Informe técnico SIC-003

Introducción al Reconocimiento de la Voz

Pedro L. Galindo Riaño

C.A.S.E.M. - Universidad de Cádiz
Pol. Rio San Pedro s/n
11510 Puerto Real (CÁDIZ)
SPAIN
Tfno : 34 – 956 – 01.64.34
Fax : 34 – 956 – 01.64.37

Introducción al Reconocimiento de la Voz

La meta principal del Reconocimiento de la Voz es desarrollar técnicas y sistemas capaces de aceptar como entrada señal hablada. La gran difusión alcanzada por los pequeños ordenadores hace que dentro de poco, cada hogar, cada centro de trabajo, cada establecimiento, etc. posea, al menos un computador. Ello hace que los sistemas de comunicación hombre-máquina sean de vital importancia. Sencillos sistemas de reconocimiento de voz permitirían acceder a información de Bases de Datos via telefono utilizando vocabularios reducidos e independientes del locutor. El estado actual de la tecnología hace que estos sistemas sean ya una realidad.

Para realizar el reconocimiento de la voz existen varios problemas básicos aún no resueltos, que podemos resumir en los siguientes :

- No existe una separación definida y constante entre las diferentes palabras o entre los diferentes sonidos cuando una persona habla.
- Cada sonido elemental (fonema) es modificado por su contexto.
- Existen una gran cantidad de parámetros variables, ya sean intra-locutor, inter-locutor, debido a la forma de hablar, al estado de ánimo, a la salud del hablante, al dispositivo de captura de la señal sonora, a las señales procedentes del entorno, etc.
- Es preciso procesar una cantidad ingente de información para contemplar todas las variantes posibles.
- Una única señal sonora lleva una gran cantidad de información, aparte de la puramente sintáctica, tales como el sexo y la identidad de la persona que habla, su humor, su acento, etc.
- No hay reglas precisas que permitan formalizar los diferentes niveles de complejidad del lenguaje(sintaxis, semántica, pragmática, etc.)

Por todas estas razones, la complejidad de realizar sistemas automáticos que reconozcan la voz hablada con una fiabilidad alta es enorme. Por supuesto, que dentro de la tecnología del Reconocimiento de la Voz, podemos distinguir diferentes dimensiones de complejidad :

- Palabras aisladas, palabras conectadas o voz contínua.
- Tamaño del vocabulario.
- Limitaciones impuestas al lenguaje y a la tarea a utilizar.
- Dependencia o independencia del locutor.
- Ambigüedad acústica.
- Ruido ambiente.

En los últimos años se han realizado una multitud de sistemas de reconocimiento. Sin embargo, la mayoría de ellos estaban limitados, bien por ser dependientes del locutor, bien por realizar el reconocimiento sobre palabras aisladas, por utilizar un vocabulario muy reducido, o una gramática muy limitada.

El presente curso pretende mostrar los principios básicos del Reconocimiento de la Voz, partiendo de los conocimientos más elementales para llegar a ser capaces de construir sencillos reconocedores de voz basados en varias técnicas, con una tasa de reconocimiento bastante aceptable.

Pedro L. Galindo

INDICE

1 Introducción

1.1 Reconocimiento Automático de la Voz

La utilización de la voz como medio habitual de comunicación constituye una de las características diferenciales más importantes de los seres humanos. A través de la voz, una persona es capaz de comunicar una inmensa cantidad de información a otra, o a muchas otras. Es capaz de transmitir, no solo información lingüística (palabras, frases,...) sino su estado de ánimo (si está contento, disgustado, nervioso, somnoliento, etc.), su salud (recordemos cuando hablamos con alguien resfriado por teléfono), su edad aproximada (si es un niño, un anciano, etc.), tal vez su nacionalidad (o al menos, si es extranjero) e incluso si es hombre o mujer.

El Procesamiento Digital de la Voz comprende múltiples áreas, entre las que podemos destacar las siguientes :

- **Síntesis de Voz :** la síntesis de la voz consiste en producir voz a partir de una cadena de caracteres escritos y de una forma independiente, es decir, sin precisar de la voz "pregrabada" de una persona. De esta forma es posible, por ejemplo, que una máquina se comunique con el usuario "hablando", que sea capaz de leer en voz alta un libro, o de pronunciar un discurso escrito en un papel, etc.
- **Reconocimiento del Locutor :** su misión es la identificación de la persona que habla. Ya en 1944, la identificación del locutor fue utilizada por las fuerzas aliadas para seguir los movimientos de las tropas alemanas analizando espectrogramas de voz de las señales de radio del tráfico enemigo.
- **Codificación :** consiste en todos aquellos procesos realizados a la señal de voz para almacenarla o transmitirla "empaquetando" la información al máximo, de tal forma que la información importante ocupe el mínimo espacio posible. Obviamente, tanto para el almacenamiento, como para la transmisión de la voz, se precisa que la señal "codificada" conserve al máximo sus propiedades cuando se "descodifica".
- **Reconocimiento de la Voz :** es el proceso por el cual, un ordenador transforma una señal acústica en texto. Es, con mucho, de todas las tareas citadas, la más compleja, la que más expectación despierta y la que más aplicaciones posee.
- **Comprensión Automática de la Voz** : es el proceso por el cual el ordenador transforma una señal acústica en alguna forma abstracta de representación del conocimiento implícito en la citada señal.

Dentro de las grandes áreas del Procesamiento Digital de la Señal, nos centraremos en el Reconocimiento Automático de la Voz Humana. La meta principal del Reconocimiento de la Voz es desarrollar técnicas y sistemas capaces de aceptar como entrada señal hablada. La gran difusión alcanzada por los pequeños ordenadores hace que dentro de poco, cada hogar, cada centro de trabajo, cada establecimiento, etc. posea, al menos un computador. Ello hace que los sistemas de comunicación hombre-máquina sean de vital importancia. Sencillos sistemas de reconocimiento de voz permitirían acceder a información de Bases de Datos via telefono utilizando vocabularios reducidos e independientes del locutor. El estado actual de la tecnología hace que estos sistemas sean ya una realidad.

Las ventajas resultantes de la utilización de sistemas de reconocimiento de voz se pueden resumir en cuatro factores principales :

- La señal de voz no requiere, para ser producida, ninguna especialización. (normalmente, todas las personas saben hablar)

- La voz transmite una información que es de 8 a 10 veces más rápida que la más veloz de las secretarias.
- La voz puede ser generada, pese a que la persona que la genere tenga ocupadas sus manos, piernas, ojos, etc., esté en movimiento o parada.
- Dado que la señal de voz se transmite con facilidad por medio de líneas teléfonicas, la información puede ser suministrada a larga distancia.

Para realizar el reconocimiento de la voz existen varios problemas básicos aún no resueltos, que podemos resumir en los siguientes :

- No existe una separación definida y constante entre las diferentes palabras o entre los diferentes sonidos cuando una persona habla.
- Cada sonido elemental (fonema) es modificado por su contexto.
- Existen una gran cantidad de parámetros variables, ya sean intra-locutor, inter-locutor, debido a la forma de hablar, al estado de ánimo, a la salud del hablante, al dispositivo de captura de la señal sonora, a las señales procedentes del entorno, etc.
- Es preciso procesar una cantidad ingente de información para contemplar todas las variantes posibles.
- Una única señal sonora lleva una gran cantidad de información, aparte de la puramente sintáctica, tales como el sexo y la identidad de la persona que habla, su humor, su acento, etc.
- No hay reglas precisas que permitan formalizar los diferentes niveles de complejidad del lenguaje(sintaxis, semántica, pragmática, etc.)

1.2 Clasificación de los Sistemas de Reconocimiento

Existen multitud de factores por los que podemos dividir los Sistemas de Reconocimiento de Voz en múltiples categorías, que definen diferentes dimensiones de complejidad :

- Dependencia/independencia del locutor.
- Palabras aisladas, palabras conectadas o voz contínua.
- Tamaño del vocabulario.
- Limitaciones impuestas al lenguaje y a la tarea a utilizar.
- Ambigüedad acústica.
- Ruido ambiente.

En función de la dependencia del locutor, se dice que un sistema es **monolocutor** cuando se diseña para que funcione con un único locutor. Estos sistemas son más fáciles de desarrollar, más baratos y tienen una mayor tasa de reconocimiento. Se dice que un sistema es **multilocutor** si funciona correctamente para cualquier locutor de un reducido grupo de hablantes (por ejemplo, 10 personas). Por último se dice que un sistema es **independiente del locutor** si funciona correctamente para cualquier hablante. Estos sistemas son los más difílices de desarrollar, los más caros, y aquellos donde la tasa de reconocimiento es menor. Los **sistemas adaptativos** adaptan su funcionamiento a las características de nuevos locutores, o sea, "aprenden" la voz del nuevo hablante, por lo que constituyen un paso intermedio, tanto en coste, rendimiento y complejidad entre los sistemas monolocutor y aquellos independientes del locutor.

En función del tipo de sentencia reconocida, se habla de sistemas de **palabras aisladas** cuando el reconocimiento se hace sobre palabras sueltas que el locutor pronuncia una a una, dejando una pausa en medio. Esta es la forma más simple de reconocimiento, ya que el comienzo y final de cada palabra es relativamente fácil de detectar. Sin embargo, su utilidad es limitada, ya que su utilización es bastante incómoda para el hablante. Se

habla de un sistema de **voz contínua** cuando el sistema es capaz de reconocer palabras unidas, no separadas por silencios entre sí. El reconocimiento de la voz contínua es bastante difícil, debido a la aparición de numerosos problemas, entre los que podemos citar la dificultad de localizar o determinar el punto en que una palabra termina y comienza la siguiente, el fenómeno de la coarticulación, donde el sonido de un fonema se ve afectado por aquellos que le preceden y que le siguen, la velocidad del hablante, que es mucho mayor que en el caso de las palabras aisladas, etc. Normalmente, los sistemas de voz contínua permiten el reconocimiento de frases pertenecientes a una gramática restringida en mayor o menor medida, y el factor que tiene mayor importancia es la comprensión de la frase completa, y no de las palabras sueltas. A caballo entre los sistemas de palabras aisladas y los de voz contínua. se encuentran los sistemas de reconocimiento de **palabras conectadas** cuando el locutor puede pronunciar una serie de palabras seguidas de entre un conjunto limitado perteneciente a un vocabulario muy limitado. En estos sistemas se da un énfasis especial a la comprensión de las palabras de forma independiente. Un ejemplo típico de estos sistemas son los sistemas de reconocimiento de números de teléfono.

En función del tamaño del vocabulario utilizado, se habla de **vocabularios pequeños** (decenas de palabras), **medianos** (cientos de palabras), **grandes** (miles de palabras) y **muy grandes** (decenas de miles de palabras). En los últimos años se han realizado una multitud de sistemas de reconocimiento. Sin embargo, la mayoría de ellos estaban limitados, bien por ser dependientes del locutor, bien por realizar el reconocimiento sobre palabras aisladas, por utilizar un vocabulario muy reducido, o una gramática muy limitada. El presente curso pretende mostrar los principios básicos del Reconocimiento de la Voz, partiendo de los conocimientos más elementales para llegar a ser capaces de construir sencillos reconocedores de voz basados en varias técnicas, con una tasa de reconocimiento bastante aceptable.

1.3 Breve revisión histórica

Dada la importancia del proceso de comunicación oral, desde finales del siglo XIII se comenzó a investigar de una forma científica. Sin embargo, hasta mediados del siglo XX, lo único que se había podido construir habían sido aparatos capaces de sintetizar sonidos sencillos. En 1938, se publican los primeros artículos sobre PCM (pulse code modulation). Ello unido a la aparición posterior de los circuítos digitales y los ordenadores electrónicos hacia los años 50, hizo que se diseñaran los primeros sistemas capaces, no solo de reproducir sonidos, sino de reconocer voz. En concreto, en 1952, en los laboratorios Bell se construye el primer sistema de reconocimiento, capaz de reconocer dígitos de forma aislada de un único locutor. El sistema estaba basado en la medición de las resonancias espectrales durante la sección vocálica de cada dígito. Durante los años 60, sistemas parecidos se realizaron en Japón, Europa y Estados Unidos.

Sin embargo, la construcción de sistemas más complejos no se produjo hasta los años 70, con la aparición de sistemas basados principalmente en reconocimiento de patrones. De gran relevancia es la aplicación de las teorías de predicción lineal al procesamiento de la voz en todas sus áreas (reconocimiento, síntesis, codificación,etc.). Se realizaron grandes avances con el desarrollo del proyecto ARPA-SUR. La idea principal era la utilización de información de nivel "superior", tal como sintaxis, semántica, prosodia, etc. en el proceso de reconocimiento. Los resultados fueron brillantes, y se realizaron diferentes sistemas, entre los que destacaremos SDC, DRAGON, HEARSAY-I y HEARSAY-II, HWIM y HARPY. Durante esta década se afianzaron los sistemas

capaces de reconocer palabras aisladas, basados en los estudios de Sakoe y Chiba, en lo que se denominó DTW (Dynamic Time Warping) en Japón, y de Itakura en Estados Unidos. Se comenzaron los estudios en reconocimiento independiente del locutor, y se vislumbró la posibilidad de realizar, incluso, sistemas de voz contínua independientes del locutor.

En los años 80, se produce la irrupción de los sistemas estadísticos, y se realiza una evolución de los sistemas basados en patrones a los sistemas basados en la modelización estadística, y más en concreto, en los Modelos Ocultos de Markov. En esta década se da un gran avance en la utilización de vocabularios mayores (varios miles de palabras), y reconocimiento de palabras conectadas. Asímismo, se introduce la utilización de gramáticas que limiten, definan y controlen el proceso de reconocimiento. Se experimentan nuevas técnicas basadas en tecnología de redes neurnales capaces de superar las deficiencias presentadas por los sistemas basados en HMM. Sin embargo, en este caso, la evolución no es tan evidente, ya que las mejoras obtenidas con sistemas realizados con redes neuronales exclusivamente no son notables, e incluso, en la mayoría de los casos, la eficacia del sistema es ampliamente superada por sistemas similares de tecnología HMM.

A lo largo de la primera mitad de los años 90, los sistemas tienden a realizarse utilizando técnicas mixtas. En concreto, la mayoría de los sistemas actuales tienden a realizar la integración de modelos estadísticos y neuronales. El interés científico se enfoca primordialmente hacia los vocabularios extremadamente grandes (varias decenas de miles de palabras), y reconocimiento de voz contínua, perteneciente a una gramática compleja.

1.4 Sistemas Comerciales

Podemos clasificar los sistemas comerciales en dos grandes categorías, intérpretes de comandos, y sistemas de dictado automático. Los intérpretes de comandos se limitan a realizar funciones que normalmente están controladas por el teclado o el ratón, sustituyendo, o complementando dichos periféricos por la introducción de comandos via micrófono. Los sistemas de dictado automático, por otro lado, transcriben la voz hablada a un procesador de textos. En la tabla 1.1 se resumen algunos de los sistemas comerciales disponibles en la actualidad para entorno WINDOWS :

Producto / Compañía	***Precio***	***Descripción***
NaturallySpeaking 1.0 Dragon Systems Inc.	$195 - $595	Software de Voz Contínua (5000 - 30000 - 60000 palabras)
Viavoice I.B.M.	$150	Software de Voz Contínua (5000 - 30000 - 60000 palabras)
Kurzweil Voice 1.0 Kurzweil Applied Intelligence	$995	Software de Dictado Automático. (60000 palabras)
Listen 2.0 Verbex Voice Systems	$99 - $139	Intérprete de Comandos por Voz. Independiente del locutor.
Say It! TimeWorks International	$229	Intérprete de Comandos por Voz. Dependiente del locutor.
Voice Companion for WordPerfect. Kolvox Communications.	$99	Intérprete de Comandos por Voz. Independiente del locutor.
Voice Mouse Interactive Products, Inc.	$79.95	Intérprete de Comandos por Voz. Dependiente del locutor.

Tabla 1.1 *Diferentes Sistemas de Reconocimiento Comerciales*

Los sistemas actuales, sin embargo, no pasan de ser medianamente útiles, salvo para personas discapacitadas. Ello se debe a que todos los sistemas comercializados ofrecen reconocimiento de palabras aisladas, lo cual obliga al locutor a realizar pausas de duración indefinida entre palabras consecutivas. La segunda gran mejora que se acometerá en el futuro será la extensión de todos los sistemas a la independencia del locutor, que permita a cualquiera utilizar el sistema, sin la necesidad de un entrenamiento previo.

1.5 Ultimos avances

Los últimos avances se corresponden con la realización de sistemas que admitan voz continua con la utilización de vocabularios grandes, la integración de gramáticas complejas, la robustez al ruido, independencia del locutor, etc.
Pese a que los últimos resultados obtenidos en el área del procesamiento de la señal son prometedores, parece ser que el mejor método para mejorar la robustez y eficacia de los sistemas de reconocimiento actuales consiste en llegar a una mejor comprensión de las diferentes fuentes de variabilidad de la voz humana. Esta variabilidad puede provenir de tres fuentes principales, la persona que habla, el entorno físico en el que se desarrolla la comunicación, y por último, el canal de transmisión de la información. Para la producción de sistemas que manejen el lenguaje hablado de una forma robusta y versátil, se han identificado varia áreas de investigación fundamentales, entre las que destacamos las siguientes :

- Modelización de la coarticulación y contexto fonético.
- Modelización de las características temporales de la voz.
- Modelización de las diferencias entre diferentes locutores.
- Estudio de la influencia del entorno acústico.
- Modelización de la percepción humana de la voz.

La realización de investigaciones y estudios encaminados al desarrollo y evolución de cualquiera de las citadas áreas será de gran utilidad para desarrollar y avanzar el proceso de expansión de los sistemas de reconocimiento automático. En cuanto al futuro, las previsiones son optimistas. La evolución del área del reconocimiento de la voz va a ser, de forma resumida, tal como se indica en la tabla 1.2

Período	***Capacidad de Reconocimiento***	***Vocabulario (Nº Palabras)***	***Aplicaciones***
1990-1995	Palabras Concatenadas Gramátic. de Estados Finitos Tareas Restringidas	10-1000	Marcado Automático Pedidos por Catálogo Tarjeta de Crédito Electr.
1995-2000	Voz Continua Modelos de Lenguaje Semánticas específicas	5000-20000	Dictado Automático Acceso Bases de Datos Asistente a la Secretaria
2000-2020	Voz Continua Gramát. Lenguaje Natural Adaptación y Aprendizaje	Ilimitado	Interacción natural Traducción Automática Integración Síntesis-Reconoc.

Tabla 1.2 *Evolución de la Tecnología del Reconocimiento de Voz*

En definitiva, el campo de investigación en el Reconocimiento Automático de la Voz no ha hecho más que empezar, y los resultados obtenidos hasta ahora permiten conservar la esperanza de que la comunicación automática hombre-máquina sea un hecho en un futuro cercano.

2 Análisis Temporal

2.1 El Aparato Fonador Humano

La figura 2.1 muestra la imagen en rayos X de una sección longitudinal del aparato vocal humano. El *tracto vocal* aparece con líneas de puntos, y es el conducto que consta de la faringe (conexión del esófago a la boca) y la cavidad oral (boca). El *tracto nasal* comienza en el *velo del paladar* (campanilla), y termina en las *fosas nasales*.

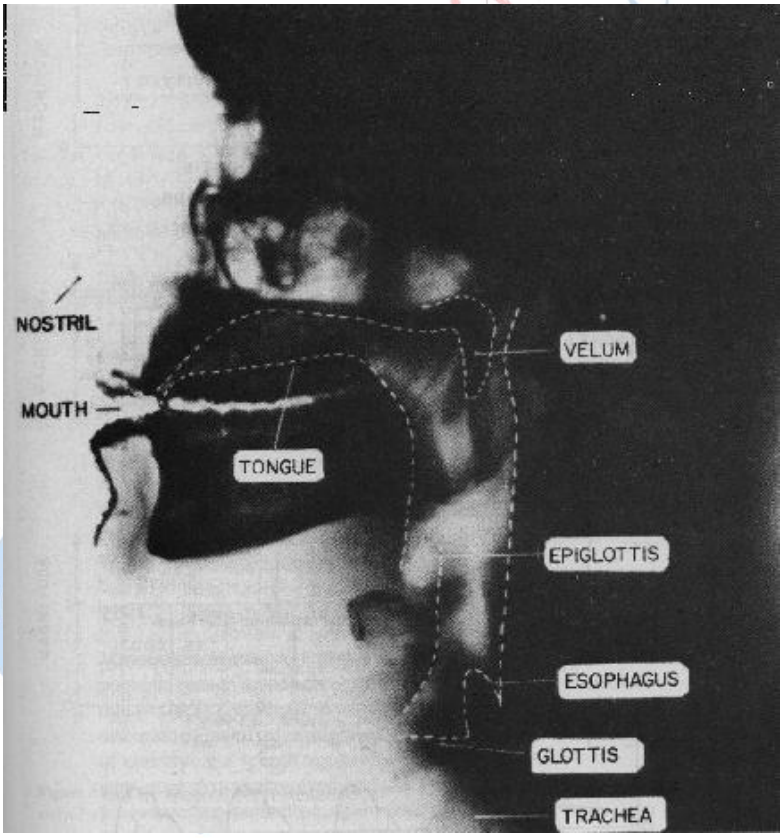


Figura 2.1. Imagen de un corte longitudinal del aparato vocal humano.

Una representación simplificada del aparato fonador humano se puede observar en la figura 2.2.

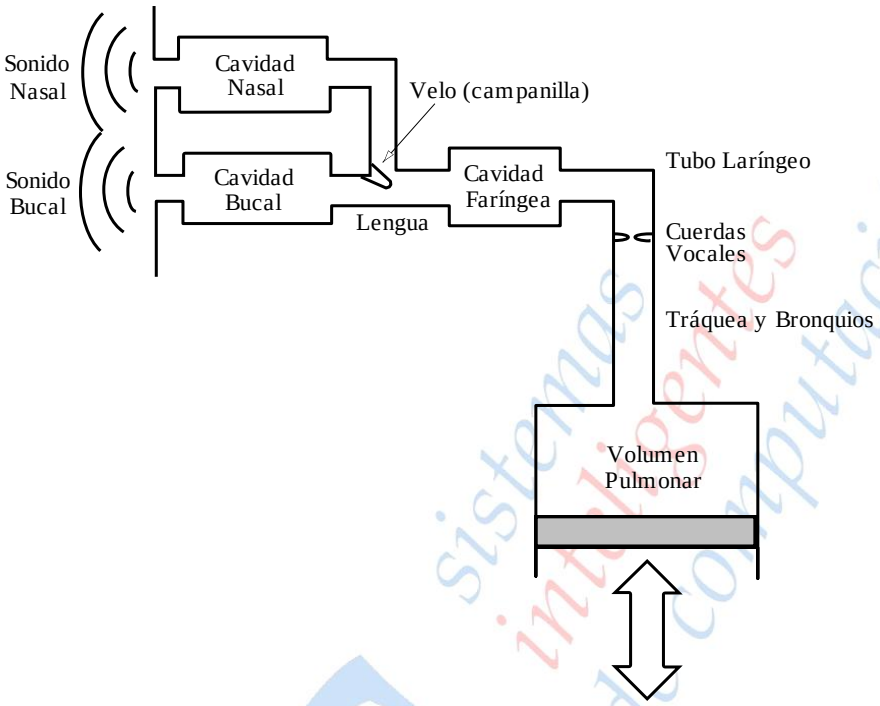


Figura 2.2. Representación del mecanismo fisiológico ***de producción de voz***

El proceso de generación de voz comienza cuando el aire entra en los pulmones mediante el mecanismo normal de la respiración. El aire sale de los pulmones a través de la tráquea, provocando que las cuerdas vocales vibren en la laringe. Las ondas provocadas por dicha vibración son moduladas al paso por la faringe, la cavidad bucal, y si el velo lo permite, por la cavidad nasal. En función de la posición de la mandíbula, lengua, velo, labios y boca, y de si las cuerdas vocales vibran o no, se producen los diferentes sonidos.

El velo actúa como una válvula para el aire que fluye a través de la laringe, e impide o permite que dicho aire pase al tracto nasal. Cuando el velo está en posición anterior, todo el flujo de aire pasa a la cavidad bucal. En caso contrario, el flujo se reparte en ambas cavidades.

Las cuerdas vocales son las que realmente producen el sonido, mediante vibración, de la misma forma que suenan las cuerdas de una guitarra. Cuando las cuerdas vibran, el sonido que se produce se denomina *sonoro*. Es el caso de las vocales y las consonantes sonoras. Cuando las cuerdas vocales no vibran, se producen un sonido *sordo.* Es el caso de las consonantes sordas. Para saber si un sonido es sonoro o sordo, basta tocar con la palma de la mano la garganta, y comprobar la vibración o su ausencia.

La voz se produce como una secuencia de sonidos. Por tanto, el estado de las cuerdas vocales, así como las posiciones, formas y tamaños de los diferentes elementos cambian a lo largo del tiempo, modificando el sonido que finalmente se produce.

2.2 El Aparato Auditivo Humano

El oído normalmente se divide en tres zonas bien diferenciadas, oído externo, medio e interno. El oído externo se encarga de la recogida del sonido, y de su transporte hasta el tímpano. El oído medio se componen de un conjunto de complicados huesecillos que transmiten la variación de presión que se produce en el tímpano, hasta la *cóclea* (caracol), que constituye el elemento fundamental del oído interno.

La cóclea recibe los impulsos procedentes de la cadena de huesecillos, y realiza un proceso de análisis frecuencial que analizaremos en un capítulo posterior.

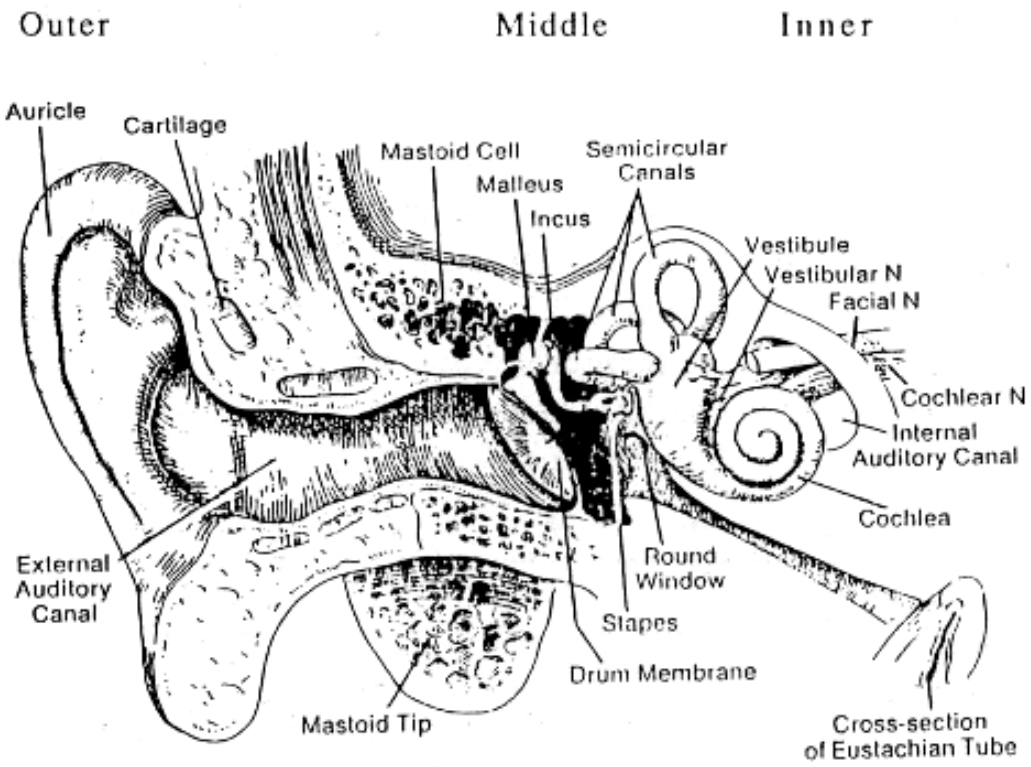


Figura 2.3. Diferentes partes del aparato auditivo humano.

En líneas generales, podemos decir que la cóclea es sensible a la periodicidad de los impulsos citados. Una vez la cóclea ha realizado su misión, se envía la información resultante al nervio auditivo, que transporta dicha información al cerebro para su procesado posterior.

2.3 Concepto de señal

La señal de voz puede ser convertida en un objeto manipulable convirtiéndolo en una señal eléctrica utilizando un micrófono. Se realiza una conversión de la señal analógica en una señal digital, fácilmente tratable a través de un ordenador.
El proceso de conversión analógico-digital consta de :

1.- Muestreo :
2.- Cuantificación
3.- Codificación

El **muestreo** consiste en convertir una señal continua en el tiempo en una señal discreta en el tiempo, midiendo el valor de dicha señal a intervalos regulares de tiempo.

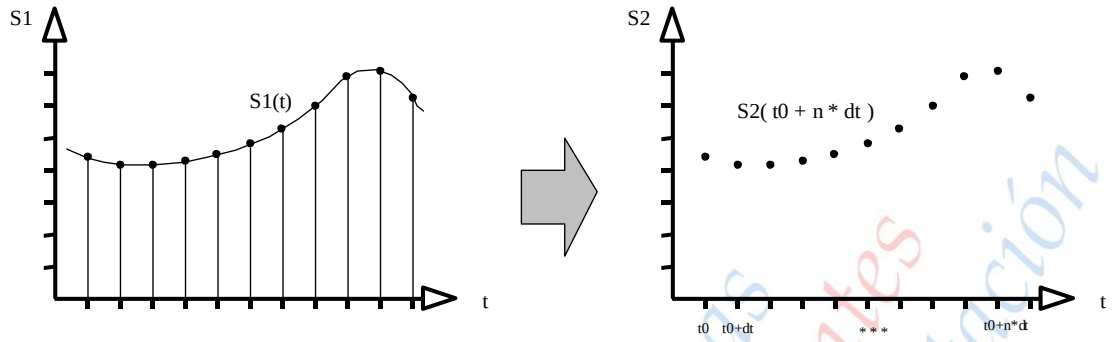


Figura 2.4. Proceso de Muestreo.

Como es bien sabido la frecuencia de muestreo de una señal es un factor determinante en el proceso de digitalización de una señal. Se sabe que para que la recomposición de la señal continua original sea posible a partir de su señal muestreada, es decir, para que no exista pérdida de información, la frecuencia de muestreo, ha de ser, al menos, mayor que el doble de la máxima frecuencia que aparece en la señal original. Este teorema se conoce como el **teorema de Shannon**.

$$F_{muestreo} \geq 2 * F_{maxima}$$

La **cuantificación** consiste en convertir una señal de amplitud continua en un conjunto de amplitudes discretas, que difiere de la señal de amplitud continua en lo que se llama **ruido de cuantificación**. Existen dos tipos básicos de cuantificación, escalar y vectorial. En este apartado nos referiremos exclusivamente a la cuantificación escalar.
Entendemos por **cuantificación escalar** al proceso de cuantificar los elementos de un conjunto de valores de forma separada. Es decir, el valor discreto asignado a un elemento es independiente del resto de los elementos del conjunto de datos a cuantificar. Un ejemplo de cuantificación escalar puede consistir en realizar un redondeo del valor de una magnitud continua (Por ejemplo, temperatura), para convertirlo en un valor entero (Por ejemplo, 27 Grados).

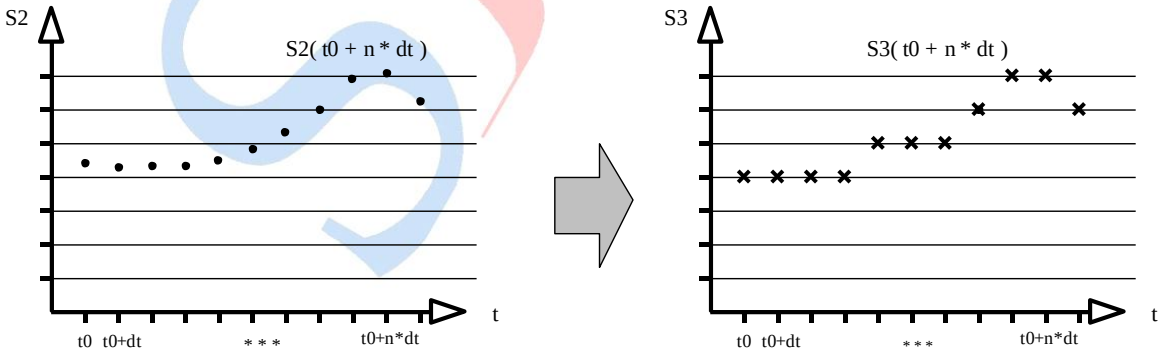


Figura 2.5. Proceso de Cuantificación Escalar

La **codificación** consiste en asignar una secuencia de bits a cada uno de los niveles en los que se ha cuantificado la señal. Si disponemos de un número de bits igual a b, el número de niveles que es posible cuantificar es igual a $N=2^b$ niveles. Por tanto, si se quiere transmitir una señal de voz por un canal, precisaremos una *velocidad de transmisión* igual a $F \cdot b$, donde F es la frecuencia de muestreo, y b es el número de bits

por muestra. Por supuesto, este valor es posible reducirlo utilizando otros esquemas de codificación, por ejemplo utilizando códigos de longitud variable.

2.4 Concepto de Trama

Una secuencia de muestras representando a una señal de voz típica se muestra en la figura 2.6. Es evidente que las propiedades de la señal cambian con el tiempo. Para poder realizar el procesamiento de la voz, se realiza la suposición de que las propiedades de esta señal cambian relativamente despacio en el tiempo. Esta suposición lleva a multitud de métodos de procesamiento en la que breves segmentos de señal se aislan y procesan de forma independiente, como si fueran fragmentos de un sonido continuo. A estos segmentos de longitud fija en los que se suele dividir la señal los denominamos *tramas.*

Los resultados de procesar cada trama generan un número (por ejemplo, podría ser el índice del fonema detectado) o bien varios números (por ejemplo, si el sonido es sonoro/sordo, la amplitud máxima de la señal, características de la misma, etc.). A este proceso se le denomina *Extracción de Características.*

Este proceso de aislar una *trama* y calcular una serie de características se realiza cada cierto tiempo, denominado *Longitud del Desplazamiento (o Solapamiento).*

Se han realizado numerosos sistemas en los que la longitud de la trama y el desplazamiento son muy diferentes. Suponiendo que N=Longitud de Trama y que M=Longitud del desplazamiento, pueden darse los siguientes casos :

- N<M : en este caso no hay solapamiento entre tramas sucesivas, perdiéndose parte de la señal. Por ello, no se suele utilizar.
- N=M : si bien no hay pérdida de señal, la inexistencia de correlación en los valores espectrales obtenidos de tramas consecutivas suele ser desaconsejable.
- N>M : es el caso más habitual. Las tramas adyacentes solapan, por lo que el análisis espectral posterior tendrá una cierta correlación entre tramas consecutivas. De hecho, si N>>M, las variaciones entre los valores obtenidos de sucesivos análisis espectrales son muy suaves.

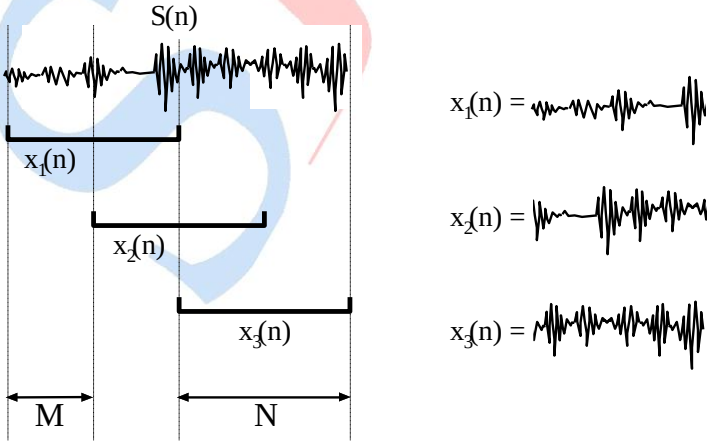


Figura 2.6. División de una señal en tramas

Una vez establecida la relación entre M y N, será preciso determinar los correspondientes valores absolutos. La selección del valor de N resulta de un compromiso que es preciso establecer entre resolución en tiempo y en frecuencia. Normalmente, los espectrogramas de banda ancha usan ventanas de una duración de

unos 3 msec., lo que permite una buena resolución temporal, útil para analizar cambios rápidos de la señal. Por otro lado, los espectrogramas de banda estrecha usan ventanas del orden de 30 msec., muy útiles para determinar el período de la señal, ya la ventana suele comprender varios períodos. Un valor óptimo de N sería el correspondiente al período de la señal. Debido a que dicho período no solo es variable de un locutor a otro, sino que es variable en el tiempo para un cierto locutor, no es posible utilizar tal medida como valor para la longitud de la trama. Un valor de N menor que el período de la señal produciría un análisis distorsionado, ya que la porción de señal analizada está truncada de forma artificial. Por ello, se suelen utilizar valores para N que comprendan varios períodos de señal (los menos posibles). Dado que los posibles valores del período de la señal varían aproximadamente entre los 40Hz. (tono grave de un hombre) y los 100 Hz.(tono agudo de mujer), un compromiso para la selección de N es :

10 msec < Longitud de Trama < 25 msec

En cuanto al valor del desplazamiento, ya hemos visto que debe ser menor que la longitud de la trama, y normalmente se selecciona como una fracción entera de la misma.

Longitud del Desplazamiento = K * Longitud de Trama

donde los valores de K suelen ser K=1/2, 1/3, 1/4, 1/5, ...
Cuanto menor sea el desplazamiento, más suaves serán las fluctuaciones sufridas por los parámetros obtenidos en el proceso de Extracción de Características.

2.5 Energía

Uno de los parámetros más sencillos de calcular es la energía.
Podemos definir la energía de una señal como :

$$E = \sum_{m=-\infty}^{\infty} x^2(m)$$

Tal cantidad da poca información sobre la evolución de dicho parámetro durante el tiempo. Por tanto, la magnitud que se suele utilizar es la denominada *Energía de Trama*, (que a partir de ahora llamaremos simplemente energía), y que se calcula como :

$$E = \sum_{m=1}^{N} x^2(m)$$

Es decir, la energía en la muestra n-sima es simplemente la suma de los cuadrados de las N muestras siguientes. Como vemos, N es el número de muestras consideradas para calcular la energía. Si N es muy pequeño(del orden de 1 período o menos), E_n fluctuará muy rápidamente, dependiendo en cada uno de los detalles de la señal. Si N es muy grande(varios períodos), E_n cambiará muy lentamente y no reflejará las propiedades cambiantes de la señal. Normalmente se suele elegir un valor de N comprendido en el rango :

10 msec. < N < 20 msec.

La energía nos permitirá distinguir con cierta fiabilidad la existencia de voz del silencio, y los sonidos sordos de los sonoros.
En el cálculo de la energía, y dado que se calcula utilizando una suma de cuadrados, existen grandes diferencias entre la energía de señales de gran amplitud, y la energía de señales de baja amplitud. Es decir, la energía es muy sensible a niveles altos de señal, enfatizando las diferencias de amplitud. Es por ello que se utiliza como unidad de medida el decibelio (dB), calculado como :

$$E_n^{'} = 10 \cdot \log(E_n)$$

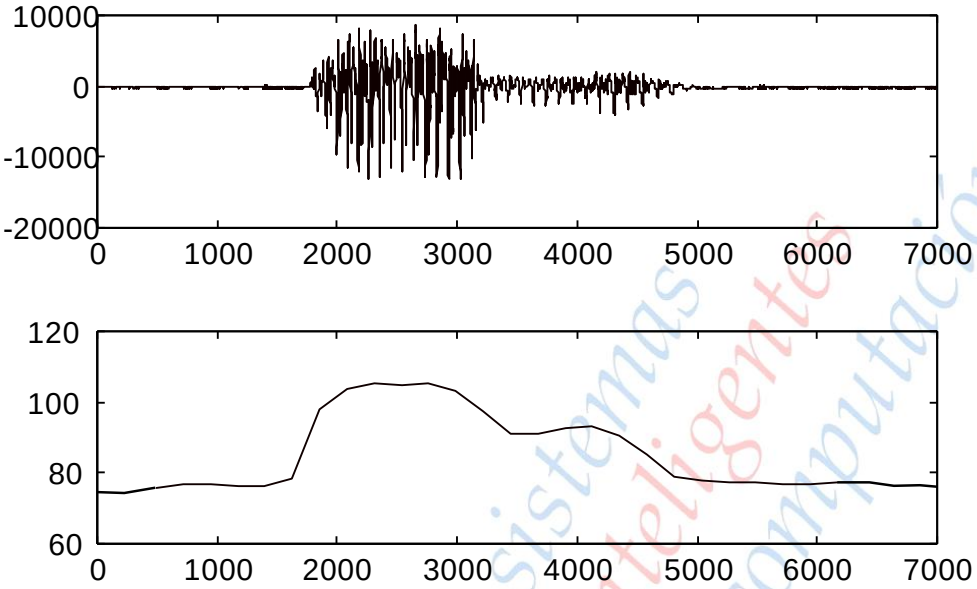


Figura 2.7. Señal de la palabra "uno", y su envolvente de energía

2.6 Densidad de cruces por cero

Un cruce por cero se produce cuando muestras consecutivas de una señal tiene diferente signo. La densidad de cruces por cero consiste en el número de cruces por cero que se producen en un cierto segmento de señal.

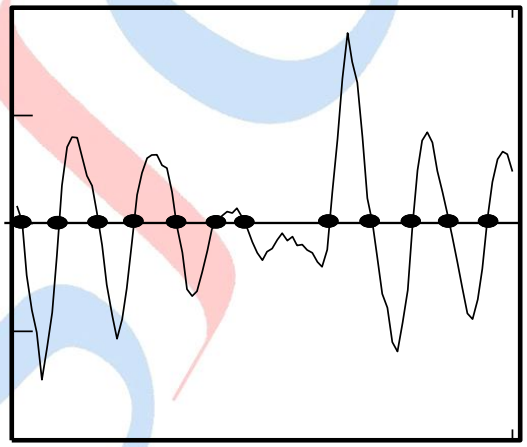


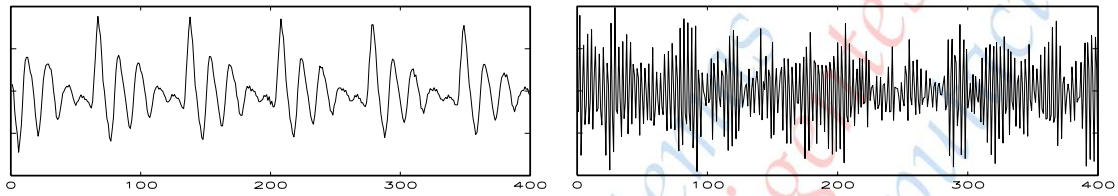
Figura 2.8. Cruces por cero de una señal

En la figura se han marcado los cruces por cero de la señal. Suponiendo que lo que se observa en la imagen sea una trama, la densidad de cruces por cero sería de 12, es decir, igual al número de cruces por cero.
De forma matemática, podemos definir :

$$ZCR = \sum_{m=1}^{N-1} \left| sgn(x(m)) - sgn(x(m+1)) \right|$$

La densidad de cruces por cero nos puede dar una idea del contenido en frecuencia de una señal. Si la medida ZCR es elevada, existen componentes de alta frecuencia, y viceversa. Las distribuciones de probabilidad de las tasas ZCR para señales sonoras o

sordas muestran que dicha magnitud es bastante fiable para la discriminación sonoro/sordo.
Como se observa en la figura, las señales sonoras son más limpias, de trazos más alargados, y por tanto, el número de cruces por cero es inferior. En cambio, las señales sordas tienen una alta componente de *fricción* que hace que el valor ZCR sea mucho mayor.



Señal sonora ***Señal sorda***

Figura 2.9. Señal sonora (menor número de cruces) y señal sorda(ZCR mayor)

La distribución de probabilidad de las tasas ZCR para señales sonoras o sordas muestran que dicha magnitud es bastante fiable para la discriminación sonoro/sordo.

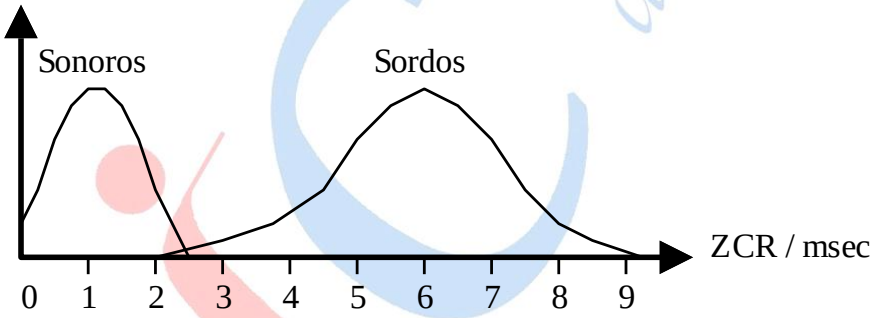


Figura 2.10 Distribución de la Sonoridad en función del índice ZCR

Es evidente que la tasa ZCR está fuertemente afectada por múltiples factores, entre ellos el nivel DC de la señal, los 50 Hz. de la línea, y cualquier ruido del sistema de digitalización. Por tanto hay que tener un cuidado extremo al utilizar esta magnitud.

2.7 Autocorrelación

Supongamos que queremos comparar dos señales x[m] e y[m] de longitud finita. Una medida de similitud es el error cuadrático entre ambas. Si consideramos un posible desplazamiento entre ambas señales, tenemos que considerar :

$$P(k) = \sum_{m=-\infty}^{\infty} (x(m) - y(m+k))^2$$

Esta función la podemos reescribir como :

$$P(k) = \sum_{m=-\infty}^{\infty} x(m)^2 + \sum_{m=-\infty}^{\infty} y(m+k)^2 - 2\sum_{m=-\infty}^{\infty} x(m)\cdot y(m+k)$$

Los dos primeros términos se corresponden a las energías de la señal primera y segunda, y por tanto no dependen del desplazamiento entre ambas señales. Por tanto, la función P se hará mínima cuando la señal siguiente se haga máxima (por el signo - de la expresión anterior)

$$R(k) = \sum_{m=-\infty}^{\infty} x(m) \cdot y(m+k)$$

Por tanto, si consideramos que x e y son la misma señal, se define la función de autocorrelación aplicada a una trama como :

$$R(k) = \sum_{m=1}^{N-k} x(m) \cdot x(m+k)$$

La función de autocorrelación tiene múltiples utilidades, como se verá a lo largo del presente documento.

2.8 Frecuencia Fundamental ó Pitch

La *frecuencia fundamental* de una señal, también denominada *pitch* se define como la frecuencia aparente que una señal tiene. Para ello basta con calcular el período aparente, y utilizar la fórmula :

$$F = \frac{1}{T}$$

Las señales sonoras secaracterizan por tener un valor de pitch muy claro. Sin embargo, las señales sordas carecen de periodicidad. Es en las transiciones donde aparecen los problemas. Es decir, la dificultad estriba en decidir *exactamente* dónde una señal deja de ser periódica, o dónde comienza a serlo. En la figura se muestran dos señales correspondientes a los sonidos /o/ y /s/ de la palabra /dos/. Se observa la evidente periodicidad de la vocal, y la irregularidad de la señal fricativa.

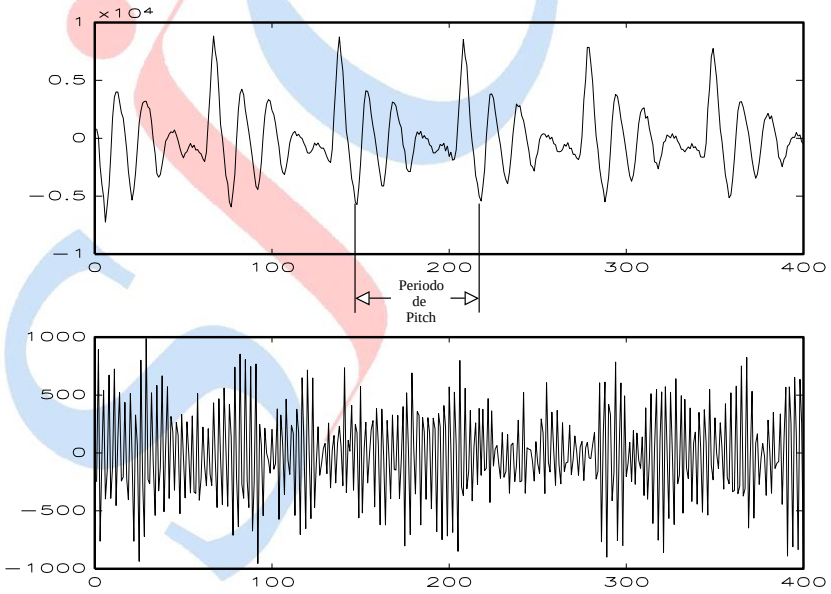


Figura 2.11. Señales temporales correspondiente a la /o/ y a la /s/ extraídas de la palabra /dos/

Existen numerosos algoritmos para calcular el *pitch*, sin embargo, no existe algoritmo conocido que lo determine con un 100% de fiabilidad, ya que, no siempre es fácil determinar si existe periodicidad en una señal o no.

El pitch varía mucho entre diferentes locutores, e incluso para el mismo locutor. De hecho, y junto con la energía, es uno de los factores determinantes en la entonación de las frases. La distribución de probabilidad de los valores de pitch muestra que el pitch

puede ser considerado como una magnitud bastante fiable para la discriminación hombre-mujer. :

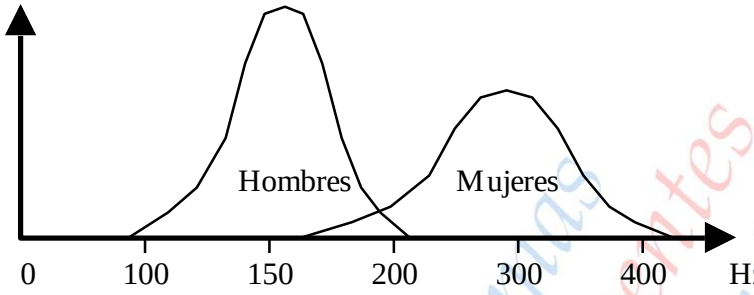


Figura 2.12. Distribución de los valores de pitch según el sexo

La autocorrelación permite determinar el pitch de una forma rápida y precisa, ya que la función de autocorrelación muestra un máximo relativo muy claro en dicho punto. Por ejemplo, en la señal correspondiente al sonido /e/ del apartado 6.1, la función de autocorrelación es la siguiente (en el intervalo correspondiente a los 100-400 Hz , es decir, con un período de 20 a 80 muestras a 8 KHz.) :

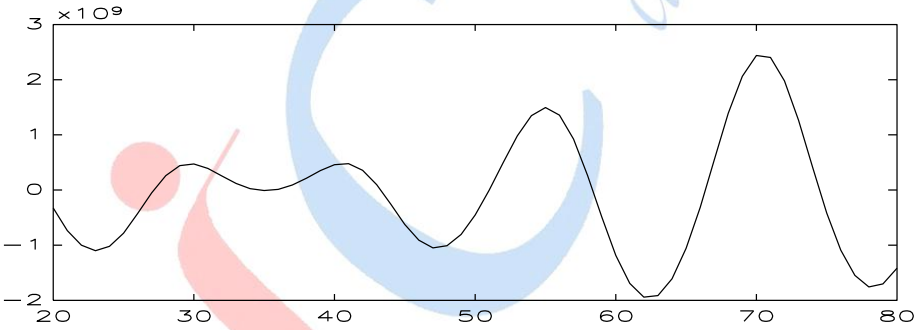


Figura 2.13. Función de autocorrelación para el segmento /e/

Como vemos, aparece un máximo en el punto correspondiente a un pitch de 70 muestras, es decir, 114 Hz.

En determinadas ocasiones, el la función de autocorrelación no es tan simple, y la determinación del máximo absoluto presenta errores. Por ello, es frecuente realizar algún tipo de preprocesado de la señal para eliminar ciertos problemas. En concreto es frecuente utilizar un filtrado paso bajo, y a continuación una técnica denominada *center-clipping.*

Center-Clipping consiste en realizar una transformación no-lineal de la señal, según el siguiente gráfico:

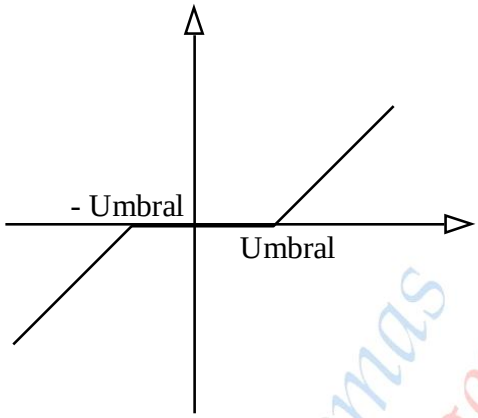


Figura 2.14. Función de Transformación Center-Clipping

Dicha función produce que se elimine toda la señal que no supera un cierto umbral, que normalmente se fija en un 30% del máximo de la señal en la trama en cuestión. En la imagen se observa la señal original, y la señal transformada, sobre la cual se realiza el cálculo de la autocorrelación para la determinación del pitch.

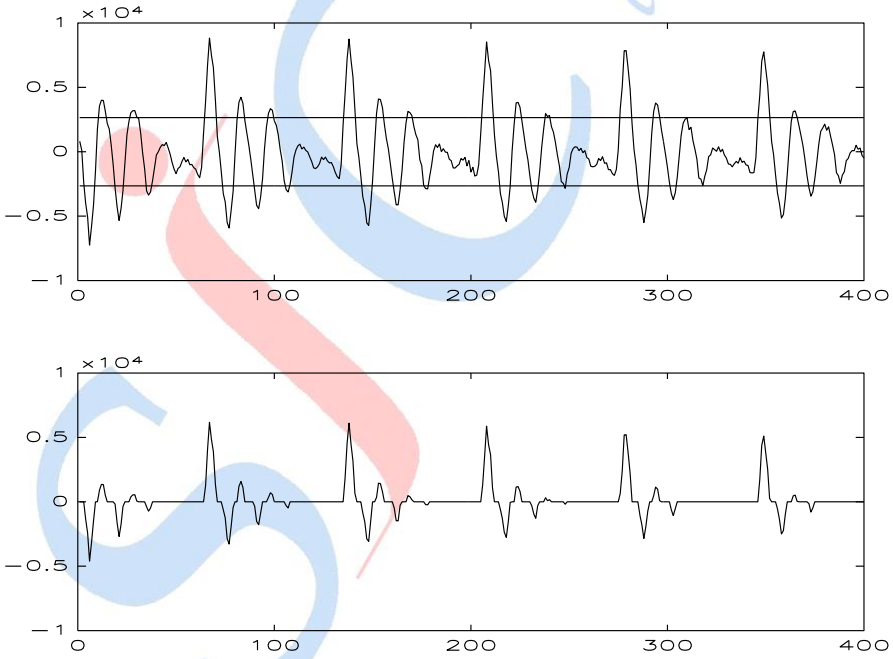


Figura 2.15. Señal original y transformada mediante Center-Clipping

2.9 Análisis Temporal de los sonidos en Castellano

Una vez analizados los parámetros más usuales en el dominio temporal, vamos a utilizarlos para estudiar los sonidos castellanos, y cómo diferenciarlos.

A) Vocales (a , e , i , o , u)

Son la [a,e,i,o,u].

Se puede realizar una clasificación fisiológica de las vocales en función de los parámetros abertura de la cavidad bucal, y de la posición de la lengua.

En función de la abertura de la cavidad bucal, podemos clasificarlas en

- Pequeña abertura, o vocales cerradas o altas

- Media abertura, o vocales medias
- Gran abertura, o vocales abiertas o bajas

En función de la abertura de la lengua distinguiremos :

- Vocales anteriores o palatales : si la lengua ocupa la posición delantera de la cavidad
- Vocales centrales
- Vocales posteriores o velares : si la lengua ocupa la posición trasera de la cavidad (junto al velo)

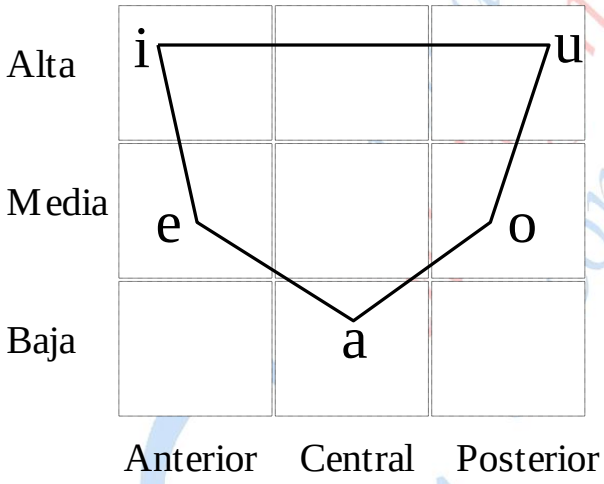


Figura 2.16. Clasificación Fisiológica de las vocales españolas

En otros idiomas, la clasificación no es tan simple. Para su detección podemos distinguir el hecho de que son sonidos de gran Energía y Sonoridad. La complejidad aumenta por el hecho de que existen vocales largas o cortas, acentuadas o inacentuadas, labializadas, etc, etc.

B) Diptongos (ia , ie , io , iu , ua , ue , ui , uo , ai , ei , oi , au , eu , ou)

Son los formados por dos vocales de diferentes características.

Ejemplos de palabras con diptongos son : rabia, agua, tiene, cuerda, labio, ruido, ciudad, antiguo, etc.

Asimismo, existen los denominados triptongos, formados por tres vocales : buey, despreciais, etc.

C) Consonantes

C.1 - Oclusivas (p , t , k , b , d , g)

Se caracterizan por una interrupción momentanea en el paso del aire a través de la cavidad bucal.

	Bilabial	Lingüodental	Lingüovelar
Sonora	b	d	g
Sorda	p	t	k

Son ejemplos de palabras con oclusivas : copa, tapa, paco, ópera, vaso, etc.

C.2 - Fricativas (f , z , s , y , j)

Se caracterizan por una interrupción contínua en el paso del aire.

El aire sale a través de la cavidad bucal.

f : foto, café, fama (fa,fe,fi,fo,fu)
θ : caza, cocer, cicerón (za,ce,ci,zo,zu)
s : casa, mesa, pesar (sa,se,si,so,su)
j : mayo, hierba, cayado (ya,ye,yi,yo,yu) (hi a principio de palabra)
x : gitano, caja, lejos, cojo (ja,je,ji,jo,ju) (ge,gi)
Los sonidos b, d, g a veces aparecen también como fricativos.

C.3 - Africadas (ch)

Se caracterizan por un momento oclusivo seguido de uno fricativo.
ch : muchacho, chico, pecho (cha, che, chi, cho, chu)
Con frecuencia, el sonido j es africado (conyuge, el hielo, etc.)

C.4 - Nasales (m , n , ñ)

Se caracterizan porque el canal rinofaringeo está abierto, y las fosas nasales actúan de cavidad de resonancia.
m : mamá, cama, loma (ma, me, mi, mo, mu)
n : nudo, cono, sartén (na,ne,ni,no,nu,én,ón,ún)
ñ : año, leña, soñar (ña, ñe, ñi, ño, ñu)

C.5 - Líquidas (l , ll , r , rr)

Poseen rasgos comunes a las consonantes y a las vocales. Se clasifican en :

- *Laterales*
  ll : llama, llave, llegar (lla, lle, llo, llu)
  l : lado, papel, tela (la,le,li,lo,lu,al,el,il,ol,ul)
- *Vibrantes*
  r : coro, pero, tara (ra,re,ri,ro,ru,ar,er,ir,or,ur)
  rr : perro, roca, torreón (rra,rre,rri,rro,rru) (ra,re,ri,ro,ru al inicio de palabra)

3 Análisis Frecuencial de señales

3.1 Conceptos básicos

Cuando tiramos una piedra en el agua, cuando hacemos vibrar un diapasón o cuando silbamos, se producen "ondas". Dichas ondas están motivadas por una presión variable con forma de vaivén que se produce en un medio transmisor (agua, aire, etc.) y que responde a una forma sinusoidal. Las señales sinusoidales aparecen con frecuencia en la naturaleza, dado que son las más simples que podemos encontrar. Una señal sinusoidal se caracteriza por tres términos, la amplitud, la frecuencia y la fase.
La representación gráfica de una señal sinusoidal es la siguiente :

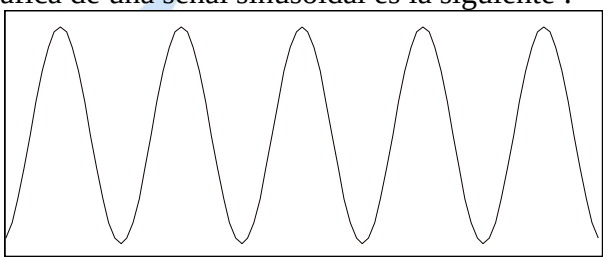
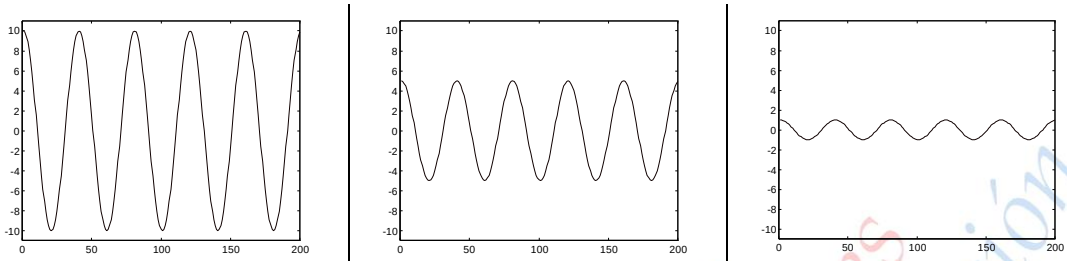


Figura 3.1 Señal sinusoidal.

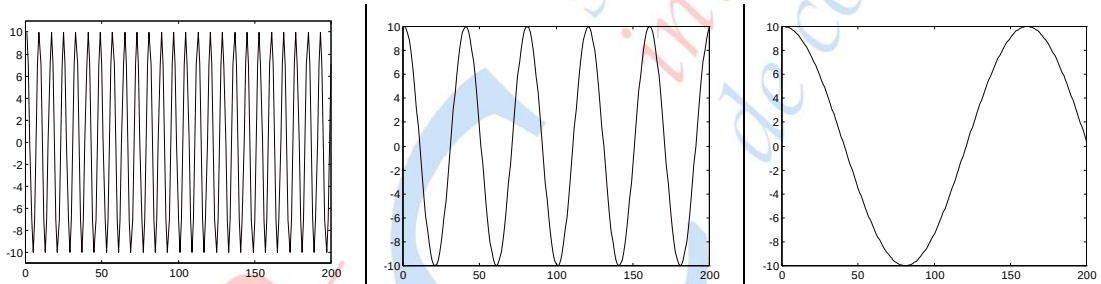
La amplitud hace referencia al "tamaño" de la señal. Si escuchásemos la señal, gran amplitud se corresponde con un volumen alto, y poca amplitud con poco volumen. En la figura se muestran tres señales de alta, media y baja amplitud.



Señal de gran amplitud | Señal de amplitud media | Señal de baja amplitud

Figura 3.2. Señales sinusoidales de diferente amplitud

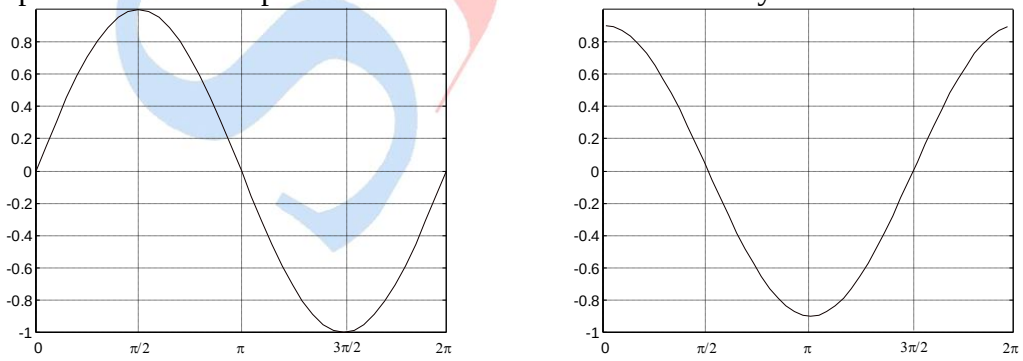
La frecuencia es el número de ciclos que hay en un segundo. La frecuencia se mide en ciclos por segundo, o Hertzios (Hz.). La frecuencia hace referencia al tono de un sonido. Alta frecuencia se corresponde con tonos agudos, y baja frecuencia con tonos graves. En la figura se muestran tres señales de alta, media y baja frecuencia.



Señal de alta frecuencia | Señal de media frecuencia | Señal de baja frecuencia

Figura 3.3. Señales sinusoidales de diferente frecuencia

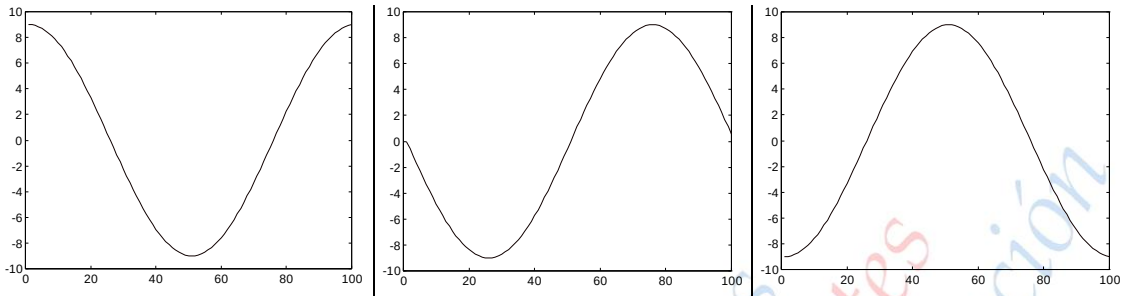
La fase hace referencia al momento en que empieza la señal. Una fase igual a 0 para un seno significa comenzar en 0 ascendentemente. Las señales sinusoidales que se utilizan son la señal seno (comienza en 0, y empieza a ascender), y la señal coseno (comienza en el pico más alto, y empieza a descender). En la siguiente figura se muestra el valor de la fase para los diferentes puntos de comienzo de una señal seno y otra coseno.



Señal seno | Señal coseno

Figura 3.4. Valores de la fase para las señales seno y coseno

En la figura siguiente se muestran diferentes posibilidades para el valor de la fase.



Señal seno de fase π/2
o bien
Señal coseno de fase 0

Señal seno de fase π
o bien
señal coseno de fase π/2

Señal seno de fase 3π/2
o bien
señal coseno de fase π

Figura 3.5. Ejemplo de señales con su correspondiente fase

3.2 Concepto de filtro

Un *filtro* consiste en un sistema que recibe una entrada x(t) y produce una salida y(t), tal y como se muestra en la salida.

Figura 3.6. Esquema de un Filtro

Un *filtro analógico* es un sistema que responde a la ecuación diferencial :

$$\sum_{k=0}^{P} a_k \frac{d^k y(t)}{dt^k} = \sum_{l=0}^{Q} b_l \frac{d^l x(t)}{dt^l} \qquad a_0 = 1$$

donde y(t) es la salida del sistema, y x(t) la entrada.

3.3 Filtros Digitales

Un *filtro digital*, responde a la ecuación recurrente

$$\sum_{k=0}^{P} a_k y\ n-k\ = \sum_{l=0}^{Q} b_l x\ n-l \qquad a_0 = 1$$

que se puede reescribir de la siguiente forma :

$$y\ n\ = -\sum_{k=1}^{P} a_k y\ n-k\ + \sum_{l=0}^{Q} b_l x\ n-l$$

o lo que es lo mismo, que la salida n-sima puede ser calculada en función de las P salidas anteriores, y de las Q entradas enteriores. Por tanto, ha de tener lo que se llama "memoria", para recordar los valores de las entradas y salidas anteriores. Un filtro se inicializa asignando el valor 0 a su memoria. En lo sucesivo trabajaremos únicamente con filtros digitales.

3.4 Tipos de Filtro por la Respuesta al Impulso

Como vemos, al definir un filtro hay que definir los coeficientes a_k y b_l. Supongamos que P y Q son finitos. Si todos los coeficientes a_k valen 0, la salida en el instante n solo

depende de los valores de x, es decir, de la entrada al sistema, y de los coeficientes b_l. Por tanto, y dado que Q es finito, la salida solo depende de las Q últimas entradas. Si, por el contrario, algún coeficiente a_k es distinto de 0, existe una recurrencia, que lleva, a que la salida del filtro pueda depender de las infinitas muestras anteriores.
Por tanto, y en función de los valores a_k existen dos tipos básicos de filtros digitales :

- Filtros de Respuesta al Impulso Finita (FIR) : todos los coeficientes a_k valen 0, salvo a_k
- Filtros de Respuesta al Impulso Infinita (IIR) : existen coeficientes a_k distintos de 0

La ventaja de los filtros FIR es que son estables. La ventaja de los filtros IIR es que obtienen mejores filtrados con menos coeficientes.

3.5 Tipos de Filtro por las Frecuencias de Filtrado

Los filtros se suelen utilizar como sistemas que dejan pasar una cierta información, eliminando ó atenuando otra. En concreto, si una cierta señal es introducida en un filtro digital, deja pasar ciertas frecuencias, y atenúa otras.
En función de esta atenuación, existen diferentes tipos de filtro :

- Paso Bajo : solo permite el paso a las frecuencias menores que la *frecuencia de corte*.

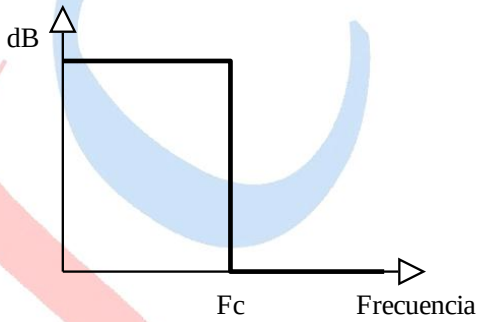


Figura 3.7. Esquema de un Filtro Paso Bajo

- Paso Alto : solo permite el paso a las frecuencias mayores que la *frecuencia de corte*.

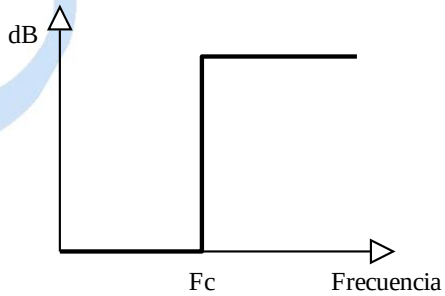


Figura 3.8. Esquema de un Filtro Paso Alto

- Paso Banda : permite el paso de las frecuencias en una cierta banda.

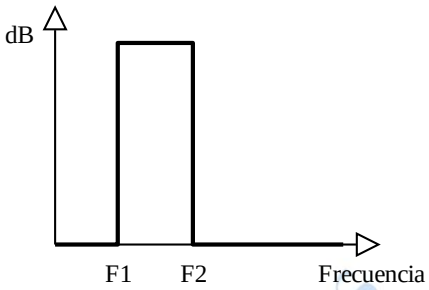


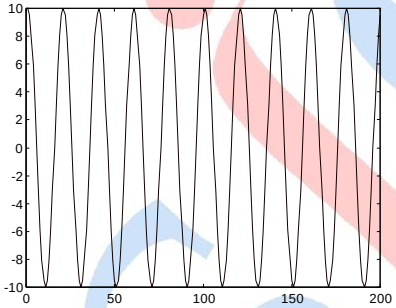
Figura 3.9. Esquema de un Filtro Paso Banda

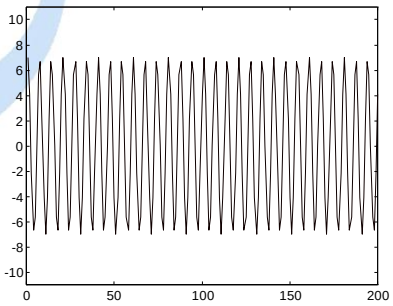
A este gráfico en el que se observa qué frecuencias "pasan" por el filtro y qué otras no, se le denomina *Respuesta en frecuencia* del filtro en cuestión.
A este gráfico en el que se observa qué frecuencias "pasan" por el filtro y qué otras no, se le denomina *Respuesta en frecuencia* del filtro en cuestión.
Supongamos una señal obtenida como la suma de dos señales sinusoidales, una de 400 Hz. y la otra de 1200 Hz. Para ello, utilizamos la función que nos permite generar un coseno de una cierta amplitud, a una cierta frecuencia, y con una cierta fase, según las siguientes instrucciones :

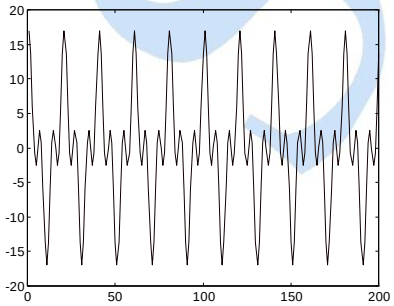
```
x1=cosgen(8000,10,400,0,200);
x2=cosgen(8000,7,1200,0,300);
x=x1+x2;
```

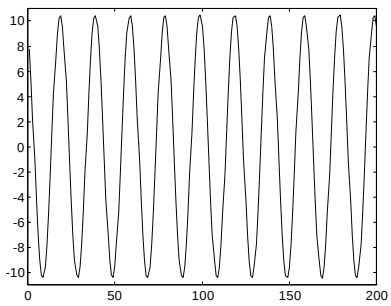
Diseñamos un filtro que solo deja pasar frecuencias por debajo de 600 Hz. y observemos qué sucede :

```
[b,a]=cheby2(20,150,0.15);
```


x1


x2


x1 + x2


filter(b,a,x1)

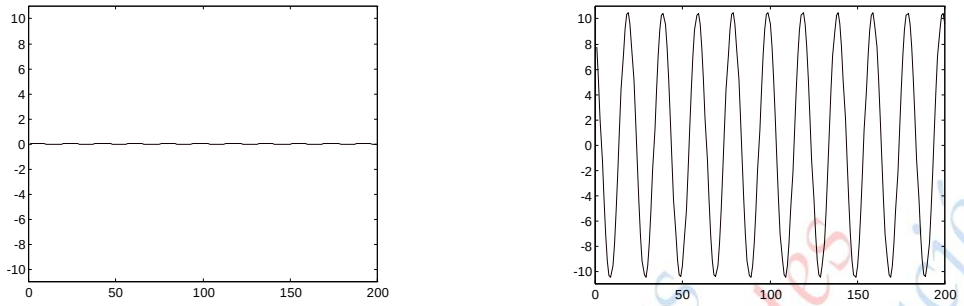


Figura 3.10. Ejemplo de filtrado de una suma de señales sinusoidales

Como vemos, el filtro solo deja pasar las señales sinusoidales menores que 600 Hz. Al filtrar x1, la señal queda inalterada. Al filtrar x2, la salida del filtro es prácticamente nula. Al filtrar la señal suma x, el filtro elimina la componente x2, y solo deja x1, como se ve en las figuras anteriores.

3.6 Respuesta en Frecuencia de un Filtro Digital

Consiste en un gráfico que indica qué frecuencias deja pasar un filtro y cuáles no, tal como las que hemos visto cuando hablamos de tipos de filtro. Se representa mediante un gráfico *db* versus *Frecuencia*. Las respuestas en frecuencia vistas hasta ahora han sido ideales (ángulos rectos por todos sitios). En la realidad aparecen efectos indeseados muy difíciles de eliminar. En la figura 3.6 se observa la respuesta en frecuencia real de un filtro paso bajo con una frecuencia de corte a 2500 Hz..

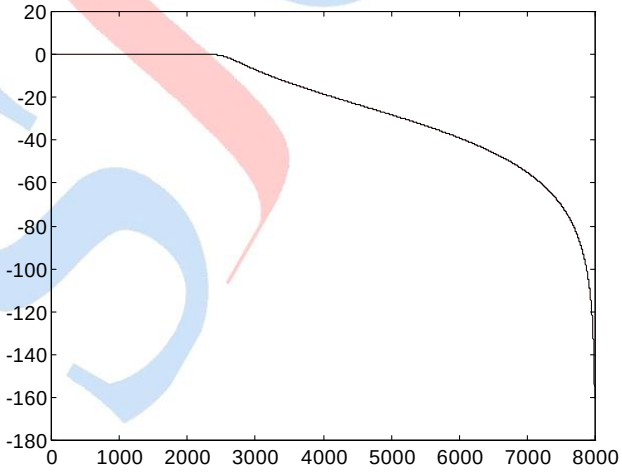


Figura 3.11. Respuesta en frecuencia de un filtro.

La respuesta mide la atenuación que un cierto filtro aplica a cada frecuencia. Es decir, consiste en obtener el gráfico *db* versus *Frecuencia*. Las respuestas en frecuencia anteriores son ideales. En la realidad aparecen efectos indeseados muy difíciles de eliminar.

La respuesta en frecuencia de un filtro se puede calcular mediante diferentes métodos. No es el objetivo de este curso profundizar más en el tema. Por tanto, referimos al lector

a la utilización de la función *freqz* para el cálculo de la respuesta en frecuencia de un filtro, dados sus coeficientes.
La forma más práctica de observar la respuesta consiste en evaluar su magnitud en dB. Por tanto, la operación a realizar es :

```
m = 10*log10(abs(freqz(b,a,N)));
```

donde N es el número de puntos deseado. El punto N-simo se corresponde con la mitad de la frecuencia de muestreo.

3.7 Energia por bandas

Es interesante el análisis de la señal de voz filtrando la misma por un *Banco de Filtros*, que consiste en una serie de filtros, por los que se hará pasar la señal para determinar la energía existente en diferentes rangos de frecuencia. Ello nos va a permitir conocer el espectro aproximado de la señal. Por ejemplo, podemos diseñar el banco siguiente :

Filtro 1 : Paso Bajo - 300 Hz.
Filtro 2 : Paso Banda 300 - 600 Hz.
Filtro 3 : Paso Banda 600 - 900 Hz.
Filtro 4 : Paso Banda 900 - 1200 Hz.
Filtro 5 : Paso Banda 1200 - 1500 Hz.
Filtro 6 : Paso Banda 1500 - 1800 Hz.
Filtro 7 : Paso Banda 1800 - 2100 Hz.
Filtro 8 : Paso Banda 2100 - 2500 Hz.
Filtro 9 : Paso Banda 2500 - 3000 Hz.
Filtro 10 : Paso Alto 3000 Hz.

Este banco de filtros nos proporciona una salida de 10 puntos por trama que nos da una idea del espectro de frecuencias existente en la señal de entrada. El proceso completo sería el siguiente :

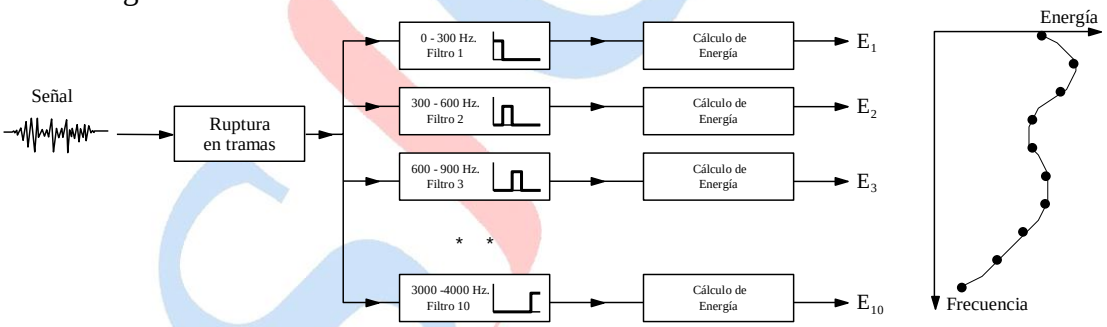


Figura 3.12. Esquema elemental de un Banco de Filtro

3.8 Respuesta en Frecuencia de un Filtro Digital

Consiste en la atenuación que un cierto filtro aplica a cada frecuencia. Es decir, consiste en obtener el gráfico *db* versus *Frecuencia*. Las respuestas en frecuencia anteriores son ideales. En la realidad aparecen efectos indeseados muy difíciles de eliminar.
La respuesta en frecuencia de un filtro se puede calcular mediante diferentes métodos. No es el objetivo de este curso profundizar más en el tema. Por tanto, referimos al lector a la utilización de la función *freqz* para el cálculo de la respuesta en frecuencia de un filtro, dados sus coeficientes.
La forma más práctica de observar la respuesta consiste en evaluar su magnitud en dB. Por tanto, la operación a realizar es :

```
m = 10*log10(abs(freqz(b,a,N)));
```

donde N es el número de puntos deseado. El punto N-simo se corresponde, por supuesto, con la mitad de la frecuencia de muestreo.

3.9 Análisis de señales

Toda señal periódica puede ser descompuesta en suma de una serie de señales sinusoidales (senos o cosenos). Por ejemplo, sea la señal siguiente que se muestra en la figura

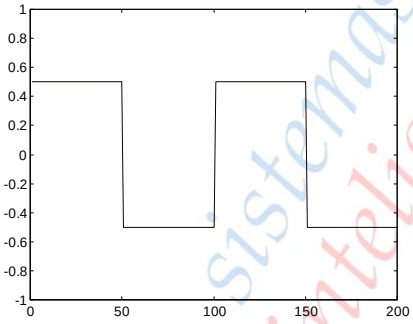


Figura 3.13. Señal original

Y observemos qué sucede cuando sumamos los siguientes cosenos :

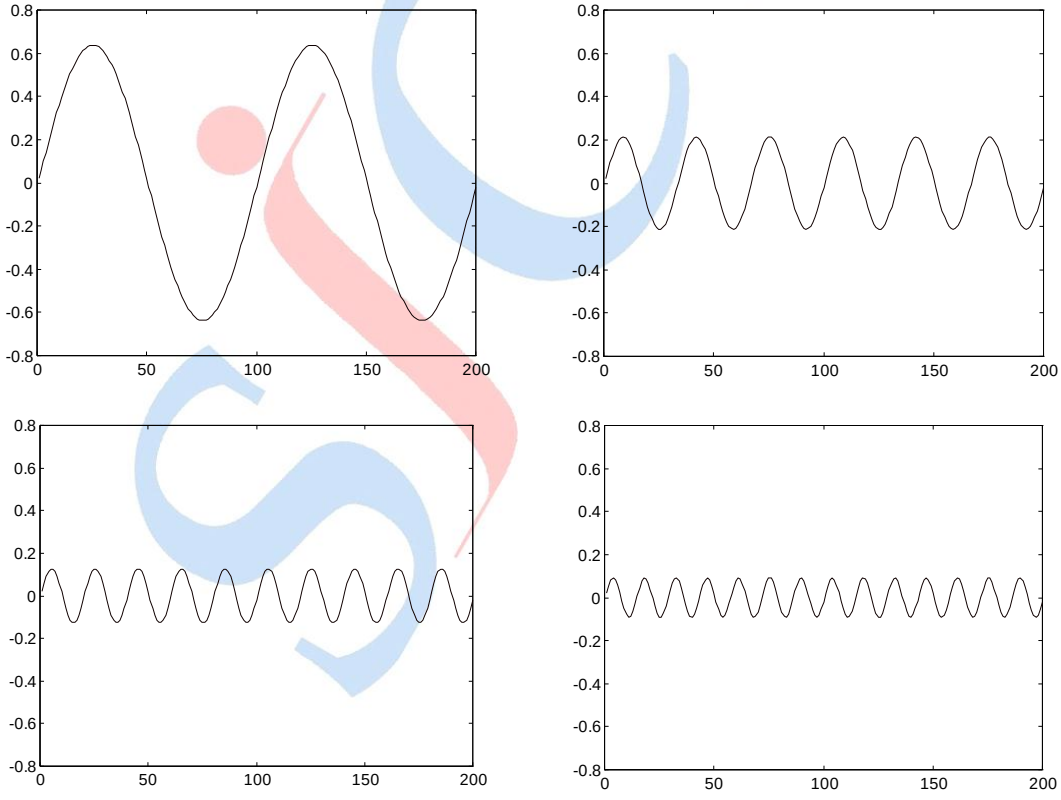


Figura 3.14. Señales componentes

Tras sumar los 4 cosenos anteriores, obtenemos la siguiente aproximación :

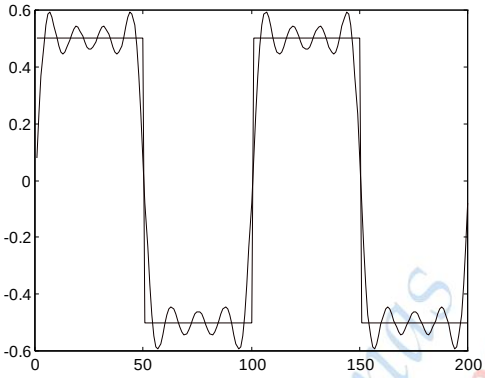


Figura 3.15. Señal resultante de la suma de 4 cosenos

Si en lugar de 4 cosenos, elegimos 20, el resultados es el siguiente :

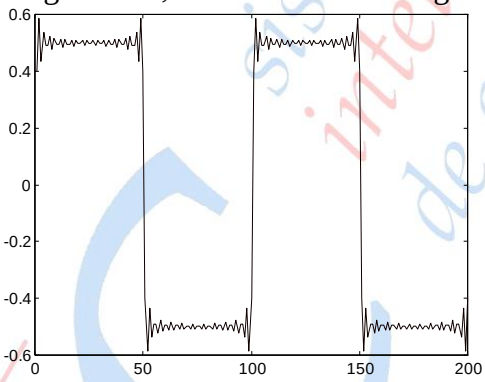


Figura 3.16. Señal resultante de la suma de 20 cosenos

Y, sin en vez de tomar 20 cosenos, tomamos 40, el resultado es :

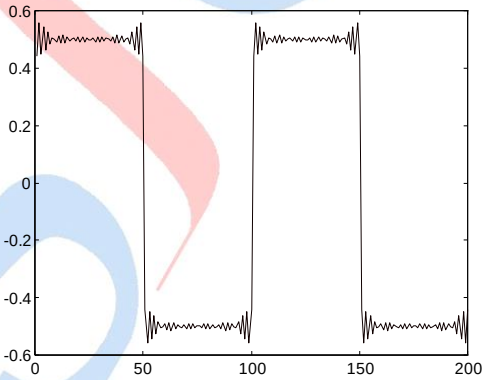


Figura 3.17. Señal resultante de la suma de 40 cosenos

Como sucede con la señal mostrada en la figura 3.17, mediante la suma de un número elevado de señales sinusoidales, la señal suma se aproxima tanto como queramos a la señal original.

3.10 La FFT

Las señales periódicas se suelen analizar en términos de componentes sinusoidales, denominadas Series de Fourier. Mientras que las sinusoides son señales de una única frecuencia, todas las señales periódicas pueden ser expresadas como una combinación lineal de sinusoides ponderadas.

La FFT (Fast Fourier Transform) consiste en un método para realizar un análisis de Fourier de una señal considerada como periódica. Es decir, dividir la señal en una serie de cosenos de diferentes amplitudes y fases, tales, que sumados, dan como resultado la señal original.

$$x(t) = \sum_{k=-\infty}^{\infty} c_k \cdot e^{\frac{2\pi k t j}{T}}$$

donde T es el período de la señal, y los coeficientes se definen como

$$c_k = \int_{t=T_0}^{T+T_o} x(t) \cdot e^{-\frac{2\pi k t j}{T}} \cdot dt$$

La FFT permite calcular los citados coeficientes de una trama de señal considerada como periódica. Para ello, supongamos una trama de una señal periódica, y tal que su longitud es un múltiplo de su período. (Es decir, que lo que debería seguir a la trama de señal coincide con la propia trama) Si realizamos la operación :

X(w)=FFT(x(t));

obtenemos una señal X(w) con las siguientes características :

- Es una señal compleja.
- La señal se puede dividir en dos mitades "casi" simétricas. Son simétricas conjugadas.
- El valor en un cierto punto (sea el punto n-simo) se correponde con la sinusoide de frecuencia n-sima.
- La frecuencia n-sima se puede calcular por una simple regla de tres :

  Si la trama tiene N puntos, la frecuencia del punto N de la transformada se corresponde con la mitad de la frecuencia de muestreo. De donde

  N ---------- FrecMuestreo / 2

  n ---------- x

  Por tanto, la frecuencia del punto n-simo se corresponde con la frecuencia FrecMuestreo*n/2/N.
- La amplitud de la sinusoide n-sima se puede calcular como la magnitud del valor complejo n-simo
- La fase de la sinusoide n-sima se puede calcular como la magnitud del valor complejo n-simo.
- Por ser una transformada coseno, la fase viene referida a señales coseno.
- La fase tiene poca importancia en el Reconocimiento de Voz. Por tanto, a partir de ahora, siempre nos referiremos a la amplitud.

Sea x una señal periódica (o cuasi-periódica), y sea y=fft(x). Elegimos los m que hacen mayores a la y, y obtenemos :

```
xi=cosgen(1000,
          2*abs(y(m))/length(x),
          1000/length(x)*(m-1),
          angle(y(m)),
          length(x));
```

donde cosgen genera a una cierta frecuencia de muestreo un cierto número de puntos de una señal coseno de una cierta amplitud, frecuencia y fase.

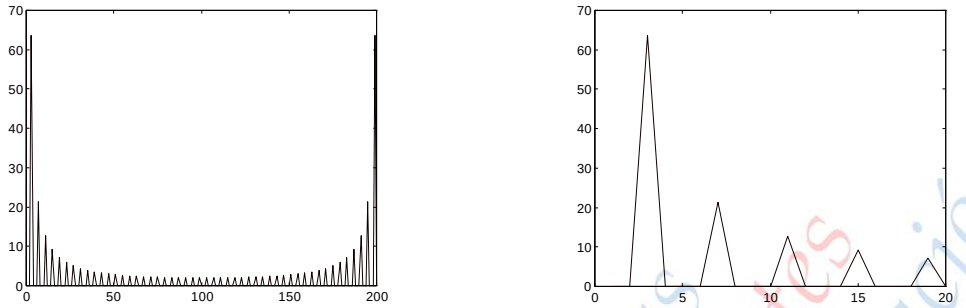


Figura 3.18. Transformada FFT de una señal (256 puntos, y 26 primeros puntos)

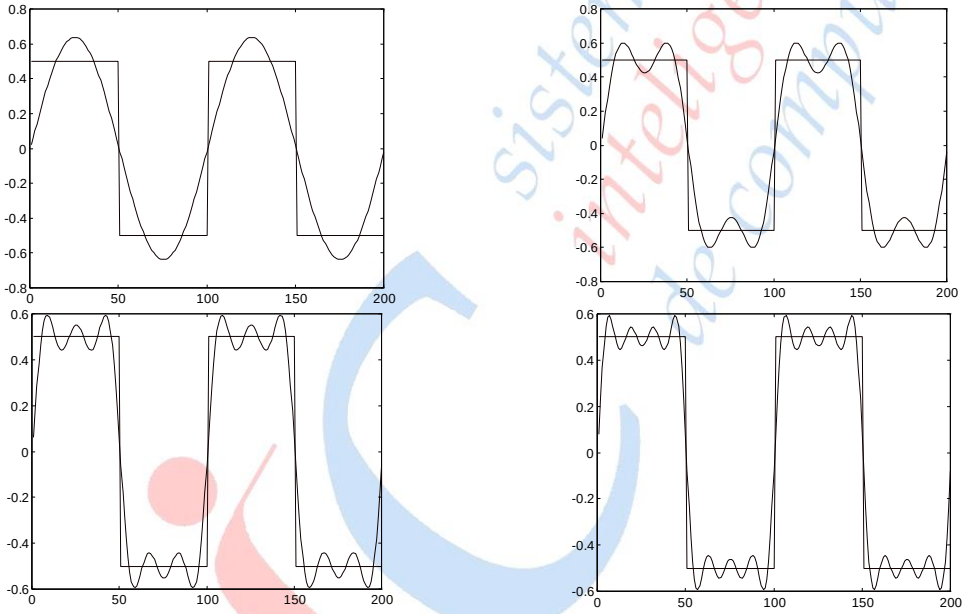


Figura 3.19. Proceso de análisis de señales mediante la FFT

3.11 Enventanado

Dado que la FFT supone que la señal de la trama en cuestión es periódica, cuando aplicamos la misma a señales no periódicas, hemos de minimizar las discontinuidades que la señal pueda tener al comienzo y final de cada trama, realizándose un enventanado. Para ello, se ejecuta la siguiente operación con cada trama :

$$X(n) = x(n) \cdot w(n) \qquad n = 1, 2, \ldots N$$

La ventana $w(n)$ suele tener un valor próximo a 0 en sus extremos, y normalmente es simétrica respecto del centro de la misma. La multiplicación de la ventana por la señal tiene dos efectos:

- Atenuar de forma gradual la señal a ambos lados de la trama seleccionada.
- Produce una convolución de la transformada de Fourier de la función de la ventana y del espectro de la señal.

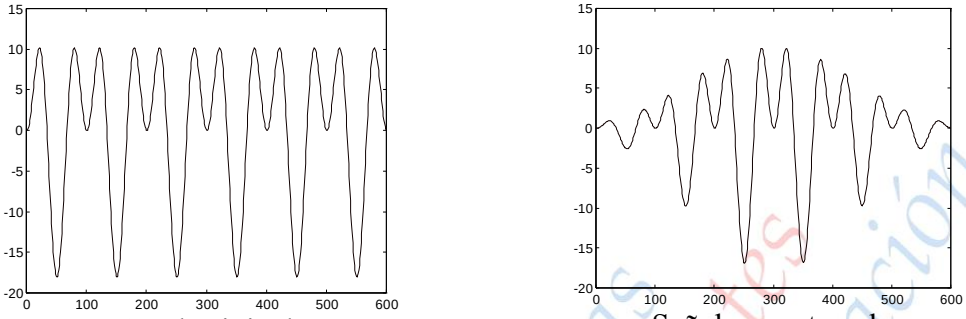


Figura 3.20. Efecto del enventanado sobre una señal

Debido al segundo efecto, la ventana debe satisfacer dos características para reducir la distorsión espectral introducida por el enventanado:

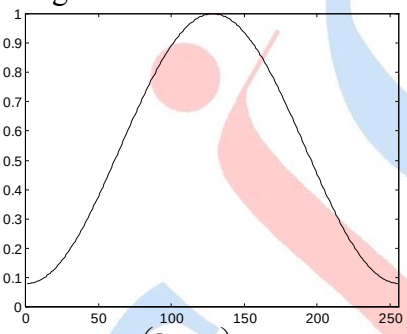
- Lóbulo principal estrecho y agudo, con buena resolución en alta frecuencia.
- Gran atenuación de los lóbulos secundarios.

Estas dos características, generalmente, son contrapuestas, y por tanto es preciso buscar un compromiso entre ambas.

La ventana más sencilla es la rectangular, que se define como :

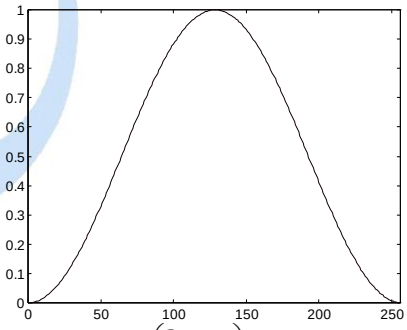
$$w(n) = 1 \qquad 0 \le n \le N\text{-}1$$

Las ventanas más comúnmente utilizadas son la de Hamming, la ventana de Hanning y la Rectangular.



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

Figura 3.21. Ventana de Hamming



$$w(n) = 0.5 - 0.5 \cdot \cos\left(\frac{2 \cdot \pi \cdot n}{N-1}\right) \qquad 0 \le n \le N\text{-}1$$

Figura 3.22. Ventana de Hanning

3.12 Tamaño de Trama

El tamaño de la trama determina la cantidad de información que se procesa en cada transformada al campo de la frecuencia. Veamos el ejemplo siguiente :

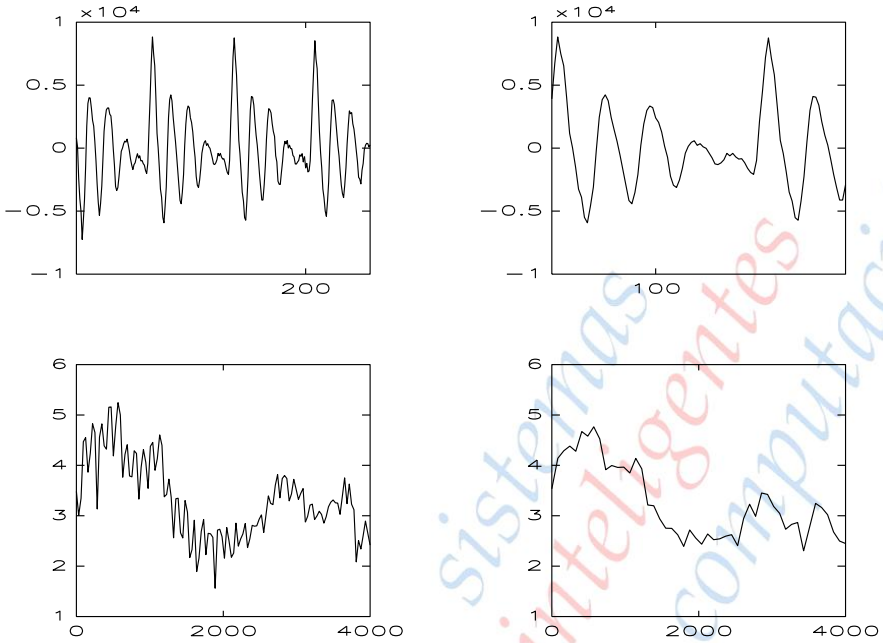


Figura 3.23. FFT de la misma señal con ventanas de 256 y 100 puntos (Fs=8KHz.)

Hemos realizado dos transformadas FFT, y hemos dibujado su magnitud. La primera se corresponde a una trama de una señal de 256 puntos, y la segunda a una trama de la misma señal, de sólo 100 puntos. Se observa que la segunda representa mejor la envolvente espectral, y no tiene en cuenta la estructura fina del espectro. Por tanto, si N es grande respecto al pitch, en el espectro se puede analizar muy bien los armónicos correspondientes al pitch, pero la envolvente espectral está "camuflada" en cierto modo. Por el contrario, si N es pequeño respecto al pitch, la señal tiene poca resolución en frecuencia, pero la envolvente espectral es muy limpia.
Una buena opción para enventanar voz sonora consistiría en una ventana rectangular con una duración igual a un período de pitch. Esto produciría un espectro de salida muy cercano a la respuesta al impulso del tracto vocal. Sin embargo, hay dos grandes problemas. Primero, es muy difícil la localización exacta y fiable del período de pitch. Y segundo, sucede que la mayoría de los períodos de pitch (sobre too en las mujeres) son más breves que la respuesta al impulso del tracto vocal, y la ventana truncaría dicha respuesta, degradando la resolución espectral.

3.13 Preénfasis

Para reducir el rango dinámico de las señales espectrales, se suele alisar el espectro para compensar los valores de las altas y de las bajas frecuencias mediante un filtro denominado Filtro de Preénfasis.
El preénfasis consiste se realiza mediante una simple diferenciación :

$$s'(n) = s(n) - a \cdot s(n-1)$$

donde *a* refleja el grado de preénfasis, y suele estar en el rango de 0.9 a 1.0
Cuanto más cercano es *a* al valor 1.0, mayor es el efecto de preénfasis. En la siguiente figura mostramos las respuesta en frecuencia típica de un filtro de preénfasis :

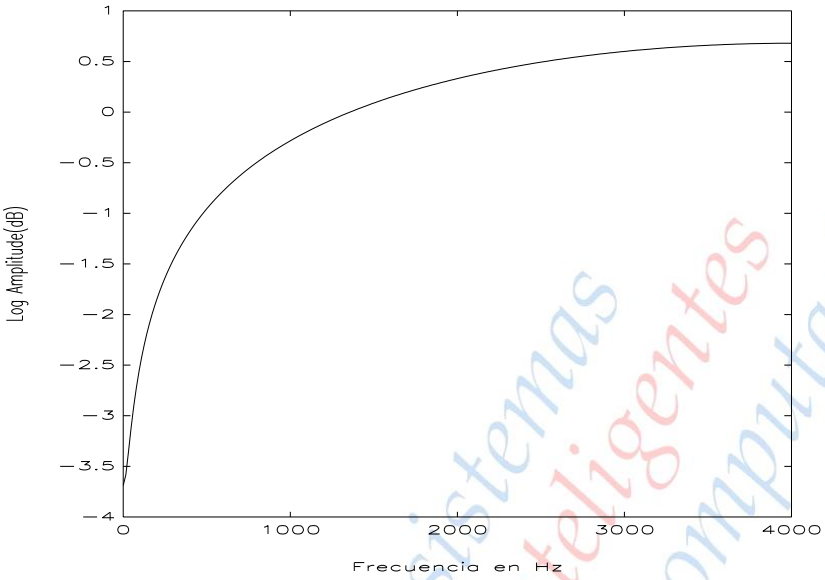


Figura 3.24. Respuesta en Frecuencia típica de un filtro de preénfasis (a=0.975)

Idealmente, el preénfasis sólo se debe aplicar a señales sonoras. Sin embargo, por la pequeña distorsión que se introduce en las señales aperiódicas, y por simplificar el sistema de análisis, la práctica totalidad de los sistemas actuales aplican el preénfasis a todo tipo de señales.

3.14 Formantes

Los formantes se corresponden (aproximadamente) con los máximos en la envolvente espectral. realmente son zonas de resonancia en las que se pone de relieve un conjunto determinado de armónicos.

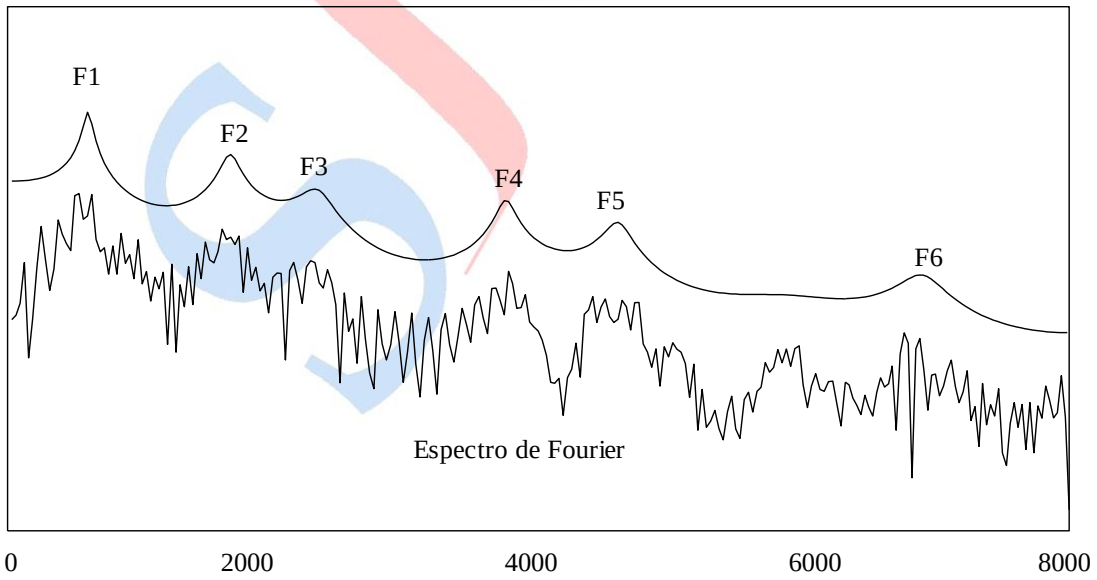


Figura 3.25. Localización de Formantes en el espectro de una señal

Al valor de pitch se le denomina normalmente F0, ó formante 0, y, como ya sabemos, depende del tono, y por tanto, del sexo del hablante. El formante F1 guarda una estrecha relación con la abertura del canal bucal. Cuando la abertura es máxima, la frecuencia de dicho formante es elevada. Cuando la abertura decrece, así lo hace F1. El formante F2

se modifica por la posición de la lengua (cuanto más elevada se halle, y más anterior, mayor será la frecuencia de F2) y por la posición de los labios (cuanto más redondeados y abocinados, más bajo será F2)
La utilidad de los formantes es determinante en la discriminación de numerosos sonidos, por ejemplo, las vocales. En las 5 figuras siguientes se muestran las envolventes espectrales correspondientes a las 5 vocales españolas :

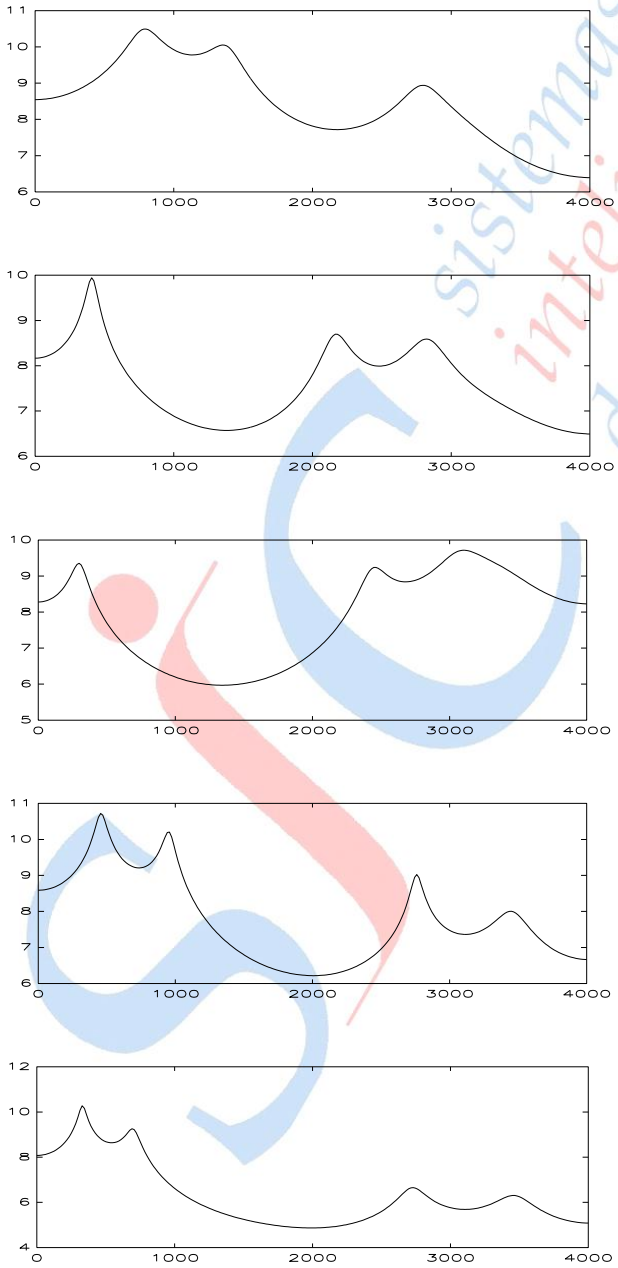


Figura 3.26. Envolventes espectrales de las vocales españolas

En concreto, simplemente utilizando los formantes F1 y F2, podemos realizar un mapa que nos permita diferenciar las 5 vocales del alfabeto entre sí :

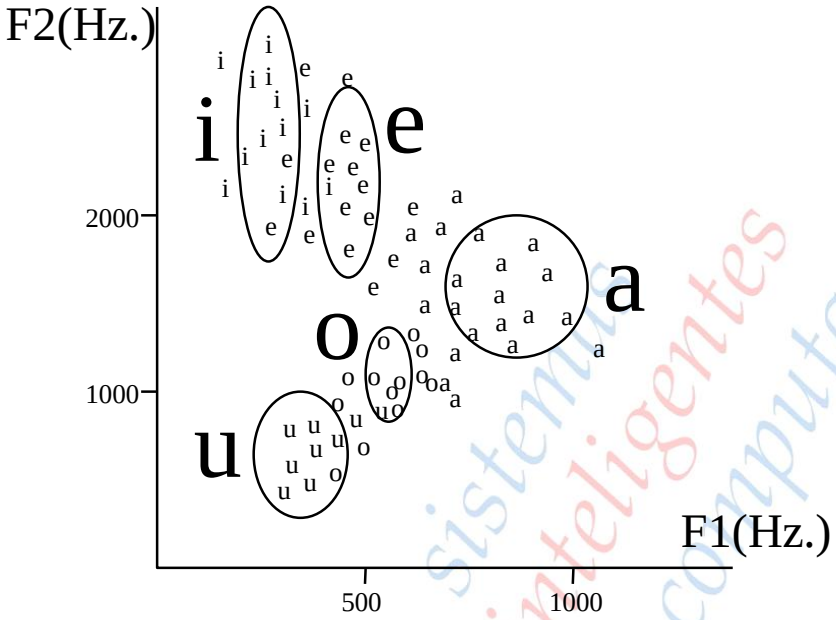


Figura 3.27. Mapa de Formantes de las 5 vocales españolas

3.15 Espectrogramas

Consisten en gráficos donde se muestran en forma bidimensional la evolución espectral a lo largo del tiempo. Normalmente se representa en un diagrama, donde el eje horizontal representa el tiempo, el eje vertical representa la frecuencia, y en escala logarítmica de tonalidades de gris (blanco=valor bajo, negro=valor alto) se muestran las amplitudes de los correspondientes espectros.

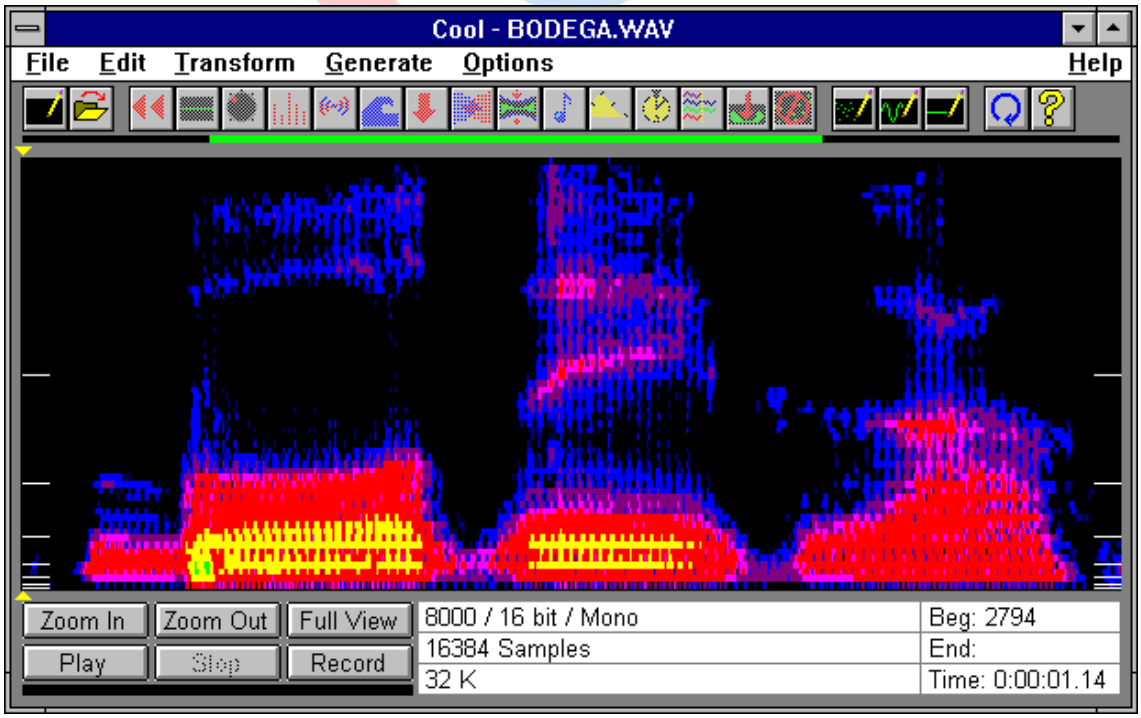


Figura 3.28. Espectrograma de la palabra " bodega "

3.16 Análisis de los fonemas castellanos

A) Detección Voz / Silencio

Se puede utilizar como elemento base la Energía y opcionalmente la Densidad de Cruces por cero.

B) Detección Sonoro / Sordo (ZCR, XCORR)

Se puede utilizar la Densidad de Cruces por Cero y la Autocorrelación.

C) Vocales (a , e , i , o , u)

Son sonidos con alta energía, claramente sonoros (es decir, periódicos), tienen una duración larga (comparadas con la mayoría de las consonantes) y están espectralmente muy bien definidas.

Existen diferentes métodos para clasificar las vocales, pero es suficiente el conocimiento para una clasificación bastante fiable entre sí.

D) Diptongos (ia , ie , io , iu , ua , ue , ui , uo , ai , ei , oi , au , eu , ou)

Son sonidos con las mismas características que las vocales, solo que la envolvente espectral varía uniformente durante el transcurso de su pronunciación, evolucionando de una vocal a otra.

E) Consonantes

Son sonidos con energía variable, pueden sonoros o sordos, tienen una duración variable y espectralmente no suelen estar bien definidas. Como vemos, la complejidad de su clasificación es bastante grande.

E.1 - Oclusivas (p , t , k , b , d , g)

Las oclusivas sordas se detectan con facilidad, dada la existencia de una zona previa a la oclusión con energía casi nula. Las oclusivas sonoras son más difíciles de detectar. La clasificación entre ellas se realiza ananlizando la envolventye espectral en la zona previa a la oclusión (oclusivas sonoras) y alrededor de la oclusión(oclusivas sordas)

E.2 - Fricativas (f , z , s , y , j)

Son sonidos eminentemente sordos, es decir, aperiódicos. Por tanto, la mayor concentración de energía se produce en las altas frecuencias. Suelen tener poca energía. Son fáciles de clasificar en función de la envolvente espectral.

E.3 - Africadas (ch)

Es un sonido sordo, con características similares a las de las consonantes fricativas.

E.4 - Nasales (m , n , ñ)

Son sonoros. Muestran una concentración de energía en la baja frecuencia, y un rango de frecuencias medias, sin picos aparentes. Su apariencia temporal se parece a la de las vocales, aunque su energía es significativamente menor. Suelen ser difíciles de clasificar entre sí.

E.5 - Líquidas (l , ll , r , rr)

Como ya sabemos, poseen rasgos comunes a las consonantes y a las vocales, y son bastante difíciles de caracterizar. Sus características dependen en gran medida del contexto en que ocurren. En general, son muy similares a las vocales y diptongos, es decir, sonoras, duración larga, la mayor parte de la energía en las bajas frecuencias, aunque su definición espectral no es muy buena. Tienen una menor energía que las vocales.

4 Predicción Lineal

4.1 Concepto

Supongamos que queremos realizar una estimación del valor de una muestra en el instante n como una combinación lineal de las p muestras anteriores.

$$s'(n) = \sum_{i=1}^{p} a_i \cdot s\ n-i$$

El conjunto de coeficientes óptimo será aquel que haga que el error cuadrático sea mínimo:

$$J = \sum_{n=-\infty}^{\infty} (s\ n\ - s'\ n\ ) = \sum_{n=-\infty}^{\infty} (s\ n\ - \sum_{i=1}^{p} a_i \cdot s\ n-i\ )^2$$

Por tanto, si hacemos que las derivadas parciales sean 0, obtenemos un sistema de ecuaciones que permite obtener el conjunto de coeficientes del predictor :

$$\frac{\partial J}{\partial a_m} = 0 \qquad 1 \le \mathrm{m} \le \mathrm{p}$$

No es difícil deducir que, en notación matricial, el sistema de ecuaciones se puede escribir en función de la función de autocorrelación como :

$$\begin{vmatrix} R[0] & R[1] & R[2] & R[3] & ... & R[p-1] \\ R[1] & R[0] & R[1] & R[2] & ... & R[p-2] \\ R[2] & R[1] & R[0] & R[1] & ... & R[p-3] \\ ... & ... & ... & ... & ... & ... \\ R[p-1] & R[p-2] & R[p-3] & R[p-4] & ... & R[0] \end{vmatrix} * \begin{vmatrix} a_1 \\ a_2 \\ a_3 \\ ... \\ a_p \end{vmatrix} = \begin{vmatrix} R[1] \\ R[2] \\ R[3] \\ ... \\ R[p] \end{vmatrix}$$

A partir de los coeficientes LPC, se puede obtener la envolvente espectral de una señal mediante la fórmula siguiente, donde G es la ganancia del filtro LPC, coefs son los coefs. LPC y n_puntos es el número de puntos deseado para el espectro.

```
espectro=log(G*abs(freqz(1,[1 -coefs'],n_puntos)));
```

El error cometido al realizar la predicción de los puntos de una trama se puede obtener mediante :

```
error=filter([1 -coefs],1,signal);
```

4.2 Recursión de Levinson-Durbin

El análisis LPC parte del análisis de autocorrelación de la señal temporal. Los coeficientes de autocorrelación se calculan como ya sabemos :

$$r(m) = \sum_{i=0}^{N-1-m} x(i) \cdot x(i+m) \qquad m = 0,1,\ldots p$$

donde *p* es el número de coeficientes LPC que deseamos calcular. Como vemos, el análisis de autocorrelación acumula en r(0) la energía de la señal analizada.
Los coeficientes LPC se pueden calcular a partir de los coeficientes de autocorrelación mediante el método de Levinson-Durbin :

$$E_0 = r(0)$$

$$k_i = \frac{r(i) - \sum_{j=1}^{L-1} \alpha_j^{i-1} \cdot r(|i-j|)}{E_{i-1}} \qquad 1 \le i \le p$$

$$\alpha_i^i = k_i$$

$$\alpha_j^i = \alpha_i^{i-1} - k_i \cdot \alpha_{i-j}^{i-1}$$

$$E_i = (1 - k_i^2) \cdot E_{i-1}$$

El número de coeficientes determina la resolución con la que el análisis LPC va a representar la envolvente espectral de la señal. Un valor reducido implica poca resolución, pero un valor excesivo implica cierta distorsión debido a que no sólo se tiene en cuenta la envolvente espectral, sino la estructura fina del mismo.

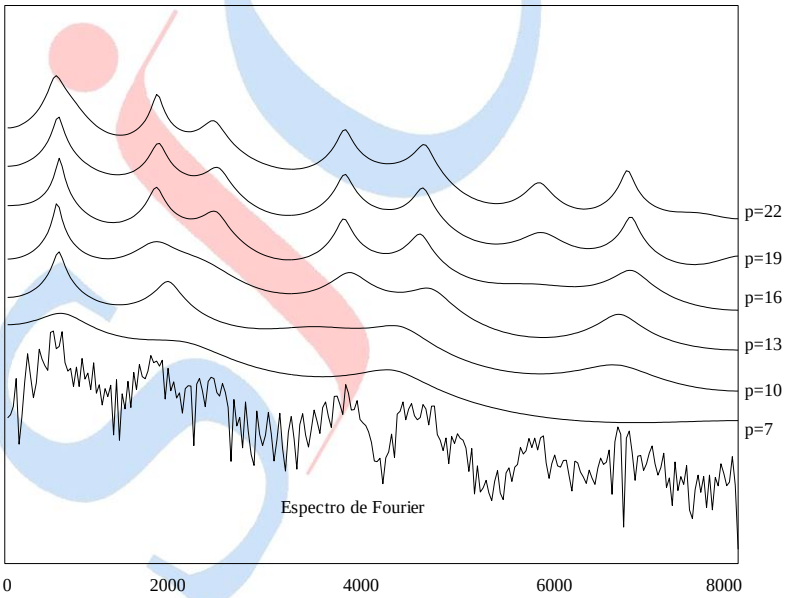


Figura 4.1. Variación del espectro LPC en función del número de coeficientes p.

En el ejemplo anterior, la frecuencia de mustreo fue de 16 KHz., y, como vemos. un número de coeficientes igual a 16 es suficiente para captar toda la información necesaria, ya que seleccionar un número mayor implica no sólo un incremento en tiempo de cálculo, sino una inserción innecesaria de formantes "falsos", ya que la envolvente se adapta en exceso a la señal.

4.3 Coeficientes CEPSTRUM

Los coeficientes cepstrales son una representación de la transformada de Fourier del logaritmo de la magnitud del espectro de la señal. Se ha demostrado que forman un conjunto de parámetros más robustos y fiables que los coeficientes de predicción lineal.
El cesptrum se calcula como la transformada inversa de Fourier del logaritmo de la magnitud de la transformada de Fourier de la secuencia de entrada. De forma gráfica podemos indicarlo así :

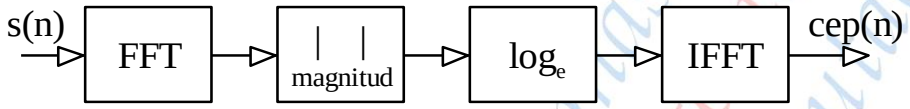


Figura 4.2. Implementación práctica para obtener el cepstrum

La función cepstrum es útil para calcular el pitch, ya que en la correspondiente posición alcanza un máximo local, que suele ser global para el rango de pitch posibles.
Además, pueden ser derivados de forma directa de los coeficientes LPC mediante las siguientes fórmulas de recurrencia :

$$c_0 = \ln\sigma^2 \qquad \sigma = \text{Ganancia}$$

$$c_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} \cdot c_k \cdot a_{m-k} \qquad m \le p$$

$$c_m = \sum_{k=1}^{m-1} \frac{k}{m} \cdot c_k \cdot a_{m-k} \qquad m > p$$

donde *p* es el número de coeficientes LPC calculados, y *m* el número de coeficientes cepstrum. No existe un criterio definido para seleccionar el número de coeficientes cepstrum adecuado.

4.4 Coeficientes diferenciales

La mayoría de los sistemas de reconocimiento actuales utilizan valores que miden modificaciones espectrales a lo largo del tiempo. Son los llamados coeficientes diferenciales, o Δcoeficientes.
El cálculo de los Δcoeficientes se puede hacer simplemente restando cada vector de coeficientes del anterior.

$$\Delta c_j(t) = c_j(t) - c_j(t-1)$$

Sin embargo, esto hace que los valores Δ obtenidos dependan exclusivamente de un segmento pequeño de voz, lo cual puede llevar a grandes errores. Por tanto, lo que se suele realizar es un cálculo de coeficientes de regresión que realiza una aproximación sobre varias tramas de la pendiente de variación de cada uno de los coeficientes. Así, cada coeficiente se calcula como :

$$\Delta c_i(t) = \frac{\sum_{n=-k}^{k} n \cdot c_i(t+n)}{\sum_{n=-k}^{k} n^2}$$

Normalmente k tiene un valor pequeño, que suele ser de 1 o de 2, totalizando 3 o 5 muestras en cada cálculo respectivamente.

4.5 Sistema LPC de Extracción de Características

Los pasos básicos en un sistema básico de extracción de carcaterísticas mediante la técnica LPC son :

1. Preénfasis : la señal se preenfatiza con mediante la ecuación en diferencias siguiente :

$$s'(n) = s(n) - a \cdot s(n-1)$$

Los valores de *a* suelen estar comprendidos en el rango 0.90 a 0.98.

2. Ruptura en tramas : la señal se divide en tramas de longitud N muestras. Este proceso se realiza cada M muestras, siendo N un múltiplo entero de M, normalmente 2, 3 ó 4.

3. Enventanado : el siguiente paso consiste en multiplicar la trama por una ventana de N puntos. Por sus características espectrales, la ventana más habitual es la ventana de Hamming, que viene dada por la fórmula :

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

4. Autocorrelación : se calculan los términos de la autocorrelación (desde 0 hasta *p*) de la trama enventanada, siendo *p* el número de coeficientes del predictor LPC que queremos utilizar.

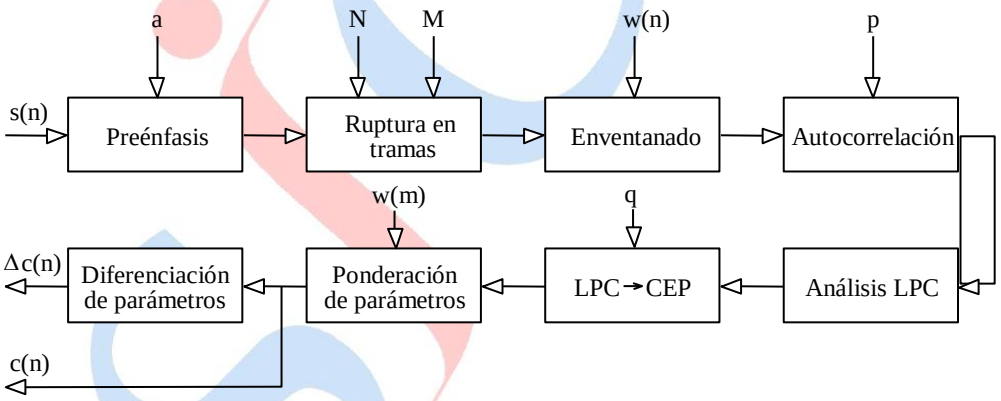


Figura 4.3. Sistema de Extracción de Características LPC

5. Análisis LPC : se calculan mediante algún método los *p* coeficientes del predictor LPC. El procedimiento normalmente utilizado es el método de Levinson-Durbin.

6. Conversión LPC a CEP : se convierten los *p* coeficientes LPC a *q* coeficientes Cepstrum, que son más adecuados para el reconocimiento robusto de la voz. Se utiliza el procedimiento recursivo mencionado en apartados anteriores.

7. Ponderación de parámetros : dado que la varianza de los diferentes coeficientes es diferente, así como su comportamiento frente a diferentes circustancias (diferentes locutores, diferentes frecuencias, etc.) se realiza un enventanado de los coeficientes cepstrum calculados. La ventana más utilizada consiste en un seno remontado, según la fórmula :

$$w(m) = 1 + \frac{q}{2} sin\left(\frac{\pi \cdot m}{q}\right) \qquad 1 \le m \le q$$

8. Diferenciación de parámetros : se calculan los coeficientes diferenciales, nrmalmente mediante el cálculo de los coeficientes de regresión sobre 3 ó 5 ventanas.

4.6 Modelo LPC del Aparato Fonador Humano

Mediante una modelización LPC del Aparato Fonador Humano, podemos realizar un sencillo sistema de síntesis de voz, siguiendo el siguiente esquema :

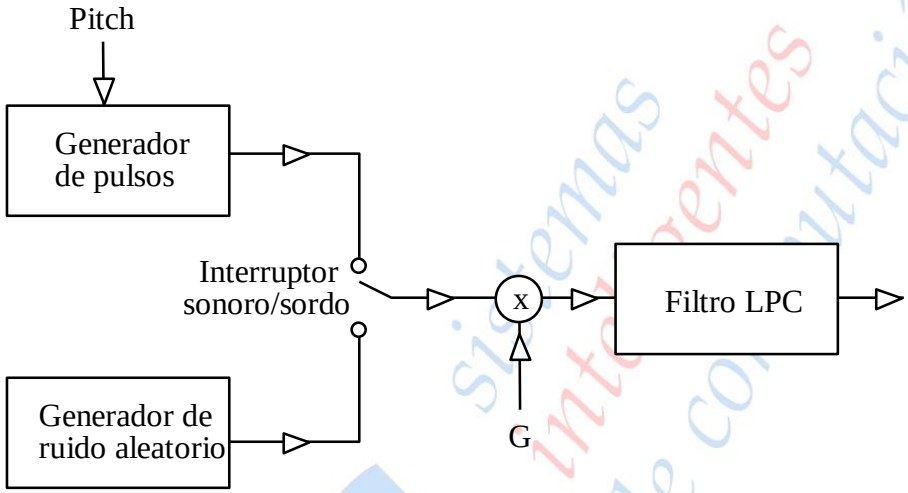


Figura 4.4. Modelo de síntesis de voz basado en LPC

En este modelo, el interruptor sonoro/sordo decide el tipo de fuente de sonido, bien un tren de pulsos a una frecuencia igual al pitch de la señal, bien una señal de ruido para sonidos sordos. El factor de ganancia G multiplica dicha señal, y, finalmente, el Filtro LPC, añade las componentes de la envolvente espectral.

5 Programación Dinámica

5.1 Concepto

El método más extendido de comparación de patrones, y dada la naturaleza eminentemente cambiante de la voz, es el de programación dinámica (*Dynamic Time Warping-DTW)*. La idea subyacente es muy simple. Se almacenan una serie de patrones prototipo de cada unidad a reconocer. El reconocimiento se realiza comparando un nuevo patrón con cada uno de los vectores prototipo, y seleccionando aquel que difiera en menor medida, según una cierta definición de distancia. Existen muchas variaciones sobre el algoritmo básico de Programación Dinámica en función de utilizar diferentes medidas de distorsión, caminos posibles y procedimientos de búsqueda.

5.2 Algoritmo Dynamic Time Warping (DTW)

Cuando una persona habla, la duración de cada sonido, de cada palabra, de cada frase es muy variable. Este hecho se acentúa aún más cuando se trata de diferentes locutores. Pese a que se trate de la misma persona pronunciando la misma palabra, se realiza una expansión compresión no lineal en el tiempo que dificulta la comparación de patrones almacenados, y de patrones nuevos.

Sea T una secuencia de vectores obtenidos del proceso de extracción de características :

T(x) = { T[1] , T[2] , T[3] , ... T[n] }

Sea R una secuencia de vectores almacenadas correspondientes a una extracción de características realizadas con anterioridad :

R(y) = { R[1] , R[2] , R[3] , ... R[m] }

Por lo general, m es diferente de n, y es difícil establecer una comparación. Una posibilidad sería establecer una función de correspondencia lineal de la siguiente manera :

$$f(y) = (y-1)\frac{n-1}{m-1} + 1$$

y realizar la comparación entre R(y)) y T(f(y)), ya que, mediante la transformación lineal, sus longitudes coinciden Sin embargo, los cambios de velocidad al hablar no son lineales, por lo que es preciso un esquema más complejo. Supongamos un patrón a reconocer de una longitud de 10, y un patrón almacenado de una longitud de 8. La función de correspondencia podría ser como la que se muestra en la figura 5.1:

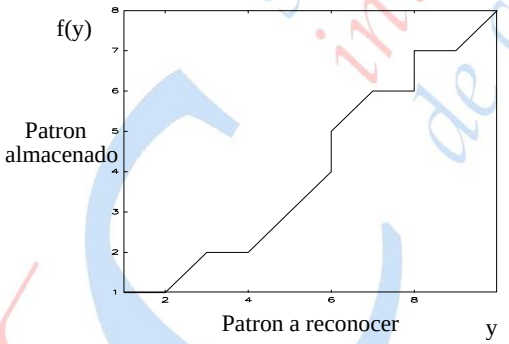


Figura 5.1. Ejemplo de función de correspondencia

El criterio más extendido consiste en definir la función de correspondencia como aquella que minimice la distorsión total, calculada como la suma de todas las distorsiones parciales

$$D_{total} = \sum d(T(y), R(f(y)))$$

Por tanto, calcularíamos la distorsión total para todas las funciones de correspondencia posibles y elegiríamos la de menor distorsión.

$$D_{optima} = min(D_{total}) \qquad \forall\ posible\ camino$$

Los posibles caminos posibles se denominan recinto de búsqueda, y se configuran como una cuadrícula :

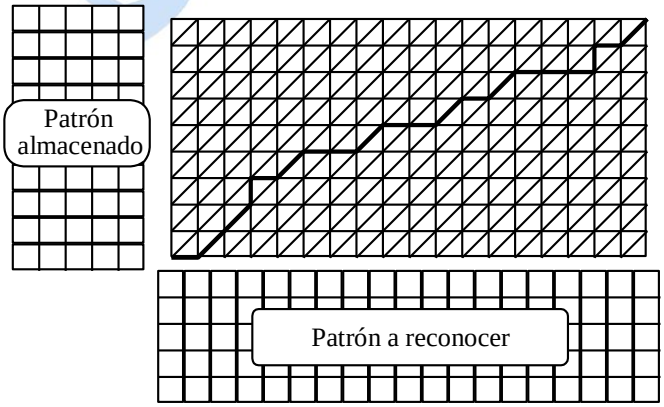
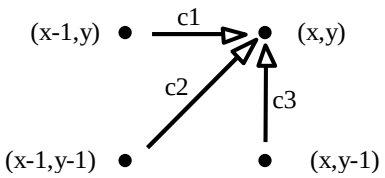


Figura 5.2. Cuadrícula de todos los caminos posibles en un cierto recinto de búsqueda

Sin embargo, el coste computacional sería prohibitivo. en su lugar se utiliza el algoritmo siguiente. Sea D(n,m) la distorsión acumulada hasta el punto (n,m). El camino óptimo también lo será localmente ya que todas las distorsiones son positivas. Por tanto, existen tres alternativas para llegar al punto (n,m), a cada una de las cuales se le asigna un coste. De esta forma, y recursivamente, podemos calcular :



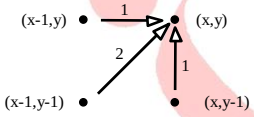
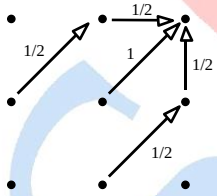
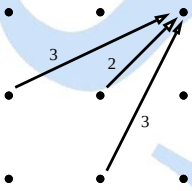
$$D(1,1) = d(T[1],R[1])$$
$$D(x,y) = \min(D(x-1,y) + c1 * d(T[x],R[y]) ,$$
$$D(x-1,y-1) + c2 * d(T[x],R[y]) ,$$
$$D(x,y-1) + c3 * d(T[x],R[y]))$$

Figura 5.3. Ejemplo de accesibilidad de un punto

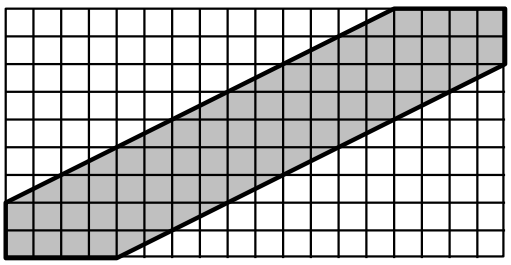
El camino recorrido se irá acumulando en cada nodo, para saber desde qué nodo se llegó al punto (x,y).
Así, una vez calculado D(n,m) se recorrerá hacia atrás para determinar el camino óptimo.

5.3 Alternativas de Diseño

A) Utilización de diferentes fórmulas de acceso a un nodo :

Accesibilidad	*Fórmulas : min(.....)*
	D(x-1,y)+d(x,y) D(x-1,y-1)+2d(x,y) D(x,y-1)+d(x,y)
	D(x-2,y)+1/2*(d(x-1,y)+d(x,y)) D(x-1,y-1)+d(x,y) D(x-1,y-2)+1/2*(d(x,y-1)+d(x,y))
	D(x-2,y-1)+3*d(x,y) D(x-1,y-1)+2*d(x,y) D(x-1,y-2)+3*d(x,y)

B) Limitación del Recinto de Búsqueda : se realiza para reducir el coste computacional.



C) Relajación de Extremos : se aplica si el algoritmo de detección de extremos no es muy fiable. En este caso, las primeras o últimas tramas de los patrones pueden ser ignoradas. Se suele acompañar de una limitación del recinto de búsqueda.
D) Normalización de las duraciones : se aplica cuando es previsible que las duraciones de los patrones a reconocer, y los patrones de test (o almacenados) sean de duración muy diferente.

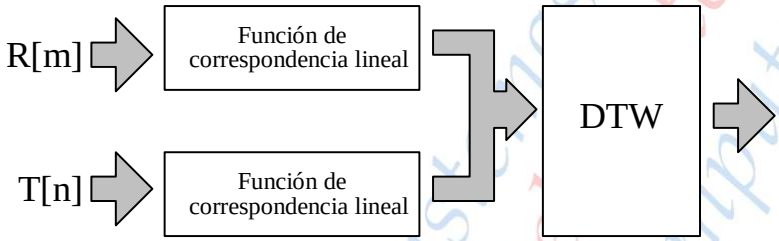


Figura 5.4. Diagrama de un sistema de normalización de duraciones

5.4 Esquema Básico de un Sistema basado en DTW

Un diagrama en bloques de un Sistema de Reconocimiento de Voz basado en la clasificación de Patrones, como es el sistema DTW se puede resumir en cuatro módulos :

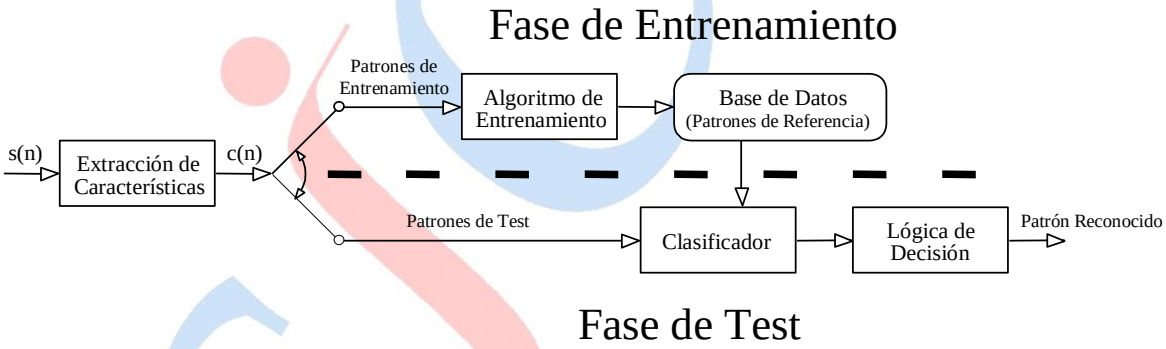


Figura 5.5. Esquema de un Sistema de Clasificación de Patrones

A) Extracción de Características : este módulo convierte la señal de entrada en un vector de carcaterísticas, por medio de un análisis espectral. Los sistemas de análisis más utilizados, son, como ya sabemos, Banco de Filtros, Análisis LPC y FFT.
B) Algoritmo de Entrenamiento : este módulo obtiene a partir de una serie de patrones de entrenamiento, uno o varios patrones representativos de los rasgos más característicos de cada clase. El elemento resultante puede ser un ejemplar de los patrones de entrada, un ejemplar obtenido por algún método de promediado, o bien un modelo matemático que caracteriza a los patrones de entrenamiento.
C) Clasificación de Patrones : el patrón de test en comparado con cada patrón de referencia y se calcula una medida de similaridad para cada uno de ellos. Se precisan dos medidas de distancia, una local, en que se define la distancia espectral entre dos tramas diferentes, y una global, normalmente a través de algún procedimiento de alineación, como DTW.
D) Lógica de Decisión : los valores de similaridad obtenidos en la clasificación de patrones se utilizan para decidir si el patrón de referencia es considerado como el patrón reconocido, o por el contrario se da un error de reconocimiento.

5.5 DTW aplicado a Palabras Concatenadas

Se han realizado numerosos estudios para mejorar la técnica original y adaptarla a voz contínua. Las técnicas de Programación Dinámica adaptada al reconocimiento de palabras conectadas se pueden resumir en tres :

- Programación Dinámica en dos niveles (*Two-Level DP*)
- Construcción de niveles (*Level Building*)
- Construcción de niveles síncrono por tramas. (*Frame-Synchronous Level Building*)

La primera técnica consiste en un procedimiento síncrono por tramas que realiza un alineamiento global de todas las posibles concatenaciones de palabras del diccionario, seleccionando aquella cuya distancia a la señal de entrada es menor. Dicha técnica posee interesantes propiedades, como escasos requerimientos computacionales, y que el cálculo, tanto de la distorsión temporal como de la secuencia óptima de palabras es desarrollada de forma síncrona, lo que permite realizar con sencillez sistemas que trabajen en tiempo real. Sin embargo, la adaptación de una gramática que limite de alguna manera todas las posibles combinaciones de palabras es de una gran complejidad.

La segunda técnica, denominada de construcción de niveles (*Level Building*), soluciona el problema de la gramática. Permite integrar una red gramatical expresada en forma de autómata formal en el algoritmo de construcción de niveles de una forma simple y eficiente.

La técnica de construcción de niveles síncrono por tramas (*Frame-Synchronous Level Building*), es, con gran diferencia, la más avanzada y compleja de las citadas, ya que incluye las ventajas de ambas. Por un lado, el procesamiento se realiza en una pasada, es decir síncrono por tramas, y por otro lado admite de forma sencilla la integración de gramáticas en el algoritmo. Esto hace que algunos de los sistemas más eficientes de Reconocimiento de Voz sean programados con la técnica descrita anteriormente.

La complejidad computacional de la programación dinámica hace que si bien es aceptable para reconocimiento dependiente del locutor en vocabularios reducidos, es impracticable si lo que pretendemos es reconocimiento de voz contínua independiente del locutor.

6 Modelos Ocultos de Markov

6.1 Introducción

Los Modelos Ocultos de Markov permiten modelizar de forma sencilla sistemas muy complejos, de ahí su interés. En concreto, y para su utilización, se considera la voz formada por un conjunto finito de subunidades, como pueden ser palabras, fonemas, etc., que, adecuadamente modelizadas, permiten reconocer correctamente una señal de voz compuesta de muchas de estas subunidades. Los resultados que se obtienen, son, como ya he dicho, los mejores conocidos hasta el momento. Aunque se están realizando investigaciones para mejorar aún más los resultados, sobre todo en el campo de las redes neuronales, todavía no se ha obtenido un sistema que dé mejores resultados.

6.2 Procesos, Cadenas y Fuentes de Markov

Un **proceso de Markov** se define como un proceso estocástico x(t) tal que para todo *n* y para los valores de tiempo :

$$t_1 < t_2 < \cdots < t_n$$

se cumple que:

$$P\ x(t_n) \le x_n \left| x(t_{n-1}),\ldots,x(t_1) \right. = P\ x(t_n) \le x_n \left| x(t_{n-1}) \right.$$

Es decir, que la densidad de probabilidad condicional de la variable estocástica X en el instante de tiempo t_n, depende únicamente del valor de la variable X en el instante t_{n-1}.

Una **cadena de Markov** es un proceso de Markov en que tanto la variable tiempo, como la variable estocástica del proceso son de tipo discreto. Es decir, verifica las siguientes propiedades :

- El rango de la variable estocástica X es un conjunto discreto de estados.
- La variable tiempo es discreta, y toma valores enteros

  t = ..., -2, -1, 0, 1, 2, ...
- El proceso es estacionario, o al menos es homógeneo en el tiempo, de forma que las probabilidades de transición dependen únicamente de la longitud del intervalo de tiempo considerado, y no de los instantes de tiempo que lo delimitan.

Una **Fuente de Markov de Primer Orden** es una cadena de Markov que modeliza una fuente de información, en que el símbolo emitido por la fuente, depende únicamente del símbolo emitido en el instante anterior. De esta manera, y suponiendo que el último símbolo emitido constituye el estado de la fuente, tenemos que el alfabeto de la fuente coincide con el conjunto de estados $S = \{ s_i \}_{i=1,\ldots n}$ del modelo.

En este último caso se definen las probabilidades de transición de la siguiente manera :

$$a_{ij} = P\ x_t = s_j \left| x_{t-1} = s_i \right. \qquad 1 \le i, j \le n$$

$$\sum_{j=1}^{n} a_{ij} = 1 \qquad 1 \le i \le n$$

De manera análoga se pueden definir Fuentes de Markov n-simo, donde el símbolo emitido depende únicamente de los símbolos emitidos en los n instantes anteriores, es decir, de los n estados anteriores.

En la figura 1 podemos observar un ejemplo de una fuente de Markov de primer orden.

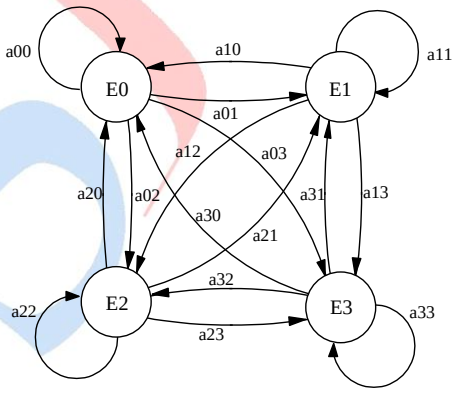


Figura 6.1. Fuente de Markov cuaternaria de primer orden.

Como vemos, aparecen cuatro estados, correspondientes a los cuatro símbolos observables, y cada flecha se corresponde con una posible transición.

6.3 Definición de Modelo Oculto de Markov

Un **Modelo Oculto de Markov** se define [RABI, 89] como una cadena de Markov doblemente estocástica, con un proceso subyacente, que no es directamente observable, y otro que sí lo es.

Podemos simplficar el concepto de la siguiente forma :

- El modelo se compone de un conjunto de N estados.
- En cada instante de tiempo se producen dos acciones:
  - — Se genera una observación, que únicamente depende del estado en que se encuentra el sistema. Nosotros consideraremos que el número de observaciones posibles es finito. (Modelos Ocultos de Markov Discretos). Una alternativa posible es la generación de una observación, pero en lugar de depender del estado del sistema, depende de la transición realizada.
  - — Se realiza un cambio de estado, que únicamente depende del estado en que se encuentra el sistema.
- Las observaciones constituyen el primer proceso estocástico, que sí es observable desde el exterior.
- La sucesión de estados constituyen el proceso estocástico oculto, ya que no es observable desde el exterior.

De lo anterior, se deduce que hemos de definir :

- Una función de probabilidades de transición entre estados. Esta función sólo depende del estado en que se encuentra el proceso en un momento dado, por tanto hemos de definir la matriz de valores

  $A = a_{ij}$

  a_{ij} = Probabilidad de transitar del estado i al estado j

  $1 \le i, j \le n$
- Una función de probabilidades de observación de cada símbolo en cada estado. Esta función sólo depende del estado en que se encuentra el proceso en un momento dado, por tanto hemos de definir la matriz de valores

  $B = b_{jk}$

  b_{jk} = Probabilidad de observar el símbolo S_k en el estado j

  $1 \le j \le n$, siendo n = numero de estados del modelo.

  $1 \le k \le K$, siendo K = numero de observaciones posibles.
- La distribución de probabilidad de estados iniciales, que nos permite conocer la probabilidad de que el estado j sea el estado inicial, y que llamaremos Π, donde :

  $$\Pi = \pi_{i \quad i=1..N}$$

  Denominaremos S_I al conjunto de estados iniciales, y N_I al número de elementos de dicho conjunto.

Por tanto, en general, un Modelo Oculto de Markov, (lo notaremos como λ), se define mediante:

$$\lambda = (A, B, \Pi)$$

donde los valores A, B y Π se corresponden con :

A = Matriz de Probabilidades de Transición

B = Matriz de Probabilidades de Observación

Π = Matriz de Probbailidades de Estado Inicial

Como vemos, un HMM puede ser considerado como una máquina de estados finitos, donde las transiciones entre estados dependen de la observación de un símbolo. Por tanto, asociado a cada estado tenemos una distribución de probabilidad de observación de los diferentes símbolos, y asociada a cada transición una probabilidad de transición, que indica la probabilidad de dicha transición.

Veamos un ejemplo de un Modelo Oculto de Markov en detalle :

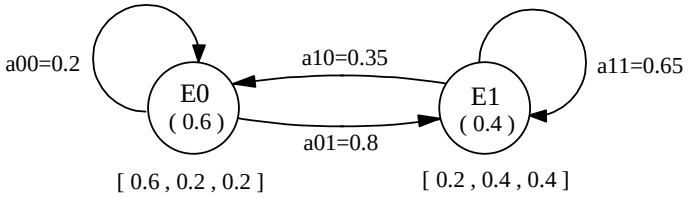


Figura 6.2. Modelo Oculto de Markov (Ejemplo)

Sea el modelo representado en la figura, y supongamos que :

- El modelo consta de **dos estados**, E0 y E1, representados como círculos.
- Suponemos que solo existen **tres posibles observaciones** (Por ejemplo, "a", "b", y "c"). Entre corchetes aparece la probabilidad de observar "a", "b", y "c" en cada estado.
  Como vemos, la suma de las probabilidades de observación es igual a 1. Es decir, si existe una observación, forzosamente corresponde a un conjunto finito llamado **vocabulario de observación**.

  Existe la posibilidad de que una de dichas probabilidades sea 0. Ello indicaría que desde dicho estado es imposible generar una cierta observación.
- Junto a cada arco aparece la probabilidad de transitar del estado i al estado j. (aij)
  De igual forma, una probabilidad de transición igual a 0 indica la imposibilidad de transitar de un estado a otro.
  La suma de todas las probabilidades de transición de un estado i a todos los estados j es 1.
- Dentro de cada estado aparece entre paréntesis la probabilidad de estado inicial. (π i).
  De igual forma, una probabilidad de estado inicial igual a 0 indica la imposibilidad de que un estado sea el inicial .
  La suma de todas las probabilidades de estado inicial es 1.

Según los supuestos anteriores, podemos calcular, por ejemplo:

A) Dada una secuencia de observaciones, por ejemplo 'a', ¿ Cual es la probabilidad de que dicho modelo haya generado dicha secuencia ?

Podremos observar 'a', si el primer estado es el 1, y se observa 'a', o bien si el primer estado es el '2', y se observa 'a'.

Por tanto,

$$P(O=' a') = \pi_1 * b_{1a} + \pi_2 * b_{2a} = 0.6*0.6 + 0.4*0.2 = 0.36 + 0.08 = 0.44$$

De la misma forma, la probabilidad de que el modelo genere una cadena de longitud 1, igual a 'b' es :

$$P(O=' b') = \pi_1 * b_{1b} + \pi_2 * b_{2b} = 0.6*0.2 + 0.4*0.4 = 0.12 + 0.16 = 0.28$$

Y de igual forma,

$$P(O=' c') = \pi_1 * b_{1c} + \pi_2 * b_{2c} = 0.6*0.2 + 0.4*0.4 = 0.12 + 0.16 = 0.28$$

Si la observación que tenemos del sistema es de longitud 2, la cosa se complica un poco. Por ejemplo,

$$P(O='ab') = \pi_1 b_{1a} a_{11} b_{1b} + \pi_1 b_{1a} a_{12} b_{2b} + \pi_2 b_{2a} a_{22} b_{2b} + \pi_2 b_{2a} a_{21} b_{1b} =$$
$$= 0.6 \times 0.6 \times 0.2 \times 0.2 + 0.6 \times 0.6 \times 0.8 \times 0.4 + 0.4 \times 0.2 \times 0.65 \times 0.4 + 0.4 \times 0.2 \times 0.35 \times 0.2 =$$
$$= 0.0144 + 0.1152 + 0.0208 + 0.0056 = 0.156$$

De esta forma, dada una cadena, podemos calcular la probabilidad de que el modelo la haya generado.

B) ¿ Cuál es la cadena de longitud dada más probable ?
De forma similar podemos calcular cuál es la cadena de una cierta longitud dada más probable. En nuestro ejemplo, la cadena de longitud 1 más probable es la cadena 'a', mientras que la probabilidad de observar las cadenas 'b' y 'c' son menores.
Por tanto, para obtener la cadena de longidud dada más probable, basta con generar todas las posibles cadenas de dicha longitud, calcular su probabilidad, y seleccionar la mayor.

C) Dada una secuencia de observaciones, y un modelo, ¿ Cual es la secuencia de estados más probable ?
Como se puede observar en el ejemplo de la cadena 'ab', de entre todas las posibles secuencias de estados (11 , 12 , 21 , 22), la más probable es (12), ya que su probabilidad de generación es 0.1152, superior a todas las demás.
Por tanto, para resolver dicho problema, basta con realizar todos los posibles recorridos de estados, y seleccionar aquel que genere una probabilidad mayor.
D) ¿ Cómo generar una secuencia de observaciones que se ajuste a un determinado modelo ?
Un Modelo Oculto de Markov puede ser utilizado como un generador de símbolos que siguen una cierta estructura de la siguiente manera:

1. Elegir un estado inicial en función de las probabilidades de estado inicial.
2. Elegir una observación de acuerdo a las probabilidades de observacion.
3. Transitar del estado actual a otro estado (incluído el actual) en función de las probabilidades de transición.
4. Si la cadena de símbolos generada es de la longitud deseada, parar. En caso contrario, volver al paso 2.

6.4 Problemas Básicos

En el capítulo anterior hemos resaltado la gran flexibilidad y potencia que los Modelos Ocultos de Markov presentan, ya que permiten modelizar sistemas de gran complejidad de forma muy sencilla.
Ahora bien, para que la modelización sea útil, hemos de buscar métodos generales que permitan realizar de forma automática lo que de forma intuítiva hemos visto en el capítulo anterior. Básicamente, estos métodos se refieren a los siguientes problemas :
Problema 1. (Problema de Evaluación)
Dada una secuencia de observaciones, y un modelo, ¿ Cuál es la probabilidad de que dicho modelo haya generado dicha secuencia ? La resolución a este problema nos permitirá conocer, dada una determinada secuencia de observaciones, y una serie de modelos diferentes, qué modelo es más probable que haya generado dicha cadena.
Problema 2. (Problema de Decodificación)
Dada una secuencia de observaciones, y un modelo, ¿ Cuál es la secuencia óptima de estados que mejor explica (o sea, la que maximiza la probabilidad de observación) dicha secuencia de observaciones ? La resolución a este problema nos permitirá conocer qué secuencia de estados es la óptima para una cierta cadena, intuyendo de alguna

manera el proceso "oculto" que se desarrolla en el interior del modelo, y que no es observable de forma directa desde el exterior.

Problema 3. (Problema de Entrenamiento)

Dada una secuencia de observaciones, y un modelo, ¿ Cómo podemos ajustar los parámetros del modelo para que maximizar la probabilidad de generar dicha secuencia ? Este ajuste nos permitirá realizar el entrenamiento de un modelo, es decir, su adaptación a un conjunto de secuencias de observaciones correspondientes a un mismo fenómeno, generando así su modelo correspondiente.

Una vez resueltos estos tres problemas, estaremos en condiciones de aplicar los Modelos Ocultos de Markov a problemas concretos.

6.5 Solución al Problema de Evaluación

La resolución general a este primer problema recibe el nombre de **Procedimiento forward-backward.**

6.5.1 Cálculo de las probabilidades "forward"

Consideremos la variable $\alpha_t(i)$, llamada probabilidad "forward" definida como:

$$\alpha_t(i) = P(O_1 O_2 \ldots O_t, q_t = i|\lambda)$$

Es decir, dado el modelo λ, y habiéndose realizado el conjunto de observaciones $O_1 O_2 \ldots O_t$, cuál es la probabilidad de encontrarnos en el estado i en el instante t.

Sean S_I el conjunto de estados iniciales, N_I el número de elementos de dicho conjunto, S_F al conjunto de estados finales, y N_F al número de elementos de dicho conjunto. Para calcular $\alpha_t(i)$, basta realizar el algoritmo inductivo siguiente :

1. Inicialización.

$$\alpha_1(i) = \pi_i b_i(O_1) \qquad 1 \le i \le N$$

Es decir, la probabilidad de que el estado inicial sea i, multiplicado por la probabilidad de observar O_1 en el estado i. Los valores de π_i son :

Si $i \in S_I$, entonces $\pi_i = \dfrac{1}{N_I}$

Si $i \notin S_I$, entonces $\pi_i = 0$

2. Inducción.

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{N} \alpha_t(i) \cdot a_{ij} \right) \cdot b_j(O_{t+1})$$
$$1 \le j \le N$$
$$1 \le t \le T-1$$

Es decir, la probabilidad de observar el símbolo correspondiente a (t+1), multiplicado por **la suma** de todos los posibles caminos de longitud (t), ponderados con los correspondientes a_{ij}.

3. Terminación.

$$\Pr(O|\lambda) = \Pr(O_1 O_2 \ldots O_T|\lambda) = \sum_{i \in S_F} \alpha_T(i)$$

Apliquemos el algoritmo al modelo que ya hemos utilizado anteriormente como ejemplo:

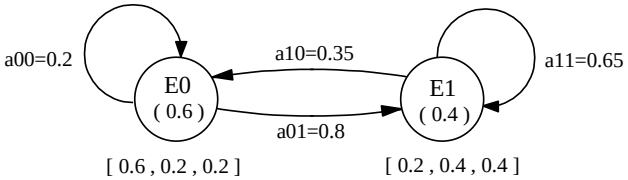


Figura 6.3. Modelo Oculto de Markov (Ejemplo)

Sea la secuencia de símbolos " a b ". ¿ Cuál es la probabilidad de que este modelo haya generado esta secuencia ?. Siguiendo los pasos del algoritmo forward-backward, obtenemos:

1. Inicialización.

$\alpha_1(1) = \pi_1 b_1(O_1) = 0.6 \times 0.6 = 0.36$
$\alpha_1(2) = \pi_2 b_2(O_1) = 0.4 \times 0.2 = 0.08$

2. Inducción.

$$\alpha_2(1) = \left\langle \sum_{i=1}^{N} \alpha_1(i) \cdot a_{i1} \right\rangle \cdot b_1(O_2) = \ 0.36 \times 0.2 + 0.08 \times 0.35 \ \times 0.2 = 0.020 \qquad (j = 1, t = 1)$$

$$\alpha_2(2) = \left\langle \sum_{i=1}^{N} \alpha_1(i) \cdot a_{i2} \right\rangle \cdot b_2(O_2) = \ 0.36 \times 0.8 + 0.08 \times 0.65 \ \times 0.4 = 0.136 \qquad (j = 2, t = 1)$$

3. Terminación.

$$P(ab|\lambda) = \sum_{i=1}^{N} \alpha_T(i) = 0.020 + 0.136 = 0.156$$

Como vemos, el resultado coincide con el que calculamos de forma intuitiva.

6.5.2 Cálculo de las probabilidades "backward"

Mediante un criterio similar, podemos realizar el proceso inverso. Consideremos la variable $\beta_t(i)$, llamada probabilidad "forward" definida como:

$$\beta_t(i) = P(O_{t+1} O_{t+2} \ldots O_T, q_t = i|\lambda)$$

Es decir, dado el modelo λ, y habiéndose realizado el conjunto de observaciones $O_{t+1} O_{t+2} \ldots O_T$, cuál es la probabilidad de encontrarnos en el estado i en el instante t.

Para calcular $\beta_t(i)$, basta realizar el algoritmo inductivo siguiente :

1. Inicialización.

Se define de forma arbitraria la probabilidad de estar en el estado i-esimo en el instante final t=T como $\beta_T(i) = \frac{1}{N_F}$ para todos los estados finales. Es decir,

$$\beta_T(i) = \frac{1}{N_F} \qquad \forall i \in S_F$$

2. Inducción.

$$\beta_t(j) = \sum_{i=1}^{N} a_{ji} \cdot b_i(O_{t+1}) \cdot \beta_{t+1}(i)$$
$$1 \le j \le N$$
$$1 \le t \le T-1$$

3. Terminación.

$$\Pr(O|\lambda) = P(O_1 O_2 \ldots O_T|\lambda) = \sum_{i \in S_I} \pi_i \cdot b_i(O_1) \cdot \beta_1(i)$$

Una propiedad interesante, y muy útil en el proceso de depuración de los programas correspondientes es que el valor $\Pr(O|\lambda)$ debe ser el mismo, tanto si se calcula a partir de las probabilidades forward, como de las probabilidades backward. Es decir,

$$\Pr(O|\lambda) = \Pr(O_1 O_2 \ldots O_T|\lambda) = \sum_{i \in S_F} \alpha_T(i) = \sum_{i \in S_I} \pi_i \cdot b_i(O_1) \cdot \beta_1(i)$$

6.6 Solución al Problema de Decodificación

Así como para el problema de evaluación hemos encontrado una solución exacta, para el problema de decodificación no existe una única solución, ya que hemos de definir el concepto de "camino óptimo".

No entraremos en este trabajo en estudiar los diferentes criterios existentes. En su lugar, estudiaremos el algoritmo más comúnmente utilizado, conocido con el nombre de **Algoritmo de Viterbi.**

6.6.1 Algoritmo de Viterbi

El algoritmo de Viterbi encuentra la mejor secuencia de estados dada una secuencia de observaciones. Para ello utiliza el conjunto de variables $\delta_t(i)$, definida de la siguiente manera:

$$\delta_t(i) = \max_{q_1, q_2, \ldots q_{t-1}} P(q_1 q_2 \ldots q_t = i, O_1 O_2 \ldots O_t|\lambda)$$

Es decir, $\delta_t(i)$ representa la probabilidad del mejor camino en el instante t, que termina en el estado i.

De esta manera, el algoritmo se define como :

1. Inicialización.

$$\delta_1(i) = \pi_i b_i(O_1) \qquad 1 \le i \le N$$

$$\Psi_1(i) = 0$$

Es decir, la probabilidad de que el estado inicial sea i , multiplicado por la probabilidad de observar O1 en el estado i.

2. Inducción.

$$\delta_{t+1}(j) = \max_{1 \le i \le N} \delta_t(i) \cdot a_{ij} \cdot b_j(O_{t+1})$$

$$\Psi_{t+1}(j) = \arg\max_{1 \le i \le N} \delta_t(i) \cdot a_{ij}$$

$$1 \le j \le N \quad , \quad 1 \le t \le T-1$$

Es decir, la probabilidad de observar el símbolo correspondiente a (t+1), multiplicado por **el maximo** de todos los posibles caminos de longitud (t), ponderados con los correspondientes aij.

3. Terminación.

$$P^* = \max \delta_T(i)$$

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

4. Calculo del mejor camino mediante la vuelta atrás.

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

Como vemos, el algoritmo de Viterbi es muy parecido al algoritmo forward-backward, con las diferencias siguientes:

- En el paso inductivo utiliza el máximo, en lugar de la suma.
- Almacena en Ψ el camino recorrido.
- Añade el último paso, para recuperar de forma iterativa el mejor camino.

A continuación, y siguiendo nuestro ejemplo, calcularemos la mejor secuencia de estados para nuestro modelo, y para la cadena "ab".

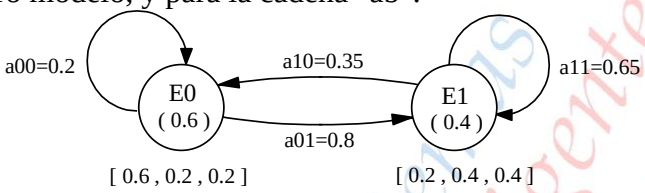


Figura 6.4. Modelo Oculto de Markov (Ejemplo)

1. Inicialización.

$\delta_1(1) = \pi_1 b_1(O_1) = 0.6 \times 0.6 = 0.36$

$\delta_1(2) = \pi_2 b_2(O_1) = 0.4 \times 0.2 = 0.08$

$\Psi_1(1) = \Psi_1(2) = 0$

2. Inducción

$$\delta_2(1) = \max_{1 \le i \le N} \alpha_1(i) \cdot a_{i1} \cdot b_1(O_2) = max\ 0.36 \times 0.2, 0.08 \times 0.35 \times 0.2 = 0.0144 \qquad (j = 1, t = 1)$$

$$\delta_2(2) = \max_{1 \le i \le N} \alpha_1(i) \cdot a_{i2} \cdot b_2(O_2) = max\ 0.36 \times 0.8, 0.08 \times 0.65 \times 0.4 = 0.1152 \qquad (j = 2, t = 1)$$

$\Psi_2(1) = 1$

$\Psi_2(2) = 1$

3. Terminación.

$P^* = max\ 0.0144, 0.1152 = 0.1152$

$q_2^* = 2$

4. Calculo del mejor camino mediante la vuelta atrás.

$q_1^* = \Psi_2(q_2^*) = 1$

Luego la secuencia de estados óptima es "E1,E2", resultando una probabilidad igual a 0.1152. Podemos observar que el valor de esta probabilidad no es exactamente igual al obtenido intuitivamente, sino algo inferior. Ello se debe a que el algoritmo de Viterbi utiliza el máximo en lugar de la suma en el paso inductivo.

6.7 Solución al Problema del Entrenamiento

El problema más difícil de resolver de los tres planteados es sin ninguna duda el del entrenamiento. Es decir, cómo ajustar los parámetros de un cierto modelo para maximizar la probabilidad de una secuencia de observaciónes.

No existe una solución analítica que resuelva dicho problema. Por tanto, es preciso aplicar algoritmos iterativos. Esto implica que las soluciones a las que se llegue pueden no constituir mínimos globales de la función que se intente optimizar. Como es lógico, es de gran importancia la determinación del criterio de optimización y ello determina la existencia de diferentes algoritmos :

- Baum-Welch (máxima verosimilitud de los modelos)

- Segmental K-means (máxima verosimilitud de la secuencia entre estados)
- Criterio MMIE (máxima información mutua)
- Entrenamiento Correctivo (mínima tasa de errores)

En concreto, vamos a estudiar el algoritmo de Baum-Welch por ser el más frecuentemente utilizado.

6.7.1 Algoritmo Baum-Welch

Consideremos un modelo λ, y la secuencia de observaciones $O = O_1 O_2 \dots O_T$. Utilizando el algoritmo forward-backward podemos calcular la probabilidad de estar en el estado i en el momento t, y estar en el estado j en el momento t+1, y que denominaremos $\xi_t(i,j)$

$$\xi_t(i,j) = \Pr(s_t = i, s_{t+1} = j | O, \lambda) = \frac{\alpha_t(i) \cdot a_{ij} \cdot b_j(O_{t+1}) \cdot \beta_{t+1}(j)}{\sum_{k \in S_F} \alpha_T(k)}$$

De igual forma podemos determinar la probabilidad de estar en el estado i en el instante t, y que denominaremos $\gamma_t(i)$

$$\gamma_t(i) = \Pr(s_t = i | O, \lambda) = \frac{\alpha_t(i) \cdot \beta_t(i)}{\sum_{k \in S_F} \alpha_T(k)}$$

Es fácil determinar que se cumple la igualdad siguiente :

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

Las fórmulas de reestimación de los parámetros del modelo pueden ser enunciadas de la siguiente manera :

a) Fórmula de reestimación de las probabilidades de transición :

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

b) Fórmula de reestimación de las probabilidades de observación :

$$\bar{b}_j(k) = \frac{\sum_{t \in O_t = v_k}^{T-1} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$$

c) Fórmula de reestimación de las probabilidades de estado inicial :

$$\bar{\pi}_i = \gamma_1(i)$$

Según las tres fórmulas de reestimación, se demuestra que el modelo así obtenido, que denominaremos $\bar{\lambda} = (\bar{\pi}, \bar{A}, \bar{B})$ cumple que :

$$\Pr(O|\bar{\lambda}) \geq \Pr(O|\lambda)$$

Por tanto, si iterativamente se aplican las fórmulas de reestimación, sustituyendo $\bar{\lambda}$ por λ, se garantiza que $\Pr(O|\lambda)$ puede ser mejorado hasta un cierto punto límite. Dicha propiedad de convergencia es de gran importancia, ya que nos asegura que en cada paso del algoritmo, la solución obtenida es mejor, o igual que la anterior.

Además, si cualquier subconjunto de parámetros (ya sea π, A o B) se fija, el resto de los parámetros pueden ser reeestimados utilizando las mismas fórmulas, y la desigualdad $\Pr(O|\bar{\lambda}) \geq \Pr(O|\lambda)$ sigue siendo válida. Dicha propiedad tiene una gran importancia

práctica, ya que permite detectar posible incorrecciones en la implementación del algoritmo.

7 Sistemas de Reconocimiento

7.1 Fases en la realización de un reconocedor

En la construcción de un Sistema de Reconocimiento podemos distinguir dos fases bien diferenciadas :

1. **Fase de Entrenamiento** : en esta fase se obtiene, a partir de una base de datos apropiada, un modelo de cada palabra (fonema, sílaba, etc.) a reconocer. La fase de entrenamiento consta de los siguientes procesos :
    - Obtención de la Base de Datos.
    - Digitalización.
    - Extracción de Características.
    - Cuantificación Vectorial.
    - Entrenamiento (Obtención de los modelos de cada palabra).

2. **Fase de Reconocimiento** : esta es la fase que permite realizar el reconocimiento automático de un segmento de voz correspondiente a una de las fases aprendidas. La fase de Reconocimiento consta de los siguientes procesos :
    - Obtención de una señal (ya sea del micrófono, de un fichero, etc.)
    - Digitalización.
    - Extracción de Características.
    - Cuantificación Vectorial.
    - Reconocimiento.

En general, podemos decir que todo sistema de reconocimiento consta de los elementos que se muestran en la figura 7.1.

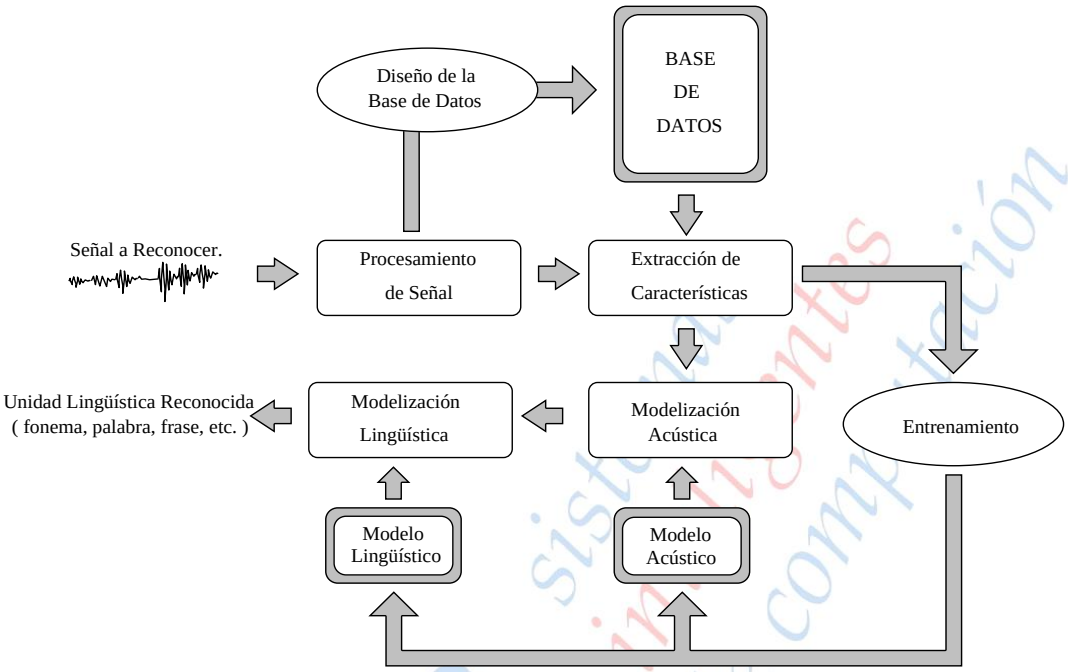


Figura 7.1 Esquema General de un Sistema de Reconocimiento de Voz.

7.2 Extracción de características

7.2.1 Análisis Espectral

Las señales de voz suelen ser analizadas utilizando rasgos espectrales, y no directamente la señal temporal. Ello se debe principalmente a que los rasgos más útiles para el análisis se encuentran en el dominio de la frecuencia. Además, se ha podido comprobar que el mecanismo auditivo humano presta mayor atención a aspectos espectrales de la señal, que a aspectos temporales.

Las tres técnicas más utilizadas para realizar el análisis espectral de una señal son las siguientes :

- **Banco de filtros :** utiliza un conjunto de filtros paso banda, cada uno de los cuales analiza un rango de frecuencias. Es un sistema que permite el análisis en tiempo real, es simple y barato. Sin embargo, el correcto diseño del mismo no es de modo alguno evidente.
- **Análisis de Fourier :** utiliza principalmente la transformada de Fourier para descomponer la señal temporal en suma de una serie de componentes sinusoidales a determinadas frecuencias. Este sistema, por sus grandes ventajas ha sido utilizando durante mucho tiempo como sistema básico de análisis de la señal.
- **Análisis LPC (Linear Predictive Coding) :** utiliza los principios de la predicción lineal, y tiene las grandes ventajas de la rapidez de cálculo, versatilidad, existencia de algoritmos simples y eficaces. Por todo ello, actualmente es el sistema de análisis espectral más extendido.

7.2.2 Análisis LPC

El análisis LPC parte del análisis de autocorrelación de la señal temporal. Los coeficientes LPC se calculan a partir de los coeficientes de autocorrelación mediante el conocido método de Levinson-Durbin.

El número de coeficientes determina la resolución con la que el análisis LPC va a representar la envolvente espectral de la señal. Un valor reducido implica poca resolución, pero un valor excesivo implica cierta distorsión debido a que no sólo se tiene en cuenta la envolvente espectral, sino la estructura fina del mismo.

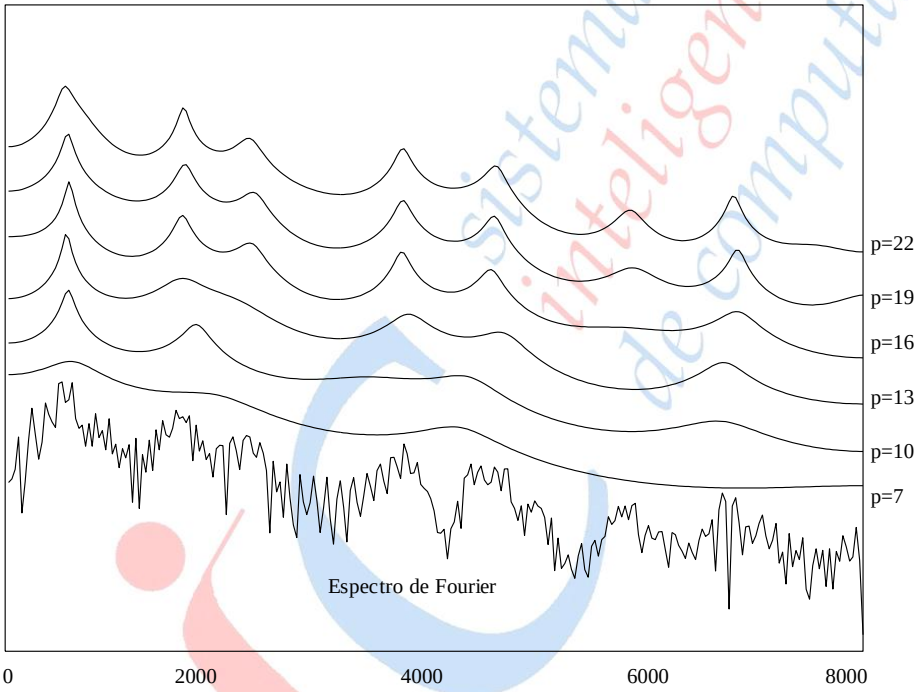


Figura 7.2 Variación del Espectro LPC en función del número de coeficientes p.

Como vemos, a partir de los 13 coeficientes, el espectro de la señal, básicamente queda igual, salvo por la estructura fina. Por ello, nuestra selección inicial en cuanto a número de coeficientes es p=14.

7.2.3 Coeficientes Cepstrum

Los coeficientes cepstrales son una representación de la transformada de Fourier del logaritmo de la magnitud del espectro de la señal. Se ha demostrado que forman un conjunto de parámetros más robustos y fiables que los coeficientes de predicción lineal. Además, pueden ser derivados de forma directa de los coeficientes LPC.

7.2.4 Composición del Vector de Características

Como hemos visto, los parámetros extraídos hasta ahora son aplicaciones sucesivas de determinados algoritmos a los valores obtenidos del análisis espectral de la señal en instantes sucesivos de tiempo.

Existen además otra serie de valores que tienen especial relevancia en el reconocimiento de la voz, y que han sido utilizados con éxito en diferentes sistemas :

Energia

Parece evidente que la energía de la trama de señal sea una buena elección como parámetro de reconocimiento. Ello, por ejemplo, nos ayudaría a diferenciar

zonas de voz y zonas de silencio, diferencia que no es tan evidente si utilizamos los parámetros derivados del análisis espectral de la señal. Las vocales tienen gran energía, las consonantes plosivas aparecen con un cambio brusco de energía, etc. Sin embargo, se ha demostrado que la energía es tan dependiente del locutor, que se hace inútil cuando se pretende utilizar en sistemas independientes del locutor.

Coeficientes diferenciales

La mayoría de los sistemas de reconocimiento actuales utilizan valores que miden modificaciones espectrales a lo largo del tiempo. Son los llamados coeficientes diferenciales, o Δcoeficientes. El cálculo de los Δcoeficientes se puede hacer simplemente restando cada vector de coeficientes del anterior.

Diferencial de la Energía

De la misma forma que los Δcoeficientes, se puede calcular la ΔEnergía. De hecho, este parámetro se presenta como uno de los más fiables en el reconocimiento de voz.

Puede parecer interesante utilizar un vector de características lo más grande posible, para poder discriminar de forma óptima las diferentes categorías. Sin embargo, la carga computacional exigida, así como la escasa mejoría de resultados hace que se intente obtener exactamente lo contrario. Es decir, realizar una reducción en el número de coeficientes, tratando de conservar las características discriminantes, descartando las redundantes o superfluas.

7.3 Modelización Acústica

El módulo de modelización acústica se encarga, en general, en evaluar la similaridad entre una cierta unidad lingüística (fonema, sílaba, palabra, etc.) y un conjunto de unidades patrón almacenadas.

Para que dicha evaluación sea posible, se deben tener en cuenta los siguientes puntos :

- **Unidad lingüística mínima :** ¿ qué unidades vamos a modelizar ?
- **Modelización acústica :** ¿ cómo vamos a modelizar dichas unidades ?

De esta forma, dado un conjunto de modelos, cada uno de ellos correspondiente a una unidad lingüística mínima, y la señal de voz cuantificada como una secuencia de vectores, podemos llegar a determinar la secuencia de Unidades Lingüísticas que se corresponden de forma óptima con la secuencia de vectores dada.

Se han propuesto numerosas alternativas para la selección de una Unidad Lingüística Mínima. La selección de **palabras** como unidad lingüística mínima ha sido utilizada muy frecuentemente en numerosos sistemas, dada su sencillez de manejo y la facilidad de aplicación a pequeños sistemas, generalmente sistemas de reconocimiento de palabras aisladas. Sin embargo, cuando el sistema crece, o se pretende reconocimiento de voz contínua, y el número de palabras del vocabulario sobrepasa las 1000 ó 2000 palabras, aparecen numerosos problemas, dado el gran tamaño del vocabulario, la dificultad de un adecuado entrenamiento, etc. Por ello, ningún sistema que aspire a utilizar un vocabulario amplio utiliza esta unidad lingüística como Unidad Mínima.

La utilización de **fonemas** como unidad lingüística trae consigo una serie de ventajas. En primer lugar, el número de fonemas diferentes para un cierto idioma es muy limitado, lo cual hace que puedan ser bien entrenados. Además, dicho número es fijo, independientemente de la aplicación que vayamos a implementar. Si somos capaces de definir un cierto vocabulario en función de los fonemas que componen cada una de sus palabras, no será necesario reentrenar nuestro sistema para que reconozca dicho vocabulario. Sin embargo, las características de un fonema pueden variar enormemente

debido al contexto en que se encuentra. Por ello, se han tratado de buscar otras soluciones alternativas.

Se han realizado un gran número de experimentos utilizando Unidades Lingüísticas de complejidad intermedia, como **sílabas**, **semisílabas**, **difonemas**, etc. Sin embargo, dado que el tamaño del vocabulario es relativamente grande en estos casos, el entrenamiento que se puede realizar suele resultar insuficiente. Además, los resultados obtenidos en los diferentes experimentos nunca han sido brillantes en el caso de sílabas o semisílabas, por lo que estas unidades no suelen ser útiles para el reconocimiento de voz. Otra opción son los **fonemas dependientes del contexto**. Dichas unidades tienen el mismo inconveniente que las unidades multifonema, es decir, un vocabulario grande, lo que lleva a dificultad en el aprendizaje. Sin embargo, los resultados obtenidos con estas unidades es realmente brillante.

En cuanto a la modelización acústica, podemos decir que la voz tiene una estructura temporal muy compleja. Por ejemplo, dos palabras iguales pronunciadas por la misma persona pueden tener apariencias muy diferentes, en función del contexto, estado de ánimo, etc. Por tanto, y dada la complejidad del problema, hemos de conseguir modelizar de alguna forma dicha variablilidad. Ello se consigue por medio de modelos matemáticos muy complejos.

Los estudios llevados a cabo sobre modelización acústica de la voz permiten definir cuatro grandes áreas de investigación :

- Sistemas basados en la Comparación de Patrones.
- Sistemas basados en el Conocimiento.
- Sistemas Estocásticos.
- Sistemas Conexionistas.

La mayoría de las recientes investigaciones muestran que dichas áreas no son independientes, ni mucho menos. La aparición de sistemas mixtos, que mezclan estrategias y extraen lo mejor de cada una de ellas están teniendo un auge cada vez mayor.

7.3.1 Sistemas Basados en la Comparación de Patrones

Los primeros sistemas de reconocimiento utilizaron la comparación de patrones. La idea subyacente es muy simple. Se almacenan una serie de patrones prototipo de cada unidad a reconocer. El reconocimiento se realiza comparando un nuevo patrón con cada uno de los vectores prototipo, y seleccionando aquel que difiera en menor medida, según una cierta definición de distancia.

El método más extendido de comparación de patrones, y dada la naturaleza eminentemente cambiante de la voz, es el de programación dinámica (*Dynamic Time Warping-DTW)*. Existen muchas variaciones sobre el algoritmo básico de Programación Dinámica en función de utilizar diferentes medidas de distorsión, caminos posibles y procedimientos de búsqueda.

Se han realizado numerosos estudios para mejorar la técnica original y adaptarla a voz contínua. Las técnicas de Programación Dinámica adaptada al reconocimiento de palabras conectadas se pueden resumir en tres :

- Programación Dinámica en dos niveles (*Two-Level DP*)
- Construcción de niveles (*Level Building*)
- Construcción de niveles síncrono por tramas. (*Frame-Synchronous Level Building*)

La complejidad computacional de la programación dinámica hace que si bien es aceptable para reconocimiento dependiente del locutor en vocabularios reducidos, es

impracticable si lo que pretendemos es reconocimiento de voz contínua independiente del locutor.

7.3.2 Sistemas Basados en el Conocimiento

Desde mucho antes que se comenzara a estudiar a fondo el desarrollo de sistemas automáticos para el reconocimiento de la voz, ya existían estudios muy serios sobre la naturaleza de la voz humana, su producción, características y percepción. Por tanto se pensó aplicar dichos conocimientos en la realización de sistemas guiados por el conocimiento de alto nivel que permitiesen realizar el reconocimiento de la voz de una forma fiable.

De gran trascendencia en este campo es el trabajo de Victor Zue, en los que se expone cómo se pueden utilizar las reglas fonéticas en el reconocimiento automático. Además, posteriores estudios demuestran que personas humanas expertas en "leer" espectrogramas son capaces de decodificarlos y traducirlos a frases. Finalmente dichos trabajos se han plasmado en la realización de un sistema completo absolutamente basado en el conocimiento.

7.3.3 Sistemas Estocásticos

Los sistemas estocásticos, y en concreto los Modelos Ocultos de Markov (*Hidden Markov Models-HMM*) han sido los sistemas más difundidos, mejor estudiados, y donde los progresos y resultados conseguidos han sido más brillantes desde un punto de vista práctico. La sólida base matemática de los mismos unido a la existencia de eficientes algoritmos de aprendizaje, hacen que la tecnología de los HMM sea la más desarrollada en la actualidad.

La actividad científica en torno a los Modelos de Markov en los últimos veinte años ha sido muy intensa, y los resultados obtenidos han sido excelentes. La popularidad viene debida a su estructura algorítmica simple, sencilla de implementar, y sobre todo, a su clara superioridad sobre otras estructuras alternativas de reconocimiento.

7.3.4 Sistemas Conexionistas

La aproximación conexionista es la más joven de las citadas anteriormente. Ya en 1943, en un famoso artículo W.S.McCulloch y W.Pitts introdujeron los principios y las bases del procesamiento conexionista. La comparación de los sistemas conexionistas con los sistemas clásicos de clasificación parece indicar la ventaja de los primeros tanto en eficacia como en capacidad de generalización.

Mención aparte merecen los sistemas denominados *Time Delay Neural Networks (TDNN)*, que representaron una de las primeras demostraciones de la eficacia de los sistemas conexionistas aplicados al reconocimiento de fonemas aislados. Se han realizado multitud de aproximaciones conexionistas al reconocimiento de la voz, entre las que destacaremos las redes de Viterbi, las redes neuronales sintácticas, las Redes Recurrentes y los Mapas Autoasociativos de Kohonen.

Pese a estar poco evolucionados aún, los resultados obtenidos por los sistemas conexionistas son prometedores. Sin embargo su superioridad frente a los sistemas estadísticos todavía no ha sido demostrada.

7.3.5 Sistemas Híbridos

Los últimos sistemas basados en la Programación Dinámica se ayudan de Redes de Neuronas Formales para obtener resultados sensiblemente mejores. Cabe citar los trabajos realizados por Sakoe bajo el nombre de Red Neuronal con Programación Dinámica (Dynamic Programming Neural Network - DPNN), y los realizados bajo el

nombre de Modelo de Predicción Neuronal (Neural Prediction Model - NPM). Ambos sistemas obtienen resultados mejores, y permiten, no solo la integración de gramáticas en los mismos, sino una buena respuesta frente al ruido, variablilidad de locutores, etc. gracias a la alta capacidad de generalización de las redes de neuronas utilizadas. Asimismo, son interesantes los híbridos HMM-LVQ, por los excelentes resultados obtenidos.

Los sistemas basados en TDNN por sí mismos no han obtenido resultados espectaculares. Sin embargo, los sistemas mixtos TDNN-HMM han conseguido alcanzar tasas de reconocimiento superiores a los basados exclusivamente en HMM.

Por último, es importante destacar la aparición de sistemas híbridos entre diferentes técnicas conexionistas, tales como mezclas de redes multicapa con Mapas Autoorganizativos aplicados a la transcripción automática de frases. Existen estudios comparativos interesantes, que comparan TDNN, LVQ2, HMM discretos y HMM contínuos.

Desgraciadamente, la evaluación comparativa de los diferentes sistemas de reconocimiento no es una tarea fácil, ya que normalmente, tanto los datos de entrenamiento y test como los criterios para la medición de las tasas de error suelen ser diferentes. La aparición de Bases de Datos de Voz (TI-Digits, TIMIT, NTIMIT, DR1, DR2, etc.) así como la definición de diferentes standards de evaluación de sistemas de reconocimiento han permitido reducir la problemática citada en gran medida.

7.4 Modelización del Lenguaje

La modelización del lenguaje incluye procesos como Léxico, Sintaxis, Semántica, Prosodia, Conocimiento del Mundo, Modelo de discurso, etc. Un buen modelo de un cierto lenguaje debe resolver problemas tales como dependencias del contexto, coarticulación entre unidades lingüísticas, discriminación entre palabras fonéticamente cercanas, etc.

La selección de la unidad lingüística mínima ha sido importante. Se han considerado, no solo palabras, o fonemas, sino unidades subfonéticas (*senones*), unidades supra-fonéticas (*alófonos*, *difonemas*, *sub-palabras*) y unidades dependientes del contexto. Asimismo, se han desarrollado sistemas automáticos para la generación de unidades fonéticas.

De una forma matemática, la modelización de un lenguaje se puede considerar como la asignación de una probabilidad a cada secuencia de unidades básicas (fonemas, palabras, etc.).

$$P(W) = P(w_1, w_2, \ldots w_n)$$

Esta probabilidad guia la búsqueda del reconocedor entre las diferentes hipótesis posibles, y constituyen un factor determinante en la determinación de la transcripción final en los sistemas de Reconocimiento de voz contínua.

7.4.1 Modelización estocástica

Siguiendo reglas elementales de teoría de la probabilidad, podemos descomponer P(W) como :

$$P(W) = w_1 \cdot P(w_2|w_1) \cdot P(w_3|w_1, w_2) \cdots P(w_n|w_1, w_2, \ldots w_{n-1}) = \prod_{i=1}^{n} P(w_i|w_1, w_2, \ldots w_{i-1})$$

Dado que es prácticamente imposible obtener todos los valores de probabilidad para todas las posibles combinaciones de palabras y para todas las longitudes, se realiza la aproximación siguiente :

$$P(w_j|w_1, w_2, \ldots w_{j-1}) \approx P(w_j|w_{j-N+1}, \ldots w_{j-1})$$

Es decir, se considera la probabilidad de observación de una palabra únicamente en función de las N anteriores. Por tanto, y en función de N, podemos distinguir gramáticas denominadas Bigram(N=2), Trigram(N=3) y gramáticas N-gram(N>3).
Asímismo, es de gran utilidad utilizar las denominadas gramáticas de pares de palabras (Word pair grammars) que definen qué pares de palabras son válidas en un cierto lenguaje. Es decir,

$$P(w_j|w_k) = \begin{cases} 1 & \text{Si } w_k w_j \text{ es un par posible} \\ 0 & \text{En caso contrario} \end{cases}$$

Una simplificación extrema nos lleva a los denominados modelos Sin Gramática (no-grammar models), donde se supone que toda pareja de palabras es posible :

$$P(w_j|w_k) = 1 \qquad \forall j,k$$

Los modelos estocásticos han sido utilizados con gran profusión, dada la facilidad de integrar dichas gramáticas en los sistemas de reconocimiento.

7.4.2 Modelización Gramátical

Diferentes alternativas a los modelos estocásticos se han propuesto, pero sin duda, la más prometedora la constituyen los modelos del lenguaje basados en la Teoría de Lenguajes Formales de Chomsky.
Se utiliza un sistema matemático para definir un lenguaje de cadenas de palabras, y asignar una estructura a dichas cadenas. La clave consiste por tanto en definir un lenguaje mediante un número finito de especificaciones, aunque el número de cadenas posibles sea ilimitado. Así, podemos distinguir entre gramáticas regulares, lineales, de contexto libre o sensibles al contexto. Para todas ellas existen dispositivos abstractos matemáticos(autómatas de estados finitos, autómatas a pilas, máquina de Turing, ...), denominados autómatas, diseñados para decidir si una determinada cadena pertenece o no a una gramática dada. Existen aplicaciones concretas, tales como compiladores gramaticales aplicados al reconocimiento de voz de palabras conectadas.
Si bien la integración de las Gramáticas Formales en los sistemas de reconocimiento es muy compleja, se han diseñado sistemas que implementan las Gramáticas de Contexto Libre. Dichas gramáticas permiten la definición de estructuras tales como las expresiones aritméticas, los lenguajes de programación y el lenguaje natural con ciertas restricciones. Por otro lado, su complejidad es lo suficientemente pequeña como para que sea abordable su implementación práctica. Recientemente se han integrado dichas técnicas a sistemas completos de laboratorio, tales como el sistema SPHINX o el sistema SPICOS, con gran éxito.

8 Bibliografía

[FURU, 89] FURUI, S. "Digital Speech Processing, Synthesis, and Recognition". Marcel Dekker, Inc. New York, 1989.

[FURU, 92] FURUI, S. y SONDHI, M.M. "Advances in Speech Signal Processing". Marcel Dekker, Inc. New York, 1992.

[GARC, 91] GARCIA, R. y GOMEZ, J. "Reconocimiento de Voz". Curso 3509-RVD. Programa de Postgrado en Sistemas y Redes de Comunicaciones. Univ. Politécnica de Madrid. 1991.

[HUAN, 90] HUANG, X.D., ARIKI, Y. y JACK, M.A. "Hidden Markov Models for Speech Recognition". Edinburgh University Press. Edimburgh, 1990.

[LEA, 80] LEA, W.A. "Trends in Speech Recognition". Prentice-Hall, Englewood Cliffs. New Jersey, 1980.

[LEE, 89] LEE, K.-F. "Automatic Speech Recognition: The Development of the SPHINX System". Kluwer Academic Publishers. Boston 1989.

[LEE, 92] LEE, K.-F. y ALLEVA, F. "Continuous Speech Recognition". in *Advances in Speech Signal Processing*, pp. 623-651. Ed. Marcel Dekker, Inc., New York 1992.

[MAKH, 75] MAKHOUL, J. "Linear Prediction: A Tutorial Review". Proc. IEEE, Vol.63, pp.561-580, Abril 1975.

[MARK, 76] MARKEL, J.D. y GRAY, A.H. "Linear Prediction of Speech". Springer-Verlag. New York, 1976.

[NAVA, 90] NAVARRO, T. "Manual de Pronunciación Española". C.S.I.C. 1990.

[O'SHA, 87] O'SHAUGHNESSY, D. "Speech Communication: Human and Machine". Addison-Wesley Publishing Company, Inc. Massachusetts, 1987.

[PAPO, 85] PAPOULIS, A. "Probability, Random Variables, and Stochastic Processes". McGraw-Hill Book Company . New York, 1985.

[QUIL, 85] QUILIS, A. y FERNANDEZ, J.A. "Curso de Fonética y Fonología Españolas". C.S.I.C. Instituto de Filología. Madrid, 1985.

[RABI, 75] RABINER, L.R. y GOLD, B. "Theory and Application of Digital Signal Processing". Prentice-Hall, Englewood Cliffs. New Jersey, 1975.

[RABI, 78] RABINER, L.R. y SCHAFER, R.W. "Digital Processing of Speech Signals". Prentice-Hall, Englewood Cliffs. New Jersey, 1978.

[RABI, 89] RABINER, L.R. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". Proceedings of the IEEE, Vol.77, No.2, pp.257-286. Febrero 1989.

[RABI, 93] RABINER, L.R. y JUANG, B.-H. "Fundamentals of Speech Recognition". Prentice-Hall, Englewood Cliffs. New Jersey, 1993.