Reconocimiento de Comandos Aislados Usando la Voz.

En la actualidad, debido principalmente al avance tecnológico tan extraordinario que ha habido en los campos de la electrónica y la computación,es posible realizar proyectos que el hombre había concebido antes solamente como fantasías. Este es el caso de los sistemas de reconocimiento de comandos hablados por medio de computadora, capaces de reconocer palabras aisladas, tomadas de un vocabulario preestablecido, en una forma automatica. Antes de analizar en detalle de la forma en que la computadora realiza el reconocimiento, es pertinente hacer mención de algunas de la caracteristicas de la voz humana. La generación de voz es un proceso físico que parece tan natural,que rara vez nos ponemos a pensar como es que se produce. Si se tomaran en cuenta todos los mecanismos del cuerpo humano que intervienen en éste proceso nos daríamos cuenta que en realidad se trata de un proceso mecánico muy complejo que involucra la interacción cordinada de pulmones,garganta,cuerdas vocales,traquea,. nariz y lengua. Para cada sonido que se emite,el cerebro coordina todos estos organos y generalmente lo hace en una forma automática. La voz se produce con una corriente de aire que sale de los pulmones. Los pulmones retienen en su interior cerca de tres cuartos de su capacidad con aire, y el diafragma puede enviar cantidades de aire controladas en un tiempo determinado y en una cantidad precisa hacia la garganta.

La cantidad y la presión del aire que se envia a la garganta (traquea) determina algunas caracteristicas del sonido emitido.Un murmullo utiliza muy poco aire, mientras que un grito requiere de una cantidad muy grande. Aunque este flujo de aire no es propiamente voz,actua como la fuente de exitación para los sonidos que componen la voz. Al salir dichas corrientes de aire con la fuerza suficiente hacen vibrar el aire que se encuentra fuera del cuerpo y estas vibraciones son ondas acusticas que son perceptibles por el oido humano. El organo que genera las vibraciones de las corrientes de aire son las cuerdas vocales. Usando estas se puede bloquear parcial o totalmente el flujo de aire y de esta forma el flujo de aire se convierte en un conjunto de pequeños zumbidos. La rapidéz con la cual las cuerdas vocales

se abren o se cierren es la frecuencia del zumbido que se produce. Pero la frecuencia del zumbido no es lo unico que distingue un sonido de voz de otro. Despues de salir de la laringe, el flujo de aire pasa hacia la traquea (faringe), la boca y la nariz. Todos ellos cambian el sonidó ligeramente. A éstos organos se les conoce como el tracto vocal. Cuando la corriente de aire pasa hacia el tracto vocal, la resonancia natural de éstas cavidades modifica la vibracion para darle otras caracteristicas adicionales al sonido emitido. Moviendo la lengua en varias posiciones dentro de la boca, cerrando y abriendo los dientes, levantando y bajando el paladar, o de alguna otra manera cambiando la forma o el tamaño del tracto vocal, se modifica la resonancia. De esta forma, el flujo de aire es modificado; a ésta modificación se le conoce como modulacion. A los ingenieros les gusta representar procesos que ocurren en el mundo fisico a través de modelos matemáticos, con los cuales se pueda experimentar bajo las condiciones de laboratorio. De esta forma se ha creado un modelo matemático que representa la forma en la que los diferentes organos vocales generan la voz. A partir de estos modelos se han desarrollado una serie de algoritmos de reconocimiento de patrones.

El sistema de reconocimiento que se desarrolló en la DEPFI-UNAM es capaz de crear los patrones de reconocimiento de cualquier palabra aislada; de éste modo, se puede tener un conjunto de patrones de acuerdo con las necesidades de cada usuario. Para la aplicación que se desarrollo se pueden reconocer los numeros del 0 al 9 y las palabras si o no. Pero se podrian tener otras palabras en este sistema dependiendo de la aplicación. La aplicación que se desarrollo era crear un conmutador telefónico que reconozca comandos hablados. Este conmutado sería capáz de hacer la conexión de una linea telefónica externa con cualquiera de las extensiones internas que maneje éste, sin la ayuda de una operadora. La extención se pide diciendo digito por digito el numero de la extensión requerida. Este sistema es una de las tantas aplicaciones de un sistema de reconocimiento de comandos, pero éstas ideas pueden ser ampliadas para tener diferentes aplicaciones, como por ejemplo, la comunicación que se podria hacer entre una institucion bancaria y sus cuenta habientes . Para hacer operaciones bancarias por teléfono se necesita en la actualidad un generador de tonos conectado al teléfono para indicar

la operación que se quiere realizar. Usando el sistema de reconocimiento de comandos hablados el generador de tonos puede ser eliminado y en su lugar se puede seleccionar la operación deseada por medio de comandos hablados. Otras aplicaciones podrian ser encontradas en la medicina, donde se podria contar con sistemas que ayuden a personas con problemas fisicos, tales como sordos, mudos, ciegos, paraliticos, etc. Por ejemplo se podria tener en un cuarto de hospital un sistema de reconocimiento que pudiera efectuar diferentes acciones para un enfermo inmovilizado, tales como tener el control de la iluminación del cuarto, tener el control de la posicion de la cama donde se encuentra el paciente, etc. Otro ejemplo seria una silla de ruedas con motor donde el paciente indique con voz la dirección que debe tomar esta (derecha, izquierda, adelante, atras). Aunque el número de aplicaciones esta limitado solo por la imaginación, todas ellas serian ligeras variaciones de la misma filosofia básica de diseño presentada en nuestro trabajo.

Modelo del Aparato Vocal

Un modelo lineal de producción de voz para sonidos no nasales, es decir, vocales, ha sido propuesto por Fant [1959, 1960], y la asumpción de éste modelo ha sido analizada en detalle [Fant, 1960; Flanagan, 1972]. El termino voz se refiere únicamente a los sonidos vocales en este estudio. Como se muestra en la figura 1-1, este modelo esta compuesto por tres filtros acusticos. La principal justificación para este modelo esta basado en la teoria del tubo acustico, medidas del volumen de velocidad y la presión de la forma de onda del sonido, los datos de los rayos-x, y los resultados obtenidos usando circuitos eléctricos para sintetizar los sonidos vocales. La mayor asumpción de este modelo es la separabilidad de los segmentos del filtro durante la generación de la voz.

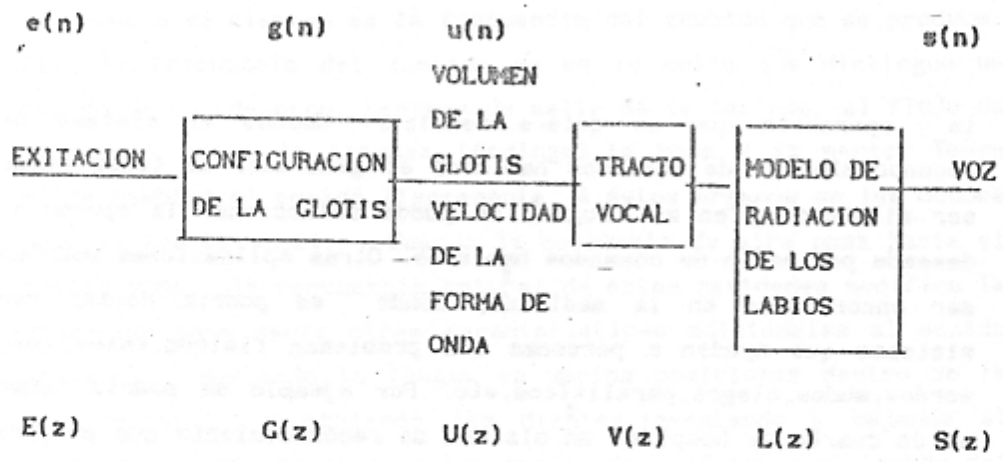

Fig.1.1.- Modelo Lineal de Producción de Voz

La exitación e(n) es modelado como una serie de impulsos unitarios escalados y espaciados con un periodo que corresponde al tono (Pitch) para sonidos llamados vocales, y es expresado como

$$e(n) = E_0 \sum_{j=0}^{\infty} \delta\,(n-JP)$$

Donde E_0 es un factor de escala, $\delta(n)$ es la función delta de Kronecker. La transformada z es igual a

$$E(z) = \frac{E_0}{1 - z^{-P}}$$

donde P es la peridicidad del tono.

El filtro de la forma de la glotis es un filtro cuya respuesta impulso esta dada por

$$g(n) = G_0(n+1)\, e^{-cnT}$$

y cuya transformada z es

$$G(z) = \frac{G_0}{(1 - e^{-cT} z^{-1})^2}$$

El filtro del tracto vocal es usado para modelar las caracteristicas de acusticas de resonancia de los espacios de aire contenidos entre la glotis y los labios. Esta aproximación del tubo acustico esta compuesto de un número especifico de secciones, cada sección tiene una area constante.

La transformada z de V(z) del filtro del tracto vocal puede ser modelado por un número pequeño finito k (generalmente 3 o 4) de polos complejos de banda angosta.

$$V(z) = \prod_{i=1}^{k} \frac{1}{(1-z^{-1}z_i)(1-z^{-1}z_i^*)}$$

o

$$V(z) = \prod_{i=1}^{k} \frac{1}{1-e^{-\pi b_i T}\cos(2\pi f_i T)z^{-1}+e^{-2\pi b_i T}z^{-2}}$$

Una resonancia espectral con frecuencia central f_i y ancho de banda de dos lados b_i es definida por cada par de polos complejos (z_i, z_i^*) , donde $z_i = \exp[-\pi b_i + j2\pi f_i T]$, z_i^* es el complejo conjugado de z_i, y T es el periodo de muestreo. Estas resonancias son también llamados formantes. Los formantes tienen sus centros de frecuencia generalmente menos de 5 kHz y anchos de banda generalmente menores a 100 Hz.

El modelo de radiación de los labios representa la transformación de una forma de onda con una con un volumen y velocidad a una forma de onda de sonido. La transformada z de esta función es igual a

$$L(z) = L_0 (1 - z^{-1})$$

donde L_0 es un factor de escala.

El modelo total lineal de laproducción de voz para S(z) es el producto de la transformada z de cada uno de los filtros mencionados anteriormente

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

o

$$= \frac{L_0 G_0 E_0 (1 - z^{-1})(1 - e^{-cT}z^{-1})^{-2}(1 - z^{-P})^{-1}}{\prod_{i=1}^{k} 1-e^{-\pi b_i T}\cos(2\pi f_i T)z^{-1}+e^{-2\pi b_i T}z^{-2}}$$

análisis de la siguiente forma. El factor en el numerador $[1-z^{-1}]$ se asume que se cancela aproximadamente por un de los factores del numerador $[1 - e^{-cT}z^{-1}]^{-1}$ ya que cT es generalmente menor que la unidad. El termino sobrante en el numerador $[1 - e^{-cT}z^{-1}]^{-1}$ es incluido en el producto de factores del denominador. Finalmente, las potencias de z son acomodadas en orden ascendente, y las constantes son combinadas para producir la forma

$$S(z) = \frac{\sigma}{\sum_{i=0}^{M} a_i z^{-i}} E(z)$$

donde a_i es real, $a_0 = 1$, $M=2k+m$. Despejando E(z)

$$E(z) = S(z) \sum_{i=0}^{M} a_i z^{-i}$$

donde ésta ecuación se conoce como el modelo de analisis.

CUANTIZACION VECTORIAL

Cuantización vectorial (CV) es una generalización de la cuantización escalar. Mientras que la cuantización escalar está ligada a la conversión analógico/digital, la CV está asociada con un procesamiento digital de señales sofisticado, donde en la mayoria de los casos, las señales de entrada ya tienen alguna forma de representación digital y la CV es usada, pero no exclusivamente, para el próposito de compresión de datos.

Un vector puede ser usado para describir casi cualquier tipo de patrón, ya sea la forma de onda de un segmento de voz o de una imagen simplemente formando un vector de muestras de la forma de onda o de la imagen. En nuestro caso estamos interesados en formar vectores de un conjunto de parámetros que representan la envolvente espectral de un sonido de voz. La cuantización vectorial puede ser vista como una forma de reconocimiento de patrones donde un patrón de entrada es aproximado por un conjunto predeterminado de patrones estándar, es decir, el patrón de entrada es igualado con uno de los patrones previamente guardados.

Un cuantizador vectorial Q de dimensión k y tamaño N es una transformación de un vector (o un punto) del espacio Euclidiano de dimensión k, R^k en un conjunto finito C que contiene N salidas o puntos de reproducción, llamados vectores de código (codevectors). Así,

$$Q: R^k \rightarrow C ,$$

donde $C = (y_1, y_2, \ldots, y_N)$ y $y_i \in R^k$ para cada $i \in J = \{1,2,\ldots,N\}$. El conjunto C es llamado el Alfabeto (Code Book) y tiene un tamaño de N, lo que significa que se tienen N elementos distintos, donde cada uno de ellos es un vector en R^k. La resolución de un cuantizador vectorial es $r = (\log_2 N)/k$ el cual mide el número de bits por componente vectorial usado para representar el vector de entrada y da una indicación de la precisión que es obtenida con el cuntizador vectorial si el Alfabeto es bien diseñado.

Asociado con cada punto del cuantizador vectorial hay una partición de R^k en N regiones o celdas, R_i para $i \in J$. La i-ésima región es definida por:

$$R_i = \{ x \in R^k : Q(x) = y_i \} \qquad (1)$$

algunas veces llamado la imagen inversa o pre-imagen de y_i con la transformación Q y puede ser denotado más concisamente como como $R_i = Q^{-1}(y_i)$. De la definición de regiones, se concluye que

$$\cup_i R_i = R^k \text{ y } R_i \cap R_j = \emptyset \text{ para } i \neq J$$

asi las regiones forman una partición de R^k. Una región que no tiene limites es llamada región de sobre-carga. Una región limitada, es decir, que tienen un volumen determinado es llamada una región granular.

Una importante propiedad de un conjunto en R^kes su convexidad. Un conjunto R^kes convexo si a y b $\in$ S implica que $\alpha a + (1-\alpha)b \in S$ donde $0 < \alpha < 1$. Un cuantizador vectorial es llamado regular si

a) Cada celda R_i es un conjunto convexo

b) Si $x \in R_i$, entonces $Q(x) = y_i$, esta contenido en R_i.

Un cuantizador vectorial Polytopal es un cuantizador regular cuyas celdas de partición están limitadas por segmentos de superficies de hiperplanos de dimension k. Equivalentemente un Polytopal es una intersección de un número finito de espacios medios de la forma $\{x \in R^k : u_v x + \beta_v \geq 0\}$. Un cuantizador vectorial puede ser descompuesto en dos operaciones, la codificación vectorial y la decodificación vectorial. El codificador E es una transformación de R^k al conjunto de indices J, y el decodificador D, transforma el conjunto de indices J en un conjunto de reproducción C. Así,

$$E: R^k \rightarrow J \text{ y } D: J \rightarrow R^k \qquad (2)$$

La operación total de CV puede ser interpretada como la composición de dos operaciones:

$$Q(x) = DC(x) = D(C(x)) \qquad (3)$$

En el contexto de un sistema de comunicaciones digitales, el codificador se encarga de seleccionar un vector de código y_i que

aproxime en cierto sentido al vector de entrada x. El índice i del vector de código seleccionado es transmitido (como una palabra binaria) al receptor, donde el decodificador realiza un procedimiento de búsqueda en tablas y genera la reproducción y_i, la aproximación cuantizada del vector de entrada original.

El principal objetivo en el diseño de un cuantizador vectorial es el de encontrar un Alfabeto, una partición y una regla de decisión que maximizará la medida del desempeño total considerando la secuencia entera de vectores a ser codificados por el cuantizador. El desempeño total puede ser medido ya sea por medio de promedios estadísticos de una medida de distorsión conveniente o por medio de la consideración del peor valor de distorsión. El promedio estadístico de la distorsión de un cuantizador vectorial $Q(\cdot)$, puede ser expresado como :

$$D = Ed(X,Q(X)) = \int d(x,Q(x))f_X(x)dx, \qquad (4)$$

donde $f_X(x)$ es la función de probabilidad conjunta del vector X y la integración es una integral múltiple sobre un espacio de k dimensiones. Cuando el vector de entrada tiene una distribución discreta, entonces

$$D = Ed(X,Q(X)) = \sum_i d(x_i,Q(x_i))q_X(x_i) \qquad (5)$$

Para un Alfabeto dado, una partición óptima es la que satisface la condición de vecino más cercano, para cada i, todos los puntos más cercanos al vector de codigo y que a cualquier otro código de vector deberá ser asignado a la región R_i. Así, para un conjunto de niveles, Y, la partición de las celdas satisface

$$R_i \subset \{ x: d(x,y_i) \leq d(x,y_j);\ \text{para toda } j \neq i$$

esto es,

$$Q(x) = y_i \text{ solo si } d(x,y_i) \leq d(x,y_j) \text{ para toda } j$$

así, dado el decodificador, el encodificador es una transformación de distorción mínima o del vecino más cercano,

$$d(x,Q(x)) = \min_{y_i \in Y} d(x,y_i) \qquad (6)$$

Otra de las condiciones de optimilidad es que para una partición dada $\{R_i;\ i=1,\ldots,N\}$, los vectores de código satisfacen:

$$y_i = cent(R_i) \qquad (7)$$

Donde el centroide cent(R) ,de cualquier conjunto $R \in R^k$,es definido como el vector y el cual minimiza la distorsión entre un punto x en R y y ,promediados sobre la distribución de X dado que X pertenece a la región R. Así

$$y^* = cent(R) \text{ si } E\{d(X,y^*)|X\in R\} \leq E\{d(X,y)|X\in R\} \qquad (8)$$

o equivalentemente

$$cent(R) = \min_{y}{}^{-1} E\{d(X,y)|X\in R\}$$

Así,el centroide es aquel vector que en cierto sentido es un representante natural o un vector central del conjunto R y la distribución de probabilidad de R.

Para una medida de distorción de error cuadrático medio, un resultado estándar es que $E(||X - y||^2)$ es minimizado cuando y es la media de X. Aplicando esto a la esperanza condicionada

$$cent(R) = E(X|X\in R) \qquad (9)$$

Así, si se asume que cada punto en R tiene una probabilidad igual, entonces para la medida del error al cuadrado se reduce al promedio aritmético:

$$cent(R) = \frac{1}{L}\sum_{i=1}^{L} x_i \qquad (10)$$

para $x_i \in R, i=1,2,3,\ldots L$

Las condiciones necesarias para optimilidad proporcionan la base para un diseño óptimo para cuantizadores vectoriales. A continuación se presenta el algoritmo más importante de diseño de cuantizadores vectoriales, que es una versión generalizada del

Algoritmo de Lloyd desarrollado para cuantización escalar. Este algoritmo se conoce como LBG (Linde,Buzo,Gray). La operación básica del algoritmo está basado en una operación de modificación iterativa del Alfabeto.

Primero se presenta el caso general cuando la función de probabilidad conjunta del vector de entrada se conoce.

a) Dado un Alfabeto, $Y_m = \{y_i\}$, encuentre la partición óptima de las celdas de cuantización, esto es, use la condición del vecino más cercano para formar las celdas.

$$R = \{ x: d(x,y_i) \leq d(x,y_j); \text{ para toda } j \neq i \}$$

b) Usando la Condición de Centroide, encuentre el Y_{m+1}, el alfabeto de reproducción optimo (Alfabeto) para las celdas acabadas de encontrar.

Esta forma de la iteración de Lloyd requiere que la función de densidad de probabilidad (FDP) sea conocida y la geometria de la partición sea especificada para el cálculo de los centroides. La computación de los centroides en el paso b) es generalmente imposible por métodos analiticos. En la practica una integración número ca seria necesaria para encontrar los centroides. Sin embargo una descripción analitica de la FDP generalmente no está disponible en la mayoria de las aplicaciones. En su lugar una distribución muestral basada en las observaciones empiricas es utilizada para generar las iteraciones del Alfabeto.

A continuación se presenta la iteración de Lloyd para datos empiricos:

a) Dado el Alfabeto, $Y_m = \{y_i\}$, particione el conjunto de entrenamiento en los conjuntos de agrupamientos S_i, usando la condición del vecino más cercano:

$$S_i = \{x \epsilon T : d(x,y_i) \leq d(x,y_j); \text{ toda } j \neq i\}$$

b) Usando la condición de centroide, calcule los centroides para el conjunto de agrupamientos encontrado, para hallar un nuevo Alfabeto, Y_{m+1}.

Ahora que ya se definio la forma de la iteración de mejoramiento del Alfabeto, el diseño del algoritmo puede quedar como sigue:

1.- Empezar con un Alfabeto inicial Y_1. Poner m=1.

2.- Dado el Alfabeto, Y_m, realice la iteración de Lloyd para generar un Alfabeto mejorado Y_{m+1}.

3.- Calcular la distorsión promedio para Y_{m+1}. Si ha cambiado por una cantidad muy pequeña desde la última iteración, alto. En caso contrario poner m+1 → m y regresar al paso 1.

Varios criterios para parar pueden ser usados, una versión particular que es común y efectiva es probar si $(D_m - D_{m+1})/D_m$ esta abajo o arriba de un nivel apropiado.

Existen varias técnicas para la inicialización de los Alfabetos, la que nosotros utilizamos es la llamada técnica de separación. Esta técnica fue introducida por Linde y compañia, la idea básica es la siguiente: El Alfabeto globalmente óptimo de resolución cero de una secuencia de entrenamiento es el centroide de la secuencia entera. El código unico , y_0, en este Alfabeto puede ser separado en dos palabras de código, $y_0+\epsilon/2$ y $y_0-\epsilon/2$, donde ϵ es un vector de norma euclidiana pequeña. Este nuevo Alfabeto tiene dos palabras de codigo, el algoritmo de Lloyd puede ser ejecutado en este Alfabeto para producir un código de resolución 1. Cuando estan completos, todos los códigos del nuevo Alfabeto pueden ser separados, formando una suposición inicial para un Alfabeto de resolución igual a 2. Asi se continua de esta manera, usando un Alfabeto de resolución r para formar las condiciones iniciales de un Alfabeto de resolución r+1 por medio de separación.

MEDIDAS DE DISTORSION

Idealmente una medida de distorsión deberá ser manejable para que permita el análisis y el diseño, además que debe ser evaluable de tal forma que sirva en el proceso de codificación para la selección del vecino más cercano o la salida con menor distorsión. La más conveniente o más ampliamente usada es el error cuadrado o la distancia Euclidiana al cuadrado entre el vector de entrada X y el vector cuantizado $\hat{X} = Q(X)$, definida como:

$$d(X,\hat{X}) = ||X - \hat{X}||^2 = \sum_{k=1}^{N} (X_1 - \hat{X}_1)^2 \quad . \qquad (11)$$

El promedio de la distorsión de error cuadrático o también llamada la distorsión promedio se define como

$$D = Ed(X,\hat{X}) = E||X - \hat{X}||^2 \quad . \qquad (12)$$

Esta medida es frecuentemente asociada con la energía o potencia de la señal de error por esta razón tiene cierto encanto en adición de ser fácil de manejar. Otras medidas de distorsión pueden ser definidas para medir la disimilitud entre la entrada y los vectores de reproducción. Muchas de las medidas de interés en CV tienen la forma

$$d(X,\hat{X}) = \sum_{i=1}^{k} d_m(X_1,\hat{X}_1) \quad , \qquad (13)$$

donde $d_m(x,\hat{x})$ es una medida de distorsión escalar. De particular interés es el caso cuando la medida de distorción escalar esta dada por $d_m(x,\hat{x}) = |x - \hat{x}|^m$ para valores enteros positivos de m. Cuando m=1, se convierte en la norma l_1 del vector de error, $X-\hat{X}$. Cuando m=2, se obtiene la medida de error discutida anteriormente. Otra medida de distorsión de particular interés es la medida de error cuadrático con ventana

$$d(x,y)=(x-y)^T W(x-y), \qquad (14)$$

donde W es una matriz de peso simétrica y definida positiva .Esta medida incluye la distorsión usual de error cuadrático en el caso especial cuando W = I, la matriz identidad. En el caso que se tenga que W es una matriz diagonal con valores diagonales $w_{ii} > 0$ se tiene

$$d(x,y) = \sum_{i=1}^{k} w_{ii}(x_i - y_i)^2 \quad . \qquad (15)$$

Todas las medidas de distorción previamente consideradas tienen la propiedad que dependen de los vectores x y $\hat{x}$ solamente a través del

vector de error $x - \hat{x}$. Tales medidas de distorción que tienen la forma $d(x,\hat{x}) = L(x-\hat{x})$ son llamadas medidas de distorción de diferencias. Medidas de distorsión que no tienen esta forma pero que dependen de x y $\hat{x}$ en una forma más complicada han sido propuestas para sistemas de compresión de datos. En sistemas de compresión de voz la medida de distorsión de Itakura y Saito tiene gran importancia.

$$d(x,\hat{x}) = (x-\hat{x})R(x)(x-\hat{x})^t , \qquad (16)$$

donde por cada x, $R(x)$ es una matriz de $k \times k$ definida positiva. Esta medida de distorsión es de la misma forma que la definida por la ecuación (3.96).

En los sistemas tradicionales de LPC, los diferentes parametros son cuantizados separadamente, pero es natural pensar en cuantizar estos parámetros usando las técnicas de cuantización vectorial. Los parámetros que describen el modelo normalizado de LPC son cuantizados juntos como vector. Ya que el primer término es igual a 1, quisiéramos cuantizar el vector $(a_1, a_2, \ldots, a_k)$. Una medida de distorsión $d(a,\hat{a})$ entre a y su reproducción $\hat{a}$, puede ser vista como una medida de distorción entre dos filtros o modelos inversos normalizados (ganancia unitaria). Tal medida de distorsión ha sido proporcionada por Itakura-Saito y tiene la forma de la ecuación (3.98) con $R(a)$ la matriz de autocorrelación $\{r_a(k-j); k=0,1,\ldots,k-1; j=0,1,\ldots,k-1\}$. Todas las medidas de distorsión consideradas dependen de la forma de onda de la voz solamente a través de sus propiedades de segundo orden, es decir, usando la autocorrelación o sus modelos espectrales. Estas medidas de distorsión son más fácilmente definidas en el dominio espectral, aunque su evaluación es mas fácil de instrumentar sin hacer referencia a tal dominio. Estas medidas de distorsión pueden ser usadas entre procesos aleatorios, lo mismo que para el caso deterministico. Una medida de distorsión espectral es una función de dos densidades espectrales, f y $\hat{f}$, el cual asigna un numero no negativo $d(f,\hat{f})$ que representa la distorsión de usar $\hat{f}$ en lugar de f. La más común de tales medidas es la medida de distorsión de diferencias donde se utiliza una norma L_p en la diferencia $f-\hat{f}$. Estas son medidas o distancias en el sentido que

ellas satisfacen requerimientos de simetria $d(f,\hat{f})=d(\hat{f},f)$ y la desigualdad del triángulo

$$d(f,g) \leq d(f,h)+d(h,g)$$

Las medidas de distorsión que nosotros utilizamos dependen del logaritmo de la diferencia de los espectros, como resultado, la división de espectros es utilizada y así medidas de distorsión usando divisiones son utilizados

$$d(f,\hat{f}) = d(1,f/\hat{f}) = d(f/\hat{f},1)$$

En la generación de los Alfabetos para codificación de voz dos medidas de distorsión son efectivas: La medida de distorsión de Itakura-Saito (d_{IS}) y la medida normalizada de Itakura-Saito (d_{GN}). Para dos espectros en potencia $f(v)$ y $\hat{f}(v)$, la distorsión de Itakura-Saito es

$$d_{IS}(f,\hat{f}) = \int_{-\pi}^{\pi} \frac{dv}{2\pi} \left[\frac{f}{\hat{f}} - \ln \frac{f}{\hat{f}} - 1\right], \qquad (17)$$

y la distorsión normalizada de Itakura-Saito es

$$d_{GN}(f,\hat{f}) = d_{IS}\left(\frac{f}{\hat{\sigma}^2}, \frac{\hat{f}}{\hat{\sigma}^2}\right).$$

La aplicación de d_{IS} a la predicción lineal se hace aparente si f es una muestra de densidad espectral de voz y $\hat{f}$ es un modelo de reproducción del espectro de la forma

$$\hat{f}(v) = \frac{\hat{\sigma}^2}{|A(z)|^2} \qquad (18)$$

donde

$$A(z) = \sum_{k=0}^{M} a_k z^{-k} \quad \text{con } a_0 = 1 \qquad (19)$$

y $z = \exp(iv)$

$$d_{IS}\left(f, \frac{\hat{\sigma}^2}{|A(z)|^2}\right) = \int_{-\pi}^{\pi} \frac{dv}{2\pi} \left[\frac{f|A(z)|^2}{\hat{\sigma}^2} - \ln \frac{f|A(z)|^2}{\hat{\sigma}^2} - 1\right], \qquad (20)$$

para el primer termino de la integral

$$\int_{-\pi}^{\pi} \frac{dv}{2\pi} \frac{f|A(z)|^2}{\hat{\sigma}^2} = \int_{-\pi}^{\pi} \frac{dv}{2\pi\hat{\sigma}^2} f \left| \sum_{k=0}^{M} a_k e^{-jkv} \right|^2 \tag{21}$$

$$= \int_{-\pi}^{\pi} \frac{dv}{2\pi\hat{\sigma}^2} \left[f \sum_{k=0}^{M} a_k e^{jkv} \sum_{l=0}^{M} a_l e^{-jlv} \right]$$

$$= \frac{1}{\hat{\sigma}^2} \left[\sum_{l=0}^{M} \sum_{k=0}^{M} a_k a_l \int_{-\pi}^{\pi} \frac{dv}{2\pi} f(v)\, e^{j(k-l)v} \right]$$

$$= \frac{1}{\hat{\sigma}^2} \left[\sum_{l=0}^{M} \sum_{k=0}^{M} a_k a_l r(k-l) \right] . \tag{22}$$

$$= \frac{1}{\hat{\sigma}^2} \Big[a_0 a_0 r(0) + a_0 a_1 r(-1) + \ldots + a_0 a_m r(-M)$$
$$+ a_1 a_0 r(1) + a_1 a_1 r(0) + \ldots + a_1 a_m r(1-M)$$
$$\vdots$$
$$+ a_m a_0 r(0) + a_m a_1 r(-1) + \ldots + a_m a_m r(-M) \Big]$$

$$= \frac{1}{\hat{\sigma}^2} \Big[\sum_{k=0}^{M} a_k a_k r(0) + 2a_0 a_1 r(1) + 2a_0 a_2 r(2) + \ldots + 2a_0 a_m r(m)$$
$$+ 2a_1 a_2 r(1) + 2a_1 a_3 r(2) + \ldots + 2a_1 a_m r(m-1)$$
$$+ 2a_2 a_3 r(1) + 2a_2 a_4 r(2) + \ldots + 2a_2 a_m r(m-2)$$
$$+ \ldots +$$
$$+ 2a_{m-1} a_m r(1) \Big]$$

$$= \frac{1}{\hat{\sigma}^2} \left[\sum_{k=0}^{M} a_k a_k r(0) + 2 \sum_{n=1}^{M} \sum_{k=0}^{M-n} a_k a_{k+n} r(n) \right]$$

$$= \frac{1}{\hat{\sigma}^2} \left[r_a(0) r(0) + 2 \sum_{n=1}^{M} r_a(n)\, r(n) \right] , \tag{23}$$

donde

$$r_a(n) = \sum_{k=0}^{M-n} a_k a_{k+n} \quad . \qquad (24)$$

Para el segundo termino de la ecuación (3.99)

$$\int_{\pi}^{\pi} \ln(f/\hat{f}) \frac{dv}{2\pi} = \ln(\sigma^2/ \hat{\sigma}^2) \quad . \qquad (25)$$

Esta propiedad es debida a Grenader y Szegó [2] la cual ha sido utilizada en la literatura de predicción lineal.
Entonces, sustituyendo en la ecuación (17) las ecuaciones (23) y (25) se obtiene los siguiente

$$d_{IS}(f,\hat{f}) = \int_{-\pi}^{\pi} \frac{dv}{2\pi} [\frac{f}{\hat{f}} - \ln \frac{f}{\hat{f}} - 1] =$$

$$= \frac{1}{\hat{\sigma}^2} \left[r_a(0)r(0) + 2 \sum_{n=1}^{M} r_a(n)\, r(n) \right] - \ln(\sigma^2/ \hat{\sigma}^2) - 1 \quad , \qquad (26)$$

y para la distorsión de Itakura-Saito normalizada

$$d_{GN}(f,\hat{f}) = d_{IS}(\frac{f}{\sigma^2}, \frac{\hat{f}}{\hat{\sigma}^2}) =$$

$$= \left[r_a(0)r(0) + 2 \sum_{n=1}^{M} r_a(n)\, r(n) \right] - 1 \quad . \qquad (27)$$

Esta ecuación es una de las mas importantes cuando se realiza el reconocimiento de comandos hablados en tiempo real.

RECONOCIMIENTO

En la codificación de la voz usando cuantización vectorial, un solo Alfabeto es diseñado para una secuencia larga de entrenamiento, que representa toda la voz que puede ser codificada por el sistema. En el caso de reconocimiento de comandos hablados, para cada palabra se generan Alfabetos separados .Se diseña cada Alfabeto de una secuencia de entrenamiento conteniendo repeticiones de una de las palabras del vocabulario.

Por ejemplo, para una palabra determinada, se genera un secuencia de Alfabetos, para cada parte en que fue dividida la palabra, usando el algoritmo de diseño de un cuantizador vectorial, en una secuencia de entrenamiento con varias repeticiones de la palabra determinada. Para clasificar la palabra desconocida, primero se códifica cada una de las secciones de la palabra usando cada uno de los Alfabetos de secciónes múltiples y la distorsión promedio para cada sección de Alfabeto es guardada. La palabra desconocida es entonces clasificada de acuerdo al Alfabeto de secciones múltiples que da un promedio de distorsión minimo.

Para ser más precisos, sea V el número de palabras en el vocabulario de reconocimiento, sea T_k el número de palabras pronunciadas en la secuencia de entrenamiento usadas para diseñar el Alfabeto C_k para la k-ésima palabra del vocabulario, donde k=1,...,V. También sea F_{qk} el número de bloques en la q-ésima palabra de entrenamiento para C_k donde q=1,...,T_k, y finalmente ,sea U_{mqk} el bloque el m-ésimo bloque de la q-ésima palabra de entrenamiento para C_k donde m=1,...,F_{qk}. Entonces hay V Alfabetos de secciones múltiples C_k. Cada uno de ellos teniendo Alfabetos por sección C_{kj}. La sección de Alfabeto C_{kj} es diseñada usando usando n bloques por cada palabra de entrenamiento de la k-ésima palabra del vocabulario. Esto es, C_{kj} es diseñada usando los bloques U_{mqk}, donde m=(j-1)n+1,...,jn, and q=1,...,T_k. En particular C_{k1} es diseñado usando los primeros n bloques de cada una de las palabras pronunciadas en el entrenamiento de la k-ésima palabra, C_{k2} usando los segundos n bloques, etc. Finalmente, sea C_{kji}, i=1,...,N_{kj} el numero de palabras de código en la sección de Alfabeto C_{kj}.

Supongamos que una nueva palabra va a ser clasificada conteniendo L bloques, y P_l es el conjunto de estimadores de la autocorrelación del l-ésimo bloque (l=1,...,L). Ahora, sea D_k la distorsión promedio resultado de codificar la palabra desconocida con el Alfabeto C_k.

$$D_k = \frac{1}{L} \sum_{j=1}^{S_k} d_{kj} \quad (28)$$

Donde S_k es el número de secciones de Alfabetos en C_k, y

$$d_{kj} = \sum_{i=(j-1)n+1}^{min(jn,L)} min\, d(P_i, C_{kji}) ,$$

es la distorsión total de codificar la j-ésima sección de la entrada con la j-ésima sección del Alfabeto C_{kj} de C_k, donde n es el número de bloques por sección.

Entonces la palabra es clasificada como la r-ésima palabra del vocabulario de reconocimiento, donde

$$D_r = \min_k D_k \quad .$$

Para nuestro caso cada palabra es dividida en cuatro secciones. Cada sección cuenta con un Alfabeto de cuatro palabras de código. Se hace una detección de inicio y final tanto para la secuencia de entrenamiento como para las secuencias de clasificación. Basicamente, la detección de inicio, se hace cuando la señal excede ciertos umbrales de energía y la detección del final se realiza cuando habiéndose detectado el inicio se encuentra que la energía de la señal esta por debajo de cierto umbral de energía. Se toman ciertas precauciones, para evitar, que cuando el ruido ambiental exceda el umbral de inicio no se determine que una palabra nueva se este pronunciando. La forma de hacer esto es contando el número de bloques que hay entre el inicio y el final de la palabra, si este número es menor a cierto número preestablecido anteriormente se decide que no hay palabra nueva y que sólamente se trataba de ruido. El úmbral de energía E_{min} es calculado de la siguiente forma

$$E = \sum_{i=1}^{W} x_i^2 \quad ,$$

donde W es el ancho de la ventana de análisis y x_i son las muestras en el dominio del tiempo, de un convertidor de 14 bits a una tasa de muestreo de 8000 Hz, además de haberles efectuado las operaciones de preénfasis y la aplicación de una ventana de Hamming.