UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Facultad de Ingeniería
División de Estudios de Posgrado

RECONOCIMIENTO DE COMANDOS DE VOZ UTILIZANDO LÓGICA DIFUSA

T E S I S

QUE PARA OBTENER EL GRADO DE
MAESTRA EN INGENIERÍA ELÉCTRICA

PRESENTA:

LETICIA ALVAREZ CASTILLO

DIRECTOR DE TESIS:
DR. ROGELIO ALCANTARA SILVA

MEXICO, D. F. 2000

A mi padre

José Jesús Alvarez García

Agradezco a Dios por sobre todas las cosas.

A mi padre José Jesús Alvarez García.

A mi madre Estela Castillo González q. e. p. d.

A mi tía Felicitas Castillo González q. e. p. d.

A mi abuela Soledad González Morales q. e. p. d.

A estas personas un agradecimiento muy especial por su guía, confianza, motivación, apoyo, paciencia y su gran amor que hicieron posible la terminación de este trabajo.

Agradezco por su presencia:

A Paco, familia y amigos.

A mi hermano Carlos Blas y familia.

A mi hermano Jesús.

A mis tías Celia Alvarez García y Alicia Castillo González.

RECONOCIMIENTOS:

Quiero expresar mi gratitud al Dr. Rogelio Alcántara Silva, por haber dirigido esta tesis, por su asesoria y paciencia durante la realización de este trabajo.

Asi mismo agradezco el apoyo del Dr. Jesús Savage Carmona y Dr. Fransisco García Ugalde.

Este trabajo se enriqueció por los valiosos comentarios del Dr. Carlos Rivera Rivera, Dr. Miguel Moctezuma Flores y Dr. Boris Escalante Ramírez,

a estas personas y a todas aquellas que en una o en otra forma colaboraron para la culminación de este trabajo, **gracias**!

ÍNDICE

RECONOCIMIENTO DE COMANDOS DE VOZ UTILIZANDO LÓGICA DIFUSA.

RESUMEN

El objetivo de esta tesis es lograr, por medio de la Lógica Difusa, incrementos en el porcentaje de reconocimiento de comandos de voz.

Consecuentemente, en este trabajo se realiza el análisis, las pruebas, la implementación y validación de tres enfoques para reconocimiento de comandos de voz. Los dos primeros métodos son enfoques ya practicados en el campo de Procesamiento Digital de Señales (DSP), que posteriormente servirán de base para innovar un tercer enfoque para reconocimiento de voz, con el fin de obtener un incremento en la tasa de reconocimiento sobre las dos primeras técnicas

El primer enfoque utiliza la técnica tradicional de Codificación por Predicción Lineal (LPC), en la que se incluye el Algoritmo de Deformación de Tiempo Dinámico (DTW) para la comparación de los vectores espectrales del patrón de prueba y el patrón de referencia.

El segundo enfoque para Reconocimiento de Voz aprovecha la técnica de Cuantización Vectorial (CV).

Una vez finalizada la implementación de los dos primeros reconocedores de voz, se inicia el desarrollo del tercer enfoque de reconocimiento, el cual utiliza Lógica Difusa, técnica que se aplica para datos imprecisos o ambiguos, como son los datos dados por el análisis espectral de una señal no estacionaria como lo es la voz. Para la implementación del método de reconocimiento difuso se requiere la determinación de las variables de entrada y salida, la fusificación de dichas variables, la determinación de reglas de inferencia y la defusificación de la salida. Con este último reconocedor de voz difuso se obtiene un incremento en la tasa de reconocimiento de 4.8% sobre la primera técnica tradicional LPC, y un incremento de 3.0% sobre el segundo enfoque de reconocimiento de voz.

Capitulo I

INTRODUCCIÓN.

I.1 Introducción

La perspectiva de hablar con una computadora ha intrigado a la gente por muchos años En el campo de ciencia ficción con cierta frecuencia observamos diálogos entre humanos y máquinas. Sin embargo a pesar de la gran inquietud y esfuerzos de investigación realizados para diseñar una máquina inteligente que pueda reconocer la palabra hablada y comprender su significado, y que además esté en condiciones de entender un discurso hablado sobre algun tema por algun usuario en cualquier ambiente, llegamos a la conclusión de que aún estamos lejos de llegar a la meta deseada. De lo anterior se deriva la reflexión de cómo podemos construir una serie de puentes que nos capaciten para avanzar sobre ambos, es decir, nuestros conocimientos así como la capacidad para construir modernos sistemas de reconocimiento de voz y que de este modo la conversación del reconocimiento de voz y el entendimiento por una máquina sea logrado.

El Reconocimiento Automático de Voz (RAV) ha aparecido en muchas aplicaciones comerciales, sabiendo que la capacidad de RAV todavía tiene un largo camino por recorrer, actualmente existen varias máquinas disponibles comercialmente que reconocen palabras de un pequeño vocabulario hasta máquinas que previamente han sido entrenadas por el locutor, esto se ha logrado por etapas que han ido evolucionando como se mencionará más adelante.

Aunque las especificaciones de los sistemas de Reconocimiento Automático de Voz (RAV) difieren en detalles de su implementación siguen estos pasos básicos en la realización del reconocimiento:

- Extracción de las características.- El patrón de prueba es procesado y produce una serie de características.
- Determinación de similaridad.- Las características extraidas del patrón de prueba son comparadas con varias series de características almacenadas, o patrones de referencia como son comunmente llamados, representando el vocabulario para ser reconocido. Un problema central en el proceso de comparación es que el tiempo de escala del patrón de prueba y el patrón de referencia no están perfectamente alineados, en algunos casos el tiempo de escala puede ser registrado por una simple expansión o compresión.
- Respuesta de decisión.- Una vez que una medida de similaridad entre un patrón de referencia y el patrón de prueba ha sido obtenida se debe tomar una decisión para la respuesta de salida.

En el área de reconocimiento de voz es necesario especificar ciertas opciones antes de que la señal sea procesada, estas son:

- Tipo de voz.- Palabras aisladas, voz continua.
- Número de locutores.- Sistemas en los que existen uno o varios locutores.

- Tipo de locutores.- Masculino, femenino, infantil.
- Ambiente de locución - Lugar público, cuarto de computación, cabina de sonidos de pruebas
- Sistema de transmisión.- Micrófono, micrófono de alta calidad, teléfono.
- Tamaño del vocabulario.- Vocabulario pequeño (1-20 palabras), vocabulario medio (20-100 palabras), vocabulario grande (mayor de 100 palabras).
- Formato de entrada para hablar.- Texto restringido, formato libre.

Las especificaciones para este trabajo son: palabra aislada, un locutor femenino, lugar público, micrófono, vocabulario pequeño, texto restringido.

El reconocimiento de comandos de voz utilizando Lógica Difusa consiste en conjugar las técnicas de Lógica Difusa y Procesamiento Digital de Señales (DSP).

El modelo de Análisis Espectral especifica paramétricamente, por medio de un *vector de coeficientes*, los *marcos* o *ventanas de voz*, los cuales son traslapados con el fin de suavizar la estimación espectral y, posteriormente, aplicar preénfasis a los marcos de voz para uniformar espectralmente las altas frecuencias [11], y una ventana de Hamming para calcular las funciones de autocorrelación de tiempo corto [14].

También se presenta un algoritmo para síntesis de voz que es una técnica que va tomada de la mano con el Reconocimiento de voz, ya que son completamente paralelas en sus primeras etapas [16]. Luego, se hace mención de lo que son las señales sonoras y no sonoras y la manera de predecir una señal en función de pasadas p muestras de voz a través de un filtro predictor

El reconocimiento de voz consiste en obtener: las medidas características, el entrenamiento del patrón, clasificación de patrones y decisión lógica.
Para la extracción de las medidas características de la señal se procede a obtener los bloques traslapados de la señal y en cada bloque aplicar un preénfasis y un ventaneo; luego obtener los coeficientes de autocorrelación y los coeficientes LPC, posteriormente los coeficientes cepstrales si así se requieren para calcular las medidas de distancias y asi tener el vector espectral.
Los parámetros de referencia son el resultado de tener las características de una o más *plantillas* de cada clase de patrón o un promedio de las características de los patrones de la misma clase, alineados en el tiempo.
Una vez que se cuenta con los patrones de referencia, se extráen las características del patrón de prueba y se procede a comparar estas últimas características con el vocabulario establecido por medio del *algoritmo de Deformación de Tiempo Dinámico* (DTW) [3], que tiene como finalidad compensar las diferencias en el tiempo de estos patrones. Consecuentemente, se puede decidir cuál patrón de referencia compatibiliza mejor con el patrón de entrada, [33].

Otro enfoque de reconocimiento de voz es por medio de la técnica de Cuantización Vectorial (CV). De un conjunto de patrones de la misma clase, se extraen sus caracteristicas espectrales con las que se forma un centroide que será perturbado con un

pequeño valor para generar dos centroides Se obtienen dos distancias a partir de los vectores de los patrones de la misma clase y los dos últimos centroides Si la distancia global no cambia, entonces se generará el doble de los centroides que se tienen, hasta llegar al número de centroides deseados. Estos centroides son llamados el Cuantizador Vectorial de una palabra o clase de patrón. Posteriormente, se podrán hacer las comparaciones de los patrones de referencias (Cuantizadores Vectoriales) y el patrón de prueba [11], [32].

Otros enfoques de reconocimiento usando Segmentación han sido implementados obteniéndose buenos desempeños, [34], [35], ver apéndice.

Otro enfoque de Reconocimiento de Voz radica en utilizar la Lógica Difusa, que consiste en determinar las entradas y salida del sistema difuso [26]. Para este sistema la salida es la palabra reconocida. Las entradas serán parámetros que en conjunto, con las reglas de inferencia podrán determinar la palabra reconocida. Se tomaron como entradas al sistema difuso, las distancias mínimas obtenidas de los reconocedores con técnicas CV y tradicional LPC, además de la potencia calculada del patrón de prueba. Para los cuatro universos o variables del sistema difuso (tres entradas y una salida), se establece el rango de los universos, así como formas, rangos y traslape de las funciones de membresía, en base a los datos tomados de los dos reconocedores y de las potencias obtenidas para estas clases de patrones o palabras del vocabulario [12]. Otras variables de entrada pudieron haber sido seleccionadas para alimentar al sistema difuso y en base a éstas, se formarían los universos de entrada y sus respectivas funciones de membresía, que podrían mejorar o simplificar este enfoque
Por último, una vez defusificada la variable de salida se obtiene la palabra reconocida por el sistema, que obtiene un porcentaje de reconocimiento superior al mejor de los dos primeros reconocedores.

De los tres reconocedores difusos el que tiene mayor porcentaje de aciertos es el reconocedor de tres entradas. *distancia CV, distancia LPC* y *potencia* (RCLD), mejorando al que más se le acerca en 1 4 % que es un reconocedor de voz difuso con dos entradas: *distancia CV* y *potencia* (RCD), superando éste en 1.2 % al que le sigue en porcentaje de aciertos, el reconocedor difuso que tiene como entradas la *distancia LPC* y la *potencia* (RLD) El RLD mejora en 0.4% el reconocedor de voz con técnica CV (RCV) y éste a su vez supera al tradicional LPC (RLPC) en 1 8% de aciertos. Por lo anterior se deduce que la diferencia entre el mejor de los reconocedores de voz RCLD y el que muestra menor porcentaje de aciertos RLPC es de 4.8 %. Como puede observarse el RCLD es el producto de los RCV y RLPC porque se apoya en sus resultados.
De lo que se concluye que los reconocedores de voz difusos proporcionan mayor porcentaje de aciertos, menor complejidad y son fáciles de implementar

I.2 Breve Historia del Reconocimiento de Voz.

Durante los últimos años hemos asistido a una explosión de trabajos relativos al análisis y síntesis del habla; sin embargo, el interés por el tema no es nuevo. En el siglo

XVIII, Wolfgang Von Kempelen construyó un dispositivo parlante totalmente mecánico [13].

Desde entonces hasta nuestros días la trayectoria en el campo de la síntesis ha sido más o menos continua, aunque el verdadero desarrollo no comienza hasta la incorporación de la tecnología electrónica, culminando en la actualidad (gracias a la informática) en sistemas comerciales capaces de realizar síntesis a partir de un diccionario dado de palabras o frases, y en sistemas comerciales que realizan de forma efectiva la síntesis partiendo de cualquier texto ortográfico cualesquiera.

En el campo del análisis, la historia es mucho más reciente pues, aunque existen desde la antigüedad tratados de fonética en los que de alguna forma se intentan estudiar los

mecanismos y propiedades de la palabra hablada, no se puede hablar de aportaciones realmente válidas hasta la década 1930-1940, en que se desarrollaron las primeras versiones del "espectrógrafo" o "sonógrafo" (fig. I.1) dispositivo que permite la obtención de un registro (sonograma) de la energía (representada por el mayor o menor ennegrecimiento) contenida en las diversas bandas de frecuencia de un palabra o frase en función del tiempo (fig I.2). Desde este momento comienza a vislumbrarse la posibilidad de la realización de sistemas para el reconocimiento del habla [13].

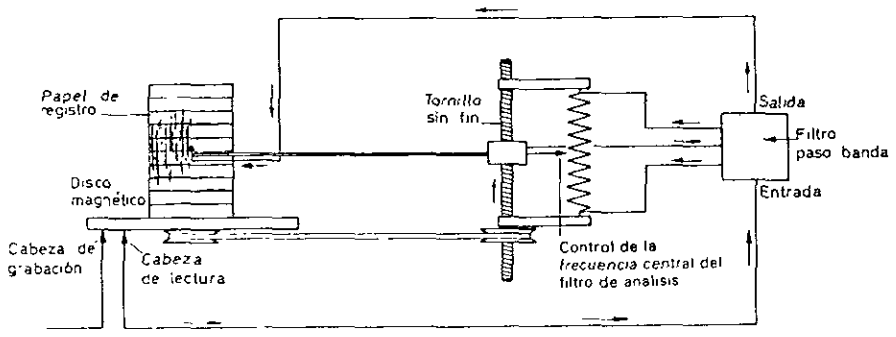


Fig. I.1 ESPECTRÓGRAFO, [13].

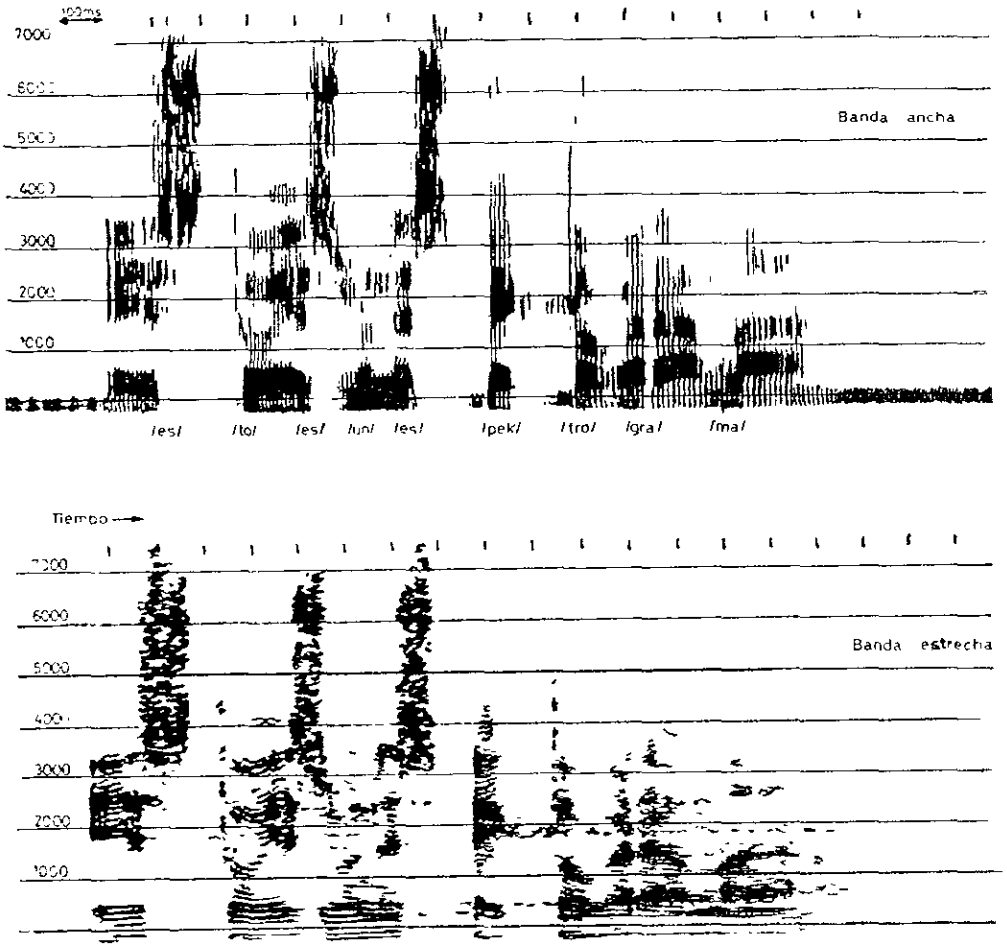


fig I 2 ESPECTROS DE BANDA ANCHA Y ESTRECHA,[13].

En 1952 en los laboratorios Bell, Davis Biddulph, y Balashek construyeron un sistema para reconocer dígitos aislados por un locutor, el sistema dependía de las medidas de resonancia espectral durante la región de la vocal de cada dígito. En 1956 en los laboratorios RCA, Olson, y Belar intentaron reconocer 10 sílabas distintas de un solo locutor, también este sistema dependía de las medidas espectrales (como las suministradas por un banco de filtro analógico, el sistema era totalmente electrónico) durante las regiones de las vocales. En 1959, en la Universidad de College en Inglaterra, Fry y Denes intentaron desarrollar un reconocedor de fonemas para reconocer cuatro vocales y nueve consonantes, en esto se uso un analizador de espectros y un comparador de patrones para tomar la decisión de reconocimiento

Los primeros trabajos que hacen uso de tecnología informática comienzan a aparecer en 1959-1960, en 1960 Deves y Mathews introducen el concepto de

normalización temporal no lineal, que permite la comparación de parámetros de palabras iguales pronunciadas a distinta velocidad. Desde estas fechas comienza la explosión de trabajos, principalmente de Reconocimiento de Palabras Aisladas.
La década de los 60's inicia con la entrada de varios laboratorios japoneses en el área de reconocimiento, construyendo hardware de propósito especial para los sistemas. Un sistema japonés descrito por Susuki y Nakata del Laboratorio de investigación de la Radio en Tokyo, fue un reconocedor de vocales en hardware, en el que se uso un banco de filtros. Otro sistema Japonés fue desarrollado por Sakai y Doshita de la Universidad de Kyoto en 1962, quienes construyeron un reconocedor de fonemas en hardware. Un segmentador de voz en hardware fue usado con un análisis de cruces por cero en diferentes regiones de la palabra de entrada para proporcionar la salida reconocida. Un tercer sistema Japonés fue el reconocedor digital en hardware de Nagata en los laboratorios NEC en 1963. Son varios los países en los que se comienza a trabajar en proyectos de esta índole (Japón, Francia, etc.), pero es en Estados Unidos donde se lanza en 1971, el mayor proyecto conocido en la historia del Reconocimiento del Habla. Se trata del 'ARPA-SUR' (Advanced Research Projects Agency -Departamento de Defensa- Speech Understanding Research), con un presupuesto de quince millones de dólares y una duración de cinco años.

Finalmente, en los 80's fue una década en la cual el principal ímpetu fue dado para vocabularios grandes, sistemas de reconocimiento de voz continua. El sistema por la Defense Advanced Research Projects Agency (DARPA) invirtió en un gran programa de investigación, aspirando a lograr una alta precisión para 1000 palabras, reconocimiento de voz continuo y tareas de manejo de base de datos. El programa DARPA continuo hasta los 90's con el reconocimiento de principio a fin del lenguaje natural, y la tarea de corrimiento para recuperación del aire de propagación de información. También la tecnología de reconocimiento de voz ha incrementado su uso dentro de las redes telefónicas para automatizarlas en sustitución de un operador.

En los años 70's las investigaciones en reconocimiento de voz fueron significantes. Primero el área de palabra aislada o reconocimiento de pronunciaciones discretas llega a ser una tecnología viable y útil basada fundamentalmente en los estudios de Velichko y Zagoruyko en Rusia, de Sakoe y Chiba en Japón, e Itakura en los Estados Unidos. Los estudios rusos ayudan al avance del reconocimiento de voz haciendo uso del reconocimiento de patrones. Los japoneses investigan cómo los métodos de programación dinámica pueden ser exitosamente aplicados. Las investigaciones de Itakura presentan la ideas de codificación de predicción lineal (LPC por sus siglas en inglés).

Así como el reconocimiento de palabras aisladas fue el foco de la investigación en los 70's, el problema de reconocimiento de palabras conectadas fue el foco de la investigación en los 80's. Las investigaciones de la voz en los 80's fue caracterizada por un corrimiento en tecnología basada en plantillas aproximadas a métodos estadísticos, especialmente el Hidden Markov Model (HMM), conocido y entendido en pocos laboratorios, y después de una publicación de los métodos y teoría de HMMs a mediados de los 80's la técnica llego a ser ampliamente aplicada en los laboratorios de investigación del mundo [11]. Otra nueva tecnología fue reintroducida a fines de los 80's, la idea de aplicar redes neuronales a los problemas de reconocimiento de voz. En los 80's, sin embargo, un profundo entendimiento de la fortaleza y limitaciones de la tecnología fue obtenido, así como también la relación de la tecnología a métodos de clasificación de

señales clásicas. Varias nuevas maneras de sistemas de implementación fueron también propuestas.

En este trabajo se trataran algunos de los métodos citados, lo que nos da la pauta para seleccionar las características que debe tener el método que se aplicará en esta tesis, por esto se hace una breve mención de los conceptos generales de procesamiento digital de señales en el capítulo II. En el capítulo III se tratarán los principios generales del reconocimiento de voz. Un reconocedor tradicional basado en la técnica LPC será tratado en el capítulo IV. Un reconocedor de cuantizacion vectorial se verá en el capítulo V. En el capítulo VI se dan los principios de lógica difusa. El reconocedor difuso será presentado en el capítulo VII. Las pruebas y validación del sistema se encuentran en el capítulo VIII y por último se muestran las conclusiones en el capítulo IX.

Capítulo II

GENERALIDADES DEL PROCESAMIENTO DIGITAL.

II.1 Introducción.

La mayoría de las señales encontradas en ciencia e ingeniería son analógicas por naturaleza Esto es que las señales son funciones de una variable continua, tal como el tiempo o el espacio y usualmente toman valores en un rango continuo. Tales señales pueden ser procesadas directamente por un sistema analógico apropiado tal como un filtro, es decir, en tal caso la señal ha sido procesada directamente en su forma analógica. El procesamiento digital se logra con la interfaz de un convertidor analógico digital (A/D).

El procesador analógico-digital puede ser una gran computadora digital programable o un pequeño microprocesador que es programado para desempeñar las operaciones deseadas sobre una señal de entrada. Existen aplicaciones prácticas donde la información requerida a la salida de los sistemas de procesamiento digital de señales también es digital, por ejemplo en señales de radar, la posición de un avión y su velocidad. Hay otras aplicaciones donde la salida se requiere en forma analógica por lo que el sistema hace uso de un convertidor digital analógico (D/A)

II.2. Filtrado.

El término filtro es comúnmente usado para describir un dispositivo que discrimina, de acuerdo a algún atributo de los objetos aplicados a sus entradas, cuando pasa a través de éste Por ejemplo, un filtro de aire permite el paso del aire a través de éste, pero elimina las particulas de polvo que están presentes en el aire que pasa a través del filtro.

Un sistema lineal invariable en el tiempo, es aquel que cumple con el teorema de superposición, el cual dice que la suma de los efectos de dos o más excitaciones es igual a considerar el efecto por la suma de las excitaciones. Este sistema lineal invariable en el tiempo también desempeña un tipo de discriminación o filtrado entre varias componentes de frecuencias a su entrada. La naturaleza de esta acción de filtrado es determinada por las características de la respuesta en frecuencia del sistema (función de transferencia $H(\omega)$), la cual cambia dependiendo de la elección de los parámetros del sistema (e.g. los coeficientes $\{a_k\}$ y $\{b_k\}$ en las diferentes ecuaciones características del sistema). De este modo, por medio de la selección apropiada de los coeficientes, se puede diseñar la frecuencia selectiva del filtro por medio del cual pasa la señal con componentes de algunas bandas de frecuencia, mientras que el filtro atenúa la señal que contiene componentes de frecuencia en otras bandas de frecuencia diferentes a la selectiva del filtro.

En general, un sistema lineal invariable en el tiempo modifica el espectro de la señal de entrada $X(\omega)$ de acuerdo a su respuesta en frecuencia $H(\omega)$ para producir una señal de

salida con espectro $Y(\omega)=H(\omega)X(\omega)$. En este sentido, $H(\omega)$ actúa como una *función de ponderación* o una *función de forma espectral* para los diferentes componentes en frecuencia de la señal de entrada. Visto en este contexto, cualquier sistema lineal invariable en el tiempo puede ser considerado como un filtro de frecuencia, consecuentemente, el termino "sistema lineal invariable en el tiempo" y "filtro" son sinónimos y se intercambian frecuentemente [15].

II.2.1 Diseño de filtros FIR de fase lineal

Uno de los mas simples tipos de filtros que se pueden diseñar, es un filtro FIR (respuesta al impulto finito) con fase lineal. Los filtros IIR (respuesta al impulso infinito) no pueden tener fase lineal. Existen muchas aplicaciones practicas como en comunicaciones digitales, donde fases (retardo) de distorsión significantes no son toleradas, y los filtros FIR son usados.

Un filtro FIR de longitud M tiene una respuesta en frecuencia

$$H(\omega) = \sum_{k=0}^{M-1} b_k e^{-j\omega k} \qquad \text{(II.1)}$$

donde los coeficientes del filtro $\{b_k\}$ son también los valores de la respuesta al impulso del filtro, esto es

$$h(n) = \begin{cases} b_n & 0 \le n \le M-1 \\ 0 & \text{de otra manera} \end{cases} \qquad \text{(II.2)}$$

La condición de fase lineal es obtenida imponiendo condiciones de simetría sobre la respuesta al impulso del filtro.

En particular consideraremos dos diferentes condiciones de simetría para $h(n)$. **La primera condición de simetría seria:**

$$h(n) = h(M-1-n) \qquad \text{(II.3)}$$

Si el filtro satisface esta condición de simetría entonces tiene fase lineal. **Si M es impar**, $M=5$, la condición de simetría es $h(0)=h(4)$, $h(1)=h(3)$, y $h(2)$ no tiene ningún termino compatible. De este modo la respuesta impulso del filtro es simétrico en el punto $h(2)$. La correspondiente respuesta en frecuencia es.

$$\begin{aligned} H(\omega) &= h(0) + h(1)\, e^{-j\omega} + h(2)\, e^{-2j\omega} + h(3)\, e^{-3\omega j} + h(4)\, e^{-4\omega j} \\ H(\omega) &= e^{-j2\omega} \left[h(2) + h(0)\, e^{j2\omega} + h(4)\, e^{-2\omega j} + h(1)\, e^{+j\omega} + h(3)\, e^{-j\omega} \right] \\ H(\omega) &= e^{-j2\omega} \left[h(2) + 2\, h(0)\, cos2\omega + 2\, h(1)\, cos\omega \right] \end{aligned} \qquad \text{(II.4)}$$

El término entre parentesis es real para todos los valores de ω y por lo tanto se denota como.

$$H_r(\omega) = h(2) + 2\,h(0)\,cos2\omega + 2\,h(1)\,cos\omega \qquad \text{(II.5)}$$
$$|H(\omega)| = |H_r(\omega)| \qquad \text{(II.6)}$$

La fase característica del filtro es

$$\Theta(\omega) = \begin{cases} -2\omega & \text{Si } H_r(\omega) > 0 \\ -2\omega + \pi & \text{Si } H_r(\omega) < 0 \end{cases} \qquad \text{(II.7)}$$

Si *M* es par, M=4, la condición de simetría es $h(0)=h(3)$, $h(1)=h(2)$, La correspondiente respuesta en frecuencia es:

$$H(\omega) = h(0) + h(1)\,e^{-j\omega} + h(2)\,e^{-2j\omega} + h(3)\,e^{-3j\omega}$$
$$H(\omega) = e^{-j3\omega/2}\,[\,h(0)\,e^{-j3\omega/2} + h(3)\,e^{-j3\omega/2} + h(1)\,e^{+j\omega/2} + h(2)\,e^{-j\omega/2}]$$

Con las relaciones de simetria, la expresión se simplifica a

$$H(\omega) = e^{-j3\omega/2}\,[\,2\,h(0)\,cos(3\omega/2) + 2\,h(1)\,cos(\omega/2)\,] \qquad \text{(II.8)}$$

En el caso donde M es impar, observamos que el término en los corchetes es real, así que $H(\omega)$ puede ser expresado como

$$H(\omega) = H_r(\omega)\,e^{-j3\omega/2}$$

donde

$$|H(\omega)| = |H_r(\omega)|$$

La fase del filtro es lineal, con fases de saltos de π radianes en la frecuencia donde $H_r(\omega)$ cambia el signo de positivo a negativo y viceversa.

$$\Theta(\omega) = \begin{cases} -3\omega/2 & \text{Si } H_r(\omega) > 0 \\ -3\omega/2 + \pi & \text{Si } H_r(\omega) < 0 \end{cases} \qquad \text{(II.9)}$$

De estos dos casos especiales, extrapolamos al caso general de un filtro con longitud arbitraria M. En general, la respuesta en frecuencia de un filtro FIR tiene una respuesta al impulso $h(n)$ que satisface la condición de simetría en la ec. (II.3) y puede ser expresada como

$$H(\omega) = H_r(\omega)\,e^{-j\omega(M-1)/2} \qquad \text{(II.10)}$$

donde

$$H_r(\omega)= h\left((M-1)/2\right) + 2\sum_{n=0}^{(M-3)/2} h(n)\cos\omega\left((M-1)/2 - n\right) \quad M \text{ impar} \qquad \text{(II.11)}$$

$$H_r(\omega) = 2\sum_{n=0}^{(M/2)-1} h(n)\cos\omega\left((M-1)/2 - n\right) \qquad M \text{ par} \qquad \text{(II.12)}$$

La característica de fase del filtro para M impar y M par es

$$\Theta(\omega) = \begin{cases} -\omega\left((M-1)/2\right) & \text{Si } H_r(\omega) > 0 \\ -\omega\left((M-1)/2\right) + \pi & \text{Si } H_r(\omega) < 0 \end{cases} \qquad \text{(II 13)}$$

La segunda condición de simetría que produce un filtro de fase lineal FIR es

$$h(n) = -h(M-1-n)$$

En este caso será llamada la respuesta al impulso *antisimetría*. Cuando M es impar el punto central de la antisimetría $h(n)$ es $n = (M-1)/2$. La condición anterior implica que

$$h\left((M-1)/2\right) = 0$$

Por ejemplo si M=5 tenemos $h(0)$=-$h(4)$, $h(1)$=-$h(3)$, y $h(2)$=0. Sin embargo, si M es par, cada termino $h(n)$, tiene un término de signo opuesto.

Es directo presentar que la respuesta en frecuencia de un filtro FIR con una respuesta al impulso antisimétrica puede ser expresada como

$$H(\omega)=H_r(\omega)\, e^{j[-\omega(M-1)/2 + \pi/2]} \qquad \text{(II.14)}$$

donde

$$H_r(\omega)= 2\sum_{n=0}^{(M-3)/2} h(n)\sin\omega\left((M-1)/2 - n\right) \quad M \text{ impar} \qquad \text{(II.15)}$$

$$H_r(\omega) = 2\sum_{n=0}^{(M/2)-1} h(n)\sin\omega\left((M-1)/2 - n\right) \quad M \text{ par} \qquad \text{(II.16)}$$

La característica de fase del filtro para M impar y M par es

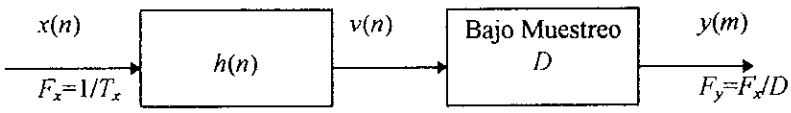


Figura II.2 DECIMACION POR UN FACTOR DE D

$$v'(n) = \begin{cases} v(n) & n = 0, \pm D, +2D, \ldots \\ 0 & \text{en otro rango} \end{cases} \qquad \text{(II.22)}$$

Puede ser vista $v'(n)$ como una secuencia obtenida por la multiplicación de $v(n)$ con un tren de impulsos periódicos $p(n)$, con período D, como se ilustra en la fig. II.3. La representación de la serie de Fourier discreta de $p(n)$ es

$$p(n) = 1/D \sum_{k=0}^{D-1} e^{j2\pi kn/D} \qquad \text{(II.23)}$$

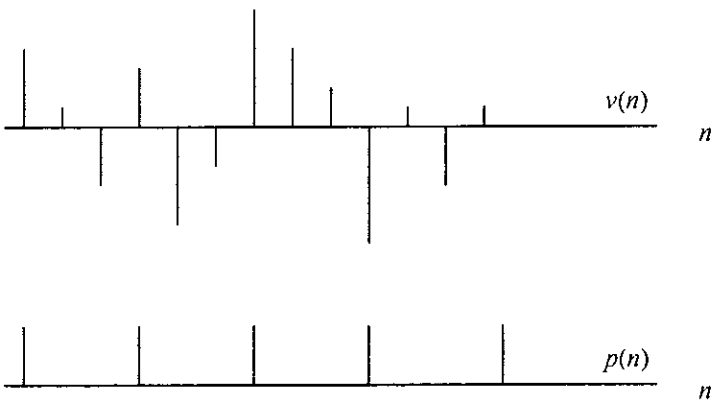


Fig. II.3 MULTIPLICACIÓN DE $v(n)$ CON UN TREN DE IMPULSOS PERIÓDICOS $p(n)$ CON PERIODO D=3.

$$v'(n) = v(n)\, p(n) \qquad \text{(II.24)}$$
$$y(m) = v'(mD) = v(mD)\, p(mD) = v(mD) \qquad \text{(II.25)}$$

La transformada z de la secuenica de salida $y(m)$ es

$$Y(z) = \sum_{m=-\infty}^{\infty} y(m)\, z^{-m} \qquad \text{(II.26)}$$

$$Y(z) = \sum_{m=-\infty}^{\infty} v'(mD)\ z^{-m} \tag{II.27}$$

$$Y(z) = \sum_{m=-\infty}^{\infty} v'(m)\ z^{-m/D} \tag{II.28}$$

Sustituyendo la ec (II 23) y (II 24) en (II.28)

$$Y(z) = \sum_{m=-\infty}^{\infty} v(m) \left[1/D \sum_{k=0}^{D-1} e^{j2\pi km/D} \right] z^{-m/D} \tag{II.29}$$

$$Y(z) = 1/D \sum_{k=0}^{D-1} \sum_{m=-\infty}^{\infty} v(m) \left(e^{-j2\pi k/D} z^{1/D}\right)^{-m} \tag{II 30}$$

$$Y(z) = 1/D \sum_{k=0}^{D-1} V\left(e^{-j2\pi k/D} z^{1/D}\right) \tag{II.31}$$

$$Y(z) = 1/D \sum_{k=0}^{D-1} H\left(e^{-j2\pi k/D} z^{1/D}\right) X\left(e^{-j2\pi k/D} z^{1/D}\right) \tag{II.32}$$

La razón de $y(m)$ es $F_y=1/T_y$, la frecuencia variable, la cual se denota como ω_y, es en radianes relativa a la razón de muestreo F_y, esto es,

$$\omega_y = 2\pi F/F_y = 2\pi F T_y \tag{II.33}$$

Las razones de muestreo son relacionados por la expresión

$$F_y = F_x/D \tag{II.34}$$

Esto hace que las frecuencias sean variables

$$\omega_x = 2\pi F/F_x = 2\pi F T_x \tag{II.35}$$

son relacionados por

$$\omega_x = D\omega_y \tag{II.36}$$

II 3.2 Interpolación por un factor de U.

Un incremento en la razón de muestreo por un factor de U puede ser realizado por la interpolación de U-1 nuevas muestras entre sucesivos valores de la señal. A continuación se describe un proceso que preserva la forma espectral de la secuencia de señal $x(n)$.

Se denota $v(m)$ como una secuencia con razón $F_y=UF_x$, la cual es obtenida de $x(n)$ adicionando U-1 ceros entre valores sucesivos de $x(n)$, De este modo

$$v(m) = \begin{cases} x(m/U) & m = 0, \pm U, +2U, \ldots \\ 0 & \text{para otro valor.} \end{cases} \qquad \text{(II.37)}$$

y su razón de muestreo es idéntica a la razón de $y(m)$. La transformada de $v(m)$ es:

$$V(z) = \sum_{m=-\infty}^{\infty} v(m)\, z^{-m}$$

$$V(z) = \sum_{m=-\infty}^{\infty} x(m)\, z^{-mU} \qquad \text{(II.38)}$$

$$V(z) = X(z^U) \qquad \text{(II.39)}$$

El espectro correspondiente de $v(m)$ será

$$V(\omega_y) = X(\omega_y U) \qquad \text{(II.40)}$$

donde ω_y es la frecuencia variable relativa a la nueva razón de muestreo F_y ($\omega_y=2\pi F/F_y$). Ahora la relación entre las razones de muestreo es $F_y=UF_x$ y la relación entre las frecuencias variables ω_x y ω_y son relacionadas de acuerdo a la formula

$$\omega_y = \omega_x / U \qquad \text{(II.41)}$$

Observamos que la razón de muestreo se incrementa por la adición de U-1 ceros muestreados entre sucesivos valores de $x(n)$ resulta en una señal cuyo espectro $V(\omega_y)$ es una repetición periódica U del espectro de la señal de entrada $X(\omega_x)$

Ya que las componentes en frecuencia de $x(n)$ en el rango $0\le\omega_y\le\pi/U$ son únicas. Las imágenes de $X(\omega)$ arriba de $\omega_y=\pi/U$ deben ser rechazadas pasando la secuencia $v(m)$ a través de un filtro paso bajas con respuesta en frecuencia $H_U(\omega_y)$, la cual idealmente tiene las características.

$$H_U(\omega_y) = \begin{cases} C & 0\le|\omega_y|\le\pi/U \\ 0 & \text{para otro valor.} \end{cases} \qquad \text{(II.42)}$$

siendo C el factor de escala para normalizar la secuencia de salida $y(m)$ apropiadamente. Consecuentemente el espectro de salida es

$$Y(\omega_y) = \begin{cases} CX(\omega_y U) & 0 \le |\omega_y| \le \pi/U \\ 0 & \text{para otro valor} \end{cases} \qquad \text{(II.43)}$$

El factor de escala C es seleccionado, así que la salida $y(m)=x(m/U)$ para $m=0, \pm U, \pm 2U,$. Por conveniencia matemática, se selecciona el punto $m=0$. De este modo

$$y(0) = 1/2\pi \int_{-\pi}^{\pi} Y(\omega_y) d\omega_y \qquad \text{(II.44)}$$

$$y(0) = C/2\pi \int_{-\pi U}^{\pi U} X(\omega_y U)\, d\omega_y \qquad \text{(II 45)}$$

como $\omega_x=\omega_y/U$ la ec. anterior queda

$$y(0) = C/U2\pi \int_{-\pi}^{\pi} X(\omega_x)\, d\omega_x \qquad \text{(II.46)}$$

$$y(0) = (C/U)x(0)$$

Por lo tanto $C=U$ es la factor de normalización deseado

La secuencia de salida $y(m)$ puede ser expresada como la convolución de la secuencia $v(n)$ con la respuesta al impulso $h(n)$ del filtro paso bajas. De este modo

$$y(m) = \sum_{k=-\infty}^{\infty} h(m-k)\, v(k) \qquad \text{(II.47)}$$

si $v(k)=0$ excepto en multiplos de U, donde $v(kU)=x(k)$ la ec. anterior llega a ser

$$y(m) = \sum_{k=-\infty}^{\infty} h(m-kU)\, x(k) \qquad \text{(II.48)}$$

II.4 Conclusiones

Términos tales como filtrado, interpolación, extrapolación entre otros son comunmente encontrados en el proceso de señales digitales El procesamiento digital de señales (DSP, por sus siglas en inglés) es una área de la ciencia e ingeniería que se ha incrementado sensiblemente durante los últimos años. Las computadoras digitales y el hardware digital de las dos últimas décadas en conjunto con los conceptos básicos de procesamiento digital han hecho posible la construcción de sistemas digitales altamente sofisticados Unos ejemplos de estos sistemas digitales en los que se puede aplicar el procesamiento digital de señales puede ser una señal natural de un electrocardiograma (ECG) tal señal provee información sobre la operación del corazón del paciente. Similarmente un electroencefalograma (EEG) provee información acerca de la actividad del cerebro. La voz es otra señal donde se aplica el DSP Estos tres ejemplos son señales que involucran solo una variable independiente, siendo esta el tiempo. La señal de voz puede ser analizada, sintetizada, comprimida o reconocida, estas son algunas de las formas en que el procesamiento digital de señales es aplicado a la señal de voz. En este trabajo nos enfocaremos al reconocimiento de voz y en el siguiente capítulo daré algunos conceptos para los procesos de voz

Capítulo III

PRINCIPIOS DEL RECONOCIMIENTO DE VOZ.

III.1 Introducción.

El reconocimiento de patrones aplicado a reconocimiento de voz es básicamente usar la señal de voz directamente, sin determinar las características explícitas (en el sentido fonético-acústico) y segmentar dicha señal. Como en la mayoría de las aplicaciones de reconocimiento de patrones, el método tiene dos pasos: el entrenamiento de los patrones de voz y el reconocimiento de patrones que se hace por medio de comparación. La voz "conocida" es llevada a un sistema por medio del procedimiento de entrenamiento El concepto anterior es que si hay suficientes versiones de un patrón para ser reconocidas (sea este un sonido, una palabra, una frase, etc.) son incluidas en un conjunto entrenado para introducirlas a un algoritmo, el procedimiento de entrenamiento debe ser capaz de caracterizar adecuadamente las propiedades acústicas del patrón (sin preferencia de algún patrón para el procedimiento de entrenamiento). Este tipo de caracterización de voz por medio de entrenamiento es llamado clasificación de patrones, porque la máquina aprende cuales propiedades acústicas de la clase de voz son confiables y repetibles a través de todas las pruebas de entrenamiento de los patrones La utilidad del método radica en la comparación de patrones, el cual compara directamente la voz desconocida (la voz para ser reconocida), con cada posible patrón aprendido previamente en la fase de entrenamiento y clasifica la voz desconocida de acuerdo a las bondades de compatibilidad de los patrones [15].

El diagrama de una aplicación de reconocimiento de patrones a reconocimiento de voz es presentado en la fig. III.1. El paradigma de reconocimiento de patrones tiene cuatro pasos:

1. Medidas características, en la que la secuencia de medidas es realizada sobre la señal de entrada para definir el "patrón de prueba". Para señales de voz las medidas características son usualmente la salida de algún tipo de técnica de análisis espectral, tal como un analizador de banco de filtros, un análisis de codificación de predicción lineal, o una transformada discreta de Fourier (DFT).
2. El entrenamiento del patrón, en el cual uno o más patrones de prueba correspondientes a los sonidos de voz de la misma clase son usados para crear un patrón representativo de las características de aquella clase. El patrón resultante, generalmente llamado un "patrón de referencia", puede ser un ejemplar o plantilla, derivada de algún tipo de técnica promedio, o puede ser un modelo que determine las estadísticas de las características del patrón de referencia
3. Clasificación de patrones, en este paso el patrón de prueba desconocido es comparado con cada clase de patrón de referencia y una medida de similaridad (distancia) entre el patrón de prueba y cada patrón de referencia es calculada. Para comparar los patrones de voz (los cuales consisten de una secuencia de vectores espectrales), se requiere de ambos: una medida de distancia local, en la cual la distancia local es definida como la "distancia" espectral entre dos vectores espectrales bien definidos, y un procedimiento

de alineamiento en el tiempo global (llamado algoritmo de deformación de tiempo dinámico), el que compensa para diferentes razones de velocidad de habla (tiempos de escala) de los dos patrones.

4 Decisión lógica, en esta fase las puntuaciones de similaridad de los patrones de referencia son usadas para decidir cual patrón de referencia (o posiblemente cual secuencia de patrones de referencia) compatibiliza mejor con el patrón de prueba desconocido.

Los factores que distinguen los diferentes reconocedores de patrones son los tipos de medidas características, la elección de plantillas o modelos para patrones de referencia, y el método usado para crear patrones de referencia y clasificar patrones de prueba desconocidos.

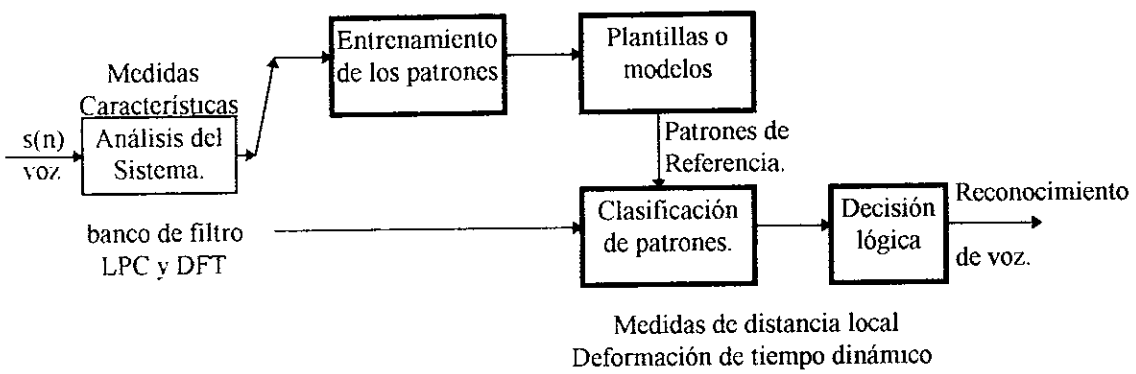


Figura III.1 DIAGRAMA DE RECONOCIMIENTO DE PATRONES APLICADO A RECONOCIMIENTO DE VOZ

Los aspectos que fortalecen y debilitan el modelo de reconocimiento de patrones son los siguientes:

- El desempeño del sistema es sensitivo a la cantidad de datos de entrenamiento disponibles para crear patrones de referencia.
- Los patrones de referencia son sensitivos a las características del ambiente y de transmisión del medio usado para crear la señal de voz; esto es porque las características espectrales de la voz son afectadas por la transmisión y el ruido de fondo.
- Ningún conocimiento específico de voz es usado explícitamente en el sistema; por lo tanto, el método es relativamente insensible a elecciones de vocabularios de palabras, sintaxis, y semántica.
- La carga computacional para el entrenamiento y clasificación de patrones es en la mayoría de los casos proporcionalmente lineal al número de patrones que están siendo entrenados o reconocidos; por lo tanto, el cálculo para un número grande de clases de sonidos puede y frecuentemente llega a ser prohibido.

III.2. Modelo de Análisis Espectral.

La aproximación del análisis de Codificación por Predicción Lineal (por sus siglas en inglés LPC) como se ilustra en la fig. III.2, desempeña un análisis espectral sobre ventanas de voz (marcos de voz) con un modelo todo-polo. Esto significa que la representación espectral resultante $X_n(e^{jw})$ es restringida para ser de la forma $\sigma/A(e^{jw})$ donde $A(e^{jw})$ es un polinomio de p^{th} orden con transformada z

$$A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - \ldots - a_p z^{-p}.$$

El orden p es llamado el orden de análisis LPC. De este modo la salida del bloque de análisis espectral LPC es un vector de coeficientes (parámetros LPC) que especifican (paramétricamente) el espectro de un modelo todo-polo, que mejor compatibiliza el espectro de la señal sobre el periodo de tiempo en el cual el marco de muestras de voz fue acumulado.

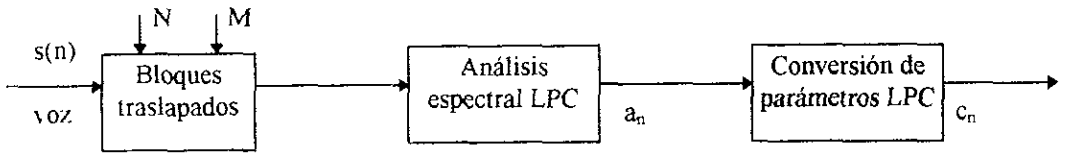


Figura III.2 MODELO DE ANÁLISIS LPC

III 2 1 Bloques traslapados.

En este paso la señal, $s(n)$ es empaquetada en marcos o bloques de N muestras, adyacentes que están siendo separados por M muestras. La fig. III.3 muestra como se traslapan los marcos donde $M=(1/3)N$. El primer marco consiste de las primeras N muestras de voz El segundo marco inicia M muestras de voz después del primer marco. y éstos dos se traslapan por N-M muestras Igualmente el tercer marco inicia $2M$ muestras después del primer marco (o M muestras después del segundo marco) y se traslapan éstos por N-$2M$ muestras. Este proceso continua hasta que toda la señal de voz es recolectada dentro de uno o más marcos. Se puede observar que si $M \leq N$, entonces los marcos adyacentes se traslapan como se muestra en la fig III 3, y la estimación espectral resultante será correlacionada de marco a marco; si $M << N$, entonces la estimación espectral LPC de marco a marco será completamente suavizada. De otra manera, si $M > N$, no existirá traslape entre los marcos adyacentes, de hecho parte de la señal de voz será totalmente perdida (i.e., nunca aparecerá en ningún marco de análisis), y la correlación entre la estimación espectral resultante LPC de marcos adyacentes contendrá una componente de ruido cuya magnitud se incrementa conforme se incremente M (i.e , porque la mayoría de la voz es omitida del análisis) Esta situación es intolerable en cualquier análisis práctico LPC para reconocimiento de voz. [11]

Es de suma importancia determinar si existe periodicidad en un frame de voz. Si existe la periodicidad, el frame es sonoro, de lo contrario el frame es clasificado como no sonoro. En caso de que el frame sea sonoro es necesario determinar el valor de la frecuencia fundamental, o también llamada frecuencia "pitch". La frecuencia fundamental se define como la frecuencia a la cual vibran las cuerdas vocales durante un sonido "sonoro". La frecuencia fundamental (f_0) es un parámetro, del cual es difícil obtener una estimación confiable a partir de la señal de voz. Normalmente, f_0 se encuentra entre 50 y 500 Hz. para voz "sonora". Para voz "no sonora", f_0 no esta definida (fig. III.5). La señal de voz contiene información tanto del pitch como de las resonancias (formantes) del conducto vocal sobre la señal de voz. Luego, existen distintos métodos para extraer el pitch intentando eliminar los efectos de la estructura formante sobre la señal de voz. Algunas de las formas para extraer el pitch son: el Método Cepstral, el Metodo SIFT (Simplified Inverse Filter Tracking) y el Método de Autocorrelación.

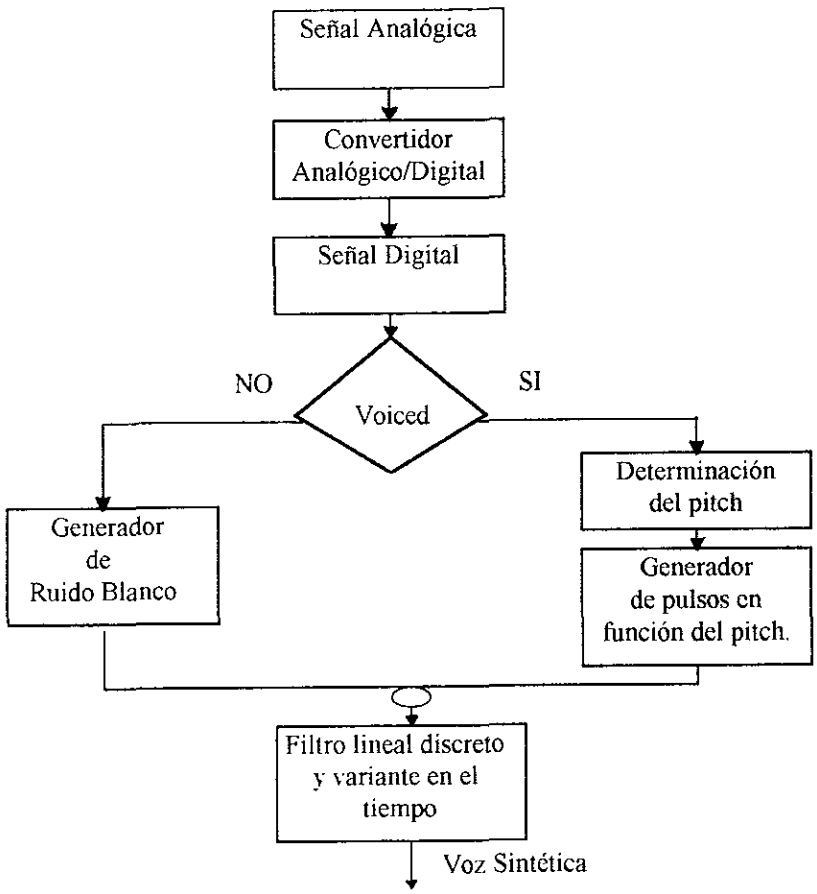


Fig. III.5 DIAGRAMA A BLOQUES PARA SÍNTESIS DE VOZ.

III.3 2. Método de autocorrelación para la obtención del pitch

Un gran número de diferentes métodos han sido propuestos para detectar el pitch. La detección del pitch por medio de la autocorrelación, es uno de los métodos más fuertes y confiables de detección de pitch y algunas de las razones son que, el cálculo es directamente hecho sobre la señal, es directo y claro.

Aunque la autocorrelación tiene algunas ventajas para detectar pitch, también existen varios problemas asociados con el uso de este método. A pesar de que la función de autocorrelación de una sección de señal de voz generalmente muestra un pico prominente en el período del pitch, los picos de la autocorrelación debido a la estructura formante detallada de la señal son frecuentemente presentados. Por lo que un problema es decidir cual de los varios picos de la autocorrelación corresponden al pico del pitch. Otro problema es el uso requerido de una ventana para calcular la función de autocorrelación de tiempo corto Un problema más, es que los picos de las formantes tienden a ser de mayor magnitud que el pico debido al período del pitch. Una dificultad final es elegir un tamaño de marco de análisis apropiado. El marco de análisis ideal debe contener de 2 a 3 períodos de pitch completo. De esta manera, para hablantes de pitch alto el análisis del marco debe ser corto (5-20ms), mientras que para hablantes de bajo pitch el marco de análisis debe ser largo (20-50ms) [5].

Para el cálculo de la autocorrelación se asume que las muestras que están fuera del marco son cero. Lo que implica que la función de autocorrelación sufre una ponderación lineal, la cual es uno para m=0 y cero para m = TAMAÑO DEL FRAME. Esta ponderación tiene el efecto de realzar el pico correspondiente al período pitch, con respecto a los picos que corresponden a múltiplos de dicho período, reduciendo de esta manera la posibilidad de que el pitch estimado se duplique o triplique [5].

III 3.2.1 Autocorrelación.

Dada la señal discreta en el tiempo que se esta procesando la llamaremos $x(n)$, para todas las n funciones de autocorrelación definida por:

$$\Phi_x(m) = \lim_{N \to \infty} (1 / 2N+1) \sum_{n=-N}^{N} x(n)x(n-m) \qquad \text{(III.1)}$$

Para una señal no estacionaria, tal como la voz, el concepto de una medida de autocorrelación de tiempo-largo es dada por la ec. (III.1), por lo que no es realmente significativa y es razonable definir una función de autocorrelación de tiempo-corto de la señal como

$$\Phi_l(m) = 1/N \sum_{n=0}^{N-1} [x(n+l)w(n)][x(n+l+m)w(n+m)] \qquad \text{(III.2)}$$

donde:

$0 <= m <= Mo-1$

$w(n)$ - es una ventana apropiada para el análisis.

N - es la longitud de la sección que esta siendo analizada.

N' - es el número de muestras de la señal usada en el cálculo de $\Phi_l(m)$

Mo - es el número de puntos de autocorrelaciones

l - es el índice de la muestra de inicio del marco.

para la detección del pitch generalmente $N' = N - m$ (III.3)

así que solamente las N muestras en el marco de análisis (por ej. $x(l)$, $x(l+1)$, . $x(l+N-1)$) son usadas en el cálculo de la autocorrelación. Los valores usados generalmente para Mo es 200 y para N es 300, correspondiendo a un período máximo del pitch de 20 mseg. (200 muestras a 10kHz. de razón de muestreo) y un tamaño del marco de análisis de 30 mseg.

III 3.2.2 Ventaneo

En todas las aplicaciones prácticas de procesamiento digital de señales, es necesario trabajar con términos cortos o frames de la señal, a menos que la señal sea de duración corta. Esto es especialmente verdadero si usamos técnicas convencionales de análisis de señales (tal como la voz) con dinámica no-estacionaria. En este caso es necesario seleccionar una porción de la señal que puede ser razonablemente asumida como estacionaria.

Recordando que una *ventana*, es decir *w(n)*, es real, y que es una longitud de secuencia usada para seleccionar un marco deseado de la señal original, es decir *x(n)*, por un simple proceso de multiplicación Algunas de las secuencias de ventanas más comúnmente usada son presentadas en la fig III 6. Por consistencia asumiremos que las ventanas son secuencias *causales* iniciando en el tiempo n=0. La duración se denotara usualmente como N. Las ventanas mas comúnmente usadas son simétricas en el tiempo $(N-1)/2$ donde este tiempo puede ser la mitad entre dos puntos muestras si N es par. Recordando que esto significa que las ventanas son secuencias de fase-lineal y por lo tanto tienen DTFTs (Transformadas Discretas de Fourier en el tiempo) que pueden ser escritas como

$$W'(\omega) = |W(\omega)|e^{-j\omega((N-1)/2)}$$

donde el término fase es una simple característica lineal correspondiente al retardo de la ventana que hace esto causal [14].

Las ventanas comúnmente usadas tienden a tener un espectro "pasa-bajas" con un "lóbulo principal" a bajas frecuencias y variaciones atenuadas en "lóbulos laterales". Esto es

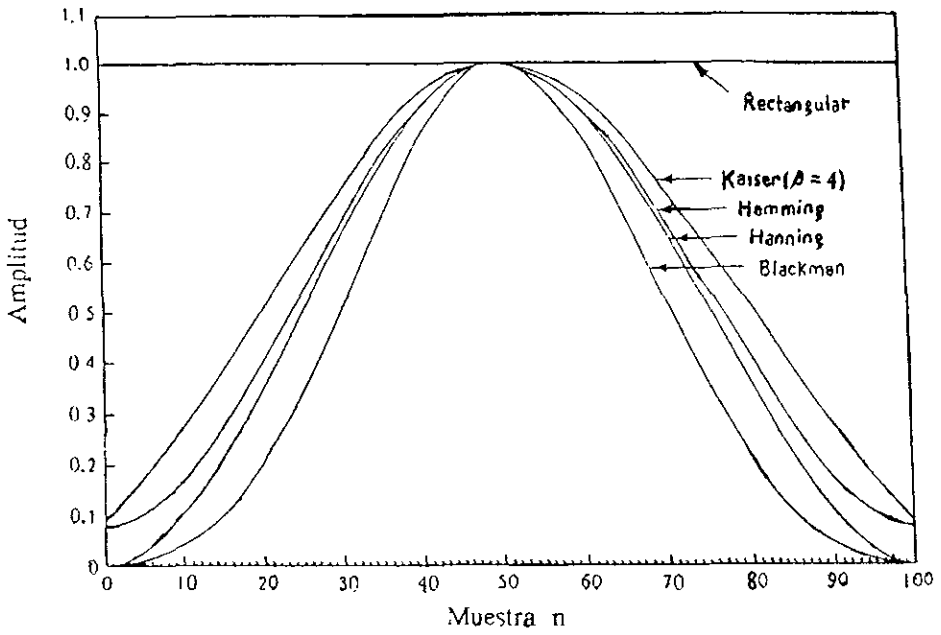


Fig III.6 DEFINICIONES Y EJEMPLOS DE TRAZOS PARA LAS VENTANAS *RECTANGULARES, KAISER, HAMMING, HANNING,* Y *BLACKMAN*. TODOS LOS TRAZOS SON PARA VENTANAS DE LONGITUD N=101, Y PARA LA VENTANA DE *KAISER*, β=4, [14].

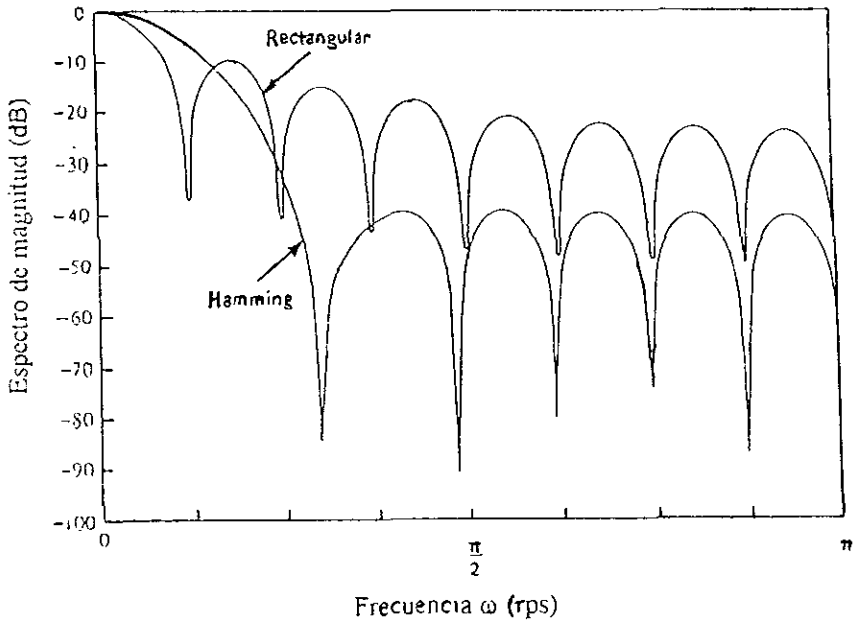


fig III 7 LA MAGNITUD ESPECTRAL DE LAS VENTANAS *RECTANGULAR* Y *HAMMING*. LA LONGITUD DE LA VENTANA N=16 ES USADA EN CADA CASO PARA CLARIFICARLO. NOTE QUE EL "ANCHO DE BANDA" NOMINAL (ANCHO DEL LÓBULO PRINCIPAL) ES $2\pi/N=\pi/8$ PARA EL CASO *RECTANGULAR* Y APROXIMADAMENTE DOS VECES PARA LA VENTANA DE *HAMMING*. SIN EMBARGO ES 20 dB MEJOR FUERA DEL PASABANDAS, [14].

III.3.2.3. Preénfasis

La señal de voz digitalizada, $s(n)$, pasa a través de un filtro digital de bajo orden (típicamente un filtro FIR de 1er orden), para que espectralmente se uniformen en magnitud las altas y bajas frecuencias de la señal, y hacer esto menos susceptible a efectos de precisión finitos más tarde en el procesamiento de la señal. Esto es, que las componentes de alta frecuencia de voz en magnitud no sean menos significativas que las componentes de baja frecuencia. En otras palabras en el procesamiento digital de la señal $s(n)$ los resultados son satisfactorios en bajas frecuencias mientras que en altas frecuencias no es así. Por lo que la señal de voz es pasada a través de un filtro de preénfasis, el cual enfatiza las altas frecuencia antes de procesar la señal [11]. El filtro de preénfasis esta dado por la siguiente ecuación:

$$s'(n) = s(n) - a\, s(n\text{-}1)$$

donde:

- $s'(n)$ es la señal preenfatizada y
- $s(n)$ es la señal de entrada.
- a se encuentra entre 0 9 y 1.0

$$s'(n) = s(n) - 0.9375\, s(n\text{-}1). \qquad \text{(III.8)}$$

A la salida del sintetizador la señal es postenfatizada por medio del filtro:

$$s'(n) = s(n) + 0.9375\, s(n\text{-}1). \qquad \text{(III.9)}$$

III 3 2 4. Procesos para reducir los efectos de la estructura formante.

Para reducir los efectos de la estructura formante como lo presenta en una forma detallada la función de autocorrelación de tiempo-corto, dos funciones de preprocesos fueron usadas antes del cálculo de la autocorrelación ec. (III 2). La fig. III.8 presenta un diagrama a bloques del proceso. La señal de voz es primero filtrada por medio de un filtro digital FIR paso-bajas de 0 a 900 Hz , de fase lineal La salida de este filtro es usada como entrada a los dos procesos, etiquetados como P1 y P2 en la fig. III.8. Los procesos P1 y P2 pueden o no ser idénticos.

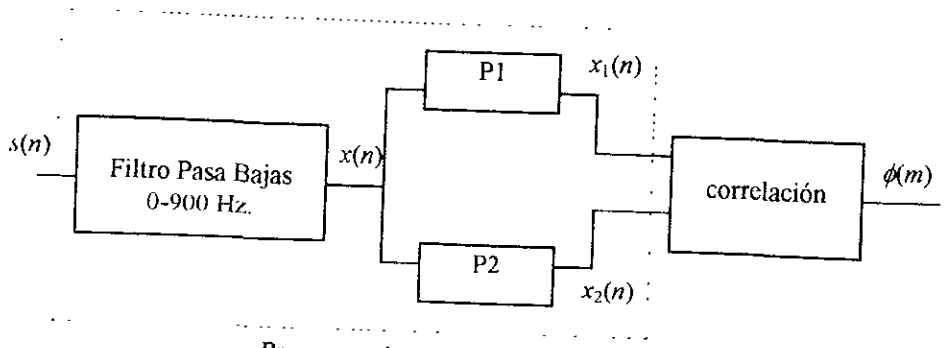


Fig III.8 DIAGRAMA A BLOQUES DEL PROCESAMIENTO DE CORRELACIÓN NO LINEAL.

Tres tipos de procesos serán mencionados a continuación. Estas son clasificadas de acuerdo a sus características de cuantización de entrada-salida de la siguiente manera:

El primer tipo de no linealidad es un recorte central comprimido cuya salida $y(n)$ obedece la siguiente relación y la entrada es $x(n)$

$$y(n) = \text{clp}\,[x(n)] = \begin{cases} (x(n) - C_L), & x(n) >= C_L \\ 0, & |x(n)| < C_L \\ (x(n) + C_L), & x(n) <= -C_L \end{cases}$$

donde C_L es el recorte de umbral para las tres no linealidades.

La segunda no linealidad es un simple recorte central con la relación de entrada-salida:

$$y(n) = \text{clp}\,[x(n)] = \begin{cases} x(n), & x(n) >= C_L \\ 0, & |x(n)| < C_L \\ -x(n), & x(n) <= -C_L \end{cases}$$

La tercera no linealidad es la combinación del centro y pico de la señal recortada con la relación de entrada-salida.

$$y(n) = \text{sgn}\,[x(n)] = \begin{cases} 1, & x(n) >= C_L \\ 0, & |x(n)| < C_L \\ -1, & x(n) <= -C_L \end{cases}$$

El recorte de umbral C_L para cada una de estas linealidades fue exactamente el método usado en [24], i.e obtener el máximo nivel de señal absoluto que exista en el primero y último tercio del marco de análisis y de estos dos valores escoger el menor para poner el recortador como un porcentaje fijo del 68% de este último valor escogido

III.3.2 5 Sonoro o no-sonoro.

El último paso en la estimación del pitch consiste en encontrar el valor máximo de la función de autocorrelación (normalizada con respecto a $r(0)$) en el intervalo de 0 al TAMAÑO DEL FRAME. Tanto la localización como el valor del pico máximo son guardados en memoria. Si el valor del pico máximo excede un umbral sonoro-no_sonoro (del orden de 0 30), el frame de voz es clasificado como sonoro, y el período pitch es igual a la posición del pico máximo. Ahora, si el pico máximo cae por debajo del valor de dicho umbral, el frame de voz es declarado como no_sonoro. En caso de que el bloque sea declarado como sonoro se genera un tren de pulsos de frecuencia igual a la del pitch, si es declarado no_sonoro se genera ruido blanco, en cualquiera de los dos casos (tren de pulsos o ruido blanco) la señal pasa a través del filtro predictor (el cual contiene los coeficientes a_i's obtenidos del método de Levinson-Durbin)

III 3.3. Filtro Predictor.

Tomando en cuenta la fig III.4, las muestras de voz sonora son linealmente predecibles, en términos de las pasadas p muestras de voz, excepto para la muestra inicial de cada pitch. Esta propiedad de las señales de voz, es utilizada para determinar los coeficientes del predictor [9].

El error de predicción E_n, se define como la diferencia entre la muestra $s(n)$ y la muestra predicha $s'(n)$,
donde

$$s'(n) = \sum_{k=1}^{p} a(k)\, s(n-k)$$

luego E_n esta dado por:

$$E_n = s(n) - s'(n)$$

Se define $\{E_n^2\}$ como el error cuadrático medio sobre las n muestras del marco de voz a analizar, excepto aquellas que se encuentran al principio de cada período pitch, esto es:

$$\{E_n^2\}_{av} = \{(s(n) - s'(n))^2\}_{av}$$

Los coeficientes predictores a_k son aquellos que minimizan el error cuadrático medio $\{E_n^2\}_{av}$. Cuando la derivada parcial de $\{E_n^2\}$ con respecto a a_k es igualada a cero, se obtiene la ecuación de Yule-Walker (ec III 11).

$$\mathbf{A}_p = \mathbf{R}_p^{-1}[k-1]\, \mathbf{r}_p\,[k] \qquad \text{(III.11)}$$

donde:

- $\mathbf{R}_p[k-1]$ es una matriz Toeplitz simétrica y cuyos elementos diagonales y subdiagonales son los mismos.
- $\mathbf{r}_p^T[k] = [\ r_y[p]\ r_y[p-1]\ r_y[p-2]\ \ldots\ r_y[2]\ r_y[1]\ ]$
- $\mathbf{A}_p$ es un vector de dimensiones p que contiene parámetros al orden p de la cual se desea encontrar la solución.

Los coeficientes a_k son elementos del vector $\mathbf{A}_p$ y son los que minimizan el error cuadrático medio $\{E_n^2\}$, los coeficientes a_k se obtienen al invertir la matriz $\mathbf{R}_p[k-1]$ con la que se ocuparían N^3 operación y solamente serían N^2 operaciones usando el Algoritmo de Levinson-Durbin. Otro de los métodos comúnmente utilizados para la obtención de los parámetros predictores en el procesamiento de voz es la recursión de Leroux-Gueguen

III.3.3.1. Estimación de los parámetros de Codificación por Predicción Lineal (LPC).

Esta es una técnica paramétrica para la representación de señales de voz, la cual trata de modelar el espectro de dichas señales como un proceso autoregresivo.

Para una representación completa del modelo LPC, los parámetros del filtro del tracto vocal (coeficientes del filtro a_i's y la ganancia) deben ser determinados, para hacer que

$$s'(n) = -a_1 s(n-1) - a_2 s(n-2) - \ldots - a_p s(n-p)$$

sea el estimado de $s(n)$ obtenido de las muestras previas.

Coeficientes Lineales Predictores

A) El algoritmo de Levinson-Durbin comienza con los coeficientes de autocorrelacion $r(i)$, $i=0,\ldots,p$ y calcula recursivamene los coeficientes del filtro a_i, por medio de las siguientes relaciones:

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

$$K_i = (-r(i)+a_1^{(i-1)}r(i-1)+ \ldots +a_{i-1}^{(i-1)}r(1)\ ) / \ E(i-1) \qquad i=1,\ldots,p$$

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

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

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

Los coeficientes $a_j^{(i)}$, $j=1, \ldots, i$ son los coeficientes del filtro de un modelo de orden *ith*. Por lo tanto, los coeficientes del modelo del orden *pth* deseado son:

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

La solución de Durbin da los parámetros K_i^f, $i=1,\ldots,p$, y $E(p)$ que es igual al cuadrado de la ganancia necesario en el modelo de síntesis:

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

Esta cantidad puede ser codificada como uno de los parámetros necesarios para síntesis. Sin embargo, dado que

$$E(p) = (1-K_1^2)(1-K_2^2)\ldots(1-K_p^2)r(0)$$

en lugar de $E(p)$ se puede decodificar y transmitir $r(0)$, que es la energía del marco de voz analizada Entonces, G es recuperado multiplicando $r(0)$ por $(1-K_1^2)(1-K_2^2)\ldots(1-K_p^2)$ durante la síntesis. Esto es preferible, porque el modelo de síntesis es menos sensible a la cuantizacion del ruido de $r(0)$ que el de G.
El grupo de parámetros K_i, i=1, .. ,p son llamados coeficientes de reflexión o coeficientes PARCOR (PARtial CORrelation), y juegan un papel importante en el método LPC, a continuación hablaremos de estos.

B) Los coeficientes de reflexión son extremadamente útiles en sistemas, donde la memoria para el almacenamiento de datos sea limitada, ya que mediante una técnica de sintesis se puede generar la señal. Otro hecho importante sobre estos coeficientes es que su obtención para un orden N, también implica que se ha obtenido la solución para modelos de orden menor que N, lo cual es de gran utilidad en sistemas de procesamiento donde se requiera de una estimación del orden del modelo.
En términos generales el modelo trata de reproducir, tan fielmente como sea posible, el espectro de la señal de voz. Al incrementar el orden del modelo, la resolución con la que el espectro de la señal es reproducido, aumenta, mientras que al disminuir el orden del modelo dicha resolución disminuye hasta que con un orden bajo reproduce solo las caracteristicas mas sobresalientes del espectro de la señal original.
Una característica muy importante de este modelo es su inexactitud en regiones donde el espectro de la señal tiene una energía baja, por lo que seria deseable limitar en su parte baja el rango dinámico de la señal. Existen diversos métodos para lograr dicha reducción del rango dinámico en un modelo LPC [9]:

- El método estabilizado de covarianza, el cual reduce el rango dinámico en el dominio de la frecuencia (espectro).
- El método de ponderación perceptual, el cual amplia ligeramente los anchos de banda del modelo LP.
- El método de autocorrelación, en el cual una pequeña cantidad de ruido es adicionada a la función de autocorrelación

Este último método es el más simple y efectivo La función de autocorrelación es modificada antes del cálculo de los coeficientes del filtro o de los coeficientes de reflexión.

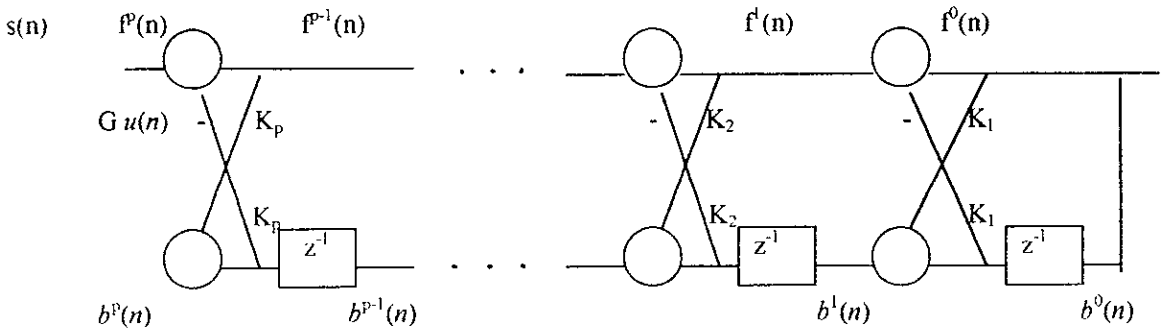


Fig III.9 FILTRO LATTICE PARA SINTETIZAR VOZ PARTIENDO DE LA EXCITACIÓN $G_u(n)$ Y LOS COEFICIENTES DE REFLEXIÓN K_i

III 3 4. Algoritmos para síntesis de voz.

Síntesis de Voz por medio de la Autocorrelación para la detección de pitch es el procedimiento que a continuación se describe.

Para la obtención de la Autocorrelación la señal de voz primeramente pasa por dos procesos para reducir los efectos de las resonancias del conducto vocal sobre la periodicidad de la señal. El proceso de recorte central consiste en suprimir la señal entre ciertos niveles, matemáticamente podemos expresarlos así

$$y(n) = \text{sgn}[x(n)] = \begin{cases} 1, & x(n) >= C_L \\ 0, & |x(n)| < C_L \\ -1, & x(n) <= -C_L \end{cases}$$

La manera de escoger el nivel de recorte (C_L) es tomar el máximo valor absoluto de la primera y última tercera parte (256/3 = 85) de la ventana (tamaño de la ventana = 256) de análisis De estos dos valores máximos escoger el menor y multiplicarlo por 0.8, ya que numerosas simulaciones demuestran que dicho valor optimiza los resultados

A continuación se calcula la autocorrelación para cada ventana

$$\Phi_x(m) = \sum_{n=0}^{256-m} x(n)x(n+m) \qquad m = M_i, M_i + 1, \ldots, M_f$$

donde

- M_i es el retraso inicial típicamente de 25
- M_f es el retraso final típicamente de 200

Los valores de 25 y 200 corresponden a un rango de Pitch de 25 a 200 Hz para una frecuencia de muestreo de 10000 Hz Adicionalmente $\Phi_x(0)$ es calculada con el objeto de normalizar la función de Autocorrelación

El siguiente paso para la estimación del pitch es localizar y guardar el valor máximo de la autocorrelación del intervalo de [M_i ,M_f] así como su posición. Si dicho valor máximo excede el umbral sonoro-no_sonoro (en el orden de 0.3). La ventana de voz es clasificada como sonora y la excitación al filtro de síntesis será un tren de pulsos en función de la posición del pitch De otra manera, si el pico máximo cae por debajo del umbral, la ventana de voz es declarada no sonora y la excitación al filtro de síntesis será un ruido blanco.

Por otro lado, al marco de análisis se le aplica la ventana de Hamming (ver sección III 3.2 2), en seguida el preénfasis (ver sección III.3.2.3), posteriormente la autocorrelación para estimar los parámetros del filtro de síntesis. Una vez obtenidos los parámetro del filtro (a_i's) y la excitación (ruido blanco o tren de pulsos) se obtiene a la salida del filtro la Voz Sintética, que se le aplicará el deenfasis para no reproducir alteración en altas frecuencias. El diagrama a bloques para síntesis de voz se muestra en la fig. III.10

III.4 Conclusiones.

En el caso de éste sintetizador se utilizan diferentes filtros en cascada, de forma que la salida de uno se convierte en la entrada del siguiente. Tienen el inconveniente de no permitir el control estrictamente independiente de cada una de las resonancias pero, por otro lado, se acomodan más a la forma física con que el tracto vocal modula la señal acústica

La codificación del habla se puede reducir a unos pocos parámetros como son: la frecuencia fundamental del segmento, la presencia o ausencia de sonoridad y el nivel energético.

La técnica de codificación por predicción lineal (LPC) introducida en el campo de la voz parte de un tratamiento temporal de la señal acústica con ciertos parámetros que permiten ahorrar la redundancia de información que se da en segmentos próximos de la voz Esta técnica constituye también una buena herramienta para la parametrización de la señal de voz y por un proceso inverso permite regenerar la señal acústicamente previamente parametrizada por un algoritmo LPC. Por otro lado la señal acústica permite extraer la información sobre el nivel de la señal acústica en cada instante.

Los procesos usados en éste capítulo como son ventaneo, preénfasis, autocorrelación, etc., son comúnmente encontrados en procesos de voz y serán utilizados en el Sistema Reconocedor de Voz de éste trabajo. Por lo que en el siguiente capítulo solamente se mencionan sin ahondar en dichos conceptos.

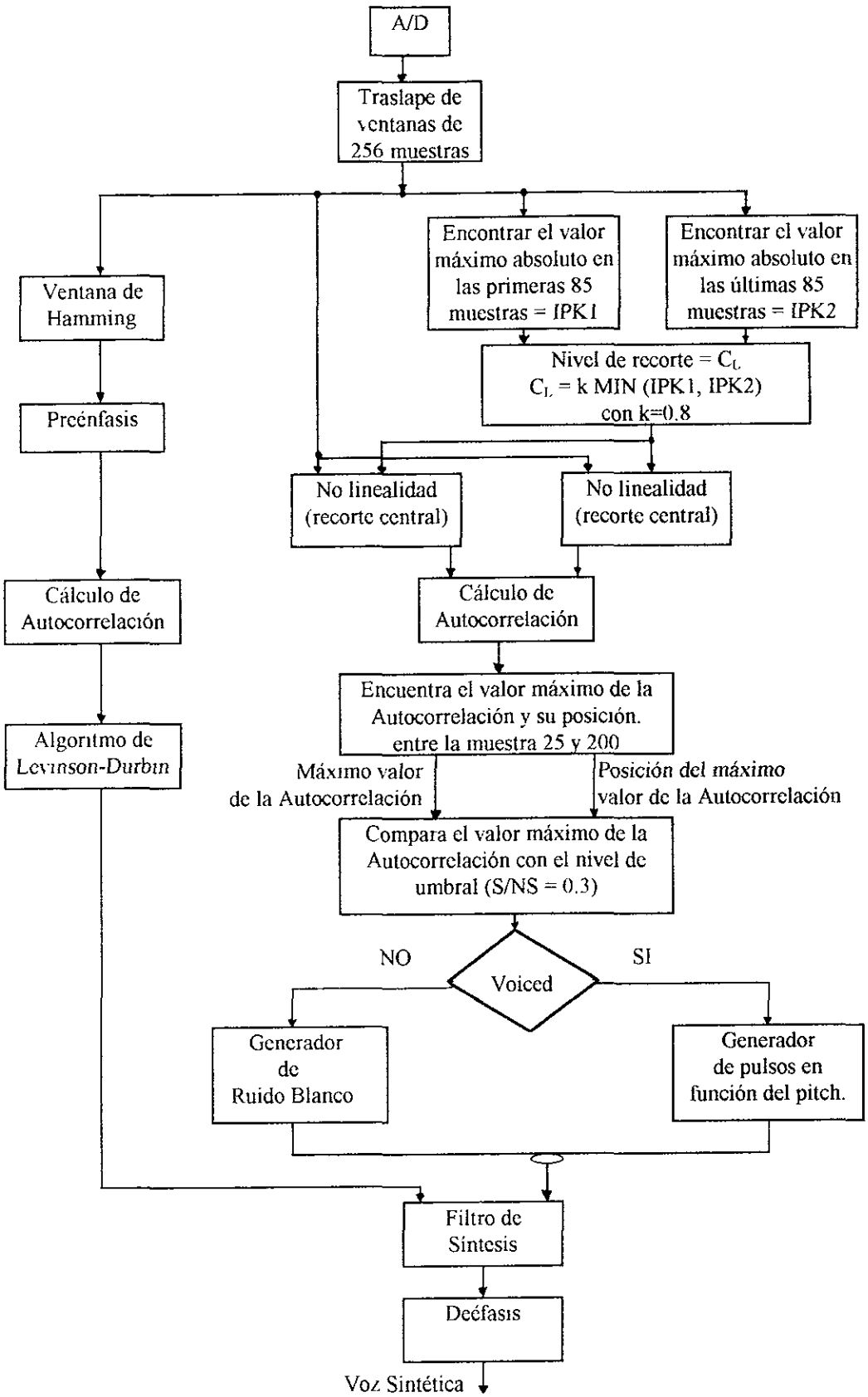


Fig. III.10 DIAGRAMA A BLOQUES DEL ALGORITMO PARA SÍNTESIS DE VOZ

Capítulo IV

RECONOCIMIENTO CLÁSICO DE VOZ.

IV.1 Introducción.

El análisis y síntesis de voz son unos de los diferentes tipos de procesamiento digital de voz, como también lo es el reconocimiento de voz. Tanto el reconocimiento como la síntesis de voz siguen procedimientos parecidos o incluso iguales hasta cierta etapa de su proceso. Es por esto que en este capítulo solo mencionaré algunas fases del proceso, ya que se trataron con detalle en el capítulo anterior.
Existen varios caminos para reconocer voz, como son Cadenas de Markov, Cuantización Vectorial, Lógica Difusa, y Codificación por Predicción Lineal (LPC), siendo el enfoque LPC el método tradicional y el que trataré en este capítulo.

IV.2 Modelo de Codificación por Predicción Lineal (LPC) para reconocimiento de voz.

El método LPC ha sido ampliamente usado por muchos años debido a que cumple con los siguientes puntos:

- LPC provee un buen modelo para la señal de voz. Esto es especialmente bueno para regiones sonoras casi-estacionarias en las cuales el modelo todo-polo de LPC provee una buena aproximación a la envolvente espectral del tracto vocal. Durante regiones no-sonoras de la señal, el modelo LPC es menos efectivo que para regiones sonoras, pero aun así proporciona un modelo aceptable para propósitos de reconocimiento de voz.
- La manera en la cual la técnica LPC es aplicada al análisis de señales de voz nos lleva a una razonable separación del tracto vocal (sonoro y no-sonoro).
- LPC es un modelo analítico, tiene una buena precisión matemática, además se implementa de una manera simple y directa ya sea en software o hardware.
- El modelo LPC trabaja bien en aplicaciones de reconocimiento.

IV.2.1 El modelo LPC

La idea básica del modelo LPC es que una muestra de voz a un tiempo n, $s(n)$, puede ser aproximada como una combinación lineal de las p pasadas muestras de voz, tal que

$$s(n)=a_1 s(n-1)+ a_2 s(n-2)+ \ldots +a_p s(n-p) \qquad \text{(IV.1)}$$

donde los coeficientes $a_1, a_2, \ldots, a_p$ son constantes en el marco o bloque de voz analisado. A continuación se muestra la ecuación anterior adicionando un termino de exitación, $Gu(n)$, dando:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n) \qquad \text{(IV.2)}$$

donde:

- $u(n)$ es una excitación normalizada
- G es la ganancia de la excitación

Expresando la ec. (IV.2) en el dominio-z, se obtiene la relación:

$$S(n) = \sum_{i=1}^{p} a_i z^{-i} S(z) + GU(z) \qquad \text{(IV.3)}$$

$$S(n) - \sum_{i=1}^{p} a_i z^{-i} S(z) = GU(z)$$

$$S(n) \left\{ 1 - \sum_{i=1}^{p} a_i z^{-i} \right\} = GU(z)$$

y la función de transferencia es:

$$H(z) = S(n) / GU(z) = 1 / \left\{ 1 - \sum_{i=1}^{p} a_i z^{-i} \right\} = 1 / A(z) \qquad \text{(IV.4)}$$

La interpretación de la ec. (IV.4) es dada en la fig. (IV.1), la cual presenta la fuente de excitación normalizada $u(n)$, siendo ponderada por la ganancia G y actuando como entrada al sistema todo-polo $H(z) = 1 / A(z)$, para producir la señal de voz $s(n)$. Sabiendo que la fuente de excitación para voz es un tren de pulsos periódico para regiones sonoras, o bien una fuente de ruido aleatorio para regiones no-sonoras

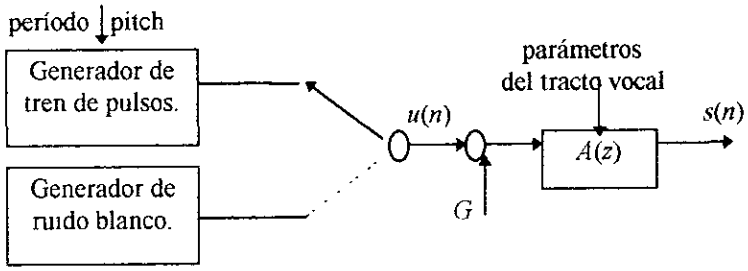


Figura IV.1 MODELO DE PREDICCIÓN LINEAL DE VOZ

IV.3 Reconocimiento de voz.

Como se mencionó en la introducción del capítulo III el reconocimiento de voz consiste básicamente de 4 etapas (ver fig. IV 2):

1. Medidas características, en la que la secuencia de medidas es realizada sobre la señal de entrada para definir el "patrón de prueba" (palabras del vocabulario o diccionario). Estas características son usualmente obtenidas de la salida de algún tipo de técnica de análisis espectral, tal como un analizador de banco de filtros, un análisis de codificación por predicción lineal o una transformada discreta de Fourier (DFT).
2. El entrenamiento del patrón, en el cual una o mas repeticiones de la misma palabra son usadas para crear un patrón representativo de las características de aquella clase o palabra El patrón resultante, generalmente llamado "patrón de referencia", puede ser una plantilla, derivada de algún tipo de técnica promedio, o puede ser un modelo que determine las estadísticas de las características del patrón de referencia.
3. Clasificación de patrones, en este paso el patrón de prueba desconocido es comparado con cada clase de "patrón de referencia" y una medida de similaridad (distancia) entre el "patrón de prueba" y cada "patrón de referencia" es calculada. Para comparar los patrones de voz (los cuales consisten de una secuencia de vectores espectrales), se requiere de ambos: una medida de distancia local, en la cual la distancia local es definida como la "distancia espectral" entre dos vectores espectrales bien definidos y un procedimiento de alineamiento en el tiempo global (llamado algoritmo de deformación de tiempo dinámico), el que compensa para diferentes razones de velocidad de habla (tiempos de escala) de los dos patrones.
4. Decisión lógica, en esta fase las puntuaciones de similaridad de los "patrones de referencia" con respecto al "patrón de prueba" son usadas para decidir cual "patrón de referencia" (o posiblemente cual secuencia de patrones de referencia) compatibiliza mejor con el "patrón de prueba" desconocido

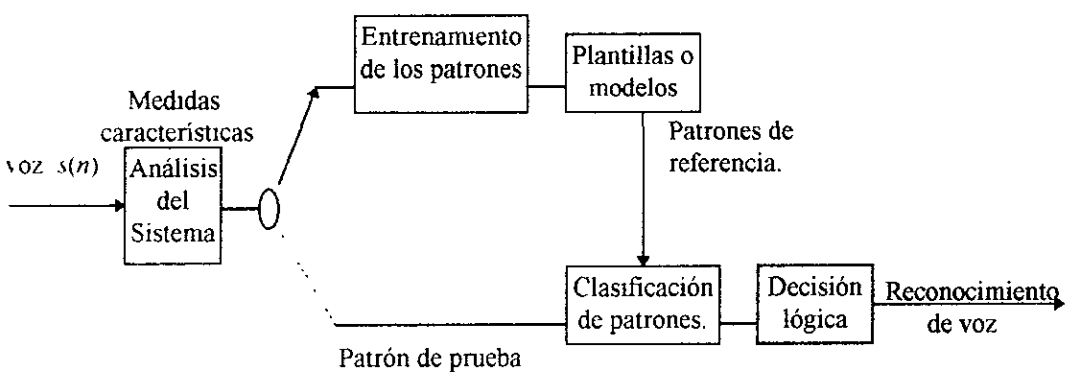


Figura IV.2 DIAGRAMA DE RECONOCIMIENTO DE VOZ.

IV.4 Detalles del Reconocimiento.

IV.4 1 Medidas características.

A continuación se mencionará cada paso del proceso de la extracción de las características de la señal. Primeramente se muestra el diagrama a bloques de esta evolución (ver fig. IV.3)

Bloques traslapados.- La señal de voz digitalizada $s(n)$ es la entrada en esta fase (ver apartado III.2.1).

Preénfasis.- Tiene como entrada la salida de *bloques traslapados* (ver apartado III.3.2.3).

Ventaneo.- La salida de la etapa anterior es la entrada para este paso (ver apartado III.3.2.2)

Coeficientes de Autocorrelación.- Este paso se detalla en el apartado III.3.2.1.

Coeficientes LPC.- El cálculo de estos coeficientes se muestra en la sección III.3.3.1.

Medidas de distancia.- Existen varios caminos para obtener las medidas de distancia y se explican algunos de ellos en la sección IV.4.1.1.

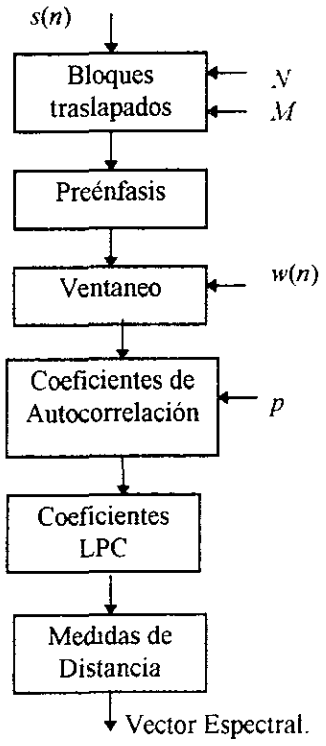


Fig IV 3 DIAGRAMA A BLOQUES PARA OBTENER LAS MEDIDAS CARACTERÍSTICAS USANDO LPC

IV 4.1 1 Medidas de distancias

Para procesamiento de voz una medida de distancia *d(x,y)* entre dos marco de datos de voz *x* y *y* deben satisfacer al menos las siguientes propiedades:

1 - Simetria *d(x,y)* *d(y,x)*.
2 - Definitividad positiva
d(x,y) *0* para *x* diferente de *y*
d(x,y) *0* para *x* igual a *y*
3.- *d(x,y)* debe tener una interpretación fisica en el dominio de la frecuencia.
4 - Debe ser posible evaluar eficientemente *d(x,y)*.

- El primer criterio asegura que una medida de distancia entre sonidos no distingue cual es el sonido de referencia y cual el sonido de prueba.
- El segundo criterio afirma que la distancia de *x* a *y* y de *y* a *x* es igual, excepto para el caso cuando *x=y* la distancia es cero.
- El tercer y cuarto criterio es para entender como la medida de distancia se relaciona con las propiedades espectrales de la voz y para evaluar tales medidas con una cantidad razonable de cálculos
- Un requerimiento adicional para hacer que *d(x,y)* sea una medida, es que debe satisfacer el triángulo de desigualdad

$$d(x,y) \leq d(x,z) + d(y,z)$$

a) Medidas espectrales y medidas espectrales logarítmicas RMS.

Considere dos modelos espectrales $\sigma/A(z)$ y $\sigma'/A'(z)$. El error o diferencia entre estos dos modelos sobre una magnitud logarítmica contra la escala de frecuencia es definido por

$$V(\theta) = \ln[\sigma^2 / |A(e^{j\theta})|^2] - \ln[(\sigma')^2 / |A'(e^{j\theta})|^2] \qquad \text{(IV.5)}$$

donde.

- θ es una frecuencia normalizada o ángulo en el plano *z*.

Un conjunto de elecciones lógicas para una medida de distancia entre los modelos espectrales es el conjunto de L_p normas o medidas definidas por d_p, donde

$$(d_p)^p = \int_{-\pi}^{\pi} |V(\theta)|^p \, d\theta/2\pi \qquad \text{(IV.6)}$$

- Para $p = 1$, se define la media absoluta de la medida espectral logarítmica.
- Para $p = 2$, se define la medida espectral logarítmica rms.
- Para casos cuando $p \rightarrow \infty$ se obtiene el pico de la diferencia espectral logarítmica

Las medidas L_P son lineales en el sentido que la multiplicación de $V(\theta)$ por una constante escalar resulta en una multiplicación de d_p por la misma constante escalar. Las medidas L_P son medidas verdaderas en el sentido que satisfacen el triángulo de desigualdad en atención a que son simétricas y positivas definidas. Las medidas L_P están relacionadas a variaciones de decibles en el domino espectral logarítmico usando el factor de multiplicación 10/ln(10)=4.34 [12].

Conforme p se incremente, los efectos de los errores grandes son más pesados que los errores pequeños, hasta el límite en que p se aproxima al infinito, solo el error máximo es incluido en la evaluación de d_p.

Aunque la ponderación no lineal en la integral de la ec. (IV.6) representa que la elección de p determina fuertemente cual porción de $V(\theta)$ hace la parte dominante del error, experimentalmente existe una fuerte correlación entre todas las medidas L_p para una amplia variedad de valores de p.

El problema básico con las medidas L_P es computacional. Dos FFT's y logaritmos son requeridos para obtener suficientes valores de $V(\theta)$ para estimar la integral de la ec. (IV 6) con una sumatoria. A continuación se presenta la medida de distancia cepstral que puede ser vista como un método eficiente para calcular la medida espectral logarítmica rms d_2.

b) Medida de distancia cepstral.

Si $A(z)$ es un polinomio en z^{-1} de orden Mth donde todas sus raíces estan dentro del circulo unitario, y $A(\infty)=1$, entonces la expansión de la serie de Taylor presenta que

$$\ln[A(z)] = -\sum_{k=1}^{\infty} c_k z^{-k}, \qquad \text{(IV.7)}$$

donde:

- $\{c_k\}$ define los coeficientes cepstrales.

Como el logaritmo de una magnitud cuadrada es dos veces la parte real del logaritmo, la expansión de la serie de Fourier para el modelo del espectro logarítmico puede ser:

$$\ln\left[\sigma^2 / |A(e^{j\theta})|^2\right] = \sum_{k=-\infty}^{\infty} c_k e^{-jk\theta}. \qquad \text{(IV.8)}$$

donde

$$c_o = \ln\left[\sigma^2\right] \qquad \text{(IV.9)}$$

y

$$c_{-k} = c_k. \qquad \text{(IV.10)}$$

Cada uno de los coeficientes cepstrales c_k y c'_k pueden ser recursivamente evaluados de los coeficientes del filtro, esto es, los coeficientes de $A(z)$ y $A'(z)$, y viceversa.

Una aplicación de la relación de Parseval para la medida de distancia L_2 da el siguiente resultado:

$$d_2^2 = \sum_{k=-\infty}^{\infty} (c_k - c_k')^2 = (c_o - c_o')^2 + 2\sum_{k=1}^{\infty} (c_k - c_k')^2. \qquad \text{(IV.11)}$$

Sin embargo, si se toma una serie truncada para definir una medida cepstral *u(L)* como

$$[u(L)]^2 = \sum_{k=-L}^{L} (c_k - c_k')^2 = (c_o - c_o')^2 + 2\sum_{k=1}^{L} (c_k - c_k')^2, \qquad \text{(IV.12)}$$

entonces *u(L)* puede ser interpretada como la distancia rms entre los espectros logarítmicos despues de que cada espectro logarítmico ha sido suavizado cepstralmente a L coeficientes. Dado que los primeros M coeficientes cepstrales (excluyendo los términos de ganancia c_o y c_o') determinan unicamente los coeficientes del filtro, es necesario que $L \geq M$ (M es el orden del filtro) para que *u(L)* sea una medida de distancia válida. Por ejemplo, si $L=M-1$, la propiedad de definitividad positiva es destruida tomando en cuenta que los primeros $M-1$ coeficientes de *A(z)* y *A'(z)* son identicos, el último término puede ser arbitrario, por ejemplo

$$d(x,y) = 0 \qquad \text{para } x \neq y$$

Conforme L se incrementa, *u(L)* se aproxima a d_2, es decir:

$$\lim_{L \to \infty} u(L) = d_2$$

Es de considerable interes saber, experimentalmente, que tan grande L debe ser antes de que *u(L)* llegue a ser una aproximación razonable de d_2. Calculando *u(L)* para valores de L iguales a M, $2M$ y $3M$, donde M=10, es sorprendente que con el mínimo número de coeficientes cepstrales (L=M), el coeficiente de correlación entre *u(M)* y d_2, es 0.98 sobre el rango de distancia desde 0-6 dB Para *u(2M)* y *u(3M)* las correlaciones fueron 0.997 y 0 999 [12].

En todos los casos los datos para *u(L)* caen abajo de d_2, lo cual es esperado, ya que *u(L)* se aproxima a d_2 desde abajo conforme L tiende a infinito. En la mayoría de los casos extremos, la máxima desviación entre d_2 y *u(L)* es aproximadamente 1.2 dB.

La alta correlación entre *u(L)* y d_2 puede ser cualitativamente explicada con referencia al dominio de la frecuencia o al dominio del tiempo. En el dominio de la frecuencia se puede notar que los anchos de banda de las formantes son generalmente mayores que 30 Hz, esto resulta del método de autocorrelación por predicción lineal. De este modo, suavisando los valores cepstrales de un espectro logarítmico, no se requerirá de un número excesivo de terminos antes de que el espectro logarítmico suavizado se aproxime al modelo del espectro logarítmico. El polinomio *A(z)* puede también ser factorizado en terminos de la forma $(1 - z_n z^{-1})$, donde z_n es la nth raíz de *A(z)*. Calculos residuales pueden ser aplicados a la ec. (IV 7) para obtener la relación entre el cepstrum y las raíces de *A(z)* como

$$c_k = -\frac{1}{k}\sum_{n=1}^{m}(z_n)^k .$$

Por lo tanto, la razón del decremento de los coeficientes cepstrales es una función directa de la raíz de magnitud más grande (el ancho de banda más estrecho). Para una razón de muestreo de 10 kHz. y un ancho de banda de 30 Hz., los valores cepstrales superan los primeros 20 términos que son reducidos a menos del 0.04 del máximo valor.

La conclusión importante de esta sección, es que la medida cepstral puede ser vista como un método muy eficiente para estimar la medida de distancia L_2 sin hacer uso de operaciones como la FFT o la DFT, ya que los coeficientes cepstrales pueden ser recursivamente obtenidos de los coeficiente de *A(z)* y *A'(z)*.

c) Razones de Verosimilitud.

Si una muestra *s(n)* es estimada por una combinación lineal de M muestras precedentes, el error predictor o residual puede ser expresado en la forma

$$e(n) = \sum_{i=0}^{M} [a_i s(n-i)] \qquad \text{(IV.13)}$$

Con $a_0 = 1$, el error cuadrático total o energía residual es dado por

$$\alpha = \sum_{n=-\infty}^{\infty} [\, e(n) \,]^2 \qquad \text{(IV.14)}$$

En el método de autocorrelación, la secuencia de datos $\{s(n)\}$ es truncada asi que $s(n)=0$ para $n<0$ y $n>N-1$. Los coeficientes $\{a_i\}$ son escogidos para minimizar α. El error α puede ser considerado para ser la salida de un filtro inverso *A(z)* donde [12]:

$$A(z) = 1 + \sum_{i=1}^{M} a_i z^{-i}$$

es el filtro que minimiza α $1/A(z)$ corresponde a una respresentación espectral suavizada de la secuencia de datos $\{s(n)\}$ Si $\{s(n)\}$ es pasada a través de un filtro inverso diferente *A'(z)* de la forma:

$$A'(z) = \sum_{i=0}^{M} a_i z^{-i}$$

el cual minimiza la energía α' para alguna otra secuencia de datos $\{s'(n)\}$, con $a'_0 = 1$, entonces el error cuadrático total o energía residual δ, debe ser mayor que el error residual

mínimo, esto es:

$$\delta = \sum_{n=-\infty}^{\infty} [\sum_{i=0}^{M} a_i s(n-i)]^2 \geq \alpha$$

manteniendose igualmente si y solo si $A(z) = A'(z)$

La posibilidad de comparar los filtros $A(z)$ y $A'(z)$ en términos de las energías residuales son ilustrados en la fig. (IV.4)

- Si $\{s(n)\}$, se define como una muestra prueba y es pasada a través de un filtro de referencia $A'(z)$, una energía residual δ es obtenida a la salida del filtro.
- La mínima energía residual α es obtenida usando la misma muestra prueba para el filtro $A(z)$.

La relación δ/α define una diferencia entre los datos de prueba y referencia o sus espectros

- A la inversa si la secuencia $\{s'(n)\}$ es definida como una muestra prueba y es pasada a través de un filtro de referencia $A(z)$, una energía residual δ' es obtenida.
- α' representa la energía residual mínima obtenida de la minimización de $A'(z)$.

Entonces la relación δ'/α' también define una diferencia entre el espectro.

En ambos casos las relaciones δ/α y δ'/α' son siempre mayores o iguales a uno y pueden igualarse a uno si y solo si los dos filtros $A(z)$ y $A'(z)$ son identicos. La única diferencia en los resultados depende de cual secuencia de datos o modelo espectral es llamado la referencia y cual es llamado la prueba.

La razones δ/α y δ'/α' son llamadas **"razones de verosimilitud",** bajo ciertas suposiciones sobre los datos y el análisis, en el que los datos asumidos son Gaussianos y el análisis de ventana es mucho mayor que la longitud del filtro inverso, estas relaciones han sido presentadas como las razones de verosimilitud. Los logaritmos de estas relaciones son llamados "**razones de verosimilitud logarítmicas**". Evaluaciones de estas relaciones pueden ser obtenidas a través del uso de las secuencias de autocorrelación. $\{r_a(n)\}$ y $\{r_s(n)\}$ denotan la secuencia de autocorrelación para los coeficientes del polinómio $A(z)$ y los datos $\{s(n)\}$, respectivamente. De una manera similar $\{r'_a(n)\}$ y $\{r'_s(n)\}$ son definidas como la secuencia de autocorrelación para los coeficientes de $A'(z)$ y la secuencia de datos $\{s'(n)\}$, respectivamente. El mínimo error residual α, puede ser calculado como

$$\alpha = \sum_{n=-M}^{M} r_a(n)\, r_s(n) \qquad \text{(IV.15)}$$

La energía residual δ es calculada de la ec (IV.16), los límites finitos en la sumatoria se deben a que $r_a(n)$ es cero para $|n| > M$

$$\delta = \sum_{n=-M}^{M} r'_a(n)\, r_s(n) \qquad \text{(IV.16)}$$

Las razones de verosimilitud δ/α y δ'/α' pueden ser eficientemente calculadas usando las ecs (IV.15) y (IV.16)

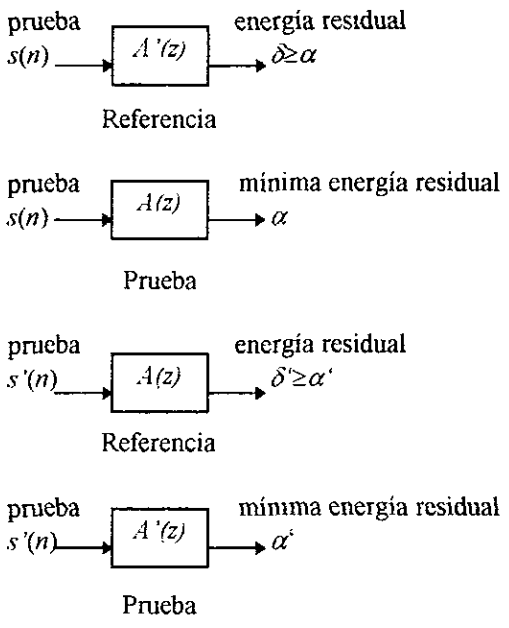


fig IV 4 POSIBLES COMBINACIONES PARA DATOS DE PRUEBA Y REFERENCIA.

d) Medida cosh.

En el estudio de máxima verosimilitud de predicción lineal, Itakura y Saito hicieron uso de la integral:

$$\Xi = \int_{-\pi}^{\pi} [e^{V(\theta)} - V(\theta) - 1]\, d\theta / 2\pi \qquad \text{(IV.17)}$$

de donde

$$\Xi' = \int_{-\pi}^{\pi} [e^{-V(\theta)} + V(\theta) - 1]\, d\theta / 2\pi \qquad \text{(IV.18)}$$

El promedio de $e^{V(\theta)}$ es encontrado para ser proporcional a la razón de verosimilitud, entonces Ξ llega a ser [12]:

$$\Xi = (\sigma / \sigma')^2 (\delta / \alpha) - 2\ln(\sigma / \sigma') - 1 \qquad \text{(IV.19)}$$

Una aproximación para obtener **una medida de distancia simétrica** es combinar los dos términos no-simetricos Ξ y Ξ' por medio de un simple promedio. Para incluir los posibles efectos de ganancia, se promedian las integrales de las ecs. (IV.17) y (IV.18) dando el resultado:

$$\Omega = \frac{1}{2}(\Xi + \Xi') = \int_{-\pi}^{\pi} \{\cosh[V(\theta)] - 1\} \frac{d\theta}{2\pi} \qquad \text{(IV.20)}$$

Sustituyendo la ec. (IV 19), y su contraparte Ξ', podemos relacionar Ω a las medidas de verosimilitud no-simétricas como:

$$\Omega = 1/2(\sigma/\sigma')^2 (\delta/\alpha) + 1/2(\sigma'/\sigma)^2(\delta'/\alpha') - 1 \qquad \text{(IV.21)}$$

Para $\sigma=\sigma'=1$, Ω se reduce a la media aritmética de las dos razones de verosimilitud no-simétricas menos uno, esto es:

$$\Omega = \frac{\delta/\alpha + \delta'/\alpha'}{2} - 1 \qquad \text{(IV.22)}$$

La integral de la ec. (IV.20) se designa como **una medida *cosh(ω)***, es simétrica en $V(\theta)$, es simpre más grande que $[V(\theta)]^2/2$ y para valores pequeños llega a ser aproximadamente igual a $[V(\theta)]^2/2$. Por lo tanto **la medida cosh(ω)** es siempre más grande que la desviación rms d_2 y se aproxima a d_2 desde arriba cuando $V(\theta)$ es pequeña en magnitud. Usando identidades trigonometricas, se puede notar que Ω es también equivalente a dos veces la media cuadrada del sinh $[V(\theta)/2]$.

La medida cosh(ω) es más apropiada para medir grandes diferencias en el espectro logarítmico más que **la medida rms**. Las grandes desviaciónes son importantes en procesos de voz y éstas usualmente corresponden a las regiones de frecuencias formantes cambiantes. Las dos medidas producen resultados idénticos para las desviaciones pequeñas.

e) Distancia de Itakura y Saito.

Una medida de distancia espectral propuesta entre dos marcos de voz representados por los coeficientes LPC a y a' y las matrices de autocorrelación aumentadas R_x y R_x' es

$$D(a,a') = a^T R_x' a / a'^T R_x' a' = E' / E \qquad \text{(IV.23)}$$

donde

- $D(a,a')$ es la distancia de Itakura y Saito.
- a son los coeficientes LPC de la señal de referencia.
- a' son los coeficientes LPC de la señal de prueba.
- R_x es la matriz de autocorrelación aumentada de la señal de referencia.
- R_x' es la matriz de autocorrelación aumentada de la señal de prueba.
- E es la energía residual mínima.
- E' es la energía residual de predicción.

La ec. (IV 23) también puede ser escrita como:

$$D(a,a') = r_a(0)r'(0) - 2\sum r_a(i)r'(i) / a^T{}'R_x{}'a' \qquad \text{(IV.24)}$$

El cálculo de la distancia $D(a,a')$ requiere de menos tiempo-máquina que comparar los espectros de la señal de entrada con los espectros de los patrones, para los que se necesitan algoritmos FFT, por lo tanto es más eficiente.

IV.4 2. Entrenamiento de un patrón.

Una variedad de técnicas han sido propuestas para la etapa de entrenamiento de un patrón en sistemas de reconocimiento de palabras aisladas, mencionaré algunas de estas técnicas a continuación.

a) Método de entrenamiento causal.

El locutor para el que será entrenado el sistema pronuncia cada palabra del vocabulario una o más veces y un patrón de referencia es creado por cada palabra hablada. Este método es muy simple pero el inconveniente es que no puede dar la confiabilidad y robustez en los patrones. Las palabras que no son reconocidas se quedan en espera de que sea obtenida una nueva plantilla más confiable. El número de patrones no confiables puede ser reducido creando dos o más plantillas para cada palabra del vocabulario. Cada replica de una palabra es separada en tiempo para minimizar la influencia externa durante el registro El precio que se paga por el incremento de confiabilidad de los patrones de una palabra es el aumento del cálculo computacional en la fase de reconocimiento del sistema.

b) Método de promedios.

El locutor dice cada palabra del vocabulario Q veces (frecuentemente en secuencia) y las Q replicas de cada palabra son promediadas para dar un único patrón de referencia. Esta técnica tiende a minimizar totalmente el riesgo de plantilla no confiable. Sin embargo este método corre el riesgo de promediar expresiones con alta discrepancia, y por lo tanto crea una plantilla de referencia que no provee una buena representación de la palabra dada. Los sistemas comerciales se han desempeñado extremadamente bien usando este método.

c) Método de agrupamiento estadístico.

El locutor dice cada palabra del vocabulario Q veces (estando Q entre 50 y 100 veces), las Q expresiones son agrupadas con base en patrones de similaridad y un patrón de referencia es obtenido de cada grupo conteniendo P o más expresiones. Generalmente, la expresión de referencia es obtenida promediando las características de los patrones alineados en el tiempo dentro de cada grupo.

Para un vocabulario de J palabras, el usuario dice cada palabra por primera vez y un patrón es medido y salvado para cada palabra. El usuario dice cada palabra por segunda vez y la distancia obtenida de un procedimiento de alineamiento en el tiempo global (DTW) entre cada nuevo patrón y los patrones previos (para aquella palabra) es calculada. Si la

mejor (mínima) distancia cae abajo de un umbral específico, una plantilla de referencia es creada para la palabra y el entrenamiento para aquella palabra se finaliza. Si la distancia es más grande, el patrón de referencia de la palabra es otra vez salvado para una tercera o subsecuente pasada. Este proceso continua hasta que todas las palabra del vocabulario son completadas, o hasta que un número máximo de repeticiones es alcanzado, esto es, del par de expresiones que den la distancia más pequeña, éstas serán usadas para obtener el patrón de referencia, promediando los dos conjuntos de características deformadas en el tiempo. Un punto fundamental en este procedimiento de entrenamiento es que para aumentar la confiabilidad de las plantillas se crean dos o más plantillas para cada palabra del vocabulario y las replicas de las palabras del vocabulario están separadas en tiempo ya que cada palabra del vocabulario es dicha una vez por pasada.

IV 4.3 Clasificación de patrones.

En esta fase se requiere de una medida de distancia local, que es definida como la "distancia espectral" entre dos vectores espectrales bien definidos y un procedimiento de alineamiento en el tiempo global (llamado algoritmo de deformación de tiempo dinámico), el que compensa para diferentes razones de velocidad de habla (tiempos de escala) de los dos patrones.

La distancia espectral o distancia local se puede obtener por medio de coeficientes cepstrales, medida espectral logaritmica o cualquier tipo de medida vista en el apartado IV 4 1 1

IV 4.3 1 Deformación de tiempo dinámico.

El algoritmo de deformación de tiempo dinámico es un proceso que compensa las diferencias en el tiempo de dos conjuntos de vectores espectrales para comparar dos patrones de voz que son dueños de los anteriores conjuntos de vectores.

Por lo tanto, el problema de deformación de tiempo dinámico puede ser formulado como el problema de encontrar un camino sobre una rejilla finita como se presenta en la fig. (IV.5). Denotamos el patrón de referencia como una secuencia de marcos, $R(n)$, $n = 1,2,...,N$, donde $R(n)$ es, en general, un vector característico multidimensional que describe las características del nth marco de la palabra hablada. En la fig. (IV.5), para mayor claridad se presentan $R(n)$ y $T(m)$ como funciones de n y m de una dimensión, (típicamente, marcos de datos consecutivos frecuentemente traslapados en tiempo). El patrón de prueba es una secuencia de marcos $T(m)$, $m=1,2,...,M$, donde $T(m)$ es también un vector característico multidimensional.

Basado en el modelo de la fig. (IV.5), el problema de DTW es encontrar un camino óptimo

$$m = w(n) \qquad \text{(IV.25)}$$

en el plano (n,m) para minimizar una función de distancia total D de la forma

$$D = \sum_{n=1}^{N} (d (R(n), T(w(n))) \qquad \text{(IV.26)}$$

donde:

- *d(R(n),T(w(n)))* es la distancia local entre el marco *n* del patrón de referencia y el marco *m=w(n)* del patrón de prueba.

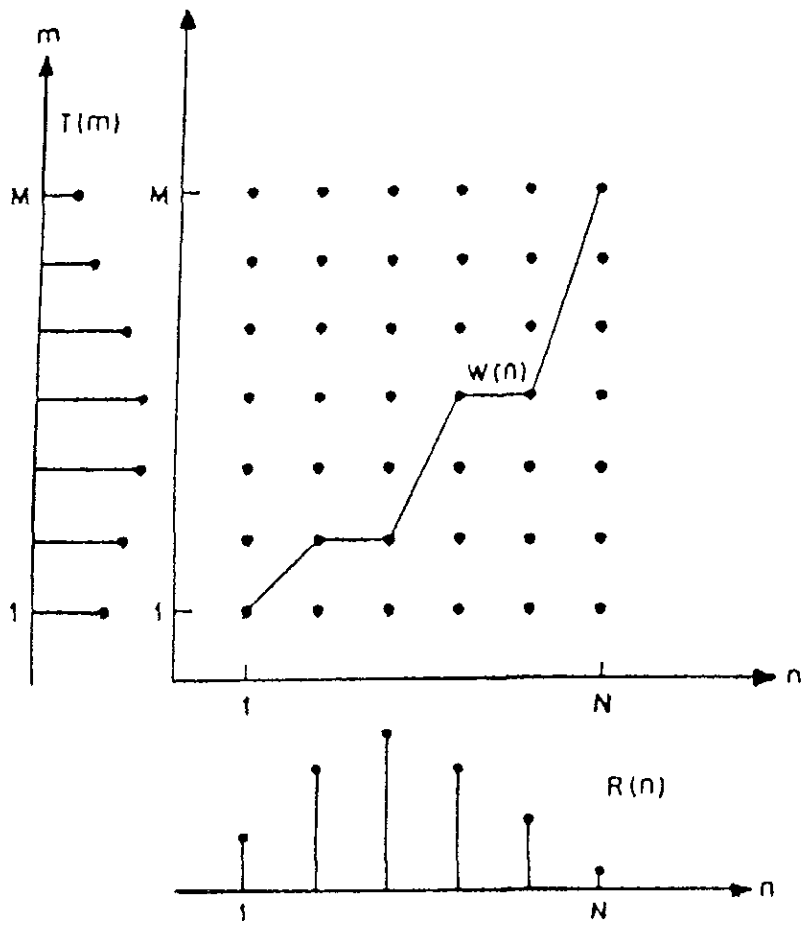


Fig. IV.5 EJEMPLO DE UNA MALLA PARA DEFORMAR *T(m)* A *R(n)* POR MEDIO DEL CAMINO *m=w(n)*, [13].

Un camino típico *w(n)* es presentado en la fig. (IV.5). Es importante notar que *w(n)* esta restringido para comenzar en el punto *n*=1, *m*=1, y pasar a través de la red de los punto *(n,m)* hasta finalizar en el punto *n=N*, *m=M*.

En la ec. (IV.25) se observa que el camino pueda ser expresado por una simple relación funcional entre *m* y *n* Es necesario crear un eje de tiempo común *k* y expresar ambos ejes de tiempo *n* y *m* como funciones de *k*, por ejemplo,

$$n = i(k), \qquad k = 1, 2, \ldots, K \qquad \text{(IV.27a)}$$

$$m = j(k), \qquad k = 1, 2, \ldots, K \qquad \text{(IV.27b)}$$

donde K es la longitud del eje de tiempo común.

La ec (IV.27) incluye la ec (IV.25), ya que si se escoge la función $i(k)$ para satisfacer la restricción, entonces

$$n = i(k) = k, \tag{IV.28a}$$

$$m = j(k) = j(n) = w(n), \tag{IV.28b}$$

Para encontrar el mejor camino en el plano (n,m), tomando como base la ec. (IV.27), varios factores del algoritmo DTW deben ser especificados, incluyendo:

a) Restricciones del punto final en el camino.
b) Restricciones de continuidad local, son los tipos posibles de movimientos (por ejemplo direcciones e inclinaciones) del camino.
c) Restricciones del camino global, son las limitaciones donde el camino puede existir en el plano (n,m).
d) Orientación de los ejes, esto es, los efectos de intercambiar los papeles de los patrones de prueba y referencia.
e) Medidas de distancia, son la medida de distancia local y la medida de distancia general usadas para determinar el camino optimo

a) Restricciones del punto final.

Determinados los puntos finales para los patrones de referencia y prueba, los puntos limites del camino paramétrico son de la forma:

$i(1)=1$, $j(1)=1$, punto inicial.

$i(K)=N$, $j(K)=M$, punto final.

b) Restricciones de Continuidad Local.

Algunas restricciones locales deben ser aplicadas para garantizar que excesivas compresiones o expansiones de las escalas de tiempo sean evitadas. Una primera restricción de este tipo es:

$$i(k+1) \geq i(k) \tag{IV.29a}$$

$$j(k+1) \geq j(k) \tag{IV.29b}$$

El conjunto de producciones (caminos que se pueden seguier para llegar a un punto determinado) para las restricciones locales del Tipo I de la fig. (IV.6) son:

$$P_1^{\,I} \rightarrow (1,0)(1,1) \tag{IV.30a}$$
$$P_2^{\,I} \rightarrow (1,1) \tag{IV.30b}$$
$$P_3^{\,I} \rightarrow (0,1)(1,1) \tag{IV.30c}$$

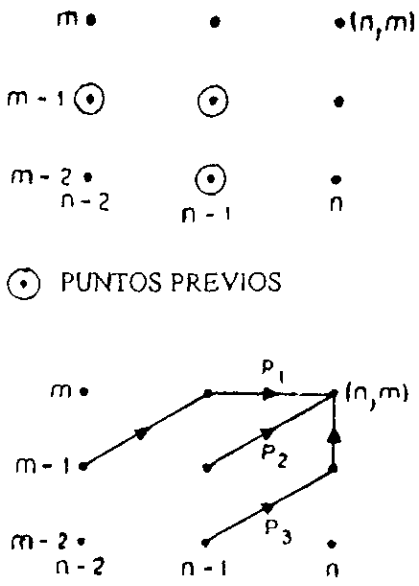


Fig. IV 6 RESTRICCIONES DEL CAMINO LOCAL TIPO I PARA ALCANZAR EL PUNTO *(n,m)* [3].

Usando las producciones de la ec. (IV.30), un camino completo del punto (N,M) al punto $(1,1)$ puede ser expresado como una secuencia de producciones. Por medio de un ejemplo, la fig (IV.7) presenta un camino señalado por las producciones de la ec. (IV.30).

$$P \rightarrow P_1^I P_1^I P_2^I P_1^I P_3^I P_2^I P_2^I.$$

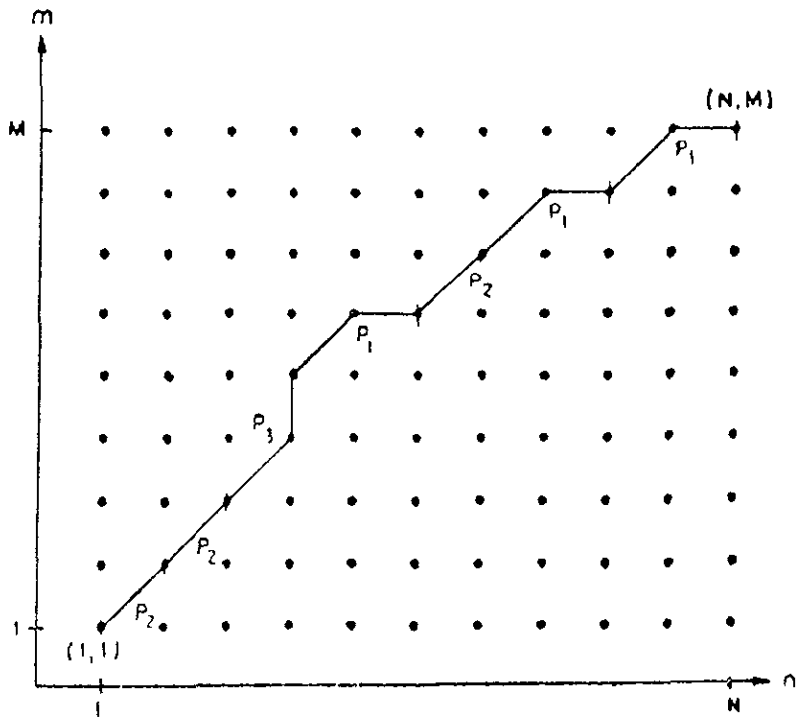


Fig IV.7 EL CAMINO Y CONJUNTO DE PRODUCCIONES PARA ILUSTRAR LA DEFORMACIÓN DE TIEMPO. [3]

El camino actual puede ser obtenido substituyendo las definiciones de las producciones para producir la secuencia:

$$P \rightarrow (1,0)(1,1)(1,0)(1,1)(1,1)(1,0)(1,1)(0,1)(1,1)(1,1)(1,1).$$

Observamos que esta secuencia ilustra, un trazo de adelante hacia atrás que es, el camino usado para alcanzar el punto (N,M) a partir del punto $(1,1)$.
Adicionalmente, simples expresiones pueden ser obtenidas para la máxima y mínima cantidad de expansión (o compresión) de las escalas de tiempo directamente de las producciones. Si denotamos la expansión máxima como E_{MAX}, y la expansión mínima como E_{MIN}, entonces

$$E_{MAX} = \max_{(r)} [\sum_{l=1}^{L(r)} \beta l^{(r)} / \sum_{l=1}^{L(r)} \alpha l^{(r)}] \qquad \text{(IV.31a)}$$

$$E_{MIN} = \min_{(r)} [\sum_{l=1}^{L(r)} \beta l^{(r)} / \sum_{l=1}^{L(r)} \alpha l^{(r)}] \qquad \text{(IV.31b)}$$

donde:

- β desplazamiento de la producción en el eje horizontal.
- α desplazamiento de la producción en el eje vertical.

Para las restricciones locales del Tipo I de la fig. (IV.6), $E_{MAX} = 2$ y $E_{MIN} = 1/2$.

La fig. (IV.8) ilustra cuatro tipos más de restricciones locales para reconocimiento de palabras aisladas. Una representación ilustrativa de los caminos locales, junto con el conjunto de producciones, y los valores de E_{MAX} y E_{MIN} para cada tipo, se muestran en dicha figura.

- Las restricciones locales del Tipo II tienen los mismos puntos iniciales y finales como las restricciones del Tipo I; sin embargo, estos caminos no van a través de puntos intermedios.
- Las restricciones locales del Tipo III son una forma generalizada de aquellas propuestas por Itakura [3]. Ambas restricciones del Tipo II y Tipo III tienen $E_{MAX} = 2$ y $E_{MIN} = 1/2$.
- Las restricciones locales del Tipo IV son una versión expandida de las restricciones del Tipo III en la cual la máxima expansión es incrementada a 3. Este conjunto fue incluido para ver si incrementos en E_{MAX} pueden ayudar o perjudicar el desempeño del algoritmo DTW.
- El último tipo de restricciónes, denotado como Tipo Itakura es el conjunto exacto propuesto por Itakura [3] La cruz en el camino denota la ponderación no lineal usada por Itakura para prevenir un camino de estado plano para dos marcos consecutivos.

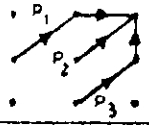
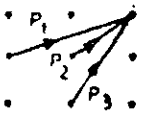
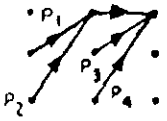
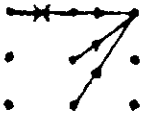
TIPO	FIGURA	PRODUCCIÓN	E_{MAX}	E_{MIN}
I		$P_1 \rightarrow (1,0)(1,1)$ $P_2 \rightarrow (1,1)$ $P_3 \rightarrow (0,1)(1,1)$	2	1/2
II		$P_1 \rightarrow (2,1)$ $P_2 \rightarrow (1,1)$ $P_3 \rightarrow (1,2)$	2	1/2
III		$P_1 \rightarrow (1,0)(1,1)$ $P_2 \rightarrow (1,0)(1,2)$ $P_3 \rightarrow (1,1)$ $P_4 \rightarrow (1,2)$	2	1/2
IV		$P_1 \rightarrow (1,0)(1,0)(1,1)$ $P_2 \rightarrow (1,0)(1,0)(1,2)$ $P_3 \rightarrow (1,0)(1,0)(1,3)$ $P_4 \rightarrow (1,0)(1,1)$ $P_5 \rightarrow (1,0)(1,2)$ $P_6 \rightarrow (1,0)(1,3)$ $P_7 \rightarrow (1,1)$ $P_8 \rightarrow (1,2)$ $P_9 \rightarrow (1,3)$	3	1/3
ITAKURA		NO HAY REGLAS DE PRODUCCIÓN	2	1/2

Fig. IV.8 TIPOS DE RESTRICCIONES LOCALES, [3].

c) Restricciones del camino global.

Ciertas partes del plano (n,m) son excluidas de la región (n,m) en las cuales el óptimo camino deformado no debe aparecer. Un simple grupo de relaciones puede ser obtenido expresando los límites de las regiones permisibles del plano (n,m). Estas relaciones son (asumiendo que $E_{MIN} = 1/E_{MAX}$)·

$$1 + (i(k)-1) / E_{MAX} \le j(k) \le 1 + E_{MAX}\,(i(k)-1) \qquad \text{(IV 32a)}$$

$$M + E_{MAX}\,(i(k)-N) \le j(k) \le M + (i(k)-N) / E_{MAX} \qquad \text{(IV.32b)}$$

La ec. (IV 32a) puede ser interpretada como limitadora del rango del camino partiendo del punto (1,1), mientras que la ec. (IV.32b) representa los límites hacia el punto (N,M).

Una restricción adicional sobre el rango global, propuesta por Sakoe y Chiba [4], es que

$$|i(k) - j(k)| \le R \qquad \text{(IV.33)}$$

donde:

- R es la diferencia de tiempo absoluta máxima permisible (en frames) entre patrones de prueba y referencia.

El efecto de esta restricción adicional es reducir el tamaño de la malla del paralelogramo cortando las esquinas, como se presenta en la fig. (IV.9).

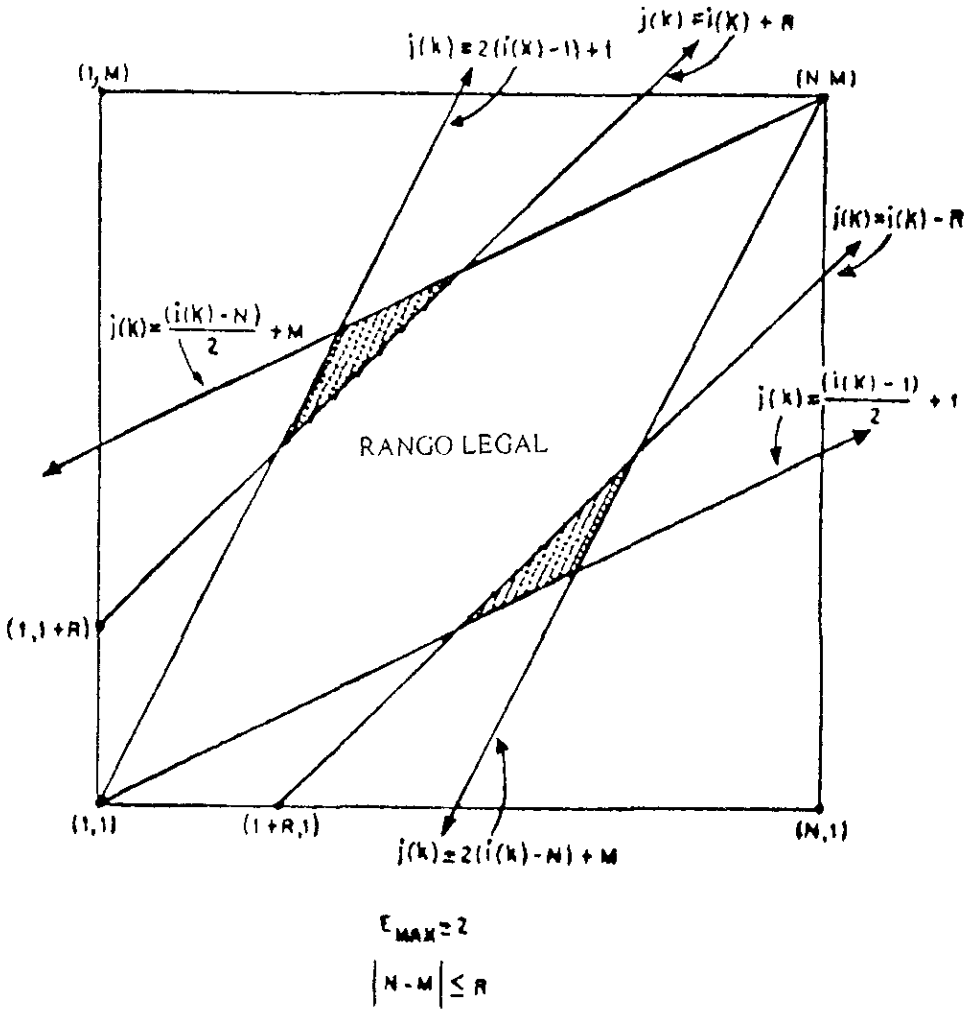


Fig IV.9 RANGO PERMITIDO PARA LOS CAMINOS DE TIEMPO DEFORMADO Y RESTRICCIONES DEL RANGO GLOBAL, [4].

- Un camino óptimo en la red (n,m) solo depende de los valores de n', m' tal que $n' \leq n$, $m' \leq m$.

Usando estos dos principios es posible crear una función de distancia acumulada parcial $D_A(n,m)$, representando la distancia acumulada a lo largo del mejor camino de (1,1) a (n,m) de la forma

$$D_A(n,m) = \min \left[\sum_{k=1}^{K'} d(i(k),j(k))\, W'(k) \right] \qquad \text{(IV.43)}$$

$D_A(n,m)$, depende solo de los caminos de (1,1) a (n,m) y puede ser definida recursivamente en terminos de un punto intermedio (n',m') donde $n'<n$, $m'<m$. Como

$$D_A(n,m) = \min \left[D_A(n',m') + d'((n',m'),(n,m)) \right] \qquad \text{(IV.44)}$$

donde d' es la distancia de ponderación de (n',m') hasta (n,m), esto es,

$$d'((n',m')\,(n,m)) = \sum_{l=0}^{L-1} d(i(K'-l),j(K'-l))\, W'(K'-l) \qquad \text{(IV.45)}$$

y en la cual L es el número de segmentos en el camino desde (n',m') hasta (n,m), y donde

$$i(K')=n \qquad j(K')=m \qquad \text{(IV.46a)}$$
$$i(K'-L)=n' \qquad j(K'-L)=m' \qquad \text{(IV.46b)}$$

Para una eficiente implementación de la ec. (IV.44), es necesario restringir el rango de (n',m') en la ec. (IV.44) para el grupo de puntos de la red los cuales usan una producción única para alcanzar (n,m) desde (n',m'). La fig. (IV 10) ilustra este punto importante con cuatro ejemplos [3]:

- El primer ejemplo usa restricciones locales Tipo I y ponderaciones (refinadas) Tipo a, los que se presentan en la fig. (IV.10a), en este caso la ec. (IV.44) llega ser

$$D_A(n,m) = \min \begin{bmatrix} D_A(n-1,m-1) + d(n,m) \\ D_A(n-1,m-2) + \tfrac{1}{2}\,(d(n,m-1) + d(n,m)) \\ D_A(n-2,m-1) + \tfrac{1}{2}\,(d(n-1,m) + d(n,m)) \end{bmatrix} \qquad \text{(IV.47)}$$

- Similarmente la fig. (IV 10b) presenta un ejemplo de restricciones Tipo II con ponderaciones Tipo d
- La fig. (IV 10c) presenta un ejemplo de restricciones Tipo III con ponderaciones del Tipo c.
- En la fig. (IV.10d) se presenta el algoritmo de DTW de Itakura. El propósito de la función g(k) es evitar caminos horizontales para más de un frame.

Siempre que $N(W')$ sea independiente del camino, o equivalentemente, cuando una solución al problema de minimización no normalizada dé una solución al problema de minimización normalizada, es posible minimizar la ec. (IV.35) como sigue:

$$D' = \min \left[\sum_{k=1}^{K} d(i(k),j(k))\, W'(k) \right] / N(W') \qquad \text{(IV.48)}$$

En tales casos se puede usar la ec (IV.44) para dar

$$D' = D_A(N,M) / N(W') \qquad \text{(IV.49)}$$

y la implementacion del algoritmo DTW es un procedimiento de tres pasos:

1. *Inicialización*: Poner $D_A(1,1) = d(1,1)W'(1)$.
2. *Recursión*: Calcular $D_A(n,m)$ recursivamente para $1 \le n \le N$, y $1 \le m \le M$.
3. *Terminación*: Poner $D' = D_A(N,M) / N(W')$

Hasta aquí se definieron todas las variables de interés en la implementación de un algoritmo de deformación de tiempo dinámico para reconocimiento de palabras aisladas.

IV.4.4 Decisión lógica.

Existen básicamente tres reglas de decisión que se aplican en Reconocimiento de palabras aisladas: la distancia mínima (DM), el vecino más proximo (Nearest Neighbor,NN), los K vecinos más proximos (K-NN).

a) La distancia mínima (DM).

Esta regla es la más simple y utilizada en reconocimiento de palabras aisladas. Cada clase del diccionario está compuesta por un solo patrón de referencia y la distancia mínima entre el patrón de prueba y el patrón de referencia de cada clase designa a la palabra reconocida.

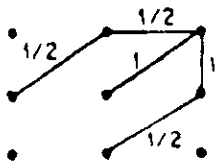
b) Vecino más próximo (NN)

Esta regla asume que el diccionario puede contener varios patrones de referencia por clase, y se suele utilizar cuando alguna palabra del diccionario admite dos o más pronunciaciones sensiblemente diferentes y también en sistemas de multilocutor. Si cada clase consta de un solo patrón de referencia, esta regla se reduce a la DM.

c) Los K vecinos más próximos (K-NN).

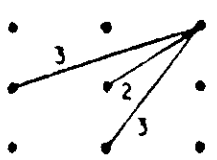
El diccionario tiene la misma estructura que en la NN pero se asume que cada clase contiene al menos k patrones de referencia. Esta regla se considera la más apropiada en aplicaciones multilocutor. La distancia entre el patrón de prueba y clases equivale a la diferencia media del patrón de prueba de los k patrones de referencia de cada clase cercanos al patrón de prueba. Esta regla se reduce a la NN en el caso de que k=1.

(a) RESTRICCIÓN DEL TIPO I FUNCIÓN DE PONDERACIÓN TIPO A REFINADA



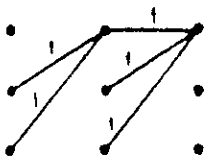
$$D_A(n,m) = \text{MIN}\Big(D_A(n-1,m-1) + d(n,m),\ D_A(n-1,m-2) + 1/2\,[d(n,m-1) + d(n,m)],\ D_A(n-2,m-1) + 1/2\,[d(n-1,m) + d(n,m)]\Big)$$

(b) RESTRICCIÓN DEL TIPO II FUNCIÓN DE PONDERACIÓN TIPO D



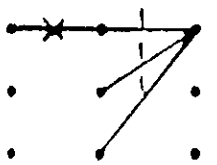
$$D_A(n,m) = \text{MIN}\Big(D_A(n-1,m-1) + 2\,d(n,m),\ D_A(n-1,m-2) + 3d(n,m),\ D_A(n-2,m-1) + 3d(n,m)\Big).$$

(c) RESTRICCIÓN DEL TIPO III FUNCIÓN DE PONDERACIÓN TIPO C



$$D_A(n,m) = \text{MIN}\Big(D_A(n-1,m-1) + d(n,m),\ D_A(n-1,m-2) + d(n,m),\ D_A(n-2,m-1) + d(n-1,m) + d(n,m),\ D_A(n-2,m-2) + d(n-1,m) + d(n,m)\Big)$$

(d) RESTRICCIÓN DE ITAKURA FUNCIÓN DE PONDERACIÓN TIPO C



$$D_A(n,m) = \text{MIN}\Big(D_A(n-1,m-2) + d(n,m),\ D_A(n-1,m-1) + d(n,m),\ D_A(n-1,m)\,g(k) + d(n,m)\Big)$$

$$g(k) = \begin{cases} 1 & j(k-1) \neq j(k-2) \\ \infty & j(k-1) = j(k-2) \end{cases}$$

Fig. IV.10 EJEMPLOS PARA CALCULAR LA DISTANCIA ACUMULADA, [3].

Certeza de la decisión.

Para los tres tipos de regla de decisión considerados, es posible la introducción de umbrales y/o reglas de "indecisión" que permitan dar una muestra como no reconocibles (es decir, que no existe suficiente certeza para asignarla a ninguna de las clases del diccionario). La más simple de utilizar es una distancia mínima de umbral, por encima de la cual no se asegura el resultado del reconocimiento

Una medida más consistente de la "certeza" o seguridad de la decisión, mediante la cual establecer una umbral de rechazo, es la relación s_r entre la distancia mínima (D^0_r) y la siguiente en magnitud (D^1_r)

$$s_r = D^0_r / D^1_r$$

como $D^0_r \leq D^1_r$ esta medida de certeza tiene un margen de variación $1 \leq s_r \leq \infty$, aunque para valores superiores a dos la decisión suele considerarse suficientemente segura en la mayoría de los casos. Si se desea una medida de certeza cuyo margen de variación este acotado en el intervalo [0,1] se puede utilizar la siguiente definición, prácticamente equivalente a la ecuación anterior.

$$s_r = (D^1_r - D^0_r) / (D^1_r + D^0_r)$$

IV 4.5 Algoritmo para reconocimiento de voz.

Reconocimiento de voz por medio de codificación por predicción lineal (LPC) es el procedimiento que a continuación se describe.

La señal se analiza por marcos traslapados y de cada marco se obtiene su potencia, preénfasis, coeficientes r'is, coeficientes a'is, coeficientes c'is y con estos últimos se determina la distancia entre el patrón de prueba y los patrones de referencia por medio de un procedimiento de alineamiento en el tiempo global. El diagrama de bloques de este sistema es mostrado en la fig. IV 11, a continuación se muestra un diagrama de bloques más detallado (fig. IV 12) y por último un diagrama de flujo (fig IV.13) para el mismo sistema

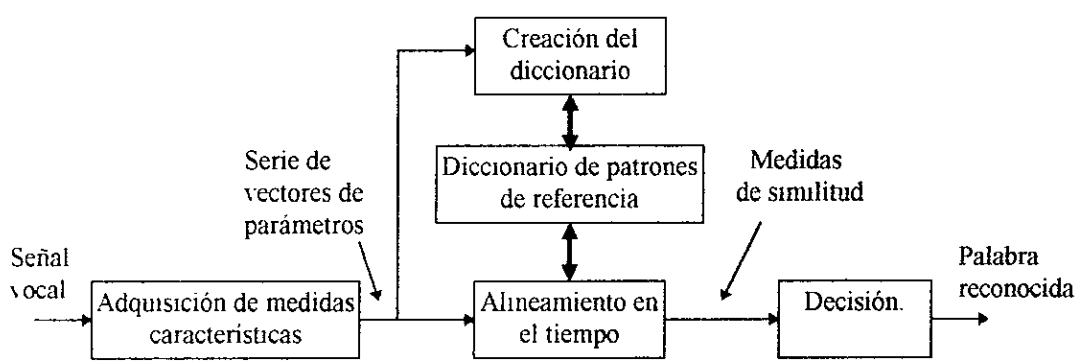


fig IV.11 DIAGRAMA DE BLOQUES PARA EL SISTEMA RECONOCEDOR DE VOZ CON EL MÉTODO LPC

a) Diagrama de Bloques para Reconocimiento de voz con el método LPC

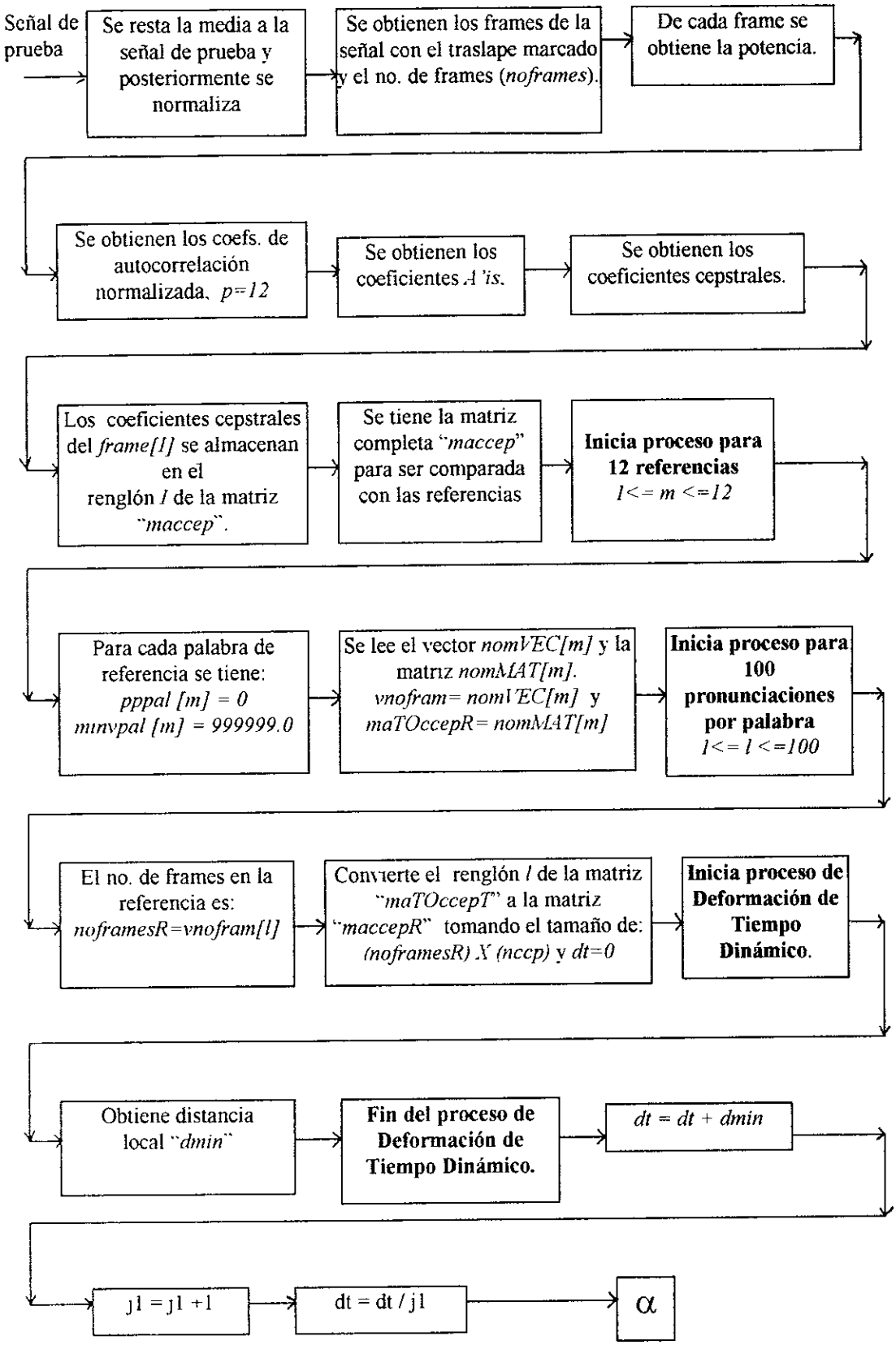
Dado un vocabulario de 12 palabras: "adelante", "alto", "atrás", "derecha", "dos", "izquierda", "lento", "no", "rápido", "si", "tres" y "uno" Se tienen 12 vectores y 12 matrices, como datos de referencia extraidos de una base de datos de 100 pronunciaciones para cada palabra, es decir 1200 pronunciaciones.
Cada vector esta compuesto de 100 elementos y cada elemento indica el número de segmentos por pronunciación de una palabra, siendo así como se compone un vector. Y la manera como son nombrados y manejados es de la siguiente forma:

vefraAD = nomVEC[1]
vefraAT = nomVEC[2]
vefraAL = nomVEC[3]
vefraDE = nomVEC[4]
vefraDO = nomVEC[5]
vefraIZ = nomVEC[6]
vefraLE = nomVEC[7]
vefraNO = nomVEC[8]
vefraRA = nomVEC[9]
vefraSI = nomVEC[10]
vefraTR = nomVEC[11]
vefraUN = nomVEC[12]

Cada matriz esta compuesta de 100 renglones (no. de pronunciaciones por palabra) por nccp*233 (nccp es el no. de coefcs cepstrales por frame y 233 es el max. no. de segmentos que puede tener una de las palabras del vocabulario para el usuario entrenado en curso). Por lo que en cada renglón se almacenan los coefcs. cepstrales de todos los frames de una pronunciación de una palabra, siendo así como se compone una matriz. Y la manera como son nombradas y manejadas es de la siguiente forma:
macceAD = nomMAT[1]
macceAT = nomMAT[2]
macceAL = nomMAT[3]
macceDE = nomMAT[4]
macceDO = nomMAT[5]
macceIZ = nomMAT[6]
macceLE = nomMAT[7]
macceNO = nomMAT[8]
macceRA = nomMAT[9]
macceSI = nomMAT[10]
macceTR = nomMAT[11]
macceUN = nomMAT[12]

Dadas las 12 matrices y 12 vectores que son las referencias para el proceso de reconocimiento, la señal de entrada o prueba sigue el siguiente proceso:

DIAGRAMA DE BLOQUES MAS DETALLADO PARA RECONOCIMIENTO DE VOZ POR EL MÉTODO LPC.



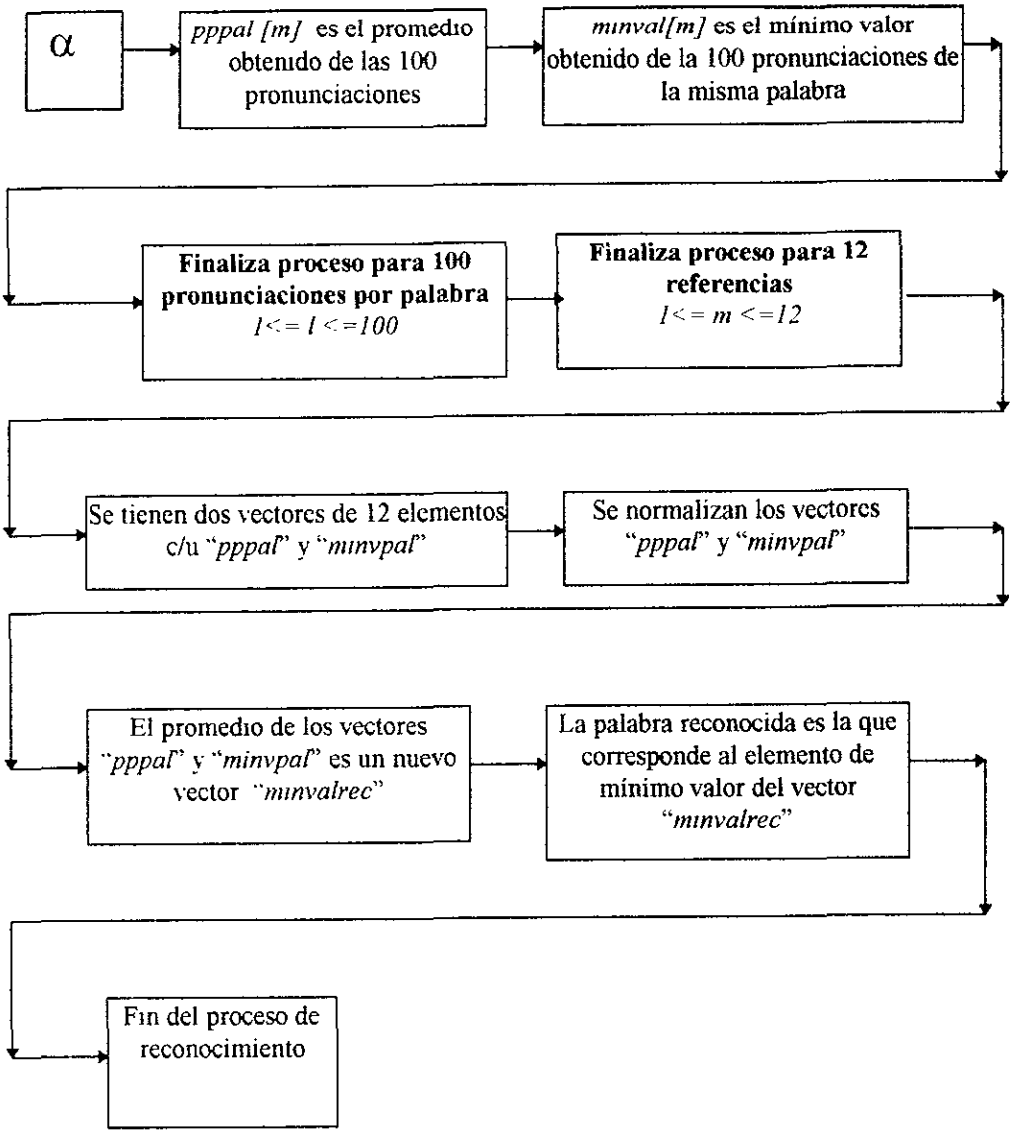
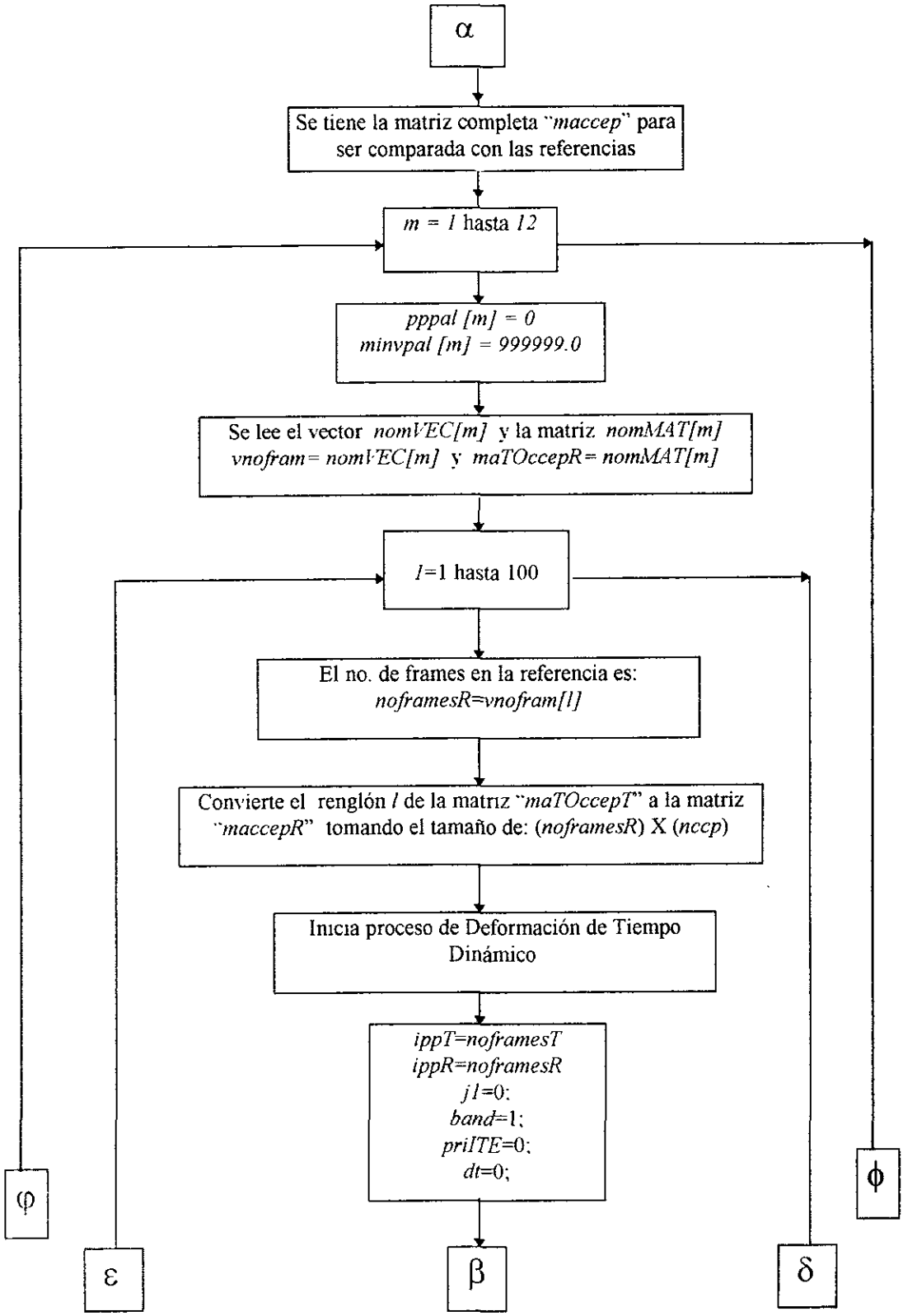
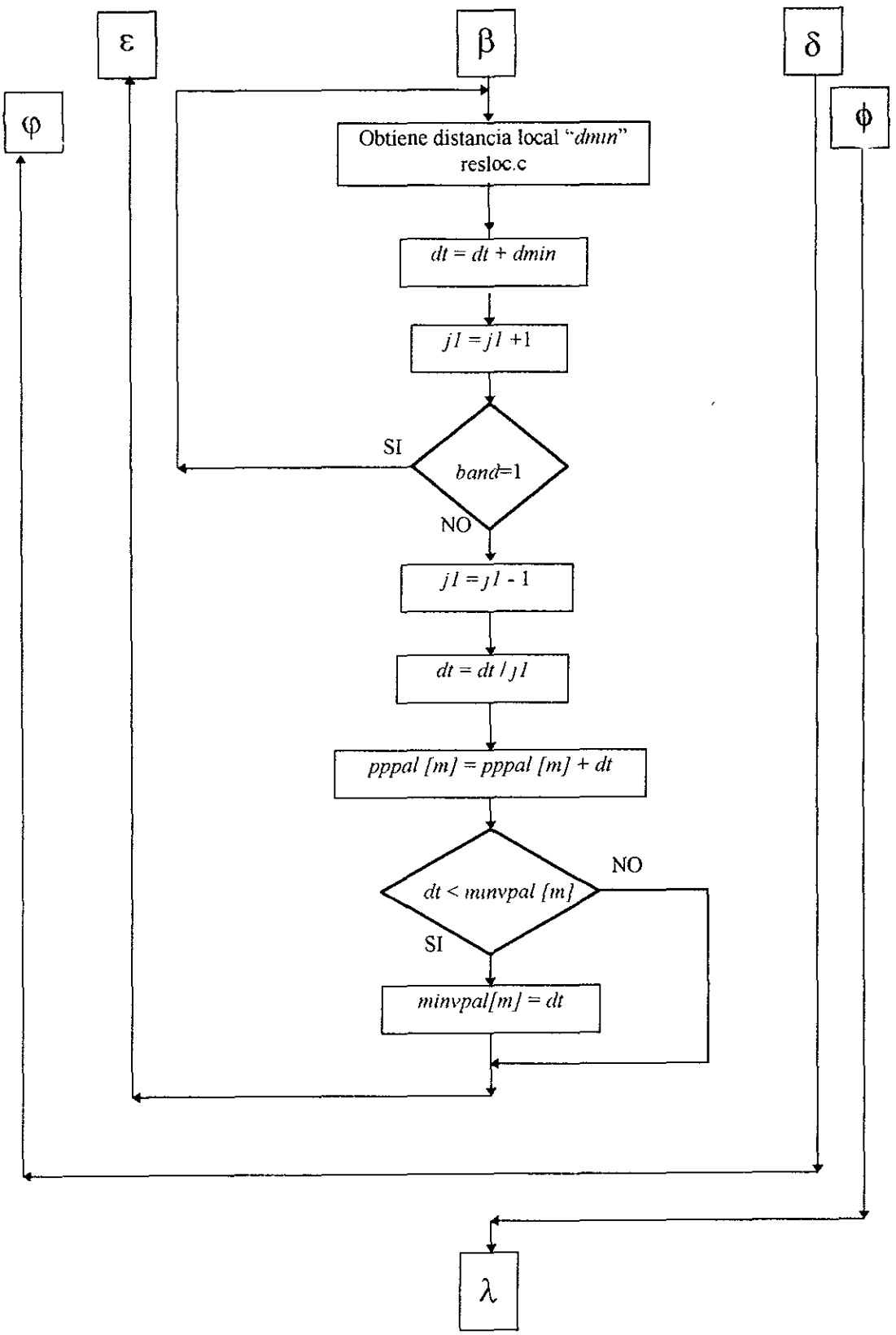


fig IV 12 DIAGRAMA DE BLOQUES MAS DETALLADO DEL RECONOCIMIENTO DE VOZ POR EL MÉTODO LPC


α
Se tiene la matriz completa "maccep" para ser comparada con las referencias
m = 1 hasta 12
pppal [m] = 0
minvpal [m] = 999999.0
Se lee el vector nomVEC[m] y la matriz nomMAT[m]
vnofram= nomVEC[m] y maTOccepR= nomMAT[m]
l=1 hasta 100
El no. de frames en la referencia es:
noframesR=vnofram[l]
Convierte el renglón l de la matriz "maTOccepT" a la matriz "maccepR" tomando el tamaño de: (noframesR) X (nccp)
Inicia proceso de Deformación de Tiempo Dinámico
ippT=noframesT
ippR=noframesR
j1=0;
band=1;
priITE=0;
dt=0;
φ
ε
β
δ
ϕ


ε
β
δ
φ
ϕ
Obtiene distancia local "dmin"
resloc.c
dt = dt + dmin
j1 = j1 +1
SI
band=1
NO
j1 = j1 - 1
dt = dt / j1
pppal [m] = pppal [m] + dt
NO
dt < minvpal [m]
SI
minvpal[m] = dt
λ

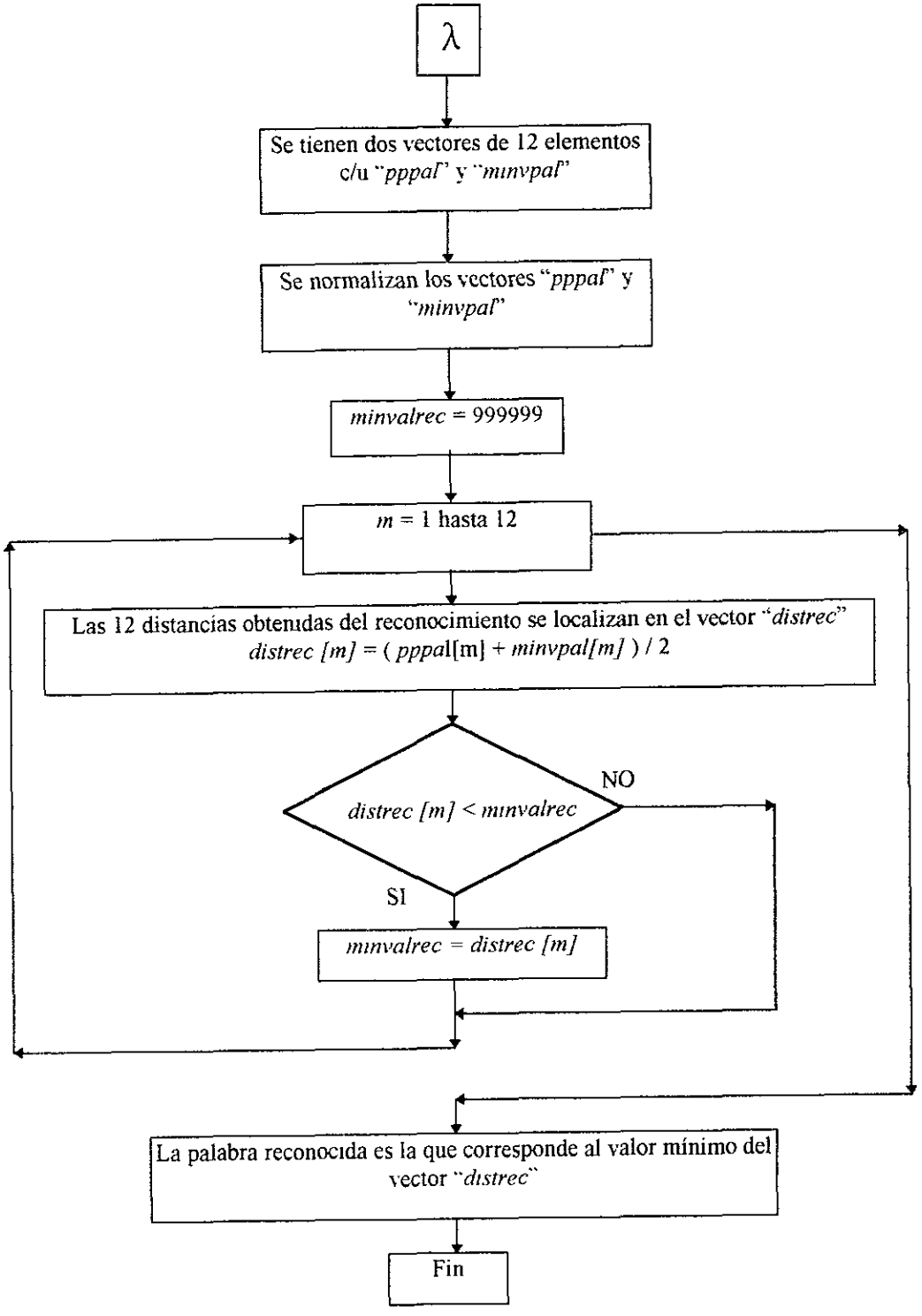
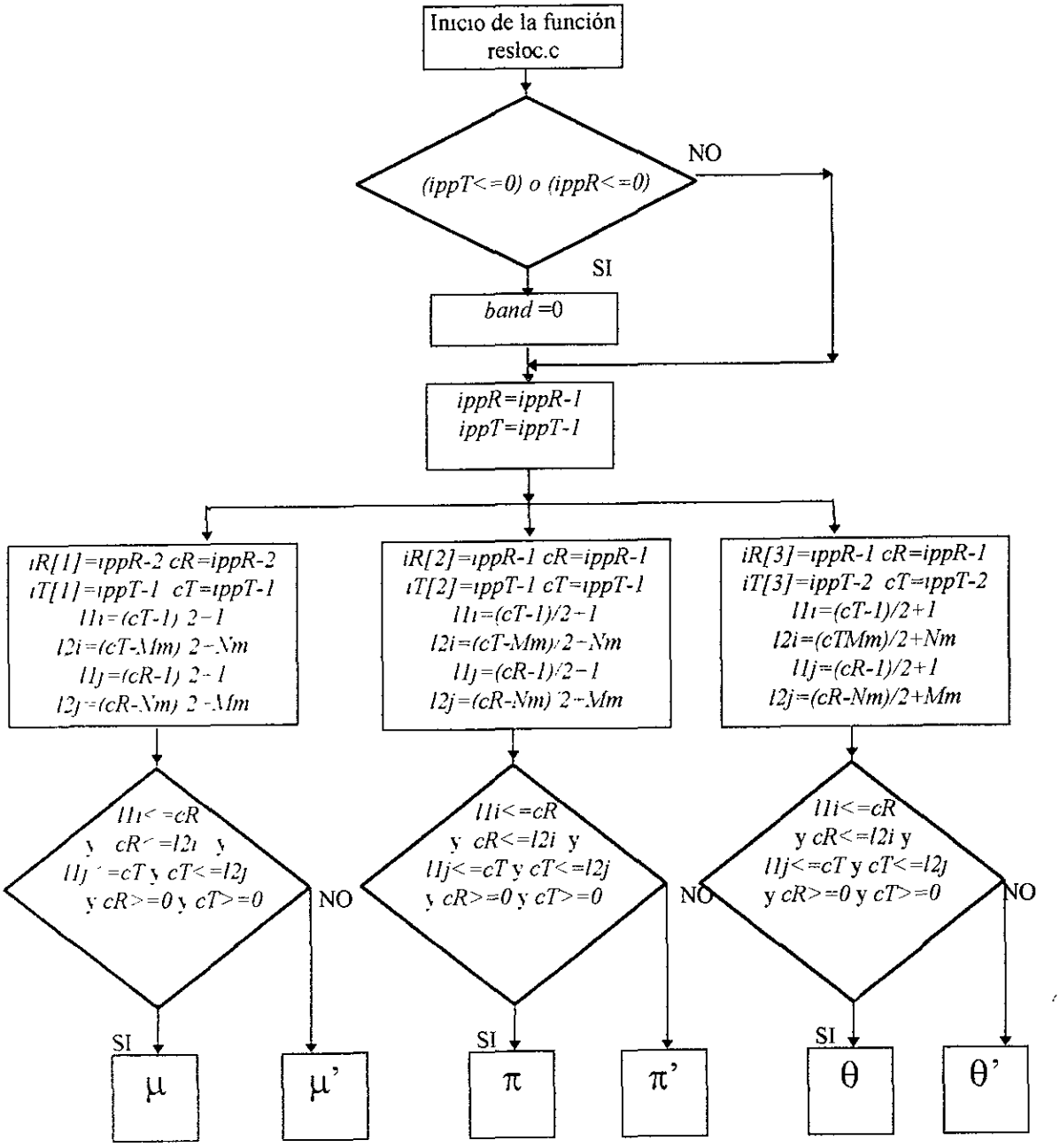


fig. IV.13 DIAGRAMA DE FLUJO PARA RECONOCIMIENTO DE VOZ POR EL MÉTODO LPC

El bloque compuesto para obtener la distancia local (dt) por medio de la función "resloc.c" esta compuesto como sigue:

- La restricción local del tipo II.
- Función de ponderación tipo "b" sin refinamiento.
- W(K)=max{i(k)-i(k-1),j(k)-j(k-1).
- El factor de normalización N(W) es función del tipo de ponderación, entonces N(W)=N(Wb).



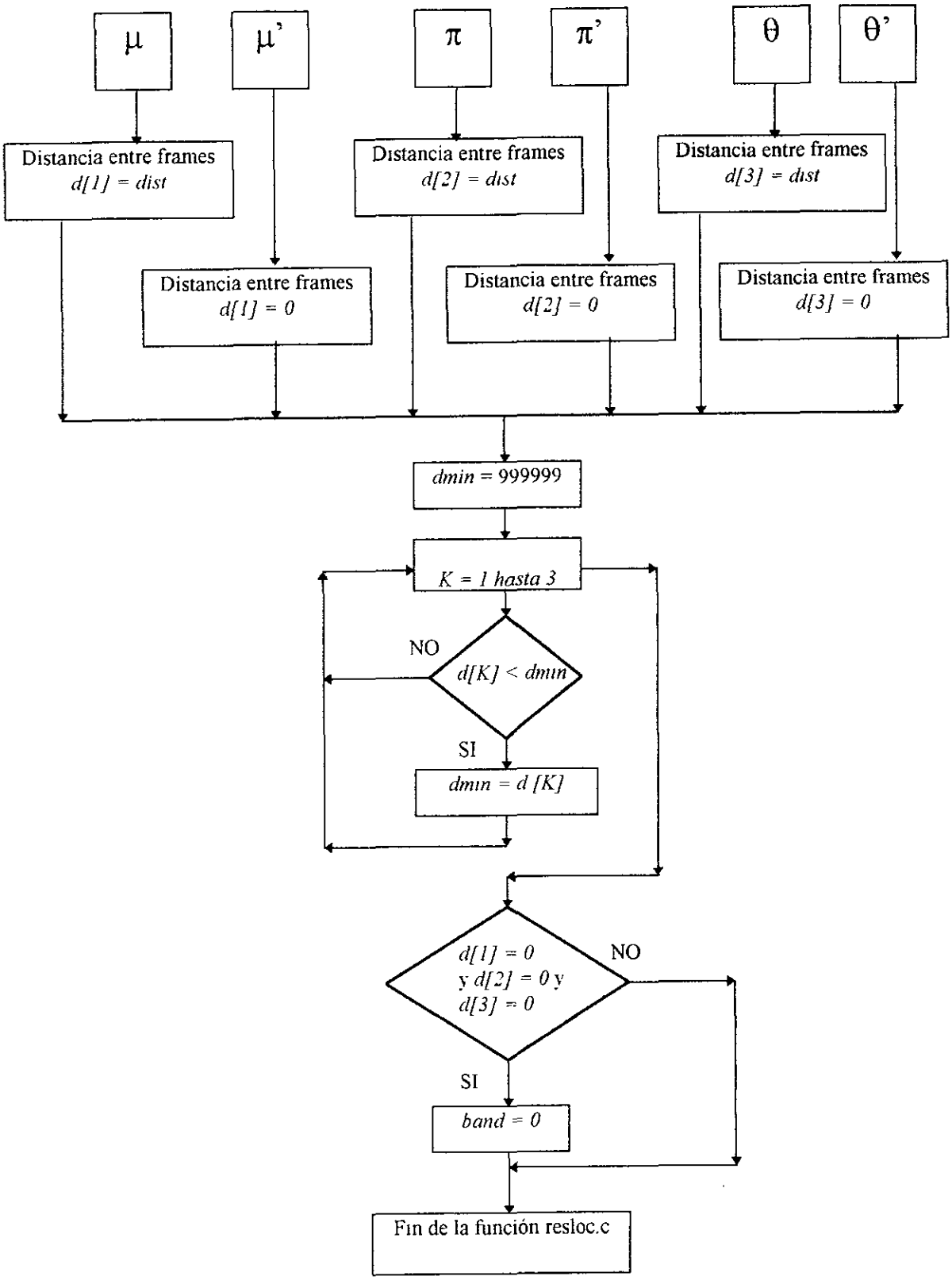


fig IV 13a DIAGRAMA DE FLUJO PARA RECONOCIMIENTO DE VOZ POR EL MÉTODO LPC. FUNCIÓN RESLOC.C

IV.5 Conclusiones.

El reconocimiento de voz por medio de la técnica LPC es el método tradicional o clásico que ha sido usado por largo tiempo.

Un sistema de Reconocimiento de Palabras Aisladas (RPA) basado en la aproximación global, descansa fundamentalmente en la diferencia entre patrones proporcionada por el algoritmo de alineamiento en el tiempo En un sistema RPA tras la parametrización y detección de principio y fin de palabra, cada palabra pronunciada queda representada como una secuencia de vectores, que puede utilizarse en dos formas: durante la *fase de entrenamiento* en donde se selecciona uno o varios patrones por clase (palabra) y su almacenamiento es para formar el "diccionario de patrones de referencia". Durante la *fase de reconocimiento*, el sistema compara la serie de vectores de parámetros de cada patrón de prueba con todos los patrones de referencia del diccionario evaluando la diferencia con cada uno de ellos mediante el algoritmo de alineamiento temporal y, posteriormente, se aplica alguna regla de decisión para obtener la palabra reconocida.

Durante el transcurso del tiempo se han desarrollado nuevas técnicas de reconocimiento de voz para palabras aisladas y en este trabajo se mencionarán dos más. La técnica de Cuantización Vectorial se contempla en el siguiente capítulo.

RECONOCIMIENTO DE VOZ POR MEDIO DE CUANTIZACIÓN VECTORIAL.

V.1 Introducción.

Una técnica de reconocimiento de voz posterior a la técnica LPC es la de Cuantización Vectorial y se tratará en este capítulo.

Si se compara la razón de información de la representación de vectores del análisis LPC (capítulo IV) con la señal de voz sin procesar, se observa que el análisis espectral reduce significativamente la razón de información. La representación ideal es tener un número finito único de vectores espectrales, cada uno correspondiente a las unidades de voz básica (i.e. los fonemas). Esto es impráctico debido a que existe mucha variedad de propiedades espectrales de cada una de las unidades de voz básica. Sin embargo, el concepto de construir un diccionario o codebook (son los parámetros que simbolizan a una plantilla o patrón que representan una palabra) de vectores de análisis distintos, aunque con más información que el conjunto básico de fonemas, esto descansa sobre la bases de un conjunto de técnicas comúnmente llamadas Cuantización Vectorial (CV). Cuantización Vectorial tiene una representación eficiente de la información espectral de la señal de voz. Esto es una de las principales razones para usar CV.

Las principales ventajas de CV son:

- Reduce el almacenamiento para información de análisis espectral.
- Reduce el cálculo para determinar la similaridad de vectores de análisis espectral. El cálculo para determinar la similaridad espectral es reducido a una tabla de referencia de similaridades entre un par de vectores.
- Representación discreta de sonidos de voz. Asociando una etiqueta fonética a cada vector del diccionario, el proceso de elegir el mejor vector del diccionario para representar un vector espectral dado llega a ser equivalente a asignar una etiqueta fonética a cada marco espectral de voz.

Las desventajas de usar un codebook en CV para representar vectores espectrales son:

- Una inherente distorsión espectral en representar el vector de análisis actual. Ya que existe solo un número finito de vectores en el diccionario, el proceso de elegir la "mejor" representación de un vector espectral dado inherentemente es equivalente a cuantizar el vector y llevarlo por definición a un cierto nivel de error de cuantización.
- El almacenamiento para vectores del diccionario no es trivial. Para tamaños de diccionario de 1000 o más vectores, el almacenamiento no es fácil, pero a medida que el codebook o diccionario se incrementa el error de cuantización se reduce. Por lo tanto existe un trueque entre error de cuantización, procesos para escoger el vector del diccionario y el almacenamiento de vectores del diccionario.

V.2 Implementación de un Cuantizador Vectorial.

1. Un conjunto grande de vectores que son el resultado de un análisis espectral v_1, v_2, ... v_L forma un conjunto entrenado, que es usado para crear el conjunto de vectores del diccionario "óptimo". Si el tamaño del diccionario del CV es de M vectores ($M=2^B$), entonces se requiere que $L>>M$ para que se pueda lograr encontrar el mejor conjunto de M vectores del diccionario de una manera robusta. En la práctica L debe ser al menos $10M$ para que el diccionario del CV trabaje bien.
2. Una distancia entre dos vectores de análisis espectral debe servir para agrupar el conjunto de vectores entrenados, así como para asociar o clasificar arbitrariamente los vectores espectrales en un diccionario único Denotando la distancia espectral $d(v_i,v_j)$ entre dos vectores v_i y v_j como d_{ij}
3 El procedimiento del cálculo del centroide: se clasifican los L vectores del conjunto entrenado en M grupos y se escogen los M vectores del diccionario como el centroide de cada uno de los M grupos.
4. Un procedimiento de clasificación para vectores de análisis espectral arbitrario, consiste en seleccionar el vector del diccionario más cercano al vector de entrada y usar el índice de ese vector como la representación espectral resultante. Esto se conoce usualmente como etiquetar al vecino más cercano, o procedimiento de codificación óptimo. El procedimiento de clasificación es esencialmente un cuantizador que acepta como entrada un vector espectral de voz y da como salida el índice del vector del diccionario que mejor compatibiliza con la entrada.

V.3 Agrupamiento del conjunto entrenado.

La manera en que un conjunto de L vectores entrenados puede ser agrupado en un conjunto de M vectores para el diccionario es como se menciona a continuación. El procedimiento es conocido como el algoritmo generalizado de Lloyd o el algoritmo de agrupamiento de K-medias [11].

1. Inicialización: Se escogen arbitrariamente M vectores (inicialmente fuera del conjunto de entrenamiento de los L vectores) como el conjunto inicial de palabras código en el diccionario.
2 Búsqueda del vecino más cercano Para cada vector de entrenamiento, encontrar el código de palabra en el actual diccionario que esta más cercano (en términos de distancia espectral) y asignar aquel vector a la celda correspondiente (asociado con el código de palabra más cercano).
3 Actualización del centroide: Actualiza el código de palabra en cada celda usando el centroide de los vectores de entrenamiento asignados a aquella celda.
4. Iteración: Repetir los pasos 2 y 3 hasta que la distancia promedio caiga abajo del nivel de umbral.

La partición de un espacio (2-dimensiones) vectorial espectral en distintas regiones, cada una de las cuales es representada por un vector centroide. La forma de cada celda

dividida es altamente dependiente de la medida de distorsión espectral y las estadísticas de los vectores en el conjunto de entrenamiento (i.e. si se usa una distancia Euclidiana, los límites de la celda son hiperplanos).

Aunque el procedimiento iterativo anterior trabaja bien, es ventajoso diseñar un diccionario de M-vectores en etapas (i.e. primero diseñando un diccionario de 1-vector, entonces usando una técnica de división en el código de palabras para iniciar la búsqueda de un diccionario de 2-vectores, y continuando el proceso de división, hasta obtener un diccionario de M-vectores deseados). Este procedimiento es llamado el algoritmo de división binaria y es formalmente implementado como sigue [11]:

1. Diseño de diccionario de 1-vector; este es el centroide de un conjunto completo de vectores de entrenamiento (por lo que ninguna iteración es requerida en este paso).
2. Doblar el tamaño del diccionario dividiendo cada diccionario actual $\mathbf{y}_n$ de acuerdo a la regla

$$y_n^+ = y_n(1+\varepsilon)$$
$$y_n^- = y_n(1-\varepsilon)$$

   donde y_n varia desde 1 hasta el tamaño actual del diccionario, y ε es un parámetro divisor que se encuentra típicamente en el rango de 0.01 a 0.05.
3. Se usa el algoritmo iterativo de K-medias para obtener el mejor conjunto de centroides para el diccionario dividido (i.e. el codebook o diccionario de dos veces el tamaño).
4. Iterar los pasos 2 y 3 hasta que un diccionario de tamaño M es alcanzado.

La fig. V 1 presenta un diagrama de flujo, para detallar los pasos de la técnica de generación de un diccionario de CV de divisón binaria.

- La caja etiquetada como "Clasificación de Vectores" es el procedimiento para la búsqueda del vecino más cercano.
- La caja etiquetada como "Búsqueda de centroides" es el procedimiento para actualizar el centroide por medio del algoritmo de K-medias.
- La caja etiquetada con "Cálculo D (distorsión)" suma las distancias de todos los vectores de entrenamiento en la búsqueda del vecino más cercano para determinar si el procedimiento esta convergiendo (i.e. D=D' de la interación anterior).

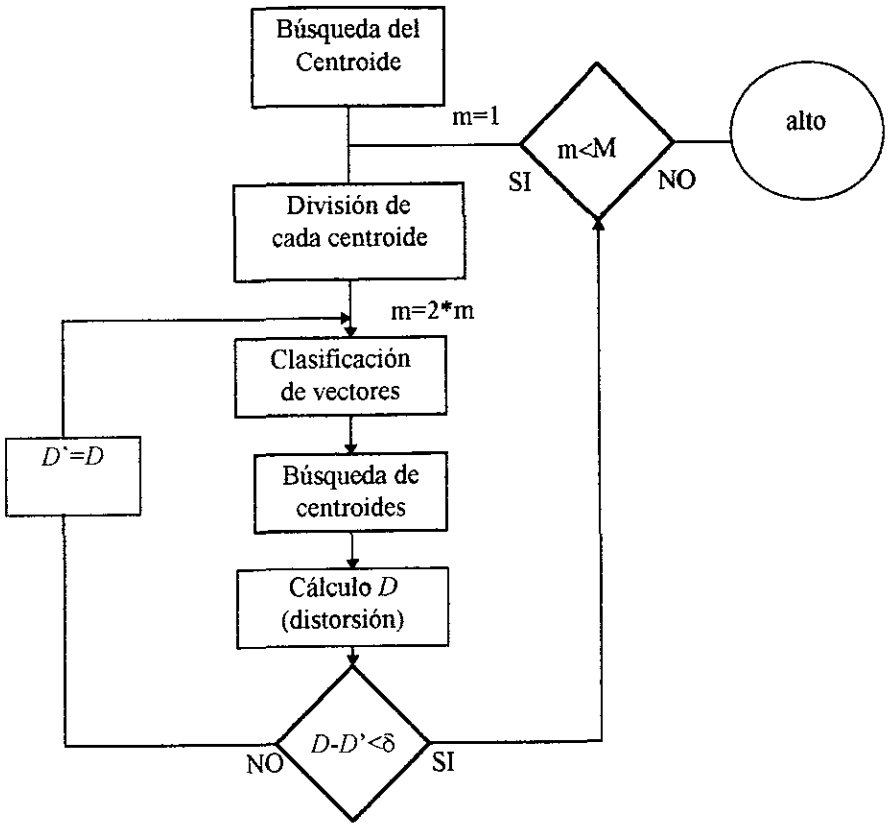


Fig. V 1 DIAGRAMA DE FLUJO PARA GENERAR UN DICCIONARIO DE DIVISIÓN BINARIA, [11].

V.4 Cálculo del Centroide.

La cuantización vectorial se aproxima al reconocimiento de voz y requiere que el diccionario para una clase de pronunciación particular sea apropiadamente diseñado para minimizar la distorsión promedio. El algoritmo de Lloyd para generación del diccionario, basado en el conjunto de entrenamiento, puede ser brevemente resumido como un procedimiento iterativo de los siguientes pasos:

- Etiquetando la distorsión mínima: Para cada vector de entrada x_t, encontrar $x_t'=y_j$, donde y_j es una entrada del diccionario C, satisfaciendo la ec. (V.1); el grupo de vectores entrenados de acuerdo a sus índices asociados al código de la palabra.

$$x_t' = \underset{y_i \in C}{arg\ min}\ d(x_t, y_i) \qquad (V.1)$$

- Cálculo del centroide: Para cada grupo de vectores con el mismo índice, cálcular el nuevo centroide el cual minimiza la distorsión promedio para los miembros del grupo (donde ellos son reproducidos por el centroide del vector).

Los dos pasos anteriores iteran hasta que algún criterio de convergencia es encontrado. El primer paso es directo ya que esto involucra solo evaluaciones sobre la distorsión. El segundo paso, cálculo del centroide, es un problema de optimización.

Considerando un conjunto de vectores $\{x_i\}^L_{i=1}$ y una medida de distorsión $d(x,y_i)$. Asumimos que estos vectores estan asignados al mismo grupo (o código de palabra) en el contexto de cuantizadores vectoriales. El centroide $\{x_i\}^L_{i=1}$ es definido como el vector $\bar{y}$ que minimiza la distorsión promedio; esto es

$$\bar{y} \cong \arg\min_{y} 1/L \sum_{i=1}^{L} d(x_i, y) \qquad \text{(V.2)}$$

La solución para el problema del centroide, es altamente dependiente de la elección de la medida de distorsión. Cuando x_i y y son vectores, $x_i=(x_{i1}, x_{i2}, \ldots, x_{iK})$ y $y=(y_1, y_2, \ldots, y_K)$ medidos en un espacio de K-dimensiones con la norma L_2 (distancia Euclidiana), el centroide obviamente es la media del conjunto de vectores,

$$\bar{y} = 1/L \sum_{i=1}^{L} x_i \qquad \text{(V.3)}$$

La solución del centroide de la ec. anterior también se aplica al caso de la distancia Euclidiana ponderada. Otro resultado bien conocido relacionado a la distancia L_1.

$$d(x_i, y) = \sum_{k=1}^{K} |x_{ik} - y_k| \qquad \text{(V.4)}$$

es que el centroide $\bar{y}$ es el vector medio de $\{\mathbf{x}_i\}^L_{i=1}$. (En este caso, cada dimensión es normalmente tratada independientemente, significando que cada y_k es el valor medio de $\{x_{ik}\}^L_{i=1}$ respectivamente)
Como se mencionó anteriormente x y y representan espectros de voz y $d(x,y)$ es la medida de distorsión espectral

V.5 Segmentación de la Cuantización Vectorial.

La aproximación estandar de cuantización vectorial que usa un único vector cuantizador para la duración completa de la pronunciación de entrada para cada palabra (clase), no preserva las características secuenciales de la clase pronunciada. Esta falta de caracterización explicita del comportamiento secuencial puede ser remediada tratando cada clase pronunciada como una concatenación de varios vectores, es decir, Ns subfuentes de

información, cada una de las cuales es representada por un diccionario de CV, llamando a esta aproximación de segmento específico de CV "Cuantización Vectorial Segmentada". Para una pronunciación $\{x_t\}^{T}_{t=1}$, la manera más simple (no necesariamente la más significativa) de descomponer esto en una concatenación de *Ns* subfuentes de información es dividir la pronunciación en *Ns* segmentos iguales: $\{x_t\}^{T/Ns}_{t=1}$, $\{x_t\}^{2T/Ns}_{t=T/Ns+1}$ ... y así sucesivamente. Este simple esquema de segmentación lineal es ilustrado en la fig.V.2. Existen otros esquemas de segmentación más sofisticados. Dado un conjunto de pronunciaciones entrenadas de una palabra (clase) conocida, *Ns* conjuntos de datos entrenados son formados y usados para diseñar los *Ns* diccionarios y tienen un orden temporal implicito porque corresponden a diferentes porciones de la pronunciación. Similar para el caso de un diccionario único, cada conjunto de *Ns* diccionarios sucesivos representan una clase y el promedio de distorsión para codificar una pronunciación desconocida con la sucesiva correspondencia de cuantizadores vectoriales [11].
El requerimiento de segmentar el CV tiene la misma complejidad computacional como el CV de pronunciación-base. La única complejidad es en el almacenamiento del diccionario; segmentar el CV es tener *Ns* diccionarios mientras que un CV de pronunciación base solo tiene uno para cada clase de pronunciación.

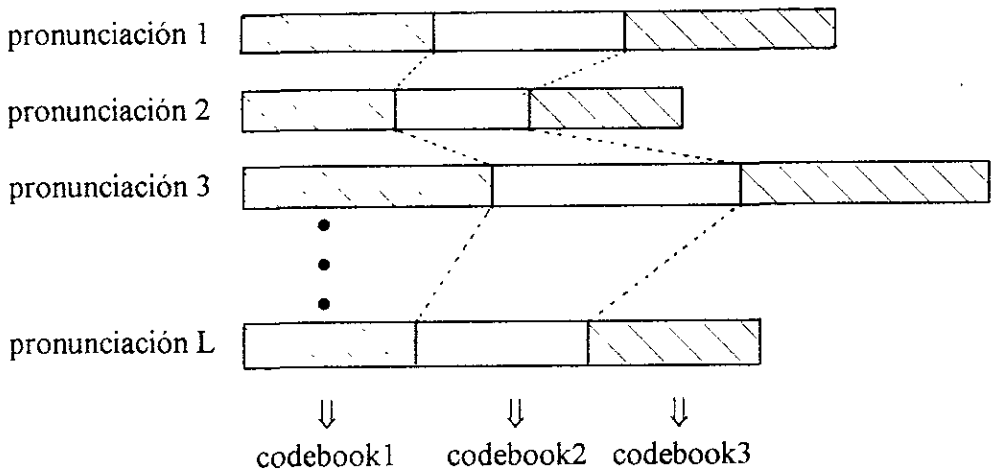


Figura V.2 ENTRENAMIENTO DEL DICCIONARIO PARA SEGMENTACIÓN EN CUANTIZACIÓN VECTORIAL, [11].

V.6 Agrupamiento.

En el reconocimiento de voz para locutores independientes, el entrenamiento de plantilla por agrupamiento es necesario para lograr una alta precisión en el reconocimiento de palabras para tareas prácticas. El agrupamiento de patrones es directo. Dando un conjunto de *L* pronunciaciones (patrones de voz), cada una de las cuales es una realización de una clase de pronunciación particular (una palabra), para ser reconocida. La tarea es agrupar los *L* patrones en *N* grupos tal que dentro de cada grupo, los patrones sean altamente

similares bajo la medida de diferencia del patrón específico escogido para el diseño de reconocimiento y por lo tanto pueda ser eficientemente representado por una plantilla típica. De este modo se crean N plantillas representativas de un conjunto de L patrones entrenados para cada clase de pronunciación. Las principales ventajas del agrupamiento de patrones son la consistencia estadística de las plantillas generadas y su habilidad para hacer frente a un amplio rango de variaciones de voz individual en un ambiente para hablantes independientes.

Los primeros algoritmos de agrupamiento para reconocimiento de palabras de voz aislada usan algunas técnicas de clasificación de patrones sofisiticada, basandose también en una intervención manual para el agrupamiento. La meta fue maximizar la razón de distancia promedio intergrupo a la de distancia promedio intragrupo. El número total de grupos por clase pronunciada fue una variable que fue interactivamente determinada examinando la razón de distancia mencionada. Estos procedimientos semiautomáticos, aunque proveen un desempeño aceptable, no fueron compatibles a un uso generalizado, por su falta de repetibilidad de los resultados y su desorden en el tiempo para completar la tarea de agrupamiento Para reducir este problema, una clase de procedimientos de agrupamiento automático no requiere de la intervención humana para ser desarrollado

V.7 Algoritmo para reconocimiento de voz.

Reconocimiento de voz por medio de Cuantización Vectorial (CV) es el procedimiento que a continuación se describe.

La señal se analiza por marcos traslapados y de cada marco se obtiene el preénfasis, coeficientes *r'is*, coeficientes *a'is*, y con estos últimos se determina la distancia entre el patrón de prueba y los cuantizadores vectoriales por medio de de la distancia de Itakura. El diagrama de bloques de este sistema es mostrado en la fig. V.3, a continuación se muestra un diagrama de bloques más detallado y por último un diagrama de flujo para el mismo sistema

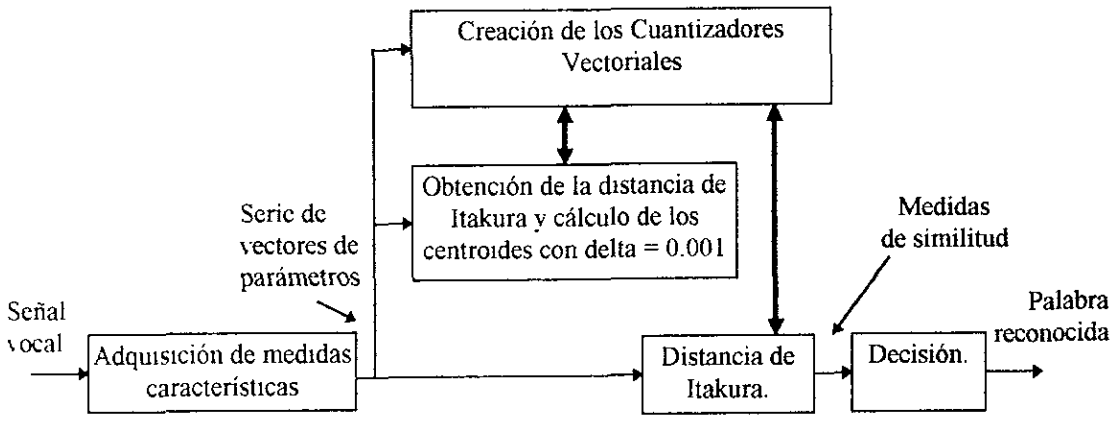


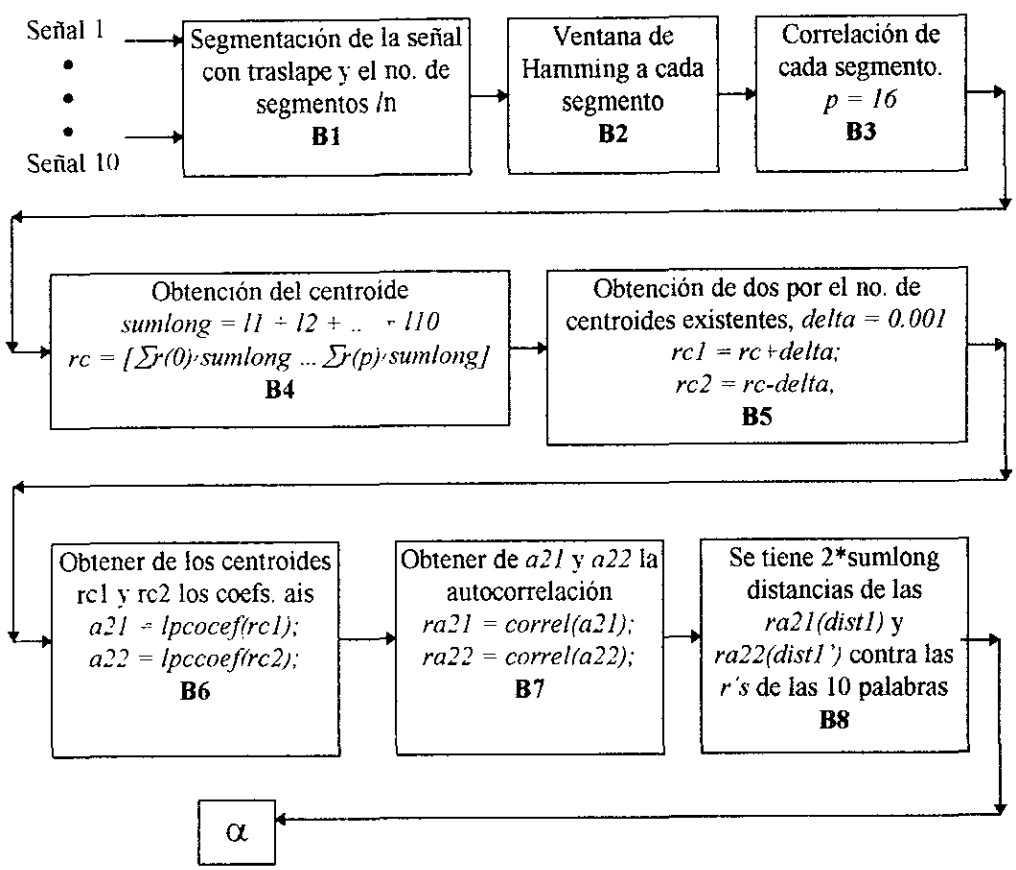
fig V 3 DIAGRAMA DE BLOQUES PARA EL SISTEMA RECONOCEDOR DE VOZ CON EL MÉTODO DE CUANTIZACIÓN VECTORIAL.

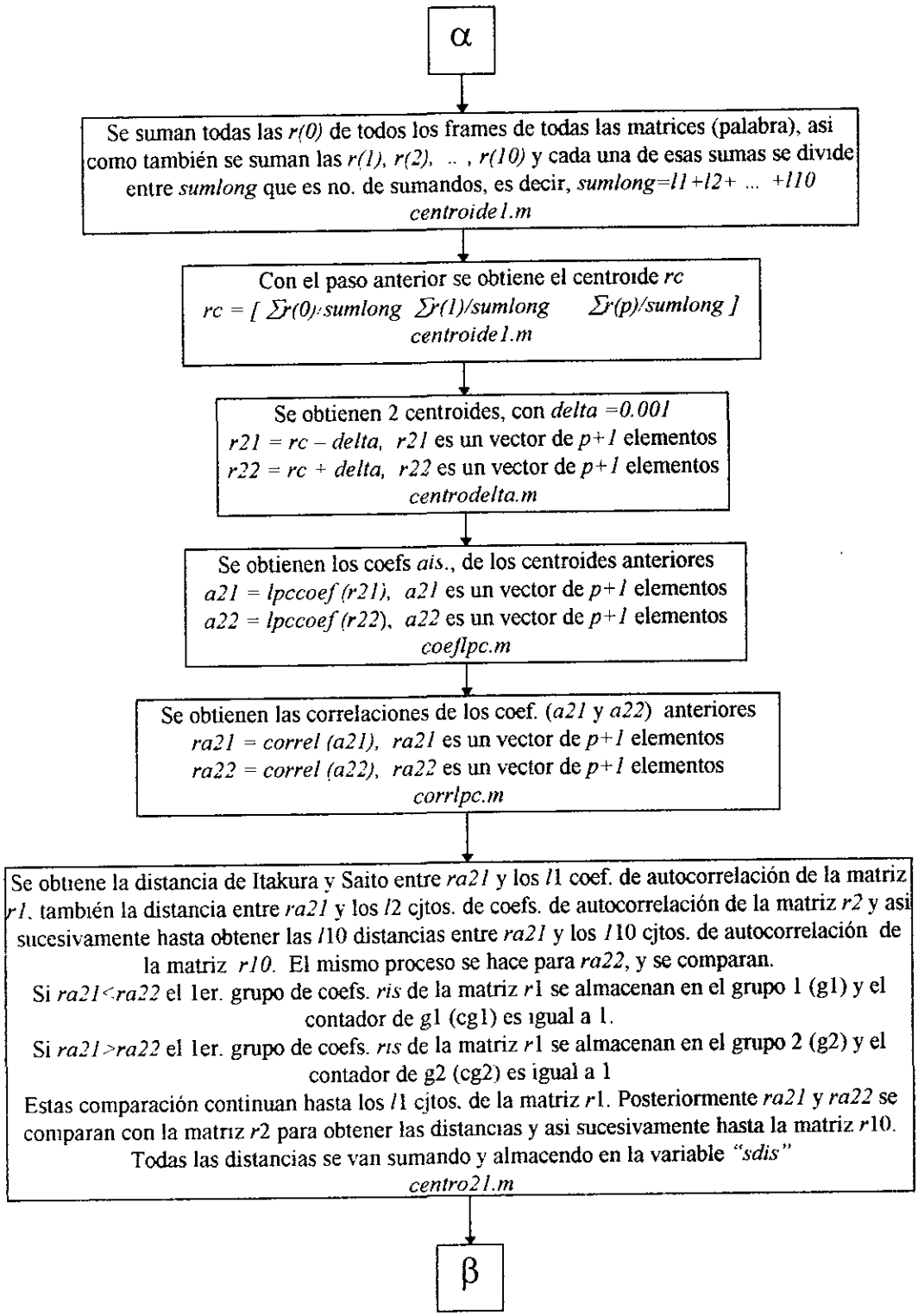
a) Diagrama de bloques para Reconocimiento de voz por el método de Cuantización Vectorial.

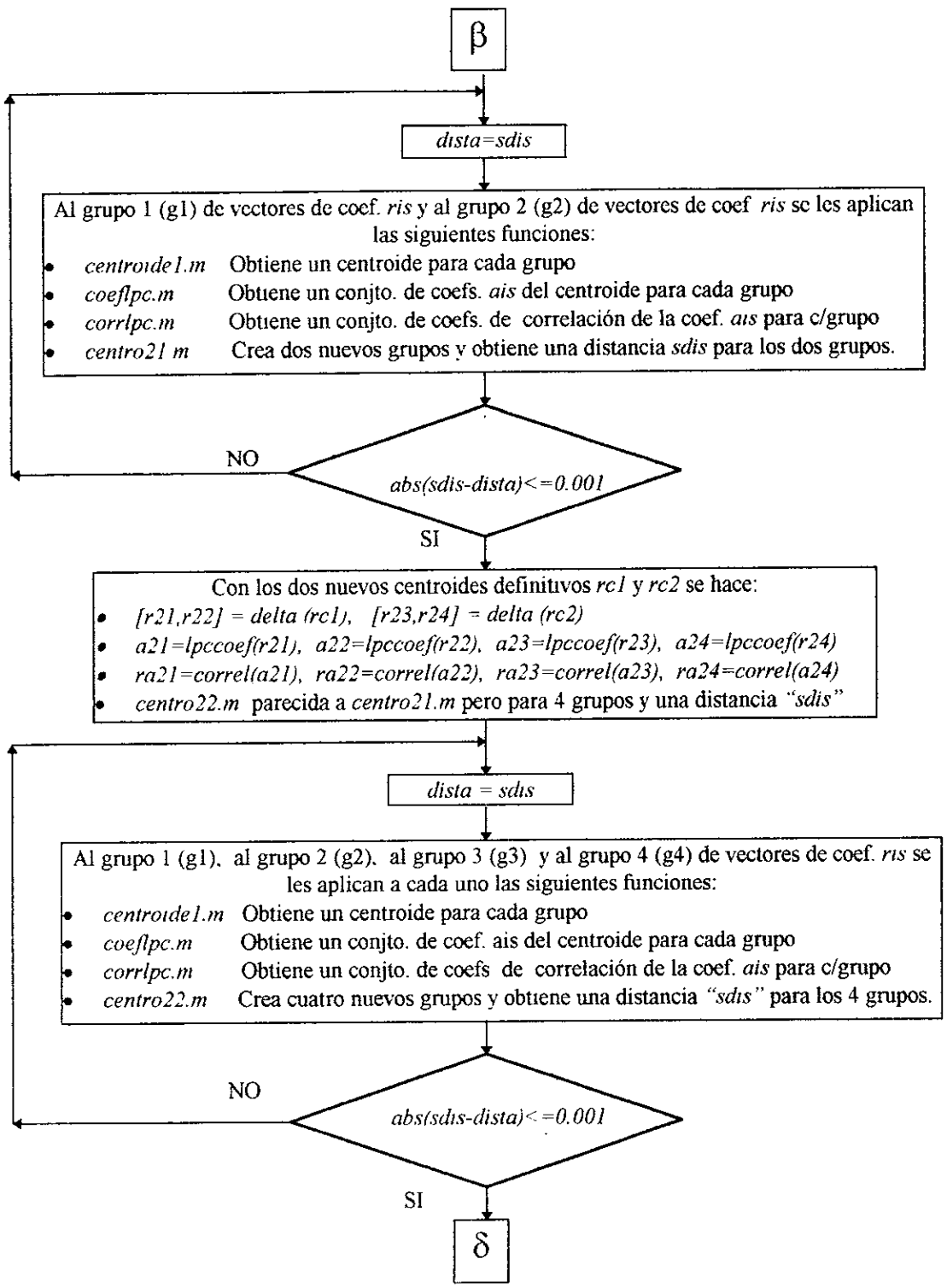
Dado un vocabulario de 12 palabras: "adelante", "alto", "atras", "derecha", "dos", "izquierda", "lento", "no", "rapido", "si", "tres" y "uno". Se tienen 10 pronunciaciones para cada palabra, es decir 120 pronunciaciones. De las diez pronunciaciones por palabra se crea un Cuantizador Vectorial para cada palabra, por lo que el siguiente diagrama de bloques debe ser repetido para cada palabra. En seguida también se muestra el diagrama de flujo.

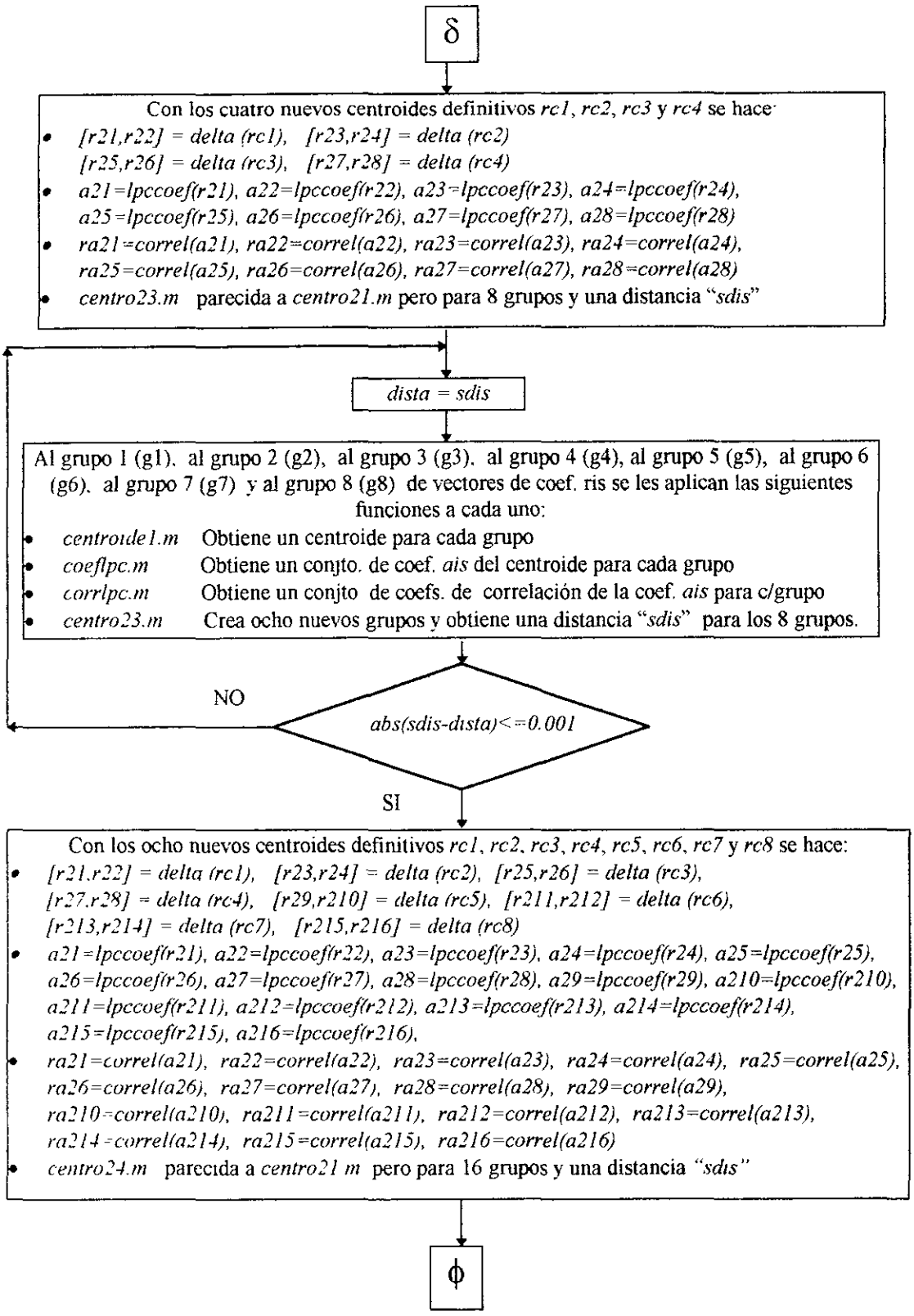
La distancia usada tanto para el proceso de reconocimiento como para la obtención de los doce cuantizadores vectoriales es la distancia de Itakura y Saito.

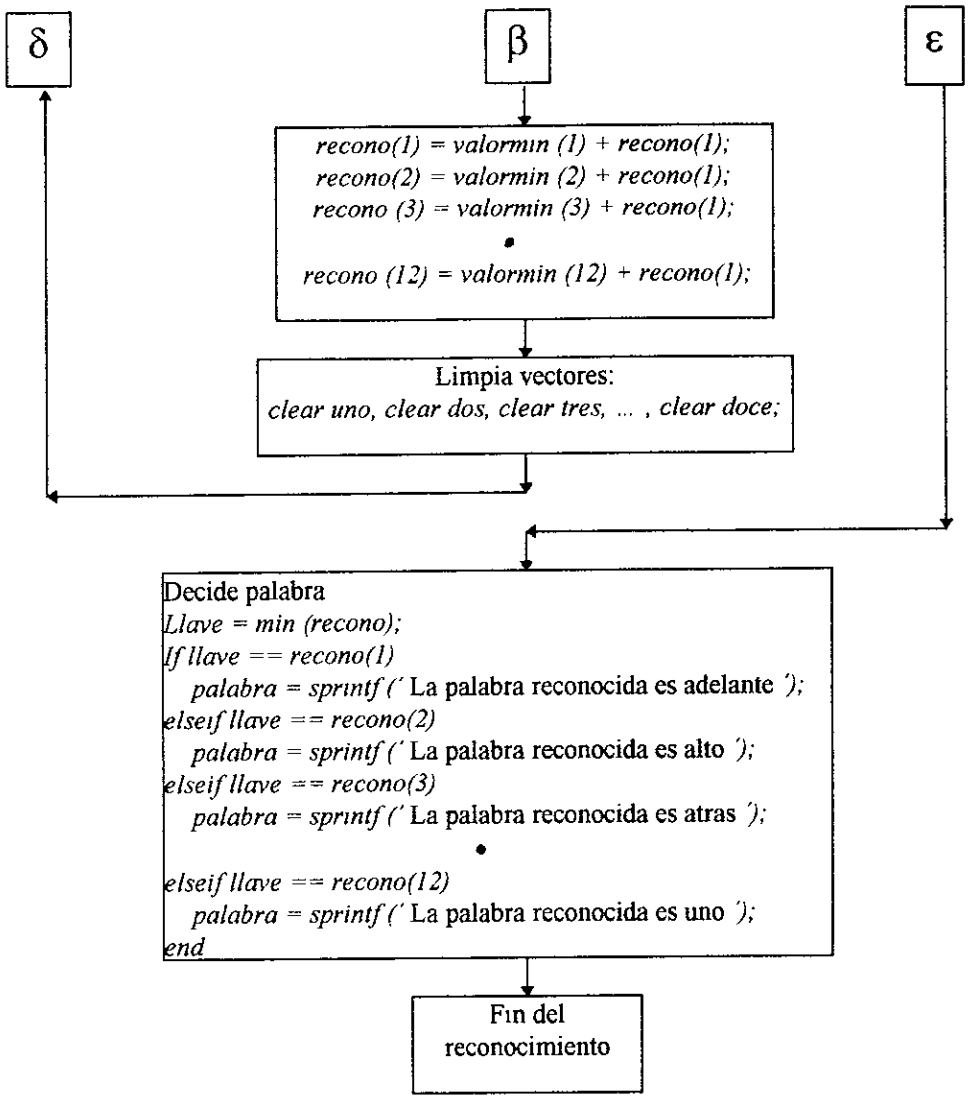
DIAGRAMA DE BLOQUES PARA LA OBTENCIÓN DE UN CUANTIZADOR VECTORIAL.




α
Se suman todas las r(0) de todos los frames de todas las matrices (palabra), asi como también se suman las r(1), r(2), .. , r(10) y cada una de esas sumas se divide entre sumlong que es no. de sumandos, es decir, sumlong=l1+l2+ ... +l10
centroide1.m
Con el paso anterior se obtiene el centroide rc
rc = [Σr(0)/sumlong Σr(1)/sumlong Σr(p)/sumlong]
centroide1.m
Se obtienen 2 centroides, con delta =0.001
r21 = rc – delta, r21 es un vector de p+1 elementos
r22 = rc + delta, r22 es un vector de p+1 elementos
centrodelta.m
Se obtienen los coefs ais., de los centroides anteriores
a21 = lpccoef (r21), a21 es un vector de p+1 elementos
a22 = lpccoef (r22), a22 es un vector de p+1 elementos
coeflpc.m
Se obtienen las correlaciones de los coef. (a21 y a22) anteriores
ra21 = correl (a21), ra21 es un vector de p+1 elementos
ra22 = correl (a22), ra22 es un vector de p+1 elementos
corrlpc.m
Se obtiene la distancia de Itakura y Saito entre ra21 y los l1 coef. de autocorrelación de la matriz r1, también la distancia entre ra21 y los l2 cjtos. de coefs. de autocorrelación de la matriz r2 y asi sucesivamente hasta obtener las l10 distancias entre ra21 y los l10 cjtos. de autocorrelación de la matriz r10. El mismo proceso se hace para ra22, y se comparan.
Si ra21<ra22 el 1er. grupo de coefs. ris de la matriz r1 se almacenan en el grupo 1 (g1) y el contador de g1 (cg1) es igual a 1.
Si ra21>ra22 el 1er. grupo de coefs. ris de la matriz r1 se almacenan en el grupo 2 (g2) y el contador de g2 (cg2) es igual a 1
Estas comparación continuan hasta los l1 cjtos. de la matriz r1. Posteriormente ra21 y ra22 se comparan con la matriz r2 para obtener las distancias y asi sucesivamente hasta la matriz r10.
Todas las distancias se van sumando y almacendo en la variable "sdis"
centro21.m
β


β
dista=sdis
Al grupo 1 (g1) de vectores de coef. ris y al grupo 2 (g2) de vectores de coef ris se les aplican las siguientes funciones:
centroide1.m Obtiene un centroide para cada grupo
coeflpc.m Obtiene un conjto. de coefs. ais del centroide para cada grupo
corrlpc.m Obtiene un conjto. de coefs. de correlación de la coef. ais para c/grupo
centro21 m Crea dos nuevos grupos y obtiene una distancia sdis para los dos grupos.
NO
abs(sdis-dista)<=0.001
SI
Con los dos nuevos centroides definitivos rc1 y rc2 se hace:
[r21,r22] = delta (rc1), [r23,r24] = delta (rc2)
a21=lpccoef(r21), a22=lpccoef(r22), a23=lpccoef(r23), a24=lpccoef(r24)
ra21=correl(a21), ra22=correl(a22), ra23=correl(a23), ra24=correl(a24)
centro22.m parecida a centro21.m pero para 4 grupos y una distancia "sdis"
dista = sdis
Al grupo 1 (g1), al grupo 2 (g2), al grupo 3 (g3) y al grupo 4 (g4) de vectores de coef. ris se les aplican a cada uno las siguientes funciones:
centroide1.m Obtiene un centroide para cada grupo
coeflpc.m Obtiene un conjto. de coef. ais del centroide para cada grupo
corrlpc.m Obtiene un conjto. de coefs de correlación de la coef. ais para c/grupo
centro22.m Crea cuatro nuevos grupos y obtiene una distancia "sdis" para los 4 grupos.
NO
abs(sdis-dista)<=0.001
SI
δ


δ
Con los cuatro nuevos centroides definitivos rc1, rc2, rc3 y rc4 se hace:
[r21,r22] = delta (rc1), [r23,r24] = delta (rc2)
[r25,r26] = delta (rc3), [r27,r28] = delta (rc4)
a21=lpccoef(r21), a22=lpccoef(r22), a23=lpccoef(r23), a24=lpccoef(r24),
a25=lpccoef(r25), a26=lpccoef(r26), a27=lpccoef(r27), a28=lpccoef(r28)
ra21=correl(a21), ra22=correl(a22), ra23=correl(a23), ra24=correl(a24),
ra25=correl(a25), ra26=correl(a26), ra27=correl(a27), ra28=correl(a28)
centro23.m parecida a centro21.m pero para 8 grupos y una distancia "sdis"
dista = sdis
Al grupo 1 (g1), al grupo 2 (g2), al grupo 3 (g3), al grupo 4 (g4), al grupo 5 (g5), al grupo 6 (g6), al grupo 7 (g7) y al grupo 8 (g8) de vectores de coef. ris se les aplican las siguientes funciones a cada uno:
centroide1.m Obtiene un centroide para cada grupo
coeflpc.m Obtiene un conjto. de coef. ais del centroide para cada grupo
corrlpc.m Obtiene un conjto de coefs. de correlación de la coef. ais para c/grupo
centro23.m Crea ocho nuevos grupos y obtiene una distancia "sdis" para los 8 grupos.
NO
abs(sdis-dista)<=0.001
SI
Con los ocho nuevos centroides definitivos rc1, rc2, rc3, rc4, rc5, rc6, rc7 y rc8 se hace:
[r21,r22] = delta (rc1), [r23,r24] = delta (rc2), [r25,r26] = delta (rc3),
[r27,r28] = delta (rc4), [r29,r210] = delta (rc5), [r211,r212] = delta (rc6),
[r213,r214] = delta (rc7), [r215,r216] = delta (rc8)
a21=lpccoef(r21), a22=lpccoef(r22), a23=lpccoef(r23), a24=lpccoef(r24), a25=lpccoef(r25),
a26=lpccoef(r26), a27=lpccoef(r27), a28=lpccoef(r28), a29=lpccoef(r29), a210=lpccoef(r210),
a211=lpccoef(r211), a212=lpccoef(r212), a213=lpccoef(r213), a214=lpccoef(r214),
a215=lpccoef(r215), a216=lpccoef(r216),
ra21=correl(a21), ra22=correl(a22), ra23=correl(a23), ra24=correl(a24), ra25=correl(a25),
ra26=correl(a26), ra27=correl(a27), ra28=correl(a28), ra29=correl(a29),
ra210=correl(a210), ra211=correl(a211), ra212=correl(a212), ra213=correl(a213),
ra214=correl(a214), ra215=correl(a215), ra216=correl(a216)
centro24.m parecida a centro21.m pero para 16 grupos y una distancia "sdis"
ϕ


δ
β
ε
recono(1) = valormin (1) + recono(1);
recono(2) = valormin (2) + recono(1);
recono (3) = valormin (3) + recono(1);
•
recono (12) = valormin (12) + recono(1);
Limpia vectores:
clear uno, clear dos, clear tres, ... , clear doce;
Decide palabra
Llave = min (recono);
If llave == recono(1)
palabra = sprintf (' La palabra reconocida es adelante ');
elseif llave == recono(2)
palabra = sprintf (' La palabra reconocida es alto ');
elseif llave == recono(3)
palabra = sprintf (' La palabra reconocida es atras ');
•
elseif llave == recono(12)
palabra = sprintf (' La palabra reconocida es uno ');
end
Fin del reconocimiento

V.7 Conclusiones.

La idea básica del CV es reducir la razón de información de la señal de voz a una razón baja a través del uso de un diccionario con un número relativamente pequeño de código de palabras. El objetivo de este método es representar la información espectral de la señal de una manera eficiente y con una manera de conexión directa para la forma de trabajo fonética-acustica. En el capítulo VII se observa otro tipo de técnica para reconocimiento de voz "Reconocimiento de voz por medio de lógica difusa", de la que se daran sus principios matemáticos en el capítulo VI.

Capítulo VI

TEORÍA DE LÓGICA DIFUSA.

VI.1 Introducción

La lógica difusa (fuzzy logic) nace en los Estados Unidos alrededor de 1965, en la Universidad de California, en Berkeley, creada por el profesor Lofti A. Zadeh.
Ha sido aplicada en diferentes áreas, como son comprensión de lenguaje, interpretación de información humana, relaciones humanas, planeación, toma de decisiones, control de robots y reconocimiento de patrones. También se utiliza en sistemas mecánicos como el reconocimiento de pinturas y de voz, operación de trenes y carros, sistemas de mantenimiento y seguridad, manejo de producción, operación de sistemas de potencia eléctrica y robots inteligentes.
La lógica difusa no resuelve todos los problemas. Los sistemas difusos trabajan mejor cuando se desea simular pensamientos y decisiones humanas y es una alternativa para controlar de otra manera complicados sistemas. Muchos científicos creen que la lógica difusa es una ciencia que es imprecisa y puede llegar a mentir y algunas veces a llegar a soluciones incorrectas, aunque los buenos resultados obtenidos contradicen esta idea.
Como se mencionó la lógica difusa elimina la complejidad de un sistema. Su aplicación no se reduce a pequeños módulos manejables como en un sistema tradicional. Una implementación difusa debe compararse con una implementación no difusa, la solución que mejor encuentre los objetivos del sistema debe ser escogida.

VI.2 Descripción de la Lógica Difusa.

La lógica difusa llega después de la lógica multivalente la cual propone 3 niveles logicos, verdadero (1), falso (0) y neutro (1/2) el cual representa la mitad verdadera y la mitad falsa, propuesta por Jon Lukasieviicz (1930). La lógica difusa abarca una infinidad de valores los cuales estarán comprendidos entre 0 (completamente falso) y 1 (completamente verdadero).
Cabe mencionar que para ubicar la lógica difusa en las aplicaciones con computadora deberemos primero dividir estas aplicaciones en tres grandes áreas como son las siguientes [27]

- Propósitos Numéricos: donde se aplican cálculos y análisis numérico, como solución de problemas matemáticos referidos a diferentes áreas del conocimiento humano
- Propósitos de Bases de Datos: donde se almacenan grandes cantidades de información.
- Ingeniería del Conocimiento· donde se encuentran los sistemas expertos, siendo esta el área más nueva.

Precisamente la lógica difusa tiene aplicaciones tanto en sistemas expertos, como en sistemas de control, aplicaciones en el área numérica, y la de Ingeniería del Conocimiento,

aunque, se pueden tener otras aplicaciones. La primera aplicación industrial en esta área se llevo a cabo en 1977 a la industria del cemento en Dinamarca [21].
Por otra parte, en la lógica difusa las reglas de subjetividad u objetividad, así como los conocimientos que se deben tener para simular la experiencia adquirida por aprendizaje de un experto para resolver un determinado problema, son expresados por sentencias o un conjunto de ellas. Las proposiciones son modelos escritos, así que podemos representar valores ambiguos. En términos generales los sistemas difusos son convenientes cuando existe incertidumbre o razonamiento aproximado. Es decir, cuando los modelos matemáticos que describen el comportamiento del sistema son demasiado complejos para modelar [29].

VI.3 Características Principales de la Lógica Difusa.

Dentro de los principales parámetros que caracterizan a la lógica difusa podemos mencionar los siguientes:

- Los métodos convencionales son buenos para resolver problemas simples en cuanto a su modelado, mientras que los sistemas basados en lógica difusa son convenientes para problemas complejos o aplicaciones que involucren descripciones humanas y pesamientos intuitivos.
- Capacidad de Aprendizaje, donde se añaden ciertos datos para configurar un mejor control
- Esta compuesta por Reglas de Inferencia, las cuales están hechas en base a un determinado conocimiento del problema, intuición y a la experiencia para resolverlo.

De esta forma la lógica difusa, maneja conceptos ambiguos como caliente, muy caliente, viejo, joven, etc. llamadas variables linguísticas. Es decir, el grado de ocurrencia, no tomando para este fin el paso del tiempo o las pruebas realizadas. **La teoría de la probabilidad mide la posibilidad de ocurrencia**, o de no ocurrencia de un evento, en tanto que **la lógica difusa mide el grado de ocurrencia** de un evento o de una determinada condición. O de forma sistematizada, podremos decir que la teoría de probabilidad esta basada en un razonamiento exacto, por manipulación simbólica y cálculos numéricos y esto a su vez por predicciones. Mientras que la Lógica Difusa está basada en razonamientos aproximados por manipulaciones simbólicas y cálculos numéricos y esto a su vez por condiciones ambiguas. De esta forma, podemos formar una estructura constituida por varios pasos para desarrollar un sistema difuso, como el siguiente [24], [27], [29]:

- Determinar si la estructura del problema a resolver amerita el uso de lógica difusa.
- Si es así, determinar el rango en donde se manejan las variables de entrada y de salida.
- Definir las funciones que constituyen las salidas y las entradas, siendo estas las posibles opciones que tomarán las entradas y salidas
- Construir las reglas que relacionarán las entradas y salidas.
- La determinación de dichas reglas requiere de pruebas exahustivas para poder verificar su correcto funcionamiento

- Los sistemas de control difuso en términos generales son estables, aunque la estabilidad de las reglas nos dan una estabilidad parcial con respecto a cada regla, para que el sistema sea estable, todas sus reglas en conjunto deberán ser estables.

VI.4 Descripción de conjuntos clásicos y conjuntos difusos.

Un *conjunto clásico* es definido por límites precisos, esto es, no existe incertidumbre en la prescripción o localización de los límites del conjunto como se presenta en la fig. VI.1a donde los límites del conjunto clásico (crisp) A es una línea no ambigua.
Por otra parte, un *conjunto difuso* es prescrito por propiedades vagas o ambiguas; por lo tanto sus límites son ambiguamente especificados, como se presenta por el límite confuso (difuso o fuzzy) para el conjunto $\underline{A}$ en la fig. VI.1b.
El punto a en la fig. VI.1a es claramente un miembro del conjunto crisp A; el punto b es claramente un no miembro del conjunto A. La fig. VI.1b presenta la vaguedad y el límite de ambigüedad de un conjunto difuso $\underline{A}$ sobre el universo X: los límites sombrados representan los límites de región de $\underline{A}$. En la región central del conjunto difuso que no esta sombreada, el punto a es claramente un miembro completo del conjunto difuso. Sin embargo, la afiliación (calidad de miembro) del punto c, el cual esta sobre la región límite, es ambigua. Si la alifiación en un conjunto (tal como el punto a) es completa entonces es representada por el número 1, y la no afiliación en un conjunto (tal como el punto b) es representada por el 0, entonces el punto c en la fig VI.1b debe tener algún valor intermedio de afiliación (afiliación parcial en el conjunto difuso $\underline{A}$) en el intervalo [0,1]. La afiliación del punto c en $\underline{A}$ se aproxima al valor de 1 conforme se acerca a la región central (no sobreada) en la fig. VI 1b [25] y [24].

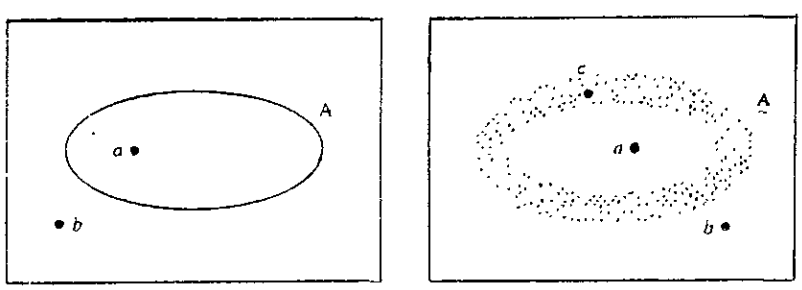


Fig. VI.1 DIAGRAMA PARA LIMITES DEL CONJUNTO CRISP (a) Y LIMITES DEL CONJUNTO DIFUSO (b).

VI.5 Conjuntos clásicos.

Un atributo útil de conjuntos y los universos sobre los cuales son definidos es una medida conocida como la cardinalidad, o el número cardinal. El número total de elementos en un universo X es llamado su número cardinal, denotado n_x, donde x es etiquetado para elementos individuales en el universo. Los universos discretos que estan compuestos de una colección finita contable de elementos tendrán un número cardinal; finito; los universos continuos estan compuestos de una colección infinita de elementos que tendrán una cardinalidad infinita. Colecciones de elementos dentro de un universo son llamados conjuntos y colecciones de elementos dentro de conjuntos son llamados subconjuntos [25].

Para conjuntos clásicos (crisp) A y B consistiendo de colecciones de algunos elementos en X, la siguiente notación es definida:

$x \in X \Rightarrow$ x pertenece a X
$x \in A \Rightarrow$ x pertenece a A
$x \notin A \Rightarrow$ x no pertenece a A
$A \subset B \Rightarrow$ A esta completamente contenido en B (si $x \in A$, entonces $x \in B$)
$A \subseteq B \Rightarrow$ A esta contenido en o es equivalente a B
$A = B \Rightarrow A \subseteq B$ y $B \subseteq A$

VI.5.1 Operaciones en conjuntos clásicos.

Sean A y B dos conjuntos en el universo X.

- La unión entre los conjuntos, denotada $A \cup B$, representa todos aquellos elementos en el universo que residen (o perteneceN a) ya sea en el conjunto A, el conjunto B, o en ambos conjuntos A y B. (Esta operación es también llamada la *lógica or*, otra forma de la union es la operación or exclusiva).
- La intersección de los dos conjuntos, denotada como $A \cap B$, representa todos aquellos elementos en el universo X que simultamente residen en (o pertenecen a) ambos conjuntos A y B
- El complemento de un conjunto A, denotado A, es definido como la colección de todos los elementos en el universo que no estan en el conjunto A.
- La diferencia de un conjunto A con respecto a B, denotada $A \mid B$, es definida como la colección de todos los elementos en el universo que residen en A y que no residen en B simultáneamente. Estas operaciones son presentadas a continuación en términos de conjunto teóricos:

Unión	$A \cup B = \{x \mid x \in A \text{ } or \text{ } x \in B\}$	(VI.1)
Intersección	$A \cap B = \{x \mid x \in A \text{ } and \text{ } x \in B\}$	(VI.2)
Complemento	$A = \{x \mid x \notin A, x \in X\}$	(VI.3)
Diferencia	$A \mid B = \{x \mid x \in A \text{ y } x \notin B\}$	(VI.4)

Ver figs. VI 2, VI 3, VI 4 y VI.5.

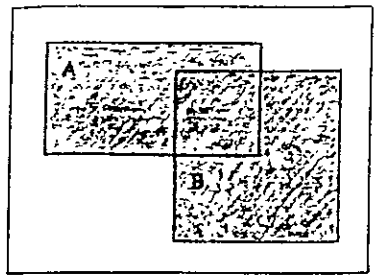


Fig. VI.2 UNIÓN DE LOS CONJUNTOS A Y B (LÓGICA OR)

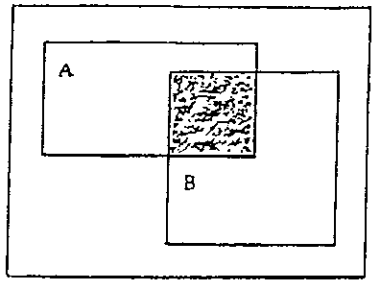


Fig. VI.3 INTERSECCIÓN DE LOS CONJUNTOS A Y B.

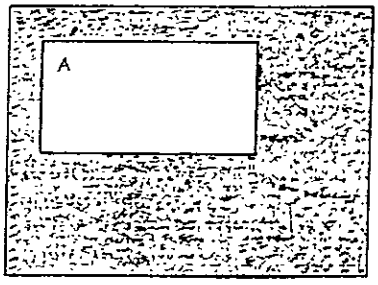


Fig. VI.4 COMPLEMENTO DEL CONJUNTO A.

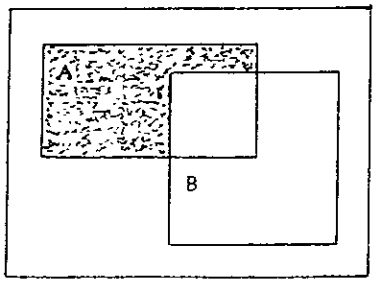


Fig VI.5 OPERACIÓN DIFERENCIA A| B DE LOS CONJUNTOS A Y B.

VI.5.2 Propiedades de conjuntos clásicos (crisp)

Ciertas propiedades de conjuntos son importantes por la manipulación matemática de conjuntos. **Las propiedades más apropiadas para definir los conjuntos clásicos y presentar su similaridad con los conjuntos fuzzy son las siguientes:**

Conmutatividad:	$A \cup B = B \cup A$	
	$A \cap B = B \cap A$	(VI.5)
Asociatividad:	$A \cup (B \cup C) = (A \cup B) \cup C$	
	$A \cap (B \cap C) = (A \cap B) \cap C$	(VI.6)
Distributividad:	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	
	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$	(VI.7)
Idempotencia:	$A \cup A = A$	
	$A \cap A = A$	(VI.8)
Identidad:	$A \cup \varnothing = \varnothing$	
	$A \cap X = A$	
	$A \cap \varnothing = \varnothing$	(VI.9)
	$A \cup X = X$	
Transitividad:	Si $A \subseteq B \subseteq C$, entonces $A \subseteq C$	(VI.10)
Involución:	$\overline{\overline{A}} = A$	(VI.11)

donde $\varnothing$ es el conjunto vacío.

El área doblemente achurada en la fig. VI.6 es un ejemplo del diagrama de Venn de la propiedad de asociatividad para intersección y las áreas doblemente achuradas en las figs. VI.7 y VI.8 son ejemplos del diagrama de Venn de la propiedad de distribución para varias combinaciones de las propiedades de unión e intersección.

Dos propiedades de operaciones conjunto son conocidas como *las leyes media exclusión* y *leyes de Morgan*. **Las leyes de media exclusión son muy importantes porque estas son las únicas operaciones conjunto descritas aquí que no son validas para ambos, conjuntos clásicos y conjuntos difusos**. Existen dos leyes de media exclusión. La primera es llamada *la ley de la exclusión media*, marcada con la unión de un conjunto A y su complemento; la segunda es llamada *la ley de contradicción*, que representa la intersección de un conjunto A y su complemento.

Ley de la exclusión media	$A \cup \overline{A} = X$	(VI.12a)
Ley de contradicción	$A \cap \overline{A} = \varnothing$	(VI.12b)

Las leyes de Morgan son importantes por su utilidad en tautología y contradicciones en lógica, así como en otras operaciones y pruebas de conjuntos. Las leyes de Morgan son desplegadas en el área achurada de los diagramas de Venn de las figs. VI.9 y VI.10 y descritas matemáticamente como se muestra a continuación:

$$\overline{A \cap B} = \bar{A} \cup \bar{B} \qquad \text{(VI.13a)}$$

$$\overline{A \cup B} = \bar{A} \cap \bar{B} \qquad \text{(VI.13b)}$$

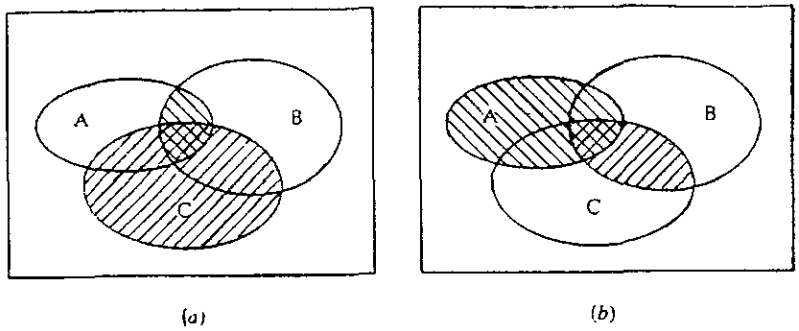


Fig. VI.6 DIAGRAMAS DE VENN PARA (a) $(A \cap B) \cap C$ Y (b) $A \cap (B \cap C)$

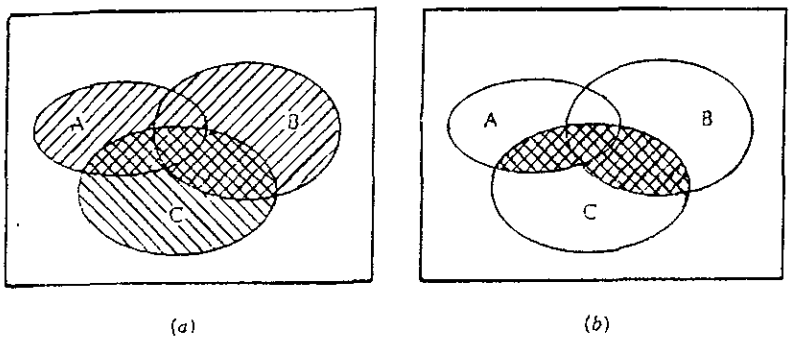


Fig VI 7 DIAGRAMAS DE VENN PARA (a) $(A \cup B) \cap C$ Y (b) $(A \cap C) \cup (B \cap C)$

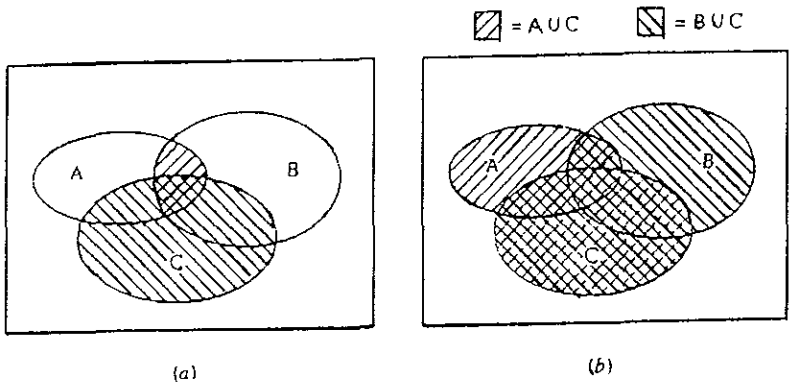


Fig. VI.8 DIAGRAMAS DE VENN PARA (a) $(A \cap B) \cup C$ Y (b) $(A \cup C) \cap (B \cup C)$.

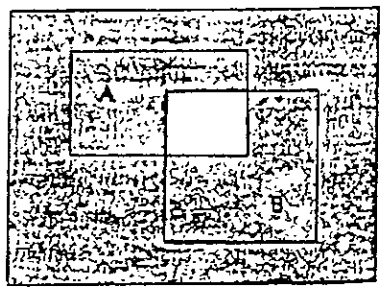


Fig. VI.9 LEY DE MORGAN ($\overline{A \cap B}$)

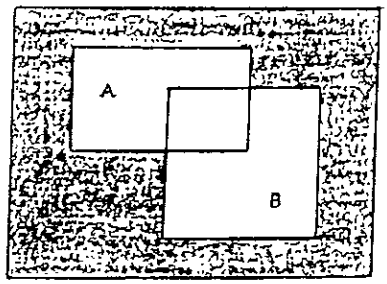


Fig. VI.10 LEY DE MORGAN ($\overline{A \cup B}$)

VI 5.3 Mapeo de conjuntos clásicos a funciones.

Sea X y Y dos universos diferentes de discurso. Si un elemento x es contenido en X y corresponde a un elemento y contenido en Y, eso es generalmente un mapeo de X a Y, o $f: X \rightarrow Y$. Como un mapeo la función característica x_A es definida por

$$x_A = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases} \qquad \text{(VI.14)}$$

donde x_A expresa la "calidad de miembro" (membresía o membership) en el conjunto A para el elemento x en el universo. Esta idea de membresía es un mapeo de un elemento x en el universo X a uno de los dos elementos en el universo Y; esto es, para los elementos 0 o 1 como se presenta en la fig. VI.11.

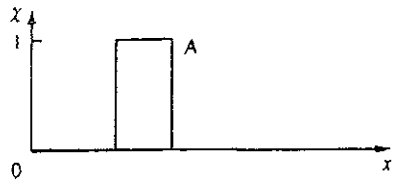


Fig. VI 11 LA FUNCIÓN MEMBRESÍA ES UN MAPEO PARA UN CONJUNTO CRISP A

Para cualquier conjunto A definido en el universo X, existe un conjunto función teórico, llamado un valor conjunto, denotado V(A), bajo el mapeo de la función característica x. Por convención, a la calidad de miembro (membresía) del conjunto nulo $\varnothing$ se asigna el valor 0 y a la calidad de miembro (membresía) del conjunto completo X se asigna el valor 1.

Si se definen dos conjuntos A y B, sobre el universo X. La unión de estos dos conjuntos en términos de función-teórica es dada como sigue (el símbolo $\vee$ es el operador máximo y $\wedge$ es el operador mínimo) Cuatro operaciones en términos función teórica son dados por:

Unión $A \cup B \rightarrow x_{A \cup B}(x) = x_A(x) \vee x_B(x) = \max(x_A(x), x_B(x))$ (VI.15)

Intersección $A \cap B \rightarrow x_{A \cap B}(x) = x_A(x) \wedge x_B(x) = \min(x_A(x), x_B(x))$ (VI.16)

Complemento $\overline{A} \rightarrow x_{\bar{A}}(x) = 1 - x_A(x)$ (VI.17)

Para dos conjuntos sobre el mismo universo, es decir A y B, si un conjunto (A) es contenido en otro conjunto (B), entonces

Contenido $A \subseteq B \rightarrow x_A(x) \leq x_B(x)$ (VI.18)

VI.6. Conjuntos difusos.

Un conjunto difuso, es un conjunto que contiene elementos que tienen grados de membresía (membership) variable. Esta idea es contrastante con conjuntos clásicos o conjuntos crisp, porque los miembros de un conjunto crisp no son miembros a menos que su membresía sea completa o total, (esto es, su membresía es asignada al valor de 1). Los elementos en un conjunto difuso, su membresía no necesita ser completa, pueden también ser miembros de otro conjunto difuso sobre el mismo universo.

Los elementos de un conjunto difuso son mapeados a un universo de valores de membrersía usando una forma de función teórica. Se denota un conjunto difuso como $\underline{A}$ para ser el conjunto difuso A. Esta función mapea elementos de un conjunto difuso A a un

valor real numerado en el intervalo de 0 a 1. Si un elemento en el universo, es decir x, es un miembro de un conjunto difuso A, entonces el mapeado es dado por $\mu_{\underline{A}}(x) \in [0,1]$. Este mapeo es presentado en la fig. VI.12 para un conjunto difuso típico.

Una convención para la notación de conjuntos difusos cuando el universo de discurso X, es discreto y finito, es como sigue para un conjunto difuso $\underline{A}$ [25]:

$$\underline{A} = \{ \mu_{\underline{A}}(x_1)/ x_1 + \mu_{\underline{A}}(x_2)/ x_2 + \ldots \} = \{ \sum_i \mu_{\underline{A}}(x_i)/ x_i \} \qquad \text{(VI.19)}$$

Cuando el universo X es continuo e infinito, el conjunto difuso $\underline{A}$ es denotado por.

$$\underline{A} = \{ \int \mu_{\underline{A}}(x)/ x \} \qquad \text{(VI.20)}$$

En ambas notaciones, la barra horizontal no es cociente, sino un delimitador. El numerador en cada termino es el valor membresía (membership) en el conjunto $\underline{A}$ asociado con el elemento del universo indicado en el denominador. En la primera notación, el símbolo sumatoria no es una sumatoria algebraica, sino denota la colección o agregación de cada elemento; por lo tanto el signo "+" no es la adición algebraica pero si es la unión función teórica. En la segunda notación el signo integral no es la integral algebraica, es la unión función teórica para variables continuas.

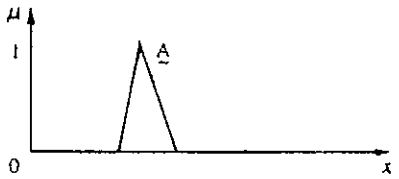


Fig. VI.12 FUNCIÓN MEMBRESÍA PARA UN CONJUNTO FUZZY A.

VI 6 1 Operaciones en conjuntos fuzzy.

Definidos tres conjuntos fuzzy $\underline{A}$, $\underline{B}$ y $\underline{C}$ sobre el universo X. Para un elemento dado x del universo, las siguientes operaciones de función-teóricas para las operaciones sobre conjuntos son definidas·

Unión $x_{\underline{A}\cup\underline{B}}(x) = x_{\underline{A}}(x) \vee x_{\underline{B}}(x)$ (VI.21)
Intersección $x_{\underline{A}\cap\underline{B}}(x) = x_{\underline{A}}(x) \wedge x_{\underline{B}}(x)$ (VI.22)
Complemento $x_{\underline{A}}(x) = 1 - x_{\underline{A}}(x)$ (VI.23)

Los diagramas de Venn para estas operaciones son presentadas en las figs. VI.13, VI.14 y VI 15

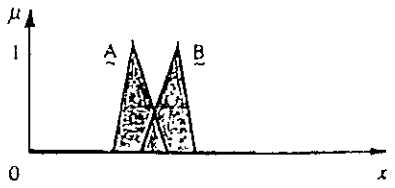


Fig VI.13 UNIÓN DE CONJUNTOS DIFUSOS A Y B

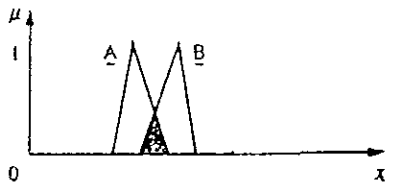


Fig VI.14 INTERSECCIÓN DE CONJUNTOS DIFUSOS A Y B

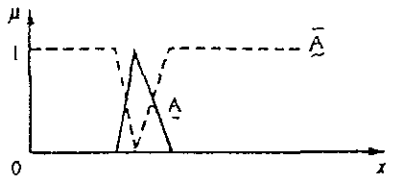


Fig VI.15 COMPLEMENTO DEL CONJUNTO DIFUSO A

Cualquier conjunto difuso A definido en el universo X es un subconjunto de aquel universo. También por definición, justo como en los conjuntos clásicos, el valor de membresía (membership) de cualquier elemento x en el conjunto nulo ∅ es 0, y el valor de membresía (membership) de un elemento x en el conjunto completo X es 1. Hay que notar que el conjunto nulo y el conjunto completo no son conjuntos difusos en este contexto. La notación apropiada para estas ideas es:

$$A \subseteq X \Rightarrow x_A(x) \leq x_X(x) \quad \text{(VI.24)}$$

$$\text{Para toda } x \in X, \quad x_\varnothing(x) = 0 \quad \text{(VI.25)}$$

$$\text{Para toda } x \in X, \quad x_X(x) = 1 \quad \text{(VI.26)}$$

La colección de todos los conjuntos difusos y los subconjuntos difusos sobre X es denotado como el conjunto potencia difuso P'(X). Debe ser obvio, basado sobre el hecho que todos los conjuntos pueden traslaparse, que la cardinalidad $n_{P(X)}$, de la potencia difusa es infinita; esto es $n_{P(X)} = \infty$

Las leyes de Morgan para conjuntos clásicos también se mantiene para conjuntos difusos, denotada por estas expresiones.

$$\overline{\underline{A} \cap \underline{B}} = \overline{\underline{A}} \cup \overline{\underline{B}} \qquad \text{(VI.27a)}$$

$$\overline{\underline{A} \cup \underline{B}} = \overline{\underline{A}} \cap \overline{\underline{B}} \qquad \text{(VI.27b)}$$

Como se mencionó anteriormente, todas las operaciones sobre conjuntos clásicos también se mantienen para conjuntos difusos, excepto las leyes de media exclusión. Estas dos leyes no se mantienen para conjuntos difusos, ya que los conjuntos difusos pueden traslaparse, un conjunto difuso y su complemento pueden también traslaparse. Las leyes de media exclusión, se extienden para conjuntos difusos y son expresadas por:

$$\underline{A} \cup \overline{\underline{A}} \neq X \qquad \text{(VI.28a)}$$

$$\underline{A} \cap \overline{\underline{A}} \neq \varnothing \qquad \text{(VI.28b)}$$

En las figs. VI.16 y VI.17 se comparan los diagramas de Venn para las leyes de media exclusión en conjuntos clásicos (crisp) y conjuntos difusos

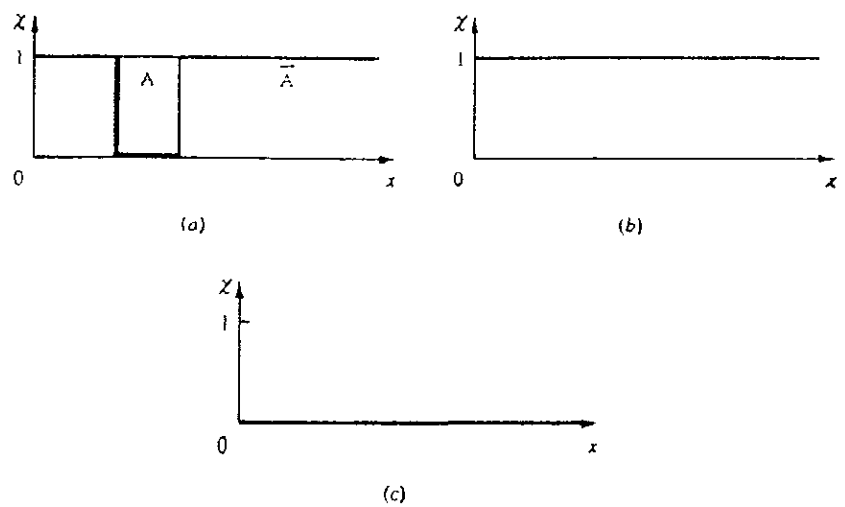


Fig VI.16 LEYES DE MEDIA EXCLUSIÓN PARA CONJUNTOS CRISP. (a) CONJUNTO CRISP A Y SU COMPLEMENTO; (b) CRISP $A \cup \bar{A} = X$ (LEY DE MEDIA EXCLUSIÓN); (c) CRISP $A \cap \bar{A} = \varnothing$ (LEY DE CONTRADICCIÓN).

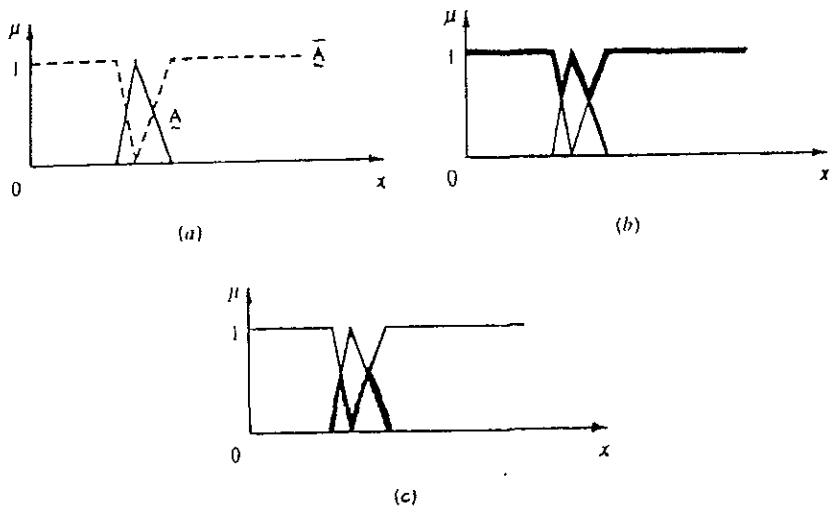


Fig. VI.17 LEYES DE MEDIA EXCLUSIÓN PARA CONJUNTOS DIFUSOS. (a) CONJUNTO DIFUSO $\underline{A}$ Y SU COMPLEMENTO; (b) DIFUSO $\underline{A}\cup\bar{\underline{A}}\neq X$ (LEY DE MEDIA EXCLUSIÓN; (c) DIFUSO $\underline{A}\cap\bar{\underline{A}}\neq\varnothing$ (LEY DE CONTRADICCIÓN).

VI.6 2 Propiedades de conjuntos difusos.

Los conjuntos difusos siguen las mismas propiedades que los conjuntos crisp. Porque los valores de membresía (membership) de un conjunto crisp son un subconjunto del intervalo [0,1], los conjuntos clásicos pueden ser clasificados como un caso especial de conjuntos difusos. Las propiedades de conjuntos difusos frecuentemente usadas son:

Conmutatividad:	$\underline{A}\cup\underline{B} = \underline{B}\cup\underline{A}$	
	$\underline{A}\cap\underline{B} = \underline{B}\cap\underline{A}$	(VI.29)
Asociatividad:	$\underline{A}\cup(\underline{B}\cup\underline{C}) = (\underline{A}\cup\underline{B})\cup\underline{C}$	
	$\underline{A}\cap(\underline{B}\cap\underline{C}) = (\underline{A}\cap\underline{B})\cap\underline{C}$	(VI.30)
Distributividad.	$\underline{A}\cup(\underline{B}\cap\underline{C}) = (\underline{A}\cup\underline{B})\cap(\underline{A}\cup\underline{C})$	
	$\underline{A}\cap(\underline{B}\cup\underline{C}) = (\underline{A}\cap\underline{B})\cup(\underline{A}\cap\underline{C})$	(VI.31)
Idempotencia:	$\underline{A}\cup\underline{A} = \underline{A}$	
	$\underline{A}\cap\underline{A} = \underline{A}$	(VI.32)
Identidad:	$\underline{A}\cup\varnothing = \underline{A}$	
	$\underline{A}\cap X = \underline{A}$	
	$\underline{A}\cap\varnothing = \varnothing$	(VI.33)
	$\underline{A}\cup X = X$	
Transitividad.	Si $\underline{A}\subset\underline{B}\subset\underline{C}$, entonces $\underline{A}\subset\underline{C}$	(VI.34)
Involución.	$\bar{\bar{\underline{A}}} = \underline{A}$	(VI.35)

Ejemplo VI.1

Para ilustrar estas ideas numéricas, se tienen dos conjuntos difusos $\underline{A}$ y $\underline{B}$ (siendo la membresía para el elemento 1 igual a 0 para ambos conjuntos difusos $\underline{A}$ y $\underline{B}$)

$$\underline{A} = \{ 1/2+.5/3+.3/4+.2/5 \} \quad \text{y} \quad \underline{B} = \{ .5/2+.7/3+.2/4+.4/5 \}$$

Podemos calcular varias de las operaciones descritas anteriormente:

Complemento	$\bar{\underline{A}} = \{ 1/1+0/2+.5/3+.7/4+.8/5 \}$
	$\bar{\underline{B}} = \{ 1/1+.5/2+.3/3+.8/4+.6/5 \}$
Unión	$\underline{A} \cup \underline{B} = \{ 1/2+.7/3+.3/4+.4/5 \}$
Intersección	$\underline{A} \cap \underline{B} = \{ .5/2+.5/3+.2/4+.2/5 \}$
Diferencia	$\underline{A} \mid \underline{B} = \underline{A} \cap \bar{\underline{B}} = \{ 5/2+.3/3+.3/4+.2/5 \}$
	$\underline{B} \mid \underline{A} = \underline{B} \cap \bar{\underline{A}} = \{ 0/2+.5/3+.2/4+.4/5 \}$
Leyes de Morgan	$\overline{\underline{A} \cup \underline{B}} = \bar{\underline{A}} \cap \bar{\underline{B}} = \{ 1/1+.0/2+.3/3+ 7/4+.6/5 \}$
	$\overline{\underline{A} \cap \underline{B}} = \bar{\underline{A}} \cup \bar{\underline{B}} = \{ 1/1+.5/2+.5/3+.8/4+.8/5 \}$
Leyes de media exclusión	$\bar{\underline{A}} \cup \underline{A} = \{ 1/1+.1/2+.5/3+.7/4+.8/5 \}$
	$\bar{\underline{B}} \cap \underline{B} = \{ .5/2+.3/3+.2/4+.4/5 \}$

Ejemplo VI.2:

Sean dos conjuntos difusos $\underline{F}$ y $\underline{D}$

$$\underline{F} = \{ 0/1+ 5/2+1/3+ 5/4+0/5 \} \qquad \text{y} \qquad \underline{D} = \{ 0/2+1/3+0/4 \}$$

y sus complementos:

$$\bar{\underline{F}} = \{ 1/1+.5/2+0/3+.5/4+1/5 \} \qquad \text{y} \qquad \bar{\underline{D}} = \{ 1/2+0/3+1/4 \}$$

La unión de $\underline{F}$ y $\bar{\underline{F}}$ es presentada en la fig. VI.18, la función de membresía (función membership) de $\underline{F}$ es representada por las lineas AB y BC, y la función membership de $\bar{\underline{F}}$ es representada por las lineas PQ y QR. La intersección de $\underline{D}$ y $\bar{\underline{D}}$ es presentada en la fig. VI 19 donde la función membership de $\underline{D}$ es representada por las lineas AB y BC, y la función membership de $\bar{\underline{D}}$ es representada por las lineas PQ y QR.

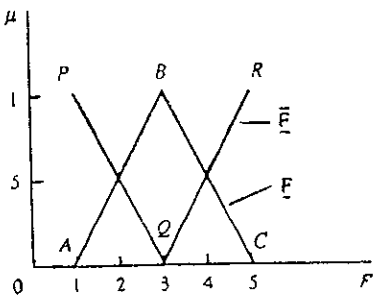


Fig. VI.18 FUNCIÓN DE MEMBRESÍA PARA LA OPERACIÓN UNIÓN

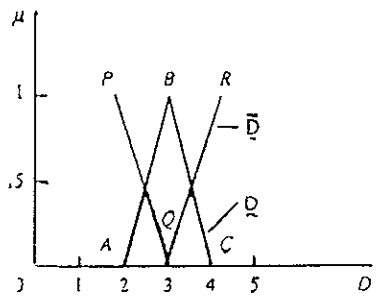


Fig. VI.19 FUNCIÓN DE MEMBRESÍA PARA LA OPERACIÓN INTERSECCIÓN

VI.7 Conjuntos como puntos en hipercubos.

Para un universo con solo un elemento, la función de membresía es definida sobre el intervalo unitario [0,1]; para un universo de dos elementos, la función de membresía es definida sobre un cuadrado unitario y para un universo de tres elementos, la función de membresia es definda sobre un cubo unitario

Por ejemplo en la fig. VI 20c el universo compuesto de tres elementos , $X=\{x_1, x_2, x_3\}$, el punto (0,0,1) representa el subconjunto crisp en tres dimensiones, donde x_1 y x_2, no tienen membresía y el elemento x_3 tiene una membresía completa, y asi sucesivamente para los otros siete vertices en la fig. VI.20c. Asi como el punto (1,1,1) donde todos los elementos en el universo tienen membresía total es llamado el conjunto completo X, y el punto (0,0,0) donde todos los elementos en el universo no tienen membresía, es llamado el conjunto nulo $\varnothing$.

Los centroides de cada uno de los diagramas en la fig. VI.20 represetan puntos únicos donde el valor de membresía para cada elemento en el universo es igual a 1/2. Por ejemplo, el punto (1/2,1/2) en la fig. VI.20b esta en el punto medio del cuadrado. Este punto medio en cada una de las figuras es un punto especial, este es el conjunto de máxima confusión ("fuzziness"). Un valor de membresía (membeship) de ½ indica que el elemento pertenece a un conjunto difuso tanto como no pertenece a este conjunto. En la fig. VI.20b,

el punto (1/4,3/4) representa un conjunto difuso donde la variable x_1, tiene 0.25 de grado membresía en el conjunto y la variable x_2, tiene 0.75 de grado de membresía en el conjunto.

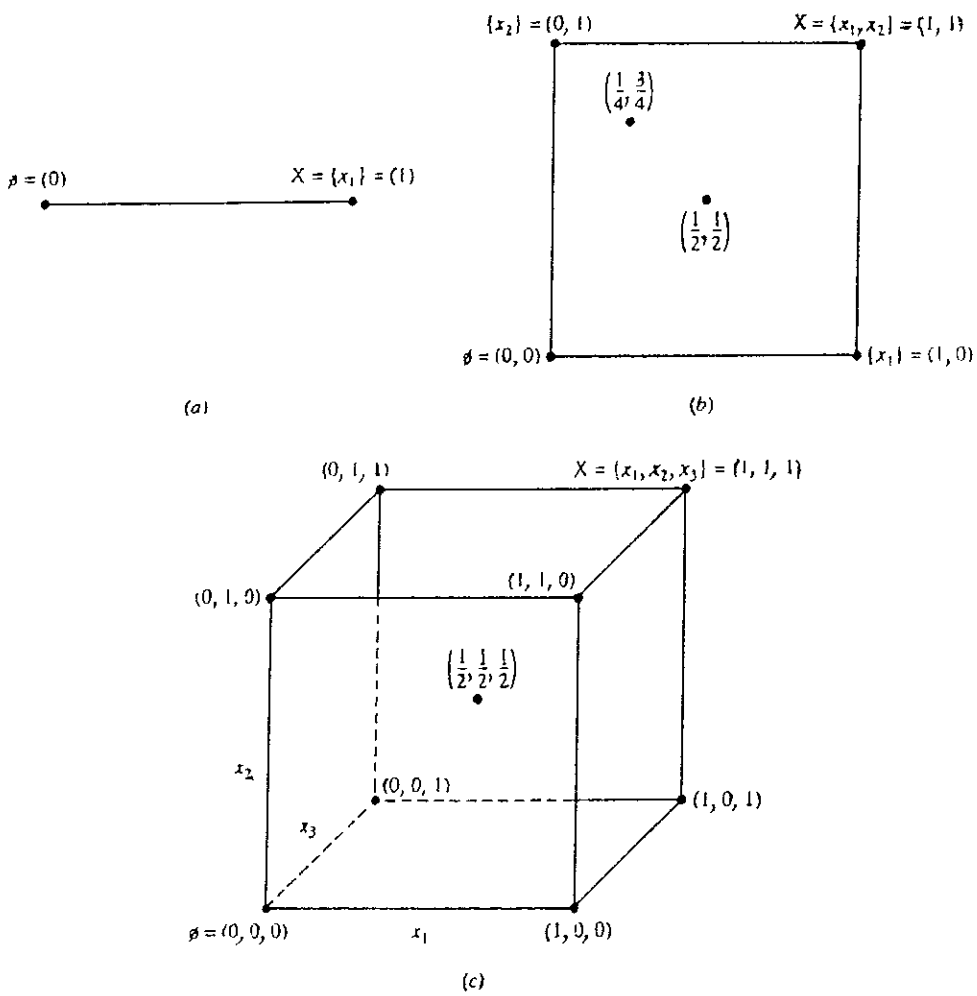


Fig. VI.20 "CONJUNTOS COMO PUNTOS" (a) UNIVERSO DE UN ELEMENTO; (b) UNIVERSO DE DOS ELEMENTOS: (c) UNIVERSO DE TRES ELEMENTOS, [25].

VI.8 Características de las funciones de membresía (funciones membership).

Una función de membresía es la que personifica un conjunto particular difuso. Ya que toda la información contenida en un conjunto difuso es descrita por sus funciones de membresía, es útil desarrollar un léxico de términos para describir varias características especiales de estas funciones Para propósitos de simplicidad, las funciones mostradas en las siguientes figuras, serán continuas; pero los términos se aplican igualmente para ambos conjuntos difusos discretos y continuos. La fig. VI.21 es una descripción de una función de membresía [12].

El centro (*core*) de una función de membresía para algún conjunto difuso $\underline{A}$ es definido como la región del universo que es caracterizada por una membresía completa en el conjunto $\underline{A}$. Esto es, el centro comprende aquellos elementos x del universo tal que $\mu_{\underline{A}}(x)=1$.

El soporte (*support*) de una función de membresía para algún conjunto difuso $\underline{A}$ es definido como la región del universo que es caracterizada por una membresía diferente de cero en el conjunto $\underline{A}$. Esto es, el soporte comprende aquellos elementos x del universo tal que $\mu_{\underline{A}}(x)>0$.

Los límites (*boundaries*) de una función de membresía para algún conjunto difuso $\underline{A}$ son definidos como aquellas regiones del universo que contenienen elementos que tienen una membresía diferente de cero pero no es completamente un membresía igual a 1. Esto es, los límites comprenden aquellos elementos x del universo tal que $0<\mu_{\underline{A}}(x)<1$. Estos elementos del universo son aquellos con algún grado de confusión (fuzziness), o solo una membresía parcial en el conjunto difuso $\underline{A}$. La figura VI.21 ilustra las regiones en el universo que contenienen el centro, soporte y límites de un conjunto difuso típico.

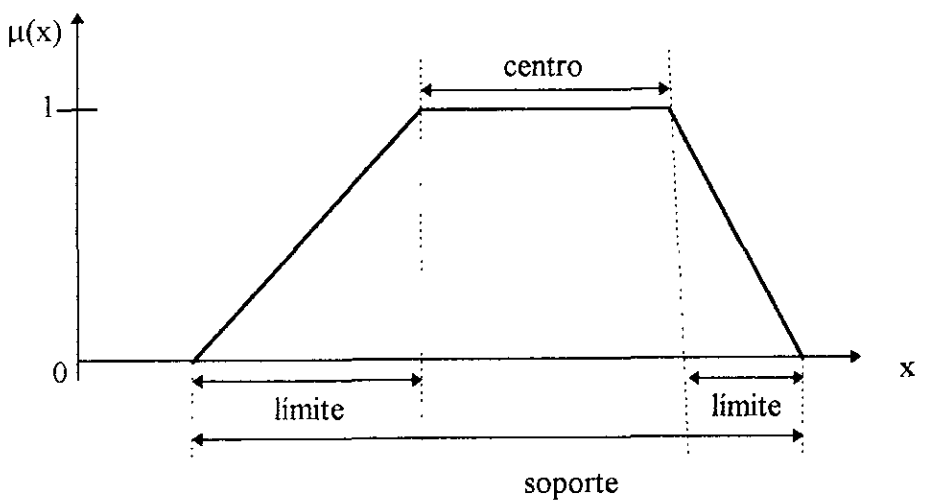


Fig. VI.21 CENTRO, SOPORTE Y LIMITES DE UN CONJUNTO DIFUSO.

Un conjunto difuso *normal* es una función de membresía que tiene al menos un elemento x en el universo cuyo valor de membresía es la unidad. Para conjuntos difusos donde uno y solo un elemento tiene una membresía igual a uno, este elemento es típicamente referido como el prototipo del conjunto. La fig. VI.22 ilustra un conjunto difuso típico normal y subnormal.

Un conjunto difuso *convexo* es descrito por una función de membresía cuyos valores de membresía se incrementan estrictamente, o cuyos valores de membresía (grados de pertenencia) se decrementan estrictamente. Es decir, para cualquier elemento x, y, z en el conjunto difuso $\underline{A}$ la relación $x<y<z$ implica que

$$\mu_{\underline{A}}(y) \geq \min [\mu_{\underline{A}}(x), \mu_{\underline{A}}(z)] \qquad \text{(VI.36)}$$

entonces $\underline{A}$ es un conjunto convexo difuso. La fig. VI.23 presenta un conjunto convexo difuso y un conjunto no convexo difuso.

Una propiedad especial de dos conjuntos convexos difusos, $\underline{A}$ y $\underline{B}$, es que la intersección de estos dos conjuntos es también un convexo difuso, como se presenta en la fig. VI 24. Esto es, para $\underline{A}$ *and* $\underline{B}$, que ambos son convexos, $\underline{A} \cap \underline{B}$ es también convexo.

Los *puntos de cruce* de una función de membresía son definidos como los elementos en el universo para el cual un conjunto particular difuso $\underline{A}$ tiene valores iguales a 0 5, esto es, para el cual $\mu_{\underline{A}}(y) = 0.5$.

La *altura* de un conjunto difuso $\underline{A}$ es el máximo valor de la función de membresía, esto es $\max \{\mu_{\underline{A}}(y)\}$. Si la altura de un conjunto difuso es menor que la unidad, el conjunto difuso se dice que es subnormal.

Si $\underline{A}$ es un conjunto difuso normal convexo con un punto único definido sobre la linea real, entonces $\underline{A}$ es frecuentemente condicionado un *número difuso*.

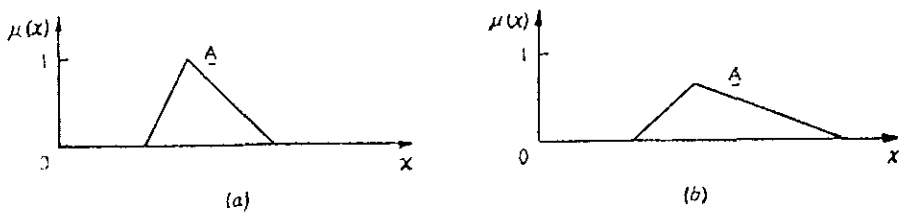


Fig. VI.22 CONJUNTOS DIFUSO (a) NORMALES Y (b) SUBNORMALES.

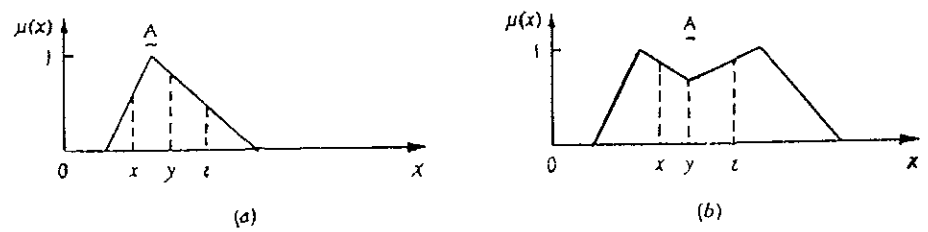


Fig. VI 23 (a) CONJUNTO DIFUSO CONVEXO NORMAL (b) CONJUNTO DIFUSO NO CONVEXO NORMAL.

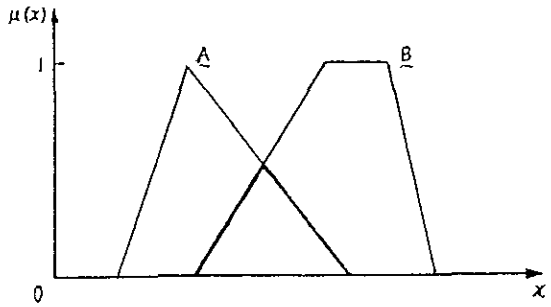


Fig. VI.24 LA INTERSECCIÓN DE DOS CONJUNTOS DIFUSOS CONVEXOS PRODUCE UN CONJUNTO DIFUSO CONVEXO

VI.9 Fuzzificación.

La fuzzificación es el proceso de convertir una cantidad precisa en una cantidad difusa. Esto se hace reconociendo que muchas de las cantidades que se consideran precisas y resueltas no son resueltas del todo. Estas cantidades llevan considerable incertidumbre, si la incertidumbre se incrementa a causa de imprecisión, ambigüedad o vaguedad, entonces la variable es probablemente difusa y puede ser representada por una función de membresía.

En el mundo real, el hardware tal como un volmetro digital genera datos precisos, pero estos datos son sujetos a errores experimentales. La información en la fig. VI.25 presenta un rango posible de errores para un voltaje leido y la función de membresía asociada puede representar tal imprecisión [26]:

La representación de datos imprecisos como conjuntos difusos son útiles pero no es un paso obligatorio cuando los datos son usados en sistemas difusos. Esta idea es presentada en la fig. VI.26, donde se consideran los datos como una lectura precisa fig. VI.26a, o como una lectura difusa presentada en la fig. VI.26b. En la fig. VI.26a podemos comparar una lectura de voltaje precisa con un conjunto difuso, es decir "voltaje bajo". En la figura se observa que la lectura precisa (crip) intersecta el conjunto difuso "voltaje bajo" en un grado de pertenencia de 0.3; esto es, que el conjunto difuso y la lectura concuerdan con el valor de membresía de 0.3. En la fig. VI.26 la intersección del conjunto difuso "voltaje medio" y la lectura del voltaje defusificado ocurre en un valor de membresía de 0.4. Se observa en la fig. VI.26b que el conjunto de intersección de los dos conjuntos difusos es un triángulo pequeño, y el valor de membresía más grande es de 0.4

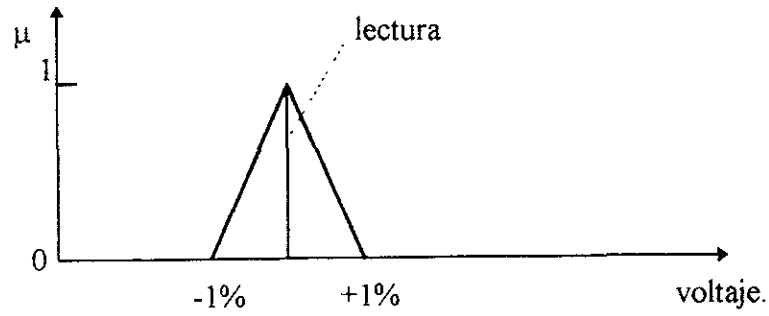


Fig. VI.25 FUNCIÓN DE MEMBRERSÍA REPRESENTANDO IMPRECISIÓN EN LA "LECTURA PRECISA DEL VOLTAJE"

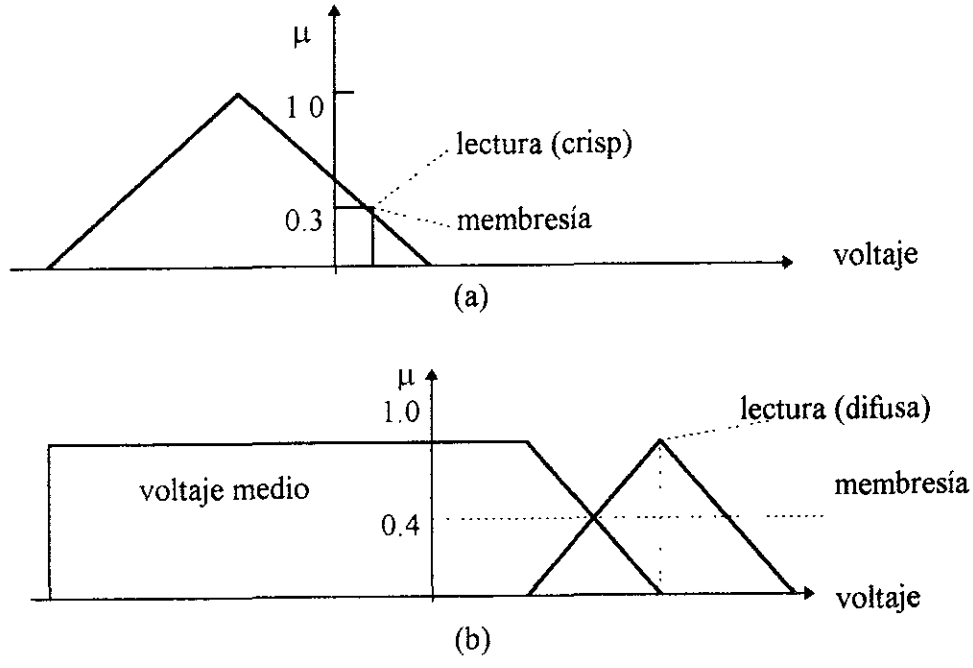


Fig. VI.26 COMPARACIÓN DE CONJUNTOS DIFUSOS Y LECTURA CRISP (PRECISIÓN): (a) CONJUNTO DIFUSO Y LECTURA PRECISA; (b) CONJUNTO DIFUSO Y LECTURA DIFUSA.

VI.10 Variables lingüísticas y razonamiento aproximado.

Una variable lingüistica es el nombre que se le da a una característica que representa "algún" conocimiento sobre "algo". Y esta variable no es otra cosa que un subconjunto del conjunto universal X, donde a dicha variable se le podrán cargar ciertos atributos, que no serán otra cosa que los conjuntos difusos, donde las funciones de membresía μ(X), indica el grado de pertenencia a cada uno de los conjuntos difusos involucrados.

Una variable lingüística podrá ser caracterizada por una función con 5 elementos (X, T(X), U, G, M), donde:

- X es el nombre de la variable
- T(X) es la característica o el término conjunto de X, que es el conjunto de nombres de valores lingüísticos o características, los cuales describirán algún conocimiento de "algo" siendo cada uno de ellos un conjunto difuso, definidos en el universo U
- U es el universo, marco en el que se encuentra comprendida la varible X.
- G es una regla sintáctica para generar nombres de valores de X.
- M es una regla semántica para asociar con cada valor el significado, M(X) denota un subconjunto de U

Por ejemplo, velocidad es interpretada como una variable lingüística donde estarán los conjuntos difusos involucrados con esta variable T(*velocidad*)={ *muy despacio, despacio, moderado, rápido*} y el universo estará definido por U=[0,180] definiendo los límites para cada conjunto difuso como se muestra en la fig. VI.27 [27].

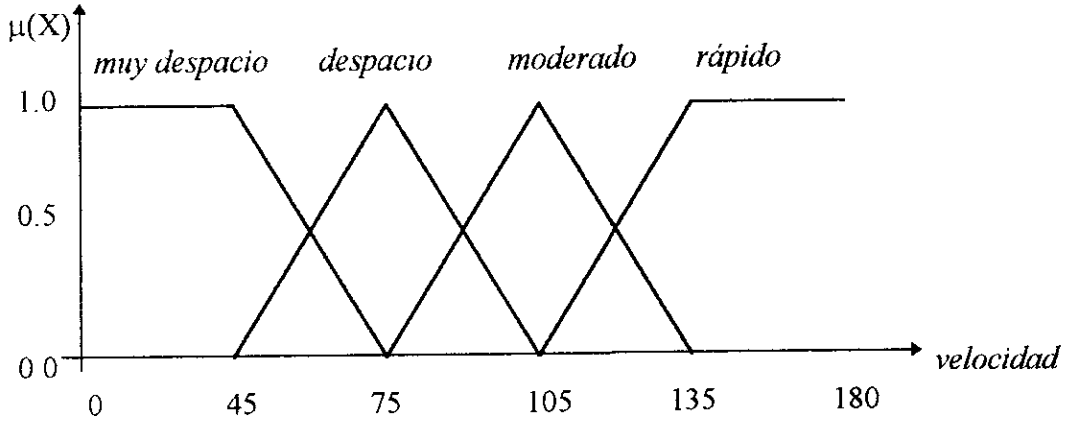


Fig. VI.27 FUNCIONES DE MEMBRESÍA PARA LA VARIABLE VELOCIDAD.

esto es:
T(*velocidad*) = { *muy despacio, despacio, moderado, rápido*}

Otros ejemplos podrían ser.

1 T(*temperatura*) = { *muy fría, fría, tibia, caliente, muy caliente*}
2 T(*altura*) = { *baja, mediana, alta, muy alta*}

En la fig VI.27 se puede apreciar que hay varios valores de velocidad que estan en más de un solo subconjunto difuso correspondiente (función de membresía). Con referencia a esto la función de compatibilidad asocia cada valor de la variable en el intervalo entre [0,1] que es el rango de las funciones miembro.

VI.11 Lógica predictiva.

En lógica predictiva clásica, una simple proposición P, es una expresión lingüística contenida dentro de un universo de proposiciones la cual puede ser identificada como siendo estrictamente verdadera o estrictamente falsa. La veracidad (verdad) de la proposición P, puede ser asignada a un valor verdadero binario, llamado T(P), justo como un elemento en un universo es asignado a una cantidad binaria para medir su membresía en un conjunto particular. Para lógica predictiva binaria (booleana), T(P) es asignado a un valor de 1 (verdadero) o 0 (falso). Si U es el universo de todas las proposiciones, entonces T es una mapeo de estas proposiciones para las cantidades binarias (0,1) o $T:U\rightarrow\{0,1\}$

Ahora P y Q serán dos proposiciones sobre el mismo universo de discurso que pueden ser combinados usando las siguientes cinco conectividades lógicas [30]:

- disyunción $\vee$
- conjunción $\wedge$
- negación -
- implicación $\rightarrow$
- igualdad $\leftrightarrow$ o $\cong$

Las cinco conectividades lógicas definidas anteriormente pueden ser usadas para crear proposiciones compuestas, donde una proposición compuesta es definida como una proposición lógica formada por dos o más proposiciónes simples conectadas logicamente.

Para una proposición P definida en el conjunto A y una proposición Q definida en el conjunto B, la implicación "P implica Q" es equivalente a tomar la unión de elementos en el complemento del conjunto A con los elementos del conjunto B.

$$(P\rightarrow Q) \cong (\bar{A}\cup B \text{ es verdadero}) \cong (\text{"no esta en A" o "en B"})$$

así que

$$T(P\rightarrow Q) = T(\bar{P}\vee Q) = \max (T(\bar{P}),T(Q))$$

Esto es lingüísticamente equivalente a la expresión "P implica Q es verdadera" cuando "no A" o "B" es verdadero. Como se observa en el diagrama de Venn de la fig. VI 28, la región representada por la diferencia A|B es el conjunto de región donde la implicación P $\rightarrow$ Q es falso. La región sombreada en la fig. VI.28 representa la colección de elementos en el universo donde la implicación es verdadera.

El área sombreada es el conjunto

$$\overline{A|B}=A\cup\bar{\bar{B}}=\overline{A\cap\bar{B}}$$

Si x esta en A y x no esta en B, entonces

$$A\rightarrow B \text{ falla} = A|B \text{ (diferencia)}$$

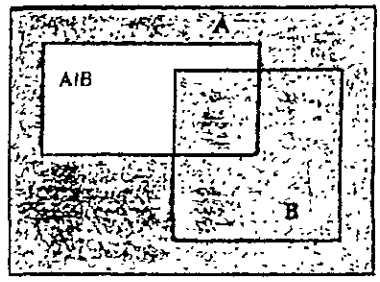


Fig. VI 28 GRÁFICA ANÁLOGA DE LA OPERACIÓN DE IMPLICACIÓN CLÁSICA; EL ÁREA SOMBREADA ES DONDE LA IMPLICACIÓN ES VERDADERA.

Suponemos que la operación de implicación involucra dos diferentes universos de discución; P es una proposición descrita por el conjunto A, el cual es definido en el universo X, y Q es una proposición descrita por el conjunto B, el cual es definido en el universo Y. Entonces la implicación $P \rightarrow Q$ puede ser representada en los términos de conjuntos teóricos por la relación R, donde R es definida por

$$\begin{aligned} &R = (AXB) \cup (\bar{A}XY) = \text{If A, Then B} \\ &\text{If } x \in A \qquad \text{donde } x \in X \text{ y } A \subset X \\ &\text{Then } y \in B \qquad \text{donde } y \in Y \text{ y } B \subset Y \end{aligned} \tag{VI.37}$$

La anterior implicación (ec. VI.37), es también equivalente a la forma de la regla lingüística: If A, Then B, La gráfica presentada en la fig. VI.29 representa el espacio cartesiano del producto XXY, presentando los conjuntos típicos A y B; y superpuestos sobre este espacio es el conjunto teórico equivalente de la implicación. Esto es,

$$P \rightarrow Q\text{: If } x \in A, \text{ Then } y \in B \qquad \text{o} \qquad P \rightarrow Q \cong A \cup B$$

las regiones sombreadas del diagrama de Venn compuesto en la fig. VI.29 representan el dominio verdadero de la implicación If A, Then B ($P \rightarrow Q$)

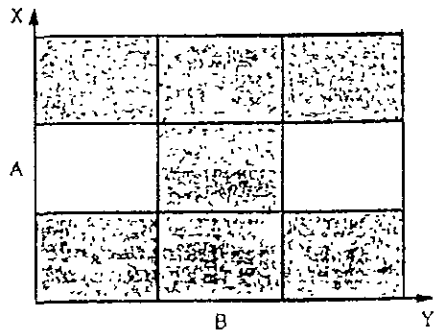


Fig VI.29 ESPACIO CARTESIANO PRESENTADO PARA LA IMPLICACIÓN If A, Then B

Otra proposición compuesta en reglas lingüísticas es la expresión

If A, Then B, Else C.

lingüísticamente, esta proposición compuesta puede ser expresada como

If A, Then B, o If $\bar{A}$, Then C.

en lógica predictiva esta regla tiene la forma

$$(P \rightarrow Q) \vee (\bar{P} \rightarrow S)$$

donde: P: $x \in A$, $A \subset X$
Q. $y \in B$, $B \subset Y$
S: $y \in C$, $C \subset Y$

El equivalente conjunto teórico de esta proposición compuesta es dado por

$$\text{If A, Then B, Else C} = (AXB) \cup (\bar{A}XC) = R = \text{relación sobre X X Y}$$

La gráfica de la fig. VI.30 ilustra la región sombreada representando el dominio verdadero de esta proposición compuesta para el caso particular donde $B \cap C = \varnothing$

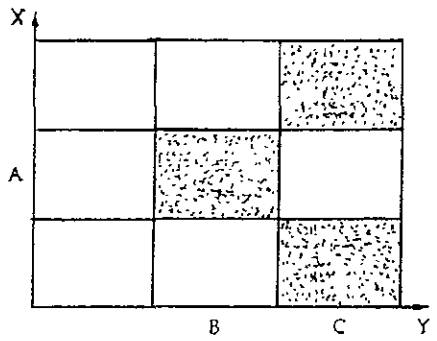


Fig. VI.30 DOMINIO VERADERO PARA If A, Then B, Else C.

VI 11 1 Tautología.

En lógica predictiva es útil considerar proposiciones compuestas que son siempre verdaderas, sin tener en cuenta los valores verdaderos de las proposiciones simples individuales. En lógica clásica las proposiciones compuestas con esta propiedad son llamadas "tautologías". Las tautologías son útiles para razonamiento deductivo y para hacer inferencias deductivas. Así, si una proposición compuesta puede ser expresada en la

forma de tautología, el valor verdadero de aquella proposición compuesta es conocida como verdadera.

Una tautología conocida como deducción *modus ponens*, es un esquema de inferencia muy común usado en sistemas expertos delanteros. Esta es una operación cuya tarea es encontrar el valor verdadero de una consecuencia en una regla de producción, dando el valor verdadero del antecedente en la regla. La deducción *modus ponens* concluye que, dadas dos proposiciones P y P→Q y ambas son verdaderas, entonces la verdad de la proposición simple Q es automáticamente inferida. Otra tautología útil es la inferencia *modus tollens*, la cual es usada en sistemas expertos retrasados. En *modus tollens* una implicación entre dos proposiciones es combinada con una segunda proposición y ambas son usadas para implicar una tercera proposición. Algunas tautologías comunes son mencionadas a continuación [25]:

$$B \cup B \leftrightarrow X$$

$$A \cup X; \qquad A \cup X \leftrightarrow X$$

$$A \leftrightarrow B$$

$$(A \wedge (A \rightarrow B)) \rightarrow B \qquad (\textit{modus ponens}) \qquad \text{(VI.38)}$$

$$(B \wedge (A \rightarrow B)) \rightarrow A \qquad (\textit{modus tollens}) \qquad \text{(VI.39)}$$

Una simple prueba del valor verdadero de la deducción *modus ponens* es dada a continuación con varias propiedades para cada uno de los pasos, con el propósito de ilustrar la utilidad de una tautología en razonamiento clásico:

$(A \wedge (A \rightarrow B)) \rightarrow B$	
$(A \wedge (A \cup B)) \rightarrow B$	Implicación
$(A \wedge A) \cup (A \wedge B)) \rightarrow B$	Distributividad
$(\varnothing \cup (A \wedge B)) \rightarrow B$	Leyes de media exclusión
$(A \wedge B) \rightarrow B$	Identidad
$(A \wedge B) \cup B$	Implicación
$(A \vee B) \cup B$	Leyes de morgan
$A \vee (B \cup B)$	Asociatividad
$A \cup X$	Leyes de media exclusión
$X \Rightarrow T(X)=1$	Identidad

Similarmente una prueba del valor verdadero de la inferencia *modus tollens* es listada a continuación

$(\bar{\bar{B}} \wedge (A \rightarrow B)) \rightarrow \bar{A}$	
$(\bar{\bar{B}} \wedge (\bar{A} \cup B)) \rightarrow \bar{A}$	Implicación
$(\bar{\bar{B}} \wedge \bar{A}) \cup (\bar{\bar{B}} \wedge \bar{\bar{B}})) \rightarrow \bar{A}$	Distributividad
$((\bar{\bar{B}} \wedge \bar{A}) \cup \varnothing) \rightarrow \bar{A}$	Leyes de media exclusión
$(\bar{\bar{B}} \wedge \bar{A}) \rightarrow \bar{A}$	Identidad
$(\bar{\bar{B}} \wedge \bar{\bar{A}}) \cup \bar{A}$	Implicación
$(\bar{\bar{B}} \vee \bar{\bar{A}}) \cup \bar{A}$	Leyes de morgan
$B \cup (A \cup \bar{A})$	Asociatividad
$B \cup X$	Leyes de media exclusión
$X \Rightarrow T(X)=1$	Identidad

VI.12 Defusificación.

Existen situaciones donde la salida de un proceso difuso necesita ser una cantidad escalar única opuestamente a un conjunto difuso. La defusificación es la conversión de una cantidad difusa a una cantidad precisa, justamente como la fusificación es la conversión de una cantidad precisa a una cantidad difusa. La salida de un proceso difuso puede ser la unión lógica de dos o más funciones de membresía difusa definidas sobre el universo de discusión de la variable de salida. Por ejemplo, suponiendo que una salida difusa esta compuesta de dos partes: la primera parte $\underline{C}_1$, una forma trapezoidal, como se presenta en la fig VI.31a y la segunda parte $\underline{C}_2$, una forma triangular, presentada en la fig. VI.31b. La unión de estas dos funciones de membresía, esto es, $\underline{C}=\underline{C}_1 \cup \underline{C}_2$, involucra el operador máximo, el cual gráficamente es la envolvente exterior de las dos formas presentadas en las figs. VI.31a y VI.31b; la forma resultante es presentada en la fig. VI.31c. De acuerdo a un proceso de salida difusa general se pueden involucrar muchas partes de salida (más de dos) y la función de membresía representa cada parte de la salida y puede tener formas trapezoidales y triangulares. Además, como se presenta en la fig. VI.31a las funciones de membresía no siempre son normales.

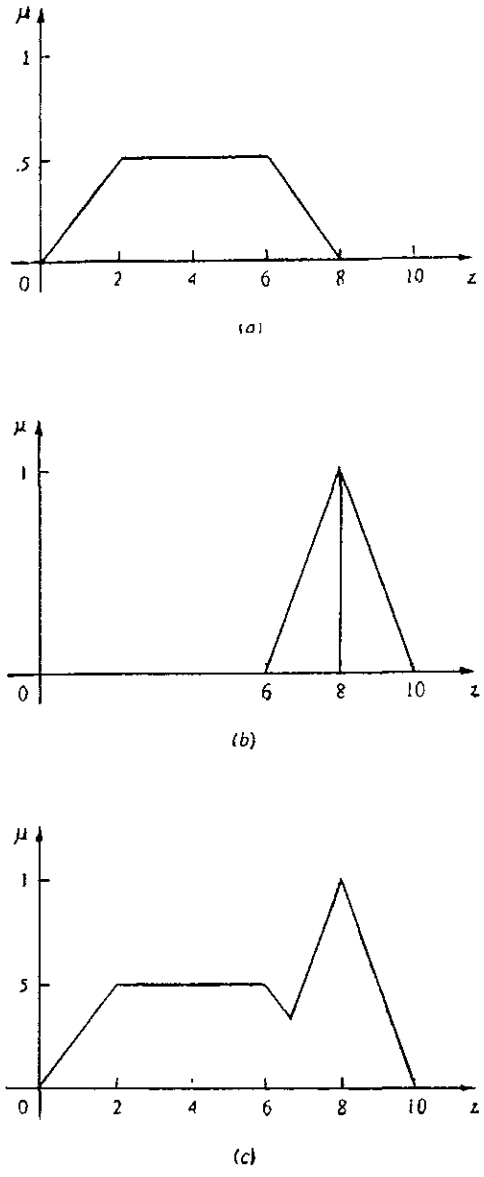


Fig VI.31 TIPICO PROCESO DE SALIDA DIFUSA. (a) PRIMERA PARTE DE SALIDA DIFUSA; (b) SEGUNDA PARTE DE SALIDA DIFUSA; (c) UNIÓN DE LAS DOS PRIMERAS PARTES.

Los métodos para defusificar funciones de salida difusa (funciones de membresía), que se mencionan a continuación son siete [30]:

1 *Principio de máxima membresía:* También conocido como el *método de altura*, este esquema es limitado a funciones de salida pico. Este método es dado por la siguiente expresión algebraica, y es presentado gráficamente en la fig. VI.32

$$\mu_{\underline{C}}(z^*) \geq \mu_{\underline{C}}(z) \qquad \text{para toda } z \in Z \qquad \text{(VI.40)}$$

siendo z* un valor preciso que se encuentra comprendido en el universo de la variable Z.

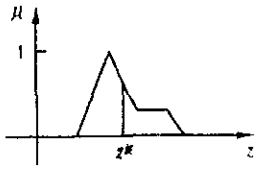


Fig. VI.33 MÉTODO DE DEFUSIFICACIÓN DEL CENTROIDE.

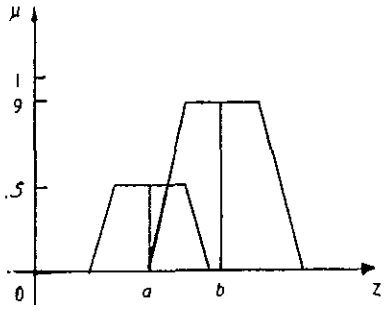


Fig. VI.34 MÉTODO DE DEFUSIFICACIÓN DE PROMEDIO PONDERADO.

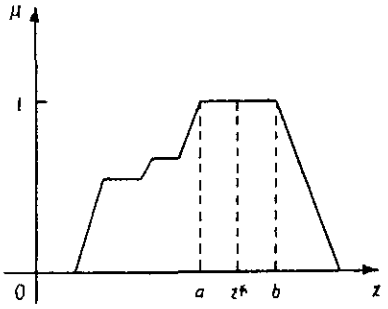


Fig. VI.35 MÉTODO DE DEFUSIFICACIÓN DE MAXIMO PROMEDIO DE MEMBRESÍA.

5. *Método de sumas de centros*: Este método es más rápido que muchos de los métodos de defusificación que existen en uso Este proceso involucra la suma algebraica de conjuntos difusos individuales de salida, es decir de $\underline{C}_1$ y $\underline{C}_2$ en lugar de su unión. Una desventaja de este método es que las áreas intersectadas son dos veces sumada. El valor z* defusificado es dado por la siguiente ecuación:

$$z^* = [\int_z z \sum^n_{k=1} \mu_{\underline{C}k}(z)\, dz] / [\int_z \sum^n_{k=1} \mu_{Ck}(z)\, dz] \qquad \text{(VI.44)}$$

Este método es similar al método de promedio ponderado de la ec. (VI.42) excepto que en el método de sumas de centros los pesos son las áreas de las funciónes de membresía respectivas, mientras que en el método de promedio ponderado son valores de membresía individuales. La fig. VI.36 es una ilustración del método sumas de centros.

6. *Centro del área más grande*: Si el conjunto de salida difusa tiene al menos dos subregiones convexas, entonces el centro de gravedad (esto es, z*, es calculada usando el método del centroide de la ec. VI.41) de la subregión difusa convexa con el área más grande es usada para obtener el valor z* defusificado de la salida, como se muestra en la fig. VI.37 y se obtiene algebraicamente como:

$$z^* = [\int \mu_{\underline{C}m}(z)\, z dz] / [\int \mu_{\underline{C}m}(z)\, dz] \qquad \text{(VI.45)}$$

donde $\underline{C}_m$ es la subregión convexa que tiene el área más grande, mostrandola como $\underline{C}_k$. Esta condición se aplica en el caso donde la salida general $\underline{C}_k$ es no convexa: y en el caso donde $\underline{C}_k$ es convexa, z* es la misma cantidad que se determina con el método del centroide o el método del centro del área más grande (porque entonces existe solo una región convexa).

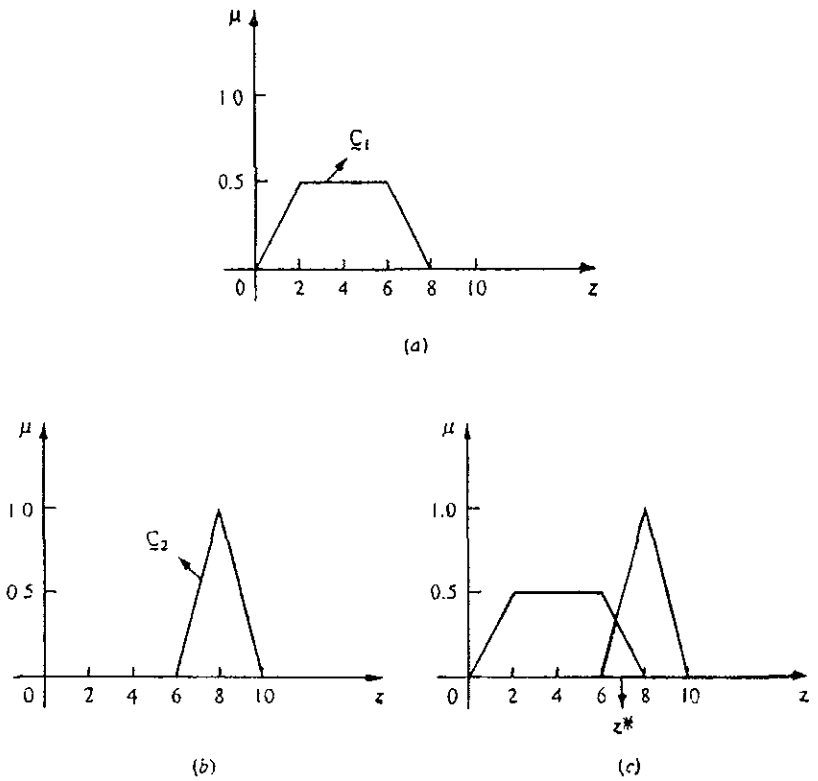


Fig. VI.36 MÉTODO SUMAS DE CENTROS (a) PRIMERA FUNCIÓN DE MEMBRERSÍA; (b) SEGUNDA FUNCIÓN MEMBRESÍA; (c) PASO DE DEFUSIFICACIÓN.

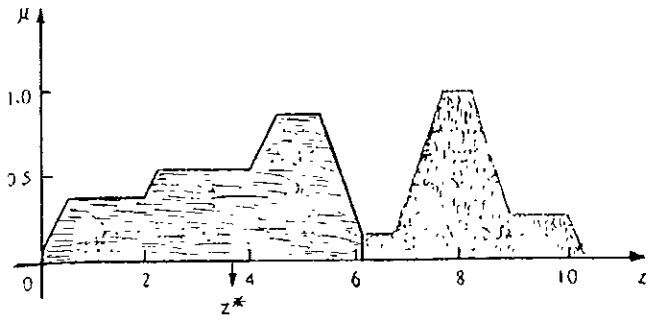


Fig. VI.37 MÉTODO DEL CENTRO DEL ÁREA MÁS GRANDE (BOSQUEJADA CON LÍNEA GRUESA). SE MUESTRA PARA $\underline{C}_k'$ NO CONVEXA.

7. *Método primero o último máximo*: Este método usa la salida general o unión de todas las salidas individuales de los conjuntos difusos $\underline{C}_k$ para determinar el valor más pequeño del dominio con un grado de membresía o pertenencia maximizado en $\underline{C}_k$. Las ecuaciones para z* son·

Primero se determina la altura más alta de la unión [denotada como $hgt(\underline{C}_k)$]

$$hgt(\underline{C}_k) = \sup_{z \in Z} \mu_{\underline{C}k}(z) \qquad \text{(VI.46)}$$

entonces el primer máximo es encontrado,

$$z^* = \inf_{z \in Z} \{ z \in Z \mid \mu_{Ck}(z) = hgt(C_k') \} \qquad \text{(VI.47)}$$

Una alternativa para este método es llamado el último máximo, y es dado por:

$$z^* = \sup_{z \in Z} \{ z \in Z \mid \mu_{\underline{C}k}(z) = hgt(\underline{C}_k) \} \qquad \text{(VI.48)}$$

En las ecuaciones VI.46, VI.47 y VI 48 el superior (sup) es el último límite superior y el inferior (inf) es el más grande límite inferior. Gráficamente, este método es presentado en la fig VI.38, donde el primer máximo es también el último máximo y, porque esto es un máximo distinto, es también el promedio máximo. Por lo tanto, los métodos presentados en las ecuaciones, VI 40 (máxima membresía), VI.43 (máximo promedio), VI.47 (primero-máximo), y VI.48 (último-máximo), todos ellos proveen el mismo valor defusificado z*, para la situación particular mostrada en la fig. VI.38

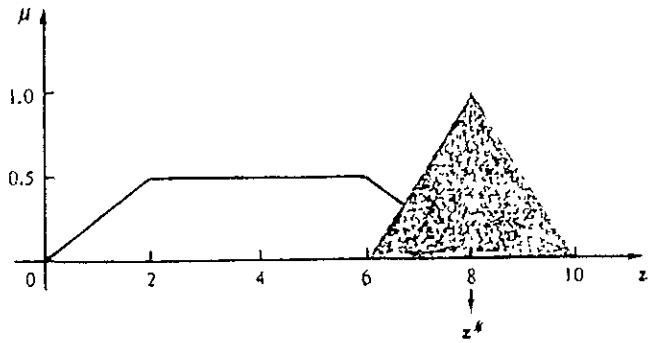


Fig. VI.38 MÉTODO PRIMERO Y ÚLTIMO MÁXIMO.

VI.13 Conclusiones.

La teoría difusa es una teoría matemática, y es llamada confusa o borrosa (*fuzziness*) tomada del aspecto de incertidumbre. La confusión es la ambigüedad que puede ser encontrada en la definición del concepto o significado de una palabra. Por ejemplo la incertidumbre en expresiones como "persona vieja", "temperatura alta", "número pequeño" pueden ser llamados fuzziness (confusiones).

La teoría de sistemas difusos es el punto de inicio para desarrollar modelos ambiguos para procesos de pensamientos y juicios. Tomando en cuenta que es ideal para representaciones intuitivas ya que simula la experiencia y los conocimientos de un operador humano, luego entonces es apropiada para controlar procesos cuyo modelo matemático es muy complejo y/o el modelo del controlador también lo es.

La teoría de conjuntos difusos fue expandida a áreas tales como lógica y medidas para teoría de sistemas, y se desarrollaron metodologías para aplicaciones tales como modelado, evaluación, optimización, toma de decisiones, control y diagnóstico. También existen aplicaciones tales como control, inteligencia artificial y reconocimiento de patrones.

En el campo de reconocimiento de patrones la herramienta que se usa es rígida para el manejo de los patrones, los cuales son usualmente distorsionados o ruidosos aún reteniendo una estructura implícita. Cuando la indeterminación de los patrones es debida a la vaguedad en lugar de la aleatoriedad, la teoría difusa puede ser una mejor herramienta para describir la mala definición de la información estructural.

Capítulo VII

RECONOCIMIENTO DE VOZ UTILIZANDO LÓGICA DIFUSA.

VII.1. Introducción.

La Lógica Difusa está expandiendose rapidamente en muchas regiones del Mundo, con una infinita gama de aplicaciones en diversas actividades de la vida humana. Desde un sistema automático del Metro en Japón hasta un sistema de transmisión automática de vehiculo en los Estados Unidos, la Lógica Difusa está transformando el diseño de los productos, muchos productos comerciales que utilizan esta técnica emplean poco menos de 20 reglas. Los expertos en la industria creen que esta disciplina llegará a ser un negocio de millones de dólares, jugando un papel importante en la implementación de la industria [22].

A pesar de su nombre paradójico, la Lógica Difusa proveé una solución práctica y económica para controlar sistemas indefinidos o complejos. La Lógica Difusa es una forma de trabajo flexible para resolver muchos tipos de problemas de control y problemas de reconocimiento de patrones entre otros. En el mundo real del reconocimiento de problemas de clasificación hay que enfrentarse a confusiones (fuzziness) en relación con etiquetas expresadas en un espacio característico, además de etiquetas de clases tomadas en cuenta en el procedimiento de clasificación.

En este capítulo se hace uso de los fundamentos teóricos de las técnicas de: Codificación por Predicción Lineal (LPC), Cuantización Vectorial (CV), y Lógica Difusa para implementar un Reconocedor de Voz. En este Sistema Reconocedor se encuentran incluidas herramientas tales como: Deformación de Tiempo Dinámico para la comparación de patrones, Medida de Distancia de Itakura, Medida de Distancia por medio de Coeficientes Cepstrales, Ventaneo, Potencia de la señal, Teoría de Lógica Difusa, Traslape entre Bloques, Obtención de Centroides y Evaluación del Vecino más Cercano entre otras herramientas que fueron brevemente tratadas anteriormente. En este capítulo sólo se mencionan y serán manejados los resultados obtenidos de algunas de estas técnicas.

VII.2 Descripción General del Sistema.

El objetivo de este sistema es reconocer doce comandos de voz siendo las palabras: adelante, alto, atrás, derecha, dos, izquierda, lento, no, rápido, si, tres y uno.

El sistema Reconocedor de Voz esta constituido de tres entradas (o universos) que son las salidas de tres modos que procesan previamente la palabra pronunciada, siendo estas: la salida del Reconocedor de Voz con la técnica LPC, la salida del Reconocedor de Voz con la técnica de CV y la potencia de la señal de entrada pronunciada. Estos tres datos son los parámetros que determinarán el reconocimiento de voz en el sistema difuso en base a las reglas de inferencia, así como también formas y rangos de los conjuntos difusos de cada entrada

Otras variables de entrada pudieron haber sido seleccionadas para alimentar al sistema difuso y en base a éstas, se formarían los universos de entrada y sus respectivas funciones de membresía, que podrían mejorar o simplificar este enfoque.

VII.2 1 Implementación del Sistema Difuso.

a) Determinación de las entradas y salidas difusas.

El sistema difuso tiene tres variables de entrada. Un sistema difuso resuelve de una manera simple lo que es un problema complejo aprovechando los resultados de experimentos anteriores que, para este caso, son las entradas al sistema difuso:

- Los resultados obtenidos del Reconocedor de Voz empleando la técnica LPC.
- Los resultados obtenidos del Reconocedor de Voz aprovechando la técnica de CV.
- La potencia de la señal que es la entrada del sistema.

Siendo la salida la variable que determina la palabra que fue pronunciada.

b) Organización de Conjuntos difusos de entrada y salida.

Las entradas y salidas (o universos de entrada y salida) son representadas por un adjetivo que las describe y son llamados conjuntos difusos.

- El primer universo de entrada es nombrado *distancia LPC*, que es la mínima distancia obtenida del Reconocedor de Voz empleando la técnica LPC.
- El segundo universo de entrada es nombrado *distancia CV*, que es la mínima distancia obtenida del Reconocedor de Voz aprovechando la técnica CV.
- El tercer universo de entrada es nombrado *potencia*, que es la potencia de la señal de entrada.
- El universo de salida es nombrado *reconoce*, que es la palabra reconocida del comando pronunciado

La fusificación u organización de conjuntos difusos de entradas y salidas es el proceso de convertir una cantidad precisa en una cantidad difusa. Esto se hace sabiendo que muchas de las cantidades que se consideran precisas y resueltas no lo son del todo. Si la incertidumbre de estas cantidades se incrementa a causa de imprecisión o vaguedad, entonces la variable es difusa y puede ser representada por una función de membresía.
Se asigna una forma trapezoidal a cada uno de los conjuntos difusos (o funciones de membresía) de cada universo de entrada. Esta forma es algo arbitraria, también se pudo haber escogido curvas exponenciales, campana o formas triangulares. La ventaja de elegir una forma trapezoidal para las funciones de membresía de los tres universos de entrada, es debida a que, en un cierto rango del valor numérico de las variables de entrada, los conjuntos difusos de cada universo tienen un rango de pertenencia constante de uno, y dos más lineales entre "0" y "1", lo que con ninguna otra forma de conjunto se podría lograr.

Hay que notar que las funciones de membresía de cada universo tienen una forma trapezoidal, según el sentido de pensamiento para el comando que corresponde al conjunto difuso. Se usa el sentido común para asignar el rango de cada universo.
Una forma triangular es tomada para las funciones de membresía del universo de salida, ya que con esta forma solo se tiene un valor numérico de la variable de salida con grado de pertenencia máximo (uno)

c) Determinación de reglas difusas.

Las reglas difusas son las fuerzas que guían al sistema. Estas son para el sistema difuso lo que es una función de transferencia para cualquier otro sistema. Estas son implementadas como una serie de expresiones IF/THEN. Las diferentes combinaciones de entrada nos dan las diferentes posibles salidas.
Por ejemplo las reglas usadas en un sistema de prevención de choques de un móvil serían como esta: "*IF la distancia es corta AND la aceleración es rápida THEN el nivel de frenado es alto*". Esto es escrito en un lenguaje de sentido común, no existen complicadas ecuaciones. Estas son expresiones simples donde se usa la intuición para escribirlas.
Algunos sistemas pueden usar una red neuronal para desarrollar éstas expresiones. Otros sistemas pueden recurrir a prueba y error. Todos estos métodos son usualmente menos caros que implementar un sistema no difuso. Los sistemas difusos son más fáciles de entender que uno no difuso.

d) Defusificación de las salidas.

La defusificación es la conversión de una cantidad difusa a una cantidad precisa. La salida de un proceso difuso puede ser la unión lógica de dos o más funciones de membresía definidas sobre el universo de discusión de la variable de salida.
Existen diferentes maneras en las que un sistema puede evaluar la exactitud difusa. Una manera es tomar el elemento de salida con la mayor parte de verdad, es decir, la salida que tenga la mayor altura de verdad o bien, la salida que esté en "1" o más cercana a "1", que es el método usado en este caso.
La Lógica Difusa remueve la complejidad de un sistema. Frecuentemente se sabe como se comporta un sistema pero no es tan fácil describirlo por ecuaciones. Sistemas como estos son ideales para ser implementados como un sistema difuso.

VII.3 Detalles del Sistema.

A continuación se muestran los lineamientos del sistema difuso en base a las reglas de inferencia, así como también formas y rangos de los conjuntos difusos de cada entrada.
En la fig VII.1 se muestran los tres universos de entrada con sus funciones de membresía (forma trapezoidal), así como también el universo de salida y sus conjuntos difusos (forma

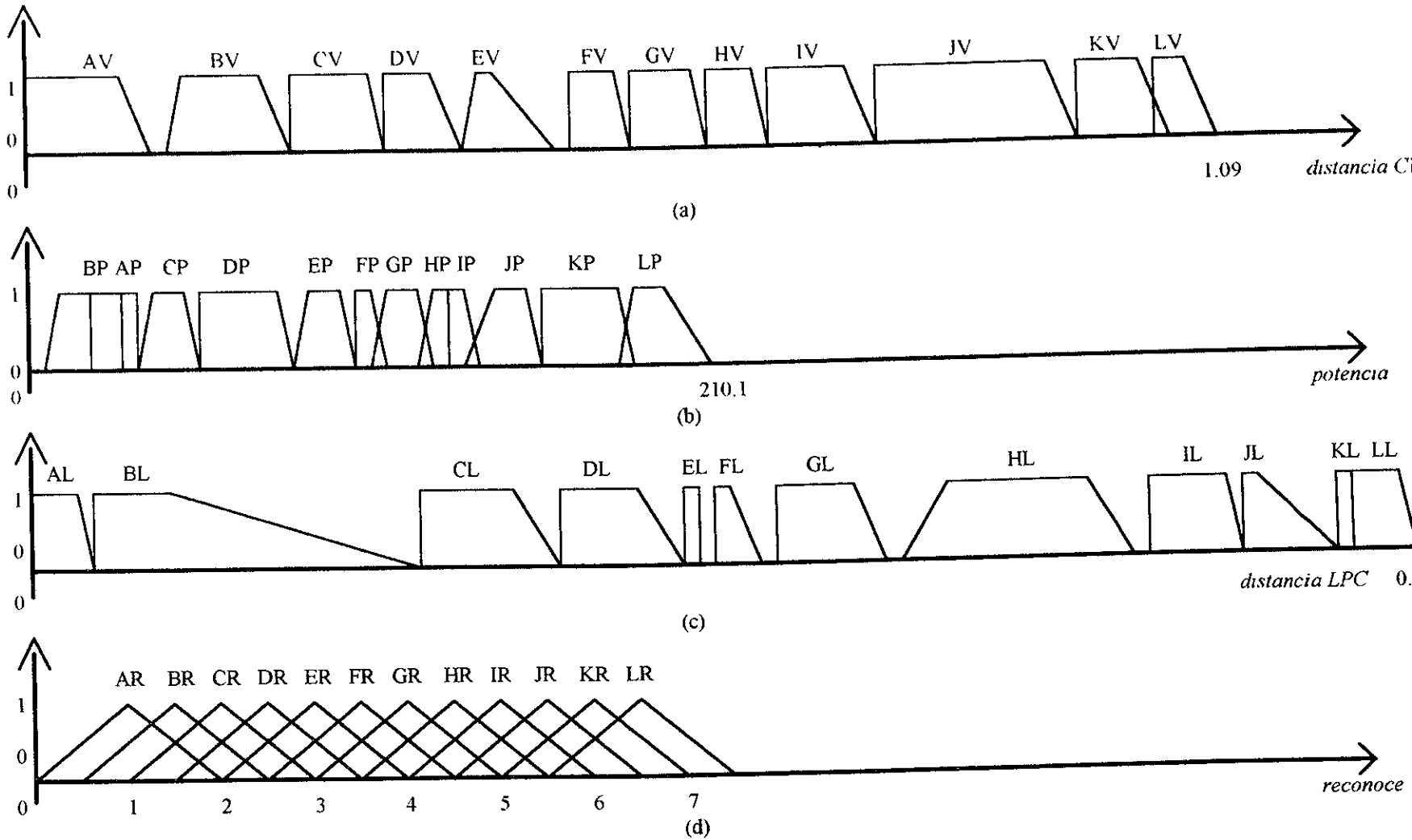


Fig. VII.1 CONJUNTOS DIFUSOS (a) ENTRADA *distancia CV* (b) ENTRADA *potencia* (c) ENTRADA *distancia LPC* (d) SALIDA *reconoce*.

triangular). Como se puede observar, de los tres universos de entrada existen regiones de las variables correspondientes donde:

- Se traslapan las funciones de membresía, es decir, regiones que pertenecen a más de un conjunto difuso, con distinto grado de pertenencia para cada uno de los conjuntos, a los que pertenece dicha región.
- Hay intervalos donde las funciones de membresía no se traslapan, esto indica que, esa región sólo pertenece a un conjunto difuso y que el grado de pertenencia cero donde finaliza una función de membresía coincide con el grado de pertenencia cero donde inicia otra función de membresía.
- También hay intervalos que no pertenecen a ningún conjunto difuso (tramos del rango de la variable) que separan las funciones de membresía.

Para la variable de salida, una vez marcado el rango de valores del universo de salida, se procede a determinar los intervalos para cada palabra del vocabulario, es decir, para sus correspondientes funciones de membresía. El traslape entre estos conjuntos difusos muestra que no existen valores que pertenezcan exclusivamente a una función de membresía. Cualquier valor numérico comprendido en el rango de valores de salida pertenece a una o a varias palabras en distintos grados de pertenencia, con esto se asegura la elección de una palabra reconocida.

a) Generación de reglas de inferencia.

Las reglas de inferencia son generadas en base a la forma en que se desea que se maneje el comportamiento de la salida, además de los experimentos obtenidos con respecto al análisis de la variable de salida. A continuación se muestra la estructura de las reglas de inferencia de este sistema (fig VII.2).

Es importante recordar que el sistema difuso se compone de tres universos de entrada (*distancia CV, potencia, distancia LPC*) y un universo de salida (*reconoce*). Cada universo consta de 12 funciones de membresía que representan los datos difusos de las 12 palabras del vocabulario:

- Para el universo de entrada *distancia CV*: AC, BC, CC, DC, EC, FC, GC, HC, IC, JC, KC, LC.
- Para el universo de entrada *distancia LPC*: AL, BL, CL, DL, EL, FL, GL, HL, IL, JL, KL, LL.
- Para el universo de entrada *potencia*: AP, BP, CP, DP, EP, FP, GP, HP, IP, JP, KP, LP.
- Para el universo de salida *reconoce*: AR, BR, CR, DR, ER, FR, GR, HR, IR, JR, KR, LR.

La fig. VII.2a muestra la matriz de reglas de inferencia de dos entradas (*distancia CV* y *potencia*) dando la siguiente salida temporal: AT, BT, CT, DT, ET, FT, GT, HT, IT, JT, KT, LT. La fig. VII.2b muestra otra matriz de reglas de inferencia con la restante entrada (*distancia LPC*) y la salida temporal de la primera matriz, dando como resultado la salida final.

distancia CV

AC	BC	CC	DC	EC	FC	GC	HC	IC	JC	KC	LC
AT											
	BT										
		CT									
			DT								
				ET							
					FT						
						GT					
							HT				
								IT			
									JT		
										KT	
											LT

AP	
BP	
CP	*p*
DP	*o*
EP	*t*
FP	*e*
GP	*n*
HP	*c*
IP	*i*
JP	*a*
KP	
LP	

(a)

distancia LPC

AL	BL	CL	DL	EL	FL	GL	HL	IL	JL	KL	LL
AR											
	BR										
		CR									
			DR								
				ER							
					FR						
						GR					
							HR				
								IR			
									JR		
										KR	
											LR

AT	
BT	
CT	*R*
DT	*e*
ET	*c.*
FT	
GT	*Tem*
HT	*po*
IT	*ral*
JT	
KT	
LT	

(b)

Fig. VII.2 REGION DE REGLAS DE INFERENCIA. (*a*)DOS ENTRADAS Y UNA SALIDA TEMPORAL (b) LA ENTRADA RESTANTE Y LA SALIDA TEMPORAL ANTERIOR DAN LA SALIDA FINAL.

b) Determinación de las alturas en base a la fusificación de las entradas.

El método a seguir se basa en utilizar el criterio de mínimos, el cual indica que se escoja el mínimo valor de las alturas correspondientes al valor numérico de las variables de entrada. Esto es, con el valor numérico de las variables de entrada intersectamos a que altura corresponde o a que alturas corresponden dicho o dichos valores numéricos (en el caso de que toque una zona con traslape). Por último, dichas alturas se comparan entre sí, (las correspondientes alturas de una variable con las alturas correspondientes de la otra variable)

c) Determinación del valor numérico de la salida en base a la defusificación de las alturas de las entradas.

Identificadas las alturas mínimas, se procede a determinar, de acuerdo a las reglas de inferencia a qué palabra corresponde la variable de salida. Asignando la altura determinada en el punto anterior al respectivo trapecio de la palabra de entrada. De estas alturas asignadas a los triángulos de salida, se trabaja con los mínimos porque corresponden a las intersecciones de los conjuntos difusos. Con éstos se determinará el centroide, que no será otra cosa que el valor numérico de la variable *reconoce* del algoritmo.

VII.4. Algoritmo para reconocimiento de voz.

A continuación se muestra el diagrama de bloques para un sistema de Lógica Difusa aplicado a Reconocimiento de Voz. El sistema consta de tres etapas previas, en cada una de ellas se procesa la palabra pronunciada (palabra que se desea reconocer) por separado, y se obtienen tres resultados o datos, uno por cada etapa. Los tres datos son las entradas del Sistema Reconocedor de Voz Difuso el cual tiene sólo una salida (fig. VII.3).

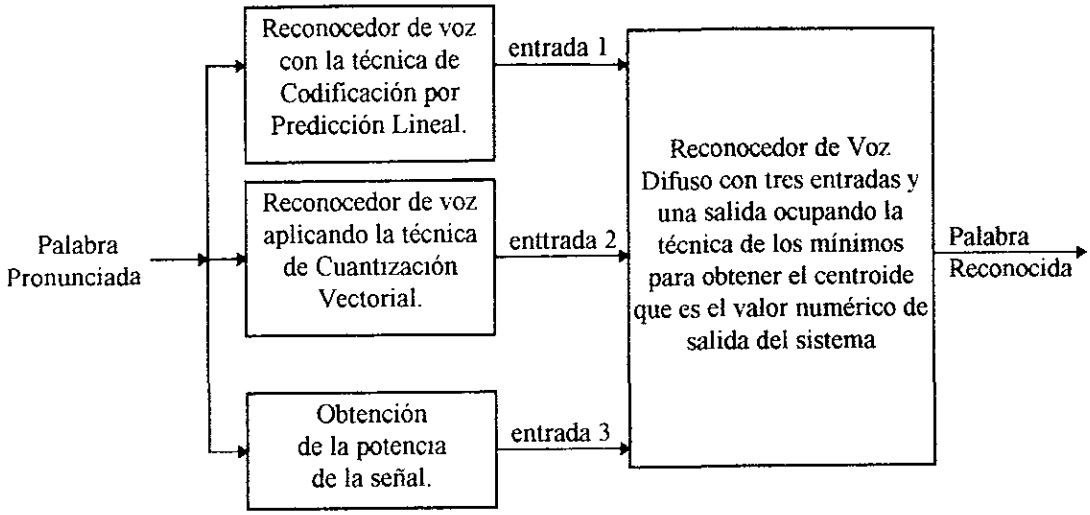


Fig. VII.3 DIAGRAMA DE BLOQUES PARA EL SISTEMA RECONOCEDOR DE VOZ UTILIZANDO LÓGICA DIFUSA

El sistema difuso es implementado en el ambiente de programación gráfico MATLAB, en la fig. VII.4 se presenta la pantalla de las variables (tres entradas y una salida) del sistema. También se muestran en la fig VII.5 a la fig. VII.8 las pantallas de las funciones de membresía de las cuatro variables.

VII.5 Conclusiones.

El Reconocedor de Voz Difuso es un sistema simple que resuelve de una manera sencilla lo que es un problema complejo. Un sistema difuso aplicado al Reconocimiento de Voz en sí, no requiere directamente de técnicas como ventaneo, traslape entre bloques, coeficientes LPC, medidas de distancia, coeficientes de autocorrelación y demás procedimientos Sólo requiere de:

- Determinar las entradas y salidas difusas
- Fusificación de entradas y salidas (organización de conjuntos difusos de entrada y salida).
- Determinación de reglas de inferencia.
- Defusificación de la salida.

El sistema difuso requiere de los resultados de experimentos anteriores que normalmente son obtenidos de numerosos ensayos, ya sea que cuenten o no con un modelo matemático para obtener los resultados. La asignación del rango de valores en las funciónes de membresía de la(s) variable(s) de entrada(s) o salida(s) depende de la intuición, de datos dados por eventos anteriores, o de métodos alternos de reconocimiento.

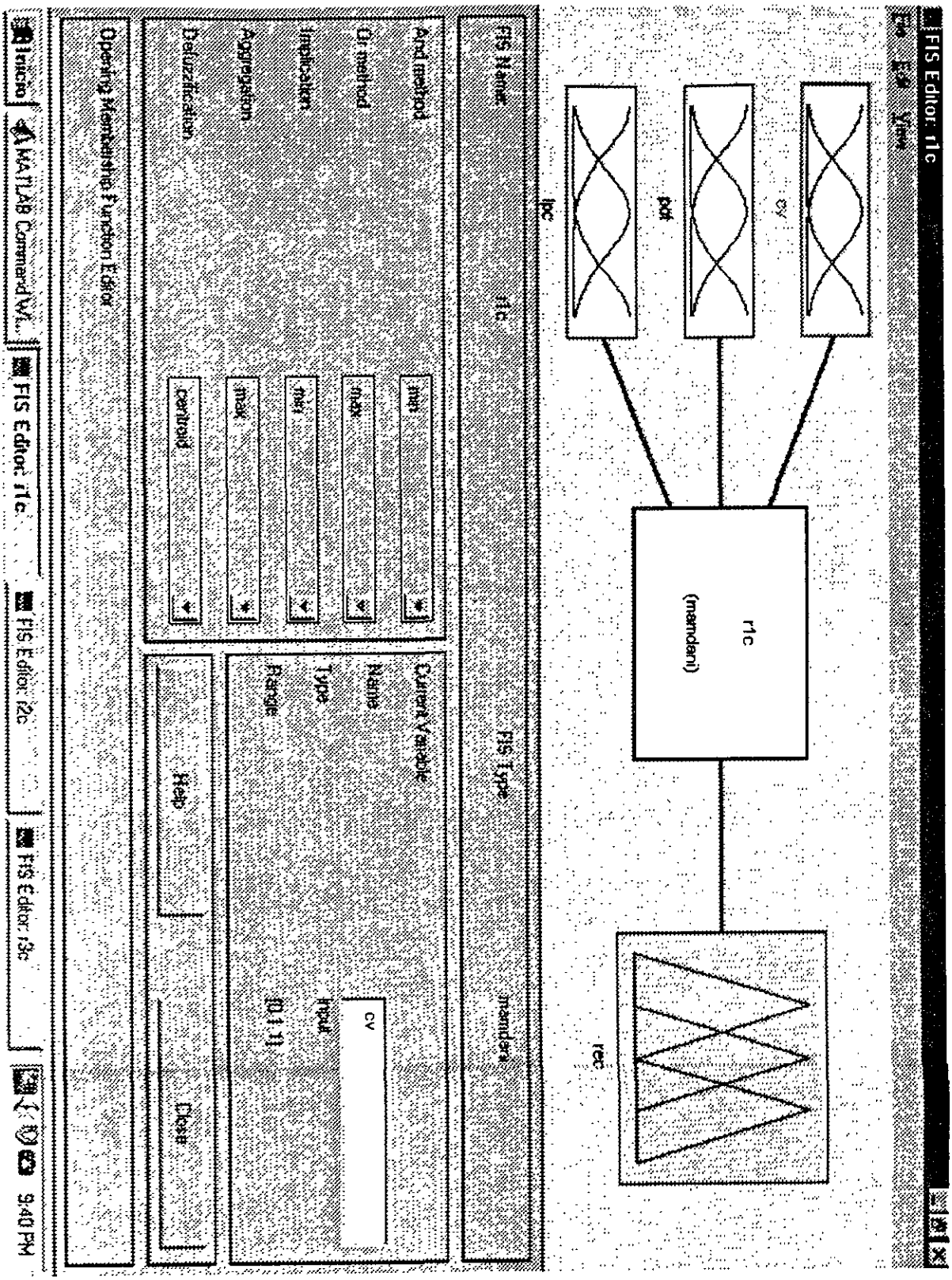


Fig VII 4 LA PANTALLA MUESTRA LAS VARIABLES DE ENTRADA Y SALIDA DEL RECONOCEDOR DIFUSO.

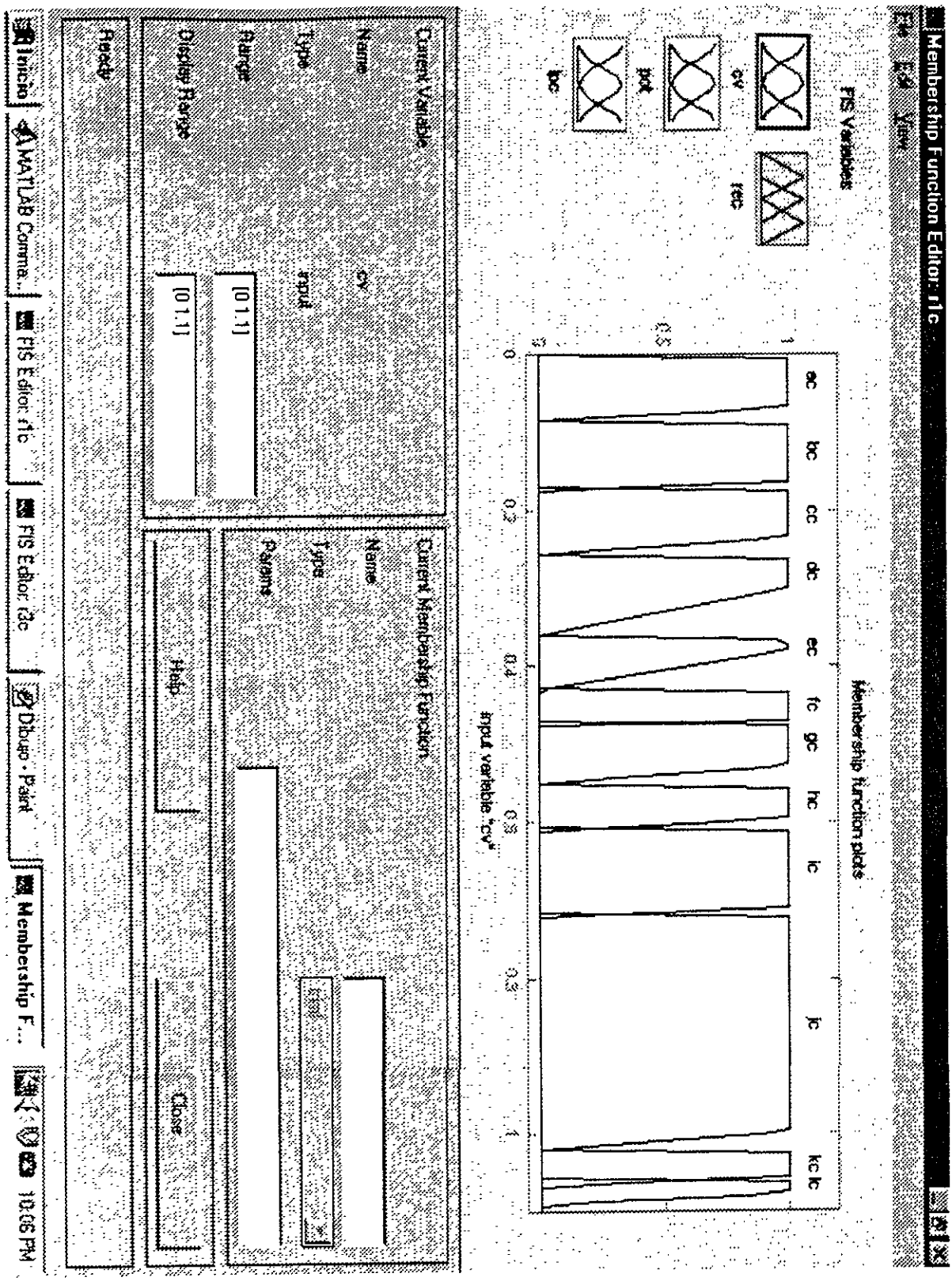


Fig. VII.5 LA PANTALLA MUESTRA LAS FUNCIONES DE MEMBRESÍA DE LA VARIABLE DE ENTRADA *distancia* CV

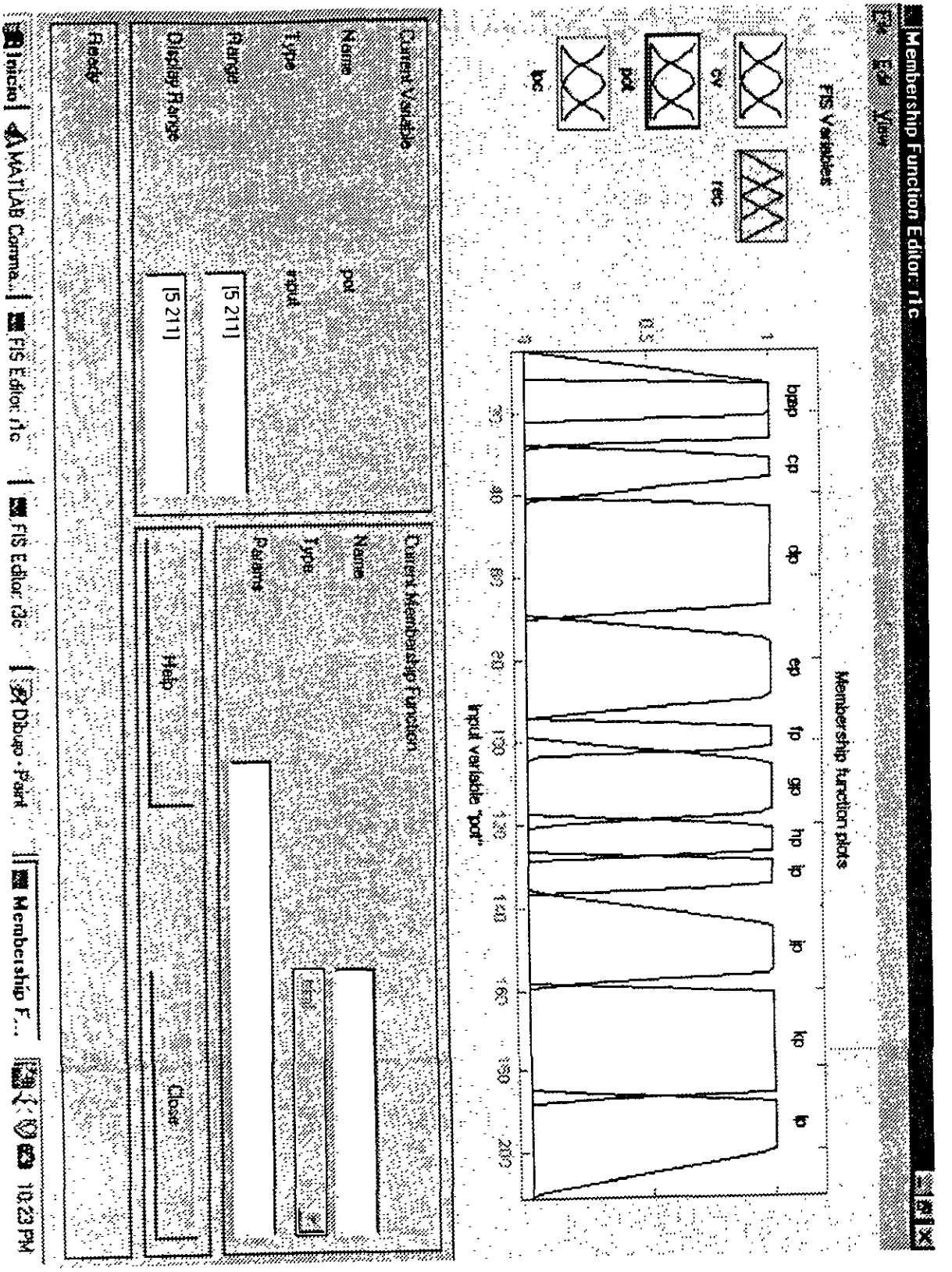


Fig. VII.6 LA PANTALLA MUESTRA LAS FUNCIONES DE MEMBRESÍA DE LA VARIABLE DE ENTRADA *potencia*

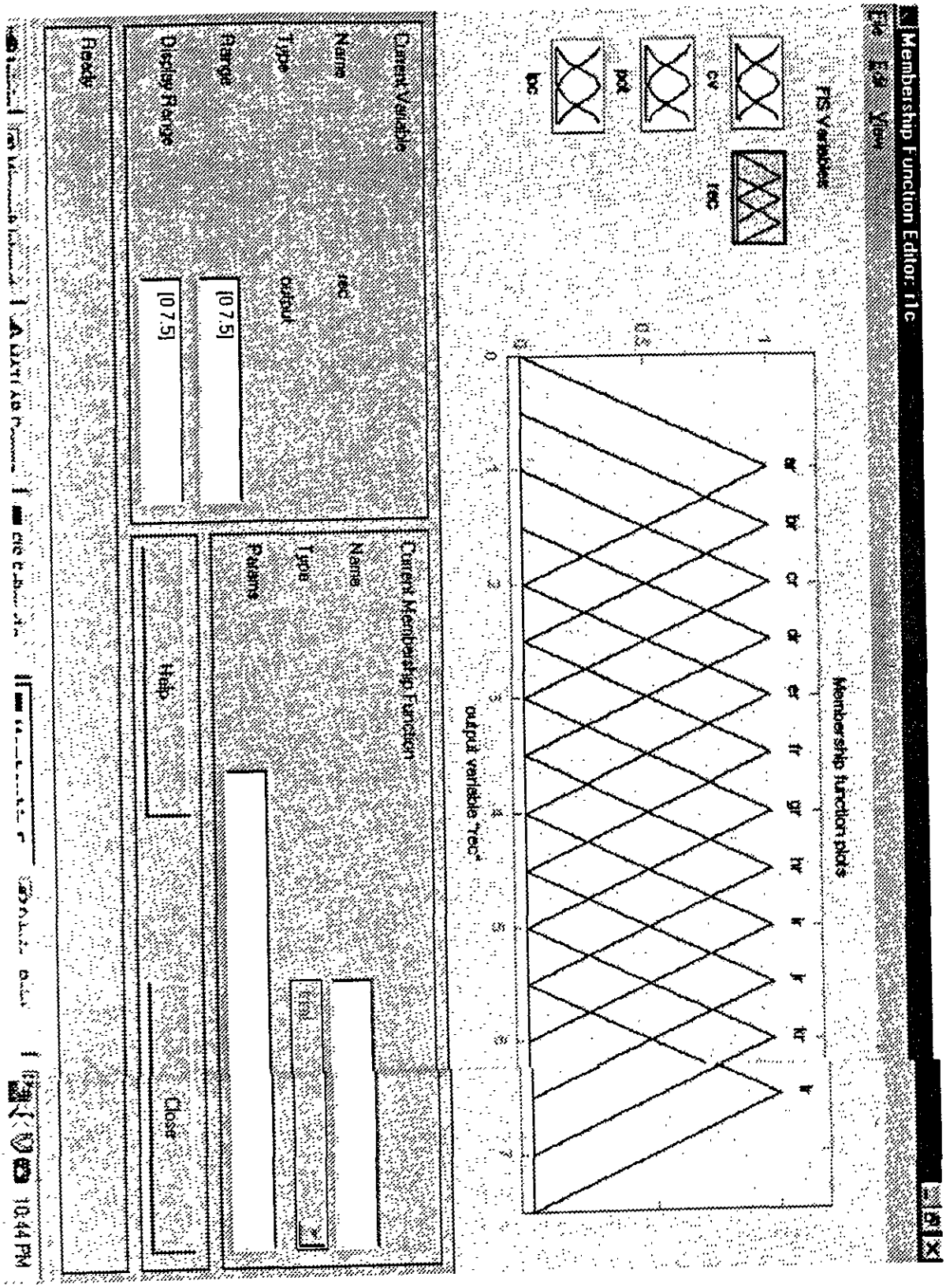


Fig VII 8 LA PANTALLA MUESTRA LAS FUNCIONES DE MEMBRESÍA DE LA VARIABLE DE SALIDA *reconoce*

PRUEBAS Y VALIDACIÓN DEL SISTEMA.

VIII.1. Introducción.

En este capítulo se describen las pruebas realizadas al sistema, así como cada una de sus partes, indicando los parámetros que se utilizaron como referencia para determinar el desempeño del sistema. Se puede comentar que para determinar las especificaciones del sistema, es necesario conocer los siguientes parámetros.

- Tamaño del bloque o marco.
- Traslape entre bloques.
- Orden de análisis LPC
- Elección de la medida para determinar la distancia entre bloques.
- Elección de la restricción local en el algoritmo DTW
- Elección de la ponderación sobre el camino local en el algoritmo DTW
- Elección de un pequeño valor para obtener el doble del número de centroides.
- Elección de entradas al sistema difuso.

Hay que hacer notar que dicho sistema fue probado variando cada uno de los parámetros anteriores y así obtener los valores que mejoran los resultados del sistema.

VIII.2. Tamaño del bloque o marco.

Con el objeto de encontrar el tamaño del marco más adecuado para analizar la señal se realizaron pruebas de variación entre 128 y 512 muestras por bloque, quedando finalmente 256 muestras por marco, con una razón de muestreo de 22050 Hz. Debido a que con un número menor de muestras por bloque se obtiene poca información espectral y con un número mayor de la cantidad seleccionada (256 muestras por marco) se pierde información, ya que la voz no es una señal estacionaria, entonces se debe analizar la palabra pronunciada por bloques para hacer que la palabra expresada sea casi estacionaria. Si los marcos son demasiado grandes, se perdería el objetivo de analizar la señal por bloques

VIII.3. Traslape entre bloques.

Si se denota a N como el número de muestras por bloque y M el número de muestras que separan bloques adyacentes. Entonces si $M \leq N$ los bloques se traslapan y la

estimación espectral resultante se correlaciona de marco a marco Luego si $M<<N$ la estimación espectral de marco a marco será completamente suavizada. Pero si $M>N$ no hay traslape entre marcos adyacentes, de hecho parte de la señal de voz será totalmente perdida y no aparecerá en ningún marco de análisis. Finalmente se tomó $M=(1/3)N$ así que $M=(1/3)256=86$

VIII.4. Orden de análisis LPC.

El orden de análisis de Codificación por Predicción Lineal (LPC) es el número de coeficientes (parámetros LPC) del análisis espectral sobre los marcos de voz de la señal. Estos parámetros especifican el espectro de un modelo todo-polo, así que un orden bajo da pobres características espectrales del marco. Un orden alto da mejores especificaciones espectrales pero conlleva un mayor tiempo de proceso Entonces se tomó un orden de 12 ($p=12$) que da buena información espectral y la información dada por un análisis espectral de mayor orden es despreciablemente mejor.

VIII.5. Elección de la medida para determinar la distancia entre bloques.

Algunos métodos fueron estudiados para este fin:

- Medidas Espectrales Logarítmicas RMS, donde se involucran dos FFT's y dos logaritmos para estimar la integral de la ec. IV.6.
- Medida de Distancia Cepstral.- Aquí se requiere del cálculo de coeficientes cepstrales que pueden ser obtenidos de los parámetros LPC
- Medida de Coseno Hiperbólico.- Implica cálculos de logaritmos y FFT's.
- Medida de Itakura y Saito.- Se requiere de los parámetros LPC y coeficientes de autocorrelación.

Por los cálculos que implican cada una de las medidas anteriores se dicidió tomar la Medida de Coeficientes Cepstrales para el Reconocedor de Voz Tradicional y la Medida de Itakura y Saito para el Reconocedor de Voz con la técnica de Cuantización Vectorial, ya que haciendo uso de estas medidas no se requiere tanto tiempo de procesamiento y son más sencillas de obtener.

VIII.6. Elección de la restricción local en el algoritmo DTW.

El objetivo del algoritmo de Deformación de Tiempo Dinámico (por sus siglas en inglés DTW) es comparar dos señales aunque éstas tengan diferente número de bloques. Este algoritmo funciona satisfactoriamente para encontrar un camino óptimo en una rejilla, colocando en el eje vertical una señal y en el eje horizontal la otra señal, formando de esta

manera dicha rejilla. Ese camino óptimo se puede obtener eligiendo uno de los cinco métodos mencionados en este trabajo y el escogido fue la Restricción Local del Tipo II, debido a que sólo tiene tres caminos posibles y son más directos para llegar al punto final.

VIII.7. Elección de la ponderación sobre el camino local en el algoritmo DTW.

La ponderación actúa sobre el camino local y hace que los caminos posibles de la restricción elegida tengan pesos distintos. La ponderación Tipo C es la elegida para este trabajo debido a que consume menor tiempo, la cual actúa sobre los caminos locales.

VIII.8. Elección de un pequeño valor para obtener el doble del número de centroides.

Este pequeño valor es el incremento y decremento que se le da al centroide o centroides existentes para perturbar al sistema y así obtener el doble de los centroides ya existentes Este valor debe ser pequeño y menor que 0.01, aquí se eligió de 0 001.

VIII.9 Elección de entradas al sistema difuso.

Los datos que son entradas al sistema difuso son:

- La mínima distancia resultante del reconocedor CV
- La mínima distancia resultante del reconocedor LPC.
- La potencia de la señal de entrada.

Otras alternativas que podrían haber sido seleccionadas para ser las entradas al sistema difuso serían la mínima distancia dada por el reconocedor de LPC usando la distancia de Itakura, la distancia cepstral y la distancia espectral logarítmica. Finalmente se eligieron las tres entradas mencionadas para hacer uso de las técnicas de CV y LPC.

VIII.10. Resultados.

Los tres reconocedores de voz (reconocedor CV, reconocedor LPC y reconocedor difuso) tienen el mismo vocabulario de doce palabras. Una vez probados los tres anteriores reconocedores se decidió hacer pruebas adicionales programando un reconocedor de voz difuso de dos entradas, siendo estas la distancia mínima dada por el reconocedor LPC y la potencia También se probó otro reconocedor de voz difuso el cual ocupa como entradas la distancia mínima dada por el reconocedor CV y la potencia. Es importante señalar que el número de pronunciaciones repetidas por palabra con las que se probaron los dos primeros

Hay que notar que entre los tres reconocedores difusos el que tiene mayor porcentaje de aciertos es el reconocedor de tres entradas mejorando al que más se le acerca en 1.4 % que es un reconocedor de voz difuso de dos entradas (*distancia CV* y *potencia*), superando este penúltimo en 1.2 % al que le sigue en porcentaje de aciertos, el reconocedor difuso que tiene como entradas la *distancia LPC* y la *potencia*. Este antepenúltimo mejora en 0.4%. el reconocedor de voz con técnica CV y este a su vez supera al tradicional LPC en 1 8% de aciertos. Por lo anterior se deduce que la diferencia entre el mejor de los reconocedores de voz (último) y el que muestra menor porcentaje de aciertos (primero) es de 4.8 %. A pesar de que el reconocedor RLD (situado en el centro de la tabla) tiene menor porcentaje de aciertos que el RCLD y el RCD, tiene la ventaja de que es menos complejo, más fácil de implementar y más rápido.

VIII.11. Conclusiones.

El algoritmo de Reconocimiento de Voz por medio de la técnica tradicional LPC, contempla lo que usualmente se ocupa en cualquier proceso de Reconocimiento de Voz como son preénfasis, ventaneo, traslape entre bloques y demás. Dicho proceso de reconocimiento es un camino que no necesita demasiada memoria ni tiempo de proceso y entrega resultados satisfactorios. Además, es la base de cualquier otro reconocimiento que utilice otro enfoque. El segundo algoritmo de Reconocimiento de Voz ocupa la técnica de CV, es un procedimiento que también hace uso de conceptos como preénfasis y ventaneo, entre otros Esta técnica utiliza un conjunto de vectores característicos que representan el diccionario de una palabra. Esta manera de Reconocimiento de Voz no emplea demasiada memoria pero si un mayor tiempo de proceso, aunque los resultados son más satisfactorios que los del primer método mencionado. Por último, el quinto algoritmo para Reconocimiento de Voz empleado en este trabajo utiliza la Lógica Difusa. El sistema difuso en sí, no se sirve directamente de los conceptos de traslape entre bloques, preénfasis, distancias y más, es decir, ocupa indirectamente estos conceptos. El Sistema Difuso es simple y sólo requiere de los resultados de experimentos anteriores, que para este caso son los resultados dados por: el proceso para obtener la potencia del comando pronunciado y los dos primeros algoritmos de Reconocimiento de Voz, finalmente el Reconocedor de Voz Difuso determinará la palabra que fue expresada Para terminar este apartado es importante mencionar que el sistema difuso de tres entradas para reconocimiento de voz es el que muestra mejor porcentaje de aciertos, es menos complicado para su implementación y es más fácil de comprender. En general los reconocedores de voz difusos cumplen con las anteriores características por las siguientes razones:

- Pueden ser vistos como una caja negra (reglas de inferencia) con entradas y una salida, de ahí su fácil comprensión e implementación. A diferencia de otros reconocedores de voz que requieren de varias cajas negras o bloques y algunos de ellos con repetidas iteraciones y retroalimentaciones.
- Las funciones de membresía de cada palabra para todos los univeros se apegan a los datos difusos de cada vocablo y por ende se obtiene más precisión en el sistema.

Capitulo IX

CONCLUSIONES

En este capítulo se encuentra el resúmen de la aportación de este trabajo, haciendo referencia sobre los puntos relevantes del mismo, formulandose las siguientes conclusiones:

La presente tésis muestra la gran ventaja de usar Lógica Difusa para Reconocimiento de Voz. Esta nueva herramienta puede ser aplicada a un sinfín de campos sin requerir de complicados sistemas ni de complejos modelos matemáticos que describan al sistema, sólo se necesita de: (1) determinar las entradas y salida del proceso, (2) defusificar dichas entradas y salida, (3) determinación de las reglas de inferencia y, (4) defusificación de la salida.
Evidentemente sería recomendable un cálculo sencillo para obtener las entradas al Reconocedor Difuso, sin embargo se dicidió utilizar como entradas las salidas del reconocedor tradicional LPC y del reconocedor CV, ya que éstos han sido ampliamente utilizados con muy buenos resultados y el primero objetivo de este trabajo fue mejorar su desempeño El evaluar un Reconocedor Difuso con entradas más simples y con desempenos equivalentes obtenidos en esta tesis, es tema de un trabajo futuro.
Defusificar las variables de entradas y salida es representar los posibles valorès de dichas variables con una función de membresía (conjunto difuso).
Las reglas de inferencia son una serie de expresiones IF/THEN escritas en sentido común.
La defusificación de la salida es encontrar un valor preciso en el universo de salida por algún método de defusificación visto en el capítulo VII.

De acuerdo a lo anterior, se concluye que el uso de un algoritmo difuso se basa fundamentalmente, en la complejidad del modelo matemático del proceso, así como del tiempo de cálculo que implica este proceso; además de la precisión en los resultados obtenidos mediante el conocimiento y la experiencia en el manejo del proceso, para de esta forma sustituir la descripción matemática del sistema, así como características que permitan modificar la respuesta a nuestro antojo, en resumen, se concluye que se deberá tener:

- Experiencia en el manejo del proceso o modelarlo matemáticamente.
- Evaluar el número de reglas de inferencia
- Tiempo de ejecución para el sistema difuso y los procesos previos que dan la entrada al sistema difuso
- Precisión en la respuesta de los sistemas previos.
- Certeza en los rangos determinados.

Para muy diversas aplicaciones es muy alentador el uso de algoritmos difusos, los cuales tienen una característica de no utilizar algún modelo matemático del proceso. Frecuentemente se hacen combinaciones de sistemas difusos y sistemas convencionales de algún tipo, que necesitan el modelo matemático en la parte del proceso donde se requiere el

sistema convencional, se pueden mencionar como ventajas el ahorro de tiempo de procesamiento para muchos casos, de lo que se concluye que no se necesita conocer el modelo matemático del proceso, sólo saber como se comporta el sistema. Pero como una desventaja se tendría que conocer de antemano el comportamiento del proceso y hacer pruebas cambiando reglas de inferencia hasta que el funcionamiento del sistema sea aceptable La segunda desventaja es que cuando se describe el comportamiento del proceso, el sistema difuso puede llegar a tener un número muy grande de reglas y por ende, puede llevar un mayor tiempo de procesamiento.

El reconocedor difuso de tres entradas es el que muestra mayor porcentaje de aciertos, menos complejidad, menor tiempo de procesamiento, es más fácil de comprender y muy fácil de implementar. Los otros dos sistemas difusos de dos entradas, cumplen con las características del reconocedor de voz de tres entradas pero muestran menor porcentaje de aciertos, debido a que al ocupar dos entradas su salida sólo depende de la certeza de dos datos, para el sistema de tres entradas su salida depende de la certeza de tres datos, es decir, la posibilidad de error en un sistema difuso de tres entradas disminuye. Es necesario mencionar que los reconocedores de voz difusos pueden ser mejorados, a fin de que las funciones de membresía de los tres universos de entrada se apeguen mejor a los datos difusos de cada palabra del vocabulario para obtener una mejor taza de reconocimiento del sistema.
De los tres reconocedores difusos el que tiene mayor porcentaje de aciertos es el reconocedor de tres entradas mejorando al que más se le acerca en 1.4 % que es un reconocedor de voz difuso de dos entradas (*distancia CV* y *potencia*), superando este penúltimo en 1.2 % al que le sigue en porcentaje de aciertos, el reconocedor difuso que tiene como entradas la *distancia LPC* y la *potencia*. Este antepenúltimo mejora en 0.4%. el reconocedor de voz con técnica CV y este a su vez supera al tradicional LPC en 1.8% de aciertos. Por lo anterior se deduce que la diferencia entre el mejor de los reconocedores de voz (último) y el que muestra menor porcentaje de aciertos (primero) es de 4 8 %.

Se concluye también que la Codificación por Predicción Lineal (LPC) es una de las más eficaces técnicas de parametrización de la señal de voz, habiéndose construido gracias a ella sistemas realmente notables.
Algo realmente importante que se debe señalar es que el análisis por Predicción Lineal proporciona, al igual que la FFT, un conjunto "autosuficiente" de parámetros que representan la señal y permite su adecuada reconstrucción. El modelo de producción de voz simula las cavidades del tracto vocal mediante un filtro lineal todo polos y modeliza los modos de excitación mediante dos tipos de señales. Estas señales son: ruido blanco que representa los segmentos no sonoros de la señal y trenes de pulsos que representan los segmentos sonoros.

Otra conclusión de importancia es la comparación entre patrones que, es un proceso en el que no se puede saber si los patrones a comparar tienen o no el mismo número de marcos o bloques, así que compararlos sin el proceso de Deformación de Tiempo Dinámico (DTW), sería eliminar algunos bloques del patrón que tenga más marcos, pudiendo ser esos bloques los que tengan las principales características del patrón o palabra pronunciada. Luego entonces. de ahí la importancia de usar el proceso de Deformación de Tiempo

Dinámico, ya que este tiene la ventaja de tomar los marcos inicial y final de los dos patrones que se van a comparar y sólo tomar los marcos o bloques con los cuales se obtiene la mínima distancia al comparar dos marcos (uno del patrón de prueba y otro del patrón de referencia). Aunque al hacer uso del algoritmo DTW se tiene la desventaja de requerir de mayor tiempo de procesamiento; es importante hacer uso de él, debido a que da una mayor precisión en el reconocimiento que cuando este algoritmo no se incluye.

De los caminos o métodos para obtener la distancia entre marcos se han evaluado tres, a saber· coeficientes cepstrales, Itakura y Saito y medidas espectrales logarítmicas. De los cuales todos mostraron buen desempeño, siendo los más sencillos y que no requieren demasiado tiempo de procesamiento la medida por medio de coeficientes cepstrales y la medida de Itakura y Saito.

Cuantización Vectorial es una técnica que representa una clase o palabra de una manera muy efectiva. Esto es, de un conjunto de la misma palabra se extraen las principales características espectrales para formar un centroide. Perturbando al sistema se pueden obtener el número de centroides que se deseé y vendría siendo el orden del cuantizador el número de centroides. Esta técnica de parametrizar una palabra tiene las ventajas de usar poca memoria para almacenar el patrón y sobre todo, de ser una muy efectiva representación de dicho patrón. Aunque tiene la desventaja de demandar más tiempo de procesamiento en la etapa de reconocimiento, así como también en la obtención de los cuantizadores.

El reconocedor de voz con técnica CV puede superar la taza de reconocimiento obtenida, perturbando los centroides con un valor más pequeño. Elegir el tipo de ponderación "*d*" para la implementación del algoritmo DTW, y realizar pruebas de la mejor medida de distancia entre los patrones de prueba y referencia, se obtendría un mayor porcentaje de aciertos en el reconocedor tradicional LPC. Independientemente de lo anterior, los conjuntos difusos de cada variable pueden ser más precisos, es decir, que cada función de membresía se apegue mejor a los datos que representa. Sin lugar a duda, las tres hipótesis anteriores llevarían a los tres sistemas difusos a un incremento en la tasa de reconocimiento, ya que éstos tienen mayor porcentaje de aciertos que el mejor de los reconocedores.

Un sistema como este, puede tener diferentes aplicaciones en el hogar como son abrir las puertas, las ventanas, las cortinas, encender la luz; también tendría aplicación sobre aparatos caseros como la televisión, la radio, y el teléfono entre otros, activar una PC, hacer dictado en una PC. Estos sistemas también son aplicables a otros usos como son las claves bancarias por teléfono; en el automóvil asi mismo se aplicaría para abrir y cerrar puertas y ventas, controlar la temperatura y operación del teléfono. Otras aplicaciones que tal vez parescan extraidas de la ciencia ficción podrían ser hablar con un robot para darle ordenes, asi como controlar en la misma forma un automovil. El reconocimiento de comandos de voz con el uso exclusivo de un usuario, podría tener un gran éxito, razones por las que es importante impulsar su desarrollo

El uso de Lógica Difusa se ha incrementado, en muchas áreas como son Control, Electrónica, Reconocimiento de Patrones y Comunicaciones Se han encontrado en muchos casos que esta herramienta o bien substituye a otro tipo de técnica o es complemento de ella. En el caso de la aplicación del Reconocimiento de Patrones, así como en otras áreas, no necesitamos un modelo matemático que nos ayude a conocer su comportamiento utilizando una técnica tradicional.

Con esta herramienta podemos ahorrar tiempo de cálculo, y como consecuencia se tendrá un software implementado en una determinada arquitectura más eficiente; además de ahorrar espacio de memoria que también tendrá como consecuencia arquitecturas más pequeñas; por tanto prototipos más baratos.

La tendencia del uso del reconocimiento de patrones utilizando este tipo de técnica, puede tener gran auge en el futuro ya que se pueden ocupar algoritmos más poderosos de reconocimiento de patrones.

REFERENCIAS

[1] Bishnu S. Atal, Lawrence R. Rabiner, "*A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Aplications to Speech Recognition*", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-24 No. 3, pp 201-212, June 1976

[2] John E. Holmgren, "*Applying Automatic Speech Recognition to Telephone Services*", IEEE Communictions Magazine, pp 31-34, November 1982.

[3] Cory Myers, Lawrence R. Rabiner, Aaron E. Rosenberg, "*Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition*", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol ASSP-28, No 6, pp 623-635, December 1980.

[4] George M. White, Richard B. Neely, "*Speech Recognition Experiments with Linear Predication, Bandpass Filtering, and Dynamic Programming*", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol ASSP-24, No. 2, pp 183-188, April 1976.

[5] Lawrence R. Rabiner, "*On the Use of Autocorrelation Analysis for Pitch Detection*", IEEE Transaction on Acoustics, Speech, and Signal Processing, Vol. ASSP-25, No. 1, pp 24-33, February 1977.

[6] W Pedrycs, "*Fuzzy sets in Pattern Recognition: Methodology and Methods Pattern Recognition*", Pattern Recognition Society, Vol. 23, No. 1/2, pp 121-146, 1990.

[7] Bart Kosko, Satoru Isaka, "*Fuzzy Logic*", Scientific American, pp 62-67, July 1993.

[8] Earl Cox, "*Fuzzy Fundamentals*", Advanced Technology/circuits, IEEE Spectrum, pp 58-61, October 1992.

[9] Papamichalis Panos E., "*Practical Approaches to Speech Coding*", Ed. Prentice Hall Inc , Englewood Cliffs, New Jersey 1987.

[10] Paul M. Embree, Bruce Kimble, "*C Lenguaje Algorithms for Digital Signal Processing*", Ed Prentice Hall, Englewood Cliffs, New Jersey 1991.

[11] Lawrence Rabiner, Biing-Hwang Juang "*Fundamentals of Speech Recognition*", Ed. Prentice Hall, Englewood Cliffs, New Jersey 1993.

[12] Augustine H, Gray, John Markel, "*Distance Measure for Speech Processing*", IEEE Trans. Acoust, Signal Processing, Vol ASSP-24, No 5, pp 380-391, October 1976,

[13] Francisco Casacuberta, Enrique Vidal, "*Reconocimiento Automático del Habla*", Ed. Marcombo Boixareu, Barcelona 1987.

[14] John R. Deller Jr , John G Proakis, John H. L. Hanson, "*Discrete Time Processing of Speech Signal*", Ed. Macmillan, New York 1993

[15] Proakis John G., Manolakis Dimitris G., "*Digital Signal Processing: principles, algorithms and applications*" Ed. Macmillan, New York 1996

[16] José Eduardo Torres F., "*Diseño y realización de un sistema de síntesis de señales de voz*", Tesis UNAM, F.I., 1995.

[17] John Makhoul, "*Spectral Analysis of Speech by Linear Prediction*", IEEE Trans. on Audio and Electroacoustics, Vol AU-21, No. 3, pp 140-148, June 1973.

[18] L R Rabiner, M. R. Sambur, C.E. Schmidt, "*Applications of a Nonlinear Smothing Algorithm to Speech Processing*", IEEE Trans. Acoust., Signal Processing, Vol ASSP-23, pp 552-557, December 1975.

[19] L. R. Rabiner, J. G Wilpon, "*A Simplified Robust Training Procedure for Speaker Trained, Isolated Word Recognition Systems*", Acoustical Society of America, Vol 68 No 5, pp 1271-1276, November 1980.

[20] Fumitada Itakura, "*Minimum Prediction Residual Principle Applied to Speech Recognition*", IEEE Trans Acoust, Speech, and Signal Processing, Vol ASSP-23, No. 1, pp 145-150, February 1975.

[21] H. Sakoe, S Chiba, "*Dynamic Programming Optimization for Spoken Word Recognition*", IEEE Trans. Acoust., Speech, Signal Processing, Vol ASSP-26, No 1, pp 194-200, February. 1978

[22] Christopher Hal, Cam Quynh Nguyen, "*Voice Command Recognition Using Fuzzy Logic Based Digital Filters*", The international Conference on Signal Processing Application and Technology, ICSPAT 94, pp 1650-1654, October 18-21, Dallas, Tx USA 1994.

[23] Leonard W. Esteves, Nasser D. Kehtarnavaz, "*A TMS320C30, Based Fuzzy Logic Recognition Systems for Spoken Digits*", Texas Instruments Fifth Annual TMS 320 Educators Conference, August 10-11, Houston, Texas USA 1995.

[24] Mahommand Jamshidi, Nader Vadiee, Timothy J. Ross, "*Fuzzy Logic and Control: Software and Hardware Applications*", Ed. Prentice Hall, Englewood Cliffs, New Jersey 1993.

[25] Toshiro Terano, Kiyoji Asai, Michio Sugeno, "*Fuzzy Systems Theory and its Applications*", Ed. Academic Press Inc., Boston 1992.

[26] Robert F. Calusdian, "*Fuzzy Reality*", IEEE Potentials, pp 20-22, April/May 1995.

[27] Selected Papers by L. A. Zade, Edited by R. R. Yager, S. Ovchinnikov, R. M. Tong, H T.Nguyen, "*Fuzzy Sets and Applications*", Ed. John Wiley and Sons, pp-29-412, USA 1987

[28] Toshinori Munakata, Yashuant Jani, "*Fuzzy Systems and Overview*", Communications of the ACM 1994, Ed. Prentice Hall, pp 30-180, New Yersey 1994.

[29] Bart Kosko, "*Neural Networks and Fuzzy Systems*", Ed.. Prentice Hall, Englewood Cliffs, New Jersey 1992

[30] Timothy J Ross, "*Fuzzy Logic with Engineering Applications*", Mc. Graw Hill, Inc. 1995

[31] José M. Tribolet, Lawrence R Rabiner, "*Statistical Properties of an LPC Distance Measure*", IEEE Transactions on Acoustics, Speech, and Signal Processing. Vol ASSP-27, No 5 october 1979.

[32] Jesús Savage Carmona, "Implementación de algoritmos de procesamiento digital de señal es con microprocesadores", Tesis UNAM-DEPFI, 1989.

[33] Miguel Comadurán Chavarría, "Sistema de reconocimiento de comandos de voz para la conducción de una silla de ruedas", Tesis UNAM-DEPFI, 1995.

[34] Antonio Fco. Mondrágon Torres, "Técnica de reconocimiento automático de voz utilizando segmentación acústica", Tesis UNAM-DEPFI, 1996.

[35] Mauricio Alberto Martínez García, "Método de reconocimiento de palabras aisladas usando segmentación acústica y cuantización vectorial", Tesis UNAM-DEPFI, 1997.

APÉNDICE

Desempeño de algunos otros esquemas de Reconocimiento de Voz

REFERENCIA	TÍTULO DE LA TESIS	AÑO	TÉCNICA		PORCENTAJE DE ACIERTOS DE VOZ
[32]	"Implementación de algoritmos de procesamiento digital de señales con microprocesadores"	1989	Cuantización Vectorial utilizando Codificación por Predicción Lineal		99.00%
[33]	"Sistema de reconocimiento de comandos de voz para la conducción de una silla de ruedas"	1995	Codificación por Predicción Lineal		86 69%
[34]	"Técnica de reconocimiento automático de voz utilizando segmentación acústica."	1996	Análisis con agrupamiento no difuso		93.46%
			Análisis con agrupamiento difuso	KM-LPCE	98.45%
				KM-LPCR	97.64%
				KM-CEPE	100.00%
				KM-CEPR	98.58%
[35]	"Método de reconocimiento de palabras aisladas usando segmentación acústica y cuantización vectorial"	1997	Sistema LPC con segmentación lineal		93.88%
			Sistema LPC con segmentación acústica		92.42%
			Sistema KLT con segmentación lineal		83.90%
			Sistema KLT con segmentación acústica.		85 00%