01170

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

DIVISIÓN DE ESTUDIOS DE POSGRADO

TESIS

TÉCNICAS DE RECONOCIMIENTO AUTOMÁTICO DE VOZ UTILIZANDO SEGMENTACIÓN ACÚSTICA

PRESENTADA POR:

ANTONIO FRANCISCO MONDRAGÓN TORRES

PARA OBTENER EL GRADO DE :

MAESTRO EN INGENIERÍA
(ELÉCTRICA)

DIRIGIDA POR:

M. EN ING. ABEL HERRERA CAMACHO

Ciudad Universitaria, Agosto de 1996

TESIS CON
FALLA DE ORIGEN

TESIS CON
FALLA DE ORIGEN

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

TESIS

COMPLETA

A la memoria de mi padre Francisco.

A la memoria de mi Abuelita Toña,
mi Tía Raquel y mi Tía Meche.

A mi madre María Guadalupe
que sé que conoce la importancia
que tienen estos momentos en mi vida.

A mi esposa Adriana
que me ha seguido en
este camino tan difícil
siempre a mi lado.

A mis hijos Antonio y Andres
que espero me den
la oportunidad
de crecer con ellos.

A todas las personas
que me han
alentado y apoyado.

Agradecimientos

Al M. en I. Jose Abel Herrera Camacho por la oportunidad de trabajar con él y el apoyo que me brindó.

Al Dr. Boris Escalante Ramírez por su apoyo, su amistad y sus conocimientos.

Al Dr. Francisco García Ugalde por su apoyo y ser un ejemplo de maestro.

Al Dr. Jesus Savage Carmona por su apoyo y sus consejos.

A todos mis profesores y compañeros que me ayudaron a completar otra fase de mi vida y que sin su ayuda hubiera sido mas dificil.

A todas aquellas personas que aunque no las mencione, saben que tienen un lugar dentro de mi y siempre les voy a estar agradecido.

Resumen

El propósito de este trabajo es el de comparar dos técnicas de agrupamiento. Una se basa en cuantización vectorial y el otro en agrupamiento con lógica difusa. La ventaja de utilizar agrupamiento difuso (suave) contra no-difuso (duro), es que cada elemento puede pertenecer a diferentes patrones en diferente grado de pertenencia.

Existen diferentes técnicas de segmentación de palabras, incluyendo matemáticas y lingüísticas. Aquí se propone el uso de otra técnica denominada segmentación en subpalabras acústicas, que separa a la palabra en segmentos de duración variable con características espectrales cuasiestacionarias.

Se utilizó una base de datos de dígitos en Inglés. Usamos 1000 repeticiones para entrenamiento (10 dígitos, 10 locutores y 10 repeticiones de cada dígito) y 1600 repeticiones para la clasificación (10 dígitos, 10 locutores y 16 repeticiones de cada dígito). Usando estos datos y la información de la segmentación acústica, entrenamos un sistema de reconocimiento de voz que utiliza un algoritmo (duro) de agrupamiento denominado K-Medias y un algoritmo (suave) denominado C-Medias difuso.

Se demostró que usando el enfoque de subpalabras acústicas, se puede generar un sistema de reconocimiento independiente del locutor y de su sexo. Entonces entrenamos ambos sistemas, y comparamos los resultados.

La primera etapa denominada Procesamiento de la señal de voz muestra como a partir de la base de datos, se procesa la señal para cambiarle su tasa de muestreo y generar los archivos que contienen la información de la segmentación acústica. Los archivos resultantes de este proceso estan muestreados a una tasa de 10,000 muestras/seg. y fueron procesadas con un filtro de preénfasis para igualar el espectro y después fueron normalizados antes de obtener los vectores de características. A las tramas de 128 muestras, se le aplicó una ventana de Hamming y existió un traslape de 20 muestras con la trama anterior.

Se generaron archivos por cada segmento acústico, conteniendo vectores de coeficientes de predicción lineal y autocorrelación por cada trama, así como se calculó la energía mínima residual utilizada en la distorsión de Itakura-Saito.

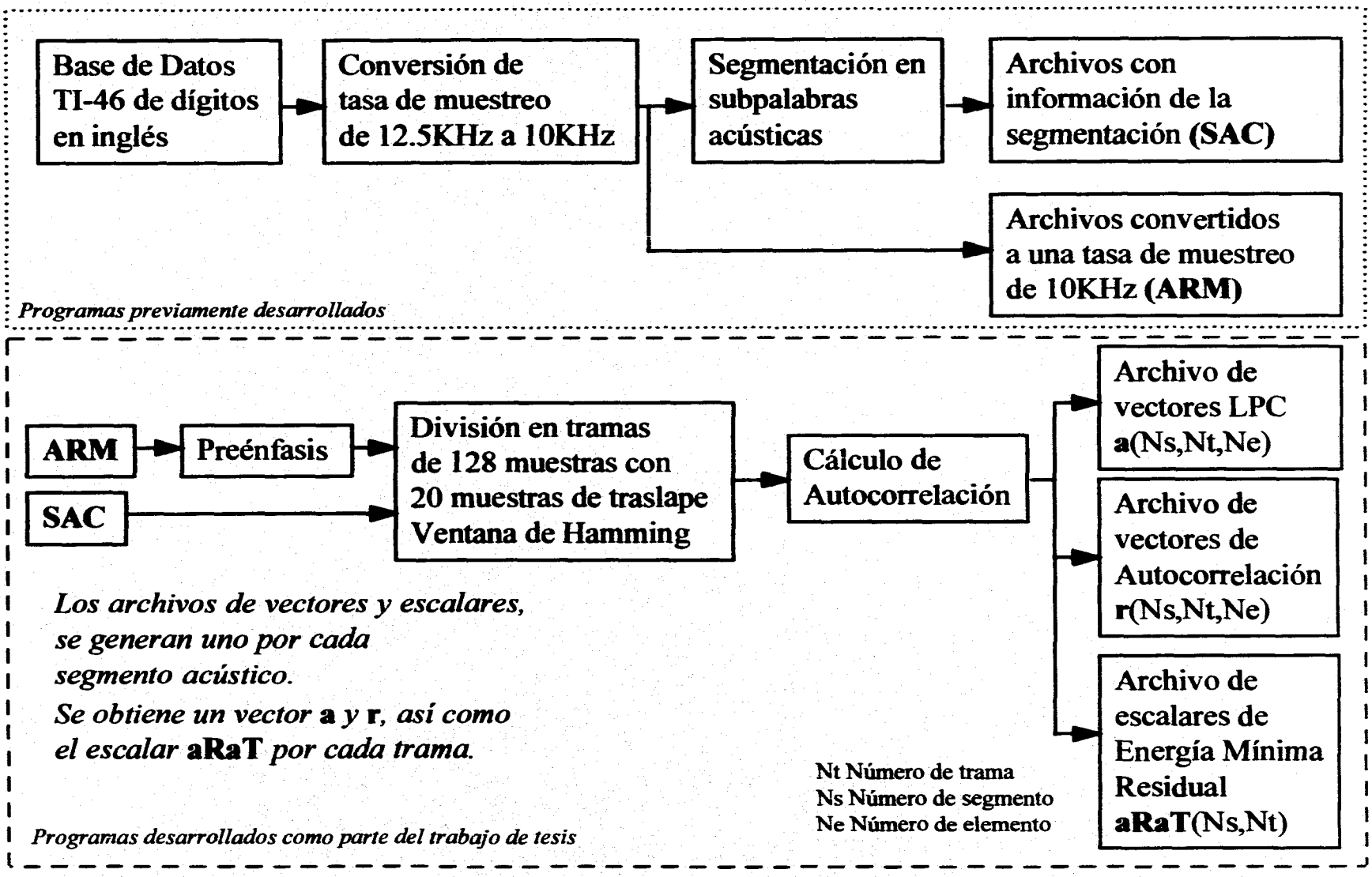
Para agrupamiento duro, usamos vectores de siete coeficientes de predicción lineal y generamos dieciséis centroides por subpalabra. Para comparación de las palabras utilizamos la distorsión de Itakura-Saito modificada. Para agrupamiento suave, usamos doce coeficientes cepstrales obtenidos a partir de los coeficientes de predicción lineal previamente calculados y también generamos dieciséis centroides por cada subpalabra. Como medida de similitud utilizamos la distancia euclidiana.

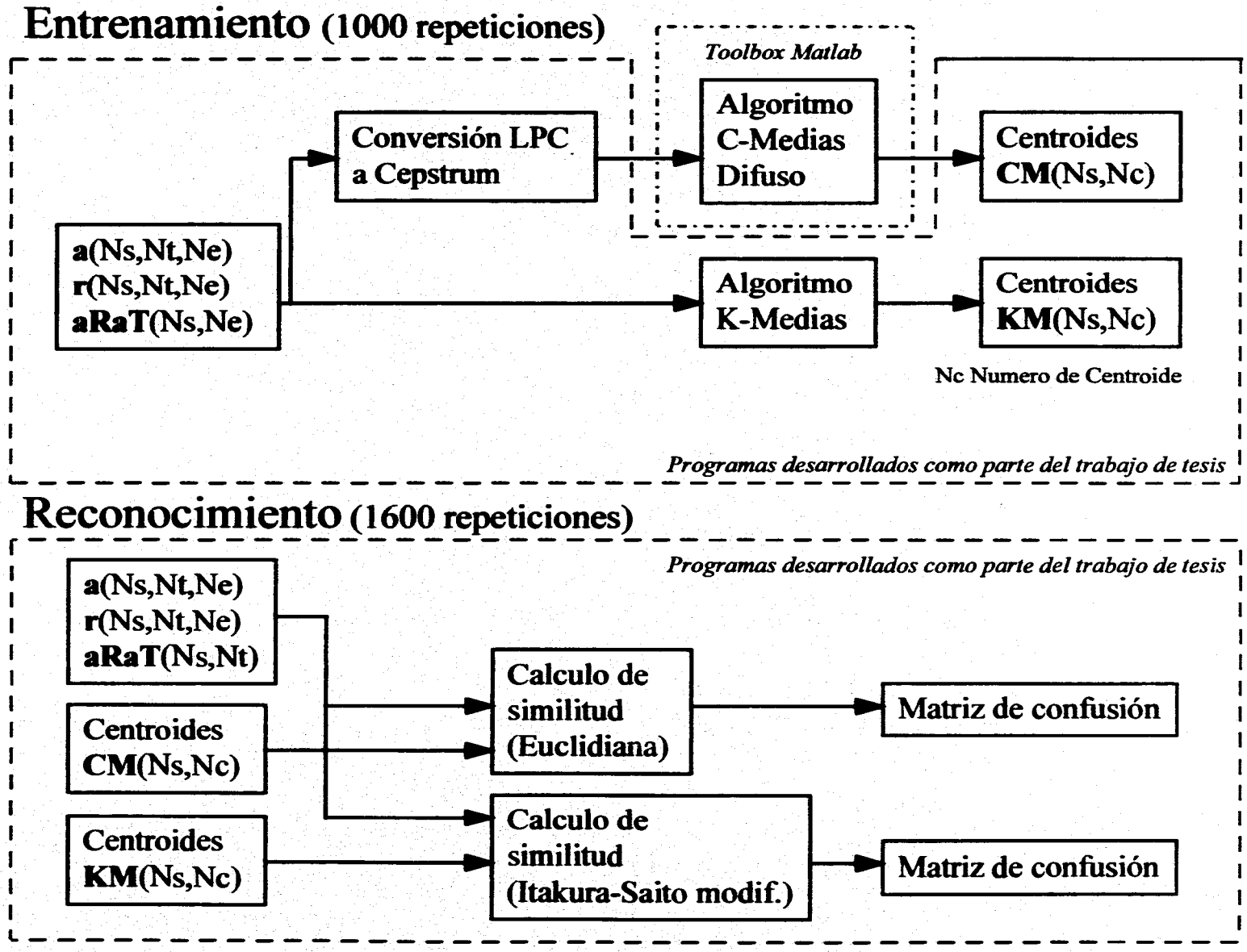
Para la fase de reconocimiento se comparó cada palabra de reconocimiento contra los centroides generados en la fase de entrenamiento y a partir de estos generamos las matrices de confusión que nos indican los resultados del reconocimiento total.

Se lista un apéndice al final del texto donde se mencionan los programas utilizados en el desarrollo de este trabajo.

A continuación se muestran dos diagramas a bloques que describen el proceso de la señal de voz, así como las fases de entrenamiento y reconocimiento.

Procesamiento de la señal de voz




Entrenamiento (1000 repeticiones)
Toolbox Matlab
Algoritmo C-Medias Difuso
Conversión LPC a Cepstrum
Centroides CM(Ns,Nc)
a(Ns,Nt,Ne)
r(Ns,Nt,Ne)
aRaT(Ns,Ne)
Algoritmo K-Medias
Centroides KM(Ns,Nc)
Nc Numero de Centroide
Programas desarrollados como parte del trabajo de tesis
Reconocimiento (1600 repeticiones)
Programas desarrollados como parte del trabajo de tesis
a(Ns,Nt,Ne)
r(Ns,Nt,Ne)
aRaT(Ns,Nt)
Centroides CM(Ns,Nc)
Centroides KM(Ns,Nc)
Calculo de similitud (Euclidiana)
Calculo de similitud (Itakura-Saito modif.)
Matriz de confusión
Matriz de confusión

Indice

Indice de Figuras

Indice de Tablas

1. Introducción

El presente trabajo pretende evaluar diferentes alternativas sobre el uso de diversas técnicas en el reconocimiento automático de voz.

Este se plantea sobre la base que el lector posee conocimientos básicos sobre señales y sistemas, así como procesamiento digital de señales determinísticas y estocásticas. Los cinco primeros capítulos tratan sobre percepción y generación de la señal de voz. En los siguientes dos capítulos se hace una analogía de las características físicas de los sistemas generación y percepción de voz a su representación matemática con el fin de justificar las herramientas que se utilizaron.

El análisis se basa en los modelos matemáticos de como se genera el habla y como es percibida por el ser humano[PARS87][RABI93][DELL87]. Se modela el tracto vocal humano con el fin de justificar el uso de coeficientes de predicción lineal que permiten una aproximación de las características físicas del medio que propaga y modela la señal de voz. Por otro lado el oído humano no tiene una respuesta lineal, de hecho se le puede caracterizar como un banco de filtros no lineal, esto es importante cuando lo que queremos imitar es el sistema humano de reconocimiento de voz[JEFF70]. Para establecer una comparación entre dos patrones de voz, utilizamos una medida, que más que geométrica, es perceptual, esto implica que dos patrones se consideran diferentes, cuando una persona los percibe como distintos.

En general, todo sistema de reconocimiento se puede clasificar como dependiente o independiente del locutor y de palabras aisladas o continuas. En este trabajo consideramos el caso de independencia del locutor, ya que se utiliza una técnica de segmentación acústica con lo que se logran aislar las subpalabras contenidas en una palabra. Al obtener características acústicas, entonces podemos entrenar al sistema independientemente del género del locutor así como de su identidad.

Para el entrenamiento del sistema utilizamos técnicas de cuantización vectorial en diferentes modalidades[TOU81]. La primera es una técnica sencilla pero poderosa, denominada "K-Medias", la segunda es una variación tanto de técnica como de concepto denominada "C-Medias", que toma sus principios de la "Lógica Difusa"[ZADE65][KOSK93] y por lo tanto las matemáticas asociadas se toman de otra disciplina diferente a las tradicionales.

Otra técnica que se podría utilizar como alternativa es "Modelos Ocultos de Markov"[RABI89], que a diferencia de los métodos anteriores que involucran una caracterización de la señal a reconocer, esta la modela estocásticamente y la decisión de que palabra es la que se reconoce es probabilística. Esta técnica se utiliza comúnmente en los sistemas de reconocimiento actuales.

Existen otras técnicas como las basadas en "Redes Neuronales"[KOSK92] que se están comenzando a utilizar como una alternativa eficaz para la resolución del problema de reconocimiento de voz. Debido a los alcances y limitaciones del presente trabajo, solamente se evaluarán y compararán las primeras dos técnicas mencionadas anteriormente.

En un principio, el proyecto se planteó en realizar el reconocimiento sobre palabras en español, para esto, un grupo de trabajo se dedicó a colectar y formar una base de datos de dígitos en español. Desgraciadamente la relación señal a ruido (RSR) era muy baja y los programas de segmentación acústica no contemplaban este tipo de RSR, por lo que la segmentación no se efectuaba en forma correcta. Debido a este factor, se tomó para el desarrollo la base de datos TI-46 de dígitos en inglés, desarrollada por Texas Instruments, que posee las características necesarias

para una correcta segmentación, ya que el trabajo inicial sobre subpalabras acústicas se hizo basado en esta base de datos.

Por último, este trabajo de tesis pretende dar al lector no muy familiarizado con el tema de procesamiento de voz, una recopilación breve pero adecuada de los temas necesarios para la comprensión del texto. Esto se hace pensando en las múltiples referencias que existen sobre el tema. Además el presente involucra una combinación de técnicas con el fin de comparar resultados y comprobar si éstas combinaciones tienen algún impacto comparado con otros sistemas de reconocimiento de voz.

1.1 Introducción a la Señal de Voz

La voz es una señal acústica que comparada con una imagen visual, puede parecer a primera vista sencilla de interpretar. Existen diferentes lenguajes dependiendo del grupo étnico. A continuación se da una definición del lenguaje.
Lenguaje. Conjunto de sonidos articulados con que el hombre manifiesta lo que piensa y siente.
El lenguaje se puede dividir en Lengua y Habla[QUIL79].

Lenguaje {	Lengua:	Modelo general y constante para todos los miembros de una colectividad lingüística.
	Habla:	Materialización de ese modelo en cada miembro de la colectividad lingüística

El habla es el medio hecho por el hombre para la comunicación de pensamientos e ideas. Las características, objetos y relaciones integradas en una señal de voz, son representaciones abstractas de conceptos inventados por la mente humana y codificados por el aparato vocal humano en una secuencia de patrones acústicos.

La dificultad para interpretarla se manifiesta en una serie de problemas que deberán de ser resueltos por cualquier sistema de reconocimiento automático de voz. Las áreas de problemas a resolver son las siguientes:

Primero, *las señales de voz tienden a ser continuas*, el habla fluida representa cambios continuos y cambiantes de patrones sin marcas específicas que indiquen el fin de un sonido y el principio de otro. Tampoco existen fronteras o límites obvios entre una palabra y la siguiente. En particular no existen pausas regulares entre una palabra y otra en una expresión.

Segundo, *las señales de voz tienden a ser altamente variables*. La voz de una persona es bastante diferente a la de otra, ya sea por origen lingüístico o por diferencias de sus sistemas vocales. También existen claras diferencias entre las voces de un hombre, una mujer y un niño; entre otras cosas por sus diferencias en tamaño físico.

Aún la voz de un solo individuo presenta variaciones bajo diferentes condiciones como cuando susurra, grita, o cuando tiene una gripa. Es virtualmente imposible para un locutor, pronunciar la misma palabra o frase, con representación analítica idéntica en dos ocasiones diferentes.

Tercero, *el habla es ambigua*. Las expresiones habladas normalmente tienen representaciones alternas. Por ejemplo "caza" y "casa" que tienen prácticamente el mismo sonido, pero el contexto es el que les da su significado dentro de una frase.

Cuarto, *las señales de voz normalmente se encuentran contaminadas*. Solamente bajo condiciones ideales, se puede asegurar que la única señal presente que entra al micrófono es la voz. Normalmente el habla se genera en un ambiente donde existe ruido ambiental. En algunos casos la magnitud de éste, puede ser mayor a la señal de voz que nos interesa. Además la señal de voz

puede haber sido enviada por un canal de comunicaciones, que pudo haber agregado más complicaciones como distorsión o retraso.

Por último, *el habla es compleja*. El lenguaje y el habla están íntimamente relacionadas y el habla es solamente una pequeña parte de un sistema simbólico de señalización para la comunicación de pensamientos e ideas entre los seres humanos. La comunicación humana esta compuesta de un comportamiento ritual que esta diseñado para una comunicación efectiva.

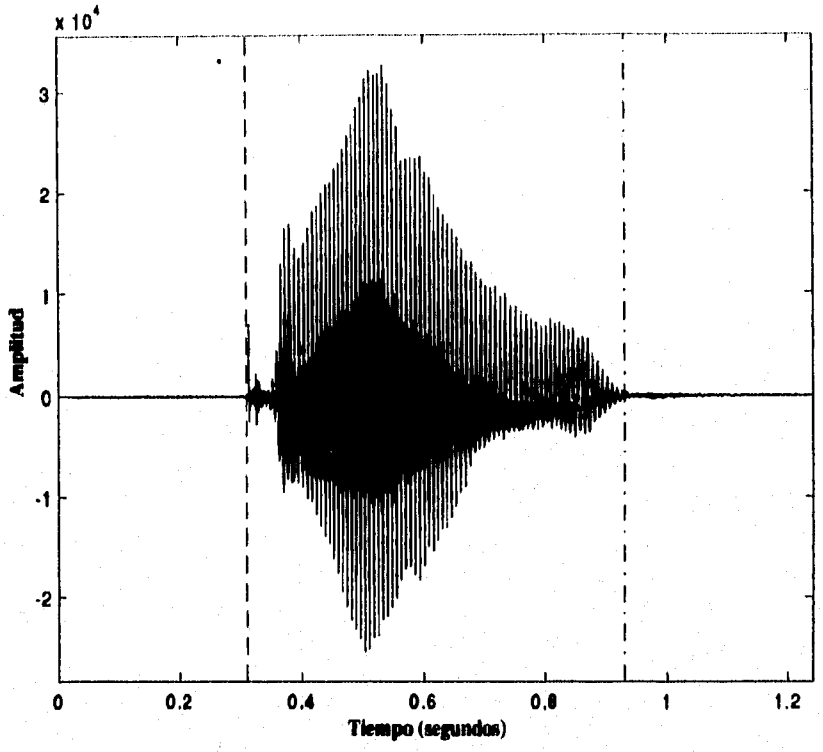


Figura 1.1 Palabra de entrenamiento "three" de la base de datos T1-46.

1.2 Implicaciones para el Reconocimiento Automático de Voz

Para que un sistema de reconocimiento sea exitoso, deberá de resolver las áreas de problemas expuestas anteriormente. Un sistema de reconocimiento automático deberá de poder reconocer palabras aún cuando estén integradas en una expresión continua. Se necesitarán algoritmos que permitan la eficiente segmentación de la expresión, normalmente la división de la expresión oral en una cadena de palabras.

Un sistema de reconocimiento, también deberá de explotar similitudes entre patrones en lugar de basarse en la repetición precisa de la misma información en diferentes ocasiones. Esto implica la necesidad de establecer "medidas" adecuadas para poder medir la similitud(o distancia) entre diferentes patrones.

Un sistema de reconocimiento deberá de hacer uso del contexto, con el fin de resolver ambigüedades. Esto significa que la identidad de una palabra, no puede ser decidida independientemente de las otras palabras integradas en la expresión.

Finalmente, un sistema de reconocimiento automático deberá de soportar señales de interferencia y ruido, además de proporcionar una buena interfaz para un proceso superior de interpretación semántica, si lo que se pretende es proporcionar capacidad más allá de un botón activado por voz.

Actualmente existen sistemas comerciales que permiten el reconocimiento de voz. Existen tres categorías: Navegación, Dictado y Desarrollo.

Los productos de navegación han estado disponibles comercialmente desde hace varios años. Estos permiten operaciones de control, como el de ejecutar y dar comandos a una aplicación mediante la voz.

Los productos de dictado, que se han introducido al mercado recientemente, permiten utilizar la voz para crear documentos de texto, introducir números a hojas de cálculo y conducir sesiones en línea. Ya que los sistemas de dictado deben de reconocer muchas más palabras que los navegadores(éstos solamente reconocen comandos del menú), su aparición en el mercado ha sido mucho más lenta. De hecho, la complejidad necesaria para el reconocimiento de voz es tan grande, que se demandan recursos de cómputo bastante significativos. Por ésta razón, los fabricantes de éste tipo de productos normalmente requieren tarjetas dedicadas de coprocesamiento.

Los productos de desarrollo permiten crear aplicaciones de propósito general y altamente especializadas que involucran la tecnología de reconocimiento de voz. Estos productos se basan prácticamente en el Lenguaje C y el Visual Basic de Microsoft, proporcionando al programador una interfase de programación de aplicaciones y bibliotecas.

La mayoría de los productos actualmente dependen del locutor, deben de ser entrenados para reconocer los patrones y la pronunciación de cada locutor. Este entrenamiento puede ser largo y tedioso. Idealmente, el reconocimiento de voz deberá de efectuarse sin importar el locutor, actualmente el mercado se está orientando en ése sentido pero falta bastante por desarrollarse.

1.2.1 Tecnologías de Entrada de Voz[1].

Reconocimiento de voz vs. Reconocimiento de locutor.
Reconocimiento de voz, es la habilidad de un sistema de reconocer palabras habladas o frases (el reconocimiento en este sentido no necesariamente es la habilidad de asociar de forma inteligente un significado). Reconocimiento del locutor (también llamado identificación de voz o identificación de locutor) es la habilidad de un sistema a reconocer la identidad de un individuo específico que está hablando, de entre una lista de posibles locutores. La verificación de voz, o verificación del locutor se enfoca a confirmar que el locutor es quién pretende ser. Con verificación de voz siempre existe un usuario hipotético que es aceptado o rechazado en función de la proximidad de la comparación. Normalmente se requiere de un sistema dependiente del texto.

Dependiente del locutor vs. Independiente del locutor.
Una tecnología independiente del locutor, normalmente es entrenada previamente, o generalizada mediante la repetición de la misma palabra por una serie de locutores. Por lo tanto un sistema independiente del locutor se puede utilizar inmediatamente en una aplicación con usuarios que difieren de edad, sexo, acentos y tonos, de tal forma que el usuario no tiene que entrenar el sistema. Las tecnologías dependientes del locutor (también denominadas reconocimiento de voz) requieren que el usuario entrene las palabras previamente antes que se puedan identificar (normalmente dos repeticiones son suficientes). Los sistemas dependientes del locutor son independientes del lenguaje, ya que el sistema no está preentrenado en ningún lenguaje o palabras. Existe un híbrido que se denomina adaptativo del locutor, en el que primero es un sistema

[1] Resumen de [MOZE96]

independiente del locutor y después se adapta a los usuarios individuales para mejorar la precisión. Este enfoque es bastante bueno y preciso, pero es dependiente del lenguaje.

Reconocimiento de voz continua vs. Reconocimiento de palabras discretas.
Esto se refiere a que tipo de flujo de información puede manejar el sistema de reconocimiento. Un sistema de reconocimiento continuo es como lo hacen los humanos, no se requiere de pausas entre las palabras, el cerebro puede procesar la información de voz tan rápido de lo que un locutor la puede pronunciar. Un sistema de reconocimiento discreto necesita analizar palabras aisladas a cada tiempo. Un sistema de voz a texto usa reconocimiento continuo y convierte los datos reconocidos a un arreglo ASCII u otro archivo de texto para ser impreso o almacenado. La búsqueda de palabras es un enfoque híbrido de escuchar continuamente y reconocimiento discreto, permite voz continua pero solamente reconoce algunas palabras específicas.

Vocabularios limitados vs. Vocabularios ilimitados.
Los enfoques a vocabularios ilimitados permiten que el locutor pronuncie un número grande de palabras en cualquier momento, buscando en todo el diccionario para identificar cada palabra pronunciada. El enfoque de vocabulario limitado restringe la cantidad y la secuencia de palabras que se pueden decir. Normalmente dividen las palabras permitidas en ramas de árbol o conjuntos manejados por menús (llamados gramática de estados finitos). En estas jerarquías, se encuentra disponible un conjunto a un tiempo dado (el conjunto activo) y cualquier palabra reconocida de este conjunto conmutará automáticamente el sistema a un nuevo conjunto activo, dando la apariencia de un sistema de vocabulario ilimitado.

1.2.2 Tecnologías de Reconocimiento de Voz[2].

Técnicas de modelado estocástico.
El modelado estocástico se refiere al proceso de hacer una secuencia de selecciones de un conjunto de alterativas. El más común de estos enfoques son los *modelos ocultos de Markov* (HMM). Un HMM describe la secuencia del espectro de sonido generado por la pronunciación de palabras, sílabas, etc, como una secuencia de transición entre estados discretos. El HMM busca información en el estado actual, analiza la información que esta entrando y toma una decisión probabilística de cual debe ser el siguiente estado. Para crear este modelo se requiere de grandes cantidades de datos de entrenamiento y bastantes cálculos, sin embargo, ya generado el modelo se puede encontrar una palabra del vocabulario en forma fácil y sin tantos cálculos.

Comparación espectral.
Las primeras técnicas de reconocimiento utilizaban comparación de patrones para determinar que tan bien se ajustaban. Con el tiempo, estos enfoques se han sofisticado y vuelto más complejos de calcular. Varios de estos enfoques analizan los patrones de energía espectral y hacen comparaciones con patrones almacenados. El modelo de cruce por cero es el más sencillo y simple de implantar, se basa en llevar cuenta de los puntos donde la amplitud cruza el eje y guarda esta serie en el tiempo como patrón de comparación. La codificación de predicción lineal (LPC) sigue los puntos previos de energía espectral para predecir los valores actuales y modela la información contra el tracto vocal humano.

Alineación en el tiempo.
Las técnicas de alineación en el tiempo pueden utilizarse en conjunto con las técnicas mencionadas anteriormente para crear vectores uniformes en tiempo para poder comparar. Por ejemplo, una alineación lineal de tiempo, puede escalar una palabra de entrada o frase para igualar la longitud del patrón, esto compensa si se habla rápido o lento. Un enfoque más sofisticado es ajuste dinánico en tiempo (DTW), este proporciona una coincidencia no lineal optimizada con el patrón. El enfoque de DTW a la alineación en tiempo mejora la precisión del ajuste, ya que los cambios temporales en la entrada y en los vectores de patrones tienden a ocurrir en formas no lineales. El tiempo en al

[2] Resumen de [MOZE96]

entrada se ajusta precisamente para coincidir con el del patrón, aunque el tiempo esté cambiando dinámicamente mientras se pronuncia la palabra.

Como se puede observar en la Figura 1.2, se tienen dos pronunciaciones de la misma palabra dichas por dos locutores distintos del sexo femenino. Como se puede observar tanto la forma de onda en el tiempo, la energía en tiempo corto, la tasa de cruces por cero y la densidad espectral de potencia son diferentes. Por lo tanto no podríamos realizar una comparación de patrones directa ni el tiempo ni en la frecuencia. De aquí la importancia de obtener características de la palabra que la hagan independiente de su posición en el tiempo o de fluctuaciones en el espectro. Además las dos palabras son segmentadas en diferente número de subpalabras acústicas.

1.3 Especificación del Trabajo

El objetivo de éste trabajo de tesis es el de generar un "Reconocedor Automático de Palabras Aisladas". Se pretende utilizar técnicas actuales en éste campo. En el caso de que éstas técnicas no satisfagan las necesidades del trabajo, entonces se desarrollarán nuevas técnicas que permitan que el sistema sea más eficaz.

De entre las técnicas de modelado y métodos de reconocimiento se investigarán las siguientes, con el fin de analizar si cumplen con los requerimientos para este trabajo.

Modelado, Caracterización y Comparación:

- Preénfasis
- Análisis de la señal en tiempo corto
- Segmentación en Subpalabras Acústicas
- Segmentaciones en tramas y uso de ventanas
- Predicción Lineal
- Cepstrum
- Distancia Euclidiana
- Distanciade Máxima Similitud

Métodos de Agrupamiento:

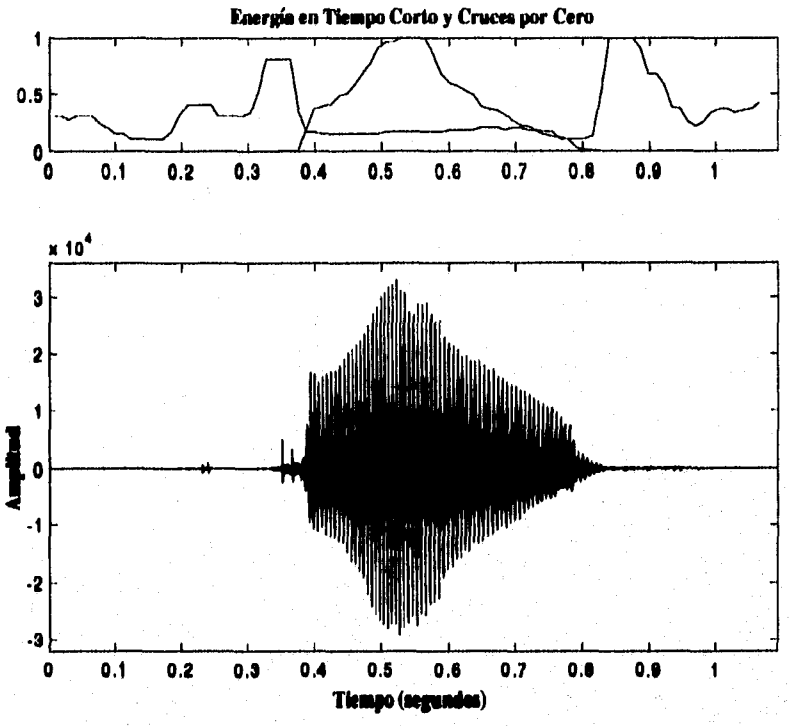
- Cuantización Vectorial
- Lógica Difusa

Cuantización Vectorial (VQ)
Se puede utilizar un vector de características de tiempo-corto para representar los rasgos esenciales de la voz de una persona. Sin embargo esa representación no es práctica cuando el número de vectores es grande. Los requerimientos de memoria para el almacenamiento y la complejidad en el cálculo son prohibitivos. Entonces, existe un método eficaz para comprimir los vectores de entrenamiento. Para esto se utiliza un diccionario de cuantización vectorial representado por centroides, que consiste de un número pequeño pero muy representativo de los vectores de características que es utilizado como un medio eficaz de caracterizar los rasgos específicos del locutor.

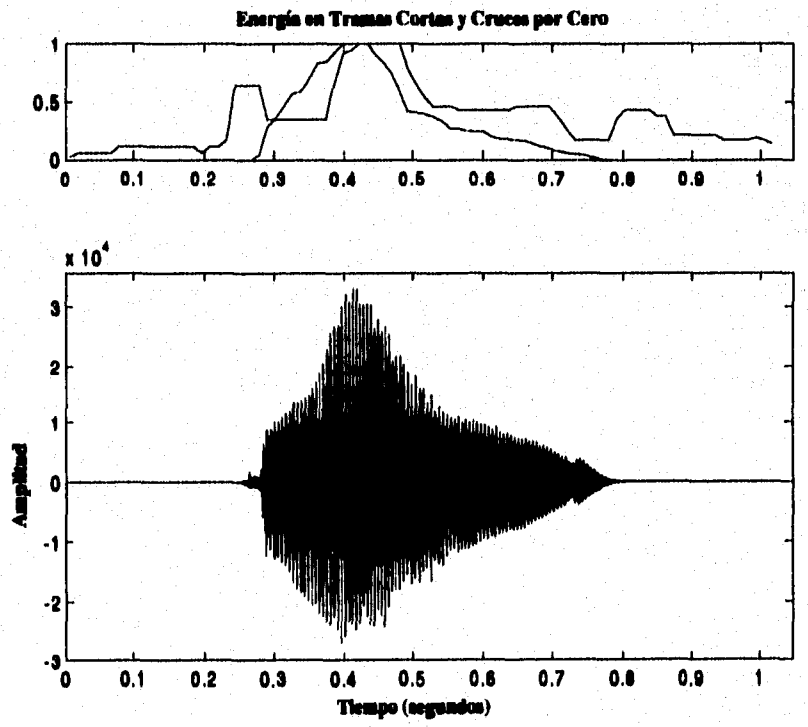
Lógica Difusa (LD)
Al igual que las Redes Neuronales, la LD es una disciplina relativamente nueva , ha sido utilizada con éxito en campos como sistemas de control, reconocimiento de imágenes, sistemas expertos, etc. Se pretende evaluar la posibilidad de integrarlo en algún bloque del sistema en caso de resultar más eficiente que las técnicas ya establecidas para llevar a cabo el reconocimiento de voz.

La LD se basa en la utilización de conjuntos difusos en los que un elemento , tiene un cierto grado de pertenencia a un conjunto, a diferencia de la teoría de conjuntos que reducen a pertenecer o no al conjunto, o la probabilidad que estima cual es la factibilidad de que ocurra un evento. En la LD un evento puede pertenecer a diferentes conjuntos simultáneamente y con diferentes grados de

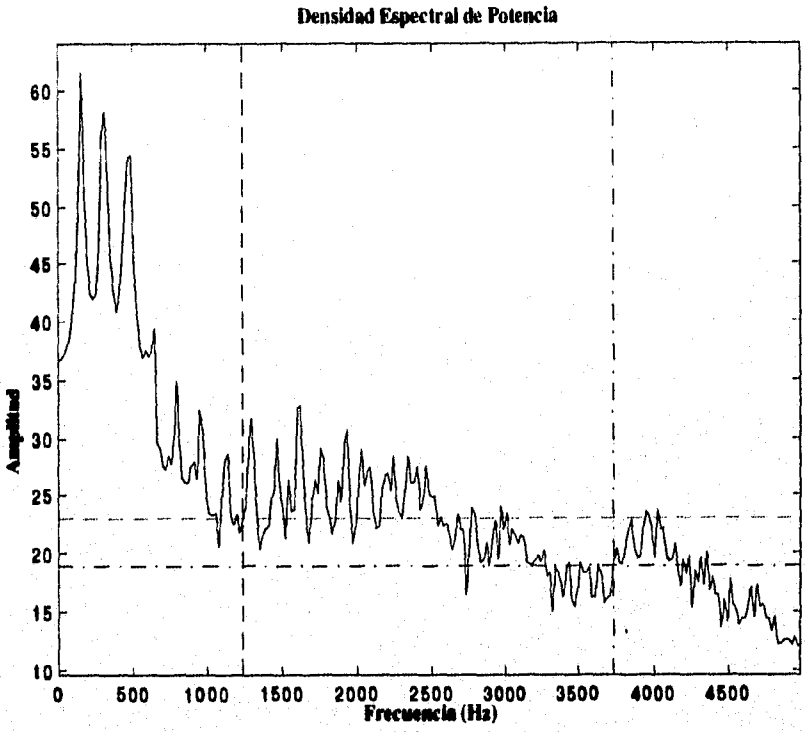
pertenencia. De aquí se piensa que puede ser útil en el reconocimiento, ya que se pueden realizar comparaciones contra una serie de patrones y aquel al cual pertenezca en mayor grado, se puede tomar como el correcto.



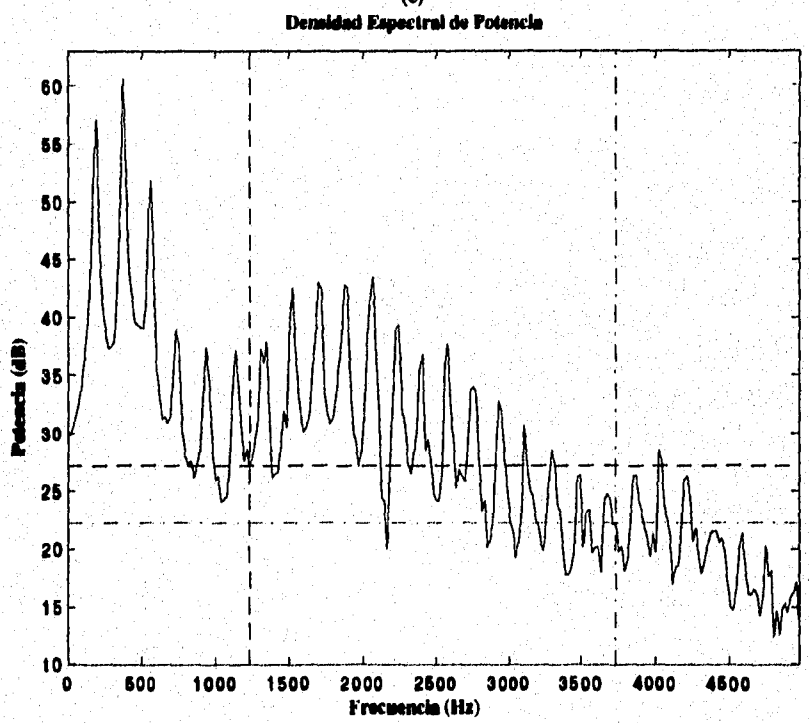
(a)



(b)



(c)



(d)

Figura 1.2 Palabra "Three" pronunciada por dos locutores del sexo femenino. Para la primera locutora, (a) representan la forma de onda en el tiempo, la energía en tiempo corto y la razón de cruces por cero, (c) representa la densidad espectral de potencia. (b) y (d) representan la misma información para la segunda locutora.

2. Ondas

2.1 Ondas Mecánicas

Una onda es cualquier perturbación de una condición de equilibrio que se mueve o se propaga de una región a otra del espacio. Las ondas más fáciles de comprender son las ondas mecánicas, propagadas a través de un medio material cuando éste se desplaza de su estado de equilibrio.

Cuando los movimientos de las partículas son perpendiculares a la dirección de propagación de la onda, ésta se denomina transversal. Si las partículas se mueven hacia atrás y hacia adelante, a lo largo de la dirección de propagación, la onda se denomina longitudinal.

La distancia entre dos máximos sucesivos (o entre dos puntos sucesivos en fase) es la longitud de onda y se designa por λ. Como la señal se propaga con velocidad constante c, avanza una distancia igual a una longitud de onda en el intervalo de tiempo de un período, se deduce que

$$c = \frac{\lambda}{\tau} \text{ o } c = f\lambda \tag{2.1}$$

Es decir, la velocidad de propagación es igual al producto de la frecuencia por la longitud de onda.

2.2 Ondas Sonoras

Las ondas longitudinales que viajan en el aire, al llegar al oído producen la sensación de sonido. El oído humano es sensible a ondas comprendidas en un intervalo de frecuencias de 20 a 20000Hz, aunque a veces el término sonido se aplica también a ondas similares con frecuencias fuera del intervalo de perceptibilidad humana.

Las ondas sonoras más sencillas son las sinusoidales con frecuencia, amplitud y longitud de onda definidas. Cuando una de éstas llega al oído causa una vibración en las partículas de aire situadas delante del tímpano, con una frecuencia y una amplitud determinadas. Esta vibración puede considerarse también en función de las variaciones de presión del aire en el mismo punto. La presión del aire se eleva sobre la presión atmosférica y después cae por debajo de ella, con movimiento armónico simple de la misma frecuencia que una partícula de aire.

2.2.1 Intensidad

Un aspecto esencial de la propagación ondulatoria de cualquier tipo es la transferencia de energía. La intensidad I de una onda que se propaga se define como la cantidad media de energía transportada por la onda, por unidad de superficie y por unidad de tiempo, a través de una superficie perpendicular a la dirección de propagación. En pocas palabras, la intensidad es la potencia media transportada por unidad de superficie.

$$I = \tfrac{1}{2}\omega B k A^2$$
$$k = \frac{\omega}{c} \tag{2.2}$$

donde:

I es la intensidad

ω frecuencia angular en rads/seg.
B es el modulo de compresibilidad
A es la amplitud

También puede expresarse en función de la Presión P como

$$I = \frac{cP^2}{2B} \tag{2.3}$$

utilizando la relación de la velocidad de onda $c = \left(\frac{B}{\rho}\right)^{\frac{1}{2}}$ también puede replantearse en las siguientes formas:

$$I = \frac{P^2}{2\rho c} = \frac{P^2}{2\sqrt{\rho B}} \tag{2.4}$$

donde

ρ es la densidad del líquido o del gas.

2.2.2 Nivel de Intensidad y Sonoridad

Debido a la gran amplitud del intervalo de intensidades a las que es sensible el oído, es conveniente utilizar una escala de intensidad logarítmica, en vez de aritmética. En consecuencia, el nivel de intensidad β de una onda sonora se define por la ecuación

$$\beta = 10 \log \frac{I}{I_o} \tag{2.5}$$

donde I_o es una intensidad arbitraria de referencia considerada igual a $10^{-12} W \bullet m^{-2}$ y que corresponde, aproximadamente, al sonido más débil que puede oírse. Los niveles de intensidad se expresan en decibeles, abreviado dB.

Si la intensidad de una onda sonora es igual a I_o o $10^{-12} W \bullet m^{-2}$, su nivel de intensidad es de 0 dB. La intensidad máxima que el oído puede tolerar es de, aproximadamente $1 W \bullet m^{-2}$ que corresponde a un nivel de intensidad de 120 dB. La Tabla 2.1 nos da los niveles de intensidad en decibeles de algunos ruidos comunes.

Origen o Descripción del Ruido	Nivel de Ruido dB's
Umbral de Dolor	120
Máquina Remachadora	95
Tren elevado	90
Calle de tránsito intenso	70
Conversación ordinaria	65
Automóvil silencioso	50
Radio con volumen bajo en casa	40
Conversación en voz baja	20
Murmullo de las hojas	10
Umbral de la sensación auditiva	0

Tabla 2.1 Niveles de intensidad (Tabla23.1 [SEAR88])

Dentro del intervalo de audibilidad, la sensibilidad del oído varía con la frecuencia. El umbral de audibilidad a cualquier frecuencia es la intensidad mínima del sonido que se puede percibir a esa frecuencia.

El término *sonoridad* se refiere a la percepción subjetiva de la magnitud de una sensación sonora. La sonoridad aumenta generalmente con la intensidad, pero debido a la sensibilidad variable del oído, no existe una relación directa entre ambas.

2.2.3 Timbre y Tono

Cuando dos señales contienen exactamente las mismas frecuencias, pero tienen una distribución distinta de intensidad, los sonidos son diferentes y se dice que difieren en *calidad* o *timbre*.

La expresión tono se refiere a las características de una sensación sonora que permite clasificarla como "alto" o como "bajo". Al igual que la sonoridad, es una cantidad subjetiva y no puede medirse con instrumentos. El tono esta relacionado con la cantidad objetiva de frecuencia, pero no existe correspondencia de un tono a otro. En un sonido puro de intensidad constante, el tono sube a medida que la frecuencia aumenta, pero el tono de un sonido puro de frecuencia constante se hace menor con el aumento del nivel de intensidad

2.2.4 Pulsaciones

Las pulsaciones son fluctuaciones de amplitud producidas por dos ondas sonoras de frecuencias ligeramente diferentes. La amplitud varía con una frecuencia llamada frecuencia de pulsación, que es diferente de ambas ondas. Si esta frecuencia es de varios hertz, se percibe como una fluctuación o pulsaciones del sonido.

2.2.5 El Efecto Doppler

Cuando una fuente sonora, una persona que escucha, o ambos, están en movimiento con respecto al aire, el tono percibido por la persona generalmente no es el mismo que cuando la fuente y ésta están en reposo. El ejemplo más conocido es el descenso brusco de tono del sonido emitido por una bocina de un automóvil que se produce al encontrar y sobrepasar a otro que avanza en dirección opuesta. Este fenómeno se llama efecto Doppler.

3. Generación de la Voz y Percepción[1]

Se requiere de un estudio de la anatomía de los órganos del habla como antecedente para la fonética articulatoria y la fonética acústica. Describiré los órganos del habla y discutiré su operación. También es necesaria la comprensión de la audición y de la percepción del habla. Esta última es necesaria en los campos de síntesis y realce de la voz, también es útil en el campo de reconocimiento automático de voz. Después describiré la estructura del oído y lo poco que actualmente es conocido o se cree acerca de la naturaleza de la percepción del habla.

3.1 Organos del habla.

Podemos dividir a los órganos vocales en tres subsistemas principalmente:

1. Pulmones y Traquea.
2. Laringe.
3. Tracto vocal.

La Figura 3.1 muestra estos subsistemas así como sus divisiones:

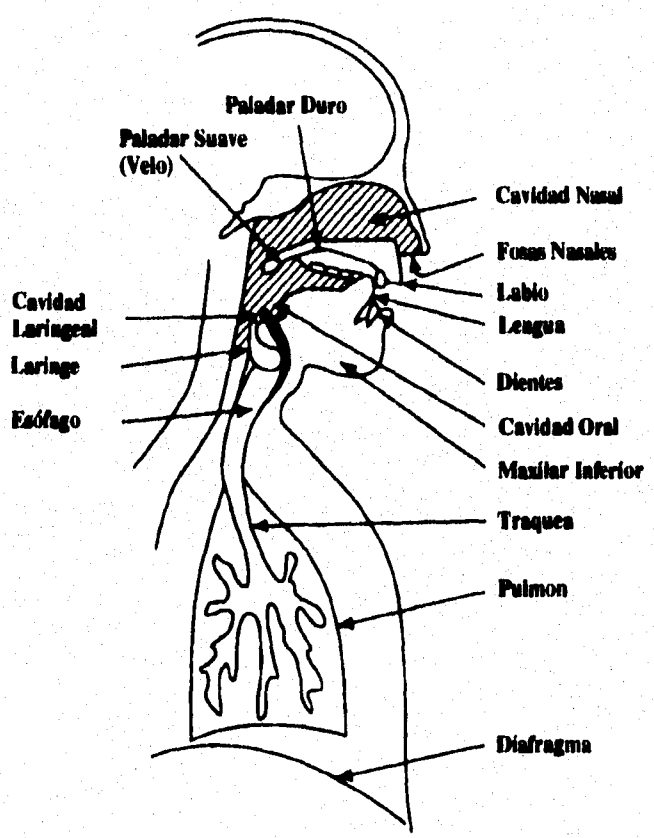


Figura 3.1 Sistema Respiratorio (Fig.2.1 [DELL87]

[1] En su mayor parte resumido del capítulo 3 de [PARS87]

3.1.1 Pulmones y Traquea

Los órganos vocales trabajan con aire comprimido; éste es proporcionado por los pulmones y entregado al sistema por medio de la traquea. Estos órganos también controlan el nivel de intensidad de habla resultante, pero raramente proporcionan una contribución audible a la voz.
La función primaria de los pulmones es intercambiar gases entre el torrente sanguíneo y el aire; se absorbe oxígeno hacia la sangre, se remueve el dióxido de carbono de la sangre y se exhala hacia la atmósfera. Los pulmones son una masa de tejido esponjoso diseñados para ofrecer una gran área sobre la cual se lleva a cabo este intercambio. Su estructura es pasiva, están contenidos dentro de una cámara denominada pleura. La pleura está contenida por las costillas y en la parte inferior por el diafragma. Se introducen o expelen gases por medio de un cambio en el tamaño de la pleura.

El diafragma es un músculo en forma de domo sujeto al fondo de la caja torácica; cuando este músculo se contrae, el domo adquiere una forma plana, el volumen de la pleura aumenta y se introduce aire hacia los pulmones. Cuando se relaja el diafragma, vuelve a adquirir su forma de domo y el proceso se revierte. Cuando se forza a expeler el aire de los pulmones (como cuando se sopla, se grita, se estornuda, o se tose), se requiere de fuerza adicional que es proporcionada por la musculatura abdominal. La inhalación se logra expandiendo la caja torácica. Cuando se realiza esto, se dice que la respiración es torácica. La respiración por el diafragma es llamada abdominal. La respiración normalmente es una combinación de ambas dependiendo de la situación.

El principal requerimiento lingüístico es proporcionar un grado de continuidad en el habla. Por eso hablamos en grupos de respiraciones; la respiración durante el habla, consiste de pequeñas inhalaciones y exhalaciones controladas de larga duración.

La tráquea es un tubo de aproximadamente 12 cm de largo por 2 cm de ancho, une a los pulmones y a la laringe. Esta conformada de anillos de cartílago unidos por tejido conectivo. Esta construcción proporciona un tubo que es rígido en la sección transversal pero que se puede doblar y torcer fácilmente en respuesta a movimientos de la cabeza. En la parte inferior, la traquea se bifurca hacia los bronquios derecho e izquierdo. La traquea y los pulmones constituyen el tracto pulmonar.

3.1.2 Laringe

Este es un sistema muy complejo de cartílagos y músculos que contienen y controlan a las cuerdas vocales. Su principales partes son:

Cartílago Cricoides
Cartílago Tiroides
Cartílago Aritenoides
Cuerdas vocales

Los cartílagos cricoides y tiroides, son básicamente estructuras de apoyo. El cartílago cricoides es esencialmente otro de los anillos que componen la tráquea, pero se encuentra en la parte posterior alta de forma que soportan la parte trasera de las cuerdas vocales. El cartílago tiroides esta localizado al frente, aproximadamente en el otro lado de la parte alta del cartílago cricoides. La forma del cartílago tiroides esta diseñada con el fin de darle fuerza para resistir la tensión de las cuerdas vocales. La forma de domo, es visible al frente de la garganta como "la manzana de Adán".

Figura 3.2 Laringe. Vista Anterior (Fig 7.4 [DIEN76]).

La cuerdas vocales, son músculos, plegados entre la parte anterior y la posterior de la laringe.

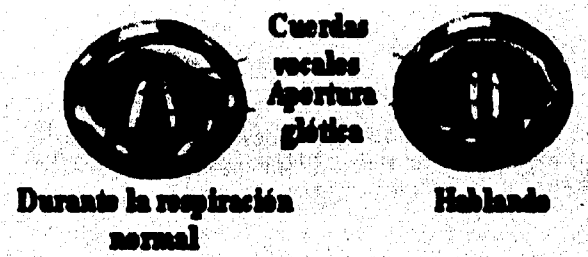


Figura 3.3 Vista superior de la caja vocal (Fig. 7.5 [DIEN76])

Las terminaciones anteriores están soportadas por el cartílago tiroides y las terminaciones posteriores por el cartílago aritenoides, que a su vez están conectados a la parte alta del cartílago cricoides. Estos cartílagos están controlados por un conjunto de músculos unidos al cartílago cricoides y pueden mover los extremos de las cuerdas vocales para juntarse o separarse. Cuando los extremos de las cuerdas se separan, se encuentran abiertas, ésta es la posición normal para la respiración (La separación entre las cuerdas se le denomina glotis). Cuando los extremos están juntos, las cuerdas se encuentran cerradas y proporcionan un sello en la parte alta del tracto pulmonar.

Las funciones de la cuerdas vocales son biológicas y acústicas. La función biológica es la de cerrar la tráquea, para proteger el tracto pulmonar o para permitir que se cree presión dentro del tórax y del abdomen. El tracto pulmonar es protegido mediante la epiglotis durante la deglución; ésta impide que las partículas de comida entren al tracto pulmonar; mediante el acto de toser, la repentina apertura de las cuerdas vocales, provoca un flujo de aire que se mueve rápidamente, de forma que se puedan desalojar partículas de la tráquea.

La función acústica de las cuerdas vocales es proporcionar la fuente principal de excitación para la voz.

3.1.3 El Tracto Vocal

Normalmente este término implica todo lo que se encuentra después de las cuerdas vocales. Se puede observar la estructura general en un corte sagital transversal en la Figura 3.4. El tracto vocal convencionalmente se divide en la siguientes regiones:

3. Laringofaringe (detrás de la epiglotis).
2. Bucofaringe (detrás de la lengua, entre la epiglotis y el velo).
3. Nasofaringe (arriba del velo, en la parte trasera de la cavidad nasal).
4. Cavidad Oral (adelante del paladar y terminado en los labios, lengua y paladar).
5. Cavidad Nasal (arriba del paladar y se extiende de la faringe hasta las ventanas nasales).

Ademas, el tracto vocal se encuentra limitado por las siguientes estructuras:

1. Epiglotis.
2. Maxilar Inferior.
3. Lengua.
4. Velo.
5. Paladar.
6. Dientes.
7. Labios.

La epiglotis es una placa de cartílago que yace arriba de las cuerdas vocales y detrás de la lengua. El maxilar inferior, o mandíbula, se usa para masticar y también soporta la terminación frontal de la lengua.

El techo de la boca se puede dividir en dos regiones principales. Al frente, está formado por un hueso llamado paladar que separa la boca de las cavidades nasales y soporta los dientes superiores. En la parte posterior del paladar, el techo se forma de músculo y tejido conectivo; ésta estructura se le llama velo o paladar blando. La úvula es un apéndice que está al final del velo. El velo puede ser levantado por un músculo y presionado contra la pared posterior de la faringe para sellar los conductos nasales del resto del tracto vocal. Al frente del paladar existe una cresta, formada por el ensanchamiento del hueso donde están insertados los dientes frontales; a este se le llama cresta alveolar. La lengua es un sistema de músculos conectados al frente al maxilar inferior y al dorso a los huesos en la garganta y la cabeza.

Las funciones del tracto vocal, como las de la laringe, son también biológicas y vocales. La biológicas incluyen la respiración, el olfato, el gusto, el masticar y el deglutir. La función vocal es la coloración y articulación de la voz; el tracto vocal también contiene los puntos principales de los cuales se irradia el habla. Con la excepción de la epiglotis, todas estas partes están involucradas en el habla. Los principales participantes en la coloración y la articulación son la lengua, los labios y el maxilar inferior.

El tracto vocal en un hombre adulto es de aproximadamente 17 cms. de longitud. Durante el trayecto de la onda acústica a través del tracto vocal, se afecta su contenido en frecuencia por la resonancia de las cavidades. Mediante el movimiento de la lengua, podemos alterar la forma de la cavidad oral y de la faringe, también podemos desacoplar la cavidad nasal del sistema mediante la elevación del velo de tal forma que selle la cavidad nasal de la faringe.

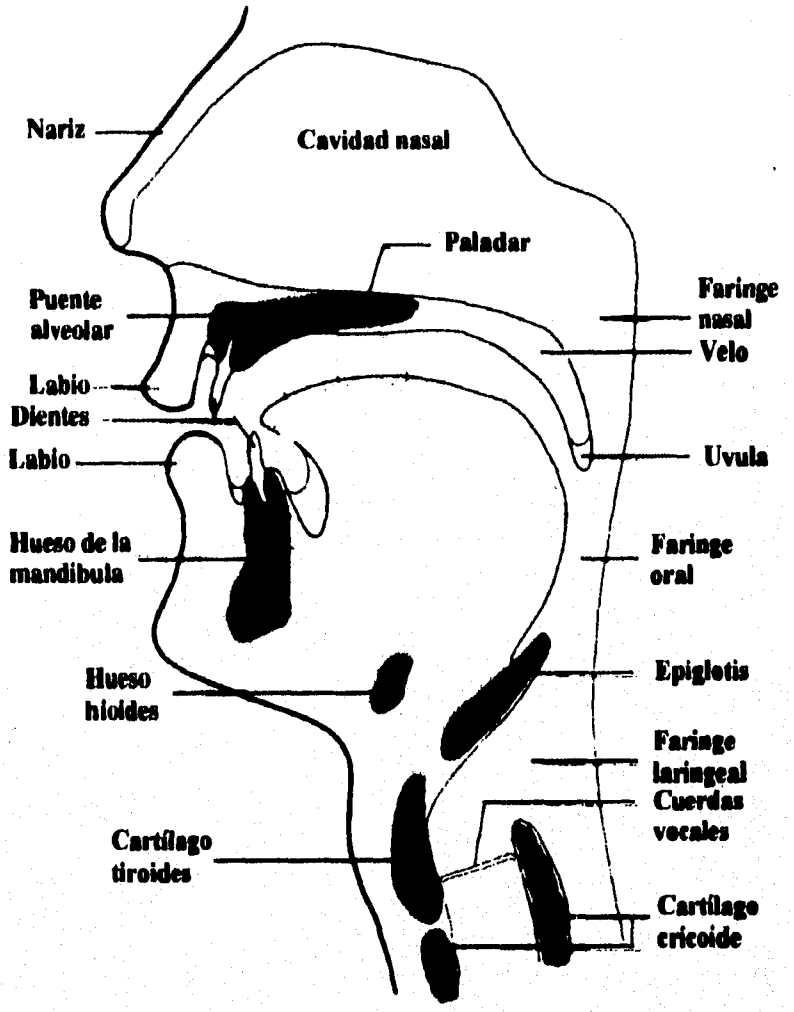


Figura 3.4 Corte sagital de una cabeza humana mostrando los principales órganos del habla (Fig. 3.3 [PARS87]

3.2 Producción de la Voz

La operación del sistema como un todo se divide básicamente en dos funciones: excitación y modulación como se muestra en la Figura 3.5. La excitación toma lugar prácticamente en la glotis pero también en otros puntos; la modulación se realiza por los distintos órganos del tracto vocal.

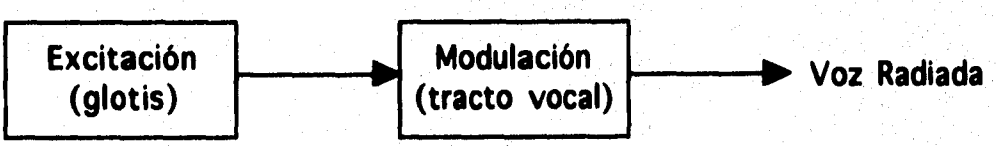


Figura 3.5 Modelo de excitación-modulación de la producción de voz (Fig3.4 [PARS87])

3.2.1 Excitación

Se realiza en diferentes formas: fonación, susurro, fricación, compresión y vibración.

3.2.1.1 Fonación

Esta es la fuente de excitación mas importante, es la oscilación de las cuerdas vocales. El cartílago aritenoides se cierra y se estiran las cuerdas vocales como se muestra en la Figura 3.6a. Cuando se forza aire a través de las cuerdas vocales, estas vibran. La oscilación es gobernada por la masa y la tensión de las cuerdas y también por el efecto Bernoulli del aire que pasa a través de estas. El abrir y cerrar de las cuerdas, corta el flujo de aire y genera una serie de pulsos como se muestra en la Figura 3.6c. La forma y el ciclo de trabajo de estos pulsos depende de las circunstancias (intensidad, timbre, etc); A la razón de repetición de los pulsos se le denomina timbre. Este es controlado principalmente por la tensión de las cuerdas vocales y regulado por la retroalimentación a través de los oídos y el cerebro. Existen diferentes modos de vibración. Los cantantes llaman a estos modos "registros". Para niveles bajos de presión del aire, las oscilaciones pueden volverse irregulares, ocasionalmente decayendo el timbre una octava al ir disminuyendo la razón de repetición a la mitad, o con pulsos que se generan en pares. Los sonidos vocales acompañados de una fonación se les denomina voceados o sonoros; a los que no se realizan a través de la fonación se les denomina no voceados o sordos.

3.2.1.2 Susurro

En éste, las cuerdas vocales están juntas, pero con una pequeña apertura triangular entre los cartílagos aritenoides como se muestra en la Figura 3-5b. El aire que fluye a través de este orificio, genera turbulencia, que causa ruido de banda ancha que sirve como señal de excitación.
Las demás fuentes de sonido, sirven como modulación así como excitación, ya que generalmente se perciben como interrupciones de la voz, formando consonantes y delimitando las sílabas.

3.2.1.3 Fricación

Si el tracto vocal es obstruido en cualquier otro punto, el flujo de aire que pasa por la obstrucción es turbulento y genera ruido de banda ancha cuyo espectro de frecuencias refleja la localización de la constricción. Los sonidos que se producen de esta forma son llamados fricativos o silbantes. La fricación puede ocurrir con o sin fonación.

3.2.1.4 Compresión

Si el tracto vocal es completamente obstruído en cualquier punto, mientras el locutor continúa exhalando, se genera una presión, que cuando es liberada, ocurre una pequeña explosión. La combinación de un silencio corto seguido de una pequeña ráfaga de ruido tiene un sonido característico. Si la liberación es abrupta y limpia, el sonido es una oclusiva o explosiva; si es gradual y turbulento, el sonido puede pasar como fricativa y se le denomina africada.

3.2.1.5 Vibración

Si se forza aire a través de cualquier oclusión que no sean la cuerdas vocales, pueden generarse vibraciones, especialmente en la lengua o en la úvula y ocasionalmente entre los labios.

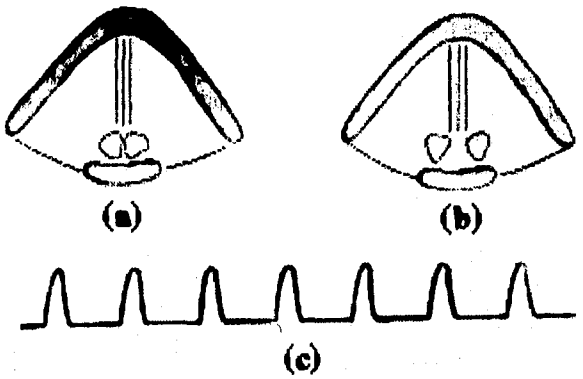


Figura 3.6 Posición de las cuerdas vocales y los cartílagos.
a) Fonación. b)Susurro. c)Pulso glotal típico. (Fig. 3.6 [PARS87])

3.2.2 Modulación

Esto es lo que realizamos para imponer información en la salida glotal. La modulación se puede analizar desde el punto de vista fisiológico;como se impone información articulatoria sobre el sonido y desde un punto de vista acústico; que hacen los órganos vocales a la señal que se emana de la glotis.

El primer punto nos lleva al dominio de la fonética articulatoria: como los órganos del habla se posicionan para producir la voz. El segundo punto nos lleva al campo de la fonética acústica: cuales son las mediciones de correlación de un sonido de voz dado y cómo corresponden las características acústicas en general a la fonética articulatoria.

Fisiológicamente el sonido es modulado por el movimiento de los órganos del habla (principalmente la lengua) de forma que se cambia la calidad de la voz y para interponer sonidos adicionales o interrupciones a la voz.

Acústicamente, la forma principal de modulación es la operación de filtrado. La forma de onda glotal es rica en armónicas y el tracto vocal como cualquier tubo acústico, tiene frecuencias naturales que son función de su forma. Estas frecuencias naturales son denominadas formantes, y son la forma principal de modular la voz. Los formantes, son tomados en cuenta para todas las vocales y algunas consonantes y también se sabe que proporcionan información importante acerca del resto de las consonantes. Otras fuentes de modulación, son las interrupciones, las obstrucciones y las ráfagas de ruido de banda ancha que forman las consonantes.

3.3 Audición y Percepción

Por audición, nos referimos al proceso por el cual se recibe un sonido y se convierte en impulsos nerviosos; por percepción nos referimos al postprocesamiento dentro del cerebro por el cuál los sonidos escuchados son interpretados y se les da un significado.

3.3.1 Audición

El oído se divide en tres partes: el oído externo, el oído medio y el oído interno.

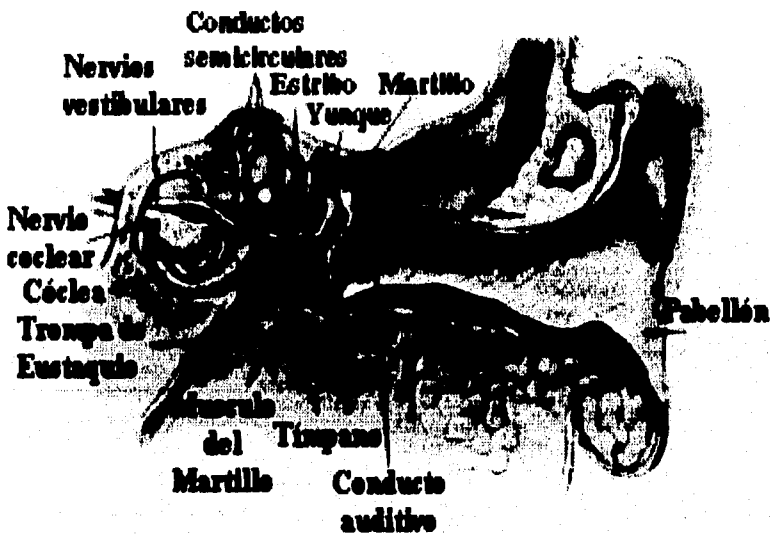


Figura 3.7 Oído (Fig. 5.27 [DIEN76]).

El oído externo consiste de el pabellón, el conducto auditivo externo y el tímpano. El pabellón protege la entrada; el conducto auditivo externo, es un tubo uniforme de aproximadamente 2.7 cms. de largo por 0.7 cms de ancho por el cual pasa el sonido hasta alcanzar el tímpano. Como todos los tubos, contiene una serie de frecuencias resonantes, de las cuales sólo una aproximadamente a 3 KHz cae dentro del rango de la voz. La membrana timpánica es una estructura cónica rígida al final del conducto auditivo externo; vibra en respuesta al sonido y es el primer enlace en una cadena de estructuras que transmiten el sonido a los transductores neuronales en el oído interno.

El oído medio es una cavidad llena de aire que se encuentra separada del oído externo por la membrana timpánica y conectada al oído interno por dos aperturas llamadas las ventanas oval y redonda. El oído medio, también esta conectado al mundo exterior por medio de la trompa de Eustaquio, que permite la igualación de la presión de aire entre el oído medio y la atmósfera que lo rodea.

El oído medio contiene tres huesos diminutos que proporcionan el acoplo acústico entre la membrana timpánica y la ventana oval. Estos huesecillos son llamados martillo, yunque y estribo. El martillo esta unido a la membrana timpánica, el estribo a la ventana oval y el yunque conecta a ambos. Estos huesos tienen dos funciones: transformación de impedancias y limitación de amplitud.

La transformación de impedancia proporciona una transferencia de energía acústica de forma más eficaz del aire al oído interno que se encuentra lleno de líquido. La transformación tiene dos componentes, la ganancia mecánica del enlace de los huesecillos y la razón de área de la membrana timpánica a la ventana oval. El incremento en la impedancia total es de aproximadanente 15 a 3.

La limitación de amplitud, protege el oído de altos niveles de sonido. Esta es realizada por los músculos del oído interno, que mediante contracción, pueden atenuar la transmisión a través de los huesecillos. A este efecto se le denomina reflejo timpánico. Específicamente, cuando se detecta una intensidad fuerte de sonido, el músculo del estribo se contrae de forma que cambia la forma de vibración del estribo en una dirección que provoca una reducción en la excitación a la ventana oval.

El oído interno, consiste del aparato vestibular, las ventana oval, la ventana redonda y la cóclea. El aparato vestibular abarca los canales semicirculares y los órganos asociados, utilizados para el balance y el sentido de la orientación.

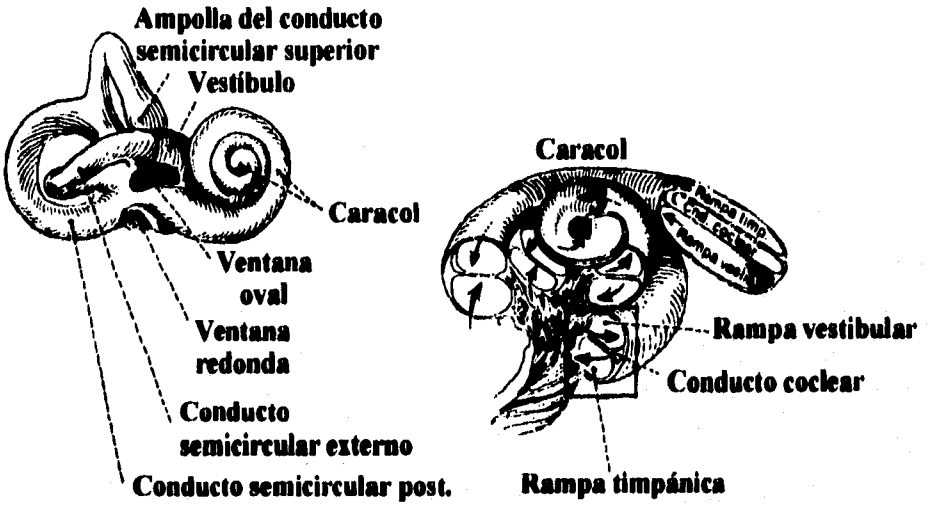


Figura 3.8 Cóclea; (Fig. 5.28 [DIEN76]).

La cóclea es un conducto óseo en forma de caracol que comunica al oído medio a través de las ventanas oval y redonda. Contiene transductores que convierten vibraciones acústicas en impulsos nerviosos. La cóclea esta dividida a la mitad por la membrana basilar y la membrana de Reissner. Esta partición divide a la cóclea en dos compartimientos, la escala vestibular y la escala timpánica.

La energía acústica entra por medio de la ventana oval, que es manejada por el estribo. El sonido viaja a través de la escala vestibular, pasa a la escala timpánica y sale por medio de la ventana redonda. La membrana basilar vibra en respuesta a este sonido.

El órgano de Corti y la membrana tectorial corren a lo largo de la membrana basilar. El órgano de Corti esta en la membrana basilar; la membrana tectorial está montada en una proyección de la pared de la cóclea y queda justo arriba del órgano de Corti. Este último contiene células ciliadas que abarcan la separación entre el órgano de Corti y la membrana tectorial. Estas células ciliares sensan la vibración de la membrana. Junto a la base de cada célula ciliar esta una sinapsis nerviosa por la cual se pasa información frecuencial hacia el cerebro.

La función de esta estructura es la de producir una dispersión espacial de componentes de frecuencia a lo largo de la membrana basilar. La cóclea actúa como un analizador de espectros mecánico neuronal. La membrana basilar parece ser el mecanismo primario para análisis en frecuencia del sonido.

Los nervios de la cóclea convergen en un nodo denominado el núcleo coclear. Los impulsos pasan de los núcleos cocleares a un segundo grupo de núcleos del bulbo. Se cree que se realiza procesamiento adicional en mas de una etapa intermedia a lo largo de la trayectoria. Existen muchos cruces en diversos puntos a lo largo de esta trayectoria, dando como resultado que el trayecto mas importante de cada oído lo lleva al hemisferio opuesto del cerebro.

3.3.2 Percepción

Primero consideraremos el desempeño de los órganos auditivos y después tomaremos lo que se sabe acerca de la percepción del habla.

3.3.2.1 Desempeño

El rango de frecuencias es de aproximadamente 20Hz a 20KHz. En el extremo inferior del rango de frecuencias, el sonido percibido se convierte en un tren de pulsos; en el extremo superior se desvanece hasta el silencio. El rango de intensidad es de 0 a 120 dB (el nivel de referencia es de 10^{-16} W/cm^2 o 0.0002 dinas/cm^2). En el extremo superior el sonido se vuelve doloroso; en el extremo inferior, se vuelve silencio.

3.3.2.2 Sonoridad

La sonoridad percibida es una función de la frecuencia y del nivel. Comparando diferentes tonos a diferentes frecuencias y amplitudes, se han detectado niveles iguales de sonoridad en forma subjetiva. La unidad de nivel de sonoridad cuando se compensa para dependencia de frecuencia es el *phon*.

3.3.2.3 Timbre

Entre los músicos y los físicos, el timbre se define relativo a alguna frecuencia de referencia. La diferencia en el timbre entre dos notas es igual a 1200 veces el logaritmo base 2 de su razón de frecuencias. La unidad del timbre es el mel.
El mel es una unidad de medida del timbre percibido o la frecuencia de un tono. Se sugiere la siguiente aproximación:

$$F_{mel} = 1000 \bullet \log_2\left[1 + \frac{F_{Hz}}{1000}\right]$$

Donde $F_{mel}(F_{Hz})$ es la frecuencia percibida (real) en mels (Hz).

3.3.2.4 Sonidos Complejos Periódicos

Las armónicas no son escuchadas como tonos separados, en su lugar, el conjunto se escucha como un tono simple cuyo timbre está en la fundamental; la presencia de armónicas superiores se percibe como dando al sonido una "calidad de tono" o "timbre". El fenómeno perceptual del timbre es importante ya que las vocales de distinguen por su contenido armónico.

En tonos complejos, el timbre es percibido aunque la fundamental no se encuentre presente. Por ejemplo, el timbre de una voz de hombre, que normalmente esta por debajo de los 120Hz, se escucha claramente sobre el sistema telefónico en el cual la respuesta en frecuencia corta abajo de los 300Hz. No es claro como se reconstruye la fundamental en el cerebro. Las siguientes teorías se han propuesto para explicar este efecto:

1. Se genera una fundamental por medio de no linealidades en el oído.
2. Se deriva la fundamental por postprocesamiento en el cerebro.
3. La fundamental resulta de la coherencia de fase en los disparo de neuronas en respuesta a las armónicas.

3.3.2.5 Enmascaramiento

Este es el fenómeno en el cual un sonido interfiere con nuestra percepción de otro[JEFF70]. El grado de enmascaramiento es una función de los niveles relativos y las frecuencias. Los tonos

cercanos son enmascarados en una forma mayor que los tonos de frecuencias que difieren ampliamente. Fletcher y Munson (1937) descubrieron que cuando un tono puro (sinusoide) era enmascarado por ruido de banda ancha, solamente una pequeña banda centrada cerca del tono, contribuía al efecto de enmascaramiento. A esto le denominaron *bandas críticas* [SCHA70]. Desde entonces, las bandas críticas han demostrado estar relacionadas a un gran numero de efectos perceptuales dependientes de la frecuencia. Estas son de ancho constante dentro de la escala mel. El ancho de una banda critica varía desde 100 mels a frecuencia central de 50 mels hasta aproximadamente 250 mels a una frecuencia central de 3600 mels.

La importancia de estas bandas recae en el efecto que los tonos tienen sobre el oído humano y que éste tiende a promediar frecuencias dentro de la misma banda haciendo que parezcan un tono de la frecuencia central de la banda crítica.

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

Tabla 3.1 Ejemplos de Ancho de las Bandas Críticas (Tabla 1 [SCHA70]).

3.3.2.6 Percepción del Habla

La principal pregunta a tratar por las teorías de la percepción del habla es el cómo la entrada acústica al oído es traducida por el cerebro en sonidos de voz.

Aparentemente el cerebro hace una distinción fundamental entre sonidos de voz y sonidos diferentes a la voz. Parece que el cerebro procesa en forma diferente los sonidos de voz que los que

no son de voz. Se cree que el centro de procesamiento de voz está localizado en el hemisferio izquierdo del cerebro.

También se cree que existe una predisposición natural del sistema nervioso humano para decodificar entradas de voz. El cerebro tiende a imponer una categorización en los sonidos de voz. Se piensa que ésta es realizada por un preprocesador que separa las señales de voz de las de no voz, separa la frecuencia fundamental de la coloración espectral debida a componentes armónicos y genera un conjunto de características. Entonces un procesador central realiza el reconocimiento utilizando estas características como entrada. El procesador asume otros conocimientos ademas de los datos de entrada: el conocimiento del lenguaje por parte del que escucha, las características del locutor y un tema que se discute es el de la combinación de las características en un todo para corregir posibles errores que pueden resultar de datos ambiguos.

La percepción es influenciada por el contexto. Se cree que existen claves léxicas, gramaticales y semánticas como ayuda en la verificación del correcto análisis de la señal de voz que está entrando. Conforme vamos escuchamos al locutor, mentalmente duplicamos lo que dice, siguiéndolo y si es posible anticipándonos. Este proceso se conoce como análisis por síntesis. Tal modelo requiere continuidad y coherencia en el habla, bajo esta hipótesis, lo que no tiene sentido es mas difícil de entender por que hace imposible mantener el modelo interno. Esta teoría de modelar internamente el habla percibida ha tenido una fuerte influencia en el desarrollo de sistemas de "entendimiento del habla".

La teoría motriz, sugiere que el habla es percibida en términos de la articulación. Esto es, la mente analiza el habla creando una simulación mental del proceso articulatorio de generación de la voz.

Ambas teorías requieren de un conocimiento del lenguaje y del proceso articulatorio por parte del que escucha.

Por ejemplo Cole y Jakimik (1980), en base a muchos experimentos en percepción, han desarrollado la siguiente teoría:

1. El conocimiento del lenguaje y del ambiente, generalmente juega una parte significativa de la percepción del habla. El habla no es reconocida por el puro análisis de la señal acústica por separado; tal análisis no es adecuado para remover ambigüedades del mensaje o para clarificar sonidos de habla no claros.
2. La voz de entrada es procesada una palabra a la vez, en el orden en que las palabras son recibidas. Se extrae información de cada una y ésta se utiliza para guiar en el análisis de las palabras subsecuentes.
3. En el reconocimiento de una palabra en particular, los sonidos que la componen son procesados en el orden de aparición, y el sonido nuevo es utilizado para reducir el número de posibilidades. La parte inicial de una palabra y particularmente la primera sílaba, es la que recibe la mayor atención por parte del que escucha. Tan pronto como se ha recibido la suficiente información para excluir todas menos una palabra, se ha realizado el reconocimiento y el que escucha pone atención superficial al resto de la palabra.

Este modelo se basa en experimentos sobre la habilidad del que escucha de detectar malas pronunciaciones de palabras; los experimentos normalmente miden el número detectado de éstas, el tiempo de reacción en su detección o la habilidad del que escucha de corregirlas. Esta afirmación se basa en que el desempeño dependerá en (1) que atención se presta a la palabra por parte del que escucha y (2) que tan esencial es la parte mal pronunciada para el reconocimiento.

El reconocimiento en este modelo es esencialmente un proceso de abajo hacia arriba, comenzando por los datos y trabajando hacia la estructura, esto va de forma contraria a un proceso de arriba hacia abajo en el cual se comienza con una sentencia hipotética y se trata de ajustar los datos a ella.

4. Fonética Articulatoria y Fonémica[1]

El principal objetivo de la fonética es el de proporcionar una descripción exacta sin ambigüedades de cada sonido conocido de voz. Se sabe que los dominios de la fonética son la anatomía y la física; entonces la fonética es independiente de cualquier lenguaje en particular. El término fonémica[2] es utilizado para el estudio de los sonidos de voz tal y como son percibidos y pensados por los hablantes de un lenguaje en particular.

La fonética articulatoria considera como es producido cualquier sonido, con un especial énfasis en los detalles anatómicos.

La fonética acústica, hace énfasis sobre las características observables que pueden ser medidas en las formas de onda de los sonidos de voz, especialmente aquellos que permiten ser distinguidos uno de otro. Un objetivo importante es el de relacionar estas características acústicas a sus correspondientes posiciones en los órganos del habla. La fonética acústica proporciona bases tanto teóricas como experimentales para el reconocimiento y síntesis de voz por medios electrónicos.

En este capítulo se hace un análisis tanto del idioma inglés como del español, ya que este trabajo se pretendía hacer en español, pero no se cuenta con una base de datos completa y con las características acústicas necesarias que permitan desarrollar la aplicación basados en ésta.

4.1 Fonética Articulatoria

La primera tarea de la fonética articulatoria es la de describir los sonidos de voz en términos de las posiciones del los órganos vocales cuando se produce un sonido. Un objetivo importante es el de proporcionar una notación común y un marco de referencia para que un lingüista pueda entender a otro y pueda reproducir con precisión cualquier expresión hablada que haya sido escrita en una "transcripción fonética detallada".

4.1.1 Alfabetos Fonéticos

Existen un número bastante grande de sonidos de voz diferentes, más de los que puede abarcar cualquier alfabeto. De aquí que los fonetistas han tenido que crear su propio sistema de notación. La notación más antigua y utilizada es el alfabeto fonético internacional (IPA[3]). Data de una época cuando la tipografía se realizaba a mano y deriva muchos de sus símbolos de la impresión de caracteres romanos de cabeza o tomándolos de alfabetos de otros lenguajes. Estos símbolos no pueden ser reproducidos en la mayoría de las impresoras de computadora; de aquí que en años recientes se ha desarrollado un substituto denominado "Arpabeto[4]". El Arpabeto se puede encontrar en dos versiones; de caracter sencillo que utiliza caracteres en minúscula para algunos sonidos y la versión de caracteres dobles que sirven en impresora que carecen de letras minúsculas.

Al final del capítulo se muestra la Tabla 4.3 que contiene los alfabetos fonéticos tanto para el inglés como para el español. También se muestran las similitudes y diferencias entre ambos lenguajes, mediante una comparación de los sonidos que existen en cada uno.

---

[1] En su mayor parte tomado del capítulo 4 de [PARS87].

[2] También llamada Fonología o Fonemática [QUIL79]. Los elementos fónicos que estudia son los fonemas.

[3] Por sus siglas del Ingles "International Phonetic Alphabet".

[4] Su nombre se deriva de la "Advanced Research Projects Agency (ARPA)" del Departamento de Defensa de los EUA que ha auspiciado una parte importante del desarrollo en la investigación de la voz.

4.1.2 Categorías

La división convencional de los sonidos de voz es en vocales y consonantes. A pesar de que estos términos también son usados en fonética, es difícil definirlos precisamente. Algunos autores[5] utilizan los términos vocoide y contoide, Las vocoides están caracterizadas por fonación y el tracto vocal prácticamente sin obstrucción y su característica más importante es el color del tono impuesto por resonancias del tracto vocal. Las contoides se caracterizan por la obstrucción del tracto vocal; la fonación no es importante y la característica más importante es la turbulencia audible o cualquier otra interrupción del flujo de la voz.

4.1.3 Contoides y Consonantes

Las consonantes están definidas en términos anatómicos. La mayoría de las consonantes están descritas por características bien definidas, éstas principalmente son:

1. Punto de articulación
4. Forma de articulación
3. Sonoridad (Fonación)

Punto de articulación. Este es la ubicación de la principal obstrucción en el tracto vocal, definida en términos de los órganos participantes la Tabla 4.1 es una lista de los principales puntos de articulación y los nombres dados a la consonantes que corresponden:

Nombre	Descripción
Bilabial‡	Entre los labios
Labiodental‡	Entre el labio inferior y los dientes superiores
Apicodental†	La punta de lengua en los dientes
Apicogingival†	La punta de la lengua en las encías
Apicoalveolar†	La punta de la lengua en la cresta alveolar
Apicodomal	La punta de la lengua en el paladar rígido
Laminoalveolar	El filo de la lengua en la parte posterior del paladar rígido
Laminodomal	El filo de la lengua en el paladar rígido
Centrodomal	La mitad de la lengua en la parte posterior del paladar rígido
Dorsovelar	La parte posterior de la lengua en el velo
Faringeal	La raíz de la lengua obstruyendo la faringe oral
Glotal	Entre las cuerdas vocales

Tabla 4.1 Puntos principales de articulación (Tabla 4.2 [PARS87]).

‡ También se les conoce como "Labiales"
† también se les conoce como "Dentales"

En inglés los fonemas consonánticos son los siguientes:

Bilabiales m
Labiodentales f,v
Interdentales T,D

[5] K. Pike, *Phonetics*, University of Michigan Press, Ann Arbor, 1966.
C. Hocket, A Manual of Phonology, Mem. 11, Indiana University Publications in Anthropology and Linguistics, 1955.
L. Brosnahan and B. Malmberg, *Introduction to Phonetics*, Cambridge University Press, Cambridge 1970.

Glotal	Q,h
Alveolares	n,t,d,s,z,l,r
Palatales	y
Velares	k,G,g

En español los fonemas consonánticos son los siguientes:

Sordos	
Bilabiales	p,b,m
Labiodentales	f
Interdentales	z
Dentales	t,d
Alveolares	s,l,n,r,rr
Palatales	ll,ñ,ch,y
Velares	k,g,j

Forma de articulación. Esta es principalmente el grado de obstrucción en el punto de articulación y la forma en que se libera hacia el siguiente sonido. Las vibrantes corresponden a la modulación por vibración. Heffner (1950) identificó cuatro tipos de vibrantes: laringeal, uvular, dental y labial. El uvular es la "r" francesa o prusiana, el dental es la "rr" española. El dental de solo un periódo se le denomina soplido; en inglés ocurre en la "dd" de algunas pronunciaciones como en la palabra "ladder".

Nombre	Descripción
Explosiva	El tracto vocal se cierra en el punto de articulación; los conductos nasales se cierran en el velo. Las explosivas tienen una liberación abrupta y limpia. También se les llama pausada.
Aspirada	El tracto vocal inicialmente se encuentra cerrado como en las explosivas; la liberación esta marcada por un soplo de aire antes que se genere el siguiente sonido.
Africada	Tracto vocal inicialmente cerrado seguido de una liberación gradual que produce turbulencia.
Fricativa	El tracto vocal está parcialmente abierto en el punto de articulación, el velo esta cerrado. Se crea ruido turbulento en el punto de articulación. También llamadas aspiradas o silbantes.
Lateral	El tracto vocal está cerrado en el punto de articulación pero abierto en los laterales.
Semivocal	El tracto vocal está parcialmente abierto en el punto de articulación sin turbulencia. (Estas son las consonantes vocoides; "w" y "j" caen dentro de esta categoría.
Nasal	El tracto vocal esta cerrado en el punto de articulación; el velo esta abierto.
Vibrante	Abertura y cierre oscilatorio en el punto de articulación.

Tabla 4.2 Principales categorías de la articulación (Tabla 4.3 [PARS87]).

Sonoridad. Esto indica la presencia o ausencia de fonación. Aún durante una pausa, es posible forzar aire a través de las cuerdas vocales por un periódo de tiempo corto. Las consonantes acompañadas de sonoridad se les denomina sonoras, a las que no se les denomina sordas.

Estas características "sonoridad", "punto de articulación" y "forma de articulación", proporcionan una terminología en la que podemos definir cualquier contoide.

Por ejemplo:

[b]	Sonora bilabial pausada.
[d]	Sonora dental explosiva.
[tʃ]	Sorda apicoalveolar africada
[θ]	Sorda apicodental aspirada
[ʔ]	Pausa glotal

Se debe de notar que la descripción de una consonante es virtualmente una fórmula que produce el sonido; es muy específico y esta relacionado directamente con la anatomía del tracto vocal.

4.1.4 Vocoides y Vocales

La primer característica de las vocales se refiere a que no están tan bien definidas como las consonantes. Esto es debido a que la lengua nunca toca otro órgano cuando produce una vocal; de aquí que no existen formas de definirlas mediante los puntos de articulación y las vocales están definidas vagamente en términos de "posición". Las vocales pueden ser descritas por estas variables:

1. Lengua en alto o bajo.
4. Lengua al frente o al dorso.
3. Labios redondeados o no redondeados.
4. Nasalizadas o no nasalizadas.

Alto y bajo se refieren a la posición de la parte más alta de la lengua. Frontal es hacia los labios y dorsal es hacia la faringe. En vocales nasalizadas, el velo está abierto de tal forma que el sonido pasa hacia la cavidad nasal así como por la boca; en vocales no nasalizadas, el velo esta cerrado y el sonido pasa solamente a través de la boca.

Diagramas vocálicos. Si consideramos solamente la posición de la lengua, podemos asignar las posiciones arriba/abajo y adelante/atrás por medio de un diagrama como el de la Figura 4.1. El espacio generado por estas dos dimensiones es llamado el espacio vocálico. Cuando movemos la lengua de (frente, alto) a (frente, bajo), existe la sensación de que la lengua se hace un poco para atrás; más aún, en la posición (frente, alto), la parte más alta de la lengua tiende a tener forma de pico, pero cuando esta (frente, bajo), tiende a ser redonda. Estas consideraciones, nos explican la línea de pendiente en el diagrama vocálico. Para poder hacer que este diagrama sea útil, debemos de mostrar algunos puntos de referencia. Esto se demuestra en el diagrama; los símbolos son los caracteres IPA para las vocales. Las vocales en los extremos se les denomina "vocales cardinales". Para pronunciar éstas, la posición de la lengua y de los labios debe de exagerarse al punto de que físicamente sea inconfortable.

Los fonetistas algunas veces distinguen entre vocales tensas y relajadas. Las vocales [i,e,o,u] son denominadas tensas, y las vocales [ɪ,ɛ,ɔ,ʊ] son denominadas relajadas.
Para producir una vocal tensa es necesario generar mayor tensión muscular y alta presión en la respiración.

Nasalización. En la pronunciación de cualquier vocal, el velo puede estar abierto o cerrado. Si el velo esta cerrado, la cavidad nasal está desconectada del sistema y el sonido vocálico está

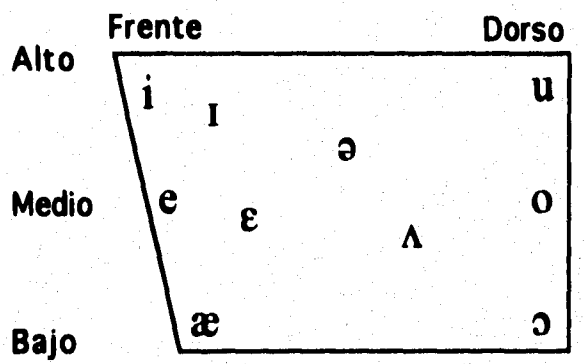
determinado exclusivamente por la posición de la lengua y los labios. Si el velo está abierto, el sonido pasa a través de la cavidad nasal. Esta cavidad tiene una acústica propia y por lo tanto le proporciona una característica de coloración a la vocal. A estas vocales se les denomina nasalizadas.

Diptongos. Es posible combinar dos sonidos vocálicos en una sola sílaba moviendo la lengua de una posición a otra. A esta combinación se le denomina diptongo.

Coarticulación. Todo lo que se ha mencionado acerca de la fonética tiene un grave error, sugiere que cada fonema es ejecutado perfectamente y uniformemente y de forma independiente al contexto. Si esto fuera cierto, el aprender un nuevo lenguaje y la síntesis y reconocimiento de voz serían tareas simples. De hecho ningún sonido de voz es producido fielmente en el contexto de otros sonidos. Cada fonema puede ser considerado como un objetivo al cual los órganos vocales pretenden llegar, pero nunca lo alcanzan. Tan pronto y como se acercan al objetivo, lo suficiente para hacerlo inteligible al que escucha, los órganos cambian su destino y se dirigen a un nuevo objetivo. Esto se realiza con el objeto de minimizar el esfuerzo requerido al hablar y permitir mayor fluidez.

En la mayoría de los casos, la producción de un fonema incluirá algunas características articulatorias provenientes del fonema anterior y algunas características anticipatorias al siguiente fonema.

Al traslapamiento de características fonéticas de un fonema a otro, se le denomina coarticulación. Este fenómeno se adiciona a los problemas de síntesis y reconocimiento de voz. Ya que la coarticulación ocurre en forma natural, cuando se genera voz sin coarticulación, el sonido generado no suena natural.



.cl.**Figura 4.1** Diagrama básico de las vocales (Fig. 4.1 [PARS87]).

4.2 Fonémica

La fonética es una visión de los sonidos del habla considerados aislados de cualquier lenguaje, la fonémica es la visión dentro de algún lenguaje en particular. La fonémica es una rama de la Lingüística Descriptiva.

4.2.1 Fonemas

Podemos pensar en un fonema, como una unidad de sonido ideal con sus correspondientes gesticulaciones articulatorias. Los sonidos que realmente se emiten se les denomina fonaciones.

En la fonética, un sonido individual es una fonación; en la fonémica, la unidad menor es el fonema.

Un fonema es la unidad de sonido más corto en un lenguaje específico que es suficiente para diferenciar un sonido de otro. Si el cambio de una fonación dentro de una expresión altera el significado de ésta, entonces la fonación es también un fonema.

De forma similar, si el cambio de una variable fonética altera el significado de la palabra, entonces esa variable marca una distinción entre dos fonemas; si no logra cambiar el significado de la palabra, entonces se dice que esa característica no es fonémica en ese lenguaje.

El número de fonemas es vasto, solamente se encuentra limitado por nuestra habilidad de distinguirlos. El número de fonemas de cualquier lenguaje es pequeño. Las Figuras 4.2 y 4.3 nos muestran los fonemas vocálicos en inglés y en español respectivamente.

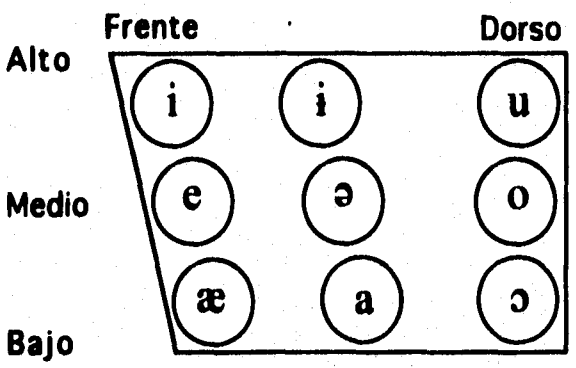


Figura 4.2 Diagrama esquemático de los fonemas vocálicos en inglés (Fig. 4.3 [PARS87]).

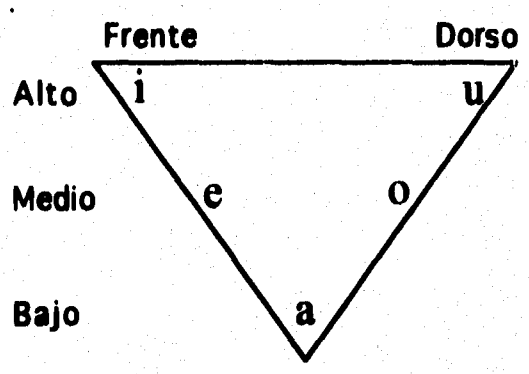


Figura 4.3 Diagrama esquemático de los fonemas vocálicos en español (Fig. 22 [QUIL79]).

4.2.2 Alófonos

Se puede llegar a establecer una relación entre los fonemas y la fonética, mediante la asociación de un conjunto de sonidos semejantes a un fonema; esto debe de ser aceptado por los locutores del lenguaje como el mismo sonido. A los miembros de este conjunto se les denomina alófonos.

Una de las razones por la cual hablamos en otro idioma con acento, es debido a que inconscientemente imponemos la organización fonémica de nuestro lenguaje, junto con su alófonos en el otro lenguaje.

4.3 Características Distintivas

La teoría de características distintivas tiene su origen en la acústica, la psicología de la percepción de la voz y el álgebra booleana.

Las características distintivas son un conjunto de 12 atributos que puede tener un fonema. Cada atributo es una característica acústica y se considera que solamente puede tomar dos valores posibles. Idealmente, cada característica es independiente de las otras. Entonces cada fonema puede considerarse como un paquete de características distintivas y en un principio puede ser definido completamente por un vector booleano o equivalentemente por un número binario de 12 bits.

Las características son las siguientes:

1. Vocálico/No-Vocálico. Se refiere a la presencia o ausencia de una estructura de formantes bien definida.
4. Consonántica/No-Consonántica. La primera se refiere a que contiene una pequeña cantidad de energía total.
3. Compacta/Difusa. Se refiere a la distribución de la energía espectral.
4. Tensa/Relajada. La primera implica que la energía total es grande con un ancho de banda amplio y duración prolongada.
5. Sonora/Sin-Sonoridad. La primera indica la presencia de componentes de baja frecuencia debido a la vibración de las cuerdas vocales.
6. Nasal/Oral. La primera muestra una distribución del energía espectral más amplia como resultado de la resonancia nasal.
7. Discontinua/Continua. Los fonemas discontinuos muestran cambios abruptos en la dispersión de energía espectral.
8. Estridente/Suave. Los fonemas estridentes tienen muchos componentes de ruido aleatorio muy intensos.
9. Restringido/No-Restringido. En fonemas restringidos,la energía aparece como una ráfaga, así como en las explosivas.
10. Graves/Agudas. Los sonidos graves son dominados por resonancias de baja frecuencia, los agudos por resonancias de alta frecuencia.
11. Plano/Liso. La diferencia estriba en la energía relativa en las resonancias de alta frecuencia: Plano es más débil, liso es más intenso.
14. Elevado/Liso. La primera indica una elevación en la frecuencia relativa de las resonancias de alta frecuencia.

Se debe de notar que estas características pueden definir 4096 fonemas; de aquí que no se espera que se representen todas las combinaciones en algún lenguaje.

Las características distintivas han sido propuestas como un modelo para la percepción humana y también como una herramienta en el reconocimiento de voz. La teoría de las características distintivas representa un primer intento para relacionar las observaciones acústicas a la percepción de los sonidos de la voz en una forma sistemática. Esta teoría ha sido utilizada como herramienta de diagnóstico.

4.4 Sílabas, Uniones y Prosódicos

Los fonemas se unen para formar sílabas. En inglés, cada sílaba contiene una vocal, conocida como núcleo o pico. Ya que las vocales están generalmente caracterizadas por una mayor energía que las consonantes, el pico silábico es usualmente también un pico de amplitud. De hecho, a nivel acústico, la definición más simple y segura de una sílaba es: un máximo de potencia en el flujo de voz. Existen varias excepciones a esto, pero es importante ya que los algoritmos de segmentación para reconocimiento de voz, normalmente toman la amplitud como un punto de partida.

Se debe de notar de una comparación con el texto que virtualmente cada sílaba está asociada con un máximo de amplitud. El problema es que cada máximo de amplitud, no esta asociado a una sílaba. Las sílabas tienden a estar separadas por sonidos no vocálicos; se puede decir que las sílabas son una alternación entre vocales y consonantes.

En inglés, las palabras normalmente están separadas por una unión. Acústicamente la unión puede ser una pausa, una extensión de la sílaba previa, un cierre momentáneo de la glotis, o una pequeña depresión en la amplitud.

Los prosódicos son un término general para aquellos aspectos de la voz que separan grupos de sílabas o palabras. Las principales variables de interés de los prosódicos son el tono y el acento de intensidad.

La variación del tono sobre una frase se le denomina entonación, es utilizada en la mayoría de los lenguajes para dar forma a la oración e indicar su estructura, a pesar de que la forma en que esto se hace varía ampliamente en los diferentes lenguajes. En algunos lenguajes, el tono se utiliza para ayudar a entender el significado de las palabras. El tono varía desde 80 y 160 Hz para locutores masculinos y entre 160 y 440 Hz para femeninos. Así como la fonación es cuantizada en cualquier lenguaje en fonemas, el tono también es cuantizado; El inglés cuenta con cuatro niveles de tonos y tres contornos terminales. Los niveles están marcados por números y algunas veces se les denomina bajo, medio, alto y extra alto; los terminales son desvanecimiento (desvanecimiento en tono y amplitud), incremento (incremento en el tono y la amplitud se mantiene constante), y sostenido (el tono y la amplitud se mantienen constantes).

El acento de intensidad, refleja el grado de énfasis con la que se dice una palabra o sílaba; las palabras enfatizadas son normalmente más fuertes, pero también son más prolongadas e intensas. Esto es el acento tiende a mover las vocales hacia los extremos del diagrama vocálico, mientras que en el inglés y otros idiomas las vocales no enfatizadas tienden a ser pronunciadas más cercanas a la posición neutra. El acento también tiende a aumentar el tono. Las irregularidades provocadas en el tono por el acento, son superimpuestas en el contorno del tono. La acentuación así como el tono puede ser cuantizado; en el inglés existen cuatro niveles de énfasis.

El acento de intensidad se aplica a palabras y a unidades mayores de voz. Las palabras individuales, normalmente tienen una sílaba que esta enfatizada o acentuada. En Inglés este es primordialmente un acento enfático, a pesar que en otros lenguajes las sílabas acentuadas están marcadas por el tono.

Símbolo IPA	Arpabeto		Ejemplos	Símbolo IPA	Arpabeto		Ejemplos
i	i	IY	heed	v	v	V	verve
ɪ	I	IH	hid	θ	T	TH	thick
e	e	EY	hayed	ð	D	DH	those
ɛ	E	EH	head	s	s	S	cease
æ	@	AE	had	z	z	Z	pizzaz
ɑ	a	AA	hod	ʃ	S	SH	mesh
ɔ	c	AO	hawed	ʒ	Z	ZH	measure
o	o	OW	hoed	h	h	HH	heat
ʊ	U	UH	hood	m	m	M	mom
u	u	UW	who'd	n	n	N	noon
ɜ	R	ER	heard	ŋ	G	NX	ringing
ə	x	AX	ahead	l	l	L	lulu
ʌ	A	AH	bud	l	L	EL	battle†
ɑɪ	Y	AY	hide	m	M	EM	bottom†
ɑʊ	W	AW	how'd	n	N	EN	button†
ɔɪ	O	OY	boy	f	F	DX	batter‡
ɸ	X	IX	roses	ʔ	Q	Q	§
p	p	P	pop	w	w	W	wow
b	b	B	bob	j	y	Y	yoyo
t	t	T	tug	r	r	R	roar
d	d	D	dug	tʃ	C	CH	church
k	k	K	kick	dʒ	J	JH	judge
g	g	G	gig	ʍ	H	WH	where
f	f	F	fife				

† l,m,n vocálica ‡ t ondulada § pausa glotal

Tabla 4.3.1 Alfabeto Fonético (ejemplos inglés) (Tabla 4.1 [PARS87]).

Fonema	Letra	Ejemplos	Fonema	Letra	Ejemplos
a	A	asa	l	L	leña
b	B	barón	ll	LL	llama
	V	varón	m	M	malo
	W	watio	n	N	nieto
ch	CH	chino	ñ	N	moño
d	D	duda	o	O	oso
e	E	meter	p	P	papel
f	F	fino	r	R	entre vocales (ora)
g	G	ante a,o,u (gato,goma,gusto)			en fín de silaba (probar)
	GU	ante e,i (guerra, guiso)	rr	R	inicial (ramo)
j	G	ante e,i (genio,gitano)		RR	despues de n,l,s (alrededor)
	J	ojo,jarro,judío			entre vocales (carro)
i	I	mina	s	S	salsa
	Y	él y tú, rey	t	T	tomate
k	K	kilo	u	U	bueno
	Q	ante e,i (queso,quiere)	y	Y	yugo
	C	ante a,o,u (cama,cosa,cura)	Ø	H	humo, hueco
z	C	ante e,i (cena,cine)			
	Z	zapato,paz			

Tabla 4.3.2 Alfabeto Fonético (ejemplos español) (Norma Lengua Nacional 5/6 pp 205).

Signo	Español Representación ortográfica	Signo	Inglés Representación ortográfica
p	par	p	pay
b	bar	b	bar
t	té	t	tea
d	dar	d	day
k	cama	k	cold
g	gana	g	go
β	saber		
f	fin	f	foot
		v	vain
θ	zumo	θ	thin
ð	codo	ð	then
s	sol	s	see
ʂ	mismo		
		z	zeal
		ʃ	show
		ʒ	measure
ǰ	ayer		
x	jota		
		h	heap
ɣ	paga		
c	chico	c	cheap
ɟ	cónyuge	ɟ	jump
m	mamá	m	make
ɱ	confuso		
n	no	n	no
ɳ	once		
ɲ	donde		
ɳ̟	concha		
ŋ	tango	ŋ	long
ɲ̊	caña		
l	lado	l	leaf
ʎ	alzar		
		ɬ	full
ɭ	toldo		
ʎ̝	calle		
r	pero	r	red
r̃	perro		
ɹ	corto		

Tabla 4.3.3 Consonantes ([QUIL79] pp. XXVIII).

Signo	Español Representación ortográfica	Signo	Inglés Representación ortográfica
j	pie	j	yes
w	cuatro	w	wine

Tabla 4.3.4 Semiconsonantes ((QUIL79] pp. XXIX).

Signo	Español Representación ortográfica	Signo	Inglés Representación ortográfica
i̯	aire		
u̯	rauda		

Tabla 4.3.5 Semivocales ((QUIL79] pp. XXIX).

Signo	Español Representación ortográfica	Signo	Inglés Representación ortográfica
i	par	i	pay
		ɪ	bar
e	té	e	tea
		ei	day
		ɛ	fell
		æ	hat
		a	father
a		a	aisle
		ʌ	cup
		ə	above
		ɔ	hot
o		o	November
		ou	go
		u	food
u		u	good
		ɚ	bird

Tabla 4.3.6 Vocales ((QUIL79] pp. XXIX).

5. Fonética Acústica[1]

El estudio de la fonética acústica data del siglo dieciocho. Sin embargo, el rápido crecimiento de la fonética acústica moderna, se desarrolló desde el invento del espectrógrafo en los laboratorios Bell en 1941. El resultado del análisis del espectrograma es una representación gráfica de la señal de voz del contenido frecuencial en el eje vertical y el tiempo en el eje horizontal. La intensidad de cualquier componente de frecuencia esta representado por el grado de obscuridad del punto en el papel.

Existen dos anchos de banda en el espectrógrafo: de banda angosta (45 Hz) y de banda ancha (300 Hz). Ya que esencialmente todas las voces tienen frecuencias fundamentales mayores a 45Hz, los espectrogramas de banda angosta muestran el tono y sus armónicas como líneas horizontales. Ya que en la mayoría de los locutores el tono es inferior a 300Hz, en el modo de banda ancha las armónicas del tono no se pueden resolver; por otro lado, los pulsos glotales individuales son visibles y los formantes (resonancias del tracto vocal) se muestran como barras obscuras. La Figura 5.1 es un espectrograma de banda ancha.

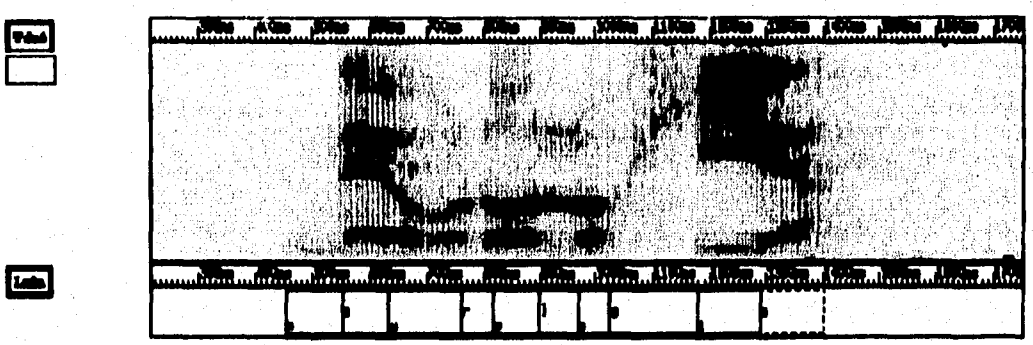


Figura 5.1 Espectrograma de banda ancha ("neurología")

5.1 Acústica del tracto vocal

Acústicamente, el tracto vocal es un tubo de sección transversal no uniforme, de aproximadamente 17 cms de largo en los hombres adultos, usualmente abierto en un extremo y casi cerrado en el otro. Existe una bifurcación casi a la mitad que es la cavidad nasal de aproximadamente 13 cms de largo, con una válvula en la bifurcación denominada velo como se muestra en la Figura 5.2. Para un análisis primario, asumiremos que el velo se encuentra cerrado; esto excluye las cavidades nasales de las consideraciones y simplifica bastante el análisis.

El tubo anterior, es una estructura de parámetros distribuidos y por lo tanto tiene varias frecuencias naturales. Si el tracto vocal tuviera una sección transversal uniforme, esta frecuencias estarían localizadas en:

$$f_n = \frac{(2n-1)c}{4l}, n = 1,2,3,... \qquad (5.1)$$

[1] En su mayor parte tomado del capítulo 5 de [PARS87].

En el aire, c=350m/s; para un tubo de longitud l =17 cms, las frecuencias naturales ocurren en múltiplos impares de ~500Hz. En realidad, el área es no uniforme; esto resulta en resonancias que no se encuentran uniformemente espaciadas, pero la densidad promedio de las resonancias del tracto vocal es aproximadamente una por kilohertz de ancho de banda, como lo indica la relación (5.1).

A estas relaciones se les conoce como formantes; estas son las bandas obscuras observadas en el espectrograma de la Figura 5.1 y son la característica más importante del tracto vocal. El tren de pulsos glotales es rico en armónicas y éstas interactúan fuertemente con las resonancias del tracto vocal para afectar la calidad del tono de la voz. Entonces, los formantes proporcionan al que escucha, la fuente principal de información acerca de la posición de los órganos del habla.

Se debe de notar que estas frecuencias resonantes corresponden a los polos de la función de transferencia. Mientras el tracto nasal se encuentre bloqueado y la glotis sea la única fuente de excitación, la función de transferencia del tracto vocal no tiene ceros de frecuencia finita. Esto es una simplificación importante.

Los formantes se identifican por número en orden creciente: F_1, F_2, etc. (Por uniformidad, al tono se le denomina F_0). En la fonética acústica clásica, solamente se consideran F1 y F2, para reconocimiento, al menos tres tienen importancia y para la síntesis se recomiendan cinco para lograr que se oiga natural.

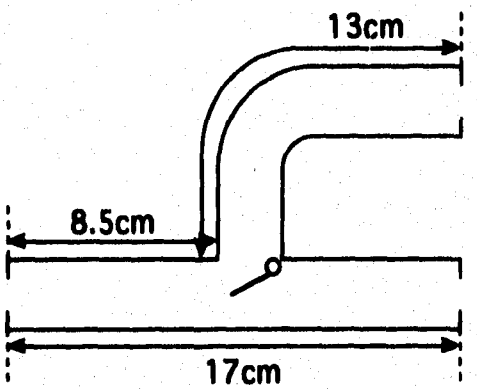


Figura 5.2 Diagrama simplificado del tracto vocal. La sección recta corre de la glotis a los labios; la sección curva corre del velo a las fosas nasales (Fig. 5.3 [PARS87]).

5.1.1 Relación a los Sonidos de Voz

Los sonidos vocálicos se producen al cambiar la forma del tracto vocal; por lo tanto esperaríamos una correspondencia entre los sonidos vocálicos y las frecuencias de los formantes. Esto es soportado por la siguiente evidencia: (1) se pueden recobrar las frecuencias de los formantes en forma consistente y mapearlas a sonidos vocálicos; (2) la voz artificial con las frecuencias de formantes adecuadas, es percibida con la calidad de vocal deseada. Si construimos un sistema coordenado usando F_1 y F_2 como base, las vocales caen en regiones específicas. La Figura 5.3 muestra la afirmación anterior y la Tabla 5.1 muestra las frecuencias típicas para algunas vocales.

	Hombres Adultos			Mujeres Adultas		
Vocal	F1	F2	F3	F1	F2	F3
[i]	255	2330	3000	340	2610	3210
[ɪ]	350	1975	2560	425	2170	2900
[ɛ]	560	1875	2550	690	2015	2815
[æ]	735	1625	2465	950	1955	2900
[ɑ]	760	1065	2550		1085	
[ʌ]	640	1250	2610	750	1300	2610
[ɔ]	610	865	2540		785	2565
[ʊ]	475	1070	2410	515	1070	2280
[u]	290	940	2180	390	995	2585

Tabla 5.1 Frecuencias de formantes típicas para alguna vocales (Tabla 5.2 [PARS87]).

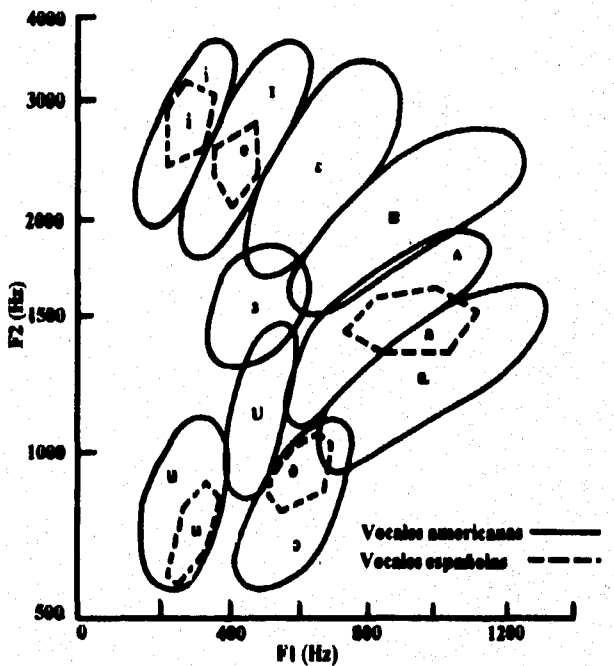


Figura 5.3 Ubicación de F1 y F2 para algunas vocales del inglés y del español (Fig 14 (BORZ80]).

Existe una pieza adicional de información, si se invierte la dirección de los ejes F1 y F2 como en la Figura 5.4, entonces se puede observar que la ubicación de las vocales corresponde a las posiciones asignadas a esta vocales dentro del diagrama básico de las vocales.

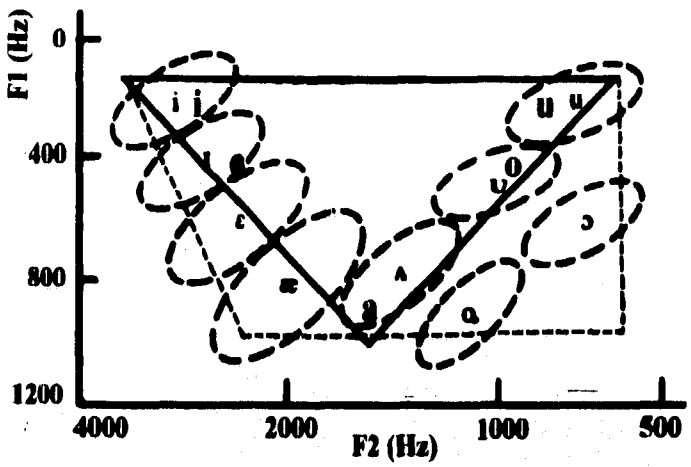


Figura 5.4 Diagrama de las vocales con la ubicación de F1 y F2 superimpuesta. Nótese la inversión de la escala de frecuencias.

5.2 Análisis de la Acústica de las Vocoides

Debido a la importancia de los formantes, se ha realizado un gran esfuerzo en analizar la acústica del tracto vocal, particularmente con una visión de relacionar las formas del tracto vocal a sus frecuencias de formantes correspondientes. El problema se complica por el hecho de que el tracto vocal varia irregularmente a lo largo de su trayecto, por lo tanto es necesario buscar una simplificación. La aproximación más utilizada es la del modelo cilíndrico a tramos que se muestra en la Figura 5.5

La ventaja es que las secciones cilíndricas son perfectamente comprendidas y son de fácil modelado y análisis; mas aún, los resultados concuerdan con las mediciones realizadas actuales con técnicas como las fotografías de rayos-x.

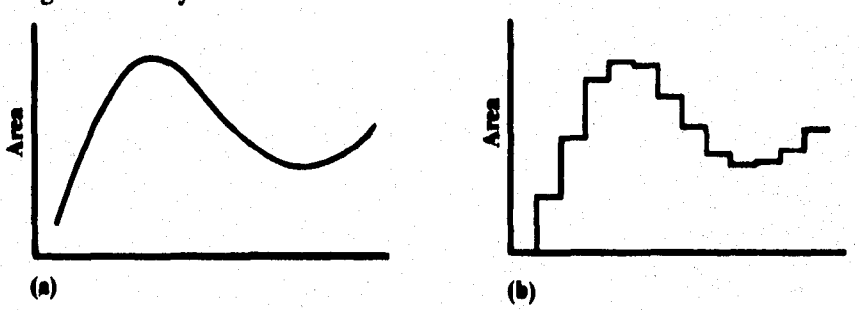


Figura 5.5 Aproximación a)suavizada, b)cilíndrica a tramos de la función área de tracto vocal; (Fig. 5.6 [PARS86]).

5.2.1 Acústica de un Tubo Cilíndrico

Un tubo de sección transversal uniforme es el análogo de una línea de transmisión, y se toma ventaja de esta analogía en el análisis del tracto vocal. Al igual que en una línea de transmisión, las ondas pueden viajar en cualquier dirección del tubo y estas ondas están gobernadas por la ecuación de onda. Asumiremos planos de onda y transmisión sin pérdidas y sin dispersión.

Primero de la ecuación de Newton, se tiene una relación entre el gradiente de presión en la dirección de la onda y la aceleración del gas:

$$\frac{\partial p}{\partial x} = -\rho_0 \frac{\partial u}{\partial t} \tag{5.2}$$

donde p = fluctuación de la presión
u = velocidad de la partícula
ρ_0 = densidad del aire

Segundo, de la consideración de la compresión y rarefacción del medio, se puede relacionar el gradiente de velocidad a la razón de cambio en presión:

$$\frac{\partial u}{\partial x} = -\frac{1}{\eta P_0} \frac{\partial p}{\partial t} \tag{5.3}$$

donde η = razón de los calores específicos de volumen constante a presión constante del aire
P_0 = presión normal atmosférica

Combinando las ecuaciones (5.2) y (5.3) resulta la ecuación de onda.

$$\frac{\partial^2 p}{\partial x^2} = \frac{\rho_0}{\eta P_0} \frac{\partial^2 p}{\partial t^2} \tag{5.4}$$

La solución de (5.4) es la suma de cualquier combinación arbitraria de ondas de presión que viajan hacia la derecha y hacia la izquierda con velocidad de movimientos $\pm c$:

$$p = f_1\left(t - \frac{x}{c}\right) + f2\left(t - \frac{x}{c}\right) \tag{5.5a}$$

La única restricción en f_1 y f_2 es que sean al menos dos veces diferenciables. Se puede verificar la solución substituyendo (5.5a) en (5.4); al realizar esto se demuestra que la velocidad de propagación debe de ser

$$c = \sqrt{\frac{\eta P_0}{\rho_0}}$$

Se puede resolver (5.4) para la velocidad de la partícula $u(x,t)$. Cuando se trata con ondas en un tubo, sin embargo, estamos más interesados en la velocidad del volumen, la razón de flujo a través del área en corte seccional S en unidades cúbicas por segundo. Si la velocidad de la partícula es u y la velocidad volumétrica es U, entonces $U = Su$. Se puede encontrar una solución para U de (5.4) y (5.2), ésta es:

$$U = \frac{S}{\rho_0 c}\left[f_1\left(t - \frac{x}{c}\right) - f2\left(t - \frac{x}{c}\right)\right] \tag{5.5b}$$

Se debe de notar que $\rho_0 c/S$ es una razón de presión a velocidad volumétrica. Por lo tanto es análogo a la impedancia eléctrica y se denomina impedancia característica del tubo Z_c.

La acústica ha tomado los siguientes términos y símbolos de la teoría de circuitos eléctricos: *Impedancia* la razón de presión a la velocidad volumétrica denotada por *Z*. Admitancia, la razón de velocidad volumétrica a la presión denotada por *Y*. Estas razones se refieren estrictamente a las transformaciones de presión y velocidad, pero también se utilizan para razones en el dominio del tiempo cuando no resulta ninguna ambigüedad.

Algunos valores típicos para c y Z_c ya se encuentran previamente calculados. Para el aire, η=1.4, ρ_0=1.14x10^{-3} gm/cm^3 y P_0=1 bar = 10^6 dyn/cm. De aquí c =5.5x10^4 cm/s y Z_c =1.11S Ω acústicos. Para S = 6cm^2 entonces Z_c=6.66 Ω acústicos.

5.2.2 Solución a la Onda Senoidal; Impedancia de Entrada

También es importante conocer la solución a la ecuación de onda para ondas senoidales, ya que esto permite análisis en el dominio de la frecuencia y la consideración de la impedancia de entrada de un tubo. Para ondas senoidales, las soluciones son

$$p(x,t) = (P_1 e^{-j\beta x} + P_2 e^{j\beta x})e^{j\omega t} \tag{5.6}$$

$$U(x,t) = \frac{1}{Z_c}(P_1 e^{-j\beta x} + P_2 e^{j\beta x})e^{j\omega t} \tag{5.7}$$

donde ω =frecuencia de la onda senoidal (en radianes por segundo)
$\beta = \omega/c$, la constante de propagación

Si el tubo tiene una longitud l y está terminado en el extremo con una impedancia Z_L (Figura 5.6), entonces $p(l,t) = Z_L U(l,t)$. Combinando esto con (5.6) y (5.7) lleva a una expresión para la impedancia acústica de entrada Z_{in}.

$$Z_{in} = Z_c \frac{Z_L \cos\omega T + jZ_c \sin\omega T}{Z_c \cos\omega T + jZ_L \sin\omega T} \tag{5.8}$$

Donde T es el tiempo requerido para que la onda se propague sobre la longitud l: $T=l/c$. Existen dos casos especiales. El primero es si el extremo del tubo se encuentra abierto $Z_L = 0$ y

$$Z_{in} = jZ_c \tan\omega T \tag{5.9a}$$

el segundo, si el tubo esta cerrado en el extremo, $Z_L = \infty$ y

$$Z_{in} = -jZ_c \cot\omega T \tag{5.9b}$$

Las admitancias correspondientes son los recíprocos de estos. Hagamos $Yc = 1/Zc$; entonces para $Z_L = 0$

$$Y_{in} = -jY_c \cot\omega T \tag{5.9c}$$

y para $Z_L = \infty$

$$Y_{in} = jY_c \tan\omega T \tag{5.9d}$$

Se debe de notar que la impedancia de entrada siempre tiene polos de frecuencia finita y ceros; en estos casos especiales, los polos y ceros caen en el eje de la frecuencia.

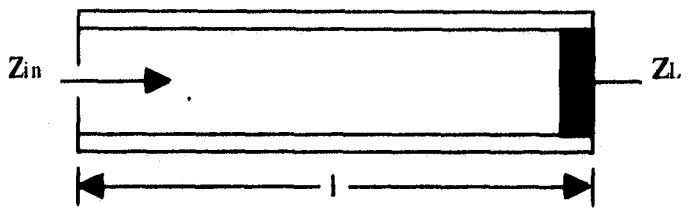


Figura 5.6 Tubo acústico terminado en una impedancia de carga (Fig. 5.7 [PARS87]).

5.2.3 Análisis del Modelo Cilíndrico del Tracto Vocal

Si el tracto vocal es dividido en un número grande de segmentos cilíndricos, se obtiene una aproximación semejante a su forma actual y por lo tanto a su función de transferencia. Si los segmentos son lo suficientemente pequeños, cada uno se puede aproximar por un sistema de parámetros concentrados; de aquí que este modelo también permite comenzar con un sistema de parámetros distribuidos y terminar con una aproximación de parámetros concentrados. Esta aproximación es mas fácilmente manejable en términos de cálculo, ademas que tiene una forma que se utilizará en el caso de predicción lineal. Como una simplificación adicional, se considerará que la transmisión a través del tracto vocal no tiene pérdidas.

La Figura 5.7 muestra la apariencia general del modelo y las convenciones de numeración que se utilizarán. Todas las secciones son de longitud Δ.

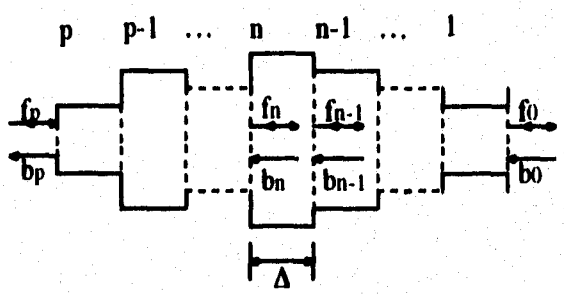


Figura 5.7 Notación y numeración de las secciones utilizada para el análisis de modelo cilíndrico a tramos del tracto vocal (Fig. 5.8 [PARS87]).

Debemos de comenzar por encontrar la relación entre las ondas que viajan a través de secciones cilíndricas adyacentes. Para cualquier sección n, f_n es la onda de velocidad volumétrica hacia adelante que parte del extremo izquierdo de la sección y b_n es la onda hacia atrás que llega del mismo punto. Ya que asumimos que no existen pérdidas y que la transmisión es no dispersiva, las ondas de presión correspondientes son $R_n f_n$ y $R_n b_n$, donde R_n es la impedancia acústica de la sección. Si la sección tiene un área lateral S_n, entonces

$$R_n = \rho_0 c / S_n .$$

Inicialmente se analizará cada sección cilíndrica como un sistema de parámetros distribuidos. Es correcto lleva a cabo este análisis en términos de las ondas hacia adelante y hacia atrás de (5.5); Esto permite que se tomen conceptos de la teoría desarrollada para las lineas de transmisión y las guías de onda. En particular se analizará la sección en términos de la matriz de transformación de la

onda. En cualquier sistema pasivo lineal de dos puertos (Figura 5.8), sean f_1 y b_1 las ondas hacia adelante y hacia atrás en el puerto 1; sean f_2 y b_2 las ondas hacia adelante y hacia atrás en el puerto 2 (La numeración está al revés con respecto a la notación mencionada). Entonces, la matriz de transformación T esta definida como sigue:

$$\begin{bmatrix} f_1 \\ b_1 \end{bmatrix} = \mathbf{T} \begin{bmatrix} f_2 \\ b_2 \end{bmatrix}$$

Esta matriz tiene la propiedad importante que la matriz T de dos o más secciones en cascada, es igual al producto de las matrices T de las secciones correspondientes. Esto permite ir de una sección sencilla a todo el modelo cilíndrico multiplicando las matrices.

Una matriz relacionada, que usualmente no es vista, es la matriz de dispersión S. Esta matriz relaciona las ondas que abandonan (o se dispersan) el sistema de dos puertos a las que entran. En nuestro caso, las ondas incidentes son f_2 y b_1 y las ondas dispersas son b_2 y f_1, entonces podemos escribir

$$\begin{bmatrix} b_2 \\ f_1 \end{bmatrix} = S \begin{bmatrix} f_2 \\ b_1 \end{bmatrix}$$

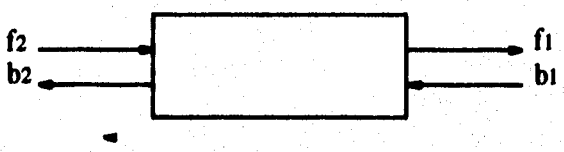


Figura 5.8 Sistema de dos puertos mostrando las ondas hacia adelante y hacia atrás (Fig. 5.9 [PARS87]).

5.2.4 Análisis de una sección simple

Dividiremos la sección como un tubo uniforme de longitud Δ, seguido por una discontinuidad donde esta unida con la siguiente sección. Para encontrar la matriz T para tal sección, tomaremos ventaja de la multiplicación de estas matrices; encontraremos la T matrices de las discontinuidades y el tubo por separado y luego las multiplicaremos.

1. En la discontinuidad, se tienen las siguientes relaciones (Figura 5.9): a la izquierda se tiene

$$\begin{aligned} p_2 &= R_2(f_2 + b_2) \quad \text{(presión)} \\ U_2 &= f_2 - b_2 \quad \text{(velocidad volumétrica)} \end{aligned} \tag{5.10}$$

y en la derecha se tiene

$$\begin{aligned} p_1 &= R_1(f_1 + b_1) \quad \text{(presión)} \\ U_1 &= f_1 - b_1 \quad \text{(velocidad volumétrica)} \end{aligned} \tag{5.11}$$

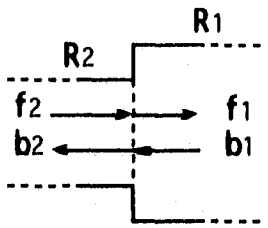


Figura 5.9 Ondas de velocidad volumétrica hacia adelante y hacia atrás en la discontinuidad entre secciones cilíndricas (Fig. 5.10 [PARS87]).

Ya que la presión y la velocidad volumétrica deben de ser continuas a través de la unión, entonces se tiene

$$f_2 - b_2 = f_1 - b_1$$
$$R_2(f_2 + b_2) = R_1(f_1 + b_1)$$

Resolviendo para f_1 y b_1 se obtiene

$$\begin{bmatrix} f_1 \\ b_1 \end{bmatrix} = \frac{1}{2R_1}\begin{bmatrix} R_2 + R_1 & R_2 - R_1 \\ R_2 - R_1 & R_2 + R_1 \end{bmatrix}\begin{bmatrix} f_2 \\ b_2 \end{bmatrix} \tag{5.12}$$

Se define el coeficiente de reflexión k como

$$k = \frac{R_1 - R_2}{R_1 + R_2}$$
$$k = \frac{S_2 - S_1}{S_2 + S_1} \tag{5.13}$$

Entonces la matriz **T** para la unión

$$T_j = \frac{1}{1+k}\begin{bmatrix} 1 & -k \\ -k & 1 \end{bmatrix} \tag{5.14}$$

2. El resto de la sección cilíndrica que precede a la unión se muestra en la Figura 5.10. Aquí existe un retardo debido a que las ondas viajan de un extremo a otro. Entonces por inspección,

$$f_1(t) = f_2\left(t - \frac{\Delta}{c}\right)$$
$$b_1(t) = b_2\left(t + \frac{\Delta}{c}\right) \tag{5.15}$$

ya que f_1 se encuentra atrasado con respecto a f_2 por Δ/c segundos y b_1 se encuentra adelantado con respecto a b_2 por la misma cantidad.

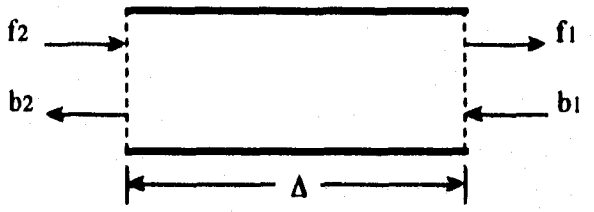


Figura 5.10 Ondas hacia adelante y hacia atrás en los extremos de la sección cilíndrica (Fig. 5.11 [PARS87])

5.2.5 Transformadas Z

En función de poder combinar las ecuaciones (5.14) y (5.15), se llevarán los resultados al dominio de la transformada Z. Si muestreamos este sistema a una razón de $F_s = c/2\Delta$, entonces un retardo de Δ/c corresponde a un factor de $z^{-1/2}$ en el dominio de la transformada Z. Entonces se puede escribir la matriz T para la sección cilíndrica como

$$T_c = \begin{bmatrix} z^{-1/2} & 0 \\ 0 & z^{1/2} \end{bmatrix} \tag{5.16}$$

Ahora se pueden multiplicar las matrices de (5.14) y (5.16) para obtener la matriz T para la n-ésima sección:

$$\begin{aligned} T_n &= \frac{1}{1+k_n}\begin{bmatrix} 1 & -k_n \\ -k_n & 1 \end{bmatrix}\begin{bmatrix} z^{-1/2} & 0 \\ 0 & z^{1/2} \end{bmatrix} \\ T_n &= \frac{1}{1+k_n}\begin{bmatrix} z^{-1/2} & -k_n z^{1/2} \\ -k_n z^{-1/2} & z^{1/2} \end{bmatrix} \\ T_n &= \frac{z^{-1/2}}{1+k_n}\begin{bmatrix} 1 & -k_n z \\ -k_n & z \end{bmatrix} \end{aligned} \tag{5.17}$$

Donde el índice se refiere al comienzo y final de la sección como se muestra en la Figura 5.11. La inversa de la matriz será utilizada posteriormente, esta es:

$$T_n^{-1} = \frac{z^{1/2}}{1-k_n}\begin{bmatrix} 1 & k_n \\ k_n z^{-1} & z^{-1} \end{bmatrix} \tag{5.18}$$

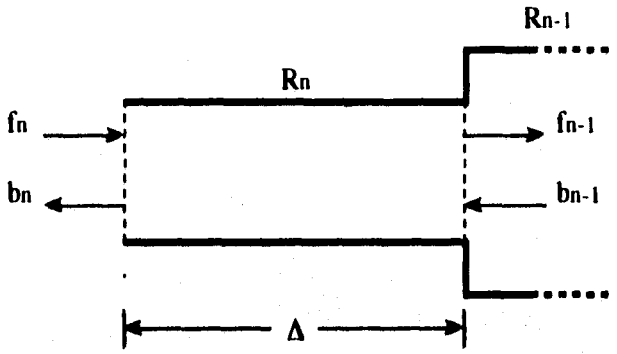


Figura 5.11 Ondas hacia adelante y hacia atrás para la sección cilíndrica completa (Fig. 5.12 [PARS87])

5.2.6 La Matriz de Dispersión

La ecuación (5.17) puede ser replanteada de forma que usa la matriz de dispersión **S**. Se debe de notar que las ondas incidentes son f_n y b_{n-1}, ya que están entrando a la sección y las ondas dispersadas son b_n y f_{n-1}. Después de alguna manipulación se tiene

$$\begin{aligned} b_n(t) &= (1+k_n)b_{n-1}\left(t-\frac{\Delta}{c}\right) + k_n f_n\left(t-\frac{2\Delta}{c}\right) \\ f_{n-1}(t) &= -k_n b_{n-1}(t) + (1-k_n) f_n\left(t-\frac{\Delta}{c}\right) \end{aligned} \tag{5.19}$$

o utilizando transformadas Z

$$\begin{bmatrix} B_n(z) \\ F_{n-1}(z) \end{bmatrix} = \begin{bmatrix} (1-k_n)z^{-1/2} & k_n z^{-1} \\ -k_n & (1+k_n)z^{-1/2} \end{bmatrix} \begin{bmatrix} B_{n-1}(z) \\ F_n(z) \end{bmatrix} \tag{5.20}$$

Estas son las ecuaciones Kelly-Lochbaum; se pueden representar por la estructura de la Figura 5.12. Kelly y LochBaum (1962) utilizaron esta estructura en un modelo del tracto vocal para generación por computadora de voz sintética

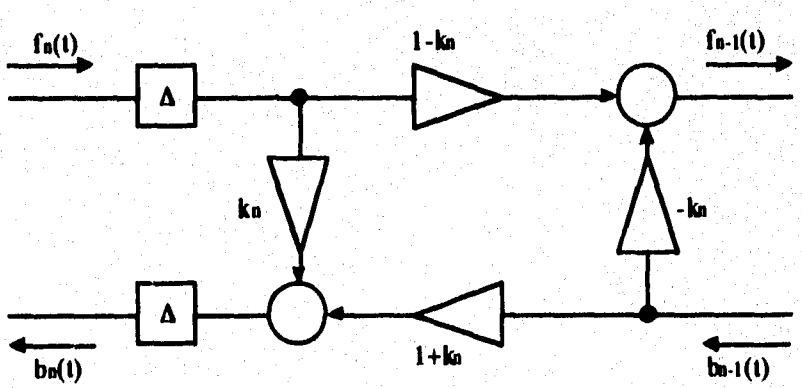


Figura 5.12 Diagrama a bloques del sistema modelado por las ecuaciones de Kelly-Lochbaum (Fig. 5.13 [PARS87]).

5.2.7 La Matriz T Completa

En este punto se puede escribir la matriz de transformación de la onda para el modelo completo; ya que las matrices **T** para las secciones en cascada se multiplican, es simplemente

$$\begin{bmatrix} F_0(z) \\ B_0(z) \end{bmatrix} = T \begin{bmatrix} F_P(z) \\ B_P(z) \end{bmatrix} \tag{5.21a}$$

donde

$$T = \prod_{i=1}^{p} T_i \tag{5.21b}$$

Después se evaluaran formas eficaces para la evaluación de esta función de transferencia. En este punto se puede utilizar (5.21) para establecer una propiedad básica del tracto vocal.

La función de transferencia de la velocidad de cualquier sistema puede ser calculad a partir de su matriz **T** como sigue:

$$A_U = \frac{F_0(z)/B_0(z)}{F_P(z)/B_P(z)} = \frac{|T|(1-k_L)}{t_{21}+t_{22}-k_L(t_{11}+t_{12})} \tag{5.22}$$

Donde k_L es el coeficiente de reflexión de la carga: $k_L=b_0/f_0$. Los elementos t_{ij} son elementos de la matriz **T**, que para nuestro caso es la matriz **T** total que acabamos de demostrar. Ahora si tratamos de evaluar **T** mediante la multiplicación de los componentes de la matriz de la forma (5.17), encontramos que los coeficientes t_{ij} son todos polinomios en z. El numerador se puede demostrar que es una constante. Esto nos lleva a la conclusión que la función de transferencia del tracto vocal es una función solo polos. Esta condición se mantiene mientras el tracto vocal se represente como un tubo simple sin bifurcaciones. En particular, estima que el velo esta cerrado y por lo tanto los conductos nasales no están acoplados al resto del tracto vocal.

5.2.8 Pérdidas

Las pérdidas que ocurren en el tracto vocal son pequeñas y sus efectos sobre la función de transferencia son mínimas, de otra forma el modelo derivado no sería útil. Existen tres fuentes principales de pérdidas de energía dentro del tracto vocal:

1. En el modelo esta implícito que las paredes del tracto vocal son rígidas. De hecho, vibran perceptiblemente en respuesta a la voz; de aquí se pierde energía.

2. La compresión y rarefacción, resultan en movimiento del aire relativo a las paredes del tracto vocal; se pierde energía por fricción asociada a este movimiento.

5. la ecuación (5.2) asume una expresión adiabática del aire; por ejemplo la transferencia de calor del gas que entra y sale durante la compresión y rarefacción es cero. De hecho existe un pequeño intercambio de calor entre el aire y las paredes del tracto vocal.

Ademas de estos efectos en el tracto, existen mas pérdidas debidas a la impedancia de radiación en los labios y la impedancia efectiva en la glotis.

El efecto de estas pérdidas en nuestro modelo es principalmente mover los polos de transmisión hacia adentro del círculo unitario. Esto significa que cada formante tiene un ancho de banda diferente de cero. En realidad, estos anchos de banda varían de formante a formante y de su frecuencia. En la practica un ancho de banda de entre 60 y 100Hz son una buena aproximación.

5.3 Propiedades de la Forma de Onda de las Vocales

5.3.1 Propiedades en el Dominio del Tiempo.

La forma de onda del pulso glotal es similar a un tren de pulsos. A esta función le denominaremos $g(t)$. Es aplicado al tracto vocal y la señal resultante de voz es la convolución de $g(t)$ con la respuesta al impulso del tracto vocal $h(t)$. Si la función de transferencia consiste únicamente de polos, entonces su respuesta al impulso es la suma de senoidales decrecientes, una por cada par de polos en $H(z)$. La función de tiempo resultante normalmente tiene la forma de la Figura 5.13. Los pulsos de mayor amplitud, representan el inicio de un nuevo pulso glotal y por lo tanto están espaciados en un intervalo regular igual al período del pulso glotal.

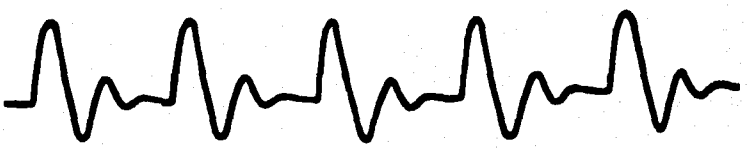


Figura 5.13 Salida del tracto vocal en respuesta al tren de pulsos glotales (Fig. 5.15 [PARS87]).

5.3.2 Características en el Dominio de la Frecuencia

El tren de pulsos $g(t)$ es por naturaleza rico en armónicas. Si consideramos $g(t)$ como un tren de impulsos convolucionado con la forma del pulso glotal, entonces el espectro será un tren de impulsos espaciado a una frecuencia igual a la frecuencia del tono y multiplicada por la transformada de la forma del pulso glotal.

La radiación de la voz desde los labios tiene la propiedad que la presión del sonido es proporcional a la derivada de la velocidad volumétrica en la boca. Este hecho introduce una amplificación de 6dB/octava en el espectro del sonido. Es conveniente reflejar esta amplificación hacia la glotis para mantener el filtro del tracto vocal libre de los efectos por radiación.

Esto nos lleva a tener un espectro "virtual" de excitación que consiste de un tren de impulsos con una caída efectiva de 6dB/octava. Después de aplicar una ventana, que es parte de cualquier análisis espectral práctico, el espectro de la excitación se ve como en la Figura 5.14.

Esta excitación es aplicada al tracto vocal. Entonces si a este espectro lo denominamos $G(f)$ y la función de transferencia del tracto vocal es $H(f)$, el espectro de la salida será $G(f)H(f)$. $H(f)$ se caracteriza por máximos que corresponden a los polos de los formantes como se muestra en la Figura 5.15.

El espectro de salida es el producto de estos dos espectros como se muestra en la Figura 5.16. La curva punteada es llamada el envolvente espectral; su forma es el producto de $H(f)$ y la envolvente

G(f). Esta envolvente espectral es una parte fundamental de muchas aplicaciones de procesamiento de voz, ya que es la fuente principal de información articulatoria. La importancia de la predicción lineal, se basa en gran parte en la capacidad de proporcionar formas rápidas, precisas y teóricamente justificables de recobrar esta envolvente.

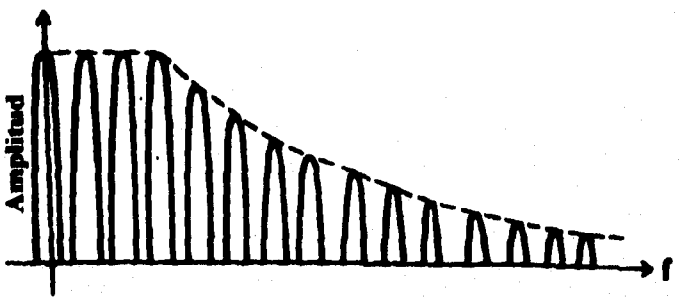


Figura 5.14 Espectro idealizado del tren pulsos glotales (Fig. 5.16 [PARS87]).

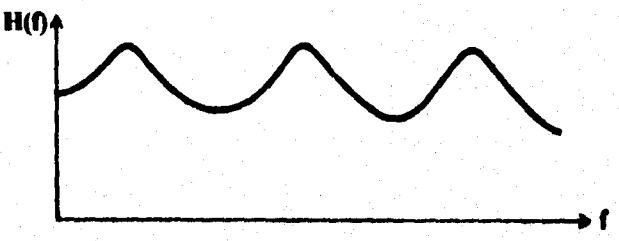


Figura 5.15 Respuesta frecuencial del tracto vocal. Los picos corresponden a los formantes (Fig. 5.17 [PARS87]).

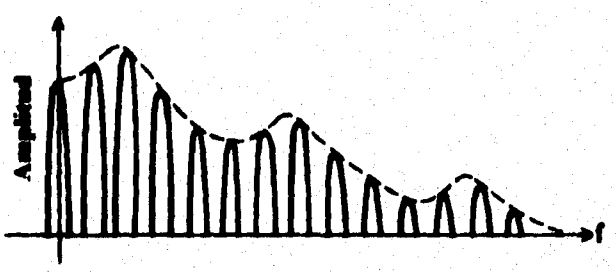


Figura 5.16 Espectro de la voz generada (Fig. 5.18 [PARS87]).

5.4 Características Acústicas de las Nasales

Entre las nasales no solamente incluimos consonantes nasales [m n] sino también vocales nasalizadas. Por ejemplo las vocales en las que el velo esta abierto y los conductos nasales están acoplados al tracto vocal. El elemento común es que existe la presencia de ún elemento acústico adicional. Utilizando la analogía con una línea de transmisión, tenemos el diagrama equivalente de la Figura 5.17. En este se representan las secciones indicadas del tracto vocal por líneas de transmisión. Ya que estas secciones están compuestas de tubos uniformes, se representan en diagrama como líneas de transmisión no uniformes. En las consonantes nasales, la boca se encuentra cerrada y $Z_m=\infty$. En las vocales nasalizadas, las dos trayectorias aparecen en paralelo.

En cualquier caso, la presencia de una trayectoria paralela significa que la función de transferencia ya no es solamente polar. La forma mas fácil de notar esto es debido a que la impedancia del tubo

adicional, bifurca al tubo principal. Esta impedancia siempre tendrá ceros de frecuencia finita. Si existe un cero en Z_n en alguna frecuencia f_z, entonces a esa frecuencia Z_n se comporta como corto circuito. De aquí que exista un cero en la función de transferencia en f_z. Las dimensiones de las cavidades involucradas son tales que típicamente existe un cero en el rango de F_1 a F_4.

En las constantes nasales, la rama principal es la cavidad nasal y la cavidad oral es la rama lateral. Un efecto adicional, que el análisis no demuestra es una atenuación de todos los formantes, debida a las pérdidas causadas por la cavidad nasal. En el caso de las vocales nasalizadas, la boca proporciona la trayectoria principal y los conductos nasales están desacoplados.

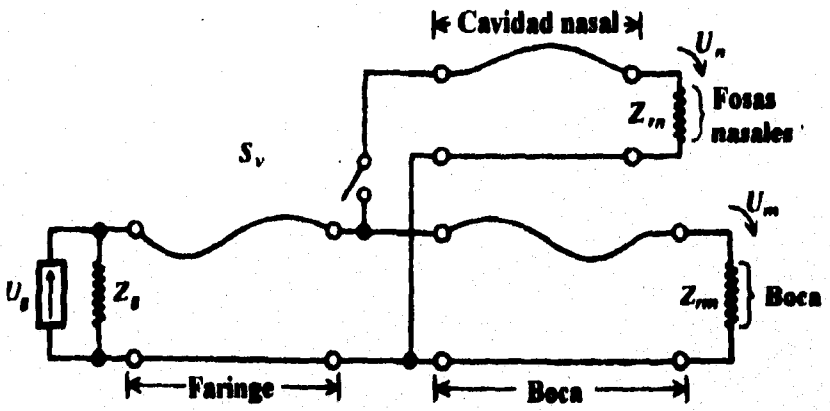


Figura 5.17 Modelo eléctrico de producción de voz, mostrando las secciones del tracto vocal modelado por líneas de transmisión no uniformes (Fig. 5.19 [PARS87]).

5.5 Características Acústicas de las Explosivas y las Fricativas.

Acústicamente, una fricativa aparece en forma similar a ruido de banda ancha y de corta duración. Una explosiva aparece como un período corto de silencio seguido de una liberación abrupta. La liberación normalmente no es limpia y usualmente se muestra como una pequeña ráfaga de ruido con un comienzo abrupto.

Si existe sonoridad durante la aparición de una explosiva o una fricativa, entonces será visible en el espectro de sonido una estructura de tipo formante.

Las características acústicas de las explosivas y las fricativas se pueden dividir en dos grupos:

1. Composición espectral del ruido.
2. Transiciones de los formantes en las vocoides adyacentes.

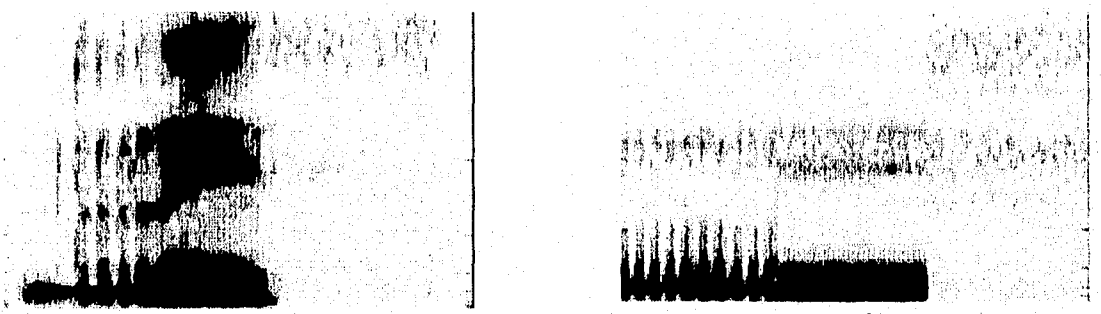
Ruido. Los siguientes puntos representan algunas conclusiones acerca de las características del ruido:

1. El ruido alveolar/dental [t,d,s] es generalmente de alta frecuencia (la concentración de energía esta por encima de los 4kHz) y de alta energía.
2. El ruido labial [p,b] es generalmente de baja frecuencia (la concentración de energía esta por debajo de los 2 kHz) y de baja energía.

3. El ruido velar [k,x] es de frecuencia media estilo formante y energía media. Los formantes resultan de la resonancia de la cavidad mas allá del punto de articulación pero es altamente dependiente de la vocal adyacente.

El ruido fricativo no es altamente dependiente del punto de articulación y por lo tanto no se ha podido distinguir las fricativas mediante la observación de su espectro de ruido.

Transición de los formantes. En los primeros experimentos con espectrogramas, se notó que las trayectorias de los formantes estaban curvadas en la vecindad con consonantes. En la Figura 5.18 se puede observar la transición real y la transición ideal de los formantes en la vecindad con consonantes. A estas curvas se les denomina transiciones de los formantes.



(a) (b)

Figura 5.18 Espectrogramas de la palabra "res" (a)real y (b) sintetizada a partir de fonemas discretos.

5.6 Modelos de Producción de Voz

Para obtener un modelo excitación modulación con más detalle, debemos tratar de incluir todo, de esta forma se obtiene el modelo de la Figura 5.19 donde,

U_g es la corriente de excitación glotal (correspondiente a la velocidad volumétrica).
V_p es el voltaje de ruido faringeal fricativo.
V_m es el voltaje de ruido oral fricativo.
S_v es el conmutador velar.

Aquí las características de transmisión de los varios segmentos del tracto vocal están representados como filtros sintonizados. La combinación de las funciones de transferencia de los filtros de la faringe y de la boca dan lugar a la matriz de transición de la ecuación (5.21); el filtro nasal proporciona la interacción mencionada en la sección 5.3.5. las fuentes de ruido fricativo están localizadas en la faringe y en la boca, como aproximaciones a la gran variedad de lugares donde se pueden originar.

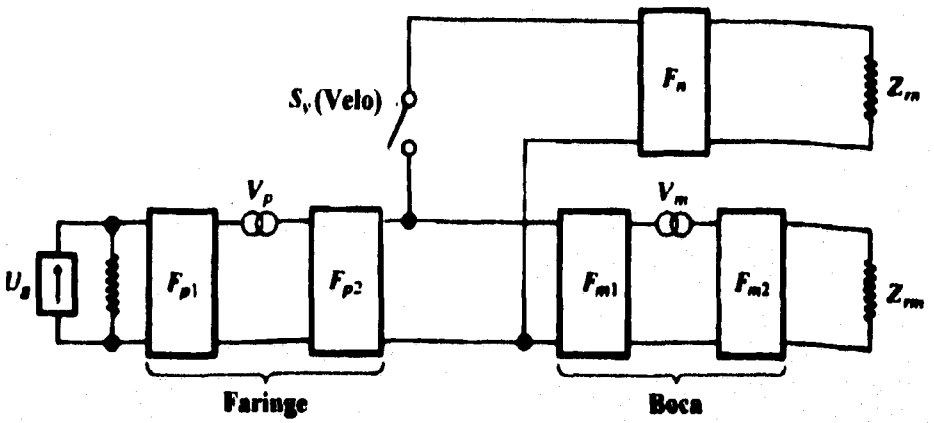


Figura 5.19 Modelo general del tracto vocal (Fig. 5.28 [PARS87]).

Normalmente se realizan las siguientes simplificaciones:

1. Se reemplazan las fuentes de ruido fricativo por una fuente equivalente en paralelo con U_g.

2. Se ignora el tracto nasal. Ya que las nasales son solamente otra función de transferencia, las podemos implantar usando el filtro oral.

Estas simplificaciones nos llevan al modelo mostrado en la Figura 5.20. Aquí U_g es la fuente glotal para sonidos sonoros y U_n es la fuente de ruido; Fg y Fn son los filtros de formado requeridos. Los conmutadores S_1 y S_2 permite la selección de las vocoides y las contoides tanto sonoras como sordas. F_v el tracto vocal y F_r toma en cuenta la impedancia de la radiación. Ahora realizaremos más simplificaciones:

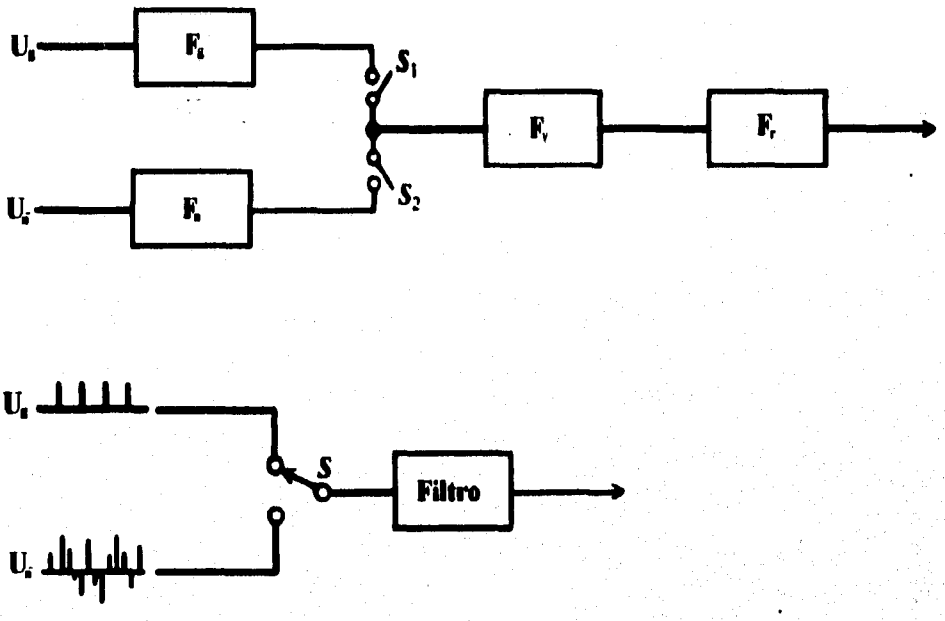
3. Reemplazamos Ug por un tren de impulsos y modificamos F_g para compensar.

4. Ignoramos la posibilidad de explosivas y fricativas sonoras y reemplazamos S_1 y S_2 por un conmutador de un polo dos tiros.

5. Consolidamos F_g, F_n, F_v y F_r en un solo filtro (Figura 5.20).

6. Ignoramos los ceros (ya que los humanos no perciben los ceros faltantes tanto como los polos faltantes); entonces el filtro incluye solamente polos.

Esta forma final se encuentra citada en diversas partes de la literatura. Es el modelo conceptual en el que la mayor parte de los análisis se basan; también es la base para el sintetizador de voz del tipo "terminal-analógico".



.c1.**Figura 5.20** Modelo simplificado del tracto vocal (Fig. 5.29 y 5.30 [PARS87]).

5.7 Estadística de las Señales de Voz

Concluimos con un breve resumen del las características mas generales de la voz. La densidad de probabilidad de la voz ha sido estudiada por McDonald(1966), encontró dos aproximaciones; la mejor de las dos es una variación de la densidad gamma

$$f_g(x) = \frac{\sqrt{k}}{2\sqrt{\pi}} \frac{e^{-k|x|}}{\sqrt{|x|}} \tag{5.23}$$

Una ecuación menos precisa es la densidad Laplaciana

$$f_L(x) = 0.5ae^{-a|x|} \tag{5.24}$$

La función de autocorrelación es fuertemente dependiente de lo que se ha dicho en ese momento, ocasionalmente es útil obtener estimados de la autocorrelación promedio de tiempo largo.

6. Procesamiento Digital de Señales[1]

6.1 Tiempo y Frecuencia Normalizada

$$s(n) = s_a(nT) = s_a(t)\Big|_{t=nT} \qquad n=...,-1,0,\ 1,\ 2,... \tag{6.1}$$

Donde n es el número de muestra

Se pierde la orientación en el tiempo en el argumento. Para recuperar la señal necesitamos conocer T. Para entender el significado del tiempo normalizado, imaginemos que escalamos el eje del tiempo real por un factor T antes de muestrear. El tiempo normalizado t' se relaciona al tiempo real como:

$$t' = \frac{t}{T} \tag{6.2}$$

y las muestras de voz son tomadas en intervalos que son exactamente "segundos normalizados" (el eje x carece de dimensiones). El intervalo entre muestras es el tiempo de muestreo, la frecuencia de Nyquist es siempre 0.5 norm-Hz o norm-rps. En general la conversión entre frecuencias "reales" F (Hz) y Ω (rps) y sus contrapartes normalizadas, f y w están dadas por

$$\begin{aligned} &f = FT \quad f = \frac{F}{Fs} \quad Fs \geq 2F_{max} \quad -0.5 \leq f \leq 0.5 \quad w = 2\pi \\ &w = \Omega T \quad f_{max} = \frac{F_{max}}{2F_{max}} = \frac{1}{2} = 0.5 \quad -0.5 \leq \frac{w}{2\pi} \leq 0.5 \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad -\pi \leq w \leq \pi \end{aligned} \tag{6.3}$$

6.2 Señales Singulares.

En el tiempo continuo, una señal singular es aquélla para la cual no existen una o más de sus derivadas en uno o más puntos en el tiempo. Aunque el concepto de su derivada no tiene significado en el tiempo discreto, utilizaremos el término singularidad para describir secuencias analógicas en el tiempo discreto. Las dos más utilizadas son:

La secuencia de impulso unitario o impulso tiempo discreto definido por

$$\delta(n) \stackrel{def}{=} \begin{cases} 1 & si\ n = 0 \\ 0 & coc \end{cases} \tag{6.4}$$

La secuencia escalón unitario definida por

[1] En su mayor parte tomado del capítulo 1 de [DELL87].

$$u(n) \overset{def}{=} \begin{cases} 1 & si \ n \geq 0 \\ 0 & coc \end{cases} \tag{6.5}$$

6.3 Señales de Energía y de Potencia.

Existen muchas formas en que una señal de tiempo discreto puede ser clasificada. Algunas de éstas son señal de energía, señal de potencia o ninguna. De acuerdo a la definición de energía de una señal discreta,

$$E_x \overset{def}{=} \sum_{n=-\infty}^{\infty} |x(n)|^2 \tag{6.6}$$

Una señal $x(n)$ puede ser llamada señal de energía si $0 < E_x < \infty$

La potencia en una secuencia de tiempo discreto es,

$$P_x \overset{def}{=} \lim_{N \to \infty} \frac{1}{2N+1} \sum_{n=-N}^{N} |x(n)|^2 \tag{6.7}$$

Una señal de potencia tiene potencia finita, pero no cero,

$$0 < P_x < \infty$$

Una señal no puede ser una señal de potencia y de energía simultáneamente, ya que si $E_x < \infty$ entonces $P_x = 0$. Pero una señal puede ser de ninguno de los dos tipos anteriores cuando $P_x = \infty$ o $E_x = 0$.

Para nuestro propósito de procesamiento de voz, es suficiente asociar la categoría de energía con dos clases de señales, éstas son:

- *Transientes*, aquéllas que decaen (usualmente exponencialmente) con el tiempo.
- *Secuencias finitas*, aquéllas que son cero fuera del intervalo finito de su duración.

Donde las señales de energía o decaen completamente de forma rápida o se detienen completamente, las señales de potencia ni decaen, ni incrementan sus envolventes. Las señales de potencia se pueden asociar con tres clases de señales. Estas son:

- Señales constantes.
- Señales periódicas.
- Realización de procesos estocásticos estacionarios y ergódicos.

Las señales que no caen en éstas categorías son la señal cero y aquéllas que en el tiempo crecen sin control.

6.4 Transformadas y Algunos Conceptos Relacionados.

Transformada de Fourier de tiempo discreto (DTFT)

$$X(\omega) = \sum_{n=-\infty}^{\infty} x(n) e^{-j\omega n} \text{ (directa)} \tag{6.8a}$$

$$x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(\omega) e^{j\omega n} d\omega \quad \text{(inversa)} \tag{6.8b}$$

Para que exista la DTFT, la secuencia $x(n)$ debe ser absolutamente sumable ,

$$\sum_{n=-\infty}^{\infty} |x(n)| < \infty \tag{6.9}$$

También debe de converger a una función en ω . Una señal que es absolutamente sumable debe ser necesariamente una señal de energía, ya que,

$$E_x = \sum_{n=-\infty}^{\infty} |x(n)|^2 \leq \left[\sum_{n=-\infty}^{\infty} |x(n)| \right]^2 \tag{6.10}$$

Aunque la DTFT es útil para análisis espectrales teóricos, no puede ser calculada en una computadora digital, porque es función de un argumento continuo. En principio también trabaja con una secuencia doble de longitud infinita, esto también limita el cómputo. Si restringimos el cálculo al caso práctico en que la secuencia es de longitud finita, entonces la transformada discreta de Fourier proporciona un mapeo entre la secuencia

$$x(n), \; n = 0,1,2,\ldots,N-1$$

y un conjunto discreto de muestras en el dominio de la frecuencia dado por,

$$X(k) = \begin{cases} \sum_{n=0}^{N-1} x(n) e^{-j(2\pi/N)kn} & k = 0,1,\ldots,N-1 \\ 0 & (\text{otra } k) \end{cases} \quad (\textit{directa}) \tag{6.11}$$

$$x(n) = \begin{cases} \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{j(2\pi/N)kn} & n = 0,1,\ldots,N-1 \\ 0 & (\text{otra } n) \end{cases} \quad (\textit{inversa}) \tag{6.12}$$

La DFT representa muestras exactas de la DTFT de la secuencia finita x(n) en N frecuencias igualmente espaciadas, $\omega k = (2\pi/N)k$, para $k \in [0, N-1]$.

La serie discreta de Fourier (DFS) está muy relacionada a la DFT computacionalmente, pero es diferente en concepto. La DFS se usa para representar una secuencia periódica de período N usando las funciones bases $e^{j(2\pi/N)kn}$ para $k = 0,\ldots, N-1$. Estas representan N frecuencias armónicas que puede estar presentes en la señal. Para una señal periódica $y(n)$ su expansión es,

$$y(n) = \sum_{k=0}^{N-1} C(k) e^{j(2\pi/N)kn} \qquad (6.13a)$$

donde los coeficientes se calculan como,

$$C(k) = \frac{1}{N} \sum_{n=0}^{N-1} y(n) e^{-j(2\pi/N)kn} \qquad (6.13b)$$

La transformada rápida de Fourier (FFT) es el nombre que se le dan a una serie de algoritmos rápidos para calcular la DFT.

La transformada Z (ZT) (bilateral) definida como,

$$x(z) = \sum_{n=-\infty}^{\infty} x(n) z^{-n} \text{ (directa)} \qquad (6.14a)$$

$$x(n) = \frac{1}{2\pi j} \oint_{\ell} X(z) z^{n-1} dz \text{ (inversa)} \qquad (6.14b)$$

Donde z es cualquier número complejo para el que la suma existe,

$$\sum_{n=-\infty}^{\infty} |x(n) z|^{-n} < \infty$$

Los valores de z para los que la serie converge, constituyen la región de convergencia (ROC) de la (ZT). La *ZT* juega un papel similar en el Procesamiento Digital de Señales (DSP) al que la transformada de Laplace lo hace en el procesamiento continuo. La interpretación de los diagramas polo-cero en el plano Z, son una herramienta fundamental para el ingeniero en procesamiento de voz.

Finalmente estableceremos las relaciones entre las transformadas de Fourier y la transformada Z. De las definiciones, es claro que, $^{DTFT}X(\omega) = {}^{ZT}X(e^{j\omega})$ para cualquier ω, la DTFT en la frecuencia ω se obtiene evaluando la *ZT* en un ángulo ω en el círculo unitario en el plano Z.

$$^{DFT}X(k) = {}^{DTFT}X\left(\omega_k = \frac{2\pi}{N}k\right) = {}^{ZT}x\left(e^{j(2\pi/N)k}\right) \qquad (6.15)$$

6.5 Ventanas y Tramas.

En las aplicaciones prácticas de procesamiento de señales, es necesario trabajar con segmentos cortos o tramas de la señal, a menos que la señal sea de corta duración. Esto es esencialmente cierto si usamos técnicas convencionales de análisis de señales con dinámica no estacionaria (como las señales de voz). En este caso es necesario seleccionar una porción de la señal que se pueda asumir estacionaria.

Ya que la ventana w(n) (en el dominio del tiempo) es real, la secuencia de longitud finita se usa para seleccionar una trama deseada de la señal original, x(n) por un proceso simple de multiplicación. Existen diferentes tipos de ventanas. Para ser consistentes, asumiremos que las ventanas son secuencias casuales y empiezan en n=0, La duración es N. Normalmente las ventanas son simétricas en el tiempo (N-1)/2, donde este tiempo puede estar a la mitad de dos muestras si N es par. Debido a que las ventanas son secuencias de fase lineal y por lo tanto tienen DTFT, pueden ser escritas como

$$W(\omega) = |W(\omega)|e^{j\omega(N-1)/2} \tag{6.16}$$

Donde el término de fase es simplemente una característica lineal que corresponde al retardo de la ventana que la hace causal.

Una trama de la señal $x(n)$ de longitud N que termina al tiempo m, digamos $f_x(n;m)$ se obtiene como,

$$f_x(n;m) = x(n)w(m-n) \tag{6.17}$$

La DTFT de $f_x(n;m)$ es

$$F_x(\omega;m) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\omega-\theta)W(-\theta)e^{-j\theta m}d\theta \tag{6.18}$$

sustituyendo 6.16 en 6.18

$$F_x(\omega;m) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\omega-\theta)W(-\theta)d\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\omega-\theta)|W(\theta)|d\theta \tag{6.19}$$

sustituimos $|W(-\theta)|$ por $|W(\theta)|$ ya que el espectro de magnitud es una función par de θ.

Ahora queremos que la ventana tenga un espectro de magnitud que aproxima a un impulso

$$|W(\theta)| \approx 2\pi\delta_a(\theta) \tag{6.20}$$

esto implicaría $F_x(\omega;m) \approx X(\omega)$, que significa que no existe ventana, esto conservaría las características temporales, originales de la señal.

Una buena ventana es aquélla que aproxima (6.20) y por lo tanto preserva las características espectrales de $X(\omega)$. Todas las ventanas tienden a tener un espectro pasabajas, con un lóbulo principal en bajas frecuencias y varios lóbulos laterales atenuados.

Para que el espectro de cualquier ventana se aproxime a $\delta_a(\omega)$, se desean dos características:

- Un lóbulo principal de banda angosta.
- Una gran atenuación en los lóbulos laterales.

Algunas ventanas muy utilizadas que proveen de corte más suave son Kaiser, Hamming[2], Hanning y Blackman. Estas tienden a distorsionar la forma de onda temporal en el rango de N puntos, pero

[2] Esta es la ventana que se utilizó en el presente trabajo.

con el beneficio de no truncar en los límites. Las propiedades espectrales de éstas ventanas se describen como sigue:

- Para un tamaño *N*, todas poseen un lóbulo principal más ancho que la rectangular, este decrece con el valor de *N*.
- Todas tienen mejor atenuación en los lóbulos laterales que la rectangular, típicamente 10- 60dB mejor.

Otro compromiso que se tiene es el tamaño de la ventana. Ya que si se tiene una ventana muy grande la señal tiende a ser no estacionaria, a este compromiso se le llama, "compromiso de Resolución Espectral-Temporal".

6.6 Sistemas de Tiempo Discreto.

Algunos conceptos importantes que se deben de manejar para el procesamiento digital de voz:

1. Linealidad.
2. Invariancia en el tiempo.
3. Ecuación diferencial lineal a coeficientes constantes que describe un sistema lineal invariante (LTI) en el tiempo discreto (DT).
6. Respuesta al impulso en tiempo discreto de un LTI en el tiempo discreto ["*h(n)*"] .
5. Suma de convolución de un Sistema LTI DT
6. Estabilidad de entrada acotada salida acotada (BIBO) y su relación a *h(n)* en un sistema LTI DT.
7. Causalidad.
8. Función de transferencia de un LTI DT ["*H(z)*"], polos y ceros.
9. Espectro de magnitud y fase de un sistema LTI DT y su relación con un diagrama de polos y ceros.
10. Relación entre una ecuación de diferencias lineal a coeficientes constantes y H(z) para un sistema LTI DT.
11. Relación entre estabilidad BIBO y *H(z)*.
12. Sistemas de respuesta al impulso finita (FIR) y respuesta infinita al impulso (IIR) y su relación con *H(z)* y la ecuación en diferencias.
13. Estructuras de cálculo canónicas para la implantación de sistemas LTI DT.

6.6.1 Realizaciones en el espacio de estados de sistemas LTI DT

Consideremos el sistema gobernado por la siguiente ecuación en diferencias a coeficientes constantes:

$$y(n) = \sum_{k=1}^{M} a(k)y(n-k) + \sum_{k=0}^{Q} b(k)x(n-k) \qquad (6.21)$$

para la cual se muestra su realización de la forma directa II en la Figura 6.1.

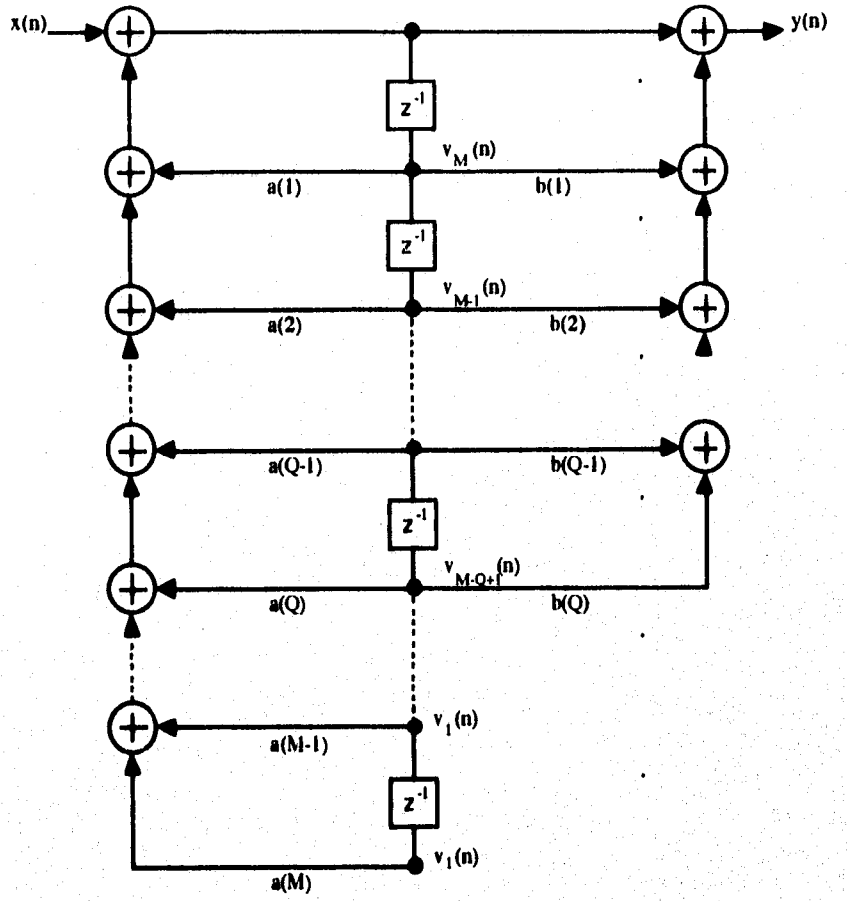


Figura 6.1 Realización en la forma directa II de un sistema discreto (Fig. 1.5 [PARS87]).

Asumimos que $Q<M$ y definimos $b(k) > 0$ para $k > Q$. El estado interno de un sistema DT en el tiempo n_0 es definido como la información cuantitativa necesaria al tiempo n_0, que junto con la entrada x(n) para $n \geq n_0$, determina de forma única la salida (n) para $n \geq n_0$. Las variables de estado del sistema son las cantidades numéricas almacenadas por el sistema que comprenden ese estado.

En la figura se han definido las variables $v_1(n),\ldots,v_M(n)$. Estas comprenden las variables de estado del sistema. Nótese que,

$$\begin{aligned} v_i(n+1) &= v_{i+1}(n) \qquad i = 1,2,\ldots,M-1 \\ v_M(n+1) &= x(n) + \sum_{i=1}^{M} a(i)v_{M-i+1}(n) \end{aligned} \tag{6.22}$$

Estas son las ecuaciones de estado del sistema. Nótese que la salida puede calcularse de las variables de estado o al tiempo n usando,

$$\begin{aligned} y(n) &= b(0)v_M(n+1) + \sum_{i=1}^{M} b(i)v_{M-i+1}(n) \\ &= b(0)x(n) + \sum_{i=1}^{M}\left[b(i) + b(0)a(i)\right]v_{M-i+1}(n) \end{aligned} \tag{6.23}$$

Que es llamada la ecuación de salida del sistema. Es claro que estas variables de estado constituyen un estado legítimo para el sistema de acuerdo a la definición. Por conveniencia las ecuaciones de estado y de salida se pueden escribir en forma vector-matriz de la siguiente forma:

$$\begin{aligned} \mathbf{v}(n+1) &= \mathbf{A}\mathbf{v}(n) + \mathbf{c}x(n) \\ y(n) &= \mathbf{b}^T\mathbf{v}(n) + dx(n) \end{aligned} \tag{6.24}$$

en donde d es el valor constante $d=b(0)$, $\mathbf{A}$ es la matriz de transición de estados $M \times M$

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 \\ a(M) & a(M-1) & a(M-2) & a(M-3) & a(M-4) & \cdots & a(1) \end{bmatrix}$$

y donde, $\mathbf{c}$ y $\mathbf{b}$ son vectores de dimensión M,

$$\mathbf{c} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}^T$$

$$\mathbf{b} = \begin{bmatrix} b(M) + b(0)a(M) \\ b(M-1) + b(0)a(M-1) \\ b(M-2) + b(0)a(M-2) \\ \vdots \\ b(1) + b(0)a(1) \end{bmatrix} \tag{6.25}$$

6.7 Señales y Sistemas de Fase Mínima, Máxima y Mixta.

Una señal con todos sus ceros dentro del círculo unitario se le denomina señal de fase mínima. Si la señal es la respuesta al impulso discreta de un sistema, entonces se le denomina sistema de fase mínima o filtro. Cuando los ceros están completamente fuera del círculo unitario, la señal (o sistema) se llama de fase no mínima. Todos los casos intermedios normalmente se les denomina de fase mixta.

Cuando los ceros están dentro del círculo unitario, se minimiza el valor absoluto de la fase negativa para una ω dada. En contraparte, cuando los ceros están fuera del círculo unitario, se maximiza la fase negativa. Existe una noción más intuitiva en el dominio del tiempo.

Ya que la fase (negativa) en ω está directamente relacionada con la cantidad de retardo temporal de una componente de banda angosta de la señal a esa frecuencia, se puede inferir que la señal de fase mínima es aquélla que para un espectro de magnitud dado, tiene un retardo mínimo para cada componente frecuencial en la concentración de energía cerca del tiempo $n=0$ de cualquier señal con el mismo espectro de magnitud. Específicamente, si $x_{\min}(n)$ es la señal de fase mínima y $E_x(m)$ representa la energía para cualquier $x(n)$ en el intervalo $n \in [0,m]$,

$$E_x(m) \stackrel{def}{=} \sum_{n=0}^{m} x^2(n) \tag{6.26}$$

entonces es cierto que,

$$E_{x_{\min}}(m) \geq E_x(m) \tag{6.27}$$

para una señal absolutamente sumable $x(n)$ con el mismo espectro de magnitud y para cualquier m. Los conceptos de fase mínima representan un papel importante en la teoría de predicción lineal y conceptos de modelado.

7. Técnicas de Modelado de Señales para Reconocimiento de Voz[1]

7.1 Introducción

El primer paso en el proceso de reconocimiento de voz, es la parametrización[2] de una señal analógica de voz. Existen algunas técnicas de análisis de señales que actualmente se toman como los estándares de la literatura. Estos algoritmos pretenden que la representación paramétrica de las señales de voz tengan un "significado perceptible" es decir: parámetros que emulan parte del comportamiento observado en los sistemas auditivos y perceptuales del ser humano. Indudablemente, estos algoritmos también están diseñados para maximizar el desempeño del reconocimiento.

En el reconocimiento de voz independiente del locutor, se tiene un gran avance en las descripciones que están siendo desarrolladas, ya que de cierta forma son invariantes a los cambios del locutor. Lo que deseamos es contar con parámetros que representen los principales espectros de energía del sonido, en vez de información sobre la voz del locutor en particular.

El modelado de señales representa el proceso de convertir las muestras tomadas a una secuencia de voz a vectores de observación representando eventos en un espacio probabilístico. La búsqueda de redes[3] [RABI89], tiene como propósito encontrar la secuencia de mayor probabilidad de estos eventos, dadas ciertas condiciones sintácticas.

7.1.1 El Paradigma del Modelado de Señales

El modelado de señales puede subdividirse en cuatro operaciones básicas: modelado de la señal, análisis del espectro, transformación paramétrica y modelado estadístico.

Existen tres fuerzas directrices en el diseño de sistemas de modelado de señales. En primer lugar, las parametrizaciones se buscan para representar aspectos trascendentales de la señal de voz, preferentemente parámetros que son análogos al sistema auditivo humano. Estos parámetros son conocidos usualmente como de significado perceptible. En segundo lugar, se desea que la parametrización sea capaz de mantenerse a las variaciones en el canal, el locutor o el transductor. Usualmente nos referimos a éste como un problema de robustez o de invariancia. Por último, se desea que los parámetros capten las variaciones del espectro en el tiempo. Con esto nos referimos al problema de correlación temporal. Con la introducción de las técnicas de modelado de Markov que permiten modelar estadísticamente la transformación en tiempo de la señal, los parámetros que incorporan mediciones absolutas y diferenciales del espectro de la señal son ahora habituales.

Actualmente el modelado de señales requiere menos del 10% del tiempo total de proceso requerido en una aplicación de reconocimiento de voz con vocabulario considerable. Las parametrizaciones que describen una señal, pueden ser procesadas fácilmente utilizando hardware de punto fijo, y ser

---

[1] En su mayor parte tomado de [PICO93].

[2] Obtención de parámetros que caracterizan la señal. Información relevante del proceso.

[3] Los modelos ocultos de Markov son ejemplo de la búsqueda de redes.

comprimidas mediante técnicas de cuantización precisas, lo cual es preferible sobre aproximaciones poco comunes.

Históricamente, la robustez hacia el ruido acústico de fondo ha sido una de las pautas que han encauzado el diseño de modelos de señales. Adicionalmente sabemos que, un modelo excelente para una aplicación, no forzosamente tiene que ser óptimo para otra.

7.1.2 Terminología

El modelo de una señal, consta de tres componentes internos: mediciones, mediciones espectrales básicas y temporales; parámetros, versiones comparadas paramétricamente y versiones suavizadas de estas mediciones; y observaciones, la resultante de alguna forma de modelado estadístico de los parámetros.

7.2 Formación de Espectro

La formación espectral involucra dos operaciones básicamente: conversión A/D, conversión de una señal, de una onda de presión del sonido a una señal digital; y filtrado digital, enfatizando las principales componentes de frecuencia de la señal.

Debido a la limitada respuesta en frecuencia de los canales de telecomunicaciones analógicos, y el uso de una frecuencia de muestreo de 8 kHz en los canales de telefonía digital, la frecuencia de muestreo más utilizada para una señal de voz en telecomunicaciones, es 8 kHz. Para otro tipo de aplicaciones, en las que los subsistemas de reconocimiento de voz cuentan con alto grado de calidad, se utilizan frecuencias de muestreo de 10, 12 y 16 kHz. Estas frecuencias de muestreo, permiten mejorar la resolución de las señales en tiempo y frecuencia.

La finalidad del proceso de digitalización es obtener información muestreada que represente una señal de voz mediante una relación señal a ruido tan alta como sea posible. Actualmente, los sistemas de telecomunicaciones entregan una relación señal a ruido por encima de los 30 dB para aplicaciones de reconocimiento de voz, suficiente para obtener un alto desempeño.

Una vez terminada la conversión, el último paso de posfiltrado digital se realiza utilizando un filtro de Respuesta al Impulso Finita

$$H_{pre}(z) = \sum_{k=0}^{N_{pre}} a_{pre}(k) z^{-k} \qquad (7.1)$$

Generalmente se utiliza, un filtro digital de un coeficiente, conocido como filtro de preénfasis

$$H_{pre}(z) = 1 + a_{pre} z^{-1} \qquad (7.2)$$

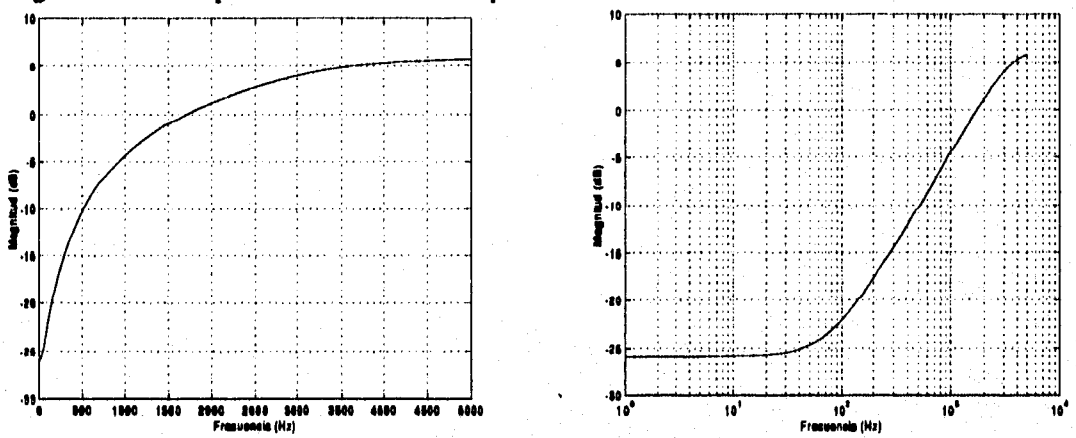
Los valores típicos para a_{pre} varían entre [-1.0, -0.4]. Los valores cercanos a -1.0 que pueden ser implementados fácilmente utilizando hardware de punto fijo, tales como -1.0 ó -(1-1/16), son frecuentes en reconocimiento de voz. Un filtro de preénfasis tiene por objeto incrementar el espectro de la señal aproximadamente 20 dB por década (la escala de incremento de magnitud en frecuencia).

Existen dos explicaciones sobre las ventajas de utilizar este tipo de filtro. En primer lugar, las secciones sonoras de una señal de voz, naturalmente presentan una pendiente espectral negativa (atenuación) de aproximadamente 20 dB por década, debido a las características fisiológicas del

sistema de producción de voz. El filtro de preénfasis ayuda a restaurar esta pendiente natural antes de realizar el análisis espectral, con la finalidad de mejorar la eficiencia del análisis.

Una explicación alternativa es que la audición, es más sensitiva arriba de la región espectral de 1 kHz. El filtro de preénfasis amplifica esta región del espectro, auxiliando los algoritmos de análisis espectral en el modelado de los aspectos perceptuales de mayor importancia del espectro de voz.

Es necesario hacer notar, que tales filtros de preénfasis también incrementan las frecuencias arriba de los 5 kHz, región en la cual el sistema auditivo es menos sensitivo. De cualquier manera, estas frecuencias son atenuadas naturalmente por el sistema de producción de voz además de tener asignado un menor peso dentro del sistema típico de reconocimiento de voz.



(a) (b)

Figura 7.1 Respuesta del filtro de preénfasis

7.3 Análisis Espectral

Clasificaremos los tipos de mediciones espectrales utilizadas en sistemas de reconocimiento de voz en dos clases: potencia — mediciones de la potencia espectral total (o temporal) de la señal; amplitud espectral —mediciones de potencia sobre un intervalo particular de frecuencia en el espectro. El conjunto típico de parámetros para reconocimiento de voz debe incluir estas dos mediciones.

7.3.1 Frecuencia Fundamental

La frecuencia fundamental se define como la frecuencia a la cual las cuerdas vocales vibran durante un sonido sonoro. La frecuencia fundamental (f_0) ha sido un parámetro difícil de estimar confiablemente a partir de la señal de voz.

Hoy en día se utilizan cuatro clases principales de algoritmos. Uno de los algoritmos principales, y uno de los mas sencillos, es el que utiliza varias mediciones de periodicidad en la señal, y las discrimina para determinar el estado de sonoridad y la frecuencia fundamental. Este algoritmo, fue conocido originalmente como el Algoritmo Gold-Rabiner.

La segunda clase de algoritmos a utilizar fue ideado por, la Agencia de Seguridad Nacional de Los Estados Unidos (NSA por sus siglas en inglés), como parte de un programa para el desarrollo de teléfonos digitales confiables basados en codificación de voz de baja tasa de transmisión, este es un algoritmo robusto para aplicaciones de telecomunicaciones. El algoritmo se basa en la función de diferencias del promedio de la magnitud, así como de un análisis discriminatorio de las distintas

mediciones de la voz. Se trata de un estándar gubernamental, disponible al público en los Estados Unidos.

La tercera clase de algoritmos, similares en su naturaleza a los antes mencionados, se basa en los conceptos de la programación dinámica. Esta clase de algoritmos utilizan un sofisticado procedimiento de optimización que evalúa distintas mediciones de correlación y cambio espectral de la señal, para llegar a un patrón óptimo de la frecuencia fundamental aunado al patrón de voz.

Por último, un algoritmo pocas veces utilizado en los sistemas de voz de tiempo real, pero frecuentemente utilizado en la investigación por experimentación, es un algoritmo que opera en el cepstrum de la señal de voz. La frecuencia fundamental es normalmente procesada en una escala logarítmica, y no una escala lineal, para relacionarla con la resolución del sistema auditivo humano. A modo de referencia, una medida de la frecuencia fundamental se define como

$$f(n) = \log_{10}\left(f_0(n)\right) \tag{7.3}$$

donde n representa tiempo discreto.

Usualmente, la frecuencia fundamental para palabras sonoras es de 50 Hz $\leq f_0 \leq$ 500 Hz. Para palabras sordas, f_0 no se define y por convención $f_0 = 0$. Frecuentemente, la frecuencia fundamental es normalizada en base al valor promedio de f_0 del locutor, o por alguna transformación fisiológica de un valor nominal durante el segmento de voz sonora correspondiente.

7.3.2 Potencia

El uso de ciertas de mediciones de potencia en reconocimiento de voz es un estándar hasta cierto punto razonable. La *Potencia* es bastante simple de calcular

$$P(n) = \frac{1}{N_s}\sum_{m=0}^{N_s-1}\left(\omega(m)s\left(n-\frac{N_s}{2}+m\right)\right)^2 \tag{7.4}$$

donde N_S es el número de muestras utilizadas para calcular la potencia, $s(n)$ denota la señal, $\omega(m)$ denota la función de ponderación, y n denota el índice de muestreo (tiempo discreto) del centro de la ventana. En lugar de utilizar la potencia directamente, muchos sistemas de reconocimiento de voz hacen uso del logaritmo de la potencia multiplicado por 10, definido como la potencia en decibeles, en un esfuerzo por emular la respuesta logarítmica del sistema auditivo humano.

La función de ponderación (7.4), es una función de ventana. La teoría del uso de ventanas fue en algún tiempo un punto muy importante en la investigación del procesamiento digital de señales. Existen muchos tipos de ventanas incluyendo la Rectangular, Hamming, Hanning, Blackman, Bartlett y Kaiser. Actualmente, en reconocimiento de voz, la ventana Hamming es la más utilizada, siendo un caso específico de la ventana Hanning. La ventana Hamming se define como,

$$\omega(n) = \frac{\alpha_\omega - (1-\alpha_\omega)\cos(2\pi n/(N_s-1))}{\beta_\omega} \tag{7.5}$$

para $0 \leq n < N_S$, y $\omega(n) = 0$ para cualquier otro valor. α_ω se define como una ventana constante en el rango de [0,1], y N_S es la duración de la ventana en muestras. Para implementar una ventana Hamming, $\alpha_\omega = 0.54$.

β_ω es la constante de normalización definida de modo que el valor RMS de la ventana sea la unidad. β_ω se define como,

$$\beta_\omega = \sqrt{\frac{1}{N_s}\sum_{n=0}^{N_s-1}\omega^2(n)} \qquad (7.6)$$

En la práctica, es deseable normalizar la ventana de manera que la potencia en la señal después de aplicarle la ventana sea aproximadamente igual a la que se tenía antes de aplicarle la ventana.

La finalidad de la aplicación de la ventana es ponderar, o facilitar, el muestreo hacia el centro de ésta. Esta característica junto con el análisis con traslape que se discutirá a continuación, desempeñan una función muy importante en la obtención de estimaciones paramétricas que varían suavemente. Es importante que el ancho de banda del lóbulo principal en la respuesta en frecuencia de la ventana, sea tan pequeña como sea posible, o el proceso al aplicar la ventana puede presentar efectos no deseables en el análisis espectral subsecuente.

La potencia, como la mayoría de los parámetros en un sistema de reconocimiento de voz (incluyendo la frecuencia fundamental mencionada en la sección anterior), es procesada utilizando el principio de trama-por-trama. La duración de la trama T_f se define como el tiempo (en segundos) durante el cual es válido un conjunto de parámetros. El periodo de la trama es un término utilizado de manera similar que denota el tiempo entre el cálculo de dos conjuntos de parámetros sucesivos. La razón de tramas, de igual forma es un término frecuentemente utilizado, es el número de tramas procesadas por segundo (Hz).

En la ecuación (7.4), n es actualizada por la duración de la trama en muestras. La duración de la trama habitualmente toma rangos de entre 20 y 10 ms, para sistemas prácticos. Los valores en este rango, representan un compromiso entre la tasa de variación del sistema y su complejidad. La misma duración de la trama finalmente depende de la velocidad de los sistemas articuladores en el sistema de producción de voz (tasa de intercambio de la forma del tracto vocal). Mientras algunos sonidos sonoros (tales como consonantes explosivas o diptongos) presentan transiciones espectrales abruptas lo que puede originar una elevación del pico espectral de hasta 80 Hz/ms, las tramas con duración menor a 8 ms, no se utilizan normalmente.

De igual importancia, es el intervalo sobre el cual se realiza el cálculo de la potencia. El número de muestras necesarias para el cálculo de la sumatoria N_S se denomina duración de la ventana (en muestras). La duración de la ventana T_ω generalmente se mide en unidades de tiempo (segundos).

La duración de la ventana controla la cantidad de promediación o suavización, utilizada para el cálculo de la potencia. La duración de la trama junto con la de ventana, controla la tasa a la cual los valores de potencia siguen los cambios dinámicos de la señal. La duración de la trama y la de ventana normalmente se ajustan por pares: la duración de ventana de 30 ms es muy utilizada con una duración de trama de 20 ms, mientras que para una duración de ventana de 20 ms se utiliza una duración de trama de 10 ms. En términos generales, dado que una trama de duración breve es utilizada para captar variaciones imprevistas del espectro, la duración de la ventana debe ser breve de manera que los detalles del espectro no sean excesivamente suavizados.

El proceso basado en el análisis de trama se ilustra en la Figura 7.2. A esta clase de análisis usualmente se le conoce como análisis con traslape, dado que a cada nueva trama, únicamente varía una fracción de los datos de la señal. La cantidad de traslape de alguna forma controla la rapidez

con la que los parámetros cambian de una trama a otra. El porcentaje de traslape esta dado por la ecuación:

$$\%\,Overlap = \frac{T_w - T_f}{T_w} \times 100\% \qquad (7.7)$$

donde T_w es la duración de la ventana (en segundos) y T_f es la duración de la trama. Si $T_w < T_f$, el porcentaje de traslape es cero.

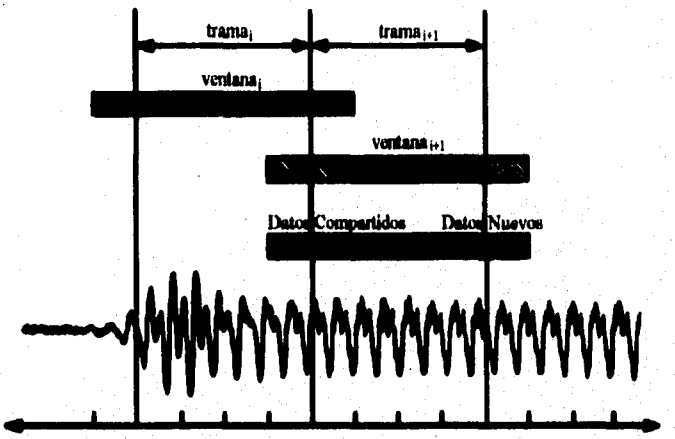


Figura 7.2 Análisis con ventanas traslapadas (Fig. 6 [PICO93]).

La combinación de 20 ms de duración de trama y 30 ms de duración de ventana corresponden al 33% de traslape. Algunos sistemas utilizan hasta 66% de traslape. Una de las metas de dichas cantidades de traslape, es reducir la cantidad de ruido introducido en las mediciones por tales artifactos como la disposición de ventanas y ruido de canal no estacionario. Por otra parte, las estimaciones excesivamente suavizadas pueden alterar cualquier variación real en la señal.

7.3.3 Análisis Espectral

En los sistemas de reconocimiento de voz, normalmente se utilizan seis algoritmos principalmente en el análisis espectral. Los procedimientos para generar dichos análisis se sintetizan en la Figura 7.3. Los métodos de banco de filtros (implementados en circuitos analógicos) fueron los primeros en utilizarse. Los métodos de predicción lineal fueron introducidos en la década de los setentas, y fueron la técnica dominante en los inicios de la década de los ochenta. Actualmente, tanto la transformada de Fourier, como las técnicas de predicción lineal son ampliamente utilizados en distintas aplicaciones de procesamiento de voz.

7.3.3.1 Banco de Filtros Digitales

Uno de los conceptos fundamentales en el procesamiento de voz es el de banco de filtros digitales. Un banco de filtros puede ser considerado como un modelo básico de las etapas iniciales de la transducción en el sistema auditivo humano. Existen dos motivaciones principales para la representación en banco de filtros. En primer lugar, la posición de máxima desplazamiento a lo largo de la membrana basilar para estímulos tales como tonos puros es proporcional al logaritmo de la frecuencia del tono. Esta hipótesis es parte de la teoría de audución llamada "place theory".

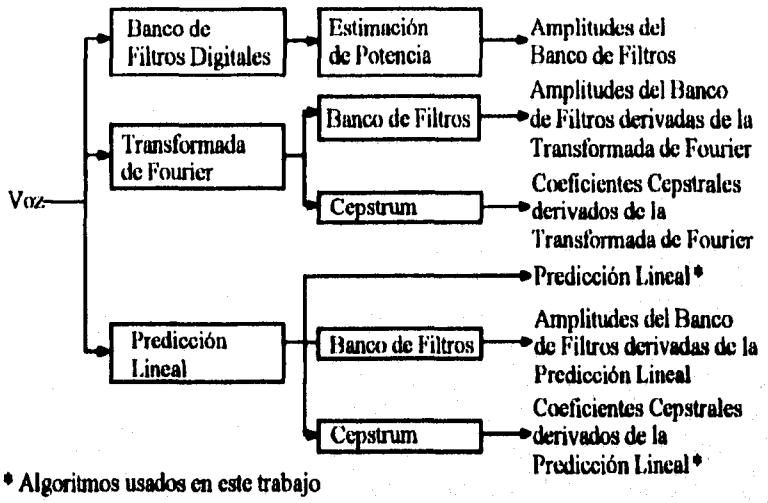


Figura 7.3 Los seis algoritmos de análisis espectral más utilizados (Fig. 6 [PICO93]).

En segundo lugar, los experimentos sobre la percepción humana han mostrado que las frecuencias de un sonido complejo en cierto ancho de banda de alguna frecuencia nominal no pueden ser identificadas individualmente. Cuando alguna de las componentes de estos sonidos caen fuera de este ancho de banda, pueden ser identificadas individualmente. Dicho ancho de banda se conoce como el ancho de la banda crítica. El ancho de la banda crítica se encuentra nominalmente de 10% a 20% de la frecuencia central del sonido.

Es posible definir un mapeo de frecuencias acústicas f a una escala de frecuencia "perceptual" de la siguiente manera:

$$Bark = 13\tan^{-1}\left(\frac{0.76f}{1000}\right) + 3.5\tan^{-1}\left(\frac{f^2}{(7500)^2}\right) \qquad (7.8)$$

Las unidades en esta escala de frecuencia perceptual se conocen como la tasa de bandas críticas, o Bark. La escala Bark se muestra en la Figura 7.4(a).

Una aproximación muy utilizada en este tipo de mapeo en reconocimiento de voz es conocido como la escala mel

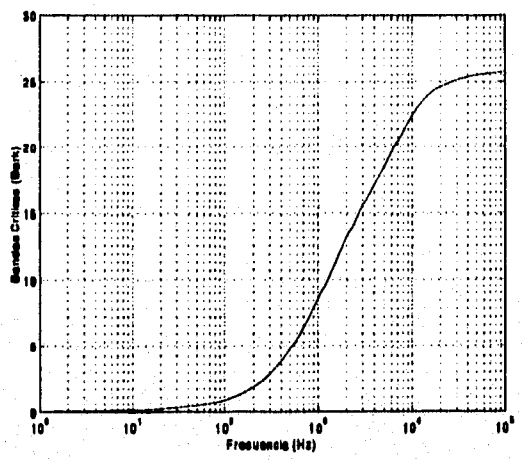
$$Frecuencia\ mel = 2595\log 10(1 + f/700.0) \qquad (7.9)$$

La escala mel pretende mapear la frecuencia percibida de un tono, dentro de una escala lineal. Este escalamiento se muestra en la Figura 7.3(b). Generalmente es aproximada como una escala lineal que va de 0 a 1000 Hz, y posteriormente una escala logarítmica que va mas allá de los 1000 Hz.

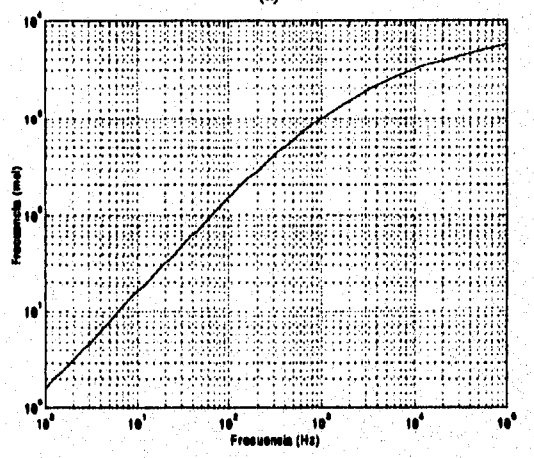
Una expresión para el ancho de la banda crítica es

$$BW_{critical} = 25 + 75\left[1 + 1.4(f/1000)^2\right]^{0.69} \qquad (7.10)$$

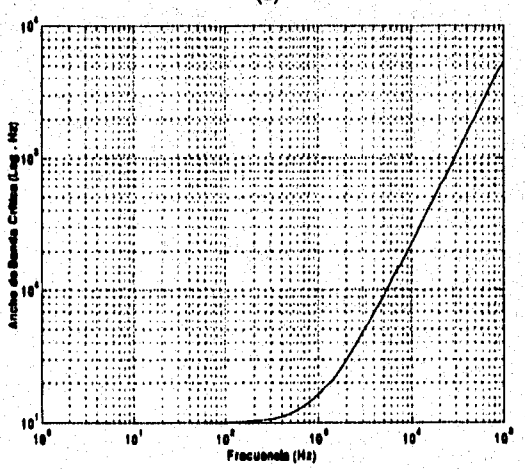
Estas transformaciones pueden ser utilizadas para calcular anchos de banda en una escala perceptual para filtros en una frecuencia determinada en las escalas Bark o mel. La función del ancho de la banda crítica se aprecia en la Figura 7.3(c).



(a)



(b)



(c)

Figura 7.4 (a) Escala Bark, (b) Escala Mel , (c) Bandas Criticas como función de la frecuencia (Fig. 9 [PICO93]).

Tanto las escalas Bark como mel pueden ser consideradas como una transformación de la escala de la frecuencia a una escala lineal de significado perceptible. La combinación de estas dos teorías dieron origen a una técnica de análisis conocida como banco de filtros de banda crítica. El banco de filtros de banda crítica es simplemente un banco de filtros pasabanda de fase lineal de respuestra al impulso finita FIR ordenados linealmente a lo largo de la escala Bark o mel. Los anchos de banda son seleccionados de modo que sean iguales al ancho de la banda crítica para la frecuencia central correspondiente.

Cada filtro en un banco de filtros digitales normalmente es implantado como un filtro de fase lineal de manera que el retraso de grupo para todos los filtros es igual a cero, y las señales de salida de los filtros se encontrarán sincronizadas en tiempo. Las ecuaciones del filtro para un filtro de fase lineal pueden ser generalizadas en la siguiente ecuación:

$$s_i(n) = \sum_{j=-(N_{FB_i}-1)/2}^{(N_{FB_i}-1)/2} a_{FB_i}(j)s(n+j) \qquad (7.11)$$

donde $a_{FB_i}(j)$ denota el j-ésimo coeficiente para el i-ésimo filtro de bandas críticas. El orden del filtro es normalmente impar para un filtro de fase lineal.

La salida de ciertos filtros puede ser correlacionada con cierta clase de sonidos de voz.

El banco de filtros digitales es el que se utiliza más frecuentemente en sistemas que pretenden emular el proceso auditivo

El resultado de este análisis es un vector de valores de potencia (o pares de potencia/frecuencia) para cada trama de datos. Estos son usualmente combinados con otros parámetros, tales como la potencia total, para formar un vector de mediciones de la señal. El banco de filtros pretende descomponer la señal en un conjunto discreto de muestras espectrales que contienen información similar a la presentada en niveles superiores de procesamiento en el sistema auditivo.

7.3.3.2 Transformada de Fourier de un Banco de Filtros

Una de las formas más fáciles y eficientes de procesar un modelo de banco de filtros espaciados no uniformemente, es simplemente realizar la transformada de Fourier de la señal, y muestrear la salida de la transformada a las frecuencias deseadas. La Transformada Discreta de Fourier (DFT) de una señal se define como

$$S(f) = \sum_{n=0}^{N_s-1} s(n)e^{-j(2\pi f/f_s)n} \qquad (7.12)$$

donde f denota la frecuencia en Hz, f_s denota la señal muestreada en frecuencia, y N_s denota la duración de la ventana en muestras.

El banco de filtros puede ser implantado utilizando la ecuación (7.12) para muestrear el espectro a las frecuencias enlistadas en la Tabla 3.1. De cualquier manera, el espectro es, en general, sobremuestreado a una mayor resolución que la descrita en la Tabla 3.1 y cada resultante del banco de filtros (magnitud espectral de potencia) es procesada como una suma ponderada de sus valores adyacentes.

$$S_{avg}(f) = \frac{1}{N_{os}} \sum_{n=0}^{N_{os}} \omega_{FB}(n) S(f + \delta f(f,n)) \qquad (7.13)$$

donde N_{OS} representa el número de muestras utilizadas para obtener el valor promedio, $\omega_{FB}(n)$ representa una función de ponderación, y $\delta f(f,n)$ representa algunas funciones que describen las frecuencias en la vecindad de f para ser utilizadas en el proceso del promedio. Es necesario notar que el método de promediación presentado es sólo un método particular de implementar una función de suavización espectral.

La Transformada Rápida de Fourier (FFT) también puede ser utilizada como un método alternativo de procesar el espectro de la señal. La FFT es una forma eficaz de calcular la DFT teniendo la condición de que el espectro debe ser evaluado en un conjunto discreto de frecuencias múltiplo de f_S/N. La principal ventaja de la FFT es su rapidez: se requieren aproximadamente $N \log N$ sumatorias así como $N \log N/2$ multiplicaciones (La DFT realiza N^2 operaciones). La principal desventaja es que los mapeos de frecuencias no lineales, tales como el banco de filtros en la Tabla 3.1, deben de ser ajustados para coincidir con las condiciones de ortogonalidad de la FFT.

Generalmente se tiene un proceso adicional, debido en gran parte, a nuestro deficiente conocimiento de la percepción humana, llegamos a la hipótesis de que las áreas de mayor amplitud del espectro, tienen una mayor ponderación en el sistema auditivo, que las regiones de baja amplitud. En un medio ambiente ruidoso, generalmente el ruido degrada desproporcionalmente nuestras estimaciones sobre las regiones del espectro de baja amplitud. Dicho de otra manera, estamos más seguros de la confiabilidad de las estimaciones sobre la áreas de mayor amplitud del espectro.

Por esta razón, generalmente marcamos un límite en el rango dinámico del espectro; este límite inferior lo conocemos como umbral del rango dinámico. En lugar de utilizar estimaciones ruidosas de las regiones de baja amplitud, simplemente descartamos las estimaciones por debajo de cierto umbral desde el pico del espectro. Para las técnicas basadas en la transformada de Fourier, esto es una implantación como una función de umbral sobre la magnitud espectral.

Es importante que la envolvente del espectro sea relativamente plana antes de implantar tales algoritmos de detección de umbral. De otra manera, ciertas porciones útiles de baja energía del espectro pueden ser erróneamente eliminadas. Recordemos que dado que el espectro de la señal de voz inherentemente decrece 20 dBs por década, un umbral basado en la energía de bajas frecuencias, donde la diferencia de la amplitud espectral pico-valle del espectro es grande, puede suprimir fácilmente señales de energía útil a frecuencias altas.

7.3.3.3 Coeficientes Cepstrales

Desde su introducción en los inicios de la década de los setentas, las técnicas homomórficas de procesamiento de señales han sido de gran interes en reconocimiento de voz. Los sistemas homomórficos son una clase de sistemas no lineales que obedecen un principio de superposición lineal generalizado. Los sistemas lineales, tales como los descritos con anterioridad, son un caso especial de sistemas homomórficos.

En procesamiento de voz, los sistemas homomórficos que buscamos, deben tener la siguiente propiedad:

$$D\left[[x_1(n)]^{\alpha} \cdot [x_2(n)]^{\beta}\right] = \alpha D[x_1(n)] + \beta D[x_2(n)] \qquad (7.14)$$

Esta es una operación de tipo superposición con respecto a la multiplicación, exponenciación y adición. Una función logarítmica obedece, sin lugar a dudas, la propiedad de superposición generalizada.

Los sistemas homomórficos fueron considerados de utilidad para el procesamiento de voz debido a que ofrecen una metodología para separar la señal de exitación de la forma del tracto vocal. Las aproximaciónes recientes de reconocimiento de voz están mayormente interesadas en modelar las caracteristicas del tracto vocal. En los modelos acústicos lineales de producción de voz, el espectro compuesto de voz, tal como es medido por la transformada de Fourier, consiste en la señal de excitación filtrada por un filtro lineal variante en el tiempo el cual representa la forma del tracto vocal.

El proceso de separar las dos componentes, generalmente conocido como deconvolución, puede ser descrito de la siguiente manera:

$$s(n) = g(n) \otimes v(n) \tag{7.15}$$

donde $g(n)$ denota la señal de excitación, $v(n)$ denota la respuesta del tracto vocal al impulso y "⊗" denota convolución. La representación en el dominio de la frecuencia de este proceso es

$$S(f) = G(f) \cdot V(f) \tag{7.16}$$

Si tomamos el logaritmo (complejo) de ambos lados, tenemos

$$\log(S(f)) = \log(G(f) \cdot V(f)) = \log(G(f)) + \log(V(f)) \tag{7.17}$$

Consecuentemente, en un domino logarítmico, la excitación y la forma del tracto vocal son superimpuestas, y pueden ser separadas utilizando un procesamiento de señales convencional.

Para procesar el cepstrum, en primera instancia se calculan las magnitudes espectrales logarítmicas. Posteriormente se calcula la transformada inversa de Fourier del espectro logarítmico

$$c(n) = \frac{1}{N_s} \sum_{k=0}^{N_s-1} \log_{10} \left| S_{avg}(k) \right| e^{(2\pi j/N_s)kn}; \quad 0 \le n \le N_s - 1 \tag{7.18}$$

$c(n)$ se define como el cepstrum. Nos referimos a los coeficientes cepstrales calculados por medio de la transformada de Fourier (o un banco de filtros analógicos) como los coeficientes cepstrales deivados de la transformada de Fourier.

Observamos que $c(0)$ en la ecuación (7.18) representa el valor promedio del espectro, o el valor rms de la señal. Inicialmente, este término era parte importante del vector de parámetros cepstrales. Posteriormente, se observó que las mediciones absolutas de potencia de la señal, eran en cierto modo poco confiables y el uso de $c(0)$ ya no fue importante.

La ecuación 18 también es reconocida como la DFT inversa del espectro logarítmico. Esto puede ser simplificado de manera conveniente haciendo notar que el espectro de la magnitud logaritmica es una función simétrica real. Por esta razón, puede ser simplificada a

$$c(n) = \frac{2}{N_s} \sum_{k=1}^{N_s} S_{avg}(I(k)) \cos\frac{2\pi}{N_s} kn \qquad (7.19)$$

$c(n)$ en la ecuación (7.19) normalmente es truncada a un orden mucho menor que N_S. $I(k)$ representa una función de mapeo que translada al entero k a las muestras correspondientes de S_{avg}. Por eficacia, S_{avg} también puede ser calculada utilizando una FFT sobremuestreada, en lugar de una DFT espaciada no uniformemente.

El cepstrum definido en la ecuación (7.19) puede ser modificado fácilmente para ser un cepstrum espaciado en escala mel muestreando la transformada de Fourier a frecuencias espaciadas apropiadamente.

7.3.3.4 Coeficientes de Predicción Lineal

Ahora pasaremos de los métodos de transformada de fourier basados en el análisis lineal espectral a una clase de técnicas de modelado paramétrico que pretende modelar óptimamente el espectro como un proceso autoregresivo. La técnica del modelo paramétrico basado en la teoría de error cuadrático mínimo, se conoce como predicción lineal (LP).

Proporcionando una señal $s(n)$, se busca modelar la señal como una combinación lineal de sus muestras anteriores. Definiendo el modelo de la señal como

$$s(n) = -\sum_{i=1}^{N_{LP}} a_{LP}(i) s(n-i) + e(n) \qquad (7.20)$$

donde N_{LP} representa el número de coeficientes en el modelo (el orden del predictor), $\{a_{LP}\}$ se definen como los coeficientes de predicción lineal (coeficientes de predicción), y $e(n)$ representa el error en el modelo (la diferencia entre el valor estimado y el valor de la medición actual).

Un atributo evidente de este modelo es que, si es exacto, deberíamos tener la capacidad de estimar los valores futuros de la señal, basandonos en un conjunto actual de mediciones. El término de error debería indicarnos algo referente a la calidad de nuestro modelo (si el error es pequeño, el modelo será exacto). De igual forma, es posible mostrar que un modelo de predicción lineal, claramente modela el espectro de la señal como un espectro suavizado.

La ecuación (7.20) puede ser reescrita en una notación de transformada Z, y ser mostrada como una operación de filtrado lineal.

$$E(z) = H_{LP}(z) S(z) \qquad (7.21)$$

donde $E(z)$ y $S(z)$ son la transformada Z de la señal de error y de la señal de voz, respectivamente, y

$$H_{LP}(z) = 1 + \sum_{i=1}^{N_{LP}} a_{LP}(i) z^{-i}$$

$$H_{LP}(z) = \sum_{i=0}^{N_{LP}} a_{LP}(i) z^{-i} \qquad (7.22)$$

donde $a_{LP}(0) = 1$. $H_{LP}(z)$ se define como el filtro de predicción lineal inverso.

Teniendo la condición de que se desea un error cuadrático medio tan pequeño como sea posible (es razonable la búsqueda de una solución que nos proporcione el menor error en la energía), los coeficientes (sin incluir $a_{LP}(0)$) en la ecuación (7.22) pueden obtenerse de la siguiente ecuación matricial:

$$\bar{a}_{LP} = \Phi^{-1}\bar{\phi} \tag{7.23}$$

donde

$$\bar{a}_{LP} = \left[a_{LP}(1),\ldots,a_{LP}(N_{LP})\right] \tag{7.24}$$

$$\Phi = \begin{bmatrix} \phi_n(1,1) & \phi_n(1,2) & \cdots & \phi_n(1,N_{LP}) \\ \phi_n(2,1) & \phi_n(2,2) & \cdots & \phi_n(1,N_{LP}) \\ \cdots & \cdots & \cdots & \cdots \\ \phi_n(N_{LP},1) & \phi_n(N_{LP},2) & \cdots & \phi_n(N_{LP},N_{LP}) \end{bmatrix} \tag{7.25}$$

$$\bar{\phi} = \left[\phi_n(1,0),\phi_n(2,0),\ldots,\phi_n(N_{LP},0)\right] \tag{7.26}$$

y

$$\phi_n(j,k) = \frac{1}{N_s}\sum_{m=0}^{N_s-1} s(n+m-j)s(n+m-k) \tag{7.27}$$

La solución presentada en las ecuaciones (7.23) a (7.27) es conocida como el Método de la Covariancia. Φ es la matriz de covariancia y $\phi_n(j,k)$ es la función de covariancia para $s(n)$.

Existen tres formas de calcular los coeficientes de predicción: métodos de covariancia basados en la matriz de covariancia (también conocidos como métodos de mínimos cuadrados puros), métodos de autocorrelación, y métodos enrejado (o armónicos). En reconocimiento de voz, se utiliza casi exclusivamente el método de autocorrelación debido a su eficacia computacional así como estabilidad inherente. El método de autocorrelación siempre produce un filtro de predicción cuyos ceros se encuentran dentro del círculo unitario en el plano-z.

En el método de autocorrelación, modificamos la ecuación (7.27) de la siguiente manera:

$$\phi_n(j,k) = \phi_n(0,|j-k|) \tag{7.28}$$

o

$$R_n(k) = \frac{1}{N_s}\sum_{m=0}^{N_s-1-k} s(n+m)s(n+m-k) \tag{7.29}$$

$R_n(k)$ se conoce como la función de autocorrelación. Esta simplificación resulta al limitar el intervalo de evaluación al rango de $[0, N\text{-}1]$, asumiendo que los valores fuera del rango, son cero.

Debido a esta condición de longitud finita, es importante aplicar una ventana en el método de autocorrelación, tal como la que se describe en la ecuación (7.5) a la señal. Habitualmente, se

utiliza una ventana de Hamming. La utilización de una ventana suprime los problemas originados por las variaciones bruscas de la señal en los límites de la ventana. Para un análisis con traslape, se garantiza una transición suave de trama a trama de los parámetros estimados.

Esta simplificación permite que los coeficientes de predicción sean calculados eficazmente utilizando la recursión de Levinson-Durbin.

Inicialización:

$$E_{LP}^{(0)} = R_n(0) \tag{7.30}$$

$$\textit{For} \quad 1 \le i \le N_{LP}$$

$$\{$$

$$k_{LP}(i-1) = -\frac{R_n(i) + \sum_{j=1}^{i-1} a_{LP}^{(i-1)}(j) R_n(i-j)}{E_{LP}^{(i-1)}} \tag{7.31}$$

$$a_{LP}^{(i)} = k_{LP}(i-1) \tag{7.32}$$

$$\textit{For} \quad 1 \le j \le i-1$$

$$\{$$

$$a_{LP}^{(i)}(j) = a_{LP}^{(i-1)}(j) + k_{LP}(i-1) a_{LP}^{(i-1)}(i-j) \tag{7.33}$$

$$\}$$

$$E_{LP}^{(i)} = \left(1 - k_{LP}(i-1)^2\right) E_{LP}^{(i-1)} \tag{7.34}$$

$$\}$$

Estas ecuaciones calculan los coeficientes de predicción con una complejidad proporcional a N_{LP}^2 y permiten realizar el cálculo completo LP con una complejidad de aproximadamente $N_s N_{LP} + 3N_s + N_{LP}^2$.

De hecho, el modelo de la señal es la inversa de $H_{LP}(z)$, y está dado por:

$$S_{LP}(z) = \frac{G_{LP}}{H_{LP}(z)} \tag{7.35}$$

G_{LP} es el modelo de la ganancia, y está dado por:

$$G_{LP} = \sqrt{E_{LP}^{(N_{LP})}} \tag{7.36}$$

El término de la ganancia también está dado por la expresión:

$$G_{LP} = \sqrt{R_n(0) \prod_{i=1}^{N_{LP}} \left(1 - k_{LP}(i-1)^2\right)} \tag{7.37}$$

El término de ganancia permite que el espectro del modelo LP coincida con el espectro de la señal de voz original. El modelo LP calculado de la ecuación (7.23) es un modelo normalizado (los valores de los coeficientes de predicción son independientes de la potencia de la señal).

Es necesario recalcar tres observaciones sobre este tipo de solución LP. En primer lugar las variables intermedias utilizadas en los cálculos $\{k_{LP}\}$ son llamados coeficientes de reflexión. Están limitados por:

$$0 \le |k_{LP}(i-1)| \le 1 \quad \forall 1 \le i \le N_{LP} \tag{7.38}$$

Este es un resultado sumamente útil para el almacenamiento y compresión de las aplicaciones que involucran modelos LP.

En segundo lugar, la solución iterativa calcula la solución para modelos de los órdenes $1 \le i \le N_{LP}$. Esto es adecuado para aplicaciones de procesamiento de señales que requieren una estimación del orden del modelo como parte del desarrollo. Generalmente, en las aplicaciones de reconocimiento de voz, el orden del modelo es un sistema de parámetros fijos.

En tercer lugar, a medida que el orden se incrementa, el modelo de la solución LP se ajusta mejor. La ecuación 7.34 representa la energía del error. De esta ecuación podemos apreciar que el error es monotónicamente decreciente a medida de que el orden aumenta. El modelo en si mismo intenta coincidir con el espectro de la mejor forma posible para el orden dado.

Es necesario hacer notar que a medida que el orden se incrementa, el modelo genera una mayor similitud con el espectro original. Con un orden menor solamente se puede capturar una forma burda del espectro. Con un orden mayor se obtiene la representación del espectro más detallado.

Como ya se ha explicado, el modelo espectral en las áreas de baja energía del espectro de la señal es frecuentemente inexacto. De cierta manera sería conveniente asignar un umbral de rango dinámico. Existen varios métodos para realizar esto en un modelo LP: un método de covariancia estabilizada que reduzca el rango dinámico en el espectro, un método de ponderación perceptual que amplía ligeramente el ancho de banda de los modelos LP, o un método de autocorrelación estabilizada en la cual una pequeña cantidad de ruido es agregada a la función de autocorrelación.

La última de estas aproximaciones es sencilla y efectiva. La función de autocorrelación de la ecuación 7.29 es modificada antes de realizar los cálculos LP de la siguiente manera:

$$R_{n\omega}(0) = (1+\gamma_{n\omega})R_n(0)$$

$$R_{n\omega}(0) = R_n(i), \quad 1 > 0 \tag{7.39}$$

El umbral de rango dinámico normalmente se especifica en decibeles.

$$\gamma_{n\omega_{dB}} = 10\log_{10}\gamma_{n\omega} \tag{7.40}$$

Un valor típico del umbral de rango dinámico es -10db.

Este proceso de estabilización es equivalente a añadir ruido blanco no correlacionado a la señal de voz antes de realizar el análisis LP. El efecto de este ruido es prevenir al modelo LP de modelar ceros en el espectro.

Es importante realizar una observación sobre los modelos LP. Añadiendo el factor de potencia y el valor de la frecuencia fundamental o tono a los coeficientes LP, es posible reconstruir una versión audible de la señal de voz. Es muy útil escuchar la versión paramétrica de la señal de voz, particularmente en los modelos de reconocimiento de voz, para el diagnostico de problemas. Algunas transformaciones paramétricas, tales como los coeficientes cepstrales, no poseen un mapeo uno a uno con la información original LP. En base a lo anterior, tienen mayor dificultad el valorar la validez del conjunto de parámetros.

Los primeros sistemas de reconocimiento de voz, utilizaban directamente parámetros LP en procesos de reconocimiento. Desde entonces, se han desarrollado transformaciones más sofisticadas de estos parámetros. De cualquier forma es importante recordar que el generar un modelo LP exacto es el primer paso importante en el análisis espectral. Dado que el análisis LP es una operación nolineal, su desempeño en un medio ambiente ruidoso puede llegar a causar problemas. Por esta razón, algunos sistemas aún utilizan análisis de banco de filtros basado en la transformada de Fourier.

En un sistema de reconocimiento de voz, normalmente se utilizan tramas con duración de 20 ms. Sin embargo, a medida que la investigación en reconocimiento de voz se ha dirigido hacia reconocimiento fonético, se han usado tramas con duración de 10 ms. El movimiento en dirección a una resolución de tiempo más fina continuará a medida que la tecnología de reconocimiento fonético madure.

7.3.3.5 Amplitudes del Banco de Filtros Derivados de LP

Se definen las amplitudes del banco de filtros como el resultado de muestrear el modelo espectral LP (en lugar del espectro de la señal) a frecuencias apropiadas del banco de filtros.

Frecuentemente, la suavización espectral inherente en el modelo LP proporciona parámetros más estables a las etapas subsecuentes del procesamiento. De cualquier forma, a medida que las técnicas de reconocimiento de voz y de procesamiento digital de señales han progresado, las diferencias en ambas aproximaciones no son tan grandes como lo fueron en alguna ocasión.

Algunas técnicas precisas para el cálculo de las amplitudes del banco de filtros involucran la evaluación directa del modelo LP.

$$S_{LP}(f) = \frac{G_{LP}}{\sum_{i=0}^{N_{LP}} a_{LP}(i) e^{-j2\pi(f/f_s)}} \tag{7.41}$$

donde f_s representa la frecuencia de muestreo. Este método requiere del orden de $4p+3$ operaciones de multiplicación/acumulación por muestra frecuencial.

Otra aproximación muy utilizada es la de calcular la potencia espectral de la autocorrelación de la respuesta al impulso de $H_{LP}(z)$. La respuesta al impulso $H_{LP}(z)$ puede calcularse directamente de los coeficientes LP.

$$R_{LP}(k) = \sum_{m=0}^{N_{LP}-|k|} a_{LP}(m) a_{LP}(m+|k|), \quad |k| \le N_{LP}$$

$$R_{LP}(k) = 0 \qquad\qquad |k| > N_{LP} \tag{7.42}$$

La densidad espectral de potencia puede ser calculada de manera eficaz a partir de la función de autocorrelación, obsevando que la función de autocorrelación es una función real par. En consecuencia, su transformada de Fourier es real, y esta dada por:

$$S_{LP}(f) = R_{LP}(k) + 2\sum_{k=0}^{N_{LP}} R_{LP}(k)\cos\left(2\pi\frac{f}{f_s}k\right) \tag{7.43}$$

Con cualquiera de los métodos, se puede implantar un espectro espaciado nolinealmente, mediante la selección adecuada de las frecuencias de muestreo de los bancos de filtros. También, aun cuando el modelo LPC proporcione un espectro suavizado que se ajuste, generalmente sigue siendo una ventaja sobremuestrear el espectro de modo que los picos pronunciados en la respuesta en frecuencia sean representados de manera exacta por el banco de filtros (el cual tiende a cuantizar en forma gruesa el espectro).

7.3.3.6 Coeficientes Cepstrales Derivados LP

Otro paso lógico en dirección del cálculo de las amplitudes del banco de filtros derivados LP, podría ser la utilización de modelo LP para calcular los coeficientes cepstrales.

Si el filtro de predicción lineal es estable (y garantiza estabilidad en el análisis de autocorrelación), el logaritmo del filtro inverso puede ser expresado como una serie de potencias:

$$\begin{aligned} C_{LP}(z) &= \sum_{i=0}^{N_c} c_{LP}(i)z^{-i} \\ &= \log H(z) \\ &= \log\left(\frac{G_{LP}}{\sum_{j=0}^{N_{LP}} a_{LP}(j)z^{-i}}\right) \end{aligned} \tag{7.44}$$

Es posible encontrar la solución de los coeficientes mediante la diferenciación de ambos lados de la igualdad con respecto a z^{-1}, igualando los coeficientes de los polinomios resultantes. Lo anterior origina la siguiente recursión:

Inicialización:

$$c_{LP}(1) = -a_{LP}(1) \tag{7.45}$$

$$\begin{aligned} &For \quad 2 \le i \le N_c \\ &\{ \\ &\qquad c_{LP}(i) = -a_{LP}(i) - \sum_{j=1}^{i-1}\left(1-\frac{i}{i}\right)a_{LP}(j)c_{LP}(i-j) \qquad (7.46) \\ &\} \end{aligned}$$

Los coeficientes $\{c_{LP}\}$ también se conocen como coeficientes cepstrales derivados LP.

ESTA TESIS NO DEBE
SALIR DE LA BIBLIOTECA

Históricamente, $c_{LP}(0)$ ha sido definido como el logaritmo de la potencia del error LP. Por ahora, podemos notar que dado que la potencia puede ser tratada como un parámetro separado, no se tiene necesidad de incluirla en las ecuaciones anteriores.

Se tiene una complicación menor en la recursión de coeficientes cepstrales. No se especifica el número de coeficientes cepstrales N_C por calcular. Dado que los coeficientes cepstrales son la transformada inversa del filtro de la respuesta al impulso del modelo LP, podemos, en teoría, calcular un número infinito de coeficientes cepstrales. De cualquier forma, el número de coeficientes cepstrales calculados puede compararse con el número de coeficientes LP: $0.75p \leq N_C \leq 1.25p$.

7.4 Segmentación Acústica

El reconocimiento de voz basado en el método acústico tiene ventajas sobre los métodos lingüísticos y matemáticos.

El método lingüístico se basa en la técnica de segmentación manual de los fonemas que componen la palabra. Esto se hace mediante la identificación visual y auditiva de los diferentes segmentos, mediante el uso de herramientas que permitan visualizar el espectrograma de la palabra así como poder escuchar el segmento previamente etiquetado como se muestra en la Figura 5.1 en la cual se muestra un espectrograma y la identificación manual de los diferentes fonemas, utilizando la técnica antes mencionada.

El método matemático divide a la palabra en un número N de segmentos denominados tramas, que se consideran estacionarios y ergódicos. Esta aseveración supone que las tramas son tan cortas que contienen información acústica consistente, pero esto no siempre es cierto, ya que se puede tener una trama que ocupe la transición entre dos fonemas o bien sea tan grande que en realidad contenga dos fonemas y esto cancelaría la condición de estacionariedad que implica que las características estadísticas de la señal no varían dentro de la trama.

En realidad existe un compromiso entre la frecuencia de muestreo y el tamaño de las tramas. El tamaño de las tramas se puede ajustar para que cumplan las condiciones de estacionariedad y ergodicidad requeridas para la mayoría de las tramas. Ahora, el método matemático espera que la señal sea un proceso estocástico y en el caso de las fonemas sordos, se cumple pero para el caso de los sonoros, como se explicó en el Capitulo 3, son señales totalmente periódicas. Esta periodicidad se explica por la excitación glotal y la acústica del tracto vocal [PARS87].

Para el caso de la segmentación acústica [HERR194][HERR294] se hace una combinación de ambas técnicas donde se buscan cambios estadísticos de la misma, por lo que en realidad se hace una segmentación de las palabras en unidades acústicas, que podría equivaler a la segmentación en fonemas o regiones cuasiestacionarias de la señal.

7.4.1 Segmentación en Subpalabras Acústicas

El algoritmo utilizado, realiza una detección de comienzo y fin de palabra. Se utiliza la metodología de clasificación de la razón de máxima similitud (MLR) para detectar cambios espectrales, utilizando una prueba de MLR con una ventana deslizante. El algoritmo de detección de comienzo y fin de palabra se trata en el siguiente inciso.

Las muestras entran a un segundo nivel pruebas iterativas de ventana deslizante donde la señal de voz es segmentada en subpalabras acústicas. Se aplican pruebas secuenciales de MLR a segmentos

de cinco tramas. Los segmentos resultantes cuasiestacionarios son separados por transiciones espectrales abruptas.

El algoritmo segmenta la palabra en segmentos de longitud variable donde cada segmento representa una región de características espectrales cuasiestacionarias. Por simplicidad se asume que $\underline{C}(n)$ (vector de bandas críticas) para la trama n, es una variable aleatoria independiente con distribución normal con media cero y de dimensión J; $\underline{C}(n) \sim N(\underline{0}, \Sigma)$.

Se deriva la formulación teórica de la prueba de MLR generalizada para la hipótesis:

$$\begin{aligned} H_0 &: \Sigma = \Sigma_0 \\ H_1 &: \Sigma \neq \Sigma_0 \end{aligned} \tag{7.47}$$

basados en un segmento muestral $[\underline{C}(1), \underline{C}(2), \ldots, \underline{C}(N)]$ con los componentes de $\Sigma, (\sigma_1^2, \sigma_2^2, \ldots, \sigma_J^2)$ desconocidos y los componentes de Σ_0 conocidos.

Cuando no se restringe a Σ, el máximo ocurre para su estimador de máxima similitud $\hat{\Sigma}$ en el cual

$$\hat{\sigma}_j^2 = \frac{1}{N} \sum_{n=1}^{N} C_{jn}^2 \tag{7.48}$$

Simplificando la expresión de la prueba de MLR, obtenemos

$$\Lambda = \ln \lambda = \frac{N}{2} \left| \sum_{j=1}^{J} \ln \frac{\hat{\sigma}_j^2}{\hat{\sigma}_{0j}^2} - \sum_{j=1}^{J} \frac{\hat{\sigma}_j^2}{\hat{\sigma}_{0j}^2} \right| \begin{matrix} H_0 \\ < \\ > \\ H_1 \end{matrix} THRESH \tag{7.49}$$

donde en el paso k

$$\Lambda(k) = \frac{N}{2} \left| \sum_{j=1}^{J} \left[\ln \frac{\hat{\sigma}_j^2(k)}{\hat{\sigma}_{0j}^2} - \sum_{j=1}^{J} \frac{\hat{\sigma}_j^2(k)}{\hat{\sigma}_{0j}^2} \right] \right| \tag{7.50}$$

en el que $\hat{\sigma}_j^2(k)$ representa la variancia muestral de la j-ésima banda de frecuencia al paso k, σ_{0j}^2 es la j-ésima variancia de Σ_0 y THRESH es un umbral que es determinado de forma experimental.

En el segundo nivel, las pruebas de MLR se aplican a segmentos de cinco tramas

$$\mathbf{Y}(n) = [\mathbf{X}(n-1), \mathbf{X}(n), \mathbf{X}(n+1)] \tag{7.51}$$

donde

$$\begin{aligned} \mathbf{X}(n-1) &= [\underline{C}(n-2), \underline{C}(n-1)] \\ \mathbf{X}(n) &= [\underline{C}(n)] \\ \mathbf{X}(n+1) &= [\underline{C}(n+1), \underline{C}(n+2)] \end{aligned} \tag{7.52}$$

y en (7.50), ahora $\sigma_{0j}^2(n)$ representa la variancia muestral de $\mathbf{X}(n-1)$ y $\sigma_{0j}^2(n)$ es la variancia muestral de $\mathbf{X}(n+1)$.

7.4.2 Detección de Voz

En la primera iteración, la señal de entrada se separa en segmentos de voz, no-voz. La prueba de MLR se aplica secuencialmente a segmentos de tres tramas como en (7.51). La prueba de máxima similitud esta dada por (7.47) donde en (7.50) $\sigma_{0j}^2(n)$ representa la variancia muestral de ruido de la j-ésima banda de frecuencia y se obtiene de los primero 60ms de la señal de entrada, que se asume como un segmento de silencio. Se utiliza un umbral de 0.25 (THRESH) en este trabajo.

8. Cuantización Vectorial[1]

8.1 Formulación del problema

Asumimos que $\mathbf{x} = [x_1, x_2, \ldots, x_N]^T$ es un vector N-dimensional cuyos componentes $\{x_k, 1 \le k \le N\}$ son variables aleatorias reales y de amplitud continua. En cuantización vectorial, el vector $\mathbf{x}$ es mapeado a otro vector $\mathbf{y}$ también real y de amplitud continua. Se dice que $\mathbf{x}$ está cuantizado como $\mathbf{y}$, donde $\mathbf{y}$ es el valor cuantizado de $\mathbf{x}$, por lo tanto podemos escribir,

$$\mathbf{y} = q(\mathbf{x}) \tag{8.1}$$

donde q() es el operador de cuantización. A $\mathbf{y}$ también se le denomina vector de reconstrucción o vector de salida que corresponde a $\mathbf{x}$. Típicamente, $\mathbf{y}$ toma un conjunto finito de valores $\mathbf{Y} = \{\mathbf{y}_i, 1 \le i \le L\}$, donde $\mathbf{y}_i = [\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{iN}]^T$. El conjunto $\mathbf{Y}$ se le conoce como el diccionario de reconstrucción o simplemente el diccionario, L es el tamaño del diccionario y $\{\mathbf{y}_i\}$ es el conjunto de vectores de código. Los vectores $\mathbf{y}_i$ también son conocidos en la literatura de reconocimiento de patrones como los patrones de referencia o plantillas. El tamaño L del diccionario también se conoce como número de niveles, término utilizado en la terminología de la cuantización escalar. Entonces se puede hablar acerca de un diccionario de L niveles. Para el diseño de éste, se particiona el espacio N-dimensional del vector aleatorio $\mathbf{x}$ en L regiones o celdas $\{c_i, 1 \le i \le L\}$ y se asocia c_i a un vector $\mathbf{y}_i$. El cuantizador entonces asigna el vector de código $\mathbf{y}_i$ si $\mathbf{x}$ está en c_i,

$$q(\mathbf{x}) = \mathbf{y}_i \quad si \quad \mathbf{x} \in c_i \tag{8.2}$$

Al proceso de diseño del diccionario también se le conoce como entrenamiento o población del diccionario. La Figura 8.1 muestra un ejemplo de particionamiento en un espacio bidimensional (N=2) como demostración de la cuantización vectorial. La región limitada por las líneas oscuras es la celda c_i. Cualquier vector $\mathbf{x}$ que cae en ésta celda es cuantizado como $\mathbf{y}_i$. Las posiciones de otros vectores de código que corresponden a otras celdas se muestran como puntos. El número total de vectores de código en éste ejemplo es L = 18.

Para N = 1, la cuantización vectorial se reduce a cuantización escalar, la Figura 8.2 muestra un ejemplo de particionar la línea real para cuantización escalar. Los valores de código (salidas o niveles de reconstrucción) se muestran como puntos. Aquí también cualquier valor de entrada $\mathbf{x}$ que cae en el intervalo c_i es cuantizado como $\mathbf{y}_i$. El número de niveles de la Figura 8.2 es de 10. La cuantización escalar tiene la propiedad especial que mientras las celdas pueden variar en tamaño, siempre tienen la misma forma, se puede decir que todas son intervalos sobre la línea real. En comparación se debe notar como en la Figura 8.1 las celdas en dos dimensiones tienen diferentes formas.

---

[1] En su mayor parte tomado de [MAKH85].

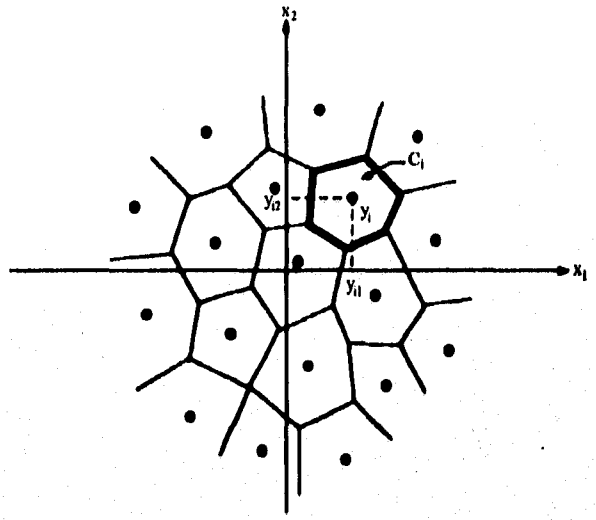


Figura 8.1 Partición de un espacio bidimensional (N=2) en L = 18 grupos (Fig 3 (MAKH85]).

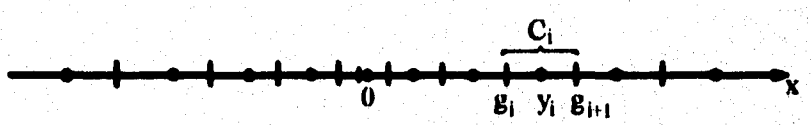


Figura 8.2 Partición de la línea real en L=10 grupos o intervalos para cuantización escalar (N=1) (Fig 4 (MAKH85]).

La libertad de tener diferentes formas de celdas en un espacio multidimensional da a la cuantización vectorial una ventaja sobre la cuantización escalar, que se expone a continuación.

Cuando $\mathbf{x}$ es cuantizada como $\mathbf{y}$, se produce un error y se puede definir una medida de distorsión $d(\mathbf{x},\mathbf{y})$ entre $\mathbf{x}$ y $\mathbf{y}$. $d(\mathbf{x},\mathbf{y})$ también es conocida como la medida de disimilitud o medida de distancia. Conforme los vectores $\mathbf{y}(n)$ se transmiten a n tiempos diferentes, se puede definir una distorsión promedio total,

$$D \underset{M\to\infty}{=} \frac{1}{M}\sum_{n=1}^{M} d[\mathbf{x}(n),\mathbf{y}(n)] \tag{8.3}$$

Si el vector proceso $\mathbf{x}(n)$ es estacionario y ergódico, el promedio muestral en (8.3) tiende a n en el límite al valor esperado.

$$\begin{aligned} D &= \mathcal{E}[d(\mathbf{x},\mathbf{y})] \\ &= \sum_{i=1}^{L} P(\mathbf{x}\in c_i)\mathcal{E}[d(\mathbf{x},\mathbf{y}_i)|\mathbf{x}\in c_i] \\ &= \sum_{i=1}^{L} P(\mathbf{x}\in c_i)\int_{\mathbf{x}\in c_i} d(\mathbf{x},\mathbf{y}_i)p(\mathbf{x})d\mathbf{x} \end{aligned} \tag{8.4}$$

donde $P(\mathbf{x} \in c_i)$ es la probabilidad discreta que $\mathbf{x}$ esté dentro de c_i, $p(\mathbf{x})$ es una función de densidad de probabilidad (PDF) multidimensional de $\mathbf{x}$ y la integral se toma sobre todos los componentes del vector $\mathbf{x}$.

Para propósitos de transmisión, cada valor de $\mathbf{y}_i$, es codificado en palabra de código $\mathbf{c}_i$ de longitud B_i bits. En general, las diferentes palabras de código tendrán diferente longitud.
La tasa de transmisión T está dada por

$$T = BFc \quad bits/s$$

donde

$$B = \lim_{M \to \infty} \frac{1}{M} \sum_{n=1}^{M} B(n) \quad bits/vector \tag{8.5}$$

es la longitud de la palabra de código promedio, $B(n)$ es el número de bits usados para codificar el vector $\mathbf{x}(n)$ en el tiempo n, y F_C es el número de palabras de código transmitidas por segundo. También es útil definir el número promedio de bits por parámetro o por dimensión

$$R = \frac{B}{N} \quad bits/dimensión$$

R tasa de bits por dimensión
B tasa de bits por vector
T tasa de bits por segundo

para un diccionario de tamaño L, el número máximo de bits necesarios para codificar un vector es

$$B_{max} = \log_2 L$$

Cuando se diseña un sistema de compresión, se intenta diseñar el cuantizador de tal forma que la distorsión a la salida sea minimizada para una tasa de transmisión dada. Una decisión muy importante en el diseño de un cuantizador es, qué medida de distorsión se utilizará.

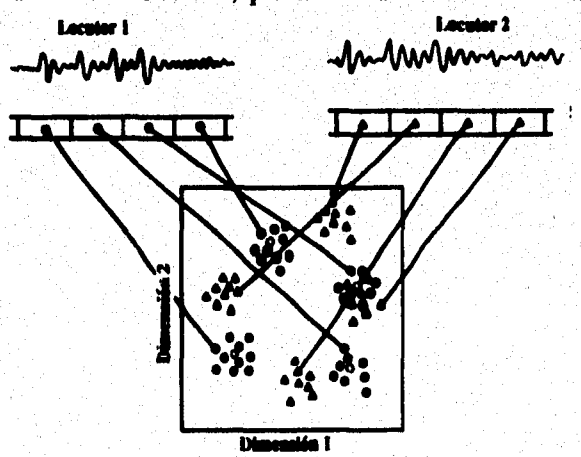


Figura 8.3 Diagrama conceptual que ilustra la cuantización vectorial en un diccionario.

8.2 Medidas de Distorsión

Para que sea de utilidad, una medida de distorsión o distancia debe ser perceptible, de tal forma que pueda ser analizada y calculada; debe ser relevante en forma subjetiva, de modo que las diferencias en los valores de la distorsión puedan ser utilizados para indicar diferencias de similitud en la calidad de voz. La mayoría de las medidas de distorsión en uso actual son perceptuales y en cierta medida relevantes subjetivamente. Se ha descubierto que si la distorsión disminuye en pocos decibeles, en algunas situaciones será perceptible en otras no. Mientras que las medidas de distorsión objetivas son herramientas necesarias y útiles en el diseño de sistemas de codificación, se requiere hacer *n* pruebas de calidad subjetiva para mejorar el desempeño del sistema.

Una medida de distancia debe obedecer a las siguientes propiedades:

1) No Negatividad

$$D(\bar{x}_1,\bar{x}_2) > 0 \quad x_1 \neq x_2$$
$$D(\bar{x}_1,\bar{x}_2) = 0 \quad x_1 = x_2$$

2) Simetría

$$D(\bar{x}_1,\bar{x}_2) = D(\bar{x}_2,\bar{x}_1)$$

3) Desigualdad del triángulo

$$D(\bar{x}_1,\bar{x}_3) \leq D(\bar{x}_1,\bar{x}_2) + D(\bar{x}_2,\bar{x}_3)$$

La Distancia Euclidiana es la distancia más utilizada que cumple éstas relaciones.

A continuación se enumeran algunas de las medidas de distorsión más utilizadas.

8.2.1. Error Cuadrático Medio (MSE).

Es por mucho la medida de distorsión más utilizada,

$$d_2(x,y) = \frac{1}{N}(x-y)^T(x-y) = \frac{1}{N}\sum_{k=1}^{N}(x_k - y_k)^2 \qquad (8.6)$$

donde la distorsión está definida por dimensión. La popularidad de MSE se basa en sus simplicidad y su seguimiento matemático. Se puede definir una distorsión más general basada en la norma L_r como

$$d_r(x,y) = \frac{1}{N}\sum_{k=1}^{N}|x_k - y_k|^r \qquad (8.7)$$

Se debe notar que (8.7) es igual a (8.6) para $r = 2$. Otros valores muy usados son $r = 1$ y $r = \infty$.

d_1 representa el error promedio absoluto y d_∞ tiende al error máximo. De hecho se puede demostrar que

$$\lim_{r\to\infty}\left[d_r(x,y)\right]^{1/r} = \max\left\{|x_k - y_k|, 1 \le k \le N\right\} \tag{8.8}$$

Minimizar D para $r = \infty$ sería equivalente a minimizar el error de cuantización máximo. Para codificación de voz, la distorsión más usada ha sido d_2 y d_1 y d_∞ .

8.2.2. Error Cuadrático Medio Ponderado

En el MSE d_2 se asume que las distorsiones contribuyen cuantizando los diferentes parámetros $\{x_k\}$ de igual forma. De manera general, se pueden introducir pesos diferentes con el fin de aportar ciertas contribuciones a la distorsión dependiendo del parámetro.

Un MSE ponderado general se define como,

$$d_w(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T \mathbf{W}(\mathbf{x}-\mathbf{y}) \tag{8.9}$$

donde $\mathbf{W}$ es una matriz positiva definida de ponderación. $\mathbf{W} = \mathbf{N}^{-1}\mathbf{I}$, donde $\mathbf{I}$ es la matriz identidad, esto resulta $d_w = d_2$.

Una selección para $\mathbf{W}$ que es muy popular en muchas aplicaciones de clasificación de patrones es $\mathbf{W} = \Gamma^{-1}$, donde Γ es la matriz de covariancia del vector aleatorio $\mathbf{x}$.

$$\Gamma = E\left[(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}})^T\right] \ , \ \bar{\mathbf{x}} = \varepsilon[\mathbf{x}] \tag{8.10}$$

En este caso d_w se reduce a

$$d_w(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T \Gamma^{-1}(\mathbf{x}-\mathbf{y}) \tag{8.11}$$

Esta se conoce como la distancia de Mahalanobis.

Además de ser positiva definida, la matriz de ponderación es simétrica. Se puede factorizar $\mathbf{W}$ como:

$$\mathbf{W} = p^T p \tag{8.12}$$

los vectores $\mathbf{x}$ y $\mathbf{y}$ se pueden transformar en un nuevo conjunto de vectores $\tilde{\mathbf{x}}$ y $\tilde{\mathbf{y}}$

$$\tilde{\mathbf{x}} = \mathbf{P}\mathbf{x} \quad \tilde{\mathbf{y}} = \mathbf{P}\mathbf{y} \tag{8.13}$$

y

$$\begin{aligned} d_w(\mathbf{x},\mathbf{y}) &= (\mathbf{P}\mathbf{x}-\mathbf{P}\mathbf{y})^T(\mathbf{P}\mathbf{x}-\mathbf{P}\mathbf{y}) \\ &= (\tilde{\mathbf{x}}-\tilde{\mathbf{y}})^T(\tilde{\mathbf{x}}-\tilde{\mathbf{y}}) \\ &= d_2(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) \end{aligned} \tag{8.14}$$

Entonces la MSE ponderada entre vectores originales es igual al MSE entre los vectores transformados. Entonces, para propósitos de cálculo, puede ser ventajoso realizar la transformación en (8.13) en todos los datos antes de realizar la cuantización vectorial.

8.2.3. Medidas de Distorsión de Predicción Lineal

En el análisis LPC, los coeficientes de predicción $\{a(k)\}$ se obtiene como resultado de la minimización de energía del residuo de la predicción. Se puede demostrar que la solución para valores de $A(z)$ óptimos es única y se calcula en función del conjunto de ecuaciones lineales simultáneas

$$\sum_{k=1}^{N} a(k)\phi(i-k) = -\phi(i) \quad 1 \le i \le N \tag{8.15}$$

donde $\{\phi(i), 0 \le i \le N\}$ se refiere a los coeficientes de autocorrelación de tiempo corto de la señal de voz sobre una trama. La ganancia G del filtro H(z) se calcula de forma que cuando es exitado por una fuente de variancia unitaria la energía de salida será igual a $\phi(0)$.
Esto se puede obtener por

$$G^2 = \phi(0) + \sum_{k=1}^{N} a(k)\phi(k) \tag{8.16}$$

que es igual a la energía residual mínima. Los parámetros de filtro H(z) se pueden calcular para cada trama, se cuantizan y se transmiten.

La ganancia G normalmente se cuantiza en una escala logarítmica y se transmite por separado. Ya que la cuantización de los coeficientes de predicción pueden generar inestabilidad del filtro polar resultante, normalmente se transforman a otro conjunto de parámetros conocidos como los coeficientes de reflexión $\{k_k, 1 \le k \le N\}$ o coeficientes de correlación parcial (PARCOR). Los coeficientes de reflexión resultan como el producto de resolver (8.15) o se pueden calcular recursivamente de los coeficientes de predicción. Para una H(z) estable, los coeficientes de reflexión tienen la propiedad de

$$|k_k| < 1 \quad 1 \le k \le N$$

Ya que los valores de $|k_k|$ cercanos a 1, los polos se acercan al círculo unitario y cambios pequeños de k_k pueden resultar en cambios significativos en el espectro. Entonces, para propósitos de cuantización, los coeficientes de reflexión son normalmente transformados a otro conjunto de coeficientes que exhiben menor sensibilidad conforme k se acerca a 1. Dos transformaciones usuales son,

$$\begin{aligned} S_k &= \frac{2}{\pi}\sin^{-1} k_k \qquad 1 \le k \le N \\ G_k &= \frac{1}{2}\log\frac{1+k_k}{1-k_k} = \tanh^{-1} k_k \quad 1 \le k \le N \end{aligned} \tag{8.17}$$

Los parámetros G_k son conocidos como las razones logarítmicas del área (LARs) de la analogía del tubo acústico del tracto vocal, y poseen la propiedad que pequeños cambios en G_k son aproximadamente proporcionales en el espectro logarítmico de H(z). El MSE d_2 y el error minmax d_∞ han sido utilizados para cuantizar G_k y S_k.

Una medida de distorsión alternativa utilizada en la cuantización de los coeficientes de predicción fue propuesta por Itakura y Saito; se deriva de los principios de máxima similitud.

Una de las primeras medidas de distancia introducidas al reconocimiento de voz fue una medida basada en el error de predicción mínimo y principios de apareamiento espectral. A ésta se le conoce como medición de máxima similitud. Esta medición calcula la energía de la diferencia en el espectro de dos conjuntos de parámetros LPC. Esencialmente evalúa la similitud de los datos de prueba que han sido generados por un modelo estadístico basado en conjunto de parámetros LPC de referencia. Esta medida de distancia está dada por,

$$D(\bar{y}_1, y_2) = \frac{a_{LP_1}^T R_2 a_{LP_1}}{a_{LP_2} R_2 a_{LP_2}}$$

donde R_2 representa la matriz de autocorrelación usada para generar los parámetros de $\bar{x}_2$.

Una versión modificada de la distorsión Itakura-Saito que es una distancia de máxima similitud entre un vector de coeficientes de predicción $\mathbf{x} = [a(1), a(2), \ldots, a(N)]^T$ y otro vector de coeficientes de predicción $\mathbf{y}$ está dado por

$$d_1(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \Phi_x (\mathbf{x} - \mathbf{y}) \tag{8.18}$$

donde

$$\Phi_x = \{\phi(i-k)/\phi(0), 0 \le i, k \le N-1\} \tag{8.19}$$

donde la matriz de autocorrelación normalizada cuyos coeficientes $\phi(i-k)$ son usados para el cálculo del vector de coeficientes de predicción $\mathbf{x}$ usados en (8.15), Ya que los coeficientes de autocorrelación en (8.19) están normalizados por $\phi(0)$, se puede demostrar que la matriz Φ_x y el vector $\mathbf{x}$ determinan de forma única el uno al otro. Es importante notar que Φ_x en (8.18) es en realidad una matriz de ponderación pero a diferencia de (8.11) donde $\mathbf{W}$ está fijas, aquí Φ_x cambia de valor al cambiar $\mathbf{x}$. Ya que $\Phi_x \neq \Phi_y$ para $\mathbf{x} \neq \mathbf{y}$, la distorsión de Itakura-Saito no es simétrica respecto a sus argumentos , por ejemplo, $d_1(\mathbf{x}, \mathbf{y}) \neq d_1(\mathbf{y}, \mathbf{x})$. La medida de distorsión no es una distancia o medida. En contraste la distorsión MSE ponderada es una distancia simétrica y una medida.

A pesar que el cálculo de d_1 en (8.18) implica la multiplicación de una matriz, el cálculo puede simplificarse considerablemente y se reduce a un producto escalar (punto).

8.3 Diseño del Diccionario

Para diseñar un diccionario de L-niveles, particionamos el espacio N-dimensional en L celdas $\{c_i, 1 \le i \le L\}$ y asociamos con cada celda c_i con un vector $\mathbf{y}_i$. El cuantizador entonces asigna el vector de código $\mathbf{y}_i$ si $\mathbf{x}$ está en c_i. Un cuantizador se dice que es un cuantizador óptimo (de distorsión mínima) si la distorsión en (8.4) es minimizada sobre todos los cuantizadores de L niveles. Existen dos condiciones para que sean óptimo. La primera condición es que el cuantizador óptimo se realiza usando una regla de distorsión mínima o del vecino más cercano

$$\mathbf{q}(\mathbf{x}) = \mathbf{y}_i \text{ , } si \text{ } y \text{ } solo \text{ } si \text{ } d(\mathbf{x},\mathbf{y}_j) \text{ , } j \neq i, \; 1 \leq j \leq L \tag{8.20}$$

Esto es, el cuantizador escoge el vector de código que resulta de la distorsión mínima con respecto $\mathbf{x}$. La segunda condición necesaria para que sea óptimo, es que cada vector de código $\mathbf{y}_i$ se escoge para minimizar la distorsión promedio en la celda c_i. Esto es $\mathbf{y}_i$ es el vector y que minimiza,

$$D_i = \xi\left[d(\mathbf{x},\mathbf{y}) | \mathbf{x} \in C_i\right] = \int_{\mathbf{x} \in C_i} d(\mathbf{x},\mathbf{y}) p(\mathbf{x}) d\mathbf{x} \tag{8.21}$$

A tal vector lo denominaremos centroide de la celda c_i, y escribimos

$$\mathbf{y}_i = cent(C_i) \tag{8.22}$$

El cálculo del centroide para una región particular dependerá de la definición de la medida de distorsión. (las celdas definidas se conocen como celdas del vecino más cercano, celdas de Veroni o regiones de Dirichlet).

En la práctica obtenemos una serie de vectores de entrenamiento $\{\mathbf{x}(n), 1 \leq n \leq M\}$. Un subconjunto de M_i de estos vectores se encontrará asignado a la celda C_i. La distorsión promedio D_i está dada por

$$D_i = \frac{1}{M_i} \sum_{\mathbf{x} \in C_i} d(\mathbf{x}, \mathbf{y}_i) \tag{8.23}$$

Para los criterios MSE o MSE ponderado, se puede demostrar que D_i es minimizada por

$$\mathbf{y}_i = \frac{1}{M_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}(n) \tag{8.24}$$

o simplemente $\mathbf{y}_i$ es la medida muestral de todos los vectores de entrenamiento contenidos en C_i. Para la distorsión de Itakura-Saito d_1, se puede demostrar que $\mathbf{y}_i$ se calcula primero mediante la promediación de la autocorrelación normalizada que corresponde a los vectores de muestra[2].

$$\phi_{\mathbf{y}_i}(k) = \frac{1}{M_i} \sum_{\mathbf{x} \in C_i} \phi_{\mathbf{x}}(k), \quad 0 \leq k \leq N \tag{8.25}$$

donde $\phi_{\mathbf{x}}(k)$ están normalizados tal que $\phi_{\mathbf{x}}(0)=1$. El vector $\mathbf{y}_i$ se obtiene como la solución a (8.15) con $\phi_{\mathbf{x}_i}(k)$ como los coeficientes de autocorrelación.

Un método para el diseño de diccionarios es un algoritmo iterativo de agrupamiento conocido en la literatura de reconocimiento de patrones como algoritmo de K-Medias. Para nuestro caso k=L. El algoritmo divide el conjunto de vectores de entrenamiento $\{\mathbf{x}(n)\}$ en L grupos C_i de tal forma que las condiciones de optimización se cumplan.

---

[2] Este es el método que se utiliza en este trabajo.

Después existe un índice de iteración m y $C_i(m)$ es el i-ésimo grupo en la iteración m, con $\mathbf{y}_i(m)$ su centroide.

8.4 Algoritmo de K-Medias

A continuación se describe el algoritmo de K-Medias también llamado algoritmo de Lloyd generalizado.

Paso 1: Inicialización: Asignar m = 0. Escoger un método adecuado para determinar un conjunto inicial de vectores de código $\mathbf{y}_i(0)$, $1 \le i \le L$.

Paso 2: Clasificación: Clasificar el conjunto de vectores de entrenamiento $\{\mathbf{x}(n), 1 \le n \le M\}$ a grupos de C_i mediante la regla del vecino más cercano

$$\mathbf{x} \in C_i(m), \text{ si y solo si } d[\mathbf{x}, \mathbf{y}_i(m)] \le d[\mathbf{x}, \mathbf{y}_j(m)], \text{ para toda } j \ne i$$

Paso 3: Actualización de los vectores de código: $m \leftarrow m+1$. Actualizar el vector de código de cada grupo, calculando el centroide de los vectores de entrenamiento en cada grupo

$$\mathbf{y}_i(m) = cent(C_i(m)), \quad 1 \le i \le L$$

Paso 4: Prueba de terminación: Si el decremento en la distorsión total $D(m)$ en la iteración m relativa a $D(m-1)$ está por debajo del umbral, se detiene; en caso contrario, volver al Paso 2.

Otra prueba de terminación equivalente puede sustituirse en el Paso 4.

El algoritmo mencionado converge a un óptimo local, pero cualquier solución en general no es única. La optimización global puede ser obtenida mediante la inicialización de los vectores de código a diferentes valores y repitiendo el algoritmo para diferentes valores de inicialización y escogiendo el diccionario que resulte en la menor distorsión de todos.

8.5 Otros Algoritmos de Agrupamiento

Existen dos variantes del algoritmo K-Medias muy utilizados que producen diccionarios estructurados; el algoritmo ISODATA[TOU81] y el algoritmo LBG [LIND80].

ISODATA (Iterative Self-Organizing Data Analysis Techniques A) es un método iterativo que va dividiendo los grupos en forma de árbol binario. Se basa en las distancias internas de los grupos para decidir los criterios de separación o de unión de los grupos. Los parámetros son el número máximo de grupos, el número mínimo de muestras en cada grupo, la desviación estándar máxima del grupo, la distancia mínima entre grupos y el número máximo de iteraciones.

LBG (por sus autores Linde, Buzo y Gray) es un método que se basa en un modelo probabilístico o en una secuencia grande de muestras. Algunas de las propiedades de este algoritmo se desarrollaron usando argumentos heurísticos.

9. Técnicas Difusas de Reconocimiento de Patrones[1]

9.1 Modelos para el Reconocimiento de Patrones.

El reconocimiento de patrones es una de las áreas de aplicación más vieja y obvia para la teoría de conjuntos difusos. El término "Reconocimiento de Patrones" abarca una gran cantidad de literatura muy diversa, incluyendo investigación en el área de inteligencia artificial, gráficos por computadora interactivos, diseño asistido por computadora, reconocimiento de patrones fisiológicos y biológicos, lingüísticos y estructurales entre otros. Se puede hacer una distinción entre reconocimiento de patrones "matemático" (antes llamado análisis de concentraciones) y "no matemático".

9.1.1 Los Datos.

Los datos se pueden obtener virtualmente de cualquier proceso o fenómeno físico. Pueden ser cualitativos o cuantitativos, numéricos, de imágenes, texturas, lingüísticos o cualquier combinación de las anteriores. La dimensionalidad puede variar de una dimensión o espacios de múltiples dimensiones. Al conjunto de datos se le denominará X.

9.1.2 Estructura o Espacio Patrón.

Se espera que los datos observados, porten información acerca del proceso que los generó o el fenómeno que representan. Por estructura, entendemos la forma en la que la información se puede organizar de tal forma que las relaciones entre las variables del proceso puedan ser identificados. La dimensionalidad del espacio patrón que contiene las propiedades estructurales, es generalmente menor que el del espacio de datos.

9.1.3 Espacio y Selección de Características.

El espacio de características es un espacio intermedio entre el espacio de datos y el proceso de clasificación. Generalmente es de una dimensión mucho menor que el espacio de datos. Esto es esencial para poder aplicar técnicas eficientes de búsqueda de patrones.

La selección de características busca la estructura interna de los datos, esto es, características o propiedades de los datos que nos permitan reconocer y desplegar sus estructura. Aquí surge una pregunta: ¿ Son las características seleccionadas lo suficientemente representativas del proceso físico que generó los datos, como para que se puedan construir grupos o clasificaciones realistas?

9.1.4 Clasificación y Espacio de clasificación.

Este espacio contiene la decisión determinada por el algoritmo de clasificación. Un clasificador en si es un dispositivo, medio o algoritmo por el cual el espacio de datos es particionado en c

[1] En su mayor parte tomado de [ZIMM90].

"regiones de decisión". La clasificación intenta descubrir las asociaciones entre subclases de una población. Es obvio que el espacio de clasificación, normalmente es una dimensión pequeña.

Bezdek en 1981, resumió las características principales como sigue:

Selección de características: La búsqueda de estructuras de los datos u observaciones $x_k \in X$

Análisis de Grupos: La búsqueda de estructura en los conjuntos de datos o muestras, $X \subset S$
Clasificación: La búsqueda de estructuras en los espacios de datos o poblaciones S.

Si se pudieran escoger características "óptimas", el agrupamiento y la clasificación serían triviales; por otro lado tratamos de descubrir las características óptimas mediante la agrupación de las variables con las características.

9.2 Agrupamiento Difuso.

9.2.1 Métodos de agrupamiento.

Asumamos que el problema importante de extracción de características, ha sido resuelto. Nuestra tarea entonces es dividir n objetos $x \in X$ caracterizados por p indicadores a c, $2 \leq c \leq n$ subconjuntos homogéneos categorizados denominados "grupos". Los objetos que pertenecen a cualquiera de los grupos debe ser similares y los objetos que pertenecen a otro grupo deben ser lo suficientemente diferentes. El número c de grupos, normalmente no se conoce de antemano.

La pregunta más importante que se tiene antes de aplicar cualquier procedimiento de agrupamiento es: que propiedades matemáticas como distancia, conectividad, intensidad, etc., deberán utilizarse en orden de identificar los grupos. Esta pregunta debe ser contestada para cada conjunto de datos por separado ya que no existe un criterio universal de agrupamiento.

Los algoritmos de agrupamiento no difusos están categorizados de acuerdo al tipo de criterio de agrupamiento en los métodos jerárquicos, grafo-teóricos y objetivos-funcionales.

Los métodos jerárquicos generan una jerarquía de particiones por medio de la unión sucesiva o separación de los grupos. Tal jerarquía puede ser representada por un dendograma, que puede ser utilizado para estimar un número apropiado de grupos c, para otros métodos de agrupamiento. En cada nivel de agrupamiento o separación, se puede utilizar una estrategia local óptima, sin tomar en cuenta las políticas usadas en niveles precedentes. La mayor ventaja de éstos métodos es su simplicidad tanto en concepción como en cálculo.

En la teoría de conjuntos difusos, este tipo de método de agrupamiento corresponde a la determinación de "árboles de similitud".

Los métodos grafo-teóricos están normalmente basados en algún tipo de conectividad de los nodos de un grafo que representa el conjunto de datos. La estrategia de agrupamiento es la de segmentar los extremos con el objeto de formar subgrafos. Si el grafo que representa la estructura de datos es un grafo difuso, entonces diferentes nociones de conectividad llevan a diferentes tipos de grupos, que a su vez pueden ser representados como dendogramas.

Los métodos objetivos-funcionales permiten la formulación más precisa de los criterios de agrupamiento. La "deseabilidad" de los candidatos a ser agrupados, es medida para cada c, que es

el número de grupos usando una función objetiva. Normalmente se definen extremos locales de la función objetiva se definen como agrupamientos óptimos.

Los algoritmos clásicos no-difusos generan particiones de tal forma que cada objeto es asignado a exactamente un grupo. A menudo, los objetos no pueden ser asignados adecuadamente a un sólo grupo (pueden ser que estén ubicados "entre" dos grupos). En éstos casos los métodos de agrupación difusos proporcionan una herramienta más adecuada para representar estructuras reales de datos.

Consideremos los métodos de agrupamiento por sí mismos. Sea el conjunto de datos $X = \{x_1,\ldots,x_n\} < \mathbf{R}^p$ un subconjunto de espacio vectorial real de p-dimensiones $\mathbf{R}^p$. Cada $x_k = (x_{k_1},\ldots,x_{k_p}) \in \mathbf{R}^p$ es denominado un vector de características. x_{k_j} es la j-ésima característica de la observación x_k.

Ya que los elementos de un grupo deben ser muy similares y los grupos lo más diferente posible, el proceso de agrupamiento es controlado por el uso de medidas de similitud. Normalmente se define la 'disimilitud' o 'distancia' de dos objetos x_k y x_l como una función de valor real $d: X \times X \to R^+$ que satisface

$$d(x_k,x_l) = d_{kl} \geq 0$$
$$d_{kl} = 0 \Leftrightarrow x_k = x_l \qquad (9.1)$$
$$d_{kl} = d_{lk}$$

Si además d satisface la desigualdad del triángulo, esto es

$$d_{kl} \leq d_{kj} + d_{jl} \qquad (9.2)$$

entonces d es una medida, esto es una propiedad que no siempre se requiere. Si cada vector de características es considerado como un punto en el espacio de p-dimensiones entonces la disimilitud d_{kl} de dos puntos x_k y x_l se puede interpretar como la distancia entre esos dos puntos.

Cada partición del conjunto $X = \{x_1,\ldots,x_n\}$ a un subconjunto no-difuso o difuso $\tilde{S}_i (i = 1,\ldots,c)$ puede ser totalmente descrito por una función indicador u_{S_i} o una función de pertenencia $\mu_{\tilde{S}_i}$ respectivamente.

En orden de establecer una terminología se determina para métodos no difusos

$$u_{S_i}: X \to \{0,1\} \qquad (9.3)$$

y para difusos

$$\mu_{\tilde{S}_i}: X \to [0,1] \qquad (9.4)$$

donde u_{ik} y μ_{ik} denotan el grado de pertenencia del objeto x_k al subconjunto $\tilde{S}_i$, esto es,

$$u_{ik} := u_{S_i}(x_k)$$
$$\mu_{ik} := \mu_{\tilde{S}_i}(x_k) \qquad (9.5)$$

Definición 9.1

Sea $X = \{x_1,\ldots,x_n\}$ un conjunto finito. V_{cn} es el conjunto de todas las matrices reales $c \times n$ y $2 \leq c \leq n$ es un número entero. La matriz $U = [u_{ik}] \in V_{cn}$ es llamada una partición no difusa en c si satisface las siguientes condiciones:

$$\begin{aligned} &1.\ u_{ik} \in \{0,1\} \qquad 1 \le i \le c,\ 1 \le k \le n \\ &2.\ \sum_{i=1}^{c} u_{ik} = 1, \qquad 1 \le k \le n \\ &3.\ 0 < \sum_{k=1}^{n} u_{ik} < n \quad 1 \le i \le c \end{aligned} \tag{9.6}$$

El conjunto de todas las matrices que satisfacen éstas condiciones es llamado M_c.

Definición 9.2

X, V_{cn} y c se definieron anteriormente. $\tilde{U} = [\mu_{ik}] \in V_{cn}$ es llamada una partición difusa en c si satisface las siguientes condiciones:

$$\begin{aligned} &1.\ \mu_{ik} \in [0,1] \qquad 1 \le i \le c,\ 1 \le k \le n \\ &2.\ \sum_{i=1}^{c} \mu_{ik} = 1, \qquad 1 \le k \le n \\ &3.\ 0 < \sum_{k=1}^{n} \mu_{ik} < n \quad 1 \le i \le c \end{aligned} \tag{9.7}$$

M_{fc} denota el conjunto de todas las matrices que satisfacen las condiciones anteriores. En contraste de la partición no difusa en c, los elementos pueden pertenecer a varios grupos y en diferentes grados. Las condiciones 2 y 3 requieren que el "total de pertenencia" de un elemento sea normalizado a 1 y que no pueda pertenecer a más grupos de los que existen,

La ubicación de un grupo está representada por su "centro de grupo" o centroide $v_i = (v_{i1},\ldots,v_{ip}) \in \mathbf{R}^p, i = 1,\ldots,c$, alrededor de éste se encuentran concentrados sus objetos.

Sea $v = (v_1,\ldots,v_c) \in \mathbf{R}^{cp}$ el vector de todos los centros de los grupos, donde v_i en general no corresponde a los elementos de X.

Uno de los criterios frecuentemente utilizados para mejorar la partición inicial es el llamado criterio de la variancia. Este criterio mide la disimilitud entre los puntos de un grupo y su centro por medio de la distancia Euclidiana. Esta distancia, d_{ik} es entonces,

$$\begin{aligned} d_{ik} &= d(x_k, v_i) \\ &= \|x_k - v_i\| \\ &= \left[\sum_{j=1}^{p} (x_{kj} - v_{ij})^2\right]^{1/2} \end{aligned} \tag{9.8}$$

El criterio de la variancia para las particiones no difusas corresponde a minimizar la suma de las variancias de todas las variables j en cada grupo i, con $|S_i| = n$, resulta:

$$\min \sum_{i=1}^{c} \sum_{j=1}^{p} \frac{1}{|S_i|} \sum_{x_k \in S_i} (x_{kj} - v_{ij})^2 \Leftrightarrow \min \frac{1}{n} \sum_{i=1}^{c} \sum_{x_k \in S_i} \sum_{j=1}^{p} (x_{kj} - v_{ij})^2 \tag{9.9}$$

como se indica en transformación anterior, el criterio de la variancia corresponde, excepto por el factor $1/n$, a minimizar la suma de las distancias Euclidianas cuadráticas. El criterio por si mismo es equivalente a resolver el siguiente problema,

$$\min z(S_1,\ldots,S_c;v)=\sum_{i=1}^{c}\sum_{x_k\in S_i}\|x_k-v_i\|^2 \tag{9.10}$$

tal que

$$v_i=\frac{1}{|S_i|}\sum_{x_k\in S_i}x_k \tag{9.11}$$

Usando la definición 9.1, el criterio de la variancia para las particiones no difusas puede escribirse como,

$$\min z(\tilde{U},v)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}\|x_k-v_i\|^2 \tag{9.12}$$

tal que

$$v_i=\frac{1}{\sum_{k=1}^{n}u_{ik}}\sum_{k=1}^{n}(u_{ik})x_k \tag{9.13}$$

Para las particiones difusas de acuerdo a la definición 9.2 el criterio de la variancia es equivalente a resolver el siguiente problema:

$$\min z(U,v)=\sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m\|x_k-v_i\|^2 \tag{9.14}$$

tal que

$$v_i=\frac{1}{\sum_{k=1}^{n}u_{ik}}\sum_{k=1}^{n}(u_{ik})^m x_k,\quad m>1 \tag{9.15}$$

Donde v_i es la media de x_k ponderada por m mediante sus grados de pertenencia. Esto significa que los x_k con mayores grados de pertenencia tienen mayor influencia en v_i que aquéllos con bajos grados de pertenencia.

Generalizando el criterio concerniente a la norma utilizada en el problema de agrupamiento no difuso, se puede plantear como sigue: Sea G una matriz de $p\times p$ que es simétrica y positiva definida. Entonces podemos definir una norma general,

$$\|x_k-v_i\|_G^2=(x_k-v_i)^T G(x_k-v_i) \tag{9.16}$$

Esto nos lleva a la formulación del problema,

$$\min z(U,v)=\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}\|x_k-v_i\|_G^2 \tag{9.17}$$

tal que

$U\in M_c$

$v\in R^{cp}$

Esto es un problema de optimización combinacional difícil de resolver, aún para valores pequeños de c y n. De hecho el número de formas distintas de particionar x en subconjuntos no vacíos es,

$$|M_c| = (1/c!)\left[\sum_{j=1}^{c}\binom{c}{j}(-1)^{c-j} j^n\right] \tag{9.18}$$

que para c=10 y n=25 son casi 10^{18} posibles particiones de 10 centros de los 25 puntos.
La definición básica para el problema de partición difuso para m>1 es,

$$\min z_m\left(\tilde{U};v\right) = \sum_{k=1}^{n}\sum_{i=1}^{c}\left(\mu_{ik}\right)^m \left\|x_k - v_i\right\|_G^2 \tag{9.19}$$

tal que

$$\tilde{U} \in M_{fc}$$

$$v \in R^{cp}$$

Diferenciando la función objetiva con respecto a v_i (para $\tilde{U}$ fijo) y a μ_{ik} (para v fijo) y aplicando la condición $\sum_{i=1}\mu_{ik} = 1$ se obtiene:

$$v_i = \frac{1}{\sum_{k=1}^{n}\left(\mu_{ik}\right)^m}\sum_{k=1}^{n}\left(\mu_{ik}\right)^m x_k, \quad i = 1,\ldots,c \tag{9.20}$$

$$\mu_{ik} = \frac{\left(\frac{1}{\left\|x_k - v_i\right\|_G^2}\right)^{1/(m-1)}}{\sum_{j=1}^{c}\left(\frac{1}{\left\|x_k - v_i\right\|_G^2}\right)^{1/(m-1)}}, \quad i = 1,\ldots,c; k = 1,\ldots,n \tag{9.21}$$

Ahora se denotará la importancia de m: Se le denomina el paso exponencial y reduce la influencia de "ruido" cuando se calculan los centros de los grupos en la ecuación (9.20) y el valor de la función objetiva $z_m\left(\tilde{U},v\right)$. m reduce la influencia de μ_{ik} pequeños (lejos de v_i) comparados con μ_{ik} grandes (cerca de v_i). Entre más grande es m>1 mayor es lo fuerte es la influencia.

Los sistemas descritos por las ecuaciones (9.20) y (9.21) no pueden ser resueltos en forma analítica. Existen algoritmo iterativos (no jerárquicos) que aproximan el mínimo de la función objetiva, comenzando de una posición dada. Uno de los mejores algoritmos para agrupamiento no difuso es el algoritmo de C-Medias o ISODATA. De forma similar el problema de agrupación difuso puede resolverse usando el algoritmo C-Medias difuso.

9.2.1.1 Algoritmo C-Medias Difuso

Para cada $m \in (0,\infty)$ se pude diseñar un algoritmo C-Medias difuso que resuelve en forma iterativa las condiciones necesarias (9.20) y (9.21) y converge a un óptimo local.
El algoritmo comprende los siguientes pasos:

Paso 1. Escoger $c(2 \le c \le n), m(1 < m < \infty)$ y la matriz G de $p \times p$ siendo G simétrica y positiva definida. Inicializar $\tilde{U}^{(0)} \in M_{fc}$, hacer $l = 0$.

Paso 2. Calcular los c centros difusos de los grupos $\{v_i^{(l)}\}$ usando $\tilde{U}^{(l)}$ de la condición (9.20).

Paso 3. Calcular la nueva matriz de pertenencia $\tilde{U}^{(l+1)}$ usando $\{v_i^{(l)}\}$ de la condición (9.21) si $x_k \neq v_i^{(l)}$. Sino hacer,

$$\mu_{ik} = \begin{cases} 1 & \text{para } j = 1 \\ 0 & \text{para } j \neq 1 \end{cases}$$

Paso 4. Escoger una norma de la matriz adecuada y calcular $\Delta = \left\| \tilde{U}^{(l+1)} - \tilde{U}^{(l)} \right\|_G \le \varepsilon$ si $\Delta > \varepsilon$ entonces $l = l + 1$ y se va al paso 2. Si $\Delta \le \varepsilon$ entonces se detiene.

Para el algoritmo de C-medios difuso se deben elegir un número de parámetros:

El número de grupos c, $2 \le c \le n$;
El peso exponencial m, $1 < m < \infty$;
La matriz G de $p \times p$ que induce a la norma;
El método de inicializar la matriz de pertenencia $\tilde{U}^{(0)}$;
El criterio de terminación $\Delta = \left\| \tilde{U}^{(l+1)} - \tilde{U}^{(l)} \right\|_G \le \varepsilon$

Así como en otros algoritmos iterativos para mejorar particiones iniciales, el número c debe escogerse adecuadamente. Si no existe información acerca de un buen valor de c, los cálculos se realizan para varios valores de c.

En un segundo paso, la mejor sobre éstas particiones se selecciona.

El peso exponencial m tiende influencia la matriz de pertenencia. Entre más grande es m, el difusor se convierte en la matriz de pertenencia de la partición final. Para $m \to \infty$, $\tilde{U}$ se aproxima a $\tilde{U} = \left[\frac{1}{c}\right]$. Esto es, una solución indeseable, ya que cada x_k es asignada a cada grupo con el mismo grado de pertenencia.

Básicamente es preferible obtener el menor número de matrices difusas ya que si se obtienen altos grados de pertenencia, es un indicativo de una mayor concentración de los puntos alrededor de los centros de grupo respectivos. No existe ninguna regla justificable para m. Normalmente se escoge m=2.

G determina la forma del grupo, que puede ser identificado por el algoritmo de C-medias difuso. Si se escoge la norma Euclidiana N_E entonces G es la matriz identidad y la forma de los grupos se asume como una hiperesfera de igual tamaño. Otras normas frecuentemente usadas son la norma

diagonal o la norma de Mahalanobis para los cuales $G_D = \left[\text{diag}\left(\sigma_j^2\right)\right]^{-1}$ y $G_M = \left[\text{cov}(x)\right]^{-1}$, respectivamente donde σ_j^2 denota la variancia de la característica j.

La partición final depende de la posición inicial escogida. Cuando se escoge apropiadamente c y existe una buena estructura de agrupación, las particiones finales generadas por el algoritmo de C-medios difuso son bastante estables.

9.2.2 Validez de los Grupos.

Los algoritmos complejos se ajustan entre los datos para los que se hipotetiza sus subestructura y las soluciones que éstos generan; de aquí que sea casi imposible transferir una hipótesis nula teórica acerca de X a $\tilde{U} \in M_{fc}$, que puede ser usada para consolidar o rechazar la validez de los grupos que se sugirieron algorítmicamente. Como resultado una serie de medidas escalares de difusión de las particiones han sido usadas como indicadores heurísticos de validez según Bezdek.

Actualmente el llamado problema de validez de los grupos concierne únicamente a la calidad o grado al cual la partición final de un algoritmo de agrupamiento aproxima la estructura real o hipotética de un conjunto de datos. Esto se reduce a la búsqueda del valor adecuado de c. La validez de los grupos es también relevante cuando se decide cuál de un número de posiciones iniciales debe ser seleccionado para mejorar el desempeño.

Para medir la validez de los grupos en agrupamiento difuso, se han adaptado algunos criterios de análisis de grupos no difusos. En particular los llamados funcionales de validez utilizados, expresan la calidad de una solución midiendo el grado de difusión. Mientras que los criterios para la validez de grupos están íntimamente relacionados a la formulación matemática del problema el criterio para juzgar la que tan apropiada es la partición final considera características reales en vez de matemáticas.

Uno de los criterios más apropiados es el valor de la función objetiva. Ya que decrece monotónicamente con un número creciente de grupos, c, esto es, llega a su mínimo para c=n, se escoge c^* para el cual se obtiene un gran decremento cuando se pasa de c^* a c^*+1.

Otro criterio es la razón de convergencia. Esto es justificable ya que la experiencia ha mostrado que, para una buena estructura de agrupamiento y una c apropiada, se obtiene una tasa de convergencia muy alta.

Ya que la porción final "óptima" depende de la inicialización de la partición inicial, esto se puede utilizar como una indicación de un número correcto de grupos c.

Los tres criterios sirven para determinar el número "correcto" de grupos. Son heurísticos en naturaleza y por lo tanto deben llevar a las particiones finales que identifican correctamente a los grupos existentes.

El siguiente criterio calcula la validez funcional de los grupos que asignan a cada partición difusa final un escalar que supuestamente indica la calidad de la solución del agrupamiento. Cuando se diseña tal criterio se asume que la estructura de los grupos es mejor identificada cuando se concentran más puntos alrededor de los centros de los grupos, esto es, el grado no difuso de la matriz de pertenencia de la partición final.

Las mejores medidas para juzgar la difusión de una solución de agrupamiento son:

Los coeficientes de partición, $F(\tilde{U};c)$,

La entropía de la partición, $H(\tilde{U},c)$, y

El exponente de proporción $P(\tilde{U};c)$.

Definición 9-3.

Sea $\tilde{U} \in M_{fc}$ una partición c difusa de n puntos de datos. El coeficiente de partición de $\tilde{U}$ es el escalar

$$F(\tilde{U};c) = \sum_{k=1}^{n}\sum_{i=1}^{c}\frac{(\mu_{ik})^2}{n}$$

Definición 9-4

La entropía de la partición de cualquier partición c difusa $\tilde{U} \in M_{fc}$ de X donde $|X| = n$, es para $1 \le c \le n$

$$H(\tilde{U};c) = \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{c}\mu_{ik}\log_e(\mu_{ik})$$

Definición 9-5

Sea $\tilde{U} \in (M_{fc}/M_{c0})$ una partición c difusa de $X;|x| = n;2 \le c \le n$ para la columna k de $\tilde{U}, 1 \le k \le n$, sea

$$\mu_k = \max_{1 \le i \le c}\{\mu_{ik}\}$$

$$\left[\mu_k^{-1}\right] = \textit{entero mayor} \le \left(\frac{1}{\mu_k}\right)$$

El exponente de proporción de U es el escalar

$$P(\tilde{U},c) = -\log_e\left\{\prod_{k=1}^{n}\left[\sum_{j=1}^{\mu_k^{-1}}(-1)^{j+1}\binom{c}{j}(1-j\mu_k)^{(c-1)}\right]\right\}$$

Las medidas mencionadas tienen las siguientes propiedades

$$\frac{1}{c} \le F(\tilde{U},c) \le 1$$

$$0 \le H(\tilde{U},c) \le \log_e(c)$$

$$0 \le P(\tilde{U},c) < \infty$$

El coeficiente de partición y la entropía de la partición son similares en el sentido que ambos obtienen sus extremos para las particiones no difusas $\tilde{U} \in M_c$:

$$F(\tilde{U},c) = 1 \Leftrightarrow H(\tilde{U},c) \Leftrightarrow \tilde{U} \in M_c$$

$$F(\tilde{U},c) = \frac{1}{c} \Leftrightarrow H(\tilde{U},c) = \log_e(c) \Leftrightarrow \tilde{U} = \left[\frac{1}{c}\right]$$

Las reglas heurísticas para seleccionar el más correcto o el mejor número de particiones son:

$$\max_c \left\{ \max_{U \in \Omega_c} \left\{ F(\tilde{U},c) \right\} \right\} \quad c = 2,\ldots,n-1$$

$$\min_c \left\{ \min_{U \in \Omega_c} \left\{ H(\tilde{U},c) \right\} \right\} \quad c = 2,\ldots,n-1$$

donde Ω_c es el conjunto de todas las soluciones óptimas para un c dado.
Las limitaciones de $F(\tilde{U},c)$ y $H(\tilde{U},c)$ son principalmente su monotonicidad y la falta de una forma correcta de evaluación que permitiera elaborar un juicio sobre la aceptabilidad de la partición final.

$H(\tilde{U},c)$ es normalmente más sensitiva con respecto a un cambio en la partición que $F(\tilde{U},c)$. Esto se cumple más al variar m.

Mientras que $F(\tilde{U},c)$ y $H(\tilde{U},c)$ dependen de todos los $c \cdot n$ elementos, el exponente de proporción $P(\tilde{U},c)$ depende del máximo grado de pertenencia de los n elementos. $P(\tilde{U},c)$ converge hacia ∞ al incrementarse μ_k y no está definida para $\mu_k = 1$.
La regla heurística para escoger una partición es,

$$\max_c \left\{ \max_{U \in \Omega_c} \left\{ P(\tilde{U},c) \right\} \right\} \quad c = 2,\ldots,n-1$$

En contraste a $F(\tilde{U},c)$ y $H(\tilde{U},c)$, $P(\tilde{U},c)$ tiene la ventaja que no es monotónica en c. No existen formas correctas de evaluación que permitan evaluar la calidad de una porción c^* del valor de $P(\tilde{U}^*,c^*)$.

La regla heurística para $P(\tilde{U},c)$ lleva una partición final "óptima" diferente a las reglas heurísticas de $F(\tilde{U},c)$ y/o $H(\tilde{U},c)$. Esto quizás requiere de otras formas de decisión derivadas de los datos en sí o de otras consideraciones.

10. Implantación de las Técnicas de Reconocimiento de Voz

Las técnicas utilizadas realizan el reconocimiento de voz de palabras aisladas e independientes del locutor. Para esto se tomo la base de datos TI-46 de dígitos en inglés. Se utilizaron diez locutores prununciando diez veces cada uno de los diez dígitos, dando un total de 1,000 repeticiones para la parte de entrenamiento. Para la parte de reconocimiento se tomaron dieciséis repeticiones en lugar de diez. Los dígitos son pronunciados tanto por hombres como mujeres y no se hace distinción del sexo durante el proceso de reconocimiento.

Los programas de segmentación acústica ya se encontraban escritos y se ejecutan bajo plataforma Sun usando el sistema operativo Sunos 4.1.3. El objetivo de estos programas es tomar una repetición de un dígito muestreada a 12,500Hz, cambiar su tasa de muestreo a 10,000Hz y sobre esta realizar la segmentación acústica. Por lo tanto después de procesar los archivos, se obtuvieron 2,600 archivos conteniendo las repeticiones de los dígitos muestreados a una frecuencia de 10,000Hz y ademas se obtuvieron 2,600 archivos con la segmentación acústica por cada repetición (extensión ".SAC"), donde el formato es el siguiente:

0	600
1	1100
2	1900
3	2300
4	2800
5	3100
6	5500
7	6200
8	7500
9	8400
10	9200

Tabla 10.1 Contenido del archivo de segmentación acústica.

El elemento 0 indica el número de muestra del comienzo de la primera subpalabra acústica, el elemento 10 indica el número de muestra final de la última subpalabra acústica. Por lo tanto esta repetición constaría de nueve subpalabras acústicas distribuidas de la siguiente forma:

Subpalabra	# Muestra Inicial	# Muestra Final
1	600	1099
2	1100	1899
3	1900	2299
...	...	...
9	8400	9199

Tabla 10.2 Interpretación del contenido del archivo de segmentación acústica

Para la identificación de los archivos se utilizó el siguiente formato:

Para entrenamiento:

Wxxyyyz

donde

xx denota el dígito siendo

20 'zero'
11 'one'
12 'two'
13 'three'
14 'four'
15 'five'
16 'six'
17 'seven'
18 'eight'
19 'nine'

yyy denota la identificación del locutor.

z el número de la repetición del '0' al '9'.

Para reconocimiento

Wxxyyyzz

donde

xx e yyy tienen la misma interpretación que para el entrenamiento.
zz el número de repetición del '00' al '15'.

10.1 Procesamiento de los Archivos de Voz

El primer paso en el reconocimiento de voz es el de obtener vectores de características de los archivos de voz por subpalabra acústica y a cada una de éstas, se les divide en tramas con traslape con la trama adyacente. Con el objeto de igualar las características espectrales de la señal en altas frecuencias se aplica un filtro digital de preénfasis.

Ya que la señal viene codificada como un entero con un rango dinámico de 12 bits, se hace una transformación y normalización de la señal de punto fijo a punto flotante con un rango dinámico que va de 1.00 a -1.00, esto permite manipular en una forma más eficaz las muestras. Al concepto de energía no se le da mucho énfasis, ya que no se pretende hacer síntesis de voz, donde la ganancia de la señal sería una función de la energía de la señal. Lo que se pretende es realizar un apareamiento espectral y generar patrones o vectores de características representativos.

10.1.1 Generación de los Vectores de Características de la Señal de Voz por Subpalabra Acústica

El vector de características que se escogió a utilizar, fue el de coeficientes de predicción lineal (LPC), debido a que este modelo nos da una buena representación de la señal de voz para tramas estacionarias de la misma. El objetivo del LPC es el de representar a la señal de voz como un modelo de un filtro lineal que contiene solo polos y es variante en el tiempo.

Suponemos que la señal es estacionaria dentro de un intervalo de 128 muestras que equivale a 12.8 ms de voz. Cada nueva trama va a tener un traslape de 20 muestras con la trama anterior o lo que es lo mismo 2ms. Esto permite que los coeficientes se vayan calculando sobre tramas que contienen información espectral de la trama anterior y por lo tanto deben de ser similares.

El número de coeficientes a utilizar es de 8, con la restricción que a_0 es igual a 1. Esto resulta en que se tengan 7 coeficientes que aportan información real sobre la trama a analizar.

Ya que se conoce el inicio y el fin de cada segmento acústico, se calcula el número de tramas a procesar (1)

$$N_T = \max \text{ int} \frac{N_{MSA}}{Tr - N_{TR}} \tag{10.1}$$

donde N_T es un entero que determina el número de tramas, N_{MSA} es el numero de muestras del segmento acústico, N_{TR} es el número de muestras de traslape entre una trama a otra y Tr es la longitud de la trama en muestras.

A cada trama se le aplica una ventana de Hamming, que es equivalente a aplicar un filtro pasa bajas a la señal contenida en la trama. Esto se hace con el objeto de que no existan transiciones abruptas que puedan generar altas frecuencias no contenidas en la señal real.

Una vez realizado el proceso anterior, se calcula un estimador sesgado de la autocorrelación (2) para la trama, con este se calculan los coeficientes LPC utilizando el método de Levinson-Durbin que es un algoritmo eficiente iterativo para la obtención de los coeficientes.

Estimador sesgado de la autocorrelación

$$r_x[l] = \frac{1}{N} \sum_{k=0}^{N-1-|l|} E\{x[k]x[k+l]\} \quad |l| = 0,1,2,\cdots,N-1 \tag{10.2}$$

Método de Levinson-Durbin

$$E^{(0)} = r(0) \tag{10.3}$$

$$k_i = \left\{ r(i) - \sum_{j=1}^{L-1} \alpha_j^{(i-1)} r(|i-j|) \right\} \Big/ E^{(i-1)} \tag{10.4}$$

$$\alpha_i^{(i)} = k_i \tag{10.5}$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}, \ 1 \le j < i \tag{10.6}$$

$$E^{(i)} = \left(1 - k_i^2\right) E^{(i-1)} \tag{10.7}$$

donde la sumatoria en (10.4) se omite para i=1. Las ecuaciones (10.4-10.7) se resuelven recursivamente para i = 1, 2, ...,p y la solución final esta dada como

$$a_m = coeficientes \ \ LPC = \alpha_m^{(p)}, \ 1 \le m \le p \tag{10.8}$$

$$k_m = coeficientes \ \ de \ \ reflexión \tag{10.9}$$

Después de realizar este proceso se obtiene la energía mínima residual a partir de los coeficientes LPC y el vector de autocorrelación. Para obtener este valor se utiliza una simplificación del procedimiento de multiplicación vectorial tomando que la matriz de autocorrelación es simétrica.

$$aRa^T = r_a(0)r(0) + 2\sum_{i=1}^{p} r_a(i)r(i) \qquad (10.10)$$

donde r_a es la autocorrelación de los coeficientes LPC y r es la autocorrelación de la señal de voz.

Para el caso de agrupamiento difuso, se optó por utilizar coeficientes cepstrales. La razón de esto se debe a que el programa de agrupamiento utilizado, calcula las distancias vectoriales utilizando la distancia euclidiana. Esta no es la única que permite el programa, pero es la más simple y manejable. Para el calculo de los coeficientes cepstrales se utilizó una formula recursiva que permite pasar de coeficientes LPC a Cepstrum. La única restricción, es que recomiendan utilizar un mayor número de coeficientes cepstrales al de LPC.

$$c_o = \ln\sigma^2 \qquad (10.11)$$

$$c_m = a_m + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right) c_k a_{m-k}, \quad 1 \le m \le p \qquad (10.12)$$

$$c_m = \sum_{k=1}^{m-1}\left(\frac{k}{m}\right) c_k a_{m-k}, \quad m > p \qquad (10.13)$$

donde σ^2 es el término de ganancia del modelo LPC. Generalmente se usa una representación con $Q > p$ donde $Q \approx \frac{3}{2}p$ es el número de coeficientes cepstrales.

10.2 Entrenamiento del sistema.

Se utilizaron dos técnicas de agrupamiento:

- K-Medias (LPC)
- C-Medias Difuso (Cepstrum)

Con cada una de estas técnicas se obtuvieron 16 centroides por cada segmento acústico, es decir si una palabra tiene N segmentos acústicos, el número de centroides es N * 16. Esto no significa que este número influye sobre el reconocimiento total, ya que cada segmento es comparado por separado, que es la parte importante de la técnica de reconocimiento automático de voz de este trabajo.

La primera técnica es común dentro de la cuantización vectorial, también se le conoce como algoritmo de Lloyd generalizado. Es un algoritmo simple, que como se mencionó anteriormente en el capitulo que trata sobre la agrupación no difusa permite establecer una serie de centroides iniciales e ir calculando la distancia menor de cada vector al centroide más cercano y asignar la pertenencia de este vector al centroide.

Para el calculo de la distancia se utilizó la distancia de Itakura-Saito modificada. Esta distancia es perceptual, esto significa que dos tramas cuya distancia es pequeña, suenan en forma similar y dos tramas cuyas distancias sean grandes, suenan diferente.

Para el calculo del centroide, se saca un promedio de la autocorrelación de todos los vectores asignados cuya distancia a ese centroide es la mínima. De esta autocorrelación promedio se calculan los coeficientes LPC y este es el valor del nuevo centroide.

La segunda técnica, se basa en el algoritmo de cuantización vectorial no difuso C-Medias o ISODATA, el cual se modificó para que cada vector pertenezca a todos los centroides, con la diferencia que la pertenencia es parcial y no total como era el caso con K-Medias [WIND83]. Ya que el objeto de la tesis era el de comparar técnicas, se utilizó un "toolbox" de Matlab desarrollado por Bogdan R. Kosanovic [KOSA95] que realiza la técnica de agrupación difusa.

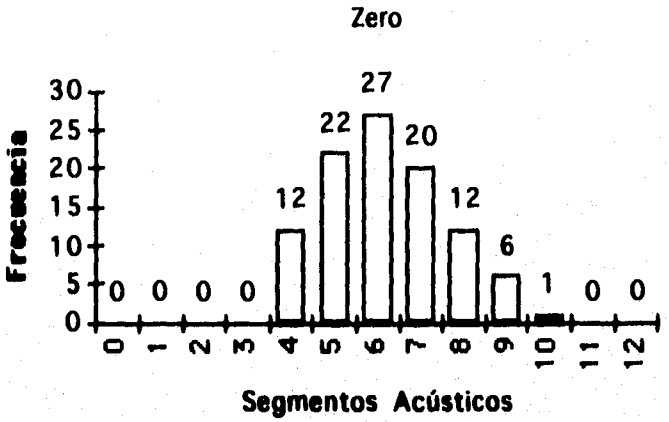
Este "toolbox" tiene varias formas de calcular las distancias entre vectores, "Euclidiana", "Diagonal" y "Mahalanobis". Se escogió la distancia euclidiana, ya que la otra opción era la de modificar el algoritmo para calcular la distancia de Itakura-Saito modificada. Se opto por probar con la distancia Euclidiana y se sabe que para los coeficientes cepstrales [DELL87] esta es una buena medida de similaridad espectral.

10.3 Sistema de Reconocimiento

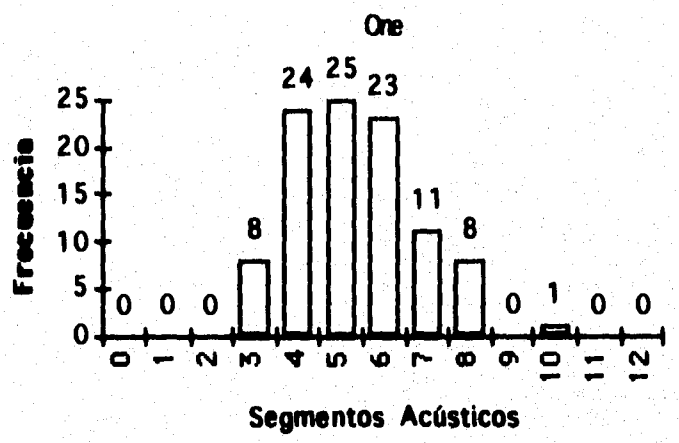
Una vez realizada la etapa de agrupamiento y habiendo obtenido los 16 centroides por cada segmento acústico, se procede a comparar cada segmento acústico que a su vez se divide en tramas y de éstas se obtiene por cada uno, un vector de características. A cada vector de características se le calcula la distancia con respecto a cada centroide y se va acumulando la distancia menor, al finalizar el cálculo de las distancias menores para cada segmento acústico, este se almacena en un archivo. Para los dos métodos de agrupamiento se utiliza el mismo programa de reconocimiento, con la diferencia que cada uno usa una distancia diferente y vectores de parámetros diferentes, aunque uno se deriva de otro.

Una vez almacenado en el archivo, se ejecuta un programa que busca para cada dígito cual fue la distancia menor de la comparación de todos los dígitos con cada uno de los centroides por cada segmento acústico, la salida de éste es una matriz de confusión que indica el número de dígitos reconocidos de forma correcta y en cuales falló. Los resultados obtenidos para ambas técnicas se muestran en el Capítulo 11.

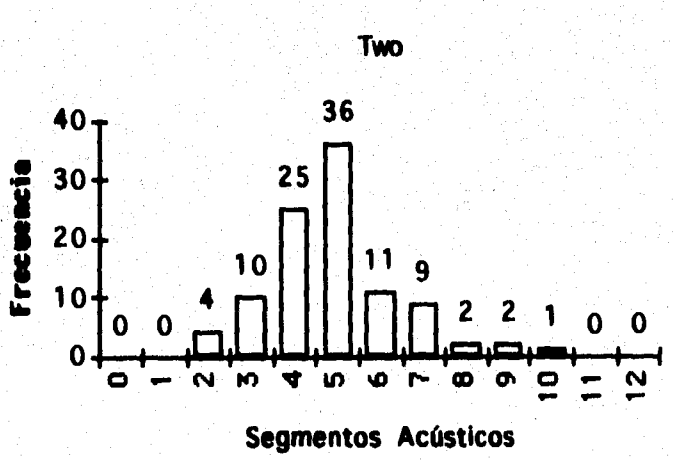
Aquí cabe aclarar un punto muy importante que se debe de tomar en cuenta cuando se interpretan los resultados, el entrenamiento de sistemas de reconocimiento involucra un gran número de palabras de entrenamiento. En este sistema se tiene un número que equivale al que se utilizaría en otros sistemas con la diferencia que la segmentación acústica da un número de palabras de entrenamiento diferente dependiendo de la segmentación acústica. Esto implica que para algunas palabras la segmentación sea consistente, pero en general tiende a tener una distribución normal, de aquí que una palabra tiende a ser segmentada en un número S de segmentos, pero también es segmentada en S-1 y S+1 segmentos y así sucesivamente. Esto provoca que el número de palabras de entrenamiento varíe grandemente, habiendo segmentos para los cuales no existen vectores para realizar el entrenamiento.



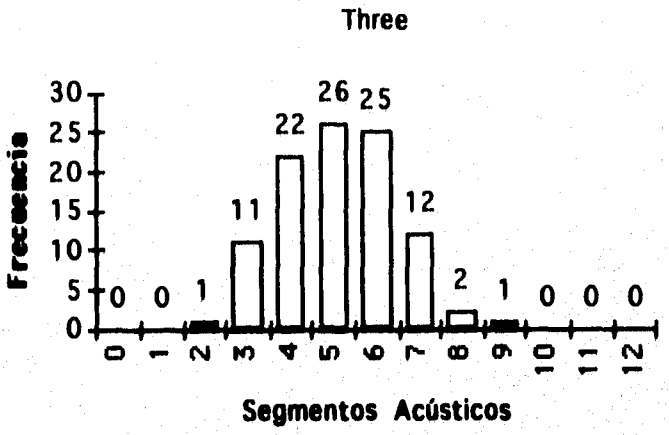
(a) Palabra "Zero"



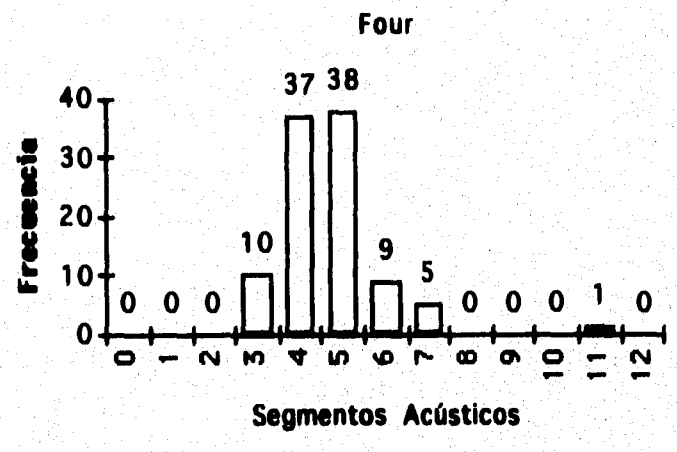
(b) Palabra "One"



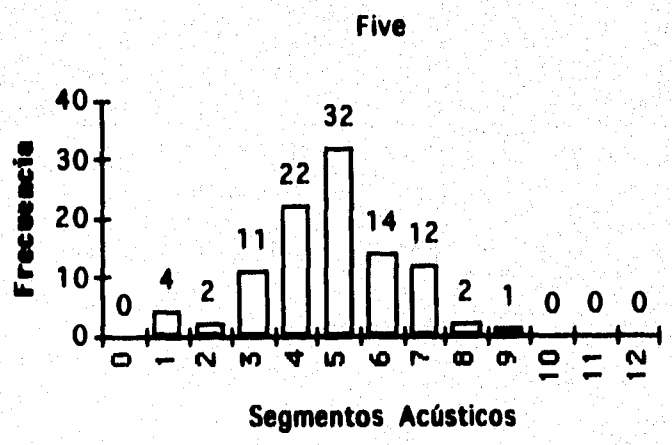
(c) Palabra "Two"



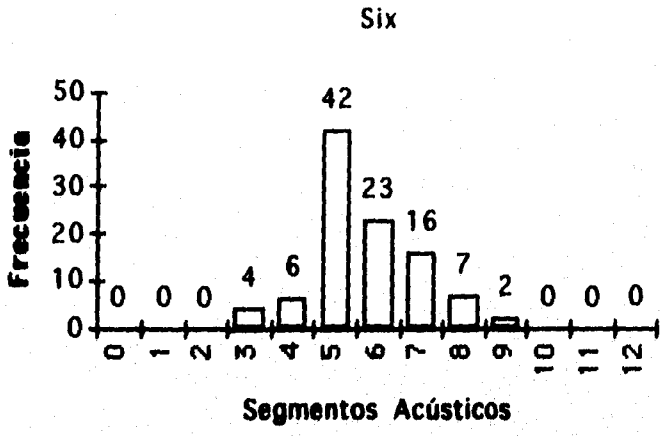
(d) Palabra "Three"



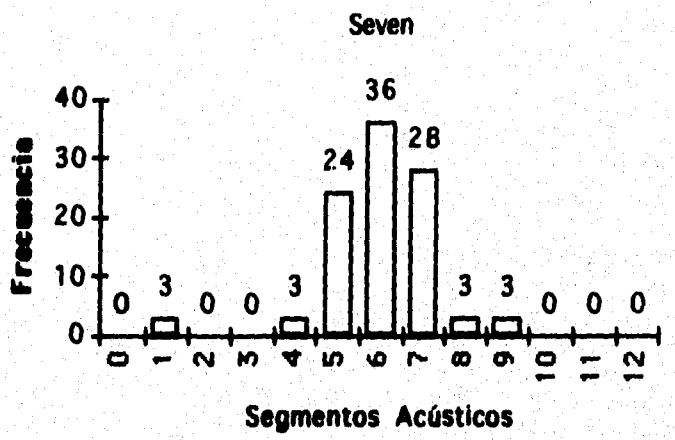
(e) Palabra "Four"



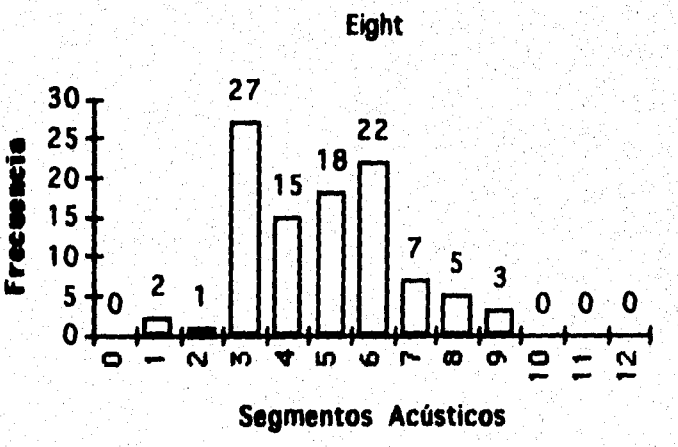
(f) Palabra "Five"



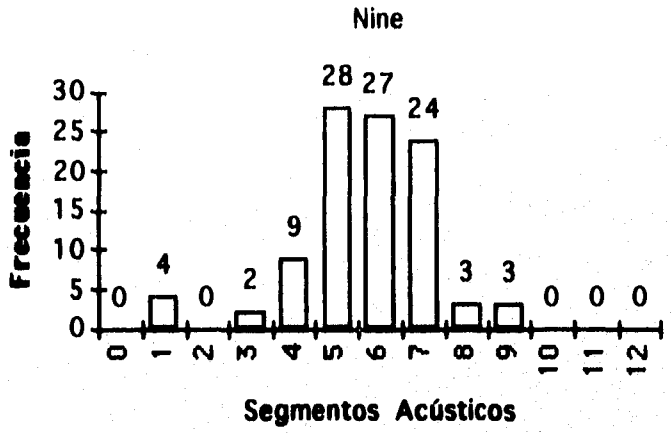
(g) Palabra "Six"



(h) Palabra "Seven"



(i) Palabra "Eight"



(j) Palabra "Nine"

Figura 10.1 Histogramas de segmentación acústica por dígito.

Ademas la comparación en el reconocimiento se realiza de forma tanto cuanto injusta ya que se puede comparar con palabras para las cuales existieron muchos vectores de entrenamiento y otras para las cuales no existió la suficiente información. A pesar de esto los niveles de reconocimiento para números de segmentación entre cuatro, cinco y seis segmentos superan el 90% en promedio, siendo muchos de estos del 100% por cada dígito.

Se sugiere la modificación del algoritmo de segmentación acústica para que genere un número preestablecido de segmentos acústicos y con esto en teoría se elevaría el promedio de reconocimiento a más del 95%, ya que existirían los suficientes vectores para entrenar el sistema.

11. Resultados y Conclusiones

11.1 Resultados

Los resultados que se presentan se basan en la combinación de técnicas de procesamiento digital de señales, segmentación acústica, cuantización vectorial, estimación paramétrica, modelado de señales, agrupamiento difuso, calculo de distancias multidimensionales, entre las más sobresalientes.

Por lo tanto la aportación de este trabajo es la asociación de conceptos de diferentes disciplinas para trabajar en conjunto y de esta forma aportar técnicas nuevas que no habían sido probadas en la literatura conocida.

El resultado fue un nivel de reconocimiento bastante aceptable dentro de los niveles estándar del reconocimiento que son superiores al 90% para sistemas independientes de locutor. Hay que tomar en cuenta como se mencionó en el capítulo anterior, que la técnica de segmentación acústica no es uniforme, es decir que el mismo dígito dicho por el mismo locutor en diferentes versiones se segmenta en diferente número para cada repetición. Esto provoca que el entrenamiento no sea consistente y por lo tanto el número de palabras para entrenar, dado un número de segmentos acústicos sea diferente dependiendo del dígito. Esto se puede observar en las tablas que contienen las matrices de confusión (11.1,11.2, 11.3 y 11.4) donde la columna NR nos da el número de repeticiones de cada dígito dentro de ese número de segmentos acústicos y TC nos da el número total de comparaciones que se realizaron para todas las repeticiones de ese dígito.

Idealmente se deberían de tener 100 repeticiones si la segmentación fuera constante. Esto significa que las comparaciones se realizan con centroides que fueron generados con diferente número de palabras de entrenamiento, entonces podemos comparar dígitos para los cuales se tuvieron una o dos palabras en el entrenamiento contra otras que pudieron haber más de treinta.

A pesar de estos inconvenientes, el sistema en general reconoce para ambas técnicas de agrupamiento (K-Medias y C-Medias Difuso) niveles superiores al 90%. Esto significa que con el número adecuado de segmentos acústicos, este número en teoría se debe de aumentar considerablemente.

Las tablas 11.1 a 11.4 muestran las matrices de confusión para la siguiente combinación de técnicas, donde el primer término se refiere a la técnica de agrupamiento, la segunda a la forma de parametrizar la señal y la tercera a la medida de comparación entre patrones:

K-Medias, LPC, Itakura-Saito Modificada

- Centroides-Entrenamiento	Palabras-Entrenamiento	(KM-LPCE)
- Centroides-Entrenamiento	Palabras-Reconocimiento	(KM-LPCR)

C-Medias Difuso, Cepstrum, Euclidiana

- Centroides-Entrenamiento	Palabras-Entrenamiento	(CM-CEPE)
- Centroides-Entrenamiento	Palabras-Reconocimiento	(CM-CEPR)

Para ambas técnicas, se generarán a partir de una serie 1000 repeticiones de los dígitos, 16 centroides (patrones) por cada segmento acústico.

11.1.1 Análisis con Agrupamiento No Difuso

Cuando se utilizan las mismas palabras en el entrenamiento y en el reconocimiento, se tienen niveles promedio de reconocimiento por número de segmentos acústicos que van de 90.79% llegando hasta el 100%. El valor para tres segmentos acústicos (KM-CEPE) de 77.26% en realidad no se toma en cuenta, ya que no existieron palabras de entrenamiento para los dígitos "zero" y "seven" y tienen un 10% de peso sobre el nivel de reconocimiento. Siendo el total promediado de 93.46%.

Cuando se utilizan las 1600 repeticiones ajenas a la fase de entrenamiento, los niveles alcanzados como máximo llegaron a 92.71% en promedio por dígito con cinco segmentos acústicos. Cabe notar que para cualquier número de segmentos acústicos, existen dígitos que se reconocen en un 100% y el número promedio baja considerablemente debido a la pequeña cantidad de dígitos de entrenamiento que se tienen. El total promedio es de 91.82%.

11.1.2 Análisis con Agrupamiento Difuso

Ademas de obtener resultados equivalentes al de K-Medias, se logró que estos niveles fueran superiores en un mínimo de 2% con respecto a su contraparte de K-Medias en el caso de reconocimiento de las palabras de entrenamiento, ademas resulta interesante el hecho que para CM-CEPE y seis segmentos se logre un nivel del 100%. El que se mencione que es interesante es el hecho que en general las palabras tienen como promedio de segmentación cuatro, cinco y seis segmentos, de acuerdo a los histogramas que muestran la segmentación por dígito mostrados en el capítulo anterior.

Para los resultados de CM-CEPR, existe un decremento en el nivel de reconocimiento con respecto a KM-LPCR, pero existe un factor que no se tomó en cuenta cuando se hizo el entrenamiento. El factor de potencia de la señal se incluyó en el análisis cepstral y en el de LPC no se tomó en cuenta. Esto hace que los patrones sean altamente dependientes de la energía total de la palabra. Se están haciendo pruebas para entrenar el sistema nuevamente, pero sin éste factor y aparentemente se esta mejorando el desempeño. Esto se comprueba mediante algunos resultados parciales que se han obtenido, que no se incluyen en éste trabajo.

En general vemos un incremento en el porcentaje de reconocimiento, esto nos indica que la técnica tiene un desempeño bastante bueno y prueba que la lógica difusa se puede aplicar al reconocimiento de voz en una forma equivalente a las técnicas convencionales de reconocimiento.

Otro factor importante es que se puede analizar si una palabra tiende a ser segmentada en forma natural en ciertos número de segmentos acústicos. Es decir hay dígitos para los cuales se requiere un número grande de segmentos y otros para los que basta con pocos segmentos.

Por lo tanto tenemos un compromiso entre el número de segmentos acústicos y el nivel de reconocimiento. Ademas cabe notar que la comparación de segmentos no es dependiente de la longitud, un segmento de un dígito puede contener un numero N de muestras y otro puede contener M, que puede se mayor, menor o igual y la comparación se realiza en función al segmento como unidad.

Esto permite que se puedan utilizar palabras de cualquier tamaño y no haya que utilizar técnicas de Ajuste Dinámico en Tiempo (DTW) [RABI93][DELL87][PARS87] antes de la comparación de los patrones.

La información que contienen las matrices de confusión se puede explicar como sigue:

El número tres, cuatro, etc, nos indica el número de segmentos acústicos en los cuales fue dividida la palabra.

Los números tanto en el renglón superior como la primera columna denotan las comparaciones. Se toma la primera columna de algún número de segmentos; digamos cuatro segmentos y comparamos el dígito "five" denotado por el renglon "5" en KM-LPCE, entonces vemos que todas las palabras "five" a reconocer, que en este caso son 22 como se puede observar en la columna NR (Número de repeticiones), se van a comparar contra los vectores de entrenamiento de todos los dígitos (0,1,..9). Vemos que al compararlo con "0" nunca se equivocó, pero al compararlo con "3", en una ocasión confundio un "five" por un "three".

Los porcentajes son evaluados tomando el número de aciertos y dividiendolo entre el número de repeticiones (NR). Como se había mencionado al principio del capítulo, TC es el número total de comparaciones para determinar el reconocimiento de un dígito en particular contra todos los demas dígitos generados en el entrenamiento. Normalmente estos son vectores que en realidad son los patrones de las características obtenidas.

Tres	0	1	2	3	4	5	6	7	8	9	TC	NR	
0	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
1	0	8	0	0	0	0	0	0	0	0	64	8	100.00%
2	0	0	9	0	0	1	0	0	0	0	80	10	90.00%
3	0	0	0	11	0	0	0	0	0	0	88	11	100.00%
4	0	0	0	0	9	1	0	0	0	0	80	10	90.00%
5	0	0	0	0	0	11	0	0	0	0	88	11	100.00%
6	0	0	0	0	0	0	4	0	0	0	32	4	100.00%
7	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
8	0	0	0	0	0	0	2	0	25	0	216	27	92.59%
9	0	0	0	0	0	0	0	0	0	2	16	2	100.00%
													77.26%

Cuatro	0	1	2	3	4	5	6	7	8	9	TC	NR	
0	10	0	2	0	0	0	0	0	0	0	120	12	83.33%
1	0	24	0	0	0	0	0	0	0	0	240	24	100.00%
2	0	0	25	0	0	0	0	0	0	0	250	25	100.00%
3	0	0	0	22	0	0	0	0	0	0	220	22	100.00%
4	0	0	1	0	34	1	0	0	0	1	370	37	91.89%
5	0	0	0	1	0	21	0	0	0	0	220	22	95.45%
6	0	0	0	0	0	0	6	0	0	0	60	6	100.00%
7	0	0	0	0	0	0	0	3	0	0	30	3	100.00%
8	0	0	0	0	0	0	0	0	15	0	150	15	100.00%
9	0	0	0	0	0	0	0	0	0	9	90	9	100.00%
													97.07%

Cinco	0	1	2	3	4	5	6	7	8	9	TC	NR	
0	19	0	2	0	0	0	0	0	0	1	220	22	86.36%
1	0	25	0	0	0	0	0	0	0	0	250	25	100.00%
2	0	0	32	0	0	3	0	1	0	0	360	36	88.89%
3	0	0	0	23	0	1	0	0	0	2	260	26	88.46%
4	1	1	1	0	29	6	0	0	0	0	380	38	76.32%
5	0	0	0	0	0	31	0	1	0	0	320	32	96.88%
6	0	0	0	1	0	0	41	0	0	0	420	42	97.62%
7	0	0	0	0	0	2	0	22	0	0	240	24	91.67%
8	0	0	0	1	0	0	1	0	16	0	180	18	88.89%
9	0	0	1	0	1	0	0	0	0	26	280	28	92.86%
													90.79%

Seis	0	1	2	3	4	5	6	7	8	9	TC	NR	
0	25	0	1	0	1	0	0	0	0	0	270	27	92.59%
1	0	23	0	0	0	0	0	0	0	0	230	23	100.00%
2	1	0	10	0	0	0	0	0	0	0	110	11	90.91%
3	0	0	0	23	0	0	0	1	0	1	250	25	92.00%
4	0	1	0	0	8	0	0	0	0	0	90	9	88.89%
5	0	1	0	0	0	12	0	0	0	1	140	14	85.71%
6	1	0	0	2	0	0	20	0	0	0	230	23	86.96%
7	0	0	0	0	0	0	0	35	0	1	360	36	97.22%
8	0	0	0	0	0	0	1	0	21	0	220	22	95.45%
9	0	0	0	0	0	1	0	0	0	26	270	27	96.30%
													92.60%

Total	0	1	2	3	4	5	6	7	8	9	TC	NR	
0	54	0	5	0	1	0	0	0	0	1	610	61	88.52%
1	0	80	0	0	0	0	0	0	0	0	784	80	100.00%
2	1	0	76	0	0	4	0	1	0	0	800	82	92.68%
3	0	0	0	79	0	1	0	1	0	3	818	84	94.05%
4	1	2	2	0	80	8	0	0	0	1	920	94	85.11%
5	0	1	0	1	0	75	0	1	0	1	768	79	94.94%
6	1	0	0	3	0	0	71	0	0	0	742	75	94.67%
7	0	0	0	0	0	2	0	60	0	1	630	63	95.24%
8	0	0	0	1	0	0	4	0	77	0	766	82	93.90%
9	0	0	1	0	1	1	0	0	0	63	656	66	95.45%
													93.46%

Tabla 11.1 Matriz de Confusión KM-LPCE

Tres	0	1	2	3	4	5	6	7	8	9	TC	NR	
0	0	0	3	0	0	0	0	0	0	0	24	3	0.00%
1	0	6	0	0	0	0	0	0	0	1	56	7	85.71%
2	0	0	5	0	1	0	0	0	0	1	56	7	71.43%
3	0	0	0	12	0	0	0	0	0	0	96	12	100.00%
4	0	2	0	0	42	2	0	0	0	0	368	46	91.30%
5	0	1	0	0	1	12	0	0	0	3	136	17	70.59%
6	0	0	0	0	0	0	2	0	0	0	16	2	100.00%
7	0	0	1	0	0	1	0	0	0	0	16	2	0.00%
8	0	0	0	0	0	0	0	0	59	0	472	59	100.00%
9	0	6	0	0	0	0	0	0	0	0	48	6	0.00%
													61.90%

Cuatro	0	1	2	3	4	5	6	7	8	9	TC	NR	
0	16	0	3	0	0	0	0	0	0	0	190	19	84.21%
1	0	26	0	0	0	0	0	0	0	1	270	27	96.30%
2	1	0	34	0	0	0	0	0	0	0	350	35	97.14%
3	0	0	0	19	0	0	0	0	0	0	190	19	100.00%
4	0	1	0	0	55	0	0	0	0	0	560	56	98.21%
5	0	0	0	0	0	30	0	0	0	0	300	30	100.00%
6	0	0	0	0	0	0	33	0	0	0	330	33	100.00%
7	0	1	2	0	0	0	0	4	0	0	70	7	57.14%
8	0	0	0	0	0	0	0	0	22	0	220	22	100.00%
9	1	1	1	0	0	1	0	0	0	16	200	20	80.00%
													91.30%

Cinco	0	1	2	3	4	5	6	7	8	9	TC	NR	
0	24	0	2	0	0	0	0	1	0	0	270	27	88.89%
1	0	40	0	0	0	1	0	0	0	3	440	44	90.91%
2	3	0	45	0	0	1	0	0	0	0	490	49	91.84%
3	0	0	0	52	0	0	0	2	0	0	540	54	96.30%
4	0	1	0	0	46	2	0	0	0	0	490	49	93.88%
5	0	0	0	0	0	40	0	1	0	0	410	41	97.56%
6	0	0	0	0	0	0	50	0	0	0	500	50	100.00%
7	0	0	0	0	0	0	0	29	0	0	290	29	100.00%
8	0	0	0	0	0	0	6	0	28	0	340	34	82.35%
9	0	1	2	0	0	2	0	0	0	29	340	34	85.29%
													92.70%

Seis	0	1	2	3	4	5	6	7	8	9	TC	NR	
0	42	0	0	0	0	0	1	0	0	0	430	43	97.67%
1	0	40	0	0	0	0	0	0	0	4	440	44	90.91%
2	5	0	28	0	0	0	0	2	0	0	350	35	80.00%
3	0	0	0	39	0	0	0	1	0	0	400	40	97.50%
4	0	0	0	0	18	0	0	0	0	0	180	18	100.00%
5	0	1	0	0	0	35	0	0	0	2	360	38	92.11%
6	0	0	0	0	0	3	40	0	0	0	430	43	93.02%
7	0	0	0	0	0	0	1	52	0	0	530	53	98.11%
8	0	0	0	0	0	0	0	0	20	0	200	20	100.00%
9	0	5	0	0	0	0	0	3	0	28	360	36	77.78%
													92.71%

Total	0	1	2	3	4	5	6	7	8	9	TC	NR	
0	82	0	8	0	0	0	1	1	0	0	914	92	89.13%
1	0	112	0	0	0	1	0	0	0	9	1206	122	91.80%
2	9	0	112	0	1	1	0	2	0	1	1246	126	88.89%
3	0	0	0	122	0	0	0	3	0	0	1226	125	97.60%
4	0	4	0	0	161	4	0	0	0	0	1598	169	95.27%
5	0	2	0	0	1	117	0	1	0	5	1226	126	92.86%
6	0	0	0	0	0	3	125	0	0	0	1276	128	97.66%
7	0	1	3	0	0	1	1	85	0	0	906	91	93.41%
8	0	0	0	0	0	0	6	0	129	0	1232	135	95.56%
9	1	13	3	0	0	3	0	3	0	73	948	96	76.04%
													91.82%

Tabla 11.2 Matriz de Confusión KM-LPCR

Tres	0	1	2	3	4	5	6	7	8	9	TC	NR	
0	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
1	0	8	0	0	0	0	0	0	0	0	64	8	100.00%
2	0	0	10	0	0	0	0	0	0	0	80	10	100.00%
3	0	0	0	11	0	0	0	0	0	0	88	11	100.00%
4	0	0	0	0	10	0	0	0	0	0	80	10	100.00%
5	0	0	0	0	0	11	0	0	0	0	88	11	100.00%
6	0	0	0	0	0	0	4	0	0	0	32	4	100.00%
7	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
8	0	0	0	0	0	0	1	0	26	0	216	27	96.30%
9	0	0	0	0	0	0	0	0	0	2	16	2	100.00%
													79.63%

Cuatro	0	1	2	3	4	5	6	7	8	9	TC	NR	
0	12	0	0	0	0	0	0	0	0	0	120	12	100.00%
1	0	23	0	0	0	0	0	0	0	1	240	24	95.83%
2	0	0	25	0	0	0	0	0	0	0	250	25	100.00%
3	0	0	0	22	0	0	0	0	0	0	220	22	100.00%
4	0	0	0	0	37	0	0	0	0	0	370	37	100.00%
5	0	0	0	0	0	22	0	0	0	0	220	22	100.00%
6	0	0	0	0	0	0	6	0	0	0	60	6	100.00%
7	0	0	0	0	0	0	0	3	0	0	30	3	100.00%
8	0	0	0	0	0	0	0	0	15	0	150	15	100.00%
9	0	0	0	0	0	0	0	0	0	9	90	9	100.00%
													99.58%

Cinco	0	1	2	3	4	5	6	7	8	9	TC	NR	
0	22	0	0	0	0	0	0	0	0	0	220	22	100.00%
1	0	25	0	0	0	0	0	0	0	0	250	25	100.00%
2	0	0	36	0	0	0	0	0	0	0	360	36	100.00%
3	0	0	0	26	0	0	0	0	0	0	260	26	100.00%
4	0	0	0	0	37	1	0	0	0	0	380	38	97.37%
5	0	0	0	0	0	32	0	0	0	0	320	32	100.00%
6	0	0	0	0	0	0	39	0	3	0	420	42	92.86%
7	0	0	0	0	0	0	0	24	0	0	240	24	100.00%
8	0	0	0	1	0	0	0	0	17	0	180	18	94.44%
9	0	2	1	0	0	0	0	0	0	25	280	28	89.29%
													97.40%

Seis	0	1	2	3	4	5	6	7	8	9	TC	NR	
0	27	0	0	0	0	0	0	0	0	0	270	27	100.00%
1	0	23	0	0	0	0	0	0	0	0	230	23	100.00%
2	0	0	11	0	0	0	0	0	0	0	110	11	100.00%
3	0	0	0	25	0	0	0	0	0	0	250	25	100.00%
4	0	0	0	0	9	0	0	0	0	0	90	9	100.00%
5	0	0	0	0	0	14	0	0	0	0	140	14	100.00%
6	0	0	0	0	0	0	23	0	0	0	230	23	100.00%
7	0	0	0	0	0	0	0	36	0	0	360	36	100.00%
8	0	0	0	0	0	0	0	0	22	0	220	22	100.00%
9	0	0	0	0	0	0	0	0	0	27	270	27	100.00%
													100.00%

Total	0	1	2	3	4	5	6	7	8	9	TC	NR	
0	61	0	0	0	0	0	0	0	0	0	610	61	100.00%
1	0	79	0	0	0	0	0	0	0	1	784	80	98.75%
2	0	0	82	0	0	0	0	0	0	0	800	82	100.00%
3	0	0	0	84	0	0	0	0	0	0	818	84	100.00%
4	0	0	0	0	93	1	0	0	0	0	920	94	98.94%
5	0	0	0	0	0	79	0	0	0	0	768	79	100.00%
6	0	0	0	0	0	0	72	0	3	0	742	75	96.00%
7	0	0	0	0	0	0	0	63	0	0	630	63	100.00%
8	0	0	0	1	0	0	1	0	80	0	766	82	97.56%
9	0	2	1	0	0	0	0	0	0	63	656	66	95.45%
													98.67%

Tabla 11.3 Matriz de Confusión CM-CEPE

Tres	0	1	2	3	4	5	6	7	8	9	TC	NR	
0	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
1	0	6	0	0	0	0	0	0	0	1	56	7	85.71%
2	0	0	4	2	0	1	0	0	0	0	56	7	57.14%
3	0	0	0	11	0	1	0	0	0	0	96	12	91.67%
4	0	0	4	8	7	4	0	0	0	0	184	23	30.43%
5	0	0	0	0	0	16	0	0	0	1	136	17	94.12%
6	0	0	0	0	0	0	2	0	0	0	16	2	100.00%
7	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
8	0	0	0	0	0	0	5	0	54	0	472	59	91.53%
9	0	4	1	0	0	0	0	0	0	1	48	6	16.67%
													56.73%

Cuatro	0	1	2	3	4	5	6	7	8	9	TC	NR	
0	16	0	1	0	0	0	0	2	0	0	190	19	84.21%
1	0	16	0	0	0	0	0	1	0	10	270	27	59.26%
2	0	0	35	0	0	0	0	0	0	0	350	35	100.00%
3	0	0	0	19	0	0	0	0	0	0	190	19	100.00%
4	0	0	0	0	55	1	0	0	0	0	560	56	98.21%
5	0	0	0	0	0	28	0	0	0	2	300	30	93.33%
6	0	0	0	0	0	0	28	0	5	0	330	33	84.85%
7	0	0	1	0	0	0	0	5	0	1	70	7	71.43%
8	0	0	0	0	0	0	0	0	22	0	220	22	100.00%
9	0	0	1	0	0	0	0	0	0	19	200	20	95.00%
													88.63%

Cinco	0	1	2	3	4	5	6	7	8	9	TC	NR	
0	20	0	2	0	0	0	0	5	0	0	270	27	74.07%
1	0	43	0	0	0	1	0	0	0	0	440	44	97.73%
2	1	0	42	0	0	1	0	5	0	0	490	49	85.71%
3	0	0	0	53	0	1	0	0	0	0	540	54	98.15%
4	7	0	0	0	35	0	0	7	0	0	490	49	71.43%
5	0	0	0	0	0	41	0	0	0	0	410	41	100.00%
6	0	0	0	0	0	0	39	0	11	0	500	50	78.00%
7	0	0	0	0	0	0	0	29	0	0	290	29	100.00%
8	0	0	0	0	0	0	0	0	34	0	340	34	100.00%
9	0	6	0	0	0	1	0	2	0	25	340	34	73.53%
													87.86%

Seis	0	1	2	3	4	5	6	7	8	9	TC	NR	
0	40	0	3	0	0	0	0	0	0	0	430	43	93.02%
1	0	33	0	0	1	0	0	0	0	10	440	44	75.00%
2	1	0	32	0	0	0	0	2	0	0	350	35	91.43%
3	0	0	0	40	0	0	0	0	0	0	400	40	100.00%
4	0	0	0	0	18	0	0	0	0	0	180	18	100.00%
5	0	0	0	0	0	37	1	0	0	0	380	38	97.37%
6	0	0	0	0	0	0	43	0	0	0	430	43	100.00%
7	0	0	4	0	0	0	0	49	0	0	530	53	92.45%
8	0	0	0	0	0	0	1	0	19	0	200	20	95.00%
9	0	4	0	1	0	4	1	2	0	24	360	36	66.67%
													91.09%

Total	0	1	2	3	4	5	6	7	8	9	TC	NR	
0	76	0	6	0	0	0	0	7	0	0	890	89	85.39%
1	0	98	0	0	1	1	0	1	0	21	1206	122	80.33%
2	2	0	113	2	0	2	0	7	0	0	1246	126	89.68%
3	0	0	0	123	0	2	0	0	0	0	1226	125	98.40%
4	7	0	4	8	115	5	0	7	0	0	1414	146	78.77%
5	0	0	0	0	0	122	1	0	0	3	1226	126	96.83%
6	0	0	0	0	0	0	112	0	16	0	1276	128	87.50%
7	0	0	5	0	0	0	0	83	0	1	890	89	93.26%
8	0	0	0	0	0	0	6	0	129	0	1232	135	95.56%
9	0	14	2	1	0	5	1	4	0	69	948	96	71.88%
													87.76%

Tabla 11.4 Matriz de Confusión CM-CEPR

KM-LPCE	KM-LPCR	CM-CEPE	CM-CEPR
92.59%	97.67%	100.00%	93.02%
100.00%	96.30%	100.00%	97.73%
100.00%	97.14%	100.00%	100.00%
100.00%	100.00%	100.00%	100.00%
91.89%	100.00%	100.00%	100.00%
100.00%	100.00%	100.00%	100.00%
100.00%	100.00%	100.00%	100.00%
100.00%	100.00%	100.00%	100.00%
100.00%	100.00%	100.00%	100.00%
100.00%	85.29%	100.00%	95.00%
98.45%	**97.64%**	**100.00%**	**98.58%**

Tabla 11.5 Tabla de valores de reconocimiento máximos.

11.2 Conclusiones

La tabla 11.5 nos muestra los niveles máximos de reconocimiento obtenidos para cualquier segmento. De aqui se puede observar que existen palabras en algún número de segmentos acústicos, para las cuales la segmentación es óptima y por lo tanto el grado de reconocimiento con respecto a otras palabras es en promedio superior al 97.64% en el peor caso.De aquí suponemos que a mayor concentración de palabras de entrenamiento, mejor desempeño se logrará.

Como se mencionó anteriormente, se probó que la segmentación en subpalabras acústicas es una técnica bastante eficaz que nos permite implantar las técnicas de reconocimiento con las siguientes características:

- Independencia del locutor y su sexo
- Independiente del número de muestras en la palabra
- Utilización eficiente de técnicas de agrupamiento

Se sugiere por un lado segmentar a un tamaño fijo y no dejar la libertad de buscar el número apropiado. Esto con el fin de que el porcentaje de reconocimiento mejore en forma sustantiva. Además se podrían utilizar otras técnicas como los Modelos Ocultos de Markov y comparar los resultados obtenidos.

Este trabajo de tesis requirió como base, conceptos de varias disciplinas y por lo tanto los nueve primeros capítulos aportan las bases para la elaboración de éste. Aunque se podrían haber dejado referencias a todos estos conceptos, siento que es importante el darle al lector la oportunidad de entender el trabajo si necesidad de recurrir a múltiples fuentes de información que harían más lento el entendimiento de este texto.

La lógica difusa, no es una disciplina que actualmente se estudie formalmente y los grupos de trabajo en ésta disciplina se consideran separados de los grupos involucrados en el procesamiento de señales determinísticas y estocásticas. Por lo que ésta fue una buena oportunidad de unir estos conceptos y generar una aplicación con un fin común y que partiera de la base de procesamiento de señales clásico hasta llegar a conceptos nuevos como la segmentación en subpalabras acústicas y la agrupación difusa.

Aunque no se menciona en este texto, se escribieron programas de procesamiento de las señales en Lenguaje C que permiten que se puedan seguir investigando estas técnicas y la combinación con

otras. En general se tiene un conjunto de herramientas útiles en el procesamiento y análisis de las señales de voz que queda a disposición de la comunidad universitaria.

Otros conceptos como el uso de las Bandas Criticas (usadas en la segmentación acústica), los Modelos Ocultos de Markov y la transformada de Karhunen-Loeve, que en un principio se habían planteado como parte de la investigación de este trabajo de tesis, quedaron fuera por cuestión de tiempo y de volumen de trabajo, aunque se realizó un trabajo importante con estos tres conceptos que no se refleja en este texto. De aquí se sugiere que se continúe esta investigación y se combinen estos conceptos con los aquí desarrollados para generar técnicas de reconocimiento más eficaces y que quizá permitan en un futuro, lograr la meta que es la de tener un sistema automático de reconocimiento que permita establecer una comunicación fluida con la computadora.

Bibliografía

Procesamiento Digital de Señales

[OPPE75] A. Oppenheim and R. Schafer,
Discrete Time Signal Processing,
Englewood Cliffs, NJ, Prentice Hall 1975

[IFFE93] E. Ifeachor and B. Jervis,
Digital Signal Processing a Practical Aproach,
Addison-Wesley, 1993

[PROA92] J. Proakis and D. Manolakis,
Digital Signal Processing, Principles, Algorithms and Applications,
Macmillan Publishing, 1992 USA

[TERR92] C. Terrien,
Discrete Random Signal and Statistical Signal Processing,
Prentice-Hall, 1992 USA

Probabilidad y Procesos Estocásticos

[PAPO91] A. Papoulis,
Probability, Random Variables and Stochastic Processes,
McGraw Hill, 1991 USA

[PEEB93] P. Peebles,
Probability, Random Variables and Random Signal Principles,
McGraw-Hill, 1993,USA

Procesamiento y Reconocimiento Voz

[PARS87] Parsons, Thomas W.,
Voice and Speech Processing,
McGraw-Hill, 1987, USA

[RABI87] Rabiner, Lawrence.,
Digital Processing of Speech Signals,
Prentice-Hall, 1978, USA

[PAPA87] P. Papamichalis,
Practical Aproaches to Speech Coding,
Englewood Cliffs, NJ, Prentice-Hall, 1987

[DELL87] Deller, John,et. al.,
Discrete-Time Processing of Speech Signals,
Prentice-Hall, 1987, USA

[RABI93] Rabiner, Lawrence, et.al.,
Fundamentals of Speech Recognition,
Prentice-Hall, 1993, USA

[HERR94] A. Herrera,
Apuntes de Procesamiento Digital de Voz,
DEPFI-UNAM, 1994, México

[ROBI96] T. Robinson,
Speech Analysis, Computer Speech and Language Processing, Course Notes
Cambridge University 1996 UK

[PICO93] J. Picone,
Signal Modelling Techniques in Speech Recognition,
Proceedings of the IEEE, Vol. 81, No. 9, September 1993

[FURU92] S. Furui and M. Sondhi,
Advances in Speech Signal Processing,
Marcel Dekker Inc., 1992 USA

[RABI89] L. Rabiner,
A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,
Proceedings IEEE, Vol. 77, No 2, pp. 257-285, Feb 1989

[FAUS96] J. Faust,
Low Cost Product Design with Speech Recognition,
Proceeedings DSPx96, San Jose CA, 1996

[MOZE96] T. Mozer,
Introduction and Overview of Low Cost Speech Recognition Technologies,
Proceeedings DSPx96, San Jose CA, 1996

[SCHA70] B. Scharf,
Critical Bands ,in Foundations of Modern Auditory,
Academic Press, Vol. I J. Tobias Ed., 1970

[JEFF70] Jeffress, LLoyd.,
Masking ,in Foundations of Modern Auditory,
Academic Press, Vol. I J. Tobias Ed., 1970

[KENT92] R. Kent and C. Read,
The Acoustic Analysis of Speech,
Singular Publishing Group Inc. 1992 USA

Segmentación Acústica

[HERR194] A. Herrera, R. Algazi, K. Brown and D. Irvine,
Subword Segmentation Alternatives for Isolated and Connected Words Recognition,
Proceedings , VII European Signal Processing Conference EUPSICO-94

[HERR294] A. Herrera, V.R. Algazi and D. Irvine,
An Acoustic Approach for Isolated Speech Recognition,
Proceedings of the International Conference on Signal Processing Applications and Technology, ICSPAT 94, Vol 2, 1677-1681

Reconocimiento de Patrones

[TOU81] J. Tou and R. Gonzalez
Pattern Recognition Principles,
Addison-Wesley, 1981, USA

Cuantización Vectorial

[ABUT90] H. Abut,
Vector Quatization,
IEEE Press 1990, USA

[MAKH85] J. Makhoul, S. Roucos and H. Gish,
Vector Quantization in Speech Coding,
Proceedings of the IEEE, Vol 73, No13, November 1985

[LIND80] Y, Linde, A. Buzo and R. Gray,
An algorithm for Vector Quantizer Design,
IEEE Trans. on Communications, Vol. COM-28, pp 84-95, January 1980

[GRAY81] R. Gray, A. Gray, G Rebolledo and J. Shore,
Rate-Distortion Speech Coding with a Minimum Discrimination Information Distorsion Measure,
IEEE Trans. on Information Theory, Vol. IT-27, pp 708-721, Nov. 1981

[ABUT82] H. Abut, R. Gray and G. Rebolledo,
Vector Quantization of Speech and Speech Like Waveforms,
IEEE Trans. on Acoustics, Speech Signal Processing, Vol. ASSP-30, pp 423-435, June 1982

[BUZO82] A. Buzo, H. Martinez, C. Rivera,
Discrete Utterance Recognition Based Upon Source Coding Techniques,
IEEE Conf. Acoust., Speech Signal Processing, Vol. 1 pp. 539-542, May 1982

Lógica Difusa

[ZADE65] L. Zadeh,
Fuzzy Sets,
Inform. Control, Vol.8, pp. 338-353, 1965

[KOSK93] B. Kosko
Fuzzy Thinking: The New Science of Fuzzy Logic.
Hyperio, New York, 1993

[KOSK92] B. Kosko
Neural Networks and Fuzzy Systems: A Dynamical Systems

Approach to Machine Intelligence.
Prentice-Hall, New York, 1992

[BEZD92] J. Bezdek and S. Pal,
Fuzzy Models for Pattern Recognition,
Methods that Search for Structures in Data,
IEEE Press, 1992 USA

[ZIMM90] H. Zimmermann,
Fuzzy Set Theory and its Applications,
Kluwer Academic Publishers, Boston/Dordrecht/London 1990

[CANN86] R. Cannon, J.Dave and J. Bezdek,
Efficient implementation of the Fuzzy C-Means Clustering Algorithms,
IEEE Transactions on Pattern Analysis and Machine Intelligence,
Vol PAMI-8, No 2, March 1986

[WIND83] M. Windham,
Geometrical Fuzzy Clustering Algorithms,
Fuzzy Sets and Systems, Vol 10, pp 271-279, 1983

[KOSA95] B. Kosanovic,
Signal and System Analysis in Fuzzy Information Space,
PhD. Dissertation, University of Pittsburgh 1995

Fonética

[QUIL79] A. Quilis, J. Fernández,
Curso de Fonética y Fonología Españolas para estudiantes angloamericanos ,
C.S.I.C, Madrid 1979

[BORZ80] A. Borzone,
Manual de Fonética Acústica,
Hachete, Argentina 1980

Anatomía

[DIEN76] C. Dienhart,
Anatomía y Fisiología Humanas,
Interamericana, 1976, México

[GANA78] W. Ganang,
Manual de Fisiología Médica,
El Manual Moderno, 1978, México

Física

[SEAR88] F. Sears, M. Zemansky, H. Young,
Física Universitaria,
Addison-Wesley Iberoamericana, 1988, México

Apéndice A

A.1 Descripción de los programas utilizados

Programa	Plataforma y Lenguaje	Descripción
SCALE_SP	Sun-Fortran	Escalamiento de la señal a un valor máximo de 16000.
CHEBY_N	Sun-Fortran	Filtro pasabajas Chebyshev de 8vo. orden Fc=4500Hz.
RESAMP	Sun-Fortran	Conversión de tasa de muestreo de 12.5KHz a 10KHz.
SUBWORD	Sun-Fortran	Segmentación en subpalabras acústicas.
LPCITA	PC-C	Convierte a archivos en formato vector de LPC, Autocorrelacion y aRaT.
KMEANITA	PC-C	Cuantización vectorial K-Medias, usa la distorsión de Itakura-Modificada modificada para el agrupamiento.
ITAKURA	PC-C	Calcula la distancia de Itakura-Saito modificada a partir de LPC y los centroides de entrenamiento.
LRA2M	PC-C	Convierte de formato vector de LPC a Cepstrum en formato de archivo M de matlab para procesarlos en el toolbox de fuzzy clustering.
FCMC	PC-Matlab	Toolbox de Matlab de agrupamiento difuso. De las tres distancias que puede utilizar se escogió la Euclidiana.
DISTFUZZ	PC-C	Calcula la distancia Euclidiana. Primero convierte de LPC a Cepstrum y después lleva a cabo la prueba de similitud.
CONFUSEN	PC-C	Programa que a partir de un archivo que es resultado de la medicion de similitud entre las palabras a reconocer y los vectores de entrenamiento, genera una linea de la matriz de confusion. La matriz completa se genera a partir de un archivo de procesamiento por lotes.
KMEANS*	PC-C	Algoritmo de K-Medias.
LPC*	PC-C	Convierte a un archivo de LPC.
DISTVECT*	PC-C	Calcula la distancia euclidiana a partir de un vector LPC y un vector de centroides de entrenamiento.
TEX2VCT*	PC-C	Convierte de formato Texas (NIST) a formato vector.
TEX2VOC*	PC-C	Convierte de formato Texas (NIST) a VOC.
VOLTEA*	PC-C	Intercambia los bytes de un archivo en formato de Texas para que la PC los pueda interpretar correctamente.
CRITBAND*	PC-C	Programa de cálculo de bandas críticas y preprocesamiento del espectrograma.
ESPECTRO*	PC-C	Despliega un espectrograma de la señal, previamente procesada.

* Usados en el desarrollo primario del proyecto y en el análisis e interpretación de resultados.