UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

DIVISIÓN DE ESTUDIOS DE POSGRADO

TESIS

MÉTODOS DE RECONOCIMIENTO DE PALABRAS AISLADAS USANDO SEGMENTACIÓN ACÚSTICA Y CUANTIZACIÓN VECTORIAL

PRESENTADA POR

MAURICIO ALBERTO MARTÍNEZ GARCÍA

PARA OBTENER EL GRADO DE

MAESTRO EN INGENIERÍA

(ELÉCTRICA)

DIRIGIDA POR

M.I. ABEL HERRERA CAMACHO

CIUDAD UNIVERSITARIA, OCTUBRE DE 1997

MÉTODOS DE RECONOCIMIENTO AUTOMÁTICO DE PALABRAS AISLADAS USANDO SEGMENTACIÓN ACÚSTICA Y CUANTIZACIÓN VECTORIAL

A mis padres Elisa y Alberto,
para quienes mi agradecimiento
es infinito.

A mi hermana, Gaby,
por todo tu apoyo y paciencia.

A mis amigos y compañeros de siempre.

De forma especial a mi amiga
Rosalía, por la gran ayuda prestada.

A la UNAM, Casa de Estudios orgullo de México.

En Especial al M.I. Abel Herrera
por el valioso apoyo brindado.

Al Consejo Nacional de Ciencia
y Tecnologia (CONACyT)
por las facilidades prestadas.

Indice

INTRODUCCIÓN

Desde la década de 1950, se habla de la "máquina de escribir fonética". El reconocedor automático de voz ha sido objeto de estudio durante 4 décadas. A pesar de los grandes esfuerzos y logros realizados por los investigadores en este campo, y de que ya es posible encontrar en tiendas convencionales de electrónica y computación sistemas capaces de reconocer palabras aisladas, aún nos encontramos lejos de lograr un sistema capaz de reconocer la palabra hablada por cualquier persona en cualquier ambiente.

Debido a lo anterior, se puede decir que aún no se conoce la forma de lograr el objetivo planteado, y lo que se propone en este trabajo, es un estudio de uno de los sistemas más utilizados en la actualidad, y una comparación con un método relativamente nuevo. Con esto se pretende presentar herramientas que sean útiles para investigaciones futuras.

Uno de los principales problemas de los sistemas actuales de reconocimiento es la necesidad que tienen de ser entrenados, es decir, de determinar un modelo adecuado para cada una de las palabras que el sistema debe reconocer. Los dos sistemas presentados aquí son de este tipo, es decir, ambos sistemas desarrollados constan de una etapa de entrenamiento, y una etapa de reconocimiento.

Las técnicas actuales de reconocimiento de voz, involucran un gran número de disciplinas, entre las que podemos mencionar: procesamiento de señales, acústica, reconocimiento de patrones, teoría de la información, lingüística, fisiología, ciencias de la computación y psicología. El desarrollo de todo sistema de reconocimiento de voz, requiere un conocimiento claro de aspectos básicos de cada una de las disciplinas involucradas, que en el caso de este trabajo, se describen en el capítulo I. El capítulo II estudia la técnica LPC que es una de las más utilizadas en la actualidad. En el capítulo III se estudia la técnica KLT, que se propone como alternativa a la anterior y finalmente, en el capítulo IV se comparan los resultados obtenidos mediante pruebas realizadas con las dos técnicas.

Capítulo I

PROCESAMIENTO BÁSICO DE VOZ

Análisis de la señal de voz

La señal de voz es una señal que varía en el tiempo de forma lenta, es decir, si la examinamos durante intervalos de tiempo lo suficientemente cortos (entre 5 y 100 milisegundos), encontraremos que tiende a ser estacionaria; si se examina por intervalos de tiempo mucho mas largos, encontraremos que la señal varía en función de los diferentes sonidos que contiene[10]. Estos sonidos se pueden clasificar de distintas formas, y la mas comúnmente utilizada es en silencios, sonidos sordos (los que son producidos sin las cuerdas vocales) y sonidos sonoros (los producidos mediante vibraciones periódicas de las cuerdas vocales). En la siguiente figura se muestra la forma de onda de una señal de voz.

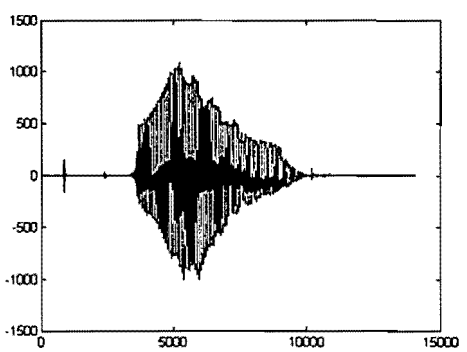
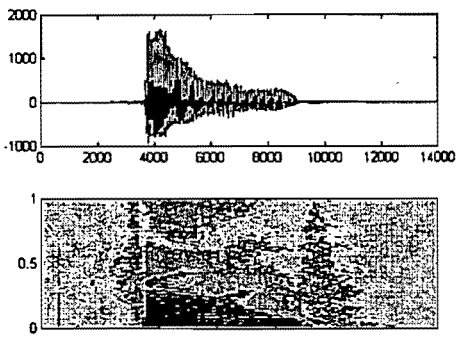


Fig. 1.1. Forma de onda de la palabra "one"

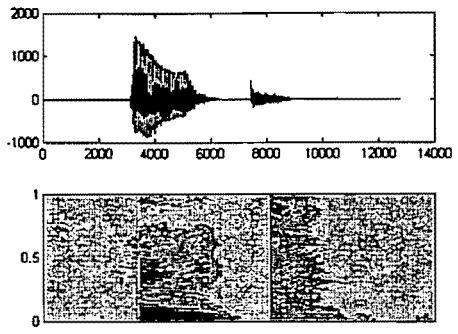
Conviene mencionar que si se desea segmentar una señal de voz en sus distintos sonidos, es difícil determinar la localización de las fronteras de forma exacta, aunque la experiencia ha demostrado que una exactitud de milisegundos es suficiente[10].

Una forma alternativa de caracterizar una señal de voz y representar la información contenida es la representación espectral. La forma mas utilizada es el espectrograma, en el que una representación tridimensional de la intensidad de la señal para las distintas bandas

de frecuencia, en un periodo de tiempo, es reducida a dos dimensiones[9]. En la siguiente figura se muestran dos señales de voz y sus espectrogramas.



(a) Gráfica de la señal de la palabra "five" y su espectrograma.



(b) Gráfica de la señal de la palabra "eight" y su espectrograma.

Fig. 1.2. Ejemplos de señales de voz y sus espectrogramas

La intensidad de la señal para cada frecuencia en particular se representa mediante la intensidad de la gráfica, de blanco (intensidad cero) a negro (intensidad máxima).

Mediante los espectrogramas es posible visualizar con facilidad las frecuencias formantes de los sonidos sonoros, y las fronteras entre sonidos sordos y sonoros. Debido a los primeros normalmente presentan componentes de frecuencias bajas, y los últimos suelen presentar componentes de alta frecuencia, se facilita enormemente la segmentación visual de la señal[9].

Existen muchas otras formas de caracterizar una señal de voz, y se puede decidir cual es la mas conveniente, dependiendo de la información que se requiera obtener de la señal. Una de las mas conocidas, debido a las aportaciones que ha ofrecido para el reconocimiento de palabras aisladas, es el modelo LPC (Linear Predictive Coding: Código de Predicción Lineal), en el que la señal de voz se caracteriza determinando los coeficientes de un filtro lineal que modela al aparato vocal. Otra forma de codificar señales de voz es mediante el uso de transformaciones. Estas técnicas se describirán con detalle en el capitulo III.

Sistema de Reconocimiento de voz

En su nivel mas elemental, un sistema de reconocimiento de voz consiste en una serie de algoritmos, incluyendo reconocimiento estadístico de patrones, teoría de comunicaciones, procesamiento de señales, combinatoria, y otros tipos. Aunque para cada tipo de sistema se confía en cada una de estas áreas, el común denominador es el procesamiento de señales, que convierte la forma de onda de la voz en una representación paramétrica para su posterior análisis y procesamiento. Como se menciono anteriormente, el objetivo de la representación paramétrica de una señal de voz es obtener información que pueda caracterizar de forma única a cada sonido o conjunto de sonidos (palabra).

Esta representación única se utiliza para formar lo que se conoce como "patrones de referencia", que se almacenan en una base de datos de palabras. La tarea de reconocimiento consiste en comparar los parámetros de la palabra que se desea reconocer con los patrones de referencia. El proceso de creación de los patrones de referencia se conoce como entrenamiento, y el proceso de comparación de la palabra con los patrones constituye el reconocimiento. En la siguiente figura se muestra el diagrama conceptual[9,10] de un sistema de reconocimiento de palabras aisladas.

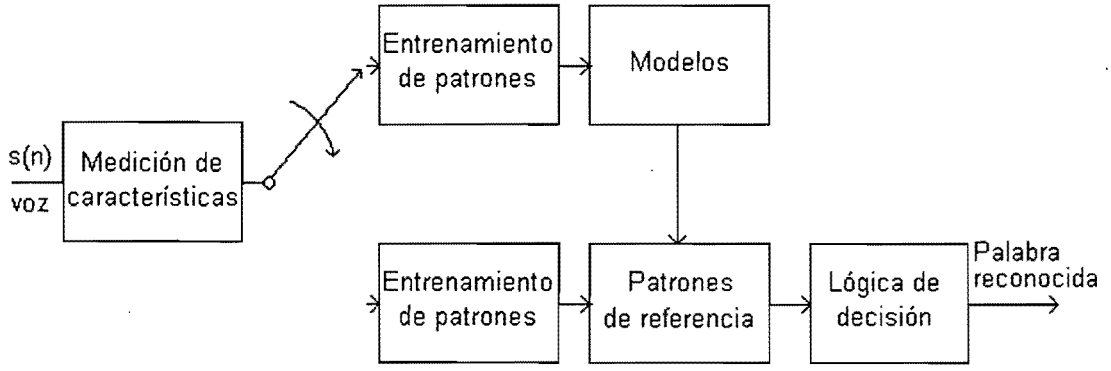


Fig. 1.3. Diagrama de bloques de un sistema básico de reconocimiento de palabras aisladas

El proceso de entrenamiento presenta el siguiente problema: dos señales que representan una misma palabra, generalmente tienen patrones distintos, incluso cuando provienen del mismo locutor y fueron creadas bajo condiciones optimas (en un estudio de grabación adecuado). Debido a esto, no es suficiente contar con una muestra de cada palabra para

crear la base de datos; es necesario contar con varias muestras de la misma palabra, y determinar dentro de que rangos se encuentran los parámetros para cada palabra. El problema es aun mayor si se desea que el sistema de reconocimiento funcione para varios locutores (sistema multiparlante), debido a que cada persona pronuncia las palabras de forma distinta, dependiendo de su sexo, nacionalidad, e incluso, estado de animo[11].

La comparación de la palabra a reconocer con los patrones de referencia, involucra el uso de una medida de distorsión o distancia, en las que, idealmente, si se comparan muestras de la misma palabra, el resultado debe ser menor que la comparación entre palabras distintas. El éxito de un sistema de reconocimiento de voz depende enormemente de la medida de distancia utilizada.

Las medidas de distancia mas conocidas son la distancia euclidiana (y algunas otras relacionadas), y la distancia de Itakura-Saitó, utilizada comúnmente en sistemas LPC.

Medidas de distancia[9,10,11]

Sean $\mathbf{x}$ y $\mathbf{y}$ definidas en un espacio vectorial X. La función de distancia es una función tal que

$0 \le d(\mathbf{x},\mathbf{y}) < \infty$

$d(\mathbf{x},\mathbf{y}) \Leftrightarrow \mathbf{x} = \mathbf{y}$

Simetría: $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$

Triángulo: $d(\mathbf{x},\mathbf{y}) \le d(\mathbf{x},\mathbf{z}) + d(\mathbf{y},\mathbf{z})$

Desplazamiento: $d(\mathbf{x}+\mathbf{z},\mathbf{y}+\mathbf{z}) = d(\mathbf{x},\mathbf{y})$

Idealmente una medida de distancia (o distorsión, si es que no cumple con todas las propiedades anteriores) debe ser lo suficientemente significativa como para permitir el análisis de la señal deseada, y debe ser computable de forma que pueda ser evaluada en un tiempo razonable (en ocasiones, tiempo real). La medida de distorsión más conveniente y más comúnmente usada es la distancia cuadrática o distancia euclidiana entre dos vectores, definida por:

$$d(X,\hat{X}) = \|X - \hat{X}\|^2 = \sum_{i=1}^{N} (X_i - \hat{X}_i)^2$$

La distorsión media cuadrática. frecuentemente llamada distorsión promedio, se define como

$$D = Ed(X,\hat{X}) = E\|X - \hat{X}\|^2$$

Esta última medida es frecuentemente asociada con la energía de una señal de error.

Numerosas medidas de distorsión se pueden definir para encontrar la similaridad entre los vectores de entrada y los vectores patrón. Normalmente presentan la forma

$$d(X,\hat{X}) = \sum_{i=1}^{k} d_m(X_i,\hat{X}_i)$$

Cualquier medida de distorsión que utiliza esta propiedad de aditividad, con la misma medida de distorsión escalar para cada componente, se denomina *medida de distorsión aditiva* y es particularmente apropiada para la codificación de formas de onda cuando cada componente tiene el mismo significado físico.

Otra medida de distorsión de particular interés es el *error medio ponderado*

$$d(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T \mathbf{W}(\mathbf{x}-\mathbf{y})$$

donde $\mathbf{W}$ es una matriz de ponderación simétrica, positiva definida, y los vectores $\mathbf{x}$ y $\mathbf{y}$ son tratados como vectores columna. En el caso de que la matriz de ponderación es igual a la matriz identidad, esta medida se reduce a la distorsión cuadrática usual. En el caso de una matriz $\mathbf{W}$ diagonal, con todos los elementos $w_{ii} > 0$, tenemos

$$d(x,y) = \sum_{i=1}^{k} w_{ii}(x_i - y_i)^2$$

que es una medida simple, pero de gran utilidad que permite dar un peso distinto a cada una de las componentes de un vector.

Nótese que la medida de distorsión de error medio ponderado puede ser vista como una medida cuadrática sin ponderación entre vectores linealmente transformados, $\mathbf{x}' = \mathbf{A}\mathbf{x}$ y $\mathbf{y}' = \mathbf{A}\mathbf{y}$ donde $\mathbf{A}$ se obtiene mediante la factorización $\mathbf{W} = \mathbf{A}^T\mathbf{A}$. En algunas aplicaciones, la matriz $\mathbf{W}$ se selecciona de acuerdo a características estadísticas del vector de entrada. En particular, si las componentes de $\mathbf{X}$ son no-correlacionadas, la matriz de ponderación

de Mahalanobis se reduce a una matriz diagonal, como se discutió antes, con $w_{ii} = \frac{1}{\sigma_i^2}$, donde σ_i^2 es la variancia de X_i.

Todas las medidas de distorsión discutidas hasta ahora son simétricas en sus argumentos $\mathbf{x}$ y $\mathbf{y}$. En algunas ocasiones es conveniente elegir una matriz $\mathbf{W}(x)$ que dependa explícitamente del vector de entrada $\mathbf{x}$ con el fin de obtener medidas de distorsión perceptualmente motivadas para compresión de señales de voz e imágenes. En este caso, la distorsión

$$d(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T \mathbf{W}(\mathbf{x})(\mathbf{x}-\mathbf{y})$$

es asimétrica. Un ejemplo de una medida de este tipo, sea $\mathbf{W}(x)$ igual a $\|\mathbf{x}\|^{-2}\mathbf{I}$ donde $\mathbf{I}$ es la matriz identidad. En este caso, la distorsión entre dos vectores es la razón de energía del error a la energía de la señal:

$$d(\mathbf{x},\hat{\mathbf{x}}) = \frac{\|\mathbf{x}-\hat{\mathbf{x}}\|^2}{\|\mathbf{x}\|^2}$$

Esto permite que la ponderación de la distorsión sea más importante para señales pequeñas que para señales grandes.

Finalmente, definimos la *distorsión máxima* o norma l_∞ por

$$d_{max}(\mathbf{x},\hat{\mathbf{x}}) = \max_i |\mathbf{x}_i - \hat{\mathbf{x}}_i|$$

donde la distorsión se determina por la componente del vector error $\mathbf{x}-\hat{\mathbf{x}}$ que contribuye al error absoluto más grande. Es bien conocido el hecho de que la norma l_m alcanza la norma l_∞ cuando $m \to \infty$, entonces,

$$\lim_{m\to\infty} \left[d_m(\mathbf{x},\hat{\mathbf{x}})\right]^{1/m} = d_{max}(\mathbf{x},\hat{\mathbf{x}})$$

Cuantización vectorial[7,9,10]

Un cuantizador vectorial Q de dimensión K y tamaño N es una transformación de un vector del espacio euclidiano de dimensión $\Re^k$ en un conjunto finito C que contiene N salidas o puntos de reproducción, llamados vectores de código (*code vectors*).

$$Q: \Re^k \to C \qquad C = \left\{ \underline{y}_1, \ldots \underline{y}_n \right\} \qquad \underline{y}_i \in \Re^k$$

C normalmente se denomina alfabeto (*codebook*).

Un vector puede ser utilizado para representar cualquier tipo de patrón, como por ejemplo, un segmento de una forma de onda de voz, o la envolvente de una señal de voz. La cuantización vectorial puede ser vista como una forma de reconocimiento de patrones donde el patrón de entrada se puede "aproximar" por un conjunto predeterminado de patrones modelo, en otras palabras, un conjunto de vectores de entrada se compara con un conjunto de vectores almacenados, o vectores patrón.

Antes de discutir los conceptos involucrados en el diseño de un sistema de cuantización vectorial, describiremos sus ventajas y desventajas[9].

Ventajas:

- Reducción de la cantidad de información en el almacenamiento de patrones
- Reducción del costo computacional para determinar la similaridad de un vector de entrada con los vectores patrón.
- Representación discreta de una señal de voz. Mediante esta representación, el proceso de elección del mejor patrón para representar una señal, es equivalente a asignar una etiqueta a cada tramo de señal. Esto es especialmente útil en la compresión de la información.

Desventajas:

- Distorsión inherente en la representación de un vector real (error de cuantización).
- El almacenamiento requerido es con frecuencia no trivial, debido al error de cuantización, el procesamiento requerido para encontrar el alfabeto (codebook), y el almacenamiento en sí, por lo que se debe buscar un buen balance entre estos tres factores.

Algoritmos de cuantización vectorial

Se han desarrollado diversos algoritmos para encontrar cuantizadores vectoriales, y podemos mencionar el algoritmo simple, el maximin, isodata, k-medias y LBG. Los métodos más comúnmente utilizados son los dos últimos, y el que se utiliza en este trabajo es el algoritmo de k-medias, que se describirá a continuación[7,9].

Algoritmo de k-medias

Criterio

$$J_e = \sum_{j=1}^{K} \sum_{\mathbf{x} \in \chi_j} \left\| \mathbf{x} - \mathbf{z}_j \right\|$$

donde:

K: Número de centroides

x_j: Centroide para el cúmulo j

χ_j: Subconjunto de muestras asociadas al centroide j.

Algoritmo:

1. Elegir K centroides iniciales $\mathbf{z}_1(1), \mathbf{z}_2(1), \cdots, \mathbf{z}_K(1)$
2.. En la iteración l, asignar las muestras a los cúmulos.

$$\text{Asignar } \mathbf{x} \text{ a } \chi_i(l) \text{ si } \left\| x - z_i(l) \right\| \leq \left\| x - z_j(l) \right\|, \quad j = 1,2,\cdots k, \quad i \neq j$$

3. Calcular los nuevos centroides para cada cúmulo

$$\mathbf{z}_i(l+1) = \frac{1}{N_i} \sum_{\mathbf{x} \in \chi_i} \mathbf{x} \quad i = 1,2,\cdots, K$$

4. Si $z_i(l+1) = z_i(l)$ para $i = 1,2,\cdots, K$ el algoritmo ha convergido. En caso contrario, ir al paso 2.

El comportamiento del algoritmo depende del número de centroides K elegidos, la forma en que se eligen inicialmente los centroides, el orden en que se toman las muestras, y las propiedades geométricas de los datos. La obtención de los resultados deseados frecuentemente requiere experimentación[9].

Un elemento muy importante de todo algoritmo de cuantización vectorial es la medida de distorsión utilizada. Las medidas de distorsión más comúnmente utilizadas fueron descritas anteriormente. Para el caso de cuantizadores con modelo LPC, se utiliza la distancia de Itakura-Saito[6].

Capítulo II

RECONOCIMIENTO DE VOZ LPC

Modelo de predicción lineal para reconocimiento de palabras aisladas

Antes de describir un sistema de reconocimiento LPC (Linear Predictive Coding), se mencionarán las razones para utilizarlo, estas son[9]:

- El modelo LPC es un buen modelo para señales cuasi-estacionarias, como las de voz, en las que el modelo LPC nos da una buena aproximación espectral de la señal. Aunque este modelo es mejor para sonidos sonoros que para los sordos, no deja de proveer una aproximación aceptable.

- El modelo LPC separa razonablemente la fuente de señal y el tracto vocal, por lo que se puede obtener una buena representación de las características del tracto vocal.

- El modelo LPC es preciso y relativamente sencillo de programar en hardware o en software.

- La experiencia ha demostrado que el reconocimiento mediante este modelo trabaja de forma aceptable.

El modelo LPC[9,10]

La idea básica de un modelo LPC es que dada una muestra de voz en el tiempo, $s(n)$, se puede aproximar como una combinación lineal de las p muestras pasadas, de forma que

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \cdots + a_p s(n-p)$$

donde los coeficientes $a_1, a_2, \cdots a_p$ se asumen constantes sobre toda la señal (que idealmente es estacionaria). Si convertimos la expresión anterior en una ecuación que incluye un término de excitación, $G \cdot u(n)$, obtenemos

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

donde $u(n)$ es una excitación normalizada y G es la ganancia de la excitación. En el dominio z tenemos

$$S(z) = \sum_{i=1}^{p} a_i z^{-i} S(z) + G \cdot U(z)$$

Convirtiendo esta expresión en función de transferencia, obtenemos:

$$H(z) = \frac{S(z)}{G\,U(z)} = \frac{1}{\sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{A(z)}$$

La interpretación de esta expresión se da en la siguiente figura, que muestra que la fuente de excitación normalizada, $u(n)$, con ganancia G, actúa como entrada al sistema para producir la señal $s(n)$.

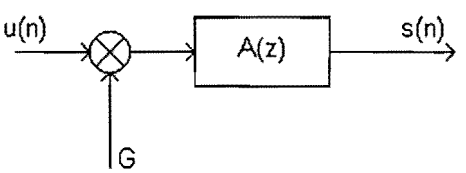


Fig. 2.1. Modelo LPC básico[10]

Sabiendo que la función de excitación real es esencialmente periódica (para producir señales de voz), el sistema apropiado para síntesis de voz se muestra en la siguiente figura. Aquí la función de excitación se elige mediante un interruptor para seleccionar entre un tren de impulsos para producir sonidos sonoros, o un generador de secuencias aleatorias para sonidos sordos. La señal de entrada con su respectiva ganancia, se introduce al filtro $H(z)$ que es controlado por los parámetros del tracto vocal del sonido que se desea producir.

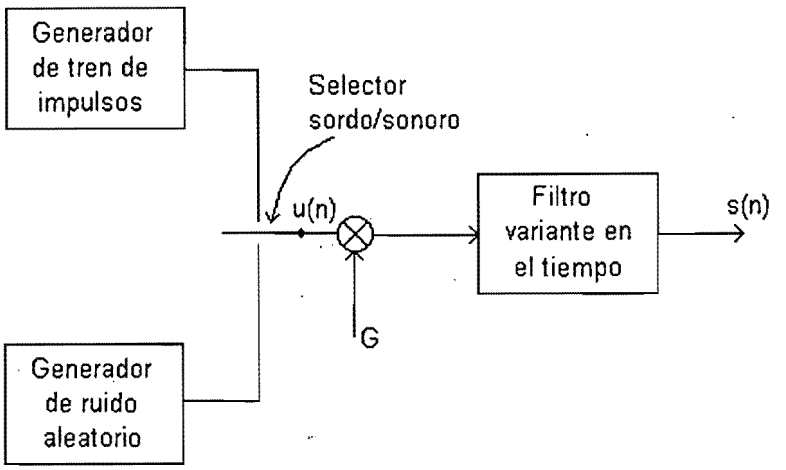


Fig. 2.2. Modelo LPC de síntesis de voz[10]

Ecuaciones de análisis LPC[9,10,11]

Basados en el modelo de la figura 2.1, la relación entre $s(n)$ y $u(n)$ es

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + G \cdot u(k)$$

Consideraremos la combinación lineal de las muestras pasadas como la estimación $\tilde{s}(n)$

$$\tilde{s}(n) = \sum_{k=1}^{p} a_k s(n-k)$$

De esta forma podemos definir un error de predicción $e(n)$

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k)$$

con función de transferencia de error

$$A(z) = \frac{E(z)}{S(z)} = 1 - \sum_{k=1}^{p} a_k z^{-k}$$

Claramente, si la señal se genera por un sistema lineal, entonces el error de predicción $e(n)$ será igual a la excitación $G \cdot u(n)$.

El problema básico del análisis de predicción lineal consiste en determinar el conjunto de valores $\{a_k\}$, a partir de la señal de voz, de forma que las propiedades espectrales del filtro

digital de la figura 2.1 sean iguales a las de la señal original. Lo que se puede hacer es encontrar los coeficientes que minimicen el error de predicción cuadrático medio:

$$E_n = \sum_m e_n^2(m)$$

Escribiendo $e_n(m)$ en términos de $s_n(m)$, obtenemos

$$E_n = \sum_m \left[s_n(m) - \sum_{k=1}^{p} a_k s_n(m-k) \right]^2$$

Para resolver el problema, necesitamos tomar la derivada con respecto a a_k de la expresión anterior e igualarla a cero, para minimizar

$$\frac{\partial E_n}{\partial a_k} = 0$$

con lo que llegamos a

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

Reconociendo que los términos de la forma $\sum_m s_n(m-i)s_n(m-k)$ son términos de la covariancia de $s_n(m)$, esto es,

$$\phi_n(i,k) = \sum_m s_n(m-i)s_n(m-k)$$

podemos expresar la última ecuación en forma compacta como

$$\phi_n(i,0) = \sum_{k=1}^{p} \hat{a}_k \phi_n(i,k)$$

que es un sistema de p ecuaciones con p incógnitas. Para resolver este sistema y obtener los coeficientes del predictor optimo, $\hat{a}_k$, es necesario calcular $\phi_n(i,k)$ para $1 \le i \le p$ y $0 \le k \le p$, y resolver el sistema de p ecuaciones resultante. En la práctica, se suelen utilizar dos métodos que se describen a continuación.

Método de autocorrelación[9,10]

Este método se utiliza para determinar los coeficientes óptimos del predictor lineal, $\hat{a}_k$, para un segmento de señal. Una forma relativamente sencilla de definir los límites del segmento es asumir que el segmento de la señal $s_n(m)$, de N muestras, es cero fuera del intervalo $0 \le m \le N-1$. Esto equivale a multiplicar la señal $s(n+m)$ es multiplicada por una ventana $w(m)$ de longitud finita, que también es cero en el intervalo $0 \le m \le N-1$. Entonces, el segmento señal se puede expresar como

$$s_n(m) = \begin{cases} s(n+m) \cdot w(m) & 0 \le m \le N+1 \\ 0 & \text{de otra forma} \end{cases}$$

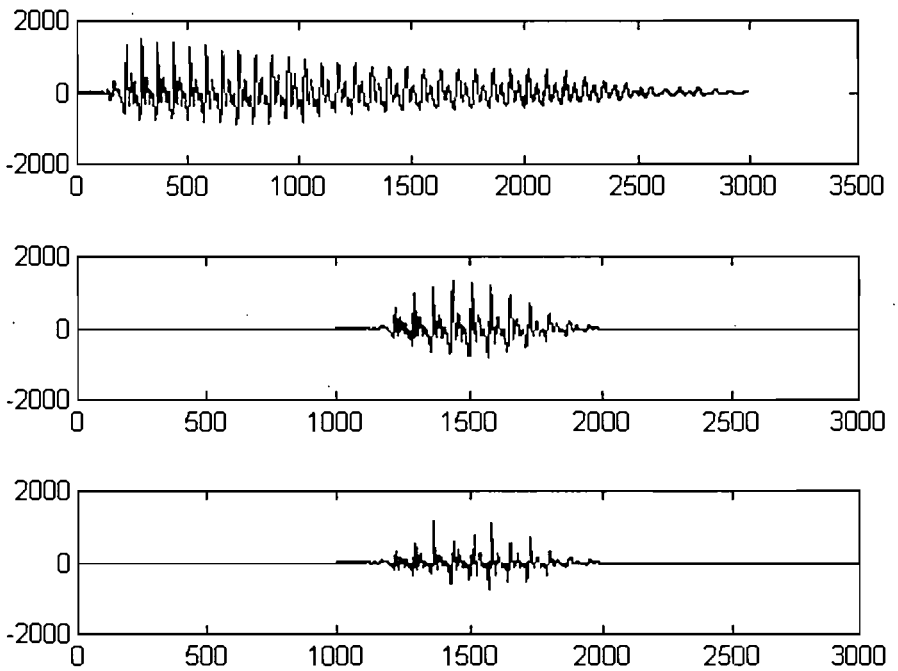


Fig. 2.3

Ilustración de (a) una muestra de señal de voz, (b) un tramo con ventana y (c) la señal estimada con LPC

Lo que se busca al ponderar la señal con una ventana como por ejemplo, la ventana de Hamming, es el minimizar la alteración en las características espectrales de la señal[10], como se mencionó anteriormente. Con esto se logra que, si aplicamos la definición de error cuadrático medio en el intervalo donde la señal ponderada es cero, el error también será cero. Como el predictor estima el valor presente de la señal en función a muestras

pasadas, es decir, el como filtro es causal, la aplicación de la ventana $w[n]$ causa equivale a forzar las condiciones iniciales a cero[10]. Si la señal analizada es estacionaria, la amplitud en los extremos del intervalo $0 \le m \le N-1$ será muy pequeña, logrando una reducción del error en los extremos del segmento, como se muestra en la figura 2.3.

Supongamos que se desean determinar los coeficientes de un filtro de orden p. Entonces, si utilizamos este predictor, con condiciones iniciales cero, el error cuadrático medio de la estimación es

$$E_n = \sum_{m=0}^{N-1+p} e_n^2(m)$$

donde $e_n(m) = s(m) - \tilde{s}(m)$,

y podemos escribir

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

o bien

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

como esta última expresión es función únicamente de $i-k$, podemos escribir

$$\phi_n(i,k) = r_n(i-k) = \sum_{m=0}^{N-1-(i-k)} s_n(m)s_n(m+i-k).$$

Como la función de autocorrelación es simétrica, esto es, $r_n(k) = r_n(-k)$, la expresión para LPC se puede escribir como

$$\sum_{k=1}^{p} r_n(|i-k|)\hat{a}_k = r_n(i), \quad 1 \le i \le p$$

y se puede expresar en forma matricial como:[9,10]

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

La matriz de este sistema de ecuaciones es una matriz de Toeplitz (simétrica, con todos sus elementos diagonales iguales) y puede res resuelto utilizando métodos bien conocidos, como el algoritmo de Levinson-Durbin, que se describirá a continuación.

Algoritmo de Levinson-Durbin[9,10]

El siguiente paso en el análisis LPC es la resolución del sistema de ecuaciones que resulta del análisis de autocorrelación. Es claro que este sistema se puede resolver por métodos tradicionales, sin embargo, debido a las propiedades que presenta la matriz de Toeplitz, es posible encontrar un método que resuelva el sistema de forma mucho más eficiente, es decir, reduciendo el número de operaciones necesarias. Este algoritmo se conoce como algoritmo de Levinson-Durbin. Debido a la eficiencia del algoritmo y a la velocidad de los microprocesadores modernos, en la actualidad es posible realizar reconocimiento de palabras aisladas, utilizando análisis LPC, en tiempo real. Este algoritmo se puede dar de la siguiente forma:

$$E^{(0)} = r(0)$$

$$k_i = \left\{ r\left(i - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} r(|i-j|) \right) \right\} \Big/ E^{(i-1)} \qquad 1 \le i \le p$$

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

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

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

Donde la solución final se obtiene como sigue:

a_m: Coeficientes LPC $= \alpha_m^{(p)}$ $\qquad 1 \le m \le p$

k_m: Coeficientes de reflexión (PARCOR: Correlación parcial)

Sistema completo de reconocimiento LPC

Pre-énfasis[8,9,10]

Si se analiza una señal de voz, podrá notarse que su representación frecuencial tiene una reducción de amplitud de unos 6 dB/octava. Esto significa que cada que se duplica la frecuencia, la amplitud de la señal se reduce en un factor de 16. En el análisis LPC es conveniente de alguna forma compensar esta característica, de forma que se logre una respuesta en frecuencia plana. Este proceso se conoce como pre-énfasis. En un sistema de procesamiento digital de señales, el pre-énfasis se puede realizar fácilmente mediante un filtro de primer orden, análogo a un filtro paso-altas con una frecuencia de corte de entre 100 Hz y 1 kHz (la posición exacta no es crítica), mediante la siguiente ecuación en diferencias:

$$y[n] = x[n] - ax[n-1]$$

donde $y[n]$ denota la salida del filtro, $x[n]$ es la muestra de entrada actual, $x[n-1]$ es la muestra anterior, y a es una constante usualmente entre 0.9 y 1. Nuevamente, el valor exacto no es crítico. Si tomamos la transformada Z del filtro, se tiene

$$Y(z) = X(z) - az^{-1}X(z) = \left(1 - az^{-1}\right)X(z)$$

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

$$H(z) = \frac{Y(z)}{X(z)} = 1 - az^{-1}$$

La magnitud de esta función es

$$\left|H\left(e^{j\omega T}\right)\right| = \left|1 - ae^{j\omega T}\right| = \sqrt{1 + a^2 - 2a\cos wT}$$

La gráfica de la respuesta en frecuencia de este filtro para a=0.9 y T=100 μs se muestra a continuación.

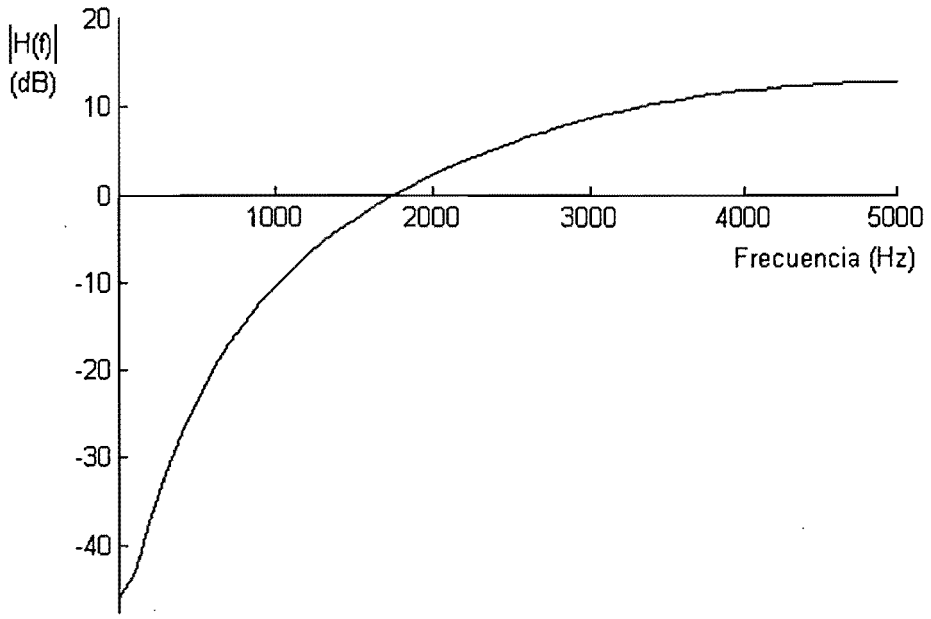


Fig. 2.4. Respuesta en frecuencia del filtro de pre-énfasis

Separación en bloques y ventanas

En este paso la señal pre-enfatizada de divide en bloques de N muestras separados por M muestras. Es claro que si $M \leq N$, entonces los bloques adyacentes tendrán estimadores LPC correlacionados. Si $M > N$ no habrá traslape entre bloques, y parte de la señal será perdida. Debido a que la simple separación de la señal en bloques alterará su respuesta en frecuencia (debido a que esto equivale a aplicar una ventana rectangular), es conveniente reducir esta alteración mediante el uso de una función conocida como ventana, que se aplica de la siguiente forma, para cada bloque de señal[10]:

$$x_w(n) = x(n) * w(n) \qquad \text{para} \quad 0 \leq n \leq N$$

donde $x(n)$ es la señal de entrada, $w(n)$ es la ventana, y $x_w(n)$ es la señal de salida.

La ventana más comúnmente utilizada, que se seleccionó para este trabajo, es la ventana de Hamming, definida por

$$w(n) = 0.54 - 0.46 * cos\left(2\pi n/(N-1)\right)$$

Distancia de Itakura-Saito[6,9]

Se basa en estimadores máximos de verosimilitud (likelihood)

$$d_i(X,\hat{X}) = \int_{-\pi}^{\pi} \left[e^{V(w)} - V(w) - 1 \right] \frac{d\omega}{2\pi} \qquad \text{donde } V(w) = \log X(w) - \log \hat{X}(w)$$

$$d_i(X,\hat{X}) = \int_{-\pi}^{\pi} \frac{X(w)}{\hat{X}(w)} \frac{d\omega}{2\pi} - \int_{-\pi}^{\pi} \log\left(\frac{X(w)}{\hat{X}(w)} \right) \frac{d\omega}{2\pi} - 1$$

El segundo término de la expresión anterior es $\log \frac{\sigma_\infty^2}{\sigma_\infty'}$ donde σ_∞^2 es la potencia del error.

$$S'(w) = \frac{\sigma^2}{\left| A\left(e^{j\omega}\right) \right|^2}$$

$$d_{IS}\left(S, \frac{\sigma^2}{\left| A\left(e^{j\omega}\right) \right|^2} \right) = \int_{-\pi}^{\pi} S(w) \frac{\left| A\left(e^{j\omega}\right) \right|^2}{\sigma^2} \frac{d\omega}{2\pi} - \log \sigma_\infty^2 + \log \sigma^2 - 1$$

De la primera integral de la expresión anterior, podemos obtener

$$\frac{1}{\sigma^2} \int_{-\pi}^{\pi} S(w) \left| A\left(e^{j\omega}\right) \right|^2 \frac{d\omega}{2\pi} = \frac{1}{\sigma^2} \int_{-\pi}^{\pi} S(w) \sum_{i=0}^{p} a_i e^{j\omega i} \sum_{k=0}^{p} a_k e^{j\omega k}$$

$$= \frac{1}{\sigma^2} \sum_{i=0}^{p} a_i \sum_{k=0}^{p} a_k \int_{-\pi}^{\pi} S(w) e^{j\omega i} e^{j\omega k}$$

$$= \frac{1}{\sigma^2} \sum_{i=0}^{p} a_i \sum_{k=0}^{p} a_k \int_{-\pi}^{\pi} S(w) e^{jw(i-k)} \frac{d\omega}{2\pi} = \frac{1}{\sigma^2} \sum_{i=1}^{p} a_i \sum_{k=1}^{p} a_k \, r(|i-k|) = \frac{1}{\sigma^2} \mathbf{a}^T \mathbf{R} \mathbf{a}$$

Entonces, la distorsión de Itakura-Saito es:

$$d_{IS}\left(S, \frac{\sigma^2}{\left| A\left(e^{j\omega}\right) \right|^2} \right) = \frac{1}{\sigma^2} \left[r(0) r_a(0) + 2 \sum_{n=1}^{p} r(n) r_a(n) \right] \qquad \text{con} \qquad r_a(n) = \sum_{i=0}^{p-n} a_i a_{i+n}$$

Esta última expresión es relativamente fácil de programar, y por lo tanto es especialmente útil en sistemas de reconocimiento de voz. Nótese que no cumple con la propiedad de simetría.

Sistema de reconocimiento[9,10,11]

A continuación se muestra el diagrama de bloques de un sistema completo de reconocimiento de palabras aisladas, utilizando análisis LPC y cuantizadores vectoriales.

Etapa de entrenamiento

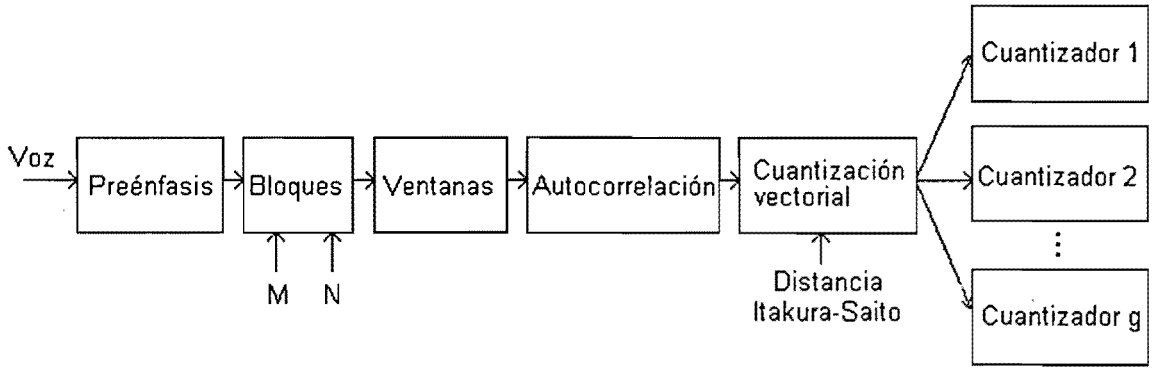


Fig. 2.5. Diagrama de bloques del sistema LPC de entrenamiento

Etapa de reconocimiento

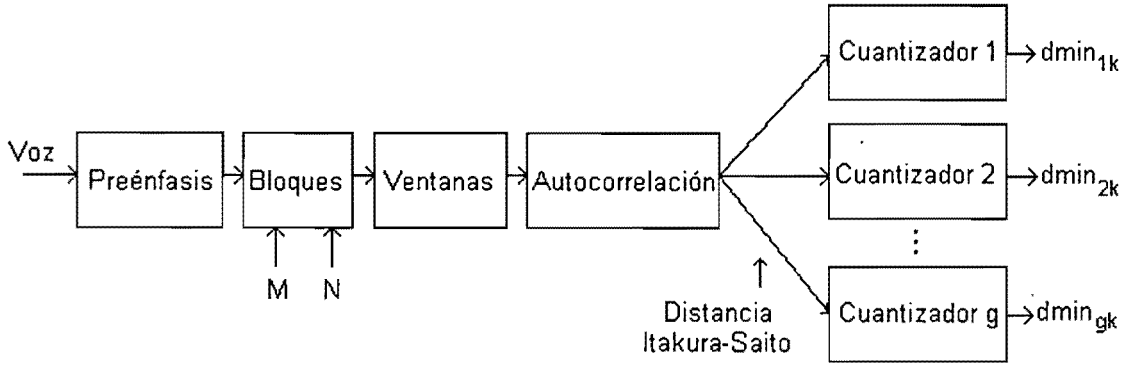


Fig. 2.6. Diagrama de bloques del sistema LPC de reconocimiento

Palabra reconocida j

$$\text{Si} \sum_{k=0}^{L} dmin_{jk} < \sum_{k=0}^{L} d_{ik} \qquad \forall\, j \neq i$$

donde

i : Número de cuantizador.

L: Número de bloques.

Capítulo III

SISTEMA DE RECONOCIMIENTO KLT

La transformada de Karhunen-Loève[1,5]

La transformada de Karhunen-Loève recibe varios nombres, como análisis de componentes principales, transformada Hotelling, y aproximación por vectores propios[5].

Se comienza con una población de vectores de la forma

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Definimos la media como

$$\mathbf{m}_x = E\{\mathbf{x}\}$$

donde $E\{\cdot\}$ es el 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:

$$\mathbf{m}_x = \frac{1}{M}\sum_{k=1}^{M}\mathbf{x}_k$$

La matriz de covariancia de una población de vectores se define como

$$\mathbf{C}_x = E\left\{(\mathbf{x}-\mathbf{m}_x)(\mathbf{x}-\mathbf{m}_x)^T\right\}$$

Esta matriz es de orden $n \times n$, real y simétrica. Entonces, cada elemento c_{ii} de la matriz es la variancia de x_i, y el elemento c_{ij} es la covariancia 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 tenemos una población de M vectores, la matriz de covariancia se puede aproximar mediante

$$\mathbf{C}_x = \frac{1}{M}\sum_{k=1}^{M}\mathbf{x}_k\mathbf{x}_x^T - \mathbf{m}_x\mathbf{m}_x^T$$

Debido a las propiedades de la matriz $\mathbf{C}_x$, siempre es posible encontrar un conjunto de valores propios ortonormales. Aquí, la idea es encontrar una base de forma que la matriz

sea lo más 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 $\mathbf{C}_x$.

Sabemos, del álgebra lineal, que λ es valor propio de $\mathbf{C}_x$ si y sólo si $\det(C_x - \lambda I) = 0$, donde I es 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 $\mathbf{C}_x$, con valor propio λ_i, para $i = 1, \cdots, n-1$. Por conveniencia, se pueden ordenar los vectores propios por orden descendente de valores propios, es decir, $\lambda_j \geq \lambda_{j+1}$ para $j = 1, \cdots, n-1$.

Sea $\mathbf{A}$ una matriz de $n \times n$ con sus renglones formados por los vectores característicos de $\mathbf{C}_x$. Entonces, el primer renglón contiene el vector propio con el valor propio más grande, y el último renglón, el vector propio con el valor propio más pequeño. Ahora podemos construir un vector $\mathbf{y}$ como sigue:

$$\mathbf{y} = \mathbf{A}(\mathbf{x} - \mathbf{m}_x)$$

esta ecuación de conoce como *Transformada de Karhunen-Loève*[5] (KLT). El vector $\mathbf{y}$ resultante tiene la propiedad $\mathbf{m}_y = 0$. La matriz de covariancia de $\mathbf{y}$ se puede obtener en términos de $\mathbf{A}$ y $\mathbf{C}_x$ mediante

$$\mathbf{C}_y = \mathbf{A}\mathbf{C}_x\mathbf{A}$$

donde $\mathbf{C}_y$ es una matriz diagonal, con los elementos de la diagonal principal iguales a los valores característicos de $\mathbf{C}_x$, es decir, $\mathbf{C}_y$ es de la forma:

$$\mathbf{C}_y = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_{21} & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{bmatrix}$$

Una propiedad importante de la transformada KLT está relacionada con la construcción de $\mathbf{x}$ a partir de $\mathbf{y}$. Debido a que los renglones de $\mathbf{A}$ son vectores ortonormales, entonces $\mathbf{A}^{-1} = \mathbf{A}^T$, de forma que:

$$\mathbf{x} = \mathbf{A}^T\mathbf{y} + \mathbf{m}_x$$

En vez de utilizar los vectores propios de $\mathbf{C_x}$, podemos utilizar únicamente los K vectores propios con los valores propios más grandes. Esto resulta en una matriz $\mathbf{A}_K$ con vectores $\mathbf{y}$ de dimensión K. El vector reconstruido por $\mathbf{A}_k$ es:

$$\hat{\mathbf{x}} = \mathbf{A}_k^T \mathbf{y} + \mathbf{m_x}$$

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

$$e_{ms} = \sum_{j=1}^{n} \lambda_j - \sum_{j=1}^{K} \lambda_j = \sum_{j=K+1}^{n} \lambda_j$$

El error puede ser minimizado seleccionando los K vectores propios con los vectores propios más grandes. La transformada KL es óptima en el sentido de que minimiza el error cuadrático medio entre el vector $\mathbf{x}$ y su aproximación $\hat{\mathbf{x}}$ [1].

Sistema completo de reconocimiento KLT

Diagrama de bloques propuesto[3]

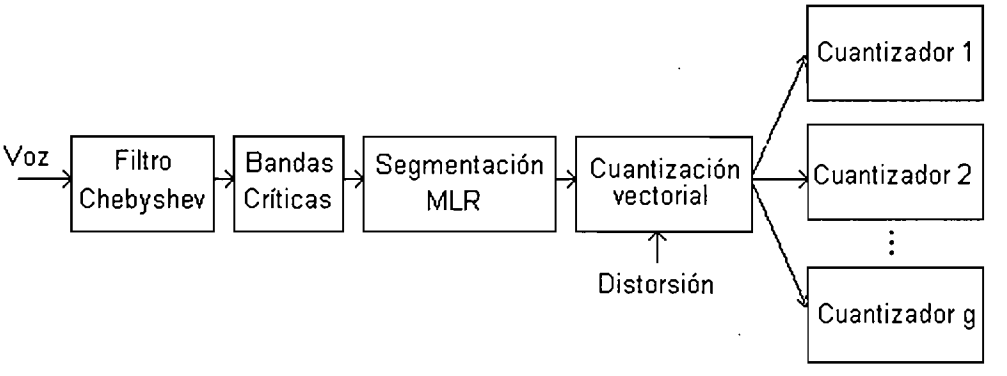


Fig. 3.1. Diagrama conceptual del sistema de reconocimiento KLT

Transformación

Filtro de Chebyshev[8]

En la primera etapa del procesamiento, se aplica a la señal de entrada un filtro digital de Chebyshev (es posible mostrar que el orden del filtro no es relevante), con el fin de limitar el intervalo de frecuencias de la señal, debido a que el algoritmo de segmentación acústica utilizado trabaja en un intervalo predefinido de frecuencias y además, como se menciono

anteriormente, la mayor cantidad de energía de una señal de voz se encuentra por debajo de los 5 kHz.

Segmentación de la señal

Se trata de un algoritmo de segmentación automática de la señal, que trabaja mediante la detección de cambios significativos en la señal de voz, por medio de comparación entre segmentos de análisis, mediante una medida de cambio en la señal.

La comparación entre los segmentos de análisis se realiza por pares, dividiendo el contenido frecuencial de la señal en M bandas y, aproximando la distribución de probabilidad a la distribución normal, para cada banda, se calcula la diferencia entre ambas mediante el algoritmo de *máxima verosimilitud* (Maximum Likelihood Ratio: MLR)[12,13]. Este método consiste en un criterio de decisión binaria, que para el caso de observaciones múltiples, se define como sigue[13]:

$$\Lambda(\mathbf{z}) = \frac{p(\bar{z}|m_2)}{p(\bar{z}|m_1)} \underset{H_2}{\overset{H_1}{\gtrless}} \eta$$

Donde $\mathbf{z}$ es un elemento del espacio de observaciones $\mathbf{Z}$, de dimensión L, H_1 y H_2 son las dos decisiones posibles, m_1 y m_2 son las señales, $p(\mathbf{z})$ es la función de densidad de probabilidad de las observaciones, y η es el umbral de decisión.

$\Lambda(\mathbf{z})$ se conoce como cociente de máxima verosimilitud, lo que indica que en la regla anterior si $\Lambda(\mathbf{z})$ es mayor que el umbral, se toma la decisión H_2, y en caso contrario de decide H_1.

Para el caso de señales de voz, se puede utilizar un vector de la forma $\mathbf{z}=[z_1, z_2, \cdots z_L]$. La probabilidad para $z \in \mathbf{z}$ está dada por[3]:

$$p(\mathbf{z}) = \prod_{i=1}^{L} \frac{1}{\sigma^2\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(\mu - z)^2}{\sigma^2}}$$

Considerando todo lo anterior, y asumiendo que la señal de voz tiene media cero, el cociente de verosimilitud se puede expresar como:

$$\Lambda(\mathbf{z}) = \frac{\prod_{i=1}^{L} \frac{1}{\sigma^2\sqrt{2\pi}} e^{-\frac{1}{2}\frac{z_{1,i}^2}{\sigma^2}}}{\prod_{i=1}^{L} \frac{1}{\sigma^2\sqrt{2\pi}} e^{-\frac{1}{2}\frac{z_{2,i}^2}{\sigma^2}}}$$

Calculando el logaritmo natural en ambos miembros de la ecuación, podemos escribir:

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

Como la media de la señal es cero, es posible mostrar que $\sum_{i=1}^{L}\frac{x_i^2}{\sigma^2} = L$, con lo que podemos simplificar la expresión anterior para obtener[3]:

$$\lambda = L\ln\frac{\sigma_2^2}{\sigma_1^2}$$

Bandas críticas[12]

La técnica de segmentación mencionada anteriormente se puede denominar *segmentación acústica* cuando los cambios de señal se detectan utilizando las denominadas bandas críticas, que podemos describir como el proceso de filtrado que tiene lugar en el sistema auditivo. Como un fenómeno puramente empírico, la banda crítica es el ancho de banda al que la respuesta subjetiva cambia abruptamente. Es decir, la intensidad de un sonido con ancho de banda determinado permanece constante al aumentar el ancho de banda, hasta cierto límite, en que la intensidad percibida comienza a aumentar. Este límite se conoce como ancho de banda crítico, y varía según la frecuencia central de la banda. En la tabla 3.1 se muestran las bandas críticas obtenidas a partir de datos obtenidos experimentalmente[12].

El algoritmo de segmentación acústica utilizado en este trabajo, combina el análisis de bandas críticas con el de máxima verosimilitud.

Medida de distancia (similitud)

Para el caso de 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[3,4]:

Tabla 3.1. Ejemplos de anchos de banda críticos.

Número	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

$$d = \frac{\boldsymbol{\lambda}^T \left(\mathbf{I} - \mathbf{A}^T \mathbf{A} \right) \boldsymbol{\lambda}}{l}$$

con los elementos de la matriz $\mathbf{A}$:

$$a_{ij} = \left\langle x_1[i], x_2[k] \right\rangle^2$$

donde:

$x_1[i]$: renglón i-ésimo de la matriz KLT para la señal 1.

$x_2[k]$: renglón k-ésimo de la matriz KLT para la señal 2.

λ :vector columna con los valores propios de la primera señal, de mayor a menor.

l : nivel de comparación (número de vectores KLT a utilizar).

Esta medida aprovecha la propiedad extracción de vectores de energía, jerárquicamente ordenados (con los valores propios correspondientes de mayor a menor), de la transformada KL. El uso de esta técnica presenta una ventaja adicional: si se utilizan vectores KLT, por ejemplo, de orden 18, se puede mostrar 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.

Cuantización vectorial

En un sistema KLT, es posible utilizar cuantización vectorial para encontrar los vectores patrón que modelen mejor a la señal de cada una de las palabras a reconocer, con la consiguiente pérdida de información (y distorsión introducida) propia de éstos métodos. Al igual que en el reconocimiento LPC, en este trabajo se utilizó el algoritmo de k-medias[7,9], utilizando la medida de distorsión descrita en la sección anterior.

Reconocimiento

Una vez obtenidos los cuantizadores vectoriales, la etapa de reconocimiento es muy similar a la utilizada en el sistema de reconocimiento LPC[9,10,11]. Una vez obtenidos los coeficientes KLT para cada segmento de la palabra que se desea reconocer, se calcula una suma de distancias mínimas a cada cuantizador patrón, para cada segmento.

Palabra reconocida j

$$\text{Si} \sum_{k=0}^{N_s} dmin_{jk} < \sum_{k=0}^{N_s} d_{ik} \qquad \forall\, j \neq i$$

donde

i : Número de cuantizador.

N_S: Número de segmentos.

Capítulo IV
RESULTADOS

En ambos sistemas, para lograr una buena consistencia de resultados y una buena comparación entre ambos métodos, se utilizó la base de datos de palabras aisladas de Texas Instruments, que fue diseñada específicamente para experimentos de procesamiento y reconocimiento de voz. Esta base de datos consta de un conjunto de palabras para entrenamiento, y un conjunto de palabras para prueba. Ambos conjuntos son los dígitos en inglés del 0 al 9. El conjunto de entrenamiento, consta de 10 muestras grabadas por 10 locutores, en total 1000 palabras (100 muestras de cada dígito). El conjunto de entrenamiento consta de 16 muestras grabadas por los mismos 10 locutores que el de entrenamiento, en total 1600 palabras (160 muestras de cada dígito). Los programas, cuyas rutinas más importantes se muestran a continuación, fueron desarrollados en lenguaje C y compilados en un sistema UNIX (SunOS).

En ambos casos se realizaron 4 pruebas.

Las dos primeras, utilizando segmentos de tamaño fijo (segmentación lineal, 4 segmentos/palabra), una con el conjunto de entrenamiento (de 1000 palabras), y la otra con el conjunto de reconocimiento (de 1600 palabras).

Las otras dos, utilizando segmentación acústica (4 segmentos/palabra), también una de ellas para cada conjunto de la base de datos.

Los resultados obtenidos se muestran a continuación en forma de tablas de confusión.

Sistema LPC

Técnica utilizada: Filtro de pre-énfasis con $a = 0.95$.

Ventanas de Hamming con $N = 128$, $M = 20$.

Medida de distorsión de Itakura-Saito.

Cuantizadores vectoriales de orden 16x16, uno para cada segmento.

Segmentación lineal, 4 segmentos por palabra.

a) Conjunto de entrenamiento

Palabra	1	2	3	4	5	6	7	8	9	0	%
1	99	0	0	1	0	0	0	0	0	0	99.00
2	0	100	0	0	0	0	0	0	0	0	100.0
3	0	0	100	0	0	0	0	0	0	0	100.0
4	0	0	0	100	0	0	0	0	0	0	100.0
5	0	0	0	0	100	0	0	0	0	0	100.0
6	0	0	0	0	0	100	0	0	0	0	100.0
7	0	0	0	0	0	0	100	0	0	0	100.0
8	0	0	0	0	0	0	0	100	0	0	100.0
9	0	0	0	0	0	0	0	0	100	0	100.0
0	0	0	1	0	0	0	0	0	0	99	99.00

En este caso una muestra de la palabra "one" fue reconocida erróneamente como "four", y una muestra de la palabra "zero" como "three". El promedio de reconocimiento es 99.80%.

b) Conjunto de reconocimiento

Palabra	1	2	3	4	5	6	7	8	9	0	%
1	142	0	0	5	5	0	1	0	6	1	88.75
2	0	152	2	0	0	4	0	0	0	2	95.00
3	0	1	156	1	0	0	0	0	0	2	97.50
4	3	0	0	155	0	0	0	0	0	2	96.88
5	0	0	0	0	149	0	0	0	9	2	93.13
6	0	3	0	0	0	154	3	0	0	0	96.25
7	0	0	1	0	0	9	150	0	0	0	93.75
8	0	0	4	0	0	1	0	150	2	3	93.75
9	7	0	0	0	6	0	1	0	146	0	91.25
0	0	2	7	0	0	2	0	1	0	148	92.50

El reconocimiento promedio en este caso es de 93.88 %.

Segmentación Acústica, 4 segmentos por palabra.

a) Conjunto de entrenamiento

Palabra	1	2	3	4	5	6	7	8	9	0	%
1	100	0	0	0	0	0	0	0	0	0	100.0
2	0	100	0	0	0	0	0	0	0	0	100.0
3	0	0	100	0	0	0	0	0	0	0	100.0
4	0	0	0	100	0	0	0	0	0	0	100.0
5	0	0	0	0	100	0	0	0	0	0	100.0
6	0	0	0	0	0	100	0	0	0	0	100.0
7	0	0	0	0	0	0	100	0	0	0	100.0
8	0	0	0	0	0	0	0	99	0	1	99.00
9	0	0	0	0	0	0	0	0	100	0	100.0
0	0	0	0	0	0	0	0	0	0	100	100.0

El promedio de reconocimiento es 99.90%.

b) Conjunto de reconocimiento

Palabra	1	2	3	4	5	6	7	8	9	0	%
1	139	0	0	0	7	0	0	0	8	0	86.88
2	0	148	3	0	0	8	0	0	0	1	92.50
3	0	1	145	0	0	0	0	1	0	13	90.63
4	0	0	0	160	0	0	0	0	0	0	100.0
5	2	0	0	0	142	0	0	0	16	0	88.78
6	0	9	3	0	0	146	0	0	0	2	91.25
7	0	1	0	0	0	6	153	0	0	0	95.63
8	0	0	3	0	0	4	0	151	0	2	94.38
9	7	0	0	0	8	0	1	0	143	1	89.38
0	0	3	5	0	1	0	0	0	0	151	94.38

El reconocimiento promedio en este caso es de 92.48 %.

Sistema KLT

Técnica utilizada: Filtro de Chebyshev paso-bajas con $f_c = 4.5 \ \ kHz$.

7 Vectores de KLT de orden 18.

Cuantizadores vectoriales de orden 16, uno para cada segmento.

Segmentación lineal, 4 segmentos por palabra.

a) Conjunto de entrenamiento: El promedio de reconocimiento es 88.80 %.

Palabra	1	2	3	4	5	6	7	8	9	0	%
1	82	0	0	1	1	0	5	0	6	5	82.00
2	1	79	16	0	0	0	1	0	0	3	79.00
3	0	3	94	0	0	2	0	1	0	0	94.00
4	0	0	0	84	2	3	6	0	0	5	84.00
5	0	0	0	0	98	1	0	1	0	0	98.00
6	0	0	0	0	0	99	0	1	0	0	99.00
7	0	0	0	0	0	0	98	1	1	0	98.00
8	0	1	1	0	0	6	1	91	0	0	91.00
9	7	0	0	0	0	0	7	0	85	1	85.00
0	7	0	0	6	1	0	8	0	0	78	78.00

b) Conjunto de reconocimiento: El promedio de reconocimiento es 83.90 %.

Palabra	1	2	3	4	5	6	7	8	9	0	%
1	108	0	0	14	3	0	3	0	29	2	67.90
2	0	151	1	0	0	1	2	2	3	0	94.30
3	0	18	125	0	1	11	0	4	0	0	78.10
4	3	2	0	149	4	0	1	0	0	1	93.10
5	6	0	0	3	129	2	5	0	15	0	80.60
6	0	0	1	0	0	148	2	8	0	0	92.50
7	6	5	0	0	3	3	133	1	7	1	83.10
8	0	1	0	0	0	4	0	152	0	0	96.80
9	11	0	4	0	19	9	6	1	108	1	67.90
0	3	0	0	9	4	0	7	0	1	136	85.00

Segmentación Acústica, 4 segmentos por palabra.

a) Conjunto de entrenamiento: El promedio de reconocimiento es 97.00 %

Palabra	1	2	3	4	5	6	7	8	9	0	%
1	87	0	0	6	0	0	2	0	4	1	87.00
2	0	99	0	0	0	1	0	0	0	0	99.00
3	0	3	96	0	0	0	0	0	0	1	96.00
4	0	0	0	100	0	0	0	0	0	0	100.0
5	0	0	0	0	100	0	0	0	0	0	100.0
6	0	0	0	0	0	100	0	0	0	0	100.0
7	1	0	0	0	0	0	97	0	2	0	97.00
8	0	1	0	0	0	1	0	98	0	0	98.00
9	0	0	0	0	1	0	1	0	98	0	98.00
0	1	1	0	3	0	0	0	0	0	95	95.00

b) Conjunto de reconocimiento: El reconocimiento promedio en este caso es de 85.00 %

Palabra	1	2	3	4	5	6	7	8	9	0	%
1	106	0	0	14	6	0	11	0	20	2	66.25
2	0	153	3	0	0	1	2	0	1	0	95.62
3	0	27	122	0	0	5	0	5	0	0	76.25
4	0	1	0	155	2	0	2	0	0	0	96.87
5	3	0	0	2	127	3	11	0	14	0	79.37
6	0	1	1	0	0	149	3	5	0	0	93.13
7	0	4	1	0	2	4	141	0	6	0	88.13
8	0	1	0	0	0	2	0	154	0	0	96.25
9	4	1	2	2	10	6	12	1	120	1	75.00
0	3	2	1	11	2	0	7	0	2	132	82.50

CONCLUSIONES

De la comparación entre las técnicas LPC y KLT realizada en este trabajo, si se analizan los resultados obtenidos en el capítulo 4, se puede concluir lo siguiente:

Con ambas técnicas es posible obtener tasas de reconocimiento altas, principalmente para el caso en que el reconocimiento se realiza con el mismo conjunto que se utilizó para el entrenamiento, en el que los resultados para ambas técnicas fueron similares. Para este caso, el análisis KLT presenta una ventaja sobre el LPC, que es la reducción de la cantidad de información necesaria para caracterizar la señal.

Lógicamente el reconocimiento real se hace con señales distintas a las del conjunto de entrenamiento. Las pruebas realizadas con la base de datos para prueba, logran tasas, naturalmente, menores que para las pruebas con el conjunto de entrenamiento.

Si se compara la tabla de KLT con la de LPC, se podría pensar, que el método KLT tiene una desventaja con respecto a LPC por haber logrado una tasa más baja. Esto no es del todo cierto, debido a que LPC divide cada subpalabra (segmento) en bloques, obteniendo un conjunto de coeficientes LPC para cada bloque, mientras que KLT solo obtiene un conjunto de coeficientes por segmento. Es decir, KLT logra una tasa de reconocimiento comparable a LPC, con menos información.

Para ambas pruebas, como se menciono, se comparó la segmentación lineal con la segmentación acústica. Nótese que en el reconocimiento LPC la tasa de reconocimiento es mayor para el caso de segmentación lineal, mientras que en KLT la segmentación acústica logró una mejora significativa.

En resumen, el método presentado aquí, KLT-VQ con segmentación acústica, ha mostrado ser un método lo suficientemente exitoso para ser utilizado, en el futuro, en sistemas de reconocimiento de palabras aisladas.

BIBLIOGRAFIA

[1] Akansu, Haddad.
Multirresolution Signal Decomposition. Transforms, Subbands, Wavelets.
Academic Press, 1992.

[2] Herrera Abel, Algazi, V.R., Brown, K.L. y Irvine,D.
Subword Segmentation Altrenatives for Isolated and Connected Words Recognition.
Proceedings, VII European Signal Processing Conference, EUPSICO-94,
vol. I, 107-110.

[3] Herrera Abel, Algazi, V.R. y Irvine, D.
An Acoustic Approach for Isolated Word Recognition.
Proceedings, International Conference on Signal Proc. Applications & Technology,
vol. II, 1677-1681.

[4] Hew, Patrick.
Positive Semidefinite Quadratic Forms.
Department of Mathematics.
University of Western Australia, 1997.

[5] Hoogenboom, R.A.
The Karhunen-Loéve Transform.
http://www.wi.leidenuniv.nl/home/rhoogenb/node27.html
1996.

[6] Itakura, F.
Minimum prediction residual principle applied to speech recognition.
IEEE Trans. Acoustics, Speech, Signal Processing, ASSP-23: 57-72.
Febrero, 1975.

[7] Martinez, H.G., Rivera,C. y Buzo, A.
Discrete utterance recognition based upon score coding techniques.
Proceedings, IEEE International Conference on Acoustics, Speech, and Signal
Processing, 1982, 539-542.

[8] Oppenheim, Schafer
Discrete-Time Signal Processing.
Prentice-Hall, 1989.

[9] Rabiner, Hwang.
Fundamentals of Speech Recognition.
Prentice-Hall, 1993.

[10] Rabiner, L., Schafer R.
Digital Processing of speech Signals.
Prentice Hall, 1978.

[11] Rowden, C.
Speech Processing.
Mc Graw-Hill, London, 1992.

[12] Sánchez C., Oscar.
Segmentación Acústica de Subpalabras.
Tesis, Maestría en Ingeniería.
UNAM, México, 1997.

[13] H. L. van Trees.
Detection, Estimation and Modulation Theory.
Part 1: Detection, Estimation and Linear Modulation Theory.
John Wiley & Sons, 1968.