Institución Universitaria

METODOLOGÍA PARA EL RECONOCIMIENTO DE COMANDOS DE VOZ BASADA EN DINÁMICAS ESTOCÁSTICAS

WILLIAM ALBERTO ACOSTA BEDOYA

Institución Universitaria Instituto Tecnológico Metropolitano ITM

Facultad de ingeniería

Medellín, Colombia

2013

METODOLOGÍA PARA EL RECONOCIMIENTO DE COMANDOS DE VOZ BASADA EN DINÁMICAS ESTOCÁSTICAS

WILLIAM ALBERTO ACOSTA BEDOYA

Proyecto de grado presentado como requisito parcial para optar al título de:

Magister Automatización y Control Industria

Director:

MSc Leonardo Duque muñoz

Línea de Investigación:

MIRP (Máquinas Inteligentes y Reconocimiento de Patrones)

Grupo de Investigación:

Automática y Electrónica

Instituto Tecnológico Metropolitano

Facultad, de ingenierías

Medellín, Colombia

2013

A mi Dios y mi amada esposa.

Agradecimientos

La entrega de este trabajo significa el punto final de una etapa, de la muchas experiencias; que me han hecho crecer tanto en lo académico, como en lo personal. Es el resultado de años de esfuerzo, de desilusiones, de alegrías, de buenos y malos momentos. Es hora pues de agradecer a todos aquellos que hicieron posible este sueño.

Agradezco infinitamente a Dios por haberme abierto los caminos para realizar este proyecto, al Espíritu Santo por darme la luz y la guía para realizar todos las actividades en esta maestría.

Agradezco a mi amada esposa María Elena Ortiz Ochoa, por su amor, por sacrificio y su paciencia.

Agradezco la paciencia a mi asesor MSc Leonardo Duque muñoz; con quien conté en todo momento para aclarar dudas y para orientarme; haciendo la transferencia de sus conocimiento, para hacer posible el desarrollo de este proyecto.

Quiero agradecer a toda mi familia: a mi madre, mis hermanos y hermanas, a mi suegra, a mis cuñados y cuñada, a todos mis sobrinos y sobrinas. porque ellos creyeron en mi y siempre me apoyaron.

Agradezco a mis amigos Cesar A. Rodríguez; por ser la persona que me impuso para realizar esta maestría. a Jorge A. Rodríguez; porque siempre conté con el mis momentos difíciles.

Por último pido disculpas si olvide dar las agracias a aluna de las muchas personas que se involucraron y me apoyaron en este proyecto.

¡A todos Dios los bendiga¡.

Resumen

En este trabajo se plantea una metodología para la identificación de comandos de voz con un diccionario reducido. El reconocimiento automático del habla, que tiene como objetivo de crear una interfaz hombre-máquina lo más natural posible, campo relativamente nuevo, y que viene en aumento sus aplicaciones, con un problema grande y es el ruido inherente a los sistemas de audio.

En el desarrollo de este proyecto se construyeron dos bases de datos: BD1, para un solo locutor con condiciones de ruido moderas, y BD2 para cinco locutores de edades y sexo diferentes, con la presencia de ruido ambiente e incluido ruido de vehículos pesados. Estas bases de datos están compuestas por las palabras comando: silla, adelante, atrás, derecha, izquierda y pare. Se realiza una representación de los comandos de voz con características frecuenciales obtenidas de la transformada de Fourier en tiempo corto.
Partiendo de los procesos estocásticos de HMM, se realiza un entrenamiento; construyendo un modelo para cada clase (palabra), para lo cual se hace una variación de mezclas Gaussianas que van desde 1 hasta 50 mezclas, y se realiza verificación utilizado un clasificadores estocásticos HMM.
En una segunda etapa del proceso, se ha agregado un filtro basado en la trasformada Wavelet y una función para retirar los punto de señal que no tiene información útil (eliminación del ruido ambiental y los silencios). Los resultados obtenidos con este método presentan una alta tasa de acierto en el orden de 98-100%, para el reconocimiento de estos comandos, con una relación de acierto- costo computacional muy significativa.
Una posible aplicación de esta herramienta será; asumir el mando de los motores acondicionados a una silla de ruedas. Lo que daría una solución al problema de movilidad; en personas con discapacidad motora, en el nivel de cuadrapléjicas, con lo cual se logrará dar una mayor autonomía en su movilidad; mejorando así su calidad de vida.

Palabras clave: Sistema de reconocimiento de habla, Modelos Ocultos de Markov, Mel Frecuency Cepstral Coefficients, transformada Wavelet, Algoritmo Forward -backward, Algoritmo de Viterbi, Algoritmo de Baum Welch, Expectation-Maximization Algorithm.

Contenido

Lista de figuras

Pág.

Lista de tablas

Pág.

Introducción

ANTECEDENTES

La experiencia en investigación de esta clase de proyecto se descarga en el grupo de investigación MIRP (Máquinas Inteligentes y Reconocimiento de patrones), que está conformado por un grupo de profesionales con formación de alta calidad, en diferentes áreas; con la posibilidad de prestar diversos servicios, este grupo realiza actividades en: procesamiento de señales, reconocimiento de patrones, control y automatización de procesos.

El Grupo de Investigación MIRP está adscrito al Centro de Investigación del ITM y en el último año ha estado desarrollando los siguientes proyectos de investigación: "Determinación de espectros de carga en pavimentos basada en análisis de vibraciones y procesamiento de señales", "Estudio teórico de las propiedades de portadores de carga separados espacialmente en hetero-estructuras semiconductoras", "Control de Turbinas de Viento", "Efectos de radiación láser sobre las propiedades electrónicas de hetero-.estructuras semiconductoras", entre otros. Entre las publicaciones recientes pueden listarse: "Sintonización óptima de un controlador para una turbina eólica con generador síncrono de magnetización permanente", "Optimal PI control of a wind energy conversion system using Particles Swarm", "Control de velocidad del motor de inducción empleando linealización por realimentación de estados", entre otras. Es un grupo enfocado hacia la investigación y solución de problemas en el área de la automatización y control industrial del país. Fomenta la participación activa de la academia en investigación y extensión dentro de la institución y crea vínculos con el sector industrial para capacitar a sus integrantes en el buen desempeño profesional, investigativo y de innovación.

JUSTIFICACIÓN

Desde tiempo inmemorial el hombre tiene el deseo de utilizar el habla como medio de comunicación con todo tipo de aparatos, cosas y animales, que se encuentra en su hábitat; de aquí toman una gran importancia las interfaces hombre-máquina.

Siempre que se hacen adelantos en el desarrollo de equipos y programas de computo para el control industrial, un ítem al que se le presta mucha importancia es la comunicación con el usuario.

Hasta este momento del desarrollo tecnológico el medio escrito (teclado) y el medio impreso (monitor); han sido de forma natural el medio de comunicación en la interfaz hombre-máquina.

Una persona con una pequeña experiencia en el teclado, puede alcanzar 20 ppm (palabras por minuto), un digitador promedio alcanza alrededor de 30 a 45 ppm, mientras que un digitador avanzado puede llegar hasta 60 ppm. Mientras que en una conversación natural y fluida una persona puede expresar alrededor de 200 ppm, y los estudios del investigador Ronald Carver han demostrado que un adulto puede escuchar con completa comprensión hasta 300 ppm, aún así, ni siquiera los subastadores pueden hablar más rápido que 250 ppm.

Con un adecuado diseño de un sistema que responda a comandos de voz; se podría dar órdenes directas a una maquina, sin requerir un teclado, esto para superar el problema de velocidad. además se podría superar otras limitaciones del orden de las variables físicas: como controlar una máquina mientras se está en la oscuridad o que el usuario tome una posición diferente a estar sentado frente al teclado.

El Reconocimiento Automático del Habla (RAH o ASR - Automatic Speech Recognition - por sus siglas en inglés), permite a los usuarios de los sistemas de control industrial, tomar el mando desde lugares o situaciones que con los controles tradicionales seria físicamente imposible o muy peligroso.

Por lo anteriormente expuesto, el reconocimiento automático del habla constituye uno de los mayores retos tecnológicos de la actualidad, al que se debe enfrentar con todas las herramientas disponibles.

Desde otra óptica en el boletín DANE, (2006, Mayo).DISCAPACIDAD – COLOMBIA, sobre el censo de 2005 (último censo realizado en el país), informa que al realizar un análisis a los resultados arrojados por la ronda de los Censos en América Latina para el decenio del 2000, se han arrojado los siguientes resultados: Venezuela (3,9%), México (1,8%) y Chile (2,2%) Ecuador (4,6%) Brasil (8,5%) y Colombia (6,4%), de estos resultados se puede observar que la población con discapacidad o deficiencias, varía entre el 1,8 y el 8,5%. Es importante destacar que un análisis comparado en términos cuantitativos y cualitativos a nivel latinoamericano es difícil, pues existente formas disímiles de abordar la discapacidad, pues algunos países se refieren a inválido, lisiado, impedido, minusválido, entre otras.

Los datos preliminares arrojados por el Censo del 2005 en Colombia, señalan que la tasa de prevalencia de discapacidad para el total de la población es del 6,4%, la cual es mayor en hombres (6,6%) que en mujeres (6,2). Por número de limitaciones, se señala que de las personas con discapacidad, el 71,2% presenta una limitación, el 14,5% dos limitaciones, el 5,7% tres limitaciones y el 8,7 % tres o más limitaciones permanentes.

En Colombia, con anterioridad a la Constitución Política de 1991, se habían dado algunas disposiciones con respecto a la atención a la discapacidad, sin embargo, a partir de su expedición, se ha venido consolidando un marco jurídico que determina los derechos de la población con discapacidad, y al mismo tiempo las obligaciones del estado y la sociedad para con ellos. Algunas de estas normas son:

* Ley 361 de 1997 "Por la cual se establecen mecanismos de integración social de las personas con limitación y se dictan otras disposiciones".

* Ley 762 de 2002, mediante la cual se aprueba la Convención Interamericana para la Eliminación de todas las formas de Discriminación contra las Personas con Discapacidad.

* Ley 100 de 1993 Por la cual se crea el Sistema de Seguridad Social Integral"

El no atender este problema, implicaría un abandono de aproximadamente 800,000 personas con discapacidad en el territorio Colombiano y unas 100,000 personas en Antioquia.

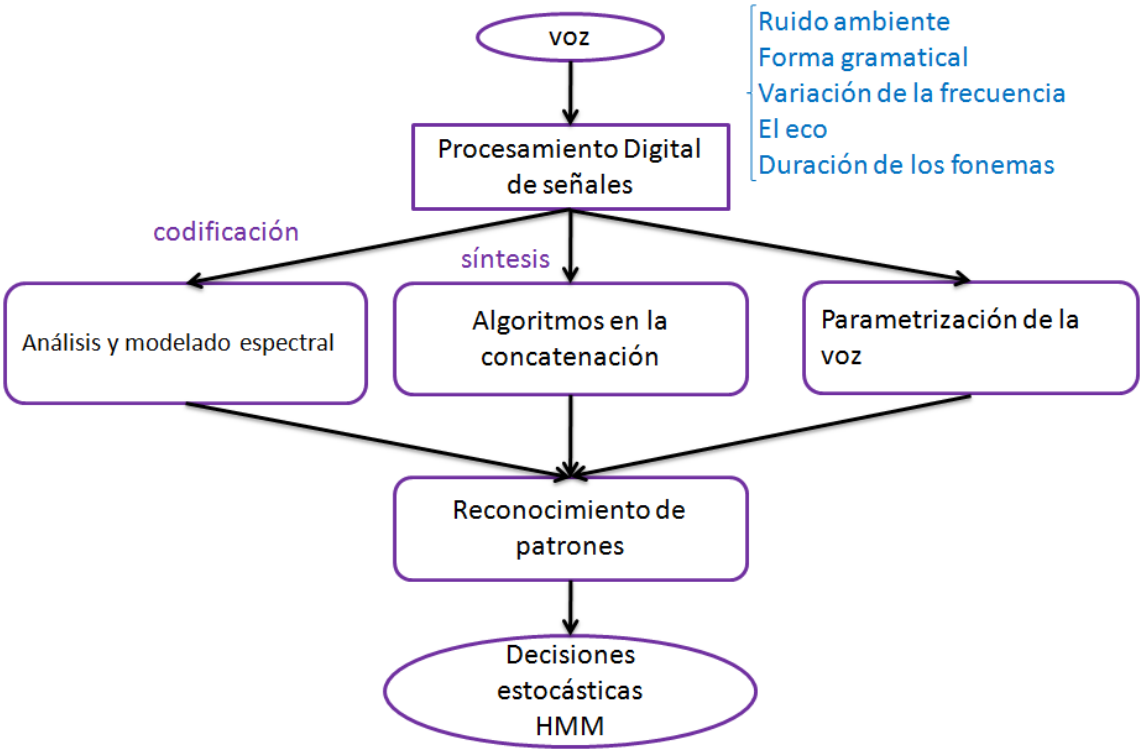
PLANTEAMIENTO DEL PROBLEMA

El campo del procesamiento de la señal de voz ha sido sujeto de estudio intenso en las últimas tres décadas, debido principalmente a los avances en las técnicas de procesamiento digital de señales y reconocimiento de patrones, además de la capacidad de proceso de los sistemas de cómputo. Su objetivo final es desarrollar interfaces hombre-máquina, que permitan que el ser humano se comunique de manera natural con los distintos dispositivos como robots, sistemas telefónicos o el computador de escritorio [1, 2, 3, 4, 5].

La comunicación entre personas tiene en cuenta gran diversidad de conocimiento, lo que permite sortear dificultades como el ruido ambiental, el acento y la concatenación de palabras, además de asuntos gramaticales [6]. El Reconocimiento Automático del Habla (RAH o ASR - Automatic Speech Recognition - por sus siglas en inglés), es un campo difícil de tratar, debido principalmente a las variaciones de fonación (los locutores no hablan igual), las ambigüedades en la señal acústica (no toda la información presente está relacionada con el habla), la falta de cuidado del hablante, la variación en la frecuencia y duración de los fonemas, y la presencia de ruido o interferencias [7, 8, 9, 10]. Los sistemas actuales están restringidos a ambientes controlados, a ser utilizados con un grupo de hablantes reducido o requieren de posicionamiento especial del micrófono resultando en interfaces poco naturales.

El procesamiento de voz se divide en tres temas de interés: codificación, síntesis y reconocimiento del habla, los cuales aun son problemas abiertos de investigación. La codificación de voz fue ampliamente estudiada en la década de los ochenta y principios de los noventa, los esfuerzos se concentraron en el desarrollo de algoritmos de parametrización de la señal [11, 12], la extracción de la frecuencia fundamental [13], y el análisis y modelado de la envolvente espectral [14]. Mientras que en la síntesis el gran paso se produjo a mediados de los noventa, cuando se introdujeron los algoritmos basados en la concatenación de unidades pregrabadas [12, 15, 16, 17]. Con respecto a los sistemas de reconocimiento del habla, los dos bloques fundamentales consisten en un sistema de parametrización de la señal de voz y un sistema de reconocimiento de patrones [7, 1, 18]. Siendo los sistemas de codificación los que han logrado un mayor grado de satisfacción.

Figura 1: Esquema de los problemas del Reconocimiento Automático del Habla



La figura 1 presenta un esquema de los inconvenientes que se requiere salvar para poder realizar un reconocimiento automático del habla.

En cara al problema de reconocimiento automático del habla se han propuesto estrategias basadas en varias aproximaciones, pero la técnica de reconocimiento de patrones que mejores resultados ha ofrecido para descifrar la señal de voz hasta el momento, es aquella basada en la teoría de decisión estadística. Esta técnica permite encontrar la secuencia de patrones que tiene la mayor probabilidad de estar asociada a la secuencia de observaciones acústicas de entrada.

Los modelos ocultos de Markov o Hidden Markov Models (HMMs por sus siglas en inglés) son modelos estadísticos, cuya salida es una secuencia de símbolos o cantidades y poseen mejores tasas de identificación de habla distorsionada y normal que aquellas basadas en plantillas o en otras aproximaciones [19]. Entre las razones de su popularidad sobresale el hecho de que la señal de voz puede ser vista como una señal

estacionaria a trozos, es decir, se puede asumir que en un corto tiempo la señal puede ser modelada como un proceso estacionario [18, 20]. Además, los HMMs pueden ser entrenados automáticamente, son viables computable [18].

SÍNTESIS DEL PROBLEMA

La elección de un método apropiado para la caracterización de la señal de voz, la extrae de información lingüística en un ambiente ruidoso o con ambigüedad de la señal acústica. Además de la definición de un modelo apropiado, que permita un buen desempeño en la clasificación. son los principales desafíos que tiene un reconocedor automático del habla.

Un RAH que permita establecer una comunicación hombre maquina, que no implique la utilización de las manos, será una solución a los problemas de limitación física.

HIPÓTESIS

Con el uso de técnicas de reconocimiento de patrones de voz y de modelos ocultos de Markov, se obtendrá un comando a partir de la identificación de palabras aisladas, para el control de una silla de ruedas.

OBJETIVOS

Objetivo General

Desarrollar una metodología de reconocimiento de voz, sobre un diccionario reducido de palabras aisladas, empleando modelos ocultos de Markov, orientado a tener el control sobre una silla de ruedas.

Objetivos Específicos

- Adecuar los registros de voz mediante técnicas de preprocesamiento de señales con el fin de reducir las perturbaciones adquiridas.
- Caracterizar la voz mediante técnicas dinámicas estocásticas con el fin de hallar la representación efectiva que permita el reconocimiento de comandos de palabras aisladas.
- Desarrollar un sistema de clasificación con base en técnicas de entrenamiento discriminativa, que esté relacionada con una medida del desempeño adecuado para el control de una silla de ruedas.

1. Marco teórico

1.1 Procesamiento del habla

1.1.1 Producción de la voz

La voz es un sonido que, producido por la laringe y amplificado por las estructuras de resonancia, nos permite la comunicación oral y alcanza en el canto su máxima expresión y belleza.

El proceso de la voz se inicia con la voluntad. En principio aparece el deseo de emitir un sonido, y éste desencadena en el sistema nervioso central un gran número de órdenes que pondrán en funcionamiento todos los elementos que producen la voz: mecanismos de la respiración, de la fonación, de la articulación, de la resonancia, de la expresión, etc.

Cuando queremos emitir un sonido, ya sea para hablar o cantar, las cuerdas vocales se cierran. En esta situación el aire espirado no encuentra vía libre para salir y se crea una presión; cuando ésta alcanza un grado determinado, vence la resistencia que ofrecían las cuerdas vocales y al pasar a través del espacio que éstas le dejan las hace vibrar, produciendo un leve sonido que será más grave o más agudo según el grado de tensión a que sean sometidas (entre otras condiciones). El sonido resultante se amplificará y se modificará al pasar por las cavidades de resonancia. Estas cavidades son espacios vacíos de la vía respiratoria (Laringe, faringe, boca y fosas nasales).

La física ha establecido que para que exista sonido se requieren tres elementos:

- Parte de la vibración penetraría en el oído.
- Parte de la vibración rebotaría sobre la cabeza y volvería en la dirección de la que procedía (reflexión).
- Parte de la vibración lograría rodear la cabeza y continuar su camino (difracción).

1. Un cuerpo que vibre.
2. Un medio elástico que vibre (las ondas sonoras son mecánicas que se propagan por la expansión y compresión del propio medio).
3. Una caja de resonancia que amplifique esas vibraciones, permitiendo que sean percibidas por el oído.

La voz humana cumple con las tres condiciones señaladas:

1. El cuerpo elástico que vibra son las cuerdas vocales.
2. El medio elástico es el aire.
3. La caja de resonancia está formada por parte de la laringe y faringe, por la boca y por la cavidad nasal. La producción de la voz se realiza a partir de diferentes procesos, estos procesos son denominados niveles de producción de la voz.

Figura 1-1:Sección del tracto vocal (tomado de ZonaForo MeriStation)

La figura 1-1 ilustra el conjunto de órganos que intervienen en la fonación; puede dividirse en tres grupos bastante bien delimitados:

1.1.2 Proceso básico en la percepción del habla

Etapas, basadas en consideraciones lingüísticas:

Análisis auditivo periférico: se produce una descodificación de las señales del habla en el sistema auditivo periférico. Los mecanismos de descodificación son de dos clases:

1. Neuroacústicos.
2. Psiocacústicos.

Análisis auditivo central: El cometido es extraer de la señal una serie de patrones espectrales (frecuencia fundamental, dirección, transiciones de formantes) y temporales (desfase entre eventos) y los almacena en la memoria ecoica. El análisis de estos patrones da lugar a claves acústicas que forman los fonemas.

Análisis acústico-fonético: Se efectúa un procesamiento lingüístico de la señal. Se trata de identificar los segmentos o fonemas del habla. Las claves acústicas se acoplan a los rasgos fonéticos que son representaciones abstractas mediadoras entre los planos físicos (acústico) y lingüístico (fonético).

1. Permite descubrir las constancias perceptivas (categorización perceptiva del habla), que nos permiten identificar sonidos discretos, resolviéndose los problemas de segmentación y variabilidad.
2. Existen en este nivel unos detectores de rasgos o mecanismos neurales especializados en la identificación de rasgos fonémicos distintivos (sonoridad, nasalidad, etc.).

Análisis fonológico: los rasgos y segmentos fonéticos identificados en la anterior etapa son convertidos en representaciones abstractas de los sonidos (segmentos fonológicos) que se combinan para formar unidades superiores, como sílabas y palabras.

1. En este nivel ciertas distinciones fonéticas se convierten en variaciones alofónicas del mismo fonema, explicando así ciertos fenómenos de asimilación o transformación fonética.
2. Como resultado del análisis fonológico aparece una secuencia lineal de fonemas, cuyos componentes están organizados jerárquicamente: el inicio u onset (grupo consonántico inicial optativo), y la rima o rime y la coda (terminación consonántica opcional).

1.1.3 Sistema auditivo periférico

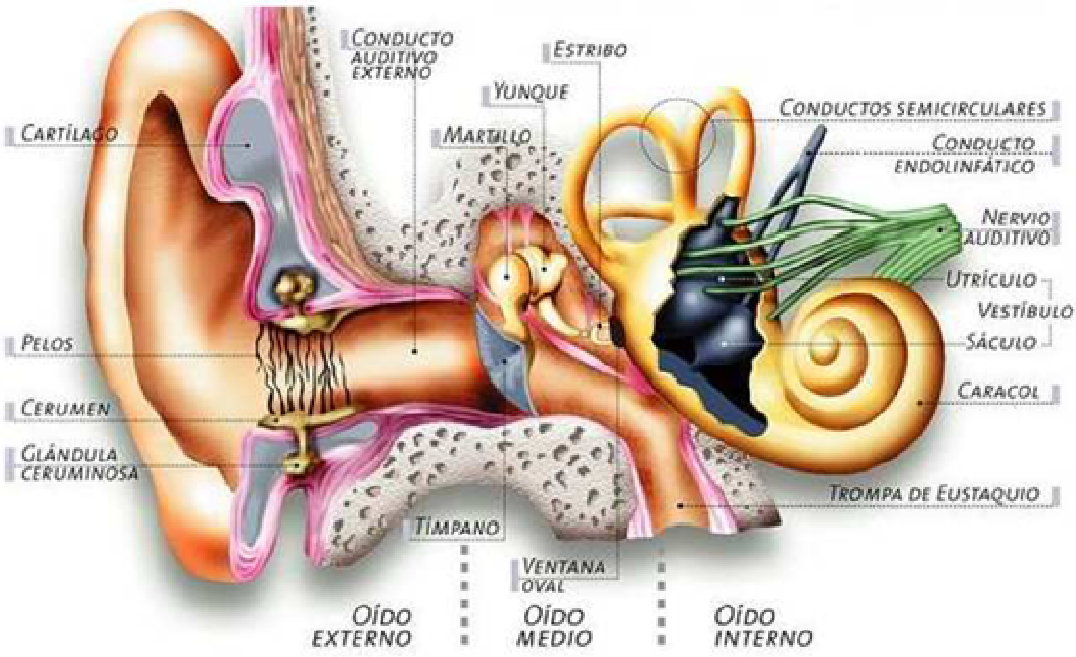
El sistema auditivo humano se puede dividir en etapas:

1. Cuando el sonido llega al oído, las ondas sonoras son recogidas por el pabellón auricular (o sistema auricular), los frentes de onda llegarían de forma perpendicularmente y el proceso de audición resultaría ineficaz (gran parte del sonido se perdería):

- Parte de la vibración penetraría en el oído.
- Parte de la vibración rebotaría sobre la cabeza y volvería en la dirección de la que procedía. (reflexión).
- Parte de la vibración y continuar su camino. (difracción).

El pabellón auricular humano es mucho menos direccional que el de otros animales (como los perros) que poseen un control voluntario de su orientación. (Los perros pueden mover las orejas a voluntad, los humanos no).

Figura 1-2:Estructura del sistema periférico auditivo (tomado de ZonaForo MeriStation).



Cuando el sonido llega al oído, las ondas sonoras son recogidas por el pabellón auricular (o aurícula), la figura 1-2 muestra la estructura del sistema periférico auditivo. El pabellón auricular, por su forma helicoidal, funciona como una especie de "embudo" que ayuda a dirigir el sonido hacia el interior del oído

2. Una vez que ha sido recogido el sonido, las vibraciones provocadas por la variación de presión del aire cruzan el canal auditivo externo y llegan a la membrana del tímpano, ya en el oído medio.

El conducto auditivo actúa como una etapa de potencia natural que amplifica automáticamente los sonidos más bajos que proceden del exterior. Al mismo tiempo, en el caso contrario, si se produce un sonido muy intenso que puede dañar el oído, el conducto auditivo segrega cerumen (cera), con lo que cierra parcialmente el conducto, protegiéndolo.

En el oído medio, se produce la transducción, es decir, la transformación la energía acústica en energía mecánica. En este sentido, el oído medio es un transductor mecánico-acústico.

Además de transformar la señal, antes de que ésta llegue al oído interno, el oído medio la habrá amplificado.

La presión de las ondas sonoras hace que el tímpano vibre empujando a los osículos, que, a su vez, transmiten el movimiento del tímpano al oído interno. Cada osículo empuja a su adyacente y, finalmente a través de la ventana oval. Es un proceso mecánico, el pie del estribo empuja a la ventana oval, ya en el oído interno.

Esta fuerza empuja a la venta oval es unas 20 veces mayor que la que empujaba a la membrana del tímpano, lo que se debe a la diferencia de tamaño entre ambas.

Esta presión ejercida sobre la ventana oval, gracias a la helicotrema penetra en el interior de la cóclea (caracol) y pone en movimiento el líquido linfático (endolinfa o linfa) que ésta contiene.

El líquido linfático se mueve como una especie de ola y, transmite las vibraciones a las dos membranas que conforman la cóclea (membrana tectorial, la superior, y la membrana basilar, la inferior).

Entre ambas membranas se encuentra el órgano de Corti, que es el transductor propiamente dicho. En el órgano de Corti se encuentran las células receptoras. Existen aproximadamente 24 000 de estas fibras pilosas, dispuestas en 4 largas filas que son las que recogen la vibración de la membrana basilar.

Como la membrana basilar varía en masa y rigidez a lo largo de su longitud su frecuencia de resonancia no es la misma en todos los puntos:

- En el extremo más próximo a la ventana oval y al tímpano, la membrana es rígida y ligera, por lo que su frecuencia de resonancia es alta.
- Por el contrario, en el extremo más distante, la membrana basilar es pesada y suave, con lo que su resonancia es baja frecuencia.

El margen de frecuencias de resonancia disponible en la membrana basilar determina la respuesta en frecuencia del oído humano, las audiofrecuencias que van desde los 20 Hz hasta los 20 kHz. Dentro de este margen de audiofrecuencias, la zona de mayor sensibilidad del oído humano se encuentra en los 1000 y los 4000 Hz.

Esta respuesta en frecuencia del oído humano, permite que seamos capaces de tolerar un rango dinámico que va desde los 0 db (umbral de audición) a los 120 dB (umbral de dolor)

El movimiento de la membrana basilar afecta las células del órgano de Corti de forma diferencial, en función de su frecuencia de resonancia. Al ser empujadas contra la membrana tectorial, las células pilosas generan patrones característicos de cada tono o (frecuencia), que al llegar aquí, al final del proceso fisiológico, son idénticas a la sonido original.

En función de estos patrones, al ser estimuladas las células pilosas producen un componente químico que genera los impulsos eléctricos que son trasmitidos a los tejidos nerviosos adyacentes (nervio auditivo y, de ahí, al cerebro), donde se producirá la percepción del sonido gracias al sistema auditivo central.

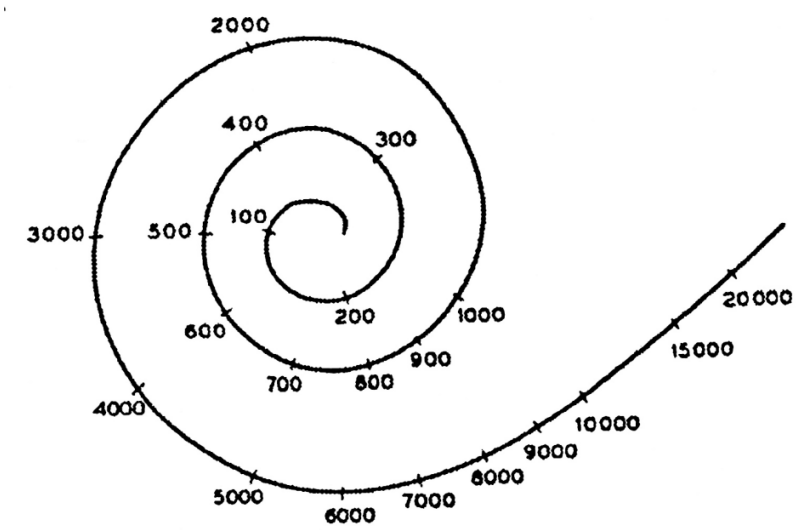
Las células del órgano de Corti, (células ciliares, capilares o pilosas), no tienen capacidad regeneradora, es decir, cuando se lesionan se pierde audición de forma irremediable. Además, con la edad, desciende la agudeza auditiva de los seres humanos.

1.1.4 La cóclea como analizador en frecuencia

La membrana basilar es una estructura cuyo espesor y rigidez no es constante; cerca de la ventana oval, la membrana es gruesa y rígida, pero a medida que se acerca hacia el vértice de la cóclea se vuelve más delgada y flexible. La rigidez decae casi exponencialmente con la distancia a la ventana oval; esta variación de la rigidez en función de la posición afecta la velocidad de propagación de las ondas sonoras a lo largo de ella, y es responsable en gran medida de un fenómeno muy importante: la selectividad en frecuencia del oído interno.

La cóclea puede ser aproximada como un banco de filtros. La figura 1-3 ilustra la distribución de frecuencias en la cóclea, los filtros correspondientes al extremo más próximo a la ventana oval y al tímpano responden a las altas frecuencias, ya que la membrana es rígida y ligera. Por el contrario, en el extremo más distante, la membrana basilar es pesada y suave, por lo que los filtros correspondientes responden a las bajas frecuencias

Figura 1-3:Distribución de frecuencias en la cóclea



1.2 Modelos Ocultos de Markov

1.2.1 Introducción

Los procesos del mundo real en general, producen resultados observables que pueden ser caracterizados como señales [18]. Las señales puede ser de naturaleza discretas (por ejemplo, los caracteres de un alfabeto finito, vectores cuantizados de un libro de códigos, etc.), o de naturaleza continua (por ejemplo, el muestreo de la voz, las mediciones de temperatura, música, etc.). La fuente de señal puede ser estacionaria (es decir, sus propiedades estadísticas no varían con el tiempo), o no estacionario (es decir, las propiedades de la señal varía con el tiempo). Las señales pueden ser puras (es decir, viene estrictamente de una sola fuente), o puede ser dañadas por otras fuentes de señal (ruido, por ejemplo) o por las distorsiones de transmisión, reverberación, etc.

Un problema de interés fundamental es caracterizar dicha señal del mundo real en términos de modelos de señal. Hay varias razones por las que se está interesado en la aplicación de modelos de señal.

En primer lugar, un modelo de señal puede servir de base para una descripción teórica de un sistema de procesamiento de señal que pueden ser utilizados para procesar la señal con el fin de proporcionar una salida deseada. Por ejemplo, si estamos interesados en mejorar una señal de voz distorsionada por el ruido y la distorsión de la transmisión, Se puede utilizar el modelo para diseñar un sistema de forma óptima para eliminar el ruido y deshacer la distorsión de la transmisión. Un segundo motivo importante del modelo de señal es que son potencialmente capaces de permitir a aprender mucho acerca de la fuente de señal (es decir, el proceso del mundo real que produjo la señal) sin tener que tener el código fuente disponible. Esta característica es importante cuando el costo de obtener señales de la fuente real es alto.

En este caso, con un buen modelo de la señal, se puede simular el origen y aprender tanto como sea posible a través de simulaciones. Finalmente, la razón más importante por la que los modelos señal son importantes es que a menudo trabajan muy bien en la práctica, y nos permitirá realizar importantes sistemas por ejemplo, los sistemas de predicción, los sistemas de reconocimiento, el sistema de identificación, etc., de una manera muy eficiente.

Estas son algunas posibles opciones para utilizar en tipos de modelos de señal para caracterizar las propiedades de una señal determinada.

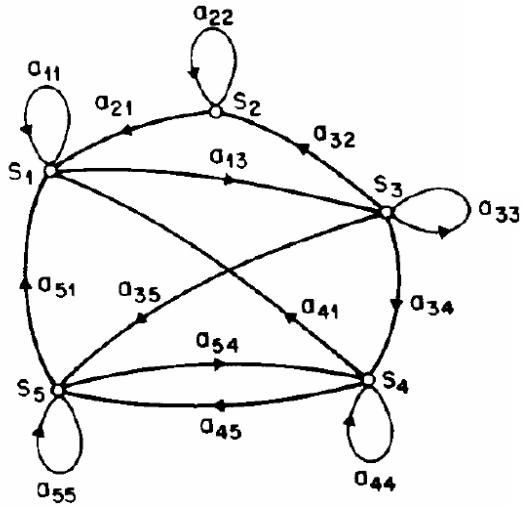
En términos generales se puede dicotomizar los tipos de modelos de señal clase de modelo determinista, y la clase de modelos estadísticos. Generalmente los modelos deterministas explotan algunas de las propiedades conocidas de la señal específica, por ejemplo, que la señal es una onda senoidal, o la suma de exponenciales, etc. En estos casos, la especificación del modelo de señal suele ser sencillo, todo lo que se requiere para determinar (estimar) los valores de los parámetros del modelo de señal (por ejemplo, la amplitud, frecuencia, fase de una onda sinusoidal, las amplitudes y las tasas de exponenciales, etc.). La segunda gama de clases de modelos de la señal es el conjunto de modelos estadísticos en los que se pretende caracterizar sólo las propiedades estadísticas de la señal. Ejemplos de tales modelos estadísticos son los procesos Gaussianos, los procesos de Poisson, procesos de Markov, y los procesos oculto de Markov, entre otros. El supuesto fundamental del modelo estadístico es que la señal puede ser bien caracterizada como un proceso paramétrico aleatorio, y que los parámetros del proceso estocástico se pueden determinar (estimado) de una manera precisa y bien definida.

1.2.2 Procesos discretos de Markov [1]

Considere un sistema que puede ser descrito en cualquier momento como en un conjunto de N estados distintos, S_1, S_2, . . . S_N, como se ilustra en la figura 1-4 (donde N = 5 para la simplicidad). Regularmente en espaciados tiempos discretos, el sistema sufre un cambio de estado (posiblemente regresar hasta el mismo estado) de acuerdo a un conjunto de probabilidades asociadas con el estado.

[1] Una buena visión general de los procesos discretos de Markov se encuentra en [21, cap. 5]

Figura 1-4: Una cadena de Markov con cinco estados (etiquetado como S_1 a S_5), con transiciones de estado seleccionado.(Tomado de Rabiner L,1989)



Denotamos los instantes de tiempo asociada a los cambios de estado como t = 1, 2, ... y denotamos el tiempo del estado actual t como q_t. Una descripción completa probabilística de dicho sistema, en general, requieren la especificación del estado actual (en el tiempo t), así como todos los estados que lo preceden. Para el caso específico de sistemas discretos, de primer orden, la cadena de Markov, esta descripción probabilística es truncada sólo al estado actual y al estado anterior, por ejemplo,

$$P[q_t = S_J | q_{t-1} = S_i, q_{t-2} = S_k, \ldots]$$

$$= P[q_t = S_J | q_{t-1} = S_i] \quad (1)$$

Además, sólo tenemos en cuenta aquellos procesos en el lado derecho de (1) es independiente del tiempo, lo que conduce al conjunto de probabilidades de transición de estados a_{ij} de la forma

$$a_{ij} = P\left[q_t = S_j \,|\, q_{t-1} = S_i \right] \quad , 1 \le i, j \le N \quad (2)$$

Con los coeficientes del estado de transición que tienen las propiedades

$$a_{ij} \ge 0 \quad (3a)$$

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{3b}$$

Debido a que obedecen las restricciones estándar estocásticas.

El proceso estocástico anterior podría ser llamado un modelo de Markov observable de donde la salida del proceso es el conjunto de estados en cada instante de tiempo, donde cada estado corresponde a un examen físico (visible, observable) del evento.

Al utilizar la notación

$$\pi_i = P\left[q_1 = S_i\right], \quad 1 \le i \le N \tag{4}$$

Permitirá indicar las probabilidades de estado inicial.

Otra cuestión interesante que se puede preguntar (y responder utilizando el modelo) es:

Dado que el modelo se encuentra en un estado conocido, ¿Cuál es la posibilidad de que permanezca en el estado para exactamente *d veces*?

Esta probabilidad puede ser evaluada como la probabilidad de la secuencia de observación

$$O = \left\{S_i, S_i, S_i, ..., S_i, S_j \neq S_i\right\}$$

1 2 3 d d+1

Teniendo en cuenta el modelo, que es

$$P\left(O \middle| Modelo, q_1 = S_i\right) = \left(a_{ii}\right)^{d-1}\left(1 - a_{ii}\right) = P_i\left(d\right) \tag{5}$$

La cantidad de *Pi(d)* es la función de probabilidad (discreta) de densidad de duración *d*, en el estado *i*. Esta densidad de duración exponencial es característica de la duración del estado en una cadena de Markov.

Sobre la base de *Pi(d)*, podemos fácilmente calcular el número esperado de observaciones (duración) en un estado, la condición de que comiencen en ese estado como

$$\bar{d}_i = \sum_{d=1}^{\alpha} d p_i(d) \tag{6ª}$$

$$\bar{d}_i = \sum_{d=1}^{\alpha} d\,(a_{ii})^{d-1}(1-a_{ii}) = \frac{1}{1-a_{ii}} \tag{6b}$$

1.2.3 Elementos de un HMM

Los ejemplos anteriores, nos dan una idea bastante buena de lo que es un HMM y cómo se puede aplicar a algunos escenarios simples. Ahora se define formalmente los elementos de un HMM, y explicar cómo el modelo genera secuencias de observación.

Un HMM se caracteriza por lo siguiente:

1) *N*, el número de estados en el modelo. Aunque los estados están ocultos, para muchas aplicaciones prácticas a menudo hay un significado físico conectado a los estados o conjuntos de estados del modelo. Por lo tanto, en los experimentos de lanzamiento de monedas, cada estado corresponde a una moneda sesgada distinta. En general, los estados están interconectados de tal manera que desde cualquier estado se puede llegar a cualquier otro estado (por ejemplo, un modelo ergódico); sin embargo, se vera más adelante en este trabajo que a menudo otras posibles interconexiones de los estados son de interés. Se denotan los estados individuales como $S = \{S_1, S_2, \ldots, S_N\}$, y el estado en el tiempo *t* como q_t.

2) *M*, el número de símbolos distintos de observación por estado, es decir, el tamaño del alfabeto discreto. Los símbolos corresponden a la observación de la salida física del sistema que está siendo modelado. Denotamos los símbolos individuales como $V = \{v_1, v_2, \ldots, v_M\}$.

3) La distribución de probabilidad transición de estado $A = \{a_{ij}\}$ en donde

$$a_{ij} = P[q_{t+1} = S_j | q_t = S_i], \qquad 1 \leq I, j \leq N \qquad (7)$$

Para el caso especial en que cualquier estado puede llegar a cualquier otro estado en un simple paso, hemos $a_{ij} > 0$ para todo *i*, *j*. Para otros tipos de HMM, tendríamos $aij = 0$ para uno o más *(I, J)* pares.

4) La observación del símbolo de distribución de probabilidad en el estado *j*, $B = \{b_j(k)\}$, donde

$$b_j(k) = P[V_k \ para\ t | q_t = S_j], \qquad 1 \leq j \leq N$$

$$1 \leq k \leq M \qquad (8)$$

5) La distribución del estado inicial $\pi = \{\pi_j\}$ donde

$$\pi_j = P[q_1 = S_j], \qquad 1 \leq j \leq N \qquad (9)$$

Dado valores apropiados de *N*, *M*, *A*, *B*, y π, el HMM puede ser utilizado como un generador para proporcionar una secuencia de observación

$$O = O_1 O_2 \ldots O_T \qquad (10)$$

(Donde cada observación *O*, Es uno de los símbolos de *V*, y *T* es el número de la secuencia de observaciones en) como sigue:

1) Seleccione un estado inicial $q_1 = S_j$ de acuerdo con la distribución de estado inicial π.

2) Establecer $t = 1$.

3) Seleccione $O_t = V_k$ de acuerdo con la distribución de probabilidad símbolo en estado S_j, es decir $b_j(k)$.

4) Transición hacia un nuevo estado $q_{t+1} = S_j$ de acuerdo con la probabilidad de transición de estado la distribución de estado para S_j es decir a_{ij}.

5) Establecer $t = t + 1$; volver al paso 3) si *t*<*T*; de lo contrario terminará el procedimiento.

El procedimiento anterior puede ser utilizado tanto como un generador de observaciones, o como generador de un modelo para una secuencia de observación dado por un HMM adecuado. Se puede apreciar desde la discusión anterior, que una especificación completa de un HMM requiere la especificación de dos parámetros del modelo (*N* y *M*), especificación de símbolos de observación, y la especificación de la probabilidad de tres medidas *A*, *B*, y *π*. Por conveniencia, se utiliza la notación compacta

$$\lambda = (A, B, \pi) \qquad (11)$$

Para indicar el juego de parámetros completo del modelo.

1.2.4 Los tres problemas básicos para los HMMs[2]

Dada la estructura de los HMM vista en la sección anterior, hay tres problemas básicos de interés que deben ser resueltos por el modelo para ser útil en aplicaciones del mundo real. Estos problemas son los siguientes:

1. Problema de Evaluación: Dada una secuencia de observaciones $O=\{ o_1, o_2, ... o_T\}$ (siendo *T* la longitud de la secuencia de observación) y el modelo $\lambda=\{A, B, \Pi\}$, el problema es cómo obtener de forma eficiente $P(O|\lambda)$, es decir, la probabilidad de obtener una secuencia de observación dado un modelo determinado.

2. Problema de Decodificación: Dada una secuencia de observaciones $O=\{ o_1, o_2, ... o_T\}$ y el modelo $\lambda=\{ A, B, \Pi\}$, encontrar una secuencia de estados $Q=\{ q_1, q_2, ... q_T)$ más probable, para la secuencia de observaciones dada.

3. Problema de aprendizaje: Como maximizar los parámetros del modelo *λ* para obtener la máxima $P(O| \lambda)$ para unas observaciones de entrenamiento *O*.

Si solucionamos el problema de evaluación, se podría evaluar como de bueno es un modelo HMM para una secuencia de observación. Además podríamos usarlo para hacer reconocimiento de patrones ya que la probabilidad $P(O|\lambda)$ determina la probabilidad de observación. Si solucionamos el problema de decodificación podremos saber la

[2] El material incluido en esta sección y en la Sección I ll se basa en las ideas de IDA presentadas por Jack Ferguson en las conferencias de los Laboratorios Bell.

secuencia de estados óptima para una secuencia de observación. En otras palabras, descubriríamos la secuencia oculta de estados. Por último la solución del problema de aprendizaje nos daría los parámetros de un modelo *λ* dado una serie de datos de entrenamiento.

1.2.4.3 Evaluación de HMM – Algoritmo Forward -backward

Para el cálculo de la probabilidad *P(O|λ)*, lo que resulta más intuitivo es el cálculo como la suma de las probabilidades de todas las secuencias de estados:

$$P(O|\lambda) = \Sigma P(O|q,\lambda) P(q|\lambda) \quad (12)$$

En otras palabras, enumerar todas las posibles secuencias de estados de longitud *T* que generen la secuencia de observación *O* y sumando sus probabilidades según el teorema de la Probabilidad Total Para ello consideremos una determinada secuencia de estados: $Q=(q_1,q_2,...,q_T)$ donde q_1 es el estado inicial. La probabilidad de la secuencia de observación *O* dada la secuencia de estados *Q* es:

$$P(O|Q,\lambda) = \prod_{t=1}^{T} P(O_t|q_t,\lambda) \quad (13)$$

Donde se asume independencia estadística de las observaciones. Por lo tanto se obtiene:

$$P(O|Q,\lambda) = b_{q1}(O_1) b_{q2}(O_2) \dots b_{qT}(O_T) \quad (14)$$

Por otra parte la probabilidad de la secuencia de estados *Q* se puede expresar como:

$$P(O|\lambda) = \pi_{q1} a_{q1q2} a_{q2q3} \dots a_{qT-1qT} \quad (15)$$

Que se interpreta como la probabilidad del estado inicial, multiplicada por las probabilidades de transición de un estado a otro.

Sustituyendo los dos términos anteriores en el sumatorio inicial (6) se obtiene la

probabilidad de la secuencia de observación:

$$P(O|\lambda) = \sum_{Q} P(O|Q,\lambda) = \sum_{q1,q2,\ldots qT} \pi \qquad (\ ) \qquad (16)$$

La interpretación del resultado obtenido es la siguiente: Inicialmente en el tiempo *t=1* nos encontramos en el estado q_1 con probabilidad π_{q1} y generamos el símbolo O_1 con probabilidad $b_{q1}(O_1)$. Al avanzar el reloj al instante *t=2* se produce una transición al estado q_2 con probabilidad a_{q1q2} y generamos el símbolo O_2 con probabilidad $b_{q2}(O_2)$.

Este proceso se repite hasta que se produce la última transición del estado q_{T-1} al estado q_T con probabilidad a_{qT-1qT} y generamos el símbolo O_T con probabilidad $b_{qT}(O_T)$.

Sin embargo, una primera aproximación al número de operaciones necesarias para calcular $P(O|\lambda)$ nos da un orden $2TN^T$ operaciones, ya que, para cada *T* se pueden alcanzar N^T posibles secuencias de estados, haciendo que el problema sea intratable incluso para pequeños valores.

Por fortuna, existe un algoritmo distinto del que antes se ha expuesto, que utiliza los cálculos intermedios para realizar posteriores operaciones de forma que se reducen el número de operaciones. Pasando a ser del $O(TN^2)$, el algoritmo consiste en los siguientes pasos:

1. Inicialización

$$\alpha_1(i) = \pi_i b_i(o_i) \qquad 1 \leq i \geq N \qquad (17)$$

En este paso se inicializan las probabilidades hacia delante como la probabilidad conjunta del estado *i* y de la observación o_1.

2. Recursión inductiva

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1}) \qquad (18)$$

Este paso también denominado paso de inducción, muestra cómo es posible alcanzar el estado *j* en el instante *t+1* desde los *N* posibles estados en el instante anterior *t*. Puesto

que $\alpha_t(i)$ es la probabilidad conjunta de observar el evento $o_1, o_2, \ldots, o_t$ y de que el estado en el instante *t* sea *i*, el producto $\alpha_t(i)a_{ij}$ es la probabilidad conjunta de que se observe la secuencia $o_1, o_2, \ldots, o_t$ y de que se alcance el estado $_j$ en el instante *t+1* a partir del estado *i* en el instante *t*.

Sumando este producto para todos los *N* posibles estados de partida en el instante *t*, se obtiene la probabilidad de estar en *j* en el instante *t+1* para todas las secuencias parciales de observación previas.

3. Finalización

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{19}$$

El cálculo final de $P(O|\lambda)$ se obtiene como suma de las probabilidades hacia delante en el último instante posible *T*, es decir, el $\alpha_T(i)$ teniendo en cuenta que por definición

$$\alpha_t(i) = P(o_1 o_2 \ldots o_t, q_t = i|\lambda) \tag{20}$$

Y que, por tanto, $P(O|\lambda)$ es la suma de las $\alpha_T(i)$.

Otro algoritmo semejante al forward es el backward que consiste en lo siguiente:

La probabilidad de observación de una secuencia en el estado *i* y con un determinado modelo es:

$$\beta_t(i) = P(O_{t+1}^t|q_t = i, \lambda) \tag{21}$$

Donde $\beta_t(i)$ es la probabilidad de generar una secuencia de observación parcial O_{t+1}^t(secuencia de observaciones desde *t+1* hasta el final) dados que el HMM está en el estado *i*, podemos obtener de forma inductiva:

1. Inicialización:

$$\beta_t(i) = \frac{1}{N} \qquad 1 \leq i \leq N \tag{22}$$

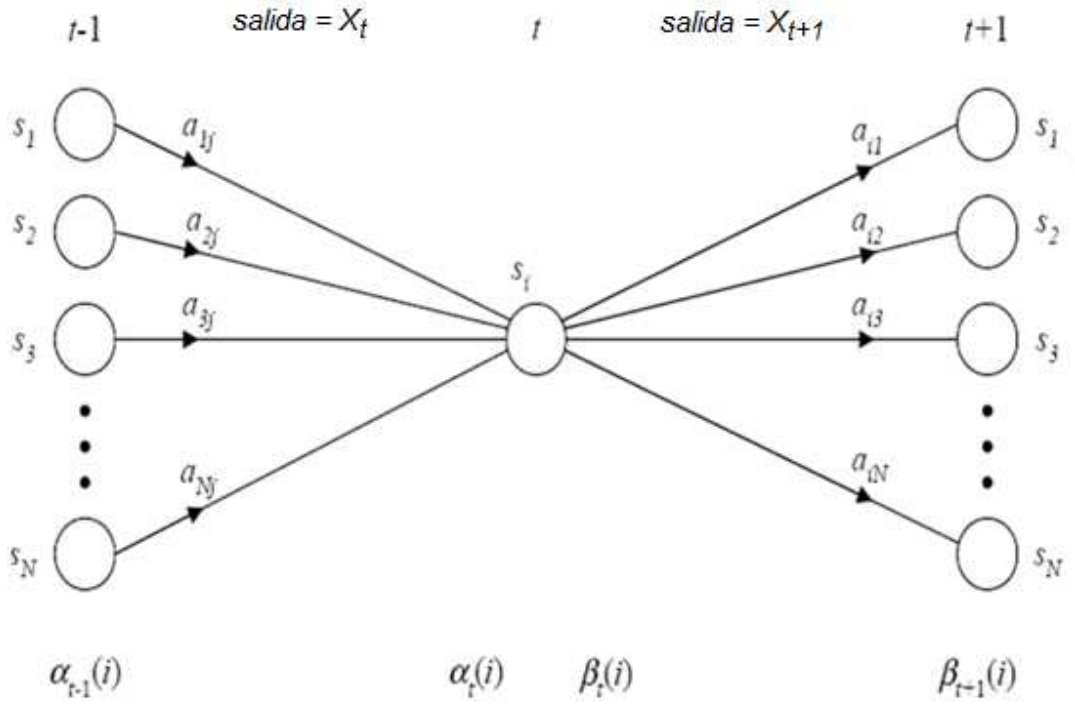
Todos los estados son equiprobables.

2. Inducción:

$$\beta_t(i) = \left[\sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)\right] \qquad t = T-1, \ldots 1;\ 1 \leq i \leq N \qquad (23)$$

La relación entre α y β adyacentes se puede observar mejor en la siguiente figura 1-5. α se calcula recursivamente de izquierda a derecha mientras β se calcula recursivamente de derecha a izquierda.

Figura 1-5:Se representa tanto el algoritmo foward como backward [Huang et al 2001]



2.2.4.3 Decodificación HMM- Algoritmo de Viterbi

Decodificar consiste en encontrar la secuencia de estados dada una secuencia de observación, lo que puede ser deseable en muchas aplicaciones de segmentación y reconocimiento de voz.

A diferencia del problema de evaluación para el que se puede dar una solución exacta, existen diferentes maneras de resolver este problema. Esto se debe a que la definición de secuencia óptima no es única, sino que existen varios criterios de optimización.

Un criterio de optimización podría ser seleccionar aquellos estados que tengan individualmente la probabilidad más alta de ocurrencia. Sin embargo, este método no parece el más acertado ya que no tiene en cuenta la probabilidad de ocurrencia de secuencias de estados. Por ejemplo, la probabilidad de transición entre determinados estados es cero *($a_{ij}=0$)*, este criterio nos podría dar como solución al problema una secuencia de estados que no fuera válida.

Este problema puede resolverse con el algoritmo de Viterbi, que es similar al algoritmo anterior (Forward), con la excepción de que en vez de tomar la suma de valores de probabilidad en los estados anteriores, se toma el máximo de las probabilidades. De esta forma se consigue no sólo dar la secuencia de observación más probable, sino el camino de máxima probabilidad, consiguiendo la secuencia de estados que da una mayor probabilidad.

Antes de definir los pasos del algoritmo de Viterbi vamos a definir las siguientes variables:

$$\delta_t(i) = \max_{q_1,q_2,\dots q_{t-1}} P[q_1 q_2 \dots q_t = i, o_1, o_2 \dots o_t|\lambda] \qquad (24)$$

Donde $\delta_t(i)$ sería el mejor candidato (máxima probabilidad) a lo largo de un camino único, en el instante *t*, que tiene en cuenta la *t* primeras observaciones y termina en el estado *i*. Por inducción tendremos

$$\delta_{t+1}(j) = \max_i \left[\delta_t(i) * a_{ij}\right] b_j(o_{t+1}) \qquad (25)$$

Para recuperar la secuencia de estados debemos seguir el argumento que maximiza la ecuación anterior para cada *t* y para cada *j*. Esto lo haremos a través de una tabla de vuelta atrás $\varphi_t(j)$.

El proceso completo para encontrar la mejor secuencia será:

1. Inicialización

$$\delta_t(i) = \pi_i b_i(o_i) \qquad 1 \leq i \leq N \qquad (26)$$

$$\phi_1(i) = 0$$

Ponemos como los caminos anteriores el *0* para una vez alcanzado el final de este algoritmo al volver por la secuencia más probable no vayamos más para atrás.

2. Inducción

$$\delta_t(j) = \max_{1 \leq j \leq N} \left[\delta_{t-1}(i) a_{ij}\right] b_j(o_t) \qquad 2 \leq t \leq T \qquad 1 \leq j \leq N \qquad (27)$$

Se guarda aquel camino que tiene mayor probabilidad,

$$\phi_t(j) = \arg\max_{1 \leq j \leq N} \left[V_{t-1}(i) a_{ij}\right] b_j(o_t) \qquad 2 \leq t \leq T \qquad 1 \leq j \leq N \qquad (28)$$

3. Finalización

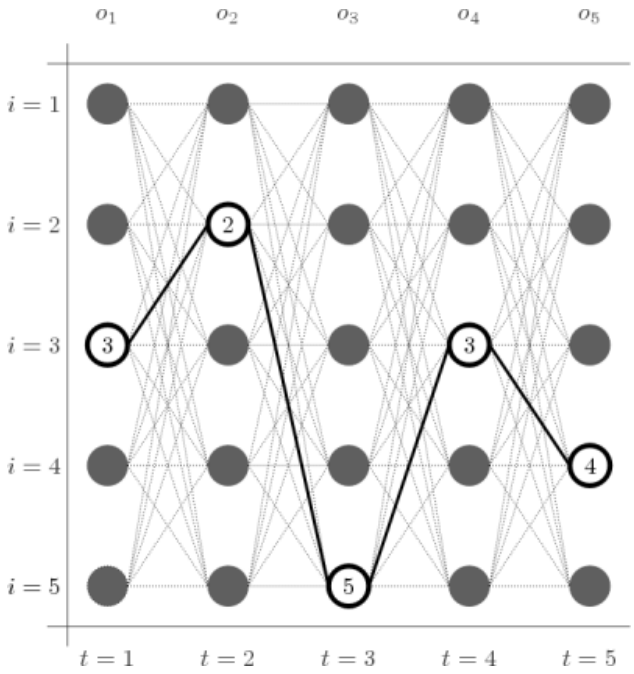
$$P^* = \max_{1 \leq i \leq N} [\delta_T(i)] \qquad (29)$$

$$q_T^* = \operatorname{argmax}_{1 \leq i \leq N} [\delta_T(i)] \qquad (30)$$

4. Seguimiento hacia atrás del camino óptimo (backtracking)

$$q_t^* = \phi_{t+1}(q_{t+1}^*) \qquad t = T-1, T-2 \ldots 1$$

Figura 1-6:Esquema del algoritmo de Viterbi (Tomado de: http://es.wikipedia.org/wiki/Algoritmo_de_Viterbi)



Como se puede observar en la figura 1-6, el algoritmo seguido es muy semejante al de avance hacia delante empleado en la fase de evaluación, y el orden de operaciones también está en torno a $O(TN^2)$.

3.2.4.3 Aprendizaje de HMM-Algoritmo de Baum Welch

Aquí el problema que se tiene es estimar los parámetros del modelo $\lambda (A, B, \Pi)$ de forma que maximicemos $P(O|\lambda)$. Sin embargo, no existe ningún método conocido que permita obtener analíticamente el juego de parámetros que maximice la secuencia de observaciones. Por otro lado, podemos determinar este juego de características de modo que su verosimilitud encuentre un máximo local mediante la utilización de procedimientos iterativos como el del método de Baum-Welch, éste no es más que un algoritmo E-M aplicado a los HMM; o bien mediante la utilización de técnicas de gradiente.

Un parámetro que debemos definir es el $\xi_t(i,j)$, como la probabilidad de encontrarnos en el estado i en el instante t, y en el estado j en el instante $t+1$, para un modelo y una secuencia de observación dados

$$\xi_t(i,j) = P(q_t = i, q_{t+1} = j|O,\lambda) \quad (31)$$

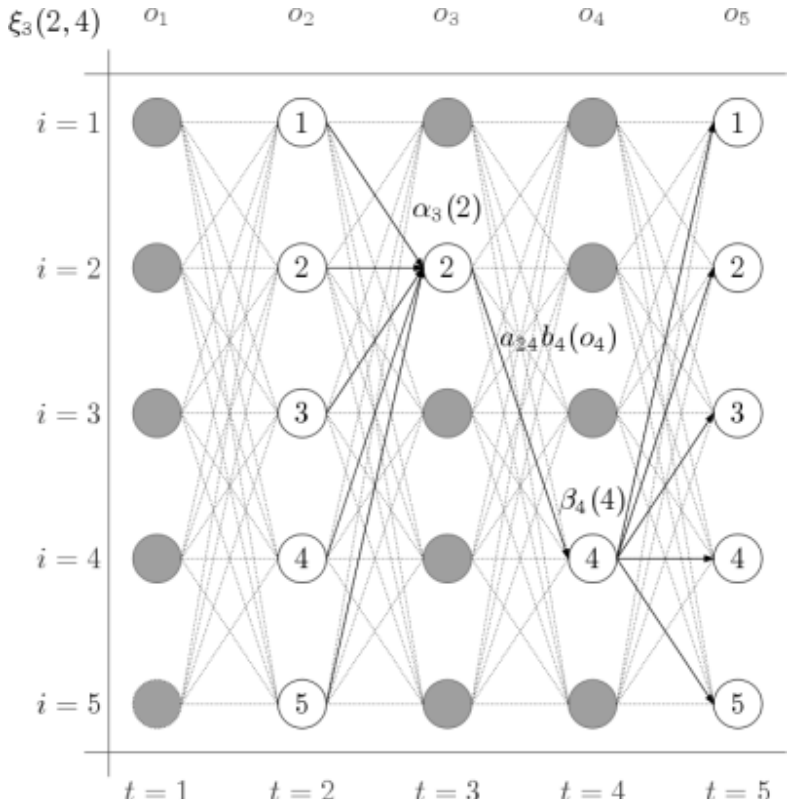
Utilizando las probabilidades de los métodos forward y backward podemos escribir $\xi_t(i,j)$ con la siguiente fórmula:

$$\xi_t(i,j) = \frac{P(q_t = i, q_{t+1} = j|O,\lambda)}{P(O|\lambda)} = \frac{\alpha_t(i).a_{ij}.b_j(o_{t+1})\beta_{t+1}(j)}{P(O|\lambda)} =$$

$$\frac{\alpha_t(i).a_{ij}.b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i).a_{ij}.b_j(o_{t+1})\beta_{t+1}(j)} \quad (32)$$

La figura 1-7 muestra un esquema parcial de los elementos necesarios para el cálculo de $\xi_t(i,j)$

Figura 1-7:Esquema parcial de los elementos necesarios para el cálculo de ξ_t (i,j).(Tomado de: http://es.wikipedia.org/wiki/Algoritmo_de_Baum-Welch)



Suponiendo $\gamma_t(i)$ la probabilidad de encontrarnos en el estado *i* en el instante *t*, para la secuencia de observaciones completa y el modelo dados; por lo tanto, a partir de $\xi_t(i,j)$ podemos calcular $\gamma_t(i)$ con sólo realizar el sumatorio para toda *j*, de la forma:

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \tag{33}$$

Realizando el sumatorio de $\gamma_t(i)$ para todo *t*, obtenemos un resultado que puede ser interpretado como el número esperado de veces (en el tiempo) que estamos en el estado *i* o de manera equivalente, número esperado de transiciones realizadas desde el estado *i* (excluyendo el instante *t=T* del sumatorio). De forma análoga, el sumatorio de $\xi_t(i,j)$ en *t* (desde *t=1* hasta *t=T-1*) puede ser interpretado como el número esperado de transiciones desde el estado *i* al estado *j*.

Con lo anterior podemos usarlo para la reestimación de los parámetros del HMM λ, quedando:

$\pi i'$=número de veces que permanecemos en el estado *i* en el instante *t=1*, $\gamma_1(i)$

Número esperado de transiciones del estado *i* al *j*

Número esperado de transiciones desde el estado *i*

$$a_{ij}^{'} = \frac{Número\ esperado\ de\ transiciones\ del\ estado\ i\ al\ j}{Número\ esperado\ de\ transiciones\ desde\ el\ estado\ i} \tag{34}$$

$$= \frac{\sum_{t-1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T} \gamma_t(i)}$$

$$b_j^{'} = \frac{Número\ esperado\ de\ instantes\ en\ el\ estado\ j\ observando\ el\ simbolo\ v_k}{Número\ esperado\ de\ instantes\ en\ el\ estado\ i}$$

$$= \frac{\sum_{t-1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T} \gamma_t(i)} \tag{35}$$

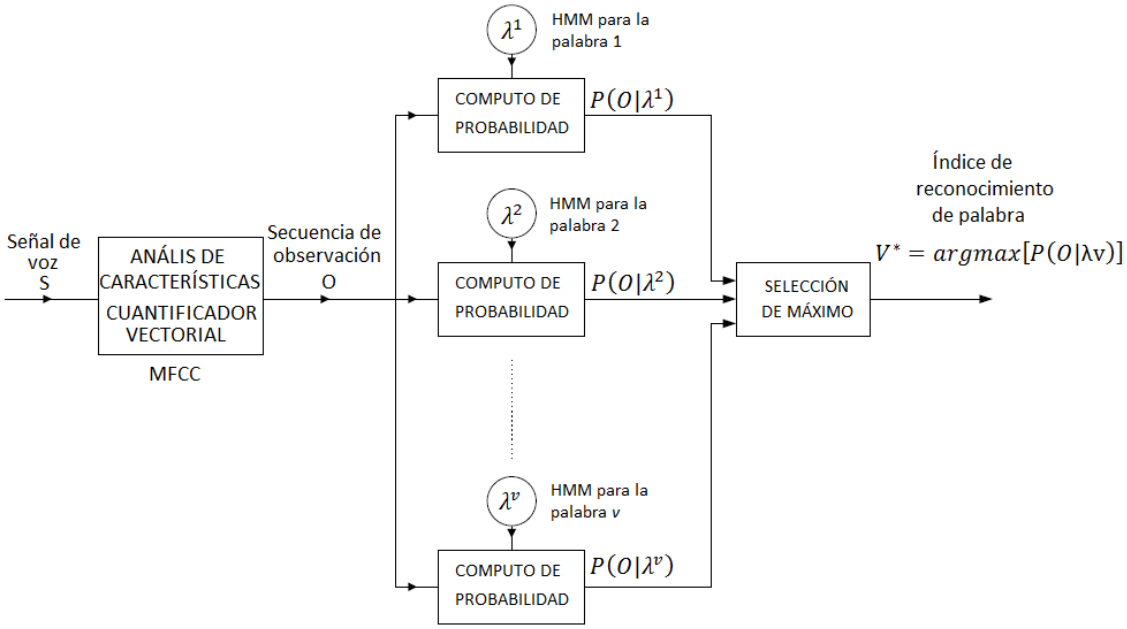
Con estos cálculos obtenemos una re-estimación de los parámetros del modelo obteniendo un nuevo modelo $\lambda'=(A',B', \pi')$. Si el modelo λ definía un punto crítico de la función de máxima verosimilitud en dicho caso tendremos $\lambda'=\lambda$, o bien el nuevo modelo que hace que se cumpla $P(O|\lambda')>P(O|\lambda)$, es decir, se ha mejorado el modelo de las secuencias de observación produciéndose con mayor verosimilitud. Por tanto, mejora la probabilidad de observar una secuencia O a partir de un modelo dado hasta llegar a un límite. Pero el principal inconveniente que tiene es que el método Baum Welch conduce de forma exclusiva a máximos locales. En la mayoría de los casos de interés la función de verosimilitud es compleja y contiene muchos de estos máximos. Los modelos que se manejan son Continuous-Density HMM con modelos de Gaussianas, las expresiones anteriores deberán pasar al caso continuo.

1.2.5 Reconocimiento de Palabras Aisladas

Considérese el uso de los HMMs para construir un reconocedor de palabras aisladas. Supóngase que se tiene un vocabulario de V palabras a ser reconocidas y que cada palabra debe ser modelada por un HMM[3] distinto. Supóngase además, que para cada palabra del vocabulario se tienen un conjunto de K ocurrencias de formación de cada palabra hablada (hablada por uno o más locutores) donde cada aparición de la palabra constituye una secuencia de observación, donde las observaciones son una representación apropiada de las características de la palabra (espectral y / o temporal). Con el fin de hacer el reconocimiento de voz de palabras aisladas, se debe hacer los siguiente procesos:

[3] Una excelente descripción de sistemas de reconocimiento de palabras aisladas de gran vocabulario, basado en unidades de sub-palabra se presenta en IBM TANCORA [39]. Otra buena referencia que compara los efectos de las densidades entre continuas y discretas, utilizando un vocabulario de palabras 60 000 es presenta en[33].

Figura 1-8:Diagrama de bloques de un reconocedor de palabras aisladas HMM.



1. Para cada *V* palabra en el vocabulario, debemos construir un HMMλ^v, es decir, hay que estimar los parámetros del modelo *(A, B,π)* que optimizan la probabilidad de los vectores de observación, del conjunto de entrenamiento para las v palabras.

2. Para cada palabra desconocida que ha de ser reconocida, el procesamiento debe realizarse como muestra la figura. 1-8, es decir, la medición de la secuencia de observación $O = (O_1, O_2 \ldots O_T)$, por medio de un análisis de características de la voz correspondiente a cada palabra; seguido por el cálculo del modelo de probabilidades para todos los modelos posibles, $P(O|\lambda^v)$, $1 \leq v \geq V$; seguido por la selección de la palabra cuyo modelo de probabilidad es el más alto, es decir,

$$V^* = argmax[P(O|\lambda v)]$$

La etapa de cálculo de probabilidad se realiza generalmente en la utilización del algoritmo de Viterbi (es decir, se utiliza la trayectoria de máxima verosimilitud) y se requiere del orden de $V.N^2.T$ cálculos. Para los tamaños de vocabulario modestos, por ejemplo, $V = 100$ palabras, con un modelo de N = estado *5*, y $T = 40$ observaciones sobre la palabra desconocida, un total de 10^5 cómputos se requiere para el reconocimiento (donde cada cálculo es una multiplicación, añadir, y un cálculo de la

densidad de observación, *b(O)*). Es evidente que esta cantidad de cálculo es modesta en comparación con las capacidades de la mayoría de los chips modernos de procesamiento de señales.

1.3 Estado del arte de los HMM en el reconocimiento de voz

1996, Hauenstein .Al comparar el rendimiento en un SRAH híbrido (HMM-NN Hidden Markov Models-Neural Networks –Cadenas Ocultas de Markov y redes neuronales–) utilizando sílabas y fonemas como unidades básicas para el modelo, encuentra que ambos sistemas presentan ventajas que se pueden aprovechar de manera combinada.

1997, Wu et al. Propusieron la integración de información al nivel de sílabas dentro de los reconocedores automáticos del habla para mejorar el rendimiento y aumentar la robustez (Wu, 1998) y (Wu et al., 1997). La razón de error alcanzada fue del 10% para un corpus de voz de dígitos del corpus de OGI (Oregon Graduate Institute). En 1998, se reportan resultados del orden del 6.8% para un corpus de dígitos proveniente de conversaciones telefónicas, haciendo uso de un sistema híbrido fonema-sílaba.

1999, Jones et al. Experimentaron con los modelos ocultos de Markov (HMM - Hidden Markov Models) para obtener las representaciones de las unidades al nivel de sílaba, encontrando que se puede mejorar substancialmente los rendimientos del SRAH en una base de datos de tamaño mediano al compararlos con modelos monofónicos. Logrando un 60% de reconocimiento que lo comparan con un 35% que se obtiene al utilizar monofonemas, dejando en claro que las aplicaciones prácticas deben de conformarse por un sistema híbrido.

1999, Fosler et al. Encontraron que una gran cantidad de fenómenos fonéticos en el habla espontánea son de carácter silábico y presentaron un modelo de pronunciación que utiliza ventanas fonéticas contextuales mayores a las utilizadas en los SRAH basados en fonemas.

2002, Manuele Bicego, Vittorio Murino, Mário A.T. Figueiredo. Una estrategia de poda secuencial para la selección del número de estados en los modelos ocultos de Markov.

Este documento aborda el problema de la selección óptima de la estructura de un modelo oculto de Markov. Se propone un nuevo enfoque, que puede tratar con inconvenientes de los métodos estándar de propósito general, como los basados en el criterio de inferencia bayesianos, es decir, requisitos de cálculo, y la sensibilidad a la inicialización de los procedimientos de formación. La idea básica es la disminución de aprendizaje, en cada sesión de entrenamiento se parte de una situación casi buena, derivada del resultado de la sesión de entrenamiento anterior por la poda el estado menos probable del modelo. Los experimentos con datos reales y sintéticos muestran que el método propuesto es más preciso en encontrar el modelo óptimo, es más eficaz en la precisión de la clasificación, mientras que presenta la reducción de la carga computacional.

2006, José Luis Oropeza Rodríguez. Algoritmos y métodos para el reconocimiento de voz en español mediante silbas. En este articulo se estable que: actualmente el uso de los fonemas tiene implícita varias dificultades debido a que la identificación de las fronteras entre ellos por lo regular es difícil de encontrar en representaciones acústicas de voz. Este trabajo plantea una alternativa a la forma en la que el reconocimiento de voz se ha estado implementando desde hace ya bastante tiempo, analizando la forma en la cual el paradigma de la sílaba responde a tal labor dentro del español. Durante los experimentos realizados fueron examinados para la tarea de segmentación tres elementos esenciales: a) la Función de Energía Total en Corto Tiempo, b) la Función de Energía de Altas Frecuencias Cepstrales (conocida como Energía del parámetro RO), y c) un Sistema Basado en Conocimiento. Tanto el Sistema Basado en Conocimiento y la Función de Energía Total en Corto Tiempo fueron usados en un corpus de dígitos en donde los resultados alcanzados usando sólo la Función de Energía Total en Corto Tiempo, fueron de 90.58%. Cuando se utilizaron los parámetros Función de Energía Total en Tiempo y la Energía del parámetro RO se obtuvo un 94.70% de razón de reconocimiento. Lo cual causa un incremento del 5% con relación al uso de palabras completas en un corpus de voz dependiente de contexto. Por otro lado, cuando se utilizó un corpus de laboratorio del habla continua al usar la Función de Energía Total en Corto Tiempo y el Sistema Basado en Conocimiento, se alcanzó un 78.5% de razón de reconocimiento y un 80.5% de reconocimiento al usar los tres parámetros anteriores. El

modelo del lenguaje utilizado para este caso fue el bigram y se utilizaron Cadenas Ocultas de Markov de densidad continua con tres y cinco estados, con tres mixturas Gaussianas por estado.

2007, Julián D. Arias Londoño, Germán Castellanos Domínguez. diseño simultaneo de una etapa de extracción de características y un clasificador basado en HMM. Presentan una metodología de diseño simultáneo de una etapa de extracción de características y un clasificador basado en modelos ocultos de Markov (HMM), por medio del algoritmo de mínimo error de clasificación (MCE). La extracción de características depende de los estados del modelo y es optimizada utilizando el mismo criterio de ajuste de parámetros del HMM. La metodología es validada en reconocimiento de patologías de voz. Los resultados muestran que el entrenamiento por medio de MCE mejora la eficiencia en comparación con el entrenamiento clásico por máxima verosimilitud. La metodología propuesta disminuye la similitud entre modelos, mejorando el desempeño.

2007, Mark J F Gales, Steve Young. The application of hidden Markov models in speech recognition. Afirman que, los Modelos ocultos de Markov (HMMs) proporcionan un marco simple y eficaz para el modelado de variables en el tiempo secuencias del vector espectrales. En consecuencia, casi todos los grandes sistemas actuales de vocabulario continuos de reconocimiento de voz (LVCSR) están basados en HMMs. Considerando que los principios básicos que subyacen LVCSR basado en HMM son bastante sencillo, las aproximaciones y suposiciones de simplificación que participan en una aplicación directa de estos principios dan lugar a un sistema que tiene poca precisión y sensibilidad inaceptable a los cambios en el entorno operativo. Por lo tanto, la aplicación práctica de los HMM en los sistemas modernos implica una considerable sofisticación.

El objetivo de esta revisión es el primero en presentar la arquitectura básica de un sistema LVCSR basado en HMM y describir las diversas mejoras que se necesitan para lograr el estado del arte de la performance. Estas mejoras incluyen la proyección de función, mejorar los modelos de covarianza, estimación de parámetros discriminativa, la adaptación y la normalización, la compensación de ruido y la combinación de múltiples sistema de paso. La revisión concluye con un estudio de caso de LVCSR para Broadcast News y la transcripción de conversación con el fin de ilustrar las técnicas descritas.

2009, Manuele Bicego, Elzbieta Pekalska, David M.J.Tax, RobertP.W.Duin Clasificación discriminativo basado en componentes para los modelos ocultos de Markov. Partiendo de que los Modelos ocultos de Markov (HMM) se han aplicado con éxito a una amplia gama de problemas de modelado de secuencia. En el contexto de clasificación, uno de los enfoques más simples es la formación de una sola clase HMM. Una secuencia de prueba se asigna a la clase cuyo HMM produce el máximo una probabilidad posterior (MAPA). Este escenario generativa funciona bien cuando los modelos se estiman correctamente. Sin embargo, los resultados pueden llegar a ser malos, cuando se emplean modelos inadecuados, debido a la falta de conocimientos previos, las estimaciones pobres, violado supuestos o datos insuficientes de formación. Para mejorar los resultados en estos casos proponemos combinar las fortalezas descriptivos de HMMs con clasificadores discriminativos. Esto se consigue por los clasificadores basados en las características de formación en un espacio vectorial HMM inducida definido por los componentes específicos de los modelos ocultos de Markov individuales. Presentan cuatro formas principales de la construcción de estos espacios vectoriales y estudian las combinaciones de entrenadores, en este contexto. Por otra parte, motivar y discuten el mérito de su método en comparación con núcleos dinámicos, en particular, para el enfoque del núcleo de Fisher.

2010, M Abushariah, R N Ainon, R Zainuddin, M Elshafei, y O Khalifa en Natural speaker-independent Arabic speech recognition system based on Hidden Markov Models using Sphinx tools En su trabajo de investigación presentan el diseño, implementación y evaluación para el desarrollo de un sistema de alto rendimiento natural, independiente del hablante para el reconocimiento de voz continua para el idioma árabe. Su objetivo es estudiar la utilidad y el éxito de un corpus del habla de nuevo desarrollo, que es fonéticamente rico y equilibrado, presentando un enfoque competitivo para el desarrollo de un sistema ASR en árabe, en comparación con el estado-del-arte para el árabe ASR investigaciones. El árabe se desarrolló como R utiliza principalmente en la Universidad Carnegie Mellon (CMU) herramientas Esfinge junto con las herramientas de HTK Cambridge. Para extraer características a partir de señales de voz, Mel-Frecuencia técnica de coeficientes cepstrales (MFCC) se aplicó la producción de un conjunto de vectores de características. Posteriormente, el sistema utiliza modelos ocultos de Markov de cinco estados (HMM) con tres emisores de estados para el modelado acústico tri-teléfono. La distribución de probabilidad de emisión de los estados fue mejor, utilizando

densidad 16 distribuciones de mezcla de gaussianas continuas. Las distribuciones estatales estaban atados a 500 senons. El modelo de lenguaje contiene uni-gramas, bi-gramas y tri-gramas. El sistema fue entrenado en 7,0 horas de corpus del habla árabe fonéticamente rico y equilibrado y probado en otra hora. Para los hablantes similares pero diferentes sentencias, el sistema obtuvo un reconocimiento de palabras precisión de 92.67% y 93.88% y una tasa de error de palabra (WER) del 11,27% y del 10,07% con y sin signos diacríticos respectivamente. Para los diferentes oradores, sino frases similares, el sistema obtuvo un reconocimiento de palabras precisión de 95.92% y 96.29% y una tasa de error de palabra (WER) del 5.78% y 5.45% con y sin signos diacríticos respectivamente. Considerando que los distintos altavoces y diferentes frases, el sistema obtuvo un reconocimiento de palabras precisión de 89,08% y 90,23% y una tasa de error de palabra (WER) del 15,59% y del 14,44% con y sin signos diacríticos respectivamente.

2011, Aymen, M. ; Abdelaziz, A. ; Halim, S. ; Maaref, H. en Hidden Markov Models for automatic speech recognition. En este articulo se analiza el problema de los Modelos Ocultos de Markov (HMM): la evaluación, la decodificación y el problema de aprendizaje. Explora un enfoque para aumentar la eficacia de los HMM en el campo de reconocimiento de voz. Si bien el modelado de Markov oculto ha mejorado significativamente el rendimiento de los sistemas de reconocimiento de voz actuales, el problema general de reconocimiento de voz independiente del hablante completamente fluido aún está lejos de ser resuelto. Por ejemplo, no existe un sistema que sea capaz de reconocer fiablemente habla conversacional sin restricciones. Además, no existe una buena manera de inferir la estructura de la lengua de un corpus limitado de frases habladas estadísticamente. Por lo tanto, se ofrecer una visión general de la teoría de HMM, discutiendo el papel de los métodos estadísticos, y señalar una serie de cuestiones teóricas y prácticas que merecen atención, y que son necesarias para entender el fin de una mayor investigación, y permitir avance en el campo del reconocimiento de palabras.

1.4 Coeficientes Cepstrales en Frecuncia escala de Mel (MFCC)

De igual manera como lo que sucede en el sistema de percepción del habla, la identificación de los sonidos se hace en el dominio de la frecuencia. para lo cual existen varia técnica como:

- **Técnicas basadas en el espectro de predicción lineal**: los coeficientes de predicción lineal (LPC), los de reflexión (RC) y los LPC-cepstrum (LPCC) son ejemplos de este grupo. Una descripción exhaustiva de estas técnicas puede encontrarse en [40].
- **Técnicas basadas en la transformada de Fourier**: p. ej., los coeficientes mel-cepstrum (MFCC) [41] y las logenergías en bandas filtradas (FFLFBE) [42].
- **Técnicas basadas en la transformada ondicular**: p. ej. [43].

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 HMM y, también, debido a la robustez de ruido superior que ofrece sobre otras técnicas alternativas de extracción de características, como, por ejemplo, LPCC [44]. Flores , J.(2010) estables que los MFCC es el parametrizador mas utilizado por los reconocedores comerciales [45].

Davis y Mermelstein (D&M) introdujeron el término "Mel Frecuency Cepstral Coefficient " en 1980 [46] cuando combinaron filtros triangulares, perceptualmente distribuidos, con la transformada discreta del coseno del logaritmo de las energías de salida de los filtros.

Además establecen que esta es la mejor parametrización para las señales de voz, que se utilizara en un reconocimiento automático del habla.

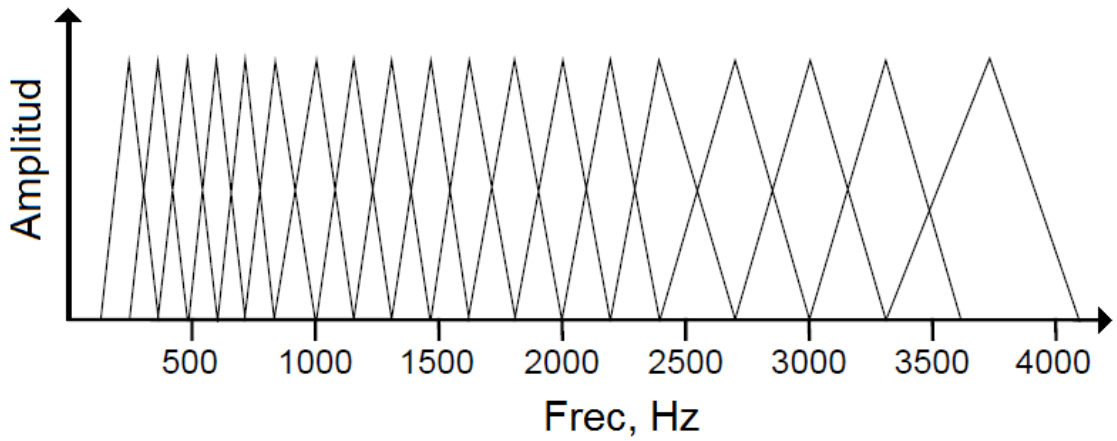
El trabajo previo de Pols [47] introdujo el espaciado entre filtros en un eje de frecuencia que imitaba la sensibilidad lineal logarítmica del sistema auditivo humano (inspirado por los diseños previos del banco de filtros de octava), y D&M extendieron el trabajo en una publicación general para consolidar su utilización. Sin embargo, D&M sólo

proporcionaron una descripción general del algoritmo (una señal era transformada a través de la DFT al dominio de la frecuencia) y cómo el espectro era escalado por un banco de filtros triangulares, distribuidos en un eje de frecuencia log-lineal. A continuación, la energía de salida de cada filtro se comprime logarítmicamente y se trasforma, haciendo uso de la transformada discreta del coseno (DCT), para obtener los coeficientes cepstrales. Sólo proporcionaron una figura del banco de filtros (figura 1-9) junto con la siguiente ecuación:

$$MFCC_i = \sum_{k=1}^{20} X_k cos\left[i\left(k-\frac{1}{2}\right)\frac{\pi}{20}\right], i = 1, 2, \dots, M$$

Donde *M* representa el número de coeficientes cepstrales y *Xk* el logaritmo de la energía de salida del filtro *k-esimo*. Como se observa en la figura 1-9, los puntos finales de cada filtro son definidos por las frecuencias de centro de los filtros adyacentes, el banco está formado por 20 filtros de los cuales, 10 están linealmente espaciados entre 100 y 1000 Hz, 5 logarítmicamente espaciados entre 1 kHz y 2 kHz y otros 5 logarítmicamente espaciados entre 2 kHz y 4 kHz.

Figura 1-9:Banco de filtros utilizado por Davis and Mermelstein en el algoritmo de extracción de características MFCC.



El ancho de banda de los filtros triangulares en MFCC, viene determinado por la distribución de la frecuencia de centro de cada filtro, siendo esta la función de la frecuencia de muestreo y el número de filtros. Es decir, si el número de filtros en el banco de filtros aumenta, el ancho de banda de cada filtro decrece.

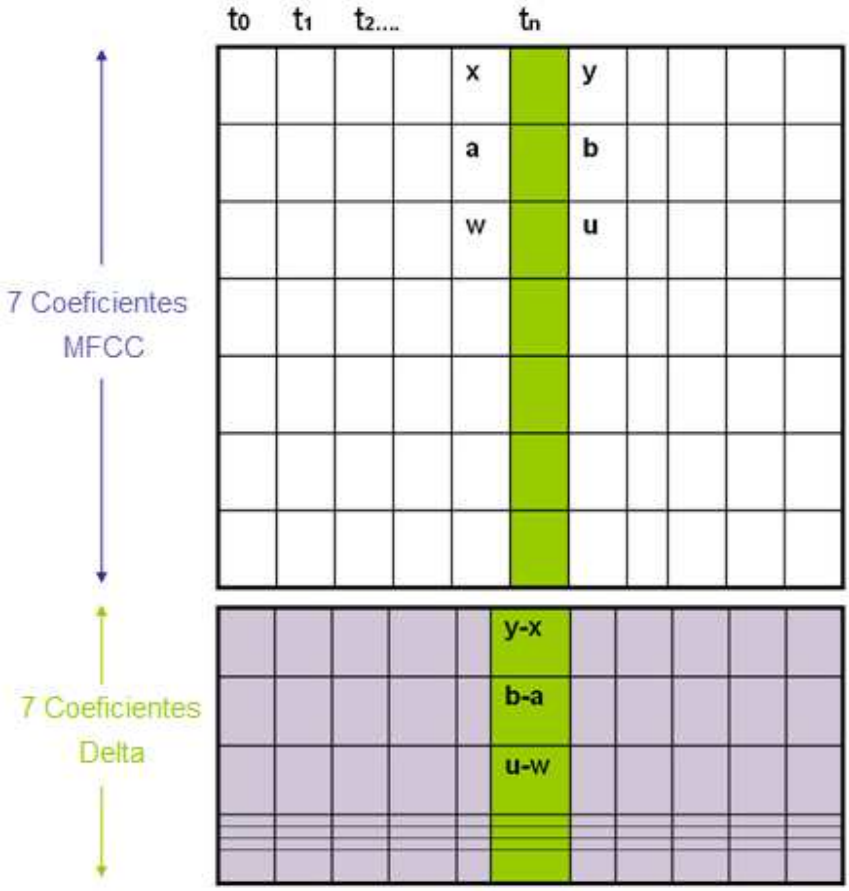
Si bien, las características del banco de filtros en MFCC se derivan del sistema auditivo humano, la descripción original de D & M no explica la elección del número de filtros ni la forma de éstos, el factor de solapamiento entre filtros adyacentes, ni sugiere cómo adaptar el diseño original para experimentos con muestras de voz a velocidades de muestreo diferentes de 10 kHz. La forma en triángulo de cada filtro utilizado en MFCC, aproxima a modelos de la banda de paso natural de las bandas críticas del sistema auditivo humano, aunque la conocida relación entre la frecuencia central y ancho de banda crítico no se utiliza para establecer el ancho de banda del filtro.

Con la descripción general del banco de filtros utilizado en MFCC dada por D&M, los investigadores han modificado el banco de filtros original (ancho de banda) para sus propios experimentos, y será a partir de este algoritmo de extracción de características del que se parta para el estudio de las mejoras en parametrización de la voz, objeto de este proyecto.

Los coeficientes MFCC representan la envolvente espectral de la señal de voz, obteniendo así importantes características identificadoras del habla. En concreto, el primer coeficiente, C_0, indica la energía de la señal y se usa o no dependiendo de la aplicación. Y el segundo coeficiente, C_1, tiene una razonable interpretación como indicador del balance global de energía entre bajas y altas frecuencias.

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 ΔΔMFCC), 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 (figura 1-10). Los coeficientes ΔΔMFCC representan la variación de los coeficientes de velocidad, por lo que son llamados coeficientes de aceleración.

Figura 1-10:Una esquematización de los Delta- Mel-Frequency Cepstral Coefficients donde se representa una posible manera de calcular los coeficientes delta.



Sin embargo, los coeficientes MFCC son difíciles de relacionar con cualquier aspecto cerrado de la producción o percepción del habla. Los detalles espectrales que contienen permiten la discriminación entre sonidos similares, pero su carencia de interpretación los hace altamente vulnerables a condiciones no lineales tales como el ruido o acentos. En particular, los MFCCs dan igual peso a las altas y bajas amplitudes en el espectro logarítmico, cuando es bien conocido que la alta energía domina en la percepción.

1.5 Transformada de wavelet

La transformada de Wavelet permite realizar un análisis de múltiple resolución (MRA), lo cual analiza la señal con resolución diferente a diferentes frecuencias.

Se diseña para producir alta resolución en el tiempo y baja resolución en frecuencia para señales de alta frecuencia, y baja resolución en el tiempo y alta resolución en frecuencia para señales de baja frecuencia. En las ecuaciones 39 y 40 se indica la forma como se obtiene la transformada de wavelet continua (CWT) de una función f(t).´

$$C(escala, posición) = \int_{-\infty}^{\infty} f(t).\, \psi(escala, posición, t)dt \qquad (39)$$

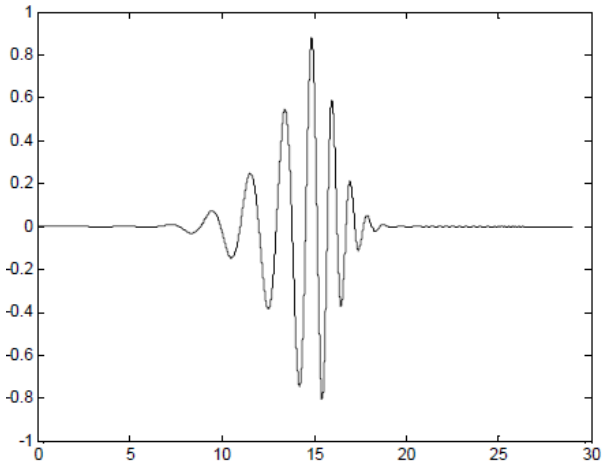
$$CWT\ (s\ ,\tau) = \frac{1}{\sqrt{|S|}} \int_{-\infty}^{\infty} f(t).\, \psi\left(\frac{t-\ \tau}{s}\right) dt \qquad (40)$$

Donde y es la función wavelet (llamada wavelet madre), la cual se debe escalar (*s*) y desplazar (*t*) sobre el eje del tiempo. El factor en el denominador (la raíz cuadrada de *s*) es utilizado como un factor de normalización de la energía.

El parámetro de escalamiento (*s*) permite comprimir o expandir la función wavelet *y* y en la transformada es utilizada en el denominador (*y* (*t/s*)). Este parámetro indica el grado de resolución con que se analiza la señal. Un valor alto de este factor (|s| *>1*) corresponde a una vista global de la señal (expansión de la wavelet) mientras que un factor de escala bajo (|s| *<1*) corresponde a ver detalles de la señal (se comprime la wavelet). El factor de posición (*t*) permite desplazar la función wavelet en el eje del tiempo(*y*(*t*)), hasta el intervalo de tiempo que se encuentre definida la función *f(t)*. El resultado que se obtiene cuando se aplica la transformada de wavelet son los coeficientes *C(s, t)* , que son función de la escala y la posición [34-36].

La wavelet utilizada en este proyecto es la Daubechies 15 (db15), la cual no es simétrica y es mostrada en la figura 1-11 [37,38].

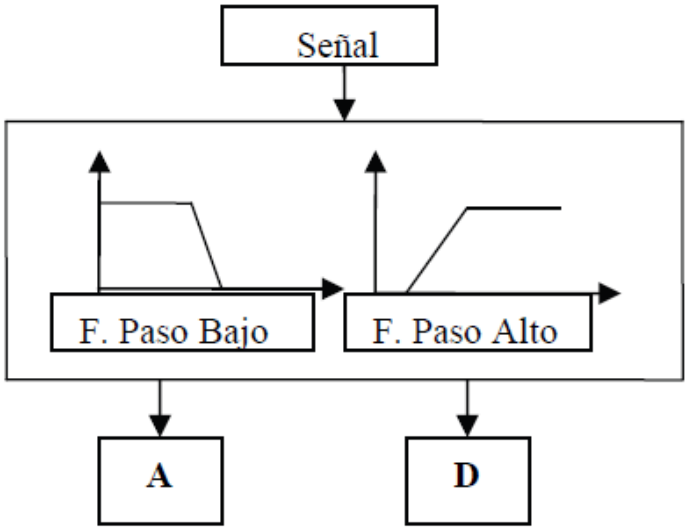
Figura 1-11:Función Wavelet db15.



1.6 Filtrado de la señal de voz

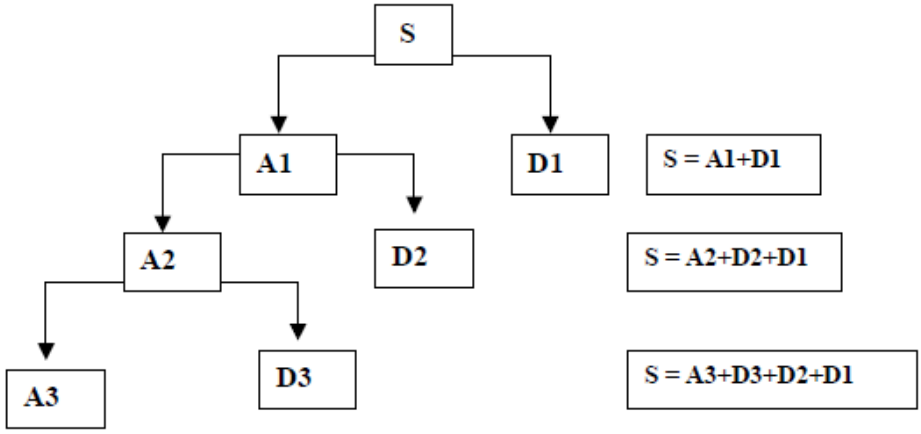
El análisis de las señales con la transformada de wavelet es equivalente a un proceso de filtrado, donde se realiza una división de los coeficientes, obteniéndose los coeficientes de aproximación (*A*) y detalle (*D*). La aproximación son los valores altos de la escala, correspondiente a las componentes de baja frecuencia de la señal, por lo tanto, está asociada a la función de escalamiento que se determina con un filtro pasa bajo. Los detalles son los valores bajos de la escala correspondientes a las componentes de alta frecuencia, y está asociada a la función wavelet que se determina como un filtro pasa alto. En la figura 1-12 se muestra un esquema del proceso de filtrado, donde la señal a procesar es pasada por los filtros paso bajo y pasa alto, los cuales son filtros complementarios y se producen dos señales [35, 37, 39, 45].

Figura 1-12:Esquema del proceso de filtrado.



El proceso de filtrado o descomposición se repite, para descomponer la señal en N niveles, cada nivel con una resolución más baja. El número de veces que es filtrada la señal viene determinada por el nivel de descomposición. En la figura 1-13 se muestra el árbol de filtrado de wavelet con tres niveles.

Figura 1-13: Proceso de descomposición con tres niveles.



Para reducir el ruido de la señal de voz se implementó un programa donde se eliminan las componentes de la señal que están por debajo de un determinado umbral (THR), el

cual es calculado según la teoría de Donoho [46-48] mediante la expresión indicada en la ecuación 41.

$$THR = \left(\sqrt{2\ln(n)}\right) \text{ x } s \qquad (41)$$

Donde:

n = longitud de la señal a analizar

s = estimado del nivel de ruido de la señal, obtenido

con la ecuación 42

$$s = \frac{mediana\left(modulo(c)\right)}{0.6745} \qquad 42$$

c = coeficientes de los detalles en el nivel uno.

El tipo de umbral a aplicar a la señal para reducir el nivel de ruido puede ser: duro (Hard) o suave (Soft). Cuando se aplica el umbral duro, si el valor absoluto del coeficiente es mayor que el umbral calculado, se mantiene el coeficiente, y en el caso contrario se iguala a cero. En la ecuación 43 se muestra la expresión utilizada para aplicar este tipo de umbral, donde *Cm(i,j)* representan los coeficientes modificados.

Cuando se aplica el umbral suave, si el valor absoluto del coeficiente es mayor que el umbral seleccionado se modifica el coeficiente, restando el umbral a su valor absoluto, en caso contrario se iguala el coeficiente a cero. En la ecuación 44 se muestra la expresión utilizada cuando se aplica este tipo de umbral. En la figura 1-14 se muestra el resultado cuando no se modifican los coeficientes, por lo tanto *Cm(i,j)* es igual a *C(i,j)*.

En las figuras 1-15a y 1-15b se muestran los resultados de las modificaciones de los coeficientes cuando se aplica el umbral duro y el suave [45, 47, 50].

$$Cm(i,j) = \begin{Bmatrix} 0 & \text{si } |C(i,j)| < thr \\ C(i,j) & \text{si } |C(i,j)| \geq thr \end{Bmatrix} \quad (43)$$

$$Cm(i,j) = \begin{Bmatrix} 0 & \text{si } |C(i,j)| < thr \\ \text{sgn}(\text{C(i,j)})|C(i,j) - thr| & \text{si } |C(i,j)| \geq thr \end{Bmatrix} \quad (44)$$

Para determinar la relación señal a ruido (*snr*) de la señal se utilizó la ecuación 45 [38].

$$snr = 10 * \log_{10}\left[\frac{\sum_{n=0}^{L-1}(x(n))^2}{\sum_{n=0}^{L-1}\left(\hat{x}(n) - x(n)\right)^2}\right] \quad (45)$$

Donde:

$\hat{x}(n)$ = señal filtrada.

$x(n)$ = señal con ruido.

snr = relación señal a ruido en decibeles

L = número de muestras de la señal

Figura 1-14: Señal original.

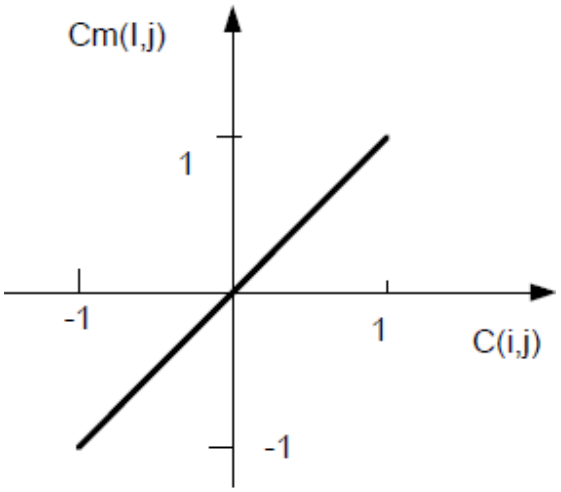
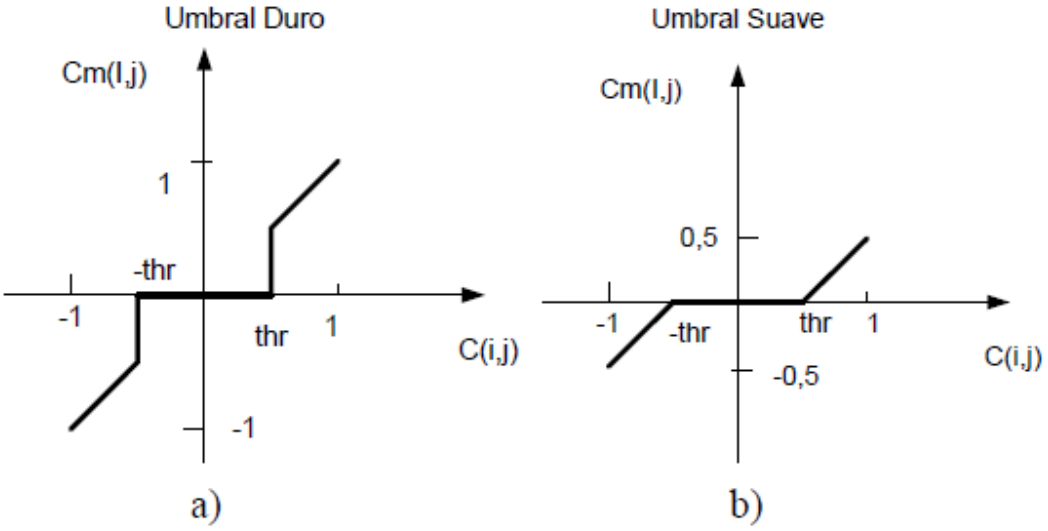


Figura 1-15: a) Umbral duro. b) Umbral suave.



2. Marco Experimental

2.1 Base de Datos

Un requerimiento para realizar este tipo de proyectos, es contar con una base de datos de audio, con sus registros en español.

Además las palabras deben de ser las apropiadas, para este proyecto: ***silla***, ***adelante***, ***atrás***, ***derecha***, ***izquierda*** y ***pare***. También se debe cumplir condiciones en un ambiente con ruido. Por no existir este tipo de base de datos, de forma académica o comercial; es una tarea de este proyecto crear dos bases de datos.

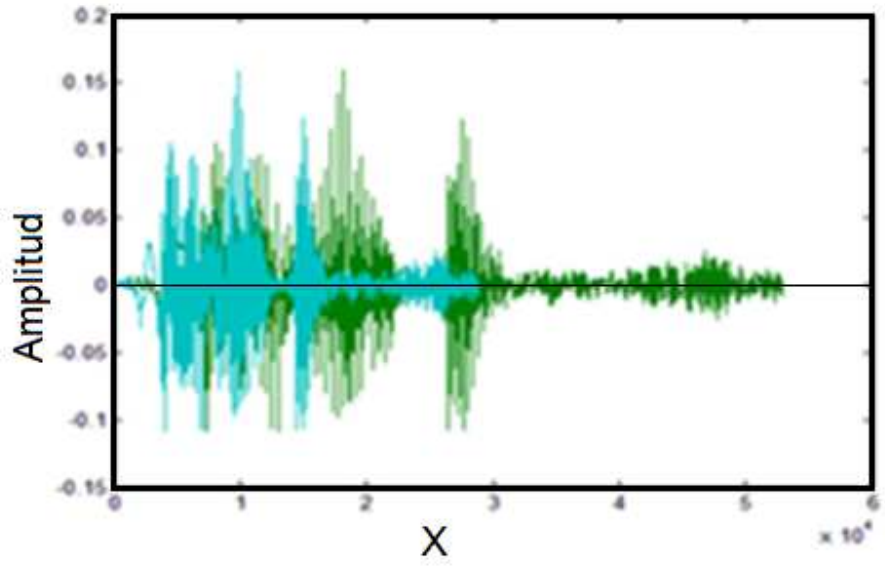
2.1.1 Base de Datos 1 (BD1)

Una primera etapa, es la construcción de una base de datos propia; que tiene las siguientes características: velocidad de muestreo de 44100Hz, a 16 bit, el tiempo de grabación para cada palabra está fijado de forma constante en 1.2 S, en un formato tipo WAV; entregando una señal normalizada, para las palabras: ***silla***, ***adelante***, ***atrás***, ***derecha***, ***izquierda*** y ***pare***. Cada clase (palabra) se ha almacenado en cien ocasiones, para obtener una base de datos de 600 palabras. Para realizar esta tarea; se dividió en subgrupos de 10 elementos de cada clase. Donde en cada procedimiento de captura, la señal está bajo condiciones controladas de ruido, con distancias entre el locutor y el elemento de captura variando en el rango de 0,2 a 0,5 m, además se hace un desplazamiento angular de ± 30°.

Todas las capturas fueron realizadas a un hombre mayor de 50 años. El recinto donde se realiza la grabación está ubicado en un sector residencial; con bajo tráfico vehicular. El elemento de captura; es el micrófono (estéreo) de la tarjeta de sonido incorporada en un PC portátil y el software utilizado, son los comandos del Matlab propios de esta aplicación.

La figura 2-1, muestra la captura de la palabra ***adelante***; se observa una relación entre la amplitud y el número de puntos utilizados para su representación. Es de anotar, que esta gráfica muestra dos señales, por ser una captura en estéreo.

Figura 2-1: Representación de la señal para la palabra **adelante**.



2.1.2 Base de Datos 2 (BD2)

En la siguiente etapa, se ha construido de una segunda base de datos (BD2); que tiene las mismas características técnicas de BD1, pero se ha cambiado la forma de captura y los locutores.

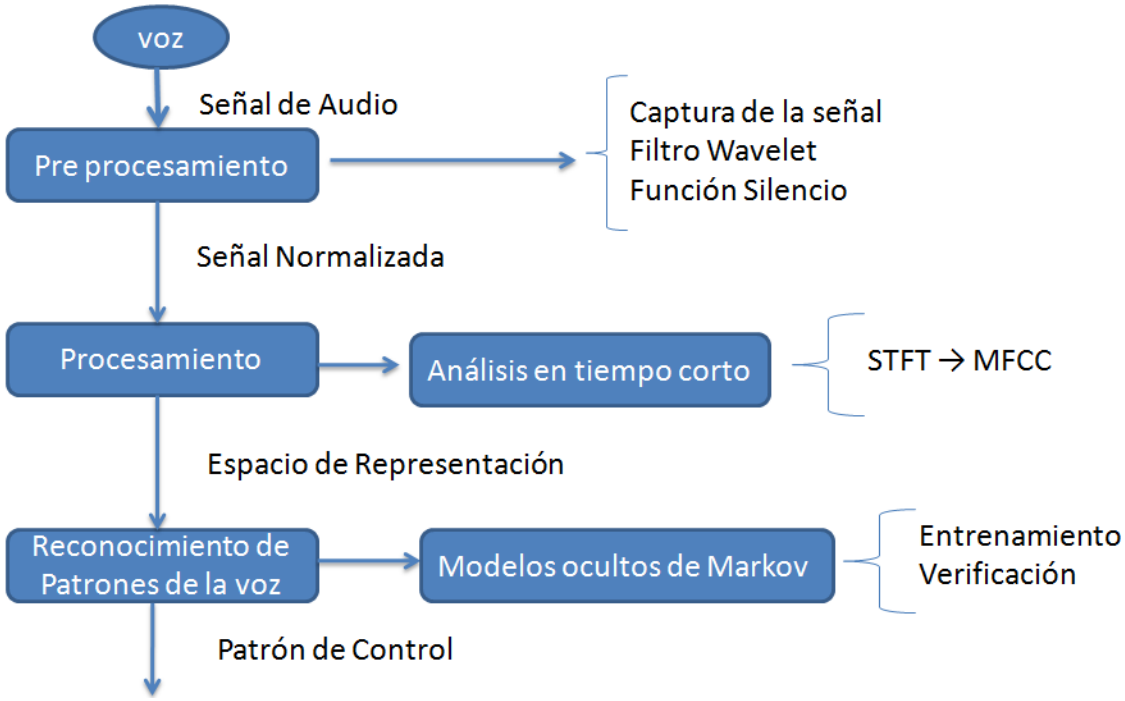
Para realizar esta tarea; el procedimiento se ha dividido en cuatro escenarios; en el primer escenario: cada procedimiento de captura de la señal esta bajo condiciones controladas de ruido, con distancias entre el locutor y el elemento de captura variando en el rango de 0,4 a 0,5 m. En el segundo escenario: cada procedimiento de captura de la señal está bajo condiciones controladas de ruido, con distancias entre el locutor y el elemento de captura variando en el rango de 0,6 a 0,8 m. En el tercer escenario la captura de la señal, está sometida a un ruido vehicular permanente, con distancias entre el locutor y el elemento de captura variando en el rango de 0,4 a 0,5 m. En el cuarto escenario, la captura de la señal está sometida a un ruido vehicular permanente, con distancias entre el locutor y el elemento de captura variando en el rango de 0,6 a 0,8 m.

Las capturas fueron realizadas a dos hombres mayores de 50 años, dos mujeres entre los 40 y 42 años y una mujer mayor de 60 años, que presenta un procedimiento quirúrgico en su garganta.

2.2 Metodología propuesta

La metodología consta de cuatro tres: 1.Preprocesamiento, 2. Procesamiento, 3. Reconocimiento de patrones de voz. Estas etapas son ilustradas en la Figura 2-2.

Figura 2-2: Diagrama a bloques del el procesamiento de la señal de voz para su identificación.



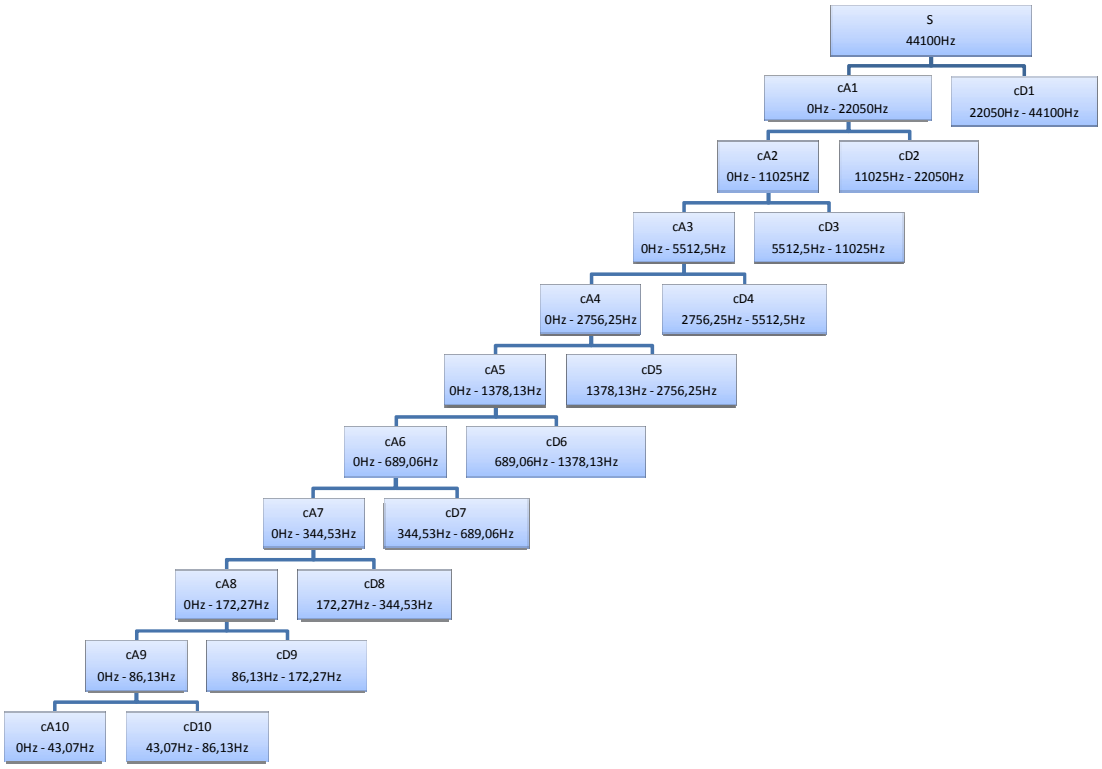
2.2.1 Preprocesamiento

Se convierte el archivo de audio en una matriz (*Y*) de 2 columnas; por ser una señal "estéreo" y 52920 filas, de muestras que forman la señal. El siguiente paso consiste en realizar una ponderación de los datos, que permite generar un vector columna de 52920 filas.

En el paso siguiente se realiza un filtrado de la sobre la señal de origen (*S*), basado en la transformada wavelet.

En este proyecto la señal *S* tiene una frecuencia de 44100Hz; en una primera etapa se divide la señal en dos partes, pasándola por un filtro pasaaltos y un filtro pasabajos, éstos deben cumplir con la condición de admisibilidad [1]. De esta manera, se obtienen dos diferentes versiones de la señal origen (*S*): una parte de la señal que corresponde al rango de aproximación (*cA1*) para los valores de: 0Hz - 22050Hz, y la otra en el rango de detalle (*cD1*) para los valores de: 22050Hz - 44100Hz. Luego se toma la parte de baja frecuencia (*cA1*) de la señal origen (*S*), para obtener de nuevo dos rangos: (*cA2*) con los valores de: 0Hz - 11025HZ y (*cD2*) con los valores de: 11025Hz - 22050Hz, este procedimiento continúa hasta obtener los 10 niveles que se proponen en esta metodología, éstos son ilustrados en la figura 2-3.

Figura 2-3: Muestra el árbol con los 10 niveles de descomposición de la señal origen (S).



Una literatura muy amplia presenta la wavelet Daubechies 16 (DB16); como una herramienta excelente en el filtraje de las señales de audio.

La familia de wavelets de Daubechies [2] está gobernada por un conjunto de *N* (entero par) cocientes *pk : k = 0; 1; . . . , N-1*. Para cada *N ϵ N* se tendrán la wavelet y la función de escala que define el orden *N*, este caso orden 16. La tabla 2-1 presenta las contantes para las DB4, DB6, DB8, DB10, DB12 Y DB16, que será la utilizada en este proceso de filtrado.

Tabla 2-1: Coeficientes de la escala Daubechies.

D4	D6	D8	D10	D12	D16
0.482963	0.332671	0.230378	0,160102	0,111541	0,0544158
0.856516	0.806892	0.714847	0,603829	0,494624	0,312872
0.224144	0.459878	0.630881	0,724309	0,751134	0,675631
-0.12941	-0.135011	-0.0279838	0,138428	0,31525	0,585355
	-0.0854423	-0.187035	-0,242295	-0,226265	-0,0158291
	0.0352263	0.0308414	-0,03224	-0,129767	-0,284016
		0.032883	0,0775715	0,0975016	0,000472485
		-0.0105974	-0,00624149	0,0275229	0,128747
			-0,0125808	-0,031582	-0,0173693
			0,00333573	0,00055384	-0,0440883
				0,00477726	0,013981
				-0,0010773	0,00874609
					-0,00487035
					0,000039174
					0,000675449
					-0,00011747

Reconstrucción wavelet; luego se realiza el proceso inverso; pasar de la transformada wavelet a la señal original.

Los pasos realizados son los siguientes:

- Primero hacer upsample tanto en la aproximación como en los detalles. Esto equivale a poner un cero entre cada dos muestras de la señal.
- Aplicar la operación inversa o de síntesis para la aproximación y para los detalles.
- Sumar la aproximación reconstruida y los detalles reconstruidos.

Con propiedad, la siguiente ecuación muestra los coeficientes de detalle que se utilizaron para la reconstrucción.

$$S = cD3 + cD4 + cD5 + cD6$$

2.2.2 Función de silencio

Para mejorar la respuesta del caracterizador MFCC y establecer un elemento diferenciador; se ha creado una subrutina para reducir el nivel de ruido y reducir el tamaño de la señal a caracterizar; lo cual llevará a una reducción en los tiempos de cómputo, tanto en el entrenamiento, como en la verificación.

Se comienza por hacer un cálculo de la energía de la señal de audio. Para esto se calcula el valor absoluto de la señal y se obtiene el pico máximo. Se divide la señal de audio por el valor antes obtenido, para independizar la forma de onda respecto de la intensidad de la señal.

La señal normalizada se eleva al cuadrado y se divide por el número de muestras de la señal con lo que se obtiene la energía promedio de la señal.

Luego, lo que se hará es dividir la señal normalizada en ventanas de un número determinado de muestras, calcular la energía de ese trozo de señal y si esa energía es mayor a un porcentaje (tomado como umbral de decisión) de la energía promedio de la señal completa, entonces dicha ventana se conserva. En caso contrario, la ventana se desecha ya que se interpreta que al no llegar su energía al umbral establecido, entonces esa ventana corresponde con un intervalo de silencio de la señal de audio. El umbral de decisión corresponde al 2% de la energía promedio de la señal entera. Este valor se definió después de varias pruebas observando cuales eran los resultados de eliminación de silencios. Las ventanas se eligieron de 400 muestras que a la frecuencia de muestreo (44100Hz) corresponde a aproximadamente 10ms por lo que no existe la posibilidad de que en el intento de eliminar silencio, se elimine algún fonema de audio (cuya duración ronda entre los 10 y 20ms).

Esta subrutina se configura con: longitudes de 9000 punto y un umbral de energía de 0.05.

Algoritmo 1: Silencio

```
Requiere: Señal: y
Requiere: Umbral: THRES=0.05
//función que elimina los períodos de silencio de la señal.

len = length( y ); // Longitud de la señal
d = max( abs(y) ); //máximo absoluto de la señal
y = y / d;  // normalización de la señal
E = sum( y.* y) / len ; // Energía promedio de la señal

for i = 1 : 400 : len – 400 // cada 10 ms
      seg = y( i : i + 399 );  // ventanas de 10ms
      e= sum ( seg.*seg)/400; // Promedio de energía de cada segmento.

      if (e > THRES*E ) // Si el promedio de energía por segmento es
      //mayor al promedio energético de la señal multiplicado por el
      //umbral

            s = [s, seg (1 : end) ] //se almacena en s o de lo contrario
            //se elimina.

      end if
end for
Regresar s
```

2.2.3 Procesamiento

En esta etapa se calcula la STFT, hallada mediante un ventaneo de Hamming de 50ms con un traslapamiento de 25 ms entre ventanas. A la representación obtenida, se le aplica un banco de filtros distribuidos en la escala de frecuencias Mel, que para este caso son 20, luego se calcula la energía de cada uno de los filtros y finalmente se aplica el logaritmo y la Transformada Coseno (DCT), obteniendo así los coeficientes MFCC y la derivada de los mismos, generando la matriz de características de cada señal, 20 coeficientes MFCC y 20 derivadas.

La figura 2-4, representa el diagrama a bloques del proceso de extracción de coeficientes ceptrales en la escala de Mel.

El número de columnas de la matriz de resultado, está determinada por el número de ventanas y el traslape entre ellas. la siguiente ecuación plante una aproximación a este cálculo.

$$número\ de\ filas = \frac{tiempo\ de\ la\ señal - tiempo\ de\ la\ ventana}{tiempo\ de\ traslapamiento} \pm 1 \quad (46)$$

$$número\ de\ filas = 46 = \frac{1200mS - 50mS}{25mS} \pm 1$$

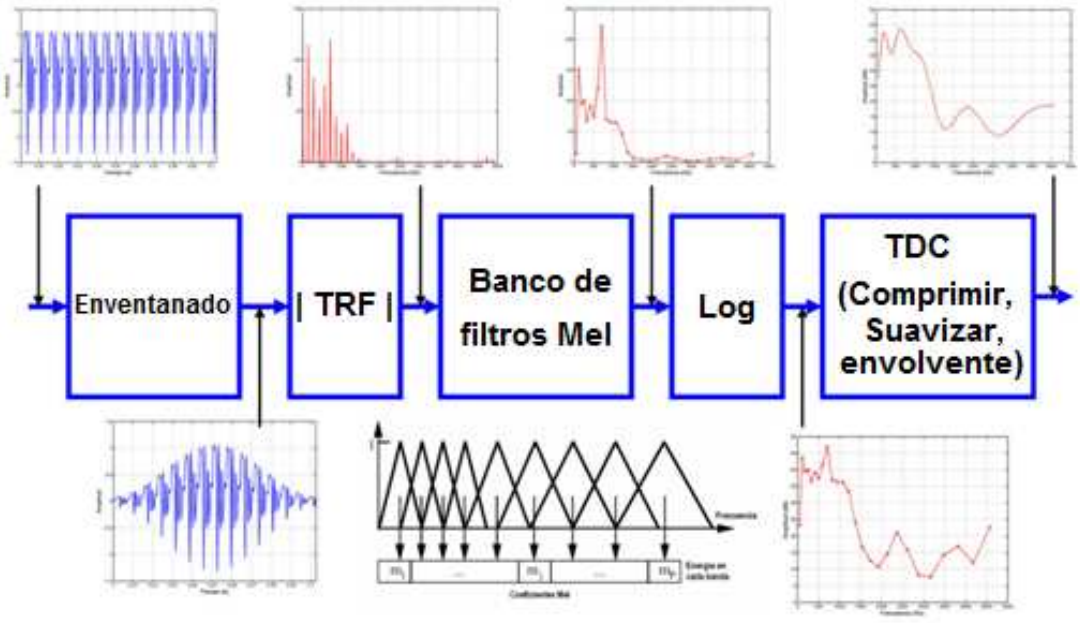
Por causa de la función silencio el tamaño de la señal se ha reducido.

$$número\ de\ filas = 7 = \frac{230mS - 50mS}{25mS} \pm 1$$

El número de columnas; está determinado por el número de coeficientes elegidos, en este proyecto es igual a 20, el valor se duplica por que se toma la derivada. este valor será igual a 40.

En la ejecución de este proceso se obtiene por palabra una matriz de 46X40; en el caso de no utilizar la función de silencio, y una matriz de 7X40, cuando se utiliza la función.

Figura 2-4: Proceso de extracción de los coeficientes MFCC.



2.2.4 Reconocimiento

Desde el proceso de caracterización, se requiere una estructura apropiada para su almacenamiento y manipulación de los datos en esta etapa; la figura 2-5 representa esta estructura.

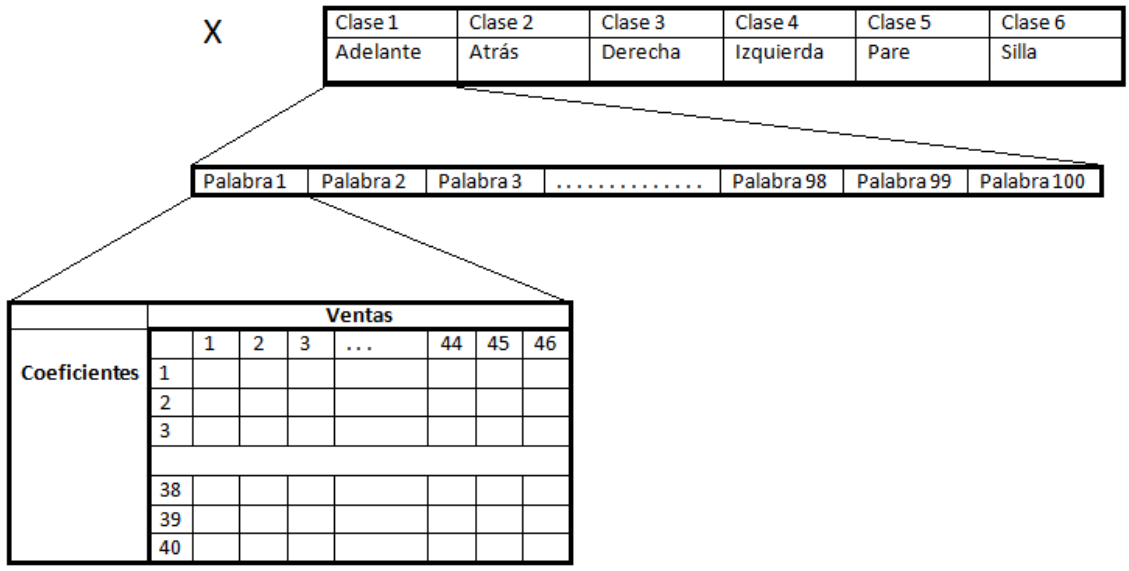
Para este proceso se utiliza una validación cruzada; tomando el 70% para realizar el entrenamiento y el 30% para realizar la validación. A continuación se realiza la normalización, con el proceso de z-score. La siguiente etapa permite realizar las configuraciones iniciales, para generar HMM con mezclas de Gaussianas, éstas son: 2 estados, y el número de mezclas Gaussianas que variará desde 1 hasta 50; la densidad inicial, la definición de la matriz de transición, definir los pesos de las Gaussianas, capturar la cantidad de características por cada palabra, se inicializa de forma aleatoria, se establece el máximo valor de iteraciones para el algoritmo EM y finalmente se elige el valor para el criterio de convergencia del EM.

El siguiente proceso, es generar el modelo para cada una de las clases con los datos almacenados: los datos son organizados de forma matricial, de una manera apropiada

que permita la inicializar las mezclas de Gaussianas, con el algoritmo de K-means. Se procede a realizar un acondicionamiento de los datos. Partiendo de un vector columna para llega a una matriz *m* x *n*. Ahora con los datos ordenados se entrena a cada uno de los modelos con el algoritmo EM; generando un modelo para cada una de las clases. En este proyecto se generan seis tipos de modelos.

La siguiente etapa es el logaritmo de evaluación, para hallar la probabilidad de que una palabra (cualquiera); comparada con cada uno de los modelos, y a través de un sistema para toma de decisión, se define la pertenencia o no a una clase determinada.

Figura 2-5: Estructura de la celda X.



2.3 Experimentos

Para la comprobación de esta metodología se realizaron seis experimentos o procesos que varían los procesos de caracterización y filtrado.

2.3.1 Proceso 1

Para este procedimiento se utilizó la base de datos ***BD1***. Se implementa una metodología con una caracterización basada en los coeficientes cepstrales en la escala de Mel (**MFCC**); lo que entrega una celda ***X***. La siguiente etapa consiste en un entrenamiento, clasificación y verificación basado en los modelos ocultos de Markov (HMM), con variaciones de 1 hasta 50 de mezclas Gaussianas. Para realizar el entrenamiento se toma el 70% de los datos 70 palabras por clase, para realizar al verificación se toma el 30% de los datos 30 palabras por clase. Este ejercicio se realiza 20 veces, lo que determina un total de 1200 pruebas de cada clase de palabra y 12000 pruebas por proceso. En el momento de iniciar los procesos de entrenamiento y verificación, se inicia un temporizador; que permite determinar el tiempo empleado en el proceso para crear otro indicador de comparación, para definir la eficiencia del proceso.

2.3.2 Proceso 2

Para este procedimiento se utilizó la base de datos ***BD1***. Se realiza un proceso de filtrado utilizando la transformada Wavelet Daubechies 16 (DB16) a 10 niveles, luego se realizan los mismos pasos realizados en el proceso 1.

2.3.3 Proceso 3

Para este procedimiento se utilizó la base de datos ***BD1***. Se realiza un proceso de filtrado utilizando la transformada Wavelet Daubechies 16 (DB16) a 10 niveles, en una siguiente etapa se realiza el proceso de la función silencio, luego se realizan los mismos pasos realizados en el proceso 1.

2.3.4 Proceso 4

Para este procedimiento se utilizó la base de datos ***BD2***. Se implementa una metodología con una caracterización basada en los coeficientes cepstrales en la escala de Mel (**MFCC**); lo que entrega una celda ***X***. La siguiente etapa consiste en un entrenamiento, clasificación y verificación basado en los modelos ocultos de Markov (HMM), con variaciones de 1 hasta 50 de mezclas Gaussianas. Para realizar el entrenamiento se toma el 70% de los datos 70 palabras por clase, para realizar la verificación se toma el 30% de los datos 30 palabras por clase. Este ejercicio se realiza 20 veces, lo que determina un total de 1200 pruebas de cada clase de palabra y 12000 pruebas por proceso. En el momento de iniciar los procesos de entrenamiento y verificación, se inicia un temporizador; que permite determinar el tiempo empleado en el proceso para crear otro indicador de comparación, para definir la eficiencia del proceso.

2.3.5 Proceso 5

Para este procedimiento se utilizó la base de datos ***BD2***. Se realiza un proceso de filtrado utilizando la transformada Wavelet Daubechies 16 (DB16) a 10 niveles, luego se realizan los mismos pasos realizados en el proceso 4.

2.3.6 Proceso 6

Para este procedimiento se utilizó la base de datos ***BD2***. Se realiza un proceso de filtrado utilizando la transformada Wavelet Daubechies 16 (DB16) a 10 niveles, en una siguiente etapa se realiza el proceso de la función silencio, luego se realizan los mismos pasos realizados en el proceso 4.

3. Resultados

Este capítulo presenta los resultados que se obtienen al cumplir con los procedimientos expuestos en el Capítulo 2.

3.1 Resultados para el procedimiento 1 con BD1

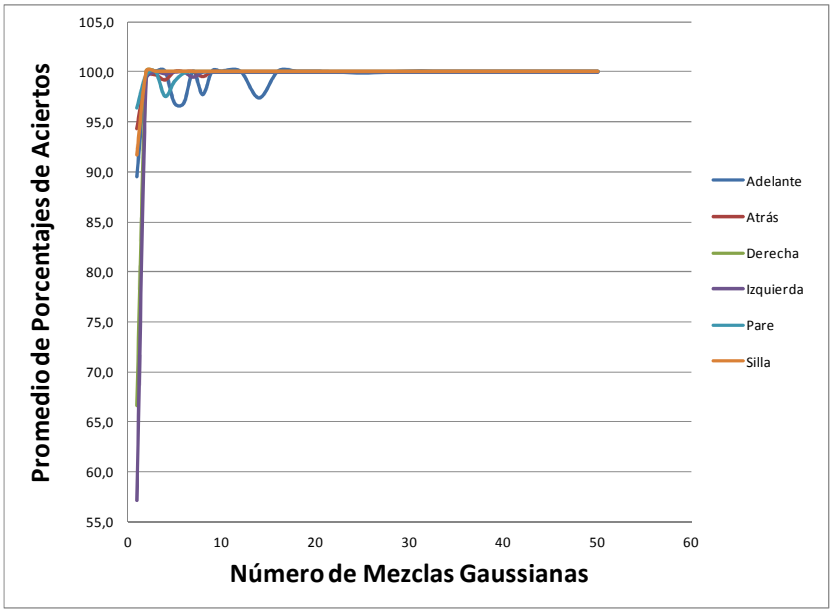
Inicialmente se utiliza una metodología basada en los algoritmos de HMM; con variaciones de mezclas Gaussianas desde 1 hasta 50 y utilizando la base de datos BD1, sin la utilización de los filtros con la transformada Wavelet, y tampoco se utilizó la función silencio, y utilizando la caracterización de coeficientes cepstrales en la escala de Mel. En la tabla 3-1 se consignan los resultados de este procedimiento. Para la verificación se realizó 1200 pruebas de cada clase de palabra y un total 12000 pruebas por proceso.

Tabla 3-1: Porcentaje de aciertos, variando el número de mezclas Gaussianas utilizando la BD1, en el procedimiento 1.

Gaussinas	Adelante	Atrás	Derecha	Izquierda	Pare	Silla
1	89,5	94,3	66,7	57,2	96,3	91,7
2	99,5	99,5	100,0	99,3	99,7	100,0
3	100,0	99,7	100,0	100,0	100,0	100,0
4	100,0	99,2	100,0	99,8	97,5	100,0
5	96,8	100,0	100,0	100,0	99,0	100,0
6	96,8	100,0	100,0	100,0	99,8	100,0
7	100,0	100,0	100,0	99,5	100,0	100,0
8	97,7	99,5	100,0	100,0	100,0	100,0
9	100,0	100,0	100,0	100,0	100,0	100,0
10	100,0	100,0	100,0	100,0	100,0	100,0
12	100,0	100,0	100,0	100,0	100,0	100,0
14	97,3	100,0	100,0	100,0	100,0	100,0
16	100,0	100,0	100,0	100,0	100,0	100,0
18	100,0	100,0	100,0	100,0	100,0	100,0
20	100,0	100,0	100,0	100,0	100,0	100,0
25	99,8	100,0	100,0	100,0	100,0	100,0
30	100,0	100,0	100,0	100,0	100,0	100,0
35	100,0	100,0	100,0	100,0	100,0	100,0
40	100,0	100,0	100,0	100,0	100,0	100,0
45	100,0	100,0	100,0	100,0	100,0	100,0
50	100,0	100,0	100,0	100,0	100,0	100,0

Al analizar la tabla 3-1 y la figura 3-1, se observa una buena respuesta alrededor de 9 mezclas Gaussianas y a partir de este valor se obtienen respuestas del 100%; es de anotar, que existe un punto de inflexión cuando se tienen 7 mezclas Gaussianas, además se observa que las palabras con mayor cantidad de aciertos son: ***derecha*** y **silla** con un porcentaje de 100% en 2 mezclas Gaussianas. La palabra que presenta mayor dificultad es ***adelante***, pues logra obtener el 100% de aciertos, con 16 mezclas Gaussianas; pero que aun en 25 mezclas Gaussianas presenta un ligero descenso.

Figura 3-1: Representación de la relación entre porcentaje de aciertos Vs el número de mezclas Gaussianas para el procedimiento 1, con la base de datos BD1.



La tabla 3-2, representa el escalafón de la clasificación por palabras para este tipo de procedimiento.

Tabla 3-2: Escalafón de la clasificación de la BD1; en el procedimiento 1.

Orden	Palabra	Porcentaje de acierto
1	Pare	99,63
2	Atrás	99,63
3	Silla	99,60
4	Adelante	98,93
5	Derecha	98,41
6	Izquierda	97,90

3.2 Resultados para el procedimiento 2 con BD1

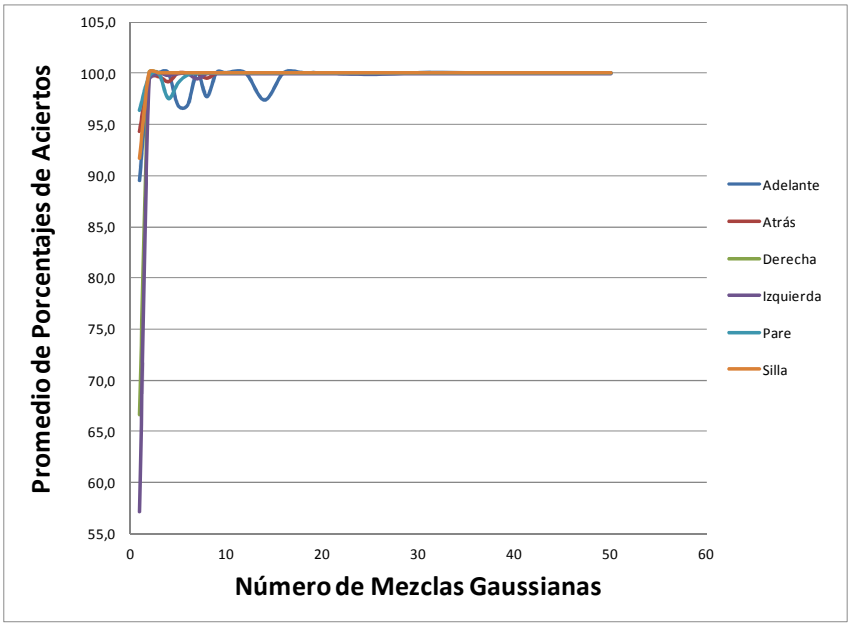
En la tabla 3-3, se recogen los resultados del segundo procedimiento, para DB1; en el cual se ha utilizado un filtro Wavelet DB16 a 10 niveles; obteniendo resultados similares a los de la primera prueba.

Tabla 3-3: Porcentaje de aciertos, variando el número de mezclas Gaussianas utilizando la BD1, en el procedimiento 2.

Gaussinas	Adelante	Atrás	Derecha	Izquierda	Pare	Silla
1	92,3	82,0	95,0	88,0	81,3	84,0
2	99,7	93,7	99,0	99,7	97,3	99,8
3	100,0	99,0	100,0	100,0	99,0	100,0
4	96,7	99,7	100,0	100,0	99,8	100,0
5	100,0	100,0	100,0	100,0	100,0	100,0
6	100,0	100,0	100,0	100,0	100,0	100,0
7	96,7	100,0	100,0	100,0	100,0	100,0
8	96,7	100,0	100,0	100,0	100,0	100,0
9	100,0	100,0	100,0	100,0	100,0	99,8
10	96,7	100,0	100,0	100,0	100,0	100,0
12	100,0	100,0	100,0	100,0	100,0	100,0
14	100,0	100,0	100,0	100,0	100,0	100,0
16	100,0	100,0	100,0	100,0	100,0	100,0
18	100,0	100,0	100,0	100,0	100,0	100,0
20	100,0	100,0	100,0	100,0	100,0	100,0
25	100,0	100,0	100,0	100,0	100,0	100,0
30	100,0	100,0	100,0	100,0	100,0	100,0
35	98,0	100,0	100,0	100,0	100,0	100,0
40	100,0	100,0	100,0	100,0	100,0	100,0
45	100,0	100,0	100,0	100,0	100,0	100,0
50	100,0	100,0	100,0	100,0	100,0	100,0

Al analizar la tabla 3-3 y la figura 3-2, se observa una buena respuesta alrededor de 5 mezclas Gaussianas y a partir de este valor se obtienen respuestas aproximadas al 100%; es de anotar que existe un punto de inflexión cuando se tienen 3 mezclas Gaussianas, y además se observa que las palabras con mayor cantidad de aciertos son derecha, ***izquierda*** y ***silla*** con un porcentaje de 100%, en 3 mezclas Gaussianas. La palabra que presenta mayor dificultad es ***adelante***, pues logra obtener el 100% de aciertos en 12 mezclas Gaussianas; no obstante en 35 mezclas Gaussianas se presenta un ligero descenso.

Figura 3-2: Representación de la relación entre porcentaje de aciertos Vs el número de mezclas Gaussianas para el procedimiento 2, con la base de datos BD1.



En la tabla 3-4, se representa el escalafón de la clasificación por palabras para el procedimiento 2 .

Tabla 3-4: Escalafón de la clasificación de la BD1; en el procedimiento 2.

Orden	Palabra	Porcentaje de acierto
1	Derecha	99,71
2	Izquierda	99,41
3	Silla	99,22
4	Pare	98,93
5	Adelante	98,89
6	Atrás	98,78

3.3 Resultados para el procedimiento 3 con BD1

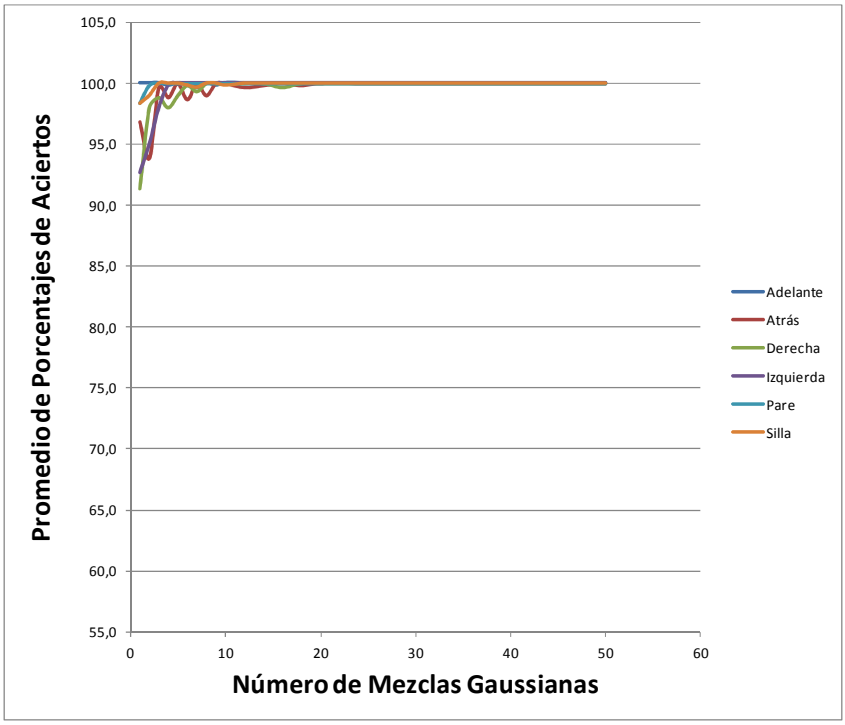
En la tabla 3-4, se recogen los resultados de un tercer procedimiento, en el cual se ha utilizado un filtro Wavelet DB16 a 10 niveles, y agregando el procedimiento de silencio.

Tabla 3-5: Porcentaje de aciertos, variando el número de mezclas Gaussianas utilizando la BD1, en el procedimiento 3.

Gaussinas	Adelante	Atrás	Derecha	Izquierda	Pare	Silla
1	100,0	96,8	91,3	92,7	98,3	98,3
2	100,0	93,8	98,0	95,0	99,8	99,0
3	100,0	99,7	98,8	98,0	100,0	100,0
4	99,8	98,8	98,0	99,8	100,0	100,0
5	100,0	100,0	99,0	100,0	100,0	100,0
6	99,8	98,7	99,8	100,0	100,0	99,8
7	99,7	100,0	99,3	100,0	100,0	99,7
8	100,0	99,0	100,0	100,0	100,0	100,0
9	99,8	100,0	100,0	100,0	100,0	100,0
10	100,0	100,0	100,0	100,0	100,0	99,8
12	100,0	99,7	100,0	100,0	100,0	100,0
14	100,0	99,8	100,0	100,0	100,0	100,0
16	100,0	100,0	99,7	100,0	100,0	100,0
18	100,0	99,8	100,0	100,0	100,0	100,0
20	100,0	100,0	100,0	100,0	100,0	100,0
25	100,0	100,0	100,0	100,0	100,0	100,0
30	100,0	100,0	100,0	100,0	100,0	100,0
35	100,0	100,0	100,0	100,0	100,0	100,0
40	100,0	100,0	100,0	100,0	100,0	100,0
45	100,0	100,0	100,0	100,0	100,0	100,0
50	100,0	100,0	100,0	100,0	100,0	100,0

Al analizar la tabla 3-5 y la figura 3-3, se observa una buena respuesta alrededor de 8 mezclas Gaussianas y a partir de este valor se obtienen respuestas aproximadas al 100%; es de anotar, que existe un punto de inflexión cuando se tienen 5 mezclas Gaussianas, y además se observa que la palabra con mayor cantidad de aciertos es ***adelante*** con un porcentaje de 100%, en 1 mezclas Gaussianas. La palabra que presenta mayor dificultad es ***derecha***, pues logra obtener el 100% de aciertos en 8 mezclas Gaussianas; no obstante en 16 mezclas Gaussianas presenta un ligero descenso.

Figura 3-3: Representación de la relación entre porcentaje de aciertos Vs el número de mezclas Gaussianas para el procedimiento 3, con la base de datos BD1.



En la tabla 3-6, se representa el escalafón de la clasificación por palabras para el procedimiento 3 .

Tabla 3-6: Escalafón de la clasificación de la BD1; en el procedimiento 3.

Orden	Palabra	Porcentaje de acierto
1	Adelante	99,96
2	Pare	99,91
3	Silla	99,84
4	Atrás	99,34
5	Izquierda	99,31
6	Derecha	99,24

3.4 Resumen para los procedimientos 1,2 y 3 con BD1

Una comparación en la tasa de aciertos, lleva a determinar una buena respuesta en los tres procedimientos. Ahora céntrese la atención en el tiempo requerido para realizar el proceso completo; entrenamiento y verificación. La tabla 3-7 recoge los tiempos requeridos para los distintos procesos.

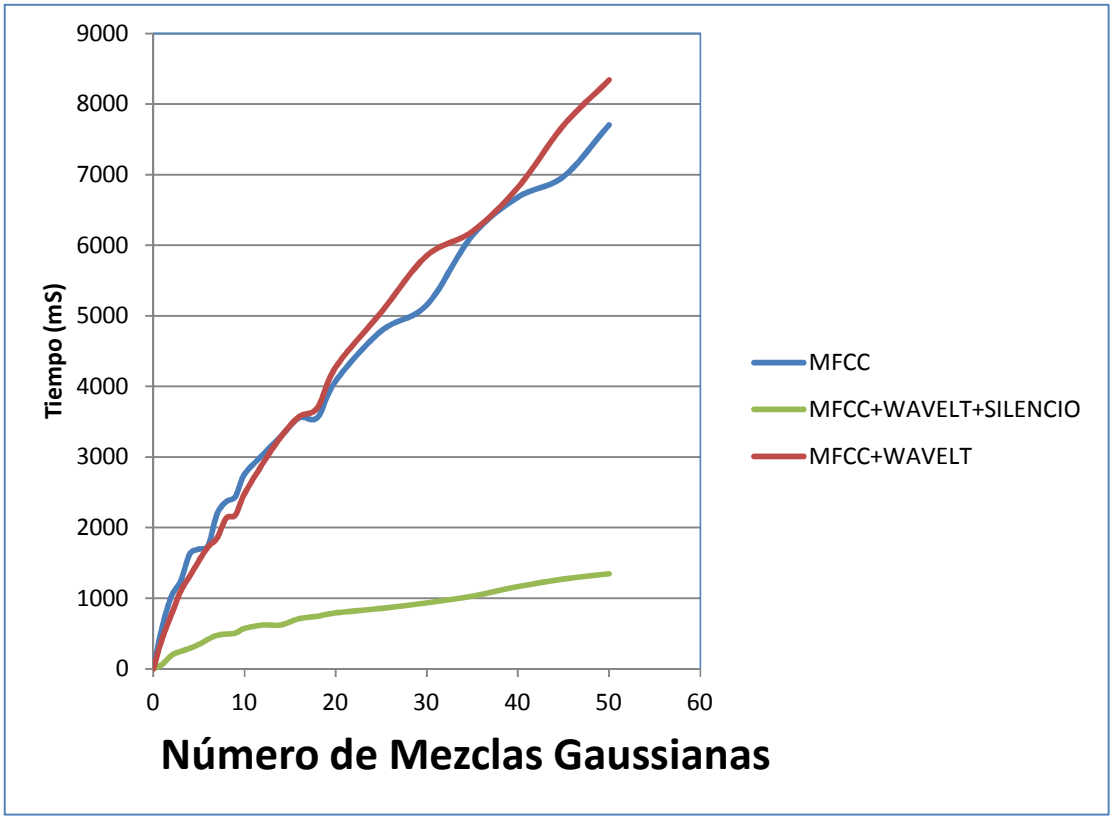
Tabla 3-7: Relaciona los tres procesos; con variaciones de las mezclas Gaussianas utilizando la BD1.

Gaussinas	MFCC (mSeg)	MFCC+WAVELT (mSeg)	MFCC+WAVELT+SILENCIO (mSeg)
0	0	0	0
1	607	437	66
2	1032	771	191
3	1245	1099	246
4	1626	1314	291
5	1698	1528	345
6	1741	1725	416
7	2196	1855	474
8	2369	2136	492
9	2441	2178	505
10	2756	2479	571
12	3031	2906	619
14	3292	3287	621
16	3555	3571	709
18	3560	3696	743
20	4077	4273	791
25	4786	5050	854
30	5155	5851	932
35	6139	6195	1029
40	6681	6818	1162
45	6974	7696	1268
50	7706	8339	1342

Al analizar la tabla 3-7 y la figura 3-4 se puede deducir que para procesos con 9 mezclas gaussianas en el primer proceso que utiliza MFCC se requiere 2441mS y en el procedimiento que utiliza MFCC mas WAVELET mas la función silencio se requiere

505mS. Presentando una respuesta del 80% menos de tiempo, que el requerido en el primer caso.

Figura 3-4: Número de mezclas Gaussianas Vs tiempo de ejecución del proceso, para los tres procedimientos, con la base de datos BD1.



Al analizar los procesos para 5 mezclas Gaussianas en el proceso que utiliza MFCC mas WAVELET se requiere 1528mS y en el procedimiento que utiliza MFCC mas WAVELET mas la función silencio se requiere 345mS. Presentando una respuesta del 77.5% menos de tiempo, que el requerido en el primer caso.

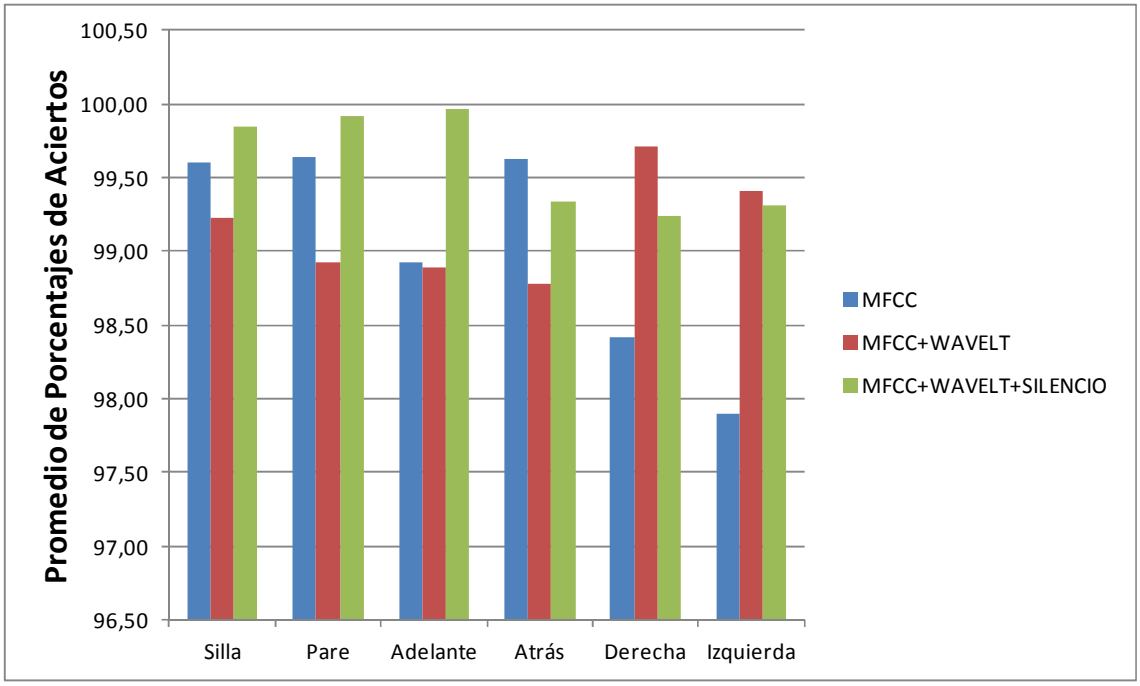
Ahora se presenta la tabla 3-8 que muestra el promedio de cada palabra, en cada uno de los proceso. Permitiendo ver en el proceso 3; una buena respuesta en todas las ocasiones, tomando valores en el rango de aceptación de 99.24% a 99.96%

Tabla 3-8: Relaciona los tres procesos; con todas las palabras de la BD1.

	Silla	Pare	Adelante	Atrás	Derecha	Izquierda	PROMEDIO POR PROCESO
MFCC	99,60	99,63	98,93	99,63	98,41	97,90	99,02
MFCC+WAVELT	99,22	98,93	98,89	98,78	99,71	99,41	99,16
MFCC+WAVELT+SILENCIO	99,84	99,91	99,96	99,34	99,24	99,31	99,60

La figura 3-5, afianza más los resultados, que están almacenados la tabla 3-8. De ésta se puede ver, que para la palabra ***adelante*** los procesos 1 y 2 presentan un relativo menor rendimiento, mientras que el proceso 3, presenta una tasa del 99.96% de aciertos.

Figura 3-5: Porcentaje de aceptación Vs palabras de la base de datos BD1.



3.5 Resultados para el procedimiento 4 con BD2

En una segunda etapa, de nuevo se utiliza una metodología basada en los algoritmos de HMM; con variaciones de mezclas Gaussianas desde 1 hasta 50 y utilizando la base de datos BD2, sin la utilización de los filtros con la transformada Wavelet, y tampoco se utilizó la función silencio, y utilizando la caracterización de coeficientes cepstrales en la escala de Mel. En la tabla 3-9 se consignan los resultados de este procedimiento. Para la verificación se realizó 1200 pruebas de cada clase de palabra y un total 12000 pruebas por proceso.

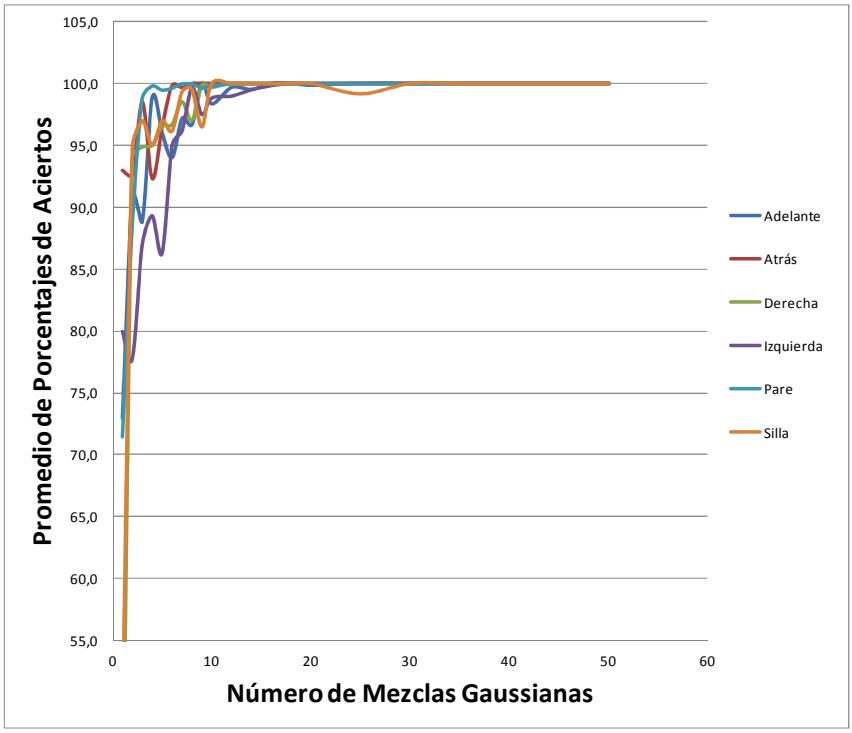
Tabla 3-9: Porcentaje de aciertos, variando el número de mezclas Gaussianas utilizando la BD2, en el procedimiento 4.

Gaussinas	Adelante	Atrás	Derecha	Izquierda	Pare	Silla
1	73,0	93,0	38,0	80,0	71,5	48,0
2	91,0	92,7	94,0	77,8	88,7	94,8
3	88,8	98,5	94,8	87,0	98,8	97,0
4	98,8	92,3	95,0	89,3	99,8	95,0
5	96,0	96,5	96,5	86,3	99,5	97,0
6	94,0	99,8	96,7	95,0	99,7	96,2
7	97,2	99,7	98,5	96,2	100,0	99,3
8	96,7	100,0	97,0	99,7	100,0	99,5
9	100,0	100,0	99,8	97,5	99,7	96,5
10	98,3	100,0	99,7	98,8	99,8	100,0
12	99,7	100,0	100,0	99,0	100,0	100,0
14	99,5	100,0	100,0	99,5	100,0	100,0
16	100,0	100,0	100,0	99,8	100,0	100,0
18	100,0	100,0	100,0	100,0	100,0	100,0
20	99,8	100,0	100,0	100,0	100,0	100,0
25	100,0	100,0	100,0	100,0	100,0	99,2
30	100,0	100,0	100,0	100,0	100,0	100,0
35	100,0	100,0	100,0	100,0	100,0	100,0
40	100,0	100,0	100,0	100,0	100,0	100,0
45	100,0	100,0	100,0	100,0	100,0	100,0
50	100,0	100,0	100,0	100,0	100,0	100,0

Al analizar la tabla 3-9 y la figura 3-6, se observa una buena respuesta alrededor de 12 mezclas Gaussianas y a partir de este valor se obtienen respuestas del 100%; es de anotar, que existe un punto de inflexión cuando se tienen 18 mezclas Gaussianas, y además se observa que la palabra con mayor cantidad de acierto es ***pare*** con un porcentaje de 100% en 8 mezclas Gaussianas. La palabra que presenta mayor dificultad es ***izquierda***, pues logra obtener el 100% de aciertos, con 18 mezclas Gaussianas. Pero

las palabras ***adelante*** y ***silla*** presentan un ligero descenso en 20 y 25 mezclas Gaussianas respectivamente.

Figura 3-6: Representación de la relación entre porcentaje de aciertos Vs el número de mezclas Gaussianas para el procedimiento 1, con la base de datos BD2.



La tabla 3-10, representa el escalafón de la clasificación por palabras para este tipo de procedimiento.

Tabla 3-10: Escalafón de la clasificación de la BD2; en el procedimiento 4.

Orden	Palabra	Porcentaje de acierto
1	Atrás	98,69
2	Pare	97,98
3	Adelante	96,80
4	Silla	96,31
5	Derecha	95,71
6	Izquierda	95,52

3.6 Resultados para el procedimiento 5 con BD2

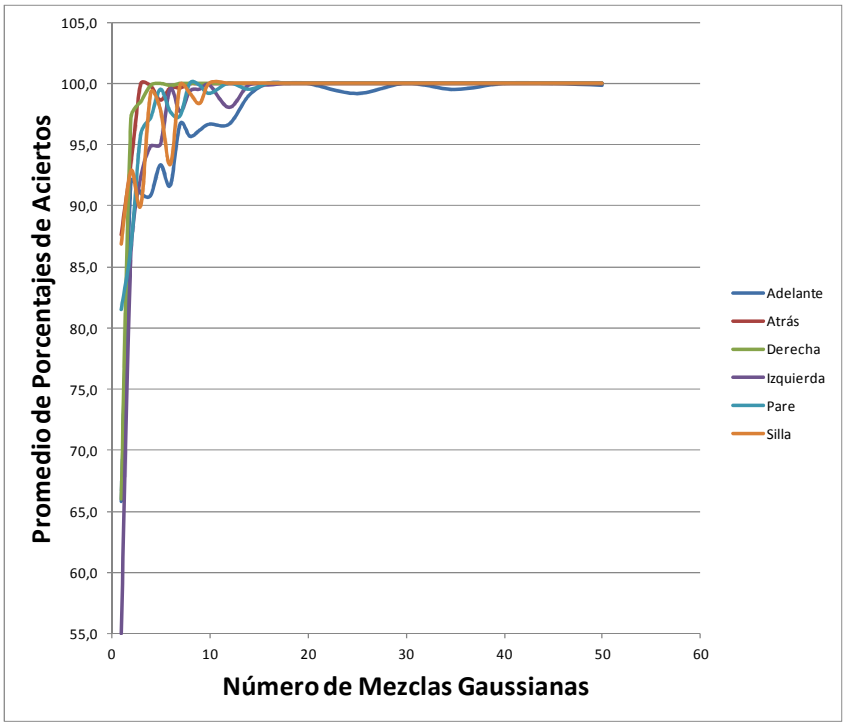
En la tabla 3-11, se recogen los resultados del quinto procedimiento, para DB2; en el cual se ha utilizado un filtro Wavelet DB16 a 10 niveles; obteniendo resultados similares a los de la primera prueba.

Tabla 3-11: Porcentaje de aciertos, variando el número de mezclas Gaussianas utilizando la BD2, en el procedimiento 5.

Gaussinas	Adelante	Atrás	Derecha	Izquierda	Pare	Silla
1	65,8	87,7	66,0	55,0	81,5	86,8
2	92,0	93,5	97,2	86,2	86,7	92,8
3	91,0	100,0	98,5	92,3	95,8	90,0
4	90,8	99,8	99,8	94,8	97,2	99,2
5	93,3	98,7	100,0	95,0	99,5	97,8
6	91,7	99,7	99,8	99,5	97,7	93,3
7	96,7	99,7	100,0	97,7	97,3	99,8
8	95,7	100,0	100,0	99,3	100,0	99,2
9	96,2	100,0	100,0	99,5	99,8	98,3
10	96,7	100,0	100,0	99,8	99,2	100,0
12	96,7	100,0	100,0	98,0	100,0	100,0
14	99,0	100,0	100,0	99,8	99,5	100,0
16	100,0	100,0	100,0	99,8	100,0	100,0
18	100,0	100,0	100,0	100,0	100,0	100,0
20	100,0	100,0	100,0	100,0	100,0	100,0
25	99,2	100,0	100,0	100,0	100,0	100,0
30	100,0	100,0	100,0	100,0	100,0	100,0
35	99,5	100,0	100,0	100,0	100,0	100,0
40	100,0	100,0	100,0	100,0	100,0	100,0
45	100,0	100,0	100,0	100,0	100,0	100,0
50	99,8	100,0	100,0	100,0	100,0	100,0

Al analizar la tabla 3-11 y la figura 3-7, se observa una buena respuesta alrededor de 14 mezclas Gaussianas y a partir de este valor se obtienen respuestas aproximadas al 100%; es de anotar que existe un punto de inflexión cuando se tienen 10 mezclas Gaussianas, y además se observa que la palabra con mayor cantidad de aciertos es ***atrás*** con un porcentaje de 100%, en 3 mezclas Gaussianas. La palabra que presenta mayor dificultad es ***izquierda***, pues logra obtener el 100% de aciertos en 18 mezclas Gaussianas; no obstante la palabra ***adelante*** presenta un ligero descenso en 25, 35 y 50 mezclas Gaussianas

Figura 3-7: Representación de la relación entre porcentaje de aciertos Vs el número de mezclas Gaussianas para el procedimiento 5, con la base de datos BD2.



En la tabla 3-12, se representa el escalafón de la clasificación por palabras para el procedimiento 5 .

Tabla 3-12: Escalafón de la clasificación de la BD2; en el procedimiento 5.

.

Orden	Palabra	Porcentaje de acierto
1	Atrás	99,00
2	Derecha	98,16
3	Silla	97,97
4	Pare	97,82
5	Izquierda	96,04
6	Adelante	95,43

3.7 Resultados para el procedimiento 6 con DB2

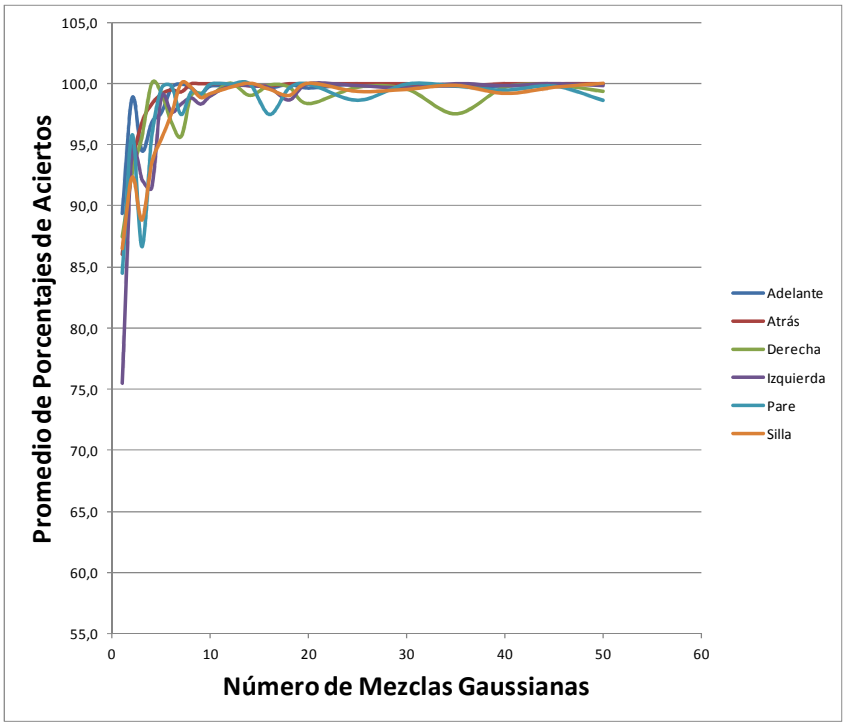
En la tabla 3-13, se recogen los resultados de un sexto procedimiento, en el cual se ha utilizado un filtro Wavelet DB16 a 10 niveles, y agregando el procedimiento de silencio.

Tabla 3-13: Porcentaje de aciertos, variando el número de mezclas Gaussianas utilizando la BD2, en el procedimiento 6.

Gaussinas	Adelante	Atrás	Derecha	Izquierda	Pare	Silla
1	89,3	86,0	87,5	75,5	84,5	86,5
2	98,8	93,2	92,7	94,7	95,8	92,3
3	94,5	96,8	95,7	92,2	86,7	88,8
4	96,8	98,3	100,0	91,5	95,7	93,5
5	97,7	99,2	99,2	99,0	99,7	95,5
6	99,7	99,5	96,8	97,7	99,8	97,5
7	100,0	99,3	95,7	98,3	97,5	100,0
8	99,7	100,0	99,0	98,8	99,3	99,7
9	99,2	100,0	99,2	98,3	99,2	98,8
10	99,8	100,0	99,0	99,0	100,0	99,2
12	99,8	100,0	100,0	99,8	100,0	99,7
14	100,0	99,8	99,0	99,8	100,0	100,0
16	99,7	99,8	99,8	99,7	97,5	99,5
18	100,0	100,0	99,7	98,7	99,7	99,0
20	99,7	100,0	98,3	100,0	100,0	100,0
25	100,0	100,0	99,7	99,8	98,7	99,3
30	99,8	100,0	99,5	99,7	100,0	99,5
35	99,8	99,8	97,5	100,0	99,8	99,8
40	99,8	100,0	99,7	99,8	99,5	99,2
45	100,0	100,0	99,8	100,0	99,8	99,7
50	100,0	100,0	99,3	99,8	98,7	100,0

Al analizar la tabla 3-13 y la figura 3-8, se observa una buena respuesta alrededor de 10 mezclas Gaussianas y a partir de este valor se obtienen respuestas aproximadas al 100%; es de anotar que existe un punto de inflexión cuando se tienen 12 mezclas Gaussianas, y además se observa que la palabra ***derecha*** es la que más rápido logra el 100% con 4 mezclas Gaussianas; no obstante en la mayoría de las pruebas no toma este valor. Se observa que para 12 mezclas Gaussianas existe una inflexión. La palabra que presenta el menor resultado es: ***izquierda***; pues sólo para 20, 35 y 45 mezclas Gaussianas toma valores del 100%.

Figura 3-8: Representación de la relación entre porcentaje de aciertos Vs el número de mezclas Gaussianas para el procedimiento 6, con la base de datos BD2.



Es de observar que este procedimiento es el que presenta de forma más desordenada los datos de respuesta.

En la tabla 3-14, se representa el escalafón de la clasificación por palabras para el procedimiento 6 .

Tabla 3-14: Escalafón de la clasificación de la BD2; en el procedimiento 6.

Orden	Palabra	Porcentaje de acierto
1	Adelante	98,77
2	Atrás	98,66
3	Derecha	97,95
4	Pare	97,71
5	Silla	97,50
6	Izquierda	97,25

3.8 Resumen para los procedimiento 4, 5 y 6 con BD2

Una comparación en la tasa de aciertos, lleva a determinar una buena respuesta en los tres procedimientos. Ahora céntrese la atención en el tiempo requerido para realizar el proceso completo; entrenamiento y verificación. La tabla 3-15 recoge los tiempos requeridos para los distintos procesos.

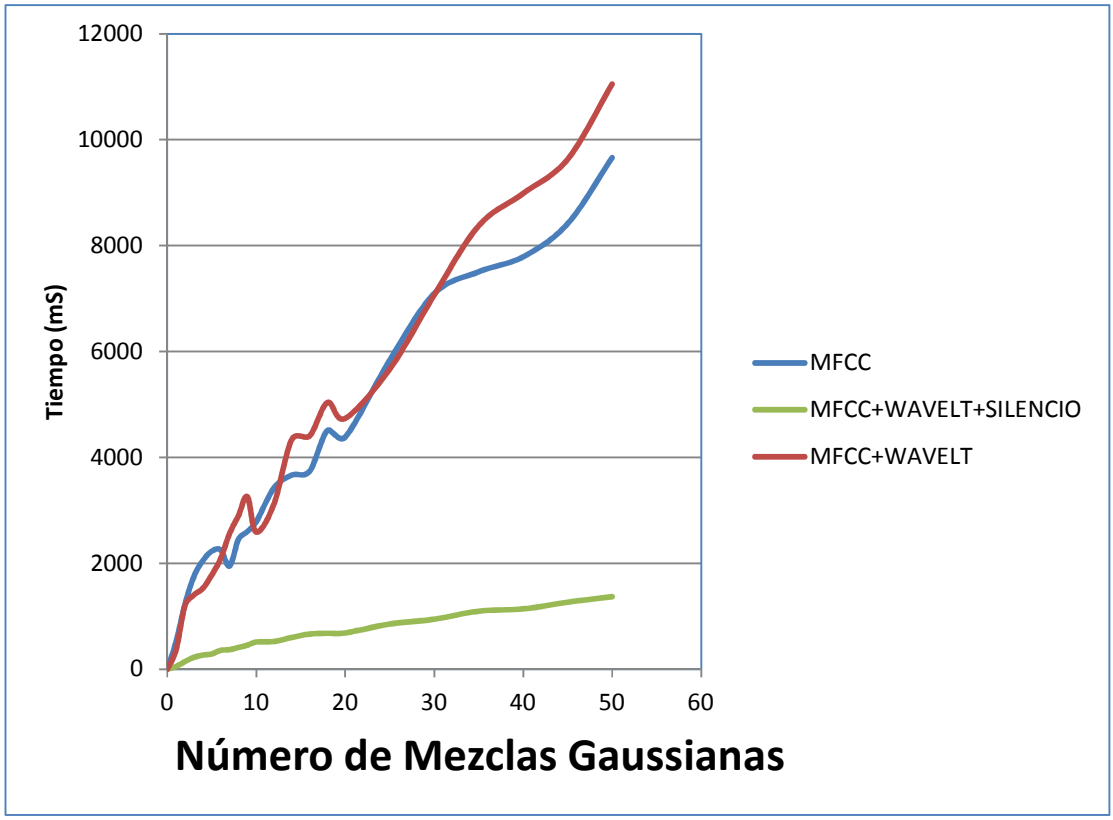
Tabla 3-15: Relaciona los tres procesos; con variaciones de las mezclas Gaussianas utilizando la BD2.

Gaussinas	MFCC (mSeg)	MFCC+WAVELT (mSeg)	MFCC+WAVELT+SILENCIO (mSeg)
0	0	0	0
1	529	376	55
2	1237	1210	147
3	1755	1403	224
4	2052	1531	267
5	2229	1780	288
6	2253	2092	355
7	1950	2562	370
8	2453	2903	411
9	2600	3258	452
10	2785	2589	513
12	3425	3124	524
14	3667	4342	601
16	3741	4406	663
18	4510	5037	678
20	4384	4739	685
25	5835	5668	853
30	7094	7066	942
35	7507	8376	1096
40	7789	8985	1138
45	8417	9632	1264
50	9662	11050	1369

Al analizar la tabla 3-15 y la figura 3-4 se puede deducir que para procesos con 18 mezclas Gaussianas en el cuarto proceso que utiliza MFCC se requiere 4510mS y en el procedimiento que utiliza MFCC mas WAVELET mas la función silencio se requiere

678mS. Presentando una respuesta del 85% menos de tiempo que el requerido en el primer caso.

Figura 3-9: Número de mezclas Gaussianas Vs tiempo de ejecución del proceso para los tres procedimientos, con la base de datos BD2.



Al analizar los procesos para 10 mezclas Gaussianas en el proceso que utiliza MFCC mas WAVELET se requiere 2589mS y en el procedimiento que utiliza MFCC mas WAVELET mas la función silencio se requiere 513mS. Presentando una respuesta del 80.2% menos tiempo que el requerido en el proceso 5.

Ahora se presenta la tabla 3-16 que muestra el promedio de cada palabra, en cada uno de los procesos; Permitiendo ver en el proceso 6, una buena respuesta en todas las ocasiones, tomando valores en el rango de aceptación de 97.25% a 98.72%. En esta tabla se puede ver que a medida que se elimina el ruido, tanto con la transformada

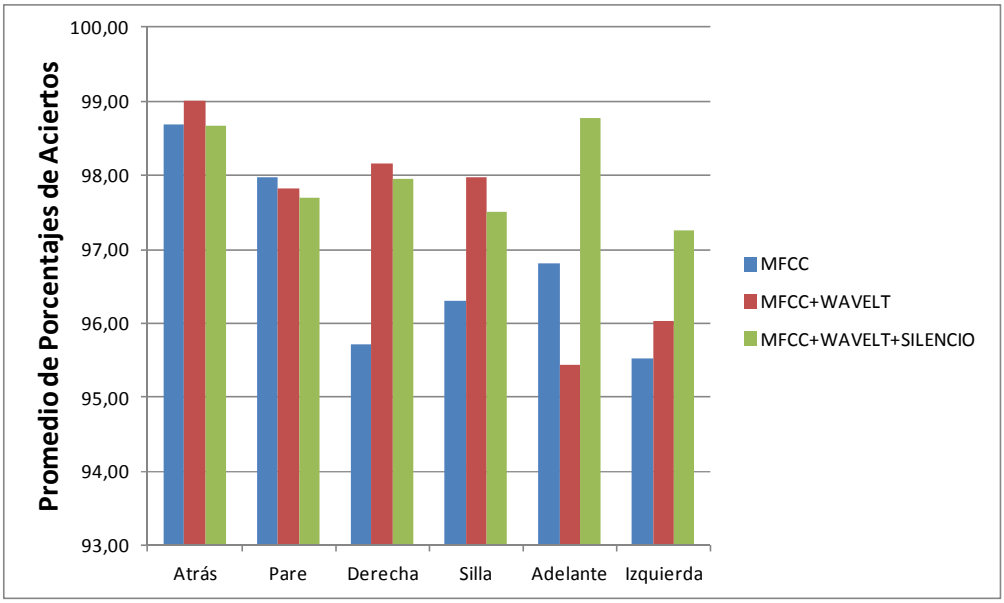
wavelet como con la combinación entre las wavelet y la función silencio, la tasa de acierto se incrementa.

Tabla 3-16: Relaciona los tres procesos; con todas las palabras de la BD2.

	Atrás	Pare	Derecha	Silla	Adelante	Izquierda	PROMEDIO POR PROCESO
MFCC	98,69	97,98	95,71	96,31	96,80	95,52	96,84
MFCC+WAVELT	99,00	97,82	98,16	97,97	95,43	96,04	97,40
MFCC+WAVELT+SILENCIO	98,66	97,71	97,95	97,50	98,77	97,25	97,97

La figura 3-10, afianza más los resultados, que están almacenados en la tabla 3-14. De ésta se puede ver que en la palabra ***adelante,*** los procesos 4 y 5 presentan un relativo menor rendimiento, mientras que el proceso 5 presenta una tasa del 98.77% de aciertos.

Figura 3-10: Porcentaje de aceptación Vs palabras de la base de datos BD2.



Hay que recordar que la base de datos BD2, hay una diversidad de hablantes y situaciones que implican la presencia de alto ruido.

4. Conclusiones y recomendaciones

En el anterior proyecto de investigación, se implemento una metodología para el reconocimiento de comandos de voz de palabras aisladas, orientado al control de una silla de ruedas, mediante el uso del algoritmo de los modelos ocultos de Markov (HMM). La caracterización de las señales de voz se ha realizado mediante los coeficientes ceptrales en la escala de Mel (MFCC), además se utilizo la reducción del ruido de fondo y la transformada wavelet; que permitieron una alta tasa de aciertos con un costo computacional razonable.

Para la adecuación de los registros de voz, se utilizaron dos algoritmos que permitieron reducir las perturbaciones, eliminando así puntos de la señal que no contienen información útil en el registro de voz. Los algoritmos utilizados fueron la eliminación del silencio y ruido de fondo. Este primer procedimiento permite acortar la señal dejando únicamente la información relacionada con la voz eliminando perturbaciones del ambiente y las pausas de las palabras, esto se logra determinando la energía promedio de la señal y la energía relativa de la misma en ventanas de tiempo. Al comparar estos dos valores y al multiplicar la energía relativa de la señal por un umbral, si el valor de la energía en la ventana de tiempo es menor que el procedimiento anterior, esto quiere decir que en esta ventana de tiempo hay silencio o ruido ambiental por lo cual esta ventana es eliminada. Por otro lado se utilizó la transformada wavelet que al funcionar como filtro, elimina aquellos canales de la señal que no pertenezcan al umbral entre 0 y 3400 Hz que es la escala de frecuencia de la voz. Estos dos procedimientos permiten eliminar las perturbaciones y ruido, dejando en el registro únicamente información relacionada con el habla, adecuando la señal de la voz para su posterior procesamiento. Se ha conseguido obtener unos modelos de la voz que proporcionan resultados adecuados en términos de su preprocesamiento; al utilizar la herramienta de las transformada Wavelet como filtro. Permitiendo presentar en el proceso de clasificación un ahorro de tiempo del 8% al 10%, con una tasa de aciertos general de 99,16% para

señales con bajo ruido y un solo locutor. Para señales con alto ruido y múltiples locutores se obtiene reducciones de tiempo en el orden del 6% al 8%, con una tasa de aciertos general de 97,4%.

En términos de la caracterización, los Coeficientes Cepstrales de Frecuencia en escala de Mel (MFCC) son una representación efectiva; porque al ser definidos como el cepstrum de una señal ventaneada en el tiempo que ha sido derivada de la aplicación de una transformada rápida de Fourier, pero en una escala de frecuencia no lineal, ss aproximan más al comportamiento del sistema de percepción del habla.

Con el empleo de un algoritmo de extracción de características, basado en los coeficientes cepstrales de la escala de Mel se ha obtenido un matriz de caracterización para cada registro: $46\,filas$ valor obtenido según la ecuación (46) y el número de columnas; está determinado por el número de coeficientes elegidos, en este proyecto es igual a 20, el valor se duplica por que se toma la derivada. este valor será igual a 40, para entregar un matriz de: 46 * 40; para los procesos 1, 2, 4 y 5, que no implican la utilización del algoritmo de la reducción de ruido de fondo (función silencio). Para los procesos 3 y 6, que implican el uso del algoritmo de la reducción de ruido de fondo (función silencio), se obtiene una matriz de caracterización de: $7filas$ y el número de columnas; está determinado por el número de coeficientes elegidos, en este proyecto es igual a 20, el valor se duplica por que se toma la derivada; este valor será igual a 40, para entregar un matriz de 7 * 40. Demostrando con esto, la disminución de los coeficientes de caracterización; permitiendo la reducción del tiempo de procesamiento y su uso en aplicaciones en tiempo real. Estas condiciones generaron resultados satisfactorios con una tasa de aciertos en el orden 99,02% para señales con bajo ruido y un solo locutor, de 96,84% para señales con alto ruido y múltiples locutores.

La metodología construyó un modelo por cada palabra del vocabulario, y la estructura de esos modelos se definió en la fase de diseño: el número de estados (*N*) se elige “a priori” según la complejidad que se pueda permitir y la calidad deseada. En este proyecto se definieron 2 estados. Lo mismo ocurre con el tipo de transiciones: la matriz *A* tendrá sólo

ciertas componentes distintas de cero, y su número es un parámetro de diseño. El tipo de funciones estadísticas que se utilizarán para modelar las probabilidades de observación de los puntos de la plantilla desde cada estado, también se fija antes de entrar en la fase de entrenamiento de los modelos, y que cada uno de los estados del fonema es representado por un determinado número de mezclas de gaussianas. Para logar los presentes resultados, fue necesario configurar el algoritmo de entrenamiento; con variaciones de las mezclas Gaussianas, desde 1 mezcla hasta 50 mezclas. El uso de un clasificador HMM, que permite el reconocimiento de comandos de voz, de palabras aisladas, para el control de dispositivos de ayuda a personas en estado de discapacidad, ha entregado unos resultados satisfactorios, tanto para señales con bajo ruido y un solo locutor, como para las señales con alto ruido y múltiples locutores. Además, presenta resultados similares con o sin el uso del algoritmo de la reducción de ruido de fondo (función silencio). Los resultados de aciertos se hallan en el rango de 96,84% a 97,97%.

El aumento del número de mezclas de Gaussianas para el modelado de cada uno de los estados, provoca un aumento en el porcentaje de palabras acertadas, de igual forma que se da un mayor incremento en la precisión de los sistemas. Por tanto, el aumento de las Gaussianas se hace con el fin de mejorar la exactitud del sistema; pero el aumento no debe ser sin límites; se puede observar en las tablas de resultados que para mezclas Gaussianas de 14 en adelante, el sistema está sobre entrenado. Con una configuración de los HMM para 10 mezclas Gaussianas; se obtuvo una tasa de aciertos adecuada (96,84% a 97,97%.) con una relación prudente entre precisión y tiempo de cómputo.

Con la unión de un filtro para una señal de voz; basado en las componentes de la trasformada Wavelet, al que se le agrega una función de silencio; luego se hace una caracterización basada en los MFCC, con la modelación y clasificación de los HMM ha sido posible crear una metodología de reconocimiento de palabras aisladas, independiente de la presencia de ruido, que puede ser utilizada por uno o varios locutores.

Bibliografía

[1] M. Grimm and K. Kroschel, Robust Speech Recognition and Understanding, 2007. 3, 8

[2] K. Homayounfar, “Rate adaptive speech coding for universal multimedia access,” IEEE Signal Processing Magazine, p. 30–39, 2003. 3

[3] D. O. Shaughnessy, “Interacting with computers by voice: Automatic speech recognition and synthesis,” ICONIP 2006, Part II, LNCS 4233,, pp. 489–498, 2006. 3

[4] B. Mellorf, C. Baber, and C. Tunley, “Evaluating automatic speech recognition as a component of a multi-input device human-computer interface,” in Proceedings of the International Conference on Spoken Language Processing, pp. 1668–1671, 1996. 3

[5] T. Kubik and M. Sugisaka, “Use of a cellular phone in mobile robot voice control,” in Proceedings of the 40th SICE Annual Conference International Session Papers SICE 2001, 2001, pp. 106–111. 3

[6] S. Shamma, “Relevance of auditory cortical representations to speech processing and recognition,” p. 5, 2005. 3

[7] B. Juang and T. Chen, “The past, present, and future of speech processing,” IEEE Signal Processing Magazine, vol. 15, pp. 24–48, 1998. 3, 7, 9, 14

[8] R. Goecke, “Current trends in joint audio-video signal processing: A review,” p. 70–73, 2005. 3, 4, 9

[9] R. Campbell, “Audio-visual speech processing,” Elsevier, pp. 562–569, 2006. 2, 3, 4

[10] ——, “The processing of audio-visual speech: empirical and neural bases,” Philosophical Transactions of The Royal Society B, no. 363, p. 1001–1010, 2008. 2, 3, 4

[11] J. Hurtado, G. Castellanos, and J. Suarez, “Effective extraction of acoustic features after noise reduction for speech classification,” in In International Conference on Modern

Problems of Radio Engineering, Telecommunications and Computer Science, (TCSET'02), 2002, pp. 245–248. 3

[12] J. Schroeter, J. Larar, and M. Sondhi, "Speech parameter estimation using a vocal tract/cord model," in In IEEE International Conference onAcoustics, Speech, and Signal Processing, (ICASSP '87), vol. 12. IEEE, 1987, pp. 308–311. 3

[13] N. Cheng, L. Mabiner, A. Rosenberg, and C. Mooonegal, "Some comparisons among several pitch detection algorithms," in IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '76, vol. 1. IEEE, 1976, pp. 332– 335. 3

[14] R. Hoory, A. Sagi, S. Shechtman, A. Sorin, Z. Shuang, and R. Bakis, "High quality sinusoidal modeling of wideband speech for the purposes of speech synthesis and modification," in In IEEE International Conference on Acoustics, Speech, and Signal Processing. (ICASSP'06), vol. 1. IEEE, 2006, pp. I877–I880. 3

[15] R. G. Termens, J. O. Lafont, F. G. Portabella, and J. M. Roca, "Síntesis de voz utilizando difonemas: uniones entre vocales," Procesamiento del lenguaje natural, no. 21, pp. 69–74, 1997. 3

[16] J. Rothweiler, "Noise-robust 1200-bps voice coding," in Proceedings of the Tactical Communications Conference: Technology in Transition, vol. 1, 1992, pp. 65–69. 3, 7

[17] J. Macres, "Real-time implementations and applications of the US federal standard CELP voice coding algorithm," in Proceedings of the Tactical Communications Conference: Technology in Transition, vol. 1, 1992, pp. 41–45. 3, 7

[18] L. R. Rabiner, "A tutorial on hidden markov models and selected application in speech recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286, 1989. 3, 4, 14, 15, 16

[19] S. Anderson and D. Kewley-Port, "Evaluation of speech recognizers for speech training applications," IEEE Transactions on Speech and Audio Processing, vol. 3, no. 4, pp. 229–241, 1995. 4, 15

[20] A. W. Drake, “Discrete-state Markov processes,” Chapter 5 in Fundamentals ofApplied Probability Theory. New York, NY: McGraw-Hill, 1967.

[21] The material in this section and in Section I l l is based on the ideas presented by Jack Ferguson of IDA in lectures at Bell Laboratories.

[22] J. K. Baker, ”The dragon system-An overview,” /€E€ Trans. Acoust. Speech Signal Processing, vol. ASP-23, no. 1, pp. 24-29, Feb. 1975.

[23] F. Jelinek, “Continuous speech recognition by statistical methods,” Proc. / € E € , vol. 64, pp. 532-536, Apr. 1976.

[24] F. Jelinek, L. R. Bahl, and R. L. Mercer, “Continuous speech recognition: Statistical methods,” in Handbook of Statistics, I/, P. R. Krishnaiad, Ed. Amsterdam,The Netherlands: North-Holland, 1982.

[25] L. R. Bahl, F. Jelinek, and R. L. Mercer, “A maximum likelihood approach to continuous speech recognition,” /€€E Trans. Pattern Anal. Machine Intel/., vol. PAMI-5, pp. 179-190, 1983.

[26] L. R. Rabiner, S. E. Levinson,and M. M. Sondhi,“On theapplication of vector quantization and hidden Markov models to speaker-independent isolated word recognition,” Bell Syst. Tech. I., vol. 62, no. 4, pp. 1075-1105, Apr. 1983.

[27] -, “On the use of hidden Markov models for speaker-independent recognition of isolated words from a medium-size vocabulary,” AT&T Tech. I., vol. 63, no. 4, pp. 627-642, Apr. 1984.

[28] R. Billi, “Vector quantization and Markov source models applied to speech recognition,” in Proc. lCASSP ’82 (Paris, France), pp. 574-577, May 1982.

[29] L. R. Rabiner, B. H. Juang, S. E. Levinson, and M. M. Sondhi, “Recognition of isolated digits using hidden Markov models with continuous mixture densities,” AT&T Tech. I., vol. 64, no. 6, pp. 1211-1222, July-Aug. 1986.

[30] A. B. Poritz and A. G. Richter, ”Isolated word recognition,” in Proc. lCASSP ’86 (Tokyo, Japan), pp. 705-708, Apr. 1986.

[31] R. P. Lippmann, E. A. Martin, and D. B. Paul, “Multistyle training for robust isolated word speech recognition,” in Proc. lCASSP ’87(Dallas, TX), pp. 705-708, Apr. 1987.

[32] D. B. Paul, “A speaker stress resistant HMM isolated word recognizer,” in Proc. lCASSP‘87(Dallas, TX), pp. 713-716, Apr. 1987.

[33] V. N. Gupta, M. Lennig and P. Mermelstein, “Integration of acoustic information in a large vocabulary word recognizer,” in Conf. Proc. /E€€ lnt. Conf. on Acoustics, Speech, and Signal Processing, pp. 697-700, Apr. 1987.

[34] K. F. Lee and H. W. Hon, “Large-vocabulary speaker-independent continuous speech recognition,” in Conf. Proc. /€€E lnt. Conf. on Acoustics, Speech, and Signal Processing, pp. 123-126, Apr. 1988.

[35] F. Jelinek, “A fast sequential decoding algorithm using a stack,” ISM]. Res. Develop., vol. 13, pp. 675-685, 1969.

[36] R. Schwartz et al., “Context-dependent modeling for acoustic-phonetic recognition of continuous speech,” in Conf. Proc. /€E€ lnt. Conf. on Acoustics, Speech, and Signal Processing, pp. 1205-1208, Apr. 1985.

[37] J. G. Wilpon, L. R. Rabiner, and T. Martin, “An improved word detection algorithm for telephone quality speech incorporating both syntactic and semantic constraints,” AT&T Bell Labs Tech. I., vol. 63, no. 3, pp. 479-498, Mar. 1984.

[38] J. G. Wilpon and L. R. Rabiner, “Application of hidden Markov models to automatic speech endpoint detection,” Computer Speech and Language, vol. 2, no. 3/4, pp. 321-341, Sept./Dec. 1987.

[39] A. Averbuch et al., “Experiments with the TANGORA 20,000 word speech recognizer,” in Conf. Proc. /€E€ lnt. Conf. on Acoustics, Speech, and Signal Processing, pp. 701-704, Apr. 1987.

[40] Deller, J. R., Proakis J. G. and Hansen, J. H. L., Discrete-Time Processing of Speech Signals, Ed. Macmillan Publishing Company, New York, 1993.

[41] Davis, S. B. y Mermelstein, P., “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”, IEEE Trans. Acoustics, Speech and Signal Proc., 28 (4), pp. 357-366, agosto 1980.

[42] Davis, S. B. y Mermelstein, P., “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”, IEEE Trans. Acoustics, Speech and Signal Proc., 28 (4), pp. 357-366, agosto 1980.

[43] Nadeu C., Hernando J. and Gorricho, M., “On the Decorrelation of Filter-Bank Energies in Speech Recognition”, Eurospeech 1995, pp. 1381-1384, 1995.

[44]. C. R. Jankowski, H. D. H. Vo, and R. P. Lippmann, “A comparison of signal processing front ends for automatic word recognition,” IEEE Trans. Speech Audio Process., vol. 3, no. 4, July 1995.

[45]. J. C. Floes, " Técnicas para el reconocimiento de voz en palabras aisladas en la lengua Náhuatil", 2010.

[46]. S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans. Acoust., Speech, Signal Processing, vol. 28(4), pp. 357–366, 1980.

[47]. L. C. W. Pols, Spectral analysis and identification of Dutch vowels in monosyllabic words, Ph.D. dissertation, Free University, Amsterdam, The Netherlands, 1977.

[48] Daubechies Ingrid. (1992) “Ten Lectures on Wavelets”.Capital City Press Filadelfia USA.

[49] Kaiser Gerald (1999) “A Friendly Guide to Wavelets” Birkhauser U.S.A.

[50] Mallat Stephane (1998) “A Wavelet Tour of Signal Processing” Academic. Press USA

[51] Chui K Charles. (1992) “An Introduction in Wavelet”.Academia. Press Inc. Reino Unido.

[52] Jiménez Carlos, Díaz J. A. y Del Pino P. (2004) “Determinación de la relación señal a ruido de la voz utilizando la transformada de wavelet”. Ingeniería UC. Venezuela.

[53] García Janer Leonard (1998) “Transformada de Wavelet aplicada a la extracción de información en señales de voz”. Barcelona España. http://www.eupmt.es/imesd/telematica/veu/thesis.pdf.

[54] Cuesta Frau, Kovak D. (2000) “Reducción del ruido en señales electrocardiográficas mediante la transformada de wavelet”. http://hpk.felk.cvut.cz/~xnovakd1/doc/wavelet.pdf

[55] Novak Daniel, Cuesta D. (2000) “Denoising electrocardiogram signal using adaptive Wavelets”.Universidad de Valencia España. http://plutarco.disca.upv.es/~jcperez/Documentos/Comg_sBrnodavid.pdf.

[56] Donoho D. (1995) “Denoising by softthresholdoing”.IEEE. Trabs. Information theory.Vol 41, Nº 3.

[57] Childers Donald (1997) “Probability and Random Processes”. McGraw Hill U.S.A.

[58] Papoulis Athanasios. (1984) “Probability Random variables and Stochastic”. Processes.McGraw Hill U.S.A.