UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

"RECONOCIMIENTO DE PALABRAS AISLADAS UTILIZANDO CUANTIZACIÓN VECTORIAL"

T E S I S

QUE PARA OBTENER EL TITULO DE
INGENIERO EN TELECOMUNICACIONES
PRESENTA

ARTURO GARDIDA DEGOLLADO

DIRECTOR: M.I. ABEL HERRERA CAMACHO

Ciudad Universitaria, Diciembre 1998

A mi Papá y a mi Mamá.

A quienes les debo todo lo que soy y todo lo que tengo.

Agradecimientos

Al Maestro Abel.
Por haberme dado la oportunidad de trabajar bajo su asesoría

Y a todas aquellas personas a las que he conocido a lo largo de mi desarrollo como persona y cómo estudiante.
Porque de todos he aprendido algo y a todos se los debo.

INDICE

INTRODUCCION

Desde la aparición del hombre sobre la tierra, siempre se ha destacado de los demás seres vivos con los cuales comparte el planeta por su capacidad de razonar, esta capacidad es la que lo ha llevado ha transformar su forma de vida y su entorno, impulsado por diversas causas como pudieran ser el lograr un avance tecnológico, la satisfacción de una necesidad o simplemente el alcanzar una vida mas cómoda.

Por otra parte, una de las cosas que mas distingue al hombre de los otros seres vivos es el poder comunicarse con sus semejantes a través del uso de la voz. La voz la utilizamos para comunicar nuestros pensamientos y nuestras emociones. Y con el advenimiento de las nuevas tecnologías tratamos nuevamente de modificar nuestro entorno, buscando poder comunicarnos con las máquinas y que ellas se comuniquen con nosotros a través del uso de la voz.

El presente trabajo de tesis sobre reconocimiento automático de palabras aisladas, intenta colaborar en la realización de esa difícil tarea. Para lo cual hace uso de la Cuantización Vectorial (VQ) como su principal herramienta.

La Cuantización Vectorial ha tomado gran auge y una gran aceptación en las áreas de procesamiento de señales y de reconocimiento de patrones, debido al gran número de aplicaciones en las que se puede emplear. La Cuantización Vectorial y el Reconocimiento se hacen utilizando tres tipos de parametrización de la voz, el primero que desde su aparición ha mostrado ser uno de los que mejor representan a las señales de voz (la Codificación de Predicción Lineal o LPC), el segundo que es uno de los menos utilizados pero que sin embargo genera buenos resultados (el Cepstrum) y, el tercero y último una transformación relativamente nueva, pero que en fechas recientes se ha empezado a utilizar con mayor frecuencia (la transformada KLT).

La organización del trabajo se realiza de tal forma, que en los primeros capítulos se describen los principios básicos utilizados, se continua con una descripción de como fueron aplicados esos principios para hacer posible el reconocimiento de palabras aisladas y se finaliza con la implementación de los sistemas de reconocimiento y la presentación de resultados.

El primer capítulo describe los antecedentes que se deben tener de la señales de voz, así como los problemas que surgen al intentar reconocerla en una forma automática. El segundo capítulo "Parametrización de las Señales de Voz" presenta los tres tipos de representación en forma paramétrica que se hacen a las señales de voz. El tercer capitulo esta dedicado completamente a la Cuantización Vectorial. En el cuarto capítulo se presentan los tipos de segmentación aplicados a las palabras aisladas. El quinto capitulo "Sistemas de Reconocimiento", reúne lo descrito en los capítulos anteriores y presenta la forma en que se utilizan los parámetros de la voz y la cuantización vectorial con el fin de lograr el reconocimiento de las palabras aisladas; incluye los resultados y la conclusiones que justifican el uso de las técnicas utilizadas.

I. ANTECEDENTES

I.1 LA SEÑAL DE VOZ Y SU ANÁLISIS BÁSICO

La voz es una característica humana que puede ser analizada, sintetizada, reconocida, comprimida, almacenada y mejorada mediante el uso de técnicas de procesamiento de señales.

Para nosotros es de lo mas normal comunicarnos usando la voz y esto nos resulta tan simple y natural que se pensó que las maquinas también lo realizarían de una forma sencilla, sin embargo esto ha resultado demasiado complicado, debido principalmente a las implicaciones que involucra la comunicación mediante el uso de la voz. Los problemas a resolver en el reconocimiento de voz son principalmente.

Las señales de voz tienden a ser continuas, la comunicación se realiza a través de frases que comprenden varias palabras en las cuales es difícil identificar el fin de una palabra y el principio de otra. Incluso dentro de una misma palabra es difícil encontrar las fronteras entre los distintos sonidos que la conforman.

Las señales de voz tienden a ser variables, ya que dos señales de voz resultan diferentes aunque representen a las misma palabra y hayan sido pronunciadas por la misma persona.

El habla es ambigua, debido a que existen palabras que se pronuncian igual como "coser" y "cocer" , pero que tiene significados distintos e inclusive una misma palabra puede tener varios significados. Y es el contexto del cual se habla, el que les da su significado.

Las señales de voz se encuentran contaminadas, en la mayoría de los ambientes existen ciertas perturbaciones que tienden a distorsionar las señales como es el caso del ruido.

El habla es compleja, ya que se encuentra sujeta a acuerdos entre los individuos, además de formar solo una parte del proceso de la comunicación

Ahora bien, el que podamos entendernos utilizando la voz no significa que no tengamos que enfrentarnos a estos problemas, sin embargo nosotros empleamos otros recursos, tales como la experiencia, el contexto, la semántica y el sentido común, los cuales hacen posible la comunicación.

Lamentablemente para nuestra causa, las maquinas no poseen ninguno de estos recursos, siendo necesario emplear otras técnicas, provenientes de varias disciplinas como: procesamiento de señales, reconocimiento de patrones, acústica, lingüística, fonética, etc. A continuación se presentan algunas de estas técnicas, empleadas en la realización de este trabajo y que en conjunto constituyen el análisis básico de la señal de voz.

Gráficas en Tiempo y en Frecuencia

Los mas básico es una representación en el tiempo de la forma de onda de la señal de voz (figura I.1). Estas representaciones son útiles, por ejemplo si visualmente se desea determinar el punto en el cual inicia o termina la palabra o si se desea saber el valor de la señal en un determinado instante.

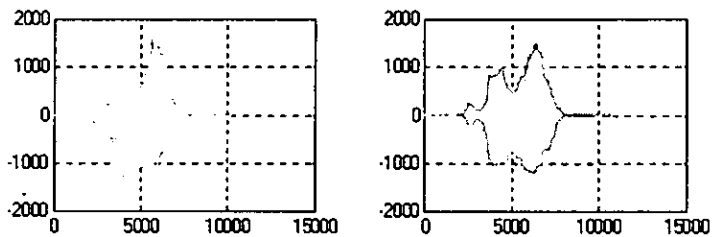

Figura I.1. Formas de onda de la palabra "Zero" (mismo locutor).

Sin embargo, dadas las variaciones que se presentan en las señales de voz, es necesaria otra forma de representarlas mejor, junto con la información que contienen; esta forma alternativa es la representación espectral, presentada en espectogramas (figura I.2). Los espectogramas son gráficas originalmente tridimensionales pero que se reducen a dos dimensiones y que representan la intensidad de la señal en un cierto ancho de banda (eje de las ordenadas) y en un intervalo de tiempo (eje de las abcisas).

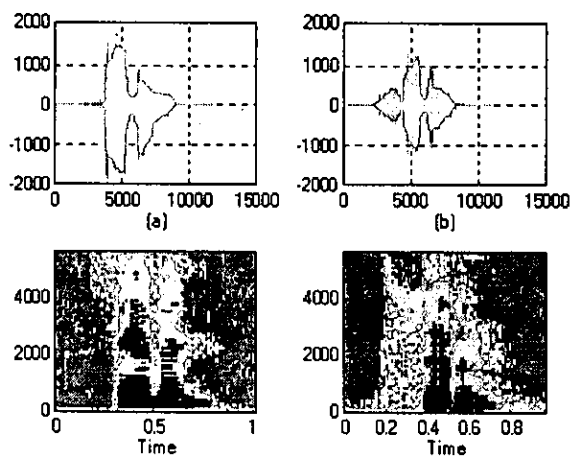

Figura I.2. Formas de onda y espectogramas de la palabra "Seven"

Pero ya que el tiempo y la frecuencia son variables relacionadas en forma inversa, no se pueden hacer gráficas que tengan una buena resolución de ambas en la misma representación. Por lo que existen dos clases de espectogramas: los de banda ancha, caracterizados por una pobre resolución en frecuencia y una buena resolución en el tiempo y los espectogramas de banda angosta, que poseen una mas alta resolución en frecuencia con su correspondiente disminución de resolución en el tiempo.

Se prefieren utilizar los espectogramas de banda ancha, por mostrar mejor las características de las señales en frecuencia. En las gráficas, la intensidad de la señal esta representada por variaciones de grises, desde el blanco (la menor intensidad) hasta el negro (la mayor intensidad).

Bandas Críticas

El fenómeno de Bandas Críticas, es un fenómeno puramente empírico descubierto por Fletcher y Munson (1937) y se puede describir como el proceso de filtrado que realiza el sistema auditivo.

La banda crítica es el ancho de banda en el cual un sonido permanece constante conforme aumenta el ancho de banda, hasta un cierto punto, en el que la intensidad que se percibe comienza a aumentar. Este punto límite es conocido como el ancho de banda crítico, y varia según la frecuencia central de la banda de acuerdo a la siguiente ecuación:

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

En la tabla I.1 se presentan las bandas críticas obtenidas a partir de datos experimentales. La importancia de las bandas críticas se encuentran en el efecto que tienen los tonos sobre el oído humano, ya que el oído tiende a promediar frecuencias dentro de la misma banda, por lo que todas las frecuencias dentro de la banda se escuchan como si fueran la frecuencia central.

Métodos en el Dominio del Tiempo

El principal objetivo en el procesamiento de las señales de voz es el de tener una representación mas conveniente y mas útil de la información que contienen dichas señales. La precisión requerida en esta representación se encuentra sujeta a las características de las señales que se deseen conservar o enfatizar.

En este apartado revisaremos las técnicas de procesamiento denominadas "Métodos en el Dominio del Tiempo", los cuales involucran los métodos aplicados directamente a la forma de onda de la señal de voz.

La importancia de estos métodos radica principalmente en su fácil implementación, además de que proveen una base útil en la estimación de algunas características importantes de las señales de voz, tales como: la frecuencia fundamental y las formantes, o pueden ser de utilidad en la separación de la señal en sonidos sordos, sonoros o silencios.

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

Tabla I.1. Anchos de banda críticos

Ventanas

La suposición básica con la que se trabaja en la mayoría de los esquemas de procesamiento de voz es que las propiedades de las señales de voz cambian relativamente muy poco con el tiempo. Lo que nos conduce a los métodos de procesamiento en tiempo corto, en los cuales solo una parte de la señal es aislada y procesada en segmentos de corta duración (tramas), considerándose que todo la trama tiene las mismas propiedades, es decir que se selecciona solo una parte de la señal que se supone estacionaria.

Esto se realiza al aplicar a las señales una ventana, seleccionando la trama de la señal original con el cual se desee trabajar por un proceso simple de multiplicación (Ec. I.2). Se asume entonces que la señal es igual a cero fuera del intervalo de interés.

El tamaño de las ventanas es del orden de 100 a 200 muestras para una frecuencia de muestreo de 10kHz (de 10 a 20 ms de duración).

$$x_w(n) = x(n)w(n) \qquad 0 \le n \le N-1 \qquad \text{I.2}$$

Existen diferentes tipos de ventanas, tales como: ventana rectangular, ventana Bartlett (triangular), ventana de Hanning, ventana de Hamming y ventana de Blackman; todas con diferentes características y aplicaciones. Para el presente trabajo resulta suficiente el uso de la ventana de Hamming definida por,

$$w(n) = 0.54 - 0.46\cos\left(2\pi n/(N-1)\right) \qquad 0 \le n \le N-1 \qquad \text{I.3}$$

Energía y Magnitud Promedio

Al observar las gráficas de las señales de voz, se puede notar que su amplitud varia considerablemente con el tiempo. Y en general la amplitud de los sonidos sordos es mucho menor que la amplitud de los sonidos sonoros. La energía en tiempo corto de la señal de voz provee una conveniente representación que refleja estas variaciones (figura I.3). La energía de una señal discreta se define como,

$$E = \sum_{m=-\infty}^{\infty} x^2(m) \qquad \text{I.4}$$

Pero esta expresión nos resulta de poca utilidad, debido a que la información que presenta nos dice muy poco acerca de las propiedades que dependen del tiempo de la señal. Por lo que mejor utilizaremos la definición de energía en tiempo corto (ecuación I.5),

$$E_n = \sum_{m=n-N+1}^{n} x^2(m) \qquad \text{I.5}$$

o reescribiendo la expresión,

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m)h(n-m) \qquad \text{I.6}$$

donde,

$$h(n) = w^2(n) \qquad \text{I.7}$$

La interpretación de la ecuación (I.6) es que la señal $x^2(m)$ es pasada a través de un filtro con respuesta al impulso h(n). El problema ahora es entonces, el de seleccionar la ventana que mejor se acople a nuestras necesidades.

La mayor aplicación de la función de energía es que nos permite distinguir entre los segmentos de voz sonora y los segmentos de voz sorda. Y para aplicaciones de voz de alta calidad también nos sirve para distinguir entre los segmentos de voz y los segmentos de silencio.

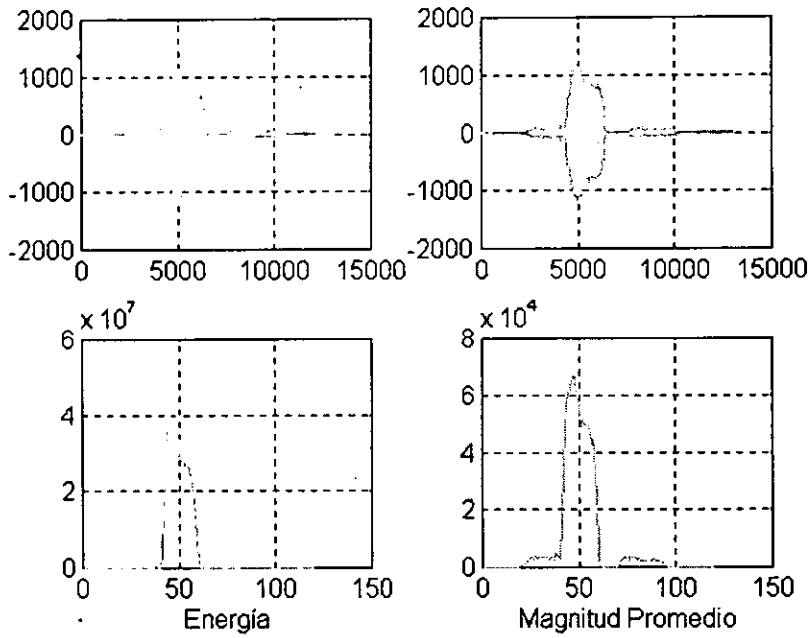

Figura I.3. Gráficas de Energía y Magnitud Promedio (misma palabra "Six")

Una dificultad que se tiene al utilizar la función de energía en tiempo corto, es que resulta muy sensible a valores grandes de la señal (por que su calculo involucra un cuadrado), enfatizándose las diferencias de valores que existen entre muestra y muestra. Una forma de disminuir estas diferencias es el utilizar la función de magnitud promedio (ecuación I.8) en vez de la función de energía,

$$M_n = \sum_{m=-\infty}^{\infty} |x(m)| w(n-m) \qquad \text{I.8}$$

en la cual se suman los valores absolutos de la señal en vez de sumar los cuadrados. Representación que resulta de igual utilidad que la función de energía (fig. I.3). Aunque los rangos dinámicos de las funciones varíen en una raíz cuadrada y que las diferencias entre voz sonora y voz sorda no sean tan pronunciados para el caso de la función magnitud promedio.

Taza de Cruces por Cero

En el estudio de señales discretas, un cruce por cero ocurre si dos muestras sucesivas tienen signos distintos. La taza de cruces por cero es una forma muy simple de medir el contenido frecuencial de una señal; lo cual resulta cierto para señales de banda angosta.

Y aunque las señales de voz son de banda ancha y la interpretación de la taza de cruces por cero en promedio es mucho menos precisa que en el caso anterior, se pueden estimar las propiedades espectrales de la señal de voz utilizando esta técnica. La taza de cruces por cero en promedio se define por,

$$Z_n = \sum_{m=-\infty}^{\infty} \left|\operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)]\right| w(n-m) \qquad \text{I.9}$$

donde,

$$\operatorname{sgn}[x(n)] = \begin{cases} 1 & x(n) \geq 0 \\ -1 & x(n) < 0 \end{cases} \qquad \text{I.10}$$

y

$$w(n) = \begin{cases} \dfrac{1}{2N} & 0 \leq n \leq N-1 \\ 0 & c.o.c \end{cases} \qquad \text{I.11}$$

Expresiones que nos indican que se verifican las muestras por pares, se suman aquellas con signos distintos y el resultado se promedia sobre N muestras consecutivas (opcional).

En procesamiento de voz la taza de cruces por cero, se utiliza bajo las consideraciones de que el modelo de voz sugiere que la energía de la voz sonora se concentra por debajo de los 3kHz, mientras que la energía de la voz sorda se encuentra en su mayoría en altas frecuencias (figura I.4). Y ya que altas frecuencias involucran una taza de cruces por cero alta, y bajas frecuencias involucran una taza de cruces baja, existe una fuerte correlación entre la taza de cruces y la distribución de la energía con la frecuencia. Por lo que una razonable generalización es que si la taza de cruces es alta, la voz es sorda, y si la taza es baja se tiene voz sonora. Aunque esta aseveración no es del todo correcta por que hace falta definir otras cuestiones, nos resultará de utilidad mas adelante.

Función de Autocorrelación

La función de autocorrelación de una señal determinística y discreta se define por,

$$\phi(k) = \sum_{m=-\infty}^{\infty} x(m)x(m+k) \qquad \text{I.12}$$

La importancia de la función de autocorrelación es que resulta una forma muy conveniente de desplegar ciertas propiedades de la señal. Por ejemplo,

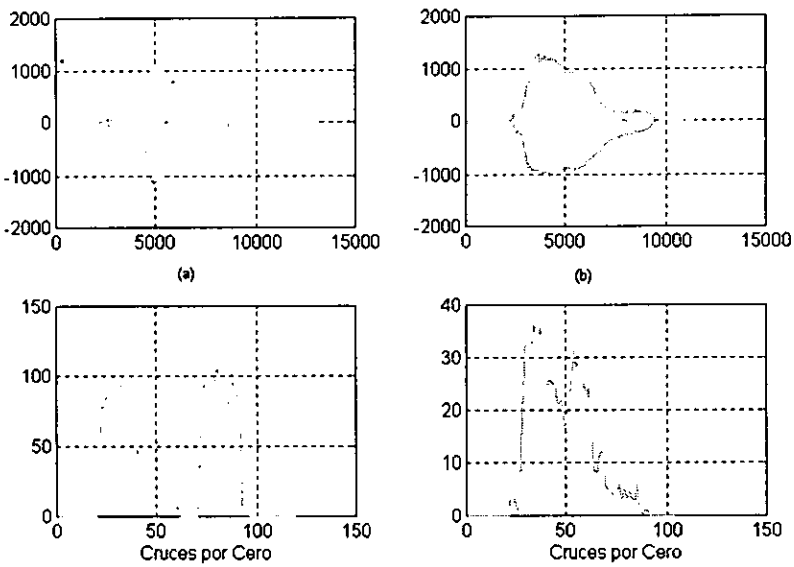

Figura I.4. Señales de voz y sus tazas de cruces por cero. (a) "Six". (b) "Nine"

1.- Si la señal es periódica con periodo P, entonces: $\phi(k) = \phi(k+P)$, es decir, la función de autocorrelación resulta igualmente periódica con periodo P.

2.- Es una función par: $\phi(k) = \phi(-k)$

3.- Tiene su máximo valor en k=0; o sea $|\phi(k)| \leq \phi(0)$ para todo k.

4.- La cantidad $\phi(0)$ es igual a la energía para señales determinísticas o es igual a la potencia promedio si la señales son periódicas o aleatorias.

Con la ayuda de estas propiedades, podemos observar que para señales periódicas, la función de autocorrelación presenta máximos en los puntos 0, ±P, ±2P, ... Esto es, que ignorando el punto de origen, el periodo de la señal puede estimarse al localizar el primer máximo de la función; y resulta aun más importante ya que nos permite estimar la periodicidad en segmentos de señales, incluyendo las señales de voz.

Al igual que para los métodos anteriores, conviene obtener una representación en tiempo corto, definiéndose la función de autocorrelación en tiempo corto por,

$$R_n(k) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)x(m+k)w(n-k-m) \qquad \text{I.13}$$

o reescribiendo la expresión,

$$R_n(k) = \sum_{n=0}^{n=N-1-m} x_w(n)x_w(n+m) \qquad m = 0,1,2,\ldots,p \qquad \text{I.14}$$

expresión que se utilizará posteriormente.

Transformada de Fourier en Tiempo Corto

En el área del procesamiento de voz, el concepto de la representación de Fourier tiene un gran importancia, ya que nos sirve para poner en evidencia ciertas propiedades que no resultan tan obvias en la señal original. Y aunque, la representación tradicional de Fourier no sea de gran utilidad en el caso de señales que cambian de manera muy marcada con el tiempo, nos basaremos en esta para definir la transformada de Fourier en tiempo corto; ya que la propiedades espectrales de la señal de voz pueden extraerse si se considera que la señal varia en forma muy lenta con el tiempo.

Una útil definición de la transformada de Fourier en tiempo corto,

$$X_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} w(n-m)x(m)e^{-j\omega n} \qquad \text{I.15}$$

En esta expresión, w(n-m) es una ventana real que determina la porción de la señal que será enfatizada en un tiempo n en particular. Esta transformada es función de dos variables: el índice n (discreto) y la frecuencia w (continua).

Las condiciones para la existencia de esta transformada se derivan de las condiciones que debe cumplir la transformada convencional. En este caso se necesita que la secuencia x(m)w(n-m) sea absolutamente sumable para todos los valores de n. Y de igual forma cumple con sus propiedades, por ejemplo la transformada de Fourier en tiempo corto es periódica con periodo igual a 2π. Otro importante hecho resulta de su facilidad de cálculo si se utilizan los algoritmos de la FFT (Transformada Rápida de Fourier).

Filtrado de Pre-énfasis

En el espectro de la voz existe una caída de -6 dB/octava, conforme la frecuencia aumenta. Esto se debe a la combinación de una caída de -12 dB/octava ocasionada por la fuente de excitación de la voz y un incremento de +6 dB/octava ocasionado por la radiación de la boca. Esto significa que, cada vez que la frecuencia aumenta al doble, la amplitud de la señal se reduce en un factor de 16. Por lo que se desea compensar este roll-off de -6 dB/octava con un pre-procesamiento de la señal de voz que de un incremento de +6 dB/octava en el rango apropiado, de manera que la medición del espectro tenga un rango dinámico similar a lo largo de todo su ancho de banda.

Esto es referido como pre-énfasis. En un sistema de procesamiento digital de señales, el pre-énfasis puede ser implementado ya sea por un filtro analógico paso altas de primer orden con una frecuencia de corte de 3 dB en algún punto entre los 100 Hz y 1 kHz (la posición exacta no es crítica), el cual precede al filtrado anti-aliasign y al convertidor A/D; o con un filtro digital paso altas que procese a la señal de voz digitalizada. Este filtrado digital puede ser implementado al usar la ecuación en diferencias:

$$y[n] = x[n] - ax[n-1] \qquad \text{I.16}$$

donde **y[n]** es la muestra actual que se obtiene a la salida del filtro de pre-énfasis, **x[n]** es la muestra actual que se tiene a la entrada del filtro, **x[n-1]** es la muestra anterior y **a** es una constante usualmente escogida entre 0.9 y 1. Calculando la transformada Z a la ecuación I.16:

$$Y(z) = X(z) - az^{-1}X(z) = \left(1 - az^{-1}\right)X(z) \qquad \text{I.17}$$

donde z^{-1} representa el operador de retardo por muestra.

La función de transferencia H(z) del filtro es:

$$H(z) = \frac{Y(z)}{X(z)} = 1 - az^{-1} \qquad \text{I.18}$$

Para el caso de segmentos de silencio, no existe la necesidad de aplicar el filtro de pre-énfasis, ya que no existen cambios espectrales que necesite ser eliminados. Sin embargo, por simplicidad, el pre-énfasis es normalmente aplicado a los segmentos de silencio también.

Detección de Inicio y Fin de Palabras Aisladas

El problema de localizar donde inicia y donde termina una palabra resulta importante en muchas áreas del procesamiento de voz, y es de particular importancia en el reconocimiento de palabras aisladas, esto con el fin de solo trabajar con las muestras que representan a la señal de voz, eliminando el ruido que se presenta al inicio y fin de la grabación.

El algoritmo presentado, llamado algoritmo de Rabiner-Sambur, esta basado en la combinación de dos mediciones realizadas en el dominio del tiempo (magnitud promedio y cruces por cero).

Como primer paso se obtienen por cada trama de n muestras la magnitud promedio de la señal y su tasa de cruces por cero. Se asume entonces que las primeras 10 tramas no contienen voz, esto con el fin de obtener una caracterización estadística del ruido de fondo. Utilizando esta estadística se calculan los umbrales que nos servirán a detectar el inicio y fin. Se define un primer umbral muy conservador ITU para obtener el intervalo en el cual la energía promedio de la señal siempre es excedido y se asume que el inicio y el final yacen fuera de este intervalo. Entonces verificando las tramas desde el punto donde se rebaso por primera vez el umbral ITU hacia el inicio (final) de la grabación, hasta un punto N1 (N2) donde la magnitud promedio cae por debajo de un umbral menor ITL y que es seleccionado tentativamente como el inicio (final) de la palabra. Resulta valido suponer que el inicio y fin no se encuentran en estos puntos, por lo que el siguiente paso es verificar hacia el inicio (final) de la grabación desde N1 (N2) la tasa de cruces por cero y compararla contra un umbral ITZC. Esto se realiza para las 25 tramas que preceden a N1 (siguen a N2). Y si la tasa de cruces por cero excede el umbral ITZC 3 o mas veces, el inicio N1 es recorrido hasta el punto en donde el umbral fue excedido por primera vez. En caso contrario el inicio se escoge en el primer punto N1. Se realiza un procedimiento similar para determinar el final de la palabra. Un ejemplo típico de este método se muestra en la figura I.5

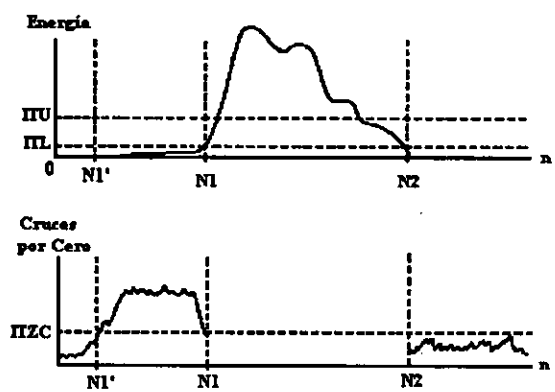

Figura I.5. Gráficas típicas de Magnitud Promedio y Cruces por Cero

I.2 SISTEMA DE RECONOCIMIENTO DE VOZ

Las áreas de investigación y desarrollo en el reconocimiento de voz se encuentran comprendidas entre los siguientes campos de estudio: reconocimiento de palabras aisladas con independencia del locutor, reconocimiento de voz independiente del locutor, verificación de locutor, reconocimiento del lenguaje, traducciones, análisis semántico y sintáctico, algoritmos de codificación y análisis de voz, y parametrización.

El presente trabajo se encuentra comprendido en el área de reconocimiento de palabras aisladas con independencia del locutor, aunque tal clasificación no es excluyente ya que se cubren otras áreas como es la parametrización.

El reconocimiento de palabras aisladas utilizando cuantización vectorial consistió en dos tareas, en la primera a partir de un conjunto de datos en forma paramétrica de las señales de voz, se obtienen los patrones de referencia.

La segunda tarea consiste en obtener de una señal de voz, la cual se desea reconocer, sus parámetros y compararlos con los patrones de referencia. La palabra será reconocida por aquel conjunto de patrones con los cuales sus parámetros tengan mayor similitud

Al proceso por medio del cual se obtienen los patrones de referencia se le denomina **entrenamiento** y al de comparación se le conoce como **reconocimiento.** En la figura I.5 se presenta un modelo conceptual del sistema de reconocimiento de palabras aisladas.

La cuantización y la comparación se realizan con el uso de una medida de distorsión (distancia) apropiada a cada tipo de parametrización; además la elección de la medida de distorsión es uno de los factores que mas influyen en el éxito de un sistema de reconocimiento de voz. Razón por la cual en el siguiente apartado se presentan las medidas de distancia o distorsión más comúnmente utilizadas.

Debido a los problemas que se presentan cuando se pronuncian dos palabras que representan a la misma señal, se necesita contar con un número considerable de repeticiones de la misma y que además sean generadas por mas de una persona, para así determinar los rangos de variación en que se encuentran los parámetros de una palabra en particular.

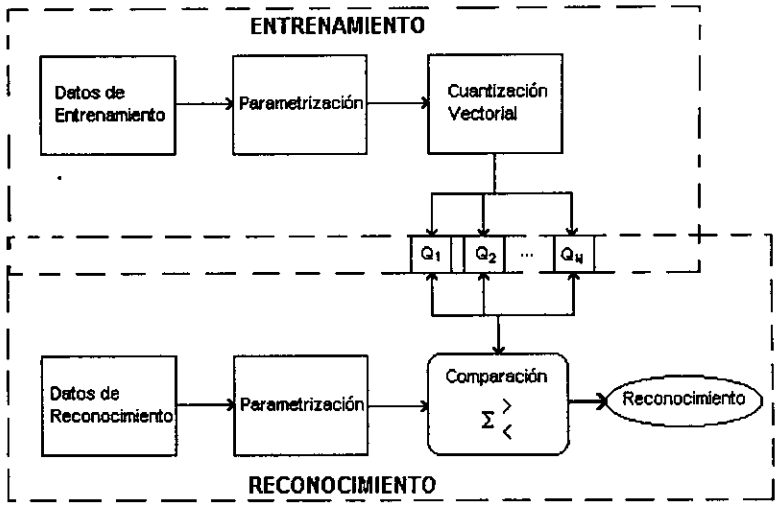

Figura I.6. Modelo conceptual del sistema de reconocimiento

I.3 MEDIDAS DE DISTORSION

La comparación de un patrón de entrada y un patrón almacenado se realiza por medio de una medida de distorsión o distancia conveniente, y que al comparar los parámetros que representan a dos señales similares deben generar una distancia menor que con respecto a lós parámetros de otras señales.

Para que sea de utilidad, una medida de distorsión debe poder ser analizada y calculada; debe ser relevante en forma subjetiva, de modo que las diferencias en los valores de la distorsión puedan ser utilizados para indicar diferencias de similitud entre los patrones a comparar. La mayoría de las medidas de distorsión cumplen en cierto grado con las características mencionadas. Entre las áreas de aplicación para las medidas de distorsión se encuentran la tasación de los codificadores de voz y el **reconocimiento automático de voz.**

Las propiedades deseables que debe cumplir una medida de distorsión $d(X_1,X_2)$ son:

1. Simetría, $d(X_1,X_2) = d(X_2,X_1)$

2. La desigualdad del triángulo, $d(X_1,X_3) \geq d(X_1,X_2) + d(X_2,X_3)$

3. No negatividad, $d(X_1,X_2) > 0$ para $X_1 \neq X_2$
   $d(X_1,X_2) = 0$ para $X_1 = X_2$

4. Computacionalmente eficiente, evaluación no muy costosa

5. Con significado físico, relacionado con alguna calidad tangible.

Una medida que satisface las tres primeras propiedades es llamada métrica, existiendo varias medidas que no lo son. A continuación se describen algunas de las medidas de mayor uso.

La medida más conveniente y ampliamente usada para calcular distancias, es el error cuadrático o distancia Euclidiana cuadrática entre dos vectores definida como,

$$d(X_1, X_2) = \|X_1 - X_2\|^2 = \sum_{j=1}^{N} (X_{1j} - X_{2j})^2 \qquad \text{I.19}$$

La distorsión del error cuadrático medio (MSE) es otra de las medidas mas utilizadas,

$$d(X_1, X_2) = \frac{1}{N}(X_1 - X_2)^T (X_1 - X_2) = \frac{1}{N}\sum_{j=1}^{N} (X_{1j} - X_{2j}) \qquad \text{I.20}$$

en la cual la distorsión esta definida por dimensión. La popularidad del MSE se basa en su simplicidad y seguimiento matemático.

Otra medida de distorsión es el error cuadrático medio ponderado. En el MSE la medida asume que las distorsiones contribuyen cuantizando los diferentes parámetros $\{X_{1j}\}$ de igual forma.

Y de manera general, se pueden introducir pesos diferentes con el fin de aportar ciertas contribuciones a la distorsión dependiendo del parámetro. El MSE ponderado general se define como,

$$d(X_1, X_2) = (X_1 - X_2)^T W (X_1 - X_2) \qquad \text{I.21}$$

donde W es una matriz simétrica y positiva definida de ponderación, y los vectores X_1 y X_2 son tratados como vectores columna.

Una selección para W ampliamente utilizada en aplicaciones de clasificación de patrones es $W = \Gamma^{-1}$, donde Γ es la matriz de covarianzas del vector aleatorio X,

$$\Gamma = E\left[(X - \overline{X})(X - \overline{X})^T\right], \qquad \overline{X} = E[X] \qquad \text{I.22}$$

En este caso la distorsión se reduce a,

$$d(X_1, X_2) = (X_1 - X_2)^T \Gamma^{-1} (X_1 - X_2) \qquad \text{I.23}$$

conocida como la distancia de Mahalanobis.

Dado que la matriz de ponderación es simétrica, se puede factorizar como: $W = P^T P$. Los vectores X_1 y X_2 se pueden transformar en un nuevo conjunto de vectores $\underline{X_1}$ y $\underline{X_2}$

$$\underline{X_1} = PX_1 \qquad \underline{X_2} = PX_2$$

y

$$d(X_1, X_2) = (PX_1 - PX_2)^T (PX_1 - PX_2) = (\underline{X_1} - \underline{X_2})^T (\underline{X_1} - \underline{X_2}) = d_{MSE}(\underline{X_1}, \underline{X_2}) \qquad \text{I.24}$$

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

Cada una de las medidas de distorsión mencionadas anteriormente resultan simétricas en sus argumentos X_1 y X_2, y aunque pueden ser aplicadas a las características derivadas del análisis de producción lineal de la voz; el uso de algunas presenta ciertas desventajas, tal es el caso de la distancia euclidiana que aunque resulte fácil de calcular no todas sus características tienen el mismo significado perceptual.

Por lo anterior, en ciertos casos resulta conveniente y efectivo escoger una matriz de ponderación $W(X_1)$ que dependa explícitamente del vector X_1, para así obtener una medida de distorsión perceptualmente motivada. En este caso, la distorsión,

$$d(X_1, X_2) = (X_1 - X_2)^T W(X_1)(X_1 - X_2) \qquad \text{I.25}$$

es asimétrica. Más adelante en los capítulos referentes a los reconocimientos de palabras aisladas, se describirán las medidas de distorsión utilizadas.

II. PARAMETRIZACIÓN DE LA VOZ

La suposición básica con la que trabajamos es que la voz puede representarse como la salida de un sistema lineal e invariante con el tiempo (LIT), cuyas características varían muy lentamente con el tiempo. Lo que nos conduce a que la voz puede ser analizada en segmentos de corta duración (tramas), y a que cada trama puede ser representada como la respuesta a una excitación de entrada a un sistema LIT. Las dos entradas que suponemos para generar la voz pueden ser un tren de impulsos cuasi-periódicos (voz sonora) o un ruido aleatorio (voz sorda). Por lo tanto el problema principal en el análisis de voz es estimar los parámetros del modelo de la voz y medir sus variaciones en el tiempo. A continuación se presenta los tres tipos de parametrización con los que se modelaron las señales de voz.

II.1 ANÁLISIS LPC

Una de las técnicas mas poderosas en el análisis de voz es el método de Análisis de Predicción Lineal. Y el modelo de predicción lineal es el más ampliamente usado para modelar al tracto vocal; ya que los modelos comunes del proceso de producción de voz usualmente tratan por separado al tracto vocal y al aire que entra a este (la "excitación").

En el análisis LPC la sección del tracto vocal de el modelo de producción de voz se representa por medio de un filtro digital lineal variante con el tiempo. Este filtro debe representar los efectos de la radiación de los labios, la forma del pulso glotal, y el acoplamiento de la cavidad nasal cada vez que se requiera. Además representa la voz en una tasa de transmisión o almacenamiento de bits baja. La importancia de esta técnica radica tanto en su capacidad de proveer unos parámetros estimadores de la voz extremadamente exactos, como en su relativa velocidad de cálculo.

La idea básica detrás de el modelo LPC es que dada una muestra de voz en un tiempo n, s(n), esta puede ser aproximada como una combinación lineal de p muestras de voz precedentes. Al minimizar la suma de las diferencias de los cuadrados (sobre un intervalo finito) entre las muestras de voz actuales y las predecidas linealmente, un conjunto único de coeficientes de predicción puede ser determinado, esto es

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \ldots + a_p s(n-p), \qquad \text{II.1}$$

donde los coeficientes a_1, a_2, ..., a_p se suponen constantes sobre la trama de voz analizada. La ecuación anterior se convierte en igualdad al incluir un termino de excitación Gu(n), quedando

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n), \qquad \text{II.2}$$

donde u(n) es la excitación normalizada y G es su ganancia. Al expresar esta ecuación en el dominio de z se obtiene la relación

$$S(z) = \sum_{i=1}^{p} a_i z^{-i} S(z) + GU(z), \qquad \text{II.3}$$

conduciéndonos a la función de transferencia,

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{A(z)} \qquad \text{II.4}$$

La interpretación de esta última ecuación esta dada en la figura II.1, donde se muestra la fuente de excitación normalizada, u(n), siendo escalada por la ganancia G, y actuando como entrada al sistema todos-polos, H(z) = 1/A(z), para producir la señal de voz, s(n).

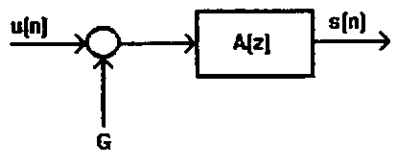

Figura II.1

Además la filosofía del modelo LPC se encuentra íntimamente relacionada con el conocimiento de que la fuente de excitación para la señal de voz es esencialmente un tren de pulsos cuasi-periódicos (para la voz sonora) o una fuente de ruido aleatorio (para la voz sorda) que sirve de excitación a un sistema lineal variante con el tiempo, y que precisamente el modelo LPC nos provee de un método robusto, realizable y preciso para estimar los parámetros que caracterizan a este sistema lineal e invariante con el tiempo.

Análisis de ecuaciones LPC

De acuerdo al modelo de la figura II.1, la relación exacta entre s(n) y u(n) es

$$s(n) = \sum_{k=1}^{p} a_k \, s(n-k) + Gu(n) \qquad \text{II.5}$$

Consideramos la combinación lineal de las muestras de voz pasadas como la estimación $\hat{s}(n)$ definida como,

$$\hat{s}(n) = \sum_{k=1}^{p} a_k \, s(n-k) \qquad \text{II.6}$$

Formamos el error de predicción, e(n), de la siguiente forma

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} a_k \, s(n-k) \qquad \text{II.7}$$

con función de transferencia del error

$$A(z) = \frac{E(z)}{S(z)} = 1 - \sum_{k=1}^{p} a_k \, z^{-k} \qquad \text{II.8}$$

Ahora el problema básico en el análisis de predicción lineal es determinar el conjunto de coeficientes de predicción, $\{a_k\}$, directamente de la señal de voz, de modo que las propiedades espectrales de la voz se mantengan constantes. Y ya que las características espectrales de la voz varían con el tiempo, los coeficientes de predicción en un tiempo dado n, deben estimarse en un segmento corto de la señal de voz alrededor de este tiempo. Entonces la idea fundamental es encontrar un conjunto de coeficientes de predicción que minimicen el error de predicción cuadrático medio sobre un segmento corto de la forma de onda de la voz. (Normalmente este análisis se realiza sobre tramas sucesivas de voz, con un espaciamiento del orden de 10 ms entre tramas).

Para empezar con las ecuaciones que se deben resolver, y así poder determinar los coeficientes de predicción, se define el segmento de voz y el segmento de error en un tiempo n como,

$$\begin{aligned} s_n(m) &= s(n+m) \\ e_n(m) &= e(n+m) \end{aligned} \qquad \text{II.9}$$

y lo que buscamos es minimizar la señal de error cuadrático medio en el tiempo n,

$$E_n = \sum_m e^2(m) \qquad \text{II.10}$$

y al usar la definición de $e_n(m)$ en términos de s(m), se puede escribir como,

$$E_n(m) = \sum_m \left[s_n(m) - \sum_{k=1}^{p} a_k \, s_n(m-k) \right]^2 \qquad \text{II.11}$$

Para resolver esta ecuación, se deriva parcialmente E_n con respecto a cada a_k y el resultado se iguala a cero,

$$\frac{\partial E_n}{\partial a_k} = 0, \quad k = 1,2,\ldots,p \qquad \text{II.12}$$

quedando,

$$\sum_m s_n(m-i) s_n(m) = \sum_{k=1}^{p} \hat{a}_k \sum_m s_n(m-i) s_n(m-k) \qquad \text{II.13}$$

Reconociendo que los términos de la forma $\sum_m s_n(m-i)s_n(m-k)$ representan las covarianzas de los segmentos $s_n(m)$, se pueden escribir como,

$$\phi(i,k) = \sum_m s_n(m-i)s_n(m-k) \qquad \text{II.14}$$

Por lo que la ecuación II.14, se puede escribir en forma mas compacta,

$$\phi_n(i,0) = \sum_{k=1}^{p} \hat{a}_k \, \phi_n(i,k) \qquad \text{II.15}$$

Ecuación que describe un conjunto de p ecuaciones con p incógnitas. Y para obtener los coeficientes de predicción óptimos tenemos que calcular $\phi_n(i,k)$ para $1 \le i \le p$ y $0 \le k \le p$ y resolver el conjunto resultante de p ecuaciones simultáneas.

Existen varias formas de calcular los coeficientes de predicción: método de covarianzas, método de autocorrelación, el método enrejado, el del filtro inverso, el de estimación espectral, el de máxima probabilidad y el método del producto interno. En reconocimiento de voz (y en este trabajo), se utiliza el método de autocorrelación debido a su eficacia computacional así como estabilidad inherente. Este método siempre produce un filtro de predicción cuyos ceros se encuentran dentro del círculo unitario en el plano Z.

Método de autocorrelación

Un forma de determinar los limites de las sumatorias de las ecuaciones es asumir que el segmento, $s_n(m)$, es igual a cero fuera del intervalo $0 \le m \le N-1$. Esto puede escribirse de la forma,

$$s_n(m) = s(m+n)w(m) \qquad \text{II.16}$$

donde w(m) es una ventana de longitud finita (usualmente ventana de Hamming), que es igual a cero fuera del intervalo $0 \le m \le N-1$.

El efecto de esta suposición en los límites de la sumatoria para las expresiones que contienen a E_n puede verse al considerar la ecuación . Claramente, si $s_n(m)$ es diferente de cero solo para $0 \le m \le N-1$, entonces el correspondiente error de predicción, $e_n(m)$, para un predictor de orden p, será diferente de cero en el intervalo $0 \le m \le N-1+p$. Entonces, para este caso E_n se expresa apropiadamente como,

$$E_n = \sum_{m=0}^{N+p-1} e_n^2(m) \qquad \text{II.17}$$

y $\phi_n(i,k)$ se puede expresar,

$$\phi_n(i,k) = \sum_{m=0}^{N+p-1} s_n(m-i)s_n(m-k), \quad \begin{matrix} 1 \le i \le p \\ 0 \le k \le p \end{matrix} \qquad \text{II.18}$$

Como esta última ecuación es sólo función de i-k (en lugar de ser función de dos variables i y k), la función de covarianza, $\phi_n(i,k)$, se reduce a una simple función de autocorrelación,

$$\phi_n(i,k) = r_n(i-k) = \sum_{m=0}^{N-1-(i-k)} s_n(m)s_n(m+i-k), \quad \begin{matrix} 1 \le i \le p \\ 0 \le k \le p \end{matrix} \qquad \text{II.19}$$

Además como la función de autocorrelación es simétrica, $r_n(-k) = r_n(k)$, las ecuaciones LPC pueden expresarse como,

$$\sum_{k=1}^{p} r_n(|i-k|)\hat{a}_k = r_n(i), \qquad 1 \le i \le p \qquad \text{II.20}$$

y se puede expresar en forma matricial,

$$\begin{bmatrix} r_n(0) & r_n(1) & r_n(2) & \cdots & r_n(p-1) \\ r_n(1) & r_n(0) & r_n(1) & \cdots & r_n(p-2) \\ r_n(2) & r_n(1) & r_n(0) & \cdots & r_n(p-3) \\ \vdots & \vdots & \vdots & & \vdots \\ r_n(p-1) & r_n(p-2) & r_n(p-3) & \cdots & r_n(0) \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \vdots \\ \hat{a}_p \end{bmatrix} = \begin{bmatrix} r_n(1) \\ r_n(2) \\ r_n(3) \\ \vdots \\ r_n(p) \end{bmatrix} \qquad \text{II.21}$$

Esta matriz de orden pxp con los valores de autocorrelación, es una matriz Toeplitz (simétrica con los elementos de la diagonal principal iguales) que puede resolverse eficientemente con el uso de varios procedimientos numéricos. Como el algoritmo de Levinson-Durbin.

II.2 ANALISIS CEPSTRAL

Sistemas Homomórficos

Se dice que un sistema es lineal cuando cumple con el principio de superposición que se utiliza en transformaciones lineales y que se expresa de la siguiente forma,

$$L\{aX_1(n)+X_2(n)\} = aL\{X_1(n)\}+L\{X_2(n)\} = aY_1(n)+Y_2(n) \qquad \text{II.22}$$

Siendo L{*} una transformación lineal. Ahora bien existen un tipo especial de sistemas que cumplen con un principio generalizado de superposición (ecuación II.22)

$$D\left\{[x_1(n)]^{\alpha}\cdot[x_2(n)]^{\beta}\right\} = \alpha D[x_1(n)]+\beta D[x_1(n)] \qquad \text{II.22}$$

A los sistemas que cumplen con esta propiedad se les denomina sistemas homomórficos. Esta denominación surge del hecho que tales transformaciones pueden ser mostradas como transformaciones homomórficas en el sentido de espacios vectoriales lineales. Un filtro homomórfico es simplemente un sistema homomórfico que tiene la siguiente propiedad: una componente (la componente deseada) pasa a través del sistema sin alterarse, mientras que la componente no deseada es eliminada.

Los sistemas homomórficos resultan de gran utilidad en el procesamiento de voz debido a que nos presentan una metodología para separar a la señal de excitación de la forma del tracto vocal, esto debido a que las señales de voz se modelan como la salida de un sistema LIT, en el cual el espectro de la señal de voz es una combinación en convolución de la señal de excitación y la respuesta del tracto vocal al impulso; el problema por lo tanto se reduce a la separación de las componentes de una convolución. Acción conocida como deconvolución.

La convolución en sistemas discretos se expresa de la siguiente forma,

$$y(k) = \sum_{-\infty}^{\infty} h(n-k)x(k) = h(n)*x(n) \qquad \text{II.23}$$

Si aplicamos la transformada de Fourier, a la ecuación II.23 resulta,

$$Y(\omega) = H(\omega)\cdot X(\omega) \qquad \text{II.24}$$

Tomando el logaritmo complejo de ambos lados

$$\log[Y(\omega)] = \log[H(\omega)]+\log[X(\omega)] = \log H + \log X \qquad \text{II.25}$$

Por lo tanto en el dominio logarítmico la excitación y la respuesta del tracto vocal al impulso se encuentran superimpuestas, y las podemos separar utilizando las técnicas de procesamiento de señales convencionales.

Al aplicar la transformada de Fourier a la ecuación II.23 y hacerla pasar a través de un filtro homomórfico se separa el espectro de potencia de el sistema generador de voz y las líneas del espectro de los armónicos de la frecuencia fundamental (pitch).

El siguiente paso es aplicar la IFT a la ecuación II.25. Al resultado de la transformada inversa de Fourier se le denomina "Cepstrum" y la variable correspondiente a su frecuencia se le denomina "quefrency. La operación de la separación de componentes dependiendo de su frecuencia se le denomina "liftering" (un filtrado en quefrency). El termino Cepstrum fue introducido por Bogert y se ha aceptado en la terminología del procesamiento de voz como la transformada inversa de Fourier del logaritmo del espectro de potencia de la señal.

Para procesar el cepstrum, primero se obtienen la magnitudes espectrales logarítmicas. Posteriormente se calcula la transformada inversa del espectro logarítmico.

$$C_n = \frac{1}{2\pi}\int_0^{2\pi} \log|Y(\omega)|e^{j\omega n}d\omega \qquad \text{II.25}$$

Se define C_n como el cepstrum. Y se representa por los coeficientes cepstral calculados por medio de la transformada de Fourier, referidos como los coeficientes cepstral derivados de la transformada de Fourier.

II.3 TRANSFORMADA KLT

La transforma KLT o transformada Karhunen-Loève, es una transformación óptima en sentido estadístico, decorrelaciona la secuencia en el dominio de la transformada, lo cual la hace ser mejor que la FFT. Las funciones son eigenvectores de la matriz de covarianzas y los elementos diagonales son las varianzas de la transformada.

Se comienza con una formación de vectores de la forma:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \text{II.26}$$

Se define la media como,

$$m_x = E\{x\} \qquad \text{II.27}$$

donde E{·} es el valor esperado.

El valor esperado de un vector se calcula tomando el valor esperado de cada elemento. La media de la población de M vectores se puede aproximar mediante la siguiente expresión:

$$m_x = \frac{1}{M}\sum_{k=1}^{M} x_k \qquad \text{II.28}$$

La matriz de covarianzas de una población de vectores se define como:

$$C_x = E\left\{(x - m_x)(x - m_x)^T\right\} \qquad \text{II.29}$$

Esta matriz es de orden nxn, real y simétrica. Entonces, cada elemento C_{ij} de la matriz es la varianza de X_i, y el elemento C_{ij} es la covarianza entre los elementos X_i y X_j de la muestra de vectores. Si los elementos son no correlacionados, entonces $C_{ij} = C_{ji} = 0$.
Si se tiene una población de M vectores, la matriz de covarianzas se puede aproximar mediante:

$$C_x = \frac{1}{M}\sum_{k=1}^{M} x_k x_k^T - m_x m_x^T \qquad \text{II.30}$$

Debido a las propiedades de la matriz C_x, siempre es posible encontrar un conjunto de vectores propios ortonormales. Aquí, la idea es obtener una base de forma que la matriz sea lo mas simple posible respecto a la base. Esto resulta en un sistema de coordenadas en el que el origen es el centro de la población de vectores, y cuyos ejes están en la dirección de los vectores propios de C_x.

Se sabe, de Algebra Lineal, que λ es el valor propio de C_x si y solo si $\det(C_x - \lambda I)=0$, siendo I la matriz identidad. Esta ecuación nos da los valores propios con sus correspondientes vectores propios. Definimos a e_i como el i-ésimo vector propio de C_x, con valor propio λ_i, para i=1, ...,n-1. Por conveniencia, se pueden ordenar los vectores propios por orden descendiente de valores propios, es decir, $\lambda_j \geq \lambda_{j+1}$ para j=1,...,n-1.

Sea A una matriz de nxn con sus renglones formados con los vectores característicos C_x . Entonces, el primer renglón contiene al vector propio con el valor propio más grande, y el último, el vector propio con el valor propio más pequeño. Ahora se puede construir un vector y como sigue,

$$y = A(x - m_x) \qquad \text{II.31}$$

Esta ecuación se conoce como la transformada Karhunen-Loève (KLT). El vector resultante y tiene la propiedad de $m_y = 0$. La matriz de covarianzas de y se puede obtener en términos de A y C_x mediante

$$C_y = AC_xA \qquad \text{II.32}$$

donde C_y es una matriz diagonal, con los elementos de la diagonal principal iguales a los valores característicos de C_x, es decir, C_y es de la forma:

$$C_y = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix} \qquad \text{II.33}$$

Una propiedad importante de la transformada KLT esta relaciona con la reconstrucción de **x** a partir de **y**. Debido a que los renglones de **A** son ortonormales, entonces $A^{-1}=A^T$ de forma que:

$$x = A^T y + m_x \qquad \text{II.34}$$

En vez de utilizar los vectores propios de C_x, podemos utilizar únicamente los K vectores propios con los valores propios mas grandes. Esto resulta en una matriz A_k con vectores y de dimensión K. El vector reconstruido por A_k es:

$$\hat{x} = A_k^T y + m_x \qquad \text{II.35}$$

Nótese que, naturalmente, esta estimación introduce un error. Este error se puede calcular de manera exacta con el error medio cuadrático:

$$e_{ms} = \sum_{j=1}^{n} \lambda_j - \sum_{j=1}^{k} \lambda_j = \sum_{j=k+1}^{n} \lambda_j \qquad \text{II.36}$$

El error puede ser minimizado seleccionando los K vectores propios con los valores propios más grandes. La transforma KLT es óptima en el sentido de que minimiza el error cuadrático medio entre el vector **x** y su aproximación $\hat{x}$.

Entre las ventajas de la transformada KLT están el decorrelacionar la secuencia dado que los coeficientes KLT son independientes y el empaquetar la mayor energía en el menor número de coeficientes de transformación.

III. CUANTIZACIÓN VECTORIAL

La cuantización vectorial (VQ) resulta una generalización de la cuantización escalar, aplicada a un vector. El cambio de una sola dimensión a varias dimensiones trae consigo un gran numero de nuevas ideas, conceptos, técnicas y aplicaciones. Ya que mientras la cuantización escalar es utilizada principalmente en la conversión analógico/digital, la cuantización vectorial trata con sofisticadas técnicas de procesamiento digital de señales, y como en la mayoría de los casos las características mas relevantes de las señales de entrada tiene representación digital, la cuantización vectorial se utiliza usualmente en la compresión de datos. Sin embargo existen ciertos paralelismos con la cuantización escalar y varios métodos se utilizan en la cuantización vectorial como una generalización.

Un vector puede utilizarse para describir prácticamente cualquier tipo de patrón, como puede ser un segmento de una señal de voz o de una imagen, simplemente al formar un vector con las muestras de la señal de voz o de la imagen. La cuantización vectorial puede aplicarse al reconocimiento de patrones, ya que una patrón de entrada es comparado y aproximado con alguno de los patrones de referencia almacenados, el reconocimiento se hace al encontrar el patrón de referencia que mas se acople al patrón de entrada. Por lo tanto la cuantización vectorial es mas que una generalización de la cuantización escalar. Y en fechas recientes se ha convertido en la principal herramienta de el reconocimiento de voz además de seguir utilizándose en la compresión de señales de voz e imágenes.

III.1 DEFINICIÓN

Partimos de un conjunto de vectores que pertenecen a un espacio K-dimensional, asumiendo que **x** es un vector perteneciente a ese conjunto cuyas componentes [x_i, $1 \le i \le K$] son variables aleatorias reales y de amplitud continua. Un cuantizador vectorial Q de dimensión K y tamaño N es una transformación de un vector **x** del espacio euclidiano de dimensión R^K en un conjunto finito C que contiene L salidas o puntos de reproducción, llamados vectores de código (code vectors).

$$Q: R^K \rightarrow C \qquad C = \left\{ \underline{y}_1, \ldots, \underline{y}_n \right\} \qquad \underline{y}_i \in R^K \qquad \text{III.1}$$

Entonces, el vector **x** es mapeado a otro vector **y** también real y de amplitud continua. Se dice que **x** esta cuantizado como **y**, donde **y** es el valor cuantizador de **x**, es decir,

$$q(x) = y_i \qquad si\ x \in C_i \qquad \text{III.2}$$

donde q(•) es el operador de cuantización. A y también se le denomina vector de reconstrucción o vector de salida que corresponde a **x**. Típicamente **y** toma un conjunto finito de valores $y=[y_i, 1 \leq i \leq L]$, donde $y_i = [y_{i1}, y_{i2}, y_{i3}, \ldots, y_{iN}]$.

El conjunto **y** es conocido como diccionario de reconstrucción o simplemente diccionario, L es el tamaño del diccionario y $[y_i]$ es el conjunto de vectores del código. Los vectores y_i son conocidos también en la literatura de reconocimiento de patrones como los patrones de referencia o plantillas. El tamaño L del diccionario también se conoce como el número de niveles, término proveniente de la cuantización escalar. Entonces se puede hablar de un diccionario de L niveles. Para el diseño de este, se particiona el espacio K-dimensional en L regiones o celdas $[C_i, 1 \leq i \leq L]$ y se asocia C_i a un vector y_i. El cuantizador entonces asigna el vector de código y_i si **x** esta en C_i.

Al proceso de creación del diccionario también se le conoce como entrenamiento o población del diccionario.

III.2 AGRUPAMIENTO

El agrupamiento es la forma en que se realiza la cuantización vectorial; consiste en lo siguiente: a partir de un conjunto de N muestras $\chi = \{X_1, X_2, X_3, \ldots, X_N\}$, se intentan separar en K subconjuntos disjuntos $\chi_1, \chi_2, \chi_3, \ldots, \chi_K$. En donde cada subconjunto representa a un grupo (clúster), y en el cual las muestras pertenecientes tienen una mayor similitud entre si, en comparación a las muestras de cualquier otro grupo.

Existen varios algoritmos de agrupamiento, tales como, agrupamiento simple, agrupamiento de distancia máxima, agrupamiento de K-Medias, agrupamiento ISODATA y agrupamiento LBG; estos dos últimos variantes del agrupamiento K-Medias pero con mayor complejidad.

III.3 ALGORITMO K-MEDIAS

Se describe a continuación el algoritmo de agrupamiento K-Medias también llamado algoritmo de Lloyd, el cual es el utilizado en este trabajo.

Su criterio de función es,

$$J_e = \sum_{j=1}^{K} \sum_{x \in \chi_j} \|x - z_j\|^2 \qquad \text{III.3}$$

donde:

K = numero de grupos
z_j = centro del grupo para el grupo j
χ_j = subconjunto de muestras asignadas al grupo j

Algoritmo:

1) Escoger K centros de grupo iniciales $z_1(1)$, $z_2(1)$, ..., $z_K(1)$.

2) En la iteración l, asignar las muestras a los grupos:

$$Asignar\ x\ a\ \chi_i(l)\ si\ \|x - z_i(l)\| \leq \|x - z_j(l)\| \quad j = 1,2,\ldots,K \quad j \neq i$$

3) Calcular los nuevos centros de grupo:

$$z_i(l+1) = \frac{1}{N_i} \sum_{x\ en\ \chi_i(l)} x \quad i = 1,2,\ldots,K$$

donde N_i es el numero de muestras asignadas a $\chi_i(l)$.

4) Si $z_i(l+1) = z_i(l)$ para i = 1, 2, ..., K el algoritmo ha convergido y debe terminarse. En caso contrario, regresar al paso 2.

Una característica de este algoritmo es que los centroides o cuantizadores obtenidos no son óptimos globales, es decir, se obtiene cuantizadores óptimos locales, que dependen de varios factores como son: la asignación inicial de centroides (principalmente), el orden en que las muestras son tomadas, las propiedades geométricas de los datos, tipo de distorsión empleada, etc. Una forma de poder llegar a los óptimos globales es probar con una gran variedad de centroides iniciales y seleccionar los cuantizadores finales que tengan la menor distorsión con respecto a los vectores a los cuales representan. Solución poco factible porque existen un gran número de combinaciones para los cuantizadores iniciales.

IV. SEGMENTACIÓN DE PALABRAS AISLADAS

Para poder realizar la cuantización vectorial y el reconocimiento de una forma que genere buenos resultados, es necesario que las palabras sean divididas en subpalabras, siendo a su vez necesario que cada una de las palabras utilizadas sea dividida en el mismo número de subpalabras, es decir, que la longitud total cada palabra sea divida en N segmentos de voz; esto con el fin de contar con patrones de referencia por cada segmento de una misma palabra.

Se utilizaron tres tipos de segmentación: segmentación lineal, segmentación no lineal (KM) y segmentación acústica (MLR).

IV.1 SEGMENTACIÓN LINEAL

La segmentación lineal es la mas simple utilizada y consiste en lo siguiente: a partir de el número total de muestras que contiene la palabra, se divide en N partes de igual longitud (figura IV.1).

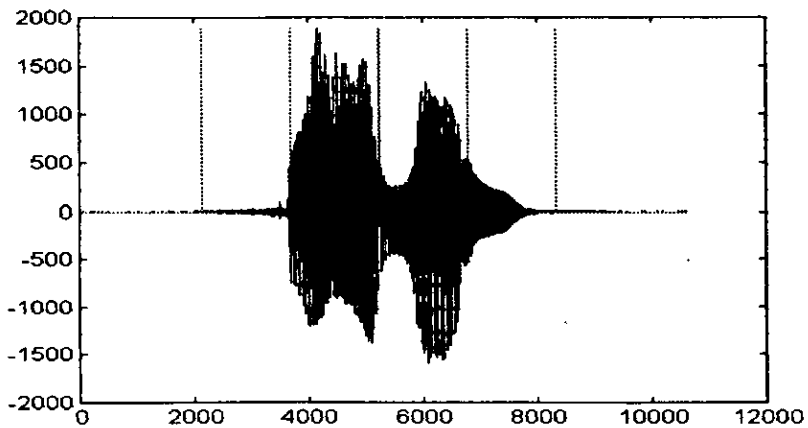

Figura IV.1. Palabra "Seven" segmentada linealmente en cuatro subpalabras

Debido a todas las diferencias que presentan dos palabras pronunciadas en diferente tiempo, aunque representen a la misma palabra y sean pronunciadas por la misma persona (figura IV.2); la segmentación lineal ha probado no generar tan buenos resultados de reconocimiento.

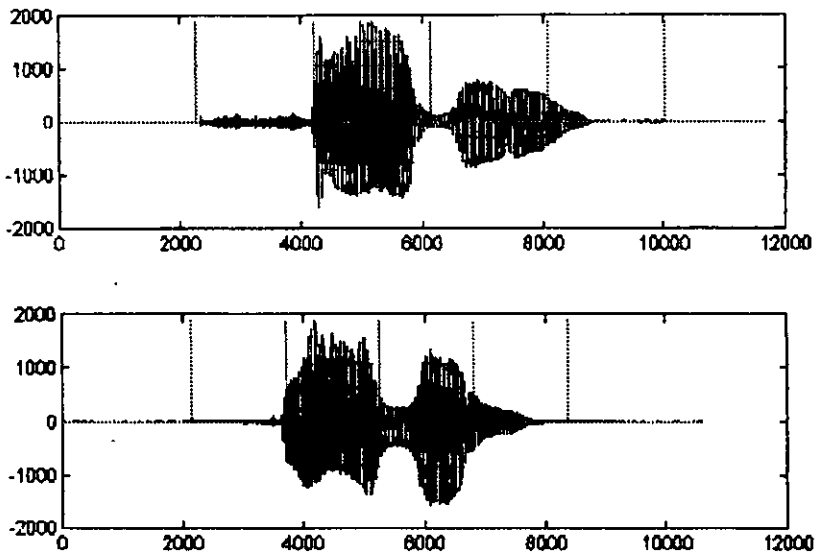

Figura IV.2. Dos palabras "Seven", segmentadas linealmente en cuatro subpalabras

Por lo tanto, fue necesario el empleo de otros tipos de segmentación, requiriéndose segmentaciones que agrupen las muestras de la señal en segmentos con características similares.

IV.2 SEGMENTACIÓN NO LINEAL (KM)

Es un tipo de segmentación experimental basada completamente en el algoritmo de agrupamiento K-Medias, con ciertas variaciones. La principal variante es que no interesa saber los cuantizadores generados, sino que se necesita conocer la ubicación de cada vector de coeficientes LPC en relación con esos cuantizadores, es decir, que vector corresponde a que clúster.

El algoritmo para segmentar consiste en lo siguiente:

1.- Calculo de inicio y fin de palabra.

2.- Obtención de coeficiente LPC y coeficientes de autocorrelación.

3.- Con los coeficientes LPC calcular 2 cuantizadores (centroides) utilizando el algoritmo K-Medias y la distancia de Itakura-Saito modificada.

4.- Agrupar los clústers de vectores que correspondan al mismo cuantizador de acuerdo a los siguientes criterios:

- Para que un clúster cuente como segmento deberá tener por lo menos tres elementos.
- Los clústers que tengan uno o dos elementos deberán considerarse como parte del mismo grupo de los vectores que los rodean y serán absorbidos.

5.- Si después de realizar el paso 4 no existen como mínimo dos clústers con tres o más elementos, se realiza otra cuantización con un centroide mas. Esto se repite hasta un máximo de 5 centroides.

6.- La segmentación se realiza en base a lo siguiente:

- Si existen solo dos clústers, a cada uno le corresponde un segmento de voz.
- Si existen tres clústers, el clúster mayor cuenta como un solo segmento y los dos restantes como otro segmento.
- Para cuatro clústers se toman en pares y a cada par le corresponde un un segmento.
- Para cinco clústers, el menor de los clústers intermedios es absorbido y se consideran como si se tuvieran solo tres.
- Si se tienen seis o mas clústers, esta parte de la señal no se puede segmentar y se considera como un solo segmento.

7.- Si el número de segmentos es igual al requerido, termina el algoritmo, en caso contrario regresar al paso 3 y realizar la cuantización con el clúster que tenga el mayor numero de elementos. El clúster mayor será eliminado del algoritmo si se llega a la última condición especificada en el paso 6 y se continua con el clúster que le sigue en tamaño.

La segmentación se basa en que los vectores LPC al agruparse quedan en un mismo clúster si tienen características similares. Este algoritmo ha demostrado tener buenos resultados y poder segmentar casi todas las señales de voz utilizadas hasta un total de 8 segmentos.

Como ejemplo y por ser una de las propuestas de la tesis, dividamos en cuatro subpalabras una repetición del dígito "eight".

Calculo de inicio y fin. Utilizando el algoritmo Rabiner-Sambur, los puntos de inicio y fin se obtuvieron en las tramas 18 y 57, respectivamente. Lo que corresponde a los puntos 2304 y 7296, ya que se utilizaron ventanas de 128 muestras sin traslape.

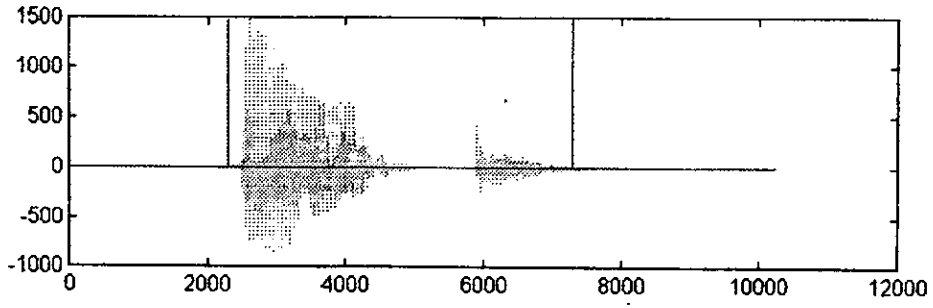

Figura IV.3. Señal con Inicio y Fin

Con lo 39 vectores LPC restantes, se obtienen 2 cuantizadores. El reporte correspondiente quedó de la siguiente forma:

1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

Con base en el punto 4, el primer elemento es absorbido y de acuerdo al punto 6 el corte se encuentran en la trama 40 (18+22) o el punto 5120.

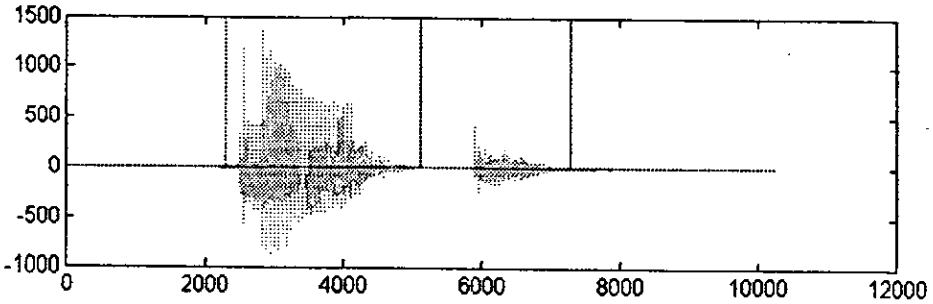

Figura IV.4. Señal segmentada en dos subpalabras

El clúster mayor es el primero puesto que tiene 22 elementos (paso 7), se realiza entonces la cuantización a dos centroides y se obtiene el siguiente reporte,

1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1

El primer y último elemento son absorbidos y como no existen 2 clústers de mas de tres elementos, se realiza la cuantización ahora con tres centroides (paso 5) resultando,

1 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 1

El primer y ultimo elementos son absorbidos (paso 4) y de acuerdo al paso 6 el corte es en la trama 26 (18+8) o bien el punto 3328.

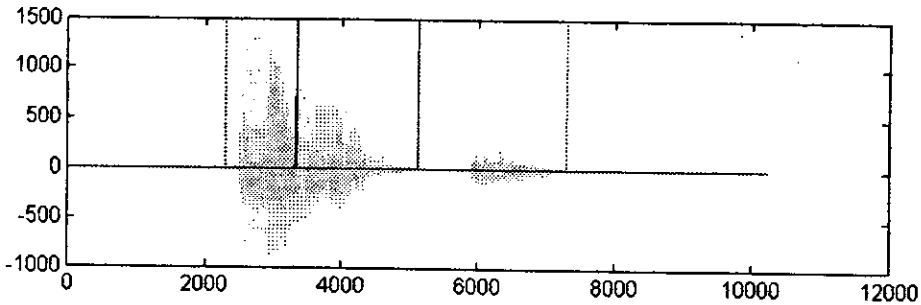

Figura IV.4. Señal segmentada en tres subpalabras

Ahora se realiza la cuantización a dos centroides para el segmento ubicado entre las tramas 40 y 57 por ser el mayor (paso 7), el reporte que se obtiene,

1 1 1 1 1 1 2 2 2 2 2 1 1 1 1 1

De acuerdo al paso 6 el corte es ahora en la trama 46 (40+6) o el punto 5888.

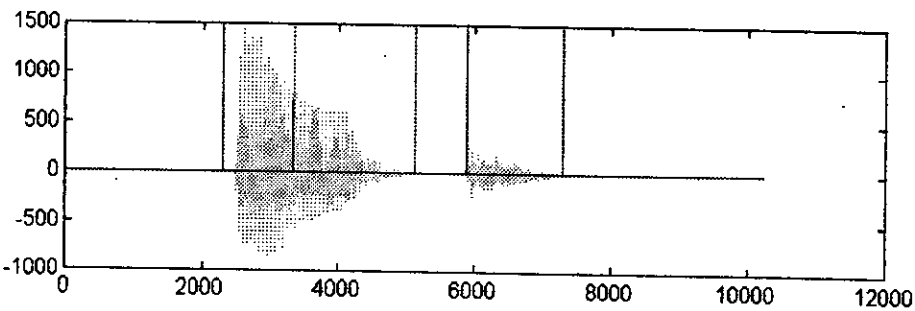

Figura IV.4. Señal segmentada en cuatro subpalabras

Como el numero de segmentos es igual al requerido, el algoritmo termina y se tiene la palabra segmentada en 4 subpalabras. Los segmentos se ubican entre los puntos: 2304, 3328, 5120, 5888 y 7296. Y como lo demuestran las gráficas, las muestras quedan agrupadas en segmentos dentro de los cuales la señal presenta características similares.

IV.3 SEGMENTACIÓN ACÚSTICA (MLR)

El tercer tipo de segmentación utilizada, segmentación acústica, es un tipo de segmentación que busca cambios estadísticos significativos de las señales de voz, realizando una segmentación de las palabras en unidades acústicas, al comparar segmentos de análisis mediante una medida de cambio en la señal y que es equivalente a segmentar la palabra en fonemas o regiones cuasi-estacionarias de la señal.

La comparación de los segmentos de análisis se realiza por pares, se divide el contenido frecuencial de la señal en L bandas críticas, y se aproxima su distribución de probabilidad a una distribución normal con media cero, para cada banda, se calcula su diferencia aplicando una metodología de clasificación de la razón de máxima verosimilitud que detecte cambios espectrales, utilizando una prueba de MLR con una ventana deslizante.

La metodología MLR esta basada en un criterio de decisión binaria, que en un espacio L-dimensional se define de la siguiente forma,

$$\Lambda(z) = \frac{p(\bar{z}|m_2)}{p(\bar{z}|m_1)} \underset{H_2}{\overset{H_1}{\gtrless}} \eta \qquad \text{IV.1}$$

En esta ecuación, **z** es un elemento del espacio de observaciones **Z** de dimensión L, H1 y H2 son las dos decisiones posibles, m1 y m2 son las señales, p(z) es la función de densidad de las observaciones y η es el umbral de comparación.

A(z) se define como el cociente de máxima verosimilitud, y si el resultado del cociente es mayor que el umbral se decide por H1, en caso contrario, cuando es menor, se decide H2. En nuestro caso H1 sería si esta parte de la señal pertenece al mismo segmento y H2 sería que no pertenece.

Para el estudio de las señales de voz, **z** es de la forma $\mathbf{z}=[z_1, z_2,\ldots, z_L]$. Y por la aproximación a una distribución normal, la función densidad de z se define como,

$$p(z) = \prod_{i=1}^{L} \frac{1}{\sigma^2\sqrt{2\pi}} \exp\left\{-\frac{(z_i-\mu_i)^2}{\sigma^2}\right\} \qquad \text{IV.2}$$

Por lo que el cociente MLR se expresa como,

$$\Lambda(z) = \frac{\prod_{j=1}^{L} \frac{1}{\sigma_1^2\sqrt{2\pi}} e^{-\frac{1}{2}\frac{z_{1,j}^2}{\sigma_1^2}}}{\prod_{j=1}^{L} \frac{1}{\sigma_2^2\sqrt{2\pi}} e^{-\frac{1}{2}\frac{z_{2,j}^2}{\sigma_2^2}}} \qquad \text{IV.3}$$

Tomando el logaritmo natural en ambos lados,

$$\ln\Lambda(z) = \lambda = \sum_{i=1}^{L}\left[\ln\frac{\sigma_2^2}{\sigma_1^2} + \frac{1}{2}\frac{z_{2,j}^2}{\sigma_1^2} - \frac{1}{2}\frac{z_{1,j}^2}{\sigma_2^2}\right] \qquad \text{IV.4}$$

Ya que la media de las señales se considera cero, se puede demostrar que $\sum_{i=1}^{L}\frac{x_i^2}{\sigma^2} = L$, por lo que la expresión anterior se reduce a:

$$\lambda = L\ln\frac{\sigma_2^2}{\sigma_1^2} \qquad \text{IV.5}$$

Expresión que nos muestra que la comparación se reduce a una prueba de varianzas. Y que la segmentación se realiza al encontrar transiciones espectrales abruptas en la señal.

V. SISTEMAS DE RECONOCIMIENTO

V.1 SISTEMA DE RECONOCIMIENTO LPC

En la figura V.1, se muestra el esquema del sistema de reconocimiento de palabras aisladas utilizando coeficientes LPC.

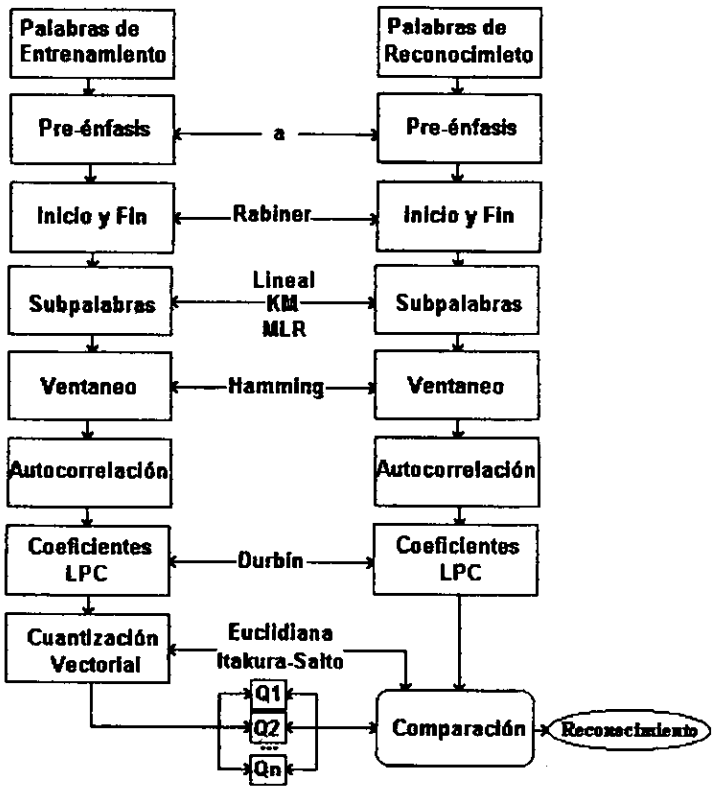

Figura V.1. Sistema de Reconocimiento LPC

Coeficientes de Predicción Lineal (LPC)

Los coeficientes LPC se obtiene al resolver el sistema de pxp ecuaciones simultáneas que resultan de el método de autocorrelación, o bien dadas las características que presenta la matriz Toeplitz, se pueden obtener con el empleo de métodos numéricos. Como lo es el algoritmo de Levinson-Durbin.

Algoritmo de Levinson-Durbin

Inicialización:

$$E_{LP}^{(0)} = r(0)$$

$$Para \quad 1 \le i \le N_{LP}$$

$$\{$$

$$k_i = \frac{\left[r(i) - \sum_{j=1}^{L-1} \alpha_j^{(i-1)} r(|i-j|) \right]}{E^{(i-1)}} \qquad 1 \le i \le p; \quad 1 \le j \le i-1$$

$$\alpha_i^{(i)} = k_i$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}$$

$$E^{(i)} = \left(1 - k_i^2\right) E^{(i-1)}$$

Las iteraciones son realizadas para i = 1, 2, 3, ..., p, y la solución final esta dada por:

$$a_m = coeficientes\ LPC = \alpha_m^{(p)}$$

Distorsión de Itakura-Saito

Una medida de distorsión alternativa que cumple con la condición de no simetría, que se utilizada en la cuantización de los coeficientes de predicción y que se deriva de los principios de máxima similitud, es la distancia de Itakura-Saito. Resulta mas costosa en su cálculo, pero esta ligada al espectro de la señal de voz:

$$d(X_1, H) = \int_{-\pi}^{\pi} \left[\frac{X_1(\omega)}{H(\omega)} \right] \frac{d\omega}{2\pi} - \int_{-\pi}^{\pi} \log \left[\frac{X_1(\omega)}{H(\omega)} \right] d\omega - 1 \qquad V.1$$

donde X_1 es el espectro de la señal de voz y H es el espectro del filtro.

Una versión modificada de la distancia de Itakura-Saito, que es una distancia de máxima similitud entre un vector de coeficientes de predicción X_1 y otro vector de coeficientes X_2 está dada por,

$$d(X_1,X_2) = (X_1 - X_2)^T \Phi_X (X_1 - X_2) \qquad V.2$$

donde

$$\Phi_X = \{ \phi(i-k)/\phi(0),\ 0 \le i,k \le N-1 \} \qquad V.3$$

es la matriz de autocorrelación normalizada cuyos coeficientes $\phi(i-k)$ son usados para el cálculo del vector coeficientes de predicción X_1.

Ya que los coeficientes de autocorrelación están normalizados por $\phi(0)$, se puede demostrar que la matriz Φ_x y el vector X_1 determinan de forma única el uno al otro. Se puede notar además que Φ_x es una matriz de ponderación, pero a diferencia de la ecuación (I.21) donde W esta fija, en esta ultima ecuación Φ_x cambia de valor al cambiar X_1. Con lo que se comprueba que la distancia de Itakura-Saito no es una distancia métrica.

V.2 SISTEMA DE RECONOCIMIENTO CEPSTRAL

Un esquema de el sistema de reconocimiento de voz utilizando Coeficientes Cepstral se presenta en la figura V.2.

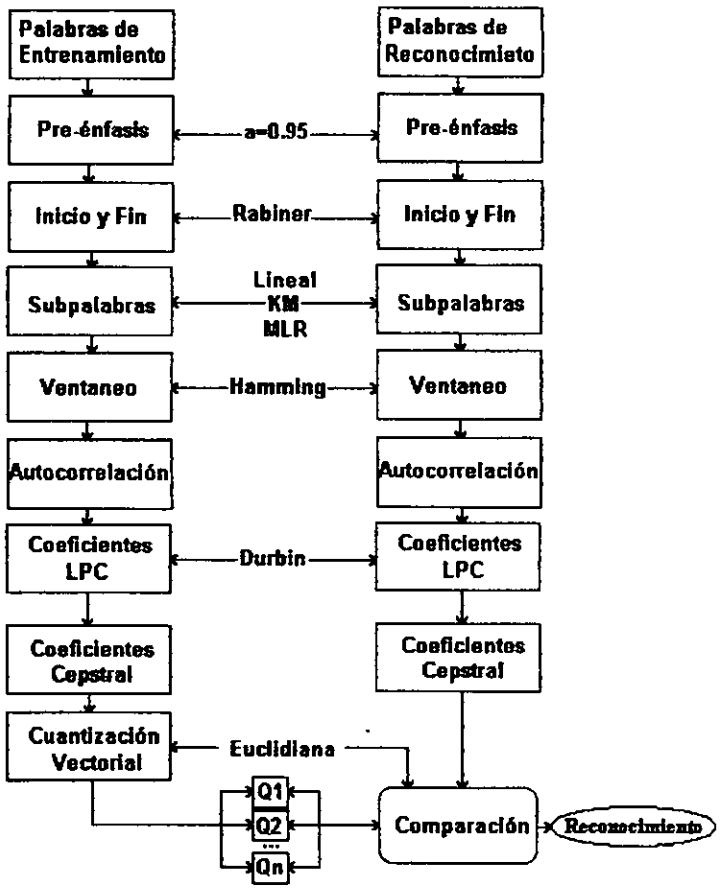

Figura V.2. Sistema de Reconocimiento Cepstral

Coeficientes Cepstral

El cálculo de los coeficientes Cepstral se realiza a partir de los coeficientes LPC, de acuerdo a la siguiente expresión,

$$c_m = a_m + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k} \qquad 1 \le k \le m \qquad \text{V.4}$$

Donde los a_m son los coeficientes LPC y los c_m son llamados coeficientes LPC Cepstral, que son los coeficientes Cepstral derivados del Análisis LPC. No son iguales a los coeficientes Cepstral obtenidos a partir de la señal de voz real, sin embargo son una buena aproximación y han mostrado generar buenos resultados, además de ser fácilmente obtenidos.

V.3 SISTEMA DE RECONOCIMIENTO KLT

Un esquema de el sistema de reconocimiento de voz utilizando Transformada KLT se presenta en la figura V.3.

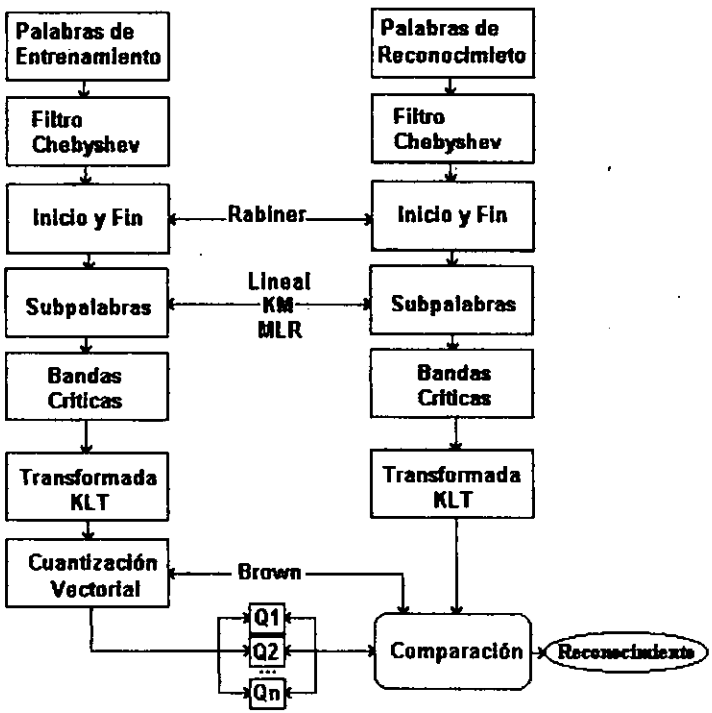

Figura V.3. Sistema de Reconocimiento KLT

Matrices KLT

Para generar las matrices resultantes de la transformada KLT, se toma el espectro de bandas críticas de una subpalabra de la señal y se realiza la transformada, así se tiene un vector de componentes aleatorias inicial de dimensión fija.

$$x = \begin{bmatrix} x_{BC1}(m) \\ x_{BC2}(m) \\ \vdots \\ x_{BC18}(m) \end{bmatrix} \qquad \text{V.5}$$

Es decir, que se considera un vector de variables aleatorias, en el cual cada variable representa una banda crítica (ecuación V.5). Por lo tanto, se tiene un vector de dimensión 18, y en el que cada elemento es a su vez otro vector cuya dimensión estará dada por el número de tramas que se consideren dentro de una subpalabra de la señal. Con esto se reduce la información con respecto a coeficientes LPC o Cepstral, ya que para estos dos casos, se generan un número determinado de vectores LPC o Cepstral por subpalabra, mientras que para KLT se genera solo una matriz por subpalabra.

Filtro Chebyshev

En la primera etapa del procesamiento, se aplica a la señal de entrada un filtro digital Chebyshev paso bajas (el orden del filtro no es relevante), con el fin de limitar el intervalo de frecuencias de la señal, debido a que la mayor cantidad de energía de una señal de voz se encuentra por debajo de los 5 kHz.

Distorsión de Brown

Para el caso de la transformada KLT, la medida de similitud de Brown, que básicamente es una medida de semejanza espectral, ha mostrado resultados satisfactorios. Se puede calcular como sigue:

$$d = \frac{\lambda^T \left(I - A^T A\right)\lambda}{l} \qquad \text{V.6}$$

con los elementos de la matriz A:

$$a_{ij} = \left\{x_1[i], x_2[j]\right\}^2 \qquad \text{V.7}$$

donde $X_1[i]$:renglón i-ésimo de la matriz KLT para la señal 1
donde $X_2[k]$:renglón i-ésimo de la matriz KLT para la señal 2
λ: vector columna con los valores propios de la primer señal, de mayor a menor
l: nivel de comparación (numero de vectores KLT a utilizar).

Esta medida aprovecha la propiedad de extracción de vectores de energía, jerárquicamente ordenados (con los valores propios correspondientes de mayor a menor) de la transformada KLT. El uso de esta técnica presenta una ventaja adicional: si se utilizan vectores KLT, de orden 18, se puede demostrar que cerca del 99% de la energía de la señal modelada se encuentra en los primeros 7 vectores, por lo que es posible caracterizar la señal con menos información que en otras transformaciones.

V.4 CUANTIZACION VECTORIAL

Los patrones de referencia o cuantizadores que serán utilizados para realizar el reconocimiento de palabras aisladas, se generan con el uso del algoritmo de agrupamiento K-Medias. Este algoritmo se aplica a los conjuntos de parámetros (vectores para LPC y Cepstrum y matrices para KLT) que se forman al reunir todos los que corresponden al mismo segmento de la señales de entrenamiento que representan a la misma palabra.

V.5 COMPARACIÓN Y RECONOCIMIENTO

La comparación se hace al medir la distorsión entre los parámetros de la señal de voz a reconocer y los patrones de referencia de cada palabra. Se comparan los parámetros por segmentos correspondientes, se toman las menores distorsiones por segmento y se suman. El reconocimiento se realiza al obtener la menor de las distorsiones totales a los conjuntos de patrones.

V.6 RESULTADOS

Los sistemas de reconocimiento descritos en este capitulo se implementaron con el uso de las bases de palabras aisladas de entrenamiento y de reconocimiento generadas en los laboratorios de Texas Instruments. Estas bases fueron diseñadas exclusivamente para efectuar pruebas en laboratorio de reconocimiento de voz.

La base para la fase de entrenamiento consiste en 100 repeticiones de cada uno de los dígitos en Inglés. Cada dígito es pronunciado 10 veces por 10 locutores distintos (hombres y mujeres) con lo que se tiene una base de 1000 palabras.

La base para la fase de reconocimiento consiste de igual forma en repeticiones de los dígitos en Inglés pronunciados por los mismo locutores, pero para esta base cada dígito es pronunciado 16 veces por locutor, con lo cual tenemos un total de 1600 palabras en la base. Con el uso de estas dos bases se genera un sistema de reconocimiento independiente del locutor.

Básicamente se realizaron 5 pruebas para cada tipo de parametrización (LPC, Cepstral y KLT). El cambio en cada prueba consistió únicamente en utilizar un numero diferente de segmentos por palabra (de 4 a 8). Sin embargo, dado que se utilizaron 3 tipos de segmentación, se tienen 15 pruebas por parametrización. Pero además como para coeficientes LPC se utilizaron dos medidas de distorsión, el total de pruebas fue: 30 con coeficientes LPC, 15 con coeficientes Cepstral y 15 con transformada KLT .

A continuación se presentan los resultados, que son colocados en matrices de confusión, en los renglones se colocan las palabras a reconocer, las columnas indican como fueron reconocidas y la diagonal principal presenta las palabras reconocidas exitosamente.

COEFICIENTES LPC

Filtro de pre-énfasis con a=0.95
Segmentación Lineal
Ventanas de 128 muestras sin traslape
Vectores LPC de orden 8
Distorsión Euclidiana cuadrática
16 cuantizadores vectoriales por segmento

4 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	154		1					5			96.2500
UNO		149	1			2		1		7	93.1250
DOS			151	2		3		4			94.3750
TRES				160							100.000
CUATRO			1		158	1					98.7500
CINCO				1	2	154		2		1	96.2500
SEIS							160				100.000
SIETE						3	1	155			97.4843
OCHO				2					157	1	98.1250
NUEVE				1		10				148	93.0818
Prom.											96.744

5 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	157		1					2			98.1250
UNO		139			1	4				16	86.8750
DOS	1		157	1		1					98.1250
TRES				159		1					99.3750
CUATRO					160						100.000
CINCO					2	157				1	98.1250
SEIS							160				100.000
SIETE						1	1	157			98.7421
OCHO				3			2		154	1	96.2500
NUEVE			1					2		157	94.3396
Prom.											96.998

6 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	155			2				3			96.8750
UNO		140			1	4		1		14	87.5000
DOS	1		152	1		3		3			95.0000
TRES				157		2		1			98.1250
CUATRO					159	1					99.3750
CINCO						155		1		4	96.8750
SEIS							160				100.000
SIETE			1			4	1	153			96.2264
OCHO				4			2		153	1	95.6250
NUEVE		1				10				148	93.0818
Prom.											95.868

7 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	155		3					2			96.8750
UNO	1	140			1	4				14	87.5000
DOS	1		155	1		1		2			96.8750
TRES				155	1	3		1			96.8750
CUATRO	1				158	1					98.7500
CINCO					1	158		1			98.7500
SEIS							160				100.000
SIETE						2	1	156			96.1132
OCHO				4			2		152	1	95.5975
NUEVE						10				149	93.7107
Prom.											96.305

8 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	154		2					4			96.2500
UNO	2	146				4				8	91.2500
DOS	1		148	3	1	1		6			92.5000
TRES	1			152		2		5			95.0000
CUATRO	2		1		156	1					97.5000
CINCO						154		4		2	96.2500
SEIS							160				100.000
SIETE			1			2	1	155			97.4843
OCHO				5			7		147	1	97.8750
NUEVE		1		1		19				138	86.7925
Prom.											94.490

Filtro de pre-énfasis con a=0.95
Segmentación Lineal
Ventanas de 128 muestras sin traslape
Vectores LPC de orden 8
Distorsión Itakura-Saito modificada
16 cuantizadores vectoriales por segmento

4 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	160										100.000
UNO		156						1		3	97.5000
DOS			156	4							97.5000
TRES				160							100.000
CUATRO					159	1					99.375
CINCO						159				1	99.3750
SEIS							160				100.000
SIETE								159			100.000
OCHO				3					157		98.1250
NUEVE						1				158	99.3711
Prom.											99.125

5 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	160										100.000
UNO		154						1		5	96.2500
DOS			157	3							98.1250
TRES				160							100.000
CUATRO	1				159						99.3750
CINCO						154		1		5	96.2500
SEIS							160				100.000
SIETE			1					158			99.3711
OCHO				5					155		96.8750
NUEVE										159	100.000
Prom.											98.625

6 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	160										100.000
UNO		156						1		3	97.5000
DOS			155	5							96.8750
TRES				160							100.000
CUATRO					158	1		1			98.7500
CINCO						157				3	98.1250
SEIS							160				100.000
SIETE								159			100.000
OCHO									156	1	97.5000
NUEVE										159	100.000
Prom.											98.875

7 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	160										100.000
UNO		155						1		4	96.8750
DOS			158	2							98.7500
TRES				159		1					99.3750
CUATRO					160						100.000
CINCO						160					100.000
SEIS							160				100.000
SIETE								159			100.000
OCHO				5					153	1	96.2264
NUEVE						1				158	99.3711
Prom.											99.059

8 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	160										100.000
UNO		155						1		4	96.8750
DOS			151	8				1			94.3750
TRES				158		1		1			98.7500
CUATRO					159			1			99.3750
CINCO						156		3		1	97.5000
SEIS							160				100.000
SIETE								159			100.000
CHO				4					156		97.5000
NUEVE						1				158	99.3711
Prom.											98.375

Filtro de pre-énfasis con a=0.95
Segmentación KM
Ventanas de 128 muestras sin traslape
Distorsión Euclidiana cuadrática
16 cuantizadores vectoriales por segmento

4 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	137	6	4		1			6		6	85.6250
UNO		128				1				31	80.0000
DOS	24		131		1	2		2		1	81.8750
TRES	1			155		2		2			96.8750
CUATRO	1	1			157	1					98.1250
CINCO	1	1		2		145		3		8	90.6250
SEIS							157	2	1		98.1250
SIETE						2		157			98.7421
OCHO				1			1		158		98.7500
NUEVE		5	1	1						152	95.5975
Prom.											92.433

5 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	151		5					1		3	94.3750
UNO		134						1		24	84.2767
DOS			145	2	1	2					92.9487
TRES				159		1					99.3750
CUATRO		2			156	1		1			97.5000
CINCO	1			1		147		4		7	91.8750
SEIS							157				100.000
SIETE			2				3	150			96.7742
OCHO				1			4		151	1	96.1783
NUEVE		4		1		2		1		151	94.9686
Prom.											94.827

6 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	158		2								98.7500
UNO		131				2				26	82.3899
DOS			153			2		2			97.4522
TRES				154		3		3			96.2500
CUATRO					157	1		1		1	98.1250
CINCO				1		153				6	95.6250
SEIS							160				100.000
SIETE			3			3	2	151			94.9686
OCHO							2		152	1	98.0645
NUEVE		1		1		4				153	96.2264
Prom.											95.785

7 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	152		5					1		1	95.5975
UNO		144				2				8	93.5065
DOS			141	1	1						98.6014
TRES				151		1		4			96.7949
CUATRO		1			155	1					98.7261
CINCO					1	150	1	2		3	95.5414
SEIS							156				100.000
SIETE			3			3		148			96.1039
OCHO							1		138		99.2806
NUEVE		1		1						156	98.7342
Prom.											97.289

8 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	152		3					3			96.2025
UNO		124				5				17	84.9315
DOS			135			3					97.8261
TRES				149			1				99.3333
CUATRO					156	1		1		1	98.1132
CINCO						153	2			5	95.6250
SEIS							160				100.000
SIETE	1		5			4		149			93.7175
OCHO							2		137		98.5612
NUEVE		1				2				156	98.1132
Prom.											96.342

Filtro de pre-énfasis con a=0.95
Segmentación KM
Ventanas de 128 muestras sin traslape
Distorsión Itakura-Saito modificada
16 cuantizadores vectoriales por segmento

4 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	158	1	1								98.7500
UNO		150								10	93.7500
DOS	3		154	3							96.2500
TRES				159				1			99.3750
CUATRO		4			156						97.5000
CINCO					1	153		1		6	95.6250
SEIS							158	1	1		98.7500
SIETE	2		1				1	155			97.4843
OCHO	1			3					156		97.5000
NUEVE		3		1						155	97.4843
Prom.											97.247

5 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	160										100.000
UNO		148						1		11	92.5000
DOS			158	2							98.7500
TRES				160							100.000
CUATRO		4			155			1			96.8750
CINCO					1	157				2	98.1250
SEIS							159		1		99.3750
SIETE	1							157		1	98.7421
OCHO				2					157		98.7421
NUEVE										159	100.000
Prom.											98.311

6 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	160										100.000
UNO		151						1		7	94.9686
DOS	1		151	4				1			96.1783
TRES				160							100.000
CUATRO		4			155			1			96.8750
CINCO						158				2	98.7500
SEIS							159		1		99.3750
SIETE			2					156		1	98.1132
OCHO									155		100.000
NUEVE										159	100.000
Prom.											98.260

7 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	159										100.000
UNO		147						1		6	95.4545
DOS			140	3							97.9021
TRES				156							99.3750
CUATRO		2			155						98.7261
CINCO						155	1			1	98.7261
SEIS							155		1		98.3590
SIETE	1		3					149		1	96.7532
OCHO									139		100.000
NUEVE										158	100.000
Prom.											98.529

8 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	157		1								99.3671
UNO		136		1						9	93.1507
DOS			138								100.000
TRES				150							100.000
CUATRO		3			155	1					97.4843
CINCO						159				1	99.3750
SEIS							158		2		98.7500
SIETE			5					154			96.8553
OCHO									139		100.000
NUEVE										159	100.000
Prom.											98.498

Filtro de pre-énfasis con a=0.95
Segmentación MLR
Ventanas de 128 muestras sin traslape
Distorsión Euclidiana cuadrática
16 cuantizadores vectoriales por segmento

4 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	133	1	5	1		2		5		3	88.6667
UNO		121	1			4				28	78.5714
DOS	1		134		1	1	1	11			89.9329
TRES				145	1	1		3			96.6667
CUATRO		2		1	140	4				1	94.5946
CINCO					3	147	2	2		2	94.2308
SEIS							160				100.000
SIETE							1	156			99.3631
OCHO				1			9		138		93.2432
NUEVE		8				14		3		133	84.1772
Prom.											91.945

5 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	136		3	1				7		2	91.2752
UNO		114	1							19	85.0746
DOS			98			1		6			93.3333
TRES				120		2		5	1		93.7500
CUATRO		1			137			1		1	97.8571
CINCO				1	2	147				3	96.0784
SEIS							159				100.000
SIETE						1	2	154			98.0892
OCHO							2		116		98.3051
NUEVE		2	1	1		5		2		140	92.7152
Prom.											94.648

6 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	134		3	1				4		4	97.7808
UNO		83								15	84.6939
DOS			64			1		2			95.5224
TRES				87		1		2			96.6667
CUATRO					112	1	1				98.2456
CINCO						127				11	92.0290
SEIS							148				100.000
SIETE	2		2			3		126			94.7368
OCHO							2		67		97.1014
NUEVE		2				1				127	97.6923
Prom.											94.847

7 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	117					1		3		1	95.9016
UNO		39						1		13	73.5849
DOS	1		27			1		1			90.0000
TRES				48		4		1			90.5660
CUATRO					68	1				1	97.1429
CINCO						109		1		2	97.3214
SEIS							122				100.000
SIETE						2		88			97.7778
OCHO							3		27		90.0000
NUEVE		1				4				111	95.6897
Prom.											92.798

8 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	82							2			97.6190
UNO		24						1		1	92.3077
DOS			15			1					93.7500
TRES				18		2				1	85.7143
CUATRO					37	1					97.3684
CINCO	1					77				3	95.0617
SEIS							90				100.000
SIETE	1		1		1	1		47			92.1569
OCHO						1	1		8		80.0000
NUEVE		4				3				76	91.5663
Prom.											92.554

Filtro de pre-énfasis con a=0.95
Segmentación MLR
Ventanas de 128 muestras sin traslape
Distorsión Itakura-Saito modificada
16 cuantizadores vectoriales por segmento

4 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	146		1			1				2	97.3333
UNO		136						1		17	88.3117
DOS			147	2							98.6577
TRES				149				1			99.3333
CUATRO		1			144	3					97.2973
CINCO						153				3	98.0769
SEIS							158		2		98.7500
SIETE	1		2			1	1	152			96.8153
OCHO									148		100.000
NUEVE		1	1			3		2		151	95.5696
Prom.											97.015

5 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	148							1			99.3289
UNO		117						1		16	87.3134
DOS			104	1							99.0476
TRES				128							100.000
CUATRO					137	1		2			97.8751
CINCO						150				3	98.0392
SEIS							159				100.000
SIETE			1					156			99.3631
OCHO									117	1	99.1525
NUEVE	1		1							149	98.6755
Prom.											97.878

6 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	143		1					1		1	97.9452
UNO		91								7	92.8571
DOS			66	1							98.5075
TRES				90							100.000
CUATRO					113	1					99.1228
CINCO						133				5	96.3768
SEIS							147		1		99.3243
SIETE			3				2	128			96.2406
OCHO							1		67	1	97.1014
NUEVE		2				1				127	97.6923
Prom.											97.517

7 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	119		1					2			97.5410
UNO		44						4		8	89.0189
DOS			30								100.000
TRES				53							100.000
CUATRO					67	1		2			95.7143
CINCO						109		1		2	97.3214
SEIS							122				100.000
SIETE			2					87		1	96.6667
OCHO									30		100.000
NUEVE		1								115	99.1379
Prom.											96.940

8 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	84										100.000
UNO		22						1		3	84.6154
DOS	2		14								87.5000
TRES				21							100.000
CUATRO					37	1					97.3684
CINCO						81					100.000
SEIS							90				100.000
SIETE			1					50			98.0932
OCHO							1		9		90.0000
NUEVE		2								81	97.5904
Prom.											95.511

COEFICIENTES CEPSTRAL

Filtro de pre-énfasis con a=0.95
Segmentación Lineal
Ventanas de 128 muestras sin traslape
Distorsión Euclidiana cuadrática
16 cuantizadores vectoriales por segmento

4 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	160										100.000
UNO		158						1		1	98.7500
DOS			158	2							98.7500
TRES				160							100.000
CUATRO					160						100.000
CINCO						156				4	97.5000
SEIS							160				100.000
SIETE							2	157			98.7421
OCHO				1					159		99.3750
NUEVE						1				158	99.3711
Prom.											99.249

5 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	160										100.000
UNO		158						1		1	98.7500
DOS			158	2							98.7500
TRES				160							100.000
CUATRO					160						100.000
CINCO						160					100.000
SEIS							160				100.000
SIETE								159			100.000
OCHO				3					157		98.1250
NUEVE						1				158	99.3711
Prom.											99.499

6 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	160										100.000
UNO		158						1		1	98.7500
DOS			155	5							96.8750
TRES				160							100.000
CUATRO					160						100.000
CINCO						160					100.000
SEIS							160				100.000
SIETE							1	158			99.3711
OCHO				4					156		97.5000
NUEVE										159	100.000
Prom.											99.249

7 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	160										100.000
UNO		158						1		1	98.7500
DOS			157	2				1			98.1250
TRES				158		1		1			98.7500
CUATRO					160						100.000
CINCO						160					100.000
SEIS							160				100.000
SIETE							1	158			99.3711
OCHO				6					153		96.2264
NUEVE						1				158	99.3711
Prom.											99.059

8 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	160										100.000
UNO		159						1			99.3750
DOS			150	10							93.7500
TRES				156		1		3			97.5000
CUATRO					160						100.000
CINCO						159		1			99.3750
SEIS							160				100.000
SIETE							1	158			99.3711
OCHO				4					156		97.5000
NUEVE						2				157	98.7421
Prom.											98.561

Filtro de pre-énfasis con a=0.95
Segmentación KM
Ventanas de 128 muestras sin traslape
Distorsión Euclidiana cuadrática
16 cuantizadores vectoriales por segmento

4 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	159			1							99.3750
UNO		157								3	98.1250
DOS			154	5						1	95.2500
TRES	1			157				1		1	98.1250
CUATRO		2			158						98.7500
CINCO						156		1		3	97.5000
SEIS							158	1	1		98.7500
SIETE							1	158			99.3711
OCHO				2			1		157		98.1250
NUEVE		1				1				157	98.7421
Prom.											98.311

5 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	159							1			99.3750
UNO		157			1			1		1	98.1250
DOS	1		158	1							98.7500
TRES				160							100.000
CUATRO		2			157			1			98.1250
CINCO						157		1		2	98.1250
SEIS							160				100.000
SIETE								159			100.000
OCHO				2					157		98.7421
NUEVE										159	100.000
Prom.											99.124

6 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	159		1								99.3750
UNO		155						1		2	98.1013
DOS			152	4				1			96.8153
TRES				160							100.000
CUATRO					159			1			99.3750
CINCO						159				1	99.3750
SEIS							160				100.000
SIETE							2	157			98.7421
OCHO				1					154		99.3548
NUEVE										159	100.000
Prom.											99.114

7 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	160										100.000
UNO		156						1			99.3631
DOS			151	4							97.4194
TRES				158							100.000
CUATRO		1			159						99.3750
CINCO						157	1			2	98.1250
SEIS							160				100.000
SIETE								159			100.000
OCHO									150		100.000
NUEVE										159	100.000
Prom.											99.428

8 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	158										100.000
UNO		141	1							4	96.5753
DOS			138								100.000
TRES				150							100.000
CUATRO					159						100.000
CINCO						160					100.000
SEIS							160				100.000
SIETE								159			100.000
OCHO									139		100.000
NUEVE										159	100.000
Prom.											99.658

Filtro de pre-énfasis con a=0.95
Segmentación MLR
Ventanas de 128 muestras sin traslape
Distorsión Euclidiana cuadrática
16 cuantizadores vectoriales por segmento

4 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	145	1	1					1		2	96.6667
UNO		145						1		8	94.1558
DOS	1		144	4							96.6443
TRES				150							100.000
CUATRO					148						100.000
CINCO						152				4	97.4359
SEIS							160				100.000
SIETE								157			100.000
OCHO									148		100.000
NUEVE		2				4		1		151	95.5696
Prom.											98.047

5 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	147		1					1			98.6577
UNO		126						1		7	94.0299
DOS			104	1							99.0476
TRES				128							100.000
CUATRO		1			137	1		1			97.8571
CINCO						151				2	98.6928
SEIS							159				100.000
SIETE								157			100.000
OCHO									118		100.000
NUEVE		1				1				149	98.6755
Prom.											98.696

6 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	144		1							1	98.6301
UNO		95								3	96.9388
DOS			66	1							98.5075
TRES				90							100.000
CUATRO					114						100.000
CINCO						136				2	98.5507
SEIS							148				100.000
SIETE								133			100.000
OCHO				1					68		98.5507
NUEVE										130	100.000
Prom.											99.118

7 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	121							1			99.1803
UNO		48						1		4	90.5660
DOS			30								100.000
TRES				53							100.000
CUATRO					70						100.000
CINCO						111				1	99.1071
SEIS							122				100.000
SIETE								89		1	98.8889
OCHO									30		100.000
NUEVE										116	100.000
Prom.											98.774

8 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	84										100.000
UNO		24						1		1	92.3077
DOS			15	1							93.7500
TRES				19		1				1	90.4762
CUATRO					38						100.000
CINCO						81					100.000
SEIS							90				100.000
SIETE								51			100.000
OCHO							1		9		90.0000
NUEVE		2								81	97.5904
Prom.											96.412

TRANSFORMADA KLT

Filtro Chebyshev a 4.5 kHz
Segmentación Lineal
Ventanas de 256 muestras
Vectores de valores característicos de orden 1x7
Vectores característicos de orden 7x18
Distorsión de Brown
16 cuantizadores vectoriales por segmento

4 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	**160**										**100.000**
UNO		**158**						1		1	**98.7500**
DOS			**159**						1		**99.3750**
TRES				**160**							**100.000**
CUATRO					**160**						**100.000**
CINCO						**159**				1	**99.3750**
SEIS							**160**				**100.000**
SIETE						1	5	**152**		1	**95.5975**
OCHO				1			3		**156**		**97.5000**
NUEVE		3				1	1			**154**	**96.8553**
Prom.											**98.745**

5 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	**160**										**100.000**
UNO		**159**						1			**99.3750**
DOS		1	**158**						1		**98.7500**
TRES				**159**			1				**99.3750**
CUATRO					**160**						**100.000**
CINCO						**158**				2	**98.7500**
SEIS							**160**				**100.000**
SIETE							1	**158**			**99.3711**
OCHO		1					1		**158**		**98.7500**
NUEVE		5				4				**150**	**94.3396**
Prom.											**98.871**

6 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	160										100.000
UNO		159			1						99.3750
DOS			159	1							99.3750
TRES				160							100.000
CUATRO					160						100.000
CINCO						159				1	99.3750
SEIS							160				100.000
SIETE							3	156			98.1132
OCHO									160		100.000
NUEVE		7				1	1			150	94.3396
Prom.											99.058

7 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	159		1								99.3750
UNO		159						1			99.3750
DOS			160								100.000
TRES				158		2					98.7500
CUATRO					160						100.000
CINCO						159				1	99.3750
SEIS							160				100.000
SIETE					1	1	2	155			97.4843
OCHO							2		158		98.7500
NUEVE		1								158	99.3710
Prom.											99.248

8 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	160										100.000
UNO		157								2	98.1250
DOS			160								100.000
TRES				160		2					98.7500
CUATRO					160						100.000
CINCO						159				1	99.3750
SEIS							160				100.000
SIETE							2	157			98.7421
OCHO							3		157		98.1250
NUEVE		3								156	98.1132
Prom.											99.248

Filtro Chebyshev a 4.5 kHz
Segmentación KM
Ventanas de 256 muestras
Vectores de valores característicos de orden 1x7
Vectores característicos de orden 7x18
Distorsión de Brown
16 cuantizadores vectoriales por segmento

4 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	155		1						1	3	96.8750
UNO		152						3	1	4	95.0000
DOS	1		157				1		1		98.1250
TRES				156					4		97.5000
CUATRO		1			159						99.3750
CINCO						159				1	99.3750
SEIS							160				100.000
SIETE					1		4	154			96.8553
OCHO				1			7		152		95.0000
NUEVE	2	10				2		5		140	88.0503
Prom.											96.616

5 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	157		1	2							98.1250
UNO	1	154						2		3	96.2500
DOS	1		156	1				1	1		97.5000
TRES				160							100.000
CUATRO		4			156						97.5000
CINCO		1			1	156				2	97.5000
SEIS		1					159				99.3750
SIETE		1					4	154			96.8553
OCHO				1			1		157		98.7421
NUEVE	1	5				1	2	5		145	91.1949
Prom.											97.304

6 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	157		1							2	98.1250
UNO		151						2		6	94.9686
DOS	1		155					1			98.7261
TRES				159					1		99.3750
CUATRO					160						100.000
CINCO						157		1		2	98.1250
SEIS							160				100.000
SIETE	1	2			1		1	154			96.8553
OCHO				1			3		151		97.4194
NUEVE		4				3	2			150	94.3396
Prom.											97.793

7 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	160										100.000
UNO		145						1		11	92.3567
DOS			155								100.000
TRES				158							100.000
CUATRO		2			158						98.7500
CINCO						158	1			1	98.7500
SEIS							159		1		99.3750
SIETE		2		2			2	152		1	95.5975
OCHO	2						3		146		96.6887
NUEVE		5								154	96.8553
Prom.											97.837

8 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	159				1						99.3750
UNO		145						1			99.3151
DOS			138								100.000
TRES			2	147			1				98.0000
CUATRO		1			158						99.3711
CINCO						159				1	99.3750
SEIS							160				100.000
SIETE		1					2	155		1	97.4842
OCHO									139		100.000
NUEVE		7				1	1			150	94.3396
Prom.											97.726

Filtro Chebyshev a 4.5 kHz
Segmentación MLR
Ventanas de 256 muestras
Vectores de valores característicos de orden 1x7
Vectores característicos de orden 7x18
Distorsión de Brown
16 cuantizadores vectoriales por segmento

4 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	136	2	3		2			5		2	90.6667
UNO	2	139	1			2		4		6	90.2597
DOS	3		141	1			1	2		1	94.6309
TRES	1		1	133			11	1	3		88.6667
CUATRO		2			141	1	3	1			95.2702
CINCO					6	145	2	1		2	92.9487
SEIS						1	159				99.3750
SIETE					1		9	145		2	92.3567
OCHO							10		138		93.2432
NUEVE	1	22	1			4	7	5	5	113	71.5189
Prom.											90.894

5 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	144		1		1					2	97.2973
UNO	1	126			1			1		5	94.0299
DOS	1		103					1			98.0952
TRES	1			120			4		3	1	93.0233
CUATRO					140						100.000
CINCO		2			4	139	5		1	2	90.8497
SEIS						1	158				99.3711
SIETE		1		2	2	5	7	140			89.1720
OCHO							1		117		99.1525
NUEVE	1	22		1		4	1	2	1	119	78.8079
Prom.											93.979

6 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	143		2							1	97.9452
UNO	2	69			1	1		1		25	69.6970
DOS			66					1			98.5075
TRES	1		1	86			1	1			95.5556
CUATRO					112	1		1			98.2456
CINCO		1				131	1	1		3	95.6204
SEIS							148				100.000
SIETE					2	2	7	122			91.7293
OCHO	1						5		63		91.3043
NUEVE	1	1				3	2	1		122	93.8462
Prom.											93.245

7 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	119							1		2	97.5410
UNO	2	19			2			1		29	35.8491
DOS			29					1			96.6667
TRES			1	45			4	1		2	84.9057
CUATRO					70						100.000
CINCO						107	2			1	97.2727
SEIS							122				100.000
SIETE							11	78		1	86.6667
OCHO							5		24		82.7586
NUEVE						5	3	2		106	91.3793
Prom.											87.304

8 Segmentos

	0	1	2	3	4	5	6	7	8	9	%
CERO	36				5	10	22	8		2	43.3735
UNO	4	1				16	2			3	04.3478
DOS	4						10	1		1	00.0000
TRES				1	1	3	11	1		4	05.8824
CUATRO	1	1			21	11		1		3	55.2632
CINCO					4	51	20			6	62.9629
SEIS	3	1			3	4	69	1		9	76.6667
SIETE	5				1	2	28	5		10	09.8039
OCHO	1					1	5	1	1	1	10.0000
NUEVE	4				2	24	27	2		24	28.9157
Prom.											29.722

CONCLUSIONES

Con base en los resultados obtenidos, concluimos lo siguiente:

- El reconocimiento de palabras aisladas utilizando cualquiera de los tres tipos de parametrización de las señales de voz (LPC, Cepstrum y KLT) genera resultados por arriba del 90%, que para sistemas de reconocimiento de voz se consideran bastante buenos; con un tiempo de cuantización de entre 45 y 60 minutos por dígito y un tiempo de reconocimiento de 4 a 6 segundos por palabra.

- El problema que se presenta en las últimas matrices de confusión utilizando KLT, fue debido a que el programa que realiza la segmentación MLR cuando se necesitaron 7 y 8 segmentos lo hizo solo para algunas palabras en forma correcta, resultando que para realizar la cuantización no se contaba con el número suficiente de matrices; entonces el numero de cuantizadores no fue siempre el mismo y los rangos de variación de los parámetros de cada palabra era muy variado.

- El porcentaje de reconocimiento se incrementa muy poco si se utiliza un número mayor de segmentos, es decir, que podríamos incrementar el número de segmentos en los cuales dividimos a las señales de voz, y los resultados no variarían mucho.

- Sin embargo, el empleo de una medida de distancia o de distorsión conveniente para cada tipo de parametrización, mejora el porcentaje de reconocimiento, como se comprueba en los resultados de coeficientes LPC al utilizar la distancia Euclidiana y la distorsión de Itakura-Saito modificada.

- Finalmente, los resultados dependen en gran medida del poder identificar el inicio y fin de la palabra con la mayor exactitud posible, ya que podemos utilizar segmentación lineal de las palabras (junto con las desventajas que presenta), y aun así obtener buenos resultados.

BIBLIOGRAFÍA

Proakis, J. and Manolakis, D., **Digital Signal Processign, Principles, Algorithms and Applications,** Macmillan Publishing, 1992 USA

Rabiner, Lawrence R. and Schafer, R.W. **Digital Processing of Speech Signals** Prentice-Hall, 1978 USA

Wheddon, C. and Linggard, R., **Speech and Language Processing** Chapman and Hall, 1990 USA

Saito, Shuzo and Nakato, Kazuo, **Fundamentals of Speech Signal Processing** Academic Press Inc. USA

Abut, H., **Vector Quantization,** IEEE, 1990 USA

A. Herrera, V. R. Algazi and D. Irvine, **An Acoustic Approach for Isolated Speech Recognition,** Proceedings of the International Conference on Signal Processing Applications and Technology, ICSPAT 94, Vol. 2

A. Herrera, V. R. Algazi, V. Brown and D. Irvine, **Subword Segmentation Alternatives for Isolated and Connected Words Recognition,** Proceeding VII European Signal Processing Conference, EUPSICO, 94

Sánchez C., Oscar, **Segmentación Acústica de Subpalabras,** Tesis, Maestría en Ingeniería, DEPFI UNAM, 1997

Mondragón T., Antonio F., **Técnicas de Reconocimiento Automático de Voz Utilizando Segmentación Acústica,** Tesis, Maestría en Ingeniería, DEPFI UNAM, 1996

Martínez G.,Mauricio, **Métodos de Reconocimiento de Palabras Aisladas Usando Segmentación Acústica y Cuantización Vectorial,** Tesis, Maestría en Ingeniería, DEPFI UNAM, 1997