmentos de las palabras a reconocer.

La eficiencia máxima se consigue cuando el sistema es capaz de procesar un conjunto amplio de audios sin producir ningún falso positivo o falso negativo. Esto se traduce en que las palabras clave detectadas son correctas en un 100% y no debe quedar ninguna palabra sin detectar. Existen diversas técnicas para medir la eficiencia de estos algoritmos, aunque una de ellas destaca sobre las demás por su amplia utilización. Ejemplo de estas técnicas son las medidas *recall* y *precision* definidas (ver figura 32 y figura 33 respectivamente) como:

$$\boldsymbol{Recall = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Negativos}}$$

Figura. 31 Obtención de la medida *Recall*

$$\boldsymbol{Precision = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Positivos}}$$

Figura. 32 Obtención de la medida *Precision*

Donde *recall* expresa la razón de palabras correctamente detectados entre todos las palabras existentes, mientras que *precision* ofrece una medida de la fiabilidad de la detección, que significa cuántos de las palabras detectadas son realmente correctas y no falsos positivos. Estas medidas normalmente se trabajan en el dominio [0,1], sin embargo se pueden utilizar en forma porcentual. Otras medidas posibles son las utilizadas por (Olson, 2008): *True Negative Rate* (TNR por sus siglas en inglés) en la figura 34 y *Accuracy* en la figura 35.

$$\boldsymbol{TNR = \frac{Verdaderos\ Negativos}{Verdaderos\ Negativos + Falsos\ Positivos}}$$

Figura. 33 Obtención de TNR

$$\boldsymbol{Acurracy = \frac{Verdaderos\ Positivos + Verdaderos\ Negativos}{Total\ de\ Segmentos}}$$

Figura. 34 Obtención de Acurracy

La primera refleja una medida de la cantidad de veces que el sistema detecta una palabra clave de manera efectiva ya sea detectando una palabra clave correctamente o incorrectamente por el sistema; mientras que la segunda métrica expresa un porcentaje de todos los segmentos donde se detectó una palabra correctamente o no.

3.3 Pruebas y experimentos

En los experimentos realizados para observar el comportamiento de las técnicas para el procesamiento de la señal de voz propuestos se utilizaron un conjunto de grabaciones de voz, las mismas se encuentran en formato WAV el cual es un formato de audio digital normalmente sin compresión de datos, en este caso

tendrán codificación PCM (sin compresión). Al no tener pérdida de calidad por el proceso de compresión puede ser utilizado en sistemas profesionales. Los audios utilizados presentan una frecuencia de 44.1 KHz con 16 bits por muestra. Primeramente se analizaran los descriptores y algoritmos para la detección de extremos. Además se implementó en MATLAB esta etapa para la detección de extremos que pertenecen al flujo de procesos para el procesamiento de la señal de habla para la detección de palabras clave sobre archivos de audio. Para cada una de estas etapas se confeccionaron varios scripts.

3.3.1 Experimentos en la detección de extremos de la señal de voz

Para realizar la detección de extremos de las señales se implementaron las funciones ***Coper*.** Se utilizó además el script ***epdByEZcr*** creado por Roger Jang que es de código libre y presenta dependencias de los scripts: ***epdParamSet*, *frame2energy***, ***frame2zcr, frame2sampleIndex*** y sus descripciones se observan en la tabla 2. La puesta en práctica de estos scripts se muestra en la figura 43 anexo 2, donde se realiza la segmentación de la palabra “energía”.

Nombre	Descripción
Coper	Se determinan los extremos de la señal vía algoritmo COPER.
epdByEZcr	Determina extremos usando el energía y la tasa de cruces por cero.
epdParamSet	Inicializa los parámetros para el procesamiento
frame2energy	Calcula el valor dela energía partir de una matriz que corresponde a los frames del audio.
frame2zcr	Calcula el valor del cruce por cero a partir de una matriz que corresponden los frames del audio.
frame2sampleIndex	Convierte de *frame* a índice de muestras para poder representar en el tiempo (en muestras).

Tabla 2 Funciones utilizadas para la detección de extremos de la señal

Para las pruebas se utilizaron 955 segmentos de audio entre los que se encuentran palabras y segmentos de oraciones de palabras. Estos 955 segmentos de audio están divididos en tres categorías de acuerdo a su velocidad de pronunciación, velocidad lenta, esta es una pronunciación pausada entre palabras pronunciadas, velocidad de pronunciación normal siempre articulando bien las palabras y artículos del

lenguaje y velocidad rápida. Un total de 350 segmentos corresponden a una pronunciación con buen pausado entre palabras, 350 de pronunciación normal y 255 segmentos de habla rápida.

En la figura 36 se muestra los resultados de estos métodos propuestos para la detección de extremos en dependencia de la velocidad de pronunciación del habla.

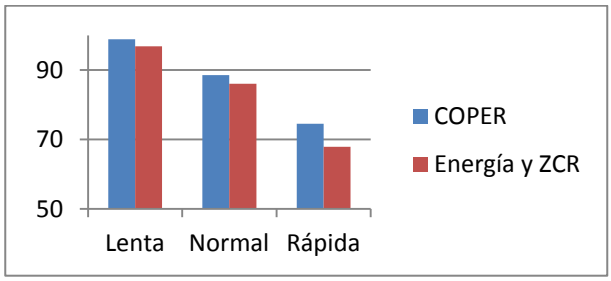


Figura. 35 Efectividad de la detección de extremos atendiendo a la velocidad de pronunciación

Como se puede observar los mejores resultados se obtienen cuando la velocidad de pronunciación es bien pausada entre palabras, llegando al 98,86 % de detección correcta de extremos con el uso del algoritmo COPER. Independientemente de la velocidad de pronunciación el algoritmo COPER en los experimentos realizados aporta mejores tasas de detección correcta que el método basado en Energía y ZCR. A medida que va aumentando la velocidad de pronunciación los índices de acierto descienden considerablemente siendo el algoritmo COPER el de mejor acierto pero con un resultado muy bajo para un rendimiento de un 74.5 % de detección correcta de los segmentos procesados.

A modo de comparación se puede decir que el algoritmo COPER es más rápido basado en el consumo computacional que los mecanismos basados en Energía y ZCR, para experimentar sobre esto se desarrolló la función en MATLAB llamada ***Tiempo_Ejecucion.m*** ver anexo 3, la cual se encarga de medir en tiempo de demora de cálculo de los métodos de detección de extremos analizados. Tras varias pruebas realizadas se comprueba el menor uso de recursos computacionales por parte del algoritmo COPER.

3.3.2 Experimentos de detección de palabras clave sobre audio

Para las pruebas de esta fase se implementaron scripts que garantizaron la utilización de los HMM. Se trabajaran con modelos HMM de cuatro estados. Los principales scripts son: ***inithmm***, ***trainhmm***, así

como ***Viterbi***, ver tabla 3. En estas funciones recae todo el peso del modelado de los vectores de datos acústicos.

Nombre	Descripción
inithmm	Esta función inicializa los parámetros definidos para un modelo HMM: la matriz de transiciones para un modelo tipología izquierda-derecha, la matriz de probabilidad inicial así como realiza la clusterización de los datos de entrada usando el algoritmo kmeans.
trainhmm	Usando el algoritmo de Máxima Estimación y como una especialización el algoritmo Baum-Welch trata iterativamente (hasta la convergencia) de obtener un modelo (previamente inicializado) que represente como máxima probabilidad el conjunto de datos de entrenamiento.
viterbi	Decodifica el modelo HMM entrenado para mediante un conjunto de datos de prueba obtener un modelo que sea el que más probabilidad tenga de haber generado dichos datos.

Tabla 3 Funciones implementadas para el uso de modelos HMM

El vector conformado presenta una estructura ***[NxC]*** donde ***N*** es la cantidad de *frame* y ***C*** los coeficientes MFCC.

Para los datos de las palabras clave a entrenar para la conformación del vocabulario se realizó el entrenamiento multi-observación. Se escogieron 9 palabras clave, para cada palabra clave se escogió 55 diferentes tipos de pronunciaciones, para la conformación de un modelo único por cada palabra. El entrenamiento multi-observación aporta flexibilidad en el reconocimiento ya que se toman en cuenta variaciones simples de la pronunciación de una palabra que puede afectar el proceso. Finalmente se obtiene el vocabulario cuando se inicializan y entrenan un modelo HMM para cada palabra clave.

La base de datos de prueba consiste en un total de 75 audios, estos fueron realizados por 2 locutores. La información almacenada en los audios de prueba corresponden a una serie de oraciones, donde se mezclan las palabras previamente entrenadas para su detección. Estos audios de prueba fueron

confeccionados en condiciones de poco ruido y realizando una buena pronunciación así como pausado de las oraciones habladas.

Se utilizarán medidas de eficiencia para observar el comportamiento de las pruebas realizadas. Los audios a procesar son analizados a nivel de palabra cada una de ellas es analizada por términos como: Falsos Positivos, Falsos Negativos , Verdaderos Positivos y Verdaderos Negativos para el cálculo de las medidas de eficiencia.

3.3.2.1 Experimentos con diferentes cantidades de coeficientes MFCC

El número de coeficientes MFCC con que se caracteriza el espectro de una señal, determina con qué precisión se aproxima la densidad espectral de potencia. Se utilizó un número de coeficientes MFCC variable dentro del rango de 10 a 16 valores, realizando pruebas de entrenamiento y clasificación con un incremento de 2 coeficientes por cada prueba realizada. Además se obtendrá el numero definitivo de coeficientes con mejores resultados a utilizar en las etapas entrenamiento y clasificación para la detección de palabras clave.

3.3.2.1.1 Experimentos con diferentes cantidades de coeficientes MFCC sin ∆MFCC-∆∆MFCC

Para la realización de este experimento se escogieron diferentes de cantidades de coeficientes MFCC, se trabajarán con 10, 12, 14 y 16 coeficientes MFCC, para esta prueba solo se utilizarán los primeros coeficientes (en dependencia de la cantidad a trabajar) sin las variaciones delta-delta.

A continuación en la tabla 4 se muestra el resultado obtenido utilizando diferentes cantidades de coeficientes MFCC.

	Resultados				Medidas de eficiencia			
No de Coeficientes MFCC	Verdaderos Positivos	Falsos Positivos	Verdaderos Negativos	Falsos Negativos	Recall	Precision	TNR	Accuracy
10	63	86	711	42	0,6	0,4228	0,8921	0,8581
12	79	67	730	26	0,7524	0,5411	0,9159	0,8969
14	71	76	721	34	0,6762	0,4830	0,9046	0,8780
16	64	96	701	41	0,6095	0,4	0,8795	0,8481

Tabla 4 Resultados usando 10, 12, 14 y 16 coeficientes MFCC sin ∆MFCC-∆∆MFCC

3.3.2.1.2 Experimentos con diferentes cantidades de coeficientes MFCC con ∆MFCC

Para la realización de este experimento se trabajará con 10, 12,14 y 16 coeficientes, y se utilizarán los coeficientes ∆MFCC que representan la variación de los coeficientes MFCC con respecto a un instante de tiempo, son denominados además como coeficientes de velocidad o primera derivada.

A continuación en la tabla 5 se muestra el resultado obtenido utilizando diferentes cantidades de coeficientes MFCC con variaciones ∆MFCC.

	Resultados				Medidas de eficiencia			
No de Coeficientes MFCC	Verdaderos Positivos	Falsos Positivos	Verdaderos Negativos	Falsos Negativos	Recall	Precision	TNR	Accuracy
10	76	36	761	29	0,7238	0,6786	0,9548	0,9279
12	96	27	770	9	0,9143	0,7805	0,9661	0,9601
14	84	29	768	21	0,8	0,7434	0,9636	0,9446
16	71	41	756	34	0,6762	0,6339	0,9486	0,9169

Tabla 5 Resultados usando 10, 12, 14 y 16 coeficientes MFCC solo con ∆MFCC

3.3.2.1.3 Experimentos con diferentes cantidades de coeficientes MFCC utilizando ∆MFCC-∆∆MFCC

Se utilizarán para las pruebas los primeros coeficientes MFCC así como se incluye ∆MFCC (delta MFCC) y ∆∆MFCC (delta-delta MFCC) que corresponden a la derivada y a la derivada de la derivada de los coeficientes MFCC extraídos.

A continuación en la tabla 6 se muestra el resultado obtenido utilizando diferentes cantidades de coeficientes MFCC incluyendo ∆MFCC y ∆∆MFCC.

	Resultados				Medidas de eficiencia			
No de Coeficientes MFCC	Verdaderos Positivos	Falsos Positivos	Verdaderos Negativos	Falsos Negativos	Recall	Precision	TNR	Accuracy
10	81	25	772	24	0,7714	0,7642	0,9686	0,9457
12	99	16	771	6	0,9429	0,8609	0,9797	0,9753
14	89	26	771	16	0,8476	0,7739	0,9674	0,9534
16	83	32	765	22	0,7905	0,7217	0,9598	0,9401

Tabla 6 Resultados usando 10, 12, 14 y 16 coeficientes MFCC con ∆MFCC-∆∆MFCC

3.3.2.1.4 Análisis de los experimentos con diferentes cantidades de coeficientes MFCC

Analizando los resultados de las tablas 4,5 y 6 se observa que los mejores resultados no solo son logrados utilizando las variaciones ΔMFCC-ΔΔMFCC (ver figura 37) sino que además son logrados utilizando una cantidad de 12 descriptores de los coeficientes MFCC (ver figura 38). Las diferencias entre las puntuaciones de efectividad del uso de 12 coeficientes con el uso de 14 coeficientes (ambos utilizando las variaciones ΔMFCC-ΔΔMFCC) es solo de 2,19 % en la detección de palabras clave. También se puede apreciar que los resultados van mejorando al aplicar el uso de las variaciones ΔMFCC-ΔΔMFCC, estas variaciones ajustan la forma en que se extraen los primeros coeficientes MFCC obteniendo una información mas precisa de las muestras.

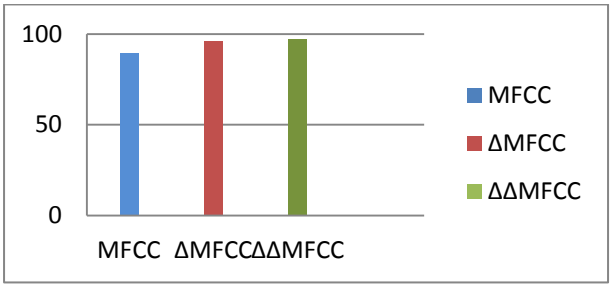


Figura. 36 Efectividad en la clasificación usando las variaciones de MFCC, ΔMFCC y ΔΔMFCC

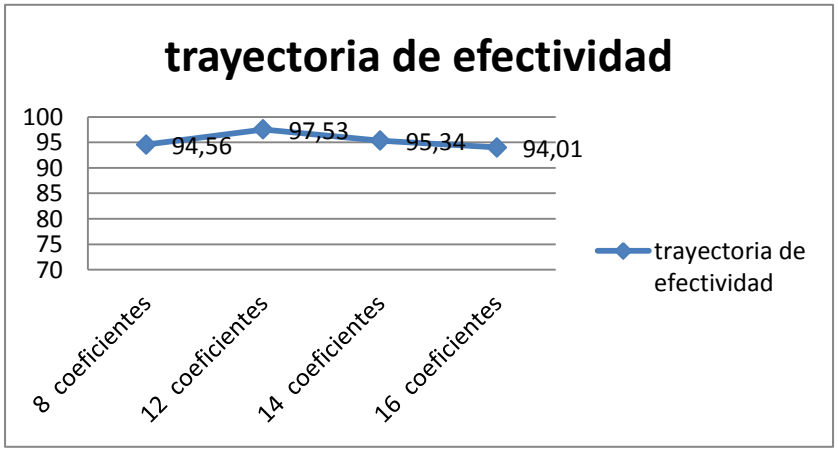


Figura. 37 Trayectoria de la clasificación usando diferentes cantidades de coeficientes MFCC

3.3.2.2 Experimentos con coeficientes LPC, LPCC y MFCC

En este experimento se utilizaron los coeficientes LPC, LPCC y MFCC para la confección de los datos espectrales que son utilizados en la etapa de entrenamiento del sistema de detección de palabras clave basado en palabras.

En la tabla 7 se muestra el resultado arrojado, al utilizar cada uno de estos descriptores en el proceso de detección de palabras clave.

	Resultados				Medidas de eficiencia			
TIPO DE DESCRIPTOR	**Verdaderos Positivos**	**Falsos Positivos**	**Verdaderos Negativos**	**Falsos Negativos**	**Recall**	**Precision**	**TNR**	**Accuracy**
LPC	**72**	**38**	**759**	**33**	**0,6857**	**0,6545**	**0,9523**	**0,9213**
LPCC	**80**	**29**	**768**	**25**	**0,7619**	**0,7339**	**0,9636**	**0,9401**
MFCC	**99**	**16**	**771**	**6**	**0,9429**	**0,8609**	**0,9797**	**0,9753**

Tabla 7 Resultados de clasificación usando los descriptores de LPC, LPCC y MFCC.

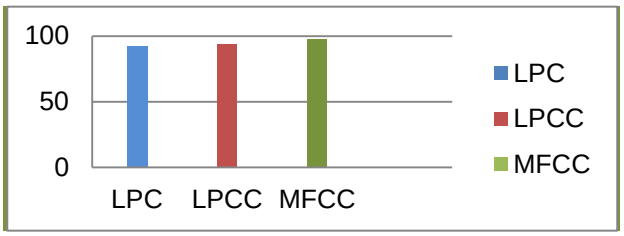


Figura. 38 Efectividad en la clasificación usando los descriptores LPC, LPCC y MFCC

Observando los resultados de la tabla 7 y la representación de la figura 39 se aprecia que los mejores resultados de la clasificación de los vectores espectrales son usando los coeficientes MFCC, como además, se había planteado en la sección (1.3.5.7) producto de las bibliografías consultadas.

3.4 Conclusiones parciales

Con la realización de los experimentos se logró comprobar el funcionamiento del flujo de la solución propuesta para la detección de palabras claves sobre archivos de audio. Se logró implementar un conjunto de funciones y scripts en MATLAB. Se emplearon medidas de eficiencia para medir el rendimiento de las técnicas propuestas un grupo de medidas de eficiencia como: *recall*, *precision*, *TNR* y *accuracy,* los resultados de estas medidas validan el funcionamiento de los procedimientos planteados. Los mejores índices de efectividad se obtuvieron usando los coeficientes MFCC, en específico una cantidad de 12,

incluyendo además las variaciones ΔMFCC-ΔΔMFCC. Además el algoritmo COPER es más efectivo y consume menos recursos computacionales para la detección de extremos que los métodos basados en energía y ZCR.

CONCLUSIONES GENERALES

Después de haber concluido se puede decir que el autor pudo poner en práctica conocimientos adquiridos en el proceso de investigación así como nuevos conocimientos adquiridos durante el proceso de implementación. El desarrollo de la presente investigación aportó al autor conocimientos relacionados a los sistemas de reconocimiento automático del habla, métodos para la extracción de características del audio, modelos probabilísticos para la modelación de las características acústicas del habla así como sistemas de detección de palabras clave y procesamiento de señales de habla. El estudio de una serie de técnicas, descriptores de audio y las técnicas propuestas para la detección de palabras claves constituyen bases teóricas para el desarrollo de la técnica propuesta. Se definió un esquema de utilización de la técnica propuesta para su utilización en los productos del departamento de Señales Digitales. De esta manera y por todo lo anterior se puede concluir que el objetivo general de la presente investigación fue cumplido, lográndose definir un flujo para la detección de palabras clave sobre audio que podrá ser aplicado en los productos del departamento de Señales Digitales.

RECOMENDACIONES

Una vez cumplidos los objetivos del presente trabajo y en correspondencia con los resultados obtenidos se recomienda como trabajos futuros:

- Continuar el estudio de técnicas para la reducción de ruido en audio.
- Aplicar cada una de las etapas del flujo propuesto en los productos del departamento de Señales Digitales de manera que se pongan en práctica las aportaciones teóricas realizadas.

- Realizar el estudio de otras técnicas de detección de palabras clave basadas en el análisis acústico a nivel de fonemas, bifonemas así como la construcción de un modelo del lenguaje con su gramática.

- Ampliar el vocabulario de palabras claves con variedades de tipos de voces, de personas de diferentes edades, géneros y lugares de residencia.

REFERENCIAS BIBLIOGRÁFICAS

Oxford University Press. 2013. oxforddictionaries. *oxforddictionaries.* [Online] 2013. [Cited: 05 21, 2013.] http://oxforddictionaries.com/es/definicion/ingles_americano/keyword .

Acevedo, Eric Simancas. 2008. *Reconocimiento de hablantes basado en Modelo de Mezclas Gaussianas.* 2008.

Adrian, Jon Ander Gómez, Lance, Marcos Calvo and Arnal, l Emilio Sanchis. 2010. *Localización de Palabras basada en Grafos de Fonemas.* 2010.

Al-Shoshan, Abdullah I. 2006. *Speech and Music Classification and Separation: A Review.* Qassim University, Saudi Arabia : Department of Computer Science, College of Computer,, 2006.

Alejandro, De Luna Ortega Carlos, César, Martínez Romo Julio and Miguel, Mora González. 2006. *Reconocimiento de Voz con Redes Neuronales, DTW y Modelos Ocultos de Markov.* 2006.

Alfonso, Ivis Rodés and Lara, José Ramón Calvo de. 2009. *Reconocimiento del locutor dependiente del texto con modelos acústicos del habla.* 2009.

Amir, A., Efrat, A. and Srinivassan, S. 2001. *Advances in phonetic word spotting.* s.l. : CIKM, 580-582,, 2001.

Babylon. 2012. Babylon LTD 9. *Babylon LTD 9.* [Online] 2012. [Cited: 05 21, 2013.] http://diccionario.babylon.com/keyword/ 2012.

Bachu, R G, et al. 2008. *Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal.* s.l. : Electrical Engineering Department, 2008.

Becerra, Adriana, et al. 2009. *Reconocimiento de voz mediante Modelos Ocultos de Markov.* 2009.

Bermejo, Sergio Lucas. 2008. *Transcripción Automática de voz espontánea para recuperación de información.* Madrid : Universidad Autónoma de Madrid, 2008.

Bhattacharje, Utpal. 2013. *A Comparative Study Of LPCC And MFCC Features For The Recognition Of Assamese Phonemes.* Arunachal Pradesh, India, : Department of Computer Science and Engineering, Rajiv Gandhi University, Rono Hills, Doimukh,, 2013.

Bilmes, Jeff A. 2000. *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models.* s.l. : International Computer Science Institute, 2000.

Bleda, J Frances and S Marini, J J Martínez. 2010. *Herramientas software para la docencia de la señal de voz en Ingeniería.* Alicante, España : s.n., 2010.

Borrero, H. and Y. Baquero, Z. Alezones. 2011. *Reconocimiento de Palabras Aisladas Utilizando LPC Y DTW, para control de navegación de un mini-robot.* 2011.

Calvo, Jose Ramón, Fernández, Rafael and Hernández, Gabriel. 2008. *Métodos de extracción, selección y clasificación de rasgos acústicos para el reconocimiento del locutor.* s.l. : RNPS No. 2142, 2008. ISSN 2072-6287.

Caravaca, Guillermo González. 2011. *Reducción de ruido en grabaciones de audio.* s.l. : Ingeniería de Telecomunicación, 2011.

Carmona, Jaime Humberto Pech. 2006. *Desarrollo de un sistema de reconocimiento de voz para el control de dispositivos utilizando mixturas gausianas.* Mexico DF : s.n., 2006.

Ceballos, Alexander, et al. 2010. *Sistema audiovisual para reconocimiento de comandos.* 2010.

CENATAV. 2012. CENATAV Advanced Tecnologies Application Center. *CENATAV Advanced Tecnologies Application Center.* [Online] 2012. [Cited: 05 19, 2013.] http://www.cenatav.co.cu/index.php/publications-cenatav/gray-series-publications. ISSN 2072-6287.

Cortizo, Enrique Alexandre and Zurera, Manuel Rosa. 2009. *Clasificación automática de voz/música utilizando discriminantes lineales de fisher.* Alcala : Departamento de la Teoria de la señal y Comunicaciones, 2009.

Cruz, Ariadna Miranda. 2010. *Procesamiento Digital de Audio para la Plataforma de Televisión Informativa PRIMICIA.* La Habana : UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS, 2010.

Cuayahuitl, H y Serridge, B. 2004. *Out-of-vocabulary word modelling and rejection for spanish keyword spotting systems.* s.l. : MICAI, 155-165, 2004.

Definición. 2008. Definición.de. *Definición.de.* [Online] 2008. [Cited: 11 13, 2012.] http://www.Definicion.de.com.

Deng, Li and Strik, Helmer. 2007. *Structure-Based and Template-Based Automatic Speech Recognition -- Comparing parametric and non-parametric approaches.* 2007.

DTEC. 2009. DTEC. *DTEC.* [Online] 12 2009. [Cited: 01 11, 2013.] http://www.dtec-bio.es/productos/reconvox/#mis_tabs.

Ecured. 2011. Ecured Enciclopedia Cubana de la red. *Ecured Enciclopedia Cubana de la red.* [Online] 2011. [Cited: 01 09, 2013.] http://www.ecured.cu/index.php/Dragon_Naturally_Speaking.

Ericson, Kate. 2009. *Recognizing Phonemes in Continuous Speech -CS640 Project.* 2009.

Facundo, Bellesi and Flavio, Ortíz. 2008. *Reconocimiento de voz para aplicación en domótica.* s.l. : Universidad Tecnológica Nacional, 2008.

Farlex Inc. 2010. thefreedictionary. *thefreedictionary.* [Online] 2010. http://es.thefreedictionary.com/t%C3%A9cnica.

Fook, Chong Yen, et al. 2010. *Comparison of speech parameterization techniques for classification of speech dysfluencies.* Malaysia : School of Mechatronic Engineering, University Malaysia Perlis (UniMAP), 2010.

G.Saráchaga, Virginia Sartori, Laura Vignoli. 2006. *Identificación automática de resumen de canciones.* 2006.

García, Msc. Ruber Hernández. 2010. *Detección de anuncios en emisiones de televisión utilizando la señal de audio.* UNIVERSIDAD DE MÁLAGA , ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA, MASTER OFICIAL EN SISTEMAS DE INFORMACIÓN AUDIOVISUAL : s.n., 2010.

García, Verónica Peña. 2008. *Reconocimiento de palabras claves en conversaciones espontáneas en castellano.* Madrid, España : s.n., 2008.

Gila, Pablo Martín. 2012. *Detección de términos orales para recuperación de información multimedia y su aplicación a videos de información turística.* 2012.

Graña, Adrian Torres and Zambrano, Reynier Blanco. 2012. *Clusterización de trazas en la minería de procesos de negocio.* La Habana : s.n., 2012.

GreenFacts. 2008. europa. [Online] Scientifics Committess, 2008. http://ec.europa.eu.

Gurlekian, Jorge A., et al. 2010. *Sistemas de análisis acústico y de reconocimiento automático en habla espontánea.* Buenos Aires, Argentina : Subjetividad y Procesos Cognitivos, Vol. 14, No 2, 2010. Pág. 89-98, ISSN impreso: 1666-244X, ISSN electrónico: 1852-7310.

Hernández, J M. 2010. jmhweb. *jmhweb.* [Online] 2010. http://jmhweb.wordpress.com/.

Hipertextual SL. 2012. appleweblog. *appleweblog.* [Online] 08 6, 2012. [Cited: 01 09, 2013.] http://appleweblog.com/2012/09/dragon-dictate-3-nuance-mac. CIF: B85439214.

Iqbal, Shumaila, Mahboob, Tahira and Khiyal, Malik Sikandar Hayat. 2011. Voice Recognition using HMM with MFCC for Secure ATM. *JCSI International Journal of Computer Science Issues.* 2011, Vol. 8, ISSN (Online): 1694-0814.

Javier Hernando, Climent Nadeu y José B. Mariño. 1997. *Técnicas robustas de reconocimiento del habla en ambientes adversos.* Barcelona : Procesamiento de lenguaje Natural, 1997. 08034.

Kola, Jonathan, Espy-Wilson, Carol and Pruthi, Tarun. 2011. *Voice Activity Detection.* 2011.

Kua, Jia Min Karen, Tharmarajah Thiruvaran and Nosratighods, Mohaddeseh. 2010. *Investigation of Spectral Centroid Magnitude and Frequency for Speaker Recognition.* Czech Republic : The University of New South Wales, Sydney, NSW 2052, Australia, 2010.

Larson, M., Eickeler, S. and Kohler, J. 2007. *Supporting radio archive workflows with vocabulary independent spoken keyword search.* s.l. : SSCS-SIGIR, 2007.

Liang, Bai, et al. 2010. *Feature Analysis and Extraction for Audio Automatic Classification.* ChangSha,R.P China : Multimedia Research &Development Center, National University of Defense and Technology, 2010.

Llerena, Y. 2006. *State of the art in Speaker Recognition.* s.l. : Facultad de Informática, Universidad de Ciego de Ávila, UNICA, 2006.

Logan, et al. 2004. *An experimental study of an audio indexing system for the web.* s.l. : ICASSP, 2, 676-679., 2004.

Lokesh and Balakrishnan, G. 2012. *Robust Speech Feature Prediction Using Mel-LPC to improve Recognition Accurracy.* s.l. : Asian Network for Scientific Information, 2012. ISSN: 1812-5638.

Londoño, Julián D. Arias, Domínguez, Germán Castellanos and Llorente, Juan I. Godino. 2007. *DISEÑO SIMULTÁNEO DE UNA ETAPA DE EXTRACCIÓN DE CARACTERÍSTICAS Y UN CLASIFICADOR BASADO EN HMM.* Universidad Tecnológica de Pereira, Colombia : cientia Et Technica, agosto, año/vol. XIII, número 035 , 2007.

López, Andrea. 2012. *Acustica Arquitectonica.* 2012.

López, Raúl Pérez. 2009. *Extracción de parámetros acústicos para la discriminación entre música y voz.* Madrid, España : s.n., 2009.

Lu, Guojun and Hankinson, Templar. 2000. *An Investigation of Automatic Audio Classification and Segmentation.* 2000.

Luengo, Iker, et al. 2009. *Detección de vocales mediante modelado de clusters de fonemas.* s.l. : Sociedad Española para el Procesamiento del Lenguaje Natural, 2009. ISSN: 1135-5948.

Martín, César San and Robeto, Carrillo. 2004. *IMPLEMENTACIÓN DE UN RECONOCEDOR DE PALABRAS AISLADAS DEPENDIENTE DEL LOCUTOR.* UTA Chile : Revista Facultad de Ingeniería, 2004.

Martineaux, Karina Silveira and Pérez, Reidelendy Fernández. 2008. *Comparación de algoritmos de clasificación y agrupamiento aplicando técnicas de Minería de datos.* s.l. : Universidad de las Ciencias Informáticas., 2008.

Martínez, F, et al. 2008. *Reconocimiento de voz, apuntes de cátedra para introducción a la inteligencia artificial.* 2008.

Martinez, S. J. 2010. *Implementación de un codificador de voz GSM .* Universidad de Alicante, Espana : s.n., 2010.

Mezcua, Belén Ruiz. 2005. *LA VOZ Y SU ESPECTRO.* 2005.

Miltre, García Miguel and Sánchez, Pérez Carolina Rocio. 2012. *Localización de silencios de una señal de voz.* Tlaxcala, México : Universidad Autónoma de Tlaxcala, Facultad de Ingeniería y Tecnología, 2012.

Mohanty, Mihir Narayan, Routray, Aurobinda and Kabisatpathy, Prithviraj. 2010. *A Statistical Approach for Voiced Speech Detection.* 2010.

Molina, María Pinto. 2011. Dossier Informático. *Dossier Informático.* [Online] 2011. [Cited: 05 21, 2013.] http://www.mariapinto.es/e-coms/recu_infor.htm.

Moreno, Yaretmy H. and Otero, Nayansi C. 2010. *Propuesta del algoritmo de reconocimiento biométrico de locutores sobre TP para la plataforma Platel.* La Habana, Universidad de las Ciencias Informáticas. : s.n., 2010.

Moyano, José María Drake. 2005. *Ruidos e Interferencias: Técnicas de reducción.* Cantabria : s.n., 2005.

NIST. 2006. *The spoken term detection (STD).* s.l. : http://www.nist.gov/speech/tests/std , 2006.

Noguerales, Javier Tejedor, Toledano, Doroteo T. and Pasamontes, José Colás. 2010. *Estado del arte en Wordspotting aplicado a los sistemas de extracción de información en contenidos de voz.* Madrid, España : I Congreso Español de Recuperación de Información (CERI), 2010.

O'Shaughnessy, D. 2008. *Invited Paper: Automatic speech recognition: History, methods and challenges.* s.l. : PR(41), No.10, pp.2965-2979, 2008.

Oliván, José Antonio Salvador and 'Aviles, Rosario Arquero. 2007. *Una aproximación al concepto de recuperación de información en el marco de la ciencia de la documentación.* s.l. : Investigación Bibliotecológica, Vol. 20, Núm. 41 pp 13-43, 2007. ISSN: 0187-358X.

Olson, D. L. & Delen, D. 2008. *Advanced Data Mining Techniques.* s.l. : Springer,Fist Edition. pp. 138., 2008.

Ong, Bee-Suan and Herrera, Perfecto. 2004. *Computing Structural Descriptions of Music through the Identification of Representative Excerpts from Audio Files.* 2004.

Otero, Dianabel Sánchez. 2010. *Conceptualización de un sistema informático para el reconocimiento y autenticación de personas por la voz.* Universidad de las Ciencias Informáticas : s.n., 2010.

Padilla, Argel González. 2012. *Nueva representación del habla para un sistema de reconocimiento.* La Habana : CUJAE, 2012.

Parada, Carolina, et al. 2011. *A Spoken Term Detection Framework for Recovering Out-of-Vocabulary Words Using the Web.* BM TJ Watson Research Center, New York, NY, 10598 : Human Language Technology Center of Excellence, Center for Language and Speech Processing, Johns Hopkins University, 2011.

Peeters, Geoffroy. 2004. *A large set of audio features for sound description.* Paris,France : s.n., 2004.

Pergaminovirtual. 2012. pergaminovirtual. *pergaminovirtual.* [Online] 2012. [Cited: 05 21, 2013.] http://www.pergaminovirtual.com.ar/definicion/Keyword.html 2012.

Ramírez, Genoveva Velásquez. 2008. *Reconocimiento de habla en MATLAB.* s.l. : Universidad de San Carlos de Guatemala, 2008.

Reyes, Fernando Peralta and Atencio, Anibal Cotrina. 2002. *Algoritmo Coper para la detección de actividad de voz.* Perú : Electrónica UNMSM, 2002.

Rincón, Jaime Varela and Pulgarin, Joan Ericloaiza. 2008. *Implementación de un reconocedor de palabras aisladas mediante redes neuronales sobre FPGA.* Pereira : s.n., 2008.

Rincón, Jaime Varela and Pulgarin, Johan Eric Loaiza. 2008. *Reconocimiento de palabras aisladas mediante redes neuronales sobre FPGA.* 2008.

Riso, Matías José Torres. 2009. *Robustez a efectos del canal en verificación de locutor, pp 22.* Santiago de Chile : universidad de Chile, 2009.

Rocamora, Martín and Herrera, Perfecto. 2007. *Comparing audio descriptors for singing voice detection in music audio files.* 2007.

Rochaix, Edmundo Carlos and Carlos, Juan Garay. 2010. Auditorias técnicas en obras en ejecución mediante la investigación de los ruidos y vibraciones. *SOCIEDAD ESPAÑOLA DE ACÚSTICA.* [Online] 2010. [Cited: 05 27, 2013.] http://www.sea-acustica.es/publicaciones/4350jh038.pdf..

Rodríguez, José Luis Oropeza. 2006. *Algorithms and Methods for the Automatic Speech Recognition in Spanish Language using syllabes.* Impreso en México : Computacion y Sistemas, 2006. CIC-IPN, ISSN 1405-5546.

Rodríguez, Luis Angel Pupo. 2011. *Implementación de un componente para la separación automática de segmentos de videos integrado al Sistema de Captura y Catalogación de Medias.* La Habana : Universidad de las Ciencias informáticas, 2011.

Roj, Valentín Sama, et al. 2005. *Utilización de medidas de confianza en sistemas de compresión del habla.* Madrid : Grupo de Tecnológica del Habla, Depto. de Ingeniería Electrónica,, 2005.

Rojo, Leticia Rueda. 2011. *MEJORAS EN RECONOCIMIENTO DEL HABLA BASADAS EN MEJORAS EN LA PARAMETRIZACIÓN DE LA VOZ.* 2011.

Romero, JESÚS Gustabo Medrano. 2012. *CONTROL DE ACCESO DE PERSONAL MEDIANTE EL USO DE MÚLTIPLES SENSORES BIOMÉTRICOS (RECONOCIMIENTO DE VOZ).* Mexico : INSTITUTO TECNOLÓGICO DE APIZACO, 2012.

Rúa, Enrique Argones, et al. 2009. *Audio-Visual Speech Asynchrony Detection using Co-Inertia Analysis and Coupled Hidden Markov Models.* 2009.

Rueda, E A and Torres, Y. 2006. *Detección de extremos en señales de voz utilizando características de energía y entropía.* Bucaramanga, Colombia : s.n., 2006.

Saravia, José Antonio Morejón. 2011. *SEGMENTACIÓN DE AUDIO Y DE LOCUTORES PARA RECUPERACIÓN DE NFORMACIÓN MULTIMEDIA Y SU APLICACIÓN A VIDEOS DE INFORMACIÓN TURÍSTICA.* s.l. : UNIVERSIDAD AUTÓNOMA DE MADRID,Ingeniería de Telecomunicación, 2011.

Sayán, Enviado por Claudia Salaverry. 2011. Monografias.com. *Monografias.com.* [Online] 2011. [Cited: 05 21, 2013.] http://www.monografias.com/trabajos84/recuperacion-informacion-internet/recuperacion-informacion-internet.shtml.

Schwarz, Petr. 2008. *Phoneme recognition based on long temporal context.* s.l. : Faculty of Information Technology Departament of Computer Graphics and multimedia, 2008.

Sintagmas, Categorías Gramaticales y. 2010. Categorías Gramaticales y Sintagmas. *Categorías Gramaticales y Sintagmas.* [Online] Categorías Gramaticales y Sintagmas, 2010. [Cited: 11 12, 2012.] http://categoriasgramaticalesysintagmas.com/que-son-los-fonemas-en-espanol-castellano.

Subramanian, Hariharan. 2004. *AUDIO SIGNAL CLASSIFICATION.* IIT Bombay : M.Tech. Credit Seminar Report, Electronic Systems Group, EE. Dept,, 2004.

Sung, Yun-Hsuan and Jurafsky, Dan. 2009. *Hidden Conditional Random Fields for Phone Recognition.* 2009.

Szoke, I., et al. 2005. *Comparison of keyword spotting approaches for informal continuous speech.* s.l. : ICSLP, 633-636, 2005.

Takuya Yoshioka, M.M. 2008. *Adaptive Supression of Non-Stationary Noise by using the Varational Bayesian Metod.* s.l. : ICASSP 2008, 2008.

Tejada, Guillermo, Reyes, Fernando Peralta and Atencio, Aníbal Cotrina. 2003. *RECONOCEDOR Y ANALIZADOR DE VOZ.* Perú : Universidad Nacional Mayor de San Marcos, 2003.

Tejedor, J., et al. 2008. *A novel two-level architecture plus confidence measures for a keyword spotting system.* s.l. : V Jornadas en Tecnología del Habla 247-250, 2008.

Tera, Jose Ignacio Puertas. 2002. *Robustez en el reconocimiento fonético de voz para aplicciones telefónicas.* Madrid,España : s.n., 2002.

Thambiratnam, K. and Sridharan, S. 2007. *Rapid yet accurate speech indexing using dynamic match lattice spotting.* s.l. : IEEE Trans. on Audio and Speech Processing, 15(1), 346-357, 2007.

Tomassi, D. R., et al. 2005. *Evaluación de técnicas clásicas de reducción de ruido en señales de voz.* s.l. : Revista de la sociedad argentina de Bioingeniería, 2005.

Tong, R., et al. 2006. *Fusion of acoustic and tokenization features for speaker recognition.* Singapore : 5th International Symposium on Chinese Spoken Language Processing, pp 566-577, 2006.

Topon, Botero and Libia, María. 2008. *Caracterización de los indicadores acústicos de la voz de los estudiantes del programa licenciatura en música.* Caldas, Colombia : s.n., 2008.

Venkateswarlu, R.L.K., Raviteja, R. and Rajeev, R. 2012. *The Performance Evaluation of Speech Recognition by Comparative Approach.* s.l. : INTECH, 2012.

Verbio Tecnologies . S. L. 2011. Verbio Speech Tecnologies. *Verbio Speech Tecnologies.* [Online] 2011. [Cited: 01 11, 2013.] http://www.verbio.com/webverbio3/. CIF / NIF: B61993135.

Vicente, Jesús de. 2007. *Contribuciones al reconocimiento robusto del habla.* LEGANES : s.n., 2007.

Welch, Lloyd R. 2003. *Hidden Markov Models and the Baum-Welch Algorithm.* s.l. : IEEE Information Theory Society Newsletter, 2003. ISSN 1059-2362.

X. Huang, A. Acero, H. Hsiao-Wuen. 2001. Spoken Language Processing. [book auth.] R. Reddy. *Prentice Hall PTR .* s.l. : ed., New Jersey, 2001.

Xin, L and Wang, B. 2001. *Utterance verification for spontaneous mandarin speech keyword spotting.* s.l. : ICH, 3, 397-401, 2001.

Zade, K.R. Aida, Ardil, C. and Rustamov, S.S. 2008. *Investigation of Combined use of MFCC and LPC Features in Speech Recognition Systems.* s.l. : World Academy of Science, Engineering and Technology, 2008.

Zanuy, M F. 2000. *Influencia en la detección voz/silencio en reconocimiento de locutor.* 2000.

Zuazua, Enrique and Llorente, Mikel Bilbao. 2005. *Temas relevantes de la matemática actual.* España : Subdirección General de Formación del Profesorado, 2005.

ANEXOS

ANEXO 1 TABLAS

Tabla 8 Resumen de las características de los modelos de clasificación

Clasificador	Ventajas	Desventajas
Distorsión Dinámica del Tiempo (DTW)	Detectan y comparan tramos fonéticos de alta estabilidad (vocales abiertas, consonantes nasales) aplicando técnicas de correlación cruzada, coherencia, entre otras, para la medida de distancias. Los sistemas DTW han sido utilizados en algunas metodologías forenses como un complemento a otros análisis clásicos.	Los principales inconvenientes de estos sistemas se relacionan con la enajenación de la información a nivel suprasegmental y la necesidad de supervisión en las tareas de segmentación.
Cuantización Vectorial(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).
Redes Neuronales (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 de clasificación.	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.
Modelos Ocultos de Markov (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. En la actualidad estos modelos son unos de los más usados en sistemas de reconocimiento de habla y tratamiento de señales de habla.	Alto costo computacional.
Modelo de mezclas de Gausianas (GMM)	Las GMM pueden representar con un alto grado de fidelidad un amplio margen de distribuciones muestrales, como es el caso con los diferentes coeficientes cepstrales que puede generar una locución. Además de las ventajas citadas, interesantes estudios comparativos sobre el rendimiento de diferentes técnicas de reconocimiento automático ante distintas circunstancias (procesos de entrenamiento, factores de degradación, etc.) han contribuido tomar como un buen sistema básico a las GMM.	Presentan un alto costo computacional implicando un tiempo considerable al crear los modelos.

ANEXO 2 FIGURAS

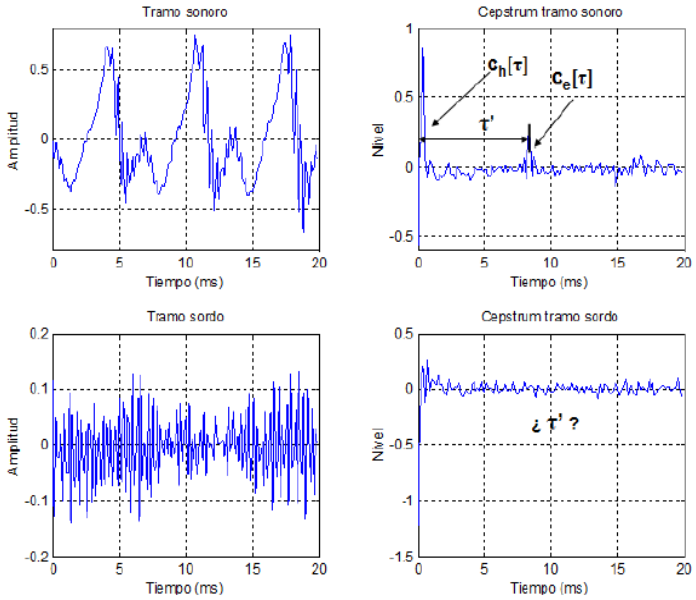


Figura. 39 Ejemplos de cálculo de cepstrum para dos segmentos de voz

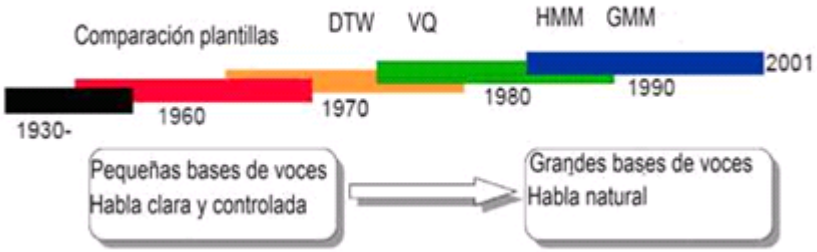


Figura. 40 Evolución de clasificadores para el reconocimiento del habla basada en patrones

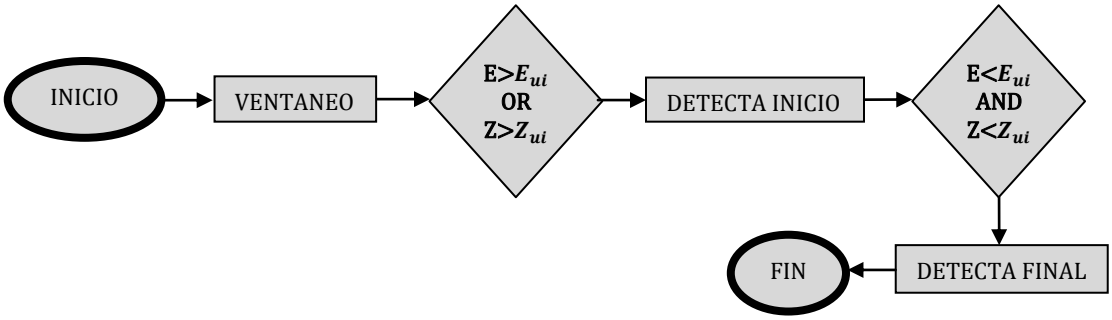


Figura. 41 Flujo para la detección de extremos mediante ZCR y energía

Donde E_{ui} y Z_{ui} corresponden a los valores de los umbrales de la energía y ZCR respectivamente.

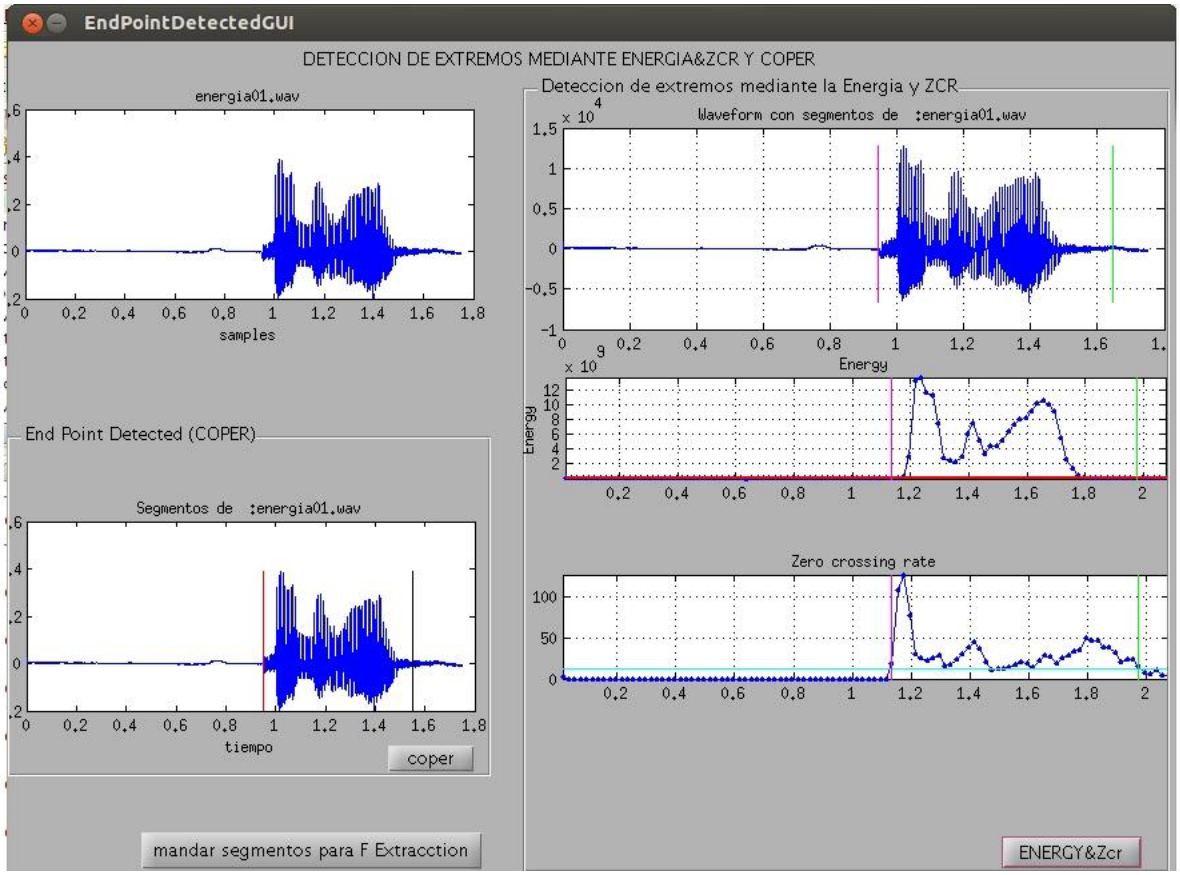


Figura. 42 Detección de extremos en MATLAB de la palabra "energía"

ANEXO 3 FUNCIONES

- **Tiempo_Ejecucion.m**

```
function [tiempo_coper,tiempo_EZCR] = Tiempo_Ejecucion(filename)
[y2,fs] = wavread(filename);
%algoritmo Coper
disp('Algoritmo Coper')
tiempo_inicio= cputime; CP2=0;
for n =2:length(y2),
   CP2 = CP2+abs(y2(n)*abs(y2(n))-y2(n-1)*abs(y2(n-1)));
  end
 tiempo_fin= cputime;
 tiempo_coper = tiempo_fin-tiempo_inicio;
 sprintf('Tiempo duracion. Tiempo=%2.2d',tiempo_coper)
%disp('Algoritmo Coper')
% algoritmo energía y ZCR
disp('Algoritmo Energía y ZCR')
 tiempoEZCR_inicio =cputime;
for n =2:length(y2),
     E(n) = (1/(length(y2))) * sum(abs(y2.^2));
   y3 = zeros(size(y2));
    y3(2:end) = y2(1:end-1);
    Z(n) = (1/(2*length(y2))) * sum(abs(sign(y2)-sign(y3)));
end
tiempoEZCR_fin =cputime;
tiempo_EZCR = tiempoEZCR_fin-tiempoEZCR_inicio;
%disp('Algoritmo Energía y ZCR')
sprintf('Tiempo duración. Tiempo=%2.2d',tiempo_EZCR)
 disp('Resultado')
if tiempo_coper > tiempo_EZCR,   sprintf('El algoritmo de E y ZCR es más rápido'),end
if tiempo_coper< tiempo_EZCR  ,  sprintf('El algoritmo COPER es más rápido'),end
if tiempo_coper== tiempo_EZCR,  sprintf('Ambos algoritmos demoran el mismo tiempo') ,end
```