Universidad Autónoma de Guadalajara

Incorporada a la Universidad Nacional Autónoma de México

Escuela de Ingeniería

"Análisis del Reconocimiento de Voz Aplicado a la Comunicación Hombre-Máquina"

TESIS PROFESIONAL

que para obtener el título de:

Ingeniero en Computación

p r e s e n t a :

Fernando Ernesto Chuw Lau

Guadalajara, Jal. 1986

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

I N D I C E.

I N T R O D U C C I O N.

La voz es el método más natural de comunicación entre los hombres y es esta la razón por la que se investiga cada vez más, en el campo de la comunicación hombre-máquina por medio de la voz. Por este motivo, el análisis y el reconocimiento de voz constituye hoy en día una importante línea de investigación de vanguardia y que ya está dando espectaculares resultados con aplicaciones en informática, robótica, ayudas a minusválidos, identificación de personas, control de marcación telefónica, comunicación de voz hombre-máquina, control de tráfico áereo, sistemas de preguntas-respuestas, manejadores de base de datos y otras aplicaciones que seria muy extenso mencionarlas todas. Si nos trasladamos unas decadas atrás, quizas el primer objetivo histórico, relativamente modesto desde el punto de vista actual, fue la construcción de canales de codificador de voz (channel vocoder), lo cual fue el primer dispositivo que intentaba tomar ventaja del módelo de filtro-fuente para la codificación de voz (Dudley, 1939). Esto era la base para una representación directa del espectro de frecuencia de una señal que puede ser obtenida por un banco de filtros pasabandas. La palabra 'vocoders' que es una contracción de 'voice coders', codificador de voz, lo cual funciona de la siguiente forma: La energía en cada banda del filtro es estimado por rectificación y supresión, y la aproximación resultante al espectro de frecuencia es transmitido o almacenado. Todo esto mencionado anteriormente tuvo su origen al querer reducir el ancho de banda teléfonico. Después de la realización de sistemas de respuesta acústica con mensajes predefinidos y almacenamiento digital llevo a sucesivas complejidades en la configuración y parametrización digital. En lo que se refiere al reconocimiento de voz por parte de una máquina es el último paso hacia la consecución de una gran simplificación en el intercambio de información. Es el proceso por el cual un operador puede usar ordenes " habladas " que pueden ser reconocidas e interpretadas por un sistema automático de reconocimiento de voz. Tradicionalmente, las comunicaciones del hombre con la máquina necesita la utilización de un determinado lenguaje, que se sintetiza en un conjunto de ordenes introducidas normalmente a través de teclados e interpretados por la máquina. Es obvio, que la utilización de ordenes habladas constituidas por palabras que forman parte del propio lenguaje del operador, simplificaria de una forma amplia la tarea de comunicación con las máquinas.

En Reconocimiento de voz, un aspecto de interés es el intento de concebir módelos para la percepción de la voz donde muchas fuentes del conocimiento como reglas fonéticas, léxicas, sintácticas, semánticas y pragmáticas son usadas para formar el mensaje verbal. Una clasificación

RECONOCIMIENTO DE VOZ.

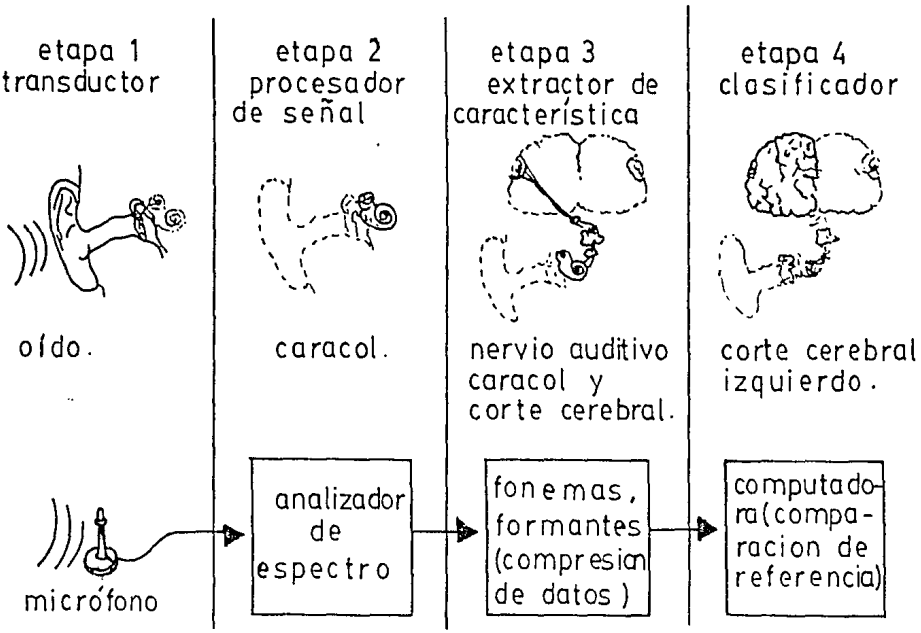


FIG. I
PROCESO DE RECONOCIMIENTO DE VOZ EN COMPARACION CON EL ORGANO AUDITIVO HUMANO.

básica de la tecnología de voz como muchos la llaman en términos de funciones de entrada y salida; puede ser dividida en entrada de voz que equivaldría a reconocimiento de voz y verificación de locutores; y salida de voz que incluye codificación-voz, sistema de respuesta de voz, aplicaciones de sintésis de voz y además de sistemas de almacenamiento y recuperación. El objetivo de la realización de este tema como tesis se debe a la necesidad de efectuar un análisis exhaustivo y descriptivo de cómo la computadora o el sistema reconoce la voz humana, al igual que conocer los métodos mediante los cuales se pueda realizar esto. Además, responder a las preguntas que han despertado en mí: ¿Cómo la computadora reconoce la voz humana?, ¿Cómo lo realiza? o ¿Cómo funciona?, en la fig. I podemos ver como es que se realiza ese proceso de reconocimento en semejanza con el órgano auditivo humano. El contenido de la tesis consta de cuatro capítulos, en la cual los dos primeros capítulos son importantes para entender o comprender el siguiente capítulo referido al análisis y reconocimiento automático de voz y por último dispositivos y sistemas de entrada de voz en la cual se expone los circuitos integrados o arquitecturas digitales que hay hoy en día en el mercado; las perspectivas en el futuro de esos dispositivos y sistemas.

I - ¿Cómo se genera la voz humana ?

1.1 Morfología del Aparato Fonador.

La voz es una onda acústica generada a partir del flujo de aire expulsado por los pulmones y modulado después por los diferentes órganos que componen el aparato fonador [47], mostrado en la fig. 1.1-1.

Las cuerdas vocales, que delimitan la glotis, es el espacio vacío que queda entre ellas, en la cual pueden cerrar o abrir el conducto laringeo. Les sigue un conducto que es, la laringe, que puede quedar ocluído o cerrado por la glotis durante una deglución pero este queda abierto para respirar o hablar. Luego de esté, sigue otro conducto tabular, la faringe, por donde se llega a las cavidades bucal y nasales. La cavidad bucal, como vemos, puede adoptar configuraciones muy variadas según la posición de la lengua, los dientes y los labios. La voz o la radiación de sonidos al exterior se realiza por la apertura labial o en las fosas nasales (sonidos nasalizados) dependiendo de la posición cerrada/abierta del velo del paladar; en otras palabras, el paso de la laringe a la cavidad nasal. Todo esto expuesto anteriormente, forman el conjunto de cavidades supraglóticas en la cual como explicamos constituye un resonador o filtro acústico que conforman el espectro de la excitación aplicada que puede ser, por ejemplo, un tren de impulsos generados por la vibración de las cuerdas vocales.

La producción de sonidos tiene su origen al excitar acústicamente el tracto vocal igual a la fig. 1.1-2, que no es otra cosa que la simulación en un símil mecánico [47] .El proceso de producción de sonidos podemos entenderlo, de forma simplificada y sencilla, como una fuente de energía: los pulmones, que origina un flujo de aire que ataca a un conjunto de cavidades, pudiendo producirse la excitación acústica, esta excitación está básicamente clasificada en tres tipos [17]:

1. Vibración de las cuerdas vocales.

   Se produce al tratar de expirar y mantener la glotis cerrada. La presión subglotal llega a ser suficiente para separar las cuerdas lo que provoca la salida de un flujo de aire al mismo tiempo que se reduce la presión y permite que los ligamentos vocales se vuelvan a cerrar, repitiéndose el ciclo. La onda generada por esta oscilación tiene una forma aproximadamente triangular, de frecuencia entre 100 y 200 Hz y ciclo de trabajo entre 0.3 y 0.7 según la tensión de las cuerdas. El espectro encontramos que se compone de la fundamental y rayas de los armonicos con amplitud decreciente de unos 12 dB/octava. Este período fundamental caracteriza la

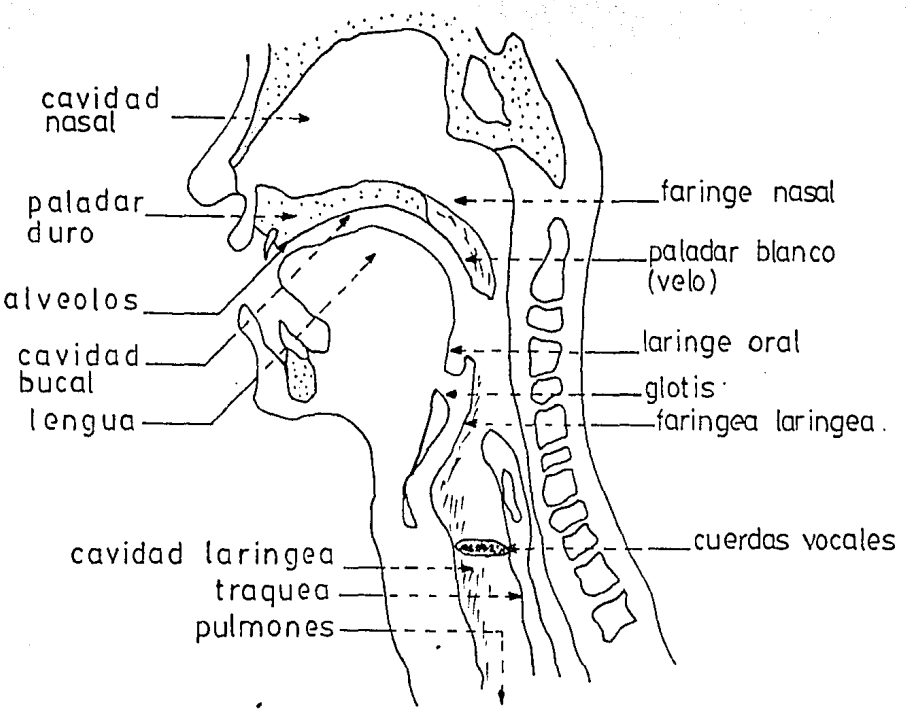
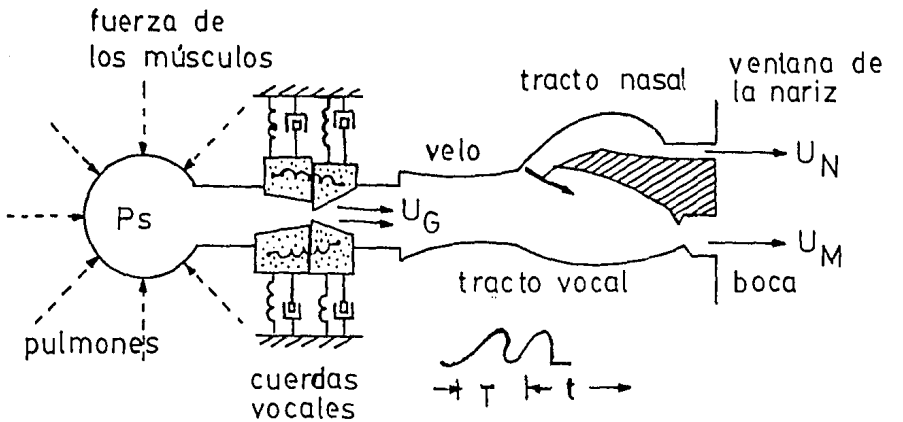


FIG. 1.1-1
EL APARATO FONADOR.



Ps - presion de aire subglotal
en los pulmones.

FIG. 1.1-2
EL TRACTO VOCAL SIMULADO EN
UN SIMIL MECANICO.

altura tonal de los sonidos articulados llamados sonoros y recibe en inglés el nombre de " pitch " lo cual en el español se conoce como tono y esto se considera uno de los parámetros más importantes en el análisis de voz.

2. Generación de una turbulencia de aire o estrechamiento, en algun punto del tracto vocal.

Esta excitación es análoga al ruido blanco (espectro plano en el márgen de audio) generado en dicho punto, corresponde a sonidos sordos como, por ejemplo, los fricativos: f, s y j.

3. por medio de sonidos oclusivos.

Este tipo de excitación es causado, si el estrechamiento del tracto vocal se convierte en estrangulamiento o cierre total seguido de un posterior relajamiento, los cuales obtenemos sonidos como la b, p o k. La apertura brusca de la oclusión produce un efecto semejante a una excitación en escalón de presión con un espectro que cae inversamente con la frecuencia.

También, son posibles durante el proceso del habla normal, que encontremos combinaciones de los diferentes tipos de excitación. Es importante mencionar que la excitación vocal es aplicada a las cavidades supraglóticas que filtran su espectro y conforman la señal. La transmisión desde la epiglotis a la apertura labial puede ser modelizada por una serie de cavidades resonantes obteniéndose una función de transferencia producto de términos resonantes de segundo orden [47]:

$$\frac{1}{(S - Sn)\ (S - Sn*)}$$

es decir, caracterizada por una serie de polos complejos conjugados correspondientes a los modos normales de transmisión. En los espectros de los sonidos articulados se observan una serie de picos a ciertas frecuencias a la que llamamos formantes. Cada formante varia en posición, amplitud y calidad con respecto al tiempo. La respuesta en frecuencia del tubo acústico(tracto vocal) con un diámetro constante es caracterizado por un número igual de espacios resonantes a las frecuencias deducidas por la formúla [47]:

$$f(n) = 340(2n - 1)/4L \text{ Hz}$$

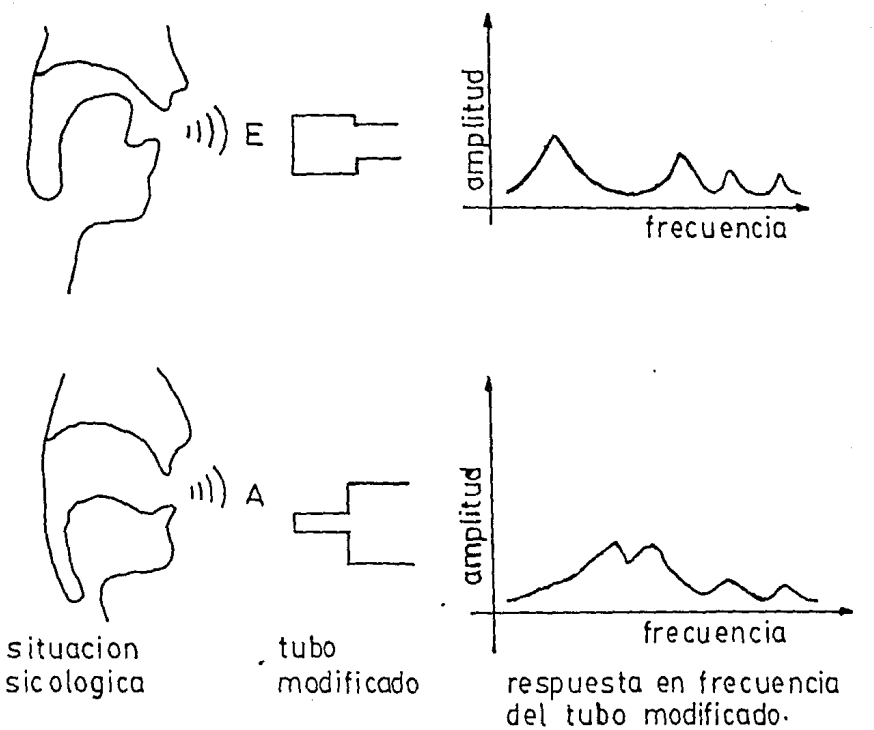


FIG. 1.1-3
EJEMPLOS DE LAS VOCALES
E Y A.

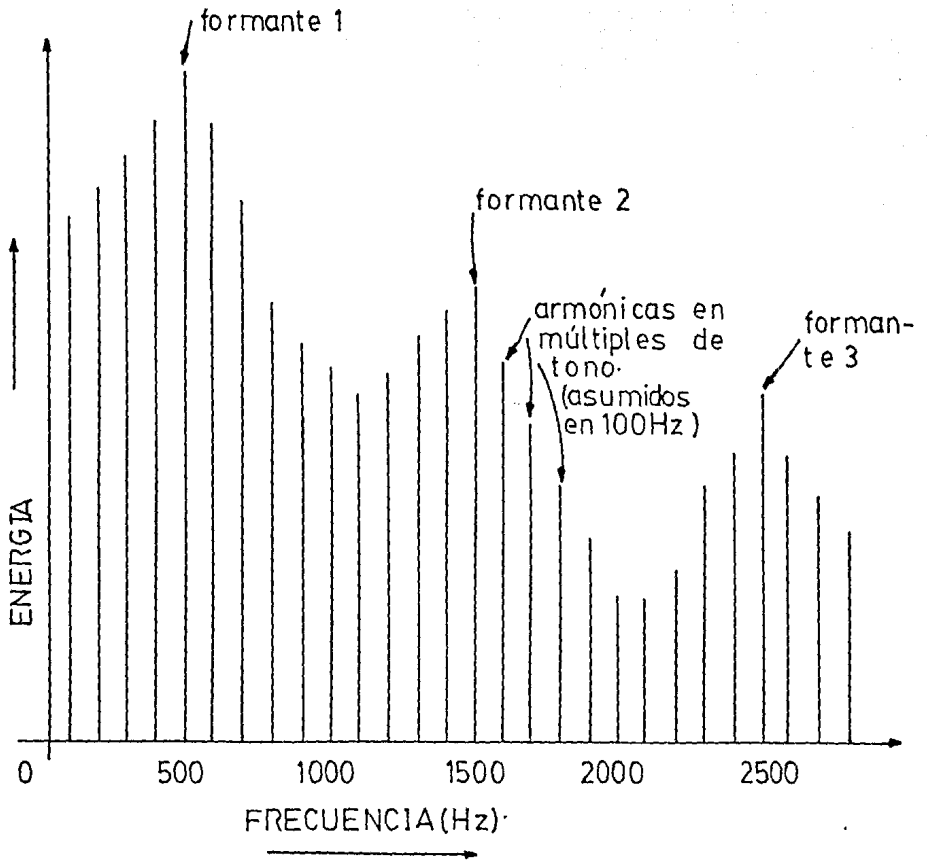


FIG. 1.1-4

ESPECTRO DE ENERGIA DE LA VOZ.

donde n=1,2,3,..,n y L es la longitud en metros. Para el hombre adulto la longitud del aparato es de unos 17 cm lo que siguiendo este módelo nos daría unos formantes situados en 800, 1500, 2500 ... Hz, que corresponden bastante bien con los observados en la materia de la vocal neutra; una e invertida, sonido en la cual no existe en el idioma español; para esto el aparato adopta la forma más tabular posible. Estos formantes son alterados modificando, mediante estrechamiento y ensanchamientos, la forma de las cavidades supraglóticas utilizando la mandíbula inferior, el cuerpo y la punta de la lengua, los dientes y los labios. Los formantes varían de posición sin guardar relación fija entre ellos. El primero lo podemos encontrar entre 200 y 1000 Hz, el segundo entre 500 y 2500 Hz y el tercero entre 1500 y 3500 Hz. Durante el habla, la configuración del tracto vocal continuamente encontramos que cambia. Por ejemplo, cuando una " E " es pronunciada, fig. 1.1-3, la cavidad faríngea es grande mientras que la cavidad oral pequeña. Por esta razón, se incrementa la frecuencia del segundo formante. Cuando una " A " es pronunciada encontramos que la situación es al revés, esto origina la reducción entre el espacio separado que hay entre el primero y segundo formante. Como sabemos, cada formante es caracterizado por un ancho de banda. Los primeros dos o tres formantes son los más importantes para entender el habla, fig. 1.1-4. En los sonidos nasales(como la m o n) la boca permanece cerrada pero el velo del paladar permite la transmisión hacia el orificio nasal por donde se produce la radiación. En estos casos aparecen también los formantes(polos de transmisión) pero además encontramos antiformantes(ceros de transmisión) debidos a que la cavidad bucal, aunque cerrada, está acoplada lateralmente a las cavidades en juego introduciendo atenuaciones importantes(cortocircuitando) a ciertas frecuencias. Podemos encontrar que algún antiformante pueda coincidir con un formante originándose que se anulen los efectos.

En la fig. 1.1-5 se muestran la excitación periódica del tracto vocal humano iniciado constantemente con el desdoblamiento vocal que ocurre repetitivamente(abriendo y cerrando).

La fig 1.1-6 muestra el resultado en un tren de pulsos de aire el cual pasa a través de las cavidades resonantes de la boca y de la cavidad nasal. Mientras que en la fig. 1.1-7 encontramos la voz compuesta de diferentes bandas de de frecuencia, estos conocidos como formantes. Se muestra la envolvente de un formante generalizado para los dos primeros formantes.

1.2 Los sonidos de la voz.

El hombre utiliza el aparato fonador para transmi-

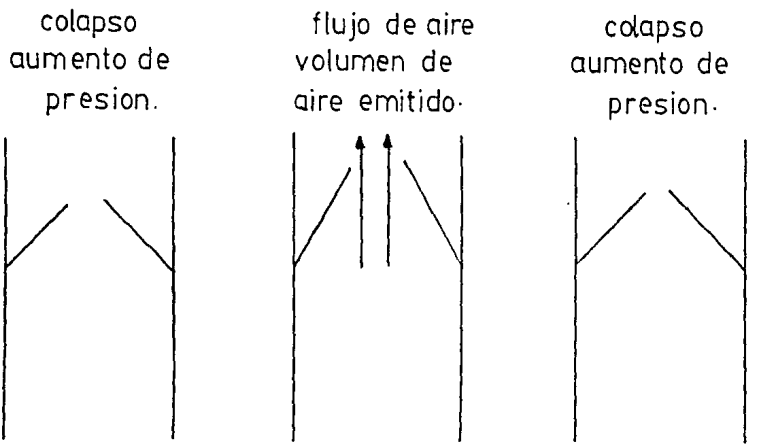


FIG. 1.1-5

REGULACION DEL FLUJO DE AIRE DESDE LOS PULMONES.

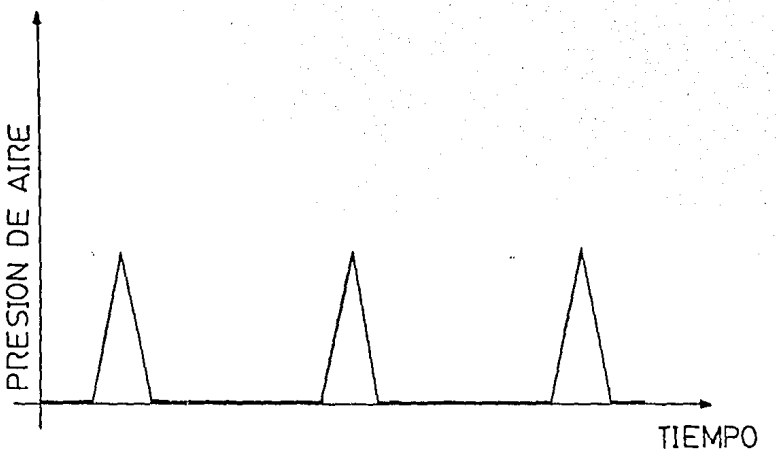


FIG. 1.1-6

RESULTADO EN UN TREN DE PULSOS DE AIRE EL CUAL PASA A TRAVES DE LAS CAVIDADES RESONANTES B LA BOCA Y DE LA CAVIDAD NASAL.

Bw - ancho de banda.
F1, F2 - amplitud pico.

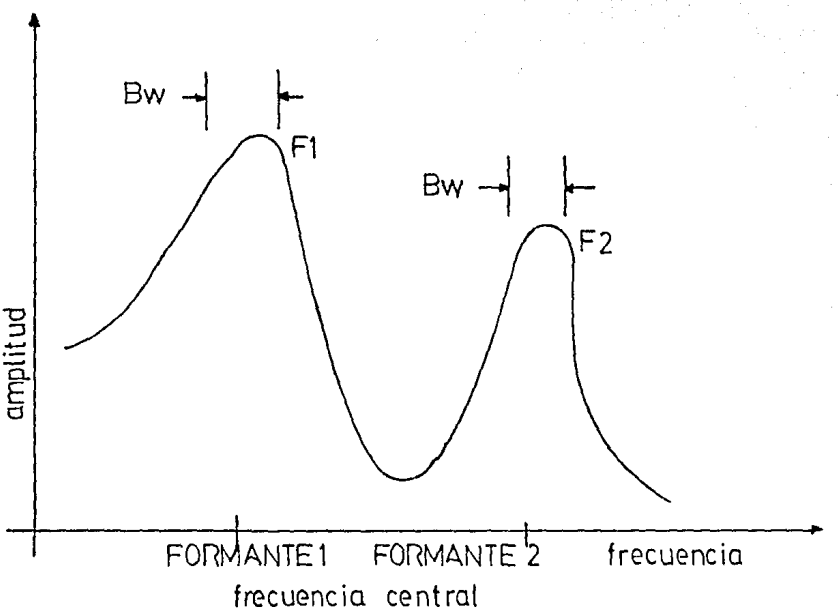


FIG. 1.1-7

EJEMPLO DE LA ENVOLVENTE DE UN FORMANTE GENERALIZADO PARA LOS DOS PRIMEROS FORMANTES.

tir ideas a sus semejantes. Por lo tanto, sabemos que dicho aparato puede producir sonidos articulados. Entonces, para cada lengua o idioma existe un conjunto discreto de estos sonidos, llamados o conocidos como fonemas. En el habla, podemos encontrar algunas variantes de los fonemas, estos son los alófonos y que completan con ellos el conjunto discreto de sonidos, alfabeto fónico, en base al cual se realiza cualquier mensaje vocal. Los sonidos lo podemos clasificar en consonantes y vocales. Las vocales se caracterizan por ser sonoras y presentar un mayor grado de abertura de los órganos articulatorios. Destacan por su intensidad con relación a las consonantes, y en español, además, son las únicas que pueden formar núcleo silábico. Encontramos, que nuestro idioma tiene un esquema vocálico bastante sencillo puesto que los cinco fonemas: /i/, /e/, /a/, /o/ y /u/. Otras lenguas tienen sólo tres pero se distinguen hasta catorce en el idioma francés.

Las consonantes se caracterizan, en general, por una mayor constricción o estrechamiento del aparato fonador y por una mayor variedad. Existen dos tipos básicos: las continuas que pueden ser pronunciadas de una forma sostenida como las vocales y que por lo tanto se producen con una configuración estable del aparato fonador y las dinámicas que requieren para su producción de mutaciones del aparato. La tabla 1.2-1 [47] recoge los fonemas(entre barras) del idioma español y se completa por simetria con algunos alófonos(entre corchetes) y los fonemas extranjeros(señalados con un arterisco).

Dinámicas son las oclusivas y las africadas (oclusión seguida de fricación). Se distinguen tres parejas de oclusivas sorda-sonora segun el lugar de oclusión y una pareja de africadas. Entre las continuas se distinguen las fricativas, nasales, y liquidas. Distinguimos según el lugar de fricación cinco parejas de fricativas opuestas por el carácter sonoro-no sonoro. Algunas de estas o no han existido o han desaparecido ya del español, dándose un ejemplo en otra lengua. Las nasales son todas sonoras y se distinguen por el lugar de la oclusión de la cavidad bucal. Finalmente, entre las liquidas también sonoras, las más parecidas a las vocales de entre las consonantes, se distinguen las laterales y las vibrantes.

Un aspecto muy importante al hablar, o al producir los sonidos que forman el habla, es la prosodia, en la cual es la parte de la gramática que enseña la recta pronunciación y acentuación de las palabras. Más adelante al entrar a la etapa de análisis y reconocimiento de voz, encontraremos que es muy importante la correcta pronunciación de las palabras para que el error(todavia no se ha logrado un reconocimiento perfecto)que involucre sea mínimo.

Tabla 1.2-1.
Clasificación de los sonidos de la Voz en Vocales y Consonantes del idioma Español.

VOCALES

	Anterior	Central	Posterior
Cerrada	i		u
Media	e		o
Abierta		a	

CONSONANTES

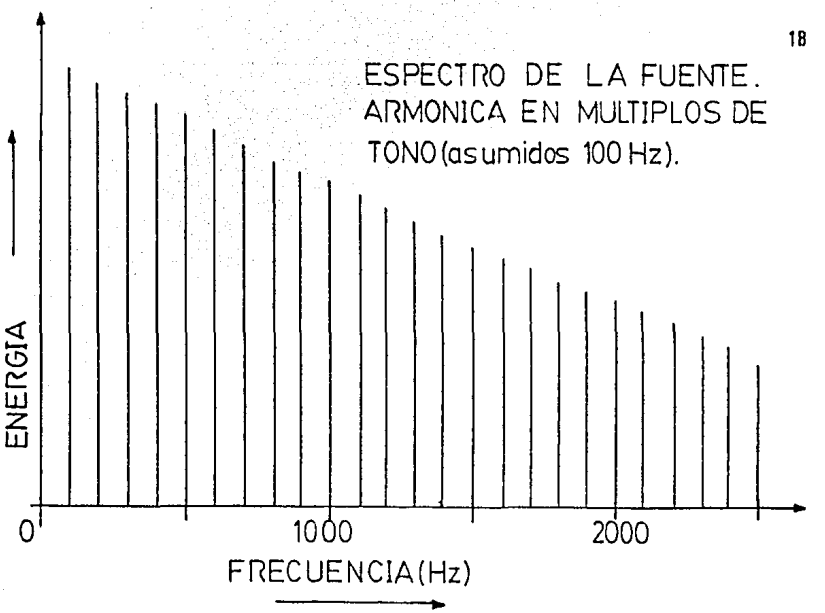
	Sorda	Sonora
Oclusivas:		
Bilabial	/p/	/b/
Linguodental	/t/	/d/ (Ej. dedo)
Linguovelar	/k/	/g/ (Ej. gana)
Africadas:		
Linguopalatal	/c/ (Ej. mucho)	[J] (Ej. cónyuge) alófono de /J/
Fricativas:		
Labiodental	/f/	/v*/ (francés: vin), alófono de /d/
Linguointerdental	/Ø/ (Ej. caza)	[ð] (Ej. dedo) alófono de /d/
Linguoalveolar	/s/	/Z*/ (inglés: zoo)
Linguopalatal	/Σ*/ (inglés: she)	/J/ (Ej. raya)
LInguoalveolar	/x/ (Ej. cajón)	[γ] (Ej. mago)
Nasales:		
Bilabial		/m/
Linguoalveolar		/n/
Linguopalatal		/ρ/ (Ej. maña)
Laterales:		
Linguoalveolar		/l/
Linguopalatal		/λ/ (Ej. calle)
Vibrantes:		
Simple		/r/ (Ej. cero)
Múltiples		/r^/ (Ej.cerro)

1.3 Módelo de Filtro-Fuente de producción de voz.

El módelo de filtro-fuente de producción de voz [40] se refiere a cuando hablamos en términos de una fuente de sonido (sordo o sonoro) por excitación de la resonancia del tracto vocal (y posiblemente el nasal). Este módelo, el cual es usado extensivamente en análisis de voz, es conocido como el módelo de filtro-fuente de producción de voz. La razón para estos sucesos es que el efecto de la resonancia puede ser modelado como un filtro de frecuencia selectivo, operación sobre una entrada lo cual es la fuente de excitación. Así, el espectro de frecuencia de la fuente es modificado por multiplicación por la característica de frecuencia del filtro (o sumadas, si las amplitudes son expresadas logarítmicamente).

Esto lo podemos comprender por la fig. 1.3-1, la cual muestra un espectro fuente y un filtro característico el cual combinado forma un espectro completo. Aunque, como se menciono antes, hay varios fricativos en la cual no son sujetos a la resonancia del tracto vocal a la misma extensión que los sonoros y los sonidos aspirados que son, ello puede ser fijamente moldeado como una fuente de ruido seguido por un filtro para darnos diferentes calidades de sonido.

El módelo de filtro-fuente es una sobre simplificación del sistema de producción del habla actual. Esto es inevitablemente, algún acoplamiento entre el tracto vocal y los pulmones, seguido por la glotis, duración del periodo cuando esta abierto. Esto efectivamente hace que el filtro característico cambie durante cada ciclo individual de excitación. Una implicación muy interesante del módelo de filtro-fuente es la característica prosódica de tono y amplitud que son propiedades importantes de la fuente, mientras un segmento es introducido por el filtro, esto se discute ampliamente en el Cap. III.


ESPECTRO DE LA FUENTE.
ARMONICA EN MULTIPLOS DE
TONO (asumidos 100 Hz).
ENERGIA
0
1000
2000
FRECUENCIA (Hz)

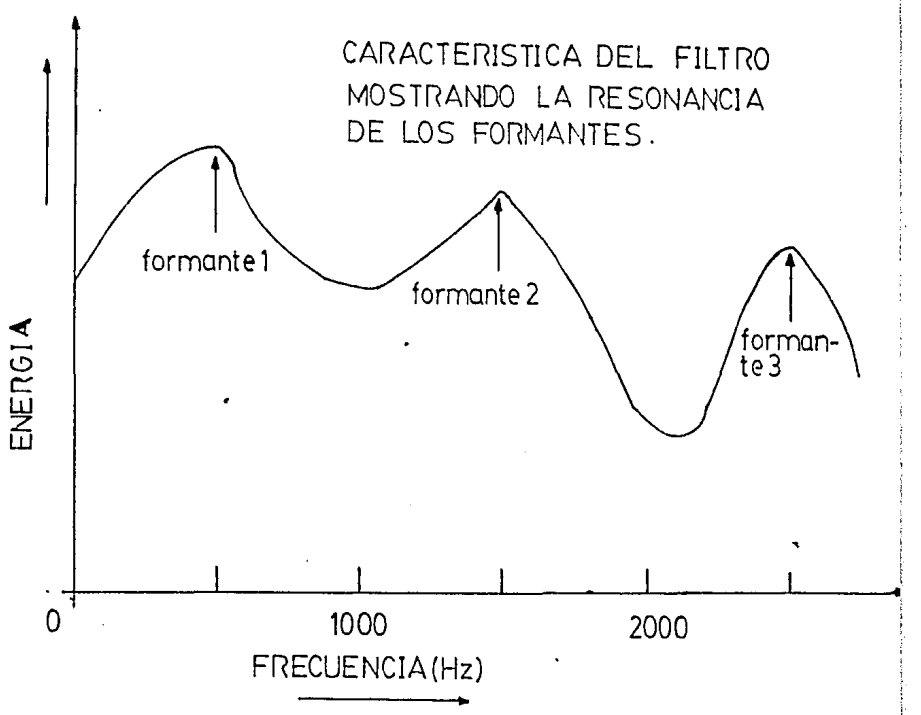

CARACTERISTICA DEL FILTRO
MOSTRANDO LA RESONANCIA
DE LOS FORMANTES.
formante 1
formante 2
formante 3
ENERGIA
0
1000
2000
FRECUENCIA (Hz)

FIG. 1.3-1

II - Tipos de modulación, codificación digital y conversión analógica/digital(ADC).

Este capítulo se refiere a explicar las técnicas de modulación de señales que no son propiamente senoidales , estas técnicas de modulación como, los sistemas de AM y FM, entre otros, están diseñados para manejar señales más complejas que una secuencia de dígitos binarios. Estas señales podria ser la voz humana, televisión, etc., como se encuentra en la transmisión normal. De igual forma, se menciona otras técnicas que se utilizan en el procesamiento de la señal de voz, como por ejemplo, el PCM(modulación por codificación de pulsos) y por ende sus derivadas(DM, ADPCM, APCM, etc.). Pero antes, de pasar al campo de las comunicaciones, se explicará primero en que consiste la codificación digital, el teorema del muestreo, la cuantización, lo cual se incluye dentro de lo que es la codificación y por último lo que es la conversión A/D.

2.1 Codificación Digital.

a) Teorema del Muestreo.

Consideremos una señal f(t), que varía continuamente, en la que deseamos convertir a su forma digital. Esto se logra primero al muestrear a f(t) a una velocidad de fc muestras por segundo, fig. 2.1-1. Supongamos que un interruptor permanece en la línea de f(t) durante los τ segundos cuando se encuentra girando a la velocidad de fc=1/T deseada veces por segundo ($\tau << T$). La salida del interruptor fs(t) es entonces una versión muestreada de f(t), fig. 2.1-2.

La señal muestreada fs(t) contiene toda la información de f(t). Además, a partir de fs(t) se obtiene f(t). ¿ Cómo podemos comprobar esto ? Suponemos en primer lugar que la señal f(t) es de banda limitada a ß Hertz, lo que nos quiere decir que se encuentra absolutamente libre de componentes de frecuencia por encima de f=ß. La transformada de Fourier [43] F(w) de este tipo de señal se puede contemplar en la fig. 2.1-3.

Las señales físicas que se presentan normalmente no tienen la característica de corte abrupto+ de frecuencia que se supone en la banda limitada. Con la señal f(t) limitada en banda a ß Hertz se demuestra fácilmente que muestreando la señal no se destruye ningún contenido de información de la misma, siempre que la velocidad de muestreo fc >= 2ß. La mínima velocidad de muestreo de 2ß veces por segundo se denomina velocidad de muestreo de Nyquist y 1/2 ß se llama intervalo de muestreo de Nyquist. para demostrar este resultado por medio del análisis de Fourier [43] hacemos lo siguiente: es evidente que la señal muestreada fs(t) puede representarse en términos de

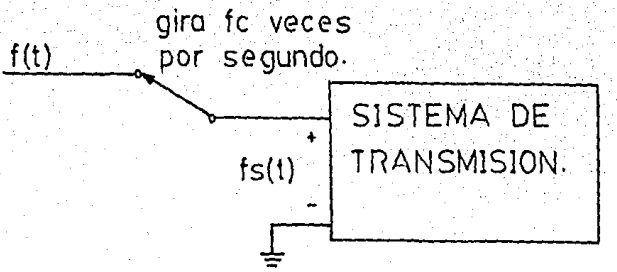


FIG. 2.1-1
MUESTREO DE UNA SEÑAL ANALÓGICA.

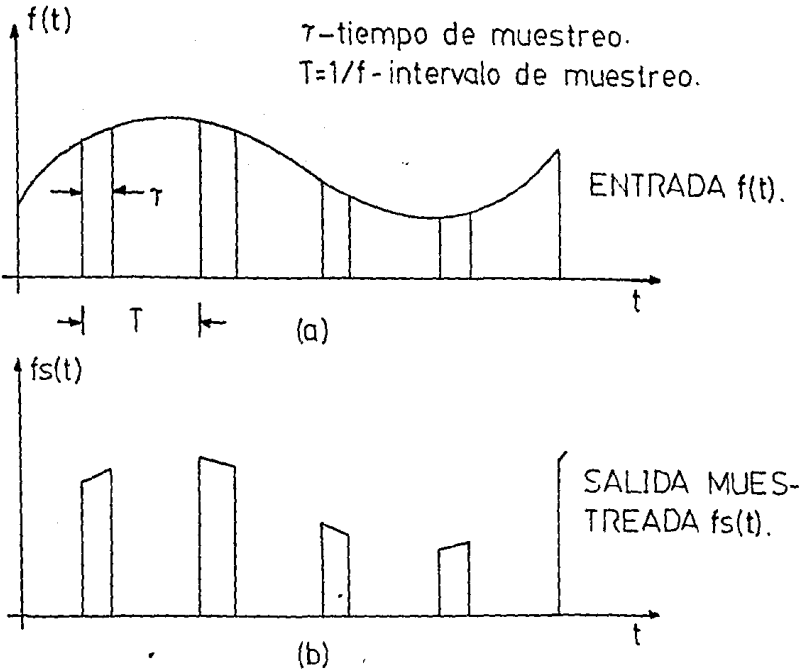


FIG. 2.1-2
PROCESO DE MUESTREO.

f(t) por medio de la relación siguiente:

$$fs(t) = f(t)S(t)$$

donde S(t) es una serie de pulsos periódicos de amplitud unitaria, de τ, ancho y periódo T=1/fc. Esta relación se utiliza para reducir el espectro Fs(w) de la señal muestreade fs(t). Para esto utilizamos el teorema de la convolución en la frecuencia, entonces tenemos:

$$Fs(\omega) = 1/2\pi\, F(\omega) * S(\omega)$$

$$(1) \qquad = 1/2\pi \int_{-\infty}^{\infty} F(x)S(\omega - x)\,dx$$

Pero la transformada de Fourier de la función de muestreo periódica S(t) es, según

$$wn = \frac{2\pi\eta}{T}$$

$$F(\omega) = \frac{2\pi\tau Am}{T} \sum_{\eta=-\infty}^{\infty} \left[\frac{\operatorname{sen}(\omega n\, \tau/2)}{\omega n\, \tau/2}\right] \delta(w - wn)$$

igual a un conjunto infinito de funciones impulso idénticamente n espaciados en frecuencia :

$$d = \tau/T \qquad wc = 2\pi/T$$

$$(2) \qquad S(\omega) = 2\pi d \sum_{\eta=-\infty}^{\infty} \left(\frac{\operatorname{sen} n\pi d}{n\pi d}\right) \delta(w - nwc)$$

En la práctica [42], se utilizan filtros pasabajas de corte abrupto, los que se introducen frecuentemente antes del proceso de muestreo para asegurar que la condición de limitación de banda se cumpla con la aproximación deseada. Si insertamos (2) en (1) y notando que

$$\int_{-\infty}^{\infty} F(x)\delta(w' - x)\,dx = F(w')$$

obtendremos:

$$Fs(\omega) = dF(\omega) + d \sum_{\eta=-\infty}^{\infty} \frac{\operatorname{sen} n\pi d}{n\pi d} F(w - nwc) \qquad n <> 0$$

que llegamos a lo que se quiso encontrar, la transformada de Fourier de fs(t). Existe una relación entre la velocidad a la cual la señal está variando y el numero de pulsos que son necesarios para reproducirla con exactitud. La velocidad a que una señal varia es, por supuesto, un valor que depende de su máxima componente de frecuencia o ancho de banda, B. Sabemos que se necesitan al menos 2B muestras uniformemente espaciadas por cada segundo para reproducir con seguridad la señal sin distorsión. Este aseveración,

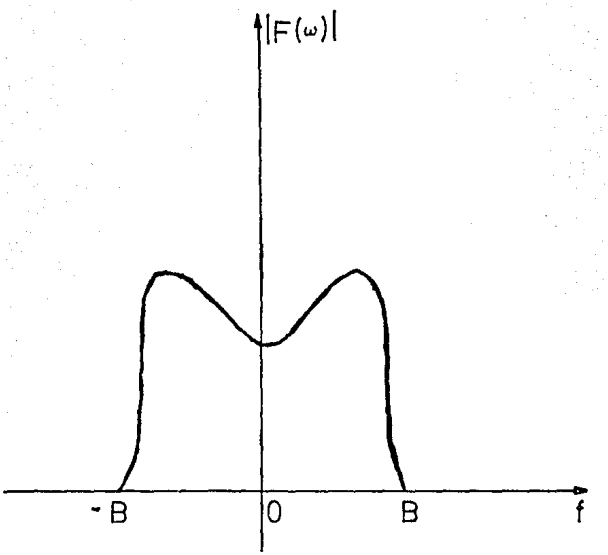


FIG. 2.1-3
SEÑAL DE BANDA LIMITADA.

que deducimos en forma natural de las consideraciones hechas sobre el espectro de una señal muestreada en forma periódica, es el famoso teorema del Muestreo. Aunque el enunciado de este teorema se deduce de la ecuación fc=2B, establece una relación del periódo del muestreo de una señal de banda limitada, el teorema [42] puede generalizarse a cualquier conjunto de muestras independientes. En esta forma, el teorema general establece:

> cualesquiera 2B muestras independientes por segundo caracterizarán por completo una señal de banda limitada. Dicho de otra forma, cualesquiera 2BT' trozos (independientes) de información son suficientes para especificar completamente una señal durante un intervalo de T' segundos de duración.

sobre esto y más información sobre Transformada de Fourier, ver Análisis de Fourier [42] [43].

b) Cuantización.

El proceso de digitalización de las señales análogicas originalmente se conoce como el proceso de cuantización, éste consiste en la subdivisión de la amplitudes de las señales en un predeterminado número de niveles discretos de amplitud. Las señales resultantes se denominan cuantizadas. La diferencia entre el proceso de muestreo y el proceso de cuantización es que en este último, se produce una pérdida irreparable de información , debido a que es imposible reconstruir la señal análogica original a partir de su versión cuantizada [42]. Ejemplo de una señal [42] en la cual cuantizamos y muestreamos simultáneamente; lo podemos apreciar en la fig. 2.1-4.

La variación total de amplitud de Ao = 7 se divide en los niveles de amplitud igualmente espaciados a = 1V. de separación. Existen asi, M = Ao/a+1 posibles niveles de amplitud, incluyendo el nivel cero. Aunque la separación entre niveles que se muestra es uniforme, con frecuencia en la práctica dicha separación se hace no uniforme con el objeto de mejorar el comportamiento del sistema al ruido. En particular, el espaciamiento de los niveles se hace disminuir con los niveles bajo de amplitud. Esto se realiza por medio de una técnica llamada compresión. Una aproximación cuantificada de una señal de 8 niveles se visualiza en la fig. 2.1-5. El ruido de cuantización (errores presente en el proceso de cuantización) puede reducirse, por supuesto, disminuyendo la separación de niveles (a) o acrecentando el número de niveles M empleados. Tipicamente, si utilizamos una conversión de 11 bits para la señal de voz, obtendremos una cuantización de 2048 niveles, y la señal enconcontramos que es adjustado a cero, lo cual significa que la mitad del pulso corresponde

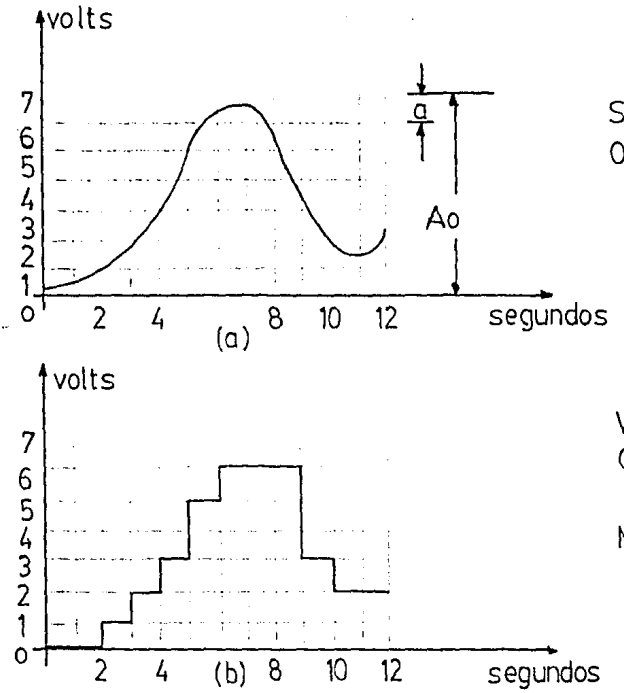


FIG. 2.1-4
CUANTIZACIÓN Y MUESTREO.

M-niveles.
a-espaciamiento de a volts.
P-valores negativos o positivos en volts.

$$a = \frac{P}{M} = \frac{2V}{M}$$

$$A = (M-1)a \text{ Volts.}$$

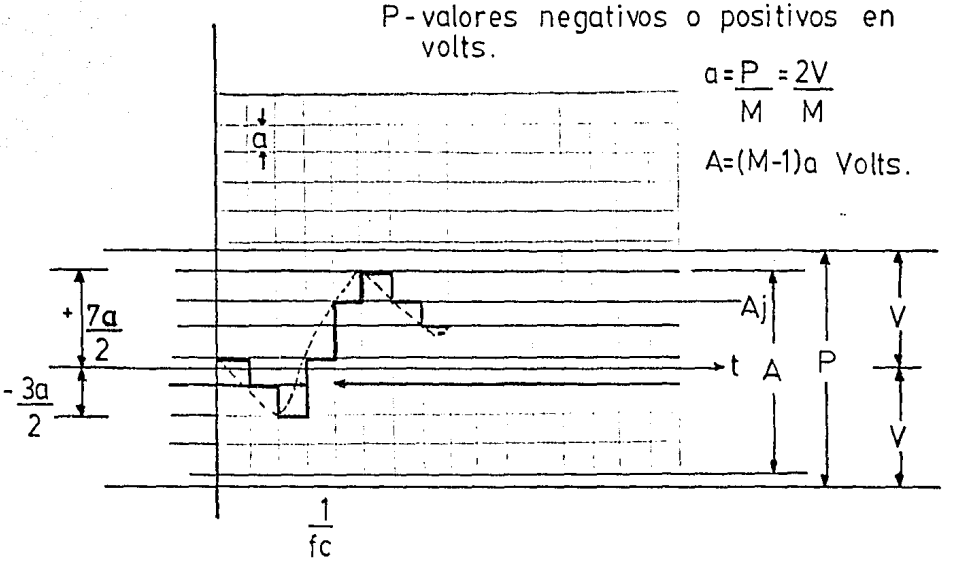


FIG. 2.1-5

APROXIMACION CUANTIZADA
DE UNA SEÑAL: OCHO NIVELES.

a un voltaje de entrada negativo y la otra parte a un voltaje de entrada positivo.

Esto es, en primera instancia, sorprendente, de modo que 11 bits son necesarios para una adecuada representación de la señal de voz. Para el propósito de procesamiento de la señal de voz, es esencial que la señal tenga una cuantización uniforme. Esto es porque todas las teorías aplicadas a sistemas lineales, y no-lineales introduce complejidades el cual no son responsable hacia el análisis. La cuantización uniforme, aunque es una operación no-lineal, es lineal en cierto caso limitado como el número de niveles que tiende a crecer y para un mayor propósito el resultado puede ser formado por presunción que la señal cuantizada es obtenida desde una análogica original por la adición de una pequeña cantidad de ruido cuantizado uniformemente distribuido, Usualmente, el ruido de cuantización es ignorado en los siguientes análisis subsecuentes.

La cuantización lineal [40] tiene un efecto desafortunado la cual el nivel de ruido absoluto es independiente del nivel de la señal, así que un número excesivo de bits debera de ser usado si el promedio o razón razonable es logrado para señales de pico. Esto puede ser entendido por la siguiente representación logarítmica:

$$y = 1 + k(\log x)$$

donde x es la señal original, y es el valor el cual es a ser cuantizado, dado un SNR el cual es independiente del nivel de la señal de entrada. Esta relación no puede ser realizada físicamente, cuando la señal es negativo, encontramos que es indefinido y diverge cuando es cero. También, encontramos que una aproximación es realizable hacia ella tal que puede retener las ventajas de la SNR constante dentro del rango usual de amplitudes de la señal.

La idea de una cuantización no-linealmente de una señal hacia el logro adecuado de la SNR para una gran variedad de amplitudes es lo que conocemos como " compansión ", lo cual también se menciona en lo que se refiere a los tipos de modulación más comunes y utilizados.

1. Códigos utilizados en el procesamiento de señales.

La codificación de niveles de amplitud en forma binaria puede ser realizada de varias maneras; uno de los procedimientos es emplear la conversión usual entre decimales y binarios, como se puede ver en la siguiente tabla [49]:

Tabla 2.1-1
Código binario.

DECIMAL	BINARIO
0	0000
1	0001
2	0010
3	0011
.	.
.	.
.	.
9	1001

Una de las dificultades que se presentan con la conversión normal de decimal a binario es que, al cambiar de un dígito decimal a otro adyacente, el código binario se modifica en un número variable de dígitos binarios. En PCM de 7 bits , donde se maneja los niveles entre 0 y 127, pueden variar hasta siete dígitos binarios lo cual hace que el código binario sea altamente susceptible de error durante la conversión A/D. El código Gray [48] [49] es popular porque sólo un dígito se modifica a la vez cuando el código decimal cambia de un nivel a otro. Para obtener el código gray a partir del código binario, tenemos las siguientes ecuaciones de conversión:

$$b1 = g1$$
$$bk = gk \ (or\text{-}excl) \ bk\text{-}1 \qquad K >= 2$$

Se suma el mismo número binario sin acarreo pero corriendo previamente el número que se va a sumar un lugar a la derecha y eliminando el dígito que sale. Ejemplo de este código:

Tabla 2.1-2
Código Gray.

DECIMAL	BINARIO	GRAY
0	0000	0000
1	0001	0001
2	0010	0011
3	0011	0010
.	.	.
.	.	.
.	.	.
9	1001	1101

Existen, también otros tipos de códigos que son utilizados en la transmisión de información y por ende, forman parte de los diferentes tipos de códigos que hay. Por ejemplo, tenemos el código llamado exceso-3 [48], en la cual es un código no-pesado; este no tiene valores asignados para cada bit de posición. El número tres es sumado al digito decimal dado y el resultado es convertido a su forma binaria. Ejemplo:

Tabla 2.1-3
Código excess-3.

DECIMAL	EXCESS-3
0	0011
1	0100
2	0101
3	0110
.	.
.	.
.	.
9	1100

La ventaja de este código es que se autocomplementa; en otras palabras, si todos los bits 0 o 1 en un número excess-3 son complementados, obtenemos los nueve complemento del número. También, tenemos el código 2421 [48], el peso es 2, 4, 2, 1 desde la posición MSB(byte más significativo) a la posición LSB(byte menos significativo). Ejemplo:

Tabla 2.1-4
Código 2421.

DECIMAL	2421
0	0000
1	0001
2	0010
3	0011
.	.
.	.
.	.
9	1111

2.2 Conversión Analógica-Digital(A/D).

El propósito de explicar los tipos de convertidores A/D es porque constituye una de las fases importantes en el análisis de reconocimiento de voz.

a) Una manera de convertir la señal análogica en digital es el método de aproximaciones sucesivas [44], mostrado en la fig. 2.2-1. Este método de aproximaciones

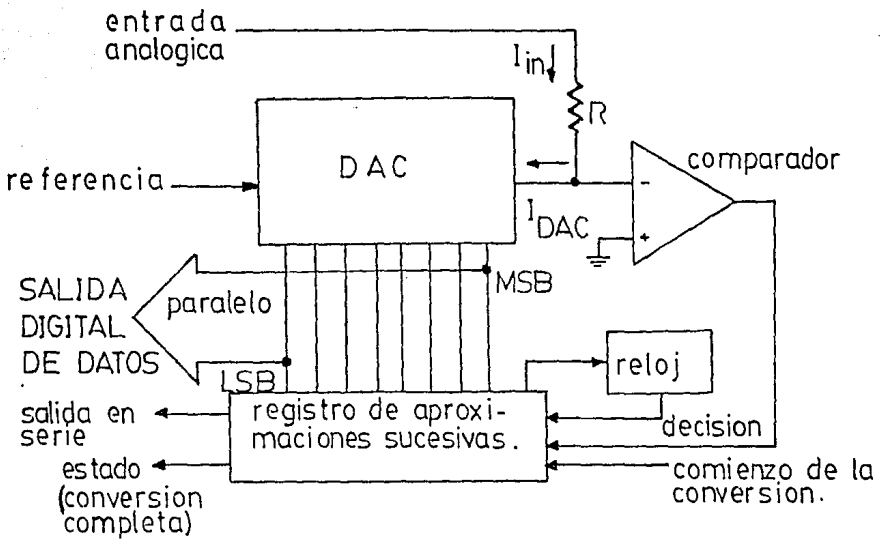


FIG. 2.2-1
ADC DE APROXIMACIONES
SUCESIVAS.

sucesivas es popular porque puede combinar resoluciones útiles hasta 12 bits y más, con un tiempo bastante breve de conversión(menos de 12 ms para la conversión de 12 bits). Otra ventaja adicional es que el tiempo de conversión es fijo e independiente de la magnitud de entrada, permitiendo una interfaz eficiente con los microprocesadores. Sus incovenientes principales son su propensión a los cambios de entrada durante la conversión(incluyendo ruidos) y tiene el costo más alto por bit, en comparación con el método de integración que se explicará más adelante.

Las aproximaciones sucesivas son similares a sopesar una masa en una balanza de precisión, utilizando pesas bien conocidas y cuyo valores forman una progresión binaria o BCD. Un convertidor de aproximaciones sucesivas consiste en un convertidor D/A, un comparador de voltajes o de corrientes, un reloj, un registrador de cambios, lógica de control y un registro de salida. La entrada básica de control es una línea de conversión de partida. Puesto que los datos de salida no son válidos hasta que se complete la conversión, una línea de estado de conversión indica que el convertidor esta ocupado(busy), mientras la conversión se encuentra en curso.

La función del convertidor A/D es la siguiente:

1. Cuando se aplica la orden de conversión, la línea de estado pasa a ocupado, se inicializa el registro, con excepción de un 1 en la posición MSB y se enciende la compuerta del reloj. El 1 en la posición MSB hace que la corriente MSB aparezca en la salida del DAC (entrada del comparador) con la señal de entrada análogica. Si la salida del DAC es menor que la entrada de señal, la salida del comparador indicara retener; si la salida del DAC es mayor, el comparador indicará rechazo. En la siguiente pulsación del reloj, el MSB se enclava en 1 o 0, dependiendo de la decisión. El bit 2 (1/4) de escala completa se agrega a la salida del DAC. Si la salida del DAC (1/4 + 1/2 ó 1/4 + 0) es menor que la señal, la decisión es retener; de otro modo, rechazar. En la siguiente pulsación del reloj, se enclava el bit 2 y se agrega el bit 3 (1/8 de escala completa) a la salida del DAC. Si la salida del DAC (1/8 + 1/4 + 1/2, 1/8 + 1/2, 1/8 + 1/4 ó 1/8 + 0) es menor que la señal, la decisión es retener; de otro modo seria rechazar. El proceso se repite hasta que el LSB se haya probado y enclavado, después de lo cual se apaga el reloj y la línea de estado indicará conversión completa. La salida de datos digitales en paralelo se encontrara disponible y el convertidor estará listo para una nueva conversión. Existen también datos en serie puesto que los bits se ciclan en serie. Cada bit en serie (no regreso a cero, NRZ) se hace válido con el borde delantero de cada pulsación del reloj.

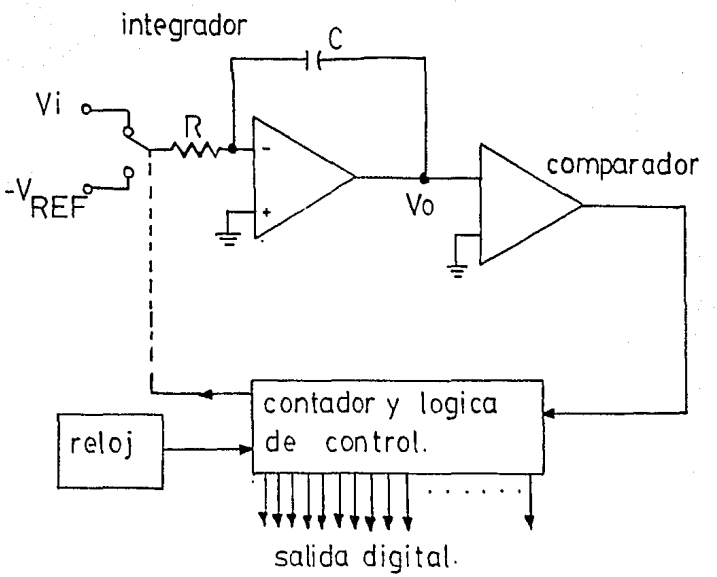


FIG. 2.2-2
ADC INTEGRADOR DE PENDIENTE DOBLE.

2. El tiempo de conversión para un convertidor de aproximaciones sucesivas de n bits es, por lo menos, igual al tiempo de n pulsaciones del reloj.

b) Otro método es el de conversión por integración.

Un ADC integrador mide el tiempo que se necesita para que una salida del integrador atraviese una gama de voltajes proporcional al valor promedio de la entrada a un indice constante(referencia), fig. 2.2-2.

Al inicio de la conversión, el integrador esta desengachado y empieza a integrar la señal de entrada Vi, al mismo tiempo, el contador empieza a contar pulsaciones del reloj. Cuando se han contado N1 pulsaciones del reloj después de un periodo t1, el contador cede y conmuta la entrada. Se aplica al integrador una referencia (-Vref) de polaridad opuesta a la señal, el integrador empieza a integrar en dirección opuesta a un índice constante y el contador empieza a contar. Cuando la salida del integrador llega al valor inicial, el comparador se dispara y la conversión está completa. El reloj se detiene y el integrador se enclava en su valor inicial. El número de conteos N2, que indica el tiempo t2, es proporcional al valor promedio de la entrada Vi.

La salida del contador puede ser binaria o de BCD. Pero el método más habitual es el de BCD por sus usos en medidores digitales de tableros (DPM) y otros dispositivos de presentación.

La ventaja y desventaja de este tipo de conversión es que los ADC de integración son mucho más lento que del anterior, el de aproximaciones sucesivas. Sin embargo, tienen una exactitud potencial mucho mayor, no omiten claves, tienen un mejor rechazo de ruidos y, puesto que se basan en menos piezas de alta precisión, tienen tendencia a ser de costo más bajo. En general, son su simplicidad, su bajo costo y su compatibilidad con la tecnología de circuitos integrados lo que los hace preferibles para los DPM.

c) Relaciones de conversión y errores.

En la fig. 2.2-3 vemos como es que podemos relacionar la conversión de una señal y el error que puede ser ocasionado al ocurrir esta. El ADC tiene un error de compensación [45]: la primera transición puede no producirse exactamente a 1.2 LSB. Además, hay un error de factor de escala (ganancia); que es la diferencia entre los valores a los que se producen en la primera transición y la última no es igual a FS (1 - 2 LBS) y un error de linealidad; que es la diferencia entre valores de transición no son todas iguales ni varian uniformemente.

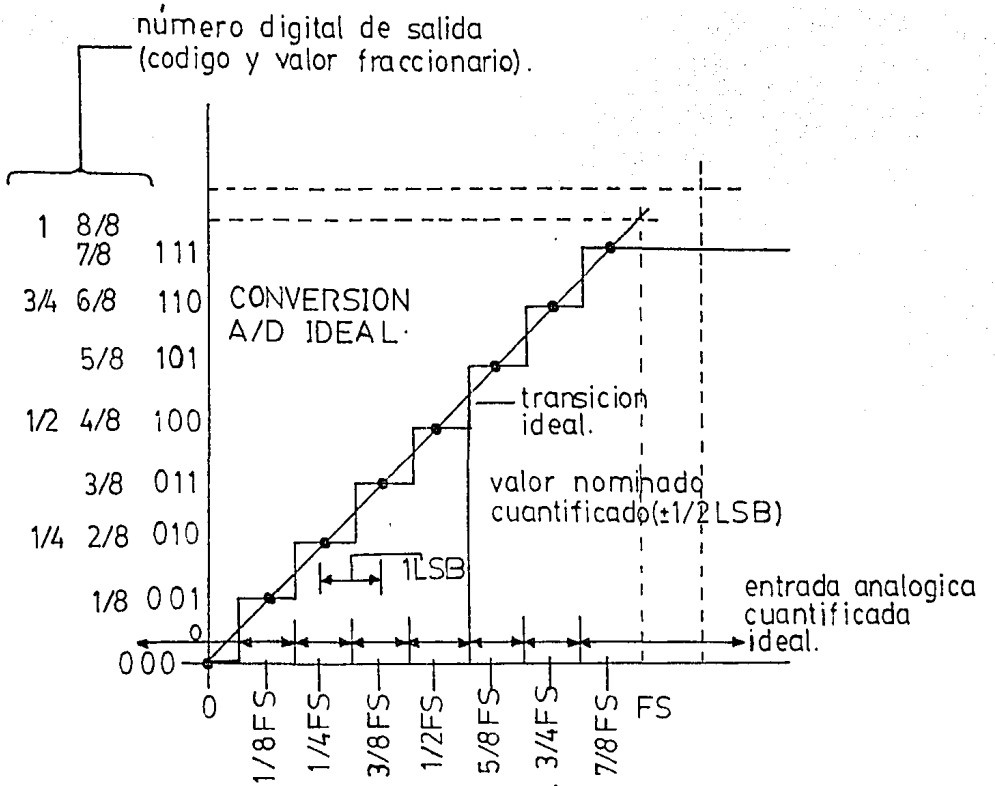
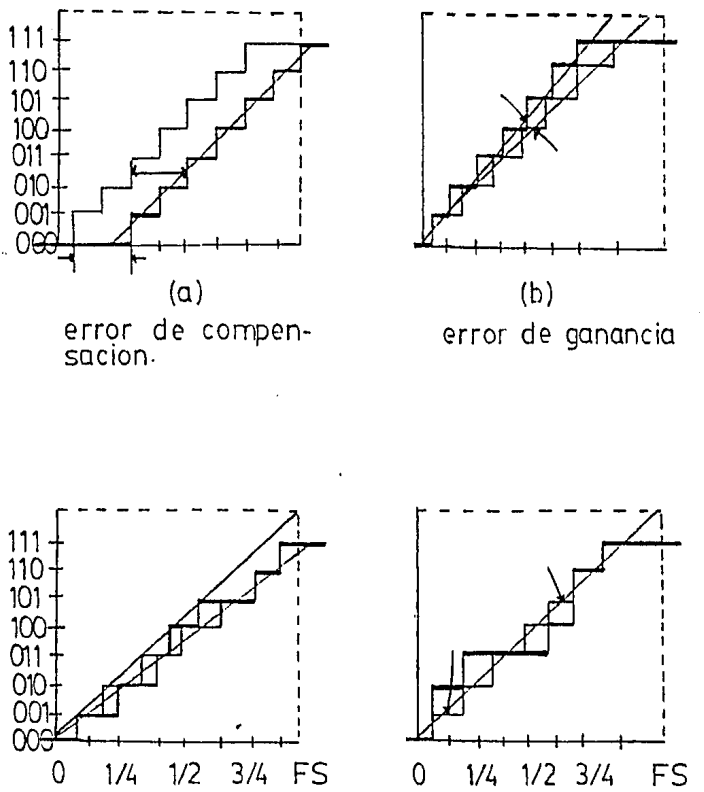


FIG. 2.2-3
RELACIONES DE CONVERSION DE UN ADC
MONOPOLAR DE 3 BITS.



(a)
error de compensacion.

(b)
error de ganancia

(c)
no linealidad

(d)
codigos omitidos debido a la no linealidad diferencial excesiva.

FIG. 2.2-4
RELACIONES DE CONVERSION DE UN ADC DE 3 BITS CON INDICACION DE LOS ERRORES.

Si la linealidad diferencial es suficientemente grande existirá la posibilidad de que se omiten uno o más códigos.

Los convertidores de integración no tienen problema de linealidad diferencial puesto que no usan DAC ni tampoco una gran cantidad de elementos de precisión. Las relaciones distintas de esos elementos provocan una no linealidad diferencial. Por consiguiente, los convertidores de integración no tienen códigos faltantes. En la fig. 2.2-4 se presentan los tipos de errores de los cuales se ha estado hablando.

d) Especificaciones que hay que considerar en todo convertidor A/D.

Generalmente, las especificaciones [46] que hay que considerar en todo convertidor A/D, como también D/A son los siguientes:

1. Precisión absoluta.
2. Precisión relativa.
3. Tiempo de conversión.
4. Convertidor de pendiente doble.
5. Alimentación transversal.
6. Ganancia.
7. Bit menos significativo(LSB).
8. Linealidad.
9. Linealidad diferencial.
10. Sensibilidad de la fuente de alimentación.
11. Convertidor de cuadrete pendiente.
12. Incertidumbre o error de cuantización.
13. Convertidor radiométrico.
14. Aproximaciones sucesivas.
15. Coeficientes de temperatura.
16. Coeficiente de temperatura de la ganancia.
17. Coeficiente de temperatura de la linealidad.
18. Coeficiente de temperatura de la compensación.
19. Coeficiente de temperatura de cero a monopolar.
20. Principio de ajuste de ganancia a cero.

2.3 Tipos de Modulación.

a) Modulación en Amplitud (AM).

En la modulación por amplitud [42] encontramos que cuando hay una variación de la envolvente, se habla del lugar geométrico de los valores picos de la portadora. En la fig. 2.3-1 muestra una señal moduladora típica f(t) y las variaciones de la envolvente de la portadora que corresponden a ella. La portadora modulada en amplitud puede describirse de la siguiente forma:

$$fc(t) = k[1 + mf(t)]\cos wct$$

$$fc(t)=K[1+mf(t)]COSwct$$

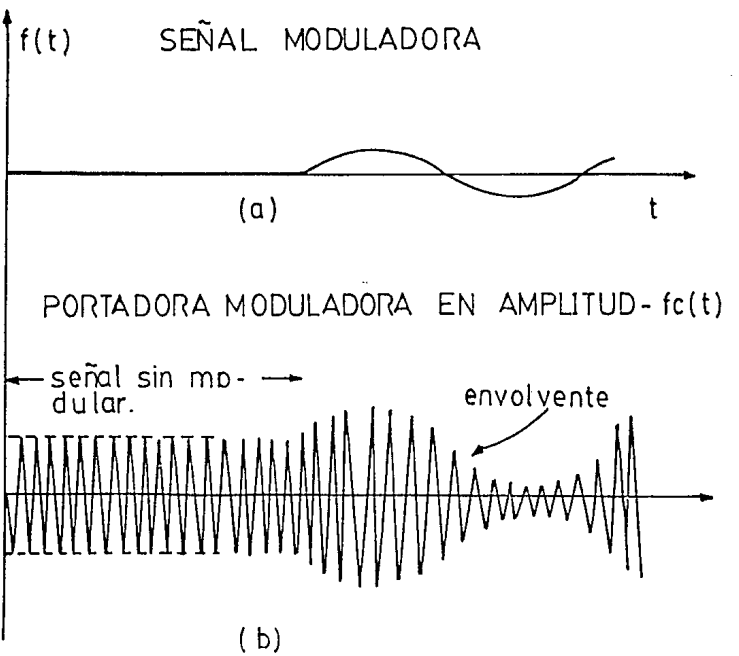
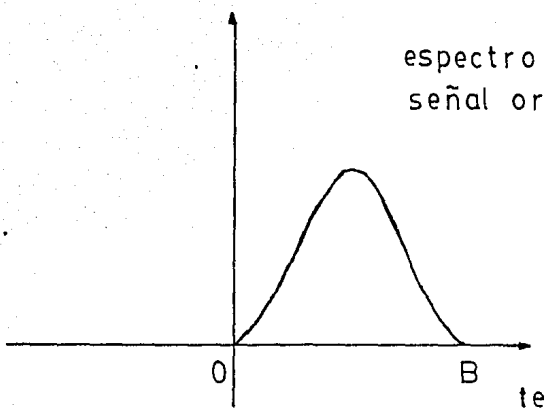


FIG. 2.3-1
MODULACION EN AMPLITUD DE UNA PORTADORA.



teorema de la modulación:

$v(t)\cos(wct)$

$1/2[V(f-fc)+V(f-fc)]$

$f = B$

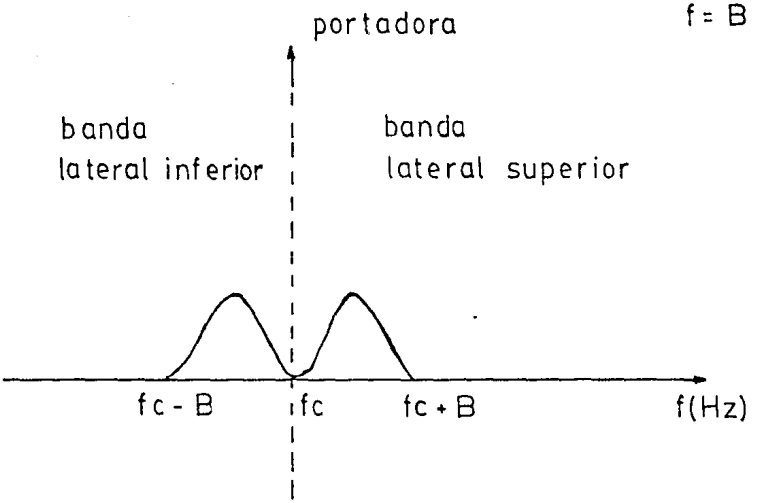


FIG. 2.3-2
ESPECTRO DE AM QUE INCLUYE - LA PORTADORA.

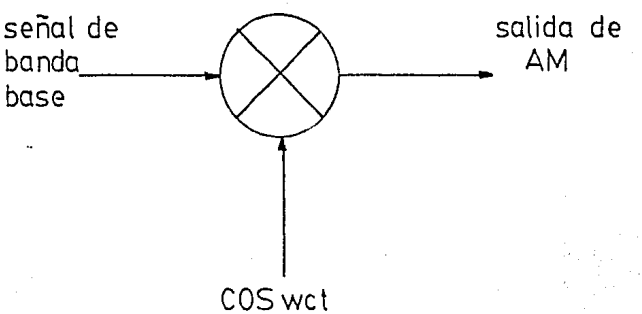


FIG. 2.3-3
MODULADOR PRODUCTO
PARA LA GENERACION
DE SEÑALES DE AM.

Sabemos, que /mf(t)/ < 1 , esto con el objeto de tener una envolvente sin distorsión. KCOS wct constituye la portadora, este término asegura la existencia de una envolvente. La razón de esto es que las señales moduladoras f(t) son más complejas, entonces agregando el término adecuado se asegura la presencia de la portadora. En el espectro de AM encontramos dos grupos de frecuencia laterales: una de banda superior y otra inferior. En la fig. 2.3-2 se muestra un espectro de AM que incluye la portadora, en la cual sólo se muestran las frecuencias positivas. Cada grupo contiene una banda de frecuencias corespondiente a la banda de frecuencias que cubre la señal original y esta se llama banda lateral. Cada banda lateral contiene todas las componentes espectrales, tanto de fase como de amplitud, de la señal original, por lo que contiene toda la información que lleva f(t).

Luego de todo esto, nos preguntamos: ¿cómo se produce fisicamente la señal de AM ? Un método podría ser, la conmutación de la señal f(t) entre encendido y apagado a la velocidad de la portadora. Este procedimiento traslada f(t) hasta todos los múltiplos armónicos de la frecuencia fc. Filtrando su pasabanda a una de estas frecuencias se obtiene la forma DSB(Doble Banda Lateral o sistema de portadora suprimida) de la señal de AM.

Para que tengamos a la salida la señal de AM, es necesario utilizar un dispositivo que realize el producto de las dos funciones de entrada f(t) y COS wct. Este dispositivo se conoce con el nombre de Modulador Producto en la cual lo apreciamos en la fig. 2.3-3.

Si la señal de banda base que se muestra aplicada a la entrada tiene la forma k[1 + mf(t)], se produce la salida de una AM normal con una portadora. Pero, si la entrada es simplemente f(t), o sea una señal de banda base sin la adición de una constante fija(o valor de cc), resulta DSB o AM de portadora suprimida a la salida. Por último, en la fig. 2.3-4 se describe los sistemas de modulación en amplitud.

b) Modulación en Frecuencia (FM).

La modulación en frecuencia [42] requiere muchos de banda más grande que los sistemas descritos anteriormente. Lo que hace popular el uso de FM es que proporciona una mejor discriminación contra el ruido y las señales de interferencia. La modulación en frecuencia la podriamos explicar de la siguiente manera, la frecuencia de la portadora se puede hacer variar de acuerdo con alguna señal específica que lleve información. Entonces la frecuencia de la portadora se escribe: wc + Kf(t) donde

(a) — AM NORMAL.
(b) — PORTADORA SUPRIMIDA O DSB.
(c) — SSB.

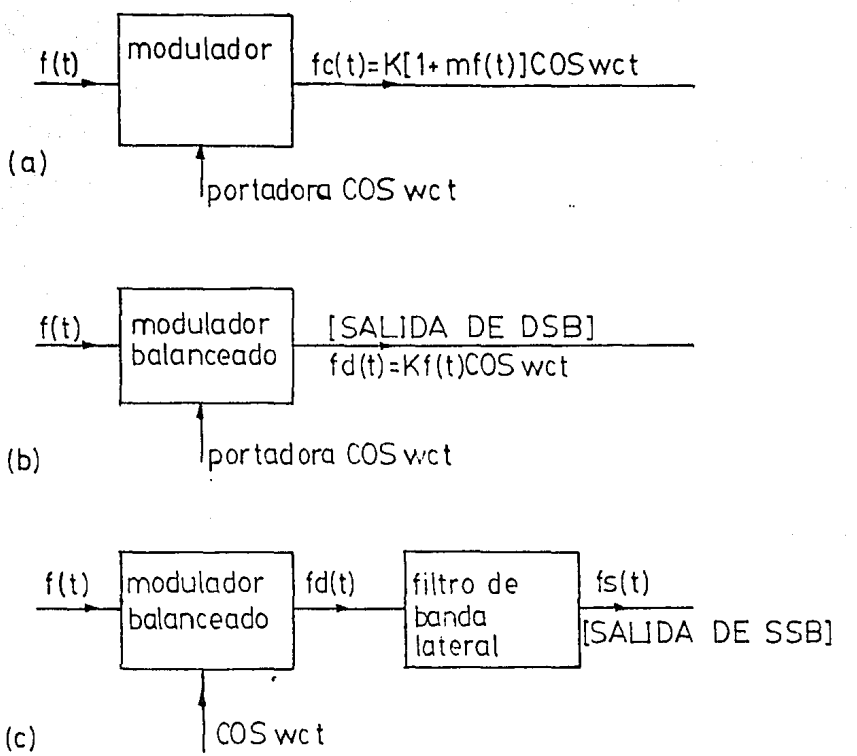


FIG. 2.3-4
SISTEMAS DE MODULACION EN AMPLITUD.-

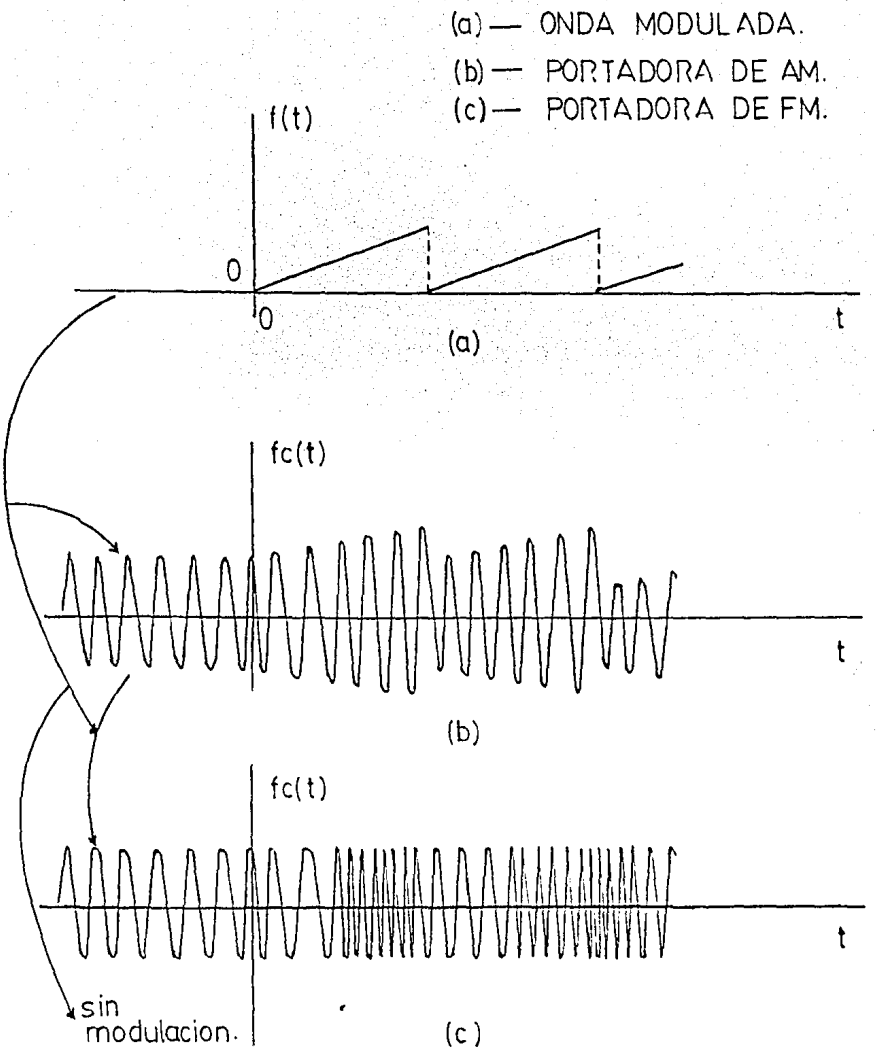


FIG. 2.3-5
MODULACION EN FRECUENCIA.

f(t) representa la señal y K es una constante del sistema. La modulación en frecuencia es un proceso no lineal, y por lo tanto, aparece nuevas frecuencias generadas. En la fig. 2.3-5, se describe la modulación en frecuencia, la cual nos muestra como esta constituida la portadora.

Como vemos, la señal de FM oscila más rápidamente con el aumento de la amplitud de la señal moduladora. El análisis del proceso de FM es, en esencia, mucho más complicado que el de AM, particularmente cuando se trata de una señal moduladora general. Esto se debe a la no linealidad del proceso de FM. La modulación en frecuencia la podemos dividir en FM de banda angosta, que es cuando la portadora senoidal modulada con $\beta << \pi/2$ y con $\beta > \pi/2$; y FM de banda ancha, consiste en cuando hay un incremento en el ancho de banda de la señal con el aumento de β. A su vez, la generación de señales de FM de banda ancha la podemos agrupar en dos clases:

1. FM indirecta.

Se utiliza primero la integración y modulación de fase para producir una señal de FM de banda angosta. Posteriormente, se emplea la multiplicación de frecuencia para incrementar el índice de modulación hasta el intervalo deseado de valores.

2. FM directa.

En este caso la frecuencia de la portadora se modula directamente o se varía conforme a la señal moduladora de entrada.

c) Modulación en Fase (PM).

Este tipo de modulación es parte dentro de la clasificación de la modulación angular, en la cual si $\emptyset(t)$ se le hace variar de alguna forma con una señal moduladora f(t); obtenemos la siguiente ecuación [49]:

$$\emptyset(t) = Wct + \emptyset o + K1f(t)$$

donde K1 es una constante del sistema, y por consiguiente con un sistema de modulación en fase. Aquí, vemos que es la fase de la portadora la que varia linealmente con la señal moduladora.

d) Modulación por Amplitud de Pulsos (PAM).

Esta técnica de modulación [51] consta de una secuencia de pulsos en la cual se considera alternativamente como una secuencia periodica de pulsos(portadora) cuya amplitud se modula de acuerdo con la información que se transmite. Esto es evidente en la forma de la expresión para los datos muestreados $fs(t)= f(t)S(t)$, en la cual la función de conmutación $S(t)$ representa la portadora de pulsos sin modulación y $f(t)$ la información que modula a la portadora. En la fig. 2.3-6 vemos un muestreo natural de la señal codificada por la amplitud de pulsos.

Esta forma de modulación permite la multicanalización en tiempo de muchos canales de información para la transmisión secuencial de ellos por un mismo canal simple.

e) Modulación por Codificación en el dominio del tiempo.

Hay diversas técnicas [40] de codificación en el tiempo para la forma de onda de la señal de voz en la cual podemos reducir el promedio de datos a una razón o proporción dada de señal a ruido(SNR) o alternativamente reducir el SNR para un promedio dado de datos. Todos estos, por supuesto requiere de más procesamiento, ambos en la codificación(para almacenar) y decodificación(para regeneración) al final del proceso de digitalización.

Encontraremos, más adelante que la simple técnica de codificación en el tiempo sera reemplazado por el más complejo método que es el de predicción lineal, la razón es, que este puede darnos una reducción substancial mayor en el promedio de datos para sólo una degradación pequeña en la calidad de voz. Esto se analizará detalladamente en el siguiente capitulo.

Antes de pasar a la clasificación general de técnicas de codificación en el dominio del tiempo, propiamente PCM(modulación por pulsos codificados, fig. 2.1-4), se explicará primero, en qué consiste el PCM. El PCM es un sistema de transmisión en la cual los digitos de la representación binaria del número son transmitidos como pulsos. En este tipo, la señal análogica $m(t)$ es muestreada y estas muestras son sujetas a la cuantización. Estas muestras cuantificadas son codificadas, luego el codificador responde a cada muestra dando como resultado un patrón de pulso binario. Fig. 2.3-7. Vemos que la combinación del codificador y el cuantificador forman lo que es el convertidor análogico-digital(ADC). Los sistemas digitales binarios constituyen la clase más común de sistemas PCM que se encuentran. Hay muchas ventajas en la utilización de este sistema y estas son las siguientes [42]:

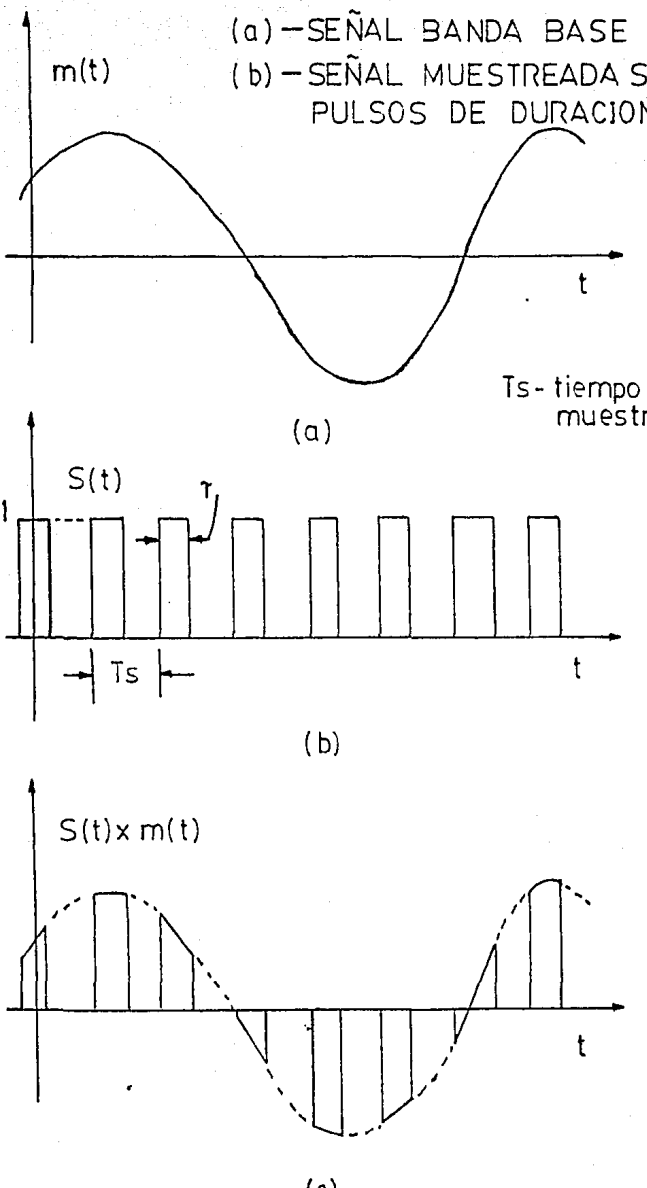


FIG. 2.3-6

MUESTREO NATURAL DE LA SEÑAL POR AMPLITUD DE PULSOS.

1.Las señales pueden regenerarse o rearreglarse periodicamente durante la transmisión, puesto que la información ya no se encuentra contenida en la amplitud continuamente variable de los pulsos, sino que consiste de simbolos discretos.

2. Toda clase de circuitos digitales puede emplearse durante la totalidad del procesamiento.

3. Las señales pueden ser procesadas digitalmente según convenga.

4. El ruido y la interferencia pueden ser propiamente minimizados mediante códigos, etc.

Ahora, que se dio una explicación general de lo que es el PCM, podemos presentar una clasificación general [40] de diferentes técnicas que son utilizadas conmúnmente en el reconocimiento de voz.

a) LINEAR PCM - Modulación por codificación de pulsos cuantizado-linealmente.

b) LOG PCM - Modulación por codificación de pulsos cuantizado logaritmicamente(compansión instántaneamente).

c) APCM - Modulación por codificación de pulsos cuantizado adaptivamente(usualmente compansión silábica).

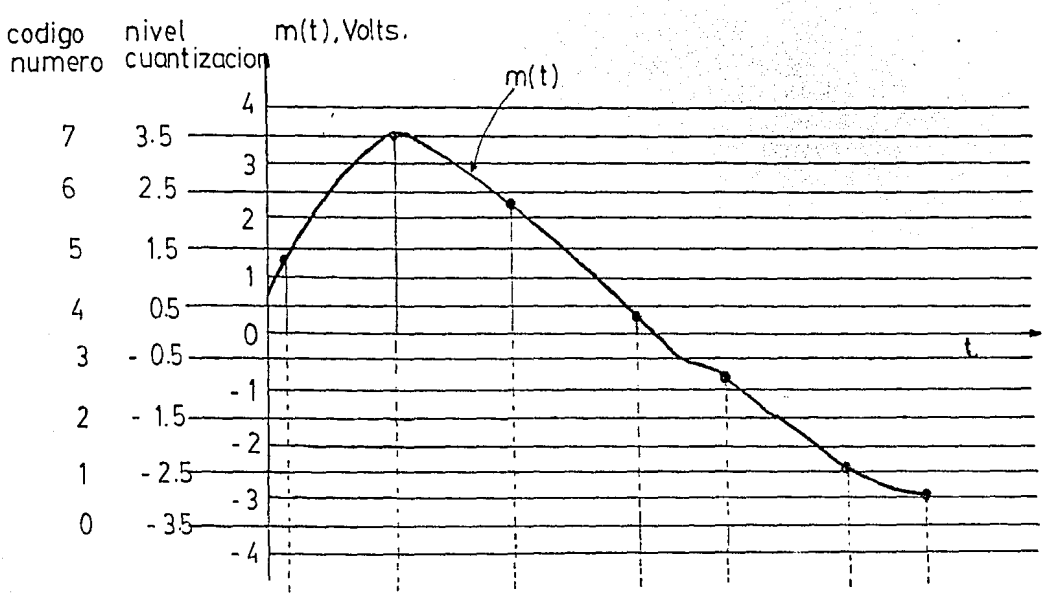
d) DPCM - Modulación por codificación de pulsos diferencial.

e) ADPCM - Modulación por codificación de pulsos diferencial con cualquiera de los dos, cuantización adaptiva o predicción adaptiva o ambos.

f) DM - Modulación delta(1 bit DPCM).

g) ADM - Modulación delta con cuatización adaptiva.

Anteriormente, habiamos mencionado una técnica de codificación en el dominio del tiempo, llamado cuantización logaritmica, o los PCM(algunas veces llamado "compansión instántanea"). Una codificación más sostificada en la amplitud total de la señal de voz y usando esta información para adjustar y cuantizar los niveles dinámicamente.



valor muestreo	1.3	3.6	2.3	0.7	-0.7	-2.4	-3.4
nivel de cuantizacion mas cerca.	1.5	3.5	2.5	0.5	-0.5	-2.5	-3.5
numero codigo	5	7	6	4	3	1	0
representacion binaria.	101	111	110	100	011	001	000

FIG. 2.3-7
CARACTERISTICAS ESENCIALES
DE UN PCM BINARIO.

Los métodos de codificación de voz basado sobre este principio son llamados sistemas de modulación por pulsos codificados adaptivo(ADPCM).; esto ocurre porque la amplitud completa cambia lentamente, y es esto suficiente para adjustar la cuantización relativamente no muy común (comparado con la razón de muestreo), y encontramos que es hecho frecuentemente a una razón de aproximación en la cual el promedio de sílabas de la velocidad del habla, es conocido con el término de "compansión silábica". Un bloque de formato de punto flotante puede ser usado, con un exponente común la cual tiende a ser almacenado cada muestras M(con M, decimos, 125 para una longitud de bloque de 100 msec a una muestra de 8 KHz), pero la mantisa puede ser almacenado como una razón de muestreo regular. Encontramos, que la energía completa en el bloque, es de

$$\sum_{n=h}^{h+M-1} x(n)^2 \qquad (M=125)$$

es usado para determinar el exponente conveniente o apropiado, y cada muestreo en el bloque nombrado $x(h), x(h+1), \ldots, x(h+M-1)$ es balanceado de acuerdo al exponente. Hay que hacer notar que para los sistemas de transmisión de voz este método necesita un retraso de muestras M al codificador, y claro algún método que se base en el exponente sobre la energía en el último bloque a evitar éste. Para el almacenamiento de voz, como siempre, el retraso es irrelevante. La forma de onda del DPCM, lo podemos ver en la fig. 2.3-8; mientras que el de ADPCM se muestra en la fig. 2.3-9.

De otra forma, no-silabica, el PCM adaptivo se refiere al continuo cambio del tamaño de medida o niveles de un cuantizador uniforme, por multiplicación de una constante de cada muestra el cual es basado sobre la magnitud del código previo de la palabra. La cuantización adaptiva aprovecha la información sobre la amplitud de la señal, y como una generalización áspera o ruda, produce una reducción de un bit por muestra en el promedio de datos para una locución de calidad-teléfonica sobre la cuantización logaritmica ordinaria, para una SNR dada. Alternativamente, para el mismo promedio de datos, un mejoramiento de 6 dB en la SNR puede ser obtenido.

En lo que se refiere a la modulación por codificación de pulsos diferencial(DPCM), en su forma más simple, se usa la muestra presente de voz como una predicción del próximo, y almacena el error de predicción: esto es, la diferencia de muestra a muestra. Esto como podemos ver es un caso simple de predicción predictiva.

Si parece admisible que el promedio de datos puede ser reducido por transmisión de la diferencia entre

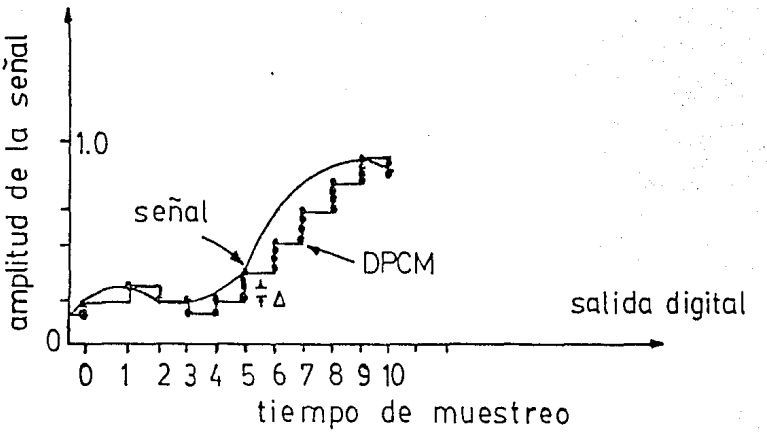


FIG. 2.3-8

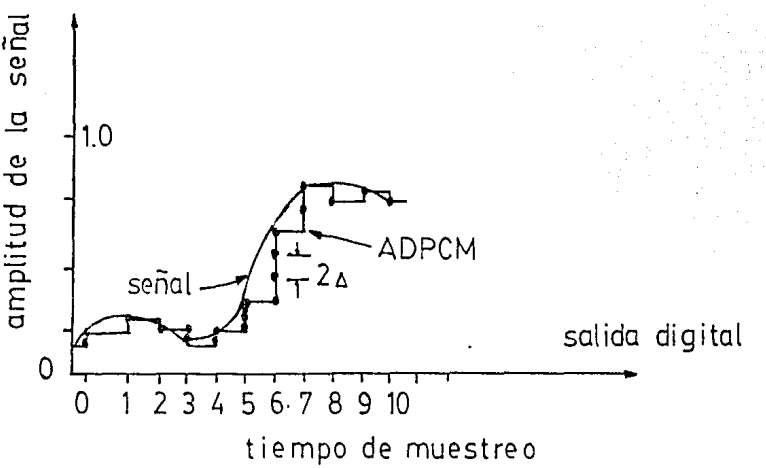


FIG. 2.3-9

ser reducido por transmisión de la diferencia entre muestras sucesivas en lugar de sus valores absolutos; entonces los bits menores son requeridos para la diferencia de señal para una exactitud completa dada porque no asume tal valores extremos como el nivel de señal absoluto.

Actualmente el mejoramiento no es tan bueno; cerca 5 dB en la de 4 a SNR, o justo debajo de un bit por muestra para una SNR dado, pero la diferencia de señal puede ser casi tan grande como el nivel de señal absoluto.

Si el DPCM es usado en conjunción con la cuantización adaptiva, dando una forma de modulación por codificación de pulsos diferencial adaptiva(ADPCM); ambos, la variación de amplitud completa y la correlación de muestra a muestra encontraremos que son utilizados, llegando a una ganancia combinada de 10-11 dB en la SNR(o justo debajo de 2 bits de reducción por muestra para una conversación de calidad-teléfonica). Otra forma de adaptación es alterar el predictor por multiplicación del valor de muestreo previo por un parámetro el cual es adjustado para una mejor representación. Entonces, la señal transmitida en un tiempo n es

$$e(n) = x(n) - ax(n-1)$$

donde el parámetro "a" es adaptado (y almacenado) sobre una escala de tiempo silábica. Esto lleva a un ligero mejoramiento en el SNR, el cual puede ser combinado con esos logros por cuantización adaptiva. Más de muchos beneficios substanciales puede ser realizado usando una suma pesada de diferentes muestreos anteriores de voz(arriba de 15), y adaptación de todos los pesos. A lo que me refiero es a una idea básica de lo que es la predicción lineal.

f) Modulación delta.

La modulación delta es otra técnica [40] en la cual la señal analógica es codificada en dígitos binarios(bits). De aquí que la modulación delta es un sistema PCM. La modulación delta, tiene el mérito de requirir la circuiteria electrónica para la modulación en el transmisor y particularmente para la demodulación en el receptor. Un diagrama de bloques de este tipo de modulación se puede apreciar en la fig. 2.3-10.

El generador de pulso provee un tren de pulsos pi(t) de amplitud y de polaridad fija. Suponemos que estos pulsos son arbitrariamente angosto pero fijo de una área finita. Ejemplo: los impulsos. Esta suposición es para simplificar la explicación de este tipo de modulación. El

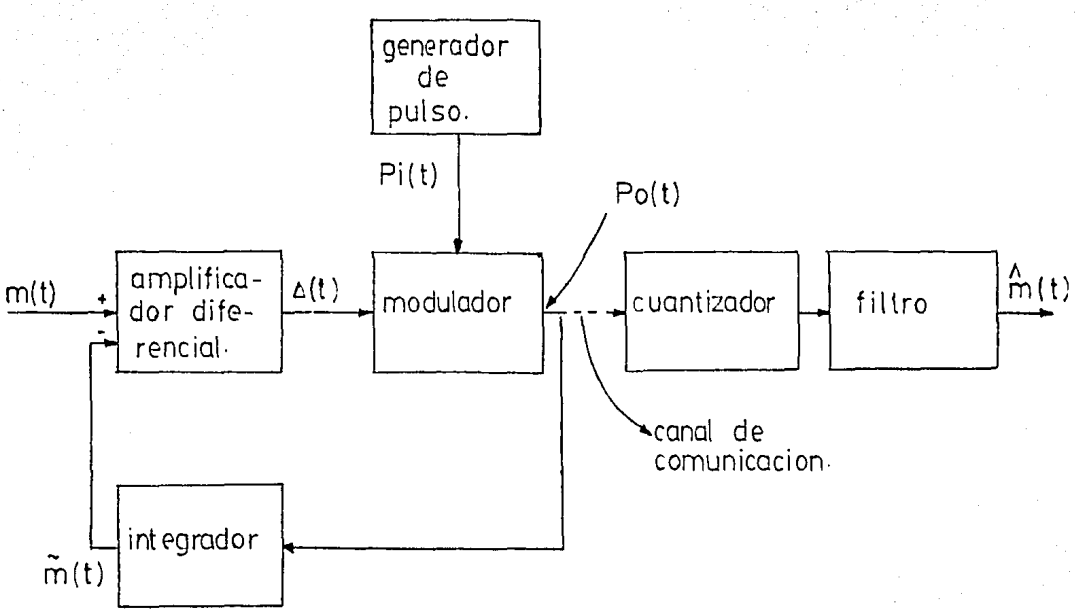


FIG. 2.3-10
SISTEMA DE COMUNICACION
MODULACION DELTA.

modulador recibe estos pulsos pi(t) de entrada como podria una señal dt. La salida del modulador po(t) es la entrada del tren de pulsos pi(t) multiplicado por +1 o por -1 dependiendo de la polaridad de dt (no de su magnitud). Si, dt es positivo cuando ocurre pi(t), la multiplicación es por +1, y si dt es negativo, entonces por -1. La forma de onda po(t) es aplicado a la salida del integrador, el cual es llamado m^(t). Cuando vemos, m^(t) es una aproximación de la señal de entrada m(t). La señal m(t) y m^(t) son comparados en un amplificador diferencial. La salida del amplificador dt es dado por dt = m(t) - m^(t). La operación de este modulador, su forma de onda se puede apreciar en la fig. 2.3-11. En esta figura, t=0, esto ocurre a la mitad del camino entre las ocurrencias del pulso. Los valores iniciales de m(t) y m^(t) son seleccionados arbitrariamente. En el tiempo t1 del primer pulso, encontramos que m(t) es más grande que m^(t). Por lo tanto, el pulso de salida del modulador es positivo. La respuesta del integrador de este pulso(impulso) es un escalón positivo abrupto como se muestra. En el tiempo t2, dt = m(t) - m^(t) es un pulso positivo fijo con el resultado de que m^(t) es nuevamente positivo. La forma de onda m^(t) continua con la misma amplitud o altura a m(t) a través del cuarto pulso, el cual el tiempo m^(t) es sobresalido. De aquí, inmediatamente después del cuarto pulso, dt es negativo, y el próximo pulso en la salida del modulador es de polaridad negativa. La modulación delta lleva su nombre precisamente a la diferencia que hay entre la forma de onda m(t) y la aproximación a m^(t). La característica principal de un sistema por modulación delta se debe a la transmisión de información sobre la señal diferencial dt.

(a) LA SEÑAL m(t) Y SU APROXIMACION $\tilde{m}(t)$.
(b) EL TREN DE PULSOS TRANSMITIDO.

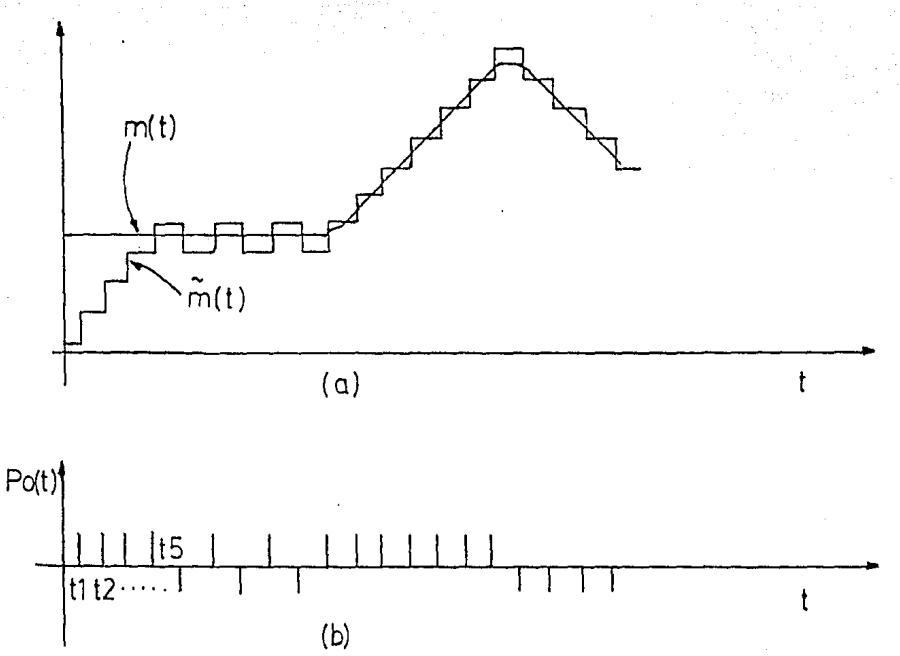


FIG. 2.3-11
FORMAS DE ONDA DEL SISTEMA DE MODULACION DELTA.

III - Análisis y Reconocimiento automático de voz.

3.1 Análisis de Voz.

En este capítulo, en la cual se describirá los métodos o técnicas más comúnes o utilizadas en el procesamiento de la señal de voz, así como los parámetros que hay que tomar en cuenta. Más adelante, nos daremos cuenta que tan complejo es el análisis de la señal de voz y que dificultad presenta ésta, como es la variación que puede tener, por ejemplo, cuando un mismo individuo no pronuncia igual la misma palabra varias veces. Vamos a encontrar que la estimación de formantes y los parámetros del tono es una de las áreas desconocidas (o que no se sabe mucho) todavía del procesamiento de la señal de voz.

Para esto encontramos que hay un trabajo alentador en la cual se menciona esto (Schroeder, 1970), titulado: 'ESTIMACION DE PARAMETROS EN EL HABLA: UNA LECCION DE HETERODOXIA ', en la cual hace enfásis que es el proceso de estimación más afortunado

> teniendo frecuentemente confianza sobre intuición basado sobre el conocimiento de la señal de voz y su producción en el aparato vocal humano antes que las aplicaciones de rutinas de métodos teóricos bien establecidos.

a) Parametrización de la señal de voz.

La representación eficiente de la información existente en la señal de voz puede realizarse actuando sobre la señal temporal o sobre la señal en el dominio transformado. Así calculando el espectro de frecuencia de una parte de la señal de voz correspondiente a una ventana de 25 ms. (espectro localizado) se puede obtener los siguientes resultados, esto lo apreciamos en la fig. 3.1-1 (a y b).

La primera de ellas corresponden a un tramo sonoro de la señal de voz, (a) de la fig. 3.1-2, y el segundo a uno sordo, tramo (b) de la misma figura. Si analizamos la envolvente de ambos espectros se observa que cada máximo corresponde a una frecuencia de resonancia(formantes) del conjunto de cavidades constituidas del aparato fonador. Observamos, en la fig. 3.1-1 que la principal diferencia entre (a) y(b) es que la estructura de rayas aparece en(a) y no en (b). Dicha estructura de rayas periódicas corresponde con la excitación, también periódicas del sonido vocalizado(fig. 3.1-1, parte a) y, que, por tanto, no aparece en los sonidos vocalizados(fig. 3.1-1 , parte b). Vemos, entonces como el análisis de la señal de voz en

el dominio de la frecuencia permite extraer información relacionada con el mecanismo de producción de la voz y, que, por lo tanto, como se vera más adelante, dependerá de la identidad de cada persona o locutor.

Los siguientes parámetros son los más utilizados dentro de los sistemas de reconocimiento de voz y estos son:

- Intensidad o energía local de la señal de voz.
  Este parámetro está relacionado con la variación más o menos monótona de la voz. Dependerá de la variación temporal de la presión subglótica y de la forma del tracto vocal y, por tanto, de la identidad del locutor. Si esta característica se varía conscientemente, entonces es útil en el alineamiento temporal entre la frase de referencia y la de prueba en la cual se vera más adelante.

- Frecuencia fundamental.
  Representa la frecuencia de oscilación de las cuerdas vocales en los sonidos vocalizados. Su valor promedio puede ser poco significativo, permitiéndose distinguir principalmente las voces de hombre de las de mujer, pero su variación a lo largo de toda la frase es una característica importante a considerar en los sistemas de reconocimiento. Además, posee la interesante propiedad de no quedar afectada por las característica de frecuencia del sistema de grabación y transmisión como sucede con la información relacionada con la envolvente espectral.

- Envolvente del espectro localizado.
  Es una representación tridimensional energía-tiempo-frecuencia de la señal de voz. Su efectividad lo encontramos en sistemas de reconocimiento basados en banco de filtros [1] por la técnica más eficiente llamada predicción lineal segun algunos investigadores.

- Coeficientes de predicción lineal.
  Por medio de cada muestra de la señal de voz se predice una suma ponderada de muestras anteriores de tiempo. Los pesos de dicha suma se denominan coeficientes de predicción y representa la envolvente espectral de cada ventana de análisis de la señal de voz [2]. El empleo de esta técnica es importante por las siguientes razones:

  1. Los coeficientes de predicción representan la envolvente espectral y con ella la información combinada de formantes, anchos de banda y forma de onda de excitación.

  2. Supone, en primera aproximación, información inde-

pendiente con la frecuencia fundamental y la energía y por último;

3. Permite desarrollos de hardware eficaces para su cálculo.

- Frecuencias y anchos de banda de formantes o resonancias de las cavidades acústicas.
  Son características eficaces en reconocimiento de voz pero poseen como principal inconveniente su dificultad de cálculo, siendo descartados en la mayoría de las aplicaciones prácticas.

- Coarticulaciones nasales.
  Se basan en el análisis de variaciones del tracto vocal durante la pronunciación de consonantes nasalizados. Su única dificultad es la determinación de la posición en la coarticulación en el contexto de la frase [3].

La representación de información de la envolvente espectral mediante los coeficientes de predicción, encontramos que existen varios conjuntos de parámetros biunívocamente relacionados con ellos. Estos son: coeficiente de respuesta impulsiva, función de autocorrelación, coeficientes de correlación parcial (PARCOR), relaciones de área y el cepstrum. En la fig. 3.1-3 se muestra la variación de algunos de estos parámetros, durante la pronunciación de la palabra " ALFREDO " por ejemplo, por dos personas diferentes.

b) Características analíticas individuales.

1. Característica prosódica.

Otra etapa de análisis la cual se incluye es la característica prosódica la cual lo podemos dividir en dos categorías básicas: en característica de calidad de voz y característica de voz dinámica. Las variaciones en la calidad de voz, el cual algunas veces lo conocemos con el nombre de 'fonemas paralingüistico' son explicados por diferencias anátomicas e idiosincracias (como la garganta inflamada) muscular, etc. Las variaciones de voz dinámica ocurren en tres dimensiones: Tono o frecuencia fundamental de voz, Tiempo y Amplitud. Dentro del primero, el patrón de variación de tono o entonación puede ser distinquido desde el rango total dentro el cual esa variación puede ocurrir. La dimensión en tiempo en compás del ritmo del habla, pausas y el ritmo completo: ya sea que es alterado lentamente o rapidamente. La tercera dimensión, que es la amplitud, es relativamente de menor importancia. La entonación y el ritmo trabajan juntos para producir comúnmente un efecto llamado 'stress'. La dificultad del stress es, desde el punto de vista acústico, una

combinación de tono, ritmo y amplitud. Aún asi, alguna característica de la calidad de voz puede ser atribuida al origen, como laringitis, aunque otras, como el paladar hendido, dentadura mal colocadas, etc., como también afecta las característica segmentales.

En el habla natural, las características prosódicas son significativamente influenciada ya sea por la pronunciación que es generada expontáneamente o por leer en voz alta. Encontramos, que las variaciones en el habla expontáneo son enormes. Todas estas emociones el cual oimos diariamente en el habla como, sarcasmos, excitaciones, rudeza, desacuerdos o discusiones, tristeza, miedo, amor, etc. Las variaciones en la calidad de voz ciertamente juega un papel importante aqui.

De todas formas, encontramos que una conversación ordinaria amistosa causal, el necesitar encontrar palabras y de alguna manera introducirla dentro de la pronunciación total nos produce una gran diversidad de estructuras prosodicas.

En la distinción de características prosódicas es particularmente útil un módelo de filtro-fuente (Cap. I), el cual son propiedades importantes de la fuente, desde un segmento, lo cual pertenece al filtro. El tono y la amplitud son propiedades originalmente primaria (fuente). El ritmo y la velocidad de locución no son, pero ninguno de ellos tienen propiedades de filtraje.

2. Estimación de Formantes.

Una vez que el espectro de frecuencia de la señal de voz se ha calculado, puede parecer simple estimar la posición de los formantes. Pero, realmente no lo es ! Una razón para esto es que solamente el análisis es realizado por sincronía del tono, el espectro de frecuencia de la fuente de excitación es mezclado con el filtro del tracto vocal. Hay muchas razones, el cual se discutirá después , el de por qué efectuar un análisis de la señal de voz involucra muchos parámetros. Pero primero, debemos considerar cómo extraer las características del filtro del tracto vocal desde el espectro combinado de fuente y del filtro. Para hacer esto, primero debemos explorar la teoria de Sistemas Lineales.

Según la teoría de Sistemas Lineales Discretos; en la fig. 3.1-4. Esto muestra una señal de entrada excitando el filtro y produce una señal de salida. Para los propósitos de análisis de voz, inmaginamos la entrada de una forma de onda "glotal",esto es el fitro del tracto vocal, y la salida una señal de voz (el cual entonces sujeto a la frecuencia alta de enfásis por radiación desde el labio). Consideramos un sistema discreto aquí, la entrada x(n) y la

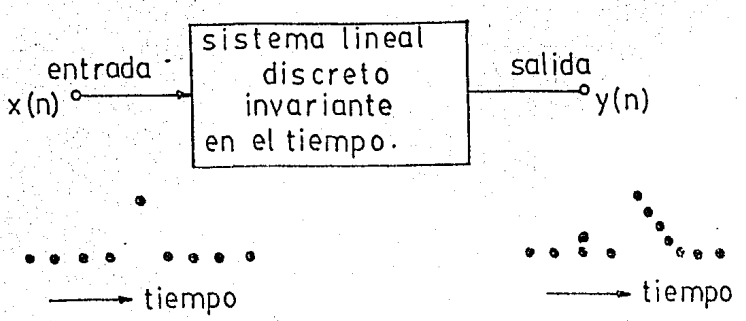


FIG. 3.1-4
SISTEMA LINEAL DISCRETO.

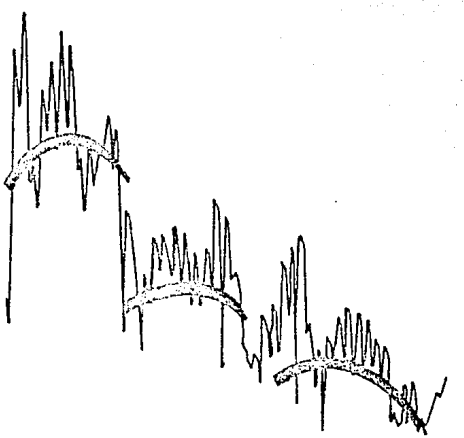


FIG. 3.1-5
ESPECTRO DE FRECUENCIA DE UNA SEÑAL TIPICA DE VOZ - SONORA.

salida y(n) y las señales muestreadas, definida sólo cuando n es integral. La teoría es completamente similar para sistemas continuos.

En la fig. 3.1-5 muestra el espectro de frecuencia de una señal típica de voz sonora. La figura completa muestra una línea curva (joroba) en la posición del formante; igual como se ha visto en el Cap. I, cuando se hablaba de formantes. Como siempre, vamos a encontrar que lo que esta sobrepuesto es una oscilación (en el dominio de la frecuencia) a la frecuencia del tono. Esto ocurre porque la transformada del filtro del tracto vocal tiende a ser multiplicado por el pulso tonal, teniendo luego componentes armónicos de la frecuencia tonal. La oscilación debe ser suprimida antes que los formantes puedan ser estimados a cualquier grado de exactitud. Una forma de eliminar la oscilación es realizar un análisis sincrónico del tono. Esto remueve la influencia de tono desde el dominio de la frecuencia por distribución (en partes) con en el dominio del tiempo. El obstáculo, es por supuesto, que eso no es fácil de estimar la frecuencia del tono: algunas técnicas son discutidas más adelante. Otra forma es usar Análisis de Predicción Lineal, lo cual realmente se deshace de la información tonal sin haber estimado el primer periodo del tono. El tercer método es remover los rizos (ondulación) del tono desde el espectro de frecuencia directamente.

3. Extracción de Tono.

Encontramos, que en muchas maneras, la extracción del tono es más importante desde el punto de vista práctico, a diferencia de la estimación de formantes. La estimación de formantes es sólo necesario si la voz es almacenada en forma de un código-formante. Para almacenamiento de predicción lineal del habla, o para sintésis desde fonéticos o textos, la estimación de formantes es innecesario; aunque por supuesto, de información general acerca de las frecuencias de formantes y las huellas o antecedentes de los formantes en el habla natural es necesario antes de la sintésis desde un sistema fonético. Encontramos, que conociendo el contorno del tono es necesario para muchos propósitos diferentes. Por ejemplo, codificación compacta de predicción lineal de voz depende sobre el tono ya estimado y almacenado como parámetros separado desde la articulación.

Otro problema la cual es ligado con la extracción del tono, es la distinción de voz sonora y no-sonora. Otro método de estimación de tono, en el cual se usa el CEPSTRUM, es el método de autocorrelación. Este método involucra una cantidad substancial de computación, y por lo tanto tiene un alto grado de complejidad. Esto se explicará más adelante. Otro método es usando predicción lineal residual, que también requiere grandes cálculos de

computación pero que también no da excelente resultados.

a) Método de autocorrelación.

La forma más confiable de estimación de tono de una señal periódica el cual es distorsionada por el ruido es examinada en corto tiempo por la función de autocorrelación. La autocorrelación de una señal x(n) con un atraso k es definido como

$$\varphi(k)\equiv\sum_{n=-\infty}^{\infty}x(n)x(n+k) \qquad \ldots.$$

Si la señal es cuasi-periódica, con un período de variación, lento, y con una distancia finita en la cual puede ser aislado con una ventana w(i), y en el cual es 0 cuando i esta afuera del rango [0,n). Iniciando esta ventana a una muestra m dado la señal de la ventana

$$x(n)w(n-m)$$

cuya autocorrelación es, la autocorrelación en corto tiempo de la señal x en el punto m, es

$$\varphi_m(k)=\sum_{n}x(n)w(n-m)x(n+k)w(n-m+k)$$

La función de autocorrelación exhibe picos atrasados lo cual corresponde a los períodos de tonos y sus múltiplos. Tal atraso, la señal es en fase con la versión de atraso de el mismo, dado una correlación alta. El tono de la voz natural tiene un rango cerca de 3 octavas, desde 50 Hz (tono bajo en el hombre) y cerca de 400 Hz (tono perteneciente al niño). Para asegurar que al menos dos ciclos de tono puedan apreciarse, aún en el bajo extremo, la ventana necesita ser menor que 40 ms de longitud, y la función de autocorrelación calculada para atraso hasta de 20 ms. Si la señal en el extremo alto del rango del tono, 400 Hz, son visualizadas hacia una ventana de autocorrelación de 40 ms, un espaciamiento considerable de résolución de tono en el dominio del tiempo es esperado. Finalmente, para la voz no-sonora, ocurrirá que habra picos no substanciales de autocorrelación.

Entonces, encontramos que la autocorrelación en corto tiempo de voz sonora exhibe picos a múltiplos de periodo del tono, claro esto no es fácilmente de detectar que cualquiera de estos picos en la función de autocorrelación que en la forma de onda de tiempo original. Tomemos un ejemplo simple, si una señal contiene la fundamental y en fase, primera y segunda armónica.

$$x(n) = a\,sen2\Pi fnT + b\,sen4\Pi fnT + c\,sen6\Pi fnT,$$

entonces, la función de autocorrelación en corto tiempo es,

$$\varphi_m(k) = \frac{a^2 \cos 2\Pi fkT + b^2 \cos 4\Pi fkT + c^2 \cos 6\Pi fkT}{2}$$

Esto no es razón a creer que la detección del periodo fundamental de la señal será de cualquier forma fácilmente en el dominio de autocorrelación que en el dominio del tiempo.

El error más común de detección de tono por análisis de autocorrelación que encontramos es la periodicidad de los formantes que son confundidos o podríamos decir equivocados con el tono. Esto típicamente conduce a repeticiones de tiempo que son como Ttono ± Tformante1, donde las T son el periodo del tono y primer formante. Afortunadamente, encontramos que son formas simples de procesamiento de señal no-linealmente que reduce el efecto de formantes sobre la estimación del tono al estar usando la autocorrelación. Una forma es el de utilizar un filtro pasabajo con la señal cortada sobre el periodo de tono máximo, por decir 600 Hz. Como vemos, el formante 1 es obtenido cerca de este valor. Una técnica diferente, el cual lo podemos usar conjuntamente con la filtración, es colocar unas grapas o clips lo que en inglés se conoce como "centre-clips", como se muestra la señal, en la fig. 3.1-e.

Esto remueve muchos de los rizos los cuales son asociados con los formantes. Para una detección de tono muy exacto, lo mejor es combinar la evidencia desde diversos métodos diferentes de análisis de la forma de onda de tiempo. La función de autocorrelación proporciona una fuente de evidencia; y el CEPSTRUM proporciona otra. Una tercera fuente viene desde la misma forma de onda en el tiempo. Un ejemplo, es el de McGonegal (1975) que describió un método semi-automático de detección de tono el cual usa un juicio humano al hacer una decisión final basado sobre estas tres fuentes de evidencia. Esto de este incorporar una exactitud altísima en el contorno del tono a costa del considerable esfuerzo humano: toma una experiencia a la persona (usuario) procesar cada segundo del habla (voz) 30 minutos [40].

b) Método de Extracción de Características.

Este método que acabas de estudiar en...

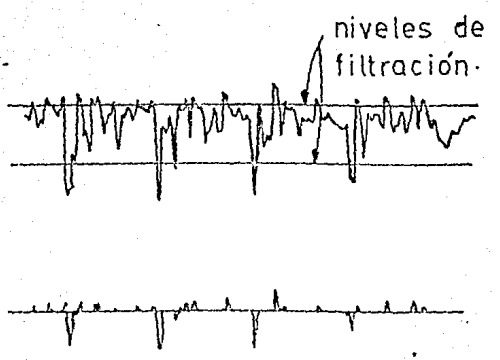


FIG. 3.1-6

FILTRACION DE LA SEÑAL DE VOZ POR MEDIO DE -- GRAPAS

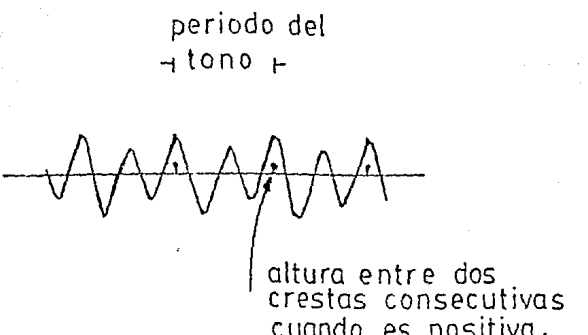


FIG. 3.1-8

FORMA DE ONDA EL CUAL NECESITA LA ALTURA ENTRE DOS CRESTAS -- CONSECUTIVAS PARA LA IDENTIFICACION CORRECTA DEL TONO.

que es el área de análisis de voz, también se menciona más adelante como una de las etapas importantes en el Proceso de reconocimiento de Patrones. Encontramos, que otras formas posibles de extracción de tono en el dominio del tiempo es intentar integrar la información desde diferentes fuentes para darnos una estimación de tono confiable. Diversas características de la señal de voz en el tiempo puede ser definido, cada cual proporciona una estimación del período del tono, y una estimación completa puede ser obtenido por voto mayoritario. Por ejemplo, supongamos sólo una característica de la señal de voz lo cual es retenido la altura y posición de los picos, donde un "pico" es definido por el criterio simplístico (simplificación extrema):

$$x(n-1) < x(n) \quad y \quad x(n) > x(n+1)$$

Teniendo establecido un pico lo cual es considerado como la representación del pulso del tono, uno puede definir un "periodo blanco" basado sobre el corriente tono estimado, sin lo cual encontramos que el próximo pulso del tono no podra ocurrir. Cuando este período ha pasado, el próximo pulso del tono es buscado. Pero, cómo lo realiza ? Primero, un criterio estricto es usado para la identificación del próximo pico como un pulso tono, pero esto podra gradualmente ser cedido si pasa el tiempo sin haber localizado un pulso apropiado. La fig 3.1-7 muestra una forma conveniente de realizar esto.

Una declinada exponencial es iniciado al final del periodo blanco y cuando un pico se acerca, es identificado como un pulso tono. Una gran ventaja de este tipo de algoritmo es que el dato es grandemente reducido por la sóla consideración del pico, el cual puede ser detectado por una simple computadora (Hardware). De este modo, puede permitir una operación de tiempo real sobre un procesador pequeño con una mínima computadora de propósito especial. Tal detector de pulso tono es excesivamente simplístico, y frecuentemente identificará el tono incorrectamente. Como siempre, puede ser usado en conjunción con otras caracteristica que nos producirá una buena estimación del tono. Por ejemplo, Gold y Rabiner [19] (1969) quienes son pioneros en esta área, utilizan 6 características, que son:

1. Altura de cada cresta.
2. Profundidad del valle.
3. Altura de un valle a la cresta siguiente.
4. Profundidad de una cresta al siguiente valle.
5. Altura entre dos crestas consecutivas cuando es es positiva (si es > que 0).
6. Profundidad de valle a valle cuando es positiva (si es > que 0).

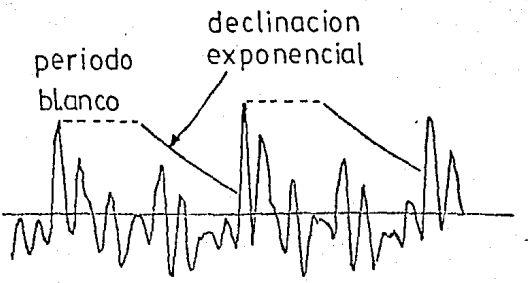


FIG. 3.1-7

DETECCION DE LOS PERIODOS --- DEL TONO DESDE PICOS USANDO UN PERIODO BLANCO Y DECLINA -- CION EXPONENCIAL.

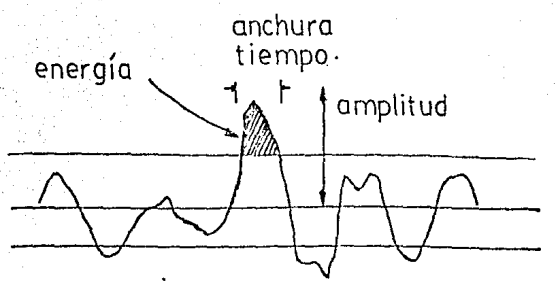


FIG. 3.1-9

TRES CARACTERISTICAS DEFINIDAS SOBRE LOS PICOS Y VALLES DE LA VOZ FILTRADA.

La características son simétricas con respecto a la cresta y al valle. La primera característica es uno de los referidos arriba, y el segundo trabaja exactamente de la misma forma. La tercera característica registra la altura entre cada valle y las crestas posteriores, y el cuarto usa la profundidad entre cada cresta y el valle posterior. El propósito de los dos detectores finales es eliminar lo secundario, pero muy grande, consideraciones desde pico o cresta.

La fig. 3.1-8 muestra el tipo de señal sobre el cual las otras características puede incorrectamente doblar el tono, pero las dos últimas características la identifica correctamente.

Gold y Rabiner también incluye los dos últimos tonos estimados desde cada detector de característica. Además, para cada característica, el presente estimado era sumado al previo para hacer el cuarto, y el previo a uno antes de hacer el quinto, y todos los tres eran sumados juntos para hacer el sexto; entonces para cada característica habia seis separados estimados de tonos. Cuál era la razón ? La razón era la siguiente, que si las tres consecutivas estimados del período fundamental son To, T1, T2; entonces si algunos picos son falsamente indentificados, el período actual podra ser cualquiera de

$$To + T1, \qquad T1 + T2, \qquad To + T1 + T2.$$

es esencial hacer esto, porque la característica de un tipo dado puede ocurrir más que una vez en un período de tono; en donde los picos secundarios usualmente existen.

Las seis características con cada contribución a seis separadores estimados, hace 36 estimaciones de tono en todo. Para la voz, los detectores de características son usualmente precedido por un filtro pasabajo a atenuar la miriada de picos por los causado formantes altos, y esto es inapropiado para aplicaciones musicales. Esto es evidencia el cual muestra que las características adicionales pueden asistir con la identificación del tono. Las características mencionadas anteriormente son todas basadas sobre la amplitud de la señal, y puede ser referidas como características secundarias desviadas de una característica primaria sencilla. Otra característica primaria puede ser fácilmente definida. Encontramos, que Tucker y Bates (1980) utilizan una forma de onda "centre-clips" y considera sólo los rizos de picos arriba de la región central. Ellos definen dos características primarias adicionales, en adición a la amplitud del pico: el ancho del pico (período por el cual se encuentra afuera del nivel de corte), y su

energía (también, afuera del nivel de corte). En la fig. 3.1-9 muestra estas características primarias. Las características secundarias son definidas, basada sobre estas tres primarias, y la estimación del tono son hechas para cada uno. Una innovacción adicional era combinar la estimación individual en una forma en la cual es basado sobre el análisis de autocorrelación, reduciendo en algun grado el propósito unico (el cual es hecho) el proceso de detección del tono.

4. Proceso de Reconocimiento de Patrones.

Las cuatros funciones que constituye el Proceso de Reconocimiento de Patrones lo cual consiste de un micrófono transductor, un preprocesador, un extractor de característica y un clasificador de nivel [9] (decisión lógica) encontraremos más adelante que son las funciones contenidas en un sistema de Reconocimiento de Voz con vocabulario limitado. En muchos intentos en Reconocimiento Automático de Voz encontramos que de una u otra forma eliminan completamente el Proceso de Extracción de característica [53] o lo utilizan como una forma simplificada de réplica módelo [16]. Indentificando y entendiendo cada una de las partes del Proceso de Reconocimiento de Patrones (fig. 3.1-10), veremos por qué son importantes todas.

a) Preprocesamiento.

Como sabemos, la señal de voz no es aperiódica ni tampoco periódica, pero debe de ser considerada como una señal cuasi-periódica lo cual las técnicas análiticas que son desarrolladas deben reflejar las características temporales de significancia como bien como características espectrales. Manteniendo este punto de vista dual durante todo el análisis, esto requerirá de una modificación como algunos la llaman, la clásica técnicas análiticas en el dominio del tiempo y en el dominio de la frecuencia. A retener ambas de estas características en un análisis de frecuencia, un método el cual produce un espectro de corta duración encontraremos que es esencial.

La representación en el dominio de la frecuencia de la señal de voz es particularmente ventajosa desde que 1) Es conocido que el sistema auditorio humano lleva a cabo un análisis de frecuencia crudo o no preparada en la periferia de sensación del auditorio y, 2) Porque tiende a ser mostrado por análisis acústico del sistema de vocalización, que una descripción exacta del sonido de voz lo cual puede ser obtenido con un concepto de frecuencia natural de un módelo de producción de voz [4]. Una función periódica de tiempo en poder del espectro de energía con cantidades finitas de energía localizado en puntos discreto en el

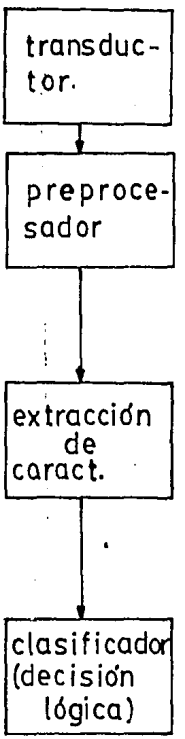


FIG. 3.1-10

PROCESO DE RECONOCIMIENTO DE PATRONES.

espectro, lo describimos conmúnmente como una linea o raya espectral. Una función aperiódica es la que contiene energía finita y es calculada en base a la Transformada de Fourier en un espectro de densidad de energía, esto es, una función continua de frecuencia. Para el análisis de señales de voz, es deseable obtener la distribución de energía espectral y sus variaciones como su función en el tiempo. Debemos mantener una resolución suficiente en ambos, en el dominio del tiempo y en el de la frecuencia para que toda las propiedades de la información que se lleve a cabo en ambos dominios puede ser detectada.

Como se ha dicho antes, el análisis de voz puede ser logrado de una u otra forma por análisis discreto del espectro analógico hacia el uso de la Transformada rápida de Fourier(FFT) o por técnicas de codificación de Predicción Linear y CEPSTRUM. En todos estos métodos, encontramos que ocurren problemas semejantes. El FFT produce un espectro discreto el cual, en el promedio de muestreo suficientemente alto, se aproxima a la Transformada de Fourier continua. Muchos diferentes tipos de ventana de datos tienden a ser utilizados en FFT. Por lo tanto, la selección de la ventana es similar a la selección de la respuesta del filtro en el analizador de espectro analógico. La separación de coeficiente de varios términos en el cálculo de la FFT o la contribución de los filtros individuales en el analizador analógico es un punto importante en el análisis del espectro que se hace pero se considera posterior a la primera etapa en el procesamiento de la señal de voz. Pero sabemos, que se requiere hacer consideraciones adicionales para lograr la detección y reconocimiento de elementos de información llevado a cabo(significa características) de la señal de voz en la cual estos elementos tienen a ser transformados en el preprocesador.

Como siempre, un análisis incorrecto en el procedimiento o subrutina en el preprocesador puede destruir características significantes anterior a cualquier procesamiento para extraer estas características. Recientemente, se ha hecho enfásis en el análisis de voz sobre el dominio del tiempo basado sobre la predictibilidad linear de las formas de onda de la voz. La codificación predictiva linear (LPC) tiene a ser ampliamente usada en la codificación eficiente para sistemas de comunicaciones de voz digital.

b) Extracción de características.

Considerado como formalismo matemático tiende a ser desarrollado para varios procesos de reconocimiento Automático de Voz. Como siempre, no existe teoría general el cual puede preseleccionar la porción de información producida por la señal de voz. Por tanto, el

diseño del extractor de característica es heurístico y debe usar "ad hoc" (cuyo significado es 'para este propósito especial sólo' ó 'con respecto a esto') o estrategias de asistencia por computadora. Sólo datos experimentales actuales pueden proveer el valor de un conjunto de características en particular. Es este el dilema en particular la cual tiene a ser resuelta en el reciente enfásis de incremento dado en investigaciones de extracción de característica para sistemas de Reconocimiento de Patrones. Encontramos, que la función de procesamiento clave en un sistema de reconocimiento de voz es el extractor de característica. Aunque, hay muchas técnicas de clasificación aceptable el cual pueden operar sobre un conjunto decaracterísticas(mediciones), entonces el esquema de no-clasificación puede compensarse para un conjunto de características inadecuadas [5]. El conjunto de característica más óptimo, generalmente es necesario que el clasificador sea menos complejo para que nos de una exactitud dada. Las varias características acústicas usadas en los sistemas de Reconocimiento de voz a ser referidos tienden a ser evaluadas extensamente con grandes locutores populares, en ruido como bien como en un cuarto silencioso, y para muchas horas de operaciones en línea con locutores no-entrenados [6] [7]. Las características son útiles para aplicaciones de voz continua [8] como los Sistemas de reconocimiento de palabras aisladas [9] [10]. El conjunto de características seleccionadas usada es suficientemente general para hacer posible adicionar nuevas palabras arbitrarias en cualquier tiempo al sistema. La selección juiciosa y extracción segura de estas características de voz crítica para virtualmente cualquier medio ambiente operacional es esencial en un práctico sistema de entrada de voz. En la fig. 3.1-11 se muestra un diagrama de bloques de un sistema de entrada de voz, en la cual podemos ver las etapas que forman el Proceso de Reconocimiento de Patrones.

Con referencia a la figura, las dos primeras funciones de procesamiento son realizadas principalmente por un reprocesador alambrado duro(hard wired) y un extractor de característica. Esto hace posible lograr un prcesamiento en tiempo-real desde que la función de clasificación es ejecutado en una minicomputadora de propósito general.

Ambos, la forma o contorno espectral y el tiempo derivado de la función envolvente espectral son derivadas en el proceso de extracción de característica. La forma espectral y su cambio con el tiempo son continuamente medido sobre el rango de frecuencia de interés. Combinaciones y secuencias de estas mediciones son procesadas para producir un conjunto de 32 características acústicas significantes, una de la cual es la estimación inicial del contorno de la palabra. Un contorno de la palabra más refinado es obtenido en la minicomputadora

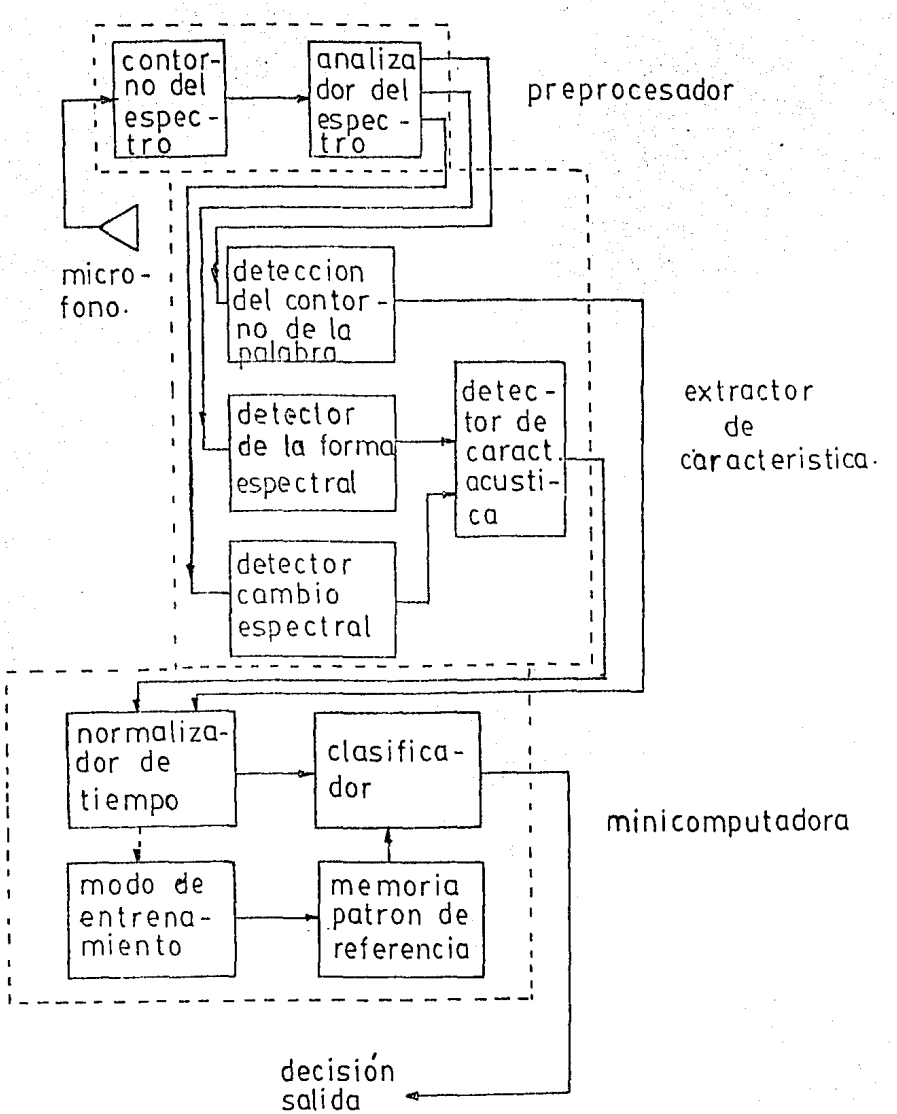


FIG. 3.1-11

DIAGRAMA DE BLOQUES DE UN SISTEMA DE ENTRADA DE VOZ.

usando la técnica de respaldo (back-up). Estas 32 características acústicas son funcionalmente similares a las descritas en los trabajos de T. B. Martin [7]. Una jerarquía de características son medidas juntas con detección lógica del contorno de la palabra. Este proceso de segmentación es ejecutado principalmente en hardware, aunque la detección final del contorno es optimizado por medio del procesamiento del software en la minicomputadora.

c) Clasificador.

La clasificación o proceso de decisión es ejecutado en software usando una minicomputadora. La minicomputadora ejecuta la multiplicidad de funciones mostradas en el diagrama de bloques o dibujo.

Para una palabra hablada, las 32 características codificadas y su tiempo de ocurrencia son almacenada en una memoria de corto-tiempo (Datos almacenados en memoria de núcleos magnéticos). Cuando el fin de la pronunciación es detectada, la duración de la palabra es dividida dentro de 16 segmentos de tiempo y las características son reconstruidas dentro de una base de tiempo normalizado. La lógica réplica-patrón subsecuentemente compara estas características de patrones ocurrentes a los patrones de referencia almacenadas para las varias palabras de vocabulario y determina el "más apto o adecuado" para una decisión de la palabra. Un total de 512 bits de información (32 características mapeadas dentro de 16 segmentos de tiempo) son requeridas para almacenar el mapa característico de cada pronunciación o expresión o patrón de referencia.

Encontramos, que desde que los sistemas de voz son adaptivos, estos deben de ser "entrenados" para locutores individuales y/o palabras. El sistema puede ser automáticamente adaptado o afinado a la característica de voz de diferentes usuarios en un simple periodo de tiempo muy corto por locución de un número pequeño de muestras de entrenamiento dentro del dispositivo a proporcionar un conjunto de característica de referencia. El sistema almacena en memoria una referencia individual del conjunto de características de la palabra para cada palabra en el vocabulario y para cada locutor en el sistema. Una vez, teniendo entrenado el sistema, nuevas palabras habladas por el usuario dentro de la operación normal del dispositivo son comparados con la referencias almacenadas y el más adecuado finalmente es seleccionado como la palabra reconocida.

Es también, posible obtener una "no-decisión" o rechazo, cuando ninguna de las características de la palabra en la memoria de referencia son cerrada o limitada a la palabra hablada. La técnica de decisión empleada puede ser

referida más simple por una corta revisión de la operación del sistema en modos de entrenamiento y reconocimiento. Estos dos modos consiste en lo siguiente: en el primero, durante este modo, el sistema automáticamente extrae una matriz característica de tiempo-normalizado para cada repetición de la palabra dada. Una matriz consistente de ocurrencias características (entre repeticiones) es requerida antes que las características sean almacenadas en las referencias de memoria o patrón. Un factor de plantilla o patrón límite es seleccionado tal que la ocurrencia de la característica (en un segmento de tiempo dado) es considerado sólo válido cuando ocurre un número mínimo de tiempo/relativo al número de muestras entrenadas. Usualmente, este factor límite es colocado entre 30 a 50 % de las ocurrencias características sin las muestras entrenadas. En la fig. 3.1-12, se muestra una matriz de características de referencia para la palabra inglesa "erase".

Mientras tanto, en el segundo modo, que es el de reconocimiento, la cual no es nada más que el modo operacional, cada nueva palabra hablada dentro del sistema es procesada de una manera analógica al procedimiento de entrenamiento. Por ejemplo: la extracción de característica, la digitalización, y el tiempo normalizado. La prueba resultante de la matriz de la palabra, entonces es comparado digitalmente con cada matriz de referencia almacenado. Similarmente e indiferentemente en cada matriz comparado son apropiadamente pesado y el resultado neto provee un producto de correlación pesada. El producto de correlación también son generado después del desplazamiento de la matriz de la palabra de entrada en un segmento de tiempo ±1. La palabra de referencia almacenada produce una alta correlación total o completa la cual es seleccionada como la palabra a prueba, proporcionada si excede un mínimo valor de correlación límite.

5. Codificacion de predicción lineal de voz (LPC).

La predicción lineal es relativamente un nuevo método de análisis-síntesis de voz, introducido en los inicios de 1970's y desde entonces ha sido utilizado extensivamente; lo cual primariamente es un método de codificación en el dominio del tiempo pero puede ser usado para darnos los parámetros en el dominio de la frecuencia, como la frecuencia de los formantes, anchos de banda y amplitud. Principalmente, en esta parte se va a exponer la parte de predicción lineal que corresponde al análisis de voz, la cual es lo que más nos interesa.

La predicción lineal lo podemos utilizar para separar las propiedades de la fuente de excitación del tono y amplitud desde el filtro del tracto vocal el cual gobierna la articulación de fonemas, o en otras palabras, separar

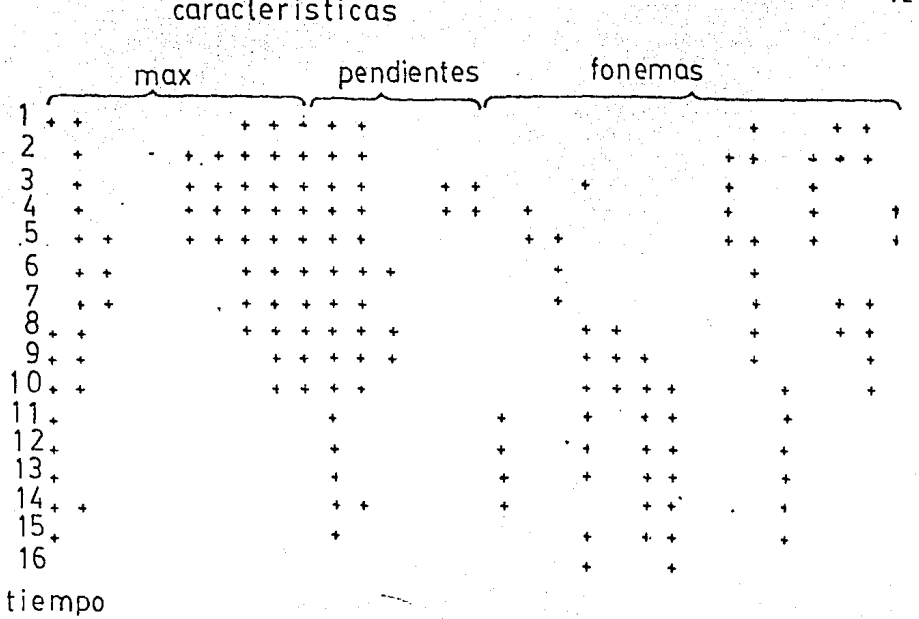


FIG. 3.1-12

MATRIZ DE REFERENCIA PARA LA PALABRA "E R A S E"

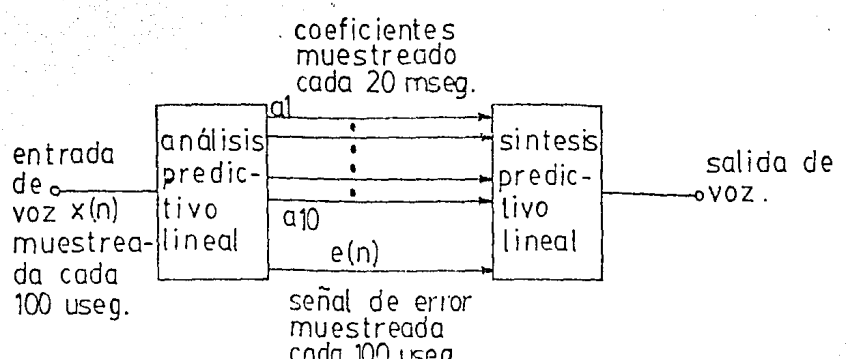


FIG. 3.1-13

PREDICCION LINEAL USANDO PCM DIFERENCIAL ADAPTIVO DE ALTO ORDEN.

mucho de lo prosódico desde la información segmental.

Sabemos, que la técnica de análisis en el dominio de la frecuencia de la Transformada discreta de Fourier [43] [42] necesariamente involucra una aproximación, por qué ? Por la razón de que éste sólo aplica las formas de ondas periódicas, y por lo tanto la operación artificial de ventanado es requerido para suprimir la aperiocidad del habla real. En constraste, la técnica LPC, siendo un método en el dominio del tiempo puede en cierta forma distribuir o repartir más racionalmente las señales aperiódicas.

La idea básica del LPC es exactamente la misma como una forma de modulación por pulsos codificados diferencial adaptivo, ver Cap. II, parte 2. Notamos que una muestra de voz x(n) puede ser predecido considerablemente de cerca por la muestra previa x(n-1). Encontramos, entonces que la predicción puede ser mejor por la multiplicación de la muestra previa por un número, digamos a1, el cual es adaptado sobre una escala de tiempo silábica. esto puede ser utilizado para la codificación de la voz por transmisión sólo del error de predicción :

$$e(n) = x(n) - a1x(n-1)$$

y usado (y el valor de a1) para reconstituir la señal x(n) al receptor (recibidor). Multiplicadores diferentes para cada uno sera necesario, en la cual podremos escribir el error de predicción como:

$$e(n) = x(n) - a1x(n-1) - a2x(n-2) - \ldots - apx(n-p)$$
$$= x(n) - \sum_{k=1}^{p} akx(n-k)$$

Los multiplicadores ak serán adaptados para minimizar la señal de error.

Todo esto que se ha estado exponiendo lo podemos ver en la fig. 3.1-13, la cual muestra la configuración para la modulación por pulsos codificados diferencial adaptivo de alto orden. Más adelante explicaremos solo la parte que nos interesa, lo cual es el análisis de predicción lineal. incluyendo los dos métodos a utilizar: 1. El método de autocorrelación (anteriormente se explicó referido a la extracción de tono) y 2. El método de covarianza.

Los 10 coeficientes utilizados en la predicción lineal, podemos observarlo en la fig. 3.1-14, lo cual es referido a una muestra o ejemplo de la señal de voz sobre una distancia de un segundo de voz (señal).

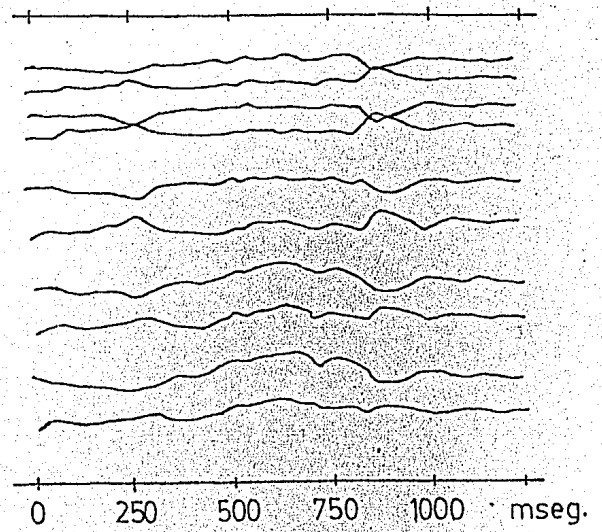


FIG. 3.1-14
COEFICIENTES PREDICTIVO LINEAL
PARA UNA MUESTRA DE LA VOZ.

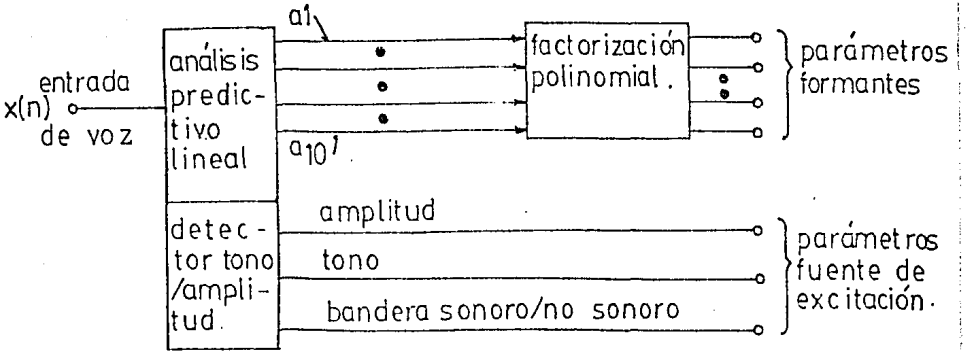


FIG. 3.1-15
PREDICCION LINEAL USADO PARA
ANALISIS DE FORMANTES.

Ya en si en la parte de análisis, utilizando la predicción lineal , encontramos que es usado para el análisis de formantes, de sus posiciones y anchos de bandas, en vez de la transmisión de los coeficientes ak; visto en la fig. 3.1-15. De acuerdo a la fig. 3.1-15, el bloque de factorización polinomial o selección de cresta podemos verlo como un polinomio

$$1 - a_1 z^{-1} - a_2 z^{-2} - \cdots - a_p z^{-p},$$

la cual cuando factorizamos dentro de los términos de un producto de segundo orden, nos da la característica de los formantes (como bien como el término de compensación espectral). La factorización es equivalente para encontrar la raíz compleja del polinomio. Consecuentemente, algoritmos de selección de cresta (peak-picking) son algunas veces utilizados. El valor absoluto del polinomio nos da el espectro de frecuencia del filtro del tracto vocal, y los formantes que aparecen como picos.

a) Método de Autocorrelación.

En este método encontramos que los coeficientes de la matriz tiene la forma

$$\sum_{n} x(n-j)x(n-k)$$

si es una suma doblemente infinita, con x(n) para ser definido como cero donde siempre n < 0, entonces, podemos hacer uso en realidad de

$$\sum_{n=-\infty}^{\infty} x(n-j)x(n-k) = \sum_{n=-\infty}^{\infty} x(n-j+1)x(n-k+1)$$
$$= \cdots = \sum_{n=-\infty}^{\infty} x(n)x(n+j-k)$$

lo cual simplifica la matriz. Esto coloca la autocorrelación en una secuencia dependiente sólo sobre el atraso a lo cual es calculado, y no sobre un tiempo absoluto.

Definiendo R(m) como la autocorrelación a un atraso m, como

$$R(m) = \sum_{n} x(n)x(n-m)$$

la matriz es convertida:

```
CONST
   N=56;
TYPE
   svec=ARRAY [0..N-1] OF REAL;
   cvec=ARRAY [1..p] OF REAL;
PROCEDURE autocorrelacion(señal:vec; ventana:svec;
                          VAR coef:cvec);

 {calcula los coeficientes de predicción lineal por el
  metodo de autocorrelacion en coef [1..p]}

  VAR
    R, temp: ARRAY [0..p] OF REAL;
    n: [0..N-1];
    i,j : [0..p];
    E: REAL;
   BEGIN  {ventanando la señal}
      FOR n:=0 TO N-1 DO
        señal:=señal[n]*ventana[n];

       {calculando el vector de autocorrelación}

      FOR i:=0 TO p DO
        BEGIN
          R[i]:=0;
          FOR n:=0 TO N-i DO
             R[i]:=R[i] + señal[n]*señal[n+i]
        END;

        {solucionando la ecuación de matrices por el metodo
         de Durbin-Levinson}

      E:=R[0];
      coef[1]:=R[1]/E;
      FOR i:=2 TO p DO
        BEGIN
          E:=(1-coef[i-1]*coef[i-1])*E;
          coef[i]:=R[i];
          FOR j:=1 TO i-1 DO
             coef[i]:=coef[i] - R[i-j]*coef[j];
          coef[i]:=coef[i]/E;
          FOR j:=1 TO i-1 DO
             temp[j]:=coef[j] - coef[i]*coef[i-j]
          FOR j:=1 TO i-1 DO
             coef[j]:=temp[j]
        END
   END;
```

FIG. 3.1-17

ALGORITMO EN PASCAL USADO PARA EL METODO DE AUTOCORRELACION.

$$
\begin{aligned}
R(0)a1 + R(1)a2 + R(2)a3 + \ldots &= R(1)\\
R(1)a1 + R(0)a2 + R(1)a3 + \ldots &= R(2)\\
R(2)a1 + R(1)a2 + R(0)a3 + \ldots &= R(3)\\
\vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad &\vdots
\end{aligned}
$$

Por supuesto, un rango infinito de sumas no puede ser usado en la práctica [40].

Para una cosa, el espectro de potencia es cambiado y sólo los datos desde un marco de tiempo corto deberá ser usado para un estimado realista de los coeficientes predictivo lineal óptimo.

He aquí un procedimiento de ventanado,

$$x(n)^{*} = Wnx(n)$$

lo cual es usado para reducir la señal a cero fuera de un rango infinito. Esto es conocido como el método de autocorrelación de parámetros predictivos calculados. Típicamente una ventana de 100 a 250 muestras es usado para el análisis de un marco de voz [40]. (usualmente el tamaño del marco seleccionado o escogido es en una región de 10 a 25 ms, podriamos decir que en un intervalo del marco es un período de tiempo)

En la fig. 3.1-16 podemos apreciar el diagrama de bloques del algoritmo para el método de autocorrelación. Este algoritmo se explica así: El cálculo de los coeficientes de la matriz R(m) es directo desde las muestras de la voz y los coeficientes de la ventana. El método de Durbin-Levinson de solución de ecuaciones de matrices opera directamente sobre el vector-R para producir los vectores coeficientes ak. El procedimiento de la fig. 3.1-17 muestra el algoritmo codificado en el lenguaje Pascal.

b) El método de covarianza.

Una de las ventajas del método predictivo lineal era que prematuramente prometía permitirnos escapar desde el problema de enventanado. Esto es, que debemos abandonar los requerimientos de los coeficientes de la matriz teniendo las propiedades de simetría de autocorrelación. En cambio, supongamos que usamos el rango de n-suma a un número fijado de elementos, decimos N, empleando en n=h, a estimar los coeficientes de predicción entre el número h de muestra y el número de muestra h+n.

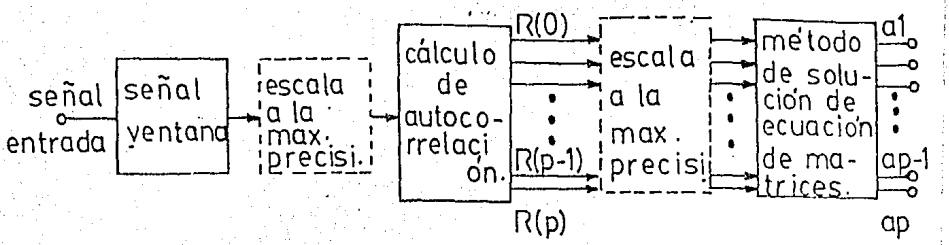


FIG. 3.1-16
ALGORITMO PARA EL METODO DE AUTOCORRELACION.

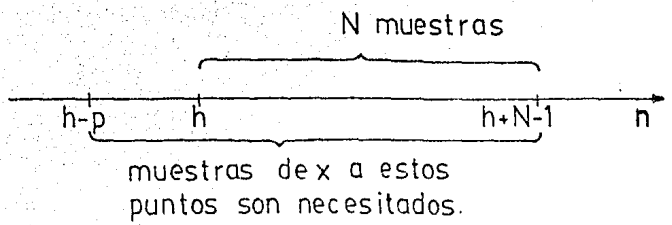


FIG. 3.1-18
PUNTOS EN EL CUAL LAS MUESTRAS DE ENTRADAS SON NECESARIAS PARA EL METODO DE COVARIANZA.

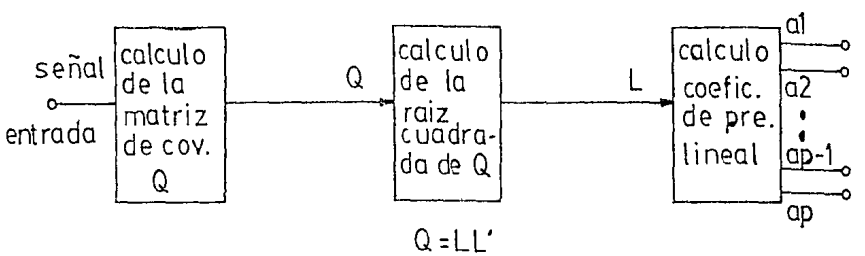


FIG. 3.1-19
ALGORITMO PARA EL METODO DE COVARIANZA.

```
CONST
   N=100;
   p=15;
TYPE
   vec=ARRAY [-p..N-1] OF REAL;
   cvec=ARRAY [1..p] OF REAL;
PROCEDURE covarianza(señal:vec; VAR coef:cvec);

 {calcula los coeficientes de predicción lineal por el
  metodo de covarianza en coef [1..p]}

  VAR
    Q: ARRAY [0..p,0..p] OF REAL;
    n: [0..N-1];
    i,j,r : [0..p];
    X: REAL;
   BEGIN

     {calculando la matriz de covarianza triangular
      superior en Q}

     FOR i:=0 TO p DO
       FOR j:=1 TO p DO
         BEGIN
           Q[i,j]:=0;
           FOR n:=0 to N-1 DO
              Q[i,j]:=Q[i,j] + señal[n-i]*señal[n-j]
         END;

     FOR r:=2 TO p DO   {calculando la raiz cuadrada de Q}
       BEGIN
         FOR i:=2 TO r-1 DO
           FOR j:=1 TO r-1 DO
             Q[i,r]:=Q[i,r] - Q[j,i]*Q[j,r];
           FOR j:=1 TO r-1 DO
             BEGIN
               X:=Q[j,r];
               Q[j,r]:=Q[j,r]/Q[j,i];
               Q[r,r]:=Q[r,r] - Q[j,r]*X
             END
       END;

     FOR r:=2 TO p DO    {calculando coeficientes [1..p]}
        FOR i:=1 TO r-1 DO
          Q[0,r]:=Q[0,r] - Q[i,r]*Q[0,i];
     FOR r:=1 TO p DO
        Q[0,r]:=Q[0,r]/Q[r,r];
     FOR r:=p-1 DOWNTO 1 DO
        FOR i:=r+1 TO p DO
          Q[0,r]:=Q[0,r] - Q[r,i]*Q[0,i];
     FOR r:=1 TO p DO
        coef[r]:=Q[0,r]
   END;
```

FIG. 3.1-20

ALGORITMO EN PASCAL USADO PARA EL METODO DE COVARIANZA.

Esto conduce a la matriz

$$\sum_{k=}^{p} a_k \sum_{n=h}^{h+N-1} x(n-j)x(n-k) = \sum_{n=h}^{h+N-1} x(n)x(n-j) \qquad j=1,2,\ldots,p.$$

alternativamente, nosotros podemos escribir

$$\sum_{k=1}^{p} a_k Q^h_{jk} = Q_{0j} \qquad j=1,2,\ldots,p;$$

donde

$$Q^h_{jk} = \sum_{n=h}^{h+N-1} x(n-j)x(n-k)$$

Notamos que algunos valores de x(n) fuera del rango h <= n < h+n son requeridos, como mostrado en la fig. 3.1-18.

Ahora $Q^h_{jk} = Q^h_{kj}$ ....., la ecuacion tiene una matriz diagonalmente simétrica: y en realidad la matriz Q^h puede ser mostrado a ser semidefinida positiva, y es al menos siempre definida positiva en la practica [40]. Acordemos con el resultado llamado teorema de Cholesky's, una matriz simétrica definida positiva Q puede ser factorizado dentro de la forma Q=LLT, donde L es una matriz triangular baja. Esto conduce a una solución eficiente del algoritmo.

Este método de coeficientes de predicción calculados tiene a ser conocido como el método de covarianza. No use enventanado de la señal de voz, y puede dar una exactitud estimada de los coeficientes de predicción con un marco de análisis más pequeño que el método de autocorrelación. Típicamente, 50 a 100 muestras de voz, tal vez sea necesario que se utilice para estimar los coeficientes, y ellos son recalculados cada 100 a 250 muestras.

La fig. 3.1-19, muestra el algoritm. para el método de covarianza, en la cual el primer bloque es el método para el cálculo de la matriz de covarianza por una estimación de

$$Q^h_{01}, Q^h_{02}, \ldots, Q^h_{0p}, Q^h_{11}, \ldots,$$

Sin embargo, la repeticiones de las multiplicaciones p veces, no es en realidad un procedimiento eficiente. El algoritmo codificado se muestra en la fig. 3.1-20, tambien en el lenguaje Pascal.

3.2 Reconocimiento Automático de Voz.

a) Sistemas de reconocimiento de habla/voz continua.

El reconocimiento de voz es formulado como un problema de máxima decodificación probable. Esta formulación requiere módelos estadísticos del proceso de producción del habla.

Productos el cual reconocen continuamente vocabularios limitados del lenguaje son ya encontrados en el mercado pero el objetivo del reconocimiento del habla continua sin restricción es todavia hoy distante desde que tiende a ser realizada.

Todas las investigaciones en la actualidad es llevado a cabo relativo al dominio del trabajo o del esfuerzo, el cual grandemente restringe las oraciones que pueden ser completadas. Estos trabajos son de dos clases: Aquellos donde las oraciones no son permitidas son describidas como **A priori** por una gramática designada por el investigador (referido a como trabajo artificial), y otros relacionados al área limitada de pronunciación natural el cual el investigador prueba el módelo desde un dato observado (referido como trabajo natural). Ejemplos de trabajos naturales son los textos de cartas comerciales, aplicaciones de patentes, reseña de libros, etc. En la fig. 3.2-1, observamos el sistema básico de reconocimiento del habla continua, la cual consiste de un procesador acústico seguido por un decodificador linqüistico.

Tradicionalmente el procesador acústico es desigando para funcionar como un fonetizador, transcribiendo la forma de onda de voz dentro de una cadena de símbolos fonéticos, mientras que el decodificador linqüistico traduce la cadena fonética posiblemente distorsionada dentro de una cadena de palabras. En muchos trabajos recientes, el procesador acústico no realiza una transcripción fonética, pero más bien produce una cadena de marcas o identificaciones en la cual caracteriza la forma de onda de la voz localizada sobre un intervalo corto de tiempo. Todo esto que se ha estado hablando se entenderá al conocer en si la función real de cada uno: Un procesador acústico(AP) funciona como un compresor de datos de la forma o señal de voz. La salida del AP es la siguiente:

1. Conservar la información importante a reconocer y,
2. Ser responsable a la caracterización estadistica.

Si la salida del AP puede ser fácilmente interpretada por la gente, es posible juzgar la extensión a la cual el AP cumple con los requerimientos del punto (1). Tipicamente, un AP es un procesador de señal, el cual transforma la señal de voz dentro de una cadena de parámetros vectoriales, seguidos por un clasificador de

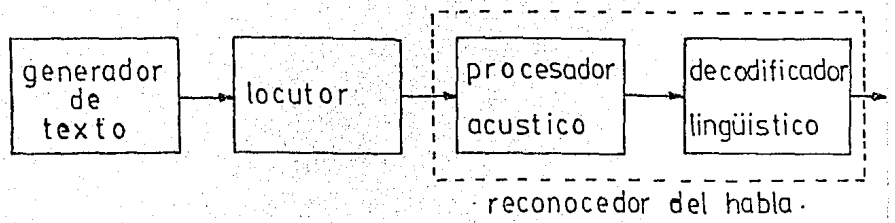


FIG. 3.2-1
SISTEMA DE RECONOCIMIENTO DE HABLA CONTINUA.

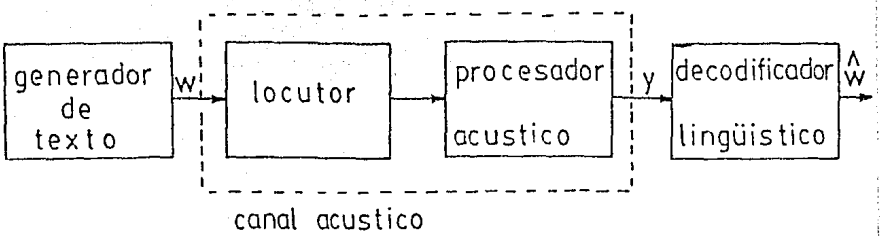


FIG. 3.2-2
VISTA DE LA TEORIA DE COMUNICACION HACIA EL RECONOCIMIENTO DEL HABLA.

patrones, el cual transforma la cadena de parámetros vectoriales dentro de una cadena de marcas o identificadores desde un alfabeto finito. Si el clasificador de patrones esta ausente, entonces el AP produce una cadena sin las marcas de los parámetros vectoriales. En una segmentación del AP, la forma de onda de la voz es segmentada dentro de eventos fonéticos precisos o claros y cada una de estas porciones de variación de longitud es entonces marcado.

Un AP de tiempo síncrono produce vectores parámetros calculados desde sucesivos intervalos de longitud fija de la forma de onda de la voz. La distancia desde los parámetros vectoriales de cada conjunto finito de vectores de parametros estandar, o prototipos, es calculado. La marca para los vectores parametros es el nombre del prototipo el cual es lo mas cercano.

En muchos AP recientes, los prototipos son obtenidos en forma automática desde los datos de la voz sin marcas. Un ejemplo típico de un AP síncrono es el " IBM Centisecond Acoustic Processor (CSAP) ". El parámetro acústico usado por el CSAP son las energías de cada banda de frecuencia de 80 en etapas de 100 Hz cubriendo el rango desde 0 hasta 8000 Hz. Estos son calculados una vez cada cien segundos (centisegundo, cs) usando una ventana de 2 cs. El clasificador de patrones tiene 45 prototipos correspondiendo más o menos a los fonemas del idioma inglés. Cada prototipo para un locutor dado es obtenido desde muestras diferentes de su voz el cual tiende a ser cuidadosamente marcado por un fonetizador. El AP produce una salida de cadena y, desde esta cadena el decodificador lingüistico(LD) hace una estimación $\hat{w}$ de la cadena de palabra w producida por un generador de texto (fig. 3.2-2). Para minimizar la probabilidad de error, $\hat{w}$ debe ser seleccionado de forma que

$$P(\hat{w}/Y) = \max_{w} p(w/Y)$$

Por la regla de Bayes, obtenemos:

$$P(w/Y) = \frac{P(w)P(Y/w)}{P(Y)}$$

donde P(Y) no depende sobre w, maximizando P(w/Y), encontramos, que es equivalente a la minimización de la probabilidad P(w,Y) = P(w)P(Y/w). Aqui P(w) es la probabilidad a priori en la cual la secuencia de palabra w será producido por un generador de texto, el cual referimos

como un módelo del lenguaje. Para muchos trabajos artificiales, el problema de modelización del lenguaje se considera algo simple.

A menudo el lenguaje es especificado por un estado finito limitado o por una gramática de contexto-libre (contex-free grammar) el cual las probabilidades pueden ser fácilmente resueltas. Para trabajos naturales, la estimación de P(w) es mucho más difícil. Los lingüistico no han progresado en este punto que puede proveer una gramática para un subconjunto medible del inglés natural, el cual es útil para el reconocimiento de voz. Además, el interés en la linqüistica tiende en especificar las oraciones del lenguaje, pero no sus probabilidades.

Para estimar P(Y/w), el otro componente de la probabilidad, el LD requiere de un módelo probabilístico del canal acústico, el cual debe contar para los locutores fonológicos y variaciones acústicas-fonéticas y para la representación del AP. Unos módelos son disponibles para el cálculo de P(w) y P(Y/w), es en principio posible para el LD calcular la probabilidad de cada oración en el lenguaje dado y lo determina lo más directamente posible w^ .

1. Sistemas automáticos de reconocimiento de voz con vocabulario limitado.

Los sistemas automático de reconocimiento de voz (SARV) tienen como objeto identificar la palabra, frase o conjunto de frases que un determinado locutor haya articulado. En el área de reconocimiento de voz existe una gran diversidad de opciones que en cada caso delimitan la característica del problema. Podemos definir los siguientes factores que afectan a la concretización de cada situación particular [11]:

- Tipo de entrada: palabras aisladas, es decir con pausa entre ellas o voz conectada, sin ningún tipo de pausa.
- Tamaño de la población: un sólo locutor, varios locutores, población ilimitada.
- Tipos de locutores: Hombre, mujer, niños.
- Características ambientales: Ruido, resonancias, etc.
- Medios de transmisión: Teléfono, micrófono con supresores de ruido, etc.
- Tamaño del vocabulario: 1-20, 20-100, >100 palabras diferentes.

- Formato de la información de entrada: Texto ajustado a reglas, texto libre, etc.

- Otros.

Como podemos deducir de la lista anterior existe una gran variedad de opciones y alternativas disponibles en la especificación de un SARV. La principal característica que se usa para definir la complejidad de un SARV se refiere al tipo de entrada. En palabras aisladas, que se explicará más adelante detalladamente, la mínima duración de la pausa que separa una palabra de otra es del orden de 200 ms. Cualquier pausa más corta de 100 ms. puede confundirse con la pausa que se produce entre sílabas en la articulación de determinadas palabras, por ejemplo: cuatro. En voz conectada, la dificultad radica en determinar donde termina una palabra y empieza la siguiente, así como en la variabilidad que se produce en la características acústicas de cada palabra dependiendo del contexto (palabras anteriores y posteriores). La segunda característica que afecta a la complejidad del sistema está relacionado con el tamaño del vocabulario que se pretende reconocer. A medida que el tamaño del vocabulario aumenta, se complica el algoritmo de reconocimiento y se debe recurrir a procedimientos más complejos que permitan diferenciar las características de cada palabra.

Un aspecto importante que debemos tener en cuenta en los SARV es el alto grado de seguridad que se les debe exigir, de tal forma que no exista disminución en la confianza o eficiencia del operador. Uno de los mayores obstaculos en el progreso de los SARV es la variación de voz para los diferentes individuos de una población que vayan a utilizar el sistema. Este hecho plantea la necesidad de utilizar métodos que permitan extraer aquellas características que resulten comunes a la población, o mejor aún, utilizar métodos que permitan la adaptación del sistema al usuario. Otra dificultad adicional existe por el hecho de que, en un mismo individuo se pueden modificar las características de su voz según sea su salud, estado de ánimo, medio ambiente, etc. Cualquier SARV con vocabulario limitado debe seguir el siguiente esquema general, mostrado en la fig. 3.2-3, como vemos, son las mismas etapas que involucra el Proceso de Reconocimiento de Patrones. La palabras o frases a reconocer son introducidas previamente en el sistema, parametrizadas, ordenadas y almacenadas en memoria. Cuando un locutor articula una palabra, que forma parte del vocabulario almacenado, es tratada por el procesador de señal que extrae el mismo tipo de parámetros que en la muestra de referencia y los compara con los correspondiente a cada una de las palabras que forman el vocabulario. El sistema debe identificar como palabra articulada, aquella cuyos parámetros sean más coincidentes con los de entrada.

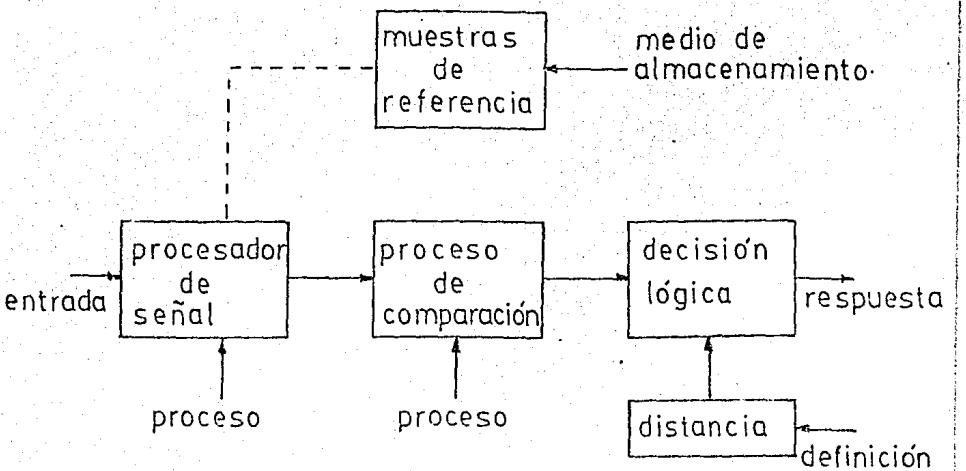


FIG. 3.2-3
ESQUEMA GENERAL DE UN SARV DE VOCABULARIO LIMITADO.

Se requiere pues un procesador de señal que extraiga un conjunto de parámetros de la señal de entrada y los ordene según un determinado vector, que representa la unidad que se pretende reconocer. Se necesita un proceso de comparación válido que permita a través de un cierto patrón de medida que podemos llamar distancia, tomar una decisión lógica lo más acertada y rápida posible. La principal referencia que se puede utilizar para caracterizar la complejidad de un SARV, de vocabulario limitado, es si se trata de voz concectada, o sea varias palabras sin pausa entre ellas, o palabras aisladas. En voz conectada, la dificultad principal radica en la obtención del comienzo y final de cada palabra, y en la gran variabilidad de parámetros en función del contexto. En resumen, la fig. 3.2-3 presenta el conjunto de problemas asociados a un SARV de vocabulario limitado.

2. Sistemas Automáticos de Reconocimiento de Palabras Aisladas.

El Reconocimiento Automático del habla, en general, ha demostrado ser un problema complejo. Por eso en la actualidad se trabaja cada vez más activamente en la aplicación y desarrollo de métodos de inteligencia artificial con lo que se espera mejorar las prestaciones de los sistemas, así como profundizar en el conocimiento de los mecanismos perceptivos y reglas de compresión del habla.

A pesar de este panorama poco alentador, para aplicaciones sencillas y específicas, existen ya métodos y técnicas suficientemente elaboradas con los que poder abordar el problema de una forma totalmente sastifactoria. Por supuesto que me refiero al Reconocimiento de Palabras Aisladas !

Como siempre, encontramos obstaculos en el desarrollo de dispositivos y uno de estos obstaculo es la elevada complejidad lo que constituye el preproceso o parametrización de la señal vocal. La tendencia tecnológica actual para aliviar esta tarea apunta a la construcción de procesadores especiales en tecnologia VLSI. Sin embargo, con la tecnologia disponible con que se cuenta hoy en día, es posible abordar el problema de forma suficientemente sastifactoria para algunas aplicaciones. En este sentido se ha propuesto un método de parametrización de la señal vocal [14] basado en la función de autocorrelación de la señal cuantificada a dos niveles, que con una complejidad matemática mínima y unos requerimientos materiales (hardware) ínfimos, ha demostrado ser suficientemente adecuado como preproceso de la señal vocal en sistemas de reconocimiento de palabras aisladas basada en la aproximación global [15] [16], que más

adelante se explicará en detalle.

1. Condiciones de Funcionamiento y Tareas básicas de un Reconocedor de Palabras Aisladas.

Debido a la extrema complejidad implicada en la tarea del reconocimiento general del habla, se hace necesario, imponer determinadas restricciones y suponer ciertas condiciones, que sin anular los objetivos que se desea, hagan factible su realización práctica. Estas consiciones deben ser las siguientes:

a) Reducir la variabilidad de la señal vocal.

Encontraremos que el sistema sera monolocutor, y trabajara en un entorno relativamente uniforme , esto significa, con las mismas condiciones de adquisición, tono y volumén de la voz sin excesiva variaciones, etc.

b) Evidenciar los problemas de segmentación.

El sistema considerara la palabra a reconocer como un todo, y no intentara subdividirla en elementos más sencillos a identificar individualmente.

c) No intentar un análisis fino de las transiciones.

Aunque gran parte de las características significativas de la señal vocal se encuentren concentradas en las transiciones de un de un estado estacionario a otro (que forman las mayoria de las consonantes), para diccionarios reducidos, del tipo de los que se emplean en la aproximación global, encontramos que es suficiente un análisis cuasi-estacionario de la señal vocal. A condición de que las palabras del léxico no sean excesivamente parecidas (/pita/, /pipa/), se consiguen con este tipo de análisis, tasas de reconocimientos sastifactorias; siempre que se cuente con un conjunto de parámetros adecuados y con métodos eficientes de evaluación de distancias.

Todo esto mencionado anteriormente, encontramos que para un sistema tendrá que realizar para cada palabra pronunciada por la persona, las siguientes tareas:

a) Un análisis de la onda de presión portadora del mensaje sonoro, que la simbolice como un conjunto de cantidades númericas, únicas que es capaz de manejar la microcomputadora que debe realizar el resto de las operaciones.

b) Una detección de los puntos de la palabra, que evite tomar como parte de ella el fondo de silencio que la rodea, evitando así cálculos inútiles y confusiones en las comparaciones. Un algoritmo simple que considere umbrales de tiempo y amplitud puede llevar a cabo esta misión.

c) Una eliminación de la redudancia de la señal vocal. Este proceso, conocido generalmente como parametrización o extracción de características, que también es una de las etapas que constituye el proceso de reconocimiento de patrones; puede realizarse mediante muy diversas aproximaciones, utilizándose aquí el cálculo de los N primeros valores de la función de autocorrelación dependiente del tiempo de la señal vocal muestreada y cuantificada de dos niveles, referenciada anteriormente [17].

d) Una normalización temporal no lineal que impida que la compara ción de la señal adquirida y parametrizada, con los patrones que forman el léxico, se vea falseada por fluctuaciones temporales, los cuales se producen siempre de una pronunciación a otra de la misma palabra. Esta normalización se lleva a cabo, clasicamente mediante un algoritmo derivado de las técnicas de programación dinámica.

e) Un cálculo de distancia entre la palabra normalizada y las que forman el diccionario, que permita decidir a quién se parece se parece más aquella, y afirmar en caso de similitud suficiente, que se ha identificado la palabra escuchada. El mismo algoritmo de programación dinámica proporciona estas distancias al tiempo que se normaliza temporalmente.

Accesoriamente, como tarea no menos importante y previa a todo el proceso de reconocimiento, el sistema deberá llevar a cabo la construcción de un diccionario óptimo de patrones, asegurándose que cada uno de estos es representativo de la palabra que simboliza. Pronunciando varias veces cada patrón, y escogiendo la mejor versión como la que más se parece en promedio a todas las demás, se consigue descartar o eliminar, en la mayoria de los casos, versiones pocos representativos o erronéas.

a) Fundamentos Teóricos: Programación Dinámica.

En la aproximación global, las palabras a reconocer (palabra-muestra), así como las palabras patrón del diccionario, se consideran como puntos en un espacio multidimensional. Definida una distancia en dicho espacio, la tarea de reconocer una palabra-muestra consiste en

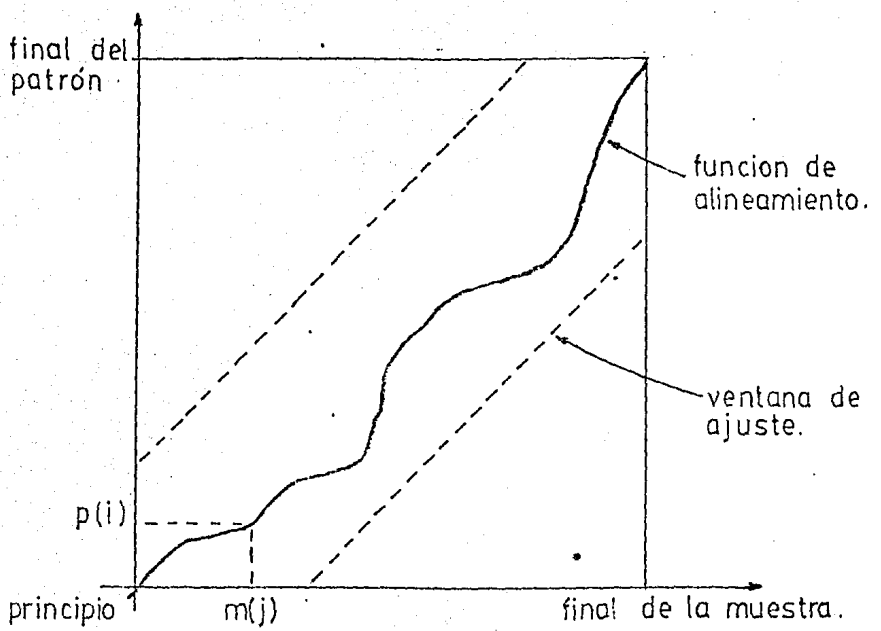


FIG. 3.2-4
FUNCION DE ALINEAMIENTO.

compararla con cada una de las palabras patrón y elegir como reconocida aquella que arroje la mínima distancia. Encontramos, entonces que la extracción de parámetros se realiza a través de una ventana temporal que se desliza sobre la señal a intervalos discretos de tiempo (típicamente de 10 a 20 ms., igual como se explicó en la primera parte correspondiente al análisis).

La definición de una distancia que mida la disimilitud entre estas dos cadenas de parámetros, es un punto crucial en esta aproximación. Como una misma palabra puede pronunciarse a distintas velocidades, y como las duraciones de los distintos segmentos de dicha palabra son subceptibles de variación independiente, es necesario un método de normalización temporal no lineal de los patrones y muestras para posibilitar la necesaria definición de distancia. Dicho método debe tolerar una variabilidad temporal con que el habla corriente se pronuncien los distintos fonemas y debe imponer tan sólo restricciones físicas de continuidad, monotonicidad, etc., en las formas de variación de los parámetros. Aunque se hayan expuesto diversas alternativas para la normalización antes citada, los métodos que finalmente se consideran como óptimos están basados en algoritmos de programación dinámica. Dichos algoritmos proporcionan simultáneamente la normalización temporal buscada, así como la definición de distancias, cuyo cómputo es asimismo realizado por el algoritmo.

Antes de utilizar el algoritmo de programación dinámica, primero se explicara la función o en que consiste; de una forma detallada el alineamiento temporal.

1. Función de Alineamiento Temporal.

Una función de alineamiento temporal relaciona cada elemento de la muestra con uno del patrón y se puede visualizar como un 'camino' entre los vertices del rectángulo obtenido de representar la serie patrón en el eje de ordenadas y la serie muestra en el eje de las abcisas, fig. 3.2-4. La pendiente del camino en un punto proporcionará el 'estiramiento' a que debe someterse la muestra en ese punto, para ajustarse al patrón. En el caso ideal en que la función de alineamiento fuese una recta, se tratara entonces de una normalización temporal lineal. La pendiente de la recta proporcionara entonces el factor de escala aplicado en la normalización. Una vez hallada la función de alineamiento óptima, la distancia patrón-muestra se puede dar como la suma normalizada de las distancias entre vectores de parámetros del patrón y de las muestras relacionadas por la función de alineamiento.

a) Alineación temporal continua.

Encontramos, que en el caso general que el patrón (P) como la muestra (M) estén representados por vectores de parámetros de variación continua en el tiempo P:P(t); en donde 0 <= t <= tp: y M:m(t); para o <= t <= Tm se puede definir la distancia patrón-muestra: D(P,M) a lo largo de una función de alineamiento temporal F:f(s) = ((u/s),v(s)) lo cual para 0 <= s <= Ts mediante la integral:

$$D_f(P,M)=\int_0^{Ts} d[p(u(s)),m(v(s))]ds/\int_0^{Ts} ds$$

donde el integrado representa la función distancia aplicada entre vectores de parámetros. La función de alineamiento óptima, que es la que interesara hallar, se define entonces como aquella que minimiza la distancia , con lo

$$D_f(P,M)$$

que la distancia patrón-muestra queda definida por:

$$D(P,M)\% \min_f[D_f(P,M)]$$

En tal caso, si la distancia entre vectores de parámetros, como es usual se define simétrica, entonces D(P,M) cumple entonces las propiedades [18]:

- Es semidefinida positiva D(P,M) >= 0, D(P,P) = 0.
- Es conmutativa D(P,M) = D(M,P).
- Es reversible con el tiempo. D(P,M) no cambia cuando se invierte la dirección del tiempo.
- Es consistente con la normalización lineal: Si Tp=Tm y U(s)=V(s), 0 <= s <= t* en el cambio óptimo, entonces

$$D(P,M)=\left(\int_0^{Tp} d[p(t),m(t)]dt\right)/Tp$$

que es una normalización lineal: Esta propiedad implica la primera.

Debe notarse que la desigualdad triangular D(A,B) + D(B,C) >= D(A,C) no es imprescindible para que esta 'distancia' pueda utilizarse en el reconocimiento de la palabra.

b) Alineación Temporal Discreta.

Cuando el patrón y la muestra, como ocurre normalmente, están representados, no por una variación continua sino por una serie de valores de los vectores de parámetros:

P:p(1), p(2), ... p(I) y M:m(1), m(2), ... m(j);

se hace necesaria una formulación discreta. Para ello, resulta conveniente definir la función de alineamiento en base a un conjunto de 'producciones'. Un par (a,b) se llama una producción simple positiva, si a y b son enteros, y si a >= 0, b >= 0. Si X = {(i,j):1 <= i <= I, 1 <= j <= J} representa el rectángulo patrón-muestra y Prod es un conjunto fijo de producciones, podemos definir un camino F sobre X como una secuencia de elemento de X. Esta secuencia se puede expresar recursivamente haciendo uso del conjunto fijo de producciones Prod. Así pues, F las podremos dar como la aplicación:

1) F(1) = (1,1).
2) F(L) = (I,J).
3) F(K) = (u(k-1) + a(k), v(k-1) + b(k))
   a(k), b(k) $\in$ Prod;
   k $\in$ {2, ... L_{F-1}}

donde, encontramos que L es la 'longitud' del camino F. Para relacionar una producción determinada con el segmento de tiempo normalizado que representa, se requiere la introducción de una función peso, w:Prod ---> R . La distancia acumulada entre P y M a lo largo del camino F se puede entonces definir como:

$$D_F(P,M) = \left[\sum_{k=2}^{t_f} d(P(u(K))), m(v(K))) \cdot W(a(K),b(K) + d[p(1), m(1)] \right] \Big/ \sum_{k=2}^{t_f} W(a(K),b(K))$$

y la distancia total patrón-muestra como la distancia acumulada a lo largo del camino óptimo Fo, tal que .

$$D(P,M) = D_{f0}(P,M) = \min_{f}(D_f(P,M))$$

2. Algoritmo de Programación Dinámica.

Como vimos antes, para el cálculo de la función de alineamiento temporal óptimo, se propusieron diversas técnicas y como son alineamiento por eventos, maximización de la correlación, etc. [19]. Sin embargo, usualmente se recurre a un algoritmo recursivo derivado de las técnicas de optimización por programación dinámica que por medios de trabajos anteriores, ha demostrado ser muy eficaz en una gran variedad de sistemas [19] [20] [21] [22]. Entonces, para que el algoritmo sea aplicable se debe exigir que la suma de los pesos a lo largo de un camino sea la misma para todos los caminos:

$$\sum_{K=2}^{t_f} W(a(K),b(K)) = N$$

En cuyo caso, la expresión de la distancia a lo largo del camino óptimo se reduce a:

$$D(P,M) = \frac{1}{N}\min_{F}\left[\sum_{K=2}^{t_f} d(p(u(K)),m(v(K)))\cdot W(a(K),b(K)) + d(p(1),m(1))\right]$$

Se han propuesto dos tipos de definición simple para los pesos, que cumplen la condición impuesta:

a) Forma Simétrica: $W(k) = a(k) + b(k)$; $N = I + J$.
b) Forma Asimétrica: $W(k) = a(k)$; $N = I$ ó
$W(k) = b(k)$; $N = J$.

donde (a,b) es la producción afectada.

La principal diferencia entre estas dos definiciones radica en que la definición simétrica conserva la conmutatividad presente en la definición general de la distancia total, mientras que la forma asimétrica pierde dicha conmutatibilidad: $D(A,B) \neq D(B,A)$. En la práctica, ambas formulaciones dan resultados equivalentes, sobre todo en el caso de que se utilicen restricciones de pendiente [23]. Con la condición exigida a los pesos, el camino óptimo se puede obtener mediante la relación de recurrencia:

$g(i,j) = \min_{(a,b)} \{ g(i-a, j-b) + d(p(i), m(j)) \cdot w(a,b) \}$ (a,b)95
Prod. que con la condición inicial: $g(1,1) = d(p(1), m(i) \cdot w(1,i)$ proporciona la expresión de la distancia: $D(P,M) = g(I,J)/N$.

. . .

Debemos observar que g(i,j) representa la distancia total acumulada para llegar a través del camino mínimo hasta el punto (i,j) del rectángulo patrón-muestra. Esta relación de recurrencia halla el camino óptimo tal como se ha definido, siempre que los pesos cumplan la condición impuesta. Sin embargo, encontraremos que en todos los casos la distancia D(P,M), pierde alguna de las propiedades mencionadas al hablar de la función de alineamiento continua. Otro tipo de formulación (Programación Dinámica Trapezoidal [18] permite conservarlas todas, asegurando una mayor aproximación a la distancia dada por la función continua ideal, pero su mayor complejidad hace que sea menos utilizada.

Para limitar el número total de puntos g(i,j) a calcular para la relación de recurrencia, se utiliza una ventana de ajuste en el plano patrón-muestra, que descarte aquellos puntos cuya pertenencia al camino mínimo es físicamente improbable. La versión que aquí se utiliza consiste en dos rectas paralelas de pendiente unidad que limita la máxima diferencia temporal entre un punto del patrón y su correspondiente de la muestra.

3. Refinamiento de Bordes y construcción de diccionario.

En reconocimiento de palabras aisladas, el fondo sobre el que está colocado el objeto (palabra) a reconocer está constituido por tramos de (teórico) silencio. Estos tramos están formados en realidad por ruido no necesariamente blanco, de (relativamente) pequeña energía, sobre el que ocasionalmente se pueden mezclar ruidos espurios de gran amplitud.

Todo el problema de la detección del principio y fin de una palabra sobre este fondo de ruido, se basa en encontrar características (parámetros) que varíen notablemente al pasar de un lugar (temporal) donde hay palabra a uno donde no lo hay. El problema general es complejo, pues la gran variabilidad de la señal vocal hace que ésta adopte en ciertos fonemas, por ejemplo en fricativos débiles, una estructura muy parecida a la de un ruido rosa, mientras que otros (explosivos) están en parte formado por zonas de silencio. Ello hace que los sistemas que deben trabajar con ruido de fondo de cierta amplitud, recurran a combinaciones complicadas de diversos parámetros más o menos elementales (densidad de cruces por cero + amplitud + etc.) [23]. Estos parámetros son a veces

extraidos ex-profeso para la detección de bordes, pero en general resulta más conveniente aprovechar la labor el parametrizador y llevar a cabo la etección después de la parametrización, utilizando los mismos parámetros que para el reconocimiento, que en teoría, son los que mejor caracterizan la señal.

a) Detección Gruesa.

Un algoritmo simple y eficaz para la detección gruesa de bordes , válido si el ruido no es excesivo, utiliza umbrales de tiempo y de energía (o amplitud) de la señal; la superación de estos umbrales indica la presencia de la palabra. El umbral de tiempo se hace necesario porque no es suficiente afirmar que hay palabra cuando hay energía: ello llevaría a tomar como palabra un ruido espurio. Una palabra presenta siempre un mínimo de energía durante un mínimo de tiempo; y viceversa, la ausencia de energía no implica fin de palabra, puesto que se terminaría al encontrar el silencio que forma parte de un fonema explosivo (/p/, /k/, ...). Una zona de silencio real tiene un mínimo de duración. Por otra parte, es evidentemente necesario considerar también como parte de la palabra el trozo de la señal inmediatamente anterior y posterior al momento del paso por el umbral de energía, pues en caso contrario podrian perderse los principios y/o finales de palabra, constituidos por señal de energía inferior al umbral (nasales, fricativas débiles, etc., iniciales y finales).

b) Detección fina.

El método descrito no permite en realidad una detección exacta del principio y final de la palabra, debido fundamentalmente a la posible existencia de fonemas inicial y/o final de muy débil de energía, los cuales, como ya se ha mencionado, obligan a preservar un segmento anterior y otro posterior de longitud fija. Lo único que asegura es que la palabra contenida en su totalidad entre las fronteras detectadas. Con todo, el método resulta siuficientemente en muchos casos, aunque un procedimiento más fino conduce no sólo a una mejora de la tasa de reconocimiento, sino a un indudable ahorro, tanto de la memoria requerida para el diccionario, como de tiempo empleado en el reconocimiento.

¿ Pero, de todo esto como lo hace en realidad, cómo funciona ? Encontramos, que el algoritmo de detección gruesa, requiere en la práctica de un buffer (lo que se traduciria a memoria tampón) cíclico, que permita trabajar a 'micrófono abierto', es decir, con tiempos de silencio iniciales indefinidamente largos. En este buffer se va almacenado cíclicamente la señal (parametrización o no), con lo que en cada momento se dispone de lo adquirido en ese instante y en los N instantes anteriores, N dependiendo

de la talla del buffer. De esta manera, en el momento que se detecta el 'principio eficaz' de la palabra se dispondrá aun del segmento de señal inmediatamente anterior en el que, como se ha dicho anteriormente, está contenido el principio real de la palabra [24]. El tamaño del buffer ciclico dependerá pues de la longitud temporal del segmento inicial anterior al 'principio eficaz' que sea necesario conservar.

El algoritmo de detección fina trabaja a partir del resultado proporcionado por el algoritmo de detección gruesa. Para el principio de la palabra busca de atrás hacia adelante a partir del momento en que se cruzó inicialmente el umbral de amplitud, el punto donde los parámetros extraidos indican la existencia del silencio. Para el fin de la palabra, también de atrás hacia adelante, busca a partir del final 'grueso', el momento en que deja de haber silencio, el algoritmo que muestra todo esto que se ha estado exponiendo se presenta a continuación:

```
principio

    inicializar   apuntador al indice del primer vector de
                  parámetros.

    mientras      vector de parámetros indicando por
                  apuntador <> silencio y apuntador > 1,
                  decrementar apuntador.

    principio-fino:=apuntador.

    inicializar   apuntador al indice del último vector de
                  parámetros.

    inicializar   final al indice del último vector de
                  parámetros cuyo parámetro amplitud >
                  umbral de amplitud.

    mientras      vector de parámetros indicado por
                  apuntador sea silencio y apuntador >
                  final, decrementar apuntador.

    final-fino:=apuntador.


fin.
```

c) Construcción del diccionario.

En reconocimiento global de palabras aisladas (más adelante se explicará), siempre es necesario una fase de aprendizaje, previa al reconocimiento propiamente dicho, cuya misión es almacenar una versión parametrizada de cada una de las palabras del diccionario a reconocer, junto con una cabecera que permita identificarla. En general, se

utilizan varias versiones de una misma palabra para construir un patrón del diccionario. En este caso se hace necesario un proceso de selección o promedio para indentificar o construir la versión más adecuada para servir como referencia. Un posible método consiste en escoger aquella versión cuya suma de distancias halladas por programación dinámica, para que las otras versiones de la misma palabra sea mínima.

Después de la fase de selección/optimización, el diccionario (que constituye el conocimiento verbal del sistema) se construye meramente reuniendo las versiones más adecuadas de cada palabra. En determinados casos puede ser necesario utilizar más de una versión por palabra a reconocer, si ésta admite dos o más pronunciaciones notablemente distintas.

Aunque sea en general posible construir el diccionario mediante promedio/selección de palabras provenientes de distintos locutores con el fin de conseguir un sistema de reconocimiento de múltiples locutores, ello lleva normalmente una fuerte disminución en las prestaciones del sistema, que sólo puede compensarse mediante una complicación del mismo a todos los niveles.

Luego de explicar en sí, lo que consiste un sistema automático de reconocimiento de palabras aisladas sencillo (un sólo locutor), podemos partir de esta base como fundamento para exponer y describir los resultados obtenidos por un investigador español [25], en la evaluación de dos procedimientos o métodos de reconocimiento de palabras aisladas, en un contexto independiente del locutor y en el lenguaje español, que es el que nos interesa. Se ha comparado el procedimiento clásico de reconocimiento de patrones, con el procedimiento cuasi-fonético en el que se utilizan, además de las características paramétricas de los sonidos, la información lexical asociada al vocabulario.

El interés de este trabajo residió en demostrar la posibilidad de diseñar un sistema de reconocimiento cuasi-fonético, que pueda igualar e incluso mejorar, los resultados obtenidos con los sistemas adaptivos, que se conciben mediante algoritmos ciegos a partir de los módelos de reconocimiento de patrones.

b) Reconocimiento global de tipo acústico.

Guiándonos por el proceso de reconocimiento de patrones que se explicó al inicio del capitulo, en la fig. 3.2-5 lo podemos aplicar al proceso de reconocer palabras aisladas. El proceso consiste en que:

a) El bloque 1 realiza el análisis de la señal para obtener un vector q de parámetros (componentes

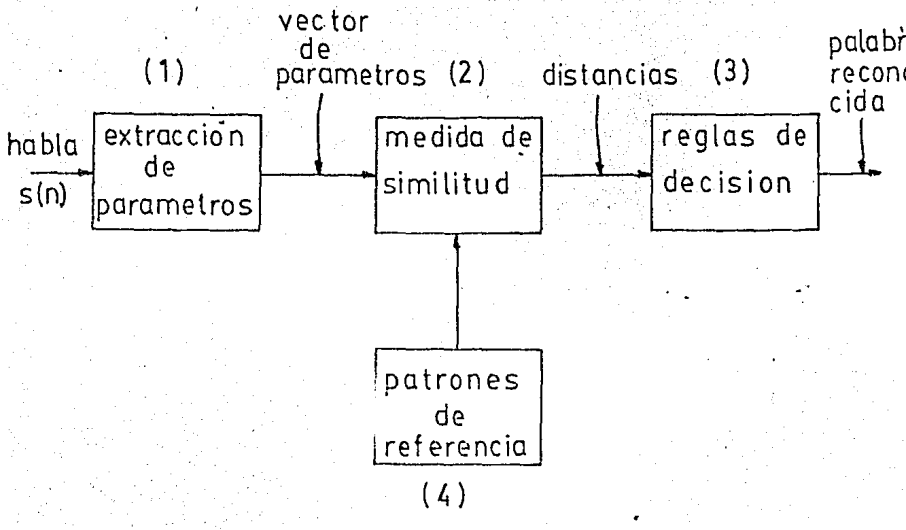


FIG. 3.2-5
MODELO CANONICO DE RECONOCIMIENTO GLOBAL DE PALABRAS AISLADAS.

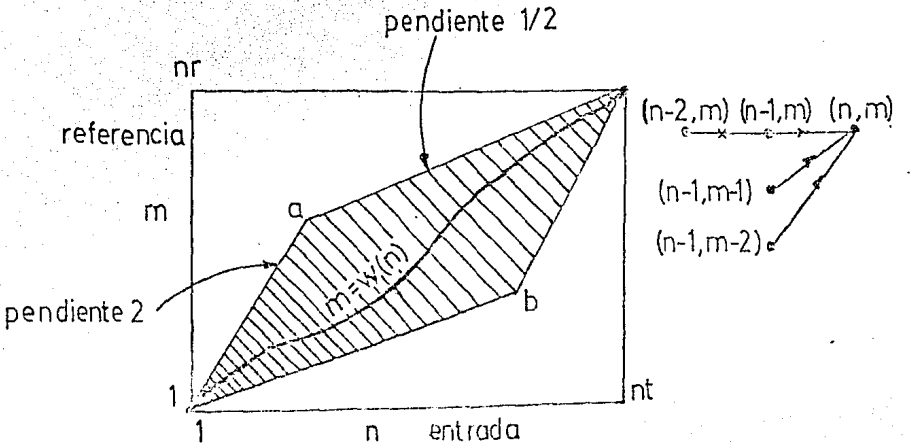


FIG. 3.2-6
ALINEAMIENTO ENTRE PALABRA A RECONOCER Y REFERENCIA.

espectrales, coeficientes LPC, etc.), para las m muestras que constituye un tramo de señal. Si tenemos que la duración de la palabra es l tramos, se podrá configurar una matriz de q x l parámetros que definen o caracterizan dicha palabra.

b) El bloque 2 se encarga de realizar una comparación, esto definido a través de una medida de distancia, entre la palabra a reconocer y cada uno de los patrones de referencia almacenados en el bloque 4.

c) El bloque 3 decide, entre todos los posibles candidatos obtenidos del bloque anterior y en base a unas determinadas reglas de decisión, el candidato con mayor probabilidad de coincidir con la palabra a reconocer.

En el siguiente diagrama que se muestra, representa un diagrama de bloques más detallado , correspondiente a una parametrización LPC de la señal. La señal de habla se pasa a través de un filtro pasabanda y se discretiza mediante muestreo a través de un convertidor A/D. El primer proceso a que se somete la señal conduce a la detección de la duración de la palabra, determinación fundamental para el correcto funcionamiento del módelo. La señal digitalizada es preacentuada usando un filtro lineal simple y es descompuesta en un conjunto de n tramos (por ejemplo de 300 muestras) con un determinado solapamiento con las tramos adyacentes (por ejemplo 200 muestras). De esta manera cada m=100 muestras se obtiene un conjunto de parámetros correspondientes a un tramo de n=300 muestras. A cada uno de estos tramos se le aplica una ventana Hamming de 300 muestras, y se realiza un análisis LPC (por ejemplo p=8) por el método de autocorrelación.

La medida de la distancia se lleva a cabo a través de un algoritmo de ajuste dinámico (DTW - Dynamic Time Warping), mediante el cual la palabra a reconocer es alineada con cada una de las referencias del vocabulario, previamente almacenadas. Pero, nos preguntamos, cuál es la necesidad de este alineamiento ? Encontraremos, que la necesidad de este alineamiento, (fig. 3.2-6) se deriva de la diferente duración de las palabras del vocabulario, aun para la misma palabra articulada por diferentes locutores o en distintos instantes de tiempo.

El criterio para realizar el alineamiento consiste en determinar un camino óptimo w(n) que minimice la distancia entre la palabra a reconocer (T) y las sucesivas referencias (R). Es decir, aquel que conduce a la determinación de:

$$D^* = \min_{w(n)} \left[\sum_n d(T(n), R(w(n))) \right]$$

La medida de la distancia d(T,R) es una de las características más eficientes del módelo. Itakura [28] demostró que, desde un punto de vista estadístico, el mejor estimador de la distancia era:

$$d(T,R) = \log \left[\frac{a_R V_T a^t_R}{a_T V_T a^t_T} \right]$$

són los vectores de coeficientes LPC y es la matriz de coeficientes de autocorrelación. Esta distancia puede ser calculada con sólo p+1 multiplicaciones y sumas, y un logaritmo, (p es el orden del predictor LPC). Una vez obtenidas las distancias entre la entrada que se ha de reconocer y las referencias que componen el vocabulario, es necesario decidir cuál de dichas referencias es el candidato óptimo para identificarlo con la entrada. Aunque los criterios de decisión pueden ser muy diversos, generalmente se usan dos reglas de decisión : la regla del vecino (NN - Neighbor rule) y el vecino más cercano (KNN - Nearest neighbor rule), típicos en la técnica de reconocimiento de patrones. Vamos a ver en que consiste cada uno de ellos:

1. Procedimiento NN.

Si existen referencias almacenadas en memoria, R , i=1,2,...,V, y para cada una se ha obtenido una distancia D se elige:

$$i^* = \underset{i}{\text{argmin}} [D^i]$$

En determinadas ocasiones, sobre todo cuando se usan informaciones adicionales para la decision, se obtienen una lista ordenada de distancias, tales que: D[1] <= D[2] <= ... <= D[V].

2. Procedimiento KNN.

Este procedimiento se utiliza cuando cada palabra del vocabulario está representada por varias referencias, en un sistema independiente del locutor por ejemplo. Si suponemos que existen P referencias para cada una de las palabras, y representamos la referencia j de la palabra i

como R , tal que 1 <= j <= v, 1 <= j <= P, y análogamente $D^{i,j}$ como la distancia a la referencia j de la palabra i; y si ordenamos ahora las P distancias a la palabra i:

$$D^{i,[1]} \leq D^{i,[2]} \leq \ldots \leq D^{i,[p]}$$

se calcula una distancia media correspondiente a las primeras K distancias:

$$r^i = \frac{1}{K} \sum_{K=1}^{K} D^{i,[k]}$$

y se escoge como candidato la referencia :

$$i^* = \underset{i}{\operatorname{argmin}}\ r^i$$

La ventaja de este procedimiento, frente al NN; se obtie ne cuando P > 6, en cuyo caso el valor de K =2 ó 3, obtiene mejores resultados.

c) Reconocimiento cuasi-fonético.

Este tipo de sistema de reconocimiento que se describe, consiste en la caracterización de las palabras del vocabulario a reconocer a través de la clasificación de sus fonemas, mediante un conjunto robusto de parámetros que lesrepresenten de una forma independiente del locutor. No se desciende a la segmentación e identificación de los fonemas, pero permite, en el entorno de un determinado vocabulario, la clasificación de los diferentes sonidos.

El algoritmo de decisión se basa en el estudio previo que se hace, para cada palabra del vocabulario, de la evolución temporal que sigue el conjunto de parámetros que la definen. Usando esta información, junto al conocimiento fonético de la estructura lexical de las palabras, permite alcanzar resultados muy sastifactorios en las tasas de reconocimiento, sobre todo para vocabularios limitados en los que no exista un alto grado de coincidencias fonéticas en las palabras que lo forman (fig. 3.2-7).

Los parámetros elegidos para la caracterización de los sonidos son:

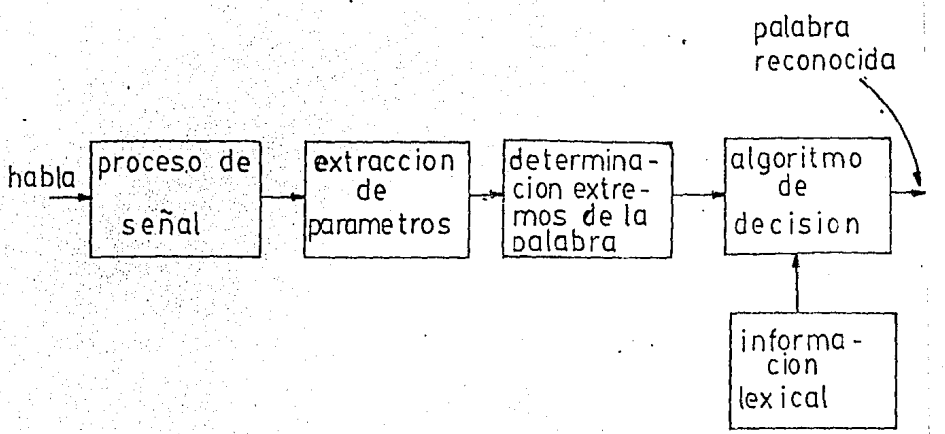


FIG. 3.2-7
DIAGRAMA GENERAL DE BLOQUES DEL SISTEMA DE RECONOCIMIENTO CUASIFONETICO.

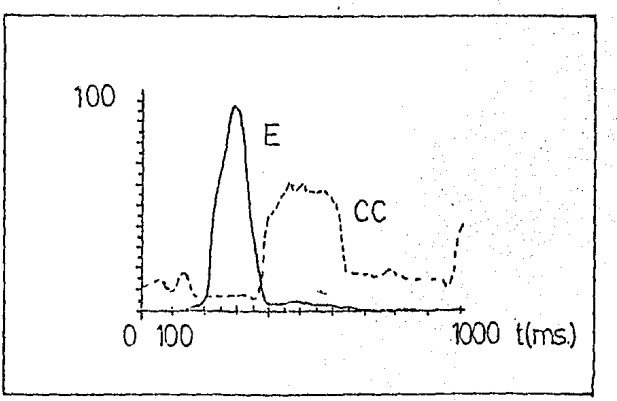


FIG. 3.2-8
DISTRIBUCION TEMPORAL DE ENERGIA Y CRUCES POR CERO DE LA PALABRA 'DOS'.

- Distribución temporal de energía (E).
- Distribución temporal de cruces por cero (CC).
- Distribución temporal del espectro de energía (F) utilizando LPC.
- Error residual del módelo LPC anterior (ERR).

El número de cruces por cero, la distribución temporal de energía y la distribución espectral de energía son parámetros adecuados para la caracterización de la señal. En efecto, los sonidos fricativos tienen una alta proporción de cruces por cero, una energía media relativamente pequeña y una distribución espectral en la que la energía se concentra en las altas frecuencias. En la fig. 3.2-8 se representa la distribución de energía normalizada (E) y la distribución de cruces por cero (CC), para los sonidos correspondientes al dígito < 2 >; y en la fig. se recoge la distribución espectral del sonido fricativo /s/, del mismo dígito. La medida de la distribución espectral de energía debe realizarse por un procedimiento que descubra las características más sobresalientes del espectro con objeto de que sean, en lo posible, independiente del locutor. Makhoul y Wolf [29] demostraron que se puede seguir utilizando en el análisis de la señal, un módelo LPC de orden p=2. En este módelo se obtienen dos polos reales, o dos polos complejos conjugados, que dan lugar a una determinada frecuencia de resonancia que se sitúa en la región del espectro con mas elevada concentración de energía. La situación en el espectro de la frecuencia de resonancia depende de la posición en el espectro de los formantes de los sonidos, así como de sus amplitudes y anchos de banda relativos. En este punto es conveniente realizar las siguientes consideraciones:

1. Para sonidos vocálicos las amplitudes de los dos primeros formantes, f1 y f2, son comparables y generalmente mayores que las de los siguientes formantes [30]. El márgen de variación de la frecuencia correspondiente al segundo formante (f2) es mayor que el del primer formante (f1). Esto lo podemos ver en la tabla 3.2-1.

   En consecuencia, la frecuencia de resonancia obtenida mediante un módelo LPC de orden p=2, tiende a seguir la evolución del segundo formante (f2). En la fig. 3.2-9 se representa la distribución espectral de energía del sonido vocálico /i/ en la palabra < cinco >, mediante un análisis FFT y dos módelos LPC; uno de orden P=15 y otro p=2. Se observa la posición de los dos primeros formantes, y la frecuencia de resonancia del módelo p=2 situada entre ellos.

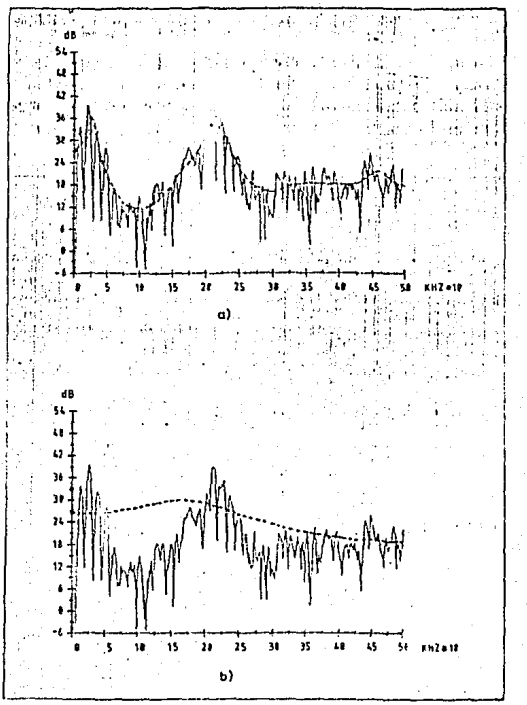


FIG. 3.2-9
DISTRIBUCION ESPECTRAL DE ENERGIA DEL SONIDO/i/ DE - LA PALABRA CINCO.

a) p=15.

b) p=2.

Tabla 3.2-1
Frecuencia (Hz) de los formantes para los sonidos vocálicos.

Hz	f1	f2	f3	f4	f5
a	733	1165	2518	3280	4180
e	444	1986	2629	3300	4350
i	286	2205	2876	2390	3660
o	553	773	2584	3085	3600
u	295	609	2446	3210	3930

2. La diferencia más notable entre los espectros de los álofonos oral y nasal de los sonidos vocálicos, es el aumento del ancho de banda del primer formante cuando el sonido esta nasalizado [30].

3. Si las frecuencias de los dos primeros formantes f1 y f2 están suficientemente separadas entre sí, o la amplitud de f1 es significativamente mayor que la de f2, la frecuencia de resonancia seguirá a f1, y obtendremos dos polos situados en el eje real.

4. Para sonidos fricativos y africados la energía se concentra en la zona superior del espectro, de una forma más o menos acusada, en función del grado de sonoridad. En la fig. 3.2-10 y 3.2-11 se representan las distribuciones espectrales del sonido fricativo /s/ de < dos > y el sonido africado sordo /c/ de < ocho >.

5. Las coarticulaciones, entre los diferentes sonidos que forman una palabra, influyen en la evolución de la posición de la frecuencia de resonancia, de tal manera que la trayectoria de la misma indica la secuencia de dichos sonidos.

El error residual de LPC (ERR) contiene información sobre la forma de distribución de la energía en el espectro. Puede demostrarse que cuando la energía está más concentrada; es menor el valor del error. Por esta razón, para los sonidos del habla, el error residual LPC es pequeño para sonidos vocálicos y aumenta a medida que la energía se distribuye más en el espectro, hasta alcanzar valores máximos en los sonidos fricativos. Por último, es conveniente indicar que no se usan umbrales fijos de comparación, para los diferentes parámetros, en la toma de decisiones. Estos umbrales dependen del locutor, de las características ambientales de ruido, etc. Una solución para resolver este problema es la utilización de una técnica de autonormalización, de tal manera que el valor de los umbrales se establece en función del 'ruido ambiental en silencio' y de las medidas realizadas directamente sobre la forma de onda de la señal a reconocer.

d) Evaluación entre el módelo global de tipo acústico y el módelo cuasi-fonético.

La diferencia entre un método y otro lo podemos hacer, evaluando cada método por separado. En el primero, encontramos que es un módelo canónico con independencia del

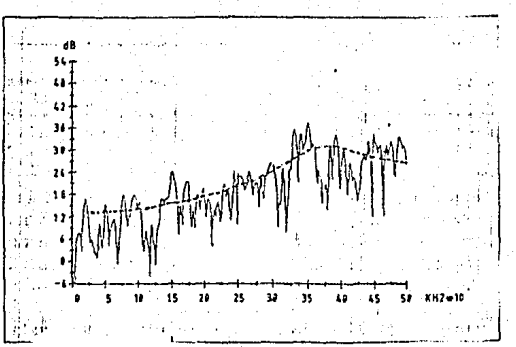


FIG. 3.2-10
DISTRIBUCION ESPECTRAL DE ENERGIA DEL SONIDO /s/ DE LA PALABRA DOS (p= 2).

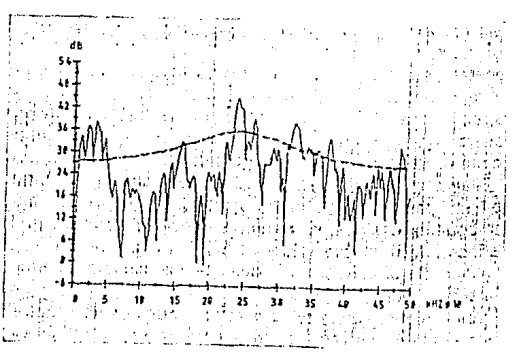


FIG. 3.2-11
DISTRIBUCION ESPECTRAL DE ENERGIA DEL SONIDO /c/ DE - LA PALABRA OCHO (p=2)

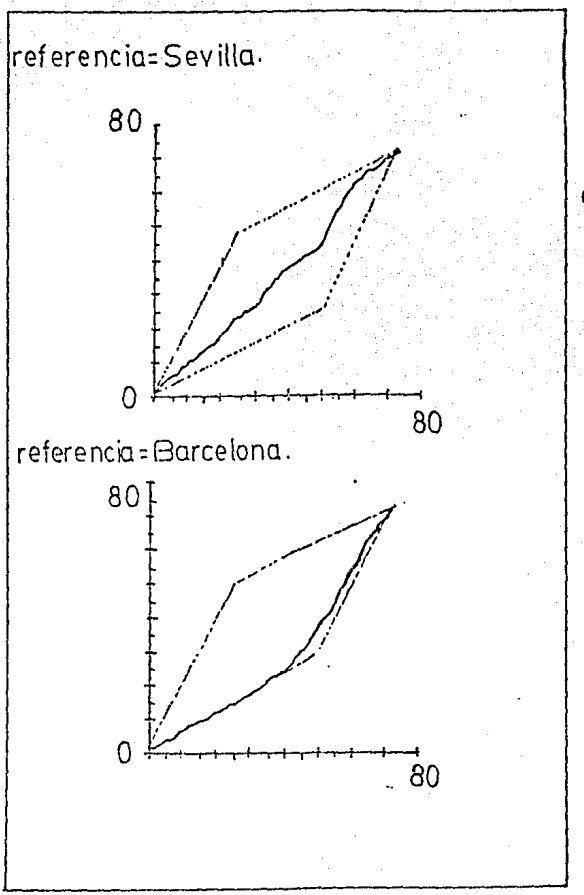


FIG. 3.2-12
ILUSTRACION DEL ALGORITMO DTW.

locutor en la cual presupone la necesidad de crear un conjunto de referencias del vocabulario a reconocer. Este razonamiento es aplicable tanto al caso de realizar un sistema dependiente o independiente del locutor; la única variación se refiere al procedimiento de creación de dichas referencias. Con objeto de comprobar la eficacia del módelo, se realizó una evaluación del sistema para un sólo locutor con dos vocabularios diferentes: vocabulario de los digitos en español y vocabulario formado por el nombre de 40 ciudades. El resultado fue que para la tasa de reconocimiento obtenida, para el vocabulario de los dígitos y para 10 repeticiones de cada dígito (10 x 10 = 100 palabras), ha sido del 100 %. En el caso del segundo vocabulario, para el mismo locutor, realizarón 10 repeticiones de los nombres de cada ciudad (10 x 40 = 400 palabras); la tasa de reconocimiento fue del 99.48 %. En la fig. 3.2-12 se muestra el algoritmo de comparación a través del mapa de DTW, para la entrada Sevilla y las referencias Sevilla y Barcelona.

En un contexto de independencia del locutor, la creación de referencias es el problema crucial. Para compensar la variabilidad entre los diferentes locutores se recurre como habitualmente se realiza en las técnicas de reconocimiento de patrones, a una técnica de agrupamiento (clustering) que asocia las repeticiones afines, es decir, aquellas que están próximas en el espacio paramétrico multidimensional que se considere. El siguiente paso consiste en elegir, para cada agrupamiento, la repetición más significativa del grupo, o tomar una referencia representativa del grupo cuyos parámetros sean el valor medio correspondiente a todas las repeticiones que forman dicho grupo. En la evaluación realizada, en un contexto independiente del locutor, se utilizarón 10 locutores y cada uno articuló 10 repeticiones cada una de las palabras que componen el vocabulario.

Para crear las referencias, correspondientes a los 10 dígitos, se procedió a considerar las cuatros primeras repeticiones articuladas por cada locutor. Es decir, en total se tiene 4 x 10 = 40 repeticiones de cada una de las palabras del vocabulario, que se va a representar por Xi(j), donde i=# de la repeticio, i=1,..,40 y j=# de orden en el vocabulario, j=1,..,40; a partir de aquí se calculó la distancia d(X(i),X(j)), entre cada repetición y todas las demás. En consecuencia, se obtiene una matriz de distancias d[i,j]; 1 <= i <= 40 y 1 <= j <= 40. La formación de grupos se realiza, por ejemplo, asociando a la repetición Xl(i) todas las Xl(j) tales que d(Xl(i),Xl(j)) < d. De esta forma se descubren los diferentes agrupamientos, al mismo que se detectan las repeticiones que no tienen ninguna afinidad con las demás. Por este procedimiento se consigue, además, una lista ordenada de grupos de acuerdo con su representatividad. El último paso consiste en determinar cuál de las repeticiones de cada grupo se debe

Tabla 3.2-2
Resultado totales de la evaluación del sistema de reconocimiento global de tipo acústico.

No. de referencias / dígito	4	5	6	7	8
0	96	98	100	100	100
1	97	100	100	100	100
2	78	95	96	100	100
3	82	92	97	100	100
4	96	100	100	100	100
5	100	100	100	100	100
6	75	90	100	100	100
7	94	100	100	100	100
8	100	100	100	100	100
9	98	100	100	100	100
TOTAL	95.4	98.4	99.3	100	100

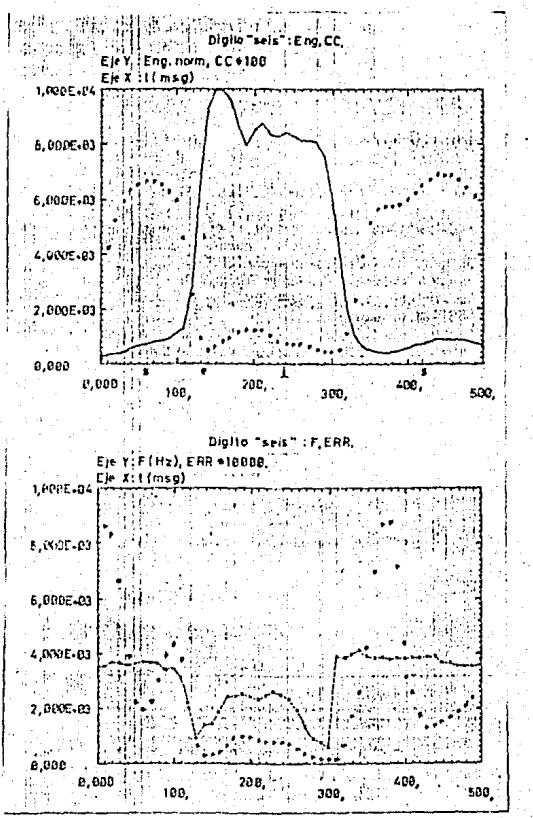


FIG. 3.2-13
CARACTERIZACION PARAMETRICA DE SEIS.

tomar como referencia representativa. El criterio seguido es:

Sea G_{kl} el grupo k de la palabra l, se dice que, $X_l^{(i)} \in G_{kl}$ es una referencia representativa del grupo si,

$$\max[d(X_l^{(i)}, X_l^{(j)})] \text{ es minima, } \forall\ x_l^{(i)} \in G_{kl};$$

Este modo de elección evita la necesidad de crear una referencia abstracta que sea el promedio de todos los componentes del grupo. En la tabla 3.2-2, se recoge los resultados totales de la evaluación, para los 10 locutores, en función del número de referencias considerado por dígito.

En el segundo módelo, que es el cuasi-fonético, se ha utilizado el vocabulario formado por los diez dígitos del español. La transcripción ortográfica-fonética y secuencia de sonidos se muestra en la tabla 3.2-3.

Un ejemplo de la información que nos proporciona la tabla: es la sécuencia de parámetros que se obtiene de cada señal a identificar, como es la secuencia temporal de los valores de energía (E) (trazo continuo), tasa de cruces por cero (CC) (línea de puntos), frecuencia (F) (línea discontinua) y error (ERR) (línea de puntos), un ejemplo es la fig. 3.2-13, correspondiente al dígito 'seis' : Pero, de que forma se realiza el reconocimiento ? ¿Cómo identifica cuál número fue el que se pronunció ? Una solución es utilizar un algoritmo que nos permita hacer una búsqueda de la palabra a reconocer. El algoritmo de decisión se ha dividido en dos fases:

a) Fase de clasificación previa.
b) Fase de Reconocimiento final.

En la primera fase se establece una clasificación que permite agrupar las palabras del vocabulario en clases afines, utilizando características generales, que permiten básicamente la eliminación de candidatos. La fase de reconocimiento final distingue el dígito articulado, profundizando en la caracterización de las secuencias paramétricas.

La forma en que se realizó la prueba, fue igual al del módelo anterior, en la cual se utilizó 10 locutores; cada locutor ha repetido diez veces cada uno de los diez dígitos (10 x 10 = 100 repeticiones). Podemos ver como resultado de esta prueba en la fig. 3.2-14 la tasa de reconocimiento para los diez locutores (trazo discontinuo) y la tasa medida de reconocimiento (trazo continuo). Estos datos aparecen resumidos en la tabla 3.2-4, y de esta tabla se deduce lo siguiente:

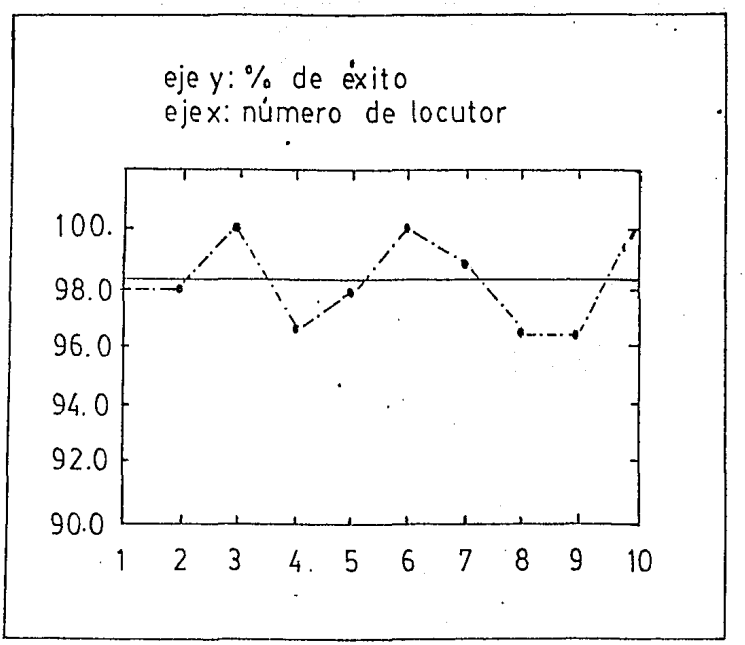


FIG. 3.2-14

TASA DE RECONOCIMIENTO DE DIFERENTES LOCUTORES.

Tabla 3.2-3
Transcripción ortográfica-fonética de los dígitos en Español.

<cero>	= [θéro]	= fricativa sorda + vocal anterior + vibrante simple + vocal posterior.
<uno>	= [úno]	= vocal nasalizada + consonante nasal + vocal posterior.
<dos>	= [dós]	= oclusiva sonora + vocal posterior + fricativa sorda.
<tres>	= [trés]	= oclusiva sorda + vibrante múltiple + vocal anterior.
<cuatro>	= [kwátro]	= oclusiva sorda + semiconsonante + vocal central + oclusiva sorda + vibrante simple + vocal posterior.
<cinco>	= [θínko]	= fricativa sorda + vocal anterior + nasal neutralizada + oclusiva sorda + vocal posterior.
<seis>	= [seis]	= fricativa sorda + vocal anterior + semivocal + fricativa sorda.
<siete>	= [sjéte]	= fricativa sorda + semiconsonante + vocal anterior + oclusiva sorda + vocal anterior.
<ocho>	= [óco]	= vocal posterior + africada sorda + vocal posterior.
<nueve>	= [nwéβ]	= consonante nasal + semiconsonante + vocal anterior + africativa sonora + vocal anterior.

Tabla 3.2-4
Resultados de la evaluación del sistema de reconocimiento de dígitos por el método cuasi-fonético.

Locutor	Aciertos	Errores	% de Exito
1. JMH	98	2	98
2. JCO	98	2	98
3. RM	100	0	100
4. FS	97	3	97
5. FGS	98	2	98
6. EC	100	0	100
7. JS	99	1	99
8. JCM	97	3	97
9. JPM	97	3	97
10. AGS	100	0	100

- La tasa total de reconocimiento obtenida en la evaluación es del 98.4 %.

- Existen tres locutores con una tasa de reconocimiento del 100 %.

- La tasa de reconocimiento inferior es del 97 %.

Para profundizar más, podemos ver la tabla 3.2-5 , en la cual representa el número de veces que un dígito articulado es identificado con cada uno de los dígitos que componen el vocabulario. De esto deducimos:

- Los dígitos 0,2, y 3 han sidos reconocidos correctamente en los 100 casos en que han sido articulados.

- El dígito 1 ha sido reconocido correctamente en 99 casos y en 1 caso ha sido confundido con el dígito 0.

3.3 Verificación e Identificación de Locutores.

El reconocimiento de personas es una de las tareas que con mayor frecuencia encontramos en los sistemas de seguridad. Tradicionalmente se resuelve este problema de una forma indirecta utilizando objetos más o menos complejo: llaves, claves numéricas, tarjetas magnéticas, etc., alcanzándose en la mayoría de los casos buenos resultados. Actuar sobre la base de una llave o tarjeta es equivalente a asociar la identidad de cada persona a la posesión del objeto. Consecuentemente, el grado de seguridad dependerá , no sólo de las características del sistema en concreto, sino también de la capacidad con que cada individuo pueda mantener la posesión del objeto base del mismo.

Encontramos, que la evolución tecnológica, especialmente de sistemas de microprocesadores, permite el diseño de sistemas más perfeccionados a menor precio, es cada vez más deseable que el reconocimiento de las personas se realice de una forma directa, basada, por lo tanto, en sus rasgos físicos, voz, huellas digitales, etc.

Anteriormente, se explicó en que consiste las áreas que involucra el reconocimiento; como reconocimiento automático de voz, el reconocimiento de locutores y la diagnosis de patología del habla. En reconocimiento de locutores, los trabajos se orientan hacia dos posibles estrategias de acción: verificación e identificación de

Tabla 3.2-5
Matriz de confusión del experimento de reconocimiento por el método cuasi-fonético.

		DIGITO ARTICULADO									
		0	1	2	3	4	5	6	7	8	9
DIGITO RECONOCIDO	0	100	1	0	0	0	3	2	0	0	2
	1	0	99	0	0	0	0	0	0	0	0
	2	0	0	100	0	0	0	0	0	0	0
	3	0	0	0	100	0	0	0	0	0	0
	4	0	0	0	0	97	0	0	0	0	0
	5	0	0	0	0	3	95	0	0	0	0
	6	0	0	0	0	0	0	97	0	0	0
	7	0	0	0	0	0	1	0	99	0	0
	8	0	0	0	0	0	0	0	0	99	0
	9	0	0	0	0	0	0	0	0	0	98
	NR	0	0	0	0	0	0	0	1	0	-
Errores individ.		0	1	0	0	3	5	3	1	1	2

NR - No Reconocido.

La verificación de locutores tiene como finalidad decidir si un locutor es o no quien dice ser; es decir, a partir de la frase pronunciada, preestablecida o no, por un locutor y la identidad presentada por él se comprueba, por comparación entre un conjunto de parámetros extraídos de la frase pronunciada y los correspondiente a la identidad presentada, que se encontrarán almacenadas en una base de datos del sistema, si el locutor es quién pretende ser. Será necesario, por tanto, una sola comparación seguida de una decisión binaria (impostor/identidad correcta). Si en lugar de presentar su identidad, el locutor simplemente pronuncia una frase, será el sistema el que tenga que decidir, cuál es su identidad comparando los parámetros de dicha frase con todos los correspondientes al colectivo de locutores identificables, siguiendo para ello algún criterio de proximidad (Punto). Por consiguiente, en este último caso, evidentemente ante un sistema de identificación de locutores que precisará realizar un número de comparaciones igual al número de locutores que lo integren.

Ahora, en términos de evaluación de un sistema de reconocimiento, bien sea de verificación o de identificación, encontramos, que hay ciertos paarámetros importantes que se debe de considerar:

1. La probabilidad de error; probabilidad de aceptación de un impostor (sistemas de verificación) o probabilidad de asociación de identidad incorrecta (sistema de identificación).

2. La probabilidad de no verificación de un locutor integrante de un sistema (sólo en sistemas de verificación).

3. La complejidad de realización.

4. El tiempo de acción necesario para la verificación o identificación.

La primera gran diferencia entre sistemas de verificación y sistemas de identificación surge al hablar en términos de probabilidad de error, pues en verificación, teóricamente, dicha probabilidad es independeiente del número de locutores, mientras que en identificación al aumentar el número de locutores, también aumenta el número de comparaciones a realizar y con él la probabilidad de error. Adicionalmente, los sistemas de verificación supondrán una complejidad de realización práctica y un tiempo de acción menor que los requeridos para el conjunto de comparaciones y decisiones necesarias en los de identificación. Por

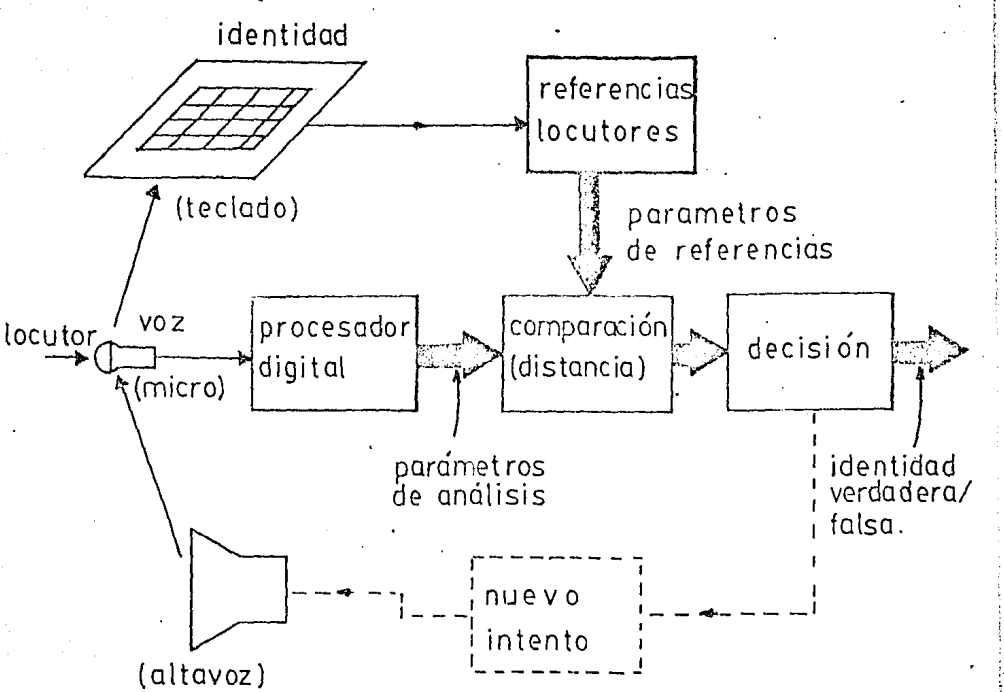


FIG. 3.3-1
SISTEMA DE VERIFICACION DE LOCUTORES.

consiguiente, los sistemas de verificación serán capaces de trabajar, teóricamente, con elevadas o bajas poblaciones de locutores solamente bajo restricciones en cuanto a memoria de almacenamiento y velocidad de acceso a los parámetros de referencia. Esta es la razón por la cual existen actualmente una fuerte tendencia hacia su desarrollo comercial; y, algunas de sus aplicaciones se centran en: autorización de transferencias bancarias y comerciales en lugares remotos, a través de canal telefónico o cualquier otro medio [31], control de accesos [32] y control de acceso a información restringida.

Después de todo esto, procederemos a analizar la estructura y característica de los sistemas de verificación de locutores. Un diagrama de bloques general de un sistema es el que se muestra en la fig. 3.3-1. El ciclo de realimentación (línea de puntos) que aparece, indica la posibilidad de requerir una nueva frase del locutor en aras de alcanzar una decisión final más precisa. Como consecuencia, no supondrá un gran incremento en cuanto a dificultad de realización, aunque si incrementará el tiempo de respuesta del sistema como contrapartida a la disminución de la probabilidad de error.

Una segunda clasificación de los sistemas de reconocimiento radica en si son dependiente de un texto o independientes. Un sistema dependiente de texto requerirá que el locutor pronuncie una o varias palabras preestablecidas en el diseño del sistema, mientras que un independiente de texto no introducirá, en principio, tal restricción. Encontramos, que aunque existe un gran número de estudios relativos a sistemas dependientes de texto, existen también algunos recientes sobre sistemas que admiten tanto dependencia como independencia de texto [33]. En general, el reconocimiento independiente de texto es más complejo ya que supone tener presente las variaciones debidas a las diferencias de texto entre la fase de prueba y la de referencia. Otra ventaja de los sistemas de texto preestablecido consiste en permitir diseños que pongan mayor enfásis en el análisis de transiciones concretas marcadas por el texto y en las cuales intervengan hábitos del habla no controlables conscientemente. Su principal desventaja es requerir una mayor cooperación por parte de los locutores ya que deberán pronunciar, en cada caso, la palabra o palabras que les marque el sistema.

Finalmente, es habitual distinguir entre dos conjuntos de sistemas de reconocimiento dependiendo de si los parámetros de análisis son invariables en el tiempo o variantes. Los primeros pueden obtenerse, bien promediando los segundos; o bien como médidas de características fijas del tramo vocal; son adecuados para los sistemas independientes del texto; requieren menor número de operaciones y memoria de alamcenamiento; y como vemos, algunos estudios [34] los muestran ventajosos en la

actualización de sistemas, necesaria para combatir la inevitables variaciones temporales en las características del habla de la mayoría de los locutores. Encontramos, que sus principales desventajas son: el no explotar los hábitos no controlables al hablar que se producen durante las transiciones de unos sonidos a otros, fuertemente dependiente de cada locutor, y al basarse en medidas promedio , ser más fácilmente susceptibles a fraude por la mimica. Los sistemas que utilicen parámetros variantes con el tiempo no presentarán las anteriores desventajas, y aun pueden clasificarse en sistemas de cálculo de parámetros de forma continua en el tiempo y sistemas que solamente calculan los parámetros en ciertos eventos temporales. Los primeros son más fáciles de realizar que los segundos pero introducen una mayor redudancia en la información que obtienen. La tabla 3.3-1, presenta un resumén de las posibles variantes, hasta aquí expuestas, para los sistemas de reconocimiento.

Como se habia dicho antes, la efectividad en el reconocimiento de la identidad de las personas en un sistema de seguridad se basa en el análisis de características propias o inherentes a cada individuo. Por lo tanto, los sistemas de reconocimiento de locutores siguen esa línea y, a partir de un primer conocimiento del mecanismo de producción de la voz (Cap. I), seleccionan un conjunto de variables o parámetros a medir, dependiente de la identidad del locutor, sobre los cuales se basan su acción.

Wolf [35] estableció como conjunto de características deseables para la elección de esos parámetros las siguientes:

1. Eficiencia para representar la parte de información de la señal de voz que depende de la identidad del locutor.

2. Facilidad de obtención.

3. Estabilidad en el tiempo.

4. Aparición frecuente y natural en la señal de voz.

5. Poca variación con el entorno.

6. Robutez frente a imitación o mimica.

Hay que hacer notar que la evaluación de la primera de estas características, la eficiencia, dependerá no sólo del conjunto de parámetros elegidos sino también de la regla de decisión utilizada, aspecto éste primordial a la hora de establecer comparaciones entre diversos esquemas de reconocimiento.

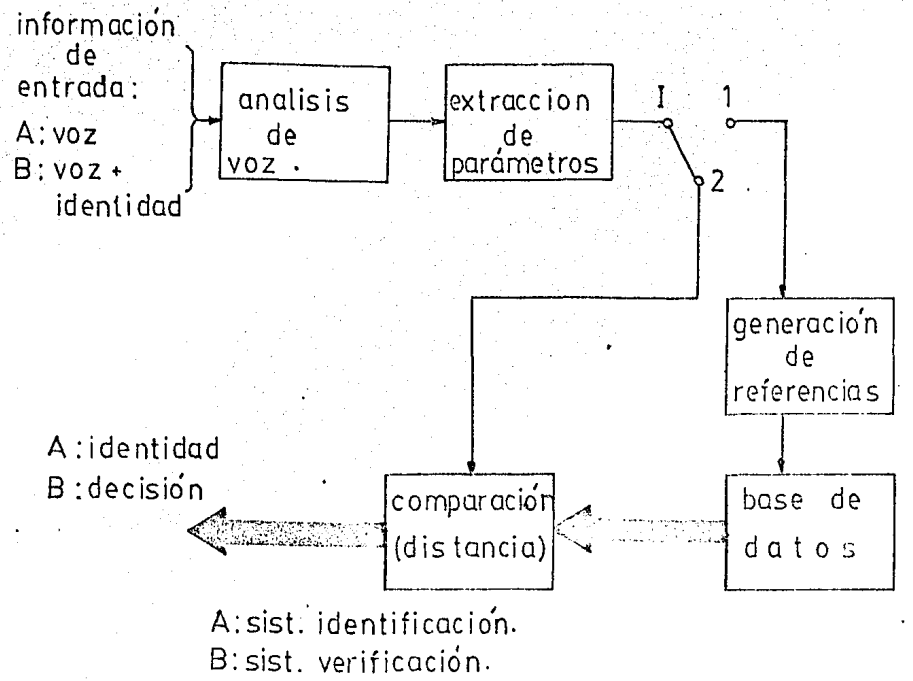


FIG. 3.3-2
SISTEMA DE RECONOCIMIENTO.

a) Sistema de verificación de locutores.

El objetivo de esta sección es entrar más en detalle en los aspectos concretos de diseño de un sistema de reconocimiento. Para ello, se presenta un diseño del sistema de verificación automática de locutores dependientes del texto, actualmente desarrollado en software en una VAX 11/750 .

Un esquema general de un sistema de reconocimiento puede ser como el que se presenta en la fig. 3.3-2. En ella se observan dos posibles situaciones según la posición que ocupe el conmutador I. En la posición 1, la finalidad del sistema será la generación de referencias y su posterior almacenamiento en la base de dato del mismo. Mientras que en la posición 2 actuará realizando realmente su misión de reconocimiento. La fase de generación de referencias es fundamental para el funcionamiento del sistema y, dada la variabilidad de las características del habla en la mayoría de los locutores con el tiempo, será preciso su actualización de forma periódica.

La primera generación de referencia requerirá la cooperación de los locutores integrantes del sistema para la grabación de un número mínimo de frases o palabras con las cuales obtener dichas referencias, mientras que para su posterior actualización se utilizará, en la mayoría de los casos, las frases resultantes de reconocimiento sastifactorios, simplificando de este modo la acción del sistema. La fig. 3.3-3, siguiendo una estructura similar a la del esquema de la figura para la posición de reconocimiento (conmutador 1 en posición 2), representa el diagrama de bloques del sistema de verificación desarrollado que se pasará a explicar seguidamente:

1. Sistema de adquisición de datos.
   Será el encargado de digitalizar la señal de la voz a la salida del micrófono. En esta etapa también se recoge la información correspondiente a la identidad del locutor (en la figura se representa por un teclado). La digitalización de la señal implica su filtrado pasabanda (300 Hz - 3 kHz) y su posterior muestreo (8 kHz) y cuantificación (12 bits); todo ello realizado emepleando el sistema de adquisición de datos de un PDP 11/23.

2. Recepción de datos.
   Los datos recogidos en la etapa anterior han de ser transmitidos y almacenados en un VAX 11/750 donde reside el sistema de verificación propiamente dicho.

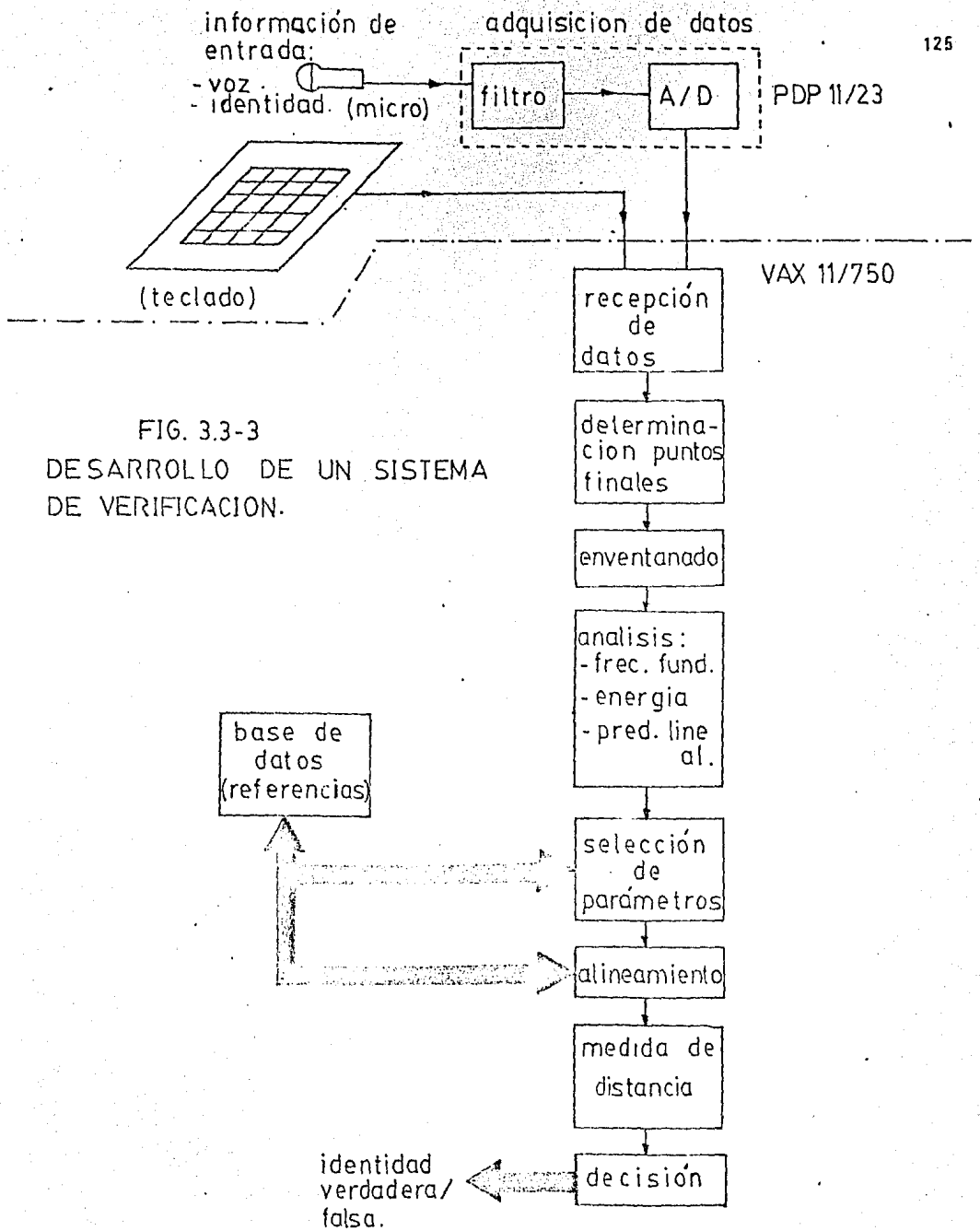


FIG. 3.3-3
DESARROLLO DE UN SISTEMA DE VERIFICACION.

3. Determinación de puntos finales.
   Antes de pasar a la extracción de parámetros de la señal de voz habrá que establecer en qué zona o zonas de la señal recibida hay realmente voz. Para ello se establecerán puntos de comienzo y final, y, en su caso (frase con varias palabras), intervalos de silencio. Los algoritmos de detección de señal de voz implantados trabajan midiendo energía y cruces por cero.

4. Enventanado.
   Anteriormente, se puso de manifiesto la necesidad de analizar no la señal de voz completa, sino enventanada (windowing) en intervalos de tiempo entre 20 y 30 ms. Este proceso se lleva a cabo utilizando una ventana de Hamming de 20 ms de duración y, adicionalmente, se mantiene una solape o superposición entre ventanas consecutivas de un 50 %. El resultado correspondiente a este bloque enventanado será, por tanto, el disponer de un número de N ventanas de señal de voz que serán analizadas una a una en la etapa siguiente.

5. Análisis.
   Para cada una de las N ventanas obtenidas se calcularán los siguientes parámetros: frecuencia fundamental, energía y 11 coeficientes de predicción lineal. La frecuencia fundamental se evaluará mediante el algoritmo SIFT y los coeficientes de predicción mediante el método de autocorrelación, ambos descritos en [36]. Como resultado a este bloque de análisis se obtendrá una representación de la frase originalmente pronunciada por el locutor formada por N vectores, cada vector formado por 13 parámetros (11 coeficientes + energía + frecuencia fundamental). Conjunto de vectores que contiene la información dependiente de la identidad del locutor sobre el cual actuará el sistema de verificación.

6. Selección de parámetros.
   Mediante transformaciones sencillas permitirá trabajar con los diferentes conjuntos de parámetros biunívocamente definidos obtenidos a partir de predicción lineal: coeficientes de predicción, coeficientes de reflexión, relaciones de área, cepstrum, etc.

7. Alineamiento referencia/prueba.
   Hasta aquí el objetivo ha sido llegar a representar la señal de voz recogida por el micrófono mediante un conjunto de vectores, que denominaremos de prueba, cuyos elementos son parámetros dependientes de la identidad del

locutor. Se tratará ahora de medir la distancia entre ese conjunto de vectores prueba y el conjunto de vectores referencia almacenados en la base de datos del sistema y correspondiente a la identidad presentada por el locutor (a través del teclado en la primera etapa). Pero antes, que un mismo locutor generalmente nunca pronuncia de igual forma la misma frase, (por ejemplo puede variar la velocidad, comenzar más rápido y terminar mas lento o viceversa). el número de vectores prueba N, no tendrá porque coincidir con el número de vectores de referencia, L (N <> L, en general). Se precisará, por tanto, previamente a la medida de distancia, un alineamiento temporal entre ambos y, que en la mayoría de los casos. supondrá realizar una transformación no lineal. todas las lineales son equivalentes para este propósito, del eje de tiempos de prueba al de referencia. (En la parte de reconocimiento global de tipo acústico se explica esto más ampliamente).

8. Medida de distancia.
Dado de que ya disponemos de dos conjuntos de L vectores, la medida de distancia entre ellos, D sera una suma de distancia entre vectores:

$$D = \sum_{i=1}^{L} d_i$$

donde, di es la distancia entre el i-ésimo vector de parámetros de prueba, X , y el i-ésimo vector de parámetros de referencia es decir

$$d_i = \text{distancia}(X_p^i, X_r^i)$$

La forma más inmediata de expresar una distancia entre vectores es a través de la distancia euclídiana:

$$d_i = (X^i - X_r^i)(X^i - X_r^i)^T$$

T - indica vector transpuesto. Sin embargo dicha distancia no es la más adecuada en el caso. pörque no toma en consideración cuál es la variabilidad de cada parámetro, ni cuál la correlación entre parámetros distintos: empleándose como alternativa [32]:

$$d_i = (X^i - X^i_\tau) W^{-1} (X^i - X^i_\tau)^T$$

. . .

Siendo W la matriz de covarianza de los parámetros: en el caso obtenido promediando valores para cada locutor y posteriormente para todos los locutores.

9. Decisión.
A partir del valor de distancia, D obtenido, se decidirá, por comparación con un umbral T. si la identidad presentada fue verdadera (D <= T) o falsa (D > T); tomándose, en este último caso. las acciones que el sistema de seguridad establezca.

La probabilidad de aceptar la identidad de un impostor como válida, o. la de rechazar la de un locutor del sistema, dependerá, en última instancia, de la elección del umbral T. Así, un valor pequeño de T hará disminuir la primera y aumentar la segunda, y uno grande lo contrario. Habitualmente, la elección de T se basa en medidas a posteriori de igualar la probabilidad de falsa verificación y falso rechazo.

3.4 Experimentos de Reconocimiento de voz con predicción lineal, filtración pasabanda, y programación dinámica.

Esta parte consiste en hacer un resumén sobre todo lo que se ha mencionado en este capitulo en base a experimentos [22] o pruebas realizadas con los algoritmos o métodos que se utilizan en el Proceso de Reconocimiento de Voz.

Se hace una comparación entre la codificación de predicción lineal [28] con la filtración pasabanda: la programación dinámica [36] con la normalización de tiempo lineal; y decodificación de una cadena de carácteres (C. S. Code [37]) con un simple espacio de filtro pasabanda de representación de la voz. Los resultados de las comparaciones se muestran más adelante en las cinco tablas. Estas tablas también reporta el efecto sobre la ejecución del sistema de 1) Vocabulario, 2) Representación parámetrica, 3) Alineación en el tiempo, 4) Código C.S., y 5) una combinación de normalización de tiempo y código C.S. También los resultados muestran la programación dinámica, el cual realiza un escalamiento de tiempo no-lineal. es extremadamente utilizado para el reconocimiento automático de voz de pronunciación de múltiples silabas.

Los sistemas de reconocimiento de voz comparados aqui tienen 3 etapas: una etapa de preprocesamiento seguido por una etapa de reducción de datos seguido por una etapa de clasificación. La etapa de preprocesamiento utiliza la codificación de predicción lineal (LPC) o la filtración pasabanda. La etapa de reducción de datos utiliza la decodificación de cadena de carácteres. La etapa de clasificación utiliza un escalamiento de tiempo lineal o escalamiento de tiempo no-lineal programación dinámica.

Similarmente, el sonido de voz es definido a ser proporcional a la distancia separando dos sonidos en un espacio vector definido por parámetros de filtros pasabanda, coeficientes predictivo lineal, o una cadena de carácteres. Los espacio de vectores de 6 y 20 dimensiones son definido por 6 y 20 canales de bancos de filtros. La codificación predictiva lineal establece un espacio de vector de dimensionalidad igual al número de coeficientes predictivos. De igual forma, mediciones con código C.S. toma lugar por medición de distancias entre puntos de referencia en vez de los puntos actuales.

Para entender el código C.S., consideremos lo siguiente: Una sucesión de puntos en un espacio vector es producido por tiempo sucesivo de muestras de la voz. Cada punto es dado en una etiqueta o marca cuasi-fonémica. Cada etiqueta es tomado desde un punto de referencia cercano o próximo (Los puntos de referencia son establecido durante una etapa de prereconocimiento "aprendiendo" durante la cual sobre 20 estados fijos o uniformes de los sonidos de voz, son decodificados como puntos de referencia). Estableciendo los puntos de referencias para el código C.S. encontraremos que es el tema de publicaciones previas [37] [38] . La sucesión de etiquetas cuasi-fonémicas producidas en esta forma son grabadas o guardadas a representar la voz y la información espectral original es descartada. Cada punto de referencia es tomado a ser una aproximación al punto espectral original. En una 'replicación', el espectro de dos expresiones que tienden a ser decodificadas en una cadena de carácteres en esta forma, las distancias son medidas entre puntos de referencias.

Las expresiones decodificadas como cadena de carácteres introduce algunas clasificaciones de errores pero también reduce grandemente el tiempo y almacenamiento necesario a procesar la voz.

Para el método de preprocesamiento de filtración pasabanda; similarmente, el sonido de voz era medido por la norma de Chebyshev (norma de valor absoluto). Esto significa sumar el valor absoluto de la diferencia de los valores coordinados entre la incógnita y la plantilla cada

10 ms. (se puede obtener [40] ligeramente, mejores resultados con la norma Euclidiana, pero no fue usado en este experimento porque era más costoso su cálculo. Un resultado similar fue obtenido por Neroth [39] quien fue comparado las normas Chebyshev y Euclidiana). Antes, de la medición de distancias, todos los filtros muestreados son normalizados dividiendo por la energía total. (Esta técnica de normalización era previamente usado por Shearme y Leach [41]).

El método de preprocesamiento LPC empleado similarmente en mediciones de sonidos de voz basado sobre Itakura en predictivo residual lineal. Una descripción detallada del predictivo residual lineal en encontrado en los apuntes de Itakura [28].

De los dos métodos de alineamiento en el tiempo, el escalamiento del tiempo lineal es el más simple. El escalamiento de tiempo lineal significa que dos representaciones de expresiones (pronunciación) a ser comparado son estiradas o comprimidas linealmente la cual son convertidas a la misma longitud. Expresiones normalizados de tiempo son también relativamente cambiados de uno a otro en orden a sobreponer el mal alineamiento esperado a la detección pobre de la iniciación y final de las expresiones. Como siempre, la traslación de tiempo no-lineal dentro del interior de una expresión causará una unión interior mal hecha. La programación dinámica, como siempre permite una traslación no-lineal, y así de este modo se logra una mejor replicación interior. También, esta requiere de más cálculo, pero significativamente mejora los resultados de reconocimiento para expresiones de múltiples sílabas.

La ecuación de programación dinámica que se utiliza es:

$$Dij = dij + \min (Dij-1, Di-1j, Di-1j-1)$$

donde dij es definido a ser la distancia entre la primera expresión a un tiempo i (time slice) y la segunda expresión a un tiempo j. Dij es la distancia total entre la primera y la segunda expresión desde su inicios o fuentes e incluyendo los tiempos i y j. La operación "min (a,b,c)" selecciona el número más pequeño desde el conjunto de números a,b, y c. Para una mayor interés o información sobre esto, se puede consultar a Itakura [28] o Sakoe y Chiba [36].

b) Condiciones experimentales.

Todo el habla utilizado en estos

experimentos fue grabado sobre una cinta analógica. Un microfóno supresor de ruido telex modelo CS-75 fue usado. Las grabaciones fueron hechas en un cuarto de laboratorio con un nivel de ruido de 65 dB (A) desde un aire acondicionado. Dos vocabularios diferentes fue grabado: el vocabulario "digito-alfábetico" fue hablado por G. White. Este contiene los nombres de las letras del alfabeto y el digito cero hasta nueve para un total de 36 palabras (el alfabeto era hablado como 'A, B, C, ...' y no 'Alfa, Bravo, Charlie, ...'). Este vocabulario fue hablado cinco veces sobre un periodo de dos días. El otro vocabulario, vocabulario de estados de Norte América, contiene 91 nombres: los 50 estados de Estados Unidos, más 10 de provincias Canadienses, más 31 de Estados Mexicanos. Estos fue pronunciado cinco veces por R. Neely en un día en cinco sesiones diferentes de grabación.

Una base de datos de voz digitalizada fue preparado desde cintas analógicas usando cualquiera de los dos métodos: LPC o análisis por bancos de filtros. Los parámetros usados a especificar la operación LPC y la operación del banco de filtros son los siguientes: El análisis LPC usa un promedio de muestreo de 10 Khz, resolución de 8 bits, un pre-enfásis de 4 dB por octava, un filtro pasabaja de 5 Khz, y una escalación de amplitud lineal. Catorce coeficientes LPC son calculados cada 12.8 ms y una ventana de Hamming de 25.6 ms fue utilizado. El análisis de banco de filtros tenia un promedio de muestreo de 100 Hz con una promediación de amplitud entre muestras, una escala de amplitud logaritmica, un pre-enfásis de 4 dB por octava, resolución de 8 bits, y cualquiera de los dos, filtros de seis un octavo o filtros de 20 un-tercio octavo. Los filtros cubren el espectro de frecuencia desde cerca de 100 Hz a 10 KHz. Los detalles del hardware usado a ejecutar estas operaciones de LPC y filtración pasabanda son dadas en el siguiente párrafo.

La voz desde una cinta analógica es pasada hacia un conjunto de filtro analógico Hewlett-Packard, módelo 8056A el cual es un conjunto de 25 filtros Chebyshev de un-tercio octavo. La formación espectral (pre-enfásis) fue obtenido por adjustación de la ganancia de cada uno de los un-tercio octavo filtros a obtener una ganancia promedio de cerca de 4 dB por octava cerca de 1000 Hz. En el experimento LPC, la voz fue filtrado pasabaja a remover todas las frecuencias cerca de 5 KHz. El filtro pasabaja fue obtenido adjustando la ganancia sobre los canales de filtro de la HP a -40 dB para las frecuencias cercanas a 5 Khz. En análisis LPC, la salida del banco de filtros es pasado directamente a un convertidor analógico/digital (ADC) de 13 bits (sólo 8 bits de resolución son actualmente usado de 13 posibles bits). En análisis de filtro pasabanda, la salida del banco de filtro es pasado a un banco de circuitos integrador-rectificador y entonces al ADC. Cada circuito integrador-rectificador rectifica el voltaje desde un canal

filtro, substrae un voltaje constante a remover los efectos de ruidos acústicos de fondo, suma los voltajes resultantes para 10 ms, y finalmente convierte el voltaje de salida desde lineal a escalas logaritmicas. El ADC entonces ejecuta la conversión analógica/digital y para los números resultantes de 8 bits a la minicomputadora. El dato digital es almacenado sobre un disco a crear un base de datos para el experimento de reconocimiento.

Las cinco repeticiones del vocabulario digito-alfabetico y del vocabulario de Estados de Norte América son usados en experimentos de reconocimiento para obtener examenes o evaluaciones de 20 vocabularios diferentes. Una primera repetición es usado a "entrenar" el reconocedor (el entrenamiento consiste simplemente de almacenamineto de un ejemplo de cada palabra en el vocabulario como un prototipo de la pronunciación). Las cuatros repeticiones que quedan o permanecen son entonces usadas a alimentar las pronunciaciones desconocidas a evaluar la exactitud del reconocedor. Las próximas diferentes repeticiones desde las 5 originales son seleccionada a generar prototipos y de nuevo las otras cuatro son usadas a evaluar el reconocedor. Este proceso de selección de una repetición desde cinco y evaluación sobre las cuatro que permanece eran repetidas cinco veces así, de esta forma se crea 5x4 examenes de vocabulario. Para las 36 palabras del vocabulario digito-alfabetico, esto resultaba en 720 examenes de pronunciación individual. De igual forma, 1820 examenes de pronunciación eran obtenido desde el vocabulario de Estados de Norte América.

c) Resultados.

En las cuatro tablas más adelante, cada 'porcentaje correcto' es basado sobre 1820 o 720 pronunciaciones dependiendo sobre cual vocabulario es usado. El vocabulario de Estados de Norte América tiene 1820 examenes de pronunciación y el digito-alfabético tiene 720 examenes de pronunciaciones.

Para el primer examén o prueba en la tabla 3.4-1, el preprocesamiento fue hecho por un banco de filtros de 20 canales, con 1/3 octavo de ancho de banda por canal; el alineamiento en el tiempo fue hecho por programación dinámica; el porcentaje correcto fue de 98 % para el vocabulario digito-alfabético y fue de 99.6 % para el vocabulario de Estados de Norte América. En la segunda prueba, lo cual difiere de la primera, sólo en la estrategia de alineamiento en el tiempo, la normalización en el 'tiempo lineal' fue logrado como sigue: Todas las pronunciaciones fueron extendidas o comprimidas linealmente a la misma longitud, es decir, 50 unidades de tiempo de longitud. Las pronunciaciones de referencias (prototipos) son comparados a las pronunciaciones desconocidas sumando

Tabla 3.4-1
Efectos del vocabulario.

Examen (prueba)	Método de preproce-samiento	Alineamiento en el tiempo	Vocabulario digito-alfa betico (% correcto)	Vocabula-rio de Es tados de Norte Amé rica %
1	20 canales	programación dinámica	98	99.6
2	20 canales	traslación lineal	98	90

El vocabulario dígito-alfabético tiene 36 palabras, en su mayor parte monosílabos. La lista de palabras de Estados de Norte América tiene 91 palabras, la mayor parte polisílabo. La primera prueba utiliza 20 canales de filtros pasabanda y programación dinámica, y encontramos resultados de reconocimiento de 98 % correcto para los Estados de Norte América. Mientras que el vocabulario sea más grande, éste nos proporcionará significativamente un promedio más alto. Pero en la segunda prueba nos da significativamente un promedio bajo. La programación dinámica no es utilizada en la segunda prueba. La traslación del tiempo lineal es utilizado. De este modo el vocabulario utilizado tiene un efecto profundo sobre la ventaja aparente de diferentes estrategias en alineamiento en el tiempo.

las 50 distancias entre sonidos de referencias y desconocidas en la misma unidades de tiempo. Entonces, las pronunciaciones desconocidas son cambiados linealmente de derecha e izquierda relativo al prototipo y la distancia total recalculada. La pequeñisima 'distancia total' es asumido a ser el resultado del alineamiento de tiempo propio y tomado como la propia medida de pronunciación y similaridad del prototipo.

Un experimento utilizando LPC recientemente reportado por Itakura [20] muestra un promedio de error de 11.4 % sobre el vocabulario dígito-alfabético la cual es comparado con el promedio de error de 3 % de White y Neely para el mismo vocabulario (tabla 3.4-2).

¿ Pero por qué la diferencia de resultados entre Itakura y el de White y Neely ? Primero, la técnica de LPC fue empleada entonces esto constituia una replicación del sistema de Itakura. Luego, esta diferencia de resultados eleva primariamente desde la diferencia de ancho de banda para entrada de voz. Itakura utiliza esencialmente 6 coeficientes LPC y un promedio de muestreo de 6 KHz para su reconocimiento de entrada de voz sobre una línea telefónica de 3.0 KHz. El sistema utilizado [22] por White y Neely utiliza 14 coeficientes LPC lo cual la entrada de voz va hacia un micrófono de supresión de ruido de 8 KHz en conexión con un promedio de muestreo de 10 KHz. También, ellos utilizarón un banco de filtros de 1/3 octavo para la forma espectral a preenfatizar las frecuencias altas. Otro es que utilizarón una estrategia de programación dinámica ligeramente diferente. Los investigadores creen que las diferencias en resultados subleva esencialmente porque el sistema de White y Neely responde mejor a las altas frecuencias necesario a distinguir algunas de las palabras confusas por el reconocedor de Itakura.

El más importante resultado del experimento (tabla 3.4-3) en el tiempo es una indicación de potencia extraordinaria de reconecimiento de programación dinámica sosbre pronunciaciones de múltiples silabas. Un segundo resultado es la normalización de tiempo lineal (con una traslación de izquierda y derecha) es como bueno como la programación dinámica sobre pronunciaciones de múltiples silabas.

El resultado más interesante mostrado en la tabla 3.4-4 es el código C.S. la cual parece reducir la ventaja de utilización de la programación dinámica relativo a la normalización de tiempo lineal.

d) Conclusiones.

La base de datos utilizado en estos experimentos es relativamente pequeña. Las conclusiones

Tabla 3.4-2
Efectos de representación parámetrica.

Método de pre-procesamiento	Método de alineamiento en el tiempo	Vocabulario dígito-alfabético (% correcto)	Tiempo de reconocimiento por expresión (s)	Promedio de datos aproximados (Bits /s)
20 canales	programación dinámica	98	30	12000
LPC	programacion dinámica	97	20	4200
6 canales	programación dinámica	96	15	3600
Código C.S.	programación dinámica	91	2	500

El preprocesamiento produce cuatro representaciones parámetricas diferentes lo cual son arregladas en orden de incremento de compresión de datos (promedio de bit más bajo). La exactitud de reconocimiento disminuye como la comprensión aumenta o va hacia arriba. Los resultados similares LPC y filtración pasabanda muestra que ellos son esencialmente equivalente.

Tabla 3.4-3
Efectos en el alineamiento en el tiempo.

Método de preprocesamiento	Método de alineamiento en el tiempo	Vocabulario	Exactitud de reconocimiento (% correcto)
20 canales	programación dinámica	Norte América	99.6
20 canales	traslación lineal	Norte América	90
20 canales	programación dinámica	Digito-alfabético	98
20 canales	traslación lineal	Digito-alfabético	98

Los métodos de alineamiento en el tiempo incluye traslación de tiempo lineal y programación dinámica. La programación dinámica produce una curvatura de tiempo no-lineal para lograr la mejor réplica entre la plantilla y las expresiones desconocidas. Note que la programación dinámica logra un 99.6 % correcto sobre las palabras de múltiples sílabas de la lista de palabras de Estados de Norte América mientras que la normalización de tiempo lineal logra sólo 90 %. Note como siempre, que no hay ventaja para palabras monosilabos del alfabeto más dígitos.

Tabla 3.4-4
Iteracción del Código C.S. con alineamiento en el tiempo.

Método de preprocesamiento	Método de alineamiento en el tiempo	Vocabulario	Exactitud de reconocimiento (% correcto)
20 canales	tiempo lineal	Dígito-alfabetico	94
20 canales	programación dinámica	Dígito-alfabetico	98
Código C.S.	tiempo lineal	Dígito-alfabético	89
Código C.S.	programación dinámica	Dígito-alfabético	91

Viendo desde alineamiento de tiempo lineal a programación dinámica para representación de 20 canales, hay un incremento desde el 94 a 98 %. Con el código C.S., la exactitud sólo va desde 89 a 91 %. Nos parecerá que el código C.S. reduce la ventaja de la programación dinámica sobre la normalización de tiempo lineal. Notamos que la 'traslación lineal' y 'tiempo lineal' difiere en que 'traslación lineal' trata diferentes alineamientos 'tiempo lineal' y preserva el mejor de todos.

presentadas por los autores proporcionan aproximadamente bases equivalentes para mediciones de la forma de onda de la voz. En otras palabras, mediciones Chebyshev o Euclidianas en un espacio de filtro pasabanda es aproximadamente equivalente al promedio logaritmico lineal predictivo residual en espacio LPC cuando los parámetros tipicos LPC y banco de filtros son usados. El error realizado por ambos LPC y sistemas de reconocimientos de bancos de filtros son casi lo mismos. Todos estos errores similares elevan la confusión entre /b/ y /d/ y entre /m/ y /n/. Estos experimentos no están concentrados sobre sonidos nasales para lo cual las diferencias son más encontradas entre progresos LPC y filtración pasabanda. Pero con esta excepción, estos trabajos constituye una seguridad en la potencia similar de estos dos progresos.

Una segunda conclusión es la popular técnica de reducción de datos que tal vez sea perjudicial o dañino a la ejecución de sistemas de reconocimiento. La tabla 3.4-2 muestra técnicas de clases de preprocesamiento de acuerdo al grado de comprensión de datos. Encontramos, que es evidente una correlación fuerte a una exactitud de reconocimiento. También, se demostró de la importancia del vocabulario en la evaluación de sistemas de reconocimiento.

¿ Por qué encontramos que el promedio de reconocimiento japonés tiende a ser más alto que el promedio americano ? Originalmente, se escogió el vocabulario de Estados de Norte Americanos en orden a evaluar esta idea. Los resultados [22] dan soporte a la idea de que la diferencia en promedios o resultados tal vez sea debido al lenguaje y diferencias del vocabulario. Itakura tiene reportado un resultado muy bueno, 97.3 % correcto sobre 200 nombrs geográficos japoneses. Cada nombre tiene un promedio de 3.5 sílabas. Entonces, para el vocabulario digito-alfabético reporta sólo el 88.6 % correcto. Todo esto nos da una suposición de que las diferencias en resultados son grandemente debido a las diferencias de vocabulario.

IV - Perspectivas en el futuro de dispositivos y Sistemas de Entrada de Voz.

Este capítulo se refiere a presentar las ventajas y desventajas que tiene los sistemas de entrada de voz, o como lo queramos llamar el Reconocimiento de voz; como también exponer algunas aplicaciones que tienen estos sitemas hoy en día, como: en la industria, en los negocios y en lo militar, en áreas como el manejo de materiales, manufacturación, control de calidad, producción, inventario y transacciones telefónicas automáticas y claro muchas aplicaciones más. También, este capítulo consiste en presentar los diferentes sistemas y circuitos integrados (chips) de Reconocimiento de voz que existen hoy en día en el mercado así como también la compañía que lo fabrica. Más adelante, también se explicará como cada día la tecnología de circuitos integrados tiende a ser fabricados con una tecnología mayor y con una integración bastantemente grande.

Nos preguntamos, ¿ por qué, los sistemas de reconocimiento de voz tiende hacia un futuro promisorio ? La respuesta es que la voz es la forma más natural que tiene el hombre para comunicarse, recibimos y transmitimos información. Ejemplo, de este avance podemos verlo en un manufacturador de Estados Unidos la cual demanda tener instalado unos sistemas de entrada de voz de más de 600 palabras aisladas (instrucciones), esto costando muchos miles de doláres cada uno. El volumén de estos son áreas donde es incoveniente la entrada de datos por medio de un teclado. Otro ejemplo, una compañía británica, lo cual un reconocedor con independencia del locutor para acceso telefónico, demanda más de 250,000 acceso de usuarios por computadora y redes telefónicas hacia el sistema. La razón es la centralización lo cual involucra lo costoso del sistema: £ 30,000 y podríamos continuar una lista amplia de usos o beneficios que tienen estos sistemas, por ejemplo: un beneficio que considero yo importante es que constituye un auxilio o una ayuda importante a personas impedidas o podríamos decir que carecen de brazos. Algunas ventajas y desventajas de estos sitemas son:

VENTAJAS	DESVENTAJAS
1. Fácil acceso por vía telefónica.	No es posible una conversación normal.
2. Libera las manos para otras tareas.	Los operadores necesitan de de una disciplina mientras que estén hablando.
3. Requerimientos de entrenamiento mínimo.	

4. Utilizado en un medio obscuro (más eficiente para entradas de control y datos).

4.1 Aplicaciones y Factores Humanos en Sistemas de Entrada de Voz.

Encontraremos, que la reacción del usuario a un sistema de entrada de voz es particularmente importante cuando el sistema es interactivo. El sistema debe de ser programado para permitir toda complejidad innecesaria y proporcionar como es natural una forma de comunicación con la máquina como sea posible.

Algunos de los factores el cual influye al usuario y causa una reacción a su personalidad, son discutidas a continuación:

a) Capacidad de realizar múltiples tareas.

Los sistemas de SARV (Sistemas Automáticos de Reconocimiento de Voz) son únicos en su habilidad de obtener datos o comandos desde un usuario cuya manos u ojos son ocupados generalmente en otras funciones. La sola alternativa en esta situación es tener los datos registrados por cinta o por un oyente, o a requerir al usuario interrumpir sus otras actividades para el propósito de entrada de datos.

Desde que es posible realizar entrada de voz simultáneamente con otras actividades, los sistemas de SARV pueden ser empleado para entrada de datos por personas que normalmente no se encuentran entrenada para estos propósitos. Los datos pueden ser verbalmente entrados a la fuente con operaciones no intermedias requeridas. La colección de datos en una fábrica es un buen ejemplo, donde la entrada de voz puede ser directamente realizado a la fuente. Desde que la entrada de datos no es usualmente su función primaria, los trabajadores de la fábrica deben sentir que los sistemas SARV proporciona beneficios de los trabajos (enriquecimiento), mejora la productividad, o no obstaculiza con sus actividades normales.

Las experiencia con operaciones de múltiples tareas las cuales comprenden entrada de voz como una de las funciones tienden a mostrar que un operador realizará un promedio de entrada de datos elevado después de 3-4 meses [12]. Después de esta cantidad de experiencia, el operador conoce el promedio propio de locución, cuando verifique los datos, y que palabras serán usadas en un instante

particular de tiempo.

b) Movilidad del operador.

Por medio de un transmisor inalámbrico, un operador puede tener completa libertad de movimiento y comunicarse tranquilamente con su sistema SARV. Un transmisor del tamaño de un paquete de cigarrillos puede proporcionar un considerable rango de operación. Como simpre, en tales casos, una retroalimentación al operador debe de ser realizado por respuesta de voz o una pantalla portátil desde que generalmente se encuentra fuera del rango visual de una pantalla fija.

c) Flexibilidad del vocabulario.

Los sistemas de SARV adaptivo es entrenado para la caracterización del habla o voz del usuario. Generalmente, esto es realizado por muestras dadas de cada palabra en el vocabulario como un conjunto de datos de referencia con el cual se comparan las futuras pronunciaciones o expresiones. Estos datos de referencia es usualmente almacenado para un uso futuro que sea necesario un re-entrenamiento para que no sea requerido cada vez que el individuo use el sistema SARV. ¿ Por qué de la característica adaptiva ? La razón es para que un usuario tenga libertad de escoger la palabra más natural de él para una función particular. Por ejemplo, la palabra escogida a verificar la exactitud de un bloque de datos que tiene que ser hablado y desplegado al usuario podrá ser cualquiera de los dos 'siga' o 'bien'. La característica adaptiva también hace posible para sistemas SARV de vocabulario limitado operar para cualquier lenguaje.

d) Retroalimentación, edición e interacción.

La retroalimentación inmediata debe ser dado al usuario del sistema de entrada de voz, cualquuiera de los dos, visualmente o por medio auditivo o ambos. La retroalimentación debe de ser inequívoco o claro y puede grandemente asistir al usuario en la realización de sus funciones de entrada de voz.

En un sistema de reconocimiento de palabras aisladas, es importante medir y mantener del usuario los espaciamientos mínimos para que las palabras no sea pronunciadas juntas (Cap. III). Esto puede ser realizado por un indicador visual o un tono audible 'leyendo'. Un usuario con experiencia de sistema de entrada de voz con palabras aisladas aprenderá rapidamente el promedio en la cual las palabras pueden ser habladas, después del cual sera necesario el indicador 'leyendo'. Pero, desde la etapa

inicial de un sistema usando entrada de voz, el indicador 'leyendo' encontramos que es un entrenamiento de mucho valor y de ayuda para el operador.

Un indicador de 'rechazo' similar al indicador de 'leyendo' puede también ser útil como cuando el operador balbucea o murmura. La indicación de 'rechazo' también puede servir al propósito de subsconcientemente entrenar al operador para hallar las palabras del vocabulario usado en una manera que pueda ser más fácilmente reconocida por el sistema SARV.

Además de las indicaciones elementales 'leyendo' y 'rechazo', todos los comandos de locución debe ser retroalimentado al operador para su verificación. Esta verificación puede tomar la forma de una indicación positiva de correción hacia la palabra de control tal como 'bien' hablada después de cada comando o de cada campo de datos, o puede simplemente ser indicado por procedimiento al proximo comando.

Las palabras de control como 'borrar', elimina el último comando y 'cancela', borra y entra un bloque de datos podrá también ser proporcionado.

La retroalimentación al operador no puede ser sólo ser usado para verificación, sino que también para indicación del usuario a través de una secuencia de entrada; revisión de sintáxis, formatos y valores esperados y preguntas especiales hechas cuando tal aplicación es requerida. En otras palabras, lo que se quiere decir, es que la entrada de voz puede ser usado para aplicaciones de requerimientos de terminales inteligentes.

e) Exactitud de reconocimiento.

El promedio de error en un sistema práctico de SARV debe ser suficientemente bajo para eliminar cualquier pérdida de confianza o eficiencia del operador. Los humanos tenemos tendencia a convertirnos insensibles o no preocuparnos a los promedios de error muy bajo en operaciones o tareas múltiples. Las correciones por voz no debe de ser suficientemente frecuente para no ser impedimento u obstáculo a la realización del trabajo especifico (propósito). Si el promedio de error es suficientemente alto que interfiere notablemente con el trabajo, el operador perderá confianza y no deseará usar el sistema de entrada de voz. En un sentido el operador hara una decisión binaria, por ejemplo, el sistema de entrada de voz, es uno u otro, 'bueno' o 'malo'. Algunas entrevistas [12] con usuarios de sistemas de entrada de voz operacional tienen a mostrar que raramente en un sistema de entrada de voz aceptado, por lo menos el promedio de error es muy bajo. Un promedio de error aceptable es criticamente

dependiente sobre la aplicación particular y el promedio de entrada de datos. Un promedio de entrada de datos de voz alto es cerca de 50 palabras o frases por minuto requerido como promedio de error bajo, en la cual estas aplicaciones donde el promedio de dato es suficientemente lento el usuario tiene tiempo de realizar correciones.

f) Estabilidad de datos de referencia.

Como mencionado previamente, un adaptivo, un sistema con vocabulario limitado realiza un procesamiento de reconocimiento por comparación de expresiones no conocidas con un conjunto de muestraas almacenadas del vocabulario de palabras obtenido desde el usuario del sistema. Estos datos de referencia deben ser estable sobre un periodo largo de tiempo para aplicaciones prácticas. Una vez que los datos de referencia tienen a ser obtenido, el operador será capaz de usar el sistema de entrada de voz con poco o nada de 're-entrenamiento'; el operador no tendrá que interrumpir frecuentemente sus operaciones normales para reentrenar palabras individuales durante el curso de sus operaciones; de ahí que viene una de las importancias de este punto.

4.2 Perspectivas de las arquitecturas digitales y sistemas de Reconocimiento de Voz.

Generalmente, los requisitos para una aplicación de reconocimiento de voz y análisis son:

- Muestreo (fm = 10 KHz).
- Filtrado de la señal (por ejemplo, filtración pasabanda de 5 KHz o 3.3 KHz en aplicaciones telefónicas).
- Conversión A/D.
- Preacentuación.
- Cálculo de los coeficientes de autocorrelación y de los coeficientes LPC.
- Cálculo de la distancia entre la mátriz de parámetros obtenida y los patrones de referencia.

En conjunto se observa que la limitación fundamental, cuando se quiere conseguir un funcionamiento en tiempo real, surge de los procesos de cálculo, debido al elevado número de multiplicaciones y sumas que deben realizarse. En este caso, debemos de ser capaz de realizar una multiplicación cada 1 us aproximadamente para el procesamiento de la señal dada.

La forma de resolver este problema, de realizar

demasiados cálculos lo podemos simplificar ya sea empleando o realizando un circuito integrado (IC) específico para la aplicación, o bien plantearse arquitecturas digitales programables más versátiles. Este último consiste en concebir una estructura en la que un procesador específico se encargue de realizar la parte más crítica (operaciones a alta velocidad) y otro de propósito general, que no posee la velocidad suficiente para realizar tales cálculos que controle el funcionamiento del procesador específico y el resto de operaciones.

Los procesadores de señales programables son concebidos como procesadores específicos para señales, las cuales han sido construidos con unas arquitecturas adecuadas para la resolución de operaciones del tipo multiplicación y suma a alta velocidad. En la tabla 4.2-1 se presentan algunas arquitecturas o dispositivos como el 2920 de INTEL en que se ofrece todo el hardware, incluido los convertidores A/D y D/A, pero cuya velocidad de cálculo, resolución y capacidad de secuenciamiento no son adecuados para la aplicación; y también procesadores específicos programables, tabla 4.2-2, en los que se obtiene mayor velocidad, pero no disponen de la interfase necesaria con el mundo analógico de forma integrada en el chip. Una última posibilidad abierta al diseñador es la de utilizar un IC's (bipolares y con tecnología MOS), disponibles comercialmente (multiplicadores, ALU, secuenciadores, memorias, etc.), para construir un procesador digital de señales a medida.

Algunas de las compañías que fabrican o producen sistemas o arquitecturas de reconocimiento de voz o también entrada de voz son a continuación [54] :

a) BELL LABORATORIES.

Los Laboratorios Bell han sidos muy activos en investigaciones de reconocimiento por mucho tiempo, hasta recientemente, con muy poco que mostrar como resultados. Sin embargo, ellos han desarrollado , un procesador de señal rápido el cual puede ser usado no sólo para el sintésis del habla pero también en el extremo frontal de un reconocedor de palabras , y un elemento de procesamiento sistólico monolítico para la clasificación de expresiones durante el reconocimiento [55].

b) INTERSTATE ELECTRONICS.

Es uno de los más grande proveedores de sistemas de reconocimiento de palabras aisladas, Interstate ofrece ambos, el chip sencillo VRC008, sistema independiente del locutor con vocabulario de 16 palabras y dos conjuntos de chips (el VRC100-1), el cual es un sistema

Tabla 4.2-1.

	Intel 2920	*AMI S2811*	*NEC µPD7720*	*TI TMS320*	*Intermetall MAA1060*	*Fujitsu MB8764*	*Hitachi HD61810*	*Toshiba 1535F 7*	*ETSITM P5M*
Instrucciones	21	40	48	60					4
Ciclo de instruccion (ns)	600	300	250	200	400	100	250	250	200
Palabra de instruccion (bits)	24	17	23	16	26	24	22	16	32
Programacion	EPROM	Mascara	Mase Eprom	Mascara	Mascara	Mascara	Mascara	Mascara	RAM
Memoria externa	No	Si	No	4K · 16	No	Si	No	Si	
RAM datos	40 · 25	128 · 16	128 · 16	144 · 16	32 · 8 128 · 24	256 · 16	200 · 16	128 · 16	1K · 16
ROM datos		128 · 16	512 · 13		128 · 8		128 · 16	512 · 16	RAM 1K · 16
ROM instrucciones	EPROM 192 · 24	256 · 17	512 · 23	1536 · 16	512 · 26	1K · 24	512 · 22	512 · 16	4K · 32
Tamaño multiplicador (bits)	No	12 · 12 · 16	16 · 16 · 31	16 · 16 · 31	16 · 8 · 24	16 · 16 · 26	12 · 12 · 16	16 · 16 · 31	16 · 16 · 32
Tiempo multiplicacion (ns)	4800	300	250	200	250	100	250	250	100
Tamaño acumulador (bits)	28	16	16	32	24	16	16 (coma fija flotante)	16	32
E S serie	No	Si	Si	No	Si				No
E S paralelo (bits)		8	8	16	16				16
E S analogicas	4 8								
Tecnologia	NMOS	VMOS	HMOS	NVOS		CMOS	CMOS	CMOS	C MSI LSI
Observaciones	, Existe 2921 con ciclo de 400ns	Existe 2921 2 Posibi Memoria externa RAM Disp	Existe 77220 CMOS Multiplicador 24 · 24 Ciclo 1 Memoria				Seleccion automatico coma fija flotante	Posibi memoria externa para datos y programa	

Tabla 4.2-2.

Manufacturer	Part no	Technology	Type	Bit rate (bps	On-board ROM?	External ROM (up to)	Audio amp reqd?	Operating voltage	Power dissipation Typical operating mW	Pins	Enquiry card no or source in [ ]*
American Microsystems	53610 53620	CMOS CMOS	LPC-10 LPC-10	~1.2k ~1.4k	20k No	? 128k	No } 30mW into No } 100Ω	~6V ~6V		24 22	421
General Instruments	SP-0250 SP-0256	NMOS NMOS	LPC-12 LPC-12	1.6-2.4 1.6-2.4	No 16k	? 491k	Yes Yes	~5V 4.5-7V		28 28	431
Hitachi	HD 61885 HD 38680	CMOS PMOS	PARCOR-10 PARCOR-10	1.25-9.9k 2.4-9.6k	32k No	128k 128k+	Yes Yes	3.6-5.5V 3.6-5.5V	2.5 350	28 or 40 28	434
ITT Semiconductors	UAA 1003	NMOS	Waveform Compression	<1.5k	27k	?	Yes	~6V		24	437 [5]
Matsushita (Panasonic)	MN 6401	NMOS	PARCOR-10	1.2-5.2k	32k	?	Yes	~5V		28	442
National Semiconductor	MM 54104	NMOS	Waveform Compression	~1k	No	128k	Yes	7-11V		40	445
Nippon Electric Company		NMOS CMOS	VSRSSS Format	700 1.2-5.6k	64k 32k		No (200m W into 30Ω) Yes	4.5-6.0V ~5V		28	[6] [7]
Sharp		CMOS	ADM		32k	128k		2.7-5.5V		48	[8]
Telesensory Systems	CRC	PMOS	Formant	<1k	No	32k+	Yes	~15V	100	40	456
Texas Instruments	TMS 5100 TMS 5200	PMOS PMOS	LPC-10 LPC-10	>1.2k >1.2k	No No	? ?	Yes Yes	~9V ~5V	180 60	28 28	457
Toshiba		CMOS CMOS	PARCOR ADM	1.2-9.6k 10-32k	No No	6m ?	Yes Yes	3-7V 3-7V	2.5 1.5	42+	[9]
Triangle Digital Services	TDS 90	NMOS	Waveform Compression	1.8-7k	No	?	No	~5V	125	40	460
Votrax	SC-01	CMOS	Phoneme	70-100	—	—	Yes	7-14	40	22	464

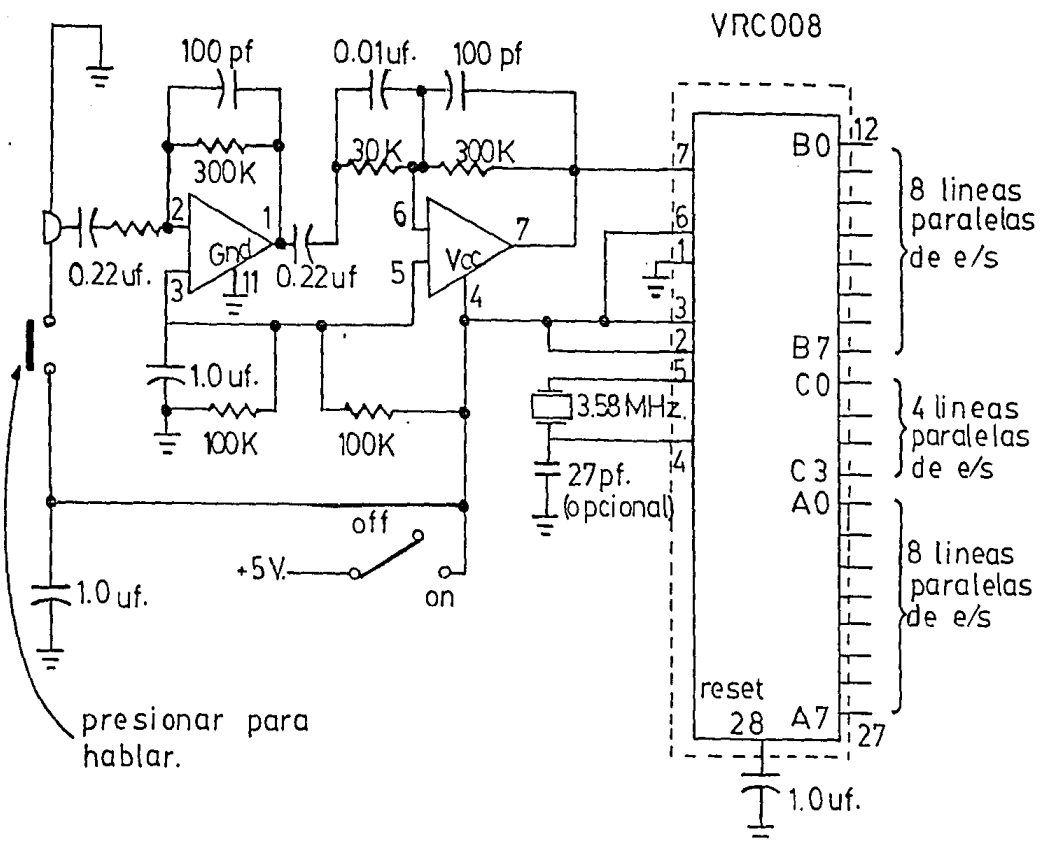


FIG. 4.2-1

EL VRC008 CON CIRCUITOS EXTERNOS PARA LA REALIZACION DE UN SISTEMA DE RECONOCIMIENTO CON INDEPENDENCIA DEL LOCUTOR.

dependiente del locutor con vocabulario de 100 palabras.

Las 28 conexiones (pins) del VRC008 es disponible en ambas versiones de NMOS y CMOS, y consume tipicamente 350 mW (NMOS) o 20 mW (CMOS) desde una alimentación sencilla de 5 Volts. Una pequeña circuiteria externa es requerida para el sistema trabajando, como lo podemos ver en la fig. 4.2-1.

El VRC100-1 consiste de dos IC's designado 100-1A y el 100-1B. El 100-1A de 28 conexiones utiliza NMOS tecnología de capacitor-conmutado con 80 OP AMP (amplificadores operacionales) para proporcionar un análisis de espectro de audio de la máxima entrada de 5 V. rms desde una fuente de impedancia de baja salida. Esto consiste de 16 filtros pasabanda, cada uno seguido por un rectificador de media onda y un filtro pasabaja de segundo orden con 25 Hz de corte. El chip también incluye un multiplexador analógico y decodificador de 16 canales.

El 100-1B es un controlador/reconocedor de 40 pins, conteniendo el algoritmo completo para reconocimiento de palabra aislada incluyendo: detección del contorno de la palabra, normalización de amplitud, comprensión de la palabra y sintaxis vocabulario programable. El vocabulario es almacenado en un ROM externo.

Interstate pretende una exactitud de reconocimiento mejor de 99 %. El tiempo mínimo entre palabras es 160 ms.

La compañía también ofrece un rango de tableros sencillo de módulos de reconocimiento y terminales de reconocimineto. El módelo VRM400 tiene un vocabulario de 100 palabras y es disponible con un manejador de software incluido sobre un disco flexible de 8" y es compatible con el sistema de software DEC utilizando RT-11. Esto con interfase a un bus de datos (Q-BUS) LSI-11. El módelo VRM041 reconoce hasta 40 palabras y tiene un RS232C o interfase entrada/salida asincrono de 20 mA. El VRM102 es un tablero similar pero tiene capacidad de 100 palabras y dos interfase RS232C serial o interfases 20 mA.

La terminal VRT101 combina el microprocesador Z80 y 48Kbytes de memoria con una pantalla o monitor de 80 carácteres por 25 líneas, teclado, unidades de disco flexible y un módulo de reconocimiento de 100 palabras.

c) NIPPON ELECTRIC.

Para los dos últimos años, NEC ha estado ofreciendo un reconocedor de voz conectado con dependencia del locutor, el DP100, pero esto ya no se esta ofreciendo actualmente en el mercado. Ellos dicen, como siempre, que un nuevo módelo, el DP200, esencialmente el mismo tipo de

máquina pero con una ejecución mucho mejor y más barato. Los DP200 son alrededormente en uso limitado en los EU, donde un costo de $22,000.00 es cotizado.

d) THRESHOLD TECHNOLOGY.

Teniendo producido la primera unidad de reconocimiento de voz hace 14 años, Threshold ahora demanda haber vendido más que 600 sistemas principalmente para aplicaciones de entrada de voz.

En el rango de compañías son el T-500/580, terminales de entrada de datos por voz el cual son sistemas dependientes del locutor con palabras aisladas permitiendo el control de procesos entrenados por locuteres, operadores atentos o dispuestos, almacenamiento de conjuntos de plantillas y la interpretación de los códigos de salida de la palabra para ser ejecutado por la comutadora. El vocabulario de 60 palabras puede ser modularmente incrementado a 340 palabras o frases. La caracteristica de terminales Threshold's "Quiktalk" lo cual reduce la longitud de las pausas requerida entre palabras para permitir un promedio de entrada de hasta 180 palabras por minuto.

AURICLE, una subsidiaria de la Threshold technology, produce el Auricle-1, sistema de reconocimiento de tablero sencillo lo cual reconoce hasta 80 palabras, y es de propósito para integración dentro del equipo de terminal a darnos una capacidad de entrada de voz, o como un periferico de computadora en sus propios derechos.

e) VERBEX.

El módelo Verbex 1800, sistema de recuperación y entrada de datos por telefóno es un independiente del locutor, palabras aisladas, sistema multicanal con respuesta de voz. El sistema puede acomodar hasta 8 usuarios simultáneamente y usar respuesta de voz para alistar al llamador. El módelo 1800 trabaja casi todos los dialectos americanos sobre líneas telefónicas seleccionada aleatoriamente.

El vocabulario de reconocimiento mínimo consiste de 10 dígitos, de cero a nueve y las palabras 'si' y 'no', pero este vocabulario puede ser expandido hasta 50 palabras.

El vocabulario de respuesta básica, incluye hasta 32 palabras o 16 segmentos de voz y puede ser expandido hasta 512 palabras, o 256 segundos de voz.

Ambos, el reconocimiento y vocabulario de respuesta rápida pueden ser adquirido como para aplicaciones

individuales.

La compañía también produce el módelo 1800-CSRS, un dependiente de locutor de voz continua, sistema de entrada por micrófono por canal sencillo, el cual proporciona una entrada rápida para una entrada de cadena de carácteres y 10 comandos de palabras aisladas.

f) VOICETEK.

COGNIVOX es una clase de reconcocimiento de voz y salida de voz para computadoras personales, tal como Rockwell AIM-65 (el Cognivox V10-1001, fue descontinuado en Junio de 1982), Pet (V10-1002 y V10-432), Apple II (V10-1003), Exidy Sorcerer (V10-132), sistema basado Z80 (V10-232) y TRS-80 (V10-332). Esto es completado con bocinas incorporadas, micrófonos y conexiones de puerto E/S en paralelo.

El reconocedor de voz es de palabras aisladas, dependiente del locutor con un vocabulario de 32 palabras, y un mínimo de 150 ms entre palabras. No hay restricción sobre el vocabulario para entrada o salida, el usuario entrena el sistema de reconocimiento con las palabras que usará y registra la salida del vocabulario de la misma forma. Los dos son completamente independiente. Los requerimientos de memoria son cerca de 4Kbytes para programa y tablas, y 1.5 Kbytes por segundo de voz para el almacenamiento del vocabulario de respuesta registrada digitalmente.

g) VOTAN.

Las series VX VOTAN consiste de 3 sistemas. El V1000 que es un sistema de reconocimiento de palabras aisladas con dependencia del locutor. El V4000, la cual es un sistema de salida de voz con la habilidad de digitalizar, comprimir, almacenar y regresar la voz inmediatamente o después. El V5000 combina la característica de otros dos sistemas.

El método de codificación tiene una variable de promedio de bit, lo cual puede ser como alto, 14,400 bps o bien como bajo, 4,800 bps.

h) GENERAL INSTRUMENTS.

En cuanto a reconocimiento de voz, la General Instruments produce el chip SP1000, de 28 pins, dispositivo de reconocimiento/síntesis de voz, para lo primero, realiza una extracción de característica por LPC sobre una señal de audio. Este dispositivo es diseñado para interfase con un

bus microprocesador estándar, con líneas de datos, líneas de dirección, línea de selección de chip, y línea de lectura/escritura. Los 8 bits de datos pueden ser leídos o grabados al chip por el procesador, siguiendo el protocolo de periferico estándar.

Las características de reconocimiento son:

- Software controlado por frecuencia de muestreo de 5.0 KHz a 15.9 KHz.
- Salida de control de ganancia automática para control de amplificador de entrada externa.
- Analizador de rejilla de LPC de 8 etapas.

Este dispositivo puede ser usado en sistemas dependiente o independiente del locutor, en reconocimiento de palabras aisladas o en voz conectada.

4.3 Conclusiones.

No se pretende en tan corto espacio dar una visión global del futuro de todos los aspectos involucrados en el hardware para el procesamiento del habla. Ello no sólo resultaría difícil, sino imposible. Apuntaremos, tan sólo algunos de los aspectos que pueden ayudar en esta mejora de prestaciones.

En primer lugar, y desde un punto de vista tecnológico, nuevos chips irán apareciendo, con características cada vez más óptimas en cuanto a velocidad y consumo, y tanto en lo que se refiere a circuitos integrados o procesadores específicos para procesamiento de señales, síntesis, o reconocimiento, como aquellos procesadores de propósito general, o circuitos estándar (multiplicadores, etc.) con los que construir procesadores a la medida. El camino, con el gran desarrollo de los circuitos integrados VLSI, no ha hecho más que empezar.

En segundo aspecto, y quizás el más importante, sera la continuación en la búsqueda de nuevas arquitecturas, más eficiente que la tradicional secuencial, para resolver problemas planteados. En lo que respecta al reconocimiento de voz, a medida que necesitemos reconocer una palabra entre un vocabulario cada vez mayor o tendamos hacia el reconocimiento continuo de voz, se necesitará entonces una gran labor de investigación en estructuras de tipo paralelo (multiprocesadores, flujo de datos, etc.); sistemas la cual seran más sostificados y que cumplan con los requerimientos dados y sobre todo lo más importante, hacer o ayudar a que el trabajo del hombre sea más fácil y productivo.

C O N C L U S I O N E S

Esta tesis ha servido para dar los primeros pasos o sentar las primeras bases acerca de Sistemas de Reconocimiento de Voz en información o material disponible en español; como también fomentar el interés y desarrollo de futuros trabajos o investigaciones en esta área. A lo largo que ha tenido el desarrollo de esta tesis me he dado cuenta que tan importante constituye hoy en día la tecnología de Reconocimiento de Voz, principalmente se han encontrado aplicaciones muy diversas, otras muy importantes y otras diria yo humanitarias o de ayuda hacia el hombre, como es, en los minusválidos o personas impedidas.

Encontramos que el proceso de Reconocimiento de Voz es en realidad un proceso muy complejo desde las características prosódicas del habla hasta los mismos sistemas en sí ya construidos. Encontramos, también factores externos involucrados en el proceso de Reconocimiento de Voz, esto es como el ruido, el estado del locutor, las características del cuarto o sala, y otras más.

Encontraremos que cada día esta tecnología esta avanzando muy rapidamente, la cual se ha logrado aumentar la cantidad de palabras a reconocer (vocabulario) y reducir el tiempo de reconocimiento como también disminuyendo el promedio de error, logrando así una confiabilidad excelente para el hombre.

Se podria dar una conclusión extensa sobre esta tesis pero sólo me limitaré a decir finalmente que es un pequeño paso e importante a la consecución de trabajos futuros.

Gracias,

Fernando Ernesto Chuw Lau.

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

[1] - Bricker P. D. Et. Al.;
STATISTICAL FOR TALKER IDENTIFICATION;
Bell System T. J., Vol. 50, pp. 1427-1454;
Abril 1971.

[2] - Markel J. D. y Gray A. H.;
LINEAR PREDICTION OF SPEECH;
Springer-Verlag, Berlin, Heildeberg, New York;
1976.

[3] - Su L., Li K. P., y Fu K. S.;
IDENTIFICATION OF SPEAKERS BY USE OF NASAL COARTICULATION.;
J. Acoust. Soc. Amer., Vol. 50, pp. 661-670;
Agosto 1971.

[4] - Flanagan J. L.;
SPEECH ANALYSIS, SYNTHESIS & PERCEPTION;
New York: Academic Press, Inc.;
1972.

[5] - Nilsson N. J.;
LEARNING MACHINES;
New York: McGraw-Hill;
1965.

[6] - Martin T. B., Zadell H., Grunza E., y Hercher M.;
NUMERIC SPEECH TRANSLATING SYSTEM in automatic pattern recognition;
Washington, D.C.:National Security Industrial Association, pp. 113-141;
Mayo 1969.

[7] - Martin T. B.;
ACOUSTIC RECOGNITION OF A LIMITED VOCABULARY IN CONTINUOS SPEECH;
Ph. D. Dissertation, Univ. Pennsylvania, Philadelphia;
Mayo 1970.

[8] - Martin T. B., Zadell H. J., Nelson A. L., y Schanne J.;
CONTINUOS SPEECH RECOGNITION AND SYNTHESIS;
Tech. Rep. AFAL-TR 6-120;
Octubre 1967 (AD821168).

[9] - Herscher M. B. y Cox R. B.;
AN ADAPTIVE ISOLATED-WORD SPEECH RECOGNITION SYSTEMS;
in Conf. Rec. 1972., Conf. Speech Communication and Processing, pp. 82-92;
1972.

[10] - Martin T. B.;
APPLICATIONS OF LIMITED VOCABULARY RECOGNITION SYSTEMS in SPEECH RECOGNITION: invited papers presented at the IEEE 1974 symposium, New York: Academic Press, Inc.;
1974.

[11] - Rabiner L. R.;
DIGITAL PROCESSING OF SPEECH SIGNALS;
Prentince-Hall, Inc.;
1979.

[12] - Martin T. B.;
PRACTICAL APPLICATIONS OF VOICE INPUT TO MACHINES;
Proceedings of the IEEE, Vol. 64, No. 4;
Abril 1974.

[13] - Reddy D. Raj;
SPEECH RECOGNITION BY MACHINE: A REVIEW;
Proceedings of the IEEE, Vol. 64, No. 4;
Abril 1976.

[14] - Rulot H., Vidal E., Casacuberta F.;
UTILIZACION DE LA FUNCION DE AUTOCORRELACION EN EL RECONOCIMIENTO AUTOMATICO DE LA PALABRA;
V Congreso de Informática y Automática en Madrid;
Mayo 1982.

[15] - Rulot H., Vidal E., Casacuberta F.;
ISOLATED WORD RECOGNITION SYSTEM BASED ON THE AUTOCORRELATION FUNCTION;
1982 Portugal Workshop on signal Processing and its Applications;
Portugal 1982.

[16] - Dudley H. y Balashek S.;
AUTOMATIC RECOGNITION OF PHONETIC PATTERNS IN SPEECH;
J. Acoust. Soc. Amer., Vol. 30, pp. 721-739;
Agosto 1958.

[17] - Morte F., Peris R., Vidal E., Casacuberta F.;
PARAMETRIZACION SIMPLE DE LA VOZ MEDIANTE MICROPROCESADOR (1). FUNCION DE AUTOCORRELACION;
Mundo Electronico, pp. 81-85, No. 141;
Junio 1984.

[18] - Okuchi M., y Sakai T.;
TRAPEZOIDAL D.P. MATCHING WITH TIME REVERSIBILITY;
Procc. Of ICASSP, pp. 1239-1242;
1982.

[19] - Rabiner L., y Levinson S;
ISOLATED AND CONNECTED WORD RECOGNITION THEORY AND SELECT APPLICATIONS;

IEEE Trans. on Communications, COM-29,5;
Mayo 1981.

[20] - Sakoe H.;
TWO LEVEL DP-MATCHING;
IEEE Trans. Acoust. Speech and Signal Proc., ASSP-27,6;
Diciembre 1979.

[21] - Kuhn M., y Tomaschewski H.;
IMPROVEMENTS IN ISOLATED WORD RECOGNITION;
IEEE Trans. Acoust. Speech and Signal Proc., ASSP-31,1;
Febrero 1983.

[22] - White G., y Neeley R.,
SPEECH RECOGNITION EXPERIMENTS WITH LINEAR PREDICTION, BANDPASS FILTERING AND DP;
IEEE Trans Acoust. Speech and Signal Proc., ASSP-24,2;
Abril 1976.

[23] - Aldefeld B., Rabiner L., Rosenberg A., y Wilpon J.;
AUTOMATIC DIRECTORY LISTING RETRIEVAL SYSTEM BASED ON ISOLATED WORD RECOGNITION;
Proc. of IEEE, Vol. 68, No. 11;
Noviembre 1980.

[24] - Morre R., Rusell M., Y Tomlinson M.;
LOCALLY CONSTRAINED PROGRAMMING IN AUTOMATIC SPEECH RECOGNITION;
Proc. of ICASSP, pp. 1270-1273;
1982.

[25] - Rubio A., y Carrion M. C.;
UN DETECTOR EXPLICITO DE PRINCIPIO Y FINAL PARA PALABRAS AISLADAS;
Rev. de Informática y Automática;
1984.

[26] - Vidal E., Casacuberta F., Sanchis E., y Rulot H.;
DIVERSAS APROXIMACIONES AL RECONOCIMIENTO DE PALABRAS AISLADAS;
Rev. de Informática y Automática;
Julio-Septiembre 1983.

[27] - Golderos A.;
RECONOCIMIENTO DE PALABRAS AISLADAS CON INDEPENDENCIA DEL LOCUTOR, APLICACION AL RECONOCIMIENTO DE DIGITOS EN ESPAÑOL;
ETSIT-UPM;
Septiembre 1983.

[28] - Itakura F.;
MINIMAL PREDICTION RESIDUAL PRINCIPLE APPLIED TO

SPEECH RECOGNITION;
IEEE Trans. Acoust. Speech and Signal Proc. ASSP-23;
Febrero 1975.

[29] - Markhoul J., y Wolf J.;
LINEAR PREDICTION AND THE SPECTRAL ANALYSIS OF SPEECH;
Report 2304, Cambridge, Mass. Bolt, Beranek and Newman Inc.;
Agosto 1972.

[30] - Santos Suarez J.;
CONTRIBUCION A LA SINTESIS POR REGLA DEL HABLA ESPAÑOLA;
Tesis Doctoral UPM-ETSIT;
Octubre 1981.

[31] - McGonegal C. A., Rosenberg A. E., y Rabiner L. R.;
THE EFFECTS OF SEVERAL TRANSMISSION SYSTEMS ON AN AUTOMATIC SPEAKERS VERIFICATION SYSTEMS;
Bell System T. J., Vol. 58, pp. 2071-2087;
Noviembre 1979.

[32] - Doddington G. R.;
PERSONAL IDENTITY VERIFICATION USING VOICE;
Proc. ELECTRO-76, pp. 22-4-1-5;
Mayo 11-14, 1976.

[33] - Furui S., Itakura F., y Saito S.;
TALKER RECOGNITION BY LONG-TIME AVERAGED SPEECH SPECTRUM;
ELECTRON, Commun. JAP., Vol. SSA, pp. 54-61;
Octubre 1972.

[34] - Furui S.;
COMPARISON OF SPEAKER RECOGNITION METHODS USING STATISTICAL FEATURES AND DYNAMIC FATURES;
IEEE Trans. Acoust. Speech and Signal Procc., Vol. 29, pp. 342-350;
Junio 1981.

[35] - Wolf J. J.;
EFFICIENT ACOUSTIC PARAMETERS FOR SPEAKER RECOGNITION;
J. Acoust. Soc. Amer., Vol. 51, pp. 2044-2055;
Junio 1972.

[36] - Sakoe H., y Chiba S.;
A DYNAMIC PROGRAMMING APROACH TO CONTINUOS SPEECH RECOGNITION;
in Proc. 7th Int. Congr. Acoustics, paper 20;

1971.

[37] - SIMPLE TECHNIQUES FOR TRANSFORMING SPEECH TO QUASI-PHONEME STRINGS;
in Proc. Speech Communication Seminar, Stockholm, Sweden, pp. 225-231;
Agosto 1-3, 1974.

[38] - White G. M.;
SPEECH RECOGNITION WITH CHARACTER STRING ENCODING;
in Procc. 1972 IEEE Conf. Decision and Control, New Orleans, LA., TJ217,117;
Diciembre 13-15, 1972.

[39] - Neroth C. C.;
AUDIO GRAPHIC PROGRAMMING SYSTEM;
Ph. D. dissertation, Univ. California, Berkeley, pp. 41-45;
1972.

[40] - Witten I.;
PRINCIPLES OF COMPUTER SPEECH;
Academic Press, Inc.;
1982.

[41] - Shearme J. N. y Leach P. F.;
SOME EXPERIMENT WITH A SIMPLE WORD RECOGNITION SYSTEMS;
IEEE Trans. Audio Electroacoust., Vol. AU-16, pp. 256-261;
Marzo 1968.

[42] - Schwartz M.;
TRANSMISION DE INFORMACION, MODULACION Y RUIDO;
McGraw-Hill, Cap. III;
1982.

[43] - Hsu H.;
ANALISIS DE FOURIER;
Fondo Educativo Interamericano;

[44] - Hoeschele D. F. Jr.;
ANALOG-TO-DIGITAL, DIGITAL-TO-ANALOG CONVERSION TECHNIQUES;
John Wiley, New York;
1968.

[45] - Hnatek E.;
A USER'S HANDBOOK OF D/A AND A/D CONVERTERS;
Wiley-Interscience, New York;
1970.

[46] - Sheingold D. H.;

ANALOG-DIGITAL CONVERSION HANDBOOK;
Analog Devices, pp. 9-3;
1975.

[47] - Golderos A., Martinez R., Nombeca J. R., Pardo M., Santos J., Muñoz E.;
COMUNICACION HOMBRE-MAQUINA POR VOZ;
Mundo Electronico, No. 96, pp. 47;
1980.

[48] - Hill F. y Peterson G.;
TEORIA DE CONMUTACION Y DISEÑO LOGICO;
Editorial Limusa;
1982.

[49] - Millman J.;
MICROELECTRONICS:Digital and Analog Circuits and Systems;
McGraw-Hill;
1983.

[50] - Kaufman M. y Seidman A.;
MANUAL PARA INGENIEROS Y TECNICOS EN ELECTRONICA;
McGraw-Hill;
1982.

[51] - Taub H. y Chilling D.;
PRINCIPLES OF COMMUNICATION SYSTEMS;
McGraw-Hill;
1983.

[52] - Flanagan J. L., Ishizaka I., y Shipley K.;
SYNTHESIS OF SPEECH FROM A DYNAMIC MODEL OF THE VOCAL CORDS AND VOCAL TRACT;
Bell System Technology J., Vol. 54, pp. 485-506;
Marzo 1975.

[53] - Sebestyen G. S.;
DECISION-MAKING PROCESSES IN PATTERN RECOGNITION;
New York:McMillan;
1963.