RECONOCIMIENTO DE FONEMAS POR COMPUTADOR

Ing. Tania Pérez
ESCUELA POLITECNICA NACIONAL

Ing. Gualberto Hidalgo
ESCUELA POLITECNICA NACIONAL

RESUMEN

El presente trabajo aborda el tema del reconocimiento automático de fonemas. Entendiéndose por fonema la unidad fonética básica de un lenguaje hablado. Primeramente se da una idea rápida del grado de complejidad que conlleva la realización de esta tarea. Luego se exponen en forma sintética las técnicas utilizadas para lograr el objetivo propueso y que son: la predicción lineal que determina los coeficientes de predicción de cada fonema, la cuantización vectorial que permite el cálculo de los "centroides" de los diversos fonémas y la distorsión de Itakura que posibilita la identificación de los fonemas por mínima distancia al "centroide". Finalmente se presentan algunos resultados parciales.

ABSTRACT

The present work is concerned with the theme automatic phoneme recognition. Phoneme meaning the basic phonetic unit of speech. Firstly the complexity of the undertaken task is briefly exposed. Then the techniques used for this purpose are syntetically explained, these techniques are: Linear Prediction for the linear prediction coefficients evaluation, Vector Quantization wich allows the phoneme "centroids" calculation, and the Itakura distorsion which permits the unknown phoneme identification by minimal "distance" to the "centroid". Finally some partial results are presented.

Introducción [1]

La tecnología de computación se ha desarrollado en forma explosiva en los últimos años, quedando sin embargo un campo que no ha seguido el mismo ritmo y es el referente al procesamiento de señales de la comunicación humana a través del lenguaje y la voz articulada.

Existe la antigua aspiración de diseñar máquinas capaces de responder a la locución humana, que es lo que se conoce como reconocimiento electrónico del habla.

La posibilidad de ingresar datos y programas al computador a través de la voz sin hacer uso de las manos es otra de las aspiraciones en este mismo campo.

Las tecnologías avanzadas de procesamiento digital de señales, la programación dinámica, los modelos ocultos de Markov, las redes neurales son muy importantes en el avance del reconocimiento electrónico del habla, pero no pueden sustituir al conocimiento básico de la lingüística, acústica y fonética. No se puede dejar al computador el conocimiento implícito del código del habla.

En el campo de la síntesis de voz se ha adelantado mucho y se pueden encontrar chips a bajo precio, capaces de emitir palabras con voz que incluso puede sonar como humana, aunque con vocabulario limitado.

En reconocimiento electrónico de voz, en cambio, los avances han sido menos notables, pero algo se ha logrado y ahora es posible añadir a la máquina la capacidad de escuchar la voz humana y reconocer que palabras se están pronunciando. Sin embargo, el vocabulario que puede ser reconocido es excesivamente limitado y la lista completa de las palabras a reconocerse debe ser declarada con anterioridad.

Existen algunas categorías de reconocimiento de voz en las que se investiga actualmente.

- Reconocimiento de locución independiente del locutor: cualquier persona puede hablar y la máquina debería ser capaz de entender el vocabulario utilizado.

- Reconocimiento de locución dependiente del locutor: se puede tener un limitado número de personas que tienen acceso a la máquina, y cuyas voces han sido registradas previamente. El computador debe ser capaz de reconocer la locución de cualquiera de ellas.

- Identificación del locutor a través de su voz: la máquina debe identificar que persona, de un grupo limitado de individuos, es la que habla.

- Verificación del locutor a través de su voz: la máquina debe estar en capacidad de distinguir si la persona que habla es o no, la que pretende ser.

La presente investigación se ocupa del reconocimiento de fonemas, es decir de aquellos sonidos básicos que constituyen el acervo fonético de un idioma. La razón para abordar el problema desde este punto de vista es que en esta forma, a partir de un número limitado de fonemas, se puede llegar al reconocimiento de un vocabulario ilimitado; mientras que si el reconocimiento se inicia a nivel de palabras el número de las mismas que es reconocible por la máquina será siempre limitado.

Una tarea compleja

Dada la capacidad de los computadores para realizar operaciones complejas y cálculos matemáticos embrollados, se pensó en un comienzo que, desarrollar programas para el reconocimiento automático del habla, sería una tarea sin mayores complejidades. Cuando han transcurrido algunos decenios de activa investigación en este campo, se ha llegado a la penosa conclusión de que la mencionada tarea era mucho más elusiva de lo que se pensaba.

Varias son las razones que explican este hecho:

En primer lugar el oído humano dispone de un entrenamiento que, para la lengua materna al menos, representa el ejercicio de toda una vida. El ser humano está dotado de una precomprensión del lenguaje que le permite anticipar palabras, sobreentenderlas o corregirlas. El computador, en cambio, está adaptado para operaciones de tipo repetitivo y debe ser programado minuciosamente para

hacer frente a cambios inesperados. La naciente técnica de Inteligencia Artificial tiende justamente a suplir esta debilidad del computador.

Precisando un poco más en las raíces de la complejidad del reconocimiento automático de locución podemos mencionar entre otros los siguientes factores:

La increible variedad de las voces humanas, que comenzando con los registros básicos de voces masculinas, femeninas e infantiles, se diversifica al infinito en los timbres, acentos y estados de ánimo individuales.

El influjo que tienen unos fonemas con otros en virtud del cual una vocal que sigue a una "p" no es enteramente igual a la misma vocal después de una "d", por ejemplo: Este fenómeno de mutua interacción entre los diversos fonemas recibe el nombre de coarticulación.

A diferencia de lo que sucede con el lenguaje escrito, en el que los diversos signos o letras están claramente individualizados como entidades discretas, el lenguaje hablado es un evento continuo en el que los fonemas se concatenan unos con otros, teniéndose sonidos, como el caso de los fonemas plosivos (k, p, t, ch) que son de caracter esencialmente cambiante, mientras que sonidos que alcanzan regularmente un estado estacionario, como son las vocales pueden llegar a perderlo si la pronunciación de las palabras es suficientemente rápida.

Métodos utilizados para el reconocimiento del habla

Dada la complejidad de la labor a efectuarse, son innumerables los métodos que se han ideado para su realización.

Método de modelos probabilísticos

Este método consiste en obtener un modelo probabilístico del lenguaje. Una vez establecido este modelo, el reconocimiento de los fonemas se realiza etiquetando el fonema dado como aquel que tiene la máxima probabilidad de ajustarse al modelo obtenido.

Métodos de distancia mínima

En este caso se caracteriza un fonema dado por un vector de parámetros, que pueden ser los coeficientes de predicción, los coeficientes de reflexión, etc. Se obtiene luego un promedio estadístico de estos vectores de parámetros denominado "centroide". Los fonemas se identifican por distancia mínima a este "centroide".

Otra forma de caracterizar los métodos de análisis del lenguaje hablado para su reconocimiento los divide en:

Métodos ascendentes En estos métodos se parte de las características acústicas de los fonemas, esto es, espectros, periodicidad, aperiodicidad, etc. y se trata de ensamblar a partir de los mismos, en forma ascendente, sílabas, palabras y frases.

Métodos descendentes En estos métodos se parte de un modelo y de una precomprensión del lenguaje y a partir de los mismos se formulan hipótesis acerca de la secuencia más probable de eventos fonéticos de una locución dada.

Se puede afirmar que dada la complejidad de la tarea a realizarse, tanto el método ascendente como el descendente, a pesar de sus méritos, son insuficientes por si solos, para arrojar resultados plenamente satisfactorios. Por tanto en la actualidad, se tiende a utilizar métodos que participan de las características de ascendentes y descendentes simultáneamente.

El presente trabajo de investigación ha utilizado, hasta el momento sólo el método ascendente. La razón es que hasta el presente se cuenta con una base muy limitada de datos que hace problemática la obtención de un modelo confiable de lenguaje hablado.

El conjunto de técnicas utilizadas para el reconocimiento de fonemas en el presente trabajo se evidencia en la siguiente figura. (Fig. 1).

Como se desprende de la misma, después de la conversión A/D el proceso se bifurca en dos: uno de aprendizaje y uno de reconocimiento.

El proceso de aprendizaje o entrenamiento comienza con la segmentación manual que da como resultado el arreglo de los datos para el cálculo de "centroides". Este cálculo concluye el proceso de aprendizaje.

El proceso de reconocimiento se inicia con el algoritmo preclasificador que da como resultado la clasificación de la onda del habla en cinco grupos:

Vocales:	a, e, i, o, u
Nasales:	m, n
Consonantes voceadas:	b, d, l, r, ll, g
Fricativas:	s, f, j
Oclusivas:	p, t, k, ch

El proceso de reconocimiento continúa con la identificación final de los fonemas, la cual se realiza por distancia mínima utilizando la distancia de Itakura.

De la breve explicación que antecede se desprende que para el presente trabajo se utilizan básicamente tres técnicas que son:

1) La predicción lineal que sirve para la obtención de los coeficientes de predicción que caracterizan a los diversos fonemas.

2) La cuantización vectorial que permite la determinación de los "centroides".

3) La distancia de Itakura que facilita la evaluación de la "distancia" del fonema a reconocerse respecto de estos "centroides".

Expondremos brevemente en que consisten estas técnicas.

Predicción Lineal [2]

Un método importante de análisis de señales discretas en el dominio del tiempo es la "predicción lineal", mediante este método la señal es obtenida a partir de la combinación lineal de sus valores anteriores y presentes y de los valores pretéritos de una entrada hipotética al sistema, cuya salida viene a constituir la señal dada.

Cualquier señal análoga en el dominio del

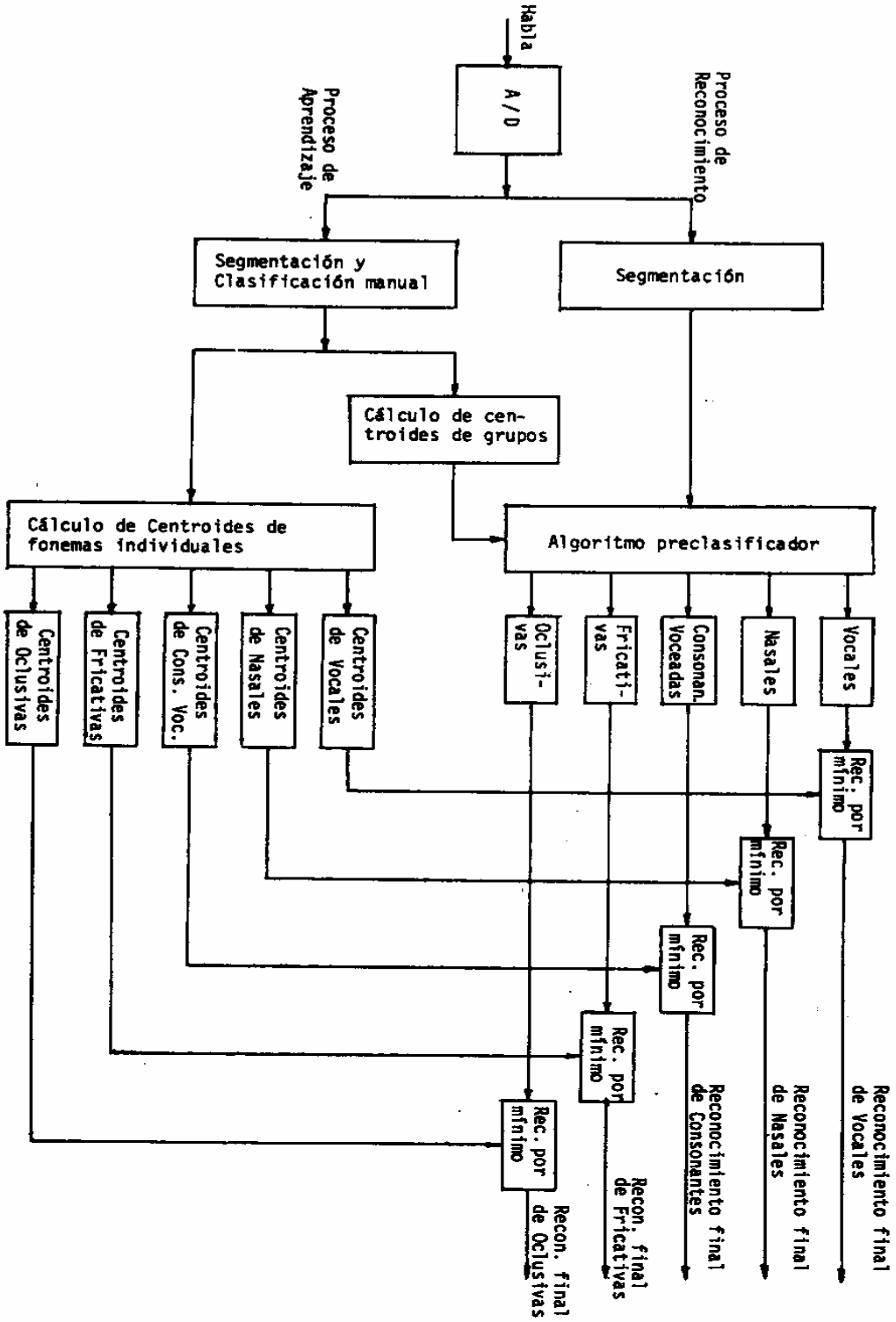
Habla
A / D
Proceso de Aprendizaje
Proceso de Reconocimiento
Segmentación y Clasificación manual
Segmentación
Cálculo de centroides de grupos
Cálculo de Centroides de fonemas individuales
Algoritmo preclasificador
Centroides de Oclusivas
Centroides de Fricativas
Centroides de Cons. Voc.
Centroides de Nasales
Centroides de Vocales
Oclusivas
Fricativas
Consonan. Voceadas
Nasales
Vocales
Rec. por mínimo
Rec. por mínimo
Rec. por mínimo
Rec. por mínimo
Rec. por mínimo
Reconocimiento final de Vocales
Reconocimiento final de Nasales
Reconocimiento final de Consonantes
Recon. final de Fricativas
Recon. final de Oclusivas

FIG. 1.

tiempo s(t) puede ser muestreada obteniéndose una sucesión de pulsos discretos s(nT) = s(n).

Uno de los modelos más utilizados es cuando se considera una señal s_n como la salida de un sistema con alguna entrada desconocida u_n, de tal forma que se cumpla la siguiente relación:

$$s_n = -\sum_{k=1}^{p} a_k s_{n-k} + G \sum_{l=0}^{q} b_l u_{n-l} \quad , \; b_0=1 \qquad (1)$$

Donde a_k , $1 \le k \le p$, b_l , $1 \le l \le q$,y la ganancia G son los parámetros del sistema hipotético. La ecuación (1) indica que la señal de salida s_n, es una función lineal de salidas pretéritas y de las entradas presentes y pasadas. Es decir, la señal s_n es predecible mediante las combinaciones lineales de entradas y salidas pretéritas. De donde se deduce el nombre de "predicción lineal".

La ecuación (1) puede especificarse en el dominio de la frecuencia tomando la transformada z a los dos lados. Si H(z) es la función de transferencia del sistema se tiene:

$$H(z)=\frac{S(z)}{U(z)} = G \frac{1+\sum_{l=1}^{q} b_l z^{-l}}{1+\sum_{k=1}^{p} a_k z^{-k}} \qquad (2)$$

Donde $S(z)= \sum_{-\infty}^{\infty} s_n z^{-n}$ es la transformada z de s_n y U(z) es la transformada de u_n. H(z) constituye el modelo polo-cero generalizado. Las raices polinomiales del numerador y denominador constituyen los ceros y polos respectivamente.

Existen dos casos especiales de este modelo que presentan interés:

1) Modelo todo ceros: $a_k = 0$ $\quad 1 \le k \le p$
2) Modelo todo polos: $b_l = 0$ $\quad 1 \le l \le q$

El modelo todo-polos es conocido como modelo autoregresivo (AR) y es el que se utilizará en el presente trabajo.

<u>Estimación de Parámetros</u>

Para el modelo todo-polos, se asume que la señal s_n está dada como una combinación lineal de sus valores pretéritos y alguna entrada u_n.

$$s_n = -\sum_{k=1}^{p} a_k s_{n-k} + G u_n$$

G factor de ganancia.

La función de transferencia H(z) se reduce a:

$$H(z)=\frac{G}{1+\sum_{k=1}^{p} a_k z^{-k}}$$

Este modelo está esquematizado en las figuras (2) y (3) en los dominios del tiempo y la frecuencia, respectivamente.

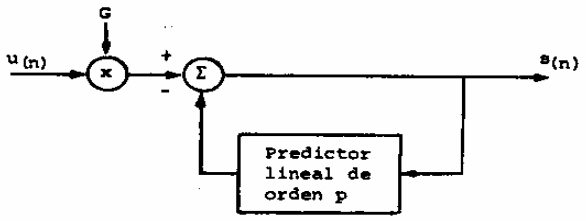

Fig. 2. Modelo AR (dominio del tiempo)

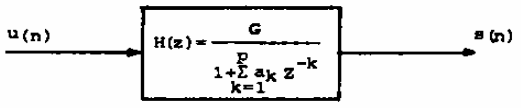

Fig. 3. Modelo AR (dominio de frecuencia)

Dada una señal particular s_n, el problema es determinar de alguna forma los coeficientes de predicción a_k y la ganancia G. Se utilizará en primer lugar el método de minimos cuadrados partiendo de consideraciones intuitivas. Se asumirá primero que s_n es una señal determinística y luego que s_n es una muestra de un proceso aleatorio.

<u>Método de minimos cuadrados</u>

Se asume que la entrada u_n es totalmente desconocida, lo cual es cierto en muchas aplicaciones (como el lenguaje por ejemplo). Entonces s_n puede ser predecible solo en forma aproximada partiendo de la suma de muestras pretéritas.

$\tilde{s}_n$ aproximación de s_n

$$\tilde{s}_n = -\sum_{k=1}^{p} a_k s_{n-k}$$

El error entre la señal verdadera s_n y el valor predicho $\tilde{s}_n$ está dado por:

$$e_n = s_n - \tilde{s}_n = s_n + \sum_{k=1}^{p} a_k s_{n-k}$$

e_n se conoce también con el término "residual"

En el método de cuadrados minimos los parámetros a_k son obtenidos como el resultado de la minimización del error cuadrático total con respecto a cada uno de los parámetros a_k.

1) Señal deterministica

El error cuadrático total E sería:

$$E = \sum_n e_n^2 = \sum_n (s_n + \sum_{k=1}^{p} a_k s_{n-k})^2 \qquad (3)$$

Todavía sin especificar el rango de la suma y minimizando E se tiene lo siguiente:

$$\frac{\delta E}{\delta a_i} = 0 \qquad 1 \le i \le p \qquad (4)$$

donde δ derivada parcial

con el resultado:

$$\sum_{k=1}^{p} a_k \sum_n s_{n-k}\, s_{n-i} = -\sum_n s_n\, s_{n-i} \qquad (5)$$

Estas se denominan ecuaciones normales. Para cualquier definición de la señal s_n, la ecuación (5) forma un conjunto de p ecuaciones con p incógnitas, el cual puede ser resuelto para los coeficientes de predicción (a_k, $1 \le k \le p$) que hacen mínimo el error en (3).

El mínimo error cuadrático total Ep se obtendría expandiendo (3) y reemplazando en (5).

$$E_p = \sum_n s_n^2 + \sum_{k=1}^{p} a_k \sum_n s_n\, s_{n-k} \qquad (6)$$

En cuanto al rango de suma n existen 2 casos de interés, los cuales darán dos métodos distintos de estimación de parámetros: el de autocorrelación y el de covarianza.

a) Método de autocorrelación. Mediante el cual se asume que el error en (3) es minimizado con una duración infinita. $-\infty < n < \infty$
Las ecuaciones (4) y (5) se reducen a:

$$\sum_{k=1}^{p} a_k\, R(i-k) = -R(i), \qquad 1 \le i \le p \qquad (7)$$

$$E_p = R(0) + \sum_{k=1}^{p} a_k\, R(k) \qquad (8)$$

donde $R(i) = \sum_{-\infty}^{\infty} s_n\, s_{n-i}$ es la función de autocorrelación de la señal s_n. R(i), es una función par de i. $R(-i) = R(i)$.

Los coeficientes R(i-k) forman la matriz de autocorrelación, la cual es simétrica y Toeplitz (Todos los elementos de sus diagonales son iguales).

La ecuación (7) expandida en forma matricial tendrá la siguiente forma:

$$\begin{bmatrix} Ro & R1 & R2 & \dots & Rp-1 \\ R1 & Ro & R1 & \dots & Rp-2 \\ R2 & R1 & Ro & \dots & Rp-3 \\ . & . & . & . & \\ . & . & . & . & \\ . & . & . & . & \\ Rp-1 & Rp-2 & Rp-3 & .. & Ro \end{bmatrix} \begin{bmatrix} a1 \\ a2 \\ a3 \\ . \\ . \\ . \\ ap \end{bmatrix} = - \begin{bmatrix} R1 \\ R2 \\ R3 \\ . \\ . \\ . \\ Rp \end{bmatrix}$$

En la práctica la señal s_n interesa en un intervalo finito. Entonces se multiplica la señal s_n por una función "ventana" w_n para obtener una señal s_n' que es cero fuera de dicho intervalo
$0 \le n \le N-1$

$$s_n' = \begin{cases} s_n\, w_n, & 0 \le n \le N-1 \\ 0 & \text{fuera del intervalo} \end{cases}$$

Para el presente trabajo se utiliza la ventana de Hamming cuya función es: $w_n = 0.56-0.46*\cos[2\pi(I-1)/(N-1)]$

donde $0 \le I \le N-1$ y N número de muestras.

La función de autocorrelación está dada entonces por:

$$R(i) = \sum_{n=0}^{N-1-i} s_n'\, s_{n+i}', \qquad i \ge 0$$

b) Método de covarianza. Por este método se asume que el error E en (3) está minimizado en un intervalo finito $0 \le n \le N-1$

2. Señal aleatoria: Para una señal aleatoria, el error e_n es también una muestra de un proceso aleatorio. Con el método de mínimos cuadrados y haciendo mínimo su valor se tiene:

$$E_p = \epsilon(s_n^2) + \sum_{k=1}^{p} a_k\, \epsilon(s_n\, s_{n-k})$$

Para un proceso estacionario se tiene:

$$\epsilon(s_{n-k}\, s_{n-i}) = R(i-k)$$

donde: ϵ valor esperado.
R(i) autocorrelación del proceso.

Cálculo de la Ganancia

Para una entrada de impulso a un filtro todo polo se puede demostrar que la ganancia debe ser igual a:

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

Cálculo de los parámetros de predicción

Los coeficientes de predicción a_k, $1 \le k \le p$ se pueden calcular resolviendo un conjunto de p ecuaciones con p incógnitas que corresponde a la ecuación (7):

$$\sum_{k=1}^{p} ak\, R(i-k) = -R(i) \qquad 1 \le i \le p$$

Existen varios métodos estandarizados para resolver estas ecuaciones. Uno de ellos es el algoritmo de Levinson y Durbin que está dado en el siguiente conjunto de ecuaciones:

a) $E_0 = R(0)$

b) $k_i = -[R(i) + \sum_{i=1}^{i-1} a_j^{(i-1)}\, R(i-1)]/E_{i-1}$

c) $a_i^{(i)} = k_i$

d) $a_j^{(i)} = a_j^{(i-1)} + k_i\, a_{i-j}^{(i-1)}$
$1 \le j \le i-1$

e) $E_i = (1-k_i^2)\, E_{i-1}$

Las ecuaciones a), b), c), d), e), se resuelven recursivamente para i = 1, 2 ....., p. La solución final está dada por

$$a_j = a_j^{(p)} \quad , \qquad 1 \leq j \leq p$$

Si todos los R(i) se normalizan dividiendo para R(0) se tienen los coeficientes de autocorrelación normalizados:

$$r(i) = \frac{R(i)}{R(0)} \leq 1$$

Si los coeficientes de autocorrelación están normalizados, el mínimo error Ei también resulta normalizado y dividido para R(0).

$$V_i = \frac{E_i}{R(0)} = 1 + \sum_{k=1}^{i} a_k \, r(k)$$

Cuantización Vectorial [3]

La cuantización vectorial es una extensión del concepto de cuantización escalar.
La cuantización escalar la podemos explicar en base a la fig. 4

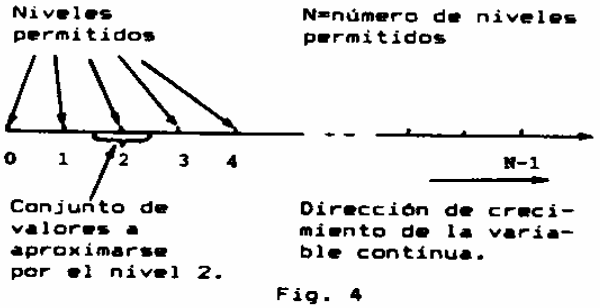

Fig. 4

De acuerdo a la fig. 4 se puede ver que la cuantización escalar consiste en establecer un conjunto de N valores discretos, generalmente equidistantes, a los cuales deberán aproximarse por distancia mínima, cualquiera de los infinitos valores que caen entre nivel y nivel. La cuantización vectorial explota el mismo concepto pero refiriéndolo al espacio multidimensional. Supongamos por ejemplo que deseamos cuantizar vectorialmente una magnitud bidimensional. Ver fig.5.

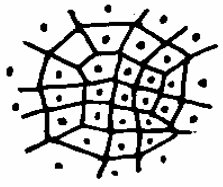

Fig. 5. Cuantización Vectorial (Vector de 2 dimensiones).

En la fig. 5 cada una de las regiones incluidas en un contorno poligonal representan los infinitos valores bidimensionales a codificarse mediante un solo vector representado por el punto interior a cada uno de los polígonos y que se conoce con el nombre de "centroide" de dichos vectores. Si extendemos este concepto al espacio n-dimensional, tendremos que en este caso dicho espacio se dividirá en "volúmenes" n-dimensionales, que delimitan regiones cerradas en los que caen infinitos vectores, todos los cuales deberán codificarse mediante un solo vector n-dimensional, interior a estos volúmenes cerrados y que representa el "centroide" de los mismos.

Así como en la cuantización escalar es el número de niveles el que determina el número de bits necesarios para la transmisión, así también en la cuantización vectorial será el número de polígonos en el plano, o de "volúmenes" en el espacio n-dimensional el que determina los niveles de cuantización.

Como se sabe la cuantización escalar al restringir el número de niveles de la variable a transmitirse permite reducir el ancho de banda de transmisión. Se comprende que este efecto será tanto más notable cuanto mayor sea el número de dimensiones del vector a transmitirse. En efecto la aproximación de los infinitos vectores que caen dentro de un "volumen" por un único punto interior que es el "centroide", produce una notable reducción de los bits de transmisión. Así es como la cuantización vectorial permite una codificación en la que el número promedio de los bits de transmisión tiende a reducirse hasta coincidir con la entropía de la fuente.

La cuantización vectorial utilizada en el presente trabajo opera con vectores (p + 1)-dimensionales, constituidos por los (p + 1) coeficientes de predicción lineal:

$$a_0, a_1, \ldots\ldots\ldots, a_p$$

La dimensionalidad de la cuantización vectorial dependerá por tanto, del número de coeficientes a utilizarse.

Para la determinación del número de coeficientes a utilizarse se aplica la regla siguiente:

Número de coeficientes = frecuencia de muestreo en kHz + 3 o 4 unidades.

Para el presente trabajo se ha adoptado la frecuencia de muestreo de 12 kHz. El número de coeficientes será por tanto de 15 o 16. O sea la cuantización vectorial utilizada, opera en el espacio de 15 o 16 dimensiones.

Cálculo de los "Centroides" [3]

Para el cálculo de los "centroides", se pueden utilizar 2 técnicas:

1. Obtener los coeficientes de predicción lineal para cada uno de los fonemas cuyo "centroide" se va a calcular. Luego evaluar la media aritmética de la suma vectorial de todos estos coeficientes. Esta media aritmética será el "centroide" de dicho fonema. Para la presentación de resultados denominaremos a este el Método 1.

2. Obtener una sola matriz de autocorrelación que sea la media aritmética de la suma de las matrices de autocorrelación

de todos los fonemas cuyo centroide se calcula. Utilizar luego esta matriz promedio para el cálculo de los coeficientes de predicción. Este vector de coeficientes será el "centroide" buscado. Para la presentación de resultados este método se conocerá con el nombre de Método 2.

Si se examina el tiempo de computación empleado por los dos métodos, es indudable que debe preferirse el segundo. En efecto este último calcula los coeficientes de predicción una sola vez, mientras que el primero los calcula tantas veces cuantos fonemas existen. En cuanto a los resultados obtenidos puede decirse que los mismos son semejantes, con una ligera ventaja para el segundo método. Esta ventaja debe entenderse en el sentido de que el segundo método permite un mayor porcentaje de reconocimiento correcto de fonemas. Ver más adelante tablas 1 y 2

Distorsión de Itakura [3] , [4]

Cuando se cuantiza escalarmente, se genera lo que se conoce como ruido de cuantización, que es igual a la diferencia entre el valor real de la muestra y el valor del nivel permitido más cercano, al que se aproxima dicha muestra. La razón por la cual una muestra se cuantiza aproximándola al valor más cercano es justamente para minimizar el ruido de cuantización. Igualmente cuando se cuantiza un vector, para que el ruido (que en este caso recibe el nombre de distorsión) sea mínimo, se escoge el valor del centroide más cercano al fonema dado.

La extensión de estos conceptos al reconocimiento de fonemas es directa. En este caso nos encontramos con vectores desconocidos que ocupan cualquier posición en el espacio n-dimensional. Para identificarlos evaluaremos su "distancia" (o distorsión) respecto de los diferentes "centroides", que previamente deberán haber sido calculados. Se considerará que el vector desconocido se identifica con el vector del "centroide" más cercano. Por lo tanto debe existir un método para evaluar la "distancia" (o distorsión) del fonema dado al conjunto preestablecido de "centroides". A primera vista podría pensarse en utilizar la distancia euclidiana:

$$d(\bar{a},\bar{\hat{a}}) = (a_0 - \hat{a}_0)^2+(a_1-\hat{a}_1)^2+\ldots+(a_p-\hat{a}_p)^2$$

donde $\bar{a} = (a_0, a_1, \ldots, a_p)$ vector de coeficientes del fonema dado.

$\bar{\hat{a}} = (\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_p)$ vector de coeficientes del centroide previamente calculado.

Sin embargo, este método no da resultados satisfactorios. F. Itakura [4] propuso la siguiente fórmula para evaluar la distorsión.

$$d(\bar{a},\bar{\hat{a}}) = (\bar{a}_i - \bar{\hat{a}}_i)^T W (\bar{a}_i - \bar{\hat{a}}_i)$$

$$0 \leq i \leq p$$

donde: $\bar{a}_i - \bar{\hat{a}}_i$, $0 \leq i \leq p$, vector columna de las diferencias de los coeficientes del mismo orden del fonema dado y de los "centroides".

T significa traspuesto.

W representa una matriz de ponderación de orden $(p + 1) \times (p + 1)$ y está constituida por la matriz de autocorrelación del fonema a identificarse. La distancia o distorsión dada por la fórmula anterior recibe el nombre de "Medida de distorsión de Itakura". Es la que se utiliza en el presente trabajo.

Resultados

Aún cuando la base de datos con que se cuenta es limitada, especialmente en lo que respecta a consonantes, se juzgó que el número de vocales era suficiente para garantizar un primer ensayo de la aplicabilidad de los conceptos anteriormente enunciados al reconocimiento de fonemas.

Se dispone de:
196 vocales a
56 vocales e
94 vocales i
122 vocales o
54 vocales u

Todas estas vocales forman parte de palabras (no son vocales aisladas) y pertenecen todas a personas del sexo masculino.

Los resultados de reconocimiento obtenidos se presentan en el cuadro a continuación:

Tabla 1. Cuadro de sustitución de vocales. Método 1

	a 196	e 56	i 94	o 122	u 54	% Rec. Correcto	% Rec. Correcto promedio total
a	192	2	0	2	0	97.96	
e	0	52	2	2	0	92.86	
i	0	0	94	0	0	100	95.98
o	0	0	0	120	2	98.36	
u	0	0	0	5	49	90.7	

Tabla 2. Cuadro de sustitución de vocales. Método 2

	a 196	e 56	i 94	o 122	u 54	% de Rec. Correcto	% Rec. Correcto promedio total
a	195	0	0	1	0	99.49	
e	1	53	0	0	2	94.64	
i	0	1	93	0	0	98.94	97.38
o	1	0	0	119	2	97.54	
u	0	0	0	2	52	96.30	

Dada la imposibilidad por el momento de obtener nuevos datos la tabla de sustitución arriba presentada, se ha realizado con los mismos fonemas que sirvieron para el cálculo de los "centroides". En cuanto se disponga de nuevos datos se procederá a evaluar la técnica anteriormente expuesta con fonemas que no pertenezcan a la población de entrenamiento, esto es, aquella que sirvió para el cálculo de centroides.

Conclusiones:

Aunque los resultados obtenidos hasta el momento deben considerarse muy parciales, permiten concluir favorablemente respecto de la viabilidad del reconocimiento de fonemas por computador.

Como se ha indicado anteriormente las técnicas aplicadas hasta el instante presente pertenecen al método ascendente: reconocimiento de fonemas a partir de sus características fonéticas. La culminación del

proyecto seguramente exigirá la complementación del mismo con el método descendente: reconocimiento de fonemas a partir de un modelo probabilistico del lenguaje hablado. Para esto se ha pensado en la utilización de la técnica conocida con el nombre: Modelos ocultos de Markov. [5].

Referencias

[1] Geoff Bristow Ed. "Electronic Speech Recognition", McGraw-Hill, 1986.

[2] John Makhoul, "Linear Prediction: A Tutorial Review", IEEE Proceedings, Vol.63, No. 4, Abril 1975, pp. 561-580.

[3] John Makhoul, Salim Roucos, Herbert Gish, "Vector Quantization in Speech Coding", IEEE Procedings, Vol.73, No. 11, Noviembre 1985, pp. 1551-1588.

[4] Fumitada Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition", IEEE Trans. Acoust., Speech, and Signal Processing, Vol. ASSP-23, No. 1, Febrero 1975, pp. 67-72.

[5] L. R. Rabiner, B. H. Juang, "An Introduction to Hidden Markov Models", IEEE ASSP Magazine, Vol. 3, No. 1, Enero 1986, pp. 4-16.

HIDALGO, GUALBERTO

Ingeniero en Electrónica y Telecomunicaciones graduado en la Escuela Politécnica Nacional en el año 1974. Obtuvo el titulo de Master en Ingenieria de Comunicaciones en el Imperial College de Londres en 1979. Actualmente es estudiante externo de la Universidad de Londres y realiza investigaciones en reconocimiento de fonemas.

PEREZ, TANIA

Ingeniero en Electrónica y Telecomunicaciones graduada en el Instituto de Comunicaciones Bonch Bruyevich. Leningrado, 1977. Actualmente realiza estudios de postgrado en Computación e Informática en la Escuela Politécnica Nacional y participa en el proyecto sobre reconocimiento de fonemas.