SISTEMA DE RECONOCIMIENTO DE COMANDOS DE VOZ PARA LA CONDUCCIÓN DE UNA SILLA DE RUEDAS

ASESOR: Dr. ROGELIO ALCÁNTARA SILVA

ALUMNO: Lic. MIGUEL COMADURÁN CHAVARRÍA

DEPFI Junio de 1995 UNAM

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

AGRADECIMIENTOS

A mi querida madre y a mi padre q.e.p.d.
A mi amada Ana Maria Salazar S.
A mis hijos Ma. Fernanda y Jose Miguel.
A mis hermanos.
A la direccion del Dr. Rogelio Alcantara Silva
A mis maestros y guias en el sendero del saber.
A mis nuevos amigos y a los de toda mi existencia.

RECONOCIMIENTOS

Se agradece la ayuda y cooperación de todas las personas que apoyaron el desarrollo de este trabajo, como fueron mis compañeros de generación y de aquellos que llegaron despues, ya que con su respaldo, no solo físico sino intelectual, pusieron su grano de arena y su presencia esta aquí. Al Dr.Rogelio Alcántara Silva, que con su ayuda se llegó a la felíz conclusión de este trabajo, por lo que le estoy muy agradecido.

Un reconocimiento a la asesoria de todos mis maestros y en particular al Dr. Francisco Garcia Ugalde, por su gran apoyo moral, ya se que siempre conté con su apoyo. También quiero agradecer a todo el personal de la DEPFI su grán ayuda para la realización de mis estudios, ya que sin reticencia me subieron y bajaron por los diversos niveles de los edificios para estar presente en mis clases.

Finalmente quiero reconocer la dirección, enseñanzas y correcciones en la redacción de este texto al Lic. Roberto Llanas.

Lic. Miguel Comadurán Chavarria.

□

ÍNDICE TEMÁTICO

CAPÍTULO CUARTO

□

INTRODUCCIÓN

Considerando que el desarrollo a nivel nacional de los sistemas de control mediante comandos de voz esta en sus inicios, se penso en un sistema que funcionara con voz, ya que su utilidad en todos los campos de interacción máquina-hombre hace de singular interés el control por voz con opción de manos libres. En particulñar sería de gran utilidad para el control de una silla de ruedas a través de un sistema inteligente de computo, en auxilio a personas que hayan perdido, aparte de la movilidad de las piernas, la de sus extremidades superiores.

El estudio de la digitalización de señales analógicas y los principios básicos de ésta, vislumbran aplicaciones de la señal voz en sistemas de control, reconociendo de entre un grupo determinado de palabras, la que ha sido voceada.

Teniendo en cuenta que los patrones digitales de voz requieren un gran almacenamiento de datos y que a frecuencias de 8kHz un registro de 1.5 segs tiene 12 000 muestras, se optó por un método de codificación de la señal, en inglés Lineal Prediction Coding (LPC), para reducir los requerimientos de memoria y almacenar la escénica de la señal en un 10% de la memoria que sería necesaria, reduciendo así el costo del sistema de detección.

En 1958, cuando empezó la éra y desarrollo de la telefonía, y debido a la necesidad del uso de sus líneas para trasmisión digital, sin tener que recurrir a nuevas inversiones, se logró la transmisión de información digital, y las investigaciones en este ramo llevaron al estudio de las comunicaciones y análisis de señales digitales.

En esa década, muchas compañías de computación querían instalar en sus sistemas el control por medio de voz, pero al ser la voz diferente en cada persona, las promesas de estas compañías no se cumplieron; es difícil encontrar dos voces iguales, pues se genera por medio de cuerdas vocales diferentes en cada caso, el sistema buco-faringeo, cuello, cabeza, etc, hacen que la voz de cada individuo sea diferente al producir ondas de resonancias particulares, reforzando unas y disminuyendo otras; por eso es difícil encontrar dos individuos con la misma voz[2].

Los estudios del análisis LPC, entropía y de señales, se desarrollaron en la decada de los 70's, con Rabiner y Schafer; toda la base matemática en el que se apoyaron estos y otros autores pertenece a la década de los 60's; se publicaron temas como "Mediciones sobre la potencia espectral" de Blackman y Tukey, "Análisis de máxima entropía espectral" de Burg. En 1967 se iniciaron las investigaciones sobre señales de voz con temas como "Codificación predictiva de señales de voz" de Atal y Schroder; en 1969, Itakura y Saito publicaron su trabajo "Sistema basado sobre los coeficientes parciales de autocorrelación en el Análisis y Sintesis de voz", en el que se apoya el método LPC [8].

Schroder publico en 1966 "Vocoders: Análisis y síntesis de voz", donde hizo un estudio de estas técnicas y una revisión de los vocoders, sistemas analógicos codificadores de voz, eficiencia y almacenamiento de señal voz, y en 1984, "Predicción lineal, entropia y análisis de señales", que revisa los conceptos fudamentales de LPC y máxima entropía (ME) para análisis espectral y su aplicación en el mundo de las señales reales. También analizo el poderoso principio de mínima entropia cruzada (MEC), que permite la incorporación a priori de información en el análisis de una señal; en dicho estudio hizo un nuevo modelado de las características de las señales de voz, reduciendo el número de polos (coeficientes de predicción) que se tienen que especificar por cada fracción de la resolución espectral, mediante una mejor información espectral.

La información espectral la da una fuente glótica y las características de radiación de los labios, además de las caractrísticas del micrófono y la respuesta en frecuencia de la transmisión; también hace uso de la información espectral de los intervalos de voz anteriores, particularmente durante los estados estacionarios y los espacios de pequeña variación, presentes en la voz.

Hacia los 80's, las investigaciones hechas por Crochiere y Flanagan en el campo de la digitalización de la señal voz, abrieron la opotunidad a su aplicación comercial. En su publicación [9] especificó las técnicas de codificación y las áreas de más desarrollo, así como la aplicación práctica del hardware con el uso de las nuevas tecnologías.

En 1983, O'Shaughnessy, publicó los últimos avances en la reproducción sintética de voz; comentó sobre los "chips" dedicados de síntesis y los requerimientos de hardware para la producción de señal sintética de voz. Una aplicación se hizo con mudos, que por medio de una computadora escriben el texto que desean reproducir vía bocina, expresando así sus ideas.

Es reconocido el esfuerzo de muchos investigadores para ayudar a minusválidos, pero en este caso, el desarrollo del diseño que se propone tiene una importancia definitiva, sirve para movilzar una silla de ruedas motorizada a través de comandos de voz, dando oportunidad a personas que no pueden usar sus manos, en su desplazamiento de un lugar a otro.

Considérense todas las aplicaciones, específicamente al caso de una persona, con las características descritas, que se encuentre sola en una recámara: no tiene que llamar a alguien para encender la luz, o cualquier aparato electrodoméstico, incluyendo un interfón, o hasta abrir una puerta automática, pues tendrá un genio cerca de él que le haga estos trabajos.

La imaginación es el límite para las aplicaciones del sistema de control por comandos de voz, pero en el caso de esta tesis se enfocó a personas que, desesperadas, no tienen oportunidad de controlar su entorno ni de satisfacer el deseo de ser un poco autosuficientes, ya sea en movimiento, o en cambio de posición en una silla de ruedas motorizada o una cama eléctrica, etc.

La investigación en este campo ha crecido recientemente debido al diseño de los procesadores digitales de señales, DSP's (Digital Signal Proccesors); con estos se ha logrado obtener una respuesta en tiempo real, ya que realizan los millones de operaciones por segundo que se requieren en el proceso de análisis de una señal.

Se puede desarrollar un sistema de control en donde se aplique el reconocimiento automático de señal voz. El sistema se ha desarrollado en la presente tesis con dos bases:

El software, relacionado al procesamiento digital de señales, que comprende todos los algoritmos de adquisición y reconocimiento de la voz vía una PC, base de esta tesis. Este se aboca a resolver el problema en una computadora personal (PC) con el fin de demostrar la viabilidad de dichos algoritmos. Con esto se establece la base para que en un futuro se pueda implantar en el trabajo de tesis presentado por el M. en I. Eduardo Castillo [12].

El software es el apoyo computacional desarrollado para la adquisición de datos y su proceso matemático; abarca diversos campos del análisis de señales como el filtrado y el método LPC de predicción lineal; se incluyen los métodos de obtención de parámetros característicos de la señal como los coeficientes de autocorrelación ri's, de reflexión ki's y los ai's del polinomio generador de la señal para la obtención de voz por síntesis.

También contiene los algoritmos de cálculo de distancias de las señales, para hacer una comparación de parámetros de la señal de prueba con cada una de las señales de la base.

Con todos esos algoritmos se aplica un criterio de correspondencia para decidir que comando se voceó.

El método de reconocimiento de señal voz se basa en los tres parámetros mencinados antes: los ai's, Ki's y ri's, donde los ai's y los Ki's dependen del valor de los ri's. El método es el siguiente:

- Filtro preenfasis, que enfatíza altas frecuencias presentes en las mujeres y en los niños, pues el método LPC tiene una pobre resolución en altas frecuencias.

- Filtrado FIR, cuya característica es la de ser un filtro pasa-banda de 150 a 4000 Hz. Este filtro tiene un corte inferior en 150 Hz con la fin de eliminar los 120 Hz de las lámparas de neón y la frecuencia de 60 Hz de la línea de alimentación, y corte superior en 4000Hz, a fin de suprimir las frecuencias que no conforman las señales de voz, y que están presentes en toda señal analógica que se desea registrar.

- Cálculo de los coeficientes de autocorrelación, base para el LPC. Las ventanas son de 30 milisegundos con traslape de 10 ms.

- Método de Levinson-Durbin, LPC.

- Cálculo de las distancias de los parámetros obtenidos con Levinson y algoritmo de detección.

Son dos las aplicaciones de los pasos anteriores; una se da en la detección de voces archivadas de antemano, y la otra en la detección directa vía micrófono; es evidente que los pasos para ámbos casos tienen algunas diferencias que no son de base, sino de concepto; es decir, que la detección directa difiere de la detección por archivo en un sistema de confirmación del comando reconocido, opción que la detección por archivos fijos no tiene.

La limitación del método LPC de predicción es que su principal aplicación es para señales estacionarias, y la voz no es estacionaria, por lo que el análisis de la señal se hace en lapsos de 30ms (milisegundos), con las características ya apuntadas, para tomar los lapsos de estacionariedad que la voz presenta en estos intervalos.

El problema inherente a la potencia de la señal se solucionó normalizandola a la potencia unitaria, lo que se hace en el momento de ser adquirida; esto proporciona un patrón uniforme de la amplitud; dado que la señal es estacionaria a tramos, se dice que existe una cuasi-estacionariedad, que es en lo que se apoya la aplicación del LPC.

La función del traslape es obtener cierta redundancia de la información de la ventana anterior y establecer un encadenamiento de la señal, que se puede perder en caso de no tomarse esta medida

El reconocimiento de la voz se dividió en cuatro áreas [2 y 8]:

Dependencia al hablar
Independencia al hablar
Hablar continuo
Voz discreta.

La dependencia al hablar confina al sistema a una sola persona. El sistema es entrenado para que responda a una voz en particular, lo que limitaría, por un lado, el propósito general de el sistema, pero por otro haría que este respondiera a una sola persona, caso deseable para el control de una silla de ruedas motorizada.

El control de la independencia al hablar es mucho más difícil, pues su análisis es más complicado, ya que se basa en una gran variedad de voces, dependiendo de la cantidad de usuarios.

La voz discreta se apoya en pausas bien definidas, y aunque el trabajo computacional se simplifica, es tedioso. El Hablar continuo, de uso cotidiano, corresponde a un medio de interacción humana, solo que el esfuerzo de análisis es devastador al tratar de separar la palabras.

Existen diversos métodos para llevar a cabo los procesos que demanda el proyecto; como la señal es analógica, se tiene que pasar por un sistema digitalizador, ya que el sistema es computacional y trabaja con números binarios. Puesto que el proceso es puramente matemático, existe cierta incertidumbre en la detección, por lo que se entrena al sistema en el reconocimiento, estableciendo un inervalo de aciertos que permiten tomar una decisión.

Considérese que los 'frames', conjunto discreto de muestras, en bloques de 256, de una misma palabra, pueden decirse con tiempos diferentes, por lo cual los valores de los marcos serán diferentes en cada caso, obteniéndose diferentes coeficientes de autocorrelación, lo que provoca una incertidumbre, aún para voces iguales.

En ese entorno, los alcances de un sistema que opere con comandos de voz están limitados por la imaginación del hombre, que van desde accionar un apagador hasta manejar un vehículo motorizado; desde abrir una puerta hasta su aplicación en un quirófano. Las características del sistema dependerán de las diversas opciones; por

ejemplo, si se desea que una sola persona controle el sistema, o que cualquiera lo controle.

Piénsese en un quirófano en donde solo el doctor puede prescribir cierta cantidad de insulina aplicable al paciente, por lo que en este caso es deseable que solo él pueda generar este comando. En el caso de la silla de ruedas, el comando debe ser dado por el conductor de esta, ya que otra persona podría accionarla sin querer.

En conclusión, es necesario un esfuerzo para que el sistema, aparte de ser inteligente, sea eficiente y perfectible por la experiencia que se adquiera en el uso de infinitas aplicaciones en todos los campos de la vida, y básicamente para personas con deficiencias motoras, o de aquéllos que deseen un control mediante la voz.

La aplicación de un sistema de control computacional en una silla de ruedas motorizada es de una delicadeza extrema, ya que un error provocaría un accidente, de ahí que se recomienda una experimentación exhaustiva en todos los procesos de control; al respecto, ya se están preparando los recursos humanos necesarios para este fin.

Por otra parte ya se han desarrollado nuevos métodos de control del tipo imagenológico, con mucha precisión, pero de costos muy elevados, ya que en lugar de los DSP's se ha recurrido a la tecnología RISC, necesaria en el procesamiento de imágenes.

Los temas de la presenrte tesis se han arreglado de la siguiente forma, en el capítulo primero se hace una reseña del procesamiento digital de señales y la aplicación de principios básicos con el propósito de interrelacionar datos de algún tipo de señal, en este caso de voz; los métodos de estimación paramétrica más usados en la actualidad se analizan con el fin de encontrar una justificación de su aplicación a un conjunto de muestras de un proceso de digitalización de una señal [1, 2 y 3]; la presencia de la tranformada de Fourier y su aplicación en la señal-voz, con una introducción a la función de autocorrelación y su relación con la ecuación en diferencias.

En el capítulo segundo se presentan los esquemas de estimación paramétrica para el reconociminto y las características de cada uno de ellos. Entre otros están los métodos de predicción lineal LPC con el apoyo de métodos recursivos, como lo son el de Levinson-Durbin y el de Leroux-Gueguen, procesos que ayudaron a la automatización y a la rapidéz en la obtención de los parámetros característicos de una señal y sus aplicaciones.

El tercer capítulo tiene en sus líneas la descripción de esquemas generales, el análisis de las palabras, los procesos de detección y las soluciones a los diferentes métodos del cálculo de las distancias, como son la distancia Euclidiana, logarítmica Euclidiana, de Itakura, Dinamic Time Warping DTW y la selección de uno de ellos por su mejor desempeño. No dejamos de reconocer que en este momento se encuentran ya nuevos métodos de análisis como son redes neuronales, métodos de análisis de Fourier de tiempo corto, métodos Fuzzy Logic y modelos de Markov, motivo de un estudio posterior que nos permita mejorar el desempeño de nuestro sistema.

En el capítulo cuarto se presentan los resultados del desempeño del sistema que se eligió como óptimo entre los diferentes sistemas de reconocimiento, en donde se muestra cómo la detección de voces de manera directa se ve contaminada con ruido, reduciendo la calidad del reconocimiento, lo mismo sucede en la aplicación con los dígitos, en donde el reconocimiento se hace en base a un diccionario de diez voces, o de doce voces para los comandos para el control de una silla de ruedas motorizada.

Se puede concluir que, en base a lo presentado, el desempeño de reconocimiento digital de señales aporta una herramienta de gran utilidad como se comentó con anterioridad; su uso sobrepasa las expectativas y puede llevar al hombre a una interacción inteligente con las maquinas. Además nuestra solución tiene otras opciones de aplicación adicionales y entre las principales se encuentran el reconocimiento directo, del que ya hablamos, grabar una voz y poderla reproducir, filtrarla y visualizarla en un ambiente gráfico de 560 muestras de ventanas continuas, toda la señal o algún intervalo de muestras en particular, lo que permite hacer un análisis de la eficiencia de los filtros digitales y de la forma de una señal ya que se presentan

las gráficas de la señal sin filtrar y la señal filtrada. También se muestra un sistema de detección de dígitos por medio de voz con la opción de rectificar si el dígito detectado de forma directa, es el correcto. Esta opción es interesante para el mercado de teléfonos públicos automáticos.

CAPÍTULO PRIMERO

EL PROCESAMIENTO DIGITAL DE SEÑALES Y SUS APLICACIONES

1.1 Análisis de señales y sistemas

Dentro de una gran variedad de campos de nuestro entorno encontramos que las señales interactúan con nuestros sensores biológicos. Estas señales se presentan en diseño de circuitos en ingeniería y como canal de comunicaciones, en astronomía, acústica y video, biomedicina, etc. Sus principales parámetros son la amplitud (diferencias en el potencial), la frecuencia y el tiempo.

Las señales se representan, matemáticamente, como una función dependiente de una o más variables independientes. Por ejemplo, una señal voz se puede representar como su presión acústica en función del tiempo; en particular, las funciones que representan la señal voz dependen de una sola variable independiente. Los transductores biológicos reciben diversos tipos de señales del medio que nos rodea, estas señales son de carácter analógico y para poder ser procesadas se transforman en señales digitalizadas y se les llama señales en el tiempo-discreto. Las señales analógicas resultan de una relación ecológica entre las diversas especies que se confunden en la naturaleza.

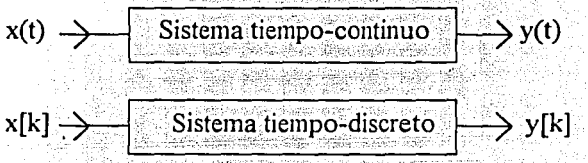
Las señales se pueden cambiar por medio de dos métodos, una transformación lineal o una no-lineal, la primera permite la superposición y el producto por un escalar y la segunda transformación no permite alguna de estas reglas de linealidad o ámbas. Se pueden distinguir entre señales pares e impares; una señal par tiene la propiedad de que $x(t) = x(-t)$ en el tiempo continuo y con $x[k] = x[-k]$ en el tiempo discreto; las señales impares se definen como $x(-t) = -x(t)$ ó $x(t) = -x(-t)$ [13], etc.

Entre las señales más importantes están las exponenciales, las logarítmicas y las polinomiales, sin dejar de considerar una, que es de utilidad especial en señales en el tiempo-discreto, la función impulso unitario $\delta(t)$ en el tiempo continuo, o $\delta[k]$ en el discreto, que se define como [15]

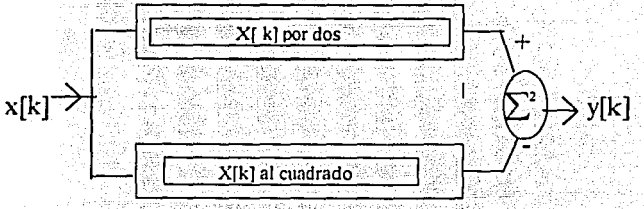
$$\delta[k] = \begin{cases} 0 & k \neq 0 \\ 1 & k = 0 \end{cases}$$

Por su parte, un sistema se visualiza como un proceso que resulta de la transformación de señales, es decir, que si una señal se pasa por un sistema, ésta será transformada por las características del sistema en otra señal, como el cambio en la calidad del tono, etc; los sistemas transforman las señales con el fin de obtener o mejorar sus propiedades.

Para nuestros dos tipos de señales, tendríamos



Un ejemplo más directo sería



Sistema para el cálculo de $y[k] = \left(2x[k] - x[k]^2\right)^2$

Las propiedades de los sistemas son

Sistemas con o sin memoria
invertibilidad e inversos, estabilidad, linealidad,
causales e invariantes en el tiempo

El ejemplo tácito de arriba es un sistema sin memoria, pues la variable dependiente se apoya en el valor de la entrada o variable independiente; uno con memoria sería

$$y[k] = x[k] - x[k-1]$$

Un sistema es invertible si conociendo su entrada se puede determinar su salida y es inverso si al invertir su salida se obtiene la entrada del sistema invertiblede un proceso anterior.

Un sistema es estable sí para pequeñas señales no se obtiene una señal divergente a una salida.

Un sistema es lineal sí cumple con el teorema de superposición y multiplicación por un escalar.

Un sistema es causal sí su salida depende de su entrada actual o de sus pasadas entradas o ámbas opciones.

Un sistema es invariante en el tiempo sí para un corrimiento en el tiempo se obtiene una salida de acuerdo a dicho corrimiento.

1.2 Análisis y síntesis de filtros digitales

Los filtros digitales se apoyan en el desarrollo que se ha tenido en los filtros analógicos [13]; para la implantación de filtros digitales tenemos que reconocer la base de los filtros analógicos. Esta base de filtros analógicos se divide en pasa-bajas, pasa-altas, pasa-banda y supresor de banda; de estos existe un concepto ideal y otro real, diferenciandose por el corte abrupto en el filtro ideal y el decaimiento relativamente suave en el filtro real.

Los principales filtros son los de Butterworth [13], los de Chebyshev y los elípticos, caracterizados por su respuesta en frecuencia. La magnitud cuadrática de ésta respuesta en frecuencia, para los filtros de Butterworth, es

$$\left|H_n(j\Omega)\right|^2 = 1/\left[1+\left(\Omega/\Omega_c\right)^{2n}\right] \qquad (1.1)$$

en donde Ω representa las frecuencias presentes en una señal y Ωc la frecuencia de corte de la señal (fig 1.1), de amplitud normalizada [13].

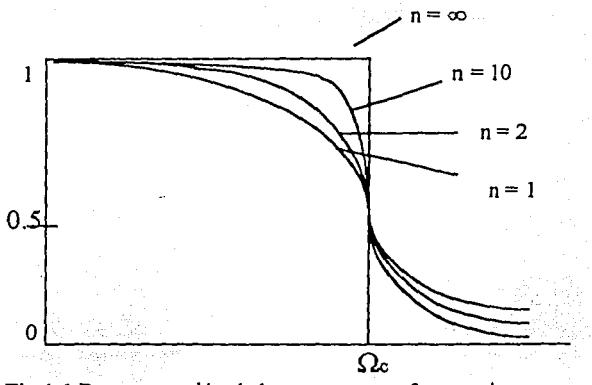


Fig 1.1 Representación de la respuesta en frecuencia

Los polinomios para filtros normalizados de Butterworth son

Orden n	Polinomios de Butterworth Bn(s)
1	$s + 1$
2	$s^2+\sqrt{2}s+1$
3	$(s^2+s+1)(s+1)$
4	$(s^2+0.76536s+1)(s^2+1.848s+1)$
5	$(s^2+0.618s+1)(s^2+0.618s+1)(s+1)$
⋮	⋮

Y los filtros en base a sus polos se forman por

$$H_n(s) = 1/B_n(s)$$

donde la dependencia de la función, s, representa su proyección en el plano complejo S.

El diseño de los filtros se basa en ciertas frecuencias críticas Ω_1 y Ω_2 y sus respectivas ganancias K1 y K2, con la ganancia en dB

$$K1 = 10\log\left|1/\left[1+\left(\Omega_1/\Omega_c\right)^{2n}\right]\right|$$

$$K_2 = 10\log\left|1/\left[1+\left(\Omega_2/\Omega_c\right)^{2n}\right]\right|$$

y el valor de n se obtiene con [13]

$$n = \left\lceil \frac{\log_{10}\left[\left(10^{-k_1/10}-1\right)/\left(10^{-k_2/10}-1\right)\right]}{2\log_{10}\left(\Omega_1/\Omega_2\right)} \right\rceil \qquad (1.2)$$

en donde $\lceil \circ \rceil$ representa el entero mayor más cercano al valor encontrado por (1.2).

Los filtros de Chebyshev, de tipo 1, contienen un rizo en la banda de paso y los de tipo 2 en la banda de rechazo. El filtro pasa-bajas de Chebyshev con ancho de banda unitario está caracterizado por la magnitud de la respuesta en frecuencia cuadrática

$$\left|H_n(j\Omega)\right|^2 = \frac{1}{1+\varepsilon^2 T_n^2(\Omega)} \qquad (1.3)$$

donde ε es la amplitud del ripple y $T_n(\Omega)$ es el polinomio de enésimo orden de Chebyshev, que se genera con la fórmula recursiva siguiente [13]

$$T_n(x) = 2x \cdot T_{n-1}(x) - T_{n-2}(x) \qquad n \geq 2$$

con

$$T_0(x) = 1 \text{ y } T_1(x) = x$$

Por otro lado, si se acepta el rizo de tipo 1 o el de tipo 2 como parte el filtro, se tienen los filtros elípticos. Los filtros elípticos pasa-bajas tienen un menor ancho de banda de transición y es óptimo en el sentido de que ningún otro filtro del mismo órden tiene ese menor ancho de banda de transición para un rizo dado en la banda de paso y en la atenuación de la banda de rechazo. El desarrollo matemático es interesante, un poco complejo e involucra integrales elípticas. La magnitud cuadrática de la respuesta en frecuencia de un filtro elíptico

pasa-bajas normalizado, tiene la misma estructura vista para los filtros de Chebyshev.

$$|H_n(j\Omega)|^2 = \frac{1}{1+\varepsilon^2 R_n^2(\Omega)} \qquad (1.4)$$

donde Rn(Ω) es una función racional de Ω de Chebyshev determinada para ciertas características de ripple. La ec 1.3 es similar a la ec 1.4, igual en estructura, esto nos lleva a considerar que existe una forma general de filtros de Chebyshev que sustituye a T(Ω) y R(Ω) por F(Ω) [13].

1.3 La transformada de Fourier y el espectro en frecuencia

Por definición, para el análisis de Fourier de pequeños intervalos en el tiempo discreto, la función para pasarlos al dominio de la frecuencia es:

$$X_n(e^{j\Omega}) = \sum_{k=-\infty}^{\infty} w[n-k]\,x[k]\,e^{-j\omega nT} \qquad (1.5)$$

Esta función permite introducir el concepto de 'ventaneo' y de otra herramienta como lo es el espectrograma, facilitando el estudio de voz mediante un análisis de Fourier de tiempos cortos, que forman la base de diversas técnicas como la DFT y la FFT.

En la función (1.5), n representa el tiempo de la muestra más reciente, por lo que la función hace una transformación del tiempo a la frecuencia, y viceversa, n ⇔ ω. La variable x[k] es la señal de entrada digitalizada en el tiempo kT, donde T es el periodo de muestreo, y ω[k] una ventana secuencial valuada en tiempo real. La ventana ω[k] determina el segmento de señal que será procesada considerando cero todo lo que se encuentre fuera de ella. Las más usuales, son la Rectangular y la de Hamming, con 8 bits/muestra, que tienen la forma

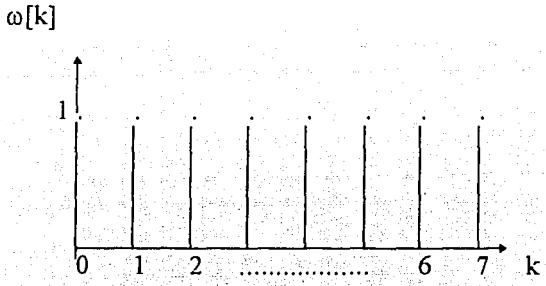


Fig. 1.1. Ventana Rectangular

Cuya función es

$$\omega[k] = \begin{cases} 1 & si\ 0 \le k \le N-1 \\ 0 & de\ otra\ forma \end{cases}$$; N=8 en el ejemplo.

y la de Hamming es

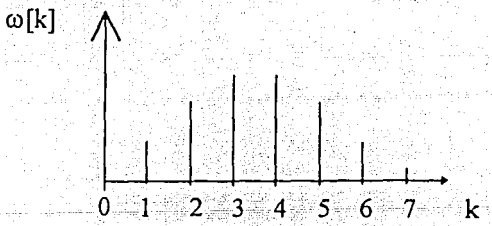


Fig 1.2 Ventana de Hamming

Cuya función es

$$\omega[k] = \begin{cases} 0.54 - 0.46\cos(2\pi k / N) & si\ 0 \le k \le N-1 \\ 0 & de\ otra\ forma \end{cases}$$

En aplicaciones de voz, es más usual la ventana de Hamming por sus mejores propiedades espectrales.

Después de este ventaneo se aplica una FFT para obtener el espectro de una muestra.

El entorno discurre de manera analógica, como la música, la voz, las fotos y películas o la señal de una videocámara. Para procesarlas con un DSP, se debe tener en consideración que las señales no se canalizan sin previo tratamiento, pues se perdería la información contenida en la señal, que después de ser procesada, con la síntesis debe excitar los transductores de los oídos, para reconocer la información transmitida.

Generalmente cuando se convierte una señal analógica en una señal digital, se muestrea a la frecuencia necesaria para satisfacer las leyes de procesamiento digital. En el teorema de Nyquist [2 y 3] se establece que la frecuencia de muestreo debe ocurrir al doble de la mayor frecuencia presente en la señal para mantener una buena fidelidad. La frecuencia típica de muestreo de voz es de 8.0 kHz, o sea más del doble de los 3.5 kHz que determina la mayor de las frecuencias presentes en la voz; esto contempla las altas frecuencias en la voz de una mujer o de un niño.

El oído humano es capáz de captar sonidos de 20 Hz a 20 kHz [8], rango de frecuencias en los que se encuentran el sonido de un violín, un piano, un sintetizador electrónico, o la voz humana.

Muchos sonidos están constituidos por frecuencias fundamentales, además de armónicas (múltiplos de la fundamental), que dán un sonido de características definidas, diferente al de una señal senoidal pura.

1.4 La función de autocorrelación

La función de autocorrelación es una función que relaciona las muestras de una señal, voz en nuestro caso, obteniendo la energía de

constitutivos, como lo son sonidos voceados, los no voceados y los silencios, siendo estos elementos, ri's, la base para el cálculo de los parámetros LPC. Los coeficientes de autocorelación pueden mostrar cuando un segmento de señal tiene una energía despreciable o nula, pudiendo con esto discriminar frames que no contengan información suficiente o que provoque que la matriz de autocorrelaciones quede mal configurada. La función de autocorrelación es

$$r_i(k) = r_i(-k) = \sum_{n}^{N-1-k} x_i[n]x_i[n+k] \qquad (1.6)$$

donde N es la longitud del frame e *i* la i-ésima ventana.

Esta función puede tener valores cercanos a la unidad y sus valores, ri's, decrecer lentamente, para una concepción de la señal en los silencios o una variación alterna de los ri's para un segmento de voz voceado. Uno de los problemas que enfrenta la función de autocorrelación es que puede adquirir valores muy próximos a cero, lo que llevaría a errores en la detección de los parámetros LPC, como son los coeficientes de reflexión y los coeficientes ai's generadores de la señal recien muestreada.

Los coeficientes de autocorrelación, ri's, son decrecientes oscilantes, su variación es decreciente cambiando de signo alternadamente, o, en algunos casos decrecientes; para una señal que esta bien conformada; estos se normalizan, tomando a r(0) como divisor común, es decir los ri's se normalizan con

$$r(k) = r(k) / r(0) \qquad k = 0, 1, \ldots, p$$

donde p es el órden del polinomio generatriz.

La propiedad decreciente se puede demostrar integrando la siguiente ecuación donde r(k) se puede expresar como

$$r(k) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} \left[x(t) - x(t+k)\right]^2 dt \qquad (1.7)$$

desarrollando esta integral se demuestra que r(0) > r(k), k = 0, 1, ... , p [17].

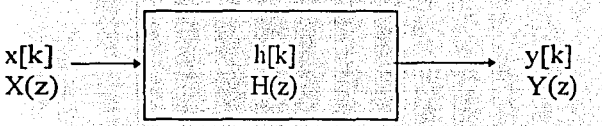
Esto justifica la principal característica de los coeficientes de autocorrelación, que son la base para el cálculo de los ai's y de los ki's.

1.5 La ecuación en diferencias

Una secuencia de muestras, x[k], que pasa a traves de un sistema, cuya función de transferencia es h[k], cumple con la siguiente ecuación

$$y[k] = \sum_{m=0}^{M-1} h[m]x[k-m]; \qquad \text{M = grado de h.} \quad (1.8)$$

Y cuyo sistema quedaría representado de la siguiente manera



representando dos opciones, la función en el tiempo y la función en la frecuencia, por lo que

$$Y(z) = H(z)X(z)$$

Los filtros FIR y IIR son funciones de transferencia basados en (1.8) y la siguiente ecuación generalíza los dos casos

$$y[k] = -\sum_{m=1}^{p} a_m y[k-m] + \sum_{n=0}^{q} b_n x[k-n] \qquad (1.9)$$

Un caso particular de (1.9) se presenta cuando en esta ecuación los valores de ai son iguales a cero, para $0 \leq i \leq p$. En este caso la función queda como

$$y[k] = \sum_{n=0}^{q} b_n x[k - n] \tag{1.10}$$

que corresponde a los filtros FIR. La ec en diferencias, (1.9), se puede desarrollar en el dominio de Z, suponiendo que q = 2, como sigue

$$H(z) = Z\{h[k]\} = \frac{(z)}{(z)} = h_0 + h_1 z^{-1} + h_2 z^{-2}$$

o en la frecuencia

$$H(\omega) = \frac{(\omega)}{(\omega)} = h_0 + h_1 e^{-j\omega} + h_2 e^{-j2\omega}$$

donde el módulo cuadrático de la respuesta en frecuencia del sistema, para $q \neq 0$ y $b_0 = 1$, en la ecuación (1.9) [7], da:

$$|H(\omega)|^2 = \frac{1}{\left(\sum_{i=0}^{p} a_i \cos(i\omega)\right)^2 + \left(\sum_{k=1}^{p} a_k \operatorname{sen}(k\omega)\right)} \tag{1.11}$$

Estos procesos de la señal y su relación con la ecuación en diferencias nos permite un exahustivo análisis de la señal en el tiempo y la frecuencia y la aplicación de los coeficientes ai's en la síntesis de la señal.

Con todo lo expuesto hasta aquí, se tienen los elementos de juicio para el reconocimiento de señal-voz, proporcionando las bases para la estimación paramétrica.

CAPÍTULO SEGUNDO
ESTIMACIÓN PARAMÉTRICA

2.1 Estimación paramétrica de modelos autorregresivos

Métodos y sistemas de predicción

Tres son los métodos principales para el análisis de las señales que se generan en el medio circundante [8]: predicción lineal (PL); Análisis espectral de máxima entropía (ME) y el criterio de mínima entropía cruzada (MEC).

Procesos Autorregresivos y Predicción Lineal

El cálculo para la obtención de los coeficientes del polinomio generatríz de una señal, ai's, con los que se hace la síntesis de voz; es uno de los procesos más populares en la detección de señales; se llama proceso autorregresivo (AR), donde se considera la existencia de una serie temporal, de órden P, definida por x[k], x[k-1], ..., x[k-p] que representa este modelo y satisface la ecuación:

$$x[k] + a_1\, x[k-1] + a_2\, x[k-2] + \ldots + a_p\, x[k-p] = v[k] \qquad (2.1)$$

NOTA: La ecuación 2.1 tiene la estructura de la ecuación (1.9), que es un SLIT por ecuación en diferencias, donde q = 0, $b_0 = 1$ y x[k] = v[k].

donde a_1 , a_2 , ... , a_p son constantes denominadas parámetros AR con $a_0 = 1$, y {v[k]} es un ruido-blanco. La ec (2.1) se puede reacomodar como,

$$x[k] = v[k] - a_1\, x[k-1] - a_2\, x[k-2] - \ldots - a_p\, x[k-p] \qquad (2.2)$$

El lado izquierdo de la ec (2.1) representa la convolución de una secuencia de entradas {x[k-i]} con los parámetros AR $\{a_i\}$, o sea,

$$\sum_{i=0}^{p} a_i x[k-i] = v[k] \qquad (2.4)$$

donde $a_0 = 1$.

Tomando la transformada Z en ambos lados de (2.4), para los parámetros a_i ,

$$H_A(z) = \sum_{i=1}^{p} a_i z^{-i} \tag{2.5}$$

y para la secuencia de entrada {x[k]}, la tranformada será,

$$U(z) = \sum_{k=1}^{\infty} x[k] z^{-k} \tag{2.6}$$

igualmente para {v[k]}

$$V(z) = \sum_{k=1}^{\infty} v[k] z^{-k} \tag{2.7}$$

La transformación total es el producto de (2.5) y (2.6) igualando a (2.7), queda,

$$H_A(z)U(z) = V(z) \tag{2.8}$$

La transformada en Z, ecuación (2.8), tiene dos interpretaciones, dependiendo del punto de vista en que se considere el proceso AR de la secuencia {x[k]}, como entrada o como salida [1, 2]:

1. Dado un proceso AR, {x[k]}, se puede usar la ec (2.1) para producir ruido-blanco, {v[k]}, como salida; esta función de transferencia se puede representar como un proceso analizador con duración finita [1]

$$H_A(z) = \frac{V(z)}{U(z)} \tag{2.9}$$

2. El ruido-blanco actuando como entrada, puede producir la secuencia de proceso AR, {x[k]}, como salida; esto representa un filtro generador, cuya función de transferencia es,

$$H_G(z) = \frac{U(z)}{V(z)} \tag{2.10}$$

Sustituyendo (2.9) en (2.10)

$$H_G(z) = \frac{1}{H_A(z)} =$$

$$= \frac{1}{\sum_{k=1}^{n} a_k z^{-k}} \tag{2.11}$$

La ec (2.9) es solo-ceros y la (2.11) es solo-polos [1], además para la respuesta en frecuencia, la solución para (2.11), las b_i's son cero con $b_0 = 1$, está dada por la ecuación (1.11).

2.2 Análisis espectral de mínima entropía cruzada (MEC)

Dadas las estimaciones a priori de un espectro S_k, para N frecuencias equidistantes, con M+1 restricciones de correlación para un espectro posterior T_k, se tiene,

$$r(n) = \sum_{k=1}^{N} T_k C_{nk} \qquad n = 0, \ldots, M \tag{2.17}$$

De lo que resulta el espectro

$$T_k = \left[\frac{1}{S_k} + \sum_{n=0}^{M} \beta_n C_{nk} \right]^{-1} \tag{2.18}$$

donde los valores de β_n, se toman de tal manera que las restricciones de (2.17) se satisfagan. Las figs 2.1 a 2.3 no solo ilustran la eficacia del MEC, sino también su ventaja de análisis sobre Máxima Entropía, para el caso de información prioritaria significativa [8].

Con una información espectral no prioritaria (S_k = constante) el método MEC tiende a igualar el espectro del método de PL, por lo que se puede escribir la ec (2.18) como el cuadrado del valor absoluto del

recíproco del polinomio predictor en la predicción lineal PL, (ec 2.2), y α es la variancia de la predicción del error,

$$T(z) = \alpha\left[\sum_{n=0}^{M} a_n z^{-n}\right]^{-2}$$

Fig 2.1 Fig 2.2 Fig 2.3

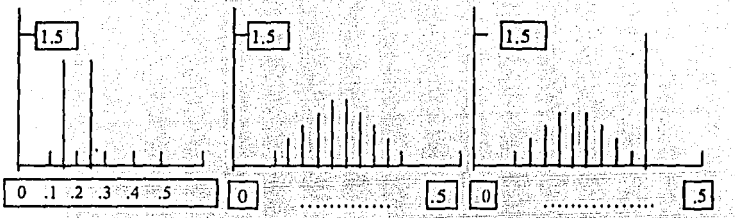


Fig 2.1) espectro original de una onda seno, + ruido-blanco[8];
Fig 2.2) espectro de máxima entropía de datos de la fig 1;
Fig 2.3) espectro de mínima entropía cruzada de la fig. 1; la potencia de la última muestra es un ruido de fondo.

ante una información espectral prioritaria no trivial con S_k = constante y (2.17 y 2.18) se pueden resolver por la introducción de los coeficientes del predictor ai's; las restricciones de correlación son,

$$\sum_{i=0}^{N-1} a_i r_{|k-i|} = \begin{cases} g & k = 0 \\ 0 & k = 1,2,..,N-1 \end{cases} \qquad (2.19)$$

donde g es el factor de ganancia que permite tener a ao = 1, por lo que (2.19) se puede escribir como,

$$\sum_{i=0}^{N-m} a_i r_{i+m} = g s_m \qquad m = M+1, \ldots, N;\ a_0 = 1 \qquad (2.20)$$

donde Sm son los coeficientes de Fourier del reciproco del espectro

principal,

$$s_m = \sum_{k=1}^{N} \frac{1}{s_k} c_{mk}$$

Las ecs (2.15 y 2.16) imponen 2N-M condiciones de las desconocidas ai y N-M de las desconocidas ri [8], para i = M+1,...,N, por el hecho de que (2.19 y 2.20) son no-lineales, el algoritmo deberá ser, en general, iterativo.

2.3 Estimación de parámetros LPC (Linear Predictive Coding)

Para la aplicación de los métodos de Predicción Lineal, PL, en un sistema de reconocimiento en base a los coeficientes de reflexión se cuenta con el método LPC para obtener una síntesis de la señal y poder aprovechar esta codificación para almacenar menos información y gastar menos memoria, con una razón de reconocimiento de patrones aceptable.

En una muestra típica de una envolvente espectral, se aprecia que para una señal voceada la envolvente tiende a cero, por lo que la presencia de las frecuencias de órden superior a los 4 kHz se anula. El modelo LPC da una buena aproximación en frecuencias bajas, pero a frecuencias altas su trabajo es malo.

Para evitar esto hay que pasar la señal voz por un filtro cuya función de transfrencia esta dada por [2]

$$H_P(z) = 1 - az^{-1} \qquad (2.21)$$

Llamada 'filtro de pre-énfasis', que enfatiza las altas frecuencias antes del proceso. Visto de otra manera, el papel de corte espectral es causado por los efectos de radiación de los sonidos bucales, por lo que este se aplica para quitar estos efectos. Los valores típicos del coeficiente 'a', están alrededor de 0.9; un valor de 'a' es de 15/16 = 0.9375. Si x[k] es la señal de entrada, la señal preenfatizada es [2,16],

$$x'[k] = x[k] - 0.9375x[k-1] \qquad (2.22)$$

Después del proceso en el sintetizador, la señal es desenfatizada por medio de,

$$x[k] = x'[k] + 0.9375x[k-1] \tag{2.23}$$

Una representación completa de los parámetros del modelo LPC la proporciona (2.1), donde x[k] depende de los ai's y de las muestras pasadas y que sumadas a v[k] en (2.1) nos da la síntesis de la señal.

Los parámetros del filtro del tracto vocal deben ser determinados (i.e. los coeficientes del filtro ai, y la ganancia G). El error se calcula con,

$$v(k) = \sum_{k} (x[k] - \bar{x}[k])^2 \tag{2.25}$$

Error que es minimizado sobre todas las muestras disponibles. La minimización del error medio cuadrático con respecto a los coeficientes ai, da el siguiente grupo de ecuaciones lineales, llamadas Ecuaciones de Winer-Hopf,

$$\begin{array}{lllll} a_1 r(0) + & a_2 r(1) & + \ldots + & a_m r(m-1) & = -r(1) \\ a_1 r(1) + & a_2 r(0) & + \ldots + & a_m r(m-2) & = -r(2) \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_1 r(m-1) + & a_2 r(m-2) & + \ldots + & a_m r(0) & = -r(m) \end{array} \tag{2.26}$$

En forma matricial, la ecuaciòn de Winer-Hopf queda como,

$$\mathbf{Ra} = -\mathbf{r} \tag{2.27}$$

donde,

$$\mathbf{r} = [\, r(1), r(2), \ldots., r(m) \,] \quad y,$$

$$\mathbf{a} = [a_1, a_2, a_3, \ldots, a_m]$$

$$\mathbf{R} = \begin{bmatrix} r(0) & r(1) & r(2) & \ldots\ldots\ldots & r(m-1) \\ r(1) & r(0) & r(1) & \ldots\ldots\ldots & r(m-2) \\ & & & \ldots\ldots\ldots & \\ r(m-1) & r(m-2) & r(m-3) & & r(0) \end{bmatrix} \quad (2.28)$$

denominada matriz de autocorrelaciones [1 y 2]. El algoritmo para el cálculo de los coeficientes de autocorrelación es la ecución (1.6), que pasamos a repetir aquí,

$$r(i) = r(-i) = \sum_{k=0}^{N-i-1} x[k]x[k+i] \quad (2.29)$$

Como **R** es una matriz no singular, de Toeplitz, se puede invertir, por lo que siempre es posible encontrar los elementos ai de la ec 2.27 de la forma [1, 2 y 16]

$$\mathbf{a} = -\mathbf{R}^{-1}\mathbf{r} \quad (2.30)$$

El método LPC de predicción lineal, usa el algoritmo de Levinson-Durbin [1, 2], se desarrolla con las siguientes funciones recursivas,

$$E(0) = r(0)$$

$$K_i = \frac{r(i) + a_1^{(i-1)} r(i-1) + \ldots\ldots + a_{i-1}^{(i-1)} r(1)}{E(i-1)} \; ; \quad i = 1, \ldots, m.$$

$$a_i^{(i)} = K_i$$

$$a_j^{(i)} = a_j^{(i-1)} + K_i a_{i-j}^{(i-1)} \quad ; j = 1, \ldots, i\text{-}1$$

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

Los $a_j^{(i)}$, j = 1, ... , i, son los coeficientes del filtro de transferencia de un sistema de i-ésimo órden y son,

$a_j = a_j^{(m)}$; j = 1, 2, ... , m donde $\alpha_i = -w_i$

y m es el nivel de proceso.

Este algoritmo nos proporciona importantes parámetros como son los coeficientes de reflexión Ki, i = 1, ..., m, y E(m), la energía del frame, como un producto colateral. La función E(m) es igual al cuadrado de la ganancia G:

$$E(m) = G^2$$

Esta cantidad puede ser codificada para cada uno de los parámetros para la síntesis. Asímismo

$$E(m) = (1 - K_1^2)(1 - K_2^2)...(1 - K_m^2)r(0)$$

De E(m) se puede codificar y transmitir r(0), que es la energía de la estructura de la voz analizada; entonces la ganancia se obtiene de multiplicar r(0) por $(1 - K_1^2)(1 - K_2^2)...(1 - K_m^2)$ mientras dure la síntesis. Esto es recomendable, ya que el modelo de síntesis es menos sensitivo al ruido de la cuantificación de r(0) que al de la cuantificación de G.

El método de Leroux-Gueguen directamente calcula los coeficientes de reflexión, Ki's [2], que aprovecha la gran dinámica y su facilidad de aplicación en maquinas de punto flotante, una PC, o los DSP's; este método soluciona el problema por medio del siguiente algoritmo recursivo,

$$v^j(i) = r(i) + a_1^{(i)} r(i-1) + ... + a_j^{(i)} r(i-j)$$

donde *r(i)* son los coeficientes de autocorrelación. El cálculo de los coeficientes de reflexión se calculan con,

$$K_i = \frac{-v^{j-1}(i)}{v^{j-1}(0)} \qquad i = 1, 2, ... , p$$

$$v^j(i) = v^{j-1}(i) + k_{j-1} v^{j-1}(j-1) \qquad i = -p + j, \quad , p$$

con la condición inicial de

$$v^0(i)=r(i) \qquad i=-p, \quad ,p$$

Este algoritmo esta en la referencia [2].

Con base en el algoritmo LPC presentado hasta aquí, se puede aplicar al reconocimiento de señales analógicas, señales que se generan en nuestro entorno y que es deseable reconocer por medio de un sistema que desarrolle la labor de reacción que hace un sujeto ante un comando de voz.

CAPÍTULO TERCERO

EL RECONOCIMIENTO DE SEÑAL-VOZ

3.1 Esquema general para el reconocimiento automático

Módulo de reconocimiento

El reconocimiento de una señal-voz se apoya en un sistema de codificación de la señal, cuya finalidad es obtener los parámetros que caracterizan la señal codificada, estos parámetros se definen con las siguientes variables, los ri's, los Ki's y los ai's.

Los ri's son los coeficientes de autocorrelación, que se calculan con la ecuación (1.6), y van desde r_0 hasta r_p, donde p es el grado del modelo AR. Con el objeto de obtener un modelo normal, los coeficientes se normalizan en base a que los ri's son decrecientes conforme i varia de 0 a p.

Los Ki's son los coeficientes de reflexión, base de un filtro lattice, filtro que al ser excitado por una señal de ruido blanco reproduce la señal de la cual se adquirieron. Los coeficientes de reflexión son los elementos principales en el reconocimiento de la señal-voz, ya que la estabilidad que los caracteriza es su principal ventaja. Los Ki's, en general, deben mantenerse en el intervalo unitario de valores, $|Ki| < 1$, para que la función de transferencia que representan sea estable.

El método de Levinson-Durbin proporciona, además de los coeficientes ai's, componentes del polinomio, solo polos, generador de la señal-voz de la que fueron obtenidos, los Ki's y como producto colateral la ganancia G; la estabilidad de los ai's no es tan precisa como la de los coeficientes de reflexión, por lo que en este proceso de detección no se emplean, pero se pueden usar en la síntesis de la señal.

Se desarrolló un programa, que se llama GRAF17.C, compilado en Turbo C, el cual puede adquirir una señal-voz vía una tarjeta instalada en una PC, en este caso SOUND BLASTER (SB16), para pasar la señal digitalizada por el proceso que se presenta en la fig 3.1

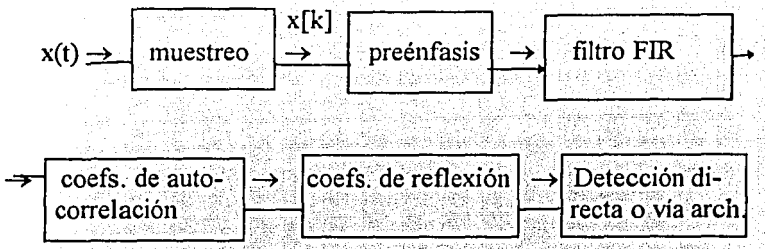


Fig.3.1 Diagrama de flujo del sistema del módulo de reconocimiento para la detección de una señal-voz.

El muestreo de la señal, vía tajeta SB16, es adquirido desde el 'prompt' del sistema operativo de la PC, en el directorio que maneja la tarjeta SOUND BLASTER, que es C:\SB16\VOCUTIL>, con el comando VREC; las opciones de grabación son,

vrec nombre-del-archivo.VOC /r:16 /a:MIC /s:6000 /t:2

donde 'r' es el nivel de cuantización, 16 bits, 'a' es el dispositivo de grabación, 's' la frecuencia de muestreo y 't' el tiempo en segundos.

Para poder reproducir una señal pregrabada se usa la opción VPLAY como sigue,

vplay nombre-del-archivo.VOC /t:2 /q

donde 'q' evita los mensajes en pantalla y agiliza la operación.

Las opciones se visualizan en caso de olvidarlas en la pantalla de ayuda al teclear VREC sin ningún parámetro; igualmente el comando VPLAY, que reproduce la señal grabada, proporciona los parámetros en pantalla al ser tecleado el solo. En este caso solo se requiere el nombre-del-archivo.VOC y el tiempo. Ambas opciones pueden no presentar mensajes mientras se ejecuta alguna de estas; se recomienda usar el parámetro, /q:, solo en VPLAY, ya que en VREC los mensajes

permiten vocear a tiempo la señal. El proceso se ejecuta con el programa GRAF17 automáticamente.

Para reconocer una señal, ver apéndice A, se ejecuta el programa GRAF17 en la opción,

[V]oz, reconocimiento y reproducción

Despues aparece un menú en el que se tienen 4 opciones y si la opción es '2' se ejecutan los siguientes pasos

Adquisición de la señal via micrófono y archivación de ésta. En este momento la señal es muestreada y almacenada en un archivo con extensión VOC, para después ser leída por el programa y sensar el principio de voz por medio de cierto nivel de la señal. A partir de este punto se eligen 5050 muestras, cantidad suficiente en consideración a la voz más larga.

Después se aplica a la señal un filtro de pre-énfasis, que enfatiza las altas frecuencias de voz, en donde el método LPC hace un trabajo un poco deficiente [2].

El siguiente objetivo es el de pasar a la señal por un filtro digital pasa-banda, con frecuencias de 150 a 4 000 Hz, que filtra la señal eliminando componentes de alta frecuencia, fuera de los rangos de la voz; este filtro es de orden ocho, en consideración a que un mayor orden del filtro, para una PC, retrasa mucho el proceso. En el momento en que se desee implantar en un DSP, el orden del filtro se puede incrementar sustancialmente, con lo que se espera una mejor limpieza de la señal.

Ya filtrada la señal, se pasa a la opción de la adquisición de los parámetros ri's con un sistema de frames de 30 ms de duración con traslapes de 10 ms, esto da 'frames' de 256 muestras con traslape de 85 muestras. Se probó con traslapes de 15 ms, pero el resultado de la detección decreció considerablemente; resultados imprecisos también se obtuvieron con traslapes menores a 10 ms.

En este momento la adquisición de los parámetros de cada frame se guardan en un archivo llamado XCOM.BAS, al que se le agregan los bloques de parámetros de cada frame codificado. Los parámetros que se guardan en este archivo son los ai's y los Ki's, en ese orden.

Las consideraciones para los Ki's que aquí se hacen son que si los Ki's caen dentro del intervalo (±1, ±1.2] se force a los coeficientes a penetrar en el intervalo unitario, por otro lado si caen fuera del intervalo previsto, la diferencia sería muy grande y este bloque queda automáticamente desechado, por lo que no es necesario descalificar toda la voz por este problema y continuar con el análisis del comando registrado.

Ya obtenidos los parámetros de la señal voceada, se pasa al proceso de comparación con la base de datos previamente formada de los comandos que se desea reconocer, en este proceso se hace una comparación de ámbos conjuntos de datos, los del comando de prueba y los del diccionario, comparandolo con cada uno de la base y considerando el criterio de la mínima distancia entre los datos, la menor distancia que se obtenga en esta operación se toma como la respuesta al reconocimiento del comando, como se vera en el inciso de distancias.

Existe cierta proximidad de algunos comandos, por los que los resultados de detección no fueron satisfactorios; teniendo en cuenta esto, se optó por comparar las mínimas distancias de cada diez coeficientes y darle un peso al comando de la mínima distancia encontrada, de tal forma que el comando con más peso será la mejor opción.

3.2 Módulo de control

La fig 3.2 muestra el diagrama de bloques del sistema de desarrollo que se requiere para la detección de los comandos de voz. En esta figura se pueden apreciar los elementos necesarios para cargar el programa que hace la detección directa de la señal de voz, vía un micrófono, el que se colocará en la cabeza del usuario; el programa se instalará en una eprom externa o en una eprom del propio 68HC11, o

bien, pregrabada en PROM, todo depende de la longitud del programa o de la produccion especifica del 'chip'.

Dicha señal será digitalizada por el ADC y enviada al DSP56001 o su equivalente, hasta lograr que el comando sea reconocido y envia la solucion al microcontrolador, 68HC11, que ejecutará las señales pertinentes en el sistema motorizado de la silla.

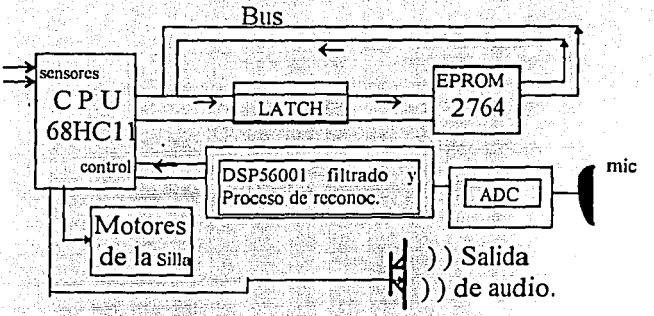


Fig 3.2 Sistema mínimo de desarrollo con EPROM externa.

El módulo de control tendrá un sistema de sensores que permitan al microcontrolador conocer el estado del sistema, como de que los motores funcionen correctamente o el control de obstáculos, control de deformaciones del terreno, banquetas, pendientes y control de potencia, fin de banquetas, luces direccionales y en el caso de terrenos inclinados compensar la dirección de la silla dándole a un motor más energía que al otro. Un problema se presenta en el momento en que alguna rueda pierde contacto con el piso, esto se debe sensar para que la silla no se salga de curso, etc.

La entrada de la señal-voz es vía micrófono, el que se espera sea de estricto direccionamiento, lo más inmune posible al ruido del medio ambiente, como son los ruidos caseros como el radio, la tele o las voces de la familia, como los niños o también los animales; en la calle, las bocinas de los carros, gritos y toda clase de ruidos callejeros.

Se pretende mantener, por medio de un comando, el estado del sistema, esto es, que si el movimiento de la silla es en línea recta, el comando propio mantendrá en este estado a la silla o si esta en reposo también mantendrá dicho estado, solo los comando de acción lateral tendrán la duración suficiente para ser ejecutados, como DERECHA 60 GRADOS, etc.

Por la salida de audio se reproduce la señal detectada, con el fin de que el usuario pueda saber que comando se detectó y poder corregirlo en caso de una detección errónea.

3.3 Cálculo de las distancias

La función comparbas() del programa de reconocimiento se apoya en una base de parámetros LPC previamente calculados y almacenados como archivos de referencia.

Los parámetros calculados de la señal se comparan con cada grupo de la base, punto a punto, como se expresó antes, obteniendose el comando con mayor peso y mejor probabilidad de acierto. En esta comparación se toma la decisión en consideración a la mejor aproximación con una eficiencia razonable en la decisión.

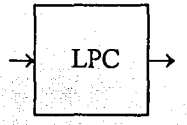
Considérese que el método LPC es perfecto para patrones fijos. Los principales métodos para el cálculo de distancias son [2 y 16],

1) Método de la Distancia Euclidiana
2) Método de la Distancia Logarítmica Euclidiana
3) Método de la Mejor Razón de Itakura.

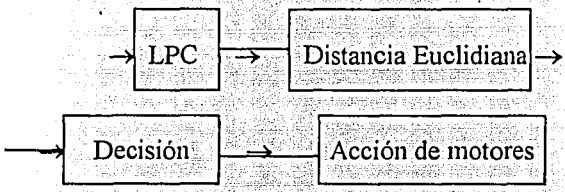
Los tres son confiables, sin dejar de reconocer la existencia de otros, como el de la razón de señal a ruido SNR y el de Distancia Logarítmica Espectral, que a fin de cuentas proponen que el cálculo debe cumplir con la mínima distancia entre las muestras.

La distancia Euclidiana

La base para cada uno de los siguientes esquemas la llamaremos 'LPC', correspondiente al diagrama de bloques de la fig 3.1, que será sustituido por,



El primer esquema estará conformado por,



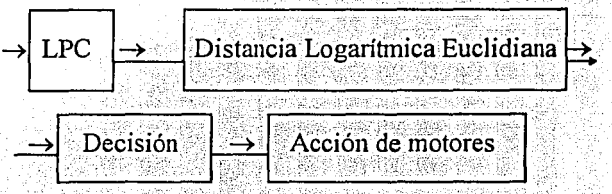
Este primer esquema se basa en la función de la distancia Euclidiana para la toma de decisiones en la comparación de la señal de prueba con la base de señales que se almacenaron como las muestras base; estas bases se formaron con los parámetros Ki's, tomando diez coeficientes para cada ventana de 256 muestras y obteniendo archivos de 290 muestras con estos parámetros, cantidad que abarca 5050 muestras que determinan una voz. La función es [2]

$$d = \sum_{i=0}^{N-1} \sqrt{(x[i] - \bar{x}[i])^2} \qquad (3.4)$$

en donde x(i) es un coeficiente de la base y $\bar{x}(i)$ los de la muestra de prueba. La distancia que resulte mínima es la que se toma como la decisión correcta.

La distancia logarítmica Euclidiana

La distancia logarítmica Euclidiana esta conformada como sigue en diagrama de bloques [2],



en el que el comparador es la distancia logarítmica Euclidiana, esta se basa en la siguente función,

$$\mathrm{LAR}i = \ln\frac{1+k_i}{1-k_i} \tag{3.5}$$

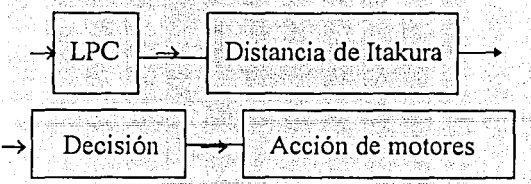
donde los Ki's son los coeficientes de reflexión y LAR*i* es la i-ésima variable que es tomada como la mejor opción de las razones de area-logarítmica, con las que la distancia logarítmica Euclidiana, para cada par, esta dada por,

$$d_{LAR} = \left[\sum_{i=1}^{P}(LAR_{1i} - LAR_{2i})^2\right]^{1/2} \tag{3.6}$$

La estimación esta muy cerca de la calidad que muestra el método de la mejor razón de Itakura.

La distancia de Itakura-Saito

Este esquema tiene el siguiente diagrama de bloques,



Para este cálculo se consideran los dos siguientes vectores de parámetros a*i*'s, en el que el primero es de la señal de prueba y el segundo de una de las bases

$$a_1 = \left[1 \; a_{11} \;\; a_{12} \;\; \ldots\ldots\ldots \;\; a_{1p}\right]$$

$$a_2 = \left[1 \; a_{21} \;\; a_{22} \;\; \ldots\ldots\ldots \;\; a_{2p}\right]$$

donde 'p' es el órden de polinomio generador de la señal en base a los ai's; y para la voz de prueba y la voz de la base, las matrices de autocorrelaciones serán $\mathbf{R_1}$ y $\mathbf{R_2}$ [2]

Los ri's son el modelo del tracto vocal Hi. Un caso como este es la comparación del modelo del tracto vocal H_1 con el H_2 con sus vectores de coeficientes y matriz de autocorrelaciones, a_1, $\mathbf{R_1}$ y a_2, $\mathbf{R_2}$ respectivamente.

Se define como la razón de más probabilidad a,

$$d_{LR_1} = \frac{a_2^T R_1 a_2}{a_1^T R_1 a_1} \tag{3.7}$$

$$d_{LR_2} = \frac{a_1^T R_2 a_1}{a_2^T R_2 a_2} \tag{3.8}$$

en donde ai y Ri son los elementos definidos arriba.

Los mejores resultados para las razones de Itakura están dadas por los logaritmos de (3.7 y 3.8), como,

$$d_{LLR_i} = 10\log(d_{LR_i}) \quad i = 1,2; \tag{3.9}$$

y se expresa en dB. Las expresiones (3.7) y (3.8) no son simétricas y se desea que los cálculos de las distancias lo sean, por lo que se puede usar,

$$d_{LRS} = \frac{(d_{LR_1} + d_{LR_2})}{2} - 1 \qquad (3.10)$$

y da la propiedad de simetría deseada.

Las funciones recursivas para la solución computacional a ésta proposición se desarrollan en base a los coeficientes de autocorrelación r(i) y a la secuencia de autocorrelación de los coeficientes del filtro $r_a(i)$ la siguiente manera [2, 16],

$$r(i) = \sum_{k=0}^{N-i-1} x(k)x(k+i) \qquad i = 0, 1, \ldots, p \qquad (3.11)$$

$$r_a(i) = \sum_{k=0}^{p} a_k a_{k+i} \qquad i = 0, 1, \ldots, p \qquad (3.12)$$

con $a_0 = 1$.

Por lo que la operación matricial, **aRa**, se calcula con la siguiente función iterativa

$$a^T Ra = r_a(0)r(0) + 2\sum_{i=1}^{p} r_a(i)r(i) \qquad (3.13)$$

Dinamic Time Warping

Este método, de dominio en el tiempo, se caracteriza por una normalización en el alineamiento de la señal de prueba que asume que la variación de la razón de locución es proporcional a la duración de las palabras y es independiente del sonido proporcionado por el locutor; por lo que la evaluación de las medidas de la distorsión se encuentra en la diagonal principal de un rectángulo formado por la normalización de los puntos de los patrones de prueba, representando cada punto de la diagonal a la distorsión [21]. Las rígidas restricciones del método de la razón de las fluctuaciones de las invariaciones del locutor no modelan

adecuadamente la situación real de las palabras voceadas por lo que se necesita un mejor alineamiento de los patrones de prueba.

En suma el método LPC presenta las mejores características de parametrización de la señal-voz [16] en su modelo de cuasi-estacionariedad y permite de una manera simple el reconocimento de estas señales.

3.4 Selección del esquema

En el análisis y pruebas de los esquemas presentados arriba se encontró que la mejor solución, por su simplicidad, precisión y características, fue el esquema en el que domina la distancia Euclidiana.

No se eligió la Distancia Logarítmica por los problemas ya comentados que presenta el caso de que algunos coeficientes de reflexión al ser calculados sobrepasan a la unidad, por lo que al calcular las razones LARi con (3.5) se incurre en un error de cálculo logarítmico. Si los parámetros Ki's son mayores que la unidad no se puede hacer esta operación y esto sucede algunas veces en el cálculo de alguna del las 29 ventanas que se obtienen para cada palabra, y en algún momento el proceso no se puede calcular; por esto se desechó este esquema pues estos cálculos producen un estado de sobreflujo.

Se puede salir de este proceso en el momento en que se encontrase que algún coeficiente de reflexión se saliera del intervalo unitario, pero esto obstaculiza la detección y muchas palabras serían eliminadas en el intento.

Con el cálculo de la mejor razón de Itakura no se obtuvo la precisión necesaria pues el porcentaje de palabras detectadas con exactitud esta bajo el 50 % y no lo hace viable. Este esquema se debe usar con una base formada por un método que promedie los coeficientes de unas 30 palabras de la misma voz para formar una buena base maestra de dicha voz.

El esquema seleccionado tiene tres particularidades

1. Al generar la base de datos se escoge la voz que detecta la mayoría de las palabras que se usaron para obtener la mejor base de datos LPC, datos que mejor se apegan a todas las voces muestreadas, es decir esta base se apega más al cálculo de la distancia de entre los parámetros LPC de todas.

2. Los pasos que se usaron para el cálculo de las bases se siguen para el de la muestra de prueba y el cálculo se hace con 250 datos de cada base, comparados de 10 en 10, de los del comando voceado, de los 290 que se tienen registrados; esto se hace para eliminar la redundancia de los finales de las palabras, como son los sonidos no voceados, como la 'ssss' la 'nnnn' y los silencios de las palabras de menor tiempo.

3. Las condiciones que las voces presentan en el cálculo de cada ventana tienen que ser condicionadas por un intervalo de selección de ventana, pues para ventanas semejantes en diferentes voces se le asignará un peso a cada una de ellas por las características de semejanza; por ejemplo las palabras ADELANTE, ATRÁS y ALTO tienen la misma vocal inicial y la mejor aproximación podría caer en cualquiera de las tres, por lo que en el primer bloque y considerando cierto rango, se le daría un peso a cada una de ellas.

Puesto que el esquema de la distancia Euclidiana presentó la solución con más precisión en la detección de las palabras, cerca del 90%, se toma este para desarrollarlo en su máxima expresión, llevándolo a su mejor configuración.

Estructura y sistema

El programa esta desarrollado en Turbo C para la demostración de su validez en una PC; posteriormente se trasladará a programa de bajo nivel por medio de un traductor para ser implantado en un DSP56001 como unidad de procesamiento digital de la señal-voz y con un 68HC11 como controlador.

Ya limpia la señal de características indeseables, se procede al reconocimiento del principio de señal, como ya se explico con anticipación; luego se le sacan a estas muestras los parámetros LPC con ventanas de 256 muestras y con un traslape del 30% (85 muestras en el caso).

En el momento en que se logra una detección se procede a la reproducción de la señal detectada por medio de una bocina. El comando se puede confirmar con un 'SI' o con un comando especifico para diferentes opciones de alguno de los primeros comandos, y al reafirmar el comando sacar vía bocina el comando que se encuentra en acción. Los primeros comandos son cinco y los comandos secundarios son siete; en la siguiente relación se presenta el orden y opciones del segundo comando,

1er comando	2o comando
ALTO	con dos opciones SI y NO
DERECHA *IZQUIERDA*	con cuatro opciones NO, UNO, DOS, TRES.
ADELANTE *ATRÁS*	con tres opciones NO, LENTO, RÁPIDO.

Esto se presenta en la pantalla de la PC con salida de audio para proporcionar evidencia suficiente para continuar con el mejoramiento del proyecto y poderlo implantar en el sistema de desarrollo[12].

Los sistemas de cuantificación dependen de los bits de la 'palabra' que maneje la tarjeta; existen tarjetas de 8, 12 o 16 bits; la tarjeta que se usó para esta tarea es de 8 y 16 bits.

Detección de fallas

En la detección de fallas del sistema o fallas del medio en el que la silla de ruedas se desenvuelve, se pueden clasificar,

Fallas de tipo prioritario (FTP).
Fallas de tipo secundario (FTS).
Fallas de tipo indiferente (FTI).

Entre las de tipo prioritario se encuentran las fallas mecánicas o las de una falsa detección, de las que depende la salud o la vida del operador; estas fallas son en sí de extrema prioridad. Para cualquiera de estas se debe contar con una segunda opción que asegure al usuario el no tener un percance.

Entre las FTS se clasifican aquellas que por fallas del sistema electromecánico se impida la correcta operación del sistema de conducción de la silla, ya sea por fallas en suministro de energía, por algún conductor desconectado, etc.

Entre las de tipo FTI se pueden encuadrar las fallas de tipo de control debido al ruido ambiental, y que por este se quiera realizar un comando no predicho.

El proceso de detección de fallas en un sistema de desarrollo se basa en la suposición de una posible falla del mecanismo de la silla. Este paso en el diseño del control por microprocesadores es de importancia básica por la delicadeza del control del sistema de que se trata. Para ello se debe aplicar en el proceso un controlador con discriminación, de tal manera que el sistema pueda saber que hacer en el caso de una falla común en alguna de las partes a controlar. Supóngase que se detecta una falla en el comando 'DERECHA'; entonces el sistema detecta que no responde a dicha orden y debe tomar la decisión de 'ALTO'; en el caso de los comandos DERECHA o IZQUIERDA se puede solucionar con tres Izquierdas de 90 grados; en caso de bloquearse se soluciona con un sistema de insistencia para acomodarse hasta dar el giro requerido.

Si existe dificultad en dar tres veces a la izquierda, el sistema no necesariamente se tiene que bloquear en 'ALTO', salvo que la falla sea de índole FTP. La respuesta a cualquier tipo de estimulación que pueda causar un ruido también se debe tomar en cuenta, ya que de ello depende del cambio que el ruido cause, además de que solo el conductor pueda dar ordenes al sistema; considérese que alguien pudiese gritar 'ALTO' en un momento indeseado, en este caso se pondría en peligro al conductor, etc, esto dependerá de la cercanía del micrófono y por lo tanto de la energía con que entre la señal, punto contemplado en la detección de principio de señal.

Envío de mensajes

Se puede disponer de un sistema de síntesis de voz o de un 'display' para el envío de mensajes en el caso de una detección de falla de tipo mecánico o de comandos alterados por la captación de 'ruido' colateral del medio, lo que ya se contemplo con la confirmación del comando con un 'SÍ'.

Se puede considerar que la palabra 'ALTO' sea muy importante; 'ALTO' es prioridad UNO y en ese orden de comprobación debe ser la primera en discriminarse. En los programas del sistema se tendrán archivo de la energía total de cada palabra y los algoritmos propios para la eliminación de la redundancia en la voz, a fin de tomar una decisión adecuada y en caso de duda decidir 'ALTO', y mediante un mensaje informar al usuario lo que sucede.

Periféricos

El apoyo del sistema mínimo está en los canales de entrada/salida de información, la que entra por micrófono o por alguno de los sensores y del principal apoyo que es el DSP56001. El micrófono debe ser direccional con objeto de captar la máxima energía del usuario. El sintetizador de voz debe transmitir por medio de audífonos un mensaje del comando que se efectuará o de un display para el caso de un minusválido auditivo. En el caso del uso del display se debe llamar la atención del usuario, minusválido auditivo, con un foco que 'centellee' y que puede simplemente decir 'NO' si no había indicado ningún comando.

Diccionario de comandos y sus prioridades

Palabras claves	Prioridad
ALTO	UNO
DERECHA	DOS
IZQUIERDA	DOS
ADELANTE	DOS
ATRÁS	DOS
SI y NO	UNO
UNO	TRES
DOS	TRES
TRES	TRES
RÁPIDO	TRES
LENTO	TRES

Estas prioridades pueden ser modificadas de acuerdo con las normas del sistema. También es factible buscar otras opciones de palabras contundentes para una mejor precisión en la detección del comando o también asignar vocales o consonantes; el ruido colateral pasaría a ser menos crítico en estas condiciones.

CAPÍTULO CUARTO

EVALUACIÓN DEL DESEMPEÑO DE RECONOCIMIENTO

4.1 Introducción

La forma de la voz está constituida por tres tipos de sonido

Voceado
No-voceado y
Silencio

Los sonidos voceados están representados por vocales y la mayoría de las consonantes y los sonidos no-voceados por la mmm..., la nnn..., la sss... que son nasales o siseados y por silencios entre las sílabas componentes de una palabra. Todo esto da cierta incertidumbre en el cálculo de los parámetros LPC, ya que en esencia, estos dependen de los coeficientes de autocorrelación.

Sabiendo que los coeficientes de autocorrelación son obtenidos por una matriz de Toeplitz, implicando invertibilidad, los coeficientes de los sonidos no-voceados quedan muy cercanos a cero, dando a la matriz de correlaciones una inestabildad que lleva a que los parámetros LPC no tengan la precisión deseada sacando a los coeficientes de reflexión del intervalo unitario, $-1 < ki < 1$, por este motivo, algunos de los métodos del cálculo de las distancias son deficientes. Los silencios se presentan entre las sílabas y su temporalidad varia de acuerdo con la frecuencia con que el locutor habla.

4.2 Los métodos aplicados

Los métodos probados para la detección de la señal se presentaron en el capítulo tercero, en donde se expuso que la precisión de algunos métodos no resultó satisfactoria, solo el método de la distancia Euclidiana pudo compensar la dificultad que representa el cálculo de los parámetros LPC de Levinson-Durbin usado en esta tesis.

En el método de la distancia de Itakura los resultados en la detección de la señal de voz fue del 50 al 60 % y en algunos casos del 30 %, por lo que fue descartado. El método de la distancia logarítmica Euclidiana

encontró su debacle en el hecho de que para el cálculo logarítmico, si los coeficientes de reflexion se salen del intervalo unitario, la función logarítmica se bloquea. En el momento en que algún coeficiente se sale del intervalo unitario hace que la diferencia resulte negativa al ser sustraída de la unidad, y el cálculo logaritmico bloquea la función del computador.

La distancia Euclidiana hace la diferencia y la eleva al cuadrado, esto permite detectar, en general, las voces sin producir un rompimiento en los cálculos del CPU por desborde; aparte de que los cálculos, comparados con los otros métodos, son mucho más simples, proporcionando al programa un mejor desempeño.

La fig 4.1 muestra el diseño del filtro pasa-banda que se le aplicó a la señal de voz, considerando que el orden del filtro es 8, se puede apreciar que el decaimiento es a partir de los 2200 Hz. dando como resultado el rechazo de componentes de la señal voz, que no se desean perder; el intervalo de frecuencias del filtro se eligió de 150 a 4000 Hz, con el fin de eliminar el ruido de la línea de alimentación comercial y de los balastros de las lamparas de neón, además de los 3.5 kHz para las máximas frecuencias de voz.

Las figs 4.2 y 4.3 son las gráficas de las voces UNO y ADELANTE, y las 4.4 y 4.5, sus respectivos espectros en frecuencia que se muestran con el objeto de apreciar sus diferencias espectrales. En la plabra ADELANTE se pueden apreciar los silencios entre la 'E' y la 'L' y entre la 'N' y la 'T'. Por la compresión de la señal al ser graficada, no se aprecia bien la forma de las ondas de las vocales y de las consonantes; para poder valorar estas formas se debe recurrir al sistema gráfico del programa GRAF17.EXE, como se explica en el apéndice A.

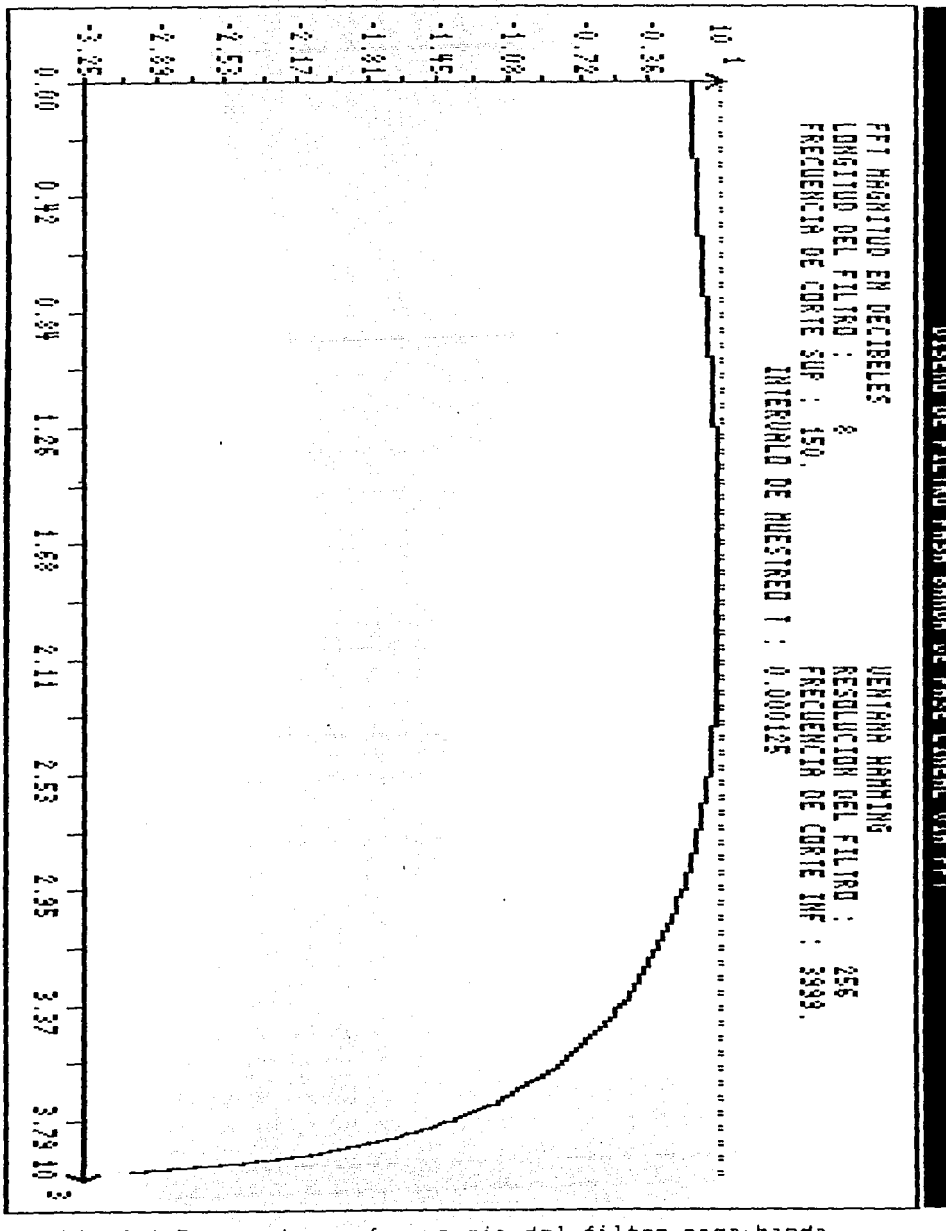


Fig 4.1 Respuesta en frecuencia del filtro pasa-banda usado en el programa GRAF17.EXE para el análisis de las señales de voz y su reconocimiento.

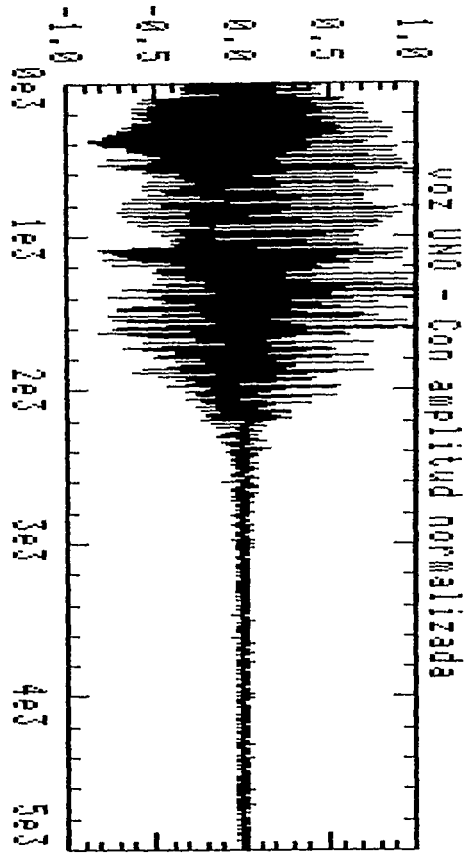


Fig 4.2 Voz UNO, de amplitud normalizada. Esta palabra se caracteriza por ser de corta duración con un silencio de 400 muestras, entre la 600 y la muestra No. 1000.

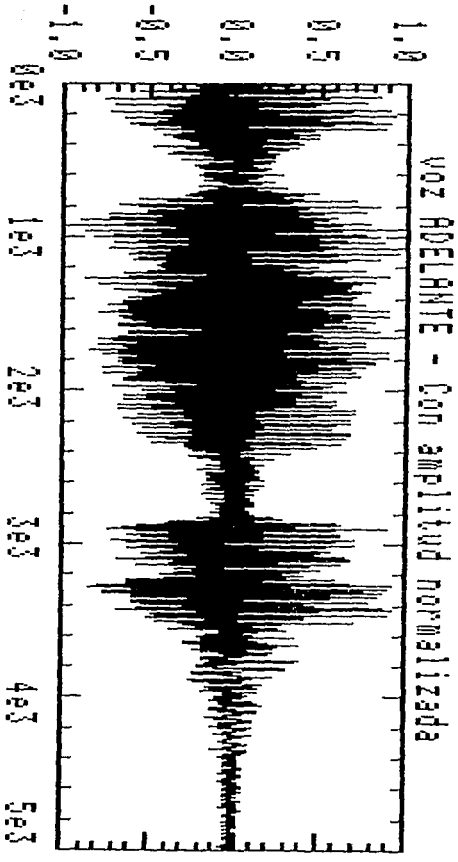


Fig 4.3 Voz ADELANTE, de amplitud normalizada. Esta VOZ se compone de cuatro sílabas. Se aprecia claramente la separación que provocan los silencios.

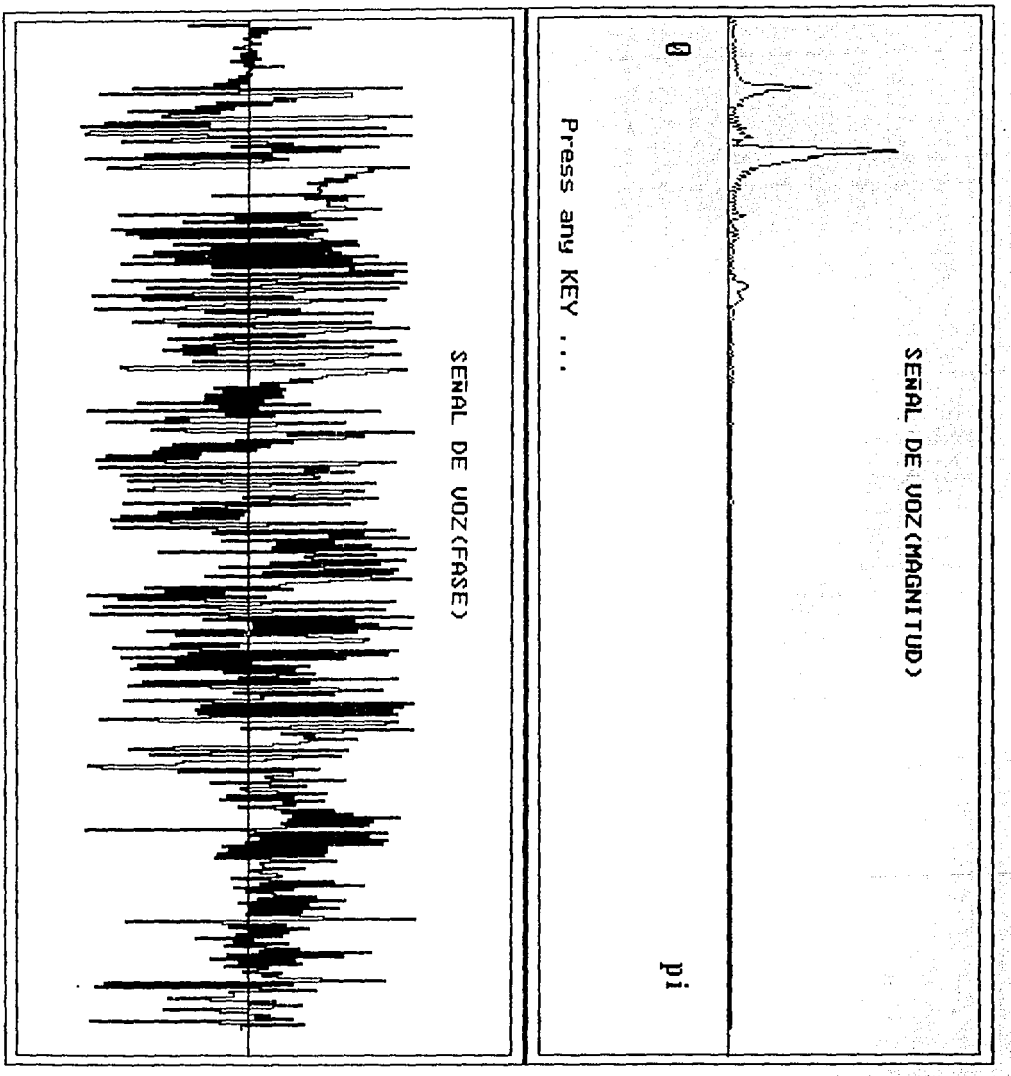


Fig 4.4 Espectro en frecuencia de la voz UNO y de su fase. Se puede apreciar la poca energía de la señal y la prescencia de sus dos principales formantes en bajas frecuencias. Voz de un hombre adulto.

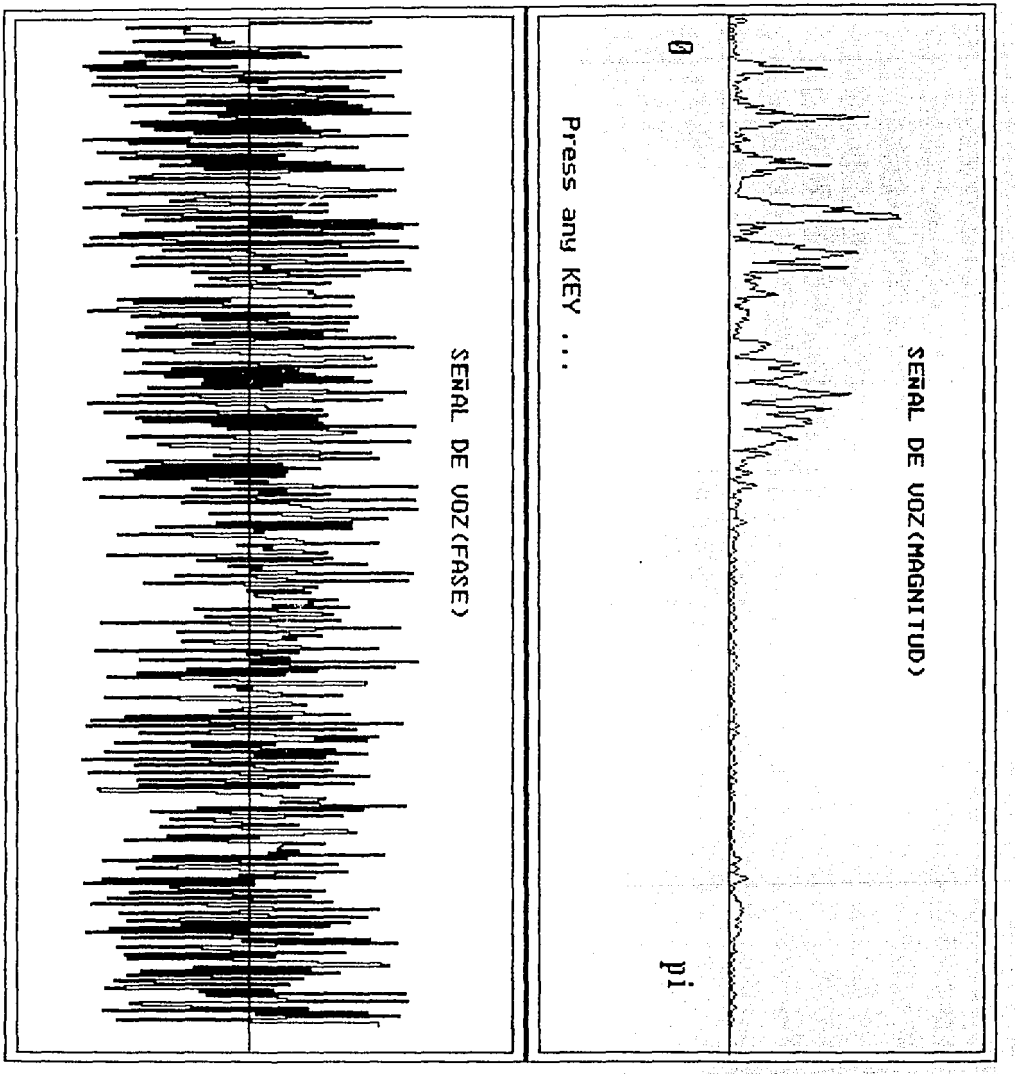


Fig 4.5 Espectro en frecuencia de la voz ADELANTE y su face. En este caso la energía de la señal es mucho mayor y la presencia de mayores frecuencias se hace presente Voz de hombre adulto.

En las gráficas del programa mencionado se aprecian los silencios, las consonantes fuertes y la especial periodicidad de las vocales, con una forma característica cada una de ellas. La periodicidad de la vocales permite definir su forma y reconocer su patrón con alta precisión. En los sonidos de las consonantes fricativas, como la 'fff...', la 'sss...', o las nasales como la 'mmm...' o la 'nnn...' el parecido con los silencios es de hacerse notar, solo que en el silencio la energía es menor, lo que da una pauta para la detección de estos elementos, componentes de la voz.

En las gráficas de las voces se pueden apreciar las diferencias de cada una de ellas, lo que es de esperarse, permitiendo establecer parámetros propios para cada voz, lo que se aprovecha en benéfico de una buena base de datos.

Las gráficas de las magnitudes de los espectros y de sus fases muestran una marcada diferencia, que se puede aprovechar en el reconocimiento de voz o para hacer una síntesis, ya que los elementos principales están bien definidos, esto reproducirá una buena señal al aplicar la transformada inversa.

4.3 Evaluación para los comandos de la silla de ruedas

La evaluación del sistema de reconocimiento por medio de una PC se baso en las siguientes proposiciones

1) La detección debe simular lo mejor posible el futuro desempeño del sistema de reconocimiento.

2) La secuencia de reconocimiento de comandos será aleatoria.

3) Las pruebas de reconocimiento se deben hacer en un medio natural; significa que las pruebas no serán protegidas contra el ruido de medio ambiente para una mejor evaluación de los comandos en su medio natural.

La evaluación de los comandos se hízo con un micrófono omnidireccional, esto lleva a que la captación de la señal se encuentre viciada por toda variedad de ruidos ambientales, caso no propio para una simulación; más, de acuerdo al sistema que se pretende implantar en la silla de ruedas, que tendrá un micrófono aislado y direccionado, en donde se espera que la energía de la señal proveniente de la boca sea definitivamente superior al ruido del medio ambiente.

Es conveniente recordar que la silla no será usada en medios excesivamente ruidosos. Esta aseveración no tiene una connotación peyorativa, sino realista; por un lado, se pretende que el usuario se encuentre en un ambiente hogareño y por el otro se considere que una persona en condiciones de cuadriparecia severa no podría deambular por las calles de la ciudad expuesta a toda clase de ruidos, que aunque son impulsivos, son de tal energía que desquiciarían el detector con los peligros correspondientes.

Con estas consideraciones se pasó a la experimentación con dos opciones

a) medio ruidoso
b) medio normal

En el medio ruidoso se encontró con un vicio de la señal que impidió un buen desempeño del detector, en donde una de las palabras más distorsionadas por el ruido fue 'IZQUIERDA', que por supuesto tuvo una tasa de error muy grande; de 30 veces que se dijo la palabra izquierda en este medio, 14 fueron errores, lo que da un 47 % de aciertos, considerado para este comando, muy bajo. En el mismo medio la palabra 'DERECHA' tuvo una tasa del 99 % de aciertos. Los resultados se presentan en la Tabla 4.1.

Las pruebas se efectuaron como sigue: Las palabras, DERECHA e IZQUIERDA, se pronunciaron tres veces cada una, para un total de 75 veces cada una, señalando que la voz DERECHA presentó una sola falla en el medio ruidoso, en el bloque de las primeras 10 columnas; la voz IZQUIERDA si fue deformada por el ruido, como se pude ver en la Tabla 4.1, en general, el desempeño de la voces en ambiente ruidoso

comandos para la silla y en los dígitos, mejorando en un medio con un ruido no muy fuerte, normal se puede decir.

Por su lado, los ventiladores provocan un desvío de las ondas sonoras y distorsionan mucho la señal

TABLA 4.1

voz	→ medio ruidoso ←									→ medio de ruido normal ←															
Derecha	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	74 a
Izquierda	1	0	2	0	1	3	2	3	1	1	1	2	3	3	2	3	3	3	2	2	3	3	2	3	3	51 a
Alto	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	24 a
Adelante	1	1	1	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	22 a
Atrás	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	25 a
Uno	0	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	43 a
Dos	1	1	0	0	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	42 a
Tres	0	1	1	1	1	2	1	2	1	1	2	2	2	2	2	2	2	2	1	1	2	2	2	2	1	40 a
Rápido	-	1	-	0	-	1	-	-	-	1	-	1	-	-	-	1	-	1	-	-	-	1	-	1	-	11 a
Lento	1	-	1	-	-	-	0	-	1	-	1	-	1	-	1	-	1	-	1	-	1	-	1	-	1	11 a
Si	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	23 a
No	2	1	1	3	3	0	1	1	1	1	1	0	1	1	1	-	-	1	2	2	-	-	2	1	-	26 a

Tabla 4.1 Los valores encasillados son aciertos, la última columna es el total de aciertos.

Las voces UNO, DOS Y TRES, que pertenecen a la confirmación de las voces DER. e IZQ. se pronunciaron dos veces, que en el medio ruidoso presentan varios errores, pero en el medio normal se comportan bien.

Para la voz ADELANTE el comando de rectificación fue RÁPIDO o LENTO alternadamente, cundo falló la detección se canceló con la voz NO, los guiones significan opción sin usar.

La voz ALTO tiene como confirmación el comando SÍ, del cual se puede ver que falló dos veces en 25 intentos y ALTO una vez, esto asegura uno de los principales comandos.

Los porcentajes para cada palabra están dados por la Tabla 2

TABLA 4.2

derecha	99 % de aciertos, 74 de 75
izquierda	68 % de aciertos, 51 de 75
alto	96 % de aciertos, 24 de 25
adelante	88 % de aciertos, 22 de 25
atrás	98 % de aciertos, 24 de 25
uno	86 % de aciertos, 43 de 50
dos	84 % de aciertos, 42 de 50
tres	80 % de aciertos, 40 de 50
rápido	88 % de aciertos, 08 de 09.
lento	92 % de aciertos, 11 de 12.
si	92 % de aciertos, 23 de 25.
no	87 % de aciertos, 26 de 30.

y el total de palabras fueron 451.

Para 451 palabras pronunciadas se obtuvo un desempeño del 86.69 %. Considerando pruebas en ruido y en silencio, se reconocieron un total de 391 palabras.

Para el medio ruidoso tenemos un total del 78.8 % de aciertos, o sea en 170 palabras se tuvieron 36 errores.

En el medio normal, o sea con poco ruido, un ruido que no altera las palabras en general, solo cuando llega con mucha energía, tenemos un desempeño del 94.5 %. De 255 palabras hubo 14 erróneas, con la característica de que se corrigieron con el comando de rectificación 'NO', el cual tuvo un desempeño, en general del 87 %, como se presentó en la Tabla 2; en medio ruidoso, de 14 veces falló 2, lo que da un desempeño de 85.7 %; por otro lado, en medio normal, se obtuvo un desempeño del 99% para la voz NO.

Es evidente que a mayor número de palabras se obtiene una mejor tasa de desempeño, esto se puede ver en el mejoramiento de la tasa general de desempeño, comparada con la tasa de desempeño en un ambiente ruidoso o con el optimismo de un ambiente silencioso.

4.3 Evaluación para los dígitos

Por su parte, el desempeño en los dígitos fue el siguiente. Para 25 voces por dígito pronunciado,

TABLA 4.3

voz	aciertos	errores	total	tasa %
uno	24	1	25	96
dos	22	3	25	88
tres	20	5	25	80
cuatro	19	6	25	76
cinco	22	3	25	88
seis	24	1	25	96
siete	21	4	25	84
ocho	18	7	25	72
nueve	17	8	25	68
cero	21	4	25	84
si	207	1	208	99.5
no	42	0	42	100

83.2 % del total de dígitos y 99.6 % para las voces si y no.

Las pruebas presentadas en este capítulo se apoyan en un proceso estocástico, con la finalidad de ser veraces en su proceso y con la honestidad de presentar resultados insesgados.

Uno de los problemas en la distorsión de la voz se debió al ventilador del laboratorio de voz y a la presencia de compañeros trabajando en las otras computadoras y otras voces de otros compañeros en los demás laboratorios. En conclusión, el ruido ambiental tiene un efecto definitivo en el proceso de adquisición de la señal, dando como resultado errores en el reconocimiento.

Sabiendo que los coeficientes de reflexión son la base del reconocimiento, y con la finalidad de lograr una mejor estabilidad en éstos se pueden forzar los coeficientes que estén fuera del intervalo

unitario a entrar a este. Por otro lado, el cálculo de las distancias determinará cual de los patrones de comparación tiene la mínima distancia y será elegido como solución al reconocimiento de dicho bloque.

CONCLUSIONES

Conclusiones

El trabajo concluye con una finalidad muy importante, la de dar al usuario del sistema una libertad de acción y que no dependa de otras personas para cualquier movimiento que tenga deseos de hacer, ya que en los casos de cuadriparecia los minusválidos no pueden con la carga del peso que representa la suma de su cuerpo, la silla y utensilios que use.

El esfuerzo realizado para que la presente tesis llegara a feliz termino ha sido grande; los trabajos se iniciaron con señales ficticias para probar las virtudes del método; por supuesto y considerando que eran 10 señales diferentes únicas, los aciertos sobre esas señales, al ser totalmente invariantes, fueron del 100%, incluso para toda la señal, unas 1500 muestras, se obtuvieron únicamente 10 coeficientes.

Las pruebas se hicieron con la SOUND BLASTER, cuyo ambiente esta desarrollado en WINDOWS y presenta un modo gráfico muy eficiente y 'amigable'; la otra opción que presenta la SB16 es la digitalizacion vía el DOS de la PC, que es el que se usa en este trabajo.

Las voces grabadas a su frecuencia óptima, en 2 segundos, proporcionan (12 000 muestras, de las que, en algunos casos, tienen 5000 muestras efectivas de señal; en base a esta observación se optó por 5050 muestras efectivas por voz y con esto asegurar un 'colchón' de 100 muestras para los 29 'frames' de 170 muestras cada uno.

Todo el proceso de ventaneo se baso en señales de voz, que en esencia no es estacionaria, y este hecho obliga a tomar decisiones drásticas en el desarrollo de los programas del cálculo de las distancias para el reconocimiento, en el que se tienen que tener en consideración, que los coeficientes de reflexión de cada 'frame' son diferentes, incluso para palabras 'iguales', desiguales en el sentido estricto, matemático se puede decir, por lo que la distancia será mínima para los 'frames' correspondientes, como los que se dan en voces que comienzan con la misma vocal, como en ALTO, ADELANTE y ATRAS, por ejemplo.

Se intento insertar muestras para que las voces semejantes fuesen equiparables en longitud, pero este proceso dio como resultado

la distorsión de la señal original, presentando muchas fallas en su reconocimiento, por lo que se desechó.

El trabajo se fue optimizando poco a poco hasta lograr el mejor resultado para obtener una tasa de reconocimiento del 80, 90 y del 100% si es posible, y que los pacientes necesitados del control de su entorno tengan libertad de movimiento, como primera opción.

Es de verdad difícil para una persona con estas limitaciones poder moverse en terrenos con pendientes de cualquier tipo, 10 % por ejemplo, por esto se trata de dar apoyo a aquéllos a los que la vida, además de no mover la piernas, les ha negado el uso de sus extremidades superiores, inmersos en una parálisis casi total.

Pero aquí esta la solución, es de hacer notar que el reconocimiento de comandos de voz por medio de un sistema inteligente de cómputo amplía los horizontes de desarrollo de estas personas, consiguiendo con este trabajo de tesis la libertad deseada, anhelo de todo ser desarrollado, pensante y actuante, dando libertad a otros seres, encargados de su cuidado la misma libertad que el sistema otorga a dichas personas. Los resultados del reconocimiento son excelentes, en el sentido de brindar una buena herramienta de desarrollo, con una buena calidad en el reconocimiento, ya que las fallas de reconocimiento se deben a la contaminación de la señal por ruidos ajenos a ésta, lo que significa que el reconocimiento es bueno.

También se encamina a desarrollar mejores algoritmos en espera de lograr tasas superiores al 85% con la aplicación de métodos en el dominio de la frecuencia, FFT de tiempo corto, de cuantización vectorial, de modelos de Markov, métodos de Fussy Logic, o de aprendizaje como lo son redes neuronales. Con este enfoque se pueden lograr aplicaciones de procesamiento en paralelo con el fin de obtener los parámetros LPC mientras se obtiene una DFT o una FFT, para enseguida calcular al mismo tiempo las diversas distancias para el reconocimiento en base a los diferentes caminos por los que transite el proceso de la señal y conjuntar todo esto en la toma de una decisión en la correcta detección de la señal-voz.

Teniendo en consideración la definitiva acción que el ruido ambiental tiene en el reconocimiento de la señal voceada, se requieren de pruebas con micrófonos estrictamente direccionales y lo más posible inmunes al ruido, a fin de obtener solo señales de voz pronunciadas por el operador y que el ruido penetre con un mínimo de energía, que no se registren amplitudes que sobrepasen los límites aceptables de ruido, en los que la palabra pronunciada por el operador sea distorsionada.

Otro punto a considerar es la de crear un estado de espera para el inicio de la detección de un comando para que la computadora no analice señales innecesarias, y que se mantenga el status del sistema o que sea prudente quitarse el micrófono cada vez que se desee establecer comunicación con algún interlocutor.

Con esta visión del sistema, se puede aplicar a diferentes aparatos caseros, como la tele, el radio, la luz y las puertas de hogar en que se desenvuelva el paciente, hasta el control de una cama plegable mecánica, encender una PC, hacer un dictado en la PC, abrir cortinas, etc.

El uso de estos controladores por comandos de voz es aplicable a todo tipo de usuarios, como son el de claves bancarias por teléfono, operación del teléfono inalámbrico en un auto, cerrar ventanas del auto, apertura de puertas y control de temperatura interna, etc.

El reconocimiento de comandos de voz y del uso exclusivo de un usuario en particular podría tener un éxito inusitado, por lo que su desarrollo es de gran prioridad.

BIBLIOGRAFÍA.

[1]. SIMON HAYKIN. ADAPTIVE FILTER THEORY. Prentice Hall ecend Edition.1991.

[2]. P. PAPAMICHALIS. PRACTICAL APROACHES TO SPEECH CODING Texas Instrument, Inc. Prentice Hsll, Inc. 1987.

[3]. LAWERENCE R. RABINER & CHARLES M. RADER. Digital Signal Processing. IEEE PRESS. A volume in the IEEE PRESS Selected Reprint Series. 1972. Libro de selecciones sobre el Procesamiento Digital de Señales.

[4]. TURBO C ++. THE COMPLET REWFERENCE. Osborne McGraw-Hill, INC. 1988.

[5]. DSP56000/56001. DIGITAL SIGNAL PROCESSORS. User's Manual. Motorola, Inc. 1990.

[6]. FIRST-GENERATION. TMS320. User's Guide. Texas Instruments, Inc. 1988.

[7]. ROGELIO ALCANTARA SILVA. Introducción al PDS, RDS, 1989, DEPFI, UNAM.

[8]. MANFRED SCHROEDER. Linear Prediction, Entropy and Signal analysis Vol. IEEE ASSP MAGAZINE, JULY 1984.

[9]. R. E. CROCHIERE and J. L. FLANAGAN, Current perspectives in digital speech. IEEE Vol. 1 No. 3. Jan 1983.

[10]. M. R. SCHROEDER. Vocoders: Analisis and Speech Control. Vol. 54 No. 5. IEEE ASSP MAGAZINE May 1966.

[11]. D. O'SHAUGHNESSY. Automatic Speech Syntesis. IEEE, Dec 1983.

[12]. EDUARDO CASTILLO FUENTES. Tesis sobre el Control de posición y velocidad para la conducción automática de una silla de ruedas. Ago. 1994. DEPFI U N A M.

[13]. LONNIE C. LUDEMAN. Fundamentals of digital signal procesing. HARPER & ROW, PUBLISHER. NEW YRK, 1986.

[14]. C68332 USER'S MANUAL, MOTOROLA INC. 1990.

[15]. ALAN V. OPPENHEIM, ALAN S. WILLSKY, IAN T. YOUNG. Signals and Sistems. Prentice-Hall, Inc. Englewood Cliffs, New Jersey 07632

[16]. LAWRENCE RABINIER, BIING-HWANG JUANG. Fundamentals of speech recognition. Prentice Hall, Inc. Englewood Cliffs, New Jersey 07632

[17]. M. FOGIEL & Staff. The electronic communications problem solver. Research and Education Association. Piscataway, New Jersey 08854. Libro de soluciones a problemas de Comunicaciones y Electrónica.

[18]. M. R. SCHROEDER, Analysis and synthesis of speech, IEEE, May 1966.

[19]. ROBERT M. GRAY, ANDRES BUZO, AGUSTINE H. GRAY, JR. And YASUO MATSUYAMA. Distortion measures for speech processing. IEEE, Vol. ASSP 28, No. 4, Aug 1980.

[20]. CORY MYERS, L. R. RABINER and AARON E. ROSENBERG. Performance tradeoffs in Dynamic Time Warping algorithms for isolated word recognition. IEEE, Vol. ASSP 28 No. 6, Dec 1980.

[21]. AGUSTINE H. GRAY JR. and JOHN D. MARKEL. Distance measures for speech processing. Vol. ASSP 24, No. 5 Oct 1976.

[22]. TOHRU IFUKUBE, TADAYUKI SASAKI and CHEN PENG. A blind aid modeled after echolocation of bats. IEEE, Vol. 38, No. 5, May 1991.

[23]. JHON E. HOLMGREN. Applaying Automatic Speech Recognition to Telephone Services. IEEE, Nov. 1982. Sin Vol.

APÉNDICE A

MANEJO DEL PROGRAMA DE ADQUISICIÓN Y ANÁLISIS DE UNA SEÑAL-VOZ

A.1 Manejo del software

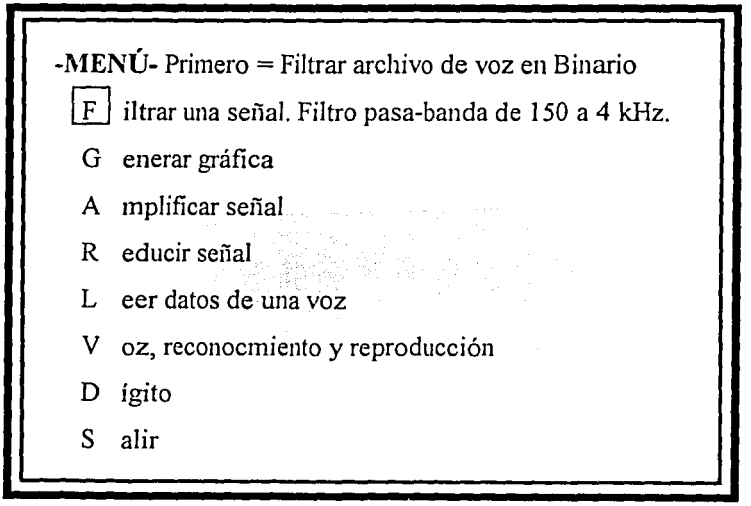
Se desarrolló un programa que, además del proceso de detección, grafica la señal voz; en él se conjuntaron todos los sistemas desarrollados, como son: el programa de generación de base de datos, reconocimiento de voz, control de datos, y el sistema de gráficas para PC.

Este conjunto de programas, al alcance de los estudiosos del análisis de señales, se puede usar como apoyo didáctico, en él se aprecia claramente el efecto del filtrado sobre la señal, para entender mejor los procesos en la obtención de los parámetros característicos de la señal.

A.2 Adquisición y análisis de señal-voz

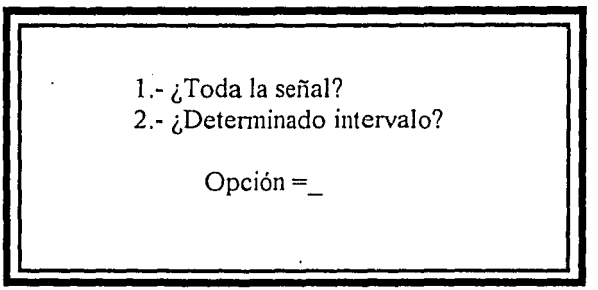
El uso del programa GRAF17.EXE, se apoya en datos de 8, o 16 bits, opciones del la tarjeta de digitalización SOUND BLASTER, almacenados en archivos en código binario.

Al arrancar el programa aparece el siguiente cuadro:



Con el cursor, que aparece como un cuadro, se selecciona primero el filtrado de la señal, la que estará en un archivo en binario generado por la opción L, que contiene datos en variable entera. Se proporciona el nombre del archivo a ser filtrado; ya filtrado se procede a su graficación con la opción G ya que automáticamente se generaron los archivos apropiados para este fin. En la gráfica respectiva aparecerán dos pantallas, una con la señal original y otra con la señal filtrada en ventanas de 560 puntos.

Al entrar en el menú de generar gráfica aparece el cuadro:



En Opción = 1, se debe recordar que se tomaron, en la opción L del menú, más de 12000 muestras, de las que las primeras 4000, más o ménos, son de ruido, pues al leer los datos a frecuencias de 6kHz a 8kHz, en en tiempo en que aparerce la señal indicadora de vocear y en lo que el locutor dice la palabra, el sistema ha leido de 2500, a 5000 muestras; después aparece la señal, que es enmarcada en cuadros de 560 muestras, con opción de salir o de continuar con más ventanas.

En Opción = 2 se puede escoger determinado "frame", ya sabiendo con la opción anterior en que número de muestra empieza la voz que permite definir el intervalo de la segunda opción y visualizar un marco determinado, por ejemplo, de 110 muestras.

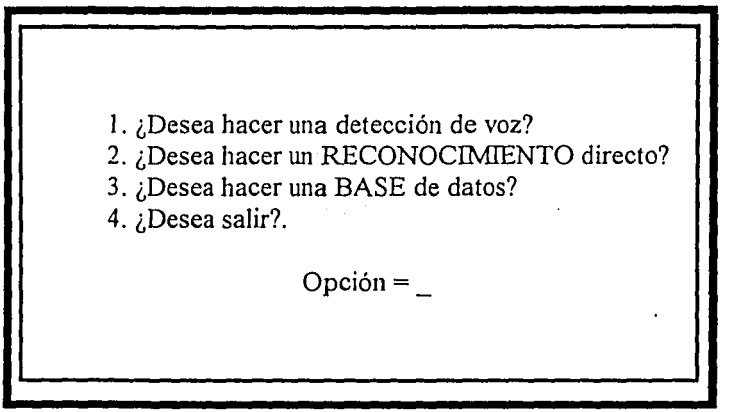
Si la señal muestra una amplitud muy grande o muy pequeña, tal que se sale de cuadro o casi no aparece, se aplica la opción del menú principal con la letra R para reducir, o con la A para amplificar, y apreciar mejor sus características; los grados de amplificación o

reducción son valores enteros desde uno, hasta cuatro veces la señal original, la amplitud que se pide al principio es dos.

La Opción L del menú principal es para el sistema de adquisición de datos de la señal, con la tarjeta Sound Blaster; si no se tiene instalada la Sound Blaster se generará un error en la PC, por lo que no se debe hacer uso de esta opción, pués para salir de este problema se tendrá que inicializar la computadora.

En la Opción T del menú principal se entrará al sistema de reproducción de algún archivo generado por la opción anterior; en este caso no se tendrá ningún problema en la reproducción de la voz, ya que la salida de la señal no esta condicionada.

En la Opción V del menú principal se encuentra el programa de reconocimiento y creación de una base de datos, por lo que aparecerá el submenú:



La Opción = 1, se usa para detectar la voz con base en los archivos de voces previamente grabados. En Opción = 2 se traslada el dominio a la tarjeta digitalizadora para la lectura de datos de una simulación real del comando de voz que se desee reconocer.

La Opción = 2 es para crear una base de datos para el futuro reconocimiento de voz; es pertinente decir aquí que la detección de la

voz es, en principio, exclusiva de un usuario, proporcionando una seguridad en el reconocimiento de las señales que son captadas por el micrófono, evitándose que otra persona genere algún comando y ponga en acción los motores de la silla de ruedas; si así no fuera, y alguien dijese, "súbelo más alto ..." en las cercanías del la silla, en ese momento se frenaría la silla, y el conductor iría a parar con todo y huesos al suelo; esto se prevee con el comando de confirmación, Sí o No, ya explicado.

En Opción = 3, se generan los archivos que servirán como diccionario en el reconocimiento; es la base de datos con la que se ha de comparar una señal que se desee reconocer, además de poder hacer una base para cada usuario. Los siguentes pasos son fáciles de seguir.

La opción del MENÚ principal D permite al usuario hacer una detección directa, vía micrófono, de 9 digitos y el cero para su aplicación a sistemas de chapas electrónicas y comandos relacionados con los dígitos o para números de cuentas bancarias, etc.

La última opción del MENÚ es la de salir con S.

En las gráficas de la señal se puede apreciar el efecto de un filtro punto por punto y como el filtrado suaviza la línea de puntos que conforman la señal digital. Se puede observar también como están conformados los silencios, las vocales y las consonantes y las distorsiones de la señal al cambiar de una sílaba a otra, proporcionando una comprensión más amplia de alguna voz en particular o apreciar diferencias con otras voces en los silencios o en los cambios de letra, etc.