UNIVERSIDAD NACIONAL AUTONOMA
DE MEXICO

FACULTAD DE INGENIERIA

RECONOCIMIENTO DE VOZ POR COMPUTADORA

T E S I S

Que para obtener el Título de
INGENIERO EN COMPUTACION
P r e s e n t a

CARLOS RIVERA RIVERA

Ciudad Universitaria, D. F., Abril 1985

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

Introducción.

El avance de la ciencia y la tecnología, ha hecho posible la construcción de maquinas procesadoras de información como son las computadoras digitales. Estas han causado un gran impacto dentro de la vida del hombre. Una de las ramas es la del procesamiento de señales de diferentes tipos, tales como : señales biomédicas, señales geofísicas, etc. Dentro de éstas una que ha presentado un reto muy fuerte es el de las señales de voz, ya que dentro de esto existen problemas como : reconocimiento de voz por computadora, reconocimiento de personas por medio de su voz y una computadora, ayuda a personas con problemas auditivos, compresión de datos, etc.

El presente trabajo está dentro del reconocimiento de voz por computadora. Dado que éste es un problema complejo se ha tratado de resolver un subproblema como es el de reconocimiento de palabras aisladas para un número reducido de personas, un ejemplo de esto sería el que una computadora reconozca comandos pronunciados por un operario y tome alguna decisión.

Las técnicas que se utilizan en este trabajo son el método de predicción lineal y el de vocoder de canal, ésta última fue una de las primeras técnicas para procesar voz mientras que la anterior últimamente ha tenido un gran impacto. También se utiliza la cuantización vectorial, la cual es una de las técnicas más poderosas para comprimir datos. Como se pretende tener una medida cuantitativa, se hace uso de la medida de distorsión de Itakura-Saito en diferentes versiones. Finalmente se hace una comparación de 6 diferentes técnicas para la solución del problema.

MODELO DEL APARATO VOCAL.

El siguiente modelo del aparato vocal fue desarrollado por Fant[2].

La entrada al sistema e(t) es un tren de impulsos con periodo P para sonidos sonoros y ruido blanco para sonidos sordos, la parte correspondiente a la glotis es modelada por un filtro pasobajas de dos polos y con frecuencia de corte de aproximadamente 100 hz. El tracto vocal es modelado con resonadores de dos polos cada uno y conectados en forma de cascada, cada una de estas resonancias es un formante con su correspondiente frecuencia central y su ancho de banda. El efecto de los labios se representa por un modelo de radiación de un solo polo.

El modelo está dado por:

$$S(Z)=E(Z)G(Z)V(Z)L(Z)$$

donde :

$$E(Z) = \sigma \sum_{n=0}^{\infty} Z^{-n} = \sigma/(1-Z^{-1}) \qquad |Z| > 1$$

$$G(Z) = 1/(1 - e^{-cT}Z^{-1})^2$$

$$L(Z) = 1 - Z^{-1}$$

El modelo del tracto esta dado por:

$$V(Z) = 1/\prod_{i=1}^{k} (1 - 2e^{-c_iT}\cos(biT)Z^{-1} + e^{-2c_iT}Z^{-2})$$

El modelo así descrito está definido para que la entrada sea un tren de pulsos o ruido y un número fijo de formantes y ancho de banda. Por lo que únicamente el estado estable de vocales o sonidos fricativos están definidos. Sin embargo es facil implementar una entrada arbitraria e(n) y los coeficientes de V(Z) se cambian en intervalos de tiempo para representar la señal de voz que es variable con el tiempo.

Combinando los efectos de los modelos anteriores se tiene :

$$G(Z)V(Z)L(Z) = (1 - Z^{-1})/(1 - e^{-cT}Z^{-1})\{\prod_{i=1}^{k} (1-2e^{-c_iT}\cos(biT)Z^{-1} + e^{-2c_iT}Z^{-2})\}$$

El término del numerador (1-Z) se cancela aproximadamente con el término [1-exp(-cT)Z] ya que por lo general cT es mucho menor que la unidad. Con esta simplificación se tiene el modelo de síntesis

$$S(Z)=E(Z)/A(Z) \qquad \text{[modelo de síntesis]} \qquad 1.1$$

donde

$$A(Z) = \sum_{i=0}^{M} a(i)Z^{-i} \qquad a(0) = 1$$

$$\approx 1/G(Z)V(Z)L(Z)$$

y $M \geq 2k+1$

El filtro A(Z) tiene únicamente ceros, se conoce como el filtro inverso, el filtro 1/A(Z) tiene polos unicamente.

Despejando E(Z) de la ecuación 1.1 se tiene el modelo de análisis

$$E(Z)=S(Z)A(Z) \qquad \text{[modelo de analisis]} \qquad 1.2$$

El disponer de un modelo parametrico de la señal de voz, nos permite la estimación de los parámetros del polinomio A(Z) y el tipo de excitación E(Z) en base a la observación de la señal S(Z), lo cual es el tema central del siguiente capítulo.

PREDICCION LINEAL

En el modelo de predicción lineal la señal s(n), se considera la salida de un sistema con una entrada u(n) y de tal forma que la siguiente relación se cumpla:

$$s(n) = -\sum_{k=1}^{p} a(k)s(n-k) + G\sum_{l=0}^{q} b(l)u(n-l) \qquad b(0)=1$$

donde a(k), $1 \leq k \leq p$, b(l), $1 \leq l \leq q$, y la ganancia G son los parámetros del sistema.

La ecuacion anterior relaciona la salida s(n) como una función lineal de salidas anteriores más entradas anteriores y actual, de ésta forma se puede considerar que se está realizando una predicción sobre la señal s(n), a esto se debe el nombre del modelo.

Tomando la transformada Z de ambos lados de la ecuación y ordenándola en una forma adecuada se obtiene:

$$H(Z) = S(Z)/U(Z) = G(1 + \sum_{l=1}^{q} b(l)Z^{-l})/(1 + \sum_{k=1}^{p} a(k)Z^{-k})$$

donde H(Z) es la función de transferencia del sistema, S(Z) y U(Z) las transformadas Z de la señal s(n) y u(n) respectivamente.

H(Z) es el modelo general, ya que posee tanto polos como ceros, donde los polos son las raices del polinomio del numerador y los ceros las del denominador.

Existen dos casos de interés especial :

1) Modelo de ceros : $a(k) = 0, \quad 1 \leq k \leq p$
2) Modelo de polos : $b(l) = 0, \quad 1 \leq l \leq q$

El modelo de ceros es conocido como movimiento promedio, y el modelo de polos como autorregresivo.

Este último modelo es el más empleado, debido a que el problema de encontrar los coeficientes a(k), conocidas la entrada y la salida del sistema (identificación del sistema), conduce a un sistema algebraico de ecuaciones lineales. El modelo utilizado en éste reporte es el de polos.

Estimación de parámetros.

En el modelo de polos la señal s(n) está dada como una combinación lineal de valores anteriores y una entrada u(n):

$$s(n) = -\sum_{k=1}^{p} a(k)s(n-k) + Gu(n)$$

La función de transferencia del modelo de polos está dada por:

$$H(Z) = G/(1 + \sum_{k=1}^{p} a(k)Z)$$

Dada una señal s(n), el problema es determinar los coeficientes a(k) y la ganancia G, tal que sean óptimos en algún criterio.

Método de mínimos cuadrados.

Se asume que la entrada u(n) es desconocida, lo cual es cierto en muchas aplicaciones, con esto se tiene que la señal s(n) es una combinación lineal de valores anteriores de s(n) y está dada por :

$$\hat{s}(n) = -\sum_{k=1}^{p} a(k)s(n-k)$$

El error entre este valor y el real está dado por

$$e(n) = s(n)-\hat{s}(n) = s(n) + \sum_{k=1}^{p} a(k)s(n-k)$$

En el método de mínimos cuadrados los coeficientes a(k) se obtienen al minimizar el valor medio cuadrático del error, con respecto a cada uno de los parámetros a(k). El análisis puede ser realizado, suponiendo a la señal s(n), como determinística o como una realización de un proceso aleatorio.

a) Señal determinística.

El valor medio cuadrático está dado por :

$$E = \sum_{n} e^2(n) = \sum_{n} \left(s(n) + \sum_{k=1}^{p} a(k)s(n-k)\right)^2 \qquad 2.1$$

minimizando E se tiene

$$\partial E / \partial a(i) = 0 \qquad 1 \leq i \leq p$$

de donde se obtiene

$$\sum_{k=1}^{p} a(k) \sum_{n} s(n-k)s(n-i) = -\sum_{n} s(n)s(n-i) \qquad 1 \leq i \leq p \qquad 2.2$$

a este sistema se le conoce como ecuaciones normales.

El error mínimo se obtiene al desarrollar las ecuaciones 2.1 y substituir el resultado de la ecuación 2.2

$$Ep = \sum_{n} s(n)^2 + \sum_{k=1}^{p} a(k) \sum_{n} s(n)s(n-k)$$

Durante todo el desarrollo anterior el rango de la sumatoria no se especificó, existen dos casos de interés, los que conducen a métodos distintos para el cálculo de los parámetros.

i) Método de la autocorrelación.

Este método minimiza el error en el rango de duración infinita $-\infty < n < \infty$, y las ecuaciones 2.1 y 2.2 se reducen a

$$\sum_{k=1}^{p} a(k)R(i-k) = -R(i) \qquad 1 \leq i \leq p \qquad 2.3$$

$$Ep = R(0) + \sum_{k=1}^{p} a(k)R(k) \qquad 2.4$$

donde

$$R(i) = \sum_{n=-\infty}^{\infty} s(n)s(n+i)$$

es la autocorrelacion de la señal s(n).

Los coeficientes R(i-k) de la ecuación 2.3 forman una matriz conocida como de autocorrelación, la cual es una matriz simétrica Toeplitz, cuya característica es que los elementos de sus diagonales son iguales.

En la práctica, la señal s(n) se conoce sólo durante un intervalo finito, una forma de resolver esto es el multiplicar la señal s(n) por una ventana con lo cual se obtiene una nueva señal s'(n) que es igual a cero fuera de un intervalo $0 \leq n \leq N-1$.

$$s'(n) = \begin{cases} s(n)w(n) & 0 \leq n \leq N-1 \\ 0 & \text{para cualquier otro valor de n} \end{cases}$$

Con esta señal la función de autocorrelación está dada por :

$$R(i) = \sum_{n=0}^{N-1-i} s'(n)s'(n+i) \qquad i \geq 0$$

ii) Método de la covariancia.

En este método el error E se minimiza sobre un intervalo finito que va desde 0 hasta N-1, las ecuaciones 2.1 y 2.2 quedan como

$$\sum_{k=1}^{p} a(k)Q(k,i) = -Q(0,i) \qquad 1 \leq i \leq p$$

$$Ep = Q(0,0) + \sum_{k=1}^{p} a(k)Q(0,k)$$

donde

$$Q(i,k) = \sum_{n=0}^{N-1} s(n-i)s(n-k)$$

es la covariancia de la señal, en el intervalo dado. Con la primer ecuación se forma una matriz de covariancias.

De esta última ecuación se nota que la matriz de covariancia es siméntrica, esto es :

$$Q(i,k) = Q(k,i)$$

y a diferencia de la matriz de autocorrelación los elementos de las diagonales no son iguales. Esto puede observarse al escribir

$$Q(i+1,k+1) = Q(i,k) + s(-i-1)s(-k-1) - s(N-1-i)s(N-i-k)$$

De esta ecuación se nota que los valores de s(n) desde $-p \leq n \leq N-1$, deben de ser conocidos, lo que son N+p muestras en total. El método de covariancia se convierte en el de autocorrelación cuando n tiende a infinito.

Señal aleatoria.

Un resultado importante es cuando s(n) se considera como una muestra de un proceso aleatorio, con lo cual las ecuaciones 2.1 y 2.2 , tomando el valor esperado quedan como :

$$E = E\{ e(n)^2 \} = E\{ s(n) + \sum_{k=1}^{p} a(k)s(n-k) \}^2 \qquad 2.5$$

$$a(k)E\{ s(n-k)s(n-i) \} = - E\{ s(n)s(n-i) \} \quad i \leq i \leq p \qquad 2.6$$

El resultado del valor esperado depende si el proceso es estacionario o no.

Caso estacionario.

Para procesos estacionarios se tiene que

$$E\{ s(n-k)s(n-i) \} = R(i-k)$$

donde R(i) es la autocorrelacion del proceso, substituyendo ésto en las ecuaciones 2.5 y 2.6 se observa que se tienen las mismas ecuaciones que para el caso determinístico. La única diferencia es que para éste caso la autocorrelación es la de un proceso estacionario en lugar de la autocorrelación de una señal determinística.

Caso no estacionario.

Para un proceso no estacionario s(n), se tiene

$$E\{ s(n-k)s(n-i) \} = R(n-k,n-i)$$

donde R(m,j) es la autocorrelación no estacionaria entre los tiempos m y j, de las ecuaciones 2.1 y 2.2 se obtiene

$$\sum_{k=1}^{p} a(k)R(n-k,n-i) = -R(n,n-i) \qquad 1 \leq i \leq p$$

$$Ep = R(0,0) + \sum_{k=1}^{p} a(k)R(0,k)$$

Cálculo de la ganancia G.

En el método de mínimos cuadrados se supuso que el error estaba dado por:

$$e(n) = s(n) + \sum_{k=1}^{p} a(k)s(n-k)$$ lo cual se puede escribir de la forma

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

como el modelo de polos está dado por

$$s(n) = - \sum_{k=1}^{p} a(k)s(n-k) + Gu(n)$$

se nota que la única señal u(n) que daría como salida de s está dada por

$Gu(n) = e(n)$

Para cualquier otra entrada la salida del filtro H(Z) será diferente de s(n).

Si se quiere que la energía de la salida iguale a la de la señal original s(n), sin importar la entrada, al menos se puede especificar la energía total en la señal de entrada. Como el filtro H(Z) es fijo, la energía u(n) debe ser igual a la energía total del error, la cual está dada por Ep.

Existen dos entradas que son de interés, el impulso determinístico y el ruido blanco.

La ganancia de la entrada se determina al hacer el análisis de la respuesta del sistema.

a) Entrada igual a un impulso.

Con la entrada siendo un impulso se tiene

$$h(n) = -\sum_{k=1}^{p} a(k)h(n-k) + G\delta(n)$$

donde h(n) es la respuesta impulso del sistema, multiplicando esta ecuación por h(n-i) y sumando sobre toda n

$$\hat{R}(i) = -\sum_{k=1}^{p} a(k)\hat{R}(i-k) \qquad 1 \leq |i| \leq \infty \qquad 2.7$$

$$\hat{R}(0) = -\sum_{k=1}^{p} a(k)\hat{R}(k) + G^2 \qquad 2.8$$

Como la condición de que h(n) debe igualar a la energía de s(n) se tiene

$$\hat{R}(0) = R(0)$$

puesto que el primer coeficiente de autocorrelación es igual a la energía total de la señal. De este resultado y notando la similitud entre 2.7 y 2.3 se tiene

$$\hat{R}(i) = R(i) \qquad 0 \leq i \leq p.$$

Esto nos dice, que los primeros p+1 coeficientes de autocorrelación de h(t) son iguales a los coeficientes de autocorrelación de la senal.

Otra forma de ver el problema de predicción lineal es el de buscar un filtro de tal forma que los primeros p+1 coeficientes de correlación de su respuesta impulso sean iguales a los p+1 coeficientes de correlación de la señal y que la ecuación 2.7 se cumpla. De la ecuación 2.8 se obtiene

$$G^2 = Ep = R(0) + \sum_{k=1}^{p} a(k)R(k)$$

Donde G^2 es la energía total de la entrada $G\delta(n)$.

b) Entrada igual a ruido blanco.

La señal de entrada es ruido blanco (muestras no correlacionadas) con media cero y variancia unitaria.

La señal de salida del sistema está dada por

$$\hat{s}(n) = -\sum_{k=1}^{p} a(k)\hat{s}(n-k) + Gu(n)$$

multiplicando por s(n-i), tomando el valor esperado y observando que u(n) y s(n-i) no están correlacionadas para i>0, con lo cual se llegan a las mismas ecuaciones de la entrada impulso, con $R(i)=E(\hat{s}(n)\hat{s}(n-i))$ la autocorrelación de la salida $\hat{s}(n)$.

Para éste caso se requiere que la energía promedio(variancia) de la salida $\hat{s}(n)$ sea igual a la variancia de la señal original s(n) o $\hat{R}(0)=R(0)$ y con el mismo razonamiento de la entrada impulso se llega al mismo resultado.

Este resultado es interesante ya que ambas entradas tienen la misma autocorrelación.

Cálculo de los parámetros del predictor.

En los resultados anteriores los coeficientes a(k) se pueden calcular al resolver el conjunto de p ecuaciones con p incógnitas, si se utiliza el método de Gauss o el de reducción de Crout. Estos métodos requieren de $p/3+O(p)$, operaciones y p localidades de memoria. Sin embargo como la matriz de el sistema es una matriz simétrica y en general positiva semidefinida, un método para resolver este tipo de sistemas requiere, 2p localidades de memoria y $p+O(p)$ operaciones, lo que representa un ahorro con respecto a los métodos generales.

La ecuación 2.3 puede ser escrita en la forma :

$$\underset{\sim}{r} = \underset{\sim}{R}\underset{\sim}{a}$$

donde

$$\underset{\sim}{r} = \begin{bmatrix} R_1 \\ R_2 \\ \cdot \\ \cdot \\ R_p \end{bmatrix} \qquad \underset{\sim}{a} = \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ a_p \end{bmatrix} \qquad \underset{\sim}{R} = \begin{bmatrix} R_0 & R_1 & \cdot & \cdot & \cdot & R_{p-1} \\ R_1 & R_0 & \cdot & \cdot & \cdot & R_{p-2} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ R_{p-1} & \cdot & \cdot & \cdot & \cdot & R_0 \end{bmatrix}$$

Se observa que la matriz R es simétrica y los elementos de sus diagonales son iguales, este tipo de matriz es llamada Toeplitz.

La solución de este sistema se pueden expresar en el siguiente algoritmo debido a Durbin [3].

$$E(0) = R(0)$$

$$K(i) = -[R(i) + \sum_{j=1}^{i-1} a(j,i-1)R(i-j)]/E(i-1)$$

$$a(i,i) = K(i)$$

$$a(i,j) = a(j,i-1) + K(i)a(i-j,i-1) \qquad 1 \leq j \leq i-1$$

$$E(i) = (1-K(i)^2)E(i-1)$$

Una observación interesante es que para resolver este sistema, la mayoría de las operaciones se deben al cálculo de las autocorrelaciones.

Del algoritmo se tiene que el error total E(i) a cada paso es menor o igual al paso anterior. Después de resolver el sistema es importante el saber si el filtro resultante es estable. Un filtro de este tipo es estable si sus polos están dentro del círculo unitario. Si los coeficientes R(i) son positivos definidos (lo cual se logra si se calcularon a partir de una señal diferente de cero) la solución del sistema da un filtro con sus polos dentro del círculo unitario.

La condición de que los coeficientes sean positivos definidos se puede perder si la palabra de la computadora con la

que se representa R(i) es pequeña, los errores debido al redondeo también afectan a la estabilidad del sistema. Para verificar que el filtro resultante es estable, se pueden calcular sus raices y ver si están dentro del círculo unitario, lo cual puede ser costoso. Otro método es ver que el error a cada paso sea positivo, ésta condición es equivalente a

$$| K(i) | < 1 \qquad 1 \leq i \leq p$$

ésto se puede ir verificando al mismo tiempo que se va resolviendo el sistema.

Estimación Espectral.

Los métodos anteriores fueron formulados en el dominio del tiempo, es importante formular el problema ahora en el dominio de la frecuencia. para conocer sus características y entenderlo mejor.

Caso estacionario.

Aplicando la transformada Z al error se tiene

$$E(Z) = [\, 1 + \sum_{k=1}^{P} a(k) Z^{-k}) S(Z) = A(Z) S(Z)$$

de aquí se observa que el error se puede obtener como el resultado de pasar la señal S(Z) por un filtro F(Z). Asumiendo la señal determinística y aplicando el teorema de Parseval, el error total a minimizar es

$$E = \sum_{n=-\infty}^{\infty} e(n)^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} | E(e^{j\omega}) |^2 dw$$

el espectro en potencia de la señal s(n) es

$$F(w) = | S(e^{jw}) |^2$$

al substituir esto en la ecuación anterior

$$E = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(w) A(e^{j\omega}) A(e^{-j\omega}) \, dw$$

Calculando $\frac{\partial E}{\partial a_i}$ e igualando a cero, se minimiza E.
El resultado de ésto es igual a el de las ecuaciones normales, pero con R(i) dado por

$$R(i) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(w) \cos(iw) \, dw$$

El error mínimo se obtiene al substituir estos resultados en la ecuación del error E, lo cual da

$$Ep = R(0) + \sum_{k=1}^{P} a(k) R(k)$$

que es el mismo resultado obtenido para las ecuaciones normales.

Aproximación del espectro.

El espectro en potencia del modelo está dado por

$$\hat{F}(w) = | H(e^{j\omega}) |^2 = G^2 / | A(e^{j\omega}) |^2$$

$$= G^2 / | 1 + \sum_{k=1}^{P} a(k) e^{-j\omega k} |^2$$

El espectro en potencia de la señal es

$$F(w) = \frac{|\ E(e^{jw})\ |^2}{|\ A(e^{jw})\ |^2}$$

Comparando las ecuaciones anteriores, se ve que si F(w) va a ser modelada por F(w), el espectro en potencia del error se modela por un espectro plano igual a . Esto significa que el error e(n) se aproxima por otra señal con espectro plano, como un impulso, ruido blanco o cualquier otra señal con este espectro.

Con las ecuaciones anteriores el error total es

$$E = \frac{G^2}{2\pi} \int_{-\pi}^{\pi} F(w) / \hat{F}(w) \quad dw$$

El minimizar el error es equivalente a minimizar la integral de $F(w)/\hat{F}(w)$.

El problema de predicción lineal se puede replantear como:

Dado un espectro F(w), se desea modelar por otro espectro $\hat{F}(w)$ tal que la integral de $F(w)/\hat{F}(w)$ se minimize.

La ganancia G se obtiene igualando las energías de los dos espectros, $\hat{R}(0)=R(0)$

donde

$$\hat{R}(i) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{F}(w) \cos(iw)\ dw$$

La forma en la cual el espectro del modelo $\hat{F}(w)$ aproxima F(w) se refleja en la relación entre las funciones de autocorrelación.

Como se tiene $\hat{R}(i) = R(i)$, $0 \leq i \leq p$ y F(w), $\hat{F}(w)$ son las transformadas de R(i), $\hat{R}(i)$ respectivamente, si el orden del modelo se incrementa, el rango donde las correlaciones son iguales se incrementa también y en el límite los dos espectros son idénticos.

Esto significa que se puede aproximar cualquier espectro tanto como se desee, por un filtro con polos únicamente.

Puesto que el análisis de predicción lineal se puede plantear como el de aproximar un espectro o autocorrelación, es importante la forma en que se calcula el espectro de la señal F(w) o su autocorrelación.

Como la mayoría de las veces la señal es pasada a través de una ventana antes de calcular las correlaciones, es importante el escoger una ventana apropiada. En este reporte la ventana que se empleó fue una de Hamming que está dada por :

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

N es el tamaño de la ventana.

se tiene entonces

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} E(w)\, dw = 1 \qquad \text{para toda } p$$

E(w) se puede interpretar como el error instantaneo entre F(w) y $\hat{F}(w)$ a la frecuencia w, también se tiene que la media aritmética de E(w) es 1, lo cual significa que existen valores mayores y menores a 1 de tal forma que su media sea 1. En términos de los dos espectros, esto representa que F(w) será algunas veces mayor y otras menor que F(w). Sin embargo la contribución al error total es mas significativa cuando F(w) es mayor que F(w) que cuando F(w) es menor a F(w), por ejemplo un radio de E(w) = 2 contribuye más al error total que un radio de 1/2.

Lo anterior conduce a la conclusión de que el modelo de predicción lineal tiende a igualar los picos o las regiones de energía relativamente grande, mejor que los valles o regiones de energía baja.

La minimización del error E da como resultado un espectro F(w) que es un buen estimado de la envolvente del espectro F(w).

Un problema equivalente es el de encontrar las a(i), $1 \leq i \leq p$ y , G tal que se minimize:

$$d(F(w),G/A(w)) = \int_{-\pi}^{\pi} \frac{dw}{2\pi}\left[\frac{F(w)}{G\ /|A(w)|} - \frac{\ln(\ F(w)\ )}{G\ /|A(w)|} - 1 \right]$$

dado que

$$\hat{F}(w) = \frac{G}{|\ A(w)\ |}$$

la expresión anterior se puede escribir como:

$$d(F(w),\hat{F}(w)) = \int_{-\pi}^{\pi} [\ (F(w)/\hat{F}(w)) - \ln(\ F(w)/\hat{F}(w)\ ) - 1\ ]dw/2\pi$$

al resolver este problema se llega al mismo conjunto de ecuaciones de predicción lineal.

La función $d(F(w).\hat{F}(w))$ recibe el nombre de medida de distorsión de Itakura y Saito, puede ser interpretada como la fidelidad con que la señal con espectro F(w) es representada por la señal con espectro F(w).

VOCODERS DE CANAL

Una manera clásica de obtener una representación aproximada de la función de transferencia del sistema vocal es por medio del uso de la transformada de Fourier. Uno de los primeros sistemas que hicieron uso de esto fue el vocoder de canal , el cual encuentra una aproximación de las diferentes componentes en frecuencia de la voz. Esto lo hace por medio de un banco de filtros en los cuales la energía de salida de cada uno de ellos representa la amplitud de la componente en frecuencia en la banda correspondiente al filtro. La estructura básica es la mostrada en la figura.

El vocoder se puede modelar de la siguiente manera;

sea $F(e^{jw})$ el espectro de la señal de voz, se tiene entonces que :

$$\hat{F}(e^{jw}) = \sum_{k=1}^{P} Tk\ \Omega_k$$

$$Tk = \int_{-\pi}^{\pi} F(e^{jw})\,\Omega_k\, dw$$

$$\Omega_k = \begin{cases} 1 & w_k \leq w < w_{k+1} \\ 0 & \text{otra } w \end{cases}$$

Se puede observar que las Tk representan la energía en la banda $[\ w_k\ ,\ w_{k+1}\ ]$, por lo general estas frecuencias están especificadas y en muchos casos uniformemente espaciadas, por lo que los únicos parámetros por especificar son los valores de las Tk. Para el cálculo de estas Tk se ha recurrido al uso de la transformada rápida de Fourier.

Medidas de Distorsión para el procesamiento de Voz.

Para realizar una comparación entre la calidad de sistemas y poder hacer mejoras, es necesario tener un criterio de fidelidad o una medida de distorsión. Esto es, se necesita asignar una distorsión o un costo de reproducir un sonido por una reproducción particular.

En matemáticas cuando se quiere encontrar la distancia entre dos puntos, se debe definir una métrica que debe cumplir con las siguientes propiedades:

1-a)	$d(f,g) \geq 0$	(no negatividad)
1-b)	$d(f,g) = d(g,f)$	(simetría)
1-c)	$d(f,g) = 0$	(significa f=g)
1-d)	$d(f,g) \leq d(f,h) + d(h,g)$	(desigualdad del triángulo)

Estas métricas tienen la propiedad de simetría y deben cumplir con la desigualdad del triángulo. Aún cuando éstas dos propiedades son deseables, en teoría de la información se ha desarrollado la medida de distorsión la cual impone menos restricciones.

Una de la razones por las cuales las propiedades anteriores no se exigen, es que en el contexto del problema el original y la reproducción se tienen bien especificados.

Para ser útil una medida de distorsión debe poseer en algún grado los siguientes atributos :

2-a) Debe ser subjetivamente significativa, esto es grandes(pequeñas) distorsiones deben corresponder a mala(buena) calidad subjetiva.

2-b) Debe ser matemáticamente tratable de tal forma que permita un análisis teórico.

2-c) Debe ser posible calcularla.

Una medida de distorsión tradicional es la del error cuadrático, esto debido a su facilidad de cálculo. Sin embargo, para sistemas de voz e imágenes esa medida de distorsión no es significativa ya que distorsiones grandes en este sentido no implican mala calidad. Debido a esto se han desarrollado medidas de distorsión más significativas que el error cuadrado.

Algo importante es el definir la medida de distorsión en el dominio de la frecuencia, del tiempo a algún otro dominio. Tradicionalmente en el estudio de la voz, se ha trabajado en la frecuencia, y la mayoría de las medidas de distorsión se han estudiado en este dominio. Si bien estas medidas pueden ser transformadas al dominio del tiempo, sus características principales están dadas en la frecuencia.

Medidas de distorsión espectral.

Una medida de distorsión espectral, es función de dos densidades espectrales, f y g, a la cual le asigna un número no negativo, d(f,g) que es la distorsión de representar f por g.

La mayoría de éstas distorsiones utilizan una norma Lp sobre la diferencia f-g. Estas son métricas ya que cumplen con la

simetría y la desigualdad del triánoulo.

Otras medidas interesantes dependen sobre la diferencia en el logaritmo del espectro, que equivale a el logaritmo de la división de los espectros.

$$d(f,g)=d(1,g/f) = d(f/g,1) \qquad 3.1$$

Es importante el conocer las diferencias entre las medidas de distorsión, intuitivamente dos medidas de distorsión son equivalentes si los resultados obtenidos no varian significatívamente.

Para dos medidas de distorsión, d1 y d2, se dice que d1 es mas fuerte que d2, si una pequeña d1 implica una pequeña d2, se exoresa como d1 => d2.

Matemáticamente para cualquier e > 0 existe un delta = delta(e) tal que si d1 < delta entonces d2 < e. Si d1 => d2 y d2 => d1 se dice que d1 y d2 son equivalentes y se escribe d1 <=> d2.

Otra noción de equivalencia es el llamado vecino más cercano, dos medidas de distorsión d1 y d2 son equivalentes en el vecino más cercano si son minimizadas por la misma densidad espectral f.

Si dos medidas de distorsión son equivalentes en los sentidos mencionados anteriormente se dice que son completamente equivalentes.

Muchas veces es conveniente tener una medida de distorsión normalizada en ganancia, ya que de ésta forma se pueden considerar por separado los efectos debidos a la ganancia y al modelo mismo.

Esta medida de distorsión está definida como :

$$d^*(f,g) = d(f/Gf,g/Gg)$$

donde Gf y Gg son las ganancias de los modelos definidos en el capítulo anterior.

Medidas de distorsión para el espectro de voz.

Una de las primeras medidas de distorsión que se utilizaron para la señal de voz, es la norma Lp de la diferencia del logaritmo de los espectros.

$$dlnp(f,g) = \| \ln(f) - \ln(g) \|p = \| \ln(f/g) \|p$$

donde

$$\| f(x) \|p = \left[\int_{-\pi}^{\pi} f(x)^p \, dx \right]^{1/p}$$

Los valores más comunes de p son 1,2 y ∞ los cuales dan : media absoluta, raiz media cuadrática y desviación máxima respectivamente.

Una propiedad de esta norma es:

$$\| f \|p < \| f \|q \qquad \text{si } 0 < p < q$$

por lo que

$$dln_\infty \geq dln2 \geq dln1$$

Debido a su facilidad de manejo matemático, la norma L2 es la más común, una de sus ventajas es que es una métrica verdadera.

Distorsión de Itakura-Saito.

Una medida de distorsión propuesta por Itakura y Saito [ ] es

$$dis(f,g) = \| (f/g) - \ln(f/g) - 1 \|_1$$

como

$$u - \ln(u) - 1 \geq 0$$

Esta distancia se puede expresar como

$$dis(f,g) = \int_{-\pi}^{\pi} (f/g) \frac{d\theta}{2\pi} - \ln(Gf/Gg) - 1$$

donde

Gf y Gg son las ganancias de los modelos de predicción lineal definidos anteriormente.

Haciendo uso de la desigualdad de Jensen (la media aritmetica es siempre mayor que la media geometrica), se tiene

$$\int_{-\pi}^{\pi} f/g \frac{d\theta}{2\pi} \geq \exp\left(\int_{-\pi}^{\pi} \ln(f/g) \frac{d\theta}{2\pi} \right) = Gf/Gg$$

de donde se obtiene

$$dis(f,g) \geq dis(Gf/Gg)$$

Esto significa que para unas ganancias espectrales dadas, un espectro constante da la distorsión mínima.

Del modelo del capítulo anterior

$$g(\theta) = Gg/| A(e^{j\theta}) |^2$$

La distancia de Itakura-Saito entre f y esta g está dada por

$$dis(f, Gg/| A(e^{j\theta}) |^2) = Tm(\underset{\sim}{a})/Gg - \ln(Gf/Gg) - 1$$

donde $Tm(\underset{\sim}{a})$ es la forma Toeplitz que está dada por:

$$Tm(\underset{\sim}{a}) = \int_{-\pi}^{\pi} \left| \sum_{k=0}^{n} a(k) e^{-jk\theta} \right| f(\theta) \frac{d(\theta)}{2\pi}$$

$$= \sum_{k=0}^{n} \sum_{l=0}^{n} a(k)a(l)r(k-l) = \underset{\sim}{a}' Rn(f) \underset{\sim}{a}$$

Donde el vector $\underset{\sim}{a}$ está formado por los coeficientes del polinomio A(Z) (el cual se definio en el capítulo anterior), y la matriz Rn(f) es la matriz de autocorrelación de la señal.

El vector $\underset{\sim}{a}$ y la ganancia Gg, se deben escoger de tal forma que minimizen la distorsión anterior. Esto significa minimizar $Tm(\underset{\sim}{a})$ que de el valor mínimo de Gf(m), que es equivalente al problema de predicción lineal.

Debido a la equivalencia anterior, se arguye que es una medida significativa para la distorsión de señales de voz.

Distorsión de Itakura.

Itakura [ ] propuso una distorsión definida como :

$$di(f,g) = d'is(f,g) = \min_{\lambda \geq 0} dis(f,\lambda g)$$

Desarrollando esto se obtiene :

$$di(f,g) = \ln\left(\int_{-\pi}^{\pi} \frac{f/Gf}{g/Gg} \frac{d\theta}{2\pi} \right)$$

di y dis están relacionadas por :

$$dis(f,g) = (Gf/Gg)\exp(di(f,g)) - \ln (Gf/Gg) - 1$$

La distorsión de Itakura se puede expresar en función de los polinomios A(Z) y $\hat{A}(Z)$:

$$d(f,g) = \ln\left(\int_{-\pi}^{\pi} | A/\hat{A} |^2 \frac{d\theta}{2\pi} \right) = \ln(\| A/\hat{A} \|_2^2)$$

Distorsión de Itakura-Saito discreta

Esta distorsión está dada por :

$$dz(f,g) = Tk/Sk - \ln(tk/Sk) - 1$$

donde f y g son filtros de la forma:

$$f(w) = \sum_{k=1}^{p} Tk\Omega k$$ o sea del tipo vocoder de canal.

Medidas de distorsión simetrizadas.

Es importante notar que las medidas de distorsión anteriores no son simétricas, esto es $d(f,g) \neq d(g,f)$, aunque para ciertos casos se tiene bien definida lo que sería el original y la reproducción, para otros se quiere tener una medida simétrica. Una forma de lograr esto es tomando la media de la medida de distorsión

$$d_{SM}^{(q)}(f,g) = \frac{1}{2} (d(f,g)^q + d(g,f)^q)^{1/q}$$

Se puede notar que :

$$d_{SM}^{(q)}(f,g) \geq \frac{1}{2} d(f,g) \text{ y}$$

$$d_{SM}^{(q)}(f,g) \Rightarrow d(f,g)$$

Las medidas de distorsión anteriores, no son las únicas que se pueden aplicar para analizar la señal de voz, su uso es debido a que se han obtenido buenos resultados con ellas [ ]. La justificación podría ser que actuaran en la misma forma que el cerebro, pero por desgracia la forma en que este procesa la voz no es conocida. El que tan buena o mala sería una de las medidas de distorsión, se hace por medios cualitativos (el que tan bien o mal

un sonido, le parezca a un conjunto de personas).

También se podría realizar por medios cuantitativos, como es el caso de las distorsiones, pero el hecho de que un sistema tenga menos distorsión con respecto a las medidas usadas no implica que sea mejor, para el objetivo final.

Dos de las distorsiones que han dado mejores resultados son las de Itakura-Saito y la de Itakura [9,14], las cuales fueron empleadas en este trabajo.

Algoritmos de Agrupamiento.

Un problema que se ha presentado en el procesamiento y transmisión de senales digitales (y tambien analogicas) es que el medio de comunicacion, tiene un ancho de banda limitado, esto quiere decir que existe un límite de la capacidad del canal, para transmitir información, por lo que es importante que la señal que va a ser transmitida tenga el mínimo de redundancia posible, o que de alguna forma en lugar de la senal original se transmita una reproducción que si bien va a tener algún grado de distorsión, no sea importante.

El problema anterior se puede formular como el de comprimir una senal o un conjunto de datos, esto es, dado una serie de puntos de longitud M, encontrar los N mejores que la reproduzcan.

Para el caso de la transmisión de señales de voz se tendría que los puntos serian segmentos de voz, los cuales se compararian con referencias(que de alguna forma se encontraron) de las cuales se transmitiría la que diera la distorsión mínima en algún criterio.

En este problema se tendrian dos puntos importantes, uno es: Que criterio se tomaría para la distorsión mínima?, el otro punto seria: En que forma se encontrarian las referencias?.

Para el primer punto se aplican las distancias vistas en el capítulo anterior, el segundo punto entra en el campo de reconocimiento de patrones.

En esta área se han desarrollado algoritmos para encontrar estas referencias, estos algoritmos están basados en la clasificación de los patrones por medio de funciones de distancia.

La razón principal para el úso de funciones de distancia en la clasificación, se debe al hecho de que la manera obvia de establecer una medida de similitud entre elementos, es determinando su proximidad.

Los dos primeros algoritmos presentados a continuación son básicamente procedimientos intuitivos, los siguientes tres están basados en un criterio de minimización.

Algoritmos para agrupamientos.

Algoritmo del encadenamiento.

El primer paso de este algoritmo es reordenar los datos. Se designa un elemento arbitrariamente, como el principio de la cadena, designado por Xs. El siguiente elemento es el vecino más cercano a Xs y se le nombra Xs[1]. En general, el (k+1) elemento de la lista es Xs[k] este proceso continúa hasta tener una lista como la siguiente :

$$Xi[0] \leqslant Xi[1] \leqslant Xi[2] \leqslant \ldots \leqslant Xi[N-1] \; ; \quad i[0] = s$$

Al k'esimo elemento se le ha asociado con una distancia d(k) :

$$d(k) = d(Xi[k-1], Xi[k]) \qquad 1 \leqslant k \leqslant N-1$$

El algoritmo del encadenamiento es una gráfica de d(k) y k.

La característica principal de esta gráfica, son los picos que se presentan en las fronteras de los agrupamientos. Este algoritmo es sensitivo al punto de inicio Xs pero es simple, por lo que se pueden tomar varios puntos de inicio.

Vecinos cercanos compartidos.

Este algoritmo se basa en que dos elementos que tienen en común al menos k vecinos más cercanos, pertenecen al mismo agrupamiento. Para realizar esto, se forma una lista como la siguiente :

$$L = \begin{bmatrix} X(1) & X(1,1) & X(1,2) & \ldots & X(1,k) \\ X(2) & X(2,1) & X(2,2) & \ldots & X(2,k) \\ \vdots & & & & \\ X(N) & X(N,1) & X(N,2) & \ldots & X(N,k) \end{bmatrix}$$

Donde cada renglón corresponde a una lista ordenada a los k vecinos más cercanos al elemento.

Suponga que $X(i) \in R(j)$ y $X(j) \in R(i)$ también que :

$$| R(i) \cap R(j) | \geqslant Ks$$

para un Ks fijo. Entonces X(i) y X(j) comparten al menos Ks vecinos, incluyéndose ellos, por lo que se les asigna a la misma clase.

Es posible que exista un X que tenga Ki vecinos con X(i), y Kj vecinos con X(j), con Ki + Kj $\leqslant$ K y Ki $>$ Ks, Kj $>$ Ks. Entonces X pertenece a i y j, o el agrupamiento i tiene un translape con el agrupamiento j.

En la práctica se utiliza un límite Lmax al número de agrupamientos a los cuales puede pertenecer un dato. Los resultados

dependeran de K, Ks y Lmax, pero como el método resulta sencillo, será fácil experimentar con varios valores de estos parámetros.

Método de las K-medias.

Este algoritmo consiste de tres pasos :
Clasificación, cálculo de los centroides de cada agrupamiento y prueba de convergencia.
Se quieren encontrar M agrupamientos, por lo que se escogen M elementos arbitrariamente que sirven como centroides iniciales.

$$Xp(i) = X(i) \qquad 1 \leq i \leq M$$

La clasificación se realiza por medio del vecino más cercano :

$$X(j) \in Wi \text{ sí y sólo si } d(X(j),Xp(i)) \leq d(X(j),Xp(k)) \quad 1 \leq k \leq M$$

después de haber aplicado lo anterior a todos los elementos, se calculan de nuevo los centroides utilizando el criterio de minimizar el máximo, esto es

$$Xp(i) = Xj(i) \quad \text{tal que} \quad \max_{k} \{ d(Xj(i),Xk(i)) \}$$

sea minimizado para $1 \leq i \leq M$

El criterio de convergencia consiste en verificar si los elementos escogidos por centroides son los mismos de la iteración anterior.
Un problema con este método es que oscile y no se llege a la convergencia.

Isodata.

La característica principal de este algoritmo es que puede separar o unir agrupamientos, por lo que el valor de M puede variar.
La parte principal del isodata es el método de las k-medias, pero el número de agrupamientos se ajusta en cada iteración de acuerdo a un criterio basado en un número de umbrales fijos y variables.
Los agrupamientos se unen, si una o más de las siguientes condiciones se cumplen:
1) El número de agrupamientos es mayor que un valor Mmax.
2) El tamaño del i'esimo |Wi| agrupamiento es menor que un valor Mmin.
3) La distancia entre los centros de los agrupamientos i y j es menor que un valor Dm.
Si M > Mmax, entonces los agrupamientos más cercanos se unen. Si |Wi| < Mmin, entonces Wi se une con el agrupamiento más cercano a él.
Existen tres condiciones bajo las cuales los agrupamientos se separan :

1) El número de agrupamientos es menor que un valor Mmin.
2) El tamaño del i'esimo agrupamiento |Wi| es mayor que un

valor Mmax.

3) El agrupamiento i'esimo es esparcido, relativo a los otros agrupamientos.

Los dos primeros criterios son similares a los del caso de la unión, para el tercer punto se calcula una distancia promedio Di para cada agrupamiento

$$Di = \sum_{X \in Wi} d(X, Xp(i))/(mi-1) \qquad 1 \leq i \leq M$$

$$\bar{D} = \sum_{i=1}^{M} miDi \ /M$$

El agrupamiento Wi se une a otro, sí:

$$Di > \max \{ \bar{D}, \theta s \}$$

para un valor de θs dado.

Si este agrupamiento va a ser separado, es dividido en dos de tal forma

$$Wi = Wi^{+} \cup Wi^{-}$$

Para llevar a cabo esto, se tienen dos métodos.

En el caso simple, se encuentran los dos puntos X^{+} y X^{-} tales que la distancia $d(X^{+}, X^{-})$ sea máxima, entonces cada punto en Wi se asigna a Wi^{+} o Wi^{-} dependiendo de cual punto X^{+} o X^{-} está más cerca. En este caso, X^{+} o X^{-} no son buenos centros de los nuevos agrupamientos.

En el otro caso, da mejores centros de los nuevos agrupamientos. Se localizan los puntos X^{+} y X^{-} con esto se calcula

$$r^{+} = d(X^{+}, Xp(i))$$

$$r^{-} = d(X^{-}, Xp(i))$$

se encuentran Xp^{+} y Xp^{-} tales que :

$$\epsilon^{+} = d(X^{+}, Xp^{+}) + d(Xp^{+}, Xp(i)) - r^{+} \quad y$$

$$\epsilon^{-} = d(X^{-}, Xp^{-}) + d(Xp^{-}, Xp(i)) - r^{-}$$

sean minimizados. Los elementos son asignados dependiendo de su proximidad a Xp^{+} y Xp^{-}.

El primer paso del algoritmo, es realizar una iteración de acuerdo a las k medias, posteriormente los agrupamientos son unidos o separados de acuerdo a las reglas anteriores. La convergencia se verifica de dos formas, una es con el criterio de las k medias y otra es calcular Ss, si este es mayor que un umbral, detenerse, si no hacer otra iteracion, este Ss está dado por :

$$Ss = \frac{\frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1}^{M} d(Xp(i), Xp(j))}{\frac{1}{M} \sum_{i=1}^{M} \frac{1}{m_i(m_i-1)} d(Xj(i), Xk(i))}$$

Algoritmo de Cuantización Vectorial.

EL algoritmo está dado por los siguientes pasos :

0) Inicialización : Dado N, número de niveles, un umbral de distorsión $\epsilon \geq 0$, un alfabeto de reproducción $\hat{A}o$ y una secuencia de entrenamiento $\{ Xj;\ j=0,\ldots,n-1\}$.

Se inicializa $m=0$ y $D_{-1}=\infty$.

1) Dado $\hat{A}m=\{Yi\ ;\ i=1,\ldots,\ N\}$, encuentre la distorsión mínima de la partición $P(\hat{A}m)=\{Si\ ;i=1,\ldots,N\}$ de la secuencia de entrenamiento : $Xj \in Si$, si $d(Xj,Yj) \leq d(Xj,Yi)$ para toda i. Calcule la distorsión promedio

$$Dm = D(\ \hat{A}m,\ P(\hat{A}m)\ ) = n^{-1} \sum_{j=0}^{n-1} \min_{Ye\hat{A}m} d(Xj,Y)$$

2) Si $(Dm-1 - Dm\ )/Dm \leq \epsilon$, se tiene el alfabeto final de reproducción $\hat{A}m$, de otra forma continue

3) Encuentre el alfabeto de reproducción óptimo $\hat{X}(P(\hat{A}m)) = \{\ \hat{X}(Si)\ ;\ i=1,\ldots,N\}$ para $P(\hat{A}m)$. Sea $\hat{A}m+1 = \hat{X}(P(\hat{A}m))$. Reemplazar m por m+1 e ir al paso 1.

Este algoritmo es parecido al de las K-medias, las diferencias son: La forma de encontrar el alfabeto de reproducción óptimo en cada paso y el criterio de convergencia.

Selección del alfabeto de reproduccion $\hat{A}$.

Un método para seleccionar el alfabeto inicial $\hat{A}o$, sería asignar los primeros N elementos de la secuencia de entrenamiento, como $\hat{A}0$, entonces aplicar el algoritmo anterior hasta encontrar los $\hat{A}m$ óptimos.

Una segunda técnica que es util para tener un nivel de distorsión deseado y con un número de niveles de cuantización variable es el siguiente :

0) Inicialización : M=1 y $\hat{A}o(1) = \hat{X}(A)$, el centroide de la secuencia de entrenamiento.

1) Perturbe el alfabeto de reproducción $\hat{A}o(m)$ que tiene M elementos $\{\ Yi\ ;\ i=1,\ldots,M\}$ con ϵ ,entonces a cada elemento Yi le corresponden : $Yi+\epsilon$, $Yi-\epsilon$ donde ϵ es un elemento fijo de perturbación.

Se tiene una nueva colección A de $\{Yi+\epsilon,\ Yi-\epsilon,\ i=1,\ldots,M\}$ que tiene 2M elementos.

Reemplace M por 2M.

2) Si M=N entonces $\hat{A}o = \hat{A}(M)$ y realize el algoritmo de cuantización vectorial por última vez, de otra forma realize el algoritmo de cuantización vectorial y ya obtenido el alfabeto óptimo para ese nivel, regrese al paso 1 de este algoritmo.

Utilizando este algoritmo, se empieza con un nivel de cuantización, el cual esta dado por el centroide de la secuencia de entrenamiento, este centroide se parte en dos y se tiene entonces un nivel dos de cuantización, cada uno de estos se parte en dos y asi sucesivamente, por lo que se tiene al final un cuantizador de 1,2,...,N niveles.

COMPARACION DE DOS TECNICAS DE RECONOCIMIENTO DE PALABRAS

Uno de los problemas interesantes que se presentan en el estudio de la voz es el del reconocimiento de palabras por medio de una computadora. Este problema se puede formular como : dado un diccionario formado por N palabras, de las cuales se han obtenido de acuerdo a un criterio las mejores características representativas, se quiere que una computadora sea capaz de poder reconocer una palabra, la cual ha sido pronunciada por alguna persona. Dentro de esto se podrian tener algunas variaciones en las cuales se tendrian sistemas para una sola persona con un número reducido de palabras dentro del diccionario y además sólo se permitiría la pronunciación de palabras que estuvieran dentro de este diccionario, o, tener un sistema general en el cual se tendría un número ilimitado de personas, así como, un diccionario grande y se permitiría la pronunciación de palabras fuera del diccionario. Este sistema debe ser capaz de indicar que palabras pertenecen o no a su diccionario.

El sistema que se empleó dentro de este estudio está más cercano al primer sistema descrito. Esto se debió en gran parte a las limitaciones en cómputo, pero principalmente a que de esta forma se tendrian menos parámetros que controlar. El sistema consta de un diccionario de 10 palabras, las cuales son los dígitos: (cero, uno, dos, tres, cuatro, cinco, seis, siete, ocho, nueve), pronunciados por una sola persona(el autor).

Los sistemas de reconocimiento que se comparan son :

Un par de sistemas basados en LPC y 4 sistemas en base al vocoder de canal. Dentro de estos sistemas se hicieron uso de las técnicas presentadas en [,] y la técnica presentada en [,]. Básicamente estas son: la primera de ellas no hace segmentación de la palabra, mientras que la segunda si realiza una segmentación. Este último sistema tiene la ventaja de que no es necesario un alineamiento en el tiempo, a diferencia de sistemas que lo requieren (time warping systems [ ]).

Experimento.

La señal de voz que se utilizó para el estudio fue grabada analógicamente en una cinta magnética y posteriormente almacenada en disco. Para esto se filtro la señal por medio de un filtro analógico paso-banda con rango de 100 hz a 3.2 khz, esta señal se muestreó a una razón de 6400 muestras por segundo por medio de un convertidor de 12 bits y rango de entrada entre +5 y -5 volts. Posteriormente la señal fue preenfatizada con un filtro de la forma :

$$x(n) = s(n) - 0.95s(n-1)$$

cuya respuesta en frecuencia se muestra en la figura, se tomaron bloques de 40 mseg y se fue corriendo una ventana de análisis cada 20 mseg, lo cual representa 256 muestras por bloque, 64 de las cuales corresponden al bloque anterior, 128 al bloque que se estaba analizando en ese momento y 64 del siguiente bloque. A estas 256 muestras se les multiplicó por una ventana de Hamming de la forma:

$$w(n) = 0.54 - 0.46\cos(2\pi n/(N-1)) \quad 0 \leq n \leq N-1$$

A los datos que se obtenían se les calculaba, para el caso de los sistemas basados en LPC 11 correlaciones y para los sistemas basados en el vocoder de canal 512 puntos de la transformada rápida de Fourier(agregandole 256 ceros a la señal), los cuales se normalizaron de tal forma que la energía total fuera igual a 1, con esto se calcularon las energías en 8 y 16 bandas uniformemente espaciadas en el rango 0 a 3.2 Khz. lo que corresponde a filtros de 400 y 200 hz de ancho de banda respectivamente.

Para el cálculo de los patrones o características de cada una de las palabras se generó una secuencia de entrenamiento, la cual consta de 10 repeticiones de cada una de las palabras y posteriormente recibieron el tratamiento señalado anteriormente. Con la señal obtenida se generaron por medio de un cuantizador vectorial los patrones de cada una de las palabras.

Los patrones se obtuvieron de la siguiente manera; para los sistemas no segmentados, a cada una de las palabras se le calcularon 8 patrones , para los sistemas con segmentación, cada palabra en la secuencia de entrenamiento fue segmentada en 4 partes y se formaron nuevas secuencias de entrenamiento, a cada una de estas se le calcularon 4 patrones, lo que da un total de 16 patrones por palabra.

La distorsión empleada para la obtención de estos patrones fue la Itakura-Saito normalizada para los sistemas con LPC y la Itakura-Saito discreta para los sistemas de vocoder de canal.

Para la comparación de los sistemas se tuvieron dos criterios, uno es simplemente el número de aciertos y errores que se obtienen, el otro trata de cuantificar el que tan bien o mal se desempeña el sistema, para esto se toma la distorsión que se obtendría al codificar la palabra bajo prueba con sus patrones y se le llama Dm, además se calcula la mejor distorsión que se obtiene de codificar la palabra por medio de los patrones de las otras palabras, esto es :

$$D' = \min_{i} Di \qquad i \neq m$$

y se calcula

$$R = (D' - Dm)/Dm$$

de esto se nota que si $R < 0$ se tiene un error y si $R > 0$ un acierto. Puede observarse que mientras mayor sea el valor de R se tiene una mejor decisión. Se calcularon tambien los valores promedio y variancia de las diferentes R. Los resultados se muestran en las tablas.

De las tablas se nota que los sistemas basados en vocoder de canal con 8,16 bandas y con segmentacion son los que tienen los mejores resultados, si bien esto podria intuirse debido a que se utiliza un mayor numero de parametros. En los otros sistemas se tienen resultados interesantes, uno de ellos es que el error cometido por el sistema de LPC no segmentado (confundir un siete por seis) se corrige con el mismo sistema pero con segmentacion, el otro resultado interesante es el hecho de que el sistema de vocoder de canal de 8 bandas no comete errores aun en el caso sin segmentacion y sus valores Rm y Rv son mejores que los del sistema de LPC el cual tiene mas parametros. Una de las desventajas que podrian presentarse es el numero de filtros requeridos en el caso de segmentacion, aunque si bien no es necesario tenerlos todos en memoria.

Aunque los resultados presentados estan en contradiccion con los resultados en [6] podria deberse a la forma en la que se calculan los valores del vocoder de canal.

Otra de las ventajas que se tienen en el sistema de bandas de energia es el hecho de que se utiliza una tecnica optima como es la transformada rapida de Fourier(esto es la razon por la cual se tomaron 8 y 16 bandas, con lo cual se evita cualquier tipo de interpolacion).

La ventaja que presenta el sistema de bandas de energia puede deberse al hecho que representan una mejor aproximacion al espectro real que la que realizan los sistemas LPC.

Estos resultados son alentadores, ya que si bien han sido para una persona y un numero pequeno de experimentos, han mostrado mejoras con respecto a sistemas anteriores.

Palabra seleccionada

Palabra pronunciada	0	1	2	3	4	5	6	7	8	9
0	25	.	.	.	.	.	.	.	.	.
1	.	23	.	.	.	.	.	.	.	.
2	.	.	24	.	.	.	.	.	.	.
3	.	.	.	24	.	.	.	.	.	.
4	.	.	.	.	22	.	.	.	.	.
5	.	.	.	.	.	22	.	.	.	.
6	.	.	.	.	.	.	21	.	.	.
7	.	.	.	.	.	.	.	23	.	.
8	.	.	.	.	.	.	.	.	23	.
9	.	.	.	.	.	.	.	.	.	20

Tabla de aciertos para los sistemas:

LPC segmentado

Vocoder de canal 8,16 bandas con y sin segmentación.

Palabra pronunciada \ Palabra seleccionada	0	1	2	3	4	5	6	7	8	9
0	25	.	.	.	.	.	.	.	.	.
1	.	23	.	.	.	.	.	.	.	.
2	.	.	24	.	.	.	.	.	.	.
3	.	.	.	24	.	.	.	.	.	.
4	.	.	.	.	22	.	.	.	.	.
5	.	.	.	.	.	22	.	.	.	.
6	.	.	.	.	.	.	21	.	.	.
7	.	.	.	.	.	.	1	22	.	.
8	.	.	.	.	.	.	.	.	23	.
9	.	.	.	.	.	.	.	.	.	20

Tabla de aciertos para el sistema LPC no segmentado.

Sistema correlaciones sin segmentar

Palabra	Rprom.	Rvar.
0	0.21069	0.04432
1	0.27071	0.05937
2	0.17424	0.02449
3	0.23374	0.06752
4	0.52925	0.24438
5	0.86546	0.65768
6	0.20499	0.01058
7	0.15475	0.02387
8	0.29362	0.07329
9	0.90900	0.72005

Sistema correlacion con segmentacion.

Palabra	Rprom	Rvar
0	0.39883	0.16028
1	0.35117	0.10139
2	0.30226	0.08208
3	0.50788	0.25956
4	0.71333	0.41159
5	1.26662	1.55347
6	0.51922	0.22708
7	0.53363	0.23629
8	0.77917	0.52252
9	1.14300	1.04956

Sistema vocoder de canal con 8 bandas. (no segmentado)

Palabra	Rprom.	Rvar.
0	1.73926	3.00412
1	0.75205	0.32729
2	0.65460	0.41452
3	1.56193	2.58297
4	1.05727	0.89082
5	1.35319	1.69453
6	1.62610	2.37394
7	1.27624	1.43894
8	0.96121	0.70810
9	2.13914	3.89802

Sistema vocoder de canal 8 bandas. (con segmentacion)

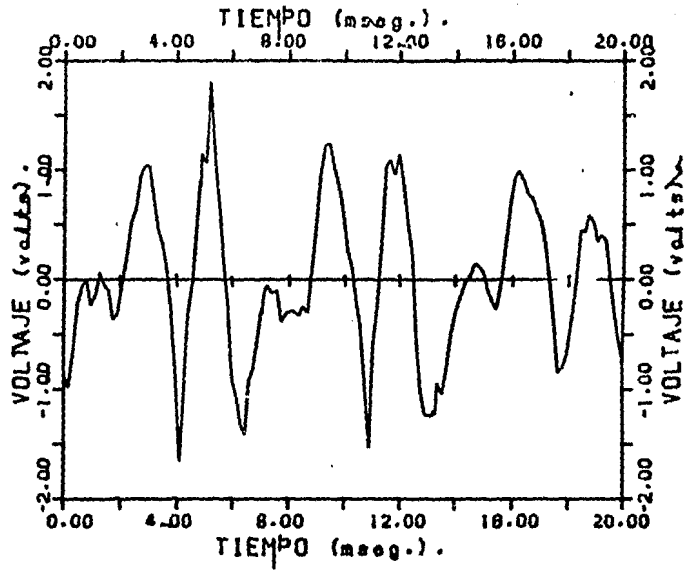
Palabra	Rprom.	Rvar.
0	1.94336	3.73821
1	1.41049	1.42382
2	0.76691	0.52036
3	3.34739	12.16320
4	1.19625	0.76705
5	1.63420	2.46033
6	2.24356	4.37261
7	2.51644	5.58091
8	1.70427	2.25502
9	2.96295	7.35784

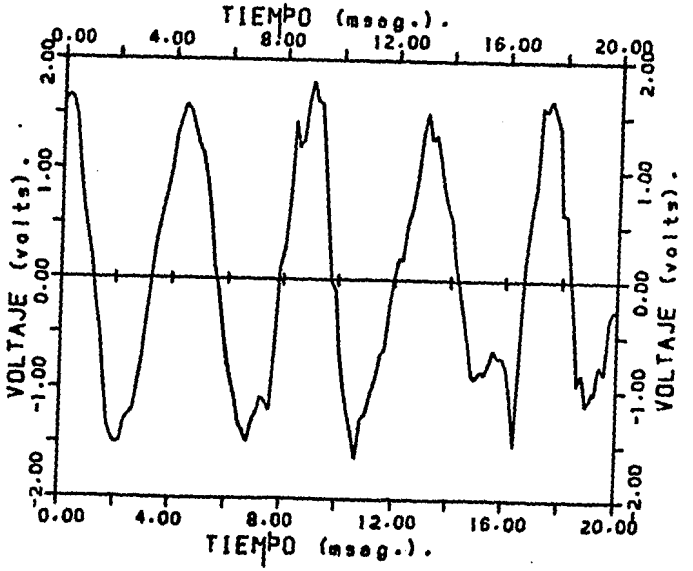
Sistema vocoder de canal con 16 bandas. (no segmentado)

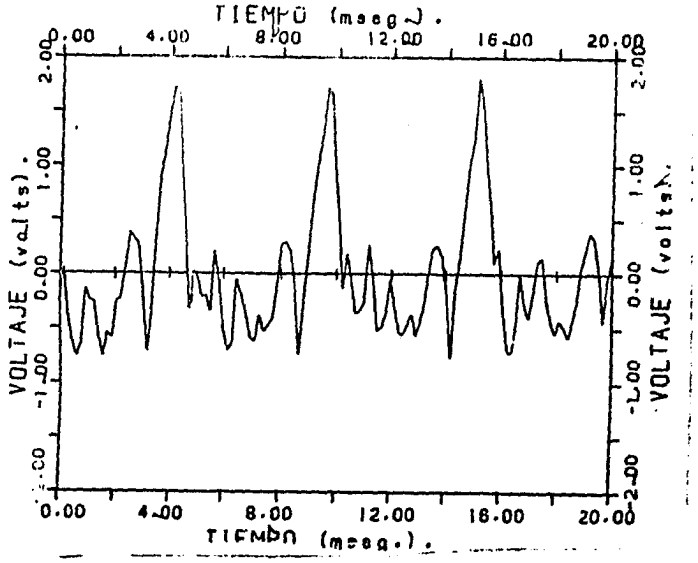
Palabra	Rprom.	Rvar.
0	1.88797	3.57489
1	1.03871	0.70190
2	0.65804	0.37680
3	1.69790	2.95307
4	0.92400	0.62049
5	1.41099	1.86888
6	1.39656	1.66477
7	1.21738	1.28564
8	0.81309	0.46443
9	2.05059	3.57580

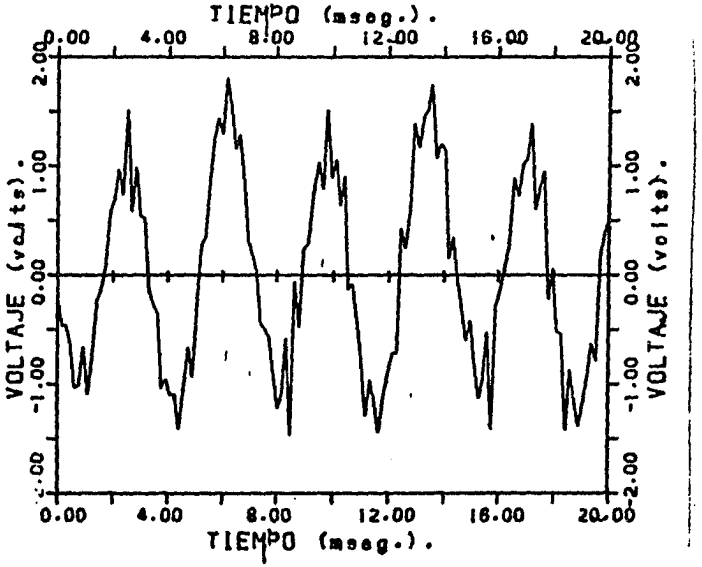
Sistema vocoder de canal 16 bandas. (con segmentacion)

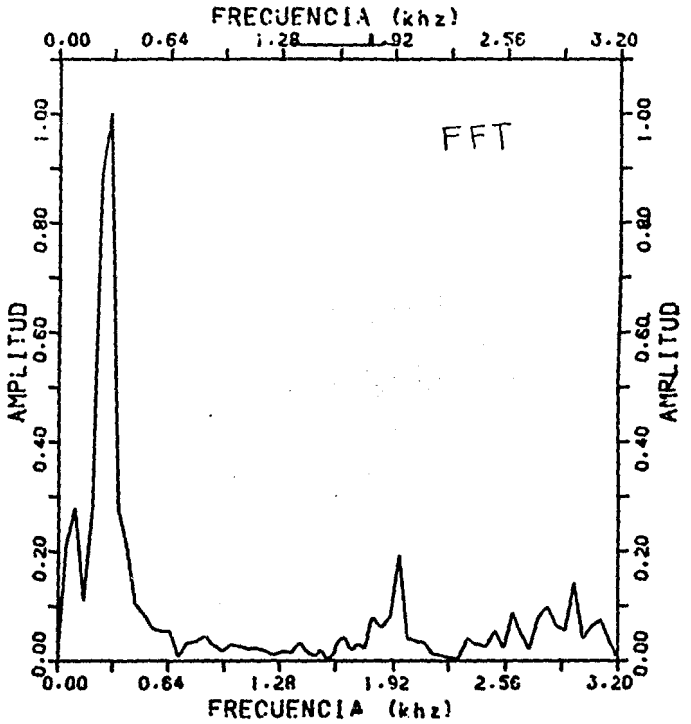
Palabra	Rprom.	Rvar.
0	2.02246	4.07200
1	1.50528	1.61758
2	0.95215	0.79181
3	2.88055	8.78173
4	1.07487	0.67527
5	1.82760	3.38950
6	1.77936	2.54038
7	2.14209	4.05593
8	1.46308	1.67001
9	2.80662	6.65809

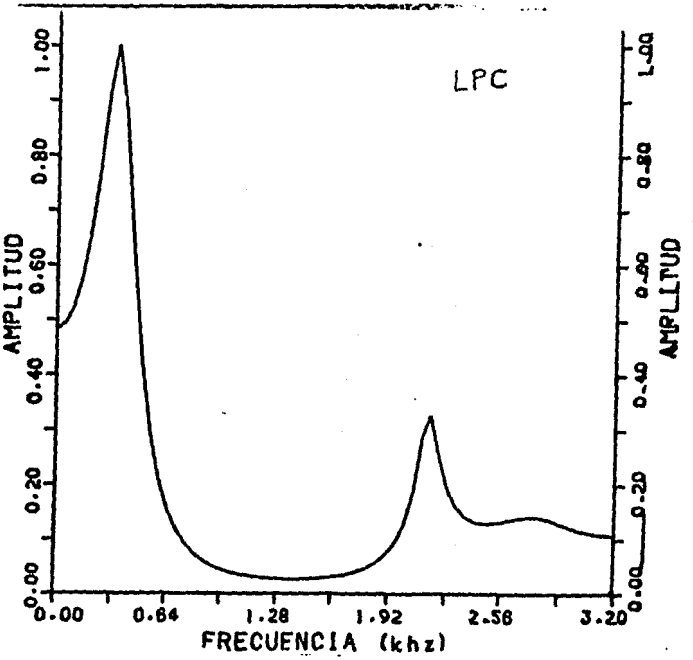

TIEMPO (mseg.).
0.00
4.00
8.00
12.00
16.00
20.00
2.00
1.00
0.00
-1.00
-2.00
VOLTAJE (volts).
TIEMPO (mseg.).

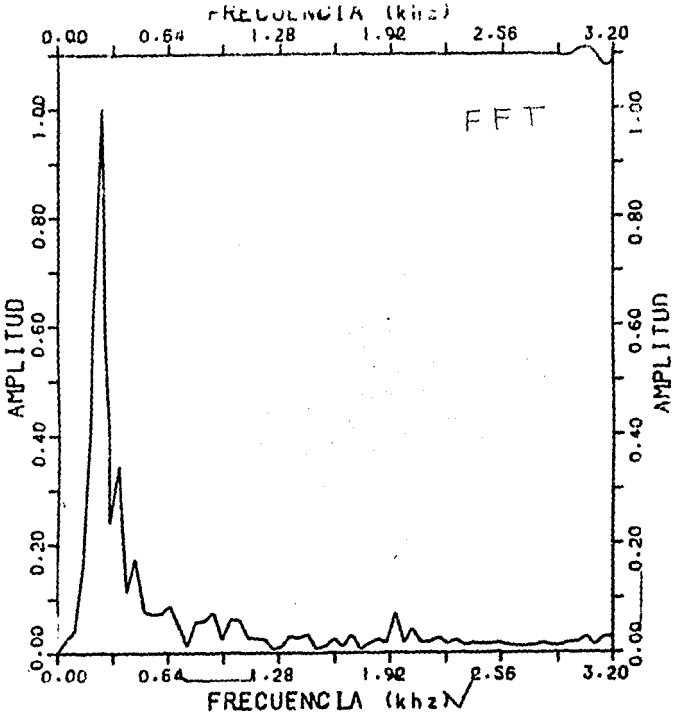

TIEMPO (mseg.).
0.00
4.00
8.00
12.00
16.00
20.00
2.00
1.00
0.00
-1.00
-2.00
VOLTAJE (volts).
TIEMPO (mseg.).

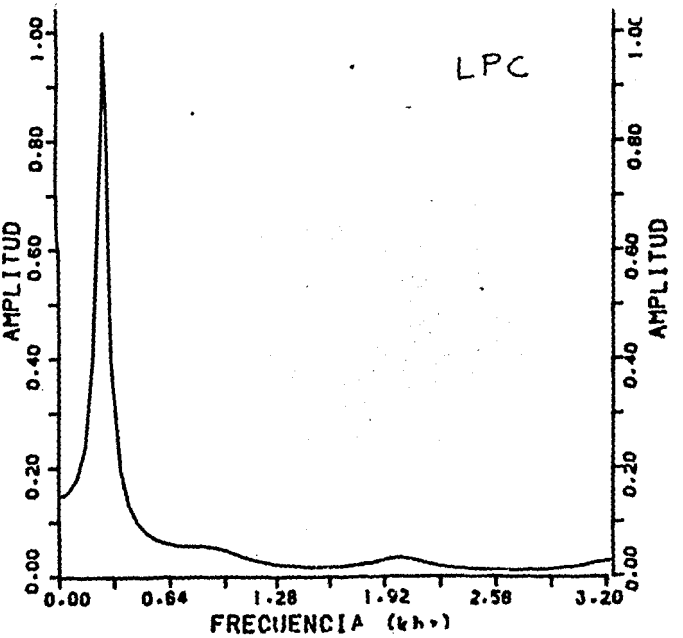

TIEMPO (mseg.).
0.00
4.00
8.00
12.00
16.00
20.00
VOLTAJE (volts).
2.00
1.00
0.00
-1.00
-2.00

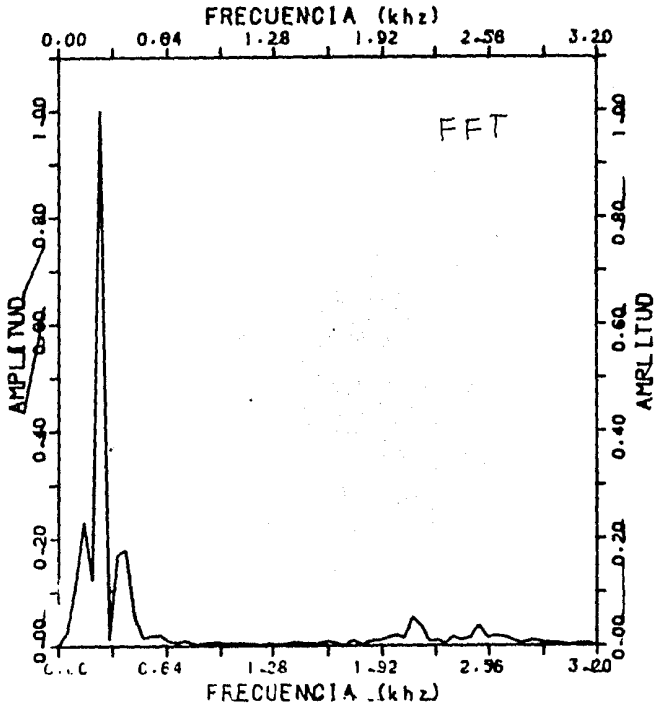

TIEMPO (mseg.).
0.00
4.00
8.00
12.00
16.00
20.00
VOLTAJE (volts).
2.00
1.00
0.00
-1.00
-2.00

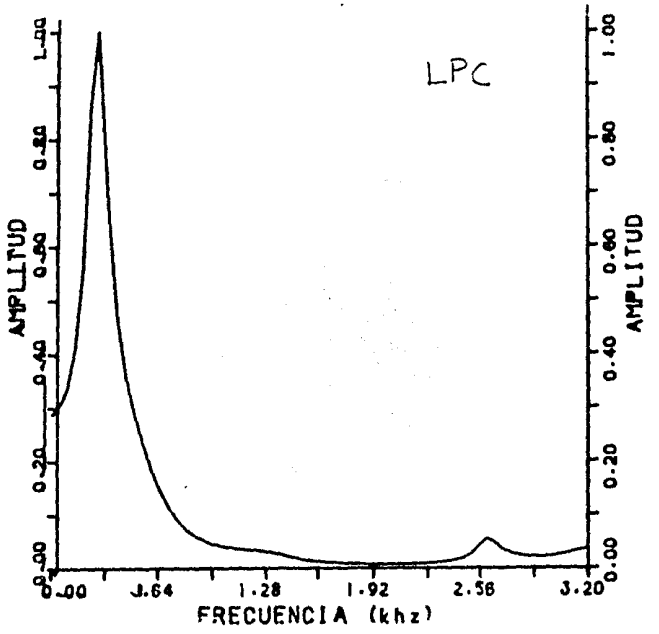

FRECUENCIA (khz)
0.00
0.64
1.28
1.92
2.56
3.20
FFT
AMPLITUD
0.00
0.20
0.40
0.60
0.80
1.00
FRECUENCIA (khz)


LPC
AMPLITUD
0.00
0.20
0.40
0.60
0.80
1.00
0.64
1.28
1.92
2.56
3.20
FRECUENCIA (khz)


FFT
AMPLITUD
0.00
0.20
0.40
0.60
0.80
1.00
0.00
0.64
1.28
1.92
2.56
3.20
FRECUENCIA (khz)


LPC
AMPLITUD
0.00
0.20
0.40
0.60
0.80
1.00
0.00
0.64
1.28
1.92
2.56
3.20
FRECUENCIA (khz)


FRECUENCIA (khz)
0.00
0.64
1.28
1.92
2.56
3.20
FFT
AMPLITUD
0.00
0.20
0.40
0.60
0.80
1.00
FRECUENCIA (khz)


LPC
AMPLITUD
0.00
0.20
0.40
0.60
0.80
1.00
0.64
1.28
1.92
2.56
3.20
FRECUENCIA (khz)

REFERENCIAS :

1) G. C. M., Acoustic Theory of Speech Production , Mouton and Co.,'s- Gravenhage, The Netherlands, 1960.

2) J. D. Markel, A. H. Gray, Jr., Linear Prediction of Speech , Springer-Verlag, Berlin, Heidelberg, New York , 1976.

4) J. Makhoul, "Linear Prediction : A Tutorial Review," Proc. IEEE 63, pag. 561-580, Abril, 1975.

5) A. Papoulis, "Maximum Entropy and Spectral Estimation A Review," IEEE Trans. Acoust., Speech, and Signal Processing, Vol. ASSP-29, No. 6, pag. 1176-1186, Diciembre, 1981.

6) B. A. Dantrich, L. R. Rabiner, T. B. Martin, "On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition," IEEE Trans. Acoust., Speech, and Signal Processing, Vol. ASSP-31, No. 4, pag. 793-806, Agosto, 1983.

7) A. Buzo, A. H. Gray Jr, R. Gray and J. D. Markel, "Speech Coding Based Upon Vector Quantization ," IEEE Trans. Acoust., Speech, and Signal Processing, Vol. ASSP-28, pag. 562-574, Octubre, 1980.

8) Y. Linde, A. Buzo and R. M. Gray, "An Algorithm for Vector Quantizer Design," IEEE Tran. on Communications, Vol. COM-28, pag. 84-95, Enero, 1980.

9) R. M. Gray, A. Buzo, A. H. Gray and Y. Matsuyama, "Distortion Measures for Speech Processing," IEEE Trans. on Acoust., Speech, and Signal Processing, Vol. ASSP-28, pag. 376-376, Agosto, 1980.

10) L. R. Rabiner and B. Gold, Theory and Application of Digital Signal Processing , Prentice-Hall, Englewood Cliffs, N.J. 1978.

11) T. Bially, W. M. Anderson, "A Digital Channel Vocoder," IEEE Trans. on Communication Tech. Vol. COM-18, No. 4, pag. 435-441, Agosto, 1970.

12) A. Buzo, C. Rivera, and H. Martinez, "Discrete utterance recognition based upon source coding techniques," Proc. of ICASSP 1982, IEEE International Conference on Acoustics, Speech, and Signal Processing, Paris, France, pag. 539-542, Mayo ,1982.

13) J. E. Shore and D. K. Burton, "Discrete utterance speech recognition without time alignment," IEEE Trans. Inform. Theory IT-29. pag. 473-491, Julio ,1983.

14) D. K. Burton, J. E. Shore, and J. T. Buck, "Isolated-Word Speech Recognition Using Multi-Section Vector Quantization Code Books," Computer Science and Systems Branch Information Technology Division, Naval Reasearch Laboratory, Washington, D. C., 13 de Julio de 1984.

15) A. Jazcilevich, Reconocimiento de Voz por Computadora, Tesis.

16) A. Buzo, H. Martinez, C. Rivera, A. Jazcilevich. "Isolated Word Recognition Based Upon Source Coding Techniques," Proc. Seventh Data Communications Symposium, pag. 268-270, Mexico, Octubre, 1981.

17) L. Rabiner, R. W. Schafer, Digital Processing of Speech Signals , Prentice-Hall, Englewood Cliffs, N.J. 1978.

18) A. V. Oppenheim, editor, Applications of Digital Signal Processing , Prentice-Hall, Englewood Cliffs, N.J. 1978.