UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

-PROYECTO FIN DE CARRERA-

MEJORAS EN RECONOCIMIENTO DEL HABLA BASADAS EN MEJORAS EN LA PARAMETRIZACIÓN DE LA VOZ

Ingeniería de Telecomunicación

Leticia Rueda Rojo

Abril 2011

Mejoras en Reconocimiento del Habla Basadas en Mejoras en la Parametrización de la Voz

AUTOR: Leticia Rueda Rojo
TUTOR: Doroteo Torre Toledano

ATVS Área de Tratamiento de Voz y Señal
(http://atvs.ii.uam.es)
Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid

Resumen

El reconocimiento automático del habla (RAH) es un campo emergente con el objetivo de crear una interfaz hombre-máquina lo más natural posible. El mayor obstáculo en RAH para un uso más amplio de esta tecnología es la robustez al ruido. Aunque los sistemas actuales de RAH superan en gran medida el reconocimiento de voz humana en condiciones de ruido, se sigue investigando en sistemas que traten de mejorar la robustez mediante la elaboración de sistemas de inspiración biológica, es decir, que intenten simular lo más fielmente posible el comportamiento del sistema auditivo humano. La mayoría de los sistemas de RAH hacen uso del algoritmo de extracción de características *Mel Frecuency Cepstral Coefficients* (MFCC), este algoritmo está basado en la implementación de un banco de filtros cuyos filtros están espaciados unos de otros según una escala de frecuencia log-lineal. Sin embargo, el ancho de banda de cada filtro se selecciona en función del espaciado entre unos y otros y no en función de ningún parámetro de inspiración biológica. La adaptación del ancho de banda de los filtros que forman el banco de filtros a otros parámetros tales como, rango de frecuencias o número de filtros, ha llevado a diferentes modificaciones en el algoritmo original. Recientemente algunos autores han introducido una nueva modificación en el algoritmo MFCC que independiza el ancho de banda de los filtros de otros parámetros del banco de filtros haciendo uso de la conocida relación entre las frecuencias de centro y el ancho de banda crítico del sistema auditivo humano. Este nuevo algoritmo llamado *Human Factor Cepstral Coefficients* (HFCC) intenta superar al algoritmo original MFCC en experimentos de RAH en condiciones de ruido. En esta línea se ha introducido una variación sobre HFCC, llamada HFCC-E, en el que el ancho de banda de los filtros se escala linealmente por un factor llamado E-factor con el fin de investigar los efectos que produce en el reconocimiento el hecho de aumentar el ancho de banda en condiciones de ruido.

Palabras Clave

Sistema de reconocimiento de habla, Modelos Ocultos de Markov, MFCC (*Mel Frecuency Cepstral Coefficients*), HFCC (*Human Factor Cepstral Coefficients*), ERB (*Equivalent Rectangular Bandwith*), CENSREC-2, HTK (*Hidden Markov Model Toolkit*).

Abstract

Automatic speech recognition (ASR) is an emerging field with the goal of creating a more natural man-machine interface. The single largest obstacle to widespread use of ASR technology is robustness to noise. Since human speech recognition greatly outperforms current ASR systems in noisy environments, ASR systems seek to improve noise robustness by drawing on biological inspiration. Most ASR front ends employ Mel Frequency Cepstral Coefficients (MFCC) which is a filter bank-based algorithm whose filters are spaced on a perceptually-motivated linear-log frequency scale. However, filter bandwidth is set by filter spacing and not through biological motivation. The coupling of filter bandwidth to other filter bank parameters such as, frequency range or number of filters, has led to variations of the original algorithm. The authors have recently introduced a novel extension to MFCC is introduced which decouples filter bandwidth from the other filter bank parameters by employing the known relationship between filter center frequency and critical bandwidth of the human auditory system. The new algorithm, called Human Factor Cepstral Coefficients (HFCC), is shown to outperform the original MFCC in ASR experiments in noise conditions. In some research the authors have introduced a variation of HFCC, called HFCC-E, in which filter bandwidth is linearly scaled in order to investigate the effects of wider filter bandwidth on noise robustness.

Key Words

Speech recognition system, HMM (*Hidden Markov Models*), MFCC (*Mel Frecuency Cepstral Coefficients*), HFCC (*Human Factor Cepstral Coefficients*), ERB (*Equivalent Rectangular Bandwith*), CENSREC-2, HTK (*Hidden Markov Model Toolkit*).

Agradecimientos

Este trabajo significa el punto y final de una etapa y de las innumerables experiencias que me han hecho crecer tanto académicamente como persona. Es el resultado de años de esfuerzo, de desilusiones, de alegrías y de buenos y malos momentos. Gracias a todos aquellos que me han acompañado en el camino.

En primer lugar mi agradecimiento a Doroteo, mi tutor, por su ayuda y su tiempo.

Gracias a mis compañeros y amigos de la universidad por los buenos momentos, las risas, los ánimos y la ayuda que en algún momento me hayan podido ofrecer. En especial gracias a ti, Cris, porque sin ti no hubiese sido lo mismo. Empezamos juntas un camino difícil que tu compañía ha hecho mucho más agradable y llevadero. Gracias por todos los momentos, por tu apoyo, tus sonrisas, tus confidencias y tu cariño. Gracias por haberme brindado tu amistad.

Gracias César, por tu apoyo incondicional, por tu confianza, tus consejos, por entenderme, por tu amor. Gracias por estar ahí siempre.

Gracias a mi hermana, Virginia, por tu encanto especial, por tus ánimos. Porque eres de esa clase de personas que dan lo mejor de si mismos sin esperar nada a cambio… Porque sabes escuchar y brindar ayuda cuando es necesario… Porque te has ganado el cariño, admiración y respeto de todo el que te conoce. Te quiero.

Y sobre todo, a quienes la ilusión de su vida ha sido convertirme en una persona de provecho. Gracias papá y mamá, por el esfuerzo que habéis hecho y el cariño que me habéis dado durante todos estos años, aún estando lejos. Muchas gracias por los ánimos, las preocupaciones y la paciencia que habéis tenido. Porque vuestro apoyo y consejos han posibilitado la conquista de esta meta. Sin vosotros este proyecto no hubiese sido posible. Os quiero.

Gracias a todos por lo que hemos logrado juntos.

Índice de contenidos

Índice de Figuras

Índice de Tablas

1 Introducción

1.1 Motivación

Con la evolución de las tecnologías asociadas a la información nuestra sociedad está cada día más conectada electrónicamente. Labores que tradicionalmente eran realizadas por seres humanos son, gracias a las mejoras tecnológicas, realizadas por sistemas automatizados. En el planteamiento general del desarrollo de los ordenadores de las próximas generaciones se prevé la comunicación hombre–máquina mediante mensajes orales. Una comunicación oral hombre-máquina debe reproducir el modelo que rige en el proceso de comunicación cotidiana entre humanos. Debemos, por tanto, facultar al ordenador para hablar y entender lo que se le dice, aunque la capacidad de entendimiento constituye hoy en día un horizonte lejano. En la mayoría de los casos la información proviene de un ser humano y finalmente también es usada por un ser humano. Por tanto, son necesarios métodos efectivos de transferencia de información entre hombres y máquinas en ambas direcciones. El habla es el medio más espontáneo y natural de comunicación entre los hombres, sin embargo, hasta el presente se puede afirmar que en su comunicación con las máquinas el hombre ha hecho uso exclusivo del lenguaje escrito. Resulta natural, por tanto, extender la capacidad de comunicación hombre-máquina al mensaje oral. Además de la naturalidad y espontaneidad aludidas, la comunicación oral hombre-máquina presenta importantes ventajas en gran cantidad de aplicaciones, como el diálogo interactivo o la entrada de grandes cantidades de datos en la máquina. Una de estas ventajas es que en la comunicación oral las manos y la vista del usuario quedan liberadas, pudiendo dedicarse a una tarea simultánea a la comunicación. Ello ofrece posibilidades muy interesantes en el gobierno de sistemas de gran complejidad en los que la atención visual sea muy importante. Una segunda ventaja importante proviene del hecho de la universalidad de la red telefónica. Aunque ésta puede ser aprovechada para la transferencia de información sin acudir al habla, la comunicación oral, al no requerir otro equipo que el teléfono, ofrece una ventaja sustancial: cualquier aparato telefónico se convierte en un enlace potencial con el ordenador y de este modo los accesos a bases de datos, las reservas y ventas de billetes de viaje, las operaciones bancarias, etc. podrían realizarse desde cualquier punto.

Se denomina reconocimiento del habla al proceso de extraer información lingüística de una señal de voz. Este proceso que el ser humano es capaz de llevar a cabo automáticamente y casi inconscientemente, es extremadamente complejo. Prueba de ello son los estudios realizados sobre la psicofísica de la percepción humana de habla que ponen de manifiesto esta complejidad y demuestran que en este proceso intervienen complejos procesos cerebrales relacionados con el lenguaje. Son estos procesos precisamente los que convierten al ser humano en el sistema de reconocimiento del habla más robusto, de hecho, el ser humano

es capaz de extraer información lingüística en ambientes ruidosos, con falta de información o incluso con información errónea.
El reconocimiento del habla es aplicable a una gran variedad de situaciones donde se requiera una comunicación hombre-máquina, redacción de textos sin teclado, telefonistas automáticas, ayuda a discapacitados físicos, etc… Sin embargo, con las técnicas actuales, sólo en el caso de palabras aisladas, cuando el vocabulario es reducido y en situaciones acústico fonéticas (variabilidad fonética, tipo de locutores, ruido y distorsiones,...) poco dificultosas existen soluciones satisfactorias. Por tanto, se ve necesaria una mejora en los métodos de extracción de características de la señal de voz para avanzar en este problema.

El reconocimiento automático de voz nace en los años 50 como respuesta científico-tecnológica al deseo del hombre de interaccionar oralmente con las máquinas. En un primer momento se fundamenta en los principios de la fonética acústica y se limita al reconocimiento de palabras aisladas de un vocabulario muy reducido, para un locutor exclusivo y utilizando para ello dispositivos electrónicos. En la década de los 70 se empieza a tener en cuenta que el conocimiento sintáctico, semántico y contextual son fuentes de información muy útiles. Se utilizan microprocesadores y se aplican las técnicas de programación dinámica al reconocimiento de palabras conectadas. Empiezan a aparecer reconocedores independientes de locutor para tareas muy concretas. Pero es en los años 80 cuando se da un giro metodológico fundamental con el modelado estadístico y el uso de los modelos ocultos de Markov o HMMs. A partir de ese momento, el reconocimiento de habla continua ha mejorado, aumentándose el tamaño de los vocabularios, diversificándose las aplicaciones y enfrentándose a situaciones cada vez más reales, en las que locutores y las condiciones del entorno de reconocimiento difieren de los que se han utilizado para entrenar el reconocedor. En la actualidad las tasas de reconocimiento más optimistas están en un orden de magnitud por debajo de las que serían atribuibles al ser humano y el reconocimiento automático de voz continua es un campo de trabajo al que la comunidad científica dedica un esfuerzo importante.

En lo que a su aplicación práctica se refiere, el reconocimiento automático del habla empezó como el particular reto científico de emular el comportamiento humano con máquinas, siendo objeto de interés para un público y unas aplicaciones bastante específicas y limitadas. La situación actual no podría ser más antagónica, habiendo sido denominada *tercera revolución industrial*. Los avances tecnológicos previos y originados por el nacimiento de la *Sociedad de la Información* con las *Tecnologías de la Información y las Comunicaciones (TICs)* asociada a ella, han provocado una revolución en la demanda de interfaces usuario-máquina lo más amigables y transparentes posibles para el usuario. El diálogo hombre-máquina aparece en este escenario como mecanismo óptimo y natural de relación de los habitantes de esa Sociedad de la Información con *sus* TICs desde varios puntos de vista:

- Desde un punto de vista **científico-tecnológico**, la inteligencia artificial y, en particular, el reto de emular la comunicación oral humana sigue siendo un estímulo atractivo. Los avances en capacidad de computación y potencia del software y hardware, así como en conocimiento del comportamiento humanos, reinventan continuamente el camino.

- Desde un punto de vista **económico**, hay que señalar tres factores:

  i. La tecnología se ha convertido en un bien de consumo.

  ii. El mercado ha cambiado sus protocolos y la compartición de información, y la globalización de los intercambios que conlleva el comercio electrónico exigen la presencia de las Tecnologías de la Información. La demanda de reconocimiento automático del habla crece de manera incesante en los sistemas de telecomunicación, en los sistemas de control y en los sistemas de entrada de datos y de acceso a bases de datos.

iii. Las personas de edad avanzada en las poblaciones del primer mundo y los inmigrantes de países con menor penetración de las TICs, son dos sectores del mercado con un alto potencial como consumidores. Estos dos sectores demandan interfaces universales, intuitivas y amigables.

- Desde un punto de vista **social**, la Sociedad de la Información ha creado oportunidades muy interesantes y, al mismo tiempo, ha generado la llamada *brecha digital* entre los usuarios de las TICs. El desarrollo de tecnologías amigables casi transparentes para el usuario, es una ayuda para combatir esta brecha.

El análisis anterior enmarca la motivación de este trabajo científico en reconocimiento automático de voz. El reconocimiento automático debe ser intrínsecamente robusto. Es decir, debe dar las máximas prestaciones posibles en las condiciones más adversas imaginables. Las condiciones adversas para un reconocedor se definen como las diferencias o desajustes que puedan existir entre los datos con los que ha sido entrenado, y los datos que debe reconocer. El robustecimiento de un reconocedor de voz se puede definir como la aportación de mecanismos que lo hagan menos vulnerable a esos desajustes de las condiciones de entrenamiento y evaluación. Existe una importante línea científica para el estudio de estrategias de robustecimiento que atacan las debilidades del reconocedor desde puntos de vista diferentes.

Los sistemas acústicos basan el reconocimiento en las características espectrales de la señal de voz. Así, todos los sistemas de reconocimiento automático del habla requieren una primera etapa en la cual, segmentos consecutivos de la señal de voz son convertidos a secuencias temporales de vectores de parámetros. Este proceso se conoce con el nombre genérico de parametrización. Su objetivo es la extracción de información relevante de la señal acústica analógica, eliminando las redundancias y la información asociada a las fuentes de variabilidad que tiene la misma. La etapa de parametrización determinará en buena parte las prestaciones del sistema, tanto en lo referente a tasas de reconocimiento como a carga computacional y requerimientos de memoria necesarios. Por tanto, se puede considerar que el problema fundamental en la parametrización es la elección de un modelo adecuado de la señal de voz. El modelo elegido para la señal debe ser capaz de estimar una envolvente útil para el sistema de reconocimiento, es decir, robusta a variaciones interlocutor o intralocutor, dependiendo de la aplicación, fenómenos de coarticulación, ruido ambiental, etc. Esta envolvente debe ser, además, susceptible de ser representada con un número reducido de parámetros, el cálculo de los cuales debe exigir la mínima carga computacional posible. En la resolución de este problema pueden distinguirse claramente dos enfoques utilizados por la mayoría de los sistemas de reconocimiento: *Linear Prediction Cepstral Coefficients* (LPCC) y *Mel Frecuency Ceptral Coefficients* (MFCC).

La principal dificultad de estos parámetros radica en la intravariabilidad inherente al proceso de producción de voz. Las representaciones actuales de la voz, aunque poco eficientes, debido a que conllevan mucha redundancia, permiten conseguir unas buenas prestaciones del reconocimiento siempre que la señal de voz se registre en condiciones favorables. Sin embargo, cuando un sistema de reconocimiento se pone a funcionar en situaciones reales se encuentra con condiciones adversas tales como cambios en el hablante (condiciones fisiológicas, emocionales, cambio en el modo de articulación debido a un fuerte ruido ambiental, entre otras) y en el entorno acústico (ruidos, reverberación y ecos) o eléctrico (como ruidos o distorsiones de la señal provocados por el micrófono o el canal de transmisión), que son irrelevantes desde el punto de vista lingüístico pero que pueden degradar en gran medida la tasa de reconocimiento. En el curso de esta investigación, algunos autores proponen una nueva forma de parametrización llamada *Human Factor Cepstral Coefficients* (HFCC) con el fin de mejorar la robustez frente al ruido. Surge así una nueva técnica encaminada a obtener un sistema más robusto, que utiliza la parametrización llevada a cabo en el algoritmo MFCC (*Mel Frecuency Cepstral Coefficients*) modificando el análisis del banco de filtros.

1.2 Objetivos

El proyecto tiene como objeto estudiar, desarrollar, implementar y documentar una nueva forma de parametrización llamada *Human Factor Cepstral Coefficients* (HFCC). De esta forma, se llevará a cabo el diseño de todo el proceso de parametrización HFCC, haciendo especial hincapié en el diseño del banco de filtros, con el fin de obtener un comportamiento más aproximado al del oído humano y así, un mejor comportamiento del reconocedor en general, intentando obtener un sistema robusto frente al ruido. Es decir, que la presencia de ruido en la señal de habla afecte lo menos posible a la tasa de reconocimiento de nuestro sistema. Se implementará, por tanto, un sistema completo de parametrización de la voz y se documentarán los resultados obtenidos. El proyecto se desarrolla apoyándose en el estado del arte actual en temas similares y se adaptan los cuantiosos estudios en parametrización a este nuevo escenario.

Las principales características de HFCC radican en que la separación entre los filtros viene dada por la escala ERB en lugar de la escala Mel que usan los MFCC. Se ha demostrado empíricamente que la escala ERB está más ajustada a las características de percepción del oído humano. Por otro lado, el ancho de banda de cada filtro que compone el banco del filtros es un parámetro de diseño libre, independiente de la separación entre filtros y vinculado con la conocida relación entre la frecuencia de centro y el ancho de banda crítico del sistema auditivo humano, lo cual nos permite una nueva línea de investigación en el reconocimiento automático de voz. La habilidad de controlar el ancho de banda de los filtros en el diseño del banco de filtros es importante por dos razones:

1. Elimina errores en el ancho de banda producidos por elecciones equivocadas en el número de filtros o en el rango de frecuencia en el banco de filtros.
2. Permite la optimización del ancho de banda.

En esta línea, se introduce una variación de HFCC llamada HFCC-E, en el que se considera un factor de escala lineal llamado E - factor con el que podemos controlar el ancho de banda de los filtros con el fin de investigar los efectos que produce en el reconocimiento una variación del ancho de banda del filtro en condiciones de ruido.

Por tanto, se analizará cómo afecta el nuevo banco de filtros diseñado en el reconocedor, así como el funcionamiento de esta parametrización en entornos ruidosos. Se analizará igualmente la variación del ancho de banda de los filtros y su efecto, realizando un análisis comparativo con todos los resultados, así como posibles mejoras que se puedan realizar. Por tanto, este trabajo se enfocará al estudio y análisis de una nueva forma de parametrización de la voz basada en el hecho de tener un comportamiento más aproximado al del oído humano, y así, intentar obtener un sistema robusto frente al ruido orientado a conseguir un mejor comportamiento del reconocedor en general.

1.3 Organización de la memoria

Esta memoria consta de los siguientes capítulos:

Capítulo 1. Introducción

Este capítulo presenta la motivación para la realización de este proyecto y los objetivos que se persiguen durante el desarrollo del mismo.

Capítulo 2. Estado del arte

Este capítulo comienza con una introducción sobre el procesamiento humano del habla, haciendo hincapié en el proceso de producción y percepción del habla. A continuación se da una visión sobre el estado del arte en reconocimiento fonético y procesamiento de señales digitales de audio. Igualmente se hace un estudio profundo en el proceso de parametrización de la señal de voz. Finalmente se explican los modelos ocultos de Markov y, para terminar, se realiza el estudio de los algoritmos de extracción de características MFCC y HFCC.

Capítulo 3. Marco experimental

Este capítulo comienza con una descripción del sistema utilizado para la realización de este proyecto. A continuación se describe la base de datos utilizada, así como los procedimientos seguidos en el diseño del proceso de parametrización objeto de este estudio. Finalmente se detallan las especificaciones de los scripts de configuración utilizados.

Capítulo 4. Pruebas y resultados

En este capítulo se analizan las pruebas realizadas y los resultados experimentales obtenidos mediante el algoritmo de extracción de características HFCC-E. Además, se evaluará la mejora relativa obtenida en comparación con los resultados MFCC proporcionados por la base de datos utilizada.

Capítulo 5. Conclusiones y trabajo futuro

Este capítulo se centrará en el análisis de las conclusiones obtenidas tras el estudio de los resultados expuestos en el capítulo 4. Además, se valorarán las posibles mejoras a realizar como trabajo futuro con el fin de mejorar el estudio realizado.

2

Estado del arte

2.1 Sonido y procesamiento humano del habla

El habla constituye la forma más natural de comunicación entre las personas, de ahí el gran interés que tiene el desarrollo de sistemas informáticos capaces de procesar el habla y generarla de forma automática. El procesamiento del habla abarca un amplio abanico de métodos y técnicas que tienen, entre muchas otras posibles, las siguientes finalidades. Por una parte, lograr que los ordenadores puedan comprender los mensajes pronunciados por los usuarios, y por otra, lograr que los usuarios puedan entender los mensajes generados por los ordenadores de forma oral.

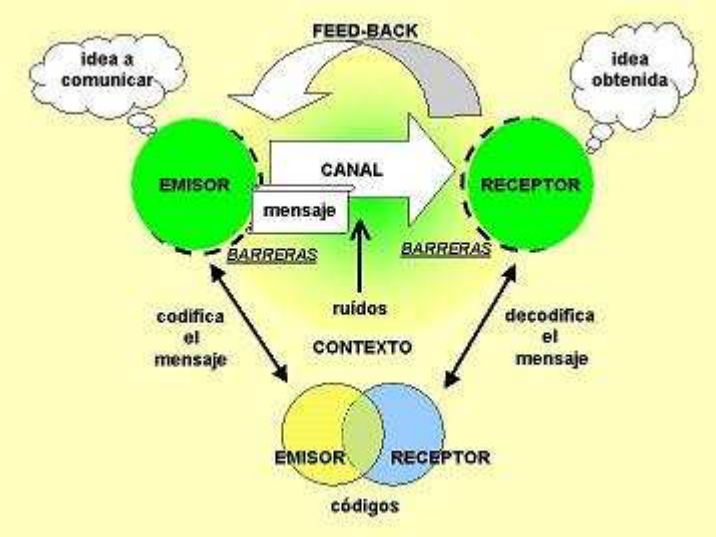

Figura 2.1 - Esquema del modelo de comunicación oral.

Cada vez es más importante tener una interacción con las máquinas cercana a la comunicación oral, a la que los humanos accedemos desde edades muy tempranas. Los primeros sentidos que se desarrollan plenamente en los humanos son los que nos permiten la comunicación oral. Es por esto y por la posibilidad de acercar las máquinas al mundo de discapacitados, tanto físicos como motrices, que la comunicación oral con las máquinas ha cobrado una importancia vital en los últimos tiempos. No obstante, si bien la voz es el medio de comunicación más usual, los humanos producimos y percibimos la misma con gran redundancia y de ella extraemos la información más relevante. Es por esto muy importante determinar cómo se produce y percibe la voz a la hora de realizar su tratamiento automático.

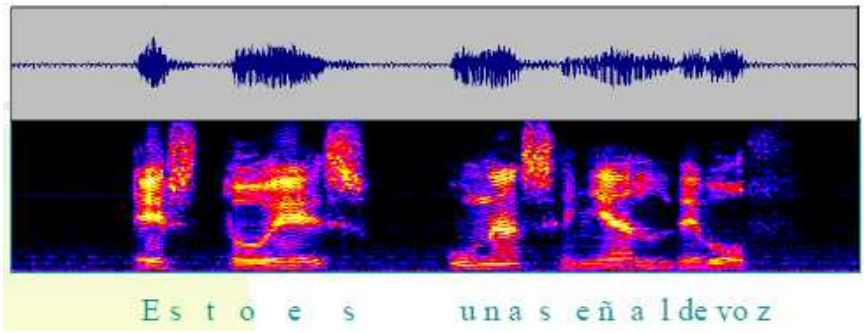

Figura 2.2 - Forma de onda y espectro de una señal de voz para la locución "Esto es una señal de voz".

A lo largo de estas páginas analizaremos el método de producción de voz y su modelado matemático. También se repasará la importancia de su tratamiento espectral, que consigue reducir de una manera importantísima la cantidad de material acústico sin perder la información que permanece en ella.

2.1.1 El sonido

Mucho de lo que aprendemos del mundo que nos rodea nos llega a través del sentido del oído. El oír es importante no solamente para aprender del mundo, sino también para comunicarse con otros humanos. La voz humana es única en su habilidad de expresar ideas abstractas.

Un sonido es una onda de presión longitudinal formada por compresiones y expansiones del aire en dirección paralela a la aplicación de energía. Las compresiones son zonas donde las moléculas de aire han sido forzadas por la aplicación de energía, dando lugar a una mayor concentración de las mismas. Y las expansiones son zonas donde la concentración de moléculas de aire es menor.

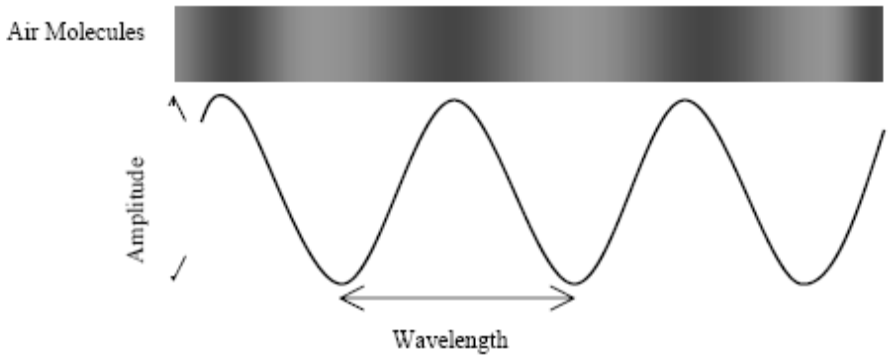

Figura 2.3 - La aplicación de energía provoca alternativamente compresión y expansión de moléculas de aire. Las áreas más oscuras significan mayor concentración de moléculas de aire.

Cuando nos referimos al sonido audible por el oído humano, lo definimos como una sensación percibida en el órgano del oído producida por la vibración que se propaga en un medio elástico en forma de ondas. El sonido audible para los seres humanos está formado por las oscilaciones de la presión del aire que el oído convierte en ondas mecánicas y finalmente, en impulsos nerviosos para que el cerebro pueda percibirlas y procesarlas.

El sonido puede representarse como una suma de curvas sinusoides con un factor de amplitud diferente que pueden caracterizarse por las mismas magnitudes y unidades de medida que cualquier sinusoidal:

longitud de onda (λ), frecuencia (f) o período (T) y amplitud. Cuando se considera la superposición de diferentes ondas es importante la fase que representa el retardo relativo en la posición de una onda con respecto a otra.

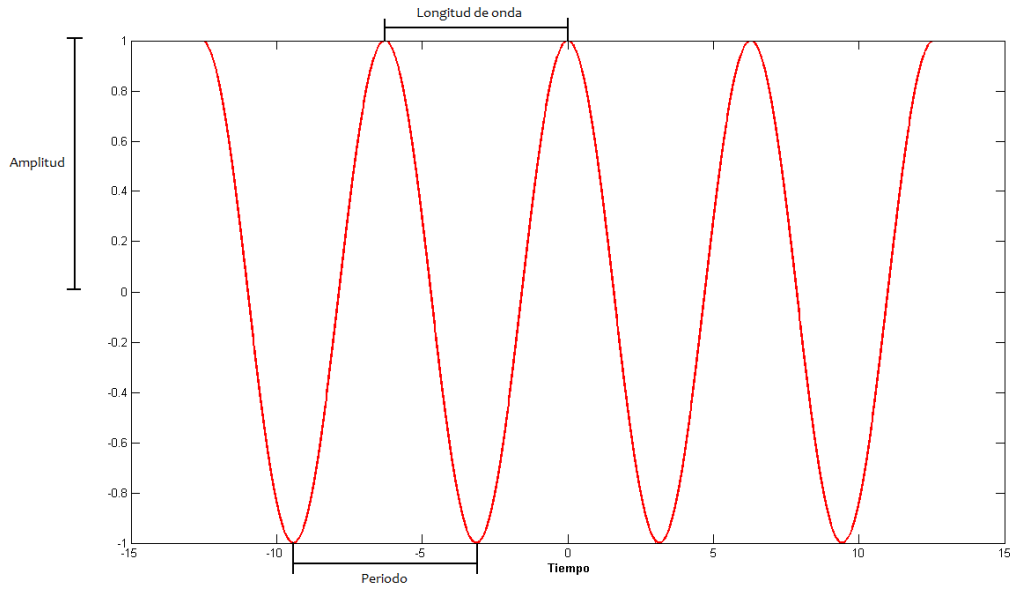

Figura 2.4 - Magnitudes de una sinusoidal.

Sin embargo, un sonido complejo cualquiera no está caracterizado por los parámetros anteriores ya que, en general, un sonido cualquiera es una combinación de sinusoidales que difieren en los cuatro parámetros anteriores. La caracterización de un sonido arbitrariamente complejo implica analizar tanto la energía transmitida como la distribución de dicha energía entre las diversas frecuencias, para ello resulta útil investigar:

- **Potencia acústica**: es la cantidad de energía por unidad de tiempo (potencia) emitida por una fuente determinada en forma de ondas sonoras. La potencia acústica viene determinada por la propia amplitud de la onda, pues cuanto mayor sea la amplitud de la onda, mayor es la cantidad de energía (potencia acústica) que genera.
- **Espectro de frecuencias:** que permite conocer en qué frecuencias se transmite la mayor parte de la energía.

Los sonidos de los que consta el habla se pueden clasificar por la forma en que se produce el sonido básicamente en tres tipos:

- **Sonoros**. Son aquellos sonidos que hacen vibrar las cuerdas vocales. Esta vibración es cuasi periódica y su espectro es muy rico en armónicos, que son múltiplos de la frecuencia de vibración de las cuerdas. A esta frecuencia de vibración de las cuerdas se le llama frecuencia fundamental. La frecuencia fundamental depende de la presión ejercida al pasar el aire por las cuerdas y de la tensión de éstas. En un hombre la frecuencia fundamental se encuentra en el rango 50-250 Hz, mientras en la mujer el rango es más amplio, encontrándose entre 100 y 500 Hz.

- **Fricativos**. En los sonidos fricativos se produce un estrechamiento del tracto vocal por el que se hace pasar el aire, lo que proporciona como resultado una excitación de ruido aleatorio.

- **Plosivos**. Estos sonidos se producen por la existencia de una obstrucción temporal al paso del aire. El sonido se produce al abrirse la obstrucción temporal produciéndose una liberación brusca de energía en forma de una pequeña explosión.

2.1.2 Producción del habla

El sistema de producción del habla no forma parte estricta del sistema sensorial humano, pero su importancia es indudable. Para determinar las operaciones de un sistema automático de reconocimiento de voz y hablante, es fundamental conocer y determinar los mecanismos que han producido un mensaje hablado, para a continuación, poder reproducirlos automáticamente. Es por ello que se van a repasar algunos conceptos fundamentales y básicos en el mecanismo de producción del habla, tanto en el órgano físico que soporta dichos mecanismos, como la producción propia del mensaje.

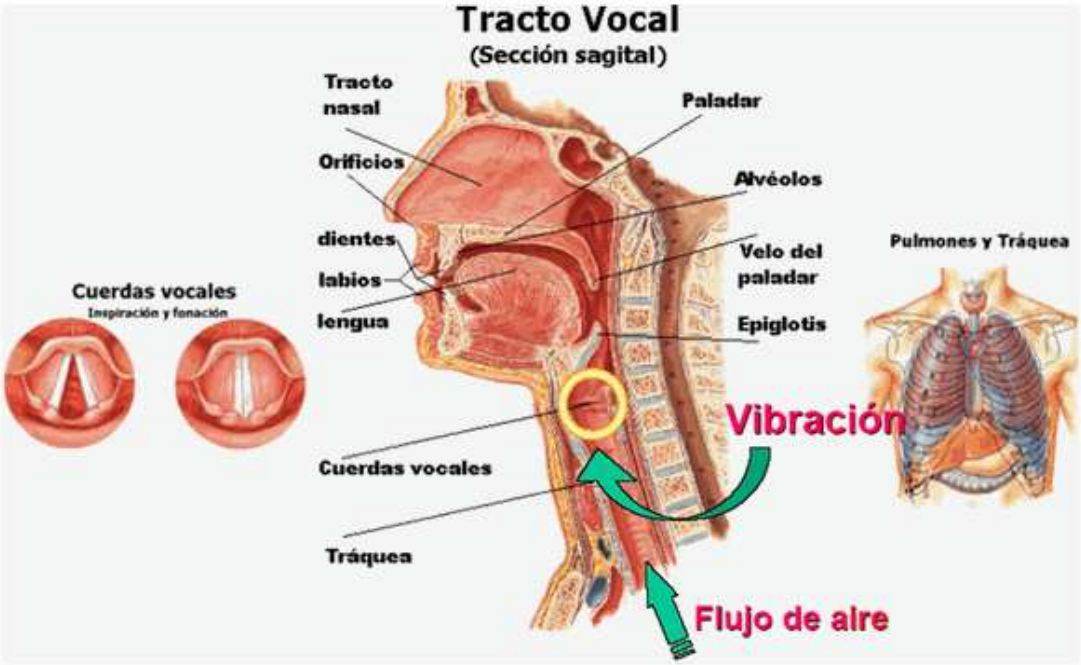

Figura 2.5 - Sección del tracto vocal.

El habla, como señal acústica, se produce a partir de las ondas de presión que salen de la boca y las fosas nasales de un locutor. El proceso comienza con la generación de la energía suficiente (flujo de aire) en los pulmones, la modificación de ese flujo de aire en las cuerdas vocales, y su posterior perturbación por algunas constricciones y configuraciones de los órganos superiores. Así, en el proceso fonador intervienen distintos órganos a lo largo del llamado tracto vocal, que en nuestro caso asumiremos que se restringe a la zona comprendida entre las cuerdas vocales y las aberturas finales: los labios y las fosas nasales.

El conjunto de órganos que intervienen en la fonación (figura 2.5) puede dividirse en tres grupos bastante bien delimitados:

1. Cavidades **infraglóticas** (sistema sub-glotal) u **órgano respiratorio**.
2. Cavidad **laríngea** u **órgano fonador**.
3. Cavidades **supraglóticas**.

2.1.2.1 Cavidades infraglóticas

Las cavidades infraglóticas constan de los órganos propios de la respiración (pulmones, bronquios y tráquea), que son la fuente de energía para todo el proceso de producción de voz.

En el proceso de inspiración, los pulmones toman aire, bajando el diafragma y agrandando la cavidad torácica. En el momento de la fonación, la espiración, provocada por la contracción de los músculos intercostales y del diafragma, aporta la energía necesaria para generar la onda de presión acústica que atravesará los órganos fonadores superiores.

2.1.2.2 Cavidad laríngea

La cavidad laríngea es la responsable de modificar el flujo de aire generado por los pulmones y convertirlo (o no, como veremos), en una señal susceptible de excitar adecuadamente las posibles configuraciones de las cavidades supraglóticas.

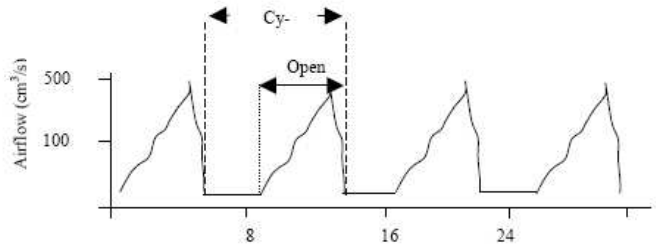

Figura 2.6 - Flujo de aire durante el ciclo de la laringe.

El último cartílago de la tráquea, el cricoides, forma la base de la laringe, cuyo principal órgano son las cuerdas vocales que son dos pares de repliegues compuestos de ligamentos y músculos. El par inferior son las llamadas cuerdas vocales verdaderas, que pueden juntarse o separarse mediante la acción de los músculos crico-aritenoides lateral y posterior, y que están protegidas en su parte anterior por el cartílago tiroides, el más importante de la laringe, abierto por su parte posterior. Finalmente, la parte superior de la laringe está unida al hueso hioides.

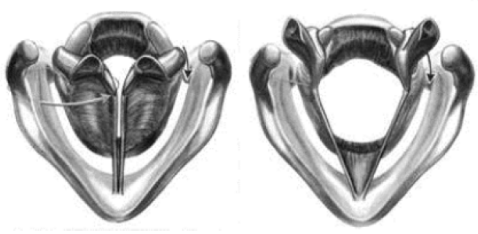

Figura 2.7 - Vista transversal de las cuerdas vocales abiertas y cerradas.

En la figura 2.7 se muestra una vista transversal simplificada de la zona en la que se encuentran las cuerdas vocales, en sus posiciones extremas: abiertas y cerradas. A la apertura que queda entre las cuerdas vocales se le denomina glotis.

La cavidad laríngea está terminada por la epiglotis, un cartílago en forma de cuchara que permite cerrar la apertura de la laringe en el acto de la deglución.

La distinción fundamental entre los sonidos se basa en su característica de sonoridad. En los sonidos sonoros, incluyendo las vocales, se observa un patrón regular tanto en su estructura temporal como en su estructura frecuencial (figura 2.8), patrón del que carecen los sonidos sordos (figura 2.9).

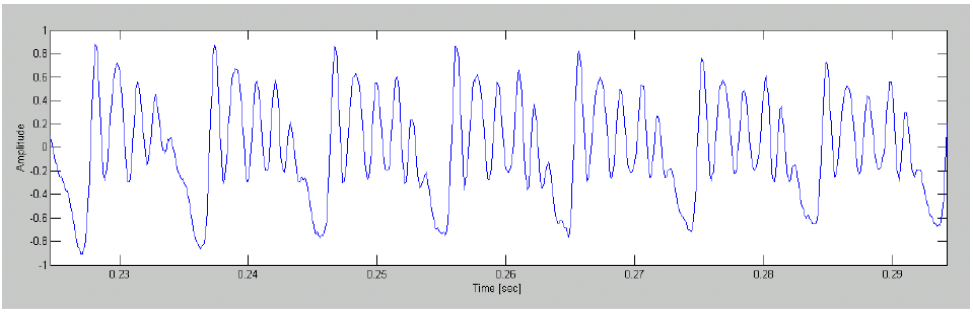

Figura 2.8 - Forma de onda de un sonido sonoro (vocal a).

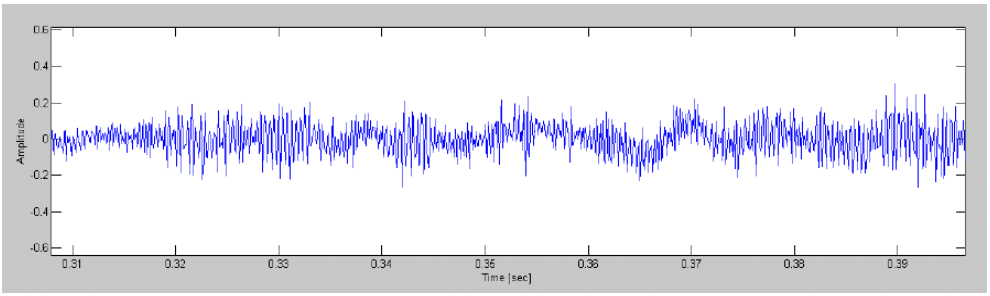

Figura 2.9 - Forma de onda de un sonido sordo (consonante s).

La cualidad de sonoridad de los sonidos sonoros se produce por la acción vibradora de las cuerdas vocales. El mecanismo de vibración se produce de la siguiente forma: si suponemos que inicialmente las cuerdas vocales están juntas, la presión subglotal se incrementa lo suficiente para forzar a las cuerdas vocales a separarse. Al separarse, el aire pasa a través de ellas y la presión subglotal disminuye, momento en el que la fuerza de los músculos hace que las cuerdas vocales vuelvan a juntarse. Cuando las cuerdas vocales se juntan, el flujo de aire disminuye y la presión subglotal aumenta de nuevo, con lo que se vuelve a reproducir el ciclo (esquematizado en la figura 2.10), y esta vibración de las cuerdas vocales produce pulsos casi periódicos de aire que excitan el sistema por encima de la laringe.

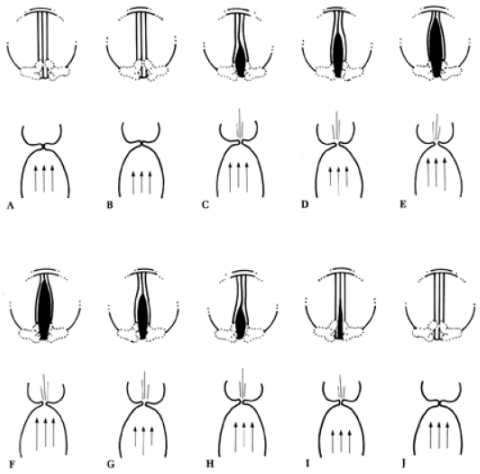

Figura 2.10 - Esquema del proceso de vibración de las cuerdas vocales.

A esta frecuencia de vibración se la denomina frecuencia fundamental, y sus valores típicos oscilan entre los 60 Hz para un hombre voluminoso, y los 300 Hz para una mujer o un niño. La señal generada en las cuerdas vocales puede variar en frecuencia e intensidad según varíe la masa, la longitud y la tensión de las mismas.

2.1.2.3 Cavidades supraglóticas

Las cavidades supraglóticas están constituidas por la faringe, la cavidad nasal y la cavidad bucal. Su misión fundamental de cara a la fonación es perturbar adecuadamente el flujo de aire procedente de la laringe, para dar lugar finalmente a la señal acústica generada a la salida de la nariz y la boca.

La faringe es una cavidad en forma tubular que une la laringe con las cavidades bucal y nasal, y que suele dividirse en tres partes: faringe laríngea, faringe bucal (boca) y faringe nasal, las dos últimas separadas por el velo del paladar. El volumen de la faringe laríngea puede ser modificado por los movimientos de la laringe, la lengua y la epiglotis mientras que el volumen de la faringe bucal se modifica por el movimiento de la lengua. La faringe nasal y las restantes cavidades nasales forman, desde el punto de vista de su acción sobre el flujo de aire procedente de la faringe, un resonador que puede o no conectarse al resonador bucal mediante la acción del velo del paladar. Según el resonador nasal esté o no conectado, el sonido será nasal u oral, respectivamente.

Si hacemos una descripción de la cavidad bucal, podemos señalar las siguientes partes:

- Los labios en el extremo.
- Los dientes.
- La zona alveolar, entre los dientes y el paladar duro.
- El paladar, en el que a su vez, y de forma simplificada, podemos distinguir el paladar duro y el paladar blando o velo.

La raíz de la lengua forma la pared frontal de la faringe laríngea, y sus movimientos le permiten modificar la sección de la cavidad bucal (movimiento vertical), adelantar o retrasar su posición frente a la de reposo (movimiento horizontal), así como poner en contacto su ápice o la parte trasera con alguna zona del paladar.

El movimiento de los labios también interviene en la articulación, pudiendo ser de apertura o cierre y de protuberancia, alargando en este último caso la cavidad bucal.

De los movimientos de los órganos supraglotales surgen los distintos modos de articulación de los posibles sonidos emitidos por un locutor. En la mayor parte de los casos es un órgano el que se mueve (activo) y otro contra el que se efectúa la articulación (pasivo). Según la pareja de órganos activo/pasivo que tengamos, tendremos una serie de posibles articulaciones.

2.1.3 Percepción del habla

La capacidad de comprender el lenguaje oral se deriva del funcionamiento de un conjunto muy complejo de procesos perceptivos, cognitivos y lingüísticos que permiten al oyente recuperar el significado de un enunciado cuando lo oye.

La percepción puede verse como un proceso que une la onda acústica y su representación conceptual por medio de una serie de niveles:

- Estructura acústica.
- Habla.
- Estructura fonética.
- Fonología.
- Estructura superficial (información fonética).
- Sintaxis.
- Estructura profunda (información sintáctica).
- Semántica.
- Representación conceptual.

Además de las diferencias en la señal, hay también diferencias marcadas en cómo un oyente procesa los sonidos de habla y los sonidos de no habla. Para los sonidos de habla: responde ante ellos como entidades lingüísticas más que como acontecimientos auditivos. El oyente aprovecha su *background* lingüístico para categorizar y etiquetar las señales de habla.

El problema fundamental es determinar cómo el estímulo acústico, que varía de manera continua, se convierte en una secuencia de unidades lingüísticas discretas de forma que sea posible recuperar el mensaje. Aunque la señal de habla sea de calidad pobre o distorsionada, el proceso de percepción se realiza perfectamente. Esto se debe a que el habla es una señal altamente estructurada y redundante de modo que las distorsiones no afectan a la inteligibilidad. La percepción también es posible porque el oyente tiene dos tipos de información disponibles, el contexto del habla (conocimiento pragmático) y el conocimiento de la lengua (sintaxis, semántica y fonología).

El mecanismo físico de la percepción del habla, al igual que la audición, se realiza por medio de dos órganos fundamentales, el Sistema auditivo periférico y el Sistema nervioso central auditivo.

El Sistema auditivo periférico es lo que vulgarmente se llama oído. En la figura 2.11 pueden observarse las 4 partes en las que se divide el sistema auditivo: oído externo, oído medio, oído interno y el Sistema nervioso central auditivo.

Los modos de funcionamiento son los siguientes:

- **Oído externo**: funciona por vibración del aire. Canaliza la energía acústica y consiste de la parte externa visible y el canal aditivo externo, de aproximadamente 2.5 cm, a través del cual viaja el sonido.

- **Oído medio**: funciona por movimiento mecánico de los huesecillos. Transforma la energía acústica en energía mecánica, transmitiéndola hasta el oído interno.

- **Oído interno**: primero el funcionamiento mecánico, por el movimiento del estribo, luego hidrodinámico por el movimiento de los líquidos interiores a la cóclea y finalmente electroquímico. Aquí se realiza la definitiva transformación de la energía mecánica en impulsos eléctricos.

- **Sistema nervioso central auditivo**: el funcionamiento es electroquímico, el movimiento de las células ciliadas provocan una reacción química que a su vez genera un impulso eléctrico.

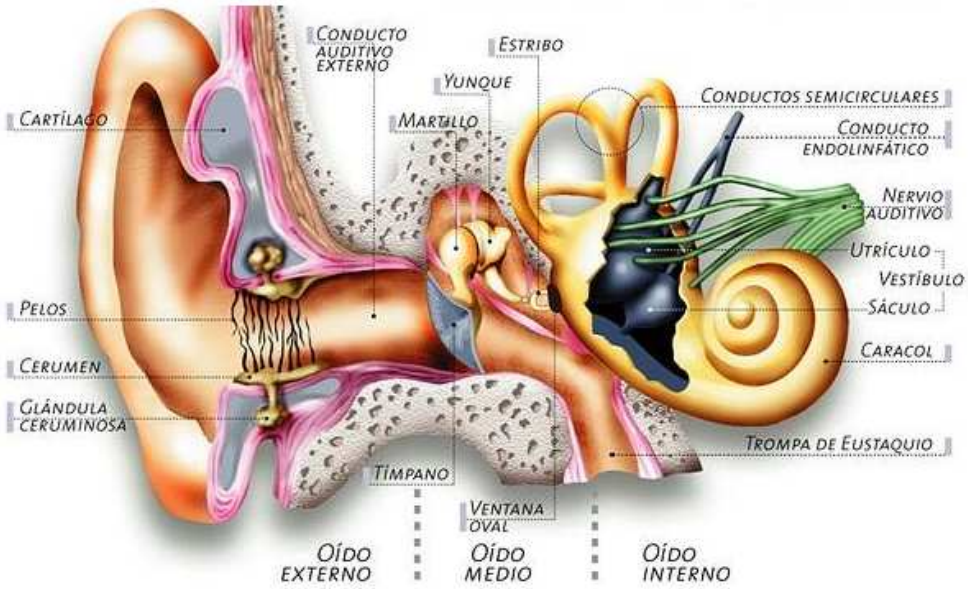

Figura 2.11 - Estructura del sistema periférico auditivo.

Cuando el sonido llega al oído, las ondas sonoras son recogidas por el pabellón auricular (o aurícula). El pabellón auricular, por su forma helicoidal, funciona como una especie de "embudo" que ayuda a dirigir el sonido hacia el interior del oído. Sin la existencia del pabellón 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 lograría rodear la cabeza y continuar su camino (difracción).

Hay que tener en cuenta que el pabellón auricular humano es mucho menos direccional que el de otros animales (como los perros) que poseen un control voluntario de su orientación.

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.

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.

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 que empuja a la ventana 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, penetra en el interior de la cóclea, la cual se comunica directamente con el nervio auditivo, conduciendo una representación del sonido al cerebro. La cóclea es un tubo en forma de espiral (de 3.5 cm aproximadamente). La espiral es dividida longitudinalmente por la membrana basilar en dos cámaras que contienen líquido linfático.

La cóclea puede ser aproximada como un banco de filtros. 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. Por ello los investigadores emprenden trabajos psicoacústicos experimentales para obtener las escalas de frecuencias que modelen la respuesta natural del sistema de percepción humano.

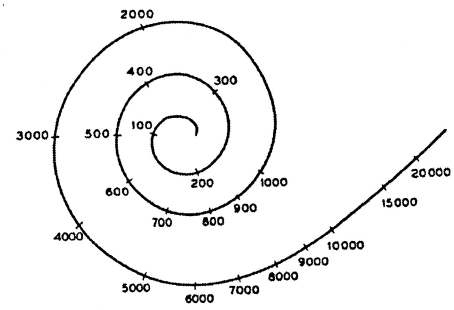

Figura 2.12 - Distribución de frecuencias en la cóclea.

El comportamiento de la cóclea como analizador en frecuencia puede resumirse en dos características:

1. La componente del espectro de la señal sonora es procesada por más de un receptor auditivo.
2. Análogamente, cada receptor auditivo procesa diversas componentes del espectro de la señal.

La forma del patrón de excitación provocado por un tono puro ilustra bien estas dos características, e indica que la selectividad en frecuencia del sistema auditivo no es infinita.

AT&T Bell Labs ha contribuido de manera muy influyente en los descubrimientos en audición, tales como banda crítica y el índice de articulación [E. Campbell, 1997]. El trabajo de Fletcher (1940) apunta a la existencia de bandas críticas en la respuesta de la cóclea. Así, el ancho de banda crítico puede interpretarse como una medida de la selectividad frecuencial del oído.

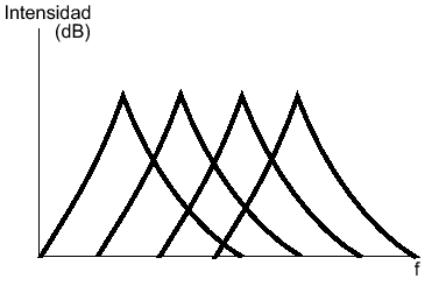

Figura 2.13 - Esquema de las bandas críticas del sistema auditivo humano.

Una característica fundamental del sistema auditivo humano es su capacidad de resolución de frecuencia e intensidad. En este aspecto, es fundamental el concepto de banda crítica. Una forma de entender el funcionamiento del sistema auditivo es suponer que contienen una serie o banco de filtros paso banda solapados conocidos como filtros auditivos (Fletcher (1940) citado en (Moore, 1998)). Estos filtros se producen a lo largo de la membrana basilar y tienen como función aumentar la resolución de frecuencia de la cóclea y así incrementar la habilidad de discriminar entre distintos sonidos. Este banco de filtros no sigue una configuración lineal, y el ancho de banda y morfología de cada filtro depende de su frecuencia central. El ancho de banda de cada filtro auditivo se denomina banda crítica (Fletcher (1940) citado en Gelfand

2004). Las bandas críticas, esquematizadas en la figura 2.13, son rangos de frecuencia dentro de los cuales un sonido bloquea o enmascara la percepción de otro sonido. Las bandas críticas conceptualmente están ligadas a lo que sucede en la membrana basilar, ya que una onda que estimula la membrana basilar perturba la membrana dentro de una pequeña área más allá del punto de primer contacto, excitando a los nervios de toda el área vecina. Por lo tanto, las frecuencias cercanas a la frecuencia original tienen mucho efecto sobre la sensación de intensidad del sonido. La intensidad percibida no es afectada, en cambio, en la presencia de sonidos fuera de la banda crítica. Es importante destacar aquí que el concepto de banda crítica es una construcción teórica y no algo físicamente comprobado.

Banda crítica	Frec. central	Ancho de banda	Frec. min.	Frec. máx.
(Bark)	(Hertz)	(Hertz)	(Hertz)	(Hertz)
1	50	-	-	100
2	150	100	100	200
3	250	100	200	300
4	350	100	300	400
5	450	110	400	510
6	570	120	510	630
7	700	140	630	770
8	840	150	770	920
9	1000	160	920	1080
10	1170	190	1080	1270
11	1370	210	1270	1480
12	1600	240	1480	1720
13	1850	280	1720	2000
14	2150	320	2000	2320
15	2500	380	2320	2700
16	2900	450	2700	3150
17	3400	550	3150	3700
18	4000	700	3700	4400
19	4800	900	4400	5300
20	5800	1100	5300	6400
21	7000	1300	6400	7700
22	8500	1800	7700	9500
23	10500	2500	9500	12000
24	13500	3500	12000	15500
25	18775	6550	15500	22050

Tabla 2.1 - Escala de Bark para estimación de las bandas críticas del sistema auditivo.

El sistema de audición lleva a cabo un análisis espectral de sonidos dentro de sus componentes de frecuencia. La cóclea actúa como si estuviese compuesta de filtros superpuestos con un ancho de banda igual al ancho de banda crítico. Una inquietud que surge de inmediato es preguntarse cuantas bandas críticas existen en el sistema auditivo y cuál es la frecuencia central de cada una. Con objeto de aproximarse a la sensibilidad del oído humano, que no tiene una respuesta lineal, existen diferentes escalas. Existe una escala de medición de las bandas críticas llamada escala de Bark, la cual se detalla en la tabla 2.1. La escala tiene un rango del 1 al 24 y corresponde a las primeras veinticuatro bandas críticas del sistema auditivo. Esta escala tiene relación con la escala Mel, que será explicada más adelante.

Nótese que las bandas críticas del oído son continuas, y un tono de cualquier frecuencia audible siempre encuentra una banda crítica que incluye dicha frecuencia. La frecuencia Bark, Z_o, puede ser expresada en términos de la frecuencia (en KHz) como:

$$Z_c = 13 arctg(0.76 f) + 3.5 arctg(\frac{f}{75})^2$$

$f[KHz]$

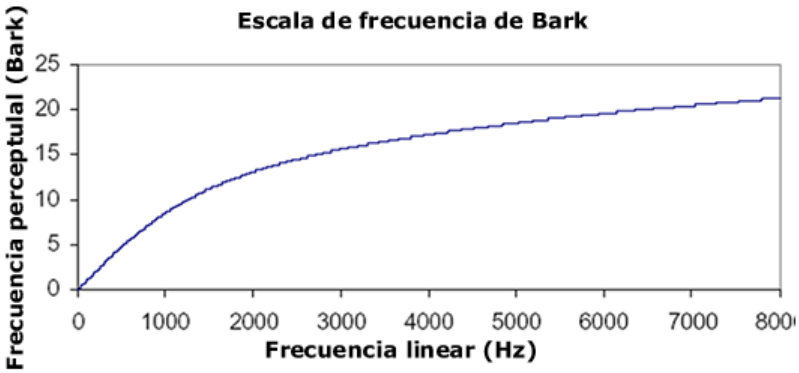

Figura 2.14 - Representación de la escala Bark.

Al igual que sucede con la intensidad perceptual, resulta útil contar con una escala perceptual de frecuencias que pueda representar de manera más fidedigna nuestra percepción de la frecuencia de un sonido. Esta escala se conoce como escala Mel, y fue propuesta por Stevens, Volkman y Newmann en 1937. El nombre Mel deriva de melodía, como una forma de explicitar que se trata de una escala basada en comparaciones entre frecuencias.

La escala Mel se construye equiparando un tono de 1000 Hz a 40 dBs, por encima del umbral de audición del oyente, con un tono de 1000 Mels. Sobre los 500 Hz, los intervalos de frecuencia espaciados exponencialmente son percibidos como si estuvieran espaciados linealmente. En consecuencia, sobre este punto, cuatro octavas en la escala lineal de frecuencias medida en Hz se comprimen a alrededor de dos octavas en la escala Mel.

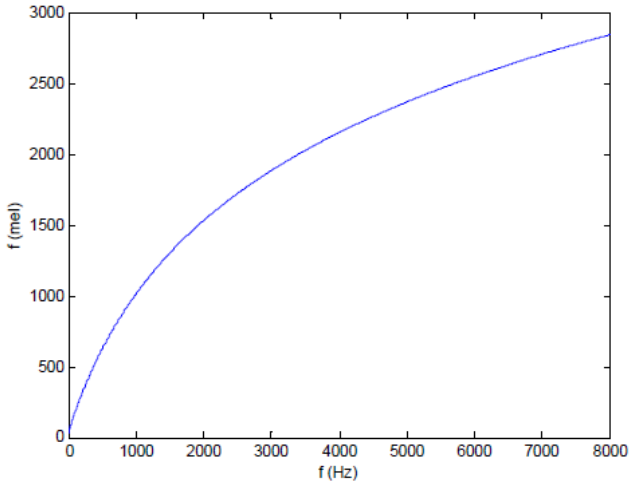

Figura 2.15 - Representación de la escala Mel.

La escala Mel ha sido ampliamente utilizada en modernos sistemas de reconocimiento de habla y puede ser aproximada en función de la frecuencia lineal como:

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

Donde f representa la frecuencia en escala lineal y $\hat{f}$ la frecuencia en escala Mel.

Según se ha explicado, la cóclea es un analizador de frecuencias donde el sonido se distribuye a través de multitud de filtros superpuestos con un ancho de banda igual al ancho de banda crítico. La mayoría de los modelos psicoacústicos realizan el intento de replicar este proceso en un esfuerzo por medir el efecto del sonido en el interior del sistema auditivo. De este estudio surge otra escala conocida como *Equivalent Rectangular Bandwidth* (ERB).

El ERB de un filtro paso banda se define como la anchura de un rectángulo cuya altura es el máximo de la respuesta en magnitud del filtro y cuyo área es igual a la respuesta de dicho filtro. La idea de esta medida es, por tanto, reemplazar una banda crítica por un rectángulo equivalente cuya área sea la misma, de manera que permita en su interior la misma cantidad de energía.

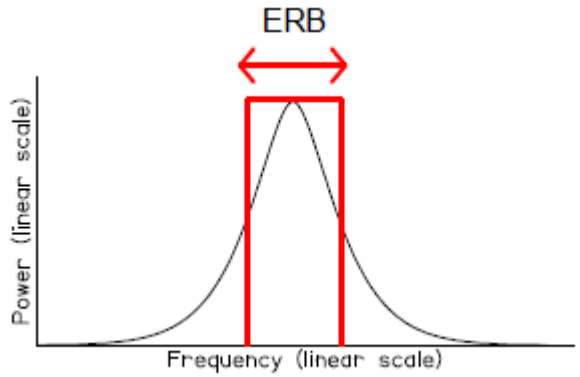

Figura 2.16 – Representación del ERB.

Para niveles de potencia moderados, podemos aproximar el ERB por la siguiente expresión:

$$ERB(f_c) = 24.7(4.37\frac{f_c}{1000} + 1)$$

Donde la expresión de ERB está en Hz y f_c es la frecuencia de centro.

Durante años se ha tratado de establecer una relación entre frecuencia y ancho de banda crítica. Al ser una medida psicoacústica, el ERB puede ser convertido a una escala que relacione la frecuencia y muestre la posición en el filtro auditivo a través de la membrana basilar. De esta forma, Moore y Glasberg (1990) introdujeron una aproximación al ancho de banda crítica utilizando la escala ERB:

$$\mathrm{ERB}(f_c) = 6.23 f_c^2 + 93.39 f_c + 28.52 \ \mathrm{Hz}$$

Donde f_c es la frecuencia de centro en KHz.

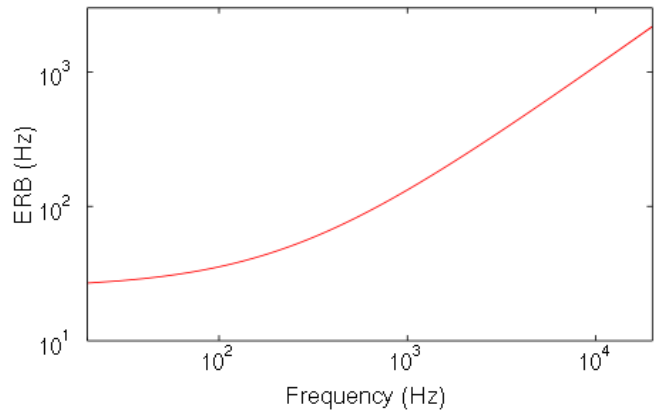

Figura 2.17 - ERB relacionado con la frecuencia de centro de acuerdo a la fórmula de Moore y Glasberg.

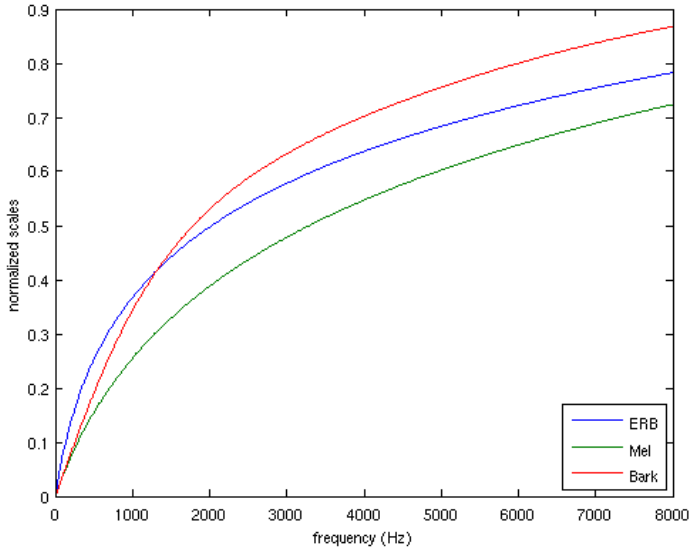

Figura 2.18 - Comparación entre la escala ERB, Mel y Bark.

2.2 Estado del arte en reconocimiento fonético

Los sonidos del habla pueden ser estudiados desde diferentes puntos de vista, articulatorio, acústico, fonético y perceptual. En esta sección se describirán desde el punto de vista fonético, acústico y articulatorio, es decir, se realizará un descripción sobre cómo se relacionan las características lingüísticas de los sonidos a posiciones y movimientos de los órganos fonatorios, así como la relación entre los fonemas y sus realizaciones acústicas interpretando la señal de voz como la salida del proceso de producción.

2.2.1 Fono y fonema

Antes de abordar propiamente el estudio de la fonética y la fonología, es conveniente definir primero los términos lengua y habla:

- La **lengua** es un modelo general y constante que existe en la conciencia de todos los miembros de una comunidad lingüística determinada, constituyendo el sistema de comunicación verbal de la misma. Es abstracto y supraindividual (compuesto por reglas).

- El **habla** es la realización concreta de la lengua (en un momento y lugar determinados) por parte de cada uno de los miembros de esa comunidad lingüística (realizaciones).

El habla puede verse como una secuencia de unidades básicas de sonidos o fonemas. Los fonemas son unidades teóricas, postuladas para estudiar el nivel fonético-fonológico de una lengua humana. Son unidades lingüísticas abstractas y no pueden observarse directamente en la señal de voz. Un mismo fonema se aplica a muchos sonidos ligeramente diferentes llamados realizaciones del fonema o alófonos. Desde un punto de vista estructural, el fonema pertenece a la lengua, mientras que el sonido pertenece al habla. La palabra <casa>, por ejemplo, consta de cuatro fonemas (*/k/, /a/, /s/, /a/*). A esta misma palabra también corresponden en el habla, acto concreto, cuatro sonidos, a los que la fonología denominará alófonos, y estos últimos pueden variar según el sujeto que lo pronuncie. La distinción fundamental de los conceptos fonema y alófono está en que el primero es una huella psíquica de la neutralización de los segundos que se efectúan en el habla.

Los fonemas no son sonidos con entidad física, sino abstracciones mentales o abstracciones formales de los sonidos del habla. Entre los criterios para decidir qué constituye o no un fonema se requiere que exista una función distintiva: son sonidos del habla que permiten distinguir palabras en una lengua. Así, los sonidos /p/ y /b/ son fonemas del español porque existen palabras como /pata/ y /bata/ que tienen significado distinto y su pronunciación sólo difiere en relación con esos dos sonidos. De esta forma, podemos decir que fonema es una unidad fonológica diferenciadora, indivisible y abstracta.

- **Diferenciadora:** porque cada fonema se delimita dentro del sistema por las cualidades que se distinguen de los demás y además es portador de una intención significativa especial. Por ejemplo, /k-o-t-a/ y /b-o-t-a/ son dos palabras que se distinguen semánticamente debido a que /k/ se opone a /b/ por la sonoridad.

- **Indivisible:** no se puede descomponer en unidades menores. Por ejemplo, la sílaba o el grupo fónico sí pueden fraccionarse. Un análisis pormenorizado del fonema revela que está compuesto por un haz de diversos elementos fónicos llamados rasgos distintivos, cuya combinación forma el inventario de fonemas. El inventario de rasgos distintivos es asimismo limitado y viene a constituir una especie de tercera articulación del lenguaje.

- **Abstracta:** no son sonidos, sino modelos o tipos ideales de sonidos. La distinción entre sonido y fonema ha sido un gran logro en los últimos tiempos en la lingüística.

Podemos clasificar los fonemas atendiendo a dos criterios: modo de articulación y punto de articulación. En el castellano se definen 24 fonemas que se clasifican en la siguiente tabla de dos entradas (tabla 2.2) atendiendo a los dos criterios enunciados. Así mismo, se indica el carácter sonoro (SN) o sordo (SR) del fonema.

Punto de articulación	Abierto		Labiales				Dentales		Alveolares		Palatales		Velares		Glotales	
Modo de Articulación			Bilabiales		Labiodentales											
	SN	SR	SN	SR	SN	SR	SN	SR	SN	SR	SN	SR	SN	SR	SN	SR
Plosivas			b	p					d	t			g	k		
Nasales			m						n			ñ				
Laterales									l			ll				
Fricativas						f				s	y			x		
Vibr simple									r							
V. Comp.									r							
Africadas												c				
Vocales	a										e,i		o,u			
Semivocales			w								j					

Tabla 2.2 - Clasificación de los fonemas del castellano.

Las vocales en castellano y en la mayoría de idiomas no se suelen clasificar de la manera anterior, sino que responden a una clasificación más sencilla atendiendo a la posición de la lengua (anterior, media o posterior) y a la abertura de la boca (cerradas, medio cerradas o abiertas), tal y como se ilustra en la siguiente tabla.

Posición de la lengua / Abertura de la boca	ANTERIORES	CENTRALES	POSTERIORES
CERRADAS	i		u
MEDIO CERRADAS	e		o
ABIERTAS		a	

Tabla 2.3 - Clasificación de los fonemas vocálicos.

Un sonido o fono se caracteriza por una serie de rasgos fonéticos y articulatorios, el número de dichos rasgos y la identificación de los mismos es tarea de la fonética. Un fono es cualquiera de las posibles realizaciones acústicas de un fonema.

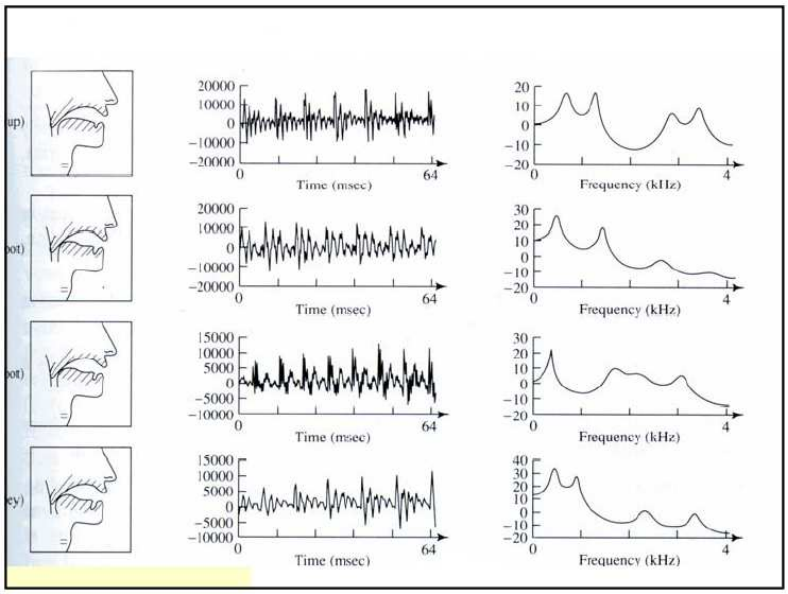

Figura 2.19 - Forma de onda y espectro de algunas vocales [*Deller et Al. 1993*].

La fonología, en cambio, no necesariamente trata entes claramente distinguibles en términos acústicos. Como realidad mental o abstracta, un fonema no tiene porqué tener todos los rasgos fonéticos especificados. Por ejemplo, en diversas lenguas la aspiración es relevante para distinguir pares mínimos, pero un fonema del español puede pronunciarse más o menos aspirado según el contexto y la variante lingüística del hablante, pero en general no está especificado el grado de aspiración. En cambio, en otras lenguas, como el chino mandarín o el coreano, un fonema tiene predefinido el rasgo de aspiración.

Dada la distinción entre fonema y fono, existe otra forma de concebir un fonema como una especificación incompleta de rasgos fonéticos. Esta relación es de hecho equivalente a la del fonema como conjunto de fonos: el fonema sería el conjunto de rasgos fonéticos comunes a todos los fonos que forman la clase de equivalencia del fonema.

Fijado un conjunto de rasgos fonéticos se pueden definir los sonidos de la lengua. En principio no hay límite a lo fina que pueda ser la distinción que establecen estos rasgos. Potencialmente la lista de sonidos puede hacerse tan grande como se quiera si se incluyen más y más rasgos, sin embargo, el número de fonemas es un asunto diferente, puesto que muchos de los anteriores sonidos serán equivalentes desde el punto de vista lingüístico. Un sistema fonológico es un par $F = (F(R))$, donde F es un inventario de fonemas abstractos definidos por unos pocos rasgos del conjunto total (las lenguas naturales oscilan entre 1 o 2 decenas hasta 4 o 5 decenas de fonemas), y R es el conjunto de reglas que en función del contexto relativo de aparición de los fonemas definen totalmente los rasgos fonéticos, así el conjunto de reglas puede pensarse como una aplicación del conjunto de secuencias admisibles de fonemas al conjunto de secuencias admisibles de sonidos: R: $P_0(F) \rightarrow P_0(S)$, donde $P_0(F)$ y $P_0(S)$ representan el conjunto de secuencias finitas de fonemas y el conjunto de secuencias finitas de sonidos.

2.2.2 Creación de modelos fonéticos

Los fonemas, tal y como se ha citado en la sección anterior, representan un nivel superior de fragmentación de palabras. La modificación de un fonema puede cambiarle el sentido a la misma. Hay que

distinguirlo de alófono, que es cada una de las pronunciaciones reales del modelo ideal que representa el fonema.
En función del grado de resolución que se quiera que presenten las unidades fonéticas se pueden obtener modelos más o menos específicos. La manera en que se hace la división en unidades fonéticas depende del contexto donde el alófono se localice:

- *Monofonemas:* Son unidades totalmente libres de contexto. Un monofonema tiene en cuenta todas las posibles realizaciones de un fonema independientemente de sus vecinos.
- *Bifonemas:* Son unidades que dependen sólo de uno de sus contextos, ya sea éste el derecho (bifonema derecho) o el izquierdo (bifonema izquierdo).
- *Trifonemas:* Son unidades que dependen de ambos contextos a la vez.
- *Trifonemas generalizados:* Dado que el número de unidades va aumentando al ir considerando más detenidamente la posición de los fonemas en su contexto, puede llegar a ser tan elevado que su entrenamiento no fuera posible. Surge el trifonema generalizado como un primer nivel de compartición, en el que varios trifonemas cercanos se agrupan para reducir el número de modelos y que el entrenamiento de éstos sea mejor.

Así, un reconocedor necesita disponer de un conjunto de unidades que permita construir cualquier palabra o frase a partir de su concatenación. Los fonemas representan este conjunto completo y reducido de unidades a partir de las cuales se puede generar cualquier palabra. Estas unidades pueden modelarse con mayor o menor resolución en función del contexto a considerar. Es necesario especificar para un determinado idioma, en este caso para el castellano, cómo son estas unidades. Además, las diferentes alternativas de entrenamiento en función del acceso a las bases de datos será un factor determinante a la hora de establecer el conjunto de unidades a entrenar.

La creación de modelos acústicos, para su posterior uso en reconocimiento de habla, se crean en dos fases: la primera es la extracción de características de la señal de voz, y la segunda es usar esas características para identificar los fonemas:

- **Extracción de las características del sonido del habla.** Por semejanza con el funcionamiento del sistema humano, la extracción de esas características, que llamaremos parámetros, se realiza en el dominio de la frecuencia. Asignación de los parámetros extraídos a las representaciones discretas de nuestro diseño (fonemas, trifonemas, palabras…) correspondientes, con el objetivo de crear un modelo para cada una de las representaciones discretas que las identifique. Tanto las técnicas para la extracción de parámetros como los distintos modelos, son definidos en las siguientes secciones. En particular, la parametrización que se ha llevado a cabo en este proyecto es de tipo *Human Factors Cepstral Coefficient* (HFCC), la cual se explicará en profundidad en la sección 2.5.3.

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

    - *HMM (Hidden Markov Model)*: las aproximaciones estadísticas toman como referencia el modelo estocástico de los datos. Se basa en la creación de modelos de fonemas en estados. Hasta la fecha este método es el que mejores resultados proporciona y el más utilizado. Éste ha sido el procedimiento seguido en este trabajo, por ello, se explicará con más detenimiento en la sección 2.4.

- *DTW (Alineamiento Temporal Dinámico)*: consiste en alinear de forma temporal los parámetros del archivo de test y los parámetros de los modelos, obteniendo la función que alinea a ambos, eligiendo la función de menor coste posible para dicha adaptación. En la siguiente imagen se ve como representar la función de adaptación.

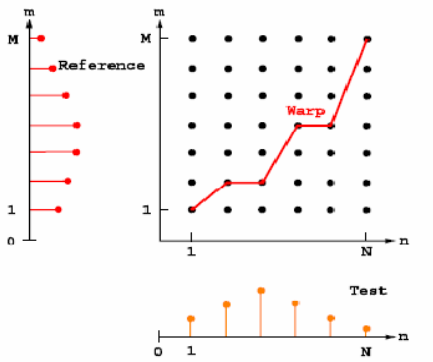

Figura 2.20 - Función de adaptación de DTW.

- *VQ (Cuantificación vectorial)*: consiste en representar las características de los fonemas como un espacio vectorial de dimensión el número de parámetros, de forma que al fonema a reconocer se le asigna el vector cuya distancia a él sea mínima. Por tanto, los fonemas quedarán representados por unos vectores determinados (centroides) de forma que todos los puntos que caigan en una zona determinada se asignarán a dicho vector. Esto puede verse en la figura 2.21 en la que el espacio es bidimensional (el número de parámetros que se emplean son dos) y en el que los puntos verdes son los vectores de test, mientras que los rojos son los vectores a los que se asignan (obtenidos de forma óptima durante el entrenamiento), siendo cada una de las regiones los fonemas posibles.

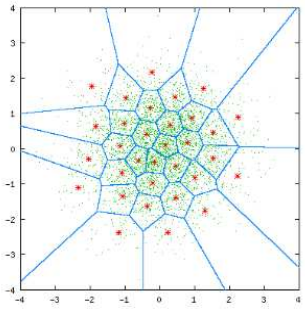

Figura 2.21 - VQ bidimensional.

2.3 Reconocimiento automático de voz

Esta sección hace un encuadre científico-tecnológico del reconocimiento automático del habla, definiendo los tipos de reconocedores, sus objetivos y las etapas de procesado de la señal de voz necesarias para implantarlos. Las diferentes posibilidades metodológicas son analizadas para cada una de dicha etapas. Se termina con una definición de los criterios de evaluación cuantitativa en los sistemas de RAH.

2.3.1 Planteamiento del problema: RAH

2.3.1.1 La comunicación oral

El proceso de comunicación oral es uno de los comportamientos humanos que más ampliamente ha sido estudiado por disciplinas como la biología, la física y la lingüística. La figura 2.22 muestra la visión esquemática de los elementos que intervienen en dicho proceso y del flujo que lo origina: en la mente del emisor se crea un mensaje que, por medio de impulsos nerviosos a los nervios motores que activan los músculos vocales, se traduce en un discurso de palabras transmitido a través de una señal acústica. La señal acústica es recibida por el receptor que lleva a cabo el proceso inverso: el movimiento de la membrana basilar en el oído del receptor se convierte en impulso eléctrico que es transmitido al cerebro mediante los nervios auditivos. En el cerebro del receptor se produce el análisis y comprensión del mensaje.

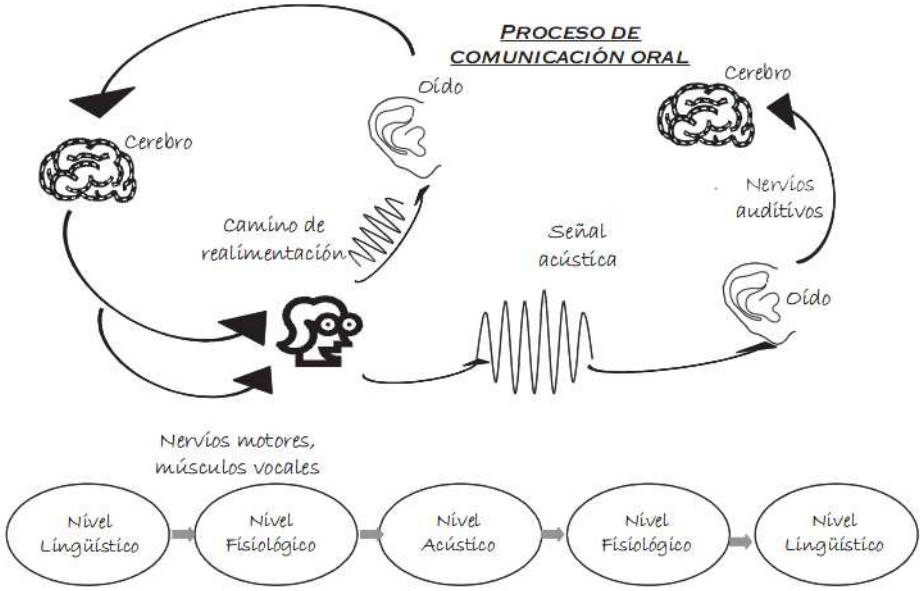

Figura 2.22 - Esquema del proceso de comunicación oral [*Rabiner y Levinson [1]*].

2.3.1.2 El reconocimiento en la comunicación entre humanos

El objetivo del reconocimiento automático del habla (*RAH*) es imitar el proceso de reconocimiento que lleva a cabo el receptor en la comunicación oral. Hay varios niveles de reconocimiento en dicho proceso humano, y los diferentes sistemas de reconocimiento automático implementan todos, algunos o sólo los más básicos dependiendo de cuáles sean su aplicación y complejidad. Se pueden distinguir ocho niveles de reconocimiento en orden de complejidad ascendente:

- **NIVEL ACÚSTICO**: la señal acústica analógica que ha enviado el emisor es recibida y traducida a un conjunto de rasgos relevantes no redundantes. En la comunicación oral, este reconocimiento se hace en el oído. Hay cuatro operaciones incluidas en este nivel, que son total o parcialmente implementadas para automatizarlo en el RAH:

- *Parametrización*: la señal analógica se transforma en una señal numérica que pueda ser tratada por la máquina digital en la que se hace el reconocimiento. Hay varios métodos de parametrización, en el dominio del tiempo y de la frecuencia, que dan lugar a diferentes parámetros de caracterización.

- *Segmentación*: determina como separar la señal analógica continua en una cadena de sonidos cuya sucesión es la señal en el tiempo. Se lleva a cabo con métodos basados en las curvas de variación de la energía o de variabilidad de la señal.

- *Extracción de la información relevante*: se busca retener sólo aquellos datos que proporcionan información útil para el reconocimiento como pueden ser los espectros de los instantes de mayor estabilidad o de los instantes de transición.

- *Información relativa a la prosodia*: estudia la variación del armónico fundamental de la voz, variación de la intensidad, y el ritmo.

- **NIVEL FONÉTICO:** la secuencia de información relevante obtenida en el nivel acústico es traducida a una secuencia de fonemas.

- **NIVEL FONOLÓGICO:** los fonemas de la lengua que hacen que el contenido fonético de las palabras se modifique en una articulación rápida o por una sucesión de términos léxicos son analizados. Las variedades dialectales son también tratadas.

- **NIVEL LÉXICO:** se identifican las palabras de la lengua en la que se produce la comunicación.

- **NIVEL SINTÁCTICO:** se detectan las reglas gramaticales que permiten describir y analizar el lenguaje, y que relacionan las palabras reconocidas a nivel léxico.

- **NIVEL SEMÁNTICO:** analiza el sentido de las palabras, buscando la comprensión del mensaje y eliminando las interpretaciones que no tengan sentido. Es el nivel de conocimiento de las palabras que da un diccionario de la lengua.

- **NIVEL PRAGMÁTICO:** estudia el sentido del mensaje recibido teniendo en cuenta el contexto de su aplicación. Reconoce la información que viene determinada por la situación en la que se produce la comunicación.

- **NIVEL PROSÓDICO:** interviene de manera paralela al resto de niveles, sin formar parte de una estructura piramidal como los demás. Este nivel detecta la información que el mensaje comunica mediante los modos de pronunciación: palabras pronunciadas con cierto nivel de insistencia para ponerlas de relieve, fronteras entre grupos de palabras, naturaleza interrogativa o declarativa de una frase, etc.

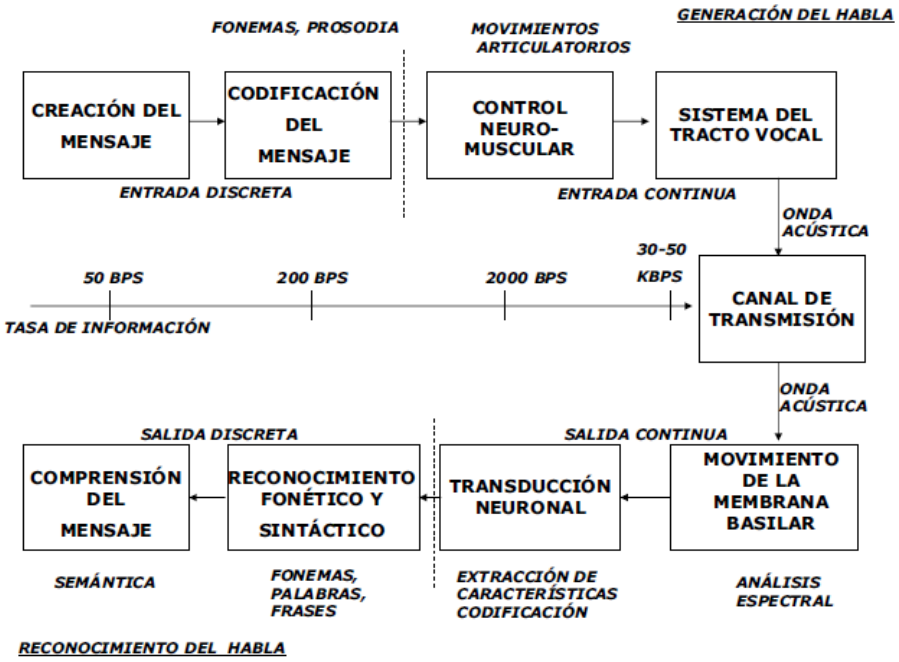

Figura 2.23 - Otra esquematización del proceso de comunicación oral [*Rabiner y Levinson [1]*].

La figura 2.23 muestra otra esquematización del modelo de producción y percepción de la voz propuesto por Rabiner y Levinson [1], en la que al igual que en la figura anterior (figura 2.22) se pueden identificar los niveles de producción y reconocimiento que se dan en la comunicación oral y las tareas necesarias para automatizar la comunicación, ya sea en la parte de producción de voz (síntesis automática de voz) o en la parte de reconocimiento (reconocimiento automático del habla).

En particular, la parte del reconocimiento del habla se automatiza implementando, en su totalidad o parcialmente, los ocho niveles de reconocimiento mencionados según la complejidad y necesidades del sistema de reconocimiento. Las posibilidades son muchas. En la tabla 2.4, se recopila una visión global de las variables que definen un sistema de reconocimiento automático del habla y el rango de valores que puede tomar:

- Hay reconocedores de palabras aisladas, de palabras conectadas y de habla continua, lo que supone un orden creciente de complejidad del reconocedor que tiene que delimitar palabras y frases.

- El habla puede ser no espontánea (leída o dirigida mediante un diálogo de opciones), o puede ser espontánea con el consiguiente incremento de la dificultad.

- El reconocedor de voz puede ser además dependiente de locutor, teniendo que discernir la información acústica para un solo hablante, puede ser adaptado al locutor, multilocutor o independiente de locutor con lo que deberá filtrar las distorsiones acústicas debidas a las peculiaridades del hablante.

- Los fines específicos o generales del reconocedor y la perplejidad y tamaño del vocabulario que reconoce, son características que aumentan la dificultad o simpleza de la tarea de reconocimiento.

- La distorsión acústica debida al ruido de canal y al ruido aditivo que acompañe a la voz incrementa la dificultad de la tarea de reconocimiento.

PARÁMETRO	RANGO
Forma de hablar	Palabras aisladas ↔ Habla continua
Estilo del habla	Texto leído ↔ Habla espontánea
Adaptación	Dependiente de locutor ↔ Independiente de locutor
Tamaño del vocabulario	Pequeño (<20 palabras) ↔ Grande (> 20.000 palabras)
Modelo de lenguaje	Estados finitos ↔ Dependiente de contexto
Perplejidad	Pequeña (<10) ↔ Grande(>100)
SNR	Alta (>30) ↔ Baja (<10)
Transductor	Micrófono de cancelación de eco ↔ Teléfono

Tabla 2.4 - Parámetros que caracterizan el sistema de reconocimiento.

Independientemente de las características que se acaban de describir, los bloques conceptuales necesarios para implementar un reconocedor son los que aparecen en la figura 2.24:

- **Datos de entrenamiento.** Son la información con la cual el sistema se estrena. Es un conjunto de datos que debe ser lo suficientemente equilibrado como para que el reconocedor *aprenda* a reconocer ese vocabulario.

- **Bloque de parametrización.** Las entradas de señal de voz pasarán por una etapa en la que se extraerán sus características representativas de ser clasificadas.

- **Bloque de reconocimiento.** La clasificación de los parámetros se hará usando los datos de entrada y las referencias con las que cuenta el sistema: el aprendizaje de los datos de entrenamiento, y las referencias acústicas, léxicas y de lenguaje.

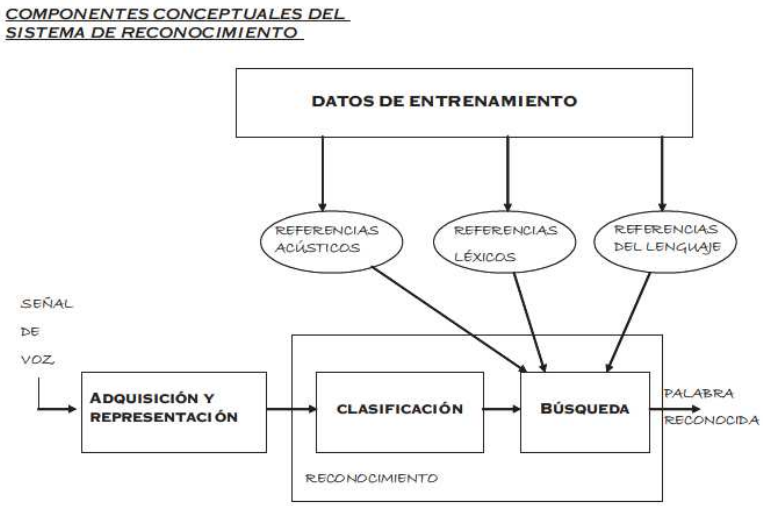

Figura 2.24- Componentes conceptuales de un sistema de reconocimiento [*Rabiner y Levinson [1]*].

2.3.2 Procesamiento de señales digitales de audio

El procesamiento digital de señales es una técnica que convierte señales procedentes de fuentes del mundo real (usualmente en forma analógica), en datos digitales que luego pueden ser analizados. Este análisis es realizado en forma digital, pues una vez que una señal ha sido reducida a valores numéricos discretos, sus componentes pueden ser aisladas, analizadas y reordenadas más fácilmente que en su primitiva forma analógica.

2.3.2.1 Señales digitales

Las señales digitales, en contraste con las señales analógicas, no varían en forma continua, sino que cambian en pasos o en incrementos discretos. Las señales en tiempo discreto son aquellas que se representan matemáticamente como una secuencia de números. Además del carácter de estar definidas en tiempo discreto, la amplitud de la señal puede ser también discreta.

2.3.2.2 Muestreo

En la mayoría de los casos, las señales en tiempo discreto surgen de tomar muestras de una señal analógica. De esta forma, el valor numérico del n-ésimo número de la secuencia es igual al valor de la señal analógica $x_a(t)$, en el instante temporal nT_s, es decir,

$$\hat{x}(n) = x_a(nT_s), \quad -\infty < n < \infty$$

La cantidad T_s se denomina período de muestreo y su inversa es la frecuencia de muestreo f_s.

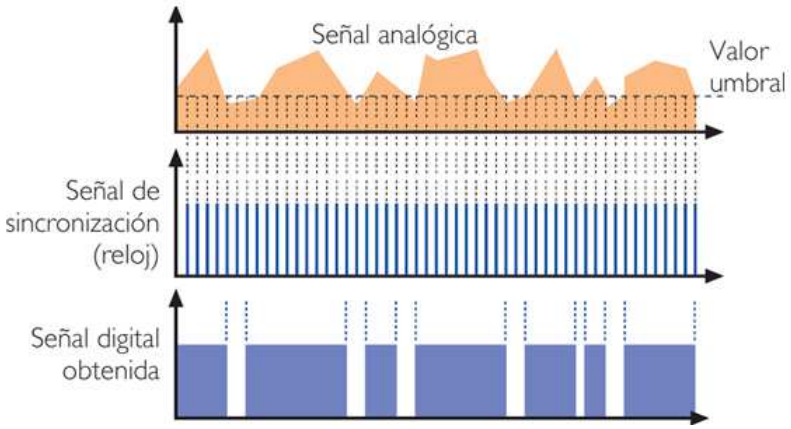

Figura 2.25 - Digitalización por muestreo de una señal analógica.

El teorema de Nyquist garantiza que, para poder reconstruir una señal a partir de sus muestras, se debe utilizar una frecuencia $N_s \geq 2\ f_N$, o sea al menos el doble de f_N. Siendo f_N la componente de más alta frecuencia de la señal.

El espectro de frecuencias del sonido audible por los humanos, es aproximadamente de 20 Hz a 20 kHz. Por esto, las señales de audio se muestrean generalmente a 44100 Hz, o sea más del doble de la máxima frecuencia audible. Éste es el caso del CD de audio. El contenido en frecuencia de las señales de voz puede abarcar hasta 15 kHz o más, pero la voz es altamente inteligible incluso con bandas de frecuencia limitadas a unos 4 kHz. Ese es el caso de los sistemas telefónicos comerciales donde la frecuencia de muestreo estándar utilizada para la voz es de 8 kHz.

En la etapa de muestreo se obtiene una señal en tiempo discreto cuyas amplitudes $\hat{x}(n)$ son valores continuos. Para digitalizar la señal resta discretizar esos valores (cuantizarlos).

2.3.2.3 Cuantización

El propósito del cuantizador es transformar la muestra de entrada $\hat{x}(n)$ en un valor $x(n)$ de un conjunto finito de valores preestablecidos. Esto se realiza redondeando los valores de las muestras hasta el nivel de cuantización más próximo.

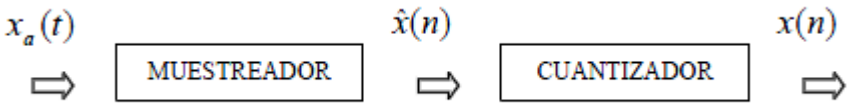

Figura 2.26 - Representación conceptual de la digitalización de una señal analógica.

La precisión de los datos dependerá del número de bits con que se codifiquen los niveles de cuantización. Por tanto, se introduce un ruido de cuantización que se asume como ruido blanco.

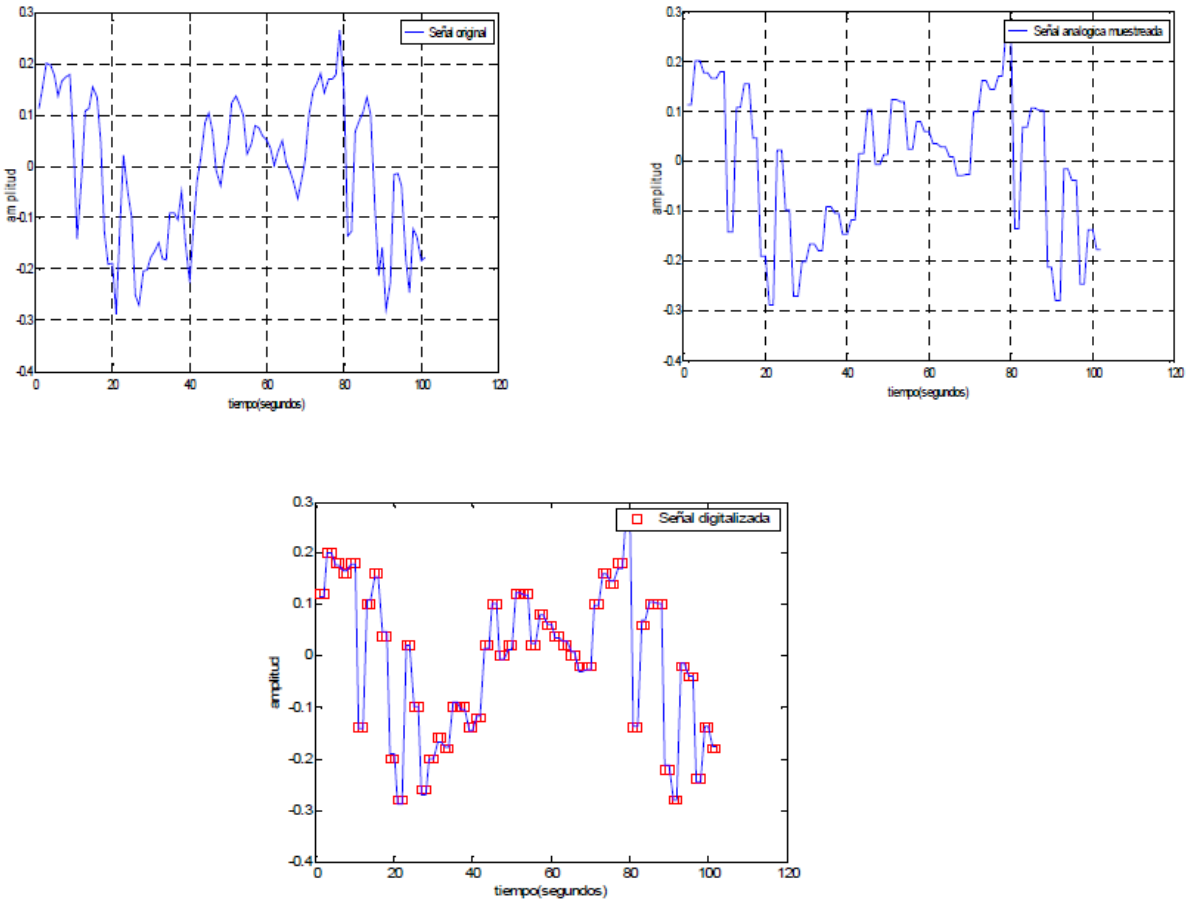

Figura 2.27 - Etapas de la digitalización. **a)** Señal original **b)** Señal muestreada con amplitudes analógicas **c)** Señal digital.

2.3.3 Parametrización de la señal de voz

El objetivo de la fase de parametrización en el proceso de reconocimiento es la extracción de información relevante de la señal acústica analógica, eliminando las redundancias y la información asociada a las fuentes de variabilidad que tiene la misma. La información relevante será aquella que permita:

- Diferenciar unos fonemas de otros. Los fonemas están caracterizados por:

  i. La envolvente espectral del fonema, determinada por los formantes que lo componen. Los formantes se definen como las frecuencias de resonancia del tracto vocal para cada fonema.
  ii. El tipo de excitación que los produce. Las vocales y consonantes sonoras están generadas mediante una excitación periódica. La frecuencia fundamental de la excitación es también una característica definitoria del fonema, aunque es variable para los diferentes hablantes y las diferentes entonaciones.
  iii. La energía de la señal. Las vocales y consonantes sonoras tienen mayor energía que las sordas, siendo la energía un buen parámetro de caracterización ya que presenta poca variabilidad para un mismo fonema una vez que ha sido convenientemente normalizada.

- Aportar datos sobre la prosodia de la frase tales como el acento, los tonos y la entonación. Esta información se obtiene analizando:

  i. Las variaciones de la frecuencia fundamental.
  ii. Las variaciones de la duración de los fonemas.
  iii. La variación en la intensidad de los fonemas diferenciados.

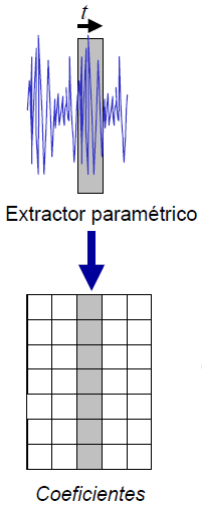

Figura 2.28 - Ilustración de la extracción de características en una señal.

Teniendo en cuenta la información expuesta como necesaria para caracterizar los fonemas y su prosodia, es razonable que la mayor parte de los sistemas de parametrización se basen en el análisis de la potencia espectral en tiempo corto [1]. Al hacer este análisis, la señal se divide en tramas lo suficientemente cortas como para poder considerar la señal cuasi-estacionaria. Siendo cuasi-estacionaria, la trama se somete a un análisis espectral y queda caracterizada por un vector de características que suele tener de 10 a 20

parámetros. La figura 2.29 muestra de manera general el proceso de parametrización con las posibles variantes en cada una de las etapas [2].

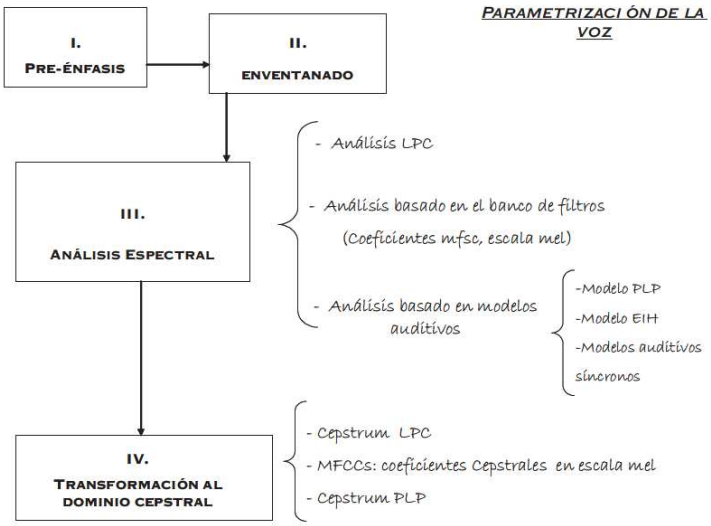

Figura 2.29 - Representación esquemática del proceso de la parametrización [*Rabiner y Levinson [1]*].

2.3.3.1 Filtro de Pre-énfasis

En primer lugar, la señal de voz muestreada pasa un filtro de pre-énfasis, típicamente un filtro FIR (*Finite Impulse Response*) de primer orden, que amplifica las altas frecuencias para compensar el efecto de los pulsos glotales y la impedancia de radiación. Generalmente, este filtro sigue la expresión:

$$H(z) = 1 - \mu \cdot z^{-1}, \text{ siendo } 0{,}95 \leq \mu \leq 0{,}98$$

El filtro de preénfasis se destina para alzar el espectro de la señal aproximadamente 20 dB por década. Hay dos explicaciones que justifican su utilización: primero, los segmentos de voz sonoros tienen una pendiente espectral negativa (aproximadamente 20 dB por década), este filtro tiende a contrarrestar esta pendiente mejorándose la eficiencia de las etapas posteriores; y, segundo, es que la audición es más sensible por encima de 1 KHz en la región del espectro. Este filtro amplifica esta zona del espectro ayudando a las etapas posteriores de análisis a modelar los aspectos más importantes del espectro de la voz.

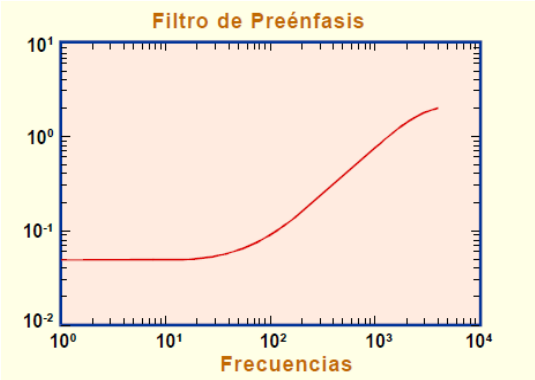

Figura 2.30 - Respuesta en frecuencia del filtro de pre-énfasis.

2.3.3.2 Enventanado

La señal de voz es un proceso aleatorio y no estacionario. Esto supone un inconveniente a la hora de analizar la señal, no obstante, es posible salvar este problema si se tiene en cuenta que a corto plazo de tiempo (del orden de *ms*) la señal es casi-estacionaria. Este hecho da lugar a un tipo de análisis donde se obtienen segmentos o tramas de la señal de pocos *ms* denominado análisis localizado. A este proceso donde se obtienen tramas o segmentos consecutivos de señal se le denomina enventanado.

El enventanado requiere que cada una de las tramas sea multiplicada por una función limitada en el tiempo de tal manera que su valor fuera de ese intervalo sea nulo. De esta forma, el enventanado consiste en agrupar las muestras de la señal *x(n)* en bloques de N elementos, y multiplicarlas por una ventana *w(n)*.

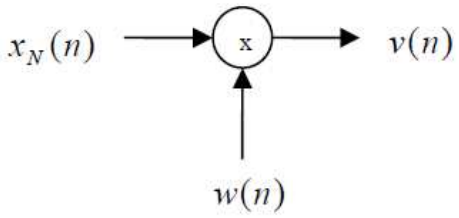

Figura 2.31 - Enventanado.

Para mantener la continuidad de la información de la señal, es muy común realizar el enventanado con bloques de muestras solapados entre sí, de esta forma no se pierden los eventos en la transición entre ventanas.

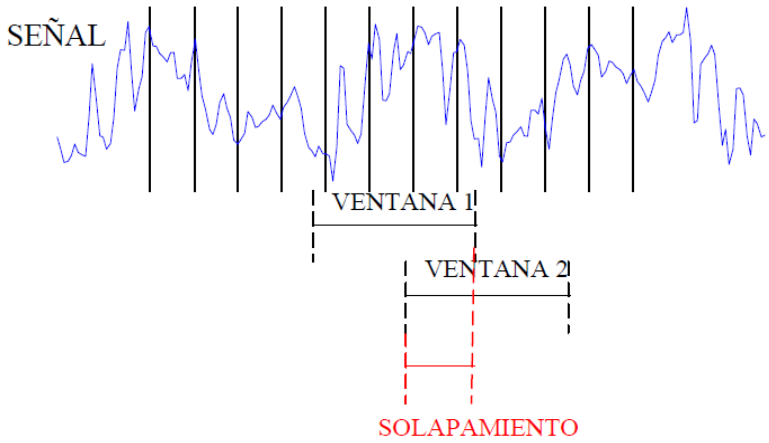

Figura 2.32 - Ilustración del solapamiento de una señal.

Cuanto más rápidamente cambien las características de la señal, más corta deberá ser la ventana para poder detectar esos cambios en el tiempo. Por otra parte, a medida que decrece la longitud de la ventana, se reduce la resolución frecuencial, es decir, la capacidad de distinguir componentes cercanas en frecuencia.

Por tanto, aparece un compromiso en la selección de la longitud de la ventana entre la resolución en tiempo y en frecuencia. Además de la longitud, se debe elegir la forma de la ventana, o más específicamente, el tipo de suavizado que se requiere en los extremos de la misma [3].

Las ventanas más utilizadas se definen para N muestras como:

- *Rectangular:*

$$w(n) = 1$$

- *Hanning:*

$$w(n) = \frac{1}{2} - \frac{1}{2} \cdot \cos\left(\frac{2\pi n}{N}\right)$$

- *Hamming:*

$$w(n) = \frac{27}{50} - \frac{23}{50} \cdot \cos\left(\frac{2\pi n}{N}\right)$$

- *Barlett:*

$$\text{Bartlett:} \begin{cases} \frac{2n}{N} & 0 < n < \frac{N}{2} \\ 2 - \frac{2n}{N} & \frac{N}{2} < n < N \end{cases}$$

- *Blackman:*

$$\text{Blackman: } w(n) = \frac{21}{50} - \frac{1}{2} \cdot \cos\left(\frac{2\pi n}{N}\right) + \frac{2}{25} \cdot \cos\left(\frac{4\pi n}{N}\right)$$

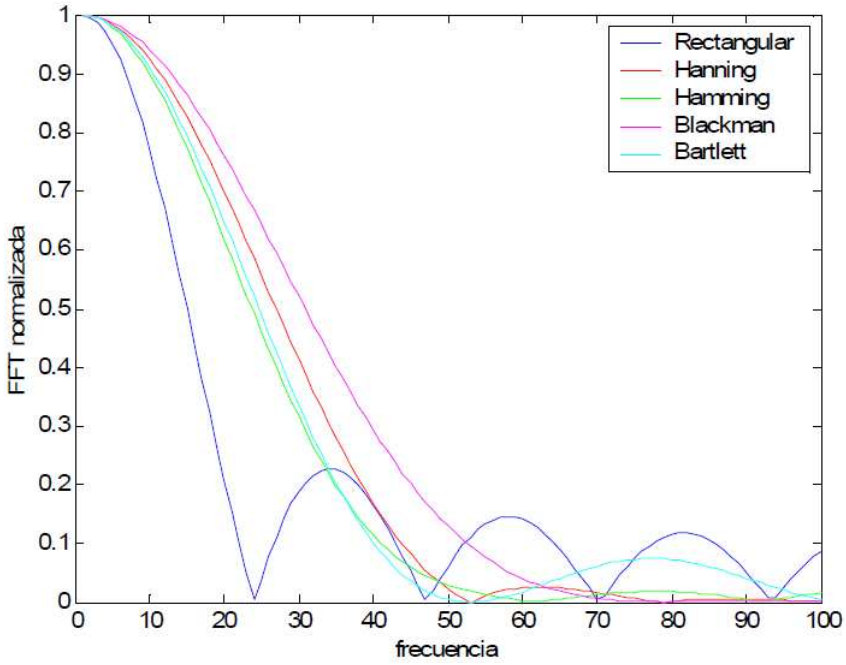

Figura 2.33 - Espectros de las distintas ventanas.

Cada ventana se caracteriza por la forma de sus lóbulos central y laterales en frecuencia. Se requiere de una ventana, que su lóbulo central sea lo más angosto posible y que los lóbulos laterales sean pequeños para tener una buena resolución en frecuencia.

La ventana rectangular posee el lóbulo central con menor ancho de banda de todos, pero sus lóbulos laterales decaen muy lentamente. Estos hacen aparecer el efecto de 'ripple' (fenómeno de Gibbs), no deseado por la distorsión armónica que generan. El resto de las ventanas tiene cada una distintas propiedades, que según la aplicación podrán ser de un modo u otro ventajosas.

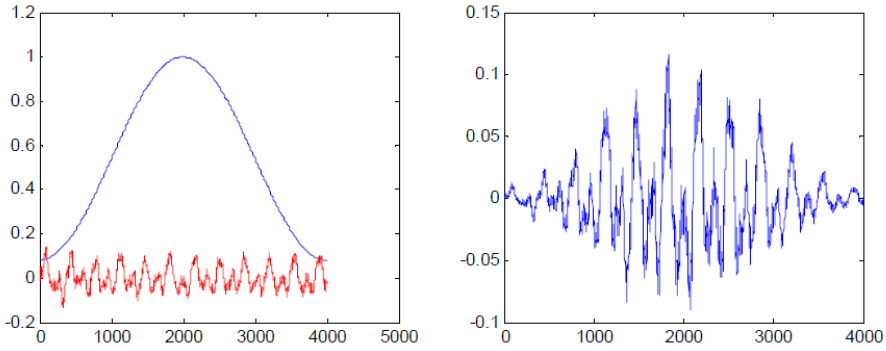

Figura 2.34 - A la izquierda, ventana de Hanning en azul y señal de audio en rojo; a la derecha la multiplicación de ambas.

2.3.3.3 Transformada discreta de Fourier

Los sistemas lineales e invariantes en el tiempo, cumplen ciertas propiedades que permiten la representación de las señales en frecuencia. Una de las propiedades es que la respuesta a secuencias sinusoidales es también sinusoidal, de igual frecuencia y con amplitud y fase determinadas por el sistema. Esta propiedad hace que las representaciones de las señales mediante sinusoides o exponenciales complejas (es decir, las representaciones de Fourier) sean muy útiles.

Para las secuencias de duración finita, se utiliza la Transformada Discreta de Fourier (DFT). Se llaman secuencias base a las exponenciales complejas que se utilizan para representar la señal.

Dada una señal en tiempo discreto *x(n)* con N muestras, su transformada *X* (*k*) está dada por:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi kn}{N}}$$

En la práctica, el costo computacional del cálculo de la DFT se reduce utilizando la transformada rápida de Fourier, FFT (*Fast Fourier Transform*), un eficiente algoritmo que permite calcular la transformada de Fourier discreta (DFT) y su inversa. La FFT es de gran importancia en una amplia variedad de aplicaciones, desde el tratamiento digital de señales y filtrado digital en general, a la resolución de ecuaciones diferenciales parciales o los algoritmos de multiplicación rápida de grandes enteros.

La evaluación directa de la fórmula anterior requiere $O(n^2)$ operaciones aritméticas. Mediante un algoritmo FFT se puede obtener el mismo resultado con sólo $O(n \log n)$ operaciones. En general, dichos algoritmos dependen de la factorización de n pero, al contrario de lo que frecuentemente se cree, existen FFTs para cualquier n, incluso con n primo.

La idea que permite esta optimización es la descomposición de la transformada a tratar en otras más simples y éstas a su vez hasta llegar a transformadas de 2 elementos donde k puede tomar los valores 0 y 1. Una vez resueltas las transformadas más simples hay que agruparlas en otras de nivel superior que deben resolverse de nuevo y así sucesivamente hasta llegar al nivel más alto. Al final de este proceso, los resultados obtenidos deben reordenarse.

Dado que la transformada discreta de Fourier inversa es análoga a la transformada discreta de Fourier, con distinto signo en el exponente y un factor 1/n, cualquier algoritmo FFT puede ser fácilmente adaptado para el cálculo de la transformada inversa.

2.3.3.4 Transformada discreta del coseno

La DCT es una transformada muy similar a la DFT (*Discret Fourier Transform*) en donde las secuencias base son cosenos y la representación de una señal real mediante esta transformada, es también real.

$$X(k) = \sum_{n=0}^{N-1} x(n) \cos\left[\frac{k\pi}{N}\left(n+\frac{1}{2}\right)\right]$$

Se utiliza en muchas aplicaciones de compresión de datos con preferencia sobre la DFT debido a una propiedad que se denomina generalmente "compactación de la energía". La DCT tiende a concentrar la mayor parte de la información de la señal en los coeficientes de baja frecuencia. Gracias a esto, se necesita un menor número de coeficientes para representarla. Esto se ejemplifica en la siguiente figura:

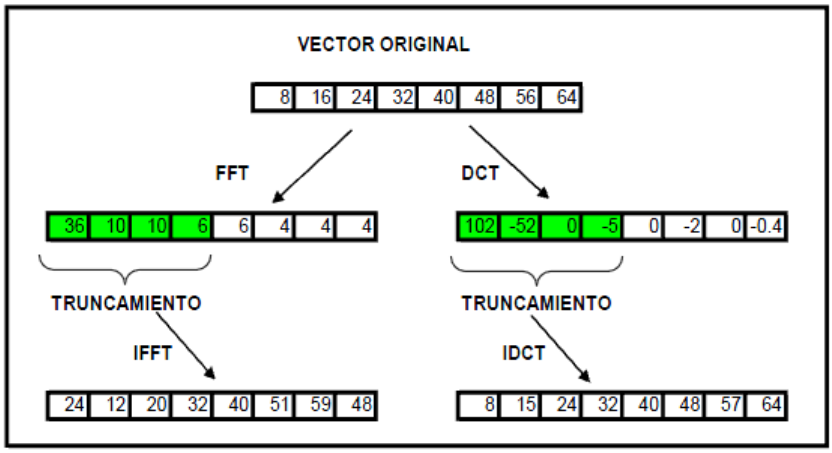

Figura 2.35 - Compactación de la DCT.

2.3.3.5 Análisis Espectral

El análisis de los segmentos del habla obtenidos se puede hacer tanto en el dominio del tiempo como en el dominio de la frecuencia. En el dominio del tiempo las magnitudes que se analizan son la energía local, la tasa de cruces por cero de la señal y su autocorrelación. Este dominio aporta un análisis de la señal rápido, sencillo y con una interpretación inmediata. Sin embargo, el análisis espectral es el utilizado por su mayor potencia para caracterizar la información de la señal de voz. Las parametrizaciones usadas en RAH se derivan en su totalidad del análisis de la potencia espectral de las tramas de voz. El análisis de la fase del espectro de frecuencias se omite debido a que los oídos son insensibles a las variaciones de la fase y en consecuencia los equipos de comunicaciones de voz y de grabación no preservan la fase original, que

también se ve alterada por factores no deseados como la acústica del entorno. El análisis de la potencia espectral se hace además en escala logarítmica por motivos prácticos:

- La escala logarítmica hace que cuando la ganancia que tiene la señal cambia, la forma del espectro de potencias se mantenga, simplemente desplazándose hacia arriba o hacia abajo.

- El filtrado lineal debido a la acústica del entorno o a variaciones en el canal, tiene un efecto convolucional en el dominio del tiempo, un efecto multiplicativo para el espectro de potencias lineal y un simple efecto de suma de una constante para los espectros logarítmicos de potencias.

- La forma de onda de la voz se puede modelar como la convolución en el dominio temporal de la excitación de una cuasi-peródica con un filtro variante en el dominio del tiempo, que está determinado por la configuración del tracto vocal para la producción de dicha señal de voz (figura 2.36). Esta configuración del tracto vocal como filtro variante en el tiempo va a ser la que nos dé la información sobre los fonemas articulados. Es deseable poder separar estas dos componentes de la forma de onda (excitación cuasi-periódica y filtro variante), siendo el dominio de la potencia espectral logarítmica el óptimo para hacerlo ya que en dicho dominio ambos componentes son aditivos.

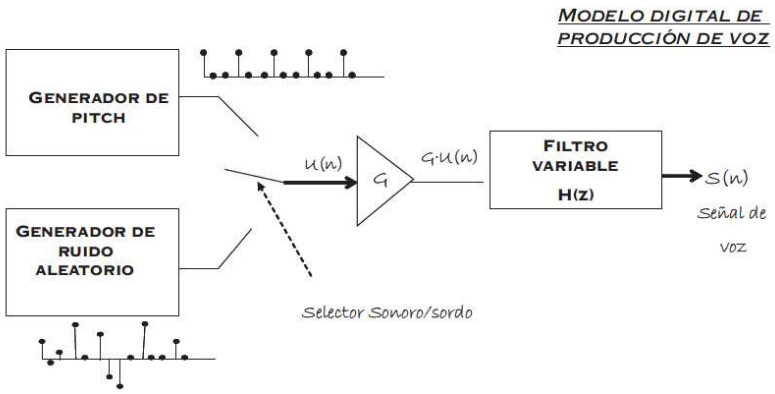

Figura 2.36 - Modelo digital de producción de voz [*Rabiner y Levinson [1]*].

La escala logarítmica presentaría problemas para valores de energía muy bajos cercanos a cero, por ello, lo normal es limitar el logaritmo a valores de unos -60dB.

Como veíamos en el esquema de la figura 2.29, hay 3 técnicas de análisis espectral que serán descritas a continuación:

- **Representaciones basadas en el modelo LPC:** (*Linear Predictive Coding*) [4],[5]. El modelo digital de producción de voz de la figura 2.36 es usado para modelar el tracto vocal como un tubo acústico a través del cual el sonido se propaga como una onda plana. Los efectos del tracto vocal en la señal de excitación son la creación de una serie de resonancias, quedando así el tracto modelado como un filtro *todo polos H*(*z*). El análisis espectral permite dar los coeficientes de dicho filtro, *H*(*z*), sin calcular el espectro explícitamente.

- **Análisis basado en el banco de filtros. Coeficientes MFSC:** (*Mel Frequency Spectral Coefficients*). Debido a que el espectro de potencias de la señal se obtiene aplicando la transformada de Fourier a las tramas de voz de ventanas que se solapan, aparecerán armónicos a frecuencias múltiplo de la frecuencia fundamental de las tramas. Este efecto se puede subsanar agrupando los conjuntos de

componentes cercanos en unas 20 bandas de frecuencias antes de hacerles el logaritmo de la potencia. Cada filtro hará un promedio pesado de las componentes espectrales presentes en su banda, caracterizado el tracto de resolución perceptual del oído humano, haciendo que las bandas que abarcan los filtros sean más anchas para frecuencias superiores a 1 KHz. Esta escala, como se ha visto anteriormente, recibe el nombre de escala Mel. El logaritmo de la energía a la salida de los filtros en escala Mel da lugar a los coeficientes MFSC.

- **Análisis basado en modelos auditivos.** Este análisis tiene en cuenta aspectos fisiológicos y psicofísicos del proceso auditivo humano que incorpora a los criterios de parametrización. Las anteriores estrategias de parametrización estaban basadas en el modelo de producción de la voz. Este grupo de técnicas se basa sin embargo en el modelo de percepción de la voz humana para parametrizar el habla, intentando reproducir el comportamiento de la membrana basilar del oído humano en la percepción. Para ello se persigue implementar mecanismos que capturen la información fisiológica que caracteriza a la percepción humana:

  - Análisis de frecuencias en canales paralelos.
  - Conservación de la estructura temporal fina del sonido.
  - Rango dinámico limitado en los canales individuales.
  - Realce de los contrastes temporales.
  - Realce de los contrastes espectrales en frecuencias adyacentes.

  Hasta los años 80, los modelos auditivos de parametrización se caracterizaban por la siguiente estructura [6]:

  i. Un banco de filtros paso banda que plasma la selectividad en frecuencias del modelo auditivo empleado. Para ello la anchura de los filtros crece de manera no lineal con la frecuencia central de los mismos. Ejemplos de escalas perceptuales para el banco de filtros son la escala *Mel,* la escala *Bark* o la escala *ERB.*
  ii. Interacciones no lineales dentro del canal y/o entre canales. Estas interacciones plasman la transducción debida a las células ciliadas del oído, y la supresión lateral entre bandas de frecuencias adyadentes.
  iii. A veces, algún mecanismo para aportar información temporal detallada como función de la frecuencia.

  Los resultados obtenidos con estas técnicas alrededor de los años 90 son bastante buenos, ligeramente mejores que los de los parámetros MFCC del dominio cepstral (que serán analizados más adelante y que son los más utilizados en la actualidad), ya que captan cierta información útil adicional:

  - La estructura temporal detallada de la señal.
  - La supresión lateral de los canales adyacentes.
  - Los contrastes temporales.
  - Otras características no lineales del proceso de audición.

Sin embargo, esta línea de investigación no siguió desarrollándose ya que, aunque los resultados eran buenos, llevaban asociado un coste computacional y de almacenamiento que no compensaba ni era factible para reconocimientos en tiempo real con coste computacional razonable. En la actualidad, existe un resurgimiento de esta línea de trabajo motivado por las capacidades de computación y almacenamiento superiores a las existentes en los años 80, por la necesidad de encontrar

parametrizaciones que mejoren MFCC y permitan enfrentarse a los actuales retos de reconocimiento. En esta línea de trabajo se introduce el estudio de una nueva forma de parametrización de la señal de voz que proponen algunos autores llamada *Human Factor Cepstral Coefficients* (HFCC), que utiliza como base la parametrización llevada a cabo con MFCC sin un coste computacional adicional.

2.3.3.6 El dominio cepstral

Las técnicas de análisis espectral que operan en el dominio de la potencia espectral logarítmica tienen la limitación de que, debido a que los espectros de los filtros en bandas adyacentes están bastante correlados, originan coeficientes espectrales también bastante correlados. Es deseable eliminar esa correlación manteniendo solo la información que sea útil para el reconocimiento. Para ello, se utiliza un filtro de decorrelación homomórfica o *Cepstrum* que, mediante la transformada inversa de Fourier del logaritmo del espectro de potencias, lleva los coeficientes cepstrales al dominio de la *cuefrencia* convirtiéndolos en coeficientes *cepstrales*. Los coeficientes cepstrales representan la señal temporal que corresponde al espectro logarítmico de potencia. El dominio de la cuefrencia es un dominio homomórfico del dominio temporal. Esto implica que las convoluciones en el dominio temporal se convierten en sumas en su dominio homomórfico de la cuefrencia. Esto será enormemente útil ya que permitirá separar las señales de voz de los ruidos convolucionales con los que estén mezcladas. Las componentes de excitación y envolvente espectral del tracto vocal aparecerán en zonas separadas del dominio transformado de la cuefrencia, que se podrán separar mediante ventanas. Haciendo un juego de paralelismo, los inventores de este operador homomórfico llamado Cepstrum (cuyo nombre crearon intercambiando la posición de las cuatro primeras letras del término *spectrum*), llamaron a ese enventanado en el dominio de la cufrencia "*liftering*" (cambiando la posición de las primeras letras del término correspondiente *filtering* del dominio spectrum). Los análisis en el dominio espectral tienen sus correspondientes homologos en el dominio de la cuefrencia que reciben los nombres de coeficentes cepstrales LPC (*Linear Predictive Coding*), MFCC (*Mel Frecuency Cepstral Coefficients*), Cepstrum PLP y ahora, en estudio en este proyecto, HFCC (*Human Factor Cepstral Coefficients*). Los coeficientes MFCC han demostrado, ser, hasta ahora, los que mejores resultados dan como técnica de parametrización teniendo en cuenta el compromiso entre coste computacional y resultados obtenidos. Tomando como referencia los resultados obtenidos tras los experimentos realizados, haremos una comparación entre los coeficientes MFCC y los HFCC con el objeto de comparar las mejoras introducidas por estos últimos en las técnicas de reconocimiento.

2.3.4 Estado del arte del reconocimiento del habla en entornos adversos

El problema del reconocimiento automático del habla en entornos adversos ha atraído la atención de muchos investigadores en los últimos años. La razón principal es que el comportamiento de los sistemas actuales de reconocimiento, que han sido diseñados suponiendo que las condiciones ambientales en que dichos sistemas van a operar no van a afectar sustancialmente la señal de voz, se degrada sustancialmente cuando las condiciones ambientales son adversas.

Las representaciones actuales de la voz, aunque poco eficientes debido a que conllevan mucha redundancia, permiten conseguir unas buenas prestaciones del reconocimiento siempre que la señal de voz se registre en condiciones favorables. Sin embargo, cuando un sistema de reconocimiento se pone a funcionar en situaciones reales se encuentra con condiciones adversas tales como, cambios en el hablante (condiciones fisiológicas, emocionales, cambio en el modo de articulación debido a un fuerte ruido ambiental, entre otras) y en el entorno acústico (ruidos, reverberación y ecos) o eléctrico (como ruidos o

distorsiones de la señal provocados por el micrófono o el canal de transmisión), que son irrelevantes desde el punto lingüístico pero que pueden degradar en gran medida la tasa de reconocimiento. Estos problemas constituyen las principales causas de degradación de los sistemas de reconocimiento automático del habla cuando se usan en la práctica. Un sistema de reconocimiento automático del habla funciona razonablemente bien en las pruebas de laboratorio, pero se producen problemas en el "mundo real".

Si la variabilidad de las características de la voz aportada por las distintas condiciones ambientales posibles fuera recogida por completo en la base de datos de entrenamiento del sistema, podríamos esperar que el resultado del reconocimiento no se degradara mucho. Pero esto sólo resulta útil cuando las condiciones ambientales en las que debe trabajar el sistema son muy específicas, y aún así, es imposible recoger todas las alteraciones que se pueden encontrar en un entorno específico, ya sea porque se desconocen, por ser demasiado numerosas o por variar con el tiempo.

Llevar el micrófono colgando o tener que mantener la posición de la cabeza frente al micrófono de sobremesa, son condiciones incómodas, pero necesarias actualmente, para que la voz captada por el micrófono sea lo bastante limpia, en especial cuando existe un ruido ambiental molesto. La situación deseable es que el micrófono se encuentre a cierta distancia del o de los hablantes y éstos puedan moverse con libertad (hands-free). En realidad, resultaría conveniente que hubiera varios micrófonos para captar señales distintas y luego procesarlas en conjunto y compararlas. De hecho, el sistema auditivo humano dispone de dos entradas de voz y dicha binauralidad le permite separar fuentes de sonido situadas en puntos distintos.

En cuanto a las condiciones adversas que pueden influir en el reconocimiento, se pueden distinguir diferentes clases de ruidos. Los más benignos para el reconocimiento del habla son los estacionarios, es decir, los que mantienen sus características estadísticas a lo largo del tiempo. Puesto que en los intervalos sin voz (silencios) aparecen aislados, se pueden determinar en ellos sus parámetros espectrales y así, teniendo los ruidos caracterizados, resulta mucho más fácil eliminarlos de los intervalos donde reside la voz. De ahí la importancia de disponer de detectores fiables de actividad oral que permitan separar los intervalos temporales de voz y de silencio. Los ruidos más difíciles de captar y eliminar suelen ser los de tipo impulsivo, tales como golpes de puerta, pitidos cortos, tos y, sobre todo, la voz de otra persona que se encuentre cerca; ésta es la situación más perjudicial para el reconocimiento, puesto que un mismo segmento de señal contiene las dos voces y resulta muy difícil separarlas para pasar a reconocer la que interesa. En cualquier caso, pero en especial cuando la base de aprendizaje no recoge las degradaciones de la voz, hay que recurrir a técnicas de reconocimiento robusto, todavía en fase de desarrollo en los laboratorios, que atacan el problema de varias formas:

1. **Obtención de una señal más limpia** (*speech enhancement*). Si se puede suponer que la señal de voz y la de ruido son aditivas y no correlativas (algo que parece realista), el espectro de la señal ruidosa es la suma de los espectros de voz y ruido, y se pueden aplicar varias técnicas de cancelación de ruido. Por ejemplo, la señal se puede procesar con un filtro de Wiener para reducir la presencia de ruido mezclado con la voz; el filtro se entrena durante los silencios para que aprenda la estadística del ruido. Una forma alternativa de eliminar (nunca totalmente) el ruido de la señal es la sustracción espectral, que estima el espectro del ruido en los silencios y luego lo sustrae del espectro de señal de voz ruidosa.

2. **Determinación de parámetros más robustos**. Es un hecho bien conocido que el sistema auditivo humano es más robusto que cualquier sistema automático, no sólo frente al ruido aditivo y las distorsiones en general, sino frente a cualquier factor de variabilidad de la voz, incluidos los cambios de articulación cuando el hablante está inmerso en un ruido intenso (en una discoteca, por ejemplo). Por tanto, sería de esperar que un sistema de reconocimiento del habla fuera más

robusto a todos estos factores si la representación de la señal de voz siguiera de cerca las características perceptivas del sistema auditivo humano. Pero los intentos realizados hasta el presente no se han visto muy favorecidos por el éxito. En esta línea de investigación irán encaminados los experimentos a realizar en este proyecto.

3. **Compensación de los parámetros distorsionados y adaptación de los modelos a las nuevas condiciones del entorno**. En estas técnicas, los parámetros son procesados a fin de que se asemejen estadísticamente a los que se hubieran obtenido con las condiciones ambientales de entrenamiento. Una alternativa es adaptar a las nuevas condiciones los modelos estadísticos de las unidades fonéticas desarrollados para habla limpia; a menudo, estas técnicas usan para ello conocimiento del ruido o la distorsión.

2.4 Reconocimiento de voz con HMMs

2.4.1 Introducción

Los modelos ocultos de Markov fueron descritos por primera vez por Baum [7]. Poco después, fueron aplicados al reconocimiento automático del habla en CMU [8] e IBM [9] [10]. En los últimos años se han convertido en la aproximación predominante en reconocimiento del habla debido a la simplicidad de su estructura algorítmica y a sus buenas prestaciones. Por ello será el sistema de reconocimiento utilizado en las pruebas experimentales realizadas en este proyecto.

2.4.2 Modelos ocultos de Markov. Definición y tipos.

Un modelo oculto de Markov o *Hidden Markov Model* (HMM) es la representación de un proceso estocástico que consta de dos mecanismos interrelacionados: una cadena de Markov de primer orden subyacente, con un número finito de estados, y un conjunto de funciones aleatorias, cada una de las cuales asociada a un estado. En un instante discreto de tiempo se supone que el proceso está en un estado determinado y que genera una observación mediante la función aleatoria asociada. Al instante siguiente, la cadena subyacente de Markov cambia de estado siguiendo su matriz de probabilidades de transición entre estados, produciendo una nueva observación mediante la función aleatoria correspondiente. El observador externo sólo "ve" la salida de las funciones aleatorias asociadas a cada estado, siendo incapaz de observar directamente la secuencia de estados de la cadena de Markov. De ahí el nombre de modelo oculto.

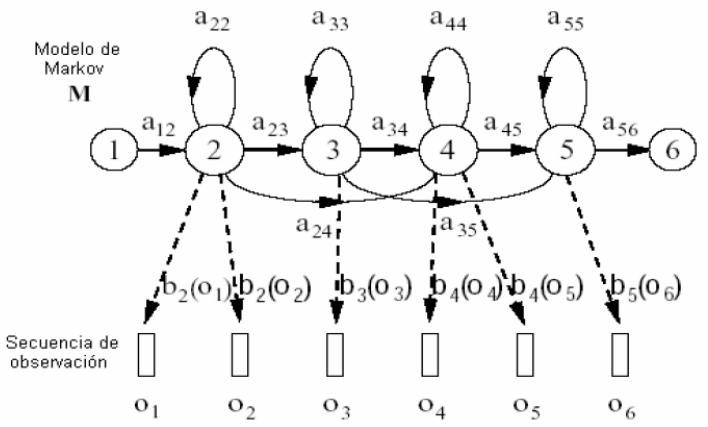

Figura 2.37 - Generación de los modelos de Markov de izquierda a derecha.

En la figura 2.37 puede observarse un ejemplo de un modelo oculto de Markov tipo Bakis o de izquierda a derecha, con dos estados no emisores y cuatro estados emisores, es decir, que emiten función de probabilidad.

Entonces, un modelo oculto de Markov es la composición de dos procesos estocásticos *(X,Y)* definidos como:

i. Una cadena oculta de Markov *X* que tiene en cuenta la variabilidad temporal, y que no es directamente observable.
ii. Un proceso observable *Y* que tiene en cuenta la variabilidad espectral y va tomando valores en el espacio de las características acústicas u observaciones.

La combinación de ambos procesos modela las fuentes de variabilidad de la señal de voz y permite reflejar una secuencia de parámetros acústicos como concatenación de los procesos elementales del modelo con la flexibilidad suficiente para hacer sistemas de reconocimiento. Los modelos ocultos de Markov usados en el reconocimiento de voz tienen dos asunciones formales características:

a) La historia de la cadena no influye en la evolución futura de la misma si existe información actual (*hipótesis de Markov de primer orden*).
b) Ni la evolución de la cadena ni las observaciones pasadas determinan la observación actual si se ha especificado la última transición de la cadena (*hipótesis de independencia de las salidas*).

Una vez hechas esas asunciones, si llamamos *y* ϵ *Y* a una variable que representa las observaciones y llamamos *i, j* ϵ *X* a las variables que representan los estados del modelo, el modelo $\lambda = (A, B, \Pi)$ queda representado por las siguientes matrices de parámetros según puede verse en la figura 2.37:

$$A \equiv a_{i,j}|i,j\epsilon X, \quad \text{probabilidades de transición}$$
$$B \equiv b_{i,j}|i,j\epsilon X, \quad \text{distribuciones de las salidas}$$
$$\Pi \equiv \pi_i|i\epsilon X, \quad \text{probabilidades iniciales}$$

Donde los términos de las matrices se definen como:

$$a_{i,j} \equiv p(X_t = j|X_{t-1} = i)$$
$$b_{i,j}(y) \equiv p(Y_t = y|X_{t-1} = i|X_t = j)$$
$$\pi_i \equiv p(X_o = i)$$

Según la naturaleza de la matriz *B* de las distribuciones de probabilidad de las salidas, los HMMs se pueden clasificar en varios tipos, [104],[25]:

i) Modelos discretos, DHMMs

Las observaciones son vectores de símbolos de un alfabeto finito de N elementos diferentes. Para cada componente de dicho vector de símbolos se define una densidad discreta:

$$\{w(k)|k = 1, ..., N\}$$

y la probabilidad del vector se calcula multiplicando las probabilidades de cada componente siendo éstos independientes entre sí.

$$b_i(y) = p(y|x_i, \lambda) = \prod_k w_{y,x_i,k}$$

ii) Modelos continuos, CHMMs

Otra posibilidad es definir las distribuciones de probabilidad en espacios de observaciones continuos, lo cual puede ser conveniente ya que la señal de voz es continua. Las distribuciones de probabilidad necesitan ciertas restricciones en este caso para que el número de parámetros del sistema sea manejable y las re-estimaciones sean consistentes: las transiciones se definen con mezclas de distribuciones paramétricas básicas caracterizadas por pocos parámetros, que suelen ser Gausianas o Laplacianas. Cada estado x_i del modelo tendrá un conjunto específico $V(x_i,\lambda)$ de funciones densidad de probabilidad. Si llamamos v_k a cada una de esas *pdfs*, la expresión de las probabilidades de las salidas será:

$$b_i(y) = p(y|x_i, \lambda) = \sum_{v_k \epsilon V(x_i,\lambda)} p(y|v_k, x_i, \lambda) \cdot P(v_k|x_i, \lambda)$$

Donde el término $P(v_k|x_i, \lambda)$ representa la probabilidad de aparición de la *pdf* v_k.

iii) Modelos semicontinuos, SCHMMs

Para modelar distribuciones complejas con la mezcla funciones paramétricas a veces es necesario un gran número de estas en cada mezcla, y un corpus de entrenamiento muy grande. Una solución efectiva es compartir las distribuciones entre diferentes transiciones del modelo. Esto es lo que hacen los modelos semicontinuos, en los que todos los estados comparten las mismas distribuciones de probabilidad con diferentes pesos:

$$V(x_i, \lambda) = V, \quad \forall x_i, \lambda$$

$$b_i(y) = p(y|x_i, \lambda) = \sum_{v_k \epsilon V} p(y|v_k) \cdot P(v_k|x_i, \lambda)$$

La técnica de HMM se usa en la actualidad en aquellos sistemas en los que el modelado tiene una dependencia del tiempo como pueden ser los sistemas reconocimiento fonético y del habla en general. Es una práctica universal usar los modelos ocultos de Markov para calcular las probabilidades acústicas debido a su capacidad de modelar estadísticamente de manera adecuada la generación de voz. Los HMM en reconocimiento de voz se utilizan teniendo en cuenta dos hipótesis:

1. La voz se puede dividir en segmentos, estados, en los que la señal de voz se puede considerar estacionaria. Es decir, en la ventana de análisis la señal mantiene la estructura de principio a fin. Se asume que las transiciones entre segmentos contiguos son instantáneas.

2. La probabilidad de observación de que un vector de características se genere depende sólo del estado actual y no de símbolos anteriores. Esta es una suposición de Markov de primer orden, denominada *hipótesis de independencia*.

Otra razón por la que los HMMs son populares, es porque pueden ser entrenados automáticamente, siendo factible realizar los cálculos en un tiempo razonable. El reconocimiento fonético es la disposición más simple posible. El modelo oculto de Markov tendrá en cada estado una distribución estadística llamada mezcla de Gaussianas de matriz de covarianza diagonal, que dé una probabilidad para cada vector observado. Cada fonema tendrá una distribución de salida. Un modelo oculto de Markov para una secuencia de fonemas se construye concatenando los modelos ocultos entrenados para los fonemas separados. El uso de los HMM permite eludir las limitaciones de algunos otros sistemas en el reconocimiento de fonemas como son los siguientes:

- DTW (*Alineamiento temporal Dinámico*) no hay posibilidad de realizar un entrenamiento estadístico, ya que se realiza comparaciones entre secuencias de vectores de parámetros.

- VQ (*Cuantificación temporal*) asignación dura entre los vectores y la clase que modela. Además tiene que respetar el compromiso entre el tamaño del codebook y el error de cuantificación.

2.4.3 Problemas a resolver para la utilización de un HMM

Los modelos ocultos de Markov se caracterizan por tres problemas que hay que resolver para que resulten modelos útiles en aplicaciones reales:

- **Evaluación.** Es el problema básico: dada una observación acústica y un modelo oculto de Markov, determinar la probabilidad de que el modelo genere esa observación, es decir, la probabilidad acústica $P(O|\lambda)$. Esta probabilidad se determina con el algoritmo *forward-backward* [11].

- **Decodificación.** Determinación de la secuencia óptima de estados $X = x_1, x_2, \ldots, x_T$ dada la observación acústica y el modelo oculto de Markov. Es decir, se busca la alineación de la observación con el modelo, asignando cada vector a un estado del modelo. Se lleva a cabo mediante el algoritmo de *Viterbi* [12].

- **Estimación o entrenamiento de los HMMs.** Consiste en el cálculo de los parámetros que caracterizan el modelo. Dados un conjunto de datos y una colección de secuencias observables, se determina el HMM que con mayor probabilidad ha generado la secuencia. Este problema se resuelve comúnmente con el algoritmo *Baum-Welch* [13].

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 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.

2.4.3.1 El problema de evaluación – Algoritmo Forward - Backward

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

$$P(O \mid \lambda) = \sum P(O \mid q, \lambda) P(q \mid \lambda)$$

Para ello consideremos una determinada secuencia de estados: Q=(q1,q2,...,qT) donde q1 es el estado inicial. La probabilidad de la secuencia de observación O dada la secuencia de estados Q es:

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

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

$$P(O \mid Q, \lambda) = b_{q1}(O_1) \cdot b_{q2}(O_2) \cdots b_{qT}(O_T)$$

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

$$P(Q \mid \lambda) = \pi_{q1} \cdot a_{q1q2} \cdot a_{q2q3} \cdots a_{qT-1qT}$$

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:

$$P(O \mid \lambda) = \sum_{Q} P(O \mid Q, \lambda) \cdot P(Q \mid \lambda) = \sum_{q1, q2, \ldots, qT} \pi_{q1} \cdot b_{q1}(O_1) \cdot a_{q1q2} \cdot b_{q2}(O_2) \cdots a_{qT-1qT} \cdot b_{qT}(O_T)$$

La interpretación del resultado obtenido es la siguiente: inicialmente en el tiempo t=1 nos encontramos en el estado q1 con probabilidad πq1 y generamos el símbolo O1 con probabilidad $b_{q1}(O_1)$. Al avanzar el reloj al instante t=2 se produce una transición al estado q2 con probabilidad aq1q2 y generamos el símbolo O2 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 $a_{qT\text{-}1qT}$ 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**
   En este paso se inicializan las probabilidades hacia delante como la probabilidad conjunta del estado i y de la observación o_1.

$$\alpha_1(i) = \pi_i b_i(o_i) \quad 1 \le i \le N$$

2. **Recursión inductiva**
   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 $á_t(i)$ es la probabilidad conjunta de observar el evento o_1, o_2,...., 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,...., 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.

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

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

$$\alpha_t(i) = P(o_1 o_2 ... o_t, q_t = i \mid \lambda)$$

   Por tanto,

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

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:

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

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**
   Todos los estados son equiprobables.

$$\beta_t(i) = \frac{1}{N} \quad 1 \le i \le N$$

2. **Inducción**
   La relación entre α y β adyacentes se puede observar mejor en la siguiente figura. α se calcula recursivamente de izquierda a derecha mientras β se calcula recursivamente de derecha a izquierda.

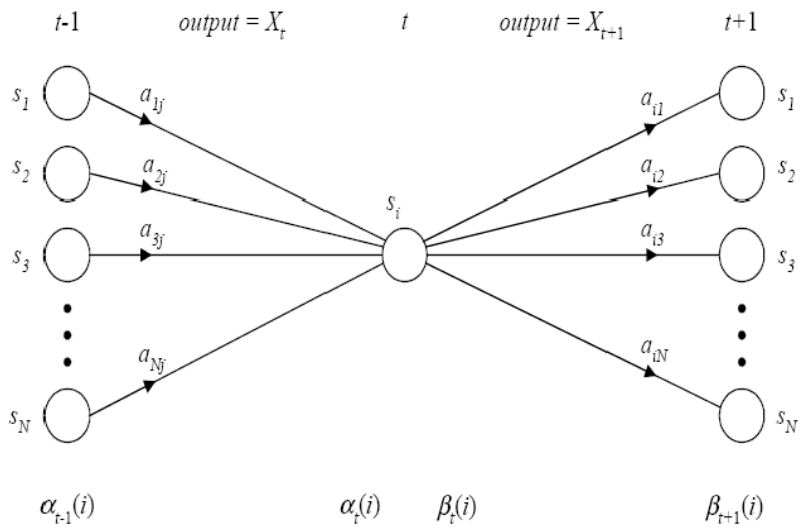

Figura 2.38 - Representación tanto el algoritmo foward como backward [*Huang et al 2001*].

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

2.4.3.2 El problema de decodificación – 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 solo 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:

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

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:

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

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 $\Phi_t(j)$. El proceso completo para encontrar la mejor secuencia será:

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

$$\delta_1(i) = \pi_i b_i(o_i) \quad 1 \le \mathrm{i} \le \mathrm{N}$$
$$\phi_1(i) = 0$$

2. **Inducción**

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

   Se guarda aquel camino que tiene mayor probabilidad,

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

3. **Finalización**

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

$$q_T^* = \arg\max_{1 \le i \le N} [\delta_T(i)]$$

4. **Seguimiento hacia atrás del camino óptimo (backtracking)**

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

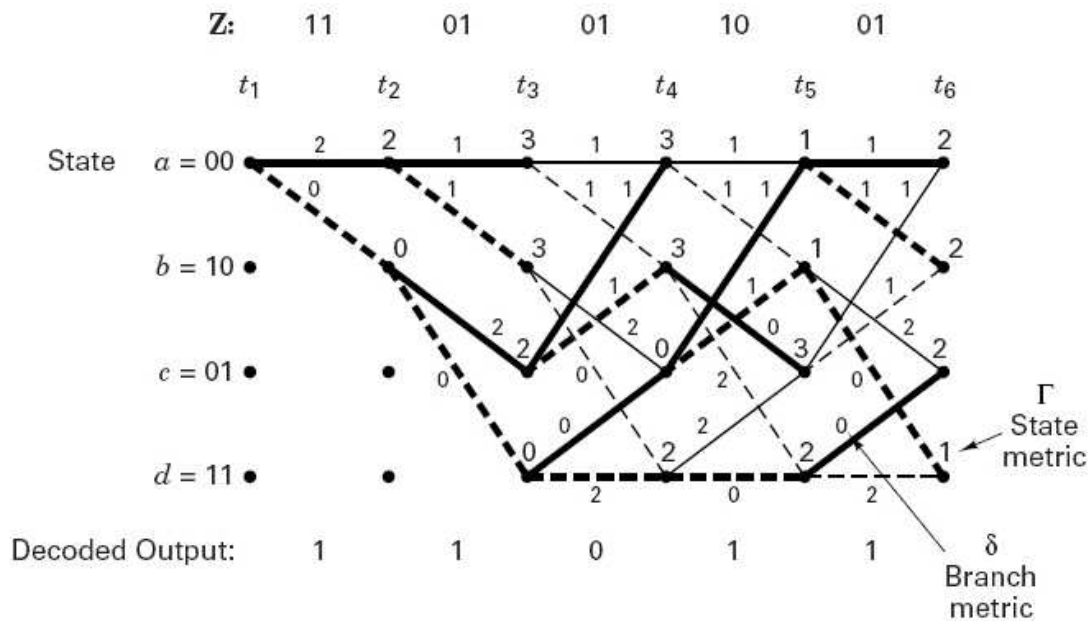

Figura 2.39 - Esquema del algoritmo de Viterbi.

Como se puede observar, 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 $O(TN^2)$.

2.4.3.3 El problema de aprendizaje – Algoritmo de Baum-Welch

Aquí el problema que tenemos es que queremos estimar los parámetros del modelo λ (A, B, Π) de forma que maximicemos *P(O|λ)*. 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, este 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:

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

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

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

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

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:

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

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:

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

$$a_{ij}' = \frac{\text{número esperado de transiciones del estado i al j}}{\text{número esperado de transiciones desde el estado i}} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T} \gamma_t(i)}$$

$$b_j'(k) = \frac{\text{número esperado de instantes en el estado j observando el simbolo } v_k}{\text{número esperado de instantes en el estado i}} = \frac{\sum_{t=1}^{T} \gamma_t(i)\,|_{o_t=v_k}}{\sum_{t01}^{T} \gamma_t(i)}$$

Con estos cálculos obtenemos una re-estimación de los parámetros del modelo obteniendo un nuevo modelo λ'=(A',B', π'). Si el modelo λ definía un punto crítico de la función de máxima verosimilitud en dicho caso tendremos λ'=λ , o bien el nuevo modelo que hace que se cumpla P(O|λ')>P(O|λ), 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.

2.5 Algoritmos de extracción de características

2.5.1 Introducción

En el reconocimiento del habla, la señal de voz, una vez digitalizada, se procesa para producir una nueva representación de la voz en forma de secuencia de vectores o agrupaciones de unos valores que

denominamos parámetros y que, como se ha explicado anteriormente, deben representar la información contenida en la envolvente del espectro. El número de parámetros debe ser reducido, puesto que la base de datos de entrenamiento siempre es limitada, por lo que cuantos más parámetros tenga la representación, menos fiables son los valores entrenados y, por otro lado, más costoso es el proceso de reconocimiento.

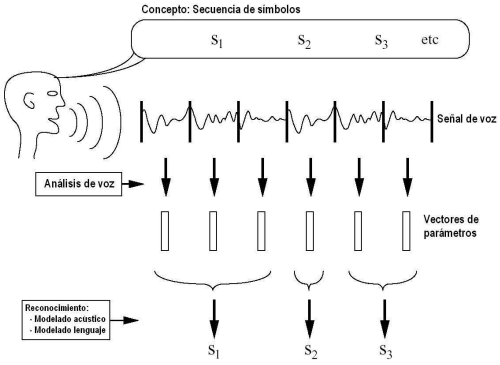

Figura 2.40 - Extracción de características de la señal de voz.

A continuación, vamos a analizar dos algoritmos de extracción de características a tener en cuenta en la realización de este proyecto.

2.5.2 Mel Frecuency Cepstral Coefficients

A imitación de lo que sucede en el sistema auditivo humano, la identificación de los sonidos se hace en el dominio de la frecuencia. 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 [14].

Davis y Mermelstein (D&M) introdujeron el término "M*el Frecuency Cepstral Coefficient* " en 1980 [15] cuando combinaron filtros triangulares, perceptualmente distribuidos, con la transformada discreta del coseno del logaritmo de las energías de salida de los filtros. El trabajo previo de *Pols* [16] 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 2.41) junto con la siguiente ecuación:

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

Donde *M* representa el número de coeficientes cepstrales y X_k el logaritmo de la energía de salida del filtro *k-ésimo*. Como se observa en la figura 2.41, 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.

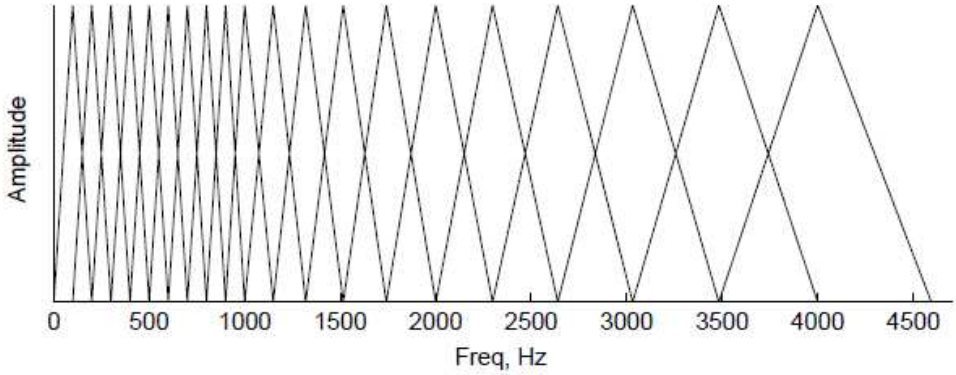

Figura 2.41 - Banco de filtros utilizado por Davis and Mermelstein en el algoritmo de extracción de características MFCC.

El ancho de banda de los filtros triangulares en MFCC viene determinado por la distribución de la frecuencia de centro de cada filtro, siendo ésta 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.

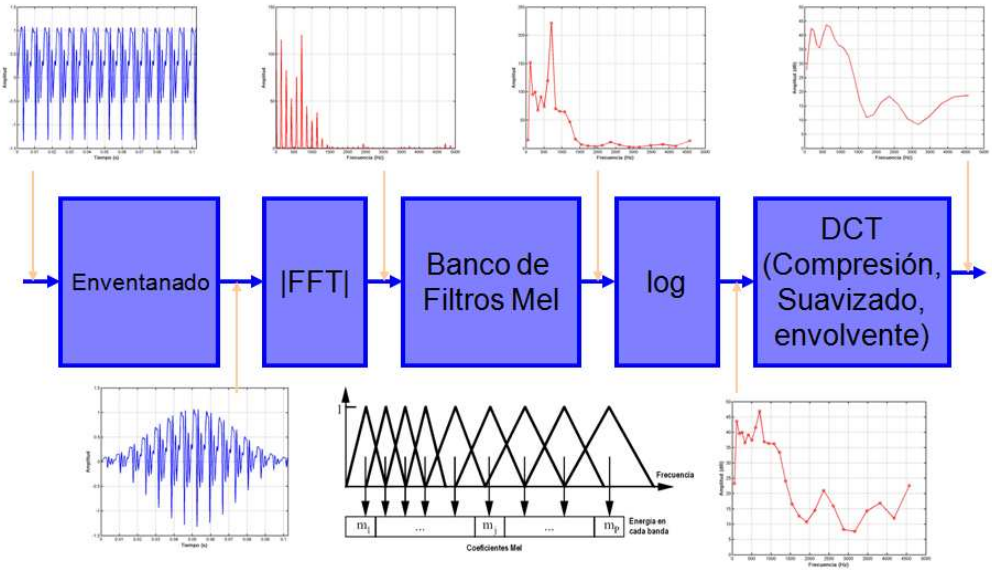

Figura 2.42 - Proceso de extracción de los coeficientes MFCC.

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, C0, indica la energía de la señal y se usa o no dependiendo de la aplicación. Y el segundo coeficiente, C1, 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 2.43). Los coeficientes ΔΔMFCC representan la variación de los coeficientes de velocidad, por lo que son llamados coeficientes de aceleración.

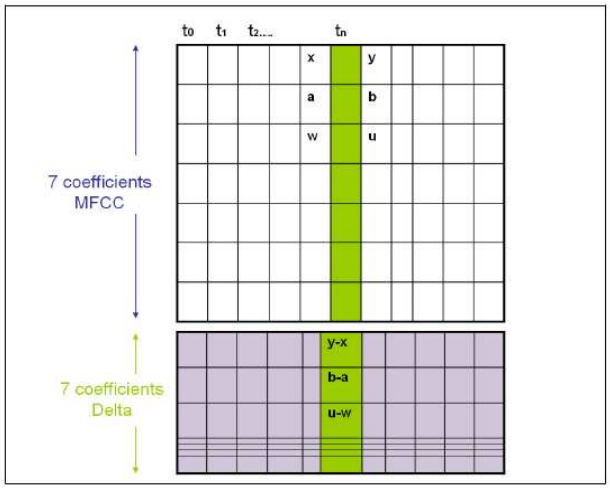

Figura 2.43 - Una esquematización de los Delta- Mel-Frequency Cepstral Coefficients donde se representa una posible manera de calcular los coeficientes delta.

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.

2.5.3 Human Factors Cepstral Coefficients

2.5.3.1 Introducción

El reconocimiento automático del habla (RAH) no ha desarrollado un uso ampliamente extendido, principalmente debido al bajo rendimiento obtenido en ambientes ruidosos. El primer paso en RAH es

obtener una representación robusta de la señal de voz, ya que, cuando un sistema de reconocimiento se pone a funcionar en situaciones reales se encuentra con condiciones adversas tales como cambios en el hablante (condiciones fisiológicas, emocionales, cambio en el modo de articulación debido a un fuerte ruido ambiental, entre otras) y en el entorno acústico (ruidos, reverberación y ecos) o eléctrico (como ruidos o distorsiones de la señal provocados por el micrófono o el canal de transmisión), que son irrelevantes desde el punto de vista lingüístico pero que pueden degradar en gran medida la tasa de reconocimiento. En el curso de esta investigación, algunos autores proponen una nueva técnica de extracción de características llamada *Human Factor Cepstral Coefficients* (HFCC), con el fin de mejorar la robustez frente al ruido. Surge así una nueva técnica encaminada a obtener un sistema más robusto que utiliza como base de estudio el proceso de parametrización llevada a cabo en MFCC, modificando el análisis del banco de filtros, añadiendo, para su diseño, información sobre el sistema auditivo humano.

2.5.3.2 Trabajo previo

Muchos son los autores que han investigado sobre los algoritmos de extracción de características utilizados en reconocimiento automático de voz que utilizan diferentes diseños del banco de filtros. Los estudios de Hermansky en predicción lineal (PLP) hacen uso de un banco de filtros en escala Bark y compresión mediante raíz cuadrada, antes de estimar los coeficientes de predicción lineal [17]. El análisis PLP utiliza aproximaciones de ingeniería a las leyes psicofísicas, tales como la igualdad de sonoridad pre-énfasis y la ley de potencia de intensidad a sonoridad, y es cierto que permite variaciones que se derivan del espectro auditivo. Chan et al. utilizan un análisis multi-resolución para el diseño de un banco de filtros distribuidos mediante la escala Mel [18]. El diseño del banco de filtros les permitió incluir el filtrado de Wiener en el proceso de extracción de características y así, se consiguió obtener una mejora en reconocimiento del habla robusto al ruido.

Otros investigadores han modificado el algoritmo MFCC con el fin de conseguir un rendimiento más robusto frente al ruido. Tchorz y Kollmeier desarrollaron un preprocesador basado en el modelo auditivo, similar a MFCC [19]. La compresión estática logarítmica utilizada en MFCC fue reemplazada por una compresión adaptativa. En un experimento de dígitos aislados, los autores demostraron una mejora de la robustez en diferentes ambientes ruidosos comparando su modelo con el de un compresor estático logarítmico, aunque la precisión del reconocimiento de voz limpia usando el de compresión adaptativa fue menor. Strope y Alwan modificaron directamente el algoritmo de MFCC añadiendo un enmascarador adaptativo antes de la aplicación de la transformada discreta del coseno (DCT) [20].

La investigación se ha centrado también en el ancho de banda de los filtros utilizados en el algoritmo MFCC. La evaluación 2000 SPINE de Singh et al., reportó el uso de filtros en MFCC con el doble de ancho de banda manteniendo las frecuencias de centro y el número de filtros [21]. Se lograron mejoras en robustez al ruido en experimentos de gran vocabulario, sin embargo, el ancho de banda del filtro no se procedió a investigar. Sinha y Umesh consiguieron mejoras en reconocimiento MFCC al aumentar el número de filtros, manteniendo el ancho de banda de filtro original, aunque no reportaron ninguna explicación para el aumento del rendimiento[22]. Además, se informó sobre mejoras en robustez al ruido en MFCC al cambiar el solapamiento entre los filtros adyacentes[23].

2.5.3.3 Algoritmo HFCC

El término *Human Factor Cepstral Coefficients* (HFCC), introducido por Skowronski y Harris (2004), representa la actualización más reciente del banco de filtros utilizado en MFCC. HFCC no pretende ser un modelo de percepción del sistema auditivo humano, sino más bien un método de extracción de características de inspiración biológica.

El ancho de banda en HFCC se determinó por resultados en experimentos psicoacústicos previos. Patterson introdujo el método "notch-noise" a mediados de los 70 con el fin de estimar la forma de los filtros del sistema auditivo humano a través de test de escucha. Diversos estudios perceptuales presentan una estructura formada por bancos de filtros para mostrar el comportamiento del sistema auditivo humano. La relación entre las frecuencias de centro de esos filtros y el ancho de banda de las bandas críticas del oído es todavía ignorada en MFCC. Sin embargo HFCC soluciona este hándicap, Moore y Glasberg [24] resumieron los experimentos de varios laboratorios usando una nueva técnica, calcularon el ancho de banda rectangular equivalente, ERB (*Equivalent Rectangular Bandwidth*), para los modelos de filtros auditivos (sugiriendo que el ERB está relacionado con el ancho de banda crítico) y definieron el ERB vs. la frecuencia de centro obteniendo como resultado la siguiente ecuación:

$$ERB = 6.23 \cdot 10^{-6} \cdot f_c^2 + 93.39 \cdot 10^{-3} \cdot f_c + 28.52$$

Donde f_c es la frecuencia de centro de cada filtro que compone el banco de filtros, en Hz. La ecuación anterior proporciona un buen ajuste a la curva experimental para el rango de frecuencias definido entre [100, 6500] Hz. De igual forma, el ancho de banda dado por esta ecuación puede ser escalado por una constante, a la que Skowronski y Harris llamaron E-factor y de la que se hablará en detalle más adelante.

Como se ha visto en la sección anterior, 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. De ahí, la diferencia más significativa en el algoritmo HFCC de Skowronski y Harris cuando lo comparamos con MFCC, es que el ancho de banda de cada filtro no está relacionado con la separación existente entre filtros. Más específicamente, el ancho de banda de un filtro utilizado en el banco de filtros de HFCC viene dado por el ancho de banda rectangular equivalente (ERB) introducido por Moore y Glasberg (1983).

Suponiendo una frecuencia de muestreo de 12500 Hz, Skowronski y Harris propusieron un banco de filtros formado por 29 filtros de igual altura, cubriendo el rango de frecuencias entre [0, 6250 Hz]. En este estudio, como se verá más adelante, se supone una frecuencia de muestreo de 16000 Hz y sólo los primeros 24 filtros que abarcan el rango de frecuencias entre [0, 4000 Hz]. Como se ilustra en la figura 2.44, en HFCC el solapamiento entre los filtros es diferente al convencional, un filtro puede solapar no sólo con sus vecinos más cercanos, sino también con filtros lejanos.

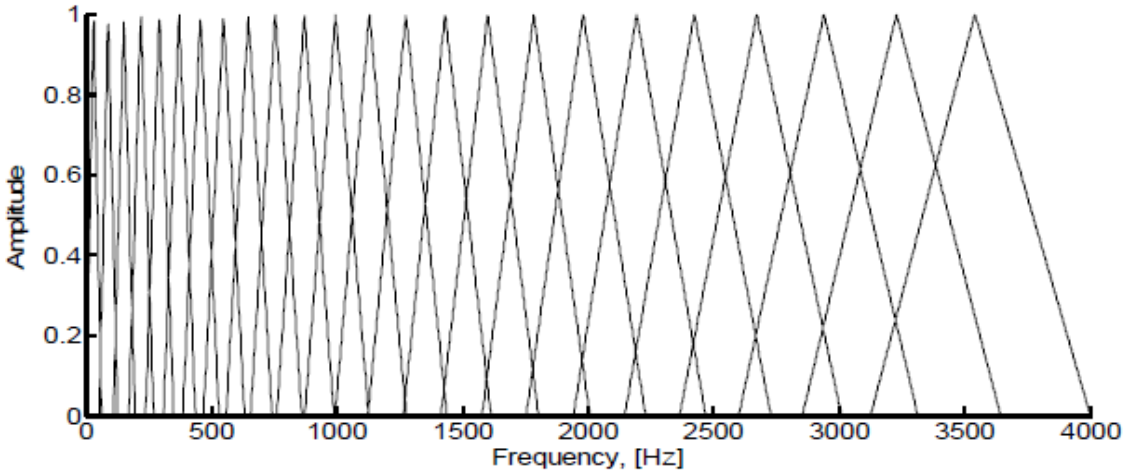

Figura 2.44 - Banco de filtros HFCC propuesto por Skowronski y Harris.

En resumen, el diseño del banco de filtros utilizado en el algoritmo HFCC, como fue descrito por Skowronski (2004), consiste en los siguientes pasos: en primer lugar se escogen las frecuencias límite, f_{low} y f_{high}, entre las que se comprende el banco de filtros, así como el número de filtros a utilizar. Las frecuencias de centro f_{c1} y f_{cM} del primer y último filtro, respectivamente, se calculan como:

$$f_{c_i} = \frac{1}{2} \cdot \left(-\overline{b} + \sqrt{\overline{b}^2 - 4 \cdot \overline{c}} \right)$$

Donde el índice *i* es 1 o M, y los coeficientes *b* y *c* se definen como:

$$\overline{b} = \frac{b - \hat{b}}{a - \hat{a}} \qquad \overline{c} = \frac{c - \hat{c}}{a - \hat{a}}$$

Recibiendo valores diferentes para cada uno de los dos casos indicados, 1 o M.

De igual forma, los valores de las constantes *a*, *b* y *c* vienen de la ecuación del ERB y son $6.23 \cdot 10^{-6}$, $93.39 \cdot 10^{-3}$, 28.52, respectivamente. Para el primer filtro, los valores de los coeficientes $\hat{a}$, $\hat{b}$, $\hat{c}$, son calculados como:

$$\hat{a} = \frac{1}{2} \cdot \frac{1}{700 + f_{low}} \qquad \hat{b} = \frac{700}{700 + f_{low}} \qquad \hat{c} = -\frac{f_{low}}{2} \cdot \left(1 + \frac{700}{700 + f_{low}} \right)$$

Para el último filtro estos coeficientes son:

$$\hat{a} = -\frac{1}{2} \cdot \frac{1}{700 + f_{high}} \qquad \hat{b} = -\frac{700}{700 + f_{high}} \qquad \hat{c} = \frac{f_{high}}{2} \cdot \left(1 + \frac{700}{700 + f_{high}} \right)$$

Una vez se calculan las frecuencias de centro del primer y último filtro, las frecuencias del resto de filtros que componen el banco de filtros se calculan fácilmente teniendo en cuenta que éstos son equidistantes entre sí en escala Mel. El paso $\Delta\hat{f}$ entre las frecuencias de centro de filtros adyacentes se calcula como:

$$\Delta\hat{f} = \frac{\hat{f}_{c_M} - \hat{f}_{c_1}}{M - 1}$$

Donde todas las frecuencias se encuentran es escala Mel.

Las transformaciones entre frecuencias lineales y frecuencias en escala Mel, $f_{ci} \rightarrow \hat{f}_{c1}$ y $f_{cM} \rightarrow \hat{f}_{cM}$, vienen dadas por la ecuación:

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

Así, las frecuencias de centro en escala Mel, $\hat{f}_{ci}$, se calculan de la siguiente forma:

$$\hat{f}_{c_i} = \hat{f}_{c_1} + (i-1)\cdot\Delta\hat{f}, \quad \text{for } i = 2,\ldots,M-1$$

El siguiente paso será realizar la transformación inversa, de frecuencias en escala a Mel a frecuencias lineales, $\hat{f}_{ci} \rightarrow f_{ci}$, haciendo uso de la siguiente expresión:

$$f_{c_i} = 700\cdot\left(10^{\hat{f}_{c_i}/2595} - 1\right)$$

Una vez obtenidas las diferentes frecuencias de centro de cada filtro, f_{ci}, calcularemos los ERB_i asociados. Finalmente, serán calculadas las frecuencias f_{lowi} y f_{highi}, frecuencia mínima y máxima respectivamente de cada filtro i-ésimo del banco de filtros. Para ello, haremos uso de las siguientes expresiones:

$$ERB_i = \frac{1}{2}\cdot\left(f_{high_i} - f_{low_i}\right)$$

y

$$\hat{f}_{c_i} = \frac{1}{2}\cdot\left(\hat{f}_{high_i} - \hat{f}_{low_i}\right),$$

La cual se transforma como sigue:

$$f_{low_i} = -(700 + ERB_i) + \sqrt{(700 + ERB_i)^2 + f_{c_i}(f_{c_i} + 1400)},$$

y

$$f_{high_i} = f_{low_i} + 2\cdot ERB_i$$

Con todos los parámetros calculados, el diseño del banco de filtros HFCC está completo. El siguiente paso será, de forma similar a la extracción de los parámetros MFCC de Davis y Mermelstein, el cálculo del logaritmo de la salida del banco de filtros y la aplicación de la DCT para la obtención de los parámetros HFCC.

Por tanto, HFCC se presenta como una modificación de MFCC cuya principal característica es que el ancho de banda de los filtros utilizados en el banco de filtros es un parámetro de diseño libre, independiente de la separación entre filtros y vinculado con la conocida relación entre la frecuencia de centro y el ancho de banda crítico del sistema auditivo humano, lo cual nos permite una nueva línea de investigación.

La habilidad de controlar el ancho de banda de los filtros en el diseño del banco de filtros es importante por dos razones:

1. Elimina errores en el ancho de banda producidos por elecciones equivocadas en el número de filtros o en el rango de frecuencia en el banco de filtros.
2. Permite la optimización del ancho de banda.

En esta línea, se introduce una variación de HFCC llamada HFCC-E, en el que se considera un factor de escala lineal llamado E-factor con el que podemos controlar el ancho de banda de los filtros con el fin de investigar

los efectos que produce una variación del ancho de banda bajo condiciones de ruido. Así, el ancho de banda de cada filtro podrá ser escalado por una constante mediante la multiplicación de ésta por la expresión de ERB.

Aunque la expresión del ERB de Moore y Glasberg describe las bandas críticas del sistema auditivo humano, HFCC es fundamentalmente un preprocesador para un clasificador artificial. Las diferentes variaciones en la expresión del ancho de banda utilizada en HFCC son acciones candidatas para promover el estudio de esta técnica.

Los resultados obtenidos en RAH haciendo uso del algoritmo de extracción de características HFCC-E serán ampliados en la sección 4.

3

Marco experimental

3.1 Sistema utilizado

La herramienta escogida para la implementación del presente proyecto ha sido MATLAB. MATLAB es un lenguaje de programación de alto nivel basado en matrices y vectores que gracias a su gran potencia de cálculo y a su agradable entorno, que incorpora la posibilidad de una visualización gráfica de los resultados, hacen que sea la herramienta ideal para el procesamiento de señales. Además, puede incorporar una gran variedad de programas denominados toolboxes que extienden la cantidad de funciones contenidas en el programa principal. Uno de los toolboxes que ha resultado muy útil para la implementación de este proyecto ha sido el de "Signal Processing Toolbox".

Con el fin de analizar los resultados obtenidos mediante MATLAB en el ámbito del reconocimiento, haremos uso de la herramienta HTK (*Hidden Markov Model Toolkit*).

HTK es un conjunto de herramientas de software para diseñar y manipular HMM. Originalmente fue creado para aplicarlo al desarrollo de sistemas ASR, ahora puede utilizarse en cualquier área del conocimiento, la única restricción es que el problema a resolver pueda ser enfocado como un proceso de modelación Estocástico Markoviano. En la actualidad es utilizado de forma exitosa en: reconocimiento y síntesis de voz, reconocimiento de caracteres y formas gráficas, análisis de vibraciones mecánicas e incluso ha sido usado con éxito en la determinación de secuencias válidas del ADN humano (*Proyecto Genoma*). Según sea el grado de complejidad de nuestro problema (nivel al que se diseñen los HMM), HTK resulta adaptable al tipo y formato de dato a utilizar y permite el diseño de diferentes tipos de reconocedores.

El desarrollo de HTK lo lleva a cabo el grupo del habla, visión y robótica del Departamento de Ingeniería de la Universidad de Cambridge (CUED), UK. Actualmente HTK es de libre distribución y su código y librería pueden ser modificados en común acuerdo con el CUED. Además la herramienta se encuentra disponible para utilizarlo en diversas plataformas o sistemas operativos, tales como: Unix, Linux, Windows XP y DOS.

En este trabajo, todo el proceso de diseño y manipulación de HMMs se ha realizado haciendo uso de la herramienta HTK, cuyas componentes son archivos ejecutables que utilizaremos bajo el sistema operativo Linux . Para el entrenamiento se ha utilizado el programa HERest, y para la clasificación el programa HVite. El programa HResults ha proporcionado los resultados que nos han permitido interpretar cómo han funcionado nuestros experimentos.

Todos los experimentos se han automatizado en la medida de lo posible mediante la creación de una serie de listas de ficheros, ficheros de configuración y scripts configurables. Igualmente, se ha tenido especial cuidado en el diseño del parametrizador dividiendo cada una de las fases explicadas en este proyecto en módulos diferenciados dependiendo de la funcionalidad, así como la realización de una programación cuidada, comentada y estructurada con el fin de poder ser utilizada y entendible por todas aquellas personas que puedan hacer uso de este trabajo en el futuro.

3.2 Base de datos

En esta sección se describe la base de datos utilizada para la realización de este proyecto. Ésta se caracteriza por ser una base de datos utilizada para la evaluación de reconocimiento de voz en el interior de un coche.

3.2.1 Especificaciones de la base de datos

3.2.1.1 Vocabulario

La tarea de reconocimiento de voz de la base de datos CENSREC-2 (*Corpus and Environments for Noisy Speech RECognition*), constituye un reconocimiento continuo de números en un entorno real de conducción en el interior de un coche. El vocabulario de CENSREC-2 se compone de 11 modelos de números en japonés ("Ichi", "ni", "san", "Yon", "ir", "roku", "nana", "Hachi", "kyu", "zero" y "maru"), un silencio ("sil") y una breve pausa ("sp"). La secuencia de números de cada expresión y la pronunciación de los números japoneses son los mismos que lo utilizados en la base de datos AURORA-2J [25]. Los diferentes locutores fueron avisados para pronunciar cada uno de los dígitos de la misma forma en la que aparece en la tabla 3.1.

Aunque no hay locuciones de seis dígitos (la fuente de datos para AURORA-2 es TIdigits, la cual no incluye locuciones de seis dígitos. Desafortunadamente, el documento de TIdigits no menciona la razón de esto.) la frecuencia de ocurrencia de locuciones de dos, tres, cuatro, cinco y siete dígitos son prácticamente iguales.

Digit	AURORA-2	AURORA-2J
1	one	/ichi/
2	two	/ni/
3	three	/saN/
4	four	/yoN/
5	five	/go/
6	six	/roku/
7	seven	/nana/
8	eight	/hachi/
9	nine	/kyuH/
0(Z)	zero	/zero/
0(O)	oh	/maru/

Tabla 3.1 - Pronunciación en japonés de los once dígitos que forman la base de datos CENSREC-2, que es igual a la empleada en AURORA-2J.

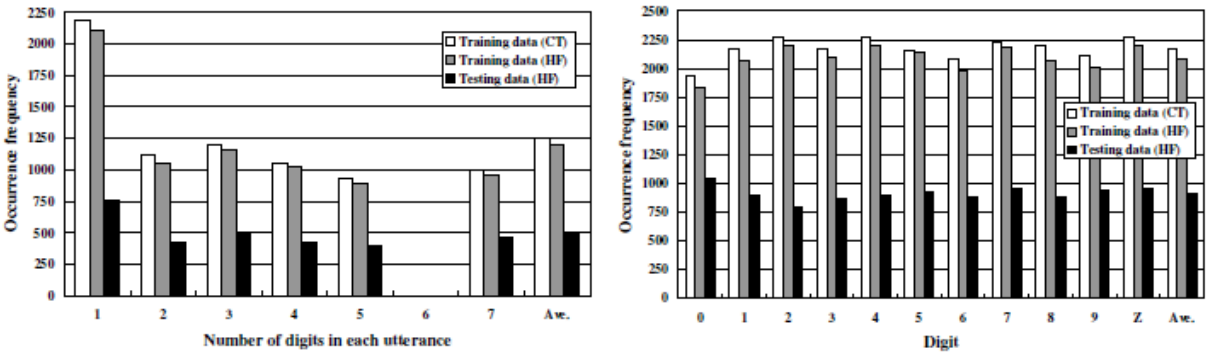

Figura 3.1 - A la derecha, representación del número de dígitos pronunciados en cada locución. A la izquierda, frecuencia de ocurrencia de cada dígito.

3.2.1.2 Adquisición de los datos de audio

Las grabaciones de audio fueron tomadas en el interior de un vehículo especialmente equipado. Se instalaron seis micrófonos tal y como se muestra en la figura 3.2. El micrófono número 1 es un micrófono de habla cercana, los micrófonos 3 y 4 fueron instalados en el salpicadero y los micrófonos 5, 6 y 7 fueron fijados al techo del vehículo. Los datos de voz grabados con el micrófono de habla cercana (CT) (nº 1: SENNHEISER HMD410 con SONY ECM77B) y con el micrófono manos libres (HF) instalado en el techo encima del asiento del conductor (nº 6: SONY ECM77B) son utilizados para la obtención de la base de datos CENSREC-2 [26].

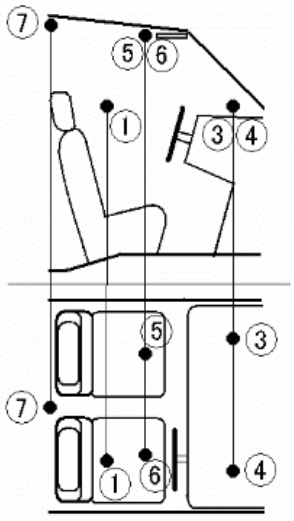

Figura 3.2 - Posiciones de los micrófonos para la obtención de los datos: en la parte de arriba la vista de lado del escenario en el interior del coche y en la parte de abajo la vista desde arriba.

Las condiciones bajo las que fueron tomadas las muestras de audio se muestran en la tabla 3.1. La base de datos fue obtenida bajo 11 condiciones ambientales haciendo uso de la combinación de tres velocidades de coche diferentes (ralentí, velocidad baja conduciendo por la calle de una ciudad y velocidad alta conduciendo por una autopista) y cuatro clases diferentes de ambientes dentro del vehículo (normal, con el aire acondicionado o ventilador encendido, con un CD reproduciéndose y con las ventanillas abiertas). Las señales de voz tomadas tanto para el entrenamiento como para las pruebas, fueron muestreadas a 16 KHz, codificadas a enteros de 16 bits y almacenadas en formato little-endian.

Car speed	In-car conditions
Idling (quiet)	Normal, Fan on, Audio on, Window open
Low speed	Normal, Fan on, Audio on, Window open
High speed	Normal, Fan on, Audio on

Tabla 3.2 - Resumen de los ambientes para la obtención de datos dentro del vehículo.

Se grabaron un total de 17.651 locuciones realizadas por 104 hablantes (52 hombres y 52 mujeres) con micrófonos CT y HF. Los datos de entrenamiento y prueba abarcan 14.687 locuciones realizadas por 73 hablantes (33 hombres y 40 mujeres) con micrófonos CT (7.492 locuciones) y HF (7.195 locuciones), y 2.964 locuciones realizadas por 31 hablantes (19 hombres y 12 mujeres) utilizando únicamente micrófonos HF.

Car speed	Microphone	In-car condition	Condition 1	Condition 2	Condition 3	Condition 4
Idling (quiet)	CT	Normal	—	—	686	686
		Fan on	—	—	686	686
		Audio on	—	—	680	680
		Window open	—	—	685	685
		Total	—	—	2,737	2,737
	HF	Normal	538	538	—	—
		Fan on	663	663	—	—
		Audio on	698	698	—	—
		Window open	498	498	—	—
		Total	2,397	2,397	—	—
	Total		**2,397**	**2,397**	**2,737**	**2,737**
Low speed	CT	Normal	—	—	685	—
		Fan on	—	—	682	—
		Audio on	—	—	690	—
		Window open	—	—	671	—
		Total	—	—	2,728	—
	HF	Normal	700	—	—	—
		Fan on	694	—	—	—
		Audio on	697	—	—	—
		Window open	666	—	—	—
		Total	2,757	—	—	—
	Total		**2,757**	—	**2,728**	—
High speed	CT	Normal	—	—	682	—
		Fan on	—	—	677	—
		Audio on	—	—	668	—
		Total	—	—	2,027	—
	HF	Normal	687	—	—	—
		Fan on	678	—	—	—
		Audio on	676	—	—	—
		Total	2,041	—	—	—
	Total		**2,041**	—	**2,027**	—
Total			**7,195**	**2,397**	**7,492**	**2,737**

Tabla 3.3 - Cantidad de datos de entrenamiento para cada condición de evaluación.

Car speed	In-car condition	Condition 1	Condition 2	Condition 3	Condition 4
Idling (quiet)	Normal	198	—	—	—
	Fan on	216	—	—	—
	Audio on	297	—	—	—
	Window open	195	—	—	—
	Total	906	—	—	—
Low speed	Normal	298	298	298	298
	Fan on	294	294	294	294
	Audio on	297	297	297	297
	Window open	291	291	291	291
	Total	1,180	1,180	1,180	1,180
High speed	Normal	293	293	293	293
	Fan on	291	291	291	291
	Audio on	294	294	294	294
	Total	878	878	878	878
Total		**2,964**	**2,058**	**2,058**	**2,058**

Tabla 3.4 - Cantidad de datos de test para cada condición de evaluación.

3.2.2 Diseño del escenario de trabajo

CENSREC-2 proporciona cuatro entornos de evaluación para reconocimiento de voz usando los datos de voz obtenidos en diferentes condiciones, descritas en la sección anterior, en el interior de un coche. Cada escenario de evaluación cumple las condiciones marcadas con un círculo (O) en las tablas 3.5 y 3.6. Para cada condición, los entornos de evaluación fueron diseñados de la siguiente forma:

- **Condición 1 (Ambiente normal):** Los datos de voz fueron obtenidos usando los mismos micrófonos y las mismas condiciones de grabación tanto para el entrenamiento como para las pruebas.
- **Condición 2 (Aire acondicionado encendido):** Los datos de entrenamiento y de prueba fueron grabados bajo diferentes condiciones de evaluación usando los mismos micrófonos.
- **Condición 3 (CD reproduciéndose):** Los datos de entrenamiento y de prueba fueron grabados bajo las mismas condiciones de evaluación usando diferentes micrófonos.
- **Condición 4 (Ventanillas abiertas):** Los datos de voz fueron obtenidos usando diferentes micrófonos y diferentes condiciones de grabación tanto para el entrenamiento como para las pruebas.

Evaluation condition	Condition 1		Condition 2		Condition 3		Condition 4	
Microphone	CT	HF	CT	HF	CT	HF	CT	HF
Idling (quiet)	—	○	—	○	○	—	○	—
Low speed	—	○	—	—	○	—	—	—
High speed	—	○	—	—	○	—	—	—

Tabla 3.5 - Resumen de los datos de entrenamiento para cada condición de evaluación.

Evaluation condition	Condition 1	Condition 2	Condition 3	Condition 4
Idling (quiet)	○	—	—	—
Low speed	○	○	○	○
High speed	○	○	○	○

Tabla 3.6 - Resumen de los datos de prueba para cada condición de evaluación. Los datos de pruebas siempre se toman con el micrófono de manos libres (HF).

3.3 Diseño

3.3.1 Adaptación de la base de datos

La base de datos CENSREC-2 está codificada en formato ".raw", por tanto, se ha realizado un script en *bash* con el fin de transformar todos los ficheros de audio a formato ".wav". Para ello se ha hecho uso del programa SOX, que nos permitirá convertir ficheros de audio de un formato a otro de manera rápida y sencilla. El script realizado tiene la siguiente estructura:

```
#Encuentra todos los ficheros .raw del directorio
Ficheros = $(find ./ -name "*.raw")

#Transforma los ficheros a formato .wav mediante SOX
for i in $ficheros;
do
    nuevo_nombre = `echo $i | sed -e "s/\.raw/.wav/"`
    sox -L -s -w -r 16000 $i $nuevo_nombre
done
```

En la llamada a la aplicación SOX, se han utilizado los siguientes parámetros para la adaptación a los parámetros definidos en la base de datos: *-L*, por estar los datos almacenados en formato little-endian, *-w*, por estar los datos codificados a enteros de 16 bits y 16000 por ser 16 KHz la frecuencia de muestreo.

3.3.2 Parametrización de la base de datos

Podemos dividir el trabajo del Reconocedor de Habla en tres etapas: parametrización, entrenamiento de los modelos y evaluación del reconocedor.

En primer lugar analizaremos la etapa de parametrización llevada a cabo mediante la herramienta MATLAB. El objetivo de este módulo de parametrización es el de representar la señal de habla de la que partimos, previamente muestreada, mediante unos parámetros que contengan la información más relevante del mensaje comprendido en la onda acústica y que eliminen la redundancia.

Tal y como se ha comentado al principio del capítulo, el proceso de parametrización de los archivos de audio se ha dividido en diferentes módulos configurables y divididos según el proceso explicado en la sección 2.3.3. A continuación pasamos a detallar el contenido de cada uno de los módulos programados.

3.3.2.1 Filtro de pre-énfasis

Cómo se explicó en la sección 2.3.4.1, debido a que la señal de voz se atenúa a medida que aumenta la frecuencia, es necesario incrementar la relevancia de las frecuencias altas. Por ello, en primer lugar, la señal de voz digitalizada pasa un filtro de pre-énfasis, típicamente un filtro FIR (*Finite Impulse Response*) de primer orden, cuya función de transferencia es $H(z) = 1 - a \cdot z^{-1}$.

El filtro de pre-énfasis utilizado ha sido el siguiente: $H(z) = 1 - 0.97 \cdot z^{-1}$
Donde a = 0,97.

Este filtro tiene un cero de transmisión que depende del valor de a, tal como se muestra en la figura 3.3. Para a = 0 el filtro es plano y para a = 1 existe un cero de transmisión en la frecuencia cero.

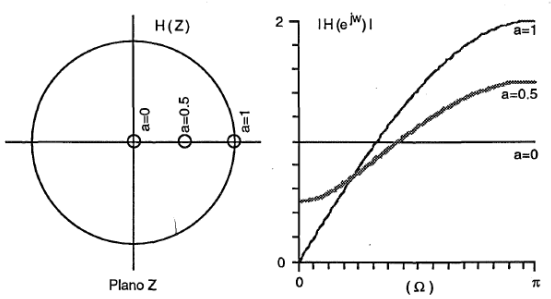

Figura 3.3 - Representación de H(Z) en el plano Z y de su módulo en el círculo unidad para distintos valores de a.

La programación del filtro utilizado se encuentra en el fichero MATLAB *enfasis.m.*

3.3.2.2 Enventanado

En la sección 2.3.4.2 se vio que el enventanado de una señal se puede considerar como la multiplicación de ésta por una señal rectangular, lo que en el espacio frecuencial se traduce en convolucionar el espectro de la señal de audio con una *sinc*. Para evitar en la medida de lo posible la aparición de ruido en la estimación espectral debido a las discontinuidades de la señal rectangular, se ha aplicado una ventana de Hamming en el proceso de enventanado de la señal.

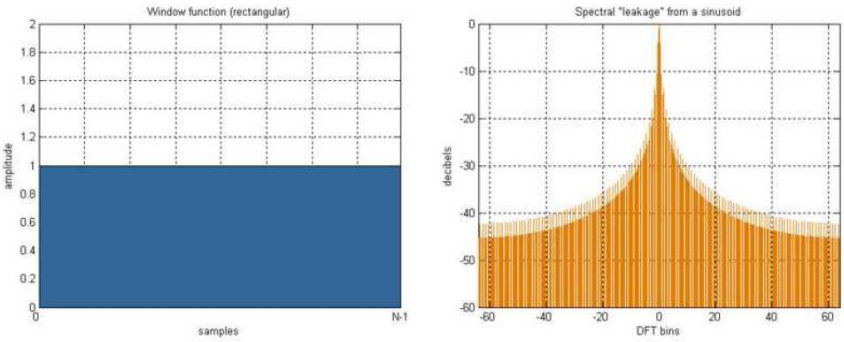

Figura 3.4 - Ventana rectangular, a la izquierda, y su espectro.

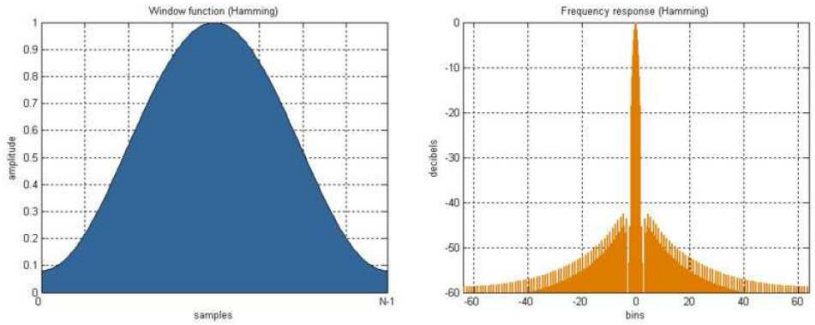

Figura 3.5 - Ventana de Hamming, a la izquierda, y su espectro.

Como observamos en las figuras 3.4 y 3.5, el espectro de la ventana de Hamming es más similar a una delta en el dominio frecuencial, por lo que la distorsión que se introduce en el espectro de la señal de audio es menor que en el caso de la ventana rectangular.

Al utilizar una ventana Hamming es preciso tener en cuenta que las muestras en los extremos de la ventana sufrirán una ponderación a diferencia de las muestras de la zona central, cuyo valor no experimentara ningún cambio. Para compensar este efecto de ponderación se hace necesario un solapamiento entre las ventanas, tal y como se muestra en la figura 3.6, de tal manera que el desplazamiento de la ventana sea inferior a la longitud de esta.

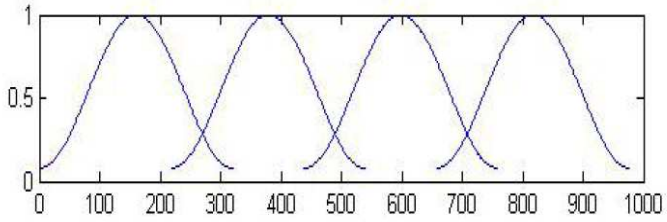

Figura 3.6 - Solapamiento entre ventanas Hamming.

Por este motivo, la longitud de la ventana Hamming utilizada ha sido de 20 ms y el solapamiento entre ellas de 10 ms.

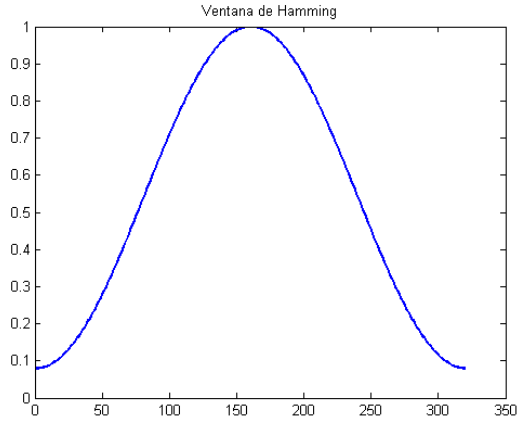

Figura 3.7 - Representación de la ventana de Hamming utilizada.

La programación del enventanado de la señal, tal y como se ha explicado en esta sección, se encuentra en el fichero MATLAB *enventanado.m.*

3.3.2.3 Aplicación de la FFT

Los coeficientes HFCC se extraen a partir de la representación de la señal de voz en el dominio espectral. Diversas investigaciones llevadas a cabo hasta la fecha han demostrado que los coeficientes obtenidos del dominio espectral representan más fielmente las características de la voz que los obtenidos del dominio temporal. Esta peculiaridad es debida a que las personas utilizan este mismo dominio para distinguir sonidos, por tanto, cabe esperar que un sistema que trabaje con características del dominio espectral se acerque más al comportamiento humano.

Para la representación en frecuencia de las señales de audio utilizadas se ha hecho uso de la transformada rápida de Fourier o FFT, ya que, según se explicó en secciones anteriores, en la práctica, este algoritmo reduce el costo computacional del cálculo de la DFT.

El tamaño de los arrays para calcular la FFT debe ser potencia de dos a fin de optimizar el cálculo de la misma. Así, en la función diseñada para este fin, se ha completado el array obtenido tras el enventanado con ceros hasta obtener uno que cumpla con esta condición. A partir de esta trama extendida se ha calculado la FFT correspondiente haciendo uso de la función *fft()* de MATLAB.

Sabemos que, dada una señal en tiempo discreto *x(n)* con N muestras, su transformada *X* (*k*), está dada por:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi kn}{N}}$$

Así, haciendo uso de la función anterior, será sencillo encontrar la correspondencia existente entre las muestras de la FFT y la frecuencia lineal, tal y como se ha representado en la siguiente expresión, lo que será de gran utilidad en el diseño del banco de filtros:

$$x[n] \text{ muestreada a } f_s \rightarrow X(e^{j2\pi f})$$

$$FFT_N\{x[n]\} = X[k], \; X[k]=X(e^{j2\pi k/Nfs}) \;\; k=0,\ldots,N/2$$

$$FFT \text{ en posición } K \rightarrow f = (k/N)\cdot f_s \;\; k=0,\ldots,N/2$$

La programación de la aplicación de la FFT sobre las muestras de audio enventanadas, tal y como se ha explicado en esta sección, se encuentra implementada en el fichero MATLAB *Aplicarfft.m*

A continuación podemos observar en la figura la representación de la relación entre las muestras de la FFT y su frecuencia lineal correspondiente.

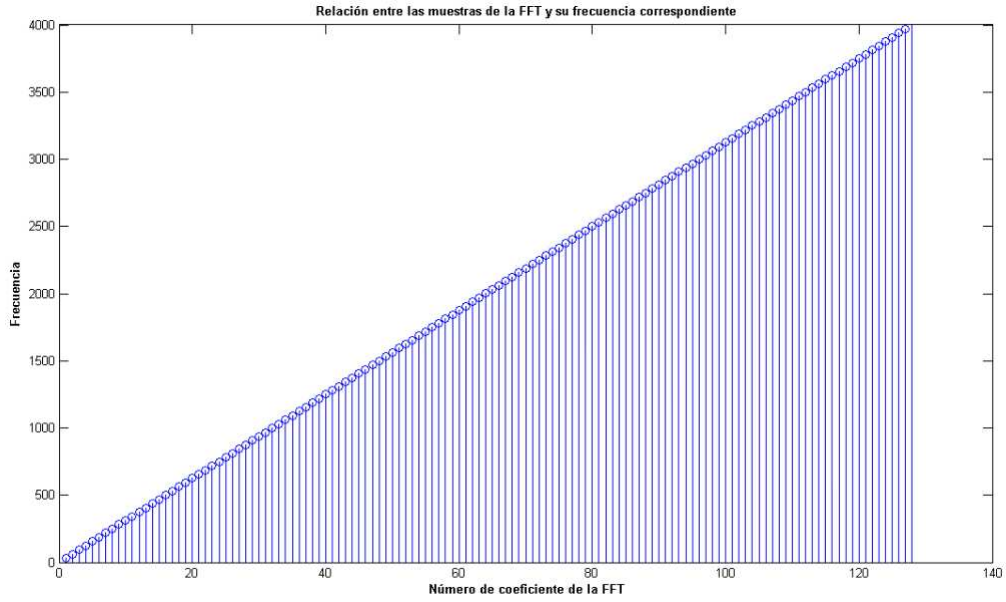

Figura 3.8 - Relación entre las muestras de la FFT y su frecuencia correspondiente.

3.3.2.4 Banco de filtros HFCC

El diseño del banco de filtros HFCC se ha llevado a cabo según lo descrito en la sección 2.5.3. Para ello, al igual que en módulos anteriores, se ha desarrollado una función en MATLAB que realiza el diseño de los filtros, así como el filtrado de la señal de audio entrante, previamente procesada por los módulos de pre-énfasis, enventanado y FFT. La cabecera de esta función presenta la siguiente estructura:

```
function[coeficientes]=BancoFiltrosHFCC(fmin,fmax,mNFilters,E_Factor,Ventana_sin_fil,fs)
```

Todos los parámetros de diseño del banco de filtros HFCC son configurables, siendo *fmin* la frecuencia mínima del banco de filtros, *fmax* la frecuencia máxima del banco de filtros, *mnFilters* el número de filtros, *E_factor* el factor de escala del ancho de banda de cada filtro, *Ventana_sin_fil* las muestras de audio procesadas por las etapas de parametrización anteriores y *fs* la frecuencia de muestreo utilizada.

Según se describió en la sección 2.5.3.3 la obtención de las frecuencias de centro, mínima y máxima de cada uno de los filtros que componen el banco de filtros se calculan como frecuencias en escala Mel a partir de las frecuencias mínima y máxima que definen el banco de filtros completo y aplicando las fórmulas definidas en esa sección. Esta conversión también se ha llevado a cabo en este módulo siendo necesario para la construcción del banco de filtros. A continuación podemos observar gráficamente cómo están relacionadas las frecuencias de centro de cada uno de los filtros en escala Mel y en Hz.

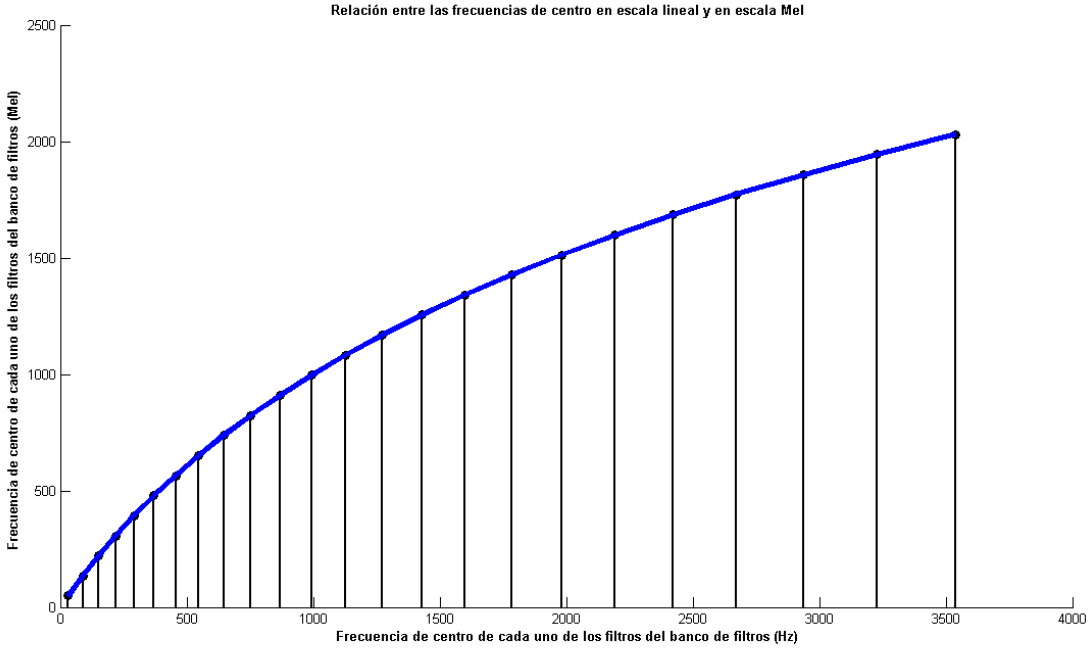

Figura 3.9 - Relación entre las frecuencias de centro en escala lineal y en escala Mel.

Según se ha detallado en secciones anteriores, suponiendo una frecuencia de muestreo de 12500 Hz, Skowronski y Harris propusieron un banco de filtros formado por 29 filtros de igual altura, cubriendo el rango de frecuencias [0, 6250 Hz]. En este estudio, se supone una frecuencia de muestreo de 16000 Hz y sólo los primeros 24 filtros que abarcan el rango de frecuencias [0, 4000 Hz].

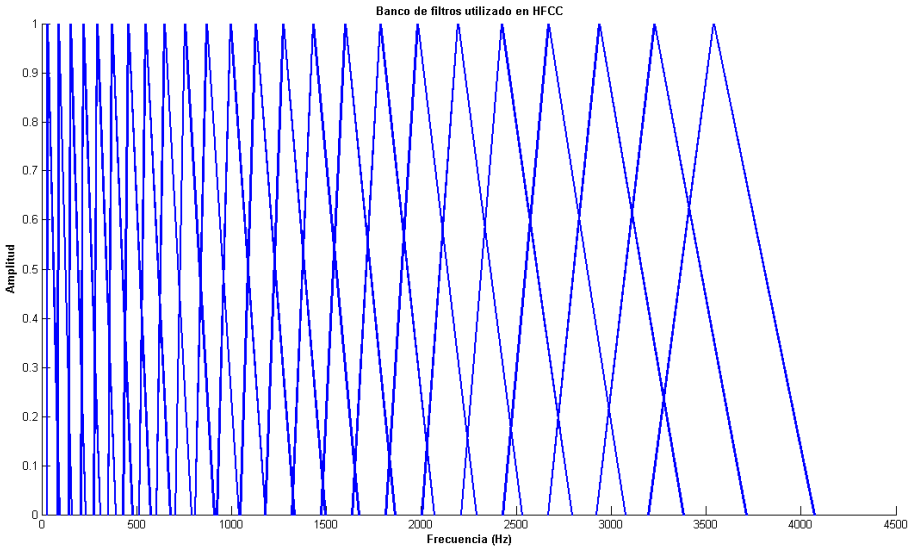

Figura 3.10 - Banco de filtros HFCC diseñado.

Como se ilustra en la figura 3.10, en HFCC el solapamiento entre los filtros es diferente al convencional, un filtro puede solapar no sólo con sus vecinos más cercanos, sino también con filtros lejanos.

De igual forma, según se ha visto, el ancho de banda de cada filtro que compone el banco de filtros, puede ser escalado por una constante, a la que Skowronski y Harris llamaron *E-factor*. Esta constante ha sido introducida como parámetro de diseño con el fin de estudiar el efecto que produce su variación en el reconocedor en general.

A continuación podemos observar gráficamente cómo varía el ERB de cada filtro del banco de filtros en función del E-factor utilizado. Para una misma frecuencia de centro, la anchura de cada filtro aumenta según lo hace el E-factor.

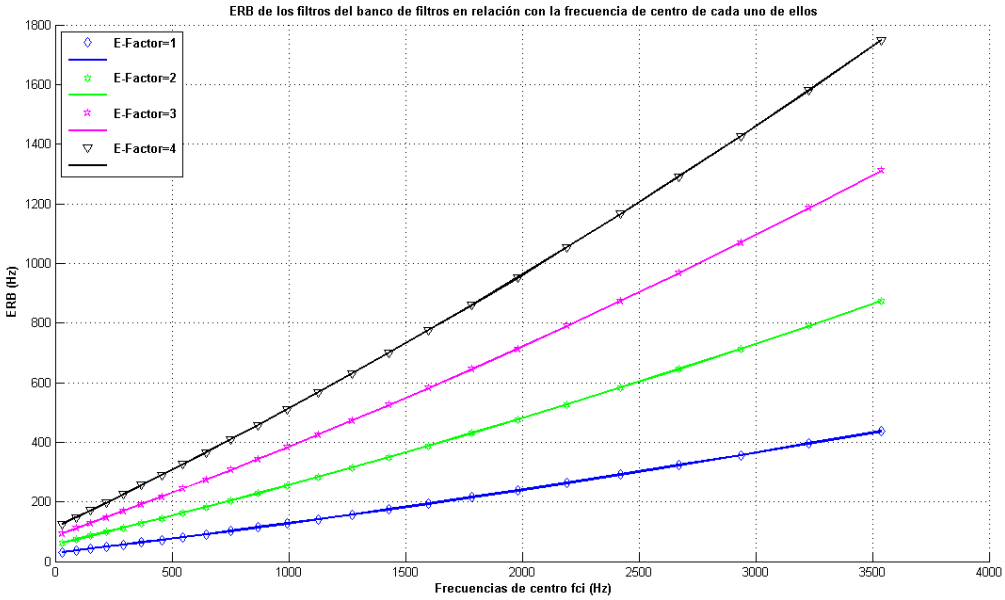

Figura 3.11 - Variación de la anchura de cada filtro en función del E-Factor.

Como la respuesta del oído no es igual para todas las bandas de frecuencia, con el fin de reducir el número de valores, la energía de la señal se agrupa en bandas de energía, siendo su tamaño variable con la frecuencia. Observando la figura 3.8, se puede ver que la primera banda está centrada aproximadamente en 250 Hz. La mayor parte de la energía que se extrae se localiza en la zona de los 250 a 3.500 Hz, captando cierta energía de baja frecuencia y muy poca de alta. Es por esta razón que en la realización del banco de filtros se ha aplicado un corte para aquellas componentes de frecuencia menores que 250 Hz, además de con el fin de adaptar nuestro filtro a las mismas condiciones que las utilizadas en la base de datos CENSREC-2.

Para finalizar el diseño del banco de filtros HFCC se aplica el logaritmo a la salida de éste. La escala logarítmica presentaría problemas para valores de energía muy bajos cercanos a cero, por ello, se ha limitado a -50dB el valor del logaritmo.

Todo el proceso de filtrado explicado en esta sección se ha desarrollado en el fichero MATLAB *BancoFiltrosHFCC.m.*

3.3.2.5 DCT

Como último paso en el proceso de parametrización, se lleva a cabo el cálculo de la DCT. Una vez calculada esta transformada con la ayuda de la función *dct()* de MATLAB, obtendremos a la salida una matriz de *MxN* coeficientes, los coeficientes HFCC. Para nuestro estudio, el vector de características estará compuesto de los 12 primeros coeficientes. Así, el proceso de parametrización completo llevado a cabo, puede ser esquematizado según se muestra en la siguiente figura:

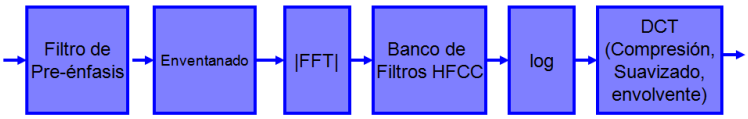

Figura 3.12 - Proceso de parametrización llevado a cabo para la extracción de los coeficientes HFCC.

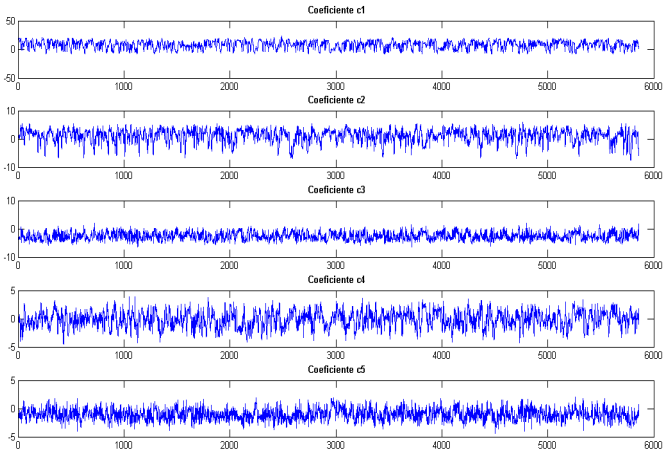

Figura 3.13 - Representación gráfica de los cinco primeros coeficientes HFCC extraídos con E-factor=1 obtenidos para el número japonés "san", con velocidad de coche a ralentí y micrófono HF.

3.3.2.6 Coeficientes ΔHFCC y ΔΔHFCC

Cuando se confecciona el vector de características para RAH, es práctica corriente considerar algunas otras variables que llevan información importante del tramo de voz considerado. Para obtener más información, como por ejemplo la de coarticulación de fonemas, es necesario introducir datos de la velocidad y aceleración de los parámetros. Por ello, al igual que lo visto en la sección 2.5.2 para los coeficientes MFCC, surgen los HFCC-Delta (ΔHFCC) y los HFCC-Delta-Delta (ΔΔHFCC), que representan la evolución temporal de los fonemas en su transición a otros fonemas. Los ΔHFCC se calculan como la variación de los coeficientes HFCC 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 y su principal misión es modelar las transiciones entre sonidos, lo que proporciona una gran cantidad de información. Los coeficientes Delta-Delta representan la variación de los coeficientes de velocidad, por lo que son llamados coeficientes de aceleración.

Para un vector de características *x(t;k)* dado, se obtienen los coeficientes delta mediante la regresión:

$$\Delta x(t;k) = \frac{\sum_{j=1}^{N_J} j\,(x(t+j;k) - x(t-j;k))}{2\sum_{j=1}^{N_J} j^2}$$

Donde N_J es utilizado para suavizar la estimación a través de los tramos (generalmente $1 \leq N_J \leq 2$). Los coeficientes de aceleración $\Delta^2 x(t; k)$ se obtienen por aplicación directa de la ecuación anterior a los $\Delta x(t; k)$.

En este estudio, haremos uso del cálculo de los coeficientes ΔHFCC, a partir de los coeficientes HFCC resultantes del proceso de parametrización, y de los coeficientes ΔΔHFCC, a partir de los coeficientes delta. El resultado, como base para el entrenamiento de nuestro reconocedor, será una matriz formada por 36 coeficientes, 12 coeficientes HFCC, 12 coeficientes ΔHFCC y 12 coeficientes ΔΔHFCC.

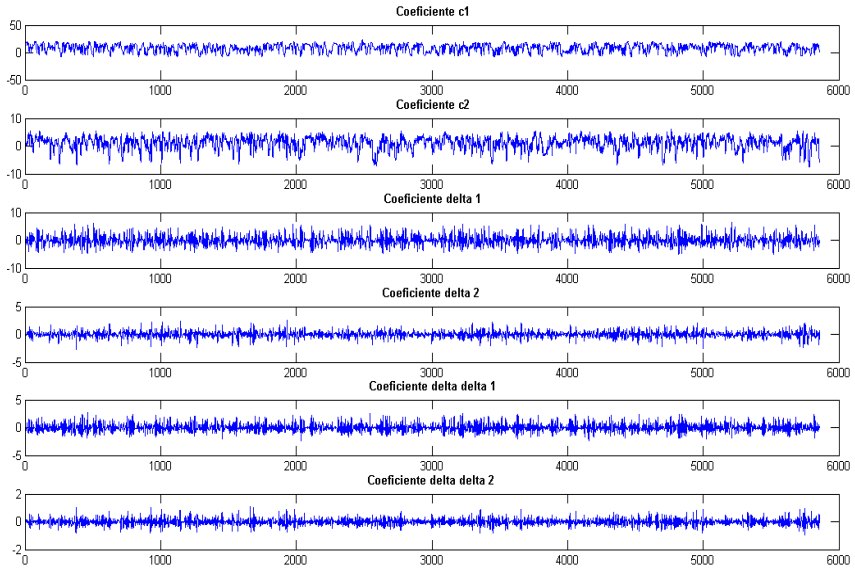

Figura 3.14 - Representación de los dos primeros coeficientes HFCC, ΔHFCC y ΔΔHFCC extraídos tras la parametrización con E-factor=1 obtenidos para el número japonés "san", con velocidad de coche a ralentí y micrófono HF.

3.3.3 Adaptación de la base de datos a formato HTK

Una vez obtenida la base de datos parametrizada, los ficheros de parámetros correspondientes deben convertirse a formato HTK para que puedan ser utilizados en los módulos de entrenamiento y test implementados con esta herramienta. Para realizar esta conversión, se ha utilizado la función *writehtk()* del toolbox *voicebox* de MATLAB. Esta función permitirá transformar las matrices obtenidas en el proceso de parametrización con los coeficientes HFCC, ΔHFCC y ΔΔHFCC a formato HTK.

Con la base de datos parametrizada en formato HTK, será necesario realizar una serie de cambios en los scripts de configuración iniciales para que todos los parámetros estén adaptados al algoritmo de extracción de características utilizado en este proyecto, tanto en el proceso de parametrización, como en el proceso de entrenamiento y pruebas.

3.4 Scripts de configuración

Los scripts de configuración proporcionados por la base de datos CENSREC-2 fueron diseñados con el fin de facilitar el entrenamiento y evaluación de los HMM en la herramienta HTK. Se ha de tener en cuenta que estos scripts fueron diseñados utilizando MFCC como algoritmo de extracción de características, por tanto, se han debido adaptar para HFCC.

El proceso a seguir en el uso de los scripts de configuración ha sido el siguiente:

1. Ajuste de las condiciones de análisis llevadas a cabo en el proceso de parametrización en los scripts **censrec2_config.pl** y **config_hcopy.**
2. Ejecución del script **init.pl** (perl/htk_baseline/init.pl).
3. Aunque en los scripts de configuración de la base de datos utilizada podemos hacer uso de un script para la extracción de características **fea _extract_htk.pl** (perl/htk_baseline/ fea_extract_htk.pl), en este caso no será necesaria su utilización, ya que la extracción de características se ha llevado a cabo mediante un proceso de parametrización propio implementado en MATLAB y descrito en la sección 3.3.
4. Ejecución del script **train.pl** (perl/htk_baseline/train.pl) para llevar a cabo el entrenamiento de los HMM. Los argumentos que se le han pasado al script han sido las diferentes condiciones de evaluación en las que se ha llevado a cabo la adquisición de la base de datos CENSREC-2, descritas en la sección 3.2.3, cond1, cond2, cond3 y cond4.
5. Ejecución del script **test.pl** (perl/htk_baseline/ test.pl) para llevar a cabo el proceso de reconocimiento. Los argumentos que se le han pasado al script han sido los mismos que para el script train.pl.
6. Los resultados del reconocimiento son almacenados en ficheros con la extensión *.res*. El formato de presentación de los resultados es un modelo de medida estandarizado por NIST (figura 3.15).

========== HTK Results Analysis ==========
---------- Overall Results ----------
SENT: %Correct=54.18 [H=1606, S=1358, N=2964]
WORD: %Corr=81.55, Acc=72.88 [H=8175, D=842, S=1008, I=869, N=10025]
==========

Figura 3.15 - Formato estandarizado por NIST para la representación de resultados en reconocimiento de habla.

Los datos mostrados en la línea encabezada por 'SENT' no son de gran relevancia, ya que indican el porcentaje de frases transcritas que coinciden con las frases originales y suelen ser bajos en reconocimiento de habla a nivel de fonemas o de palabras. Los datos representados en la línea encabezada por 'WORD' hacen referencia a medidas a nivel de palabras o fonemas, dependiendo del tipo de reconocimiento. Los valores mostrados son los siguientes:

- H: número de fonemas/palabras correctos/as.
- D: número de fonemas/palabras borrados/as.
- S: número de fonemas/palabras sustituidos/as.
- I: número de fonemas/palabras insertados/as.
- N: número de fonemas/palabras totales en la transcripción original.
- %Corr es el porcentaje de fonemas/palabras correctos/as:

$$\%Corr = \frac{H}{N}100$$

- %Acc es la precisión del sistema:

$$\%Acc = \frac{H-I}{N}100$$

En este proyecto se busca un sistema de reconocimiento con un alto valor de %Corr, sin descuidar el valor de %Acc, es decir, un alto porcentaje de aciertos que no sea debido a la inserción de un gran número de fonemas/palabras espurios. El número de inserciones se ajusta mediante un parámetro de penalización en la herramienta HTK utilizada para la obtención de las trascripciones.

3.5 Entrenamiento y evaluación de HMM

La herramienta HTK ha permitido el entrenamiento de los modelos acústicos, el reconocimiento y la extracción de resultados. El objetivo en esta parte del proyecto ha sido la creación de unos modelos, los cuales serán entrenados y probados con la base de datos CENSREC-2. HTK está basado en el algoritmo de Viterbi y dado un conjunto de parámetros de entrada elige la secuencia (de palabras o fonemas) de mayor probabilidad. Para el entrenamiento se ha utilizado el programa HERest, y para la clasificación el programa HVite. El programa HResults ha proporcionado los resultados que nos han permitido interpretar cómo han funcionado nuestros experimentos. La estructura de evaluación ha sido diseñada de la siguiente manera:

- El reconocimiento de voz se lleva a cabo usando palabras completas HMM. En el proceso de reconocimiento, se han definido un diccionario de pronunciación estándar y una gramática de reconocimiento en scripts de referencias descritos por la sintaxis de notación EBNF (*Extended Backus-Naur Form*), tal y como se muestra en la figura 3.16. En estos scripts de referencias, que tienen por objeto la evaluación de los modelos acústicos, "|" denota alternativas, "<>" se refiere a una o más repeticiones y "[]" contiene opciones. Esta gramática genera repeticiones arbitrarias de cada uno de los dígitos seguidos por pausas cortas o terminaciones en silencio, que también son permitidas.

```
$digit = one | two | three | four | five | six | seven | eight | nine | zero | oh ;

( [sil] < $digit [sp] > [sil] )
```

Figura 3.16 - Gramática escrita en notación EBNF.

- El vocabulario de CENSREC-2 se compone de 11 modelos de números en japonés ("Ichi", "ni", "san", "Yon", "ir", "roku", "nana", "Hachi", "kyu", "zero" y "maru"), un silencio ("sil") y una breve pausa ("sp"). Se lleva a cabo un entrenamiento de modelos HMM formado por 18 estados con 16 distribuciones de salida para cada dígito numérico, "sil" tiene cinco estados con tres distribuciones y "sp" tiene tres estados con una distribución. La distribución de salida para "sp" es la misma que la del tercer estado de "sil." Cada distribución numérica HMM consta de 20 Gaussianas y las de "sil" o "sp" tienen 36.

- El vector de características está compuesto de 12 HFCCs así como de sus correspondientes coeficientes delta y de aceleración. Las condiciones de análisis definen un filtro de pre-énfasis ($1 - 0{,}97 \cdot z^{-1}$) y enventanado aplicando ventana de Hamming de longitud 20 ms y solapamiento de 10 ms.

- En el análisis del banco de filtros, se ha aplicado un corte para aquellas componentes de frecuencia menores que 250 Hz para adaptar el filtro diseñado a las mismas condiciones de análisis que las expuestas en la base de datos utilizada.

4

Pruebas y resultados

4.1 Pruebas realizadas

En este capítulo se describen las pruebas experimentales más importantes realizadas en este trabajo para evaluar y comparar los distintos algoritmos de extracción de características, descritos en capítulos anteriores, en el sistema de reconocimiento implementado.

Para evaluar el comportamiento de dichas técnicas de extracción de características se ha optado por la utilización de la base de datos CENSREC-2, que implementa un sistema de reconocimiento de números en japonés mediante modelos ocultos de Markov. Los modelos ocultos de Markov son los que en estos momentos proporcionan unas mejores prestaciones en todos los sistemas en desarrollo. El hecho de que el sistema sea de palabras aisladas permitirá prescindir de las implicaciones de los niveles de conocimiento superiores: sintáctico, semántico, pragmático,... La base de datos realiza pruebas multilocutor, con un vocabulario pequeño y de poca confusibilidad como es el de los dígitos, y en diferentes condiciones ambientales en el interior de un coche.

Según se ha descrito en la sección 3.4 del capítulo anterior, el entrenamiento inicial ha sido realizado con vectores de características de 36 elementos compuestos por coeficientes HFCC (12 coeficientes), ΔHFCC (12 coeficientes) y ΔΔHFCC (12 coeficientes) extraídos mediante MATLAB y entrenados mediante modelos HMM en la herramienta HTK. Las condiciones de análisis definen un filtro de pre-énfasis ($1 - 0{,}97 \cdot z^{-1}$) y enventanado aplicando ventana de Hamming de longitud 20 ms y solapamiento de 10 ms. En el análisis del banco de filtros, se ha aplicado un corte para aquellas componentes de frecuencia menores que 250 Hz.

A continuación se describen las distintas pruebas realizadas, así como los resultados obtenidos de las mismas mediante las herramientas proporcionadas por HTK y haciendo uso de los datos de evaluación proporcionados por la base de datos.

4.2 Resultados experimentales

4.2.1 HFCC-E

Según se ha descrito en secciones anteriores, diversos estudios perceptuales presentan una estructura formada por bancos de filtros para mostrar el comportamiento del sistema auditivo humano. HFCC modifica el algoritmo original empleado en MFCC, relacionando las frecuencias de centro de esos filtros y el ancho de banda de las bandas críticas del oído mediante la función original dada por Moore y Glasberg, que utiliza el ERB para determinar la anchura de banda de los filtros auditivos.

Autores como Skowronski y Harris han demostrado que, utilizando la aproximación de Moore y Glasberg del ERB para definir el ancho de banda del filtro, HFCC muestra una mejora en la robustez al ruido en los experimentos de reconocimiento automático de voz sobre MFCC. En este trabajo se realiza un estudio sobre el efecto de variación del ancho de banda del banco de filtros en HFCC mediante la investigación de los efectos de la ampliación del ancho de banda lineal ERB y su respuesta en ambientes ruidosos. Según se ha visto en la sección anterior, esta variación del algoritmo HFCC se llama HFCC-E, por considerar un factor de escala lineal del ERB llamado E-factor.

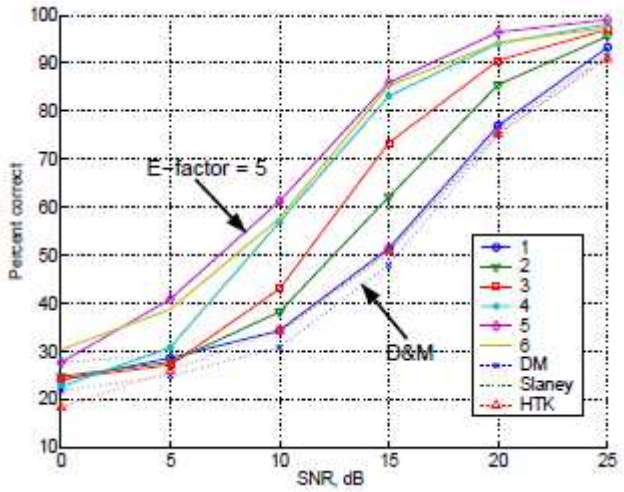

Figura 4.1 – Resultados obtenidos por Skowronski y Harris en reconocimiento de voz para diferentes realizaciones de HFCC-E y MFCC en condiciones de ruido blanco.

	SNR=5	10	15	20 dB
E-factor=1	3.7±1	3.5±1	3.5±3	1.9±2
2	2.8±2	7.4±2	14.1±5	10.4±2
3	2.3±2	12.2±3	25.4±6	15.5±5
4	5.9±4	26.1±7	35.2±6	19.0±4
5	**16.0±4**	**30.4±6**	**38.0±4**	**21.4±6**
6	13.9±2	26.9±4	37.3±3	19.2±6
Slaney	4.3±1	3.1±1	2.5±3	1.0±2
HTK	1.1±3	3.5±2	2.9±4	0.3±2

Tabla 4.1 - Resultados obtenidos por Skowronski y Harris para HFCC-E y variaciones de MFCC con ruido blanco respecto a MFCC (mejores resultados en negrita).

Tanto la tabla 4.1 como la figura 4.1, muestran los resultados obtenidos en reconocimiento por Skowronski y Harris haciendo uso del algoritmo HFCC-E con factores de escala del ERB comprendidos entre 1 y 6, y MFCC (variaciones sobre el algoritmo original: Slaney y HTK [1]) bajo condiciones de ruido blanco y gaussiano. Estos resultados están expresados en porcentaje de mejora sobre MFCC (algoritmo original introducido por Davis y Mermelstein, D&M). En general se observa una mejora en todos los experimentos realizados, concretamente, en la tabla 4.1 se observa que para un SNR de 15dB, HFCC-E, con E-factor=5, obtiene unos resultados de precisión en el reconocimiento de 38.0 ± 4 puntos porcentuales por encima de MFCC. De igual forma, de la gráfica de la figura 4.1 puede extraerse como en la región de transición entre el 40 y el 80% de reconocimiento correcto, la curva de D&M está situada aproximadamente 7dB a la derecha de la curva obtenida con HFCC-E y E-factor=5. Por tanto, los resultados obtenidos por Skowronski y Harris muestran que HFCC-E, con un E-factor=5, mejora el reconocimiento en 7dB SNR sobre MFCC. La tabla y la figura 4.2 muestran los mismos resultados pero bajo condiciones de ruido rosa.

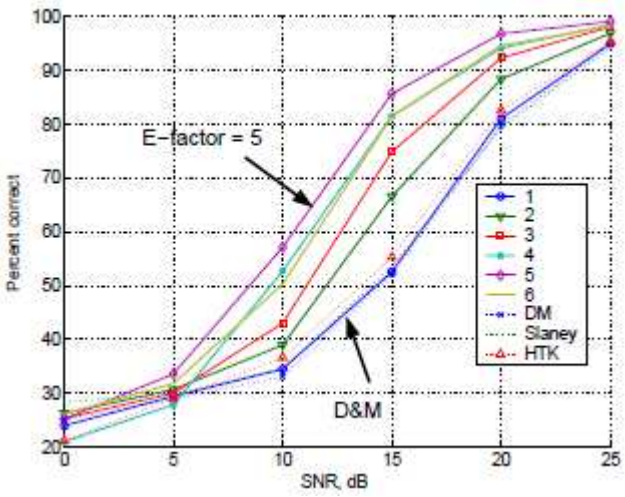

Figura 4.2 - Porcentaje de reconocimiento correcto para diferentes realizaciones de HFCC-E y MFCC vs condiciones de ruido rosa.

	SNR=5	10	15	20 dB
E-factor=1	0.5± 0.6	1.4±2	0.4±1	1.3±2
2	1.7± 1	5.9±2	14.4±4	8.6±2
3	1.1± 1	9.9±3	22.8±5	12.6±4
4	−1.1± 2	19.6±7	29.4±5	14.4±4
5	**4.7± 3**	**24.0±7**	**33.5±3**	**17.1±6**
6	2.9± 2	17.2±4	29.3±4	14.9±5
Slaney	0.9± 0.7	1.7±1	1.3±2	−0.4±2
HTK	0± 2	3.6±1	3.1±1	0.5±2

Tabla 4. 2 - Resultados obtenidos por Skowronski y Harris para HFCC-E y variaciones MFCC con ruido rosa respecto a MFCC.

En el siguiente apartado se analizarán los resultados experimentales obtenidos tras la obtención de los coeficientes HFCC para la base de datos CENSREC-2. Con el fin de analizar las posibles mejoras obtenidas en robustez frente al ruido, se han realizado experimentos para diferentes realizaciones HFCC-E con E-factors comprendidos entre 1 y 6, por ser estos valores los especificados por Skowronski y Harris en su análisis de este estudio.

4.2.2 Evaluación de resultados HFCC-E vs MFCC

En la evaluación de resultados se han utilizado como referencia los datos proporcionados por la base de datos CENSREC-2 obtenidos mediante el algoritmo de extracción de características MFCC (realización mediante HTK). De igual forma, se han comparado las transcripciones obtenidas mediante HFCC con las transcripciones reales de los datos mediante la herramienta HTK.

A continuación, en las siguientes tablas, se presentan los resultados obtenidos en RAH para cada uno de los cuatro entornos de evaluación descritos en la sección 3.2. Se han realizado experimentos para diferentes realizaciones de HFCC-E, con E-factors comprendidos entre 1 y 6, con cada una de las citadas condiciones. La primera tabla muestra los resultados de la evaluación HFCC y MFCC (*baseline*) para el porcentaje de palabras correctas, "*Word Correct*". En la siguiente tabla se muestran los resultados de HFFC y *baseline* en valores de "*Word Accuracy*", es decir, una interpretación de la precisión que tiene nuestro reconocedor. Como conclusión, en la tercera tabla se obtiene la mejora relativa de nuestro sistema, "*Relative Improvement*", respecto a los resultados de evaluación proporcionados obtenidos mediante MFCC.

La mejora relativa obtenida se obtiene de acuerdo a la siguiente expresión:

$$\text{Relative improvement} = \frac{\%\text{Acc} - \%\text{Acc of baseline}}{100 - \%\text{Acc of baseline}} \times 100\ (\%)$$

En este proyecto se busca un sistema de reconocimiento con un alto valor de "*Word correct*" (%Corr), sin descuidar el valor de "*Word accuracy*" (%Acc), es decir, un alto porcentaje de aciertos que no sea debido a la inserción de un gran número de palabras espurias. El número de inserciones se ajusta mediante un parámetro de penalización (utilizada para la obtención de las trascripciones) en la herramienta HTK, por lo que muchas inserciones supondrán una tasa de reconocimiento con valores de %Acc bajos.

CENSREC-2 HFCC Evaluation Results Word Correct (%)

	Condition 1	Condition 2	Condition 3	Condition 4	Average
Baseline	**87,01**	**80,71**	**61,53**	**46,12**	**68,84**
E=1	81,55	76,19	53,52	39,97	**62,81**
E=2	81,66	78,17	52,24	38,34	**62,6**
E=3	82,67	78,72	52,5	39,95	**63,46**
E=4	83,13	78,41	53,04	38,53	**63,28**
E=5	83,29	79,8	52,81	39,97	**63,97**
E=6	83,13	78,83	52,33	38,53	**63,21**

Tabla 4.3 – Resultados obtenidos de palabras correctas con HFCC-E para cada una de las condiciones de análisis y para E-Factors comprendidos entre 1 y 6 (en verde los resultados de la mejor realización para cada condición).

CENSREC-2 HFCC Evaluation Results Word Accuracy (%)					
	Condition 1	**Condition 2**	**Condition 3**	**Condition 4**	**Average**
Baseline	**80,23**	**68,32**	**60,46**	**43,03**	**63,01**
E=1	72,88	36,92	53,31	39,13	**50,57**
E=2	73,01	40,19	51,98	37,67	**50,71**
E=3	74,49	40,03	52,3	39,1	**51,49**
E=4	75,59	39,06	52,55	37,82	**51,26**
E=5	75,86	45,31	52,18	39,13	**53,12**
E=6	75,29	45,21	51,72	37,66	**52,47**

Tabla 4.4 - Resultados obtenidos en cuanto a precisión en el reconocimiento con MFCC y HFCC-E para cada una de las condiciones de análisis y para E-Factors comprendidos entre 1 y 6.

Si se analizan las tablas 4.3 y 4.4, se observa que los mejores resultados en reconocimiento los obtenemos, en general, con un E-factor=5 para la condición 1, en la que siempre se utilizan los mismos micrófonos (HF) y las mismas condiciones de grabación tanto para el entrenamiento como para las pruebas. Los resultados muestran una media del 83,29% de palabras correctas para las diferentes combinaciones de velocidades de coche y condiciones en el interior de éste, y una alta precisión, 75,86%.

El porcentaje de palabras correctas disminuye según complicamos las condiciones de análisis, obteniendo los resultados más bajos en reconocimiento para la condición 4, en la que los datos de voz fueron obtenidos utilizando diferentes micrófonos y diferentes condiciones de grabación tanto para el entrenamiento como para las pruebas. En este caso, para un E-factor=5, se ha obtenido un porcentaje de palabras correctas del 39,97% y una precisión en el reconocimiento del 39,13%.

Profundizando en las condiciones bajo las que se han obtenido las locuciones de voz, se producen un mayor número de inserciones durante el reconocimiento en aquellas condiciones en las cuales se han utilizado micrófonos HF tanto para el entrenamiento como para las pruebas (ver tablas anexo A). Este hecho es de destacar, ya que las inserciones son un parámetro de penalización en la herramienta HTK utilizada para la obtención de las trascripciones, lo que lleva a un reconocimiento con una menor precisión. Esto puede observarse en las condiciones 1 y 2, donde los resultados obtenidos en %Acc son peores que los obtenidos en %Corr en comparación con las condiciones 3 y 4 para cualquiera de las realizaciones de HFCC-E. Es de destacar la condición 2, en la que podemos observar un alto porcentaje de palabras correctas, 79,80%, y sin embargo, se obtiene un porcentaje mucho menor en la precisión en el reconocimiento, 45,31%, ya que realiza un gran número de inserciones. El hecho de utilizar micrófonos HF y condiciones totalmente diferentes en el entrenamiento y las pruebas puede haber influido en este comportamiento. Tal y como se ha descrito en otras secciones, el reconocimiento automático debe ser intrínsecamente robusto, es decir, debe dar las máximas prestaciones posibles en las condiciones más adversas imaginables. Las condiciones adversas para un reconocedor se definen como las diferencias o desajustes que puedan existir entre los datos con los que ha sido entrenado y los datos que debe reconocer, por lo que los resultados obtenidos para la condición 2 podrían indicar que los coeficientes HFCC son menos robustos frente a determinados cambios en las condiciones de evaluación que los coeficientes MFCC.

De igual forma, se produce un mayor número de palabras borradas (D) en aquellas condiciones en las que se han utilizado micrófonos CT para el entrenamiento y HF para las pruebas. Esto penaliza el

reconocimiento en general, obteniendo unos valores más bajos de %Corr y %Acc, como puede observarse en las condiciones 3 y 4.

CENSREC-2 HFCC Evaluation Results Relative Improvement Corr(%)					
	Condition 1	**Condition 2**	**Condition 3**	**Condition 4**	**Average**
E=1	-42,03%	-23,43%	-20,82%	-11,41%	-19,37%
E=2	-107,78%	-210,06%	-24,82%	-15,68%	-58,19%
E=3	-33,41%	-10,32%	-23,47%	-11,45%	-17,28%
E=4	-29,87%	-11,92%	-22,07%	-14,09%	-17,86%
E=5	-28,64%	-4,72%	-22,67%	-11,41%	-15,65%
E=6	-29,87%	-9,75%	-23,91%	-14,09%	-18,09%

Tabla 4.5 - Mejora relativa obtenida con HFCC-E respecto a los resultados proporcionados por la base de datos CENSREC-2 en su realización *baseline* en función del %Corr.

CENSREC-2 HFCC Evaluation Results Relative Improvement Acc (%)					
	Condition 1	**Condition 2**	**Condition 3**	**Condition 4**	**Average**
E=1	-37,18%	-99,12%	-18,08%	-6,85%	-33,64%
E=2	-36,52%	-88,79%	-21,45%	-9,41%	-33,25%
E=3	-29,03%	-89,30%	-20,64%	-6,85%	-31,15%
E=4	-23,47%	-92,36%	-20,01%	-9,15%	-31,78%
E=5	-22,10%	-72,63%	-20,94%	-6,85%	-26,74%
E=6	-24,99%	-72,95%	-22,10%	-9,43%	-28,49%

Tabla 4. 6 - Mejora relativa obtenida con HFCC-E respecto a los resultados proporcionados por la base de datos CENSREC-2 en su realización *baseline* en función del %Acc.

En la tabla 4.6 se resume la mejora relativa obtenida con el nuevo método de extracción de características respecto a MFCC en función del %Acc. Se observa una mejor precisión en el reconocimiento en general para los resultados proporcionados por la base de datos CENSREC-2 haciendo uso de MFCC, siendo un 26,74% mejor que los resultados extraídos con HFCC para su mejor realización, E=5. Hay que resaltar que esta mejora se ve muy perjudicada por los resultados extraídos en la condición 2, que como se ha visto en la tabla 4.4, obtiene un porcentaje muy bajo en la precisión en el reconocimiento, 45,31%, muy por debajo de la realización baseline, 68,32%. La mejora relativa obtenida en la condición 4 es la más aproximada, en general, a la obtenida mediante MFCC con un 6,85% de diferencia entre ambas para la mejor realización de HFCC (E=5).

Tomando como referencia los estudios de Skowronski y Harris, se ha visto que los mejores resultados en HFCC los obtuvieron para un E-Factor=5 (tablas 4.1 y 4.2). Los resultados extraídos para E<5, resultaban muy alejados de la mejor realización, de igual forma, para E=6 se observaba como los resultados de la evaluación volvían a disminuir. A la vista de los resultados obtenidos en este proyecto, podemos afirmar que, al igual

que en los estudios de Skowronski y Harris, se han obtenido los mejores resultados, en general, para HFCC con E=5, resultando las evaluaciones para E<5 peores que ésta y apreciando para E=6 como los resultados obtenidos en el reconocimiento comienzan de nuevo a disminuir. Por tanto, se puede afirmar que el hecho de incrementar el ancho de banda de los filtros en HFCC, aumenta la robustez del reconocimiento de habla en entornos adversos.

Tal y como se muestra en la tabla 4.5, la mejora relativa obtenida con HFCC-E en función del %Acc no es muy optimista. Según los datos proporcionados en la evaluación *baseline* de la base de datos CENSREC-2, se obtiene un reconocimiento mejor, tanto en %Corr como en %Acc, con MFCC, siendo los mejores resultados obtenidos con HFCC-E un 26,74% peores en función del %Acc y un 15,65% en función del %Corr . A la vista de esta conclusión, se ha de tener en cuenta que Skowronski y Harris evaluaron sus resultados en condiciones de ruido blanco y gaussiano, mientras que en este proyecto se utiliza ruido real procedente de la toma de muestras en el interior de un vehículo bajo condiciones adversas. Si se discrimina en las tablas 4.1 y 4.2, obtenidas por Skowronski y Harris, los resultados en reconocimiento obtenidos en función del valor de SNR, se observa que para un SNR de 5dB, es decir, para aquella realización en la que la fuente de ruido es mayor, son mucho peores que los obtenidos con valores para el SNR de 10, 15 y 20 dB. Lo que podría indicar una degradación en el comportamiento de los coeficientes HFCC cuando las condiciones empeoran.

A la vista de los resultados extraídos en los experimentos de HFCC-E realizados por Skowronski y Harris eran de esperar unos resultados relativamente buenos al ser extrapolados a nuestra base de datos. Parece que el hecho de utilizar ruido real, en lugar de ruido blanco, penaliza mucho la función del reconocedor, resultando los coeficientes HFCC menos robustos que los coeficientes MFCC. Este hecho es de destacar, puesto que la motivación de este estudio ha sido el de encontrar un algoritmo de extracción de características que, gracias a un diseño más aproximado al del comportamiento del sistema auditivo humano, resultase más robusto en condiciones adversas. Estos hechos pueden interpretarse como que la relación señal/ruido, si bien da indicaciones globales respecto a las características de la señal limpia comparada con la señal ruidosa, no siempre refleja con precisión los efectos del ruido sobre las bandas críticas. El análisis realizado debe complementarse con un trabajo futuro que permita establecer cuál es la vinculación entre la distorsión de las características tiempo-frecuencia y el grado de deterioro de la inteligibilidad de la señal de voz cuando ésta se contamina con ruido, medida a través de la opinión de grupos de oyentes u otras medidas objetivas adecuadas.

Aunque la mayoría de los parámetros de diseño han sido los mismos que los descritos en la base de datos utilizada, otros han sido elegidos teniendo en cuenta otros estudios en este campo por no concretarse en las especificaciones de ésta, como por ejemplo, el número de filtros utilizado en el diseño del banco de filtros, 24, así como el rango de frecuencias en el que este filtro ha sido diseñado [0 – 4000 Hz]. De igual forma, el número de coeficientes extraídos en la parametrización de la señal ha sido diferente al utilizado con HFCC. En MFCC se han añadido al vector de características, además de los propios coeficientes MFCC, ΔMFCC y $\Delta\Delta$MFCC, los coeficientes de energía, obteniendo un total de 39 coeficientes, en lugar de los 36 que se han extraído en este proyecto. Aunque el hecho de incluir el coeficiente c0 es casi equivalente a la energía, el hecho de no normalizarlo para compensar variaciones de energía debidas a la proximidad al micrófono u otros efectos colaterales indeseados ha podido influir negativamente en las tareas de reconocimiento, ya que la etapa de parametrización determina en buena parte las prestaciones del sistema, tanto en lo referente a tasas de reconocimiento como a carga computacional y requerimientos de memoria necesarios. Por tanto, se puede considerar que el problema fundamental en la parametrización ha sido la elección de un modelo inadecuado de la señal de voz. El hecho de partir de la hipótesis de estar utilizando un algoritmo más robusto al ruido que MFCC, avalado por los resultados experimentales en este campo, ha sido fundamental a la hora de intentar elegir una representación de la señal mediante un vector de parámetros con un número reducido de parámetros, el cálculo de los cuales debe exigir la mínima carga

computacional posible. Es por esta razón por lo que no se introdujeron en un principio mayor número de coeficientes en el vector de parámetros. A la vista de los resultados extraídos hubiese sido interesante el hecho de incluir más coeficientes, pero, una vez realizada la evaluación, y debido al gran tamaño de la base de datos, resulta muy costoso computacionalmente el hecho de parametrizar de nuevo todas las locuciones de voz para las diferentes realizaciones de HFCC así como su entrenamiento. Hay que tener en cuenta que el proceso completo de parametrización se ha llevado a cabo en MATLAB haciendo uso de un ordenador personal con capacidades limitadas, por lo que este proceso podía durar varios días, incluso semanas. De todas formas, se incluye como un posible trabajo futuro en esta línea de investigación con el fin de conseguir un reconocimiento lo más robusto posible.

5

Conclusiones y trabajo futuro

5.1 Conclusiones

Durante el desarrollo del proyecto y a la vista de los resultados obtenidos, extraemos las siguientes conclusiones:

- La principal conclusión es que se ha hecho uso de una nueva forma de parametrización que intenta aproximarse el comportamiento del sistema auditivo humano con el fin de obtener unos resultados en reconocimiento del habla más robustos en entornos adversos que MFCC. Estos resultados, aunque mejorables, son suficientemente buenos como para desarrollar reconocedores de voz en el estado del arte como se ha visto en el capítulo 2, así como sentar una base de estudio que intente mejorar las condiciones de análisis con el fin de obtener un reconocimiento con mayores tasas, tanto en términos de porcentaje de palabras correctas (%Corr) como de precisión del sistema (%Acc).

- Se ha conseguido obtener unos modelos acústicos que proporcionan unos buenos resultados en el reconocimiento de habla en términos de porcentaje de palabras correctas, %Corr, y de precisión del sistema, %Acc, aunque con un bajo porcentaje de mejora relativa respecto a MFCC, resultado estos últimos un 26,74% mejores en términos de precisión del sistema.

- Los malos resultados, en cuanto a precisión del sistema, obtenidos en la condición 2 para cada una de las realizaciones HFCC-E han sido determinantes a la hora de disminuir el valor obtenido en la mejora relativa del sistema desarrollado. El gran número de inserciones realizadas durante el reconocimiento, posiblemente motivado por el hecho de utilizar micrófonos HF y condiciones totalmente diferentes en el entrenamiento y las pruebas, han influido negativamente en los valores obtenidos. Esto podría indicar que los coeficientes HFCC son menos robustos frente a determinados cambios en las condiciones de evaluación que los coeficientes MFCC, ya que el reconocimiento automático debe ser intrínsecamente robusto, es decir, debe dar las máximas prestaciones posibles en las condiciones más adversas imaginables, que en un reconocedor se definen como las diferencias o desajustes que puedan existir entre los datos con los que ha sido entrenado y los datos que debe reconocer.

- El hecho de utilizar ruido real, en lugar de ruido blanco y gaussiano, como en los estudios de Skowronski y Harris, penaliza mucho la función del reconocedor. Estos hechos pueden interpretarse como que la relación señal/ruido, si bien da indicaciones globales respecto a las características de la señal limpia comparada con la señal ruidosa, no siempre refleja con precisión los efectos del ruido sobre las bandas críticas.

- En los resultados aportados por la base de datos CENSREC-2, el número de coeficientes extraídos en la parametrización de la señal ha sido diferente al utilizado con HFCC. En MFCC se han añadido al vector de características, además de los propios coeficientes MFCC, ΔMFCC y $\Delta\Delta$MFCC, los coeficientes de energía, , obteniendo un total de 39 coeficientes, en lugar de los 36 que se han extraído en este proyecto. Aunque el hecho de incluir el coeficiente c0 es casi equivalente a la energía, el hecho de no normalizarlo para compensar variaciones de energía debidas a la proximidad al micrófono u otros efectos colaterales indeseados ha podido influir negativamente en las tareas de reconocimiento, ya que la etapa de parametrización determina en buena parte las prestaciones del sistema, tanto en lo referente a tasas de reconocimiento como a carga computacional y requerimientos de memoria necesarios. Por tanto, se puede considerar que el problema fundamental en la parametrización ha sido la elección de un modelo inadecuado de la señal de voz.

- Otro hecho a tener en cuenta es que, aunque la mayoría de los parámetros de diseño han sido los mismos que los descritos en la base de datos utilizada, otros han sido elegidos teniendo en cuenta otros estudios en este campo por no concretarse en las especificaciones de ésta, como por ejemplo, el número de filtros utilizado en el diseño del banco de filtros, 24, así como el rango de frecuencias en el que este filtro ha sido diseñado [0 – 4000 Hz]. Esto, aunque a nivel de reconocimiento HFCC no resulta un problema, si que podría ver perjudicada la mejora en el reconocimiento, al realizar una comparación con otro método de extracción de características, en este caso MFCC, que ha empleado otros parámetros de diseño.

5.2 Trabajo futuro

Dados los experimentos realizados y las conclusiones extraídas sería interesante:

- A la vista de los resultados en otros trabajos de RAH parece que el aumento del número de Gaussianas para el modelado de los estados de los distintos modelos o la eliminación de silencios de cada palabra pronunciada, mejora los resultados en términos, tanto de porcentaje de fonemas/palabras y frases correctas, %Corr, como de precisión del sistema, %Acc. De esta manera, podría obtenerse un sistema de mayor exactitud y mayor tasa de aciertos sin necesidad de inserciones espurias.

- Se ha visto que la relación señal/ruido, si bien da indicaciones globales respecto a las características de la señal limpia comparada con la señal ruidosa, puede que no siempre refleje con precisión los efectos del ruido sobre las bandas críticas. El análisis realizado debe complementarse con un trabajo futuro que permita establecer cuál es la vinculación entre la distorsión de las características tiempo-frecuencia y el grado de deterioro de la inteligibilidad de la señal de voz cuando ésta se contamina con ruido, medida a través de la opinión de grupos de oyentes u otras medidas objetivas adecuadas.

- A la vista de los resultados extraídos hubiese sido interesante el hecho de incluir mayor número de coeficientes en el vector de parámetros, pero, una vez realizada la evaluación, y debido al gran tamaño de la base de datos, resulta muy costoso computacionalmente el hecho de parametrizar de nuevo todas las locuciones de voz para las diferentes realizaciones de HFCC así como su entrenamiento. Por tanto, sería interesante el hecho de modificar el proceso de parametrización realizado incluyendo más coeficientes y ver su influencia en el reconocedor en general.

- La elección de parámetros de diseño diferentes a los del reconocimiento realizado con MFCC, nos lleva a la realización de pruebas alternativas cambiando el número del banco de filtros o el rango de frecuencias, con el fin de observar cómo influye la elección de estos parámetros en el reconocedor en general.

Glosario de Acrónimos

- **RAH:** Reconocimiento Automático del Habla
- **GMM:** Gaussian Mixture Model
- **HMM:** Hidden Markov Model
- **HTK:** Hidden Markov Model Toolkit
- **LPCC:** Linear Prediction Cepstral Coefficients
- **MFCC:** Mel Frequency Cepstral Coefficients
- **HFCC:** Human Factors Cepstral Coefficients
- **DTW:** Dynamic Time Warping
- **PLP:** Perceptual Linear Predictive
- **ERB:** Equivalent Rectangular Bandwidth
- **CENSREC:** Corpus and Environments for Noisy Speech RECognition

Bibliografía

[1]. M. D. Skowronski and J. G. Harris, "Increased MFCC filter bandwidth for noise-robust phoneme recognition," in *Int. Conf. Acoust., Speech, Sign. Process.*, Orlando, Florida, 2002, pp. 801–4, IEEE.

[2]. A. de la Torre, A. Peinado, and A. Rubio. *Reconocimiento Automático de Voz en condiciones de ruido*. Monografías del Departamento de Electrónica, nº 47. Departamento de Electrónica y Tecnología de Computadores, Universidad de Granada, 2001.

[3]. Alan V. Oppenheim, Ronald W. Schafer. (1999) "Tratamiento de Señales en Tiempo Discreto".

[4]. J.D. Markel and A.H. Gray. *Linear Predictionm of Speech*. Springer-Verlag, 1976.

[5]. J. Makhoul. Spectral analysis of speech by linear prediction. *IEEE Transactions on Acoustic, Speech, Signal Processing*, 21, 3:140–148, 1973.

[6]. A. Cole, J. Mariani, H. Uzkoreit, A. Zaenen, and V. Zue. *Survey of the State of the Art in Human Language Technology*. National Science Foundation, Directorate XIII-E of the Commision of the European Commnunities Center for Spoken Language Understanding, Oregon Graduate Institute, 1995.

[7]. L.E. Baum, "An Inequality and Associated Maximization Technique ¡n Statistical Estimation of Probabilistic Functions of Markov Processes", Inequalities, vol. 3, pp. 1-8, 1972.

[8]. U.K. Baker, "The DRAGON System - An Overview", IEEE Trans. ASSP, vol. 23, n9 1, pp. 24-29, Febrero 1975.

[9]. R. Bakis. "Continuous Speech Recognition via Centisecond Acoustic States", en 91st Meeting of the Acoustical Society of America, Abril 1976.

[10]. F. Jelinek, "Continuous Speech Recognition by Statistical Methods", Proc. IEEE, vol. 64, n9 4, pp. 532-556, 1976.

[11]. L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of IEEE,*, 77, 2:257–286, 1989.

[12]. Andrew J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13,2:260–269, 1967.

[13]. L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, 41:164–171, 1970.

[14]. 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.

[15]. 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.

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

[17]. H. Hermansky, "Perceptual linear prediction (PLP) analysis for speech," *J.Acoust. Soc. Am.*, vol. 87, pp. 1738–1752, 1990.

[18]. C. P. Chan, P. C. Ching, and T. Lee, "Noisy speech recognition using denoised multiresolution analysis acoustic features," *J. Acoust. Soc. Am.*, vol. 110, no. 5, pp. 2567–2574, November 2001.

[19]. J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 2040–2050, October 1999.

[20]. B. Strope and A. Alwan, "A model of dynamic auditory perception and its application to robust word recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 451–464, September 1997.

[21]. R. Singh, M. L. Seltzer, B. Raj, and R. M. Stern, "Speech in noisy environments: robust automatic segmentation, feature extraction, and hypothesis combination," in *Int. Conf. Acoust., Speech, Sign. Process.*, Salt Lake City, Utah, 2001, pp. 273–276, IEEE.

[22]. R. Sinha and S. Umesh, "Non-uniform scaling based speaker normalization," in *Int. Conf. Acoust., Speech, and Sign. Process.*, Orlando, Florida, 2002, pp. 589–592, IEEE.

[23]. Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*. Prentice Hall PTR, 1993.

[24]. B. C. J. Moore and B. R. Glasberg, "Suggested formula for calculating auditory-filter bandwidth and excitation patterns," *J. Acoust. Soc. Am.*, vol. 74, pp. 750–753, 1983.

[25]. S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka, T. Nishiura, A. Sasou, M. Mizumachi, C. Miyajima, M. Fujimoto, and T. Endo, "AURORA-2J, An Evaluation Framework for Japanese Noisy Speech Recognition," *IEICE Transactions on Information and Systems*, Vol. E88-D, No. 3, pp. 535-544, Mar. 2005.

[26]. K. Takeda, H. Fujimura, K. Itou, N. Kawaguchi, S. Matsubara, and F. Itakura, "Construction and Evaluation a Large In-Car Speech Corpus," IEICE Transactions on Information and Systems, Vol. E88-D, No. 3, pp. 553-561, Mar. 2005.

[27]. Skowronski, Mark D., Harris, John G., *"Exploiting independent filter bandwidth of human factor ceptral coefficients in automatic speech recognition"*, June 2004.

[28]. Skowronski, Mark D., Harris, John G., *"Improving the filter bank of a classic speech feature extraction algorithm"*, IEEE Intl Symposium on Circuits and Systems, Bangkok, Thailand, vol IV, pp 281-284, May 2003.

[29]. Skowronski, Mark D., Harris, John G., *"Human Factor Cepstral Coefficients"*, December 2002.

[30]. Todor Dimitrov Ganchev, *"Speaker recognition"*, November 2005.

Tablas de resultados

A.1 Resultados evaluación HFCC con E-factor=1

```
========================== HTK Results Analysis ==========================
----------------------- Speaker Results Condition 1 ------------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-------------------------------------------------------------------------
cahf:  83.85( 76.99)  [H= 831, D= 81, S= 79, I= 68, N= 991]  55.44 [N=294]
cmhf:  79.50( 63.82)  [H= 791, D= 72, S=132, I=156, N= 995]  38.38 [N=297]
cnhf:  90.24( 81.08)  [H= 906, D= 21, S= 77, I= 92, N=1004]  61.07 [N=298]
cwhf:  72.52( 63.13)  [H= 710, D=154, S=115, I= 92, N= 979]  43.30 [N=291]
eahf:  73.01( 68.61)  [H= 714, D=148, S=116, I= 43, N= 978]  49.14 [N=291]
emhf:  73.20( 51.37)  [H= 721, D=109, S=155, I=215, N= 985]  29.25 [N=294]
enhf:  77.55( 64.90)  [H= 760, D= 96, S=124, I=124, N= 980]  44.03 [N=293]
iahf:  90.47( 89.13)  [H= 674, D= 21, S= 50, I= 10, N= 745]  77.31 [N=216]
imhf:  76.80( 71.60)  [H= 768, D=125, S=107, I= 52, N=1000]  52.86 [N=297]
inhf:  95.51( 93.91)  [H= 659, D=  7, S= 24, I= 11, N= 690]  85.86 [N=198]
iwhf:  94.54( 93.66)  [H= 641, D=  8, S= 29, I=  6, N= 678]  86.67 [N=195]
----------------------------- Overall Results -----------------------------
SENT: %Correct=54.18 [H=1606, S=1358, N=2964]
WORD: %Corr=81.55, Acc=72.88 [H=8175, D=842, S=1008, I=869, N=10025]
=========================================================================
```

```
========================== HTK Results Analysis ==========================
----------------------- Speaker Results Condition 2 ------------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-------------------------------------------------------------------------
cahf:  83.35( 64.38)  [H= 826, D= 53, S=112, I=188, N= 991]  44.90 [N=294]
cmhf:  76.98( 52.56)  [H= 766, D= 64, S=165, I=243, N= 995]  27.95 [N=297]
cnhf:  88.25( 56.97)  [H= 886, D= 16, S=102, I=314, N=1004]  32.89 [N=298]
cwhf:  72.52( 38.61)  [H= 710, D= 63, S=206, I=332, N= 979]  27.15 [N=291]
eahf:  71.17( 12.37)  [H= 696, D= 20, S=262, I=575, N= 978]  11.34 [N=291]
emhf:  68.22( 14.82)  [H= 672, D= 61, S=252, I=526, N= 985]   9.52 [N=294]
enhf:  72.45( 17.76)  [H= 710, D= 27, S=243, I=536, N= 980]   1.71 [N=293]
----------------------------- Overall Results -----------------------------
SENT: %Correct=22.25 [H=458, S=1600, N=2058]
WORD: %Corr=76.19, Acc=36.92 [H=5266, D=304, S=1342, I=2714, N=6912]
=========================================================================
```

```
========================== HTK Results Analysis ==========================
----------------------- Speaker Results Condition 3 ------------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-------------------------------------------------------------------------
cahf:  54.99( 54.69)  [H= 545, D=355, S= 91, I=  3, N= 991]  38.10 [N=294]
cmhf:  56.48( 56.18)  [H= 562, D=330, S=103, I=  3, N= 995]  38.38 [N=297]
cnhf:  67.53( 67.33)  [H= 678, D=226, S=100, I=  2, N=1004]  48.99 [N=298]
cwhf:  47.60( 47.60)  [H= 466, D=423, S= 90, I=  0, N= 979]  35.05 [N=291]
eahf:  47.24( 47.24)  [H= 462, D=414, S=102, I=  0, N= 978]  30.93 [N=291]
emhf:  48.22( 47.92)  [H= 475, D=407, S=103, I=  3, N= 985]  29.59 [N=294]
enhf:  52.14( 51.84)  [H= 511, D=355, S=114, I=  3, N= 980]  35.84 [N=293]
----------------------------- Overall Results -----------------------------
SENT: %Correct=36.73 [H=756, S=1302, N=2058]
WORD: %Corr=53.52, Acc=53.31 [H=3699, D=2510, S=703, I=14, N=6912]
=========================================================================
```

```
=========================== HTK Results Analysis ===========================
---------------------- Speaker Results Condition 4 -------------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-----------------------------------------------------------------------------
cahf:  45.61( 45.41)  [H= 452, D=410, S=129, I=  2, N= 991]  29.93 [N=294]
cmhf:  45.33( 45.23)  [H= 451, D=416, S=128, I=  1, N= 995]  28.28 [N=297]
cnhf:  57.77( 55.08)  [H= 580, D=267, S=157, I= 27, N=1004]  33.56 [N=298]
cwhf:  35.96( 35.85)  [H= 352, D=509, S=118, I=  1, N= 979]  28.52 [N=291]
eahf:  31.29( 31.29)  [H= 306, D=496, S=176, I=  0, N= 978]  20.27 [N=291]
emhf:  33.81( 33.50)  [H= 333, D=495, S=157, I=  3, N= 985]  22.11 [N=294]
enhf:  29.49( 27.24)  [H= 289, D=490, S=201, I= 22, N= 980]  15.36 [N=293]
---------------------------- Overall Results --------------------------------
SENT: %Correct=25.46 [H=524, S=1534, N=2058]
WORD: %Corr=39.97, Acc=39.13 [H=2763, D=3083, S=1066, I=56, N=6912]
=============================================================================
```

Tabla A 1 - Resultados obtenidos para HFCC-E con E=1 para cada una de las cuatro condiciones de análisis.

A.2 Resultados evaluación HFCC con E-factor=2

```
=========================== HTK Results Analysis ===========================
---------------------- Speaker Results Condition 1 -------------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-----------------------------------------------------------------------------
cahf:  82.44( 75.78)  [H= 817, D= 77, S= 97, I= 66, N= 991]  53.74 [N=294]
cmhf:  81.21( 62.61)  [H= 808, D= 69, S=118, I=185, N= 995]  37.71 [N=297]
cnhf:  91.53( 84.66)  [H= 919, D= 21, S= 64, I= 69, N=1004]  67.45 [N=298]
cwhf:  71.71( 61.39)  [H= 702, D=165, S=112, I=101, N= 979]  38.49 [N=291]
eahf:  72.80( 68.61)  [H= 712, D=150, S=116, I= 41, N= 978]  49.48 [N=291]
emhf:  74.11( 48.93)  [H= 730, D= 88, S=167, I=248, N= 985]  30.27 [N=294]
enhf:  78.57( 70.61)  [H= 770, D= 97, S=113, I= 78, N= 980]  47.10 [N=293]
iahf:  91.41( 89.53)  [H= 681, D= 25, S= 39, I= 14, N= 745]  75.46 [N=216]
imhf:  75.70( 70.70)  [H= 757, D=125, S=118, I= 50, N=1000]  50.51 [N=297]
inhf:  94.64( 93.48)  [H= 653, D=  8, S= 29, I=  8, N= 690]  81.82 [N=198]
iwhf:  93.95( 92.92)  [H= 637, D=  9, S= 32, I=  7, N= 678]  84.10 [N=195]
---------------------------- Overall Results --------------------------------
SENT: %Correct=53.74 [H=1593, S=1371, N=2964]
WORD: %Corr=81.66, Acc=73.01 [H=8186, D=834, S=1005, I=867, N=10025]
=============================================================================
```

```
=========================== HTK Results Analysis ===========================
---------------------- Speaker Results Condition 2 -------------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-----------------------------------------------------------------------------
cahf:  85.17( 67.31)  [H= 844, D= 45, S=102, I=177, N= 991]  50.00 [N=294]
cmhf:  80.60( 56.18)  [H= 802, D= 56, S=137, I=243, N= 995]  30.64 [N=297]
cnhf:  89.94( 63.25)  [H= 903, D= 11, S= 90, I=268, N=1004]  40.60 [N=298]
cwhf:  73.54( 40.45)  [H= 720, D= 69, S=190, I=324, N= 979]  27.15 [N=291]
eahf:  72.80( 12.78)  [H= 712, D= 31, S=235, I=587, N= 978]   9.62 [N=291]
emhf:  69.75( 15.13)  [H= 687, D= 49, S=249, I=538, N= 985]   6.12 [N=294]
enhf:  75.00( 25.20)  [H= 735, D= 54, S=191, I=488, N= 980]   3.41 [N=293]
---------------------------- Overall Results --------------------------------
SENT: %Correct=24.00 [H=494, S=1564, N=2058]
WORD: %Corr=78.17, Acc=40.19 [H=5403, D=315, S=1194, I=2625, N=6912]
=============================================================================
```

```
=========================== HTK Results Analysis ===============================
---------------------- Speaker Results Condition 3 ------------------------
------------------------------------------------------------------------------
cahf:  54.59( 54.29)  [H= 541, D=355, S= 95, I=  3, N= 991]  38.78 [N=294]
cmhf:  54.17( 53.87)  [H= 539, D=354, S=102, I=  3, N= 995]  37.04 [N=297]
cnhf:  65.14( 65.04)  [H= 654, D=239, S=111, I=  1, N=1004]  47.65 [N=298]
cwhf:  47.29( 47.09)  [H= 463, D=420, S= 96, I=  2, N= 979]  33.68 [N=291]
eahf:  45.50( 45.40)  [H= 445, D=418, S=115, I=  1, N= 978]  30.58 [N=291]
emhf:  47.51( 47.31)  [H= 468, D=403, S=114, I=  2, N= 985]  29.59 [N=294]
enhf:  51.12( 50.51)  [H= 501, D=358, S=121, I=  6, N= 980]  34.13 [N=293]
---------------------------- Overall Results -----------------------------
SENT: %Correct=35.96 [H=740, S=1318, N=2058]
WORD: %Corr=52.24, Acc=51.98 [H=3611, D=2547, S=754, I=18, N=6912]
==============================================================================
```

```
=========================== HTK Results Analysis ===============================
---------------------- Speaker Results Condition 4 ------------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
------------------------------------------------------------------------------
cahf:  43.59( 43.09)  [H= 432, D=429, S=130, I=  5, N= 991]  28.23 [N=294]
cmhf:  43.32( 43.02)  [H= 431, D=429, S=135, I=  3, N= 995]  25.25 [N=297]
cnhf:  53.49( 51.49)  [H= 537, D=306, S=161, I= 20, N=1004]  31.88 [N=298]
cwhf:  35.75( 35.55)  [H= 350, D=509, S=120, I=  2, N= 979]  27.84 [N=291]
eahf:  30.67( 30.47)  [H= 300, D=509, S=169, I=  2, N= 978]  20.27 [N=291]
emhf:  33.40( 33.20)  [H= 329, D=511, S=145, I=  2, N= 985]  20.75 [N=294]
enhf:  27.65( 26.43)  [H= 271, D=522, S=187, I= 12, N= 980]  17.06 [N=293]
---------------------------- Overall Results -----------------------------
SENT: %Correct=24.49 [H=504, S=1554, N=2058]
WORD: %Corr=38.34, Acc=37.67 [H=2650, D=3215, S=1047, I=46, N=6912]
==============================================================================
```

Tabla A 2 - Resultados obtenidos para HFCC-E con E=2 para cada una de las cuatro condiciones de análisis.

A.3 Resultados evaluación HFCC con E-factor=3

```
=========================== HTK Results Analysis ===============================
---------------------- Speaker Results Condition 1 ------------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
------------------------------------------------------------------------------
cahf:  83.55( 78.30)  [H= 828, D= 78, S= 85, I= 52, N= 991]  59.86 [N=294]
cmhf:  81.61( 65.63)  [H= 812, D= 78, S=105, I=159, N= 995]  43.10 [N=297]
cnhf:  90.74( 87.35)  [H= 911, D= 31, S= 62, I= 34, N=1004]  72.82 [N=298]
cwhf:  72.11( 61.80)  [H= 706, D=142, S=131, I=101, N= 979]  39.52 [N=291]
eahf:  74.64( 65.64)  [H= 730, D=145, S=103, I= 88, N= 978]  43.64 [N=291]
emhf:  75.53( 51.88)  [H= 744, D= 76, S=165, I=233, N= 985]  28.91 [N=294]
enhf:  80.82( 75.82)  [H= 792, D= 96, S= 92, I= 49, N= 980]  58.02 [N=293]
iahf:  92.75( 91.14)  [H= 691, D= 18, S= 36, I= 12, N= 745]  81.94 [N=216]
imhf:  76.50( 69.30)  [H= 765, D=112, S=123, I= 72, N=1000]  50.84 [N=297]
inhf:  95.80( 94.49)  [H= 661, D=  3, S= 26, I=  9, N= 690]  85.35 [N=198]
iwhf:  95.58( 93.95)  [H= 648, D=  6, S= 24, I= 11, N= 678]  84.10 [N=195]
---------------------------- Overall Results -----------------------------
SENT: %Correct=56.65 [H=1679, S=1285, N=2964]
WORD: %Corr=82.67, Acc=74.49 [H=8288, D=785, S=952, I=820, N=10025]
==============================================================================
```

```
=========================== HTK Results Analysis ===============================
---------------------- Speaker Results Condition 2 ------------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
---------------------------------------------------------------------------
cahf:  84.36( 65.29)  [H= 836, D= 46, S=109, I=189, N= 991]  43.88 [N=294]
cmhf:  79.50( 55.08)  [H= 791, D= 55, S=149, I=243, N= 995]  30.98 [N=297]
cnhf:  90.64( 63.35)  [H= 910, D= 12, S= 82, I=274, N=1004]  36.58 [N=298]
cwhf:  75.18( 42.39)  [H= 736, D= 41, S=202, I=321, N= 979]  30.93 [N=291]
eahf:  74.54( 11.04)  [H= 729, D= 13, S=236, I=621, N= 978]   6.53 [N=291]
emhf:  70.86( 17.06)  [H= 698, D= 50, S=237, I=530, N= 985]   9.52 [N=294]
enhf:  75.61( 25.00)  [H= 741, D= 51, S=188, I=496, N= 980]   6.83 [N=293]
---------------------------- Overall Results ------------------------------
SENT: %Correct=23.66 [H=487, S=1571, N=2058]
WORD: %Corr=78.72, Acc=40.03 [H=5441, D=268, S=1203, I=2674, N=6912]
===========================================================================
```

```
=========================== HTK Results Analysis ===============================
---------------------- Speaker Results Condition 3 ------------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
---------------------------------------------------------------------------
cahf:  54.29( 53.99)  [H= 538, D=364, S= 89, I=  3, N= 991]  39.12 [N=294]
cmhf:  54.67( 54.37)  [H= 544, D=347, S=104, I=  3, N= 995]  38.05 [N=297]
cnhf:  66.14( 65.94)  [H= 664, D=241, S= 99, I=  2, N=1004]  48.99 [N=298]
cwhf:  45.97( 45.86)  [H= 450, D=434, S= 95, I=  1, N= 979]  32.30 [N=291]
eahf:  45.50( 45.40)  [H= 445, D=416, S=117, I=  1, N= 978]  28.52 [N=291]
emhf:  47.92( 47.82)  [H= 472, D=404, S=109, I=  1, N= 985]  29.59 [N=294]
enhf:  52.65( 52.35)  [H= 516, D=347, S=117, I=  3, N= 980]  34.13 [N=293]
---------------------------- Overall Results ------------------------------
SENT: %Correct=35.86 [H=738, S=1320, N=2058]
WORD: %Corr=52.50, Acc=52.30 [H=3629, D=2553, S=730, I=14, N=6912]
===========================================================================
```

```
=========================== HTK Results Analysis ===============================
---------------------- Speaker Results Condition 4 ------------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
---------------------------------------------------------------------------
cahf:  43.39( 43.29)  [H= 430, D=407, S=154, I=  1, N= 991]  25.51 [N=294]
cmhf:  45.53( 45.13)  [H= 453, D=385, S=157, I=  4, N= 995]  24.92 [N=297]
cnhf:  54.88( 52.59)  [H= 551, D=277, S=176, I= 23, N=1004]  31.54 [N=298]
cwhf:  36.36( 36.16)  [H= 356, D=479, S=144, I=  2, N= 979]  24.40 [N=291]
eahf:  32.72( 32.52)  [H= 320, D=471, S=187, I=  2, N= 978]  18.90 [N=291]
emhf:  36.04( 35.63)  [H= 355, D=444, S=186, I=  4, N= 985]  20.07 [N=294]
enhf:  30.20( 28.16)  [H= 296, D=488, S=196, I= 20, N= 980]  17.06 [N=293]
---------------------------- Overall Results ------------------------------
SENT: %Correct=23.23 [H=478, S=1580, N=2058]
WORD: %Corr=39.95, Acc=39.10 [H=2761, D=2951, S=1200, I=56, N=6912]
===========================================================================
```

Tabla A 3 - Resultados obtenidos para HFCC-E con E=3 para cada una de las cuatro condiciones de análisis.

A.4 Resultados evaluación HFCC con E-factor=4

```
=========================== HTK Results Analysis ===========================
----------------------- Speaker Results Condition 1 -----------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-------------------------------------------------------------------------
cahf:  85.57( 82.44)  [H= 848, D= 74, S= 69, I= 31, N= 991]  67.69 [N=294]
cmhf:  79.70( 66.13)  [H= 793, D= 75, S=127, I=135, N= 995]  42.42 [N=297]
cnhf:  91.43( 88.05)  [H= 918, D= 28, S= 58, I= 34, N=1004]  74.50 [N=298]
cwhf:  73.14( 65.07)  [H= 716, D=144, S=119, I= 79, N= 979]  44.33 [N=291]
eahf:  76.18( 65.95)  [H= 745, D=134, S= 99, I=100, N= 978]  42.96 [N=291]
emhf:  76.45( 54.21)  [H= 753, D= 89, S=143, I=219, N= 985]  32.65 [N=294]
enhf:  81.63( 76.94)  [H= 800, D= 91, S= 89, I= 46, N= 980]  57.68 [N=293]
iahf:  92.48( 90.60)  [H= 689, D= 20, S= 36, I= 14, N= 745]  77.78 [N=216]
imhf:  76.30( 68.60)  [H= 763, D=111, S=126, I= 77, N=1000]  49.16 [N=297]
inhf:  95.65( 94.06)  [H= 660, D=  4, S= 26, I= 11, N= 690]  84.34 [N=198]
iwhf:  95.72( 94.25)  [H= 649, D=  7, S= 22, I= 10, N= 678]  87.18 [N=195]
---------------------------- Overall Results ----------------------------
SENT: %Correct=57.93 [H=1717, S=1247, N=2964]
WORD: %Corr=83.13, Acc=75.59 [H=8334, D=777, S=914, I=756, N=10025]
=========================================================================
```

```
=========================== HTK Results Analysis ===========================
----------------------- Speaker Results Condition 2 -----------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-------------------------------------------------------------------------
cahf:  85.27( 64.88)  [H= 845, D= 48, S= 98, I=202, N= 991]  45.92 [N=294]
cmhf:  79.60( 56.28)  [H= 792, D= 56, S=147, I=232, N= 995]  33.33 [N=297]
cnhf:  90.34( 63.35)  [H= 907, D= 15, S= 82, I=271, N=1004]  37.58 [N=298]
cwhf:  75.69( 38.82)  [H= 741, D= 33, S=205, I=361, N= 979]  29.55 [N=291]
eahf:  73.93(  5.21)  [H= 723, D= 16, S=239, I=672, N= 978]   6.53 [N=291]
emhf:  70.66( 19.90)  [H= 696, D= 53, S=236, I=500, N= 985]  12.93 [N=294]
enhf:  73.06( 23.88)  [H= 716, D= 85, S=179, I=482, N= 980]   6.83 [N=293]
---------------------------- Overall Results ----------------------------
SENT: %Correct=24.73 [H=509, S=1549, N=2058]
WORD: %Corr=78.41, Acc=39.06 [H=5420, D=306, S=1186, I=2720, N=6912]
=========================================================================
```

```
=========================== HTK Results Analysis ===========================
----------------------- Speaker Results Condition 3 -----------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-------------------------------------------------------------------------
cahf:  54.49( 53.99)  [H= 540, D=343, S=108, I=  5, N= 991]  37.76 [N=294]
cmhf:  56.08( 55.88)  [H= 558, D=323, S=114, I=  2, N= 995]  35.35 [N=297]
cnhf:  65.94( 64.94)  [H= 662, D=227, S=115, I= 10, N=1004]  47.32 [N=298]
cwhf:  46.58( 46.27)  [H= 456, D=403, S=120, I=  3, N= 979]  29.55 [N=291]
eahf:  46.42( 45.71)  [H= 454, D=390, S=134, I=  7, N= 978]  28.52 [N=291]
emhf:  48.63( 48.43)  [H= 479, D=378, S=128, I=  2, N= 985]  28.91 [N=294]
enhf:  52.76( 52.24)  [H= 517, D=321, S=142, I=  5, N= 980]  34.81 [N=293]
---------------------------- Overall Results ----------------------------
SENT: %Correct=34.65 [H=713, S=1345, N=2058]
WORD: %Corr=53.04, Acc=52.55 [H=3666, D=2385, S=861, I=34, N=6912]
=========================================================================
```

```
============================ HTK Results Analysis ==============================
----------------------- Speaker Results Condition 4 ------------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-------------------------------------------------------------------------
cahf:  42.38( 41.88)  [H= 420, D=412, S=159, I=  5, N= 991]  26.19 [N=294]
cmhf:  44.12( 43.82)  [H= 439, D=411, S=145, I=  3, N= 995]  25.59 [N=297]
cnhf:  53.78( 51.99)  [H= 540, D=274, S=190, I= 18, N=1004]  33.22 [N=298]
cwhf:  34.32( 34.22)  [H= 336, D=497, S=146, I=  1, N= 979]  23.02 [N=291]
eahf:  30.57( 30.37)  [H= 299, D=480, S=199, I=  2, N= 978]  17.18 [N=291]
emhf:  33.50( 33.10)  [H= 330, D=471, S=184, I=  4, N= 985]  17.01 [N=294]
enhf:  30.51( 28.88)  [H= 299, D=460, S=221, I= 16, N= 980]  16.72 [N=293]
----------------------------- Overall Results -----------------------------
SENT: %Correct=22.74 [H=468, S=1590, N=2058]
WORD: %Corr=38.53, Acc=37.82 [H=2663, D=3005, S=1244, I=49, N=6912]
=========================================================================
```

Tabla A 4 - Resultados obtenidos para HFCC-E con E=4 para cada una de las cuatro condiciones de análisis.

A.5 Resultados evaluación HFCC con E-factor=5

```
============================ HTK Results Analysis ==============================
----------------------- Speaker Results Condition 1 ------------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-------------------------------------------------------------------------
cahf:  85.97( 82.74)  [H= 852, D= 70, S= 69, I= 32, N= 991]  69.39 [N=294]
cmhf:  80.60( 65.33)  [H= 802, D= 76, S=117, I=152, N= 995]  40.74 [N=297]
cnhf:  91.83( 89.54)  [H= 922, D= 29, S= 53, I= 23, N=1004]  76.85 [N=298]
cwhf:  73.95( 66.60)  [H= 724, D=133, S=122, I= 72, N= 979]  45.70 [N=291]
eahf:  76.18( 67.28)  [H= 745, D=126, S=107, I= 87, N= 978]  43.99 [N=291]
emhf:  76.65( 53.81)  [H= 755, D= 68, S=162, I=225, N= 985]  34.35 [N=294]
enhf:  81.84( 76.73)  [H= 802, D= 88, S= 90, I= 50, N= 980]  56.66 [N=293]
iahf:  91.41( 90.20)  [H= 681, D= 20, S= 44, I=  9, N= 745]  77.78 [N=216]
imhf:  75.80( 67.60)  [H= 758, D=122, S=120, I= 82, N=1000]  49.83 [N=297]
inhf:  95.07( 94.35)  [H= 656, D=  9, S= 25, I=  5, N= 690]  85.86 [N=198]
iwhf:  96.31( 95.13)  [H= 653, D=  4, S= 21, I=  8, N= 678]  86.67 [N=195]
----------------------------- Overall Results -----------------------------
SENT: %Correct=58.60 [H=1737, S=1227, N=2964]
WORD: %Corr=83.29, Acc=75.86 [H=8350, D=745, S=930, I=745, N=10025]
=========================================================================
```

```
============================ HTK Results Analysis ==============================
----------------------- Speaker Results Condition 2 ------------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-------------------------------------------------------------------------
cahf:  87.49( 68.31)  [H= 867, D= 41, S= 83, I=190, N= 991]  45.58 [N=294]
cmhf:  79.70( 56.78)  [H= 793, D= 68, S=134, I=228, N= 995]  30.30 [N=297]
cnhf:  91.63( 66.04)  [H= 920, D= 11, S= 73, I=257, N=1004]  41.95 [N=298]
cwhf:  75.79( 47.70)  [H= 742, D= 51, S=186, I=275, N= 979]  30.58 [N=291]
eahf:  74.34( 17.08)  [H= 727, D= 30, S=221, I=560, N= 978]   4.12 [N=291]
emhf:  72.08( 28.53)  [H= 710, D= 59, S=216, I=429, N= 985]  12.93 [N=294]
enhf:  77.24( 31.12)  [H= 757, D= 53, S=170, I=452, N= 980]   7.85 [N=293]
----------------------------- Overall Results -----------------------------
SENT: %Correct=24.83 [H=511, S=1547, N=2058]
WORD: %Corr=79.80, Acc=45.31 [H=5516, D=313, S=1083, I=2391, N=6912]
=========================================================================
```

```
=========================== HTK Results Analysis ===========================
----------------------- Speaker Results Condition 3 -----------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-------------------------------------------------------------------------
cahf:  53.18( 52.77)  [H= 527, D=344, S=120, I=  4, N= 991]  37.41 [N=294]
cmhf:  55.68( 55.08)  [H= 554, D=324, S=117, I=  6, N= 995]  37.37 [N=297]
cnhf:  64.94( 64.24)  [H= 652, D=229, S=123, I=  7, N=1004]  45.30 [N=298]
cwhf:  47.40( 46.88)  [H= 464, D=385, S=130, I=  5, N= 979]  30.58 [N=291]
eahf:  46.63( 45.71)  [H= 456, D=389, S=133, I=  9, N= 978]  27.84 [N=291]
emhf:  48.93( 48.63)  [H= 482, D=375, S=128, I=  3, N= 985]  29.59 [N=294]
enhf:  52.55( 51.63)  [H= 515, D=321, S=144, I=  9, N= 980]  34.13 [N=293]
---------------------------- Overall Results ----------------------------
SENT: %Correct=34.65 [H=713, S=1345, N=2058]
WORD: %Corr=52.81, Acc=52.18 [H=3650, D=2367, S=895, I=43, N=6912]
=========================================================================
```

```
=========================== HTK Results Analysis ===========================
----------------------- Speaker Results Condition 4 -----------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-------------------------------------------------------------------------
cahf:  42.79( 42.18)  [H= 424, D=402, S=165, I=  6, N= 991]  26.87 [N=294]
cmhf:  45.13( 44.52)  [H= 449, D=387, S=159, I=  6, N= 995]  28.28 [N=297]
cnhf:  55.28( 53.69)  [H= 555, D=272, S=177, I= 16, N=1004]  34.56 [N=298]
cwhf:  36.26( 36.06)  [H= 355, D=462, S=162, I=  2, N= 979]  22.34 [N=291]
eahf:  30.88( 30.57)  [H= 302, D=470, S=206, I=  3, N= 978]  16.49 [N=291]
emhf:  35.94( 35.53)  [H= 354, D=436, S=195, I=  4, N= 985]  18.03 [N=294]
enhf:  33.06( 30.92)  [H= 324, D=418, S=238, I= 21, N= 980]  16.38 [N=293]
---------------------------- Overall Results ----------------------------
SENT: %Correct=23.32 [H=480, S=1578, N=2058]
WORD: %Corr=39.97, Acc=39.13 [H=2763, D=2847, S=1302, I=58, N=6912]
=========================================================================
```

Tabla A 5 - Resultados obtenidos para HFCC-E con E=5 para cada una de las cuatro condiciones de análisis.

A.6 Resultados evaluación HFCC con E-factor=6

```
=========================== HTK Results Analysis ===========================
----------------------- Speaker Results Condition 1 -----------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-------------------------------------------------------------------------
cahf:  85.27( 83.45)  [H= 845, D= 82, S= 64, I= 18, N= 991]  71.77 [N=294]
cmhf:  79.10( 62.31)  [H= 787, D= 69, S=139, I=167, N= 995]  37.04 [N=297]
cnhf:  91.33( 88.55)  [H= 917, D= 24, S= 63, I= 28, N=1004]  74.83 [N=298]
cwhf:  73.85( 66.60)  [H= 723, D=146, S=110, I= 71, N= 979]  47.77 [N=291]
eahf:  74.74( 65.95)  [H= 731, D=141, S=106, I= 86, N= 978]  44.67 [N=291]
emhf:  77.66( 51.17)  [H= 765, D= 74, S=146, I=261, N= 985]  32.31 [N=294]
enhf:  82.35( 78.16)  [H= 807, D= 90, S= 83, I= 41, N= 980]  58.70 [N=293]
iahf:  92.21( 90.20)  [H= 687, D= 23, S= 35, I= 15, N= 745]  75.93 [N=216]
imhf:  76.40( 68.60)  [H= 764, D=118, S=118, I= 78, N=1000]  49.83 [N=297]
inhf:  96.09( 94.20)  [H= 663, D=  7, S= 20, I= 13, N= 690]  86.36 [N=198]
iwhf:  95.13( 93.95)  [H= 645, D=  8, S= 25, I=  8, N= 678]  85.64 [N=195]
---------------------------- Overall Results ----------------------------
SENT: %Correct=58.37 [H=1730, S=1234, N=2964]
WORD: %Corr=83.13, Acc=75.29 [H=8334, D=782, S=909, I=786, N=10025]
=========================================================================
```

```
============================ HTK Results Analysis =================================
----------------------- Speaker Results Condition 2 ------------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-----------------------------------------------------------------------------
cahf:  85.87( 74.97)  [H= 851, D= 62, S= 78, I=108, N= 991]  56.12 [N=294]
cmhf:  78.09( 53.97)  [H= 777, D= 70, S=148, I=240, N= 995]  27.95 [N=297]
cnhf:  90.24( 60.86)  [H= 906, D= 13, S= 85, I=295, N=1004]  36.58 [N=298]
cwhf:  76.51( 48.31)  [H= 749, D= 41, S=189, I=276, N= 979]  33.68 [N=291]
eahf:  73.62( 30.88)  [H= 720, D= 35, S=223, I=418, N= 978]  19.59 [N=291]
emhf:  71.78( 24.77)  [H= 707, D= 51, S=227, I=463, N= 985]  12.59 [N=294]
enhf:  75.41( 22.65)  [H= 739, D= 34, S=207, I=517, N= 980]   6.14 [N=293]
------------------------------ Overall Results -------------------------------
SENT: %Correct=27.55 [H=567, S=1491, N=2058]
WORD: %Corr=78.83, Acc=45.21 [H=5449, D=306, S=1157, I=2317, N=6912]
=============================================================================
```

```
============================ HTK Results Analysis =================================
----------------------- Speaker Results Condition 3 ------------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-----------------------------------------------------------------------------
cahf:  52.88( 52.57)  [H= 524, D=348, S=119, I=  3, N= 991]  38.10 [N=294]
cmhf:  54.77( 54.37)  [H= 545, D=325, S=125, I=  4, N= 995]  35.69 [N=297]
cnhf:  66.04( 65.54)  [H= 663, D=215, S=126, I=  5, N=1004]  46.64 [N=298]
cwhf:  47.29( 46.78)  [H= 463, D=386, S=130, I=  5, N= 979]  30.58 [N=291]
eahf:  45.19( 44.27)  [H= 442, D=390, S=146, I=  9, N= 978]  27.84 [N=291]
emhf:  49.04( 48.22)  [H= 483, D=363, S=139, I=  8, N= 985]  28.57 [N=294]
enhf:  50.71( 49.90)  [H= 497, D=320, S=163, I=  8, N= 980]  31.06 [N=293]
------------------------------ Overall Results -------------------------------
SENT: %Correct=34.11 [H=702, S=1356, N=2058]
WORD: %Corr=52.33, Acc=51.72 [H=3617, D=2347, S=948, I=42, N=6912]
=============================================================================
```

```
============================ HTK Results Analysis =================================
----------------------- Speaker Results Condition 4 ------------------------
spkr: %Corr( %Acc )  [ Hits, Dels, Subs, Ins, #Words] %S.Corr [ #Sent ]
-----------------------------------------------------------------------------
cahf:  41.07( 40.67)  [H= 407, D=409, S=175, I=  4, N= 991]  26.19 [N=294]
cmhf:  44.62( 43.92)  [H= 444, D=391, S=160, I=  7, N= 995]  28.96 [N=297]
cnhf:  52.99( 51.69)  [H= 532, D=280, S=192, I= 13, N=1004]  31.21 [N=298]
cwhf:  35.44( 34.73)  [H= 347, D=454, S=178, I=  7, N= 979]  20.96 [N=291]
eahf:  31.29( 30.47)  [H= 306, D=471, S=201, I=  8, N= 978]  16.49 [N=291]
emhf:  34.01( 33.50)  [H= 335, D=449, S=201, I=  5, N= 985]  17.35 [N=294]
enhf:  29.80( 28.16)  [H= 292, D=453, S=235, I= 16, N= 980]  16.04 [N=293]
------------------------------ Overall Results -------------------------------
SENT: %Correct=22.50 [H=463, S=1595, N=2058]
WORD: %Corr=38.53, Acc=37.66 [H=2663, D=2907, S=1342, I=60, N=6912]
=============================================================================
```

Tabla A 6 - Resultados obtenidos para HFCC-E con E=6 para cada una de las cuatro condiciones de análisis.

B

Presupuesto

1) Ejecución Material

- Compra de ordenador personal (Software incluido)..... 800€
- Material de oficina ..... 100€
- Total de ejecución material ..... 900€

2) Gastos generales

- Sobre Ejecución Material ..... 144€

3) Beneficio Industrial

- Sobre Ejecución Material ..... 54€

4) Honorarios Proyecto

- 1800 horas a 7€ / hora ..... 12600€

5) Material fungible

- Gastos de impresión ..... 200€
- Encuadernación ..... 200€

6) Subtotal del presupuesto

- Subtotal Presupuesto ..... 14998€

7) I.V.A. aplicable

- 18% Subtotal Presupuesto ..... 2699,64€

8) Total presupuesto

- Total Presupuesto ..... **17697,64€**

Madrid, Abril de 2011
El Ingeniero Jefe de Proyecto

Fdo.: Leticia Rueda Rojo
Ingeniero Superior de Telecomunicación

C

Pliego de condiciones

Pliego de condiciones

Este documento contiene las condiciones legales que guiarán la realización de este proyecto. En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

Supuesto que la utilización industrial de los métodos recogidos en el presente proyecto ha sido decidida por parte de la empresa cliente o de otras, la obra a realizar se regulará por las siguientes:

Condiciones generales

1. La modalidad de contratación será el concurso. La adjudicación se hará, por tanto, a la proposición más favorable sin atender exclusivamente al valor económico, dependiendo de las mayores garantías ofrecidas. La empresa que somete el proyecto a concurso se reserva el derecho a declararlo desierto.

2. El montaje y mecanización completa de los equipos que intervengan será realizado totalmente por la empresa licitadora.

3. En la oferta, se hará constar el precio total por el que se compromete a realizar la obra y el tanto por ciento de baja que supone este precio en relación con un importe límite si este se hubiera fijado.

4. La obra se realizará bajo la dirección técnica de un Ingeniero Superior de Telecomunicación, auxiliado por el número de Ingenieros Técnicos y Programadores que se estime preciso para el desarrollo de la misma.

5. Aparte del Ingeniero Director, el contratista tendrá derecho a contratar al resto del personal, pudiendo ceder esta prerrogativa a favor del Ingeniero Director, quien no estará obligado a aceptarla.

6. El contratista tiene derecho a sacar copias a su costa de los planos, pliego de condiciones y presupuestos. El Ingeniero autor del proyecto autorizará con su firma las copias solicitadas por el contratista después de confrontarlas.

7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirvió de base para la contratación, a las modificaciones autorizadas por la superioridad o a las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con arreglo a los cuales, se harán las modificaciones y la valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados. Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna clase, salvo en los casos de rescisión.

8. Tanto en las certificaciones de obras como en la liquidación final, se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.

9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.

10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata, se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos precios convenidos por uno u otro procedimiento, se sujetarán siempre al establecido en el punto anterior.

11. Cuando el contratista, con autorización del Ingeniero Director de obras, emplee materiales de calidad más elevada o de mayores dimensiones de lo estipulado en el proyecto, o sustituya una clase de fabricación por otra que tenga asignado mayor precio o ejecute con mayores dimensiones cualquier otra parte de las obras, o en general, introduzca en ellas cualquier modificación que sea beneficiosa a juicio del Ingeniero Director de obras, no tendrá derecho sin embargo, sino a lo que le correspondería si hubiera realizado la obra con estricta sujeción a lo proyectado y contratado.

12. Las cantidades calculadas para obras accesorias, aunque figuren por partida alzada en el presupuesto final (general), no serán abonadas sino a los precios de la contrata, según las condiciones de la misma y los proyectos particulares que para ellas se formen, o en su defecto, por lo que resulte de su medición final.

13. El contratista queda obligado a abonar al Ingeniero autor del proyecto y director de obras así como a los Ingenieros Técnicos, el importe de sus respectivos honorarios facultativos por formación del proyecto, dirección técnica y administración en su caso, con arreglo a las tarifas y honorarios vigentes.

14. Concluida la ejecución de la obra, será reconocida por el Ingeniero Director que a tal efecto designe la empresa.

15. La garantía definitiva será del 4% del presupuesto y la provisional del 2%.

16. La forma de pago será por certificaciones mensuales de la obra ejecutada, de acuerdo con los precios del presupuesto, deducida la baja si la hubiera.

17. La fecha de comienzo de las obras será a partir de los 15 días naturales del replanteo oficial de las mismas y la definitiva, al año de haber ejecutado la provisional, procediéndose si no existe reclamación alguna, a la reclamación de la fianza.

18. Si el contratista al efectuar el replanteo, observase algún error en el proyecto, deberá comunicarlo en el plazo de quince días al Ingeniero Director de obras, pues transcurrido ese plazo será responsable de la exactitud del proyecto.

19. El contratista está obligado a designar una persona responsable que se entenderá con el Ingeniero Director de obras, o con el delegado que éste designe, para todo relacionado con ella. Al ser el Ingeniero Director de obras el que interpreta el proyecto, el contratista deberá consultarle cualquier duda que surja en su realización.

20. Durante la realización de la obra, se girarán visitas de inspección por personal facultativo de la empresa cliente, para hacer las comprobaciones que se crean oportunas. Es obligación del contratista, la conservación de la obra ya ejecutada hasta la recepción de la misma, por lo que el deterioro parcial o total de ella, aunque sea por agentes atmosféricos u otras causas, deberá ser reparado o reconstruido por su cuenta.

21. El contratista, deberá realizar la obra en el plazo mencionado a partir de la fecha del contrato, incurriendo en multa, por retraso de la ejecución siempre que éste no sea debido a causas de fuerza mayor. A la terminación de la obra, se hará una recepción provisional previo reconocimiento y examen por la dirección técnica, el depositario de efectos, el interventor y el jefe de servicio o un representante, estampando su conformidad el contratista.

22. Hecha la recepción provisional, se certificará al contratista el resto de la obra, reservándose la administración el importe de los gastos de conservación de la misma hasta su recepción definitiva y la fianza durante el tiempo señalado como plazo de garantía. La recepción definitiva se hará en las mismas condiciones que la provisional, extendiéndose el acta correspondiente. El Director Técnico propondrá a la Junta Económica la devolución de la fianza al contratista de acuerdo con las condiciones económicas legales establecidas.

23. Las tarifas para la determinación de honorarios, reguladas por orden de la Presidencia del Gobierno el 19 de Octubre de 1961, se aplicarán sobre el denominado en la actualidad "Presupuesto de Ejecución de Contrata" y anteriormente llamado "Presupuesto de Ejecución Material" que hoy designa otro concepto.

Condiciones particulares

La empresa consultora, que ha desarrollado el presente proyecto, lo entregará a la empresa cliente bajo las condiciones generales ya formuladas, debiendo añadirse las siguientes condiciones particulares:

1. La propiedad intelectual de los procesos descritos y analizados en el presente trabajo, pertenece por entero a la empresa consultora representada por el Ingeniero Director del Proyecto.

2. La empresa consultora se reserva el derecho a la utilización total o parcial de los resultados de la investigación realizada para desarrollar el siguiente proyecto, bien para su publicación o bien para su uso en trabajos o proyectos posteriores, para la misma empresa cliente o para otra.

3. Cualquier tipo de reproducción aparte de las reseñadas en las condiciones generales, bien sea para uso particular de la empresa cliente, o para cualquier otra aplicación, contará con autorización expresa y por escrito del Ingeniero Director del Proyecto, que actuará en representación de la empresa consultora.

4. En la autorización se ha de hacer constar la aplicación a que se destinan sus reproducciones así como su cantidad.

5. En todas las reproducciones se indicará su procedencia, explicitando el nombre del proyecto, nombre del Ingeniero Director y de la empresa consultora.

6. Si el proyecto pasa la etapa de desarrollo, cualquier modificación que se realice sobre él, deberá ser notificada al Ingeniero Director del Proyecto y a criterio de éste, la empresa consultora decidirá aceptar o no la modificación propuesta.

7. Si la modificación se acepta, la empresa consultora se hará responsable al mismo nivel que el proyecto inicial del que resulta el añadirla.

8. Si la modificación no es aceptada, por el contrario, la empresa consultora declinará toda responsabilidad que se derive de la aplicación o influencia de la misma.

9. Si la empresa cliente decide desarrollar industrialmente uno o varios productos en los que resulte parcial o totalmente aplicable el estudio de este proyecto, deberá comunicarlo a la empresa consultora.

10. La empresa consultora no se responsabiliza de los efectos laterales que se puedan producir en el momento en que se utilice la herramienta objeto del presente proyecto para la realización de otras aplicaciones.

11. La empresa consultora tendrá prioridad respecto a otras en la elaboración de los proyectos auxiliares que fuese necesario desarrollar para dicha aplicación industrial, siempre que no haga explícita renuncia a este hecho. En este caso, deberá autorizar expresamente los proyectos presentados por otros.

12. El Ingeniero Director del presente proyecto, será el responsable de la dirección de la aplicación industrial siempre que la empresa consultora lo estime oportuno. En caso contrario, la persona designada deberá contar con la autorización del mismo, quien delegará en él las responsabilidades que ostente.