UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

Facultad de Ingenieria

REALIZACION DE UN SISTEMA TRADUCTOR VOZ - TEXTO

T E S I S

Que para obtener el título de :

INGENIERO EN COMPUTACION

Presenta :

ANDRES GOYTIA IRALA

Director: Luis Andres Buzo De la Peña

México, D. F. 1985

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

I N D I C E

CAPITULO 1

INTRODUCCION

Desde tiempos ancestrales, la voz ha sido el medio natural mediante el cual cada ser humano se ha comunicado con sus semejantes, la cual subsiste hasta nuestros tiempos. El advenimiento de nuevas tecnologías, tales como el micrófono, la radio y demás medios para la captación, transformación, difusión, recepción y reconstrucción de la voz, han permitido ampliar esta capacidad de comunicación.

Actualmente, la importancia de la voz no ha disminuído y, podemos mencionar algunas áreas en que adquiere particular importancia, como son:

- Interacción Hombre-Máquina
- Ayuda a minusválidos
- Comunicaciones

El número de aplicaciones prácticas que tiene la comunicación hombre-máquina, por medio de voz, ha estimulado tanto al mundo científico como al tecnológico. Ejemplo de esto es la existencia de máquinas que se comunican con el hombre mediante voz sintética. Otra forma de interacción hombre--máquina es el poder tener máquinas que entiendan el medio natural de comunicación del hombre, para que de esta forma deje de ser en el lenguaje de las máquinas (botones, palancas, etc.) La solución de este problema está basada en el análisis de -- las señales de voz, lo cual consiste en la extracción de algunas características de la señal para que de esta forma se puedan identificar los distintos sonidos emitidos por el hombre y, posteriormente aplicar reglas de concatenación de sonidos y reglas gramaticales para así obtener la palabra o comando dado en forma oral. Esta forma de comunicación con las máquinas ha sido elegida para la interacción con la que será la computadora de la quinta generación [1, 2, 3] , en la que --

esta función será llevada a cabo por el sistema básico de -- aplicación, compuesto a su vez por varios sistemas, entre -- los que podemos destacar al sistema para entender voz, cuya_ función es la identificación del usuario, como parte de su - sistema de propósito general activado por voz.

Para los minusválidos que han perdido alguna de sus capa cidades motoras, la voz resulta ser una buena alternativa -- para mejorar su situación, pues pueden controlar una silla - de ruedas, abrir o cerrar puertas y realizar algunas otras - actividades, todo ello accionado mediante su voz. Los sis temas de procesamiento de voz pueden también emplearse en el campo de la educación de infantes con problemas auditivos, - los cuales a falta de la realimentación natural del oido, -- tardan en aprender a emitir sonidos que al ser concatenados_ forman palabras, una solución es realizar la realimentación_ auditiva mediante una computadora, convirtiéndola a una realimentación visual, la cual le muestra al niño, por medio de figuras, que tan diferente es el sonido que emite con respec to a un sonido de referencia, tal procedimiento puede llevar se a cabo en forma interactiva.

Uno de los problemas que han adquirido gran relevancia den tro de las comunicaciones, es el de la transmisión a la me-- nor tasa posible, ya que de esta manera se emplea con mayor_ eficiencia un canal de comunicación, además involucra menor espacio de almacenamiento para los datos. Una de las señales que con mayor porcentaje se presenta en las comunicacio nes es la voz humana, esto debido al uso tan extenso que se_ hace de la telefonía; por otra parte, las comunicaciones se_ están transformando cada día más en comunicaciones de tipo - digital, por los menores costos y por la amplia gama de procesamiento que se pueden realizar sobre una señal digitalizada, por esto es que se han desarrollado varias técnicas de codificación digital de voz, con las cuales se han conseguido los rendimientos que muestra la figura 1.1

Como se puede observar, los métodos para la codificación - de voz pueden ser clasificacos en tres diferentes categorías, codificación de la forma de onda, modelado de la fuente - -- (VOCODER), y técnicas paramétricas ó híbridas [4,5] . El_

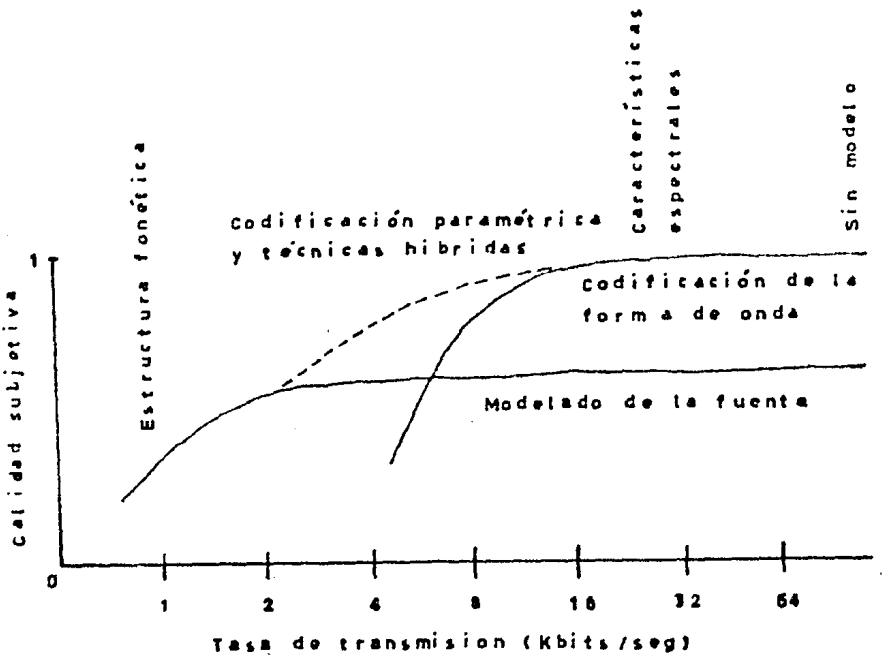


Figura 1.1

objetivo de los codificadores de la forma de onda es tratar - de obtener un facsimil de la señal, mediante la codificación con la menor cantidad de bits posible; la calidad lograda es_ la mejor que se pueda lograr con cualquier otra técnica digital, pero la tasa de transmisión es superior a los 10 Kbits/ seg. La codificación por modelado de la fuente emplea un - modelo matemático del aparato productor de la voz, del que se calculan los parámetros para que emita un segmento de voz dado, y se transmiten, en el receptor se carga el modelo con -- estos parámetros y de esta manera se reproduce la voz; aunque con una mediana calidad, la voz reproducida es facilmente entendible y se obtienen tasas de transmisión hasta de 2 Kbits/ seg. La codificación híbrida, como su nombre lo indica, -- combina elementos de las técnicas descritas anteriormente con

el fin de mejorarlas en el intervalo de 2 a 8 Kbits/seg.

El presente trabajo es la realización de un sistema para procesamiento de señales de voz que, aprovechando la estructura fonética de la voz, permite obtener símbolos asignados a cada fonema y que concatenados forman un texto que de alguna manera representa lo dicho por alguna persona. Para desarrollar este sistema se hizo uso de técnicas del área de codificación digital de voz, particularmente por un modelo de la fuente, y técnicas del área de reconocimiento de patrones, estas técnicas son expuestas brevemente en los capítulos 2 y 3, donde ademas se expone la codificación mediante cuantización vectorial. En el capítulo 4 se expone la forma en que quedaron integradas las diferentes técnicas empleadas, para tener completamente armado el sistema y en operación; finalmente, en el capítulo 5 se exponen algunos resultados, además de que se dan conclusiones y se proponen algunas aplicaciones y mejoras al sistema desarrollado.

CAPITULO 2

CODIFICACION POR PREDICCION LINEAL

Las técnicas para la codificación de voz mediante un modelo de la fuente, consideran a la voz como la respuesta de un sistema lentamente variante en el tiempo a una excitación periódica o a una excitación con ruido [6,7].

Especificacmente, el modelo de producción de voz consiste esenciąlmente de un tubo acústico, el ducto vocal, excitado por una fuente apropiada para generar el sonido deseado.

La forma de onda de la voz como una onda acústica de presión se origina por movimientos fisiológicos voluntarios de las estructuras mostradas en la figura 2.1. El aire es expelido de los pulmones hacia la traquea y forzado a pasar por entre las cuerdas vocales.

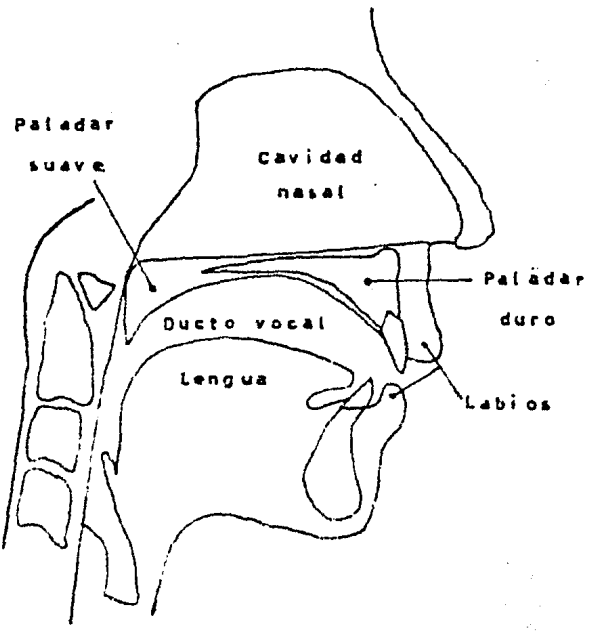


Figura 2.1

Durante la generación de sonidos "sonoros", como las vocales, el aire presionado hacia los labios causa que las cuerdas vocales se abran y se cierren a una frecuencia que depende de la presión del aire en la traquea y de los ajustes fisiológicos de las cuerdas vocales. Estos ajustes incluyen cambios en la longitud, grosor, y tensión de las cuerdas vocales; a mayor tensión, la frecuencia fundamental de la voz se eleva. A la apertura entre las cuerdas vocales se le llama glotis. La presión de aire preglótica y las variaciones del área de la glotis en el tiempo determinan la velocidad y el volúmen del flujo de aire expelido hacia el ducto vocal.

El ducto vocal es un tubo acústico no uniforme que se extiende desde la glotis hasta los labios y varía su forma en función del tiempo. Los principales componentes anatómicos que causan esta variación son los labios, la mandíbula, lengua y paladar.

Durante la generación de sonidos no nasales, el paladar suave separa al ducto vocal de la cavidad nasal. La cavidad nasal constituye un tubo acústico adicional para la transmisión del sonido.

Los sonidos "sordos" como la /f/ son generados manteniendo abiertas las cuerdas vocales y forzando el aire por entre ellas.

Para el modelado del sistema se toma en cuenta el proceso que lleva a cabo la glotis y el ducto vocal; por lo que respecta a la glotis, se modela como una fuente de excitación periódica (sonido "sonoro") ó ruido (sonido "sordo"), simulando las variaciones de la presión del aire al pasar por ella.

El modelo para el ducto vocal es un tubo acústico que aproxima al ducto vocal como un conjunto interconectado de secciones de tubo acústico de igual longitud, como se muestra en la figura 2.2. Se supone que la propagación del sonido a través de cada sección se realiza como una onda plana, y que se desprecian las pérdidas internas y el efec-

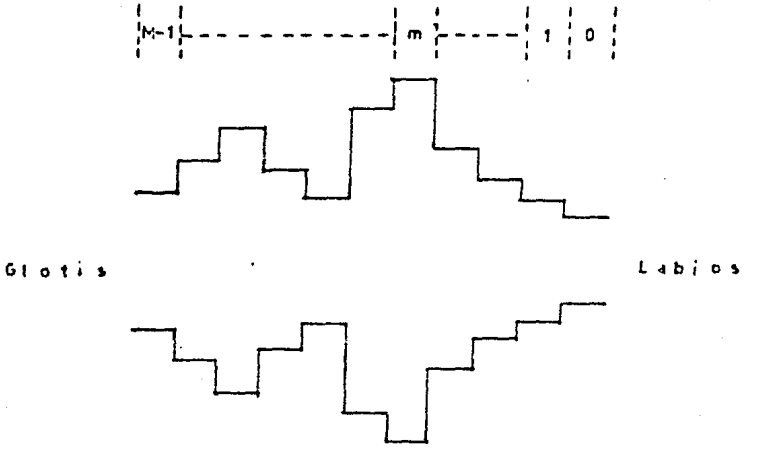
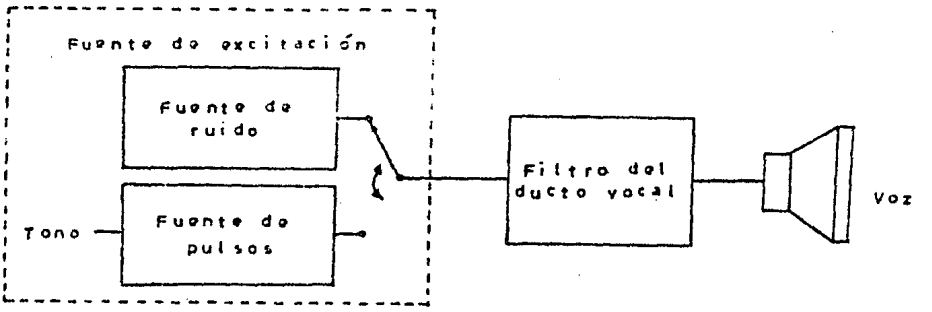


Figura 2.2

to de la cavidad nasal y el acoplamiento entre el ducto vocal y la glotis.

Por lo anteriormente detallado se tiene el modelo mostrado en la figura 2.3; de acuerdo a este modelo, la voz es generada por una fuente de excitación sonora que produce pulsos durante regiones "sonoras" de voz, cuando las cuerdas vocales están vibrando, ó es ruido durante regiones "sordas" de voz debido a la turbulencia en puntos de constricción en el ducto vocal. Esta excitación es modulada espectralmente por el ducto vocal, el cual actúa como un filtro acústico variante en el tiempo.



Los parámetros del filtro determinan la identidad (características espectrales) del sonido particular de los dos tipos de excitación; para aplicar el análisis de series de - - tiempo para la obtención de estos parámetros, cada señal analógica (voz) s(t) se muestrea para obtener una señal discreta en el tiempo s(nT), también conocida como serie de tiempo, donde n es una variable entera y T es el periodo de muestreo (Desde ahora emplearemos s_n en vez de s(nT)).

Podemos considerar que la señal s_n es la salida de algún sistema con una entrada desconocida u_n tal que se cumple la relación:

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

donde a_k, $1 \leq k \leq p$, b_l, $1 \leq l \leq q$, y la ganancia G son los parámetros del sistema hipotético. La ecuación (1) muestra que la salida s_n es una función lineal de salidas anteriores y - de entradas presentes y pasadas. Esto es, la señal s_n es predictible a partir de combinaciones lineales de salidas y entradas anteriores. De ahí el nombre de predicción li--- neal [8, 9, 10, 11].

Aunque la estimación (ó predicción) lineal por mínimos - cuadrados surge a partir de Gauss en 1795, parece ser que el primer uso específico del término predicción lineal fue en - el libro de Norbert Wiener "The Linear Predictor for a Single Time Series" en 1949. Los primeros investigadores que aplicaron directamente las técnicas de la predicción lineal (ó equivalentes) al análisis y síntesis de voz fueron Saito e Itakura en 1966 y Atal y Shroeder en 1967.

La ecuación (1) se puede especificar en el dominio de la frecuencia mediante la transformada Z de ambos lados de (1). Si H(Z) es la función de transferencia del filtro que repre-

senta al ducto vocal, entonces tenemos:

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

donde

$$S(z) = \sum_{n=-\infty}^{\infty} s_n z^{-n} \qquad (3)$$

es la transformada Z de s_n, y U(Z) es la transformada Z de u_n. La ecuación (2) es el modelo general de polos y ceros. Hay dos casos especiales del modelo y que son de interés:

1) Modelo únicamente con ceros: $a_k=0$, $1 \leq k \leq p$
2) Modelo únicamente con polos : $b_1=0$, $1 \leq 1 \leq q$.

Al modelo únicamente con ceros se le conoce en la literatura de estadística como el modelo de promedio movible (moving average, MA), y al modelo únicamente con polos se le conoce como el modelo autorregresivo (AR). Al modelo con polos y ceros se le conoce como el modelo de promedio movible autorregresivo (autorregresive moving average, ARMA).

Un segmento de voz es suficientemente complejo que no podemos esperar que cumpla exactamente con el modelo de la ecuación (2), mucho menos un modelo más simplificado, como un modelo únicamente de polos ó únicamente de ceros. Pero debemos tomar en cuenta un atributo importante de la función de transferencia del ducto vocal, y es que está caracterizado principalmente por resonancias, las cuales son bien representadas por polos.

En el modelo de polos, suponemos que la señal s_n está

dada como una combinación lineal de valores anteriores y - alguna entrada u_n:

$$s_n = -\sum_{k=1}^{p} a_k s_{n-k} + G u_n \tag{4}$$

La función de transferencia se reduce a una función de transferencia solo con polos

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{5}$$

Dada una señal en particular s_n, el problema es determinar los coeficientes del predictor a_k y la ganancia G de alguna manera.

Para la determinación de los coeficientes suponemos -- que la entrada u_n es totalmente desconocida. Por lo tanto, la señal s_n puede ser predicha solo aproximadamente a - partir de una suma ponderada de muestras anteriores. Sea la aproximación de s_n indicada por $\tilde{s}_n$, donde

$$\tilde{s}_n = -\sum_{k=1}^{p} a_k s_{n-k} \tag{6}$$

El error entre el valor actual s_n y el valor predicho - $\tilde{s}_n$ está dado por

$$e_n = s_n - \tilde{s}_n = s_n + \sum_{k=1}^{p} a_k s_{n-k} \tag{7}$$

e_n es conocido como el residuo. Con el método de mínimos cuadrados los parámetros a_k se obtienen como el resultado - de la minimización del error cuadrático medio o total con - respecto a cada uno de los parámetros.

Se emplea comunmente dos procedimientos. El primero, conocido como el método de autocovariancia, minimiza el error cuadrático medio en un segmento de longitud finita de s_n. El segundo método, conocido como el método de autocorrelación, minimiza el error en un intervalo de duración infinita, pero como s_n solamente es conocida en un intervalo finito, o estamos interesados en la señal durante un intervalo finito (intervalos de señal cuasiestacionaria), se multiplica la señal s_n por una función ventana w_n para obtener otra señal s'_n que es cero fuera del intervalo donde la función ventana es no nula. Con cualquiera de los dos métodos, el proceso de análisis se aplica a segmentos sucesivos de la señal de voz tal que los coeficientes del modelo sean continuamente actualizados para reflejar la naturaleza variante en el tiempo del ducto vocal.

Los dos métodos han sido empleados para el análisis de voz, pero el método de autocorrelación es el método más aceptado. Una de sus principales ventajas es que la matriz de coeficiente es Toeplitz; esto es, los elementos sobre las diagonales son iguales.

Además existen métodos eficientes para la solución de las ecuaciones (8); una consideración adicional es que la matriz Toeplitz garantiza ser no singular, y, a pesar de inexactitudes numéricas, el filtro resultante es estable.

$$\sum_{k=1}^{p} a_k R(i-k) = -R(i) \quad , 1 \le i \le p \tag{8}$$

donde

$$R(i) = \sum_{n=-\infty}^{\infty} s_n s_{n+i} \tag{9}$$

es la función de autocorrelación de la señal s_n; con el uso de la función ventana, (9) se vuelve:

$$R(i) = \sum_{n=0}^{N-1-i} s'_n s'_{n+i} \quad , i \ge 0 \tag{10}$$

donde

$$(11) \qquad s'_n = \begin{cases} s_n w_n & ,0 \le n \le N-1 \\ 0 & ,\text{c. c.} \end{cases}$$

Con los coeficientes óptimos se obtiene el error mínimo cuadrático:

$$(12) \qquad E_p = R(0) + \sum_{k=1}^{p} a_k R(k)$$

El método más rápido para resolver (8) es el método recursivo de Durbin:

$$(13a) \qquad E_0 = R(0)$$

$$(13b) \qquad k_i = -\left[R(i) + \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j)\right] / E_{i-1} \qquad a_i^{(i)} = k_i$$

$$(13c) \qquad a_j^{(i)} = a_j^{(i-1)} + k_i a_{i-j}^{(i-1)} \qquad ,1 = j = i-1$$

$$(13d) \qquad E_i = (1-k_i^2) E_{i-1}$$

Las ecuaciones (13) se resuelven recursivamente para -- i= 1, 2, ..., p. La solución final está dada por:

$$(13e) \qquad a_j = a_j^{(p)} \qquad ,1 \le j \le p$$

Para el cálculo de la ganancia G se reordena la ecua--- ción (7):

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

comparando (4) y (14) se ve que la única señal de entrada - que resulta como la señal s_n a la salida es la que cumple - $Gu_n=e_n$, por lo que la energía de la señal de entrada Gu_n -- debe ser igual a la energía del error, la cual está dada -- por E_p de la ecuación (12), entonces:

$$G^2=E_p=R(0)+\sum_{k=1}^{p}a_k R(k) \qquad (15)$$

Con el conjunto de coeficientes del filtro que representa al ducto vocal y la ganancia, tenemos una representación paramétrica de un segmento de voz, con lo que podemos definir patrones para la voz.

C A P I T U L O 3

RECONOCIMIENTO DE PATRONES

Un patrón es la descripción de un objeto; cuando una persona percibe un patrón, hace una inferencia y asocia esta percepción con algunos conceptos generales, los cuales ha derivado de su experiencia, estimando las posibilidades de que los datos de entrada puedan ser asociados con una de un conjunto de poblaciones estadísticas conocidas [12] .

En forma sencilla, el reconocimiento de patrones puede ser definido como la categorización de los datos de entrada en clases identificables mediante la extracción de razgos significativos o atributos de los datos, despreciando detalles irrelevantes.

Una clase de patrones es una categoría caracterizada por algunos atributos en común. Un patrón es la descripción de algún miembro de una categoría representante de una clase de patrones; cuando un conjunto de patrones caen dentro de clases disjuntas, es deseable categorizar estos patrones dentro de sus clases respectivas mediante el uso de algún dispositivo automático.

El diseño de un sistema automático para reconocimiento de patrones involucra varios problemas. El primero concierne a la representación de los datos de entrada, los cuales pueden ser medidos de los objetos que se desea reconocer, cada cantidad medida describe una característica del patrón o del objeto y puede ser acomodada en la forma de un vector de mediciones o vector representativo del patrón ;

$$X = \begin{vmatrix} x_0 \\ x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_{k-1} \end{vmatrix}$$

Al conjunto de patrones pertenecientes a la misma clase - corresponde un grupo de puntos diseminados dentro de alguna -- región del espacio de mediciones; por ejemplo, en la figura -- 3.1 se muestran dos clases de patrones, denotadas por ω_1 y - ω_2, cada patrón está caracterizado por dos mediciones, por lo que cada patrón puede verse como un punto en un espacio de dos dimensiones.

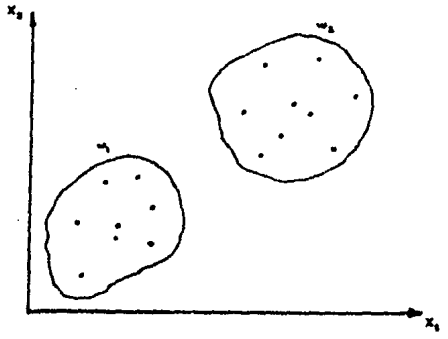


Figura 3.1

El segundo problema es la extracción de razgos característicos o atributos de los datos recibidos y la reducción de la dimensionalidad de los vectores representativos del patrón, ya que si se cuenta con un conjunto completo de razgos discriminatorios para cada clase de patrón, el reconocimiento y la clasificación de los patrones presentará poca dificultad.

El tercer problema en el diseño de un sistema para reconocimiento de patrones involucra la determinación de procedimientos de decisión, los cuales son necesarios en el proceso de -- identificación y clasificación; después de que los datos observados de los patrones han sido expresados en forma de vectores en el espacio de patrones, el problema puede verse como el de generar fronteras de decisión, las cuales separan a las diferentes clases de patrones.

Desde el punto de vista de la realización del sistema para reconocimiento de patrones, tenemos el problema de poblar el espacio de patrones con los arquetipos de cada clase, cosa que generalmente se hace en forma manual, que requiere que cada patrón arquetipo esté perfectamente caracterizado y se debe obtener una cantidad adecuada de éstos patrones para que puedan reflejar las propiedades de la población a la que representan.

La clasificación de patrones mediante funciones de distancia es uno de los conceptos más antiguos en el reconocimiento automático de patrones; esta técnica es una herramienta efectiva para la solución de problemas en los cuales cada clase de patrones tiende a agruparse alrededor de una cierta región del espacio de patrones.

Un clasificador por distancia mínima calcula la distancia entre un patron X de clasificación desconocida y los prototipos de cada clase, y asigna al patrón nuevo la clase de la cual está más próximo, X es asignado a la clase ω_i si $D_i < D_j$, para toda $j \neq i$; los empates se resuelven arbitrariamente.

Consideremos un conjunto de muestras de patrones de clasificación conocida $\{S_1, S_2, \ldots, S_n\}$, donde se supone que cada patrón corresponde a una de las clases $\omega_1, \omega_2, \ldots, \omega_n$. Podemos definir una regla de clasificación mediante el vecino más cercano (nearest neighbor), la cual asigna un patrón X de clasificación desconocida a la clase de su vecino más cercano, donde decimos que $S_i \in \{S_1, S_2, \ldots, S_n\}$ es el vecino más cercano a X si:

$$D(S_i, X) = \min_l \{D(S_l, X)\} \quad , \; l = 1, 2, \ldots, N$$

donde D es cualquier distancia definible sobre el espacio de patrones.

Una aplicación del método de clasificación mediante la regla del vecino más cercano es la cuantización vectorial. Un

cuantizador de N niveles y k dimensiones es un mapeo q, que asigna a cada vector de entrada, X de k dimensiones, un vector de reproducción $\hat{X} = q(x)$, tomado de un conjunto de patrones de clasificación conocida, $\hat{A} = \{Y_i;\ i = 1,\ldots, N\}$. El cuantizador q está completamente descrito por el conjunto de patrones de reproducción $\hat{A}$ junto con la partición, $S = \{S_i;\ i = 1,\ldots,N\}$, del espacio de patrones en conjuntos $S_i = \{X : q(x) = Y_i\}$ de vectores de entrada que quedan mapeados en el i -ésimo patrón de reproducción, en la figura 3.2 se muestra conceptualmente un cuantizador vectorial.

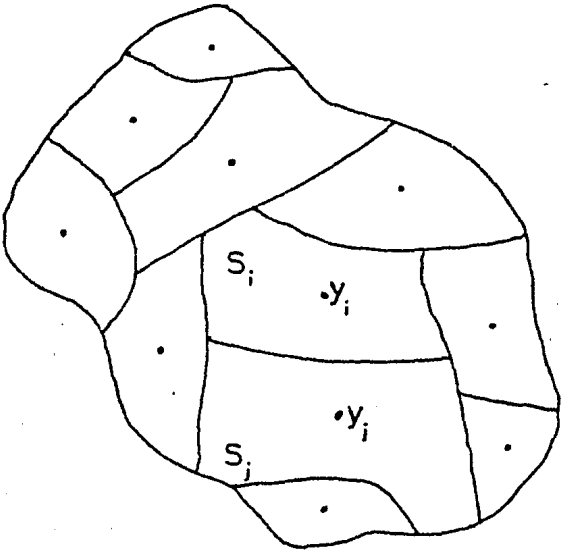


Figura 3.2

Al codificar mediante un cuantizador vectorial se supone que la distorsión provocada al reproducir un vector X por un vector de reproducción $\hat{X}$ está dada por una medida de distorsión no negativa, la cual se emplea para realizar la clasificación -

mediante el vecino más cercano; para que una medida de distorsión sea aplicable en cuantización vectorial de voz, debe ser analíticamente tratable, calculable a partir de muestras de datos y debe tener un significado subjetivo. La medida de distorsión de Itakura-Saito parece satisfacer los requerimientos antes mencionados, debido a que aproxima razonablemente bien la forma en que el oido humano percibe los sonidos [14, 15].

La codificación de voz mediante cuantización vectorial emplea muestras de la función de autocorrelación de la señal, medidas en un marco de análisis (aprox. 20 ms. para tener marcos con señales de naturaleza casi estacionaria), la forma del espectro de voz en cada marco es codificada usando un cuantizador vectorial, en función de un conjunto conocido de parámetros de predicción lineal llamado palabra (codeword); a la colección de posibles palabras se le llama diccionario (codebook). Sea S_j el conjunto de parámetros de predicción lineal que resultan de analizar el j-ésimo marco de voz a codificar, entonces el marco es codificado mediante la palabra que mejor representa a S_j de acuerdo a la regla del vecino más cercano.

Los diccionarios de cuantización vectorial son diseñados para minimizar la distorsión promedio que resulta de codificar una secuencia larga de marcos de voz de entrenamiento; por ejemplo, un diccionario para la palabra "alto" puede ser diseñado corriendo el algoritmo para el diseño del cuantizador vectorial con una secuencia de entremaniento consistente de varias repeticiones de la palabra "alto", hasta lograr la mínima distorsión. [13].

C A P I T U L O 4

REALIZACION DEL SISTEMA

Debido a que este sistema se diseñó con la idea de que sea un prototipo, se limitó a las señales de entrada a palabras y frases en las cuales únicamente aparecieran sonidos del conjunto Λ, definido por:

$$\Lambda = \{ a, e, i, o, u, l, m, n, r, s, . \}$$

donde el punto (.) representa un periodo de silencio.

Primeramente se procedió a entrenar el sistema con los fonemas del conjunto Λ, para lo cual se segmentaron y etiquetaron señales con estos fonemas, con el fin de obtener una secuencia de entrenamiento para cada fonema y así poder generar un diccionario constituído por palabras (codeword) representativas de variaciones del mismo fonema. Como resultado del entrenamiento, cada fonema queda representado por un diccionario de cuantización vectorial como se muestra conceptualmente en la figura 4.1. [16,17]

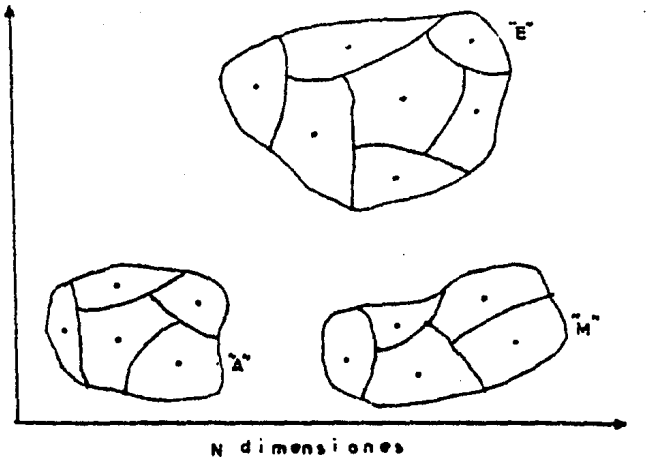


Figura 4.1

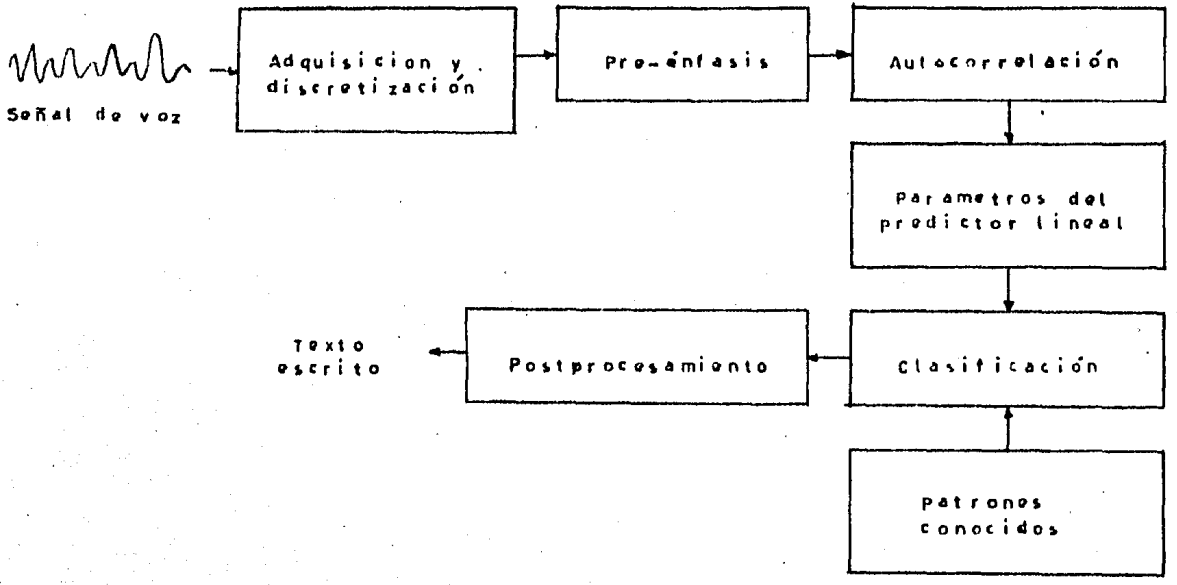

Señal de voz
Adquisicion y discretización
Pre-énfasis
Autocorrelación
Parametros del predictor lineal
Clasificación
Patrones conocidos
Postprocesamiento
Texto escrito

Figura 4.2

Una vez que se tiene el conjunto de patrones conocidos, obtenidos por entrenamiento, ya se está en posibilidad de -- realizar la traducción, según el diagrama que se muestra en_ la figura 4.2.

La primera etapa del procesamiento, al igual que la de_ entrenamiento, consiste en el muestreo de la señal a 6.4 Khz y una conversión analógico-digital uniforme con 12 bits por_ muestra; posteriormente se pasa por una etapa de pre-énfasis, la cual es un filtrado del tipo:

$$x[n] = S[n] - 0.95\ S[n-1]$$

Para el cálculo de las autocorrelaciones se emplearon - ventanas rectangulares que abarcan 256 muestras, de las cuales, se emplearon 128 muestras centrales (20 ms.) como marco de análisis, además de un traslape con los marcos contiguos_ de 64 muestras a ambos lados del marco de análisis actual.

Empleando 11 coeficientes de autocorrelación de la señal $(r_x(0),\ldots,r_x(10))$, se aplica el algoritmo de Durbin para obtener los coeficientes del predictor lineal de orden 10 que - mejor representa al marco actual.

La etapa de clasificación compara al patrón recibido con cada uno de los patrones conocidos, empleando la medida de -- distorsión de Itakura-Saito y selecciona a los cinco patrones que estén más cercanos al patrón que acaba de entrar; debido_ a que cada uno de los patrones conocidos tiene asignado un -- símbolo, cada marco de análisis queda representado por cinco_ símbolos, algunos de ellos repetidos.

La parte final es la etapa llamada de postprocesamiento, cuya finalidad es asignar un solo símbolo a una secuencia de_ marcos de análisis, cada uno de ellos con cinco símbolos; para lo cual se emplea el siguiente algoritmo:

1.- Lectura de 5 símbolos (un renglón, 20 ms. de señal).
2.- Eliminación de símbolos repetidos y asignación de peso en función del número de repeticiones de cada símbolo.
3.- Ordenación de las parejas (símbolo, peso), de mayor a me-

nor peso.

4.- Si el símbolo de mayor peso es silencio, pasar a 5; en - caso contrario poner indicador de detección de palabra. Almacenar las parejas de mayor peso. Pasar a 1.

5.- Si es verdadero el indicador de detección, eliminar las primeras y las últimas parejas almacenadas.

5.1 Contar el número de renglones contiguos (1) en que aparece cada símbolo y se suman los pesos de estos renglones.

5.2 Se obtiene el símbolo con el máximo peso y si 1 excede un umbral preestablecido, se imprime el símbolo. Pasar a 1.

Se empleó una minicomputadora PDP-11/40 para la realización del sistema, el cual está formado por varios programas, cada uno de ellos está especializado en una función en particular, como son los que realizan el muestreo, conversión de las muestras a valores de voltaje, autocorrelaciones, etc. Cada uno de ellos se corre por separado y se - comunica con los demás mediante archivos.

Algunos de estos programas fueron realizados íntegramente en FORTRAN o en PASCAL y algunos otros fueron realizados en FORTRAN con subrutinas en lenguaje ensamblador.

C A P I T U L O 5

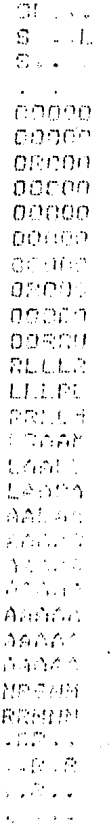
RESULTADOS Y CONCLUSIONES

El sistema fue alimentado con algunas palabras y una frase, mediante los cuales se certificó su operación. A continuación se presentan algunos resultados obtenidos.

Se presenta la salida de la etapa de clasificación, en la que en cada renglón se presentan cinco símbolos tomados del alfabeto Λ y que son los que mejor representan a un marco de análisis de 20 ms. de duración, por lo que podemos ver al eje del tiempo como una línea vertical que crece hacia la parte de abajo de la página. Posteriormente se presenta el texto final como resultado de la etapa de postprocesamiento.

Texto hablado: HOLA

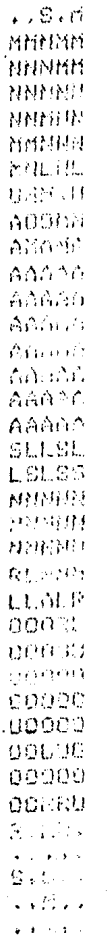
Salida del clasificador:



Salida del postprocesamiento

Texto hablado: MANO

Salida del clasificador:

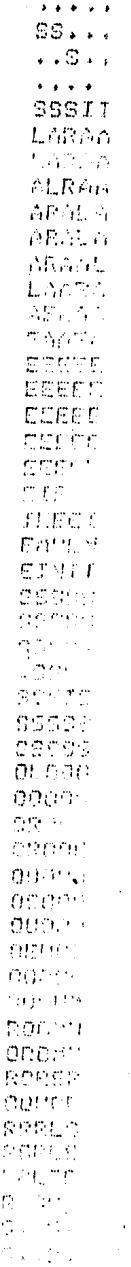


Salida del postprocesamiento:

MANO

Texto hablado: REZO

Salida del clasificador:

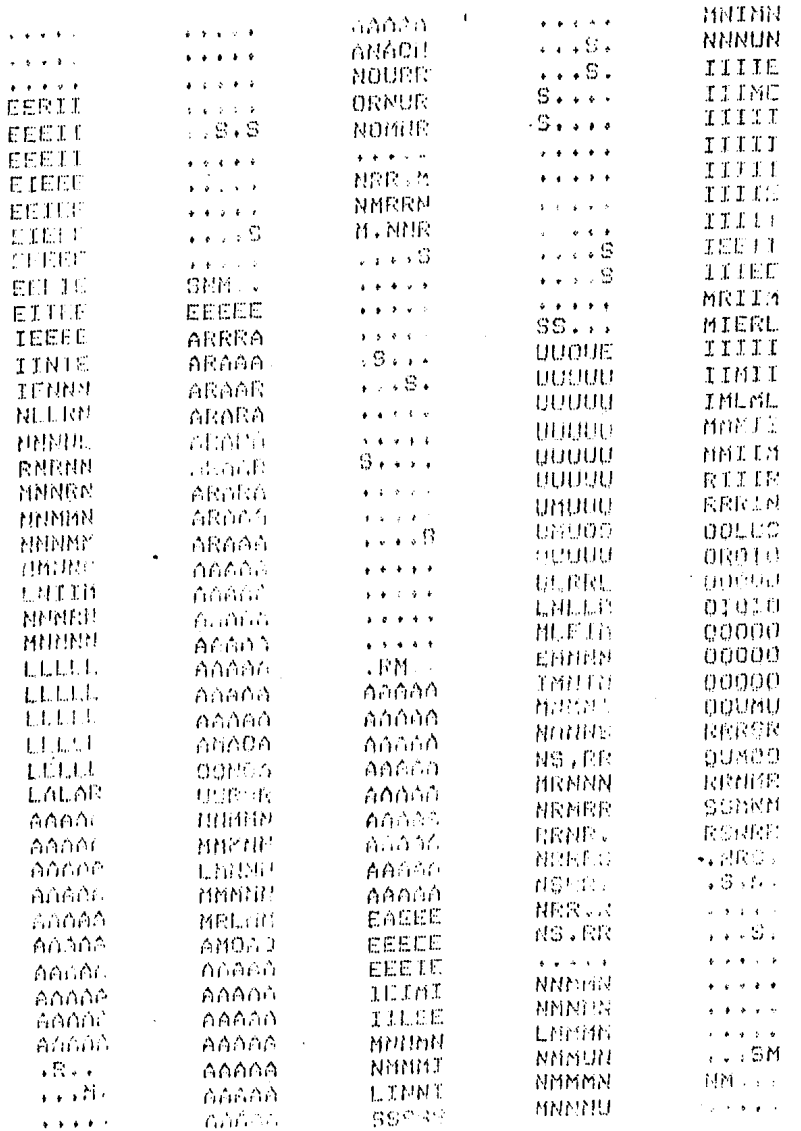


Salida del postprocesamiento:

REZO

Texto hablado: EN LA RAMA HAY UN MIRLO

Salida del clasificador



Salida del postprocesamiento:

ENLA
RAMA
AY
UN
MIR

El sistema desarrollado, en su primera etapa, tuvo resultados bastante alentadores. De los sonidos asociados al conjunto Λ se pudo observar en los experimentos realizados, un 90% de aciertos en el reconocimiento de las vocales, un 99% en el caso de los periodos de silencio y un 70% en el caso de las consonantes.

Considerando que este es el esquema más sencillo con que se puede realizar un traductor de las características mencionadas, se puede ver un panorama amplio de sistemas de procesamiento de voz, con reconocimiento de palabras pero sin un diccionario limitado, para ello es necesario contar con un modelo de la voz a base de fonemas y de reglas de concatenación entre ellos.

El sistema desarrollado permite descubrir partes del procesamiento que son de importancia para el desarrollo de un mejor sistema y que ponen de relieve la necesidad de investigar sobre algunas áreas del campo de la teoría de la información.

Se espera que la factibilidad de realizar un sistema de este tipo sea algo que motive a algunas personas a trabajar dentro de este campo, además de la motivación debida a los proyectos y sus aplicaciones descritos en el capítulo 1 de este trabajo.

B I B L I O G R A F I A

1. Philip C. Treleaven e Isabel Gouveia Lima, "Japan's -- Fifth Generation Computer Systems", IEEE COMPUTER, Agosto 1982, pags. 79-88.

2. Raj Reddy y Victor Zue, "Recognizing continuos speech - remains an elusive goal", IEEE SPECTRUM, Noviembre 1983, número dedicado a las computadoras de la quinta generación.

3. T. Moto-oka et al., "Keynote speech, challenge for ---- Knowledge Information Processing Systems", Reporte Preliminar sobre los Sistemas de Computación de la Quinta Generación, Seminario sobre la Computadora de la Quinta Generación, Agosto 1984, P.U.C.-U.N.A.M.

4. R.E. Crochiere y J.L. Flanagan, "Current perspectives in digital speech", IEEE Communications Magazine, Enero - - 1983, pags. 32-40.

5. J.L. Flanagan, M.R. Schroeder, B.S. Atal, R.E. Crochiere, N.S. Jayant y J.M. Tribolet, "Speech Coding", IEEE - - - Transactions on Communications, Abril 1979, pags. 710-736.

6. A.V. Oppenheim, "Digital Processing of Speech", Aplications of Digital Signal Processing, Editado por A.V. Oppenheim, 1978.

7. L.R. Rabiner y R.W. Schafer, "Digital Processing of Speech Signals", 1968, Prentice-Hall.

8. John Makhoul, "Linear Prediction: A Tutorial Review", - -- Proceedings of the IEEE, Abril 1975, pags. 561-580.

9. J.D. Markel y A.H. Gray Jr., "Linear Prediction of Speech", 1976, Springer-Verlag.

10. D. Fernando Fernández-Baca P., "Codificación de voz empleando Predictores Lineales con Excitación Residual", Tesina de Licenciatura (Computación), 1980, Facultad de Ingeniería, - U.N.A.M.

11. Manfred R. Schoeder, "Linear Prediction, Entropy and Signal Analysis", IEEE Acoustics, Speech, and Signal Processing Magazine, Julio 1984, pags. 3-11

12. J.T. Tou y R.C. González, "Pattern Recognition Principles", 1979, Addison-Wesley Publishing Company.

13. Y. Linde, A. Buzo y R.M. Gray, "An algorithm for Vector Quantizer Design", IEEE Transactions on Communications, Enero 1980, pags. 84-95.

14. A. Buzo, A.H. Gray Jr., R.M. Gray y J.D. Markel, "Speech Coding Based Upon Vector Quantization", IEEE Transactions on Acoustics, Speech and Signal Processing, Octubre 1980, pags. 562-574.

15. R.M. Gray, A. Buzo, A.H. Gray Jr. y Y. Matsuyama, "Distortion Measures for Speech Processing", IEEE Transactions on Acoustics, Speech and Signal Processing, Agosto 1980, pags. 367-376.

16. J.E. Shore y D. Burton, "Discrete Utterance Speech Recognition Without Time Normalization", Computer Science and Systems Branch, Information Technology Division, Naval Research Laboratory, Mayo 1982.

17. A. Goytia, F. Kuhlmann, A. Buzo y C. Rivera, "Procesamiento de señales de voz: Traducción de voz a texto", Memorias Decimo Congreso Academia Nacional de Ingeniería, Septiembre 1984.