94
21

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE INGENIERIA

SEGMENTACION ACUSTICA DE VOZ.

T E S I S

QUE PARA OBTENER EL TITULO DE:

INGENIERO EN COMPUTACION

P R E S E N T A :

OSCAR SANCHEZ CACIQUE

DIRECTOR: M.I. JOSE ABEL HERRERA CAMACHO.

CIUDAD UNIVERSITARIA. MEXICO D.F. 1997

A mi padre *Cándido*
por la educación y apoyo que me dio,
y por que sabe la importancia que tiene para mí
la conclusión de esta fase de mi formación.

A mi madre *Alicia*
a quien le agradezco el cariño y la motivación
que me ha brindado siempre,
gracias a su invaluable ayuda
logré superar esta difícil etapa.

A mis padres
que siempre me dieron su confianza y creyeron en mí,
a quienes debo lo que ahora soy,
les dedico con todo cariño y gratitud esta tesis.

A mis hermanos *César* y *Yadira*
que sin su apoyo y compañía en todo momento
hubiera sido muy difícil terminar esta etapa de mi vida.

A mi abuelita *Petra*
a quien debo toda mi motivación
para seguir luchando gracias a su gran vida.

A mis tías *Teresa, Guadalupe, Severa,*
Angeles, Zoila y a mi tío *Eusebio*
de quienes agradezco parte de mi formación como persona,
gracias a sus ejemplares muestras de apoyo y sencilles.

A mis familiares *Lorena, Alfonso, Coni, Erik, Elizabeth, Mireya,*
Yarmil, Ulises, Wendy, Israel, Nendi, Ana, Isaura y Nei;
y a todos aquellos que aunque no estan nombrados en esta hoja,
les dedico con agradecimiento este trabajo.

A mis amigos de siempre *Omar y Rogelio*
y a sus respectivas familias,
con los que he compartido bellos momentos,
y me han brindado su amistad incondicional.

A *Lizbeth e Iliana* y a toda su familia
por la amistad y confianza que me dieron,
además de sus firmes consejos y motivación diaria.

A todos mis compañeros de la Facultad de Ingeniería
especialmente al grupo de los luminosos:
Erika, Adriana, Juan Manuel, Juan Antonio, Héctor,
Genaro, Carlos, Fernando, Ivan y Alejandro
a los que agradezco su gran amistad
y sus siempre valiosos consejos.

A mis sobrinos *Cindel, Itzel, Fernando y Ruben*
por sus agradables juegos que siempre motivan
y ayudan a distraer la tensión del trabajo diario.

A todos mis amigos del estado de Morelos,
especialmente a los de Jonacatepec,
en los que siempre he encontrado un gran apoyo.

En memoria de *Ana Claudia*(♀) ,

de mi tío *Petronilo*(♀)

y de mi abuelita *Paula*(♀) .

AGRADECIMIENTOS

*A la **Universidad Nacional Autónoma de México** y especialmente a la **Facultad de Ingeniería** por la formación que recibí en sus aulas.*

*Al **M. en I. José Abel Herrera Camacho** por darme la oportunidad de trabajar con él, por brindarme su tiempo, confianza y apoyo incondicional, además de la paciencia y motivación que tuvo en la realización de esta tesis.*

*A la **División de Estudios de Posgrado de la Facultad de Ingeniería** por las facilidades que me brindó.*

A todos los maestros de la Facultad de Ingeniería por su colaboración para mi formación profesional.

A todas aquellas personas que siempre me apoyaron y alentaron para la conclusión de esta etapa importante de mi vida.

INDICE.

Introducción.

INTRODUCCION.

Actualmente existen múltiples aplicaciones reales en el área de reconocimiento automático de voz. Tales aplicaciones consideran una gran variedad de enfoques en cuanto a las solución de sus problemas. No existe una normalización en cuanto a los caminos a seguir para generar los sistemas de reconocimiento; y en ocasiones no existe un criterio de comparación fijo entre estos sistemas, debido a que cada uno está creado para resolver una necesidad en especial. Depende de los objetivos que se pretendan abarcar es como se plantean las soluciones de los problemas.

Un sistema reconocedor de voz universal, que reconozca cualquier tipo de voz de cualquier lenguaje, y que tenga muy pocas restricciones de operación es el gran objetivo de cualquier investigador del área. Pero hoy existen muchas limitaciones para su realización (alto tiempo de procesamiento, gran consumo de memoria, etc.). Muchos investigadores han hecho sus trabajos enfocados en ciertas necesidades especiales, sin abarcar la generalización del sistema universal. Aunque ciertamente desde los 80's han desarrollados múltiples aplicaciones comerciales que utilizan varias técnicas .

El desempeño de los sistemas de reconocimiento de voz se debe medir tomando en cuenta su *"perfil de capacidades"*, que son los múltiples atributos que caracterizan al sistema. En este rubro se clasifican principalmente en : Palabras continuas o palabras separadas, dependencia del parlante o independencia del parlante, vocabulario pequeño o vocabulario amplio. Existen otros perfiles , pero generalmente los antes mencionados son los más importantes.

La voz continua es como hablamos normalmente, sin pausas marcadas entre cada palabra, pero el esfuerzo de análisis para este tipo de voz es grande para tratar de separar cada palabra . En cambio las palabras separadas o discretas, aunque no es la forma normal de comunicación , simplifican bastante el proceso de cómputo del reconocimiento.

Los sistemas que dependen de un solo parlante obliga al sistema a funcionar solo con una sola persona, limitando el uso para cualquier otro parlante. En contraparte el sistema de reconocimiento es más fácil, en contraste con un sistema multiusuario e independiente del parlante, pero en cambio se incrementan también las dificultades de realización.

Obviamente el reconocimiento de un vocabulario grande de palabras hace que se incremente tanto el espacio de almacenamiento como tiempo de procesamiento de computadora. En cambio un vocabulario pequeño además de reducir las desventajas anteriores permite menos conflictos al momento de la identificación de cada palabra.

Es pertinente limitar las funciones de las aplicaciones de acuerdo a las necesidades que se tengan, es decir acotar el desempeño de nuestro sistema mediante el perfil de capacidades. Iniciamos nuestro trabajo apoyados en las investigaciones de un pequeño sistema reconocedor de voz [HerrAlgIrv94], para palabras discretas o separadas, vocabulario pequeño (10 dígitos) e independencia del parlante. Decir que en el área de reconocimiento de voz existe la mejor solución es muy difícil, lo que se hace el utilizar lo investigado hasta ahora y se trata de ver lo más óptimo o que simplemente funcione para resolver las necesidades que se tengan. Por esta razón nos basamos en las necesidades planteadas para realizar nuestro trabajo.

Las señales de voz, aunque consideradas como no estacionarias, por su gran variabilidad en el tiempo, tienen una estructura bien definida, relacionada con segmentos lingüísticos y cuasiestacionarios. Estas propiedades son utilizadas en nuestro trabajo para procesar la señales de voz a reconocer. Comúnmente los sistemas de procesamiento de voz se dividen en tres categorías : *sistemas matemáticos, sistemas acústicos y sistemas lingüísticos*.

Generalmente los sistemas basados en cálculos puramente matemáticos, dividen una frase o palabra en una secuencia de segmentos de longitud fija y extraen el mismo tipo de características de cada segmento. Tales sistemas no explotan el gran nivel de correlación que existe entre los segmentos acústicos, y al ser siempre fija la metodología de

procesamiento, el tiempo requerido para reconocimiento siempre será parecido, impidiendo que este disminuya.

Los sistemas lingüísticos dividen una frase o palabra en segmentos definidos en este campo, ya sean fonemas , sílabas, etc. Generalmente estos trabajos también son muy grandes y requieren sesiones de entrenamiento controlados por humanos para determinar las características necesarias de las unidades lingüísticas definidas. Como resultado de la variabilidad de la articulación al producir la voz no existe una correlación directa entre segmentos acústicos y fonemas o sílabas.

En cambio, presentamos una técnica de segmentación acústica que incorpora en gran medida la estructura de la voz relacionada con un procesamiento lingüístico y a la vez incorporamos las características acústicas de la misma voz.

En el sistema que se propone en primer lugar, para cada palabra se calcula la envoltura espectral en frecuencias agrupada en bandas críticas. Esta es una eficiente representación que preserva la estructura espectral suficiente para el reconocimiento, además de utilizar menos espacio de almacenamiento que un espectrograma normal.

Usando la representación en bandas críticas, aplicamos el criterio de decisión *MLR (Cociente de Máxima Similitud)* [VanTrees68], para dividir cada palabra en segmentos acústicos, donde cada subpalabra es una región homogenea en el espectro de frecuencias en bandas críticas.

Lo importante del sistema propuesto previamente y lo que originó este trabajo son las siguientes características y problemas :

- Una de las partes primordiales y básicas del sistema de reconocimiento es el segmentador acústico. Del buen funcionamiento de segmentador dependen los buenos resultados de reconocimiento. En [HerrAlgIrv94] sugieren que para un mejor desempeño del sistema de reconocimiento de voz se mejore y perfeccione el segmentador acústico. Tratando de que se obtenga un segmentador acústico con las siguientes características deseables: Se pueda aplicar a cualquier tipo de palabras ; los segmentos acústicos

resultantes sean los más significativos, o sea las regiones más homogéneas y distinguibles; el número de segmentos este restringido en un intervalo muy pequeño y fijo de valores; además que el tiempo de procesamiento no aumente mucho.

- La independencia del sistema al tipo de parlante es debido a que el análisis es acústico. Si la palabra "one" es pronunciada por distintas personas, conserva básicamente las mismas características acústicas. Lo que generalmente varía es la intensidad y el tiempo de pronunciación, pero sus propiedades en las frecuencias se conservan. Un segmentador acústico tampoco depende del lenguaje, debido a que los tipos de procesamientos son en la frecuencia, y en el lenguaje que sea, siempre se obtendrán los mismos segmentos para cada palabra.

Basado en la importancia de un segmentador acústico y sus actuales deficiencias motivaron la realización de este trabajo, teniendo como *objetivos primordiales de nuestra tesis* los siguientes:

a) Analizar y mejorar el método de segmentación acústica propuesto por [HerrAlgIrv94], tratando de resolver los problemas que ahí se plantean.

b) El objetivo es mejorar el método de segmentación de manera que divida cualquier palabra sola en segmentos de longitud variable, subpalabras acústicas; cada segmento deberá tener casi las mismas características acústicas similares.

c) El número de segmentos por cada palabra deberá estar limitado en un número pequeño que caracterice a las palabras. Todas las palabras deberán tener un número fijo de segmentos que sean los de mayor importancia.

d) Lograr que el segmentador sea lo más automático posible, es decir que dependa en muy poco de la señal a procesar. Nos referimos a los niveles de decisión para determinar los segmentos importantes, que sean fijos y previamente determinados mediante un análisis matemático.

Principalmente el objetivo de nuestra tesis no es comparar el método de segmentación acústica con otros segmentadores, debido a que realmente no existe una amplia documentación de los mismos. Además como se explicó anteriormente las aplicaciones de reconocimiento se desarrollan de acuerdo a las necesidades que se tengan. Trataremos de mostrar el funcionamiento del segmentador despues de perfeccionarlo, cuidando los objetivos antes descritos.

En el capítulo primero de esta tesis presentamos una breve información sobre el *Procesamiento Digital de Señales*, que en la actualidad ya es demasiado extenso. Nosotros seleccionamos los temas de este capítulo para fundamentar su aplicación posterior en el método de segmentación. La transformada de Fourier en tiempo corto y la descomposición en bandas críticas son dos técnicas fundamentales en el método de segmentación y en esta sección se presentan sus fundamentos.

Debido a que uno de los objetivos del segmentador es que debe considerar las propiedades acústicas de la señal de voz, en el capítulo segundo desarrollamos la información necesaria para interpretar el enfoque acústico. Se presentan los diferentes esquemas que caracterizan a las señales de voz y la forma en que el humano las percibe. Información relevante para entender como es que nuestro sistema auditivo procesa las señales acústicas y visualizar una posible analogía con el procesamiento de un sistema reconocedor de voz.

El capítulo tercero describimos en su totalidad el método de segmentación acústica basado principalmente en dos técnicas importantes: la descomposición en b*andas críticas* y la prueba del *Cociente de Máxima Similitud (MLR)*. Este método esta basado en las investigaciones realizadas por [HerrAlgIrv94], y nosotros realizamos su documentación y mejoramiento de acuerdo a las necesidades antes planteadas. En esta sección describimos todo el proceso por el que pasa la señal para generar los resultados finales, que es la segmentación de cada palabra. Aquí se describe como es que se generan los tipos de pruebas MLR para la detección de segmentos acústicos, y para la determinación del inicio y fin de cada palabra, que es fundamental en el sistema.

El capítulo de la *presentación de resultados* es el cuarto, y es importante recalcar que el método de segmentación no se comparará con algún otro, simplemente intentará observar el funcionamiento de nuestro segmentador y que considere las necesidades planteadas como objetivos. Lo resultados mostrados son derivados con el segmentador acústico que ya incluye nuestras mejoras descritas explícitamente en el capítulo anterior. Creemos que mostrando las tablas y las gráficas de funcionamiento del segmentador se observará hasta que punto se han cumplido los objetivos previstos.

En el capítulo cinco hacemos un breve estudio de los sistema reconocedores de voz actualmente documentados. Tal vez nuestro segmentador acústico no es en si una aplicación real de reconocimiento automático de voz, pero basados en el sistema generado por [HerrAlgIrv94], y la información aquí presentada, mostramos varios enfoques de como es posible utilizar el segmentador en otro tipo de aplicaciones. Mencionamos distintas técnicas documentadas de reconocimiento, tanto las más generales enfocadas a reconocer patrones (segmentos acústicos por ejemplo) y una idealización de una aplicación posible utilizando como base al método de nuestra tesis.

CAPITULO 1.

Procesamiento Digital De Señales.

CAPITULO 1.

PROCESAMIENTO DIGITAL DE SEÑALES.

INTRODUCCION.

Actualmente existen diferentes técnicas y métodos encargados de procesar señales digitales. Se han desarrollado por diferentes investigadores teorías para procesamiento. Debido a que el objetivo de este trabajo es generar aplicaciones relacionadas con el reconocimiento automático de voz, utilizamos algunas de las técnicas generalmente recomendadas por trabajos previos. Tratamos de que tales técnicas sean prácticas y que realmente sirvan como herramientas para obtener buenos resultados. Uno de los objetivos primordiales es encontrar segmentos acústicos en un intervalo definido, para ello necesitamos la señal de voz con ciertas características. Los temas de este capítulo no fueron elegidos aleatoriamente, sino que se han desarrollado como una necesidad del proceso que seguirá cada archivo de voz para encontrar sus segmentos acústicos.

El enfoque de cada tema está muy relacionado con ideas para que inmediatamente se utilicen en la computadora. Pero solo hasta el capítulo tres es donde se explica la forma más práctica de su aplicación. Cada tema tiene una función primordial en la transformación que sufre la señal de voz original. En el preénfasis lo que se intenta es encontrar un filtro que realce las frecuencias altas, para hacer que la voz sea uniforme en todo el intervalo de frecuencias. El filtrado digital es útil para seleccionar solamente el intervalo de frecuencias que aporta información de la señal de voz; cambiar la frecuencia de muestreo ayuda a comprimir la señal original de voz, y además normaliza el formato de todas la señales para que tengan una tasa común de muestreo de 10,000 Hz.

Para realizar la transformada discreta de Fourier y obtener la información en el campo de las frecuencias, se utiliza el análisis espectral en tiempo corto, empleando pequeñas ventanas de tiempo. Esto se realiza aprovechando la propiedad de que la voz es estable y cuasiestacionaria en un pequeño intervalo de tiempo, aproximadamente 0.01 segundos. La descomposición en bandas críticas resulta una herramienta básica para nuestro método de segmentación acústica, y en este capítulo explicamos una forma rápida de su utilización.

1.1. PREENFASIS.

Para señales de voz debido a las dimensiones y propiedades de la boca y tracto vocal, existe una disminución de amplitud conforme aumentan las frecuencias de la señal. Esto significa que, para cada doble de frecuencias, la amplitud de la señal es reducida por un factor específico de 16 [Owens93]. Por esto es necesario compensar esta disminución preenfatizando la señal original de voz.

En el preénfasis, se procesa la señal de voz para incrementar el valor del intervalo apropiado del espectro, para que tenga en toda la señal de voz niveles de amplitud homogéneos en todos las frecuencias analizadas. Abarcando a todas las frecuencias audibles.

Preenfatizar significa levantar o darle amplitud a algunas frecuencias específicas . En el procesamiento de voz las frecuencias altas son las que necesitan mayor énfasis.

En un sistema de procesamiento digital de señales, el preénfasis puede ser aplicado con un filtro analógico de características paso altas de primer orden, con frecuencia de corte entre 100 Hz y 1 kHz [Owens93]. El filtro analógico es posible utilizarlo al momento de recibir la señal de voz original, pero para señales que ya están almacenadas digitalmente en algún medio, es necesario aplicarles un proceso posterior.

Lo indicado es utilizar un filtrado digital paso altas que seguirá inmediatamente después del proceso de digitalización. La ecuación de un filtro digital paso alta es (1.1) [Owens93]:

$$Y[n] = X[n] - a\,X[n-1] \qquad (1.1)$$

En la ecuación (1.1), $Y[n]$ es la salida del filtro en el tiempo *"n"* o la muestra *"n"*, $X[n]$ es el valor de la muestra *"n"* antes de ser filtrada. $X[n-1]$ es el valor de la muestra anterior a la actualmente procesada. Cuando $X[n]$ es la primer muestra, $X[n-1]$ toma el valor igual a cero. *"a"* es una constante que usualmente tiene un valor entre 0.9 y 1 [Owens93].

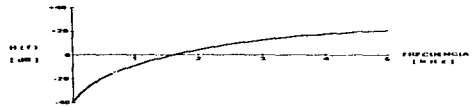

Figura 1.1 Respuesta en frecuencia del filtro digital preénfasis.

1.2. FILTROS DIGITALES.

Para los fines de este trabajo, no pretendo profundizar en el tema de los filtros digitales, que es muy extenso, pero intentaré dar un enfoque simplificado y práctico a esta herramienta muy utilizado en el procesamiento digital de señales. El objetivo es mostrar la aplicación y la forma de diseño de los filtros digitales.

Un filtro es un sistema por el que pasa el contenido espectral de una señal entrante en una banda específica de frecuencias. En otras palabras, la función de transferencia del filtro forma una " ventana" en el dominio de las frecuencias, a través del cual una porción del espectro entrante se le deja pasar, como se muestra en la figura 1.2 [Stearns90].

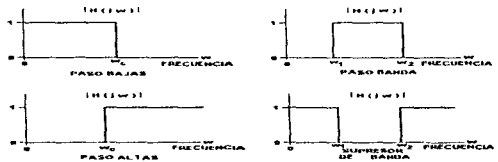

Figura 1.2 Amplitud característica ideal de los cuatro filtros básicos.

Existe una gran variedad de métodos para diseñar filtros digitales, pero nos enfocaremos a un tipo de diseño práctico. Filtrar señales resulta de gran importancia, pues permite seleccionar las partes espectrales que interesen de verdad, ya sea eliminando ruido incrustado en la señal o algún contenido en las frecuencias no deseado.

1.2.1. FILTROS DIGITALES POR TRANSFORMACION BILINEAL.

Aprovechando los métodos de diseño de filtros analógicos, se pueden utilizar para los filtros digitales. La transformación bilineal es vista como una transformación matemática de un sistema analógico (plano de la variable s) a un sistema discreto (plano de la variable z). Realizando la conversión del filtro analógico previamente diseñado a un filtro digital deseado.

Dos métodos utilizados para diseñar filtros analógicos son los filtros de Butterworth y Chebyshev, de los cuales existe bastante literatura y puede ser utilizada como referencia para el diseño de los mismos. Nosotros utilizamos la aproximación de Chebyshev para diseñar el filtro analógico con las características acordes a nuestras necesidades. No pretendemos profundizar en la teoría de filtros analógicos y solamente describimos el procedimiento general para diseñar un filtro Chebyshev descrito en [Stearns90].

La ecuación (1.2) es la función de transferencia de una filtro paso bajas de orden n.

$$H_A(s) = \frac{H_0}{(s-s_1)(s-s_2)(s-s_3)\ldots(s-s_n)} \qquad (1.2)$$

Donde H_0 es el factor de amplificación del filtro, y las variables s_n son los polos respectivos de la función de transferenica y se calculan con las siguientes relaciones:

$$s_n = \omega_c (senh\ \alpha\ cos\ \beta_n + j\ cosh\ \alpha\ sen\ \beta_n) \qquad (1.3)$$

$$\alpha = (1/N)\ senh^{-1}(1/\varepsilon) \qquad (1.4)$$

$$\beta_n = \frac{(2n + N-1)\ \pi}{2N} \qquad (1.5)$$

$$\varepsilon = \sqrt{10^{0.1\delta} - 1} \qquad (1.6)$$

$n = 1, 2, \ldots\ldots N$

N es el orden del filtro, ε es un valor característico de cada filtro, y depende de δ, que es el valor de rizo en dB . En la figura 1.3 se observa la respuesta característica de un filtro Chebyshev con frecuencia de corte en ω_c y su respectivo valor de rizo expresado en función de ε.

Figura 1.3 Respuesta en la frecuencia general de un filtro Chebyshev .

Para diseñar un filtro Chebyshev, primero se debe fijar un valor de rizo δ , la frecuencia de corte ω_c y la frecuencia máxima de amplitud ω, deseadas. Con esta información y la relación (1.7) se obtiene el orden del filtro.

$$N > = \frac{\cosh^{-1}(\delta / \varepsilon)}{\cosh^{-1}(\omega_r / \omega_c)} \qquad (1.7)$$

De esta forma se calculan los polos de la función de transferencia mediante las ecuación (1.3), y se obtiene la función de transferencia de un filtro analógico con las características de diseño.

El procedimiento recomendado para diseñar un filtro digital es el siguiente:

1.- Fijar los valores de las frecuencias de corte ω_c y ω_r. Encontrar los valores de ω_c' y ω_r' usando la ecuación (1.8) :

$$\omega' = \tan (\omega T / 2) \qquad (1.8)$$

2.- Diseñar un filtro analógico con la función de transferencia $H_A(s)$ de la ecuación (1.2) sustituyendo ω_c' en lugar de ω_c en la ecuación (1.3).

3.- Transformar $H_A(s)$ a H(z) mediante la relación (1.9).

$$s = \frac{z-1}{z+1} \qquad (1.9)$$

H(z) es la función de transferencia del filtro digital deseado.

4.- Implantar en la computadora la función de transferencia del filtro digital. Multiplicando o afectando la señal a ser filtrada.

1.3. CAMBIO DE LA FRECUENCIA DE MUESTREO.

En el procesamiento digital de señales es necesario en ocasiones cambiar la tasa de muestreo. Algunas veces al momento de capturar la señal de voz, se muestrea con una tasa baja, pero se necesita reconstruir la señal pues tal aplicación lo necesita. En algunos métodos de compresión de señales de voz, se necesita tener cierta tasa de muestreo para mayor eficiencia de los mismos. Pero para poder realizar la conversión de las tasas de muestreo, se debe tomar en cuenta el *Teorema de Nyquist*:

Una señal que contiene frecuencias no mas grandes que f_m, puede ser muestreada o transformada a una tasa f_s ,siempre y cuando f_s sea mayor o igual a $2f_m$ [Rowden92].

Teorema de Nyquist. $f_s >= 2\ f_m$. (1.10)

Al proceso de aumentar la frecuencia de muestreo se le llama Interpolación y a la reducción se le denomina Decimación [Oppenheim89].

Decimación.

La ecuación que define un sistema que disminuya la tasa de muestreo es :

$$X_d[n] = X[nM] = X_c(nMT) \qquad (1.11)$$

Llamado decimador o compresor. La tasa de muestreo puede ser reducida por un factor M si la tasa original de muestreo es por lo menos M veces la tasa de muestreo de Nyquist o si el ancho de banda de la secuencia es primero reducida por un factor de M por medio de un filtro ideal paso bajas, con frecuencia de corte p / M [Oppenheim89].

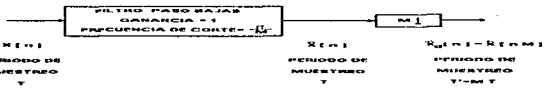

Figura 1.4 *Sistema general para la reducción de la tasa de muestreo por un factor M..*

Incrementar la tasa de muestreo involucra operaciones análogas a un convertidor digital a continuo (D/C) [Oppenheim89]. En el proceso de conversión de digital a continuo, se trata de interpolar valores intermedios entre dos muestras definidas con la frecuencia de muestreo anterior, al momento de aumentar la tasa de muestreo tendrán que encontrarse los valores aproximados que sean los más adecuados entre este par de muestras.

La ecuación de un sistema expansor o interpolador es:

$$X_e[n] = \begin{cases} X[n/L] & n = 0, \pm L, \pm 2L \\ 0 & \text{Otro caso.} \end{cases} \qquad (1.12)$$

El sistema se muestra en la figura 1.5, funciona de forma similar a un convertidor digital a continuo (D/C), primero se crea un tren de impulsos $Xe[n]$ y entonces aplicamos un filtro paso bajas para reconstruir la secuencia original [Oppenheim89].

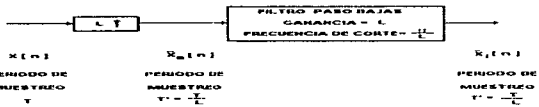

Figura 1.5 Sistema general para incrementar la tasa de muestreo por un factor L.

El aumento del periodo de muestreo puede ser desarrollado de manera simple por medio de una interpolación lineal, que se realiza el cálculo de los valores intermedios mediante la reglas lineales de las muestras en los extremos [Openheim89].

Figura 1.6 Interpolación lineal.

1.4. ANALISIS ESPECTRAL EN TIEMPO CORTO.

Una propiedad de las señales de voz es que son estacionarias y uniformes relativamente en algunos intervalos cortos de tiempo (desde 10 mseg a 100 mseg). Pero cambia considerablemente cuando el intervalo excede 0.5 segundos. A esta característica muy particular de las señales de voz se le llama cuasiestacionaria [Saito85].

Figura 1.7 *La voz es cuasiestacionaria y uniforme en un intervalo de tiempo muy corto.*

El análisis espectral en tiempo corto significa tomar un pequeño intervalo de muestras de la señal de voz y calcular su Transformada de Fourier o aplicarle algún tipo de procesamiento a este pequeño intervalo de voz. De esta forma se puede analizar todo la señal segmento por segmento hasta cubrir el total de la misma.

Existen dos razones para usar análisis espectral en tiempo corto o espectral promedio:

1) El espectro se calcula mucho más fácil utilizando pequeñas ventanas de intervalos de tiempo estacionario (análisis cuasiestacionario) [Saito85].

2) Los cambios lentos en el espectro se ven acentuados utilizando análisis en tiempo corto. Estos cambios tienen importancia relevante en procesamiento de voz [Saito85].

La transformada de Fourier esta planteada para ser calculada en un intervalo infinito de valores, pero para su implantación práctica se puede sustituir por una número finito de valores.

$$F_T(w) = \int_{-\infty}^{+\infty} f(t)\, w(t)\, e^{-jwt}\, dt = \int_{-T}^{+T} f(t)\, w(t)\, e^{-jwt}\, dt \qquad (1.13)$$

En la ecuación (1.7) $w(t)$ es conocida como función ventana-tiempo, y debe tener ciertas características para garantizar el mejor cálculo del espectro. Algunas ventanas usadas típicamente en el análisis en tiempo corto son:

1) Ventana Rectangular.

$$w(n) = 1 \qquad 0 <= n <= N-1 \qquad (1.14)$$

2) Ventana Barlett (Triángulo).

$$w(n) = \begin{cases} \dfrac{2n}{N-1} & 0 <= n <= N-1/2\ ; \\ \dfrac{2-2n}{N-1} & N-1/2 <= n <= N-1\ ; \end{cases} \qquad (1.15)$$

3) Ventana Hanning .

$$w(n) = 1/2\ \{\ 1 - \cos(2\pi n / N-1)\ \} \qquad 0 <= n <= N-1 \qquad (1.16)$$

4) Ventana Hamming.

$$w(n) = 0.54 - 0.46 \cos(2\pi n / N-1) \qquad 0 <= n <= N-1 \qquad (1.17)$$

5) Ventana Blackman.

$$w(n) = 0.42 - 0.5 \cos(2\pi n / N-1) + 0.03 \cos(4\pi n / N-1) \quad 0 <= n <= N-1 \qquad (1.18)$$

La ventana generalmente usada por sus propiedades es la ventana Hamming cuya función es (1.17). La ventana de Hamming evita los efectos de las aportaciones de la señal en los extremos de la ventana. En la figura 1.8 se muestra cuando las ventanas de análisis se traslapan al momento de realizar el cálculo espectral, existen pequeñas aportaciones en los extremos de las ventanas que no son deseados, y la ventana de Hamming, por sus propiedades evita esta transposición de ventanas ("Aliasing").

De acuerdo con experimentos anteriores [Saito85], la longitud recomendable de la ventana será de 10 a 30 mseg, y el desplazamiento de cada ventana de análisis puede ser de 10 a 20 mseg, permitiendo cierto traslape entre cada ventana.

Utilizando ventanas más grandes en tiempo, se obtiene más resolución en las frecuencias, y cuando se utilizan ventanas en tiempo corto, el análisis del espectro puede mostrar cambios rápidos temporales de la señal, a expensas de la resolución de las frecuencias [Saito85].

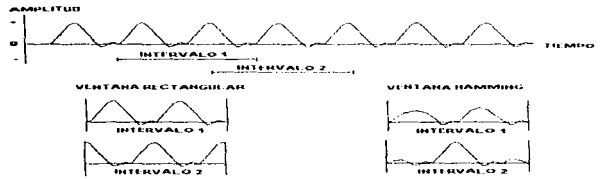

Figura 1.8 Reducción del efecto de las orillas por la ventana de Hamming.

1.5. TRANSFORMADA RAPIDA DE FOURIER.

El análisis de Fourier no es un concepto nuevo, se han desarrollado una gran cantidad de libros y artículos sobre este tema. Sin embargo el análisis de Fourier es visto como una interesante y útil transformación matemática que genera resultados en el dominio de las frecuencias.

No es necesario saber los detalles de las ecuaciones y variantes del análisis, pero si es necesario tener claros los conceptos básicos para poder interpretar los resultados de la transformación. Empleamos la Transformada Rápida de Fourier (FFT) como un instrumento de medida y para analizar las señales de voz desde un punto de vista acústico mediante sus componentes frecuenciales.

La transformada Discreta de Fourier (DFT) es el cálculo de la transformada normal de Fourier aplicada a las señales digitales mediante la adaptación se las variables de tiempo y frecuencias. En el que se limita el cálculo a un conjunto finito de puntos [Schafer75].

$$X(k w) = \sum_{-n}^{+n} X(n T)\, e^{-jkwnt} \qquad (1.19)$$

En (1.19) T es el periodo de muestreo de la función temporal y w es la resolución del espectro discreto. La resolución nos dice que tan cercanos están los valores de las

frecuencias encontradas por la transformada. Entre más cercanas estén, la calidad de la transformada es mejor, pero implica más tiempo.

En el análisis discreto de Fourier los límites de integración están comprendidos entre N muestras en el tiempo, para los que se pueden calcular N muestras en la frecuencia [Schafer75]:

$n = 0, 1, 2, \ldots\ldots\ldots N-1.$ *Dominio del tiempo.*

$k = 0, 1, 2, \ldots\ldots \ldots N-1$ *Dominio de la frecuencia.*

La resolución en el dominio de las frecuencias es:

$$w = 2\pi / NT \tag{1.20}$$

Por conveniencia, los términos T y w se sustituyen usualmente por los índices, quedando la ecuación de la transformada como:

$$X(k) = \sum_{n=0}^{N-1} X(nT)\, e^{-2\pi jnk/N} = \sum_{n=0}^{N-1} X(n)\, w^{-nk} \tag{1.21}$$

Donde:

$$w = e^{2\pi j/N} = \cos(2\pi/N) + j\,\mathrm{sen}(2\pi/N) \tag{1.22}$$

Se han desarrollado muchos algoritmos para el cálculo digital eficiente de la DFT, usando un número mínimo de multiplicaciones y utilizando un almacenamiento óptimo de los datos y coeficientes involucrados.

En general, los algoritmos de la Transformada Rápida de Fourier hacen uso de las redundancias que tienen lugar en la Transformada Discreta Normal de Fourier, para reducir las operaciones aritméticas. El número de muestras es una propiedad particular de los datos; la mayoría de los algoritmos de la Transformada rápida operan sobre N muestras, donde N se recomienda que sea igual a 2 elevado a cualquier potencia de un valor entero [Ramirez85].

Un método recomendado que se muestra esquemáticamente en la figura 1.9, es el de una ordenación y separación previa de los datos a ser transformados. Los datos se agrupan y en cada paso se separan en pequeños subgrupos. Esto ocurre hasta que la

transformación es completada cuando se obtiene un dato por grupo. Esta operación es conocida como decimación en frecuencia [Ramirez85].

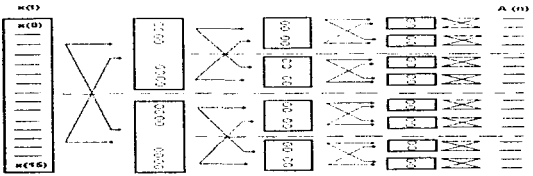

Figura 1.9 *Organización general de un algoritmo de reagrupación en frecuencias.*

1.6. DESCOMPOSICION EN BANDAS CRITICAS.

Una importante técnica en el análisis o síntesis de voz y audio es la descomposición en bandas críticas. Lo que se intenta con esta técnica es imitar mediante un modelado matemático al sistema auditivo humano, donde las bandas críticas refleja la percepción auditiva humana.

Es conocido por investigaciones y experimentos que el sistema auditivo funciona como varios filtros paso banda con distintas frecuencias y anchos de banda. Estos filtros se encuentran dispuestos uno después de otro para cubrir todas las frecuencias de la señal de audio.

La restricción impuesta por el modelo auditivo son mejor conocidas como sistema no uniforme de análisis y síntesis de voz, en el cual los anchos de banda de los canales y frecuencias centrales se incrementan cuando las frecuencias aumentan [Atal93].

Las bandas críticas han sido medidas a través de los intervalos auditivos. El ancho de banda para una frecuencia central de 200 Hz es de alrededor de 100 Hz y para la frecuencia central de 5000 Hz es de aproximadamente 1000 Hz. Estas mediciones

condujeron a encontrar una relación entre la escala de frecuencia a otra conocida como escala de los barks (Un bark es una banda crítica). La relación encontrada en los experimentos [Ainsworth88] es (1.23):

$$f = 650 \sinh (x / 7) \qquad (1.23)$$

f = *frecuencia en Hertz.*

x = *Banda Crítica en Barks.*

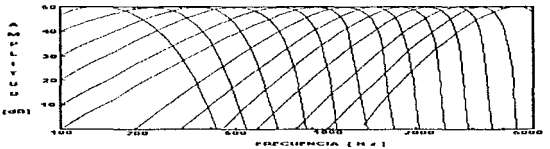

Figura 1.10 *Bandas críticas. Notar que la escala de las frecuencias es logarítmica, y que las formas de las bandas son idénticas.*

Con la relación entre Hertz y Barks se pueden transformar los resultados generados por la transformada de Fourier. De esta forma se agrupa el intervalo de valores desde 0 Hz a 5000 Hz en cada una de las 18 bandas críticas correspondientes.

Las ventajas de la descomposición es que se obtiene un arreglo por ventana con menor número de subíndices y de similar cantidad de información. Además de imitar el mismo procedimiento que se realiza en el oído humano.

1.7. ESPECTROGRAMAS.

Una de las formas más efectivas para observar las características en la frecuencia de una señal de voz, es mediante el cálculo y despliegue de un espectrograma. Un espectrograma es una ilustración en tres dimensiones de la forma en que las aportaciones en la frecuencia de la señal de voz varían con el tiempo.

Un espectrograma es una valiosa forma de obtener patrones visualmente de una palabra o frase de voz [Owens93].

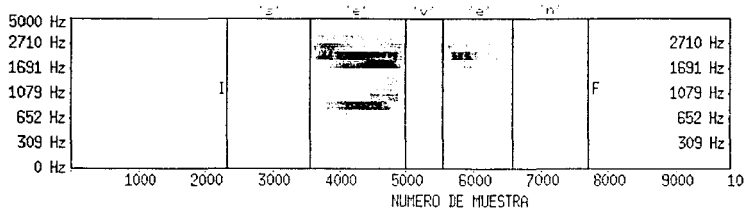

Figura 1.11 *Ejemplo de un espectrograma de la palabra seven con cinco segmentos acústicos.*

En la figura 1.11 se muestra un espectrograma, donde la axisa horizontal representa la variable tiempo y la vertical es la variable frecuencia. La tercera dimensión es dada por la obscuridad de los puntos, e indica la energía que aporta la señal en un punto definido por los valores de tiempo y frecuencia. La obscuridad de un punto es proporcional a la energía aportada en ese instante de tiempo y su respectiva frecuencia.

Dos frases o más que tienen sonidos similares, tienen espectrogramas muy parecidos. Esta es una de las características utilizadas en el reconocimiento automático de voz, pues se pueden comparar distintos espectrogramas, de esta forma se pueden clasificar o referenciar la palabra entrante con su respectiva palabra previamente almacenada, y que es con la que el parecido es mayor. Aunque hacer la comparación en forma gráfica resulta un poco complicado y es bastante laborioso.

Obtener patrones visualmente mediante la ayuda de expertos humanos es difícil, y diseñar un algoritmo de computadora para emular este proceso es substancialmente un problema mucho mas complicado [Owens93].

CAPITULO 2.

Percepción De La Voz.

CAPITULO 2.

PERCEPCION DE LA VOZ.

INTRODUCCION.

Desde los inicios de estudio del reconocimiento automático de voz, han aparecido múltiples enfoques sobre como desarrollar aplicaciones referentes al tema. Nadie ha normalizado este tipo de enfoques, pues todos tiene sus ventajas y desventajas, es por esto que tenemos la libertad de elegir que información se utiliza para el desarrollo de un sistema reconocedor de voz.

En mi caso, para desarrollar el segmentador de voz se utilizó y trató de aplicar un enfoque basado en la forma en que los humanos percibimos la voz. En cada tema se destacan algunas características básicas que son útiles para los fines del trabajo.

Las características acústicas de la voz representan información constante del tipo de voz generada, y bien utilizada son una herramienta muy útil en el desarrollo de los sistemas de procesamiento de voz.

Iniciamos con una introducción sobre las partes funcionales del sistema auditivo, se logra entender las propiedades de nuestro oído, que es en cierta medida donde se realiza parte del reconocimiento de la voz. En el oído existen limitaciones físicas que generan temas como las bandas críticas, el enmascaramiento de señales, la selectividad de frecuencias entre otros. La selectividad de frecuencias es ampliamente utilizada en este trabajo, por lo que debe ser realmente entendida.

En cuanto a los subtemas de la producción de voz, es para fundamentar el procesamiento que se le da a la señal de voz, cuidando sus características y no alterando sus propiedades.

2.1. SISTEMA AUDITIVO HUMANO.

Para darle un enfoque acústico a un sistema reconocedor de voz, es importante conocer las funciones y características físicas del oído humano. Debido a las dimensiones y órganos del sistema auditivo se tienen ciertas propiedades acústicas aprovechables en cualquier sistema de procesamiento de voz.

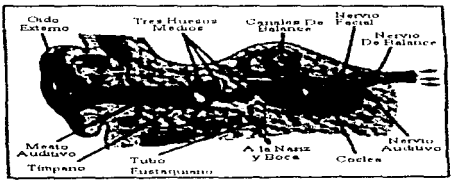

Figura 2.1 Corte transversal del sistema auditivo humano.

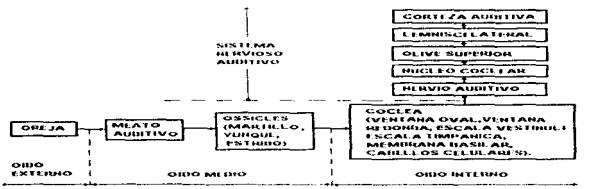

Figura 2.2 Diagrama de bloques del sistema auditivo humano.

El oído está dividido en tres partes principales: oído externo, oído medio y oído interno como se muestra en las figura 2.1 y 2.2. El oído externo está formado por la oreja, la cual tiene la función de direccionar las señales acústicas del exterior hacia el interior del oído. Al oído externo también lo compone el meato auditivo, que es un canal uniforme con dimensiones variables, dependiendo de cada persona. El meato auditivo también se encarga de conducir los sonidos hacia el oído medio y tiene forma de un tubo. Como cualquier tubo

este tiene una frecuencia de resonancia, en aproximadamente 3 kHz las frecuencias empiezan a decaer en amplitud [Parsons87]. Al final del meato auditivo se encuentra la membrana timpánica o tímpano. Su función es vibrar cuando el sonido choca con esta membrana y es el primer contacto físico que encuentra la señal acústica.

Más adentro de la membrana acústica se encuentra el oído medio, compuesto de una cavidad llena de aire donde existen tres pequeños órganos llamados respectivamente martillo, yunque y estribo. Proporcionan el acoplamiento acústico, permitiendo una transformación de impedancias entre la vibración del tímpano y el oído interno, además de que limitan la amplitud de la señal entrante amortiguando vibraciones [Parsons87]. La transformación de amplitudes es debida a dos funciones, la función mecánica ejercida por el martillo y el yunque, ayudados del área del tímpano. La relación de transformación es de aproximadamente 15 a 1 [Parsons87]. La función principal del oído medio es proteger de los niveles altos de los sonidos. La limitación es hecha por contracción de los músculos internos del oído.

El oído interno está compuesto del aparato vestibular, la ventana oval y circular, y el caracol o coclea. El caracol es el pasaje de comunicación entre el oído medio por conducto de las ventanas oval y circular. El caracol contiene transductores que convierten vibraciones acústicas a impulsos nerviosos. En el caracol la energía acústica entra por la ventana oval y viaja a través del mismo y vuelve a salir por la ventana circular. En su trayecto las vibraciones hacen mover a la membrana basilar, alejandose o acercandose al estribo en función de la frecuencia de la señal entrante, como se muestra en la figura 2.3. La membrana basilar tiene conectados una cantidad grande de cabellos nerviosos, que se encargan de sensar las vibraciones de la membrana basilar, cada cabello está conectado a una sinapsis nerviosa y así es como se transmite la información frecuencial de la señal auditiva al cerebro. En la figura 2.4 se muestra la función en la frecuencia necesaria para hacer vibrar a la membrana basilar, note que en las frecuencias altas se necesita mayor amplitud de las señales para producir una respuesta en la membrana.

La función del caracol es producir una dispersión espacial de los componentes frecuenciales mediante toda la longitud de la membrana basilar. Actúa como un analizador de espectro mecánico-neural. [Parson87].

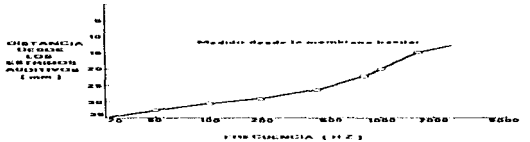

Figura 2.3 Posición de la máxima amplitud a lo largo de la membrana basilar como una función de la frecuencia aplicada".

En este punto no se sabe si todo el procesamiento espectral es realizado en el cerebro y en algún órgano intermedio entre el caracol. Pero para fines de este trabajo es un hecho que se realiza interiormente una descomposición frecuencial del sonido, para su posterior interpretación.

2.2. ACUSTICA DE LA PRODUCCION DE LA VOZ.

Cuando una persona genera voz, se producen sonidos en una secuencia de distintas frecuencias, estas frecuencias son llamadas armónicas naturales y es sabido que cada una de estas frecuencias es múltiplo de la frecuencia fundamental [Rowden92]. Pero nunca escuchamos esta combinación de frecuencias, pues al pasar por el tracto vocal algunas frecuencias resuenan con mayor intensidad y otras no lo hacen tanto.

Figura 2.4 Respuesta en frecuencia de la membrana basilar.

2.2.1. FRECUENCIAS FORMANTES.

Debido a que el tracto vocal y nasal no son de dimensiones fijas, y cambian por el movimiento que describen las distintas articulaciones al momento de producir voz. Si observamos una posición particular de las articulaciones al momento de producir voz, es posible reconocer distintas frecuencias características (*"frecuencias formantes"*). La frecuencia característica más baja es llamada frecuencia formante (f1), la siguiente es la segunda frecuencia formante (f2) y así sucesivamente.

Cada uno de los sonidos vocales tiene una muy definida característica de frecuencias formantes, sin hacer caso de quien es la persona que generó la voz. Por ejemplo, para la vocal /e/, se tienen valores típicos para la frecuencia formante f1 y f2 de 300 Hz y 2100 Hz respectivamente. Al producir la voz, la lengua está cerca de la boca y la segunda formante resulta del pequeño tamaño de la cavidad del tracto vocal [Rowden92].

La frecuencia fundamental variará dependiendo de la persona parlante, sus modismos y el énfasis, pero esta magnitud y la relación de las frecuencias formantes hacen de cada sonido de voz una fácil clasificación y reconocimiento [Rowden92].

2.3. MODELO DE PRODUCCION DE VOZ "FUENTE FILTRO."

Para poder entender la forma de producción de voz y procesamiento de señales de voz es mediante algún modelo simplificado que sea una analogía de lo que realmente ocurre en el tracto vocal y las distintas partes de la boca.

En la figura 2.5 se muestra el modelo usualmente utilizado para modelar el proceso de producción de voz, el de fuente-filtro. Es normalmente simulado en forma electrónica mediante una señal entrante y producida por un generador de pulsos con una variedad de frecuencias armónicas, y a la vez una señal de ruido de ancho de banda definido.

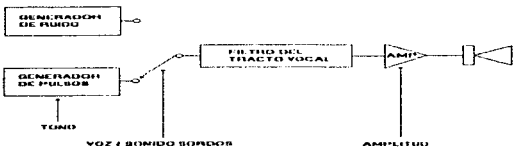

Figura 2.5 Modelo de producción de voz fuente-filtro.

La señal es pasada por el filtro que tiene características muy similares al tracto vocal. Los parámetros del filtro pueden ser claramente no constantes, pero esta variedad es debida a las modificaciones que también ocurren en el tracto vocal debido al movimiento de las articulaciones . El tono de la voz está sujeta a la amplitud de la señal entrante al filtro.

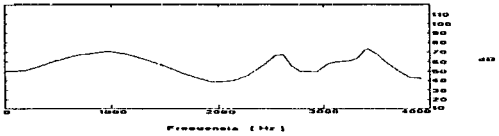

Figura 2.6 Envoltura espectral del filtro del tracto vocal.

Las formas de la respuesta en frecuencia mostradas en la figura 2.6 son debidas a las características del filtro, las cuales son aplicadas a la señal en el dominio de la frecuencia. Típicamente las características del filtro consisten de curvas donde los picos o polos representan las frecuencias formantes y cambian por las variaciones del sonido [Rowden92] .

2.4. PERCEPCION DEL TONO.

El tono se define relativamente a alguna frecuencia de referencia. La diferencia de tono entre dos notas es igual a 1200 veces el logaritmo de base dos de sus frecuencias. La unidad del tono musical es el céntimo [Parsons87].

El tono en el cual el sonido es producido depende de muchos factores tales como la frecuencia de excitación de las cuerdas vocales, el tamaño de la caja vocal y la longitud de las cuerdas vocales. El tono también varía con las palabras, al dar más énfasis en ciertas sílabas.

2.5. PERCEPCION DE LA SONORIDAD.

La sonoridad es medida como el nivel de intensidad en que es producida la voz. Depende de varias circunstancias como las emociones del parlante. Otros factores que también influyen en la sonoridad de la voz es que tan lejos queremos que se perciba, o el ruido del medio. El ajuste de la intensidad depende de la situación en la que nos encontremos.

Las variaciones de la intensidad son producidos por los músculos de la laringe, las cuales permiten un gran o un mínimo flujo de aire. La intensidad o sonoridad es también afectada por el flujo de aire desde el pulmón, que es el órgano principal de la generación de voz.

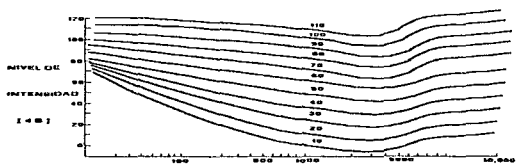

Figura 2.7 Curvas de nivel. El número en las curvas indica el tono en fons .

Para lograr percibir la intensidad se toman en cuenta dos factores, la frecuencia y el nivel de intensidad de cada frecuencia. Si se comparan tonos en diferentes frecuencias y amplitudes se generan las llamada curvas de nivel [Parsons87]. Si se observa la figura 2.7, en la que se muestran las curvas de nivel para cada tono, existe una decadencia de todas las curvas en 3 kHz a 4 kHz, lo que indica un incremento de sensibilidad por parte del oído humano en esta región, ya que es necesaria menor intensidad de la señal para ser percibida.

2.6. ADAPTACION, ASIMILACION Y ARTICULACION.

Al producir voz normalmente, las articulaciones son hechas rápida y consecutivamente. Es por ello que cada articulación no esta discretizada debido a que es modificada por las articulaciones de las sílabas o palabras vecinas.

La adaptación de un fonema resulta de la influencia de sonidos cercanos, los cuales cambian la forma del tracto vocal y la posición de las articulaciones. Cuando hablamos más rápido, la adaptación es más notoria porque la lengua no alcanza las mismas posiciones que generalmente ocupa al hablar despacio.

En casos extremos de adaptación, un sonido de voz puede cambiar tanto, que toma algunas características de los sonidos vecinos, denominándosele a este fenómeno como asimilación [Parsons87].

La articulación es debida a dos movimientos articulatorios al momento de producir la voz en el mismo tiempo para la generación de dos diferentes fonemas.

Al combinar la adaptación y la articulación se obtiene un sonido de mejor calidad, y es más impermeable al ruido porque la información generada es una articulación simple, tarda más tiempo y se percibe con mayor claridad [Parsons87].

2.7.SELECTIVIDAD DE FRECUENCIAS.

Al escuchar sonidos de voz es necesario oír las diferentes aportaciones de las distintas frecuencias formantes, para así hacer una fácil identificación de los sonidos vocales. Fletcher (1940) propuso la teoría de que la percepción de los componentes frecuenciales en un sonido es detectado por una serie de filtros paso - bandas centrados continuamente en diferentes frecuencias en todo el intervalo auditivo humano [Rowden92]. Es decir, la forma de ver al sistema auditivo humano combinado con los procesos cerebrales es mediante un analizador espectral con distintos filtros paso-bandas centrados en alguna frecuencia especial y que en conjunto abarcan todo el espectro audible posible. La habilidad para discriminar entre dos sonidos simultáneos los cuales tienen frecuencias similares o muy juntas, se limita a el ancho de banda de cada filtro.

A cada frecuencia central de los filtros auditivos se les denomina Banda Crítica [Rowden92]. Existe un fenómeno comúnmente presentado y que se puede explicar con fundamento en las bandas críticas, y es el llamado enmascaramiento. Este fenómeno es característico cuando un sonido cualquiera, no puede ser escuchado debido a la interferencia de otro sonido máscara.

2.8. BANDAS CRITICAS.

Anatómicamente existe una membrana en el oído que funciona respondiendo de forma variable dependiendo de las frecuencias de la señal entrante. Esta es la llamada membrana basilar, que cumple con la función de analizar espectralmente el sonido.

La membrana basilar vibra de forma diferente dependiendo de las frecuencias del sonido que se escuchen. es decir, cumple con la tarea de analizar espectralmente la señal de entrada compleja, detectando los diferentes componentes en frecuencia en diferentes puntos a través de su longitud [Owens93]. Cada punto de la membrana basilar puede considerarse como un filtro paso-bandas con cierta frecuencia central y un ancho de banda, y que en conjunto realizan un análisis espectral de la señal entrante.

Las bandas críticas han sido determinadas en un amplio intervalo de frecuencias para experimentos acústicos, y pueden verse la figura 2.8[OweFran93].

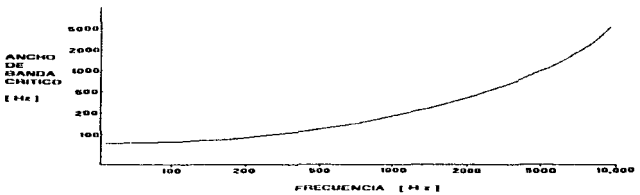

Figura 2.8 Bandas críticas en un amplio intervalo de frecuencias.

En la figura 2.8 a cada frecuencia central con su respectivo ancho de banda se le denomina Banda Crítica, y son las frecuencias que los humanos pueden discriminar al escuchar la voz y otros sonidos. Si los componentes frecuenciales de una señal de voz coinciden con una banda crítica, entonces esta frecuencia es percibida por el oído, en caso de que esta componente no entre dentro de esta banda, entonces no es percibida.

2.9. ENMASCARAMIENTO.

Es el nombre que se le da al fenómeno que ocurre cuando el sonido específico interfiere en la percepción de otro sonido diferente. El grado en el cual ocurre el enmascaramiento es una función relativa a las características del sonido, como son la intensidad y sus valores frecuenciales.

El fenómeno de enmascaramiento es importante en el estudio de la voz, pues la consecuencia más generalizada de este fenómeno, es el levantamiento o alteración de ciertas frecuencias de un sonido por una máscara de sonido no relevante. Es decir, la alteración del sonido original por otro sonido puede ser de vital importancia en un sistema

reconocedor de voz, por la posible interpretación equivocada de la señal de voz entrante debido al enmascaramiento.

En reconocimiento de voz ante este fenómeno no se puede hacer mucho, pues una vez que el sonido presenta enmascaramiento, es muy difícil discriminar y eliminar la máscara intercalada en el sonido. Lo más óptimo sería al momento del reconocimiento, evitar ruidos extremos y una pronunciación clara. Aunque también el enmascaramiento puede originarse por una pronunciación rápida, donde sonidos anteriores permanecen en el medio enmascarando a los sonidos actuales. Es notorio que no se puede hacer demasiado para evitar el enmascaramiento.

2.10. PERCEPCION GENERAL DE LA VOZ.

Como percepción de voz en este trabajo nos referimos al proceso cerebral para darle interpretación a la voz. Es cierto que aún no se sabe con seguridad cual es el proceso que se sigue en la percepción de voz, pero varios investigadores han hecho experimentos para tratar de encontrar la forma en que percibimos.

Quizá al conocer el proceso de percepción no sea totalmente utilizado en un sistema de reconocimiento automático de voz, pero puede ser de gran ayuda para desarrollar algunas partes del sistema, fundamentándose en las funciones realizadas en la percepción de voz por los humanos.

Parece que el cerebro hace una distinción fundamental entre sonidos de voz y sonidos que no lo son. Aparentemente el cerebro procesa en forma diferente a los sonidos de voz. Es creíble que existe una predisposición innata en el sistema nervioso humano para decodificar entradas de voz [Parsons87].

La importancia del contexto en el que se genera la voz influye en la facilidad y precisión de la percepción. Nosotros percibimos la voz por modelación del parlante internamente. Duplicamos mentalmente la voz, siguiendo y anticipándonos a la misma [Parson87]. Tal modelo requiere coherencia y continuidad de la voz. Esta teoría de modelación interna es de gran influencia para desarrollar sistemas reconocedores de voz.

La forma en que algunos sistemas reconocedores de voz procesan la información es basándose en la idea de que la percepción se basa principalmente en la habilidad del humano para reconocer características distinguidas en los sonidos recibidos, sin tomar en cuenta muchos factores del medio como el contexto. Es decir, solo tomar en cuenta características físicas y acústicas de la señal de voz para realizar el proceso de detección de voz.

Existen algunas teorías más simples que intentan explicar el proceso de percepción de voz [Parsons87]:

1- La voz humana no es entendida solo por sus características acústicas o análisis puro de señales sino que involucra el conocimiento del lenguaje y el medio.

2- La voz entrante es procesada palabra por palabra en el orden en que sean recibidas. La identificación de cada palabra ayuda para el reconocimiento de las siguientes palabras, limitando las posibilidades de la palabra subsecuente.

3- En el reconocimiento de voz el oyente pone más interés en detectar y reconocer la primer parte del sonido, para así eliminar posibilidades del resto de la palabra y hacer más fácil el reconocimiento.

La experiencia de investigadores al escribir programas para reconocimiento de voz se ven influenciados por algunos modelos de percepción de voz. Muchos reconocedores de voz usan estas teorías como fundamentos. Algunos inician un análisis de la señal entrante, después de que la primer palabra es reconocida, se auxilian de este primer reconocimiento para la identificación de las palabras restantes.

Saber con seguridad cual es el proceso que se realiza en el cerebro para la percepción de la voz es muy difícil. Lo único que se tiene son posibles teorías que sugieren algunas formas. En un sistema reconocedor se pueden tomar esas ideas y aplicarlas. Quizá no sea la forma en que se realiza en el cerebro, pero para una aplicación práctica son útiles, como puede observarse en los distintos reconocedores de voz.

CAPITULO 3.

Segmentación Acústica De Voz.

CAPITULO 3

SEGMENTACION ACUSTICA DE VOZ.

INTRODUCCION

Basados en las investigaciones hechas en el artículo [HerrAlgIrv94], en donde realizan un pequeño sistema de reconocimiento automático de voz, para palabras separadas, independencia de parlante y un número pequeño de palabras a reconocer. Plantean como básico para el funcionamiento del sistema la etapa de segmentación acústica, que es el encargado de dividir cada palabra en pequeños bloques que tengan cada uno casi las mismas características en el dominio de la frecuencia. De este trabajo surge la necesidad de mejorar el segmentador considerando las siguientes características deseables:

a) El segmentador sea lo más general posible, es decir que divida a cualquier palabra en sus respectivas subpalabras. Y además sea lo más automático, que no dependa de cambiar valores para su funcionamiento.

b) El número de segmentos esté entre un pequeño intervalo de valores (facilita el reconocimiento de voz [HerrAlgIrv94]). Considerando que los segmentos resultantes sean los más representativos de cada palabra.

En este capítulo describimos el *método de segmentación acústica* propuesto, y explícitamente en cada una de las etapas donde sea necesario realizamos las modificaciones pertinentes, para tratar de cumplir con los objetivos propuestos. Además de que verificamos el funcionamiento del método.

Existen dos etapas realmente importantes, donde se fundamenta el método de segmentación, estos son la descomposición en *bandas críticas y la prueba del cociente de máxima similitud (MLR)*. La descomposición en bandas críticas simplifica la representación en el dominio de las frecuencias de la señal de voz, facilitando la extracción de los patrones necesarios para la segmentación. La prueba de máxima similitud es básica para realizar la segmentación acústica, y utiliza la información del espectro en bandas críticas. Es decir, es

un método que procesa la señal matemáticamente y además considera las características auditivas de la señal de voz.

El orden de este capítulo esta basado en como se procesa la señal original de voz. Cada tema es un pequeño subproceso de todo el sistema segmentador. La señal original ya digitalizada es preprocesada mediante filtros para extraer solo la información relevante, posteriormente calculamos su representación en bandas críticas mediante un análisis de ventanas. Después encontramos las dos funciones de máxima similitud para detectar los valores de inicio y fin de la señal, y la otra para la detección de todos los segmentos acústicos posibles. Por último la parte de selección de los segmentos , tomando en cuenta las características de las señales de voz.

3.1. PREPROCESAMIENTO DE LA SEÑAL DE VOZ.

La señal de voz necesita tener ciertas características para facilitar la extracción de información. Previamente a esta etapa, la señal de voz ya ha sido digitalizada y almacenada en la computadora. El preprocesamiento aplicado a cada una de las palabras guardadas incluye un preénfasis, filtrado paso bajas y cambio de su tasa de muestreo.

Es importante notar que realmente el preprocesamiento no forma parte del método de segmentación acústica, pues se puede utilizar como preámbulo de cualquier aplicación donde se procesa voz, ya que este preprocesamiento ayuda a transformar la señal para que se facilite la obtención de sus características y mejoren los resultados finales.

3.1.1. PREENFASIS.

La señal de voz original presenta poca influencia de las aportaciones de las frecuencias más altas (aproximadamente 3000 Hz en adelante). Por esta razón se debe dar mayor amplitud a estas aportaciones, para homogeneizar las amplitudes en todo el intervalo frecuencial, como se observa en la figura 3.1.

La fórmula digital utilizada en el preénfasis es [Owens93]:

$$y[n] = x[n] - a\,x[n-1] \qquad (3.1)$$

Figura 3.1 Efecto del filtro preénfasis en el dominio de las frecuencias.

Donde cada muestra leída del archivo a preenfatizar se almacena en el arreglo *x[n]*. *x[n-1]* es una muestra anterior en el tiempo. *y[n]* es la salida del preénfasis en el tiempo *n*. Se eligió el valor de $a = 0.95$, siguiendo las recomendaciones de [Owens93].

3.1.2. FILTRADO.

En la señal de voz pueden existir aportaciones de frecuencias que no son útiles , debido en algunos casos a ruidos externos, ruidos asociados al proceso de producción de voz que no son relevantes, etc.. El filtrado selecciona las frecuencias que se desean escuchar y utilizar, dejando fuera las que no se quiera procesar.

Filtramos la señal con un paso bajas con frecuencia de corte ω_c en 4500 Hz, pues la señal de voz generalmente se encuentra en el intervalo de 100 Hz y 4500 kHz, la frecuencia de máxima amplitud ω_r en 4510 Hz , el valor de rizo en $\delta =$ 0.3 dB y la frecuencia de muestreo de las palabras es de 12500 Hz . Calculamos el orden necesario del filtro para cumplir las condiciones antes fijadas, mediante las ecuaciones (1.6) y (1.7) descritas en el capítulo 1.

$$\varepsilon = 0.2674309$$
$$N >= 7.3310$$

Por lo tanto el orden del filtro deberá ser de 8. Los polos de la función de transferencia analógica, calculados con la ecuación (1.3) son:

$s_0 = -0.2126108339 + j\ 4.3034035776$
$s_1 = -0.6054644296 + j\ 3.6482493933$
$s_2 = -0.9061415544 + j\ 2.4376823104$
$s_3 = -1.0688668417 + j\ 0.8560001934$
$s_4 = -1.0688668417 - j\ 0.8560001934$
$s_5 = -0.9061415544 - j\ 2.4376823104$
$s_6 = -0.6054644296 - j\ 3.6482493933$
$s_7 = -0.2126108339 - j\ 4.3034035776$

Sustituyendo los polos en la función de transferencia $H_A(s)$ (1.2) del filtro analógico, y aplicando la transformación bilineal mediante la relación (1.19) , obtenemos al simplificar, la función del filtro digital en diferencias:

$$y(t) = [x(t)+8\,x(t-T)+28\,x(t-2T)+56\,x(t-3T)+70\,x(t-4T)+56\,x(t-5T)+28\,x(t-6T)+8\,x(t-7T)+x(t-8T)] - [2.17\,y(t-T)+3.30\,y(t-2T)+2.68\,y(t-3T)+1.82\,y(t-4T)+0.63\,y(t-5T)+0.3\,y(t-6T)+0.04\,y(t-7T)+0.05\,y(t-8T)] \quad y(t) \qquad (3.2)$$

En (3.2), cuando $t = 0$ los valores de "x" y "y" antes de ese tiempo se asumen como cero. Con este filtrado se garantiza que lo único que entrará al método de segmentación es la parte de la señal de voz. Eliminando las parte supuestamente de ruido. En la figura 3.2 se observa la respuesta en frecuencia del filtro digital con valor de rizo igual a 0.3 dB, frecuencia de corte ω_c de 4500 Hz y frecuencia de máxima amplitud ω, de 4510 Hz.

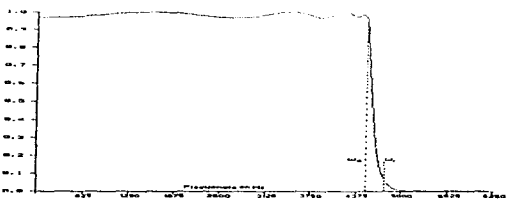

***Figura 3.2** Respuesta en la frecuencia filtro digital paso bajas con ganancia igual a 1.*

3.1.3. CAMBIO DE LA FRECUENCIA DE MUESTREO.

Para facilitar los cálculos del tamaño de las ventanas de análisis que posteriormente se utilizarán en el método, se utiliza un cambio de la tasa de muestreo, ajustándola a un valor más cómodo. Cualquiera que fuera la frecuencia de muestreo original, procesamos la señal para que en todos los casos sea de *10,000 Hz*, que cumple con el *teorema de Nyquist*. Pues la frecuencia máxima audible de una señal de voz se puede considerar en *5,000 Hz*, el doble es *10,000 Hz* que es la frecuencia mínima a la que podemos remuestrear. En el caso de los archivos originales que utilizamos para probar el método,

tenían una frecuencia de muestreo de *12,500 Hz*, por lo que el procesamiento fue para reducir la tasa de muestreo.

La forma práctica de cambiar la frecuencia de muestreo, es utilizando una interpolación lineal, tomando valores vecinos al que se desea aproximar [Oppenheim89].

$$X_a(t) = \sum_{-\infty}^{+\infty} X_a(nT) \left[\frac{sen\ \pi (t - nT) / T}{\pi (t - nT) / T} \right] \qquad (3.3)$$

3.2. BANDAS CRITICAS.

El proceso de agrupación en bandas críticas esta compuesto de varias etapas importantes, que influyen en el desempeño del método de segmentación. El objetivo final es encontrar un arreglo bidimensional que contenga los valores de las bandas críticas para cada conjunto de muestras. Las bandas críticas agrupan ciertas frecuencias específicas relacionándolas a un promedio matemático de las mismas.

Uno de los objetivos y ventajas del cálculo de bandas críticas es que a partir de la señal original, se encuentra una representación en la frecuencia que además de ser muy simple y ocupe poco espacio en la memoria del computador, además que sea útil para los procesos posteriores.

3.2.1.VENTANAS.

Antes de procesar la señal que ya fue previamente preprocesada con filtros, ahora se le aplica un análisis por ventanas de tamaño fijo. Para calcular la Transformada de Fourier, es necesario hacerlo con un número fijo de valores. Tomamos ventanas de 256 muestras y se calcula su transformada.

Matemáticamente se puede deducir que el número de ventanas para toda la señal de *"N"* cantidad de muestras es igual a N/20. El valor de *256* muestras por ventana, y el desplazamiento de las mismas de *20* muestras, es debido a la eficiencia del procesamiento. Lo ideal sería un desplazamiento de *"1"* muestra entre cada ventana, y hacer un traslape de casi el 100 %, pero el tiempo de procesamiento y la cantidad de espacio para almacenar

los resultados se incrementarían en gran medida. Los valores fueron fijados después de haber realizado algunas pruebas de tiempo y cálculos de espacio en memoria requeridos.

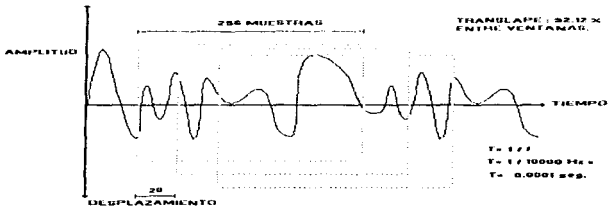

Figura 3.3 *Procesamiento de la señal de voz por medio de ventanas de análisis.*

El tamaño de la ventana esta relacionada con la definición de la *Transformada de Fourier*, entre más grande, se mejora los resultados de la Transformada, pero el tiempo de cálculo se incrementa. Se eligió el valor de 256 muestras por ventana al realizar algunas pruebas de desempeño de la transformada.

$2^7 = 128$ *MUESTRAS*	$2^8 = 256$ *MUESTRAS*	$2^9 = 512$ *MUESTRAS*
Poco Tiempo	*Poco Tiempo*	*Mucho Tiempo*
Poca Definición FFT	*Buena Definición FFT*	*Buena Definición FFT*

Con ventanas cortas se pueden capturar transiciones de voz y de silencio, con ventanas más grandes se destacan las transiciones más específicas de voz como son los diptongos y voces nasales.

Como se mencionó en los capítulo anteriores, al momento de segmentar en ventanas que se traslapan como se muestra en la figura 3.3, es un hecho que se presente una alteración de los resultados por efecto de las ventanas vecinas (*Efecto de Aliasing*). Por ello se utilizan las ventanas de Hamming que por sus características mostradas en la figura 3.4, evita este efecto no deseado.

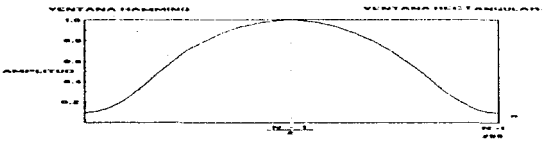

Figura 3.4 Ventana Hamming.

$$VENTANA\ HAMMING = 0.54 - 0.46\ COS\ (2\ \pi\ n\ /\ N-1) \qquad (3.4)$$

$$0 <= n <= N-1$$

n : n-ésima muestra en la ventana.

N : Número de muestras por ventana.

Al multiplicar la ventana original por la ventana de Hamming se obtiene una ventana con valores de la señal pero con propiedades de la ventana de Hamming.

3.2.2. FFT POR VENTANA.

Cada ventana tiene dimensión fija desde 0 hasta 255 muestras. Se les aplica el algoritmo de la transformada rápida de Fourier para obtener el espectro en frecuencia de cada ventana. El resultado de este proceso son dos arreglos de igual dimensión que la ventana original (256 muestras). Uno contiene la parte real de la transformada de Fourier y el otro la parte imaginaria. A nosotros solo nos interesa el espectro en magnitud, por lo que generamos otro arreglo con la magnitud de cada muestra, calculada con la siguiente ecuación:

$$MAGNITUD\ [\ i\ ] = \sqrt{REAL\ [\ i\ ]^2\ +\ IMAGINARIA\ [\ i\ ]^2} \qquad (3.5)$$

Teniendo los valores utilizados en el dominio del tiempo se puede determinar la distancia entre cada valor en el dominio de la frecuencia y también hasta que valor máximo alcanza la transformada de Fourier.Conocemos la cantidad de muestras por ventana *(256)* y la distancia en tiempo entre cada muestra *($1x10^{-4}$ segundos)*. La distancia entre cada valor de la transformada de Fourier, y la frecuencia máxima resultante, es según [Ramírez85] calculada por las ecuaciones (3.6a) y (3.6b).

$$\Delta f = 1 / N \Delta t = 1 / (256)(1x10^{-4} \text{ segundos}) = 39.06 \text{ Hz} \quad (3.6a)$$

$$\text{Frecuencia máxima} = 39.06 \text{ Hz} * (256 / 2) = 5000 \text{ Hz}. \quad (3.6b)$$

Se debe destacar que de los *256* valores de la Transformada, los primeros *128* son simétricos a los otros *128*, por lo que solo se debe tomar en cuenta los primero *128*. Es decir en nuestro análisis se obtiene el espectro desde *0 Hz* hasta *5000 Hz* que es el intervalo que caracteriza a las señales audibles de voz humana.

En el arreglo *MAGNITUD[i]* están almacenados los distintos valores de las aportaciones de cada frecuencia separados por *39.06 Hz* entre cada valor.

MAGNITUD [0] = *APORTACION DE LA FRECUENCIA 0 Hz*

MAGNITUD [1] = *APORTACION DE LA FRECUENCIA 39.06 Hz*

MAGNITUD [127] = *APORTACION DE LA FRECUENCIA 5000 Hz*

La definición frecuencial es aceptable para fines de procesamiento de señales de voz, ya que realmente no logramos percibir cambios de señales de *39 Hz* , pues percibimos cambios mucho mayores.

3.2.3. BANDAS CRITICAS.

Las señales originalmente procesadas tienen aproximadamente 10,000 muestras en promedio, lo que origina que se tengan de 500 a 1000 ventanas por cada señal de voz. Si a estas 1000 ventanas se les asocia 256 valores de su Transformada de Fourier se está hablando de 256000 celdas en memoria de números en punto flotante para almacenar el espectro en magnitud.

Una forma de reducir espacio sin disminuir información es agrupando el espectro en magnitud original en Bandas Críticas.

Banda 1 { Intervalo 1 de Frecuencias }

Banda 2 { Intervalo 2 de Frecuencias }

.

.

Banda 18 { Intervalo 18 de Frecuencias }

Con *18* bandas y *1000* ventanas como máximo se originan *18000* celdas que contendrán el arreglo del Espectro Frecuencial mapeado en bandas críticas. Una forma fácil para distribuir el Espectro en las distintas bandas es mediante un mapeo lineal. A cada Banda le corresponderán casi el mismo intervalo de frecuencias que agrupar.

El sistema auditivo humano tiene una pobre resolución en las altas frecuencias que para las bajas. Es recomendable que los sonidos sean preprocesados incrementando las propiedades de la resolución a las frecuencias del oído (Hertz a Bark). Esto es debido a que el oído no es igualmente sensitivo a la energía de distintas frecuencias [Shihua91] .

El reporte expuesto por [HeAlBrIr94] proponen y usan la expresión derivada empíricamente dada por (3.7) , cuya gráfica se muestra en la figura 3.5.

$$z = 13 \arctan(f/1316) + 3.5 \arctan(f/7500)^2 \qquad (3.7)$$

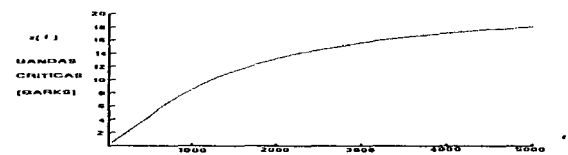

Figura 3.5 *Función propuesta por [HeAlBrIr94] para la distribución en bandas críticas.*

Donde *z* es en Bark y *f* en Hertz. Cada unidad en la escala de los Barks corresponde a una Banda Crítica. La forma simple es utilizar una función de distribución lineal, en donde a cada Banda Crítica le corresponda el mismo número de intervalos de frecuencias, pero no se toman en cuenta las propiedades de la percepción de la voz. Por esta razón utilizamos la distribución en bandas críticas aplicando la fórmula anterior.

Distribución en bandas críticas:

BANDA	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*
PUNTOS DE LA FFT	*3*	*3*	*3*	*3*	*3*	*3*	*3*	*4*	*4*	*5*	*5*	*6*	*7*	*8*	*10*	*13*	*15*	*30*

El resultado del procesamiento de bandas críticas es un arreglo BC[18] que contiene los valores promedios de los puntos de la FFT que le corresponden a cada Banda. Note la distribución no lineal debida a la teoría del sistema auditivo humano.

BC [0] { Promedio de los primeros 3 puntos de la FFT }
BC [1] { Promedio de los siguientes 3 puntos de la FFT }
.
.
BC [17] { Promedio de los últimos 30 puntos de la FFT }

En el diagrama de flujo de la figura 3.6 se observa el procedimiento general para determinar el espectro en bandas críticas de cada señal de voz .

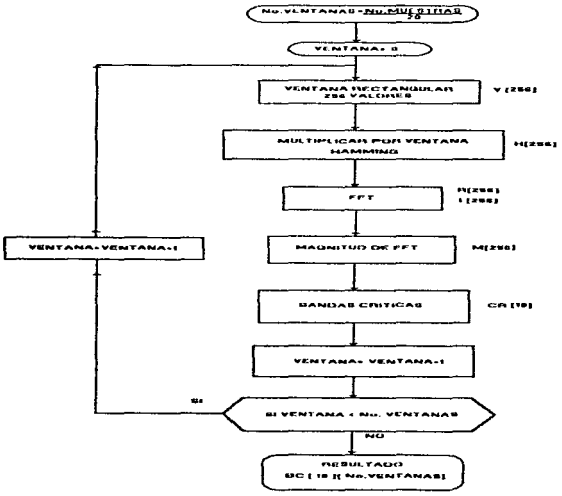

Figura 3.6 *Diagrama de flujo para la obtención del espectro en bandas críticas.*

El arreglo que es el resultado de cada ventana se almacena en otro que es bidimensional y que contendrá todo el espectro de la señal procesada.

$$BC[\,BANDA\,][\,NUMERO\ DE\ VENTANA\,] = MAGNITUD\ ESPECTRAL$$
$$BANDA = [\,0,\,1,\,2,\ldots\ldots\ldots,\,17\,]$$
$$NUMERO\ DE\ VENTANA = [\,0,\,1,\,2,\ldots\ldots,\,No.\ DE\ VENTANAS\ DE\ LA\ SEÑAL\,]$$

El resultado final del procesamiento descrito en la figura 3.6 es un arreglo bidimensional de 18 renglones por el número de ventanas de la señal de voz, y depende directamente del número de muestras de la señal a analizar. El ciclo para obtener las bandas críticas se hace procesando cada ventana a la vez. Primero se obtiene la ventana de las muestras originales de voz, tomando las primeras 256 muestras. Calculamos su transformada de Fourier y se obtiene la magnitud del espectro. Por último agrupamos las frecuencias en las *18* bandas críticas de forma alineal. Así sucesivamente hasta obtener el espectro de todas las ventanas agrupadas en el arreglo *BC[18][No. Ventanas]*. Este arreglo que contiene el espectro en bandas críticas está listo para los procesamientos de la segmentación acústica posterior.

3.3. COCIENTE DE MAXIMA SIMILITUD (MLR).

Para la detección de la voz y el silencio de la señal de voz, y también para la segmentación acústica, utilizamos como método de decisión el *cociente de máxima similitud (MLR)*.

Fundamentados en la teoría de decisión, nos basamos para generar las funciones necesarias que se utilizan para la discriminación. Un sistema básico de un simple problema de decisión es el propuesto en la figura 3.7 por [VanTrees68].

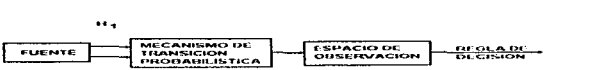

Figura 3.7 Componentes de un sistema de decisión .

Relacionando con las necesidades del método, es un problema clásico de decisión, pues las observaciones consisten de un conjunto N de números y pueden ser representados como puntos en un espacio N-dimensional [VanTrees68].

En un sistema clásico de decisión de la figura 3.7, existe en primer lugar una fuente cualquiera, en este caso la señal de voz, se generan dos hipótesis a probar (voz o silencio, inicio de segmento o no). Estas hipótesis son etiquetadas como H_0 y H_1. El mecanismo de transición probabilística se encarga de determinar que hipótesis es verdadera; basado en este conocimiento, se genera un punto en el espacio de observaciones acorde a alguna ley de probabilidad. Posteriormente con las reglas de decisión se da forma total al problema de decisión.

El criterio de decisión es la parte importante para decidir cual de las hipótesis es la verdadera; la forma más común es mediante el uso de una prueba del *cociente de máxima similitud* [VanTrees68].

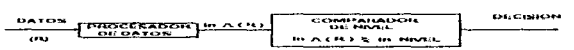

Figura 3.8 *Procesador del cociente de máxima similitud.*
Λ (R) : Cociente de Similitud Probabilistica.

$$\Lambda(R) = \frac{P_{rH1}\,(R / H_1)}{P_{rHO}\,(R / H_O)} = \frac{\textit{Probabilidad condicional de } H_1}{\textit{Probabilidad condicional de } H_O} \qquad (3.8)$$

El *criterio de Bayes* conocido como *Prueba del cociente de máxima similitud* es la relación (3.9)[VanTrees68] .

$$\Lambda(R) \underset{H_O}{\overset{H_1}{>\!<}} \textit{Nivel} \quad o \quad \ln \Lambda(R) \underset{H_O}{\overset{H_1}{>\!<}} \ln \textit{Nivel} \qquad (3.9)$$

Toda esta información se puede utilizar primero encontrando la función de probabilidad, acoplándola a la aplicación pertinente. Posteriormente se debe fijar un nivel de decisión *"n"* de acuerdo a la función antes generada.

Básicamente el algoritmo usa la prueba *MLR* como metodología de clasificación para detectar en el espectro de la voz los cambios en el espectro. Detecta las subpalabras empleando las ventanas deslizantes *MLR* en dos iteraciones. En la primera pasada, la señal entrante se divide en intervalos de voz e intervalos de silencio, mientras que en la segunda iteración la señal es segmentada acústicamente en subpalabras. La estructura de la señal de voz es usada para adaptar la estrategia de detección.

Asumimos por simplicidad que $C(n)$, el vector de bandas críticas para la ventana *"n"*, es de dimensión *"J" (Número de bandas críticas)*, Media igual a cero ($\mu = 0$), por las propiedades de las señales de voz y con variable aleatoria independiente de distribución normal $C(n) = N(Q, \Sigma)$. La generalización de la prueba MLR para las hipótesis H_0 y H_1 es la representada con las ecuaciones (3.10).

$$H_0 : \Sigma = \Sigma_0 \qquad H_1 : \Sigma \neq \Sigma_0 \tag{3.10}$$

Basandose de un segmento de muestras $(C(1), C(2), \ldots\ldots, C(N))$, con los componentes de Σ, $(\sigma_1^2, \sigma_2^2, \ldots\ldots, \sigma_J^2)$ desconocidos y los componentes de Σ_0 conocidos o de referencia. La función de probabilidad para el segmento de muestra, asumiendo independencia entre cada ventana es (3.11).

$$L(\mu, \Sigma) = (2\pi)^{-JN/2} |\Sigma|^{-N/2} \exp\left[-1/2 \sum_{n=1}^{N} C(n)^T \Sigma^{-1} C(n) \right] \tag{3.11}$$

Y el cociente de máxima similitud es (3.12).

$$\lambda = \frac{L(Q, \Sigma_0)}{\Sigma^{max} L(Q, \Sigma)} \tag{3.12}$$

El numerador es la función de probabilidad para (μ, Σ) con la restricción de los parámetros $(\mu = 0, \Sigma = \Sigma_0)$, y el denominador el máximo de la función de probabilidad restringido por los parámetros $(\mu = 0, \Sigma$ *valor positivo)*. El máximo valor de la prueba MLR ocurrirá cuando el estimador de máxima similitud Σ sea la variancia (3.13) de los segmentos muestreados.

$$\bar{\sigma}^2_j = 1/N \sum_{n=1}^{N} C^2_{jn} \qquad (3.13)$$

Insertando el estimador (3.13) en (3.11) y simplificando (3.12) , se genera la función para detectar cambios espectrales y diferencias entre regiones de la voz (3.14).

$$\Lambda = ln\,\lambda = \frac{N}{2}\left| \sum_{j=1}^{J} ln \frac{\bar{\sigma}^2_j}{\sigma^2_{oj}} - \sum_{j=1}^{J} \frac{\bar{\sigma}^2_j}{\sigma^2_{oj}} \right| \underset{H_1}{\overset{H_0}{><}} Nivel \qquad (3.14)$$

En donde en la ventana de análisis k se deriva la función (3.15):

$$\Lambda(\kappa) = \frac{N}{2}\left| \sum_{j=1}^{J} \left[ln \frac{\bar{\sigma}^2_j(\kappa)}{\sigma^2_{oj}} - \frac{\bar{\sigma}^2_j(\kappa)}{\sigma^2_{oj}} \right] \right| \qquad (3.15)$$

En donde $\bar{\sigma}^2_j(\kappa)$ representa la variancia muestreada de la banda crítica *j* en el paso *k*, σ^2_{oj} *es* la variancia de la banda crítica *j* de Σ_o, y el *nivel* es determinado experimentalmente de acuerdo a la aplicación específica. Mediante esta función podemos detectar variaciones o similitudes entre dos parámetros de prueba, que son las variancia de referencia contra la variancia de alguna muestra. Esta función generará un valor específico que indicará si se sobrepasa el nivel de decisión o no, para así poder comprobar alguna de las dos hipótesis.

3.4. DETECCION DE INICIO Y FIN DE VOZ.

Antes de segmentar la señal de voz, es necesario saber en que parte comienza y finaliza la señal de voz. Existen muchos métodos para la determinación de los mismos, pero aprovechamos las ventajas del espectro en bandas críticas, el cual refleja perfectamente las partes significativas de la voz.

La prueba MLR es aplicada secuencialmente a un segmento de tres ventanas del arreglo de bandas críticas mediante la ecuación (3.16).

$$Y(n) = [\, C(n-1),\ C(n),\ C(n+1) \,]$$

$$\bar{\sigma}^2{}_i(n) = 1/3 \sum_{k=n-1}^{n+1} C^2{}_{i,k} \qquad (3.16)$$

Variancia de un segmento de prueba.

La variancia (3.16) de este segmento de análisis se comparará contra la variancia muestreada del silencio (3.17), que se obtiene de las primeras 6 ventanas de toda la señal, las cuales se asumen como silencio.

$$\bar{\sigma}^2{}_{0i} = 1/6 \sum_{n=1}^{6} C^2{}_{i,n} \qquad (3.17)$$

Variancia de un segmento considerado como silencio.

Las dos hipótesis de la prueba *MLR* son:

$H_0(n) : \Sigma(n) = \Sigma_0$: *La variancia del segmento de prueba es similares a la del silencio.*

$H_1(n) : \Sigma(n) \neq \Sigma_0$: *La variancia del segmento de prueba es diferente a la del silencio.*

En la figura 3.9 se muestra el diagrama de flujo para obtener el cociente de máxima similitud entre el área de voz y la del silencio de la señal. Las dos variancias (3.16) y (3.17), se sustituyen en la ecuación (3.15) encontrada para la prueba MLR, y se realiza el cálculo hasta comparar toda la señal en grupos de 3 ventanas. Al final de este proceso se obtiene un arreglo de una dimension denominado *LAMBDA[i]* .

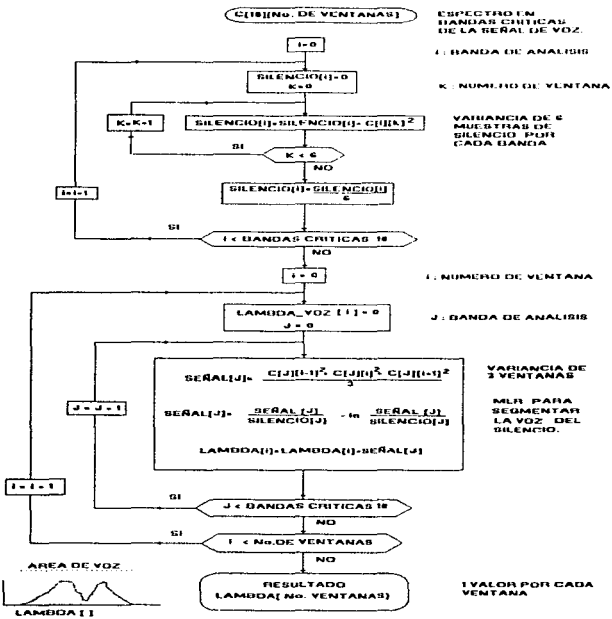

Figura 3.9 Cálculo de LAMBDA para detectar la parte de voz de la señal.

Cuando se compare una parte de la señal que contenga voz, contra el segmento de silencio, el valor de la prueba MLR resultará grande, en cambio cuando se compare silencio de la señal con el silencio de referencia la prueba MLR origina valores pequeños cercanos a la unidad.

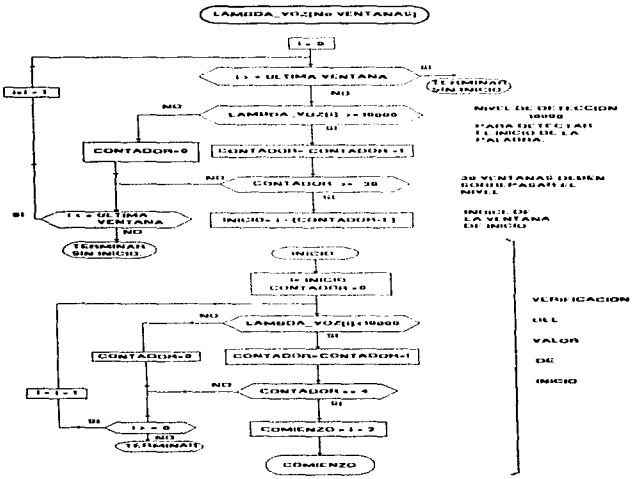

Figura 3.10 Detección y verificación del valor de la ventana inicial de voz.

En la detección, la función *LAMBDA* se compara con niveles de decisión al inicio de la señal de voz y al final de la misma como se observa en el diagrama de flujo de la figura 3.10. Al posicionarse en el valor inicial de LAMBDA, se hace un recorrido ascendente para encontrar el valor que supere el nivel prefijado.Además realizamos un proceso de verificación para asegurar el lugar de inicio. El nivel para aceptar el inicio de la palabra en nuestro algoritmo es cuando esta función sobrepase el valor de 10,000. El mismo valor se utilizó para la detección del final de la palabra, pero en este caso la función *LAMBDA* debe bajar de este nivel. El valor del nivel se obtuvo al promediar distintos valores observados de diferentes señales de voz.

La función *LAMBDA[i]* sirvió intencionalmente también para detectar segmentos de silencio intercalados en la señal de voz, aprovechando los cálculos y haciendo más eficiente nuestro método. La figura 3.11 representa un ejemplo general característico de la función *LAMBDA[i],* donde de distinguen las distintas áreas de selección de la voz.

Figura 3.11 *Forma general de LAMBDA[i] para la detección de silencio y voz de cada palabra.*

La detección del punto final de voz de la figura 3.12 es similar al de inicio, las únicas diferencias son el recorrido desde el final de la señal de voz, hacia adelante o los primeros valores.

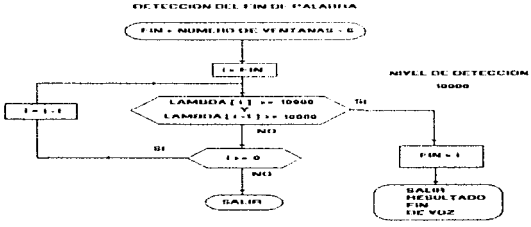

Figura 3.12 Detección del punto final de cada palabra.

Al final de este proceso se obtiene el valor de dos puntos que acotan la señal de voz, estos puntos son INICIO y FIN. Posteriormente formarán parte de la segmentación pues serán los marcos de nuestro análisis.

3.5. DETECCION DE SILENCIO DE LA SEÑAL.

Si se han encontrado anteriormente los valores de inicio y fin de la señal de voz, esto no excluye que dentro de estos intervalos no existan segmentos con poca aportación energética. Estos segmentos de silencio también forman parte de las palabras .

Aunque no todas las palabras tienen segmentos de silencio, algunas presentan más de uno intercalado entre toda la palabra. Nosotros para su detección aprovechamos la función *LAMBDA[i],* calculada para la detección de inicio y fin de voz, y la utilizamos también para detectar los segmentos de silencio. La función *LAMBDA [i]* presenta valores muy bajos en las partes de poco aporte energético, visibles en la figura 3.11.

Lo único que almacenamos en el proceso son los índices de inicio y fin del silencio, de esta forma como no estamos procesando la señal original, solo se necesita saber los valores de las acotaciones del silencio para la segmentación. Los primeros segmentos acústicos que encontramos son los de silencio. En el diagrama de flujo de la figura 3.13 simplemente hacemos un recorrido desde el INICIO de cada palabra hasta encontrar que *LAMBDA[i]* disminuya el valor de 10,000; posteriormente *LAMBD[i]* lógicamente debe ascender el mismo valor, si es que es un segmento de silencio. Además el silencio deberá durar cierto tiempo para que sea de consideración. Etiquetamos con *SILENCIO* al índice de comienzo de silencio y con *VOZ* a la terminación del mismo, para su posterior almacenamiento.

La verificación tambien mostrada en la figura 3.13, se refiere a que después de encontrar el segmento de silencio, este debe descender el valor de 1000 , que es un valor muy bajo para ser considerado como silencio legítimo. Los niveles de decisión se ajustaron de acuerdo a varias muestras de palabras procesadas y se promediaron tales valores.

Cabe señalar que estos segmentos pueden ser detectados en la etapa posterior de segmentación acústica, pero simplemente solo sabríamos que es un segmento como cualquier otro, y no se identificaría como silencio. Quizá para algunas aplicaciones es bueno saber en donde están las partes de silencio de cada palabra.

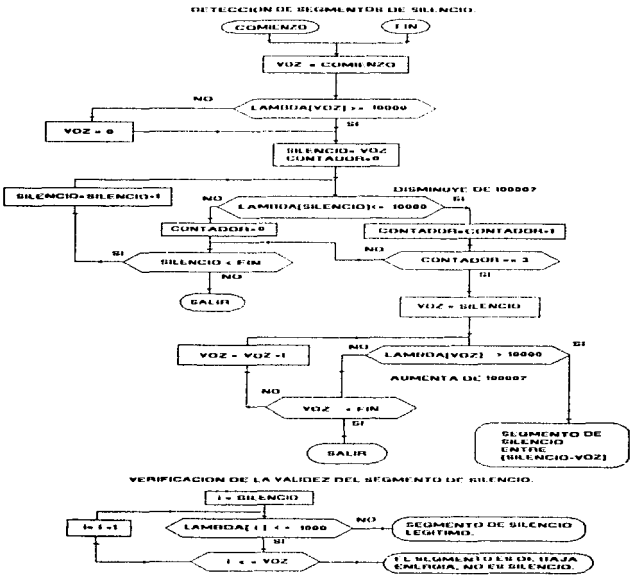

Figura 3.13 *Diagrama de flujo para la detección de segmentos de silencio y su verificación.*

3.6 SEGMENTACION ACUSTICA.

En este momento del proceso tenemos algunos datos que facilitarán la segmentación acústica. En primer lugar está un arreglo que contiene el espectro en frecuencias, agrupado en bandas críticas *(BC [18][No. Ventanas])*. Es la forma simplificada que conserva toda la información en el dominio de la frecuencia de la señal de voz. Además contamos con los valores de *INICIO* y *FIN*, que indican donde comienzan y termina la señal de voz. También tenemos etiquetados los segmentos de silencio, que posteriormente se contarán como parte de la segmentación.

Un segmento acústico para los fines de este trabajo es una región en el espectro frecuencial que presenta casi las mismas características. Es decir, es un periodo de tiempo que tiene pocos cambios en las frecuencias. Estos segmentos pueden representar a cada palabra caracterizándola muy bien, y tienen una relación directa con los propiedades acústicas de cada palabra.

El objetivo del procesamiento es encontrar todos los segmentos posibles dentro de la señal de voz. Esto se obtendrá por medio del cálculo de una nueva función *LAMBDA* que en base a la prueba *MLR*, se genera a partir del arreglo de bandas críticas. Con esta nueva función se detectarán tales segmentos comparándola con unos niveles prefijados y que se calculan posteriormente.

Se obtendrán varios segmentos de cada palabra, pero solo algunos son los más significativos, por esta razón los clasificamos y eliminamos los menos importantes. Una de las ventajas de este método de segmentación es que puede fijarse el intervalo de segmentos aproximados para cada palabra. Esta es una ventaja para aplicaciones posteriores, pues con un valor fijo de segmentos, facilitará las comparaciones con los patrones de referencia que también tendrán el mismo número de segmentos representativos.

En la sección donde se seleccionan los segmentos acústicos, es donde se aplica mayor sentido común y conocimientos de los espectros de voz. Pero al final se obtendrán los valores de los índices donde comienzan y terminan cada segmentos.

El proceso completo que utilizamos para realizar la segmentación acústica es el que se muestra en el siguiente diagrama de flujo de la figura 3.14.

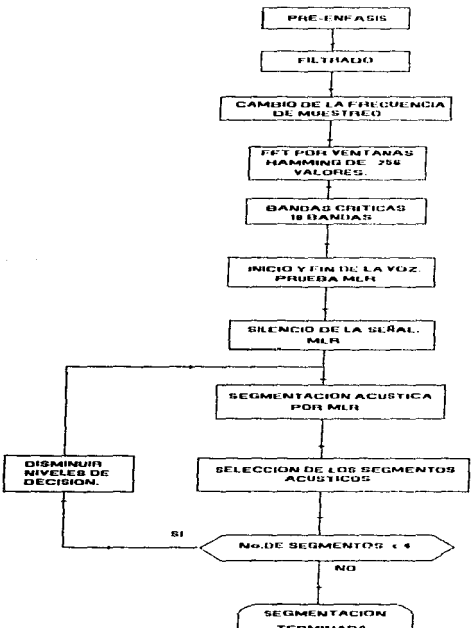

Figura 3.14 Método completo para la segmentación acústica.

En la práctica encontramos que algunas palabras no tienen tan marcados los segmentos acústicos , y para detectarlos es necesario disminuir los niveles de decisión para encontrar estos pequeños cambios. Esta característica la notamos al procesar varias señales de voz, por esta razón añadimos esta función al método, que es la de bajar automáticamente los niveles en caso de que la palabra analizada no tenga bien marcados sus cambios espectrales. Esta propiedad de selección automática de niveles vuelve al método muy eficiente en cuanto a forzarlo a encontrar todos los posibles segmentos acústicos de cada señal de voz.

3.6.1.FUNCION MLR PARA SEGMENTOS ACUSTICOS.

El objetivo de todo segmentador es que sea lo más simple y rápido posible. Generamos una función simple que nos ayuda a mostrar los cambios espectrales, y que además es fácil de interpretar. Para encontrar las diferencias entre los segmentos acústicos, aprovechamos la estructura natural de la voz, en la que los cambios espectrales son precedidos de múltiples ventanas de un valor estadístico y seguidos por múltiples ventanas con diferente valor estadístico. La prueba MLR es aplicada secuencialmente a un segmento de 5 ventanas:

$$Y(n) = [\, Z(n-1), Z(n), Z(n+1) \,] \qquad (3.18)$$

$$Z(n-1) = [C(n-2),\ C(n-1)]$$

$$Z(n) = [C(n)]$$

$$Z(n+1) = [C(n+1), C(n+2)]$$

Haciendo la analogía con la fórmula de la *prueba MLR* (3.15):

$H_0\ (n) : \Sigma(n) = \Sigma_0$:La variancia del segmento anterior a "n" es similar a la del segmento posterior.
$H_1\ (n) : \Sigma(n) \neq \Sigma_0$: La variancia del segmento anterior a "n" es diferente a la del segmento posterior.

$$\Lambda(n) = \frac{N}{2} \left| \sum_{j=1}^{J} \left[ln\ \frac{\bar{\sigma}^2_j\ (n)}{\bar{\sigma}^2_{0j}\ (n)} - \frac{\bar{\sigma}^2_j\ (n)}{\bar{\sigma}^2_{0j}\ (n)} \right] \right| \qquad (3.19)$$

LAMBDA(k) es de la forma $| \ln x - x | = x - \ln x$

Donde $\bar{\sigma}^2_{o_i}(n)$ es la variancia muestreada del segmento $Z(n-1)$, se calcula con la fórmula (3.20), y $\bar{\sigma}^2_i(n)$ cuya fórmula es (3.21), es la variancia respectiva del segmento $Z(n+1)$. O sea son las variancias de las ventanas anteriores y posteriores a un punto de análisis que es $Z(n)$. Y equivale a analizar una columna completa del arreglo de bandas críticas.

$$\bar{\sigma}^2_i(n) = 1/2 \sum_{k=n+1}^{n+2} C^2_{jk} \qquad (3.20)$$

$$\bar{\sigma}^2_{o_i}(n) = 1/2 \sum_{k=n-2}^{n-1} C^2_{jk} \qquad (3.21)$$

La prueba *MLR* en cada paso *"n"*, detecta si ocurren transiciones espectrales en el punto de análisis. Es decir si las variancias de las ventanas anteriores Z(n-1), difiere de la variancia de la ventana subsecuente $Z(n+1)$, entonces el resultado de la prueba MLR es un número mayor a *"1"*. Es decir existirá una función *LAMBDA[n]* que almacene las diferencias en cada punto. Donde existan *cambios espectrales se reflejarán en esta función, pues las transiciones espectrales son caracterizadas por valores altos de LAMBDA*. La implantación del proceso para calcular la función *LAMBDA[n]* se muestra en la figura 3.15.

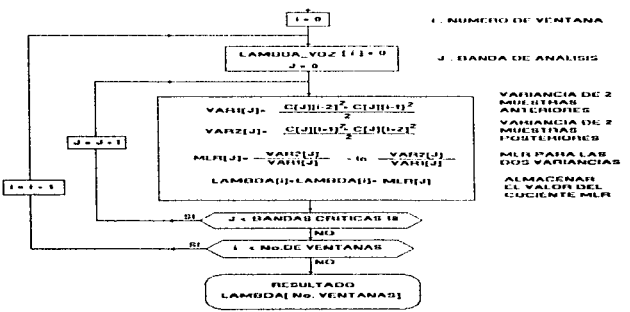

***Figura 3.15** Cálculo de LAMBDA para detectar los segmentos acústicos.*

La función *LAMBDA[n]* tiene la forma *(n - ln n)* mostrada en la figura 3.16, y es básica en la *prueba MLR,* pues nos da la pauta para el buen funcionamiento del segmentador. Donde realmente *"n"* es una comparación del espectro anterior a un punto con el espectro después del punto de análisis. Si existe alguna diferencia se reflejará en la función de prueba. Siempre generará valores positivos mayores a *"1"*. Sirve para valorar las diferencias espectrales, resultando un valor mayor a *"1"* si la diferencia es significativa. En caso de que la comparación de los espacios espectrales sea "1" la respuesta de la función permanece también en *"1, LAMBDA* dice realmente en donde hay cambios en el espectro y la utilizamos para la segmentación acústica.

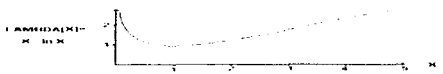

Figura 3.16 Función para acentuar los posibles cambios espectrales.

3.6.2.NIVELES DE DECISION.

El arreglo *LAMBDA[i]* contiene los valores de las comparaciones en cada punto de la señal, basada en el espectro en b*andas críticas* (18 bandas por punto de análisis). En caso de que en uno o más bandas críticas existiese una diferencia, la función de comparación de variancias originaría que el valor resultante sea mayor a *18* como se observa en la figura 3.17.

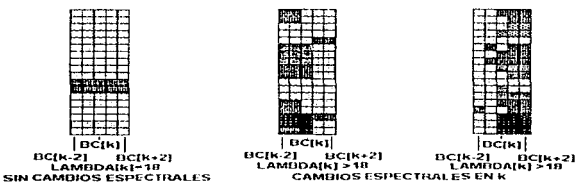

Figura 3.17 Prueba MLR para las 18 bandas críticas del espectro.

El nivel de decisión lo fijamos a partir de 20. Es notorio que cuando existan grandes cambios en los puntos de análisis el valor de *LAMBDA* será mucho mayor a *20*. A nosotros solo nos interesa detectar todos los cambios, por lo que fijamos el *nivel en 20*, aunque obtengamos en primer termino bastantes segmentos por palabra; posteriormente eliminamos los menos significativos tomando en cuenta los valores de la función de probabilidad.

3.6.3. DETECCION DE LOS SEGMENTOS ACUSTICOS.

Como se mencionó anteriormente, *LAMBDA[i]* contendrá los posibles cambios del espectro de la señal. Generalmente son impulsos a lo largo de toda la señal. El método de segmentación se encarga de encontrar estos puntos y asignarlos como acotadores de los segmentos acústicos. Entre dos puntos de cambio existe por regla un segmento acústico, una región con casi las mismas propiedades en la frecuencia.

La parte de detección se simplifica al manipular solo en *LAMBDA* para la elección de los segmentos acústicos. Se trata de encontrar todos los impulsos que sobrepasen cierto nivel fijado en 20 por las causas antes vistas. En la figura 3.18 se observa el proceso de detección de los impulsos en *LAMBDA* y donde existe una relación directa entre los índices de la señal original de voz y los índices de la función *LAMBDA*, entonces donde se encuentren los picos, ahí están los cambios de segmentos en la señal original.

Figura 3.18 *Función LAMBDA característica de una señal de voz para detección de segmentos.*

Un impulso con valor alto significa que en ese punto existe un cambio visible en las regiones espectrales, es decir, en ese lugar se puede distinguir auditivamente el cambio de segmento acústico. Los impulsos anchos indican un cambio espectral no muy notorio, y

más lento en el tiempo, y por lo tanto no es tan fácil distinguirle auditivamente. Pueden existir combinaciones de las características de los impulsos o crestas, pero un parámetro que no falla como criterio de selección es su magnitud. Las crestas más grandes son las que consideramos como principales puntos de cambio espectral de las palabras.

El procedimiento general para la detección de segmentos acústicos es el siguiente, y en la figura 3.19 mostramos el diagrama de flujo de este proceso.

1. - **ENCONTRAR EL LUGAR DONDE LAMBDA INCREMENTE NIVEL.** Avanzando en el arreglo LAMBDA se compara con el valor de NIVEL esperando que exista un pico mayor a 20.
2. - **ESPERAR A QUE LAMBDA DESCIENDA DE NIVEL.** Después de localizar un punto de ascenso, debe existir posteriormente un descenso de NIVEL, en caso de no llegar al fin de la palabra. En este punto se tienen dos valores que indican las cotas de un pico.
3. - **PARTE DE ESTABILIDAD.** Después de un pico generalmente se presenta una parte estable que no presenta otros picos o variaciones. Deberá existir esta parte estable aunque sea muy corta para considerar el pico válido.
4. - **DETECCION DEL VALOR DE LAMBDA MAXIMO.** Encontramos el valor máximo en la cresta antes detectada. Solo interesa el valor más grande del intervalo. Este valor servirá posteriormente para la discriminación de segmentos no significativos.
5. - **ALMACENAR LAS ACOTACIONES DE LOS SEGMENTOS.** Cada vez que se encuentra un punto de cambio de segmentos válido, debe ser almacenado para su posterior selección.
6. - **ASCIENDE NIVEL.** Después de cada pico, por lógica deberán existir otros más antes de encontrar el fin de la palabra. Entonces esperamos un ascenso de NIVEL que indique otro pico.
7. - **VERIFICACION DE FIN DE PALABRA.** Si no se a llegado al final de la palabra continuamos buscando más picos regresando al número 2.
8. - **GUARDAR TODOS LOS PUNTOS DE CAMBIO.** Al final del proceso se tendrá un arreglo conteniendo los puntos de cambio con sus respectivos valores de LAMBDA[ ] que es el valor que los difiere de los demás (El arreglo es SEG[No. de segmentos]). En este punto se tienen seleccionados todos los segmentos acústicos y falta la selección de los de mayor amplitud.

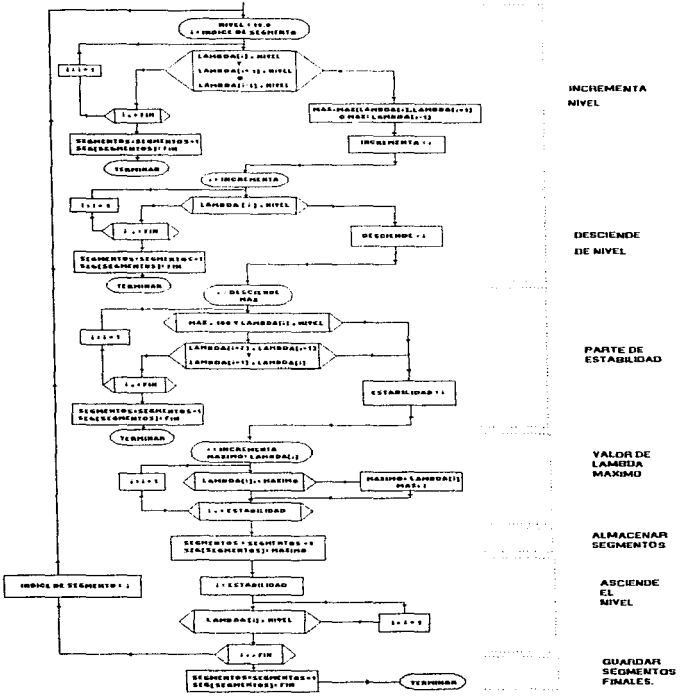

Figura 3.19 Diagrama general del método de detección de segmentos acústicos.

3.7. CRITERIOS DE SELECCION DE LOS SEGMENTOS ACUSTICOS.

Existe la necesidad de limitar el intervalo de segmentos encontrados. En esta parte del proceso existen todos los segmentos posibles, y basados en la experiencia de las observaciones y por las características de las palabras , generamos algunos parámetros y criterios de selección de los mejores segmentos:

- Entre mayor sea el valor de *LAMBDA[ ]* en el punto de análisis, significa que es donde existe un cambio más notorio del espectro de voz. Considerando esto, se seleccionan los de mayor valor.

- Existen puntos de cambio muy juntos. Estos picos juntos son reflejo de un mismo cambio en el espectro, debido a que nuestro análisis es cada 20 muestras de la señal original y no es en cada muestra, por lo que en pequeños espacios de análisis pueden existir varios picos que sean parte de un mismo lugar de cambio. La medida que tomamos es seleccionar el pico de mayor valor cuando se encuentren varios en un espacio muy reducido de tiempo.

- Los segmentos de silencio encontrados en procesamientos anteriores también se consideran parte de la señal. Los incrustamos como segmentos acústicos de gran peso.

- Si el número de segmentos al final es menor a un número prefijado, o son muy pocos, quizá sea necesario bajar el nivel de decisión para encontrar cambios aunque sean muy pequeños.

- Una sola palabra tiene múltiples segmentos acústico, pero solo consideramos aquellos que son notorios y caracterizan a cada palabra, para tratar de cumplir el objetivo de limitar el número de segmentos acústicos.

Todas estas características de selección tratan de ser una generalización para cualquier tipo de palabras separadas. Tal vez tengan que adaptarse este tipo de criterios para algún tipo de aplicación especial, pero para los fines de este trabajo se utilizaron los criterios anteriores.

3.8. GENERACION DE RESULTADOS.

En cierta forma el segmentador acústico es solo una parte de algunos sistemas reconocedores de voz, pero de alguna manera necesitamos ver el desempeño de nuestro método aplicando la segmentación a algunas palabras separadas.

Utilizamos la base de datos de *Texas Instruments Inc. (TI-46)*. Son palabras separadas de dígitos en Ingles, del número cero al nueve. Usamos en total 2600 palabras para la prueba del método, de las cuales originalmente la base de datos las tiene organizadas para alguna aplicación de reconocimiento de la siguiente forma: 1000 repeticiones para entrenamiento (10 dígitos, 10 parlantes diferentes y 10 repeticiones de cada dígito), 1600 repeticiones para clasificación (10 dígitos, 10 parlantes y 16 repeticiones de cada dígito). Pero a nosotros solo nos interesa el desempeño del segmentador, no tomamos en cuenta la distribución de tales palabras, segmentamos las 2600 solo considerando cuales son las del mismo dígito para obtener una generalización de los números de segmentos característicos de cada número.

Cada palabra se proceso de la misma forma con el método propuesto en este capítulo y se generaron los archivos resultantes de la segmentación. Las palabras están muestreadas a una cierta tasa, por lo que en el tiempo las caracterizan el número de muestras de cada una. Para indicar en que puntos está segmentada la palabra, lo único es guardar el número de la muestra que se encontró como punto de cambio. Es decir, como resultado general obtenemos un archivo por cada palabra que nos dice en que punto se inicia la voz y en que punto termina, además entre estos valores se guardan los puntos de cambio de cada segmentos.

Forma general de una archivo con los resultados de la segmentación:

```
1600    "Punto de Inicio"
2700
3600                        Segmentos = No. de Valores - 1 = 6 - 1 = 5 Segmentos.
4000
7000
7500    "Punto Final".
```

Con esta información es suficiente para indicar la segmentación completa, además no se altera la señal de voz original y los resultados son generados en poco espacio de memoria del computador. En paralelo con la generación de estos archivos , también se obtiene el archivo que almacena el espectro en bandas críticas de la señal, pero debido a que conserva solo 18 bandas críticas el espacio por cada espectro es pequeño.

Programé el método de segmentación acústica en el lenguaje *ANSI "C"* y Las pruebas del proceso las realicé en la estación de trabajo de *Sun (SPARCstation5)*. Aunque debido a que el programa está en *ANSI "C"*, puede funcionar en computadoras personales realizando algunas pequeñas modificaciones al mismo. Los archivos de resultados son generados por el programa, pero me apoye del programa *GNUPLOT (version 3.5 para unix)* el cual sirve para realizar las graficas de funciones mediante algunos comandos. *GNUPLOT* lo utilicé para desplegar algunas muestras gráficas de palabras segmentadas por nuestro método. (Ver capítulo 4).

Los espectrogramas son derivados de un programa que tambien realicé en lenguaje "C" en la estación de trabajo de *Sun*, cuya función principal es la de desplegar en forma gráfica, el espectro de la señal procesada con sus respectivos divisiones de segmentos. (Ver capítulo 4).

CAPITULO 4.

Presentación De Resultados.

CAPITULO 4.

PRESENTACION DE RESULTADOS.

INTRODUCCION.

Verificar el método de segmentación con las 2600 palabras almacenadas resulta una tarea complicada de realizar, pues no existe un patrón real de comparación en cuanto a los resultados de la segmentación de cada una de las palabras, ya que cada una genera resultados diferentes. Aplicando la información incluida en este trabajo, generamos un programa que segmenta acústicamente cada palabra, sin importar cual sea de las 2600, detectará como mínimo 4 segmentos, los más representativos y además tiene la opción de selecionar un número fijo de segmentos.

Los resultados de la detección de inicio y fin, no podemos verificarlos para todas las palabras procesadas, pero nos basamos en los resultados obtenidos en la tesis [RamYam96], donde evaluan distintos métodos de detección de inicio y fin de señales de voz. En su trabajo comprueban que el método MLR ofrece buenos resultados, en comparación de los demas métodos probados en ese trabajo. En este capitulo mostramos algunos ejemplos de la detección de inicio y fin de algunas palabras por el método MLR.

De igual forma, comprobar cada una de las palabras es una taréa muy laboriosa, por esta razón trataremos de mostrar, mediante algunos ejemplos representativos los resultados de nuestra segmentación acústica. Realmente no existe una prueba que garantice el buen desempeño del método de segmentación, debido a la gran variabilidad de la voz y a la gran cantidad de palabras que procesamos, pero de alguna forma nuestro trabajo presenta un fundamento teórico importante del que nos basamos para generar nuestros resultados.

4.1 . DETECCION DE INICIO Y FIN DE PALABRAS.

Los valores tanto de *inicio* y *fin* de las palabras, son valores que indica el número de muestra de la señal en donde comienzan y terminan las partes significativas de la voz, dejando fuera las partes de silencio que cada palabra presenta antes y después de la misma. El método que utilizamos es el cociente de máxima similitud explicado en el capítulo 3, que aunque no se garantiza la exactitud al determinar los puntos de inicio y fin de las 2600 palabras, es el que aporta mejores resultados que otros métodos [RamYam96].

Presentamos dos tipos de gráficas, los espectrogramas y la señal en magnitud contra número de muestra. Las partes claras de los espectrogramas es donde se considera poca energía y realmente es la región no audible para los humanos. En los dos tipos de gráficos se observan las acotaciones del inicio y fin de cada palabra etiquetadas con 'I' y 'F' respectivamente. Destaca la importancia de determinar la parte de *inicio* y *fin* de cada palabra, para así eliminar todas las partes de silencio que no forman parte de la señal de voz. A continuación mostramos algunos ejemplos en los que se observa la determinación de los puntos de *inicio* y *fin* de cada palabra.

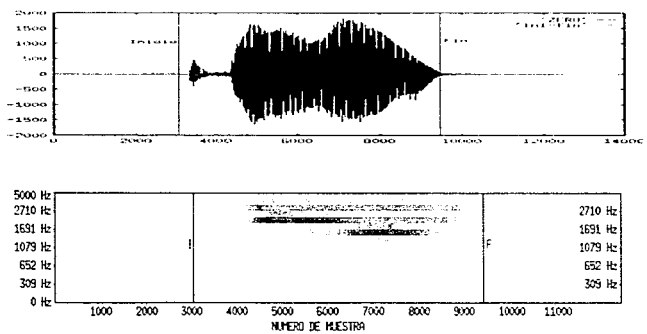

***Figura 4.1** Gráfica de amplitud y espectrograma de la palabra "zero" con inicio y fin.*

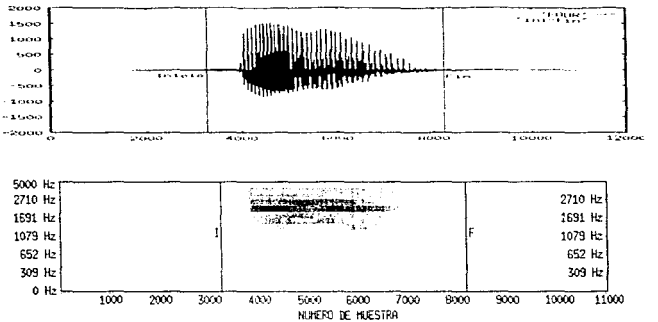

Figura 4.2 Gráfica de amplitud y espectrograma de la palabra "four" con inicio y fin .

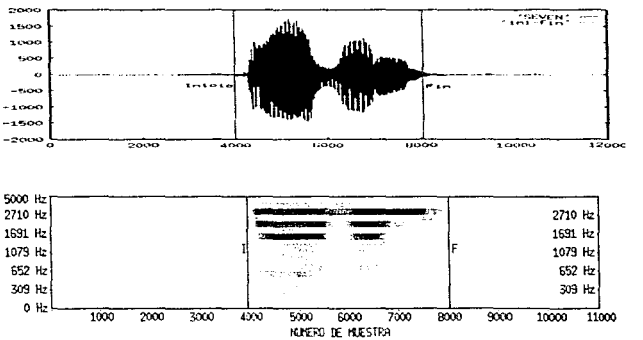

Figura 4.3 Gráfica de amplitud y espectrograma de la palabra "seven" con inicio y fin .

4.2 . SEGMENTACION ACUSTICA.

El objetivo final de nuestro trabajo es dividir cada palabra en sus respectivos segmentos acústicos, tomando en cuenta que estos sean los más representativos e importantes de cada una de las palabras. Nos basamos en la prueba MLR para fundamentar una segmentación lo más general posible, que sirva para procesar cualquiera de las palabras con las que probamos el método. Tomando en cuenta que la voz tiene características muy variables, pues depende del proceso de producción de la misma y de múltiples factores ambientales, crear un método general bien fundamentado que segmente eficientemente las palabras no es una labor sencilla. Por ello tratamos de crear un programa que abarcara muchas de las propiedades de la voz y que no solo funcionara con una sola palabra, sino que fuera lo más general posible.

En esta sección se muestran los resultados que se obtuvieron para algunas palabras en particular, en las que se ilustra la forma en que logramos generalizar el procedimiento de segmentación acústica. Para ello nos auxiliamos de las gráficas de las señales en amplitud contra número de muestra, las gráficas de la función MLR y de los espectrogramas.

Es evidente, que una región homogénea en los espectrogramas tiene una relación directa con algún segmento acústico, que a su vez lo podemos relacionar con un valor lingüístico, como lo tratamos de ilustrar en las gráficas, donde se puede observar la relación entre cada segmento acústico y su valor lingüístico. Cada grupo de espectrogramas ilustran las etapas de discriminación de los segmentos acústicos realizada automáticamente por nuestro programa.

Los espectrogramas son una herramienta útil para poder visualizar los distintos segmentos acústicos, lo que no siempre puede observarse en las gráficas de magnitud contra número de muestra, y en ocasiones, tampoco son distinguibles exactamente auditivamente. La palabra "seven" mostrada en la figura 4.4 es un ejemplo evidente de señal en la que se pueden observar sus segmentos acústicos en la señal de amplitud contra número de muestra, lo que no sucede con la palabra "nine" de la figura 4.12, en donde es más difícil encontrar los cambios entre cada segmento en la señal. Pero en cambio con la ayuda de los espectrogramas resultan más evidentes las transiciones acústicas.

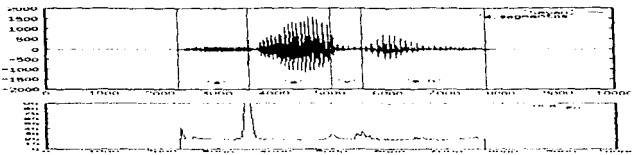

Figura 4.4 *Segmentación acústica final y función MLR de la palabra "seven".*

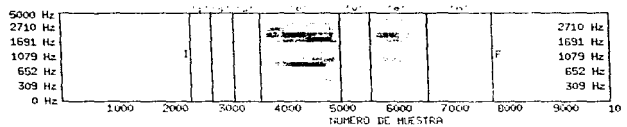

Figura 4.5 *Segmentación acústica de la palabra "seven"sin limitación de número de segmentos.*

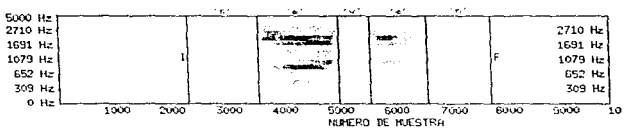

Figura 4.6 *Segmentación acústica de la palabra "seven" limitada en 5 segmentos.*

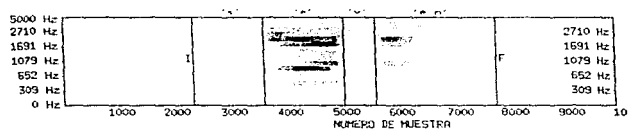

Figura 4.7 *Segmentación acústica final de la palabra "seven" limitada en 4 segmentos.*

Existe una relación directa entre las crestas de la función MLR y los cambio espectrales, como se observa en las figuras 4.4,4.5,4.6 y 4.7. Además se distingue la evolución de la selección de los segmentos acústicos importantes de la palabra.

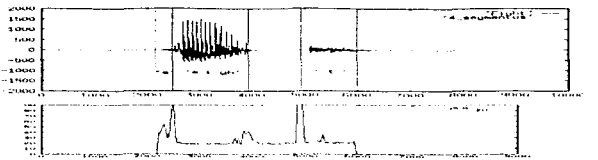

Figura 4.8 Segmentación acústica final y función MLR de la palabra "eight".

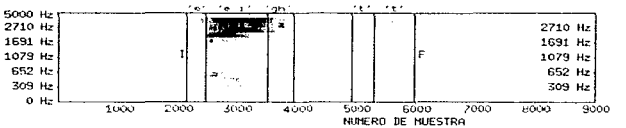

Figura 4.9 Segmentación acústica de la palabra "eight" sin limitación de número de segmentos.

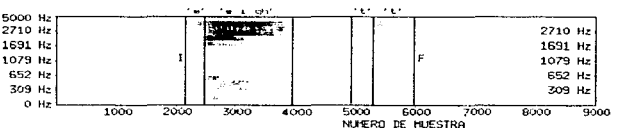

Figura 4.10 Segmentación acústica de la palabra "eight" limitada en 5 segmentos.

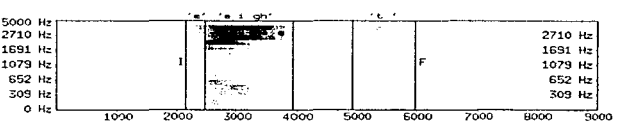

Figura 4.11 Segmentación acústica final de la palabra "eight" limitada en 4 segmentos.

Note que en la segmentación final en la figura 4.11, se mantienen las regiones más homogéneas en el espectro, equivalen a una región con propiedades acústicas similares.

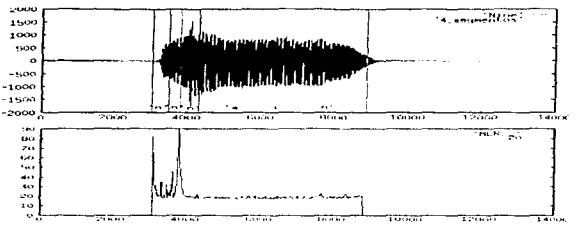

Figura 4.12 Segmentación acústica final y función MLR de la palabra "nine".

En la figura 4.12 se observa que para la primer parte de la pronunciación de la palabra inglesa "nine" , cuya correspondencia acústica es 'nain', existe una división de la primer 'n' en tres segmentos, y esto quizá no resulte evidente en la señal en magnitud contra número de muestra, pero en la figura 4.13 , que es el espectrograma de la misma palabra, se distingue que realmente existen tres regiones diferentes que conforman a toda la pronunciación de la letra inicial 'n'. Es decir, nuestro método no segmenta en fonémas o sílabas, sino que divide cada palabra en regiones auditivas similares, como se observa en la evolución de las figuras 4.14,4.15 y 4.16.

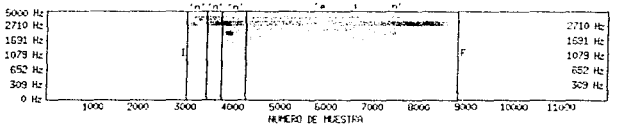

Figura 4.13 Segmentación acústica final de la palabra "nine" limitada en 4 segmentos.

Además en el resultado final de las figuras 4.12 y 4.13, se tiene como segmento acústico a la parte pronunciada como 'ain', de la misma palabra "nine", y esto es debido a que realmente es una región similar acústica.

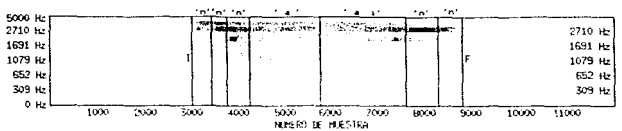

Figura 4.14 *Segmentación acústica de la palabra "nine" sin limitación de número de segmentos.*

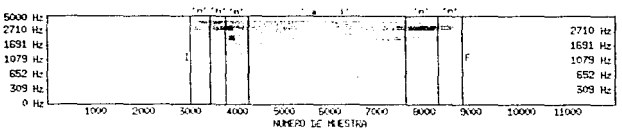

Figura 4.15 *Segmentación acústica de la palabra "nine" limitada en 6 segmentos.*

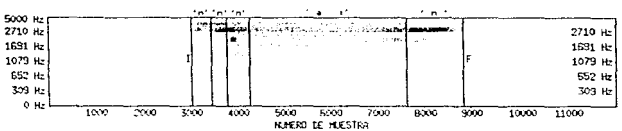

Figura 4.16 *Segmentación acústica de la palabra "nine" limitada en 5 segmentos.*

La selección de los segmentos acústicos mostrada en la secuencia de las figuras 4.14,4.15 y 4.16, se realizan automáticamente tomando como referencia a la función MLR característica de la misma palabra "nine" mostrada en la figura 4.12. Nosotros observamos la relación directa entre la magnitud de las crestas de dicha función y los cambios espectrales de las palabras, como una característica importante al momento de generalizar el procedimiento de segmentación acústica general.

CAPITULO 5.

Posibles Aplicaciones.

CAPITULO 5.

POSIBLES APLICACIONES.

INTRODUCCION.

Quizá nuestra investigación sobre los *segmentos acústicos* , y su detección no sean precisamente una aplicación inmediata en el reconocimiento automático de voz, pero estoy convencido de que si representa el fundamento para realizar distintos sistemas de reconocimiento completo.

En este capítulo tratamos de dar ideas de como utilizar nuestro segmentador acústico de distintas formas. Para poder entender su importancia, doy una breve descripción de distintos enfoques generales de sistemas reconocedores de voz, que en la actualidad se usan. Complementando con técnicas también usadas, tratamos de compatibilizar estas técnicas con nuestro subsistema. En la parte donde más se observa la real utilización del segmentador es en los sistemas de reconocimiento de palabras separadas. Mencionar otras ideas de este tipo de sistemas hacen ver en que lugar se ubica a nuestro segmentador acústico.

Pero la idea que nosotros aportamos la describimos en la planificación de la posible aplicación, en donde explico más detalladamente como podría realizarse un sistema simple de reconocimiento automático de voz de palabras separadas. Basado obviamente en nuestro segmentador, y tomando ideas de las técnicas antes descritas en este capítulo.

Por último se hacen algunas especulaciones de las direcciones de las aplicaciones de reconocimiento de voz, fundamentado en las investigaciones y lecturas que involucra la realización de este trabajo. Tal vez parezcan muy idealistas las ideas de este capítulo , pero llevarlo a la práctica sería cuestión de mucho más tiempo y por lo cual no es el objetivo de esta tesis.

5.1. CLASIFICACION DE LOS SISTEMAS RECONOCEDORES DE VOZ.

Los sistemas de reconocimiento de voz han tenido distintos enfoques de realización; depende de las posibles variantes que rodean al sistema. Pero básicamente se clasifican tomando en cuenta lo siguiente:

- *Persona Parlante:* Específica o no Específica.
- *Patrones De Referencia* . (Unidades de reconocimiento): Palabras, fonemas, segmentos acústicos, frases, etc.
- *Tipo De La Voz A Procesar*: Discreta (Palabras separadas) o voz continua.
- *Medio En Que Se Habla*: Cuarto de la computadora, lugar público, etc.
- *Sistema De Transmisión*: Micrófono de gran calidad, teléfono, etc.
- *Tipo De Entrenamiento* : Sin entrenamiento, entrenamiento continuo, previo entrenamiento, etc.
- *Tamaño Del Vocabulario* : Pequeño (1-20 palabras), Mediano (20-100 palabras), Vocabulario alto (Mayor a 100 palabras).

Combinando cada una de las variantes, se pueden generar muchas posibilidades de los sistemas de reconocimiento. Lo ideal de un sistema reconocedor de voz es que no discrimine a la persona que habla, y que con cualquiera reconozca la frase hablada, claro, también depende de la aplicación o las necesidades que se tengan. El patrón de referencia no tiene importancia relevante desde el exterior del sistema, mientras este responda eficientemente, el enfoque que se haya seguido no tiene mucha importancia para los usuarios.

Si la voz es continua como generalmente nos comunicamos, y el sistema tiene un alto porcentaje de reconocimiento, entonces se le puede calificar como muy bueno, pero actualmente es lo más difícil de realizar. Es una tarea muy compleja , pero es una de las inquietudes de los investigadores del área. Obviamente un sistema reconocedor de voz universal es el objetivo de cualquier investigación, pero existen bastantes limitaciones y no se han superado del todo.

Para el trabajo que realizamos en nuestro segmentador acústico, nos enfocamos en una clasificación específica de reconocedor de voz. Resultaría quizá inoportuno mencionar todas las posibilidades de los sistemas y compararlos, pero existen varios parámetros que observar. El *segmentador acústico* funciona como la base para un sistema reconocedor con las siguientes características:

1- Vocabulario De Palabras Separadas. (Palabras segmentadas).
2- Ilimitada Población De Parlantes.
3- Hablar En Un Cuarto Cerrado, con baja interferencia de ruido.
4- Sistema De Entrenamiento al inicio del sistema.
5- Vocabulario Pequeño (1 - 20 Palabras).
6- Palabras Simples.

5.2. ENFOQUES DE LOS SISTEMAS RECONOCEDORES DE VOZ.

En los años 50's y 60's los sistemas reconocedores se basaban en los principios de comparación de patrones, en cambio en los años 70's cambian a ingeniería de conocimiento o investigaciones basadas en reglas [Rowden92]. Posteriormente surgieron nuevas ideas para realizar sistemas reconocedores de voz, pero en gran medida involucran a las dos técnicas anteriores para generar los sistemas. Existió una maduración de múltiples ideas y algoritmos los cuales están ahora comenzando a proporcionar buenas soluciones.

En la actualidad no existe algún enfoque forzoso a seguir para realizar las aplicaciones relacionadas con el reconocimiento de voz, y existe una libertad para dar cualquier enfoque, pero es importante conocer los múltiples conocimientos que se han generado en esta área.

5.2.1. COMPARACION DE PATRONES.

La teoría básica de estos sistemas es que extraen en primer lugar, la información de todas las palabras a reconocer. Extraen los datos necesarios, ya sea los fonemas, segmentos acústicos, etc. Características primordiales de cada palabra. Las almacena en plantillas y posteriormente las utiliza para compararlas con las características de las palabras entrantes.

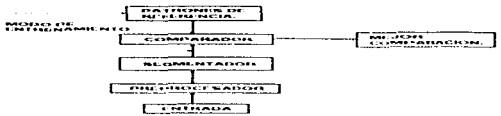

Figura 5.1 *Estructura básica de un sistema de reconocimiento de voz de comparación de patrones.*

Si se observa la figura 5.1, una parte básica es el segmentador, el cual se encarga de dividir la señal continua en patrones discretos, pequeños segmentos representativos de la señal. Existe una dificultad asociada con el tipo de segmentador y el funcionamiento del mismo, debido a las características de la voz. Investigadores tienen dificultades con la segmentación adecuada [Rowden92].

Este trabajo está dedicado completamente al desarrollo de un segmentador acústico debido a la importancia que tiene este en los sistemas reconocedores de voz. Después de encontrar que tipo de segmento que se desea comparar, el trabajo de reconocimiento a realizar se reduce en un buen porcentaje.

5.3. TECNICAS DE RECONOCIMIENTO DE PALABRAS SEPARADAS .

Muchos investigadores han desarrollado ideas muy importantes para tratar de resolver sistemas como el que estudiamos, y es por esta razón que incluimos este capítulo, para asociar nuestra investigación con otras ya existentes.

Reconocer palabras de diferentes parlantes tiene múltiples aplicaciones, entre ellas pueden ser empleadas para reconocer comandos simples de voz, reconocer números separados para cuentas de bancos, tarjetas de crédito, etc. Para crear un sistema de acceso a un local por medio de voz, etc. Realmente sistemas donde sea necesario o indispensable utilizar un pequeño número de simples comandos.

5.3.1. RECONOCIMIENTO.

El reconocimiento de palabras separadas es usualmente resuelto con el método de comparación de plantillas.

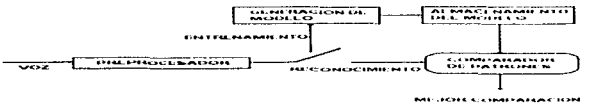

Figura 5.2 *Técnica típica de comparación de plantillas para reconocimiento de palabras separadas.*

Existe una etapa de entrenamiento, donde se almacenan las plantillas representativas de las palabras a reconocer. Las plantillas dependen de que tipo de patrones se manejen en el sistema. En nuestro enfoque, son plantillas de cada número conteniendo los tipos de sus segmentos acústicos. Para posteriormente en el método de reconocimiento, se extraen los segmentos acústicos de las palabras entrantes y se comparan con las plantillas previamente almacenadas. La comparación de mayor fidelidad es en donde se clasifica y se reconoce a la palabra entrante.

Varias técnicas se han desarrollado para realizar la comparación de plantillas, distancia métrica, y ajuste dinámico en el tiempo (DTW). Cada uno con sus ventajas y desventajas, pero el de mayor uso actualmente es DTW debido a la flexibilidad y adaptabilidad en las aplicaciones.

Cabe señalar que previamente a la comparación está la extracción de características (patrones), que es donde las cosas se complican. Nuestro segmentador obtiene pequeños bloques característicos de cada palabra. Gran parte del trabajo está realizado, pues solo resta elegir algún método de comparación para concluir el sistema reconocedor de voz de palabras separadas.

5.3.2. DISTANCIA METRICA.

La comparación del patrón de entrada y la plantilla del patrón almacenado, es por medio del cálculo de una distancia métrica *D(x,y)*, el cual mide matemáticamente las similitudes entre el patrón '*x*' y el patrón '*y*'.

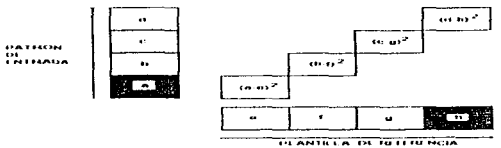

Figura 5.3 Distancia Métrica Euclidiana

$$D(x,y) = \sqrt{(a-e)^2 + (b-f)^2 + (c-g)^2 + (h-d)^2} \qquad (5.1)$$

Cada bloque puede relacionarse a algún valor numérico asignado a un segmento acústico. La palabra que reúna el menor valor al calcular la distancia es en donde se le clasifica. Obviamente lo importante en este tipo de comparación es encontrar que parámetros se desea introducir a la fórmula, que son ya sea segmentos acústicos o algún otro patrón. Después de este detalle la comparación y clasificación puede ser tarea muy simple.

5.3.3. AJUSTE DINAMICO EN EL TIEMPO (DTW).

Una característica de los patrones que se comparan con la distancia métrica, es que deben ser de la misma longitud. Esto es una característica poco frecuente en señales de voz, debido a que cada palabra aunque sea la misma, no presenta la misma longitud siempre. Para resolver este problema, se utilizan algoritmos de ajuste de los patrones almacenados como plantillas, alargándolos o acortándolos en el tiempo para lograr una comparación adecuada.

Como se observa en la figura 5.4, el patrón plantilla es alargado o comprimido en una manera no lineal para hacer que secciones similares de ambos patrones tengan la

misma duración. Aunque este método es tardado en cuanto a tiempo de máquina. La única diferencia con la comparación en distancia métrica es la posible modificación en el tiempo de los patrones antes de la comparación.

Figura 5.4 Ajuste dinámico en el tiempo para comparación de simples plantillas.

Relacionando esta técnica con el segmentador acústico, el método de alineación en el tiempo resulta muy compatible. Considere que la palabra es segmentada en bloques de longitud variable, pero que conservan las características acústicas de las palabras. Se puede comparar la sucesión de segmentos acústicos con otros previamente almacenados, pero como es de esperarse, no van a coincidir en el tiempo por lo que será necesario aumentar o disminuir las plantillas de comparación para realizar la compatibilidad.

5.3.4. MODELOS OCULTOS DE MARKOV.

El principio del modelado de las cadenas de Markov está basado en un modelo de patrones de voz como una secuencia de vectores derivados de una función probabilística de primer orden o *modelo de Markov* . Los estados en tal modelo son conectados por transiciones probabilísticas y cada estado es identificado con una apropiada función de probabilidad. [Rowden92].

Para propósitos de reconocimiento primero se necesita obtener la probabilidad de una secuencia de observaciones. Es decir se necesita un proceso de entrenamiento teórico, en el que se tiene que deducir las posibilidades probabilísticas mediante el cálculo de una función. Los patrones acústicos de una señal de voz consisten típicamente de una secuencia de vectores que se derivan de la señal de voz usando algún tipo de preprocesamiento. Puede ser por medio de análisis de Fourier, u otro método.

La importancia de los modelos ocultos de Markov para reconocimiento automático de voz es que puede ser usado en una variedad diferente de formas, dependiendo del tipo de modelado de los patrones de voz. Puede modelarse la voz en segmentos como lo son las palabras, subpalabras, fonemas o segmentos acústicos.

Figura 5.5 Estructura de un modelo de Markov de cuatro estados.

La aplicación instantánea de la *cadena de Markov* utilizando el modelo que nosotros elegimos se basa en los segmentos acústicos. El primer fundamento de compatibilidad es que cada palabra separada presenta segmentos acústicos definidos. Por lo general una misma palabra tendrá casi el mismo número de segmentos acústicos definidos en el mismo orden. Por ejemplo, la palabra "one" del vocablo ingles, puede tener según nuestras pruebas entre 4 y 5 segmentos bien marcados; cada segmento puede representar una transición entre los estados en el modelo *de Markov*.

La idea anterior sirve para visualizar la compatibilidad de las posibles utilizaciones del modelo de *Markov*. Antes de iniciar el proceso de reconocimiento, se deben conocer las posibles rutas de las palabras a identificar. Me refiero a rutas, como los segmentos que componen generalmente a cada palabra. Es decir, el sistema tendrá una cadena general que contendrá los distintos caminos posibles que recorrerá cada palabra. Se llegará a un reconocimiento cuando se alcance un punto terminal de la cadena, el cual especifique que palabra es la más probable.

Muchos sistemas basados en subpalabras emplean modelos ocultos de Markov por sus gran adaptabilidad y modalidades [Rowden95].

5.4. PLANIFICACION DE UNA POSIBLE APLICACION.

Presentamos algunas sugerencias bien fundamentadas de como realizar un reconocedor automático de palabras separadas. El diagrama general puede ser el mostrado en la figura 5.6.

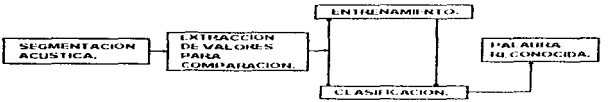

Figura 5.6 Sistema general de reconocimiento automático para palabras separadas.

Después de obtener los valores de cada segmento acústico se inicia realmente el proceso de reconocimiento. Se deberá escoger algún método de extracción de valores para ser comparados posteriormente. En el caso de [HerrAlgIrv94], propone un análisis por medio de la *Transformada Karhunen- Loeve (KLT),* en donde cada subpalabra es mapeada en un conjunto de eigenvectores y eigenvalores. Esta representación describe la información lingüística contenida en cada palabra mediante algunos cuantos valores, los cuales se utilizan posteriormente en la etapa de entrenamiento y clasificación del sistema.

Pero esta etapa de extracción de características queda a libertad de elección. También proponemos comparar segmentos acústicos íntegros en representación del espectro en frecuencias, o tal vez usar *Codificación Lineal Predictiva (LPC)* [Picone93], para extraer vectores que representan a cada palabra. Es decir el segmentador es compatible con varios métodos actuales empleados en el reconocimiento.

En el modo de entrenamiento en el caso de [HerrAlgIrv94], los eigenvectores de las subpalabras acústicas son almacenadas en la memoria del computador. En el modo de reconocimiento las palabras se clasifican de alguna manera, las entradas y las palabras antes almacenadas se comparan usando algún método adecuado (Distancia Métrica, DTW, etc.).

Realizar todo el sistema de reconocimiento es buena propuesta de algún otro trabajo de tesis, y ahí se elegirán los caminos que seguirá el procedimiento después del segmentador acústico.

5.5. FUTURAS APLICACIONES Y PERSPECTIVAS.

En el presente existen varios sistemas reconocedores de voz con distintas características. Tanto sistemas reconocedores de palabras separadas como continuas, pero debido a que no existe una normalización en la forma de enfoque de los reconocedores de voz, cada uno presenta sus características, por lo que no se puede llegar a una comparación estricta. Pero cualquier sistema debe ser valorado dependiendo de sus características de desempeño, mediante alguna metodología para medir ésta y generar una clasificación [Rowden92]. Generalmente las aplicaciones de reconocimiento de voz exhiben un *porcentaje de desempeño*. Pero este parámetro no es una medida concreta de desempeño del sistema. Debido a que puede tener diferentes enfoques que hacen que los sistemas no se puedan comparar, como por ejemplo, que algunos sistemas analizan palabras separadas en contra de los que analizan voz continua. Pues es notorio que crear un sistema que reconozca palabras separadas es más fácil que un sistema para voz normal continua.

Lo importante, para las aplicaciones y las perspectivas de investigación deben caracterizar su desempeño en forma que el *porcentaje de reconocimiento* refleje sus capacidades verdaderas en término de sus propiedades correspondientes. Entonces el desempeño de un dispositivo de reconocimiento de voz puede ser caracterizado en términos de su 'perfil de capacidades' [Rowden92], a través de un número de características que lo representan en la figura 5.7.

Como puede verse no se comparará el desempeño de un sistema con vocabulario grande y con parlante dependiente con un sistema de pequeño vocabulario e independencia del parlante. No existe una prueba fija sobre los tipos de evaluaciones. Lo que queremos decir, es que actualmente existen muchos sistemas, pero también existen muchos problemas que resolver en el área de reconocimiento. Por lo que nosotros esperamos sistemas que se hayan creado de acuerdo a ciertas necesidades, claro que lo óptimo pero lo más difícil es crear un sistema de reconocimiento universal.

Nuestro trabajo se enfocó simplemente a segmentos acústicos, pues es un buen análisis para sistemas de palabras separadas, con parlante independiente, y el cual es el objetivo primordial al que queríamos llegar. Se han desarrollado dispositivos electrónicos que ayudan en el procesamiento de voz, pero todavía con muchas limitantes. Un sistema electrónico para reconocer 50 palabras fue posible en 1986 [Rowden92]. Realmente se esperan nuevos desarrollos , aplicaciones efectivas en nuevas campos y a bajos costos en pocos años.

Grado De Dificultad	
Fácil	*Difícil*
Un Solo Parlante	*Muchos Parlantes*
Palabras Separadas	*Voz Continua*
Parlantes Consistentes	*Parlante Inconsistente*
Vocabulario Pequeño	*Gran Vocabulario*
Comandos Simples	*Lenguaje Natural*
Entrenamiento Previo	*Sin Entrenamiento*

Figura 5.7 *Perfiles de capacidades.*

Pero hoy, debido a los múltiples enfoques y soluciones, es deseable seleccionar una arquitectura del sistema óptimo, de acuerdo a la complejidad real de las necesidades que se tengan. Es decir, si solo se desea un reconocedor de voz para un sistema de acceso en alguna puerta, la arquitectura tal vez sea más simple que para un sistema cuyo requerimiento sea interpretar automáticamente la voz continua y ejecutar las ordenes. Es decir dependiendo de la aplicación es la metodología a usar.

En el futuro se intentará crear reconocedores de voz considerando las siguientes dificultades actuales: Que reconozca diferentes estilos de voz, cuidando acentos, énfasis y efectos posibles de la voz. Hacer que el medio donde se pronuncie cada palabra no influya ni interfiera en el desempeño del reconocedor. Además que identifique un gran vocabulario. Y que trate de imitar la forma en que el humano escucha, para así que no existan diferencias entre hablar con un humano y hablar con una máquina.

Conclusiones.

CONCLUSIONES.

Obtuvimos una metodología general para segmentar acústicamente palabras de forma automática, aunque es cierto que no podemos comprobar, ni comparar los resultados del método, debido a la gran cantidad de palabras, nos basamos en nuestros fundamentos teóricos para decir que al final de cada proceso de segmentación acústica, se generan como resultados, las divisiones de las regiones más homogéneas de cada palabra. Además de que limitamos el número de subpalabras acústicas, seleccionando aquellas que sean las más representativas.

Cada palabra separada presenta características individuales que dependen de la situación en que fueron producidas, por esta razón también sería muy difícil comprobar cada segmento acústico de cada palabra. Pero como se puede observar en nuestros resultados, la prueba MLR ofrece información general importante que facilita la selección no arbitraria , y la clasificación automática de los segmentos acústicos.

Es importante señalar que el método obtiene las regiones homogéneas más importantes de cada palabra, condicionado a que dichas palabra esten acotadas con sus valores de inicio y fin correctamente. En caso de no existir una adecuada limitación del inicio y fin, es probable que la segmentación acústica no sea adecuada. Pero esto no quiere decir que sea errónea, pués como está plantedo nuestro método, se puede aplicar a regiones específicas de voz y se obtienen los diferentes segmentos acústicos, así sean regiones de silencio, o de ruido de la señal.

Los objetivos planteados al inicio de nuestro trabajo de limitar automáticamente el número de segmentos acústicos para cada palabra y además de que fueran los más importantes, los cumplimos apoyados en la información que aporta la prueba MLR, que es confiable debido a su planteamiento teórico.

Creemos que nuestro método de segmentación acústica facilitará algunas aplicaciones del reconocimiento automático de voz, ya que debido a que es una técnica que se basa en propiedades acústicas de la voz, y cuyos procesamientos son independientes del locutor y su sexo, es independiente también del número de muestras de cada palabra, además es una metodología automática general que puede adaptarse a alguna aplicación específica en el área de reconocimiento automático de voz.

Existen limitantes de la técnica que planteamos, básicamente es el tiempo de procesamiento, que es alto considerando los múltiples cálculos para obtener los segmentos acústicos. Claro que nosotros preprocesamos las señales de voz para eliminar información que no es deseada, por lo que consumimos más tiempo, pero en sí, solo existen dos procesos básicos de la segmentación acústica, la descomposición en bandas críticas y la prueba MLR. Gran parte del procesamiento es debido a la descomposición en bandas críticas, y tal vez en aplicaciones futuras este proceso se simplifique, ya sea por software o hardware. En cambio, la prueba MLR ofrece resultados aunque no en tiempo real, pero no consumen gran tiempo como el dedicado para el preprocesamiento y la descomposición en bandas críticas de la señal.

BIBLIOGRAFIA.

[Ainsworth88] *Ainsworth, William Antony.* ; ***Speech Recognition by Machine.*** P.peregrinus on Behalf of the Institutions of Electrical Engineers. London 1988.

[Atal93] *Atal Bishnu S ; Cuperman , Vladimir ; Gersho, Allen.* ***Speech and Audio Coding for Wireless and Network Applications.*** Kluwer Academic. Boston 1993.

[HerrAlgIrv94] *Herrera, Abel ; Algazi, V.R.; Irvine, D.* ***An Acoustic Approach for Isolated Speech Recognition.*** Proceedings of the International Conference on Signal Processing. Aplications and Technology. ICSPAT 94. Vol 2. pp. 1677-1681.

[HeAlBrIr94] *Herrera, Abel ; Algazi, V.R. ; Brown, K.L ; Irvine, D.;* ***Subword Segmentation Alternatives For Isolated And Conected Words Recognition.*** Proceedings of VII European Signal Processing Conference EUPSICO-94.

[Oppenheim89] *Oppenheim, Alan V. ; Schafer, Ronald W.* ***Discrete-time Signal Processing.*** Prentice Hall, Inc; USA 1989.

[Owen93] *Owen, Frank, J.* ; ***Signal Processing of Speech;*** McGraw-Hill ; México 1993.

[Parsons87] *Parsons, Thomas W* ; ***Voice And Speech Processing.*** McGraw-Hill ; México 1987.

[Picone93] *Picone J.; **Signal Modeling Techniques in Speech Recognition.*** Proceedings of the IEEE. Vol 81. No. 9. September 1993.

[Rabiner78] *Rabiner, Lawrence R ; Schafer, Ronald W. coaut.* ***Digital Processing of Speech Signals.*** Prentice-Hall; USA 1978.

[Ramirez85] *Ramirez, Robert W . ;**The FFT: Fundamentals an Concepts.*** Prentice-Hall . Englewood Cliffs; USA 1985.

[RamYam96] *Ramos Alvarez Alfredo; Yamasaki Granados Karin.* ***Tesis Detección de Principio y Fin de Señales de Voz.*** Facultad de Ingenieria . UNAM.1996.

[Rowden92] *Rowden, Chris ; **Speech Processing.*** McGraw-Hill ; London, México 1992.

[Saito85] *Saito, Shuzo ; Nakata, Kazuo .* ***Fundamentals of Speech Signal Processing.*** Academic; Tokyo 1985.

[Schafer75] *Schafer, Ronald W. ; Rabiner, Lawrence R.*
Digital Representations of Speech Signals. Proceedings of the IEEE. 1975.

[Schwab86] *Schwab, Eileen C; Nusbaum, Howard.*
Pattern Recognition By Humans and Machines. Volume1. Speech Perception. Academic Press series in Cognition and Perception. USA 1986.

[Stearns90] *Stearns, Samuel D. ; Dan R. Hush ,* ***Digital Signal Analysis.*** Prentice-Hall; 2da. Edición ; New Jersey 1990.

[Shihua91] *Shihua Wang ; Andrew Sekey ; Allen Gersho.*
Auditory Distortion Measure for Speech Coding. IEEE. Signal Processing Society ; ICASP91. International Conference on Acoustic, Speech & Signal Processing. New York 1991.

[VanTrees68] *Van Trees ; Harry L. ;* ***Detection, Estimation and Modulation Theory*** Wiley ; USA 1968. Part 1.

Aliasing: Término en inglés, nos indica las distorsiones tipo escalera que sufre una señal cuando se disminuye su frecuencia de muestreo, o existe translape de las ventanas de análisis en tiempo corto.

ANSI: *American National Standard Institute;* Instituto Nacional Americano de Normas. Organismo encargado de crear normas para el intercambio de información.

Bandas Críticas: Representación de las características frecuenciales de las señales de voz agrupadas en intervalos de frecuencias audibles para los humanos.

DTW: *Dinamic Time Warping;* Ajuste Dinámico en el Tiempo. Técnica de comparación de patrones, en donde el patrón de referencia es modificado en el tiempo para una comparación adecuada de características.

D/C: *Digital / Continuo*; Convertidor de señales digitales a señales continuas.

DFT: *Discrete Fourier Transform*; Transformada Discreta de Fourier. Técnica matemática para transformar una señal discreta en el tiempo al dominio de la frecuencia discreta.

Espectrograma: Gráfica de tres dimensiones que muestra el comportamiento de las señales de voz en el dominio de las frecuencias.

FFT: *Fast Fourier Transform*; Transformación Rápida de Fourier. Técnica matemática derivada de la Transformada Discreta de Fourier (DFT), que mediante el aprovechamiento de las redundancias de la DFT, reduce la cantidad de cálculos.

Fonema: Un fonema es la unidad más pequeña de la voz.Secuencia de segmentos acústicos.
Frecuencia de Muestreo: Es la periodicidad en la que son tomadas las muestras de una señal.

KLT: *Karhunen Loeve Transform;* Transformada de Karhunen Loeve Técnica matemática para la transformación de bloques de muestras de voz desde un espacio a otro de tal manera que se elimine la correlación que existe entre ellos y se obtenga una representación característica de los mismos.

LPC: *Linear Predictive Coding*, Codificación Lineal Predictiva. Técnica de procesamiento digital de señales que representa a una señal de voz por una serie de coeficientes de una forma predictiva, que es el mejor modelo en segmentos pequeños de voz.

MLR: *Maximum Likelihood Ratio;* Cociente de Máxima Similitud. Criterio de decisión probabilística para la verificación de hipótesis mediante las comparaciones de probabilidades de las señales de voz.

Modelo Oculto de Markov: Método estadístico para estimar la probabilidad del siguiente valor de una cantidad, comenzando de los valores previos.

Muestra: Es un ejemplo (digitalizado) de una señal para un periodo corto de tiempo.

Palabra Separada o Discreta: Cada expresión tratada como si fuera una entrada en el diccionario, si el diccionario contiene frases, entonces la expresión no necesita ser una palabra sola. En este trabajo el diccionario contiene palabras simples.

Reconocimiento de Voz: Termino general que se le da a la acción de tomar a las señales de voz e interpretar las palabras contenidas o identificar al orador.

Segmento Acústico: Región homogénea de las señales de voz con casi las mismas características frecuenciales audibles.

Voz: Patrones acústicos en conjunto con interpretación semántica de la información hablada por los humanos.