1
291.

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

RECONOCIMIENTO AUTOMÁTICO DE VOZ

T E S I S
QUE PARA OBTENER EL TÍTULO DE
INGENIERO EN TELECOMUNICACIONES
P R E S E N T A:

ALEXANDRE FREDERIC BOUCHET LOPEZ

DIRECTOR DE TESIS M. C. ABEL HERRERA C.

CIUDAD UNIVERSITARIA 1997

Dedico esta tesis a mis padres, quienes me han dado más de lo que he querido creer durante muchos años.

Agradecimientos

Agradezco a Sonia por tantas cosas...

Al M. en Ingeniería Abel Herrera Camacho por haberme iniciado al fascinante mundo de la investigación.

A mi familia por haber soportado días de humor variable y por su apoyo constante a lo largo de la elaboración de este trabajo escrito.

A la familia Cheng por su cariño y apoyo constante.

A Ricardo Ibarra con quien descubrí tantas veces el hilo negro.

A Valerie, Guarina, F.J. Chavo, Federico ("porque sino se enoja") y a todos los que me apoyaron de alguna u otra forma en la elaboración de esta tesis.

A todos los compañeros de la Facultad, a la familia Talavera, y en especial a Javier, Fabricio y Alfredo por su amistad y su apoyo constante, discreto e incondicional.

Indice

INTRODUCCIÓN		1
CAPITULO I.	ELEMENTOS DE ACÚSTICA, DIGITALIZACIÓN Y CARACTERÍSTICAS DE LA VOZ	3
I.1.	Elementos de acústica	3
I.2.	Elementos de digitalización	5
I.3.	Elementos de acústica vocal y características de la voz	7
I.3.1.	El modelo fuente-filtro	7
I.3.2.	Sonidos sonoros	9
I.3.3.	Sonidos sordos	10
I.3.4.	Energía de una señal de voz	11
I.3.5	Tasa de cruces por cero de una señal de voz	14
CAPITULO II	EL MODELO LPC	16
II.1.	La hipótesis autorregresiva	16
II.2.	Obtención de los coeficientes LPC	18
II.3.	El método recursivo de Levinson-Durbin	21
II.4	La distancia de Itakura	22
II.5	Importancia del preénfasis	22
CAPITULO III	[illegible]	24
III.1.	Palabras aisladas	25
III.1.1.	[illegible]	25
III.1.2.	[illegible]	27
III.2.	[illegible]	30

III.2. Palabras conectadas 33
III.2.1. Concatenación de patrones de referencia 34
III.2.2. Algoritmo de una sola etapa para reconocimiento de palabras conectadas 34

CAPITULO IV. DESCRIPCION DE LOS EXPERIMENTOS Y RESULTADOS 37

IV.1. Generalidades acerca de los experimentos 37
IV.2. Experimento I. Verificación experimental de la validez del modelo fuente-filtro LPC 37
IV.3. Experimento II. Reconocimiento de palabras aisladas 42
IV.4. Experimento III. Reconocimiento de palabras conectadas 47

CAPITULO V. ANALISIS Y DISCUSION DE LOS RESULTADOS 48

V.1 Experimento I 48
V.2 Experimento II 49
V.3. Experimento III 50

CONCLUSIONES 53

BIBLIOGRAFIA 55

Introducción

La meta de la investigación en el reconocimiento automático de voz es desarrollar técnicas y sistemas que permitan que la voz sea aceptada por las computadoras como señal de entrada. Este problema ha sido estudiado ampliamente desde años atrás, y sin embargo existe hoy, a nivel mundial, un número sumamente limitado de sistemas comerciales que resuelven este problema de forma satisfactoria. En todos los casos, estos sistemas funcionan bajo una serie de restricciones que permiten limitar la dificultad del problema. La experiencia de años de investigación en esta área ha logrado definir los siguientes factores determinantes en la dificultad que involucra este reconocimiento:

- Grado de conexión de las palabras
- Tamaño del vocabulario de trabajo
- Restricciones impuestas por la tarea específica y el lenguaje
- Dependencia e independencia del parlante
- Grado de ambigüedad acústica
- Ruido presente en el medio ambiente

El grado de conexión de las palabras determina si se desea reconocer palabras aisladas o cadenas de palabras pronunciadas en forma extensiva. En este último caso, que llamaremos palabras conectadas, la dificultad es mayor ya que no se tiene ningún conocimiento previo acerca de la ubicación de cada palabra dentro de la grabación. Además, en palabras conectadas un fenómeno de variabilidad de pronunciación aun mayores que en el caso de palabras aisladas. La dificultad crece al aumentar las variabilidades de tiempos, destacando la presencia en el vocabulario de pares que maximizan las probabilidades de confusión entre palabras, probabilidades que aumentan en el contexto de cadena. El grado de ambigüedad acústica, o variabilidad acústica entre palabras pertenecientes al vocabulario de trabajo, es una medida importante para la tarea específica del usuario. Los algoritmos [illegible] tamaño del vocabulario, [illegible] [illegible]

cualquier instante. Los sistemas dependientes de un solo parlante funcionan en general mejor que los sistemas independientes de la identidad del parlante. Esto se debe a que las características espectrales de la voz cambian de forma muy marcada de parlante a parlante. El nivel de ruido puede afectar gravemente a sistemas de reconocimiento ya que sus características se mezclan con las características de interés, distorsionando así la información útil para el reconocimiento.

El ajuste dinámico del tiempo es una técnica de comparación no lineal que ha demostrado ser indispensable cuando las palabras son representadas como secuencias de parámetros extraídos de la misma. En el caso de reconocimiento de voz, la representación de las señales por secuencias de coeficientes LPC ha comprobado ser una valiosa herramienta de compresión y extracción de información.

Los objetivos de este trabajo fueron los siguientes:

- Realizar una investigación exhaustiva y un estudio completo del modelo digital fuente-filtro de la voz basado en coeficientes LPC.

- Realizar una investigación exhaustiva y un estudio completo de las diferentes variantes de ajuste dinámico del tiempo con sus aplicaciones a palabras aisladas y palabras conectadas.

- Determinar para ambos casos cuál es el método adecuado y determinar experimentalmente sus parámetros óptimos.

El presente capítulo describe brevemente los sistemas numéricos, los conceptos básicos de digitalización de señales analógicas, el teorema de Nyquist, la cuantización y la codificación PCM. Los sistemas numéricos de ahora en adelante, mencionados como el cuarto capítulo describe la tecnología empleada en dos experimentos. El primero empleando exclusivamente una determinada categoría de sus características, explicadas de manera detallada y discutida su importancia.

Capitulo I

Elementos de Acústica, Digitalización y Características de la Voz

I.1. Elementos de Acústica

El sonido proviene de alguna perturbación en un medio elástico, generalmente el aire, que causa una alteración de presión y un desplazamiento de sus partículas. La sensación auditiva se genera en un intervalo de frecuencias de 20 Hz a 20 Khz aproximadamente. Mediciones efectuadas indican que las máximas variaciones toleradas por el oído humano son del orden de 280 dinas/cm^2 mientras que variaciones menores de $3*10^{-4}$ dinas/cm^2 no producen sensación auditiva. La propagación del sonido es esencialmente longitudinal, es decir que la dirección del movimiento de las partículas es la misma que la de la onda. La velocidad de propagación del sonido en el aire es de alrededor de 340 m/s, valor que varía según la presión atmosférica y la temperatura ambiente.

La intensidad de una onda sonora se define como la potencia media transportada por unidad de superficie y está expresada a

$$I = \frac{P^2}{2\rho v} \qquad (1.1)$$

donde P es la amplitud de variación de presión y v la velocidad de propagación. En general se emplea una escala logarítmica para expresar la intensidad a través del nivel de intensidad β, con

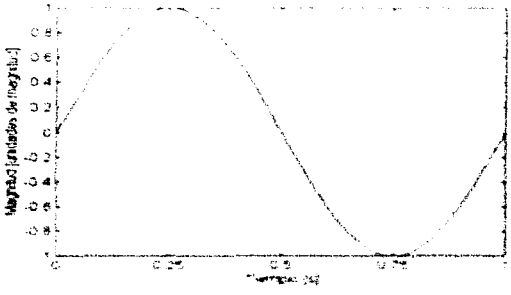
$$\beta = 10 \log_{10} \frac{I}{I_{ref}} \qquad (1.2)$$

donde I_0 es el nivel de referencia empleado, que en general es de 10^{-16} a 10^{-12} W/cm^2, el rango del umbral mínimo audible. La siguiente tabla indica algunos niveles de intensidad representativos calculados con una referencia $I_0 = 10^{-12}$.

Decibeles	Fuente Acústica
200	Transbordador espacial despegando
190	
180	
170	Avión de propulsión a chorro con post-combustión
160	
150	Avión de hélices
140	
130	Concierto amplificado de Rock
120	Orquesta de 75 instrumentos (pico máximo)
110	Piano (pico máximo)
100	Autopista en una autopista
90	Voz gritando
80	
70	Voz de conversación
60	
50	
40	
30	Susurro
20	
10	
0	

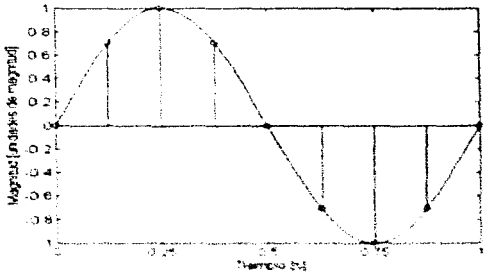
1.2. Elementos de Digitalización

El sonido, como muchas señales naturales, es una señal continua en el tiempo y en amplitud. Esto quiere decir que la señal puede tomar cualquier valor dentro de un intervalo continuo (que contiene una infinidad de valores posibles) y que puede medirse para cualquier tiempo dentro de un intervalo de tiempo continuo (que contiene una infinidad de valores posibles), como lo ejemplifica la siguiente figura.

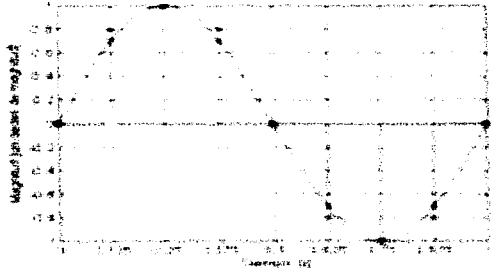


Para poder procesar esta señal en una computadora digital, es necesario limitar el número de valores que puede tomar tanto la amplitud como el tiempo ya que este cuenta número finito de bits únicamente se puede representar cierto número finito de valores. Se realiza entonces el proceso de digitalización que consiste dos pasos: el de muestreo en el tiempo y el de cuantización en la amplitud. Un solo periodo de cualquier señal representada mediante PCM (por Pulse Code Modulation) que es la representación digital más antigua y fácil de entender y otra más compleja de este representación), más sencilla de su análisis, por lo que será utilizada a continuación a pesar de que en la práctica no es el caso de ningún sistema actual de voz.

El primer paso consiste en tomar una muestra del valor de la señal cada cierto tiempo, como lo ejemplifica la siguiente figura.



El segundo paso consiste en aproximar el valor de la amplitud de la señal en cada muestra con el valor más cercano o aproximadamente diferente de un conjunto finito de valores. Este proceso se conoce como cuantificación.



Es común que ocurran ambos pasos de forma simultánea en sistemas reales. En este ejemplo se tomó una muestra cada 0.125s y el número de niveles de cuantización es 11. El número de valores posibles, niveles de cuantización, queda determinado por el número de bits empleados para cada muestra mientras que el numero de muestras tomadas por unidad de tiempo depende de la frecuencia de muestreo. La relacion entre el numero de niveles existentes (N) y el numero de bits empleados para representarlos (n) es

$$N = 2^n \qquad (1.4)$$

y la frecuencia de muestreo (f_m) debe cumplir con el teorema de Nyquist

$$f_m \geq 2f_{max} \qquad (1.5)$$

donde f_{max} es la frecuencia maxima que contiene la señal por muestrear.

1.3. Elementos de Acustica Vocal y Caracteristicas Importantes de la Voz

1.3.1. El modelo fuente filtro

El tracto vocal es un tubo acústico no uniforme que se extiende desde la laringe a los labios. Otro tubo, el tracto nasal, puede ser acoplado o desacoplado de este mediante movimiento del velo. A grandes rasgos podemos decir que el tracto vocal generado por la laringe [illegible] es modificado por [illegible] diferentes.

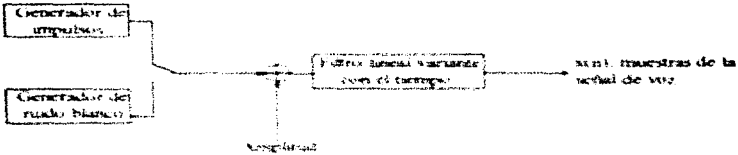
Los sonidos producidos en [illegible] vocales [illegible] pueden clasificarse en sordos y con sonoridad en el [illegible] que sea necesaria para la vibración de las cuerdas vocales y sonidos [illegible] en [illegible].

-Los sonidos fricativos se producen al formar una constriccion en algún lugar del tracto, creando así turbulencias que producen una fuente de sonido de banda ancha que excita el tracto vocal. Los sonidos fricativos pueden producirse con o sin fonación.

-Los susurros se generan en la laringe al producirse una turbulencia por el paso del aire por una pequeña apertura triangular entre los cartílagos cricoide, tiroide y aritenoide.

Todas estas fuentes crean una excitación de banda relativamente ancha del tracto vocal que a su vez se comporta como un filtro lineal, variante en el tiempo, que impone sus propiedades de transmisión sobre el espectro de la fuente. Se ha observado que la variación de este filtro es lenta y que por lo tanto se puede considerar invariante en periodos de alrededor de 10 ms.

Porque la fuente de sonido y la forma del tracto son relativamente independientes, es razonable modelarlas en forma separada, como en la siguiente figura.

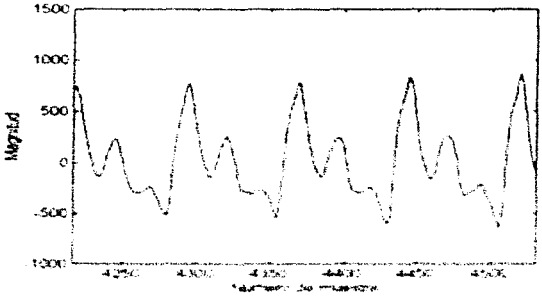


En este modelo digital, las muestras de la señal de voz son la salida de un filtro digital variante en el tiempo, que aproxima las propiedades del tracto vocal y cuya excitación es [illegible]

realidad representa las propiedades de radiación, del tracto vocal y de los pulsos de fonación.

I.3.2. Sonidos sonoros

La característica principal de los sonidos sonoros es que son cuasiperiódicos como se ve claramente en la siguiente figura:



La vibración de las cuerdas vocales con el paso del aire generan una excitación cuasiperiódica del tracto vocal que modifica el espectro de dicha excitación según su respuesta en frecuencia. Observamos sonidos cuasiperiódicos puesto que la misma frecuencia fundamental y sus armónicos son ligeramente diferentes. Otra característica de estos sonidos es que su potencial energético que los acompaña suele ser elevado. Las formantes determinan el timbre del sonido y dentro de periodo existe una variabilidad dependiendo de la respuesta en frecuencia del tracto vocal en particular que corresponda a esa determinada fonación. Estos sonidos corresponden a todas las vocales y consonantes sonoras del español. Ejemplos claros a

continuación presentamos el espectrograma (aproximación gráfica del contenido espectral instantáneo de una señal) de las vocales /a/ /e/ /i/ /o/ /u/.

[illegible]

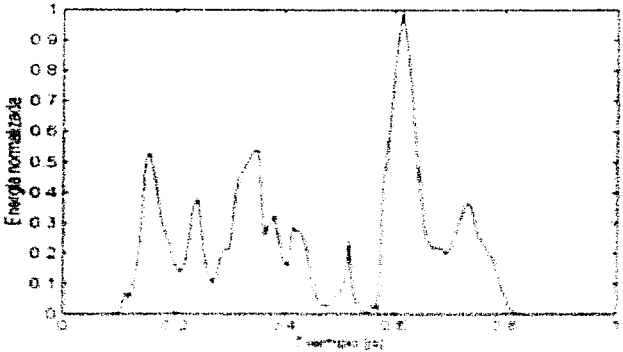
Una de las características más importantes de una señal es su energía. En el caso de una señal discreta real, la energía se define como:

$$E = \sum_{n} x^2[n] \qquad (3.8)$$

Para señales con duración infinita, la suma será infinita. Consecuentemente, la señal tendrá energía finita sólo si la energía es acotada.

$$E < \infty \Leftrightarrow \sum_{n} x^2[n] < \infty \qquad (3.9)$$

donde w(m) es una ventana que selecciona un segmento de x(n) y N es el número de muestras de la ventana. Si se desea dar el mismo peso a todo el intervalo analizado, se emplea una ventana rectangular, mientras que si se desea dar mayor peso a la sección central del intervalo se puede emplear una ventana de pesos variables, como una ventana de Hamming. La selección de la longitud de la ventana es muy importante. Si se selecciona una ventana menor al periodo (en el caso de voz sonora), E(n) fluctuará muy rápido dependiendo de los detalles precisos de la forma de onda; en cambio si se selecciona una ventana de longitud mayor a varios periodos, E(n) tendrá muy poca variación y no reflejara las propiedades variantes de la señal de voz. Una longitud de 10-20 ms ofrece buenos resultados para señales de voz. La siguiente figura muestra el perfil de energía normalizada para la palabra normalizada

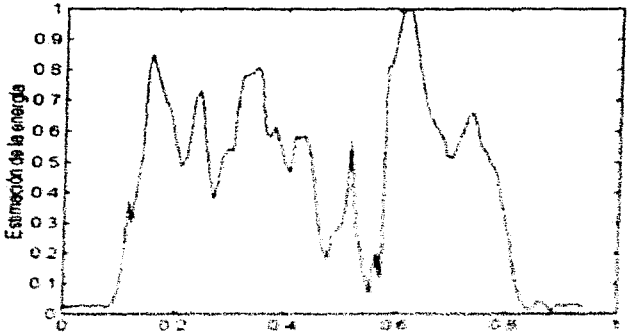


Una dificultad que existe con la medición de energía directa es [illegible] que es extremadamente sensible a niveles altos de señal porque esto implica [illegible]

Se puede evitar este problema de una simple forma, usando como estimación de la energía la función

$$E(n)=\sum_{m=0}^{N-1}|w(m)x(n-m)|. \qquad (1.5)$$

La siguiente figure muestra esta estimación normalizada de la energía de la misma palabra /normalizada/.

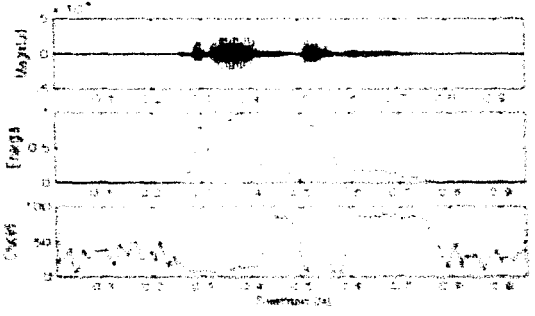


I.3.5. Tasa de cruces por cero de una señal de voz

Otra caracteristica interesante de una señal de voz es su tasa de cruces por cero. En una señal digital, un cruce por cero ocurre entre dos instantes de muestreo n y n-1 si

$$\text{signo}[x(n)] \neq \text{signo}[x(n-1)] \qquad (1.6)$$

Esta caracteristica permite por ejemplo estimar las posiciones de sonidos sonoros y sonidos sordos. En efecto, la energia de los sonidos sonoros esta concentrada debajo de 3kHz mientras que la energia de los sonidos sordos lo esta arriba de 3 kHz, y por lo tanto la tasa de cruces por cero sera mas alta para sonidos sordos que para sonidos sonoros. Ademas, los sonidos sonoros generalmente tienen mayor energia que los sordos, lo que ayuda tambien en la localizacion de estos tipos de sonidos. La siguiente figura muestra dichas diferencias de tasas y de energia para una grabacion de la palabra "cruces"



Los numerosos cruces por cero observables antes y después de la palabra se deben al alto nivel de ruido en la grabación.

Para que la estimación de la tasa de cruces por cero sea correcta, es necesario que la señal no contenga componentes de frecuencias bajas como ruido de 60Hz o niveles de DC, ya que su efecto es desastroso en esta medición. Realizando una medición de cruces por la media de la ventana en cuestión se puede solucionar fácilmente este inconveniente. En dicho caso, es aconsejable restar primero la media del contenido de la ventana de análisis a las muestras de dicha ventana y posteriormente realizar un análisis de tasa de cruces por cero convencional

Capitulo II
El Modelo LPC

II.1. La hipótesis auto-regresiva

Los metodos mas exitosos hasta la fecha para realizar análisis de la voz son los que se basan en el concepto de prediccion lineal, porque son precisos y requieren de poco cómputo.

La idea basica detras de la codificacion lineal predictiva (LPC por Linear Predictive Coding) es que una muestra de voz puede ser aproximada por una combinacion lineal de las p muestras anteriores, donde p es el orden del modelo. Minimizando el error cuadratico medio entre las muestras reales y las predichas linealmente se pueden determinar los coeficientes del predictor, es decir los pesos de la combinacion lineal. El uso del analisis lineal predictivo fue inspirado por el modelo digital de la voz mencionado.

Suponiendo que las muestras de una señal de voz son producidas por el modelo de la siguiente figura, donde fuente filtro de la voz:

donde el sistema tiene la siguiente función de transferencia.

$$H(z)=\frac{A}{1-\sum_{k=1}^{p} a_k z^{-k}} \qquad (2.1)$$

Para sonidos sonoros el sistema es excitado por un tren de impulsos cuasiperiódicos, y para sonidos sordos es excitado por ruido blanco. El análisis por predicción lineal se basa en la observación que para un tal sistema, las muestras x(n) están relacionadas con la excitación δ(n) por la siguiente ecuacion en diferencias.

$$x(n)=\sum_{k=1}^{p} a_k x(n-k)+G\delta(n) \qquad (2.2)$$

Supongamos que procesamos esta señal con un predictor lineal

$$x(n)=\sum_{k=1}^{p} a_k x(n-k) \qquad (2.3)$$

Entonces el error de predicción se define como

$$e(n)=x(n)-x(n)=x(n)-\sum_{k=1}^{p} a_k x(n-k) \qquad (2.4)$$

Comparando (2.2) y (2.4) vemos que si $a_k = a_k$, y si la señal realmente obedece al modelo de (2.2), entonces $e(n)=G\delta(n)$. Por lo tanto, existe una diferencia de excitación de la voz sonora, el error de predicción debería ser muy pequeño si los coeficientes del predictor a_k son iguales a los parámetros a_k de la función de transferencia del tracto vocal. Para sonidos sordos, el error sería equivalente a la ganancia de la señal. Por lo tanto el procesamiento del predictor

$$P(z)=\sum_{k=1}^{p} a_k z^{-k} \qquad (2.5)$$

es una buena aproximación del denominador de la función de transferencia del tracto vocal.

II.2 Obtención de los coeficientes LPC

Una forma de obtener los coeficientes del predictor se basa en la minimización del error cuadrático medio de predicción dentro de un periodo corto: se buscan los valores α_k que minimicen

$$E = \sum_{n=0}^{N} [x_s(n) - \hat{x}(n)] \qquad (2.6)$$

$$E = \sum_{n=0}^{N} \left[x(n) - \sum_{k=1}^{p} \alpha_k x(n-k) \right]^2 \qquad (2.7)$$

donde $x_s(n)$ es un segmento de voz seleccionado en la vecindad de la muestra l,

$$x_s(n) = x(n+l) \qquad (2.8)$$

y al resolver $\partial E / \partial \alpha_i = 0$ (igualamos a cero las derivadas parciales para que E sea mínimo) para $i = 1, 2, \ldots, p$ obtenemos las ecuaciones

$$\sum_{n=0}^{N} x(n-i)x(n) = \sum_{k=1}^{p} \alpha_k \sum_{n=0}^{N} x(n-i)x(n-k) \qquad \text{para } i = 1, 2, \ldots, p \qquad (2.9)$$

Sin derechos reservados

$$\varphi_i(i,k)=\sum_{n=0}^{N-1} x_i(n-i)x_i(n-k) \qquad (2.10)$$

entonces podemos reescribir (2.9) como

$$\sum_{k=1}^{p}\alpha_k\,\varphi_i(i,k)=\varphi_i(i,0), \qquad i=1,2,\ldots,p. \qquad (2.11)$$

Este sistema de p ecuaciones con p incógnitas puede ser resuelto calculando $\varphi_i(i,k)$ para $1\le i\le p$ y $1\le k\le p$. Sustituyendo variables reescribimos (2.10) como

$$\varphi_i(i,k)=\sum_{n=-i}^{N-1-i} x_i(n)x_i(n+i-k)$$

$$\varphi_i(i,k)=\sum_{n=-k}^{N-1-k} x_i(n)x_i(n+k-i) \qquad (2.12)$$

de donde claramente vemos que $\varphi_i(i,k)=\varphi_i(k,i)$. También observamos de (2.12) que valores de $x_i(n)$ son requeridos fuera del intervalo $0\le n\le N-1$. Para no proveer los valores fuera de este intervalo es necesario emplear alguna ventana west de duración finita para reducir los efectos de límite que surgen al tratar de predecir las p primeras muestras basándose en muestras cero, valor en cero por estar fuera del intervalo $0\le n\le N-1$. Existe para estos casos límite un error de predicción muy grande y es necesario introducir entonces una ventana que reduzca suavemente la señal hasta cero en los extremos del intervalo $0\le n\le N-1$. Obtenemos entonces

$$x_i(n)=\begin{cases} w(n)\,x_i(n), & 0\le n\le N-1 \\ 0, & \text{en otros casos.} \end{cases} \qquad (2.13)$$

Utilizando estas definiciones de $x_i(n)$, (2.12) se convierte en

$$\varphi_t(i,k)=\sum_{n=0}^{N-1-(i-k)} x_t(n)x_t(n+i-k)$$

$$\varphi_t(i,k)=\sum_{n=0}^{N-1-(k-i)} x_t(n)x_t(n+k-i)$$

$$\varphi_t(i,k)= r_t(i-k)= r_t(k-i) \qquad (2.14)$$

En este caso (2.11) se transforma en

$$\sum_{k=1}^{p}\alpha_k r_t(|i-k|)= r_t(i), \qquad i=1,2,\ldots,p. \qquad (2.15)$$

Tomando en cuenta que la definicion de la autocorrelación en tiempo corto es

$$R_t(m)= \frac{1}{N}\sum_{n=-\infty}^{\infty} x_t(n)w(n)x_t(n+m)w(n+m) \qquad (2.16)$$

queda claro que $r_t(i)=NR_t(n)$ Lo anterior implica que para obtener los coeficientes LPC es necesario calcular primero los coeficientes de autocorrelación en tiempo corto de la ventana de interés y luego resolver un sistema de ecuaciones $\sum_{k=1}^{p}\alpha_k r_t(|i-k|)= r_t(i)$, que supone un sistema de p ecuaciones con p incognitas que puede ser expresado de la siguiente forma

$$\Phi \bullet a = \Psi \qquad (2.17)$$

Este sistema puede ser resuelto empleando diversos procedimientos, primero para resolver sistemas de ecuaciones lineales. Sin embargo, es importante notar que la matriz Φ en este caso es una matriz simetrica y todos sus elementos de la diagonal principal son mayores que cero, una matriz de Toeplitz y el sistema puede resolverse por un método mas eficiente. Basandose en estas propiedades, Durbin propuso un procedimiento recursivo para resolver el sistema de

II.3. Método recursivo de Levinson-Durbin

De (2.15) vemos que el sistema por resolver es el siguiente:

$$\begin{bmatrix} r(0) & r(1) & r(2) & \cdots & r(p-1) \\ r(1) & r(0) & r(1) & \cdots & r(p-2) \\ r(2) & r(1) & r(0) & \cdots & r(p-3) \\ \vdots & & & & \\ r(p-1) & r(p-2) & r(p-3) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} r(1) \\ r(2) \\ r(3) \\ \vdots \\ r(p) \end{bmatrix} \qquad (2.18)$$

Los pasos del método recursivo de Levinson-Durbin son los siguientes:

$$\begin{aligned} & E^{(0)} = r(0) \\ & k_i = \frac{\left\{ r(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} r(i-j) \right\}}{E^{(i-1)}} \qquad 1 \le i \le p \\ & \alpha_i^{(i)} = k_i \\ & \alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \qquad 1 \le j \le i-1 \\ & E^{(i)} = (1 - k_i^2) E^{(i-1)} \end{aligned} \qquad (2.19)$$

Los coeficientes $\{\alpha_j^{(p)}\}$ son los valores de salida después de realizar los pasos descritos para todos los valores válidos de i y j. Estos valores definen totalmente las características del filtro digital que representa al tracto vocal en el modelo fuente-filtro descrito. Estos parámetros corresponden con exactitud a los de la red de dispersión de Itakura. Una señal corregida de voz pasada por la red, nos representaría exactamente los mismos parámetros formando así una pareja de coeficientes LPC.

Una característica interesante del modelado por análisis LPC es que una medida de distancia ,o de disimilitud, entre dos modelos LPC ha sido desarrollada por Fumitada Itakura. Esta medida ha demostrado ser extremadamente eficiente.

II.4. La distancia de Itakura

Supongamos que tenemos dos señales T y R cuyos coeficientes a_T y a_R LPC han sido determinados. V_T es la matriz de autocorrelaciones de T. La distancia entre los modelos LPC definidos para T y R es la siguiente

$$d(T,R) = \log\left[\frac{a_R V_T a_R}{a_T V_T a_T}\right] \qquad (2.20)$$

Esta distancia relaciona el error cuadrático medio de predicción de la señal T por el modelo a_T con el error cuadrático medio de predicción de la señal T por el modelo a_R. Si la señal T, cuyo modelo LPC es a_T, corresponde al modelo a_R de la señal R, entonces ambos errores de predicción serán similares, su cociente será cercano a 1 y el logaritmo del cociente será cercano a cero. Por otra parte si T no se asemeja a R, el error de predicción de T mediante el modelo de R será mayor al error de predicción de T mediante su propio modelo y el cociente será mayor que 1 dando un logaritmo mayor que cero.

II.5. Importancia del preénfasis

La importancia matemática antes que nada, para el cálculo de nuestros coeficientes, el ingeniero hace una tendencia general de 6dB por octava considerando nuevamente la frecuencia. Dado que en este caso la información de frecuencia más importante se encuentra en las diferentes bandas formantes, no necesitamos. Debido a esta razón nos preocupamos para que la información que obtengamos utilice [illegible] más [illegible]. Esto [illegible] proceso efectuamos con un filtro para [illegible] de [illegible] [illegible] [illegible] de la [illegible] de [illegible]

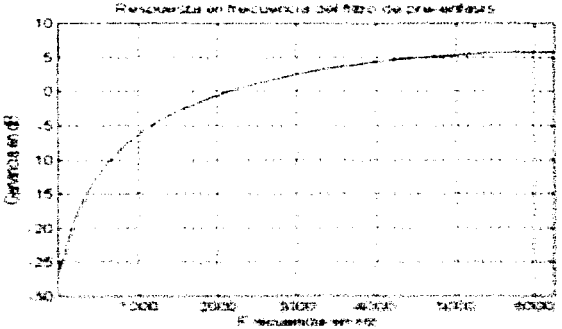
no es crítico ya que se busca principalmente realzar las frecuencias altas. Un tal filtro puede ser implementado digitalmente empleando la ecuación en diferencias

$$y(n)=x(n)-ax(n-1) \tag{2.21}$$

donde y(n) es la salida presente del filtro, x(n) su entrada presente y x(n-1) su entrada anterior. El valor de a debe ser tal que $0.95 \le a \le 1$. Otra vez, el valor de a no es muy crítico. Efectuando una transformación discreta Z obtenemos

$$Y(z)=X(z)-az^{-1}X(z)=(1-az^{-1})X(z) \tag{2.22}$$

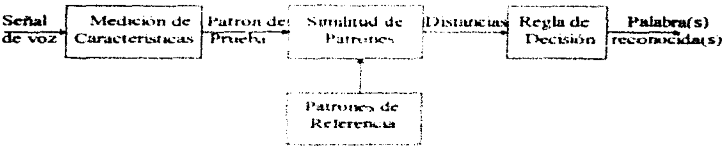
A continuación presentamos la respuesta en frecuencia de dicho filtro cuando a=0.95 y la frecuencia de muestreo es de 12.5 kHz



Capitulo III

El ajuste dinámico en el tiempo.

La siguiente figura muestra el modelo de reconocimiento de patrones **utilizado en la mayoría de los sistemas** de reconocimiento de palabras.



Este modelo consta de tres pasos

- Medición de características.
- Determinación de similitud con los diferentes patrones de referencia.
- Decisión de la palabra reconocida, basada en alguna regla.

El primer paso es básicamente alguna técnica de reducción de datos. Las características medidas pueden ser de muchos tipos diferentes, como mediciones de energía, tasas de cruces por cero [illegible]

representación por LPC la señal quedará representada por una secuencia de grupos de coeficientes (patrón de coeficientes) donde cada grupo corresponde a un periodo corto (alrededor de 10 ms). El número de grupos de coeficientes que representen a cierta señal dependerá por lo tanto de la longitud de dicha señal. La regla empleada para determinar a qué palabra corresponde la señal analizada requiere de las distancias obtenidas entre la señal de prueba y todos los patrones de referencia. Se escoge la referencia cuya distancia con la señal de prueba es menor, o se escoge la clase (porque varias referencias pueden representar una misma palabra, y por lo tanto pertenecer a la misma clase) cuya presencia sea mayor dentro de los n primeros candidatos. Por ejemplo si n = 5 y el mejor candidato es cierta palabra a y los segundo y terceros candidatos son otra palabra b, se reconoce la señal analizada como b.

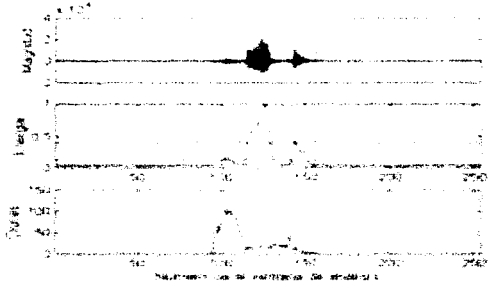
Nos interesaremos primero al caso mas simple donde se desea comparar una señal de prueba con una sola señal o patrón de referencia a la vez, caso general en el reconocimiento de palabras aisladas.

III.1. Palabras aisladas

III.1.1. Detección de inicio y fin de palabra

En un sistema de reconocimiento de palabras aisladas el primer paso consiste en el de detectar la palabra que se quiere reconocer dentro de el espacio de grabación. Esto permite reducir el tiempo de procesamiento ya que sin silencios correspondientes al inicio y al fin de grabación serán descartados. Por otra parte en la mayoría de los sistemas es necesario realizar dicha detección para poder asegurar que realmente se esta comparando una palabra con otra, y no una palabra y silencio o silencio con otro. Existen diversos métodos para detectar el inicio y fin de palabra. El método mas comúnmente utilizado por la mayoría de los sistemas es el mas preciso y versátil.

Este método de detección de inicio y fin de palabra se basa en dos mediciones en el dominio del tiempo: la energía de la señal y su tasa de cruces por cero. Se efectúa un análisis de estos dos parámetros por ventanas de alrededor de 10ms obteniendo así dos patrones de variación. Es necesario para este algoritmo que los primeros 100ms no contengan voz para que se pueda obtener así una caracterización estadística del ruido de fondo. Usando esta caracterización y el nivel máximo de energía se calculan umbrales de tasa de cruces por cero y de energía. Después se analiza el patrón de energía para determinar el intervalo en el que la señal siempre excede un umbral de energía muy conservador (ITU). Se hace entonces la suposición que los puntos de inicio y fin se encuentran fuera de este intervalo y se busca en qué puntos la energía de la señal cae debajo de otro umbral, menor a ITU, que llamaremos ITL. Estos puntos provisorios de inicio y fin se denominan N1 y N2. El último paso consiste en analizar las 25 ventanas (250ms) anteriores y posteriores a los puntos N1 y N2 respectivamente para determinar cuantas veces el umbral de cruces por cero, IZ, es superado en ese intervalo. Si el umbral es superado 3 o más veces, los puntos de inicio y/o fin se desplazan al primer/último punto dentro de estos 250ms donde el umbral fue superado. La siguiente figura ejemplifica el método para la palabra 'siete', que representa uno de los casos más difíciles puesto que contiene un silencio dentro de la palabra.



III.1.2. Definición de distancia entre los patrones LPC de dos palabras

Una vez efectuado el análisis de inicio y fin de palabra, se procede a un análisis LPC determinando así para cada palabra de referencia el patrón de coeficientes LPC correspondiente. Cada vez que se desea reconocer alguna palabra que llamaremos palabra prueba, se debe detectar su inicio y su fin y luego calcular su patrón de coeficientes LPC.

Una vez obtenidos los patrones de características, el siguiente paso es el de determinar la similitud o disimilitud entre patrones de prueba y patrones de referencia. Porque las variaciones de velocidades del habla afectan no solamente la longitud de las palabras, sino también la localización interna de eventos para una misma palabra, la determinación de similitud involucra no solamente el cálculo de distancias, sino también algún tipo de ajuste dinámico de los ejes temporales. Por ajuste dinámico de los ejes temporales entendemos o nos referimos a algún mecanismo que pueda alinear las características de las palabras para que puedan así ser comparadas características correspondientes.

Supongamos que tenemos un patrón de prueba $T=\{T(1),T(2),\ldots T(N)\}$ y un patrón de referencia $R=\{R(1),R(2),\ldots R(M)\}$. El problema de ajuste dinámico del tiempo es de comparar las $R(i)$'s adecuadas con las $T(i)$ adecuadas de tal forma que se minimice la distancia total entre los patrones T y R. Sea C la función que relaciona a los elementos de T con los elementos de R

$$C=\{c(1),c(2),\ldots c(K)\} \qquad (3.1)$$

donde cada elemento de C es un par de apuntadores que indican los elementos de R y T por comparar se

$$c(k)=\{r(k),t(k)\} \qquad (3.2)$$

El problema es entonces el de encontrar la función C que minimice la distancia entre T y R. Para cada c(k) tenemos una función de distancia o costo d[c(k)]. La distancia entre dos elementos en comparacion es, en el caso de análisis LPC, la distancia de Itakura entre los dos modelos LPC. (En este caso los elementos de R y T son vectores de coeficientes LPC). La función de ajuste dinamico del tiempo debe entonces minimizar el costo o la distancia total:

$$D(C)=\sum_{k} d[c(k)] \qquad (3.3)$$

Existen ciertos requerimientos impuestos por las caracteristicas de las señales de voz sobre la funcion c(k). Estos son:

- La funcion debe ser monotona creciente:

$$\begin{aligned} i(k) &\geq i(k-1) \\ j(k) &\geq j(k-1) \end{aligned} \qquad (3.4)$$

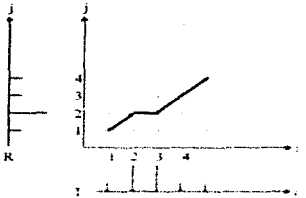
- La funcion debe cumplir las siguientes condiciones en los limites:

$$\begin{aligned} i(1) &= i(k) \\ i(K) &= NT \\ j(k) &= NR \end{aligned} \qquad (3.5)$$

- La funcion debe considerar todos los elementos de los patrones involucrados:

$$\begin{aligned} i(k) &= i(k-1)+1 \\ j(k) &= j(k-1)+1 \end{aligned} \qquad (3.6)$$

Encontrar la función de ajuste temporal es equivalente a encontrar el camino de mínimo costo a través de una rejilla de puntos como lo indica la siguiente figura.



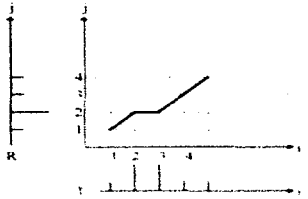
Los elementos 2 y 3 de T se asocian y comparan con el elemento 2 de R ya que representan una misma característica (magnitud grande). En el caso de análisis LPC, la medida de similitud no es la magnitud, sino la distancia de Itakura entre los coeficientes LPC involucrados.

A primera vista puede parecer que DWT tenga que ser evaluada para un número demasiado grande de posibles trayectorias de caminos posibles. Sin embargo, la programación dinámica controla este problema al tomar en cuenta que el mejor camino desde el punto (1,1) hasta cualquier otro punto es independientemente de lo que ocurra después de ese punto. Por lo tanto, el costo total del camino que larga hasta el punto (i,j) es el costo de ese punto más el costo del camino hasta su predecesor apropiado:

$$[DWT]_{ij} = d(i,j) + \min[DWT]_{i-1,j-1}, [DWT]_{i-1,j}, [DWT]_{i,j-1} \quad (3.9)$$

Los arcos horizontales y verticales tienen una ponderación que corresponde con todas las eventualidades mencionadas.

Encontrar la función de ajuste temporal es equivalente a encontrar el camino de mínimo costo a través de una rejilla de puntos como lo indica la siguiente figura.



Los elementos 2 y 3 de I se asocian y comparan con el elemento 2 de R ya que representan una misma característica (magnitud grande). En el caso de análisis LPC, la medida de similitud no es la magnitud, sino la distancia de Itakura entre los coeficientes LPC involucrados.

A primera vista puede parecer que [illegible] tenga que ser evaluada para un número demasiado grande o poco práctico de caminos posibles. Sin embargo, la programación dinámica controla este problema al tomar en cuenta que el mejor camino desde el punto (1,1) hasta cualquier otro punto es independiente de lo que ocurra después de ese punto. Por lo tanto, el costo total del camino que llega hasta el punto (ak, jak) es el costo de ese punto más el costo del camino hasta su predecesor óptimo.

[illegible] (3.7)

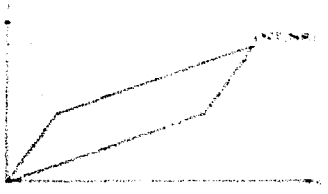
Los valores [illegible] pueden ser especificados mediante [illegible] que restringen [illegible] de manera arbitraria.

La programación dinámica funciona por etapas. En su aplicación al ajuste dinámico del tiempo, cada etapa corresponde a una columna del enrejado. Para cada columna, empezando por i=1, se calculan las distancias óptimas acumuladas para los puntos de dicha columna. Dado que el cálculo de la siguiente columna está basado en las distancias óptimas para la columna anterior, podemos asegurar que a su vez, las distancias obtenidas para esta nueva columna son las óptimas.

A continuación presentamos un resumen de una serie de restricciones adicionales en la trayectoria de C con las que diferentes investigadores han trabajado. Algunas permiten no solamente mejorar el éxito de reconocimiento, sino también disminuir el número de cálculos necesarios.

III.1.3. Restricciones comunmente empleadas en la función de ajuste dinámico del tiempo.

- I. Rabiner propone definir la siguiente región de validez o de libertad en forma de paralelogramo cuyos bordes sean rectas con pendientes de 1/2 y 2

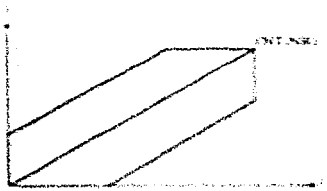


La función de ajuste dinámico debe necesariamente encontrarse dentro esta región. Esta restricción impide que la pendiente máxima y mínima de cualquier segmento, pase

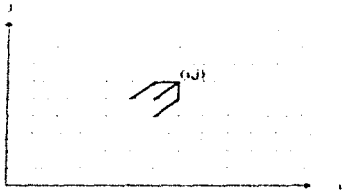
cualquier par de palabras es 2 y que la mínima es ½. Esto implica a su vez que patrones cuya relación de distancias sea menor a ½ o mayor a 2 no podrán ser comparados. Esta restricción es válida porque las variaciones de velocidad del habla son pequeñas, y por lo tanto grabaciones de longitudes considerablemente distintas muy probablemente correspondan a palabras distintas. El número de puntos calculados disminuye ya que únicamente se calculan los puntos internos a la región definida. Los puntos predecesores válidos para algún punto (i,j) son (i-1,j-1), (i-1,j), (i,j-1).

- El Itakura propone una función de ajuste que depende únicamente del índice i. Esto trae como consecuencia que transiciones verticales de la función de ajuste no son permitidas. Por lo tanto los predecesores válidos para un punto (i,j) son (i-1,j-1) y (i-1,j). En este caso se emplea también la región de validez descrita por L. Rabiner.

- El estudio más completo realizado sobre restricciones útiles en el ajuste dinámico del tiempo fue realizado por H. Sakoe. Primero define una región de validez, o de libertad, limitada por dos rectas paralelas a la recta de ajuste lineal. La región de validez queda por lo tanto definida por un parámetro r que indica el número de ventanas válidas arriba y abajo de la recta de ajuste lineal.



En su estudio determinó de manera experimental que la mejor restricción que hay que imponer sobre la pendiente local de la función de ajuste es que no sea mayor a 2 ni menor a ½. Esta restricción difiere de las anteriores porque se refiere a la pendiente local y no a la global como en el caso del paralelogramo. La definición de predecesores difiere en este caso. Los caminos válidos que nos lleven hasta un punto (i,j) son:



y de esta forma la distancia mínima acumulada en (i,j) es:

$$g(i,j)=M\,[N\begin{bmatrix} g(i-1,j-2)+2d(i,j-1)+d(i,j) \\ g(i-1,j-1)+2d(i,j) \\ g(i-2,j-1)+2d(i-1,j)+d(i,j) \end{bmatrix}] \qquad (3.8)$$

donde g(a,b) es la distancia mínima acumulada hasta el punto (a,b) y d(a,b) es la distancia local del punto (a,b).

En todos los casos la distancia local entre el patrón T y el patrón R es la distancia acumulada en el punto (NT,NR) que corresponde a la ecuación fuente especificada en (3.5).

III.2. Palabras conectadas

Sorprendentemente, una de las aplicaciones más importantes de las técnicas para reconocimiento de palabras aisladas ha sido en el área de reconocimiento de palabras conectadas. En el reconocimiento de palabras conectadas, la señal de entrada es una secuencia de palabras pertenecientes a cierto vocabulario definido y el reconocimiento se basa en la comparación del patrón resultante de entrada con patrones de palabras aisladas del mismo vocabulario. La mayor dificultad en este tipo de sistemas reside en un fenómeno conocido como coarticulación. En efecto, en el caso de palabras conectadas las características del principio de cada palabra se ven afectadas por las características del final de la palabra anterior. Esto trae una distorsión de las palabras que al ser comparadas con palabras aisladas de referencia acumularían distancias mayores, disminuyendo así la diferencia de distancia entre una comparación entre dos señales equivalentes, que representan la misma palabra, y dos señales no equivalentes. Ejemplos típicos de reconocimiento de palabras conectadas incluyen el reconocimiento de dígitos conectados donde el vocabulario es el conjunto de los 10 dígitos (cero ... nueve) o el reconocimiento de letras conectadas donde el vocabulario es el alfabeto.

El sistema general de reconocimiento de palabras conectadas empleando patrones es muy similar al sistema de reconocimiento de palabras aisladas. El único paso que varía es el de comparación del patrón de entrada con los patrones de referencia, porque en este caso el patrón de entrada contiene varias palabras cuyas fronteras dentro de la señal no son conocidas. Las técnicas de programación dinámica son especialmente adecuadas para este tipo de problemas ya que pueden resolver de forma simultánea los problemas de detección de límites entre palabras, almacenamiento y ajuste de tareas del usuario y reconocimiento. Esto descarta la posibilidad de errores causados por una segmentación errónea de la señal de prueba. Existen dos enfoques principales para diseñar algoritmos de programación la

concatenación de patrones de referencia para crear patrones de referencia de palabras conectadas y la generalización de las reglas de transición descritas por (3.4), (3.5) y (3.6) para generar un algoritmo de una sola etapa.

III.2.1. Concatenación de patrones de referencia.

Este primer enfoque considera que la cadena de palabras por reconocerse debe ser comparada con las diferentes cadenas posibles formadas por la concatenación de patrones de referencia. La dificultad que esto involucra se debe a la cantidad de comparaciones necesarias para poder realizar una búsqueda exhaustiva entre todas las posibilidades de concatenación. Diferentes autores (L. Rabiner, H. Sakoe) encuentran soluciones aproximadas que minimizan, sujetando ciertos criterios, el número de cálculos requeridos. Dichas optimizaciones se basan tanto en las propiedades del algoritmo dinámico como en la definición de la zona de libertad. Es necesario dividir el problema de optimización en diferentes etapas, por lo que son algoritmos más complejos.

El problema principal que hemos notado que coexiste con estos métodos es que, si bien si se compara el patrón de prueba con todas las concatenaciones posibles de patrones, no se analizan, explícita o implícitamente, todos los caminos, o funciones de ajuste dinámico que relacionen al patrón de prueba con las cadenas de prueba. Por otra parte, para asegurar su buen funcionamiento estos métodos requieren de parámetros determinados experimentalmente, y su eficiencia decae en función del número de palabras que contengan las señales de prueba.

III.2.2. Algoritmo de una sola etapa para reconocimiento de palabras conectadas.

Este algoritmo fue originalmente desarrollado para señales aisladas, aceptando [illegible] preparativos hasta que H. Sakoe lo reformuló de una manera más versátil, aceptando expresiones. Este algoritmo es una generalización del algoritmo empleado para señales aisladas. En el caso de palabras aisladas, la función de ajuste [illegible] dinámico [illegible] el patrón

de entrada está compuesto por i=1,.....,N ventanas de análisis y cada patrón de referencia está compuesto por j=1,.... J(k) ventanas de análisis. k es el identificador del patrón de referencia. En el caso de reconocimiento de palabras conectadas permitiremos que la función de ajuste tenga tres dimensiones, respetando ciertas características impuestas por la naturaleza de las señales involucradas:

$$W=(w(1),w(2),\ldots,w(L)) \qquad \text{donde } w(l)=(i(l),j(l),k(l)) \qquad (3.9)$$

Tomando en cuenta que las distancias locales ahora pueden tener tres argumentos, d(i,j,k), entonces el problema de reconocimiento puede ser tratado como el siguiente problema de minimización

$$\underset{w}{MIN} \sum d(w(l)) \qquad (3.10)$$

Este problema es el de minimizar la distancia global con respecto a todos los caminos posibles. Como ya vimos, este problema puede resolverse mediante programación dinámica. Falta sin embargo definir las reglas de transición deseables e impuestas por las características de las señales de voz.

- Regla de transición en el interior de una referencia

Si $w(l)=(i,j,k)$ y $i<I$, entonces

$$w(l+1)=(i+1,j,k),(i+1,j+1,k),(i+1,j+2,k) \qquad (3.11)$$

Esta regla de transición es idéntica al caso de palabras aisladas.

- Regla de transición entre referencias

si w(l)=(i,1,k), entonces,

$$w(l-1)\in\{(i-1,1,k),(i-1,j(k'),k')\ k'=1,\ldots\ldots,K\} \qquad (3.12)$$

Esta regla implica que se puede llegar a la referencia k desde la última ventana de análisis de cualquier referencia, incluyendo la misma referencia k.

Mediante programación dinámica se puede determinar fácilmente la distancia mínima acumulada. En este caso no nos interesa tanto la distancia acumulada como la trayectoria que sigue la función de ajuste dinámico. En efecto, una vez encontrada esta distancia, se puede determinar la trayectoria que la generó, y determinar así por qué referencias pasó y en qué instantes ocurrió transición de una referencia a otra, resolviendo así simultáneamente los problemas de segmentación y de reconocimiento.

Capitulo IV

Descripción de los experimentos y resultados

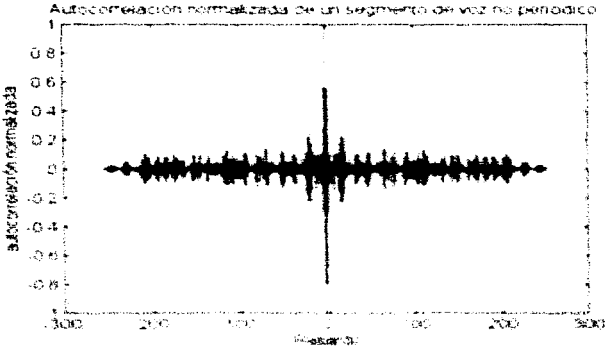
IV.1. Generalidades acerca de los experimentos

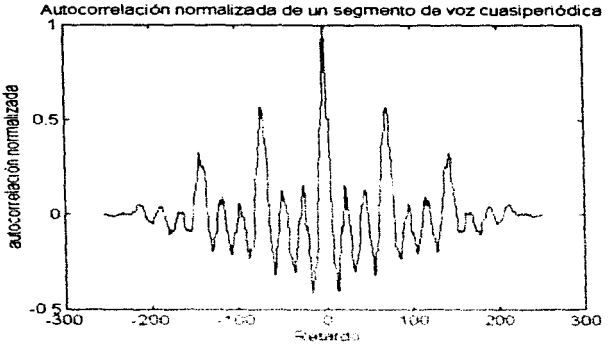
Todos los experimentos que a continuación se presentaran involucraron el uso de algún lenguaje de programacion. Se opto por emplear el ambiente de programación MATLAB porque ofrece un gran diversidad de opciones de visualizacion y flexibilidad en el manejo de todo tipo de variables. La contraparte de estas cualidades es que los tiempos de ejecucion fueron mayores a los que hubiesemos observado al emplear el lenguaje de programacion C. Se opto por MATLAB tomando en cuenta que el objetivo de este trabajo no es producir un sistema de reconocimiento de palabras que funcione en tiempo real, sino analizar exhaustivamente los metodos de analisis dinamico del tiempo y la representacion por modelado de tipo LPC. Las herramientas graficas del ambiente mencionado resultaron fundamentales en el estudio de los conceptos analizados. Se empleo la version 1.0 del compilador para MATLAB que permite compilar rutinas escritas en MATLAB (extension .m) en rutinas mas rapidas llamadas ejecutables de MATLAB (extension .mex) mediante el compilador XNSF WATCOM version 10.0

IV.2. Experimento 1. Verificacion experimental de la validez del modelo fuente-filtro LPC

En este experimento se hizo determinar si en realidad es lo suficientemente confiable como para poder efectuar una resintesis de una determinada señal de voz. Se escribieron los programas que aparecen comentados en el apendice A (LPCMEX.m) para versiones de LPC resintesis cada vez mas exactas. Para cada versión el programa calcula la energia y cantidad de coeficientes LPC que debera usar el compilador. Por otro lado, el programa determina el error cuadratico medio entre la señal original y la resintetizada. Para lograr

esta última estimación es necesario emplear ventanas más largas que contengan más de un periodo y el programa analiza por lo tanto ventanas de 256 muestra cada 64 muestras. El análisis de periodicidad esta basado en la autocorrelación de la señal. Existen métodos complejos para determinar la periodicidad de una señal, pero en este caso se determinó el procedimiento a seguir basándonos en observaciones experimentales ya que el rango de frecuencias fundamentales posibles es pequeño. Primero se limita el análisis al rango de tonos fundamentales que tiene una voz humana convencional (100-300Hz). Dentro del rango de periodos equivalentes se determina si alguna autocorrelación tiene como magnitud por lo menos el 40% de la magnitud máxima. Si existen varios picos mayores a este umbral se escoge el pico máximo. Si no existen picos mayores a este umbral se determina que el segmento no es periódico. En efecto se observó que en el caso de no periodicidad las autocorrelaciones son menores a este umbral mientras que en el caso de periodicidad la autocorrelación correspondiente al periodo es mayor a este umbral, como se ve a continuación

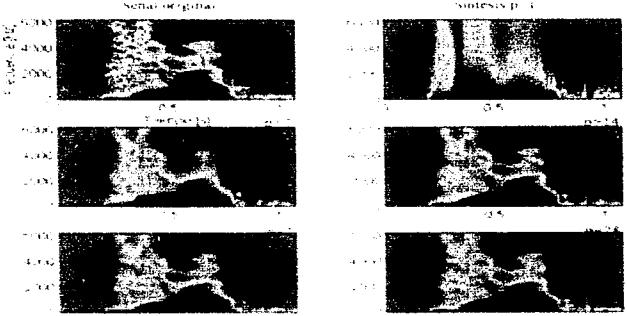




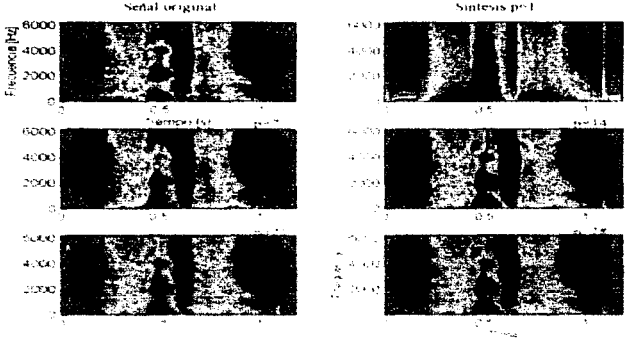
El programa resintetiza después una señal, basándose en los patrones de parametros obtenidos en el analisis. Para efectos de continuidad en la fase de la frecuencia fundamental se genera primero una señal excitadora completa, lo que incluye preservar dicha continuidad cuando se llega a un fin de ventana y cambio de frecuencia fundamental. Despues se procede a filtrar esta señal mediante los coeficientes LPC correspondientes. Se obtiene entonces una señal sintetica cuya energia se determina con analisis por ventanas de 128 muestras cada 64 muestras para determinar la energia y esta se normaliza a 1. El paso final consiste en multiplicar los segmentos de señal por los valores adecuados para que el perfil final de energia de la señal sintetizada corresponda al perfil de energia de la señal original.

Mediante pruebas auditivas se comprobó que señales sintetizadas con residuales de orden superior a 7 son inteligibles y que señales resintetizadas con residuales de orden superior a 21 demuestran gran similitud a la original. Hasta un punto de que es casi imposible distinguir. La mejor prueba se efectúa escuchando alternadamente la señal original y las señales sintetizadas. Cuando la señal sintetizada es inteligible, pero notablemente diferente

espectrogramas de una señal original y de las señales resintetizadas con diferentes ordenes de modelado para la palabra /zero/. La distribución de los espectrogramas es la siguiente: de derecha a izquierda y de arriba hacia abajo, señal original, señal sintética de orden 1, señal sintética de orden 7, señal sintética de orden 14, señal sintética de orden 21 y señal sintética de orden 28.



Esto es la misma distribución de espectrogramas para la palabra /six/.



En algunas pruebas experimentales resultó que la palabra reconstruida tendría algo similar a un fantasma que no existía en la grabación original.

IV.3. Experimento II: Reconocimiento de palabras aisladas

Para realizar este experimento se empleo una base de datos que contiene grabaciones de los diez digitos en ingles (0-9). Diez parlantes repiten cada digito un total de 26 veces cada uno,10 para entrenamiento y 16 para prueba, generando así una base de 2600 grabaciones. Se contó por lo tanto con 1000 grabaciones de referencia con las que se reconocieron 1599 palabras de prueba (un archivo de prueba de la palabra 'nine' resulto estar vacio por lo que se decidio descartarlo). Se efectuo primero una deteccion de inicio y fin de palabra y se determinaron experimentalmente los siguientes umbrales:

ITL=(maximo de energia)*0.4

ITH=(media de la energia del ruido de fondo)+2*(desviacion estandar de la energia del ruido de fondo)

ITZ=(media de los cruces por cero del ruido de fondo)+2*(desviacion estandar de los cruces por cero del ruido de fondo)

Despues cada grabacion fue transformada a un patron de coeficientes LPC de orden p=7, analizando ventanas de 128 muestras cada 196 muestras. Para cada patron de prueba se determinaron las distancias mínimas con respecto a todas las referencias, asignando una distancia infinita a las referencias cuyas longitudes eran mayores a 4/3 de la longitud de la prueba o menores a 3/4 de la longitud de la misma. El algoritmo empleado fue el propuesto por H. Sakoe por ser el que ha demostrado mejores tasas de reconocimiento.

Se obtuvieron las siguientes estadísticas de reconocimiento (cada renglón corresponde a 160 pruebas y las columnas corresponden al resultado del reconocimiento):

prueba 1:
ajuste lineal (r=0)

	one	two	three	four	five	six	seven	eight	nine	zero
one	111	2	1	1	4	0	1	1	38	1
two	1	143	1	1	6	1	2	2	1	2
three	4	9	116	0	10	2	3	4	11	0
four	10	4	10	130	6	0	1	10	0	9
five	10	1	1	2	108	0	5	0	41	0
six	1	8	8	2	4	127	10	6	0	0
seven	1	9	0	6	1	1	121	0	3	18
eight	4	1	2	0	0	7	0	146	0	0
nine	27	3	1	5	8	0	0	0	111	8
zero	3	2	0	2	1	0	18	0	1	1136

359 errores de 1599 pruebas. Tasa de reconocimiento [illegible]

prueba2:

Sin aplicar presentadas a las señales empleando una ventana de ventana n=14 empleado por H

Salida sin normalizar y sin explicar

	one	two	three	four	five	six	seven	eight	nine	zero
one	158	0	0	1	0	0	0	0	1	0
two	0	159	0	0	0	0	0	0	0	1
three	0	0	156	0	0	0	0	0	0	0
four	0	0	0	156	0	0	0	0	0	0
five	4	0	0	0	148	0	0	0	8	0
six	0	1	0	0	0	156	0	3	0	0
seven	0	0	0	0	0	0	156	0	0	0
eight	0	0	0	0	0	0	0	160	0	0
nine	2	0	0	0	0	0	0	0	158	0
zero	0	1	0	0	0	0	0	0	0	159

24 errores de 1599 pruebas. Tasa de reconocimiento [illegible]

prueba 3:

r=14 con preenfasis después de la detección de inicio y fin

	one	two	three	four	five	six	seven	eight	nine	zero
one	159	0	0	1	0	0	0	0	0	0
two	0	160	0	0	0	0	0	0	0	0
three	0	0	157	0	0	0	0	0	0	0
four	1	0	0	159	0	0	0	0	0	0
five	1	0	0	0	157	0	0	0	2	0
six	0	0	0	0	0	158	2	0	0	0
seven	0	0	0	0	0	0	160	0	0	0
eight	0	0	0	0	0	0	0	160	0	0
nine	0	0	0	0	0	0	0	0	159	0
zero	0	0	0	0	0	0	0	0	0	160

10 errores TASA=99.37%

prueba 4

r=14 con preenfasis antes de la detección de inicio y fin

	one	two	three	four	five	six	seven	eight	nine	zero
one	160	0	0	0	0	0	0	0	0	0
two	0	159	1	0	0	0	0	1	0	0
three	0	0	160	0	0	0	0	0	0	0
four	1	0	0	159	0	0	0	0	0	0
five	1	0	0	0	159	0	0	0	0	0
six	0	0	0	0	0	160	0	0	0	0
seven	0	0	0	0	0	1	159	0	0	0
eight	0	0	0	0	0	0	0	160	0	0
nine	0	0	0	0	0	0	0	0	160	0
zero	0	0	0	0	0	0	0	0	0	160

4 errores TASA=99.75%

Una vez demostrada la utilidad del preentasis aplicado antes de la detección de inicio y fin de palabra se realizaron pruebas con diferentes valores de r, el ancho de la región de validez, aplicando dicho preentasis, para determinar experimentalmente el valor optimo.

prueba 5:
r=18

	one	two	three	four	five	six	seven	eight	nine	zero
one	160	0	0	0	0	0	0	0	0	0
two	0	159	0	0	0	0	0	1	0	0
three	0	0	160	0	0	0	0	0	0	0
four	1	0	0	159	0	0	0	0	0	0
five	1	0	0	0	159	0	0	0	0	0
six	0	0	0	0	0	160	0	0	0	0
seven	0	0	0	0	0	1	159	0	0	0
eight	0	0	0	0	0	0	0	160	0	0
nine	0	0	0	0	0	0	0	0	159	0
zero	0	0	0	0	0	0	0	0	0	160

4 errores 1 ASA=99.75%

prueba 6:
r=10

	one	two	three	four	five	six	seven	eight	nine	zero
one	160	0	0	0	0	0	0	0	0	0
two	0	159	0	0	0	0	0	1	0	0
three	0	0	160	0	0	0	0	0	0	0
four	1	0	0	159	0	0	0	0	0	0
five	1	0	0	0	159	0	0	0	0	0
six	0	0	0	0	0	160	0	0	0	0
seven	0	0	0	0	0	1	159	0	0	0
eight	0	0	0	0	0	0	0	160	0	0
nine	0	0	0	0	0	0	0	0	159	0
zero	0	0	0	0	0	0	0	0	0	160

4 errores 1 ASA=99.75%

prueba 7:

r=6

	one	two	three	four	five	six	seven	eight	nine	zero
one	160	0	0	0	0	0	0	0	0	0
two	0	159	0	0	0	0	0	1	0	0
three	0	0	159	0	0	1	0	0	0	0
four	1	0	0	159	0	0	0	0	0	0
five	1	0	0	0	159	0	0	0	0	0
six	0	0	0	0	0	160	0	0	0	0
seven	0	0	0	0	0	1	159	0	0	0
eight	0	0	0	0	0	0	0	160	0	0
nine	0	0	0	0	0	0	0	0	159	0
zero	0	0	0	0	0	0	0	0	0	160

5 errores TASA=99.69%

prueba 8

r=2

	one	two	three	four	five	six	seven	eight	nine	zero
one	160	0	0	0	0	0	0	0	0	0
two	0	159	0	0	0	0	0	1	0	0
three	0	0	159	0	0	1	0	0	0	0
four	1	0	0	159	0	0	0	0	0	0
five	1	0	0	0	159	0	0	0	0	0
six	0	0	0	0	0	160	0	0	0	0
seven	0	1	0	0	0	0	159	0	0	0
eight	0	0	0	0	0	0	0	160	0	0
nine	0	0	0	0	0	0	0	0	159	0
zero	0	0	0	0	0	0	0	0	0	160

8 errores TASA=99.[illegible]%

prueba 9:

r=1

	one	two	three	four	five	six	seven	eight	nine	zero
one	160	0	0	0	0	0	0	0	0	0
two	0	159	0	0	0	0	0	1	0	0
three	0	0	159	0	0	1	0	0	0	0
four	1	0	0	159	0	0	0	0	0	0
five	1	0	0	0	159	0	0	0	0	0
six	0	0	0	0	0	160	0	0	0	0
seven	0	1	0	0	0	0	159	0	0	0
eight	0	0	0	0	0	0	0	160	0	0
nine	0	0	0	0	0	0	0	0	159	0
zero	0	0	0	0	0	0	0	0	0	160

12 errores TASA=99.[illegible]%

IV.4. Experimento III: Reconocimiento de palabras conectadas.

En este experimento se empleó una base de datos que contiene dos repeticiones de cada dígito en inglés (0-9 y la segunda versión del cero en inglés pronunciada /o/) por catorce parlantes distintos. Estas grabaciones fueron empleadas como patrones de referencia. Se seleccionaron manualmente regiones de ruido de fondo para generar los patrones de referencia para dicho ruido. Esto permite mayor flexibilidad en el caso en que la señal de prueba contenga una región de silencio entre dos dígitos. Por otra parte, la base contiene grabaciones de cadenas de dígitos de hasta cinco dígitos y de siete dígitos pronunciadas por los mismos catorce parlantes, que fueron empleadas como señales de prueba. Se empleó el algoritmo descrito por H. Sakoe por ser el más elegante y el que menor número de cálculos requiere, preservando un tasa de reconocimiento igual o superior a las registradas con los otros algoritmos. Ensayos preliminares con resultados desastrosos demostraron la necesidad de emplear la restricción sobre la pendiente local como se hizo en el caso de palabras aisladas. A continuación presentamos los resultados para cada longitud de cadena (en número de dígitos contenidos).

Longitud de la cadena	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
2 dígitos	[illegible]	18	2	[illegible]	[illegible]	[illegible]	[illegible]	0
3 dígitos	[illegible]	26	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	12
4 dígitos	[illegible]	50	4	[illegible]	[illegible]	[illegible]	[illegible]	0
5 dígitos	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
7 dígitos	[illegible]	[illegible]	8	[illegible]	4	[illegible]	[illegible]	[illegible]

Tabla IV.4.1 Errores de reconocimiento en palabras conectadas (experimento III).

Longitud de la cadena	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
2 dígitos	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
3 dígitos	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
4 dígitos	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
5 dígitos	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
7 dígitos	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Capitulo V

Análisis y discusión de los resultados

V.1. Experimento 1

En este experimento se pudo demostrar el buen funcionamiento práctico del modelo fuente-filtro de la voz basado en los coeficientes LPC. En algunos espectrogramas se observaron ciertas estructuras en forma de bandas verticales que no existen en la señal original. Despues de un analisis del patron de periodicidad se determinó que dichas estructuras aparecen al existir errores en la deteccion de periodicidad. En efecto, unicamente ocurre este fenomeno en ausencia de señal de voz y ocurre cuando por algun motivo el ruido aleatorio de fondo presenta algun tipo de periodicidad en un periodo corto. Esto no es grave ya que dichas estructuras tienen una energia muy pequeña y practicamente no son audibles.

Podemos constatar aqui la validez del mandamiento del ruido de fondo que se efectuo. En efecto, el modelo fuente filtro esta diseñado para representar una señal de voz y no silencios o ruidos de fondo. De hecho, todo la señal considerada y sintetizada por el mismo modelo ya que en la mayoria de los casos el ruido aparente representado por una señal de entrada aleatoria y el filtro determinado por los coeficientes LPC estima de aproximarse el contenido espectral del sistema original. Tomando esto en cuenta, asi como el hecho que de todas formas los silencios o ruidos de fondo presentados tendrian una energia muy pequeña, se justifica el empleo del mismo modelo.

El efecto de [illegible] en algunos casos, se necesitaron un mayor numero, fue mas intrigante y requerio de mayor reflexión. Despues de un analisis se determino que dicho efecto se debe a que la señal sintetizada es ligeramente [illegible]. Esto se debe a errores de [illegible] en el procesamiento del programa. La sintetizacion de señal con una frecuencia fundamental ligeramente diferente a la original ocasiona [illegible] estructuras como periodicidades [illegible]. Estas [illegible] [illegible] [illegible] [illegible]

frecuencias armónicas a la fundamental de la señal original, determinadas por los coeficientes del filtro. Como consecuencia escuchamos dos sonidos, en vez de uno solo con tono y timbre propio en el caso de la señal original. Para mayor información sobre la organización de la percepcion del sonido recomendamos consultar (A. S. Bregman, 1996).

V.2. Experimento II

El primer paso del experimento consistio en realizar reconocimiento con ajuste lineal del tiempo para tener algun dato de comparacion. Observamos claramente la necesidad de emplear preenfasis en las señales, resaltando asi la informacion que nos permite distinguir entre sonidos sonoros. Las pruebas 2-4 respaldan esta afirmacion, mientras que las pruebas 3 y 4 demuestran que el preenfasis tambien trae un mejoramiento de la deteccion de inicio y fin de palabra. Explicamos esto de la siguiente manera.

Recordando que la mayoria de los errores de deteccion de inicio y fin ocurren en palabras que inician y/o terminan con sonidos sordos, y que la energia de estos sonidos esta concentrada por encima de 4-5 kHz, observamos que al aplicarles preenfasis su energia aumenta en comparacion con la energia del resto de sonidos. Esta energia esta generalmente repartida por todo el espectro. Esto se debe a que el preenfasis es de orden igual un filtrado pasa alto. De esta forma, los sonidos sordos se distinguen mas claramente del ruido de fondo cuando se le aplica preenfasis a la señal, y esto definitivamente facilita la deteccion de inicio y fin de palabra.

Las pruebas 4-7 demuestran que el valor optimo para el umbral de silencio es [illegible]. En efecto, con este valor se obtienen los mejores porcentajes, siendo de 94.4 % para [illegible] y [illegible] con una tolerancia de [illegible] cuadros. Cabe destacar que el valor de [illegible] es el mas versatil de usarse, ya que consistentemente nos proporciona porcentajes altos para una gran cantidad de cuadros, y por lo tanto, siempre se [illegible] un reconocimiento de al menos [illegible] (para [illegible] cuadros). Esto se debe a que [illegible] determinado la pertenencia de ciertos cuadros a [illegible] [illegible] Para [illegible] ensanchar de las [illegible] [illegible]

De estas pruebas podemos concluir que las variaciones temporales en el habla son pequeñas, de alrededor de +- 10 ventanas de analisis en este caso, lo que equivale aproximadamente a 90 milisegundos. Sin embargo, y a pesar de que estas variaciones sean pequeñas, el ajuste dinamico del tiempo es fundamental para un sistema de reconocimiento ya que mejoro en este caso la tasa de reconocimiento desde 77.55% hasta 99.75% en el caso optimo, para el cual se registraron unicamente 4 errores. Una observación posterior demostro que estos errores se deben a errores en la detección de inicio y fin de palabra, por lo que concluimos que un aumento en el orden del modelo LPC empleado no mejoraria el funcionamiento del sistema.

Determinamos que las caracteristicas optimas para un sistema de reconocimiento de palabras aisladas basado en ajuste dinamico del tiempo y empleando una representacion de la señal por patrones de coeficientes LPC son las siguientes:

- Un preenfasis debe ser aplicado a todas las señales previo cualquier procesamiento.
- El algoritmo debe ser el propuesto por H. Sakoe con r=20(ms) y un orden de modelo LPC de 7.

V.3 Experimento III

Los resultados obtenidos anteriormente arrojan excelentes porcentajes de reconocimiento de palabras correctamente aisladas en el vocabulario de prueba utilizado, lo encontramos sin embargo resaltar las deficiencias de reconocimiento antes dichas experimentando para diferentes usos del vocabulario por diferentes personas, en diversas condiciones. Fue así pruebas con palabras aisladas de propio uso hacia el entrenamiento y de palabras de diferentes personas para cada grupo de cada dígito. Para cada uno de estos experimentos, en adición a las pruebas con palabras correctamente aisladas de hablantes originales existentes anteriormente, se probaron uno de cada dígito para cada uno de nuevos hablantes. Finalmente, en el caso de palabras erróneamente se

consideraron las dos versiones del cero en inglés, /zero/ y /o/, generando así un vocabulario más grande y con mayor ambigüedad ya que /o/ es idéntica al último sonido de /zero/.

Debemos tomar en cuenta estos factores fundamentales que modificaron la precisión de cualquier sistema de reconocimiento de palabras. En efecto, al contar con menos palabras de referencia se tiene una representación mucho menos precisa de las diferentes posibilidades de pronunciación de cada palabra, aumentando así la posibilidad de confusión entre dos palabras. Este fenómeno fue más claro ya que en su gran mayoría los errores de reconocimiento se debieron a confusión entre las palabras 'five' y 'nine' y entre 'one' y 'nine'. Otro gran número de errores fue causado por la transformación de la palabra 'zero' en la palabra 'zero' seguida de la palabra 'o', resultando así en la inserción de un dígito. Esto se debe a la enorme similitud entre el último fonema de la palabra 'zero' y la palabra 'o'. Notamos también que la mayoría de los errores ocurrieron en los archivos de menor tamaño. Examinando con detalle estos archivos nos percatamos que la duración de algunos de sus dígitos es menor a la mitad de la duración de la mayoría de los correspondientes referencias. Como se empleó una limitación sobre la pendiente local de la función de ajuste esta función no pudo realizar un ajuste suficiente entre dichos dígitos y sus referencias, generando así el alto número de errores. Podríamos pensar que con una gramática esta restricción se podría resolver este problema, pero sin embargo definitivamente consideramos que esta restricción es imprescindible.

La cuarta columna [illegible] la palabra originalmente que en esta experimentación la deficiencia para que se [illegible] en base de reconocimiento [illegible] obtenidos para palabras aisladas [illegible] permitía obtener mejores porcentajes para palabras conectadas. Esta tabla debemos considerar un [illegible] de [illegible] para cada dígito representando las diferentes [illegible] palabras [illegible] número [illegible] de [illegible] para [illegible] palabras.

La creación de una misma base de referencia para reconocimiento de palabras aisladas y de palabras conectadas podría además permitir una comparación de los resultados obtenidos. Es importante mencionar que la creación de tal base de datos es un proceso que requiere de mucho tiempo y que podría ser, por ejemplo un trabajo, de servicio social extremadamente útil ya que además de ofrecer las ventajas mencionadas, permitiría trabajar con palabras en español. Por las características de los algoritmos empleados podemos anticipar que los resultados no deberían variar mucho, pero es importante sin embargo verificarlo prácticamente.

Conclusiones

De las experiencias obtenidas mediante el desarrollo de este trabajo podemos afirmar que el modelo fuente-filtro basado en los coeficientes LPC es una modelización razonable de la voz. Este puede ser empleado tanto para síntesis de señales como para reconocimiento. En el caso de síntesis, es deseable emplear un orden de filtro mayor a 21 para obtener señales de calidad, aunque señales comprensibles se obtienen con órdenes menores. En el caso de reconocimiento de palabras aisladas un orden p=7 fue suficiente para obtener resultados excelentes.

La determinación de un buen de reconocimiento orientado en el caso de palabras conectadas se debe principalmente al empleo de una base de referencia poco adecuada. Es importante recalcar aquí que una base de datos de referencia para nuevos ejemplos de palabras conectadas debe incluir además de diferentes pronunciaciones que representen las diferentes entonaciones para cada dígito, pronunciaciones que representen las otras variedades de pronunciación observadas en la pronunciación de cualquier de ellos. Por estos motivos podemos anticipar un aumento significativo en las tasas de reconocimiento de palabras conectadas, mediante el uso de una base de datos más adecuada.

Las siguientes [illegible] para corregir de forma [illegible] de palabras [illegible] una medida de [illegible] dígitos [illegible] del [illegible] Una [illegible] de destacadas propiedades para [illegible] En el [illegible] el [illegible] En el [illegible]

La creación de una base de datos completa en español es necesaria y urgente para poder continuar con la investigación en esta área en la Facultad. Esta base podría ser creada a través de un programa de servicio social ya que se trata de una tarea que requiere de mucho tiempo.

Bibliografía

Libros:

Thomas W. Parsons. Voice and Speech Processing. McGraw-Hill, 1987

L. R. Rabiner and R. W. Schafer. Digital Processing of Speech. Prentice-Hall, 1978

F. J. Owens. Signal Processing of Speech. McGraw-Hill, 1993

J. R. Deller, J. G. Proakis and J. H. Hansen. Discrete-Time Processing of Speech Signals. MacMillan, 1993

Alex Waibel and Kai-Fu Lee. Readings in Speech Recognition. Morgan Kaufmann Publishers, 1990

Albert S. Bregman. Auditory Scene Analysis. MIT Press, 1990.

Artículos:

D. R. Reddy. Speech Recognition by Machine: A Review. IEEE Proceedings, April 1976

R. W. Schafer and L. R. Rabiner. Digital Representations of Speech Signals. The Institute of Electrical and Electronics Engineers, 1979

L. R. Rabiner and [illegible]. Isolated and Connected Word Recognition-Theory and Selected Applications. [illegible]

F. Itakura. Minimum Prediction Residual Principle Applied to Speech Recognition. IEEE, 1974

H. Sakoe and S. Chiba. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. IEEE, 1978

H. Sakoe. Two-Level DP-Matching--A Dynamic Programming-Based Pattern Matching Algorithm for Connected Word Recognition. IEEE, 1979

H. Ney. The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition. IEEE, 1984