CENTRO DE INVESTIGACIÓN CIENTÍFICA Y DE
EDUCACIÓN SUPERIOR DE ENSENADA

RECONOCIMIENTO AUTOMÁTICO DE LOCUTOR Y REALIZACIÓN DE UN SISTEMA EXPERIMENTAL

TESIS
MAESTRÍA EN CIENCIAS

HERNANDO SILVA VARELA

Ensenada, Baja California, México, febrero de 1994

RESUMEN de la tesis de Hernando Silva Varela presentada como requisito parcial para la obtención del grado de MAESTRO EN CIENCIAS en FÍSICA APLICADA con opción en ELECTRÓNICA Y TELECOMUNICACIONES. Ensenada, Baja California, México. Febrero de 1994.

RECONOCIMIENTO AUTOMÁTICO DE LOCUTOR Y REALIZACIÓN DE UN SISTEMA EXPERIMENTAL.

Resumen aprobado por: ___

Se describe el problema de reconocimiento automático de locutor (RAL) y las modalidades que éste presenta. Se analizó el mecanismo de generación de la señal de voz, su modelado, y los parámetros de ésta que han sido aplicados en RAL. Fueron consideradas, para su estudio, tres técnicas que han sido planteadas en la literatura para la solución del problema y que han producido resultados relativamente buenos, éstas son: alineamiento dinámico en el tiempo (ADT), cuantificación vectorial (CV) y modelos ocultos de Markov (MOM). Se describe también el diseño y la realización del sistema experimental de reconocimiento automático de locutor (SERAL), el cual es dependiente del vocabulario y está basado en una técnica de alineamiento dinamico en el tiempo (ADT). SERAL está constituido por equipo electrónico y programas, el núcleo del equipo electrónico lo forman una computadora personal compatible con los modelos de IBM y la tarjeta de adquisición, procesamiento y síntesis de señales (TAPS), la cual fue diseñada *exprofeso* utilizando el procesador digital de señales DSP32 y el circuito de interfaz analógica TLC32044.

Fue desarrollado un programa en lenguaje de alto nivel "C", encargado de administrar al sistema y de servir como interfaz entre el usuario y el mismo. También se desarrollaron programas en lenguaje de bajo nivel (ensamblador DSP32) para las tareas de adquisición y reproducción de señales de voz, detección de inicio-fin de palabras aisladas, parametrización de los segmentos de voz detectados, y alineamiento y comparación de patrones de locutor. La modularidad del sistema permite que cualquiera de los tres algoritmos anteriores pueda ser modificado y/o utilizado de manera independiente en futuras realizaciones de sistemas similares.

SERAL puede trabajar en los modos de identificación y verificación automática de locutor (IAL y VAL). El sistema fue utilizado en experimentos con 10 locutores que lo entrenaron y probaron usando primero la palabra "HOLA" y después la palabra "NUMERO", los resultados mostraron una tasa de equierror (TEE) de 3.01% y 2.57% para la primera palabra y 3.59% y 1.51% para la segunda, para umbral general e individual respectivamente.

ABSTRACT of the MsC thesis presented by Hernando Silva Varela in order to obtain the MASTER of SCIENCE degree in APPLIED PHYSICS with the option of ELECTRONICS and TELECOMMUNICATIONS. Ensenada, Baja California, México. February 1994.

AUTOMATIC SPEAKER RECOGNITION
AND THE IMPLEMENTATION OF AN EXPERIMENTAL SYSTEM.

Abstract approved by:

The automatic speaker recognition (ASR) problem an its modalities are described. The speech production mechanism, its modeling, and the speech parameters that have been applied in ASR were analyzed. Three techniques described in the literature were considered for study; these techniques have shown relatively good results and are known as: dynamic time warping (DTW), vector quantization (VQ) and hidden Markov modeling (HMM). The design and the implementation of the experimental system for automatic speaker recognition (SERAL), which is vocabulary dependent, and is based in a dynamic time warping (DTW) approach, is also described. SERAL is composed by hardware and software; hardware is based on an IBM compatible personal computer, and an acquisition, processing and signal synthesis (TAPS) card, which was designed *exprofeso* by using the digital signal processor DSP32 and the analog interface circuit TLC32044.

System managing and user interface were developed in a high level language (C) program. Low level language programs (DSP32 assembler) for the acquisition and replay of speech, isolated word boundaries detection, parametrization of the detected speech segments, and alignment and comparison of speaker templates were also developed. The system modularity allows the independent modification and/or utilization of the three algorithms mentioned above in the future implementations of similar ASR systems.

SERAL can work in automatic speaker identification and verification modes (ASI and ASV respectively). Several experiments were carried out with 10 speakers who trained and tested the system by using the Spanish words "HOLA" (Hello) and "NÚMERO" (number). The results have shown a 3.01% and 2.57% equal error rates (EER) for the first word, and 3.59% and 1.51% for the second one, by using general and individual threshold respectively.

CENTRO DE INVESTIGACIÓN CIENTÍFICA Y
DE EDUCACIÓN SUPERIOR DE ENSENADA.

DIVISIÓN DE FÍSICA APLICADA.
DEPARTAMENTO DE ELECTRÓNICA Y TELECOMUNICACIONES.

RECONOCIMIENTO AUTOMÁTICO DE LOCUTOR Y REALIZACIÓN DE UN SISTEMA EXPERIMENTAL.

TESIS

que para cubrir parcialmente los requisitos necesarios para
obtener el grado de MAESTRO EN CIENCIAS presenta:

HERNANDO SILVA VARELA.

Ensenada, Baja California, México. Febrero de 1994.

DEDICATORIA

A mis cuatro ojos verdes.

A mi familia.

A áquel lugar del centro,
espacio cerrado de media luz;
viejo laberinto de piedra
con pretensiones de grandeza.

A los que siguen creyendo;
porque los envidio.
A los que ya no creen;
porque los comprendo.

A Pedro Mendizábal de Cabral.

AGRADECIMIENTOS

Al Dr. Ciro A. Martínez G.-M. por su apoyo a éste trabajo.

A los miembros del comité de tesis: Dr. Enrique Mitrani A., Dr. Joaquín Álvarez G., Dr. J. Apolinar Reynoso H., y M.C. Vicente Ferreira B. por sus comentarios y sugerencias en la realización de éste trabajo.

A los técnicos Rogelio Reyes S., Raúl Moreno B. e Isaac Fimbres quienes diseñaron y elaboraron el circuito impreso de la primera versión de la TAPS.

A Carlos Merla V., Jorge Peña G. y Miguel A. Gómez (v.g., Los brazos) quienes implantaron la interfase DSP32-TLC32044 en un arreglo lógico programable.

A todos mis profesores, compañeros y amigos, en especial a los zacatecos Alejandro I, Alejandro II, Conrado, Ernesto, Humberto, Reynaldo y Víctor; a los del grupo de PDS Julio y Santiago; a mis compañeros de generación Gerardo, Horacio, Miguel y Rafael (v.g., Teenager mutant ninja turtles).

A todos los locutores que accedieron a cooperar en los experimentos.

Al señor Edward F. Mackin, su esposa Mary Mackin, y su familia en San Francisco, California, USA; por su hospitalidad.

Al centro de investigación científica y de educación superior de Ensenada (CICESE) y al consejo nacional de ciencia y tecnología (CONACyT) por su apoyo y soporte logístico.

CONTENIDO

		Página
I	INTRODUCCIÓN.	1
II	ANÁLISIS DEL PROBLEMA DE RAL	5
	II.1 Presentación del problema.	5
	II.2 Verificación e identificación de locutor.	5
	II.3 Dependencia e independencia del vocabulario.	8
	II.4 Entrenamiento y prueba (reconocimiento).	9
	II.5 Reconocimiento subjetivo de locutor.	11
	II.5.1 Reconocimiento de locutor por escuchas.	12
	II.5.2 Reconocimiento de locutor usando espectrogramas.	13
III	MODELOS Y PARÁMETROS DE LA SEÑAL DE VOZ UTILIZADOS EN RECONOCIMIENTO AUTOMÁTICO DE LOCUTOR (RAL).	21
	III.1 El proceso de producción de la voz.	21
	III.2 Modelo digital para la señal de voz.	23
	III.2.1 Modelo análogo a la salida.	23
	III.2.2 Cálculo de los parámetros por medio de codificación predictiva lineal (CPL)	27
	III.3 Cepstrum	35
	III.4 Efectividad de los parámetros de la señal de voz en reconocimiento automático de locutor (RAL).	40
IV	TÉCNICAS DE RECONOCIMIENTO AUTOMÁTICO DE LOCUTOR (RAL)	44
	IV.1 Generalidades.	44
	IV.2 Alineamiento dinámico en el tiempo (ADT).	45
	IV.3 Cuantificación vectorial (CV).	53
	IV.3.1 Resumen de cuantificación vectorial (CV).	53
	IV.3.2 Aplicación de cuantificación vectorial (CV) en reconocimiento automático de locutor (RAL).	58

CONTENIDO (Continuación).

		Página
IV.4	Modelos ocultos de Markov (MOM).	63
	IV.4.1 Cadenas de Markov.	64
	IV.4.2 Extensión a modelos ocultos de Markov (MOM).	65
	IV.4.3 Aplicación de modelos ocultos de Markov (MOM) en reconocimiento automático de locutor (RAL).	69
V	REALIZACIÓN DEL SISTEMA EXPERIMENTAL DE RECONOCIMIENTO AUTOMÁTICO DE LOCUTOR (SERAL).	73
V.1	Discusión.	73
V.2	Infraestructura física utilizada.	74
	V.2.1 Tarjeta de adquisición, procesamiento y síntesis de señales (TAPS)	74
	V.2.2 Sistema físico integrado.	77
V.3	Infraestructura programática desarrollada.	78
	V.3.1 Adquisición y reproducción de señales.	79
	V.3.2 Algoritmos implantados.	80
	V.3.2.1 Detección de inicio y fin de palabras aisladas.	81
	V.3.2.2 Extracción de parámetros.	90
	V.3.2.3 Alineamiento, generación y comparación de patrones de locutor	94
	V.3.3 Archivos auxiliares y de configuración.	99
VI	EXPERIMENTOS Y RESULTADOS	101
VI.1	Base de datos.	101
VI.2	Estimación de umbrales.	102
VI.3	Experimentos y resultados.	107
VII	CONCLUSIONES Y RECOMENDACIONES.	117
REFERENCIAS		121

CONTENIDO (Continuación).

Página

apéndice A - Teoría acústica de la producción de voz y modelos de tubos sin pérdidas. 124

A.1 Teoría acústica de la producción de la voz. 124

A.1.1 Acústica básica de la propagación del sonido. 124

A.1.2 Tubo uniforme sin pérdidas. 126

A.2 Modelo de concatenación de tubos sin pérdidas. 130

A.2.1 Propagación de ondas en tubos concatenados sin pérdidas. 131

A.2.2 Condiciones de frontera. 133

A.2.3 Interpretación digital. 134

apéndice B - Glosario de términos y acrónimos español-inglés. 139

apéndice C - Ejemplo de modelos evidentes de Markov (MEM). 141

apéndice D - Diagramas esquemáticos de la TAPS. 142

apéndice E - Diagramas de flujo interactivos computadora-DSP32. 145

apéndice F - Lista de programas de SERAL. 150

LISTA DE FIGURAS

Figura		Página
1	Esquema general de reconocimiento de locutor	10
2	Esquema general de identificación de locutor.	10
3	Esquema general de verificación de locutor.	11
4	Forma de onda de la frase "Two plus seven is less than ten" (escala en segundos). Debajo de la forma de onda se muestra la transcripción fonética alineada temporalmente [Oppenheim y Schafer, 1989].	16
5	(a) Espectrograma de banda ancha de la forma de onda de la figura 5. (b) Espectrograma de banda angosta [Oppenheim y Schafer, 1989].	19
6	Esquema del sistema vocal humano.	21
7	(a) Representación del modelo de tubos sin pérdidas; (b) Modelo análogo a la salida.	24
8	Modelo típico análogo a la salida para la producción de voz.	26
9	Diagrama simplificado del modelo para la producción de voz.	28
10	Realización práctica para la obtención del *cepstrum*	37
11	(a) *cepstrum* y (b) logaritmo del espectro de magnitud para segmentos secuenciales de voz sonora [Oppenheim, 1989].	38
12	Porcentaje de exactitud en la identificación de locutor como función de la trama utilizada [Atal, 1974].	43
13	Normalización lineal temporal para dos secuencias de diferente longitud.	47
14	Región permisible de la ruta función de alineamiento.	48
15	Ilustración de la efectividad del alineamiento dinámico en el tiempo para el contorno de energía de dos versiones de la palabra inglesa /seven/.	52
16	Partición de un espacio vectorial en subespacios, donde cada subespacio es representado por un centroide.	55

LISTA DE FIGURAS (continuación)

Figura		Página
17	Diagrama a bloques de la estructura de un cuantificador vectorial	56
18	Diagrama conceptual ilustrando la obtención del *libro de códigos*.	59
19	Diagrama a bloques del sistema de reconocimiento automático de locutor basado en cuantificación vectorial (CV) [Soong *et al*, 1987].	60
20	Tasa de error en identificación de locutor contra tasa de *libro de códigos* para secuencias de 1, 2, 4 y 10 dígitos.	61
21	Porcentaje de error de identificación contra dígitos hablados por el locutor. repetidos y diferentes [Soong *et al*, 1987].	62
22	Mediana de la tasa de equierror (TEE) contra número de dígitos de prueba en experimentos de verificación de locutor [Soong *et al*, 1987].	62
23	Cadena de Markov de 5 estados con transiciones de estado seleccionadas.	64
24	Cadena de Markov de 6 estados para el modelo de un dado.	66
25	*Modelo oculto de Markov (MOM)* de 2 estados para dos dados.	67
26	Diagrama a bloques del sistema de RAL basado en MOM.	70
27	Diagrama a bloques de la tarjeta de adquisición, procesamiento y síntesis de señales (TAPS).	74
28	Diagrama a bloques del circuito de interfase analógica (CIA) TLC32044.	76
29	Diagrama esquemático del sistema fisico integrado.	77
30	Diagrama esquemático del proceso de *identificación de locutor*.	80
31	Diagrama esquemático del proceso de *verificación de locutor*.	81
32	Forma de onda de la palabra /fases/.	82
33	Diagrama a bloques del algoritmo de *detección de inicio y fin*.	83

LISTA DE FIGURAS (continuación)

Figura		Página
34	Diagrama ilustrativo de la detección de pulsos de energía.	85
35	Diagrama ilustrativo de la asociación de pulsos de energía.	87
36	Despliegue gráfico de SERAL ilustrando la *detección de inicio y fin* para la palabra /fases/ de la figura 32.	88
37	Despliegue gráfico de la *detección de inicio y fin* para la secuencia de palabras /tres seis siete nueve dos/.	89
38	Diagrama a bloques del procedimiento de extracción de parámetros.	90
39	Despliegue gráfico de SERAL mostrando el *contorno de coeficientes cepstrum* y *el contorno de pendientes* para la palabra /fases/ de la figura 32.	94
40	Diagrama de flujo del procedimiento para el ADT de dos *contornos*.	95
41	Despliegue gráfico de SERAL mostrando el alineamiento y la comparación de los patrones de dos locutores al utilizar la palabra /NÚMERO/.	97
42	Despliegue gráfico de SERAL mostrando la información contenida en un patrón típico de locutor.	98
43	Distribuciones típicas de *distancias intralocutor e interlocutor*.	103
44	Esquema experimental de cómputo para la estimación de las TEE.	106
45	Resultados experimentales resumidos gráficamente para la TEE del umbral general y el promedio de las TEE de los umbrales individuales.	115
46	(a) Tracto vocal como un tubo no uniforme. (b) Función área [Rabiner y Schafer, 1978].	126
47	Tubo uniforme sin pérdidas con terminaciones ideales [Rabiner y Schafer, 1978]	127
48	(a) Respuesta en frecuencia. (b) Localización de los polos para un tubo uniforme sin pérdidas [Rabiner y Schafer, 1978].	129

LISTA DE FIGURAS (continuación)

Figura		Página
49	Modelo del tracto vocal como una concatenación de tubos acústicos sin pérdidas.	130
50	Flujo de señal en la juntura entre dos tubos sin pérdidas.	132
51	Gráfica de flujo de señal para un modelo completo de dos tubos.	133
52	Concatenación de N tubos sin pérdidas de longitud igual.	135
53	(a) Gráfica de flujo de señal para el modelo del tracto vocal como tres tubos sin pérdidas concatenados; (b) Sistema equivalente en tiempo discreto; (c) Sistema equivalente de tiempo discreto usando únicamente retardos completos.	137
54	Diagrama esquemático de la sección de adquisición, ésta incluye la interfase DSP32-TLC32044	142
55	Diagrama esquemático del módulo de procesamiento, éste incluye el DSP32 y su memoria externa [Martínez y Félix, 1993].	143
56	Diagrama esquemático de la sección de interfase al ducto ISA.	144
57	Diagrama de flujo interactivo computadora-DSP32 para la adqusición de señales.	145
58	Diagrama de flujo interactivo computadora-DSP32 para la síntesis de señales.	146
59	Diagrama de flujo interactivo computadora-DSP32 para la detección de inicio y fin de palabras aisladas.	147
60	Diagrama de flujo interactivo computadora-DSP32 para la extracción de parámetros	148
61	Diagrama de flujo interactivo computadora-DSP32 para el alineamiento, comparación y promediación de contornos.	149

LISTA DE TABLAS

Tabla		Página
I	Resultados de una simulación para la obtención de los coeficientes de predicción lineal para 360 datos de valor 1.01.	33
II	Parámetros variados en los experimentos y sus valores.	108
III	Resultados variando el orden de predicción.	109
IV	Resultados variando la longitud de trama.	110
V	Resultados variando el recorrimiento de trama.	111
VI	Resultados al agregar información transicional.	112
VII	Resultados al variar el parámetro δ del algoritmo de ADT.	112
VIII	Resultados para varias muestras de entrenamiento.	113
IX	Resultados óptimos para las palabras "HOLA" y "NÚMERO".	114

RECONOCIMIENTO AUTOMÁTICO DE LOCUTOR Y REALIZACIÓN DE UN SISTEMA EXPERIMENTAL.

CAPÍTULO I INTRODUCCIÓN

La comunicación oral es la transferencia de información de una persona a otra por medio de la voz, la cual consiste en variaciones de presión generadas por el tracto vocal del locutor. Tales variaciones de presión se propagan como ondas a través del aire y alcanzan el oído de los escuchas, quienes descifran las ondas convirtiéndolas en un mensaje. Los gestos y movimientos del locutor realizados en una comunicación "cara a cara" no se reflejan en las ondas de voz, y no pueden por lo tanto ser tomados en cuenta al estudiar las señales de voz.

A la cadena de eventos que van desde la concepción de un mensaje en el cerebro del locutor, hasta el arribo del mensaje al cerebro del escucha, se le conoce como "cadena del habla". Esta cadena está formada por un mecanismo productor del habla localizado en el locutor, un medio de transmisión que puede ser el aire, y un proceso de percepción realizado por los oídos y cerebro del escucha.

En muchas aplicaciones de procesamiento del habla, una parte de la cadena se implanta por medio de un dispositivo de simulación. La síntesis o generación automática de habla por medio de algoritmos puede tomar el papel del locutor, mientras que en reconocimiento automático del habla o del locutor, un algoritmo toma el papel del escucha y descifra las ondas de voz con el fin de obtener el mensaje textual subyacente, o una hipótesis concerniente a la identidad del locutor.

Una frase bastante utilizada entre científicos dedicados al tratamiento de la voz, es que "El problema del procesamiento de voz es como manejar un tubo de pasta dental; la síntesis de voz es como exprimir la pasta fuera del tubo, mientras que el reconocimiento de voz es como tratar de ponerla adentro [Bristow, 1986]". La ciencia ficción nos ha inducido a creer que eventualmente habrá máquinas que puedan entender cualquier mensaje, hablado por cualquier persona, en cualquier idioma, y que además lo harán usando únicamente voces

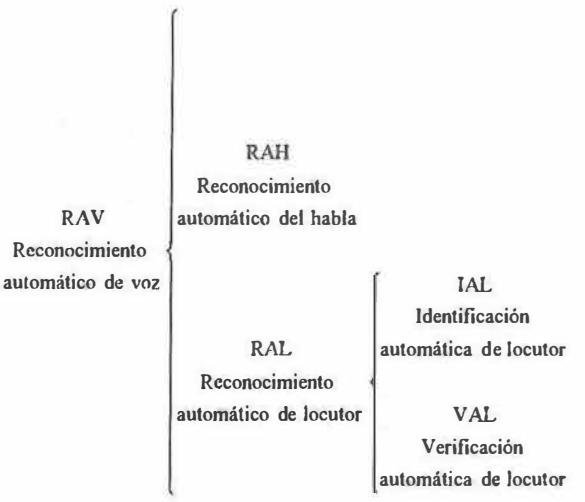
mecánicas y monótonas. Los autores de tal falacia, la han planteado de esta manera para que no quede duda acerca de quién habla, un robot o un humano, y al mismo tiempo sin el deseo de limitar el vocabulario del diálogo hombre-máquina. Sin embargo, la analogía del tubo de pasta dental no nos deja duda de que la situación es a la inversa.

Con la tecnología actual es posible darle a una máquina la capacidad de detectar (reconocer) en que momento son habladas ciertas palabras; esta tarea es conocida como "RECONOCIMIENTO AUTOMÁTICO DEL HABLA (RAH)". Sin embargo, el número de palabras que se pueden reconocer hasta ahora es limitado, y se requiere además un proceso de entrenamiento previo al uso del sistema.

Si se requiere que cualquier locutor pueda usar el sistema de reconocimiento del habla, se estará hablando de un sistema *independiente del locutor*, caracterizado por tener un vocabulario severamente limitado. Si por otra parte, existe sólo un pequeño grupo de usuarios del sistema, la tasa de error será mucho menor que en el caso anterior, lo cual permitirá aumentar el tamaño del vocabulario. Estos sistemas de reconocimiento automático del habla son *dependientes del locutor*.

Por otra parte, con la tecnología actual también es posible darle a una máquina la capacidad de hacer un juicio razonablemente exacto acerca de quién, de una lista de potenciales locutores, está hablando (Identificación automática de locutor (IAL)) o si un locutor es realmente quien dice ser (Verificación automática de locutor (VAL)).

Para propósitos de nomenclatura, "Reconocimiento Automático de Voz (RAV)" se puede dividir en "Reconocimiento Automático del Habla (RAH)" y "Reconocimiento Automático de Locutor (RAL)". A su vez, esta última tarea se subdivide en "Identificación Automática de Locutor (IAL)" y "Verificación Automática de Locutor (VAL)". El esquema general para RAV queda de la siguiente manera:



Ha habido mucho menos trabajo de investigación en RAL que en RAH, debido a que se han encontrado más aplicaciones para el segundo y existe menor conocimiento acerca de cuales aspectos de la señal de voz identifican a un locutor, comparado con los aspectos acústico-fonéticos de RAH [O'Shaughnessy, 1987].

Este trabajo tiene como propósito analizar el problema de reconocimiento automático de locutor y las técnicas que han sido utilizadas para tratar de resolverlo. La meta de este trabajo de tesis es diseñar un sistema de RAL que funcione en tiempo real, a partir de una plataforma basada en un procesador digital de señales (PDS) integrado en una tarjeta prototipo para computadora personal. Debido a la naturaleza multidisciplinaria del problema, queda fuera del alcance de esta tesis profundizar en cada uno de los aspectos colaterales del RAL.

En el capítulo II se analiza el problema de RAL en sus diferentes modalidades, así como sus aplicaciones; además de algunos métodos subjetivos para reconocimiento de locutor como marco de comparación para los métodos automáticos.

En el capítulo III se analiza el proceso de producción de la voz, así como los modelos y parámetros de la señal de voz que han sido utilizados en RAL, incluyendo aquellos que

pueden ser obtenidos por medio de predicción lineal. En el capítulo IV se analizan tres técnicas que han sido aplicadas en RAL y que han producido resultados satisfactorios.

En el capítulo V se describe el diseño de un sistema experimental para RAL *dependiente del vocabulario*, el cual está basado en una técnica de alineamiento dinámico en el tiempo (ADT) analizada en el capítulo IV; el sistema, llamado SERAL, utiliza como parámetros los coeficientes *cepstrum* expuestos en el capítulo III.

En el capítulo VI se describen los experimentos y resultados obtenidos con el sistema desarrollado para una base de datos (locutores). En el capítulo VII se describen las conclusiones de este trabajo y se ofrecen algunas sugerencias para el trabajo futuro con el sistema desarrollado o con otros por desarrollar.

CAPÍTULO II ANÁLISIS DEL PROBLEMA DE RAL.

II.1 Presentación del problema.

El reconocimiento automático de locutor es un ejemplo de **identificación biométrica personal,** término empleado para diferenciar técnicas que basan la identificación en ciertas características intrínsecas de la persona, tales como la voz, huellas dactilares, huellas retinales o estructura genética, de aquellas que usan artificios tales como llaves, emblemas, tarjetas magnéticas o claves de acceso memorizadas [Doddington, 1985].

Esta distinción le confiere a las técnicas biométricas una mayor confiabilidad y quizá hasta infalibilidad, ya que se presume que las características biométricas intrínsecas son únicas en cada ser humano; además, de que los atributos biométricos no pueden ser perdidos u olvidados, y no necesitan por lo tanto ser recordados. Así, la motivación principal para estudiar **reconocimiento de locutor** es poder realizar una identificación personal más confiable, usando la voz como medio de identificación, teniendo la ventaja de que ésta puede realizarse aún a distancia (v.g., por teléfono).

El reconocimiento de locutor puede ser aplicado en sistemas de seguridad tales como: control de acceso físico (la cerradura de una puerta o un sistema de ignición de automóvil activados por voz), control de acceso a datos de computadora o control automático de transacciones telefónicas (reservaciones aéreas o acceso bancario por teléfono). Existen también aplicaciones que requieren de la identificación de una persona, tales ejemplos caen en el campo de la criminalística y el procesamiento automatizado de información de reconocimiento.

II.2 Verificación e identificación de locutor

Reconocimiento de locutor es un concepto genérico que se refiere a cualquier tarea encaminada a realizar una discriminación de la identidad de las personas, basada en las características de la voz; dentro de esta descripción general caben dos tareas que han sido

estudiadas exhaustivamente; éstas son conocidas como **identificación de locutor** y **verificación de locutor**. Ambas usan una base de datos con patrones o modelos de referencia almacenados para una población de *N* locutores; estos locutores son llamados "clientes ", mientras que todos aquellos locutores que no tienen dicho modelo o patrón son llamados "impostores".

La tarea de **identificación de locutor** consiste en clasificar y etiquetar (asígnar un nombre o apelativo) una muestra de voz como "perteneciente a" o "siendo hablada por" un locutor, entre los N posibles locutores de referencia, mientras que **verificación de locutor** es la tarea de decidir si una muestra de voz no etiquetada (desconocida), "pertence a" o "fue hablada por" un locutor específico cuyo modelo o patrón de referencia se encuentra en la base de datos. **Verificación de locutor** sólo puede dar dos resultados:

1) La muestra de voz verificada es **aceptada** como "perteneciente" al locutor que afirmó ser su "dueño", esto es: El locutor si es quién dice ser.
2) La muestra de voz verificada es **rechazada** como "perteneciente"al locutor que afirmó ser su dueño, esto es: El locutor no es quién dice ser.

Por otra parte, **identificación de locutor** dará como salida el nombre, apelativo o etiqueta del locutor cuyo modelo de referencia, entre los N posibles, se parezca más a la muestra de voz que se está identificando.

Debe notarse que esta concepción de *identificación de locutor* siempre dará como resultado un nombre, apelativo o etiqueta, aún si la muestra en cuestión fue hablada por un "impostor", por lo tanto, está connotación de *identificación de locutor* debe ser utilizada solamente para identificar "clientes". En el caso de desear hacer *identificación de locutor* tanto con "clientes" como con "impostores", se debe incluir una respuesta de "no identificado" para aquellas muestra de voz habladas por impostores [Soong *et al*, 1987].

Las técnicas de análisis y decisión utilizadas en identificación y verificación de locutor

rechazo. Un *falso rechazo* ocurre cuando un "cliente" que está siendo verificado es rechazado, mientras que una *falsa aceptación* ocurre cuando en respuesta a la verificación de un "impostor" se da una "aceptación". Los errores en *identificación de locutor* son **falsa identificación** y **falsa no identificación**. La *falsa identificación* ocurre cuando se identifica a una persona equivocada, ya sea "cliente" o "impostor"; la *falsa no identificación* ocurre cuando se responde "no identificado" a un "cliente".

Con el fin de establecer un parámetro normalizado para medir la tasa de error, muchos investigadores ajustan el umbral de decisión de tal manera que las dos clases de errores, tanto en *identificación* como en *verificación*, ocurran en igual proporción, lo cual genera una tasa de equierror (TEE). La eficiencia de la tarea de reconocimiento de locutor se mide con la tasa de error, que en general es más alta para identificación que para verificación [Doodington, 1985].

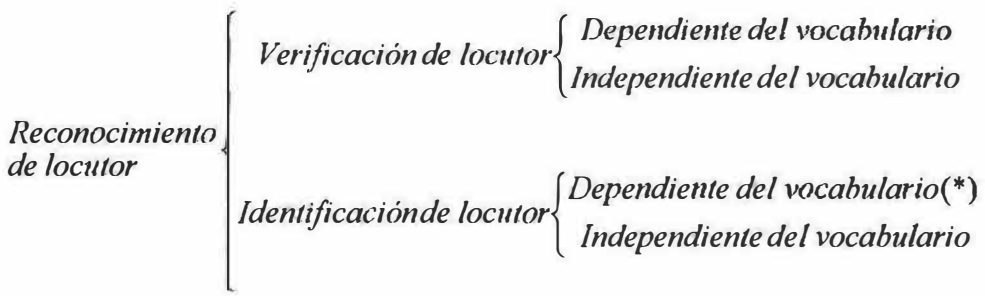
II.3 Dependencia e independencia del vocabulario.

El grado de control sobre la forma en que es generada la muestra de voz que se utiliza para reconocer al locutor es muy importante. Entre los puntos por considerar destaca el que la muestra corresponda siempre a una palabra o frase establecida con anterioridad, o bien, que ésta contenga cualquier palabra o frase posible. En aquellos casos en que el locutor es cooperativo y desea ser reconocido, es posible pedirle que pronuncie una frase establecida previamente y que ha de ser utilizada siempre que el locutor desee ser reconocido, esta modalidad del *reconocimiento de locutor* se designa **dependiente del vocabulario**; en aquellos casos en que tal situación no es posible, ya sea porque el locutor no es cooperativo, o porque el reconocimiento debe hacerse subrepticiamente, es necesario realizar reconocimiento de locutor con vocabulario libre. Esta última modalidad es llamada **independiente del vocabulario**, y en ella el locutor puede pronunciar en teoría cualquier palabra o frase; sin embargo, en la práctica esto no se aplica y los experimentos de

reconocimiento de locutor independiente del vocabulario se realizan empleando un conjunto finito de palabras o frases (v.g., los dígitos [Soong *et al*, 1987][Tishby, 1991]).

En general, el *reconocimiento de locutor dependiente del vocabulario* exhibe una mayor eficiencia que su contraparte *independiente del vocabulario*, ésto debido a que la capacidad para controlar el vocabulario se extiende generalmente también al locutor y al ambiente que lo rodea.

La característica de *dependencia del vocabulario* es muy empleada en *verificación de locutor* y raramente en *identificación de locutor* [O'Shaughnessy, 1987], además de que si se desea hacer *identificación de locutor dependiente del vocabulario*, es necesario que todos los locutores utilicen la misma palabra o frase con el fin de que la variabilidad debida a diferentes frases no ejerza influencia alguna, y la tarea pueda ser condiderada puramente de *identificación de locutor*. Lo anterior puede ser sintetizado en el siguiente esquema:



(*) Sólo si todos los locutores usan la misma palabra o frase.

II.4 Entrenamiento y prueba (reconocimiento).

En todas las tareas de reconocimiento de locutor están presentes dos etapas: **entrenamiento y prueba**. La etapa de *entrenamiento* siempre se realiza antes que la de *prueba* y consiste en crear u obtener los modelos o patrones de referencia que forman la base de datos de "clientes"; esto se hace utilizando una o varias frases habladas por los

mismos. La etapa de *prueba*, por su parte, es propiamente la de reconocimiento (en sus modalidades de verificación e identificación) y muy frecuentemente es utilizada para hacer una caracterización de la eficiencia de la tarea en términos de tasa de error.

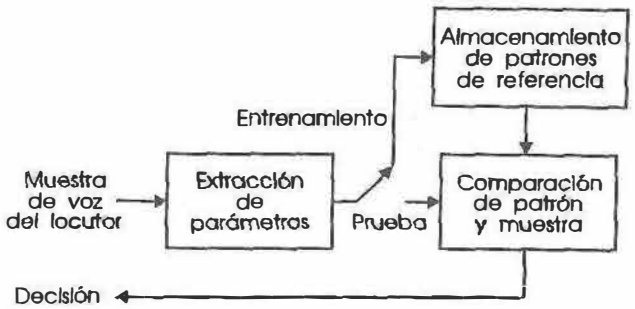


Figura 1. Esquema general de reconocimiento de locutor.

La figura 1 muestra el esquema general para reconocimiento de locutor, mientras que las figura 2 y 3 muestran los esquemas particulares para identificación y verificación de locutor respectivamente, en las tres figuras mencionadas se hace referencia a las etapas de entrenamiento y prueba.

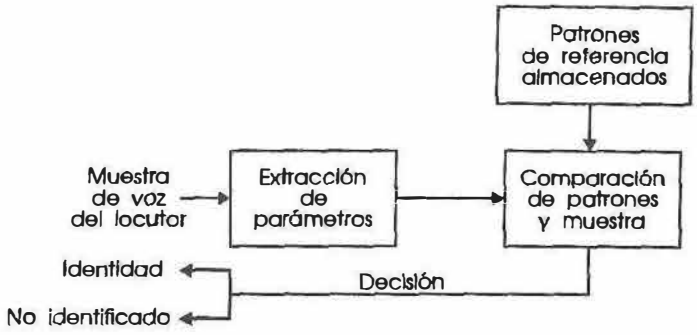


Figura 2. Esquema general de identificación de locutor.

A causa de la gran variabilidad acústico-fonética de las diferentes palabras o frases, se requiere más material de entrenamiento y prueba para *reconocimiento de locutor independiente del vocabulario* que en el caso *dependiente del vocabulario*, para obtener

tasas de error similares; la duración neta de las muestras de entrenamiento frecuentemente excede los 30 segundos en entrenamiento y 5 segundos en prueba para el primer caso, mientras que en este último 2 segundos de entrenamiento y 0.5 segundos de prueba producen resultados satisfactorios.

Por otra parte, para ambos casos *dependiente e independiente del vocabulario*, el tipo de sonidos contenidos en los vocablos de entrenamiento y prueba está correlacionado con la eficiencia de la tarea de reconocimiento de locutor, en particular, los vocablos con sonidos nasales han mostrado una tasa de error más baja en varios reportes [Atal, 1974][Soong *et al*, 1987][Tishby, 1991], incluyendo los resultados de los experimentos de esta tesis reportados en el capítulo V.

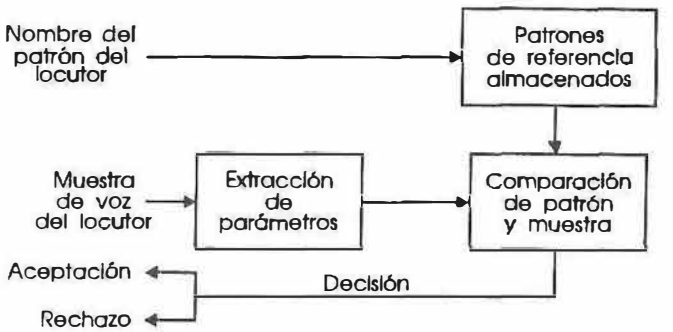


Figura 3. Esquema general de verificación de locutor.

II.5 Reconocimiento subjetivo de locutor.

Entre las fuentes de información que identifican a un locutor se incluye información de alto nivel tal como dialecto, contexto o materia y estilo de hablar (uso de estructuras sintácticas o lingüísticas específicas); esa información de alto nivel es muy valiosa como ayuda para el reconocimiento de locutor por escuchas, pero no ha sido posible utilizarla en sistemas automáticos de reconocimiento de locutor, debido a las dificultades prácticas para adquirirla y manipularla [Doddington, 1985].

Los sistemas de RAL se enfocan en características de bajo nivel tales como: amplitudes espectrales, frecuencia fundamental, frecuencias formantes, anchos de banda, etc. (las características de bajo nivel más utilizadas serán analizadas en el siguiente capítulo). Estas variables pueden ser medidas como función del tiempo o promediadas estadísticamente para ser establecidas como parámetros de un locutor.

Entre las características anatómicas que distinguen a los locutores destacan la forma del tracto y de las cuerdas vocales; sin embargo, a pesar de que la mayoría de los parámetros o características utilizadas en procesamiento de voz contienen información útil para la identificación de ambos, locutor y mensaje hablado, la señal de voz no conlleva características anatómicas detalladas de manera explícita, ésto distingue al reconocimiento de locutor de la identificación de huellas digitales o patrones retinales; ya que éstas son características físicas estáticas, mientras que el reconocimiento de locutor y de la firma caligráfica, usan características de eficiencia dinámica que dependen del acto de hablar y firmar respectivamente[Doddington, 1985].

II.5.1 Reconocimiento de locutor por escuchas.

El reconocimiento de locutor por escuchas (humanos) ha sido estudiado ampliamente con la intención de conocer la manera en la que los escuchas pueden reconocer locutores, qué variables afectan la eficiencia de los escuchas, y cuáles son las bases perceptuales del reconocimiento de locutor. Estas evaluaciones pueden servir como marco de referencia para estimar la eficiencia de los sistemas automáticos.

Sin embargo, es muy difícil establecer comparaciones objetivas entre escuchas y métodos automáticos cuando se trata de reconocimiento de locutor, sobre todo tomando en cuenta que el escucha puede hacer una selección bastante amplia de la estrategia y el ambiente de desempeño, por ejemplo, los escuchas pueden reconocer locutores sin una comparación explícita de dos muestras de voz (la cual es la forma natural de reconocer

locutores con la que todos estamos familiarizados), y más aún, no requieren que el locutor pronuncie una frase determinada, además de que pueden hacer uso de la información de alto nivel mencionada anteriormente.

Debe tomarse en cuenta además, que los escuchas son robustos reconocedores de locutores aún bajo condiciones adversas tales como ruido o distorsión espectral, que reconocen con más facilidad a los locutores que les son más familiares, y que si bien se ha determinado que la banda octava de frecuencia de 1 a 2 kHz es la más útil para los escuchas, éstos sólo han duplicado la tasa de error cuando la señal es severamente filtrada pasa-altas a 2 KHz o pasa-bajas a 1 KHz [Doddington, 1985].

Se puede afirmar, sin embargo, que bajo algunas condiciones especiales tales como: dependencia del vocabulario, bajo nivel de ruido, no degradación de la señal, y pocas muestras de entrenamiento, los métodos automáticos presentan una eficiencia mayor o igual a la de los escuchas [Atal, 1976] y [Doddington, 1985].

II.5.2 Reconocimiento de locutor usando espectrogramas.

Otro de los métodos subjetivos que se han aplicado para reconocer locutores es el análisis visual de espectrogramas. Para conocer el concepto de espectrograma es necesario primero analizar el concepto de **transformada de Fourier dependiente del tiempo (TFDT)**, a partir de la **transformada discreta de Fourier (TDF).**

Considérese una secuencia de longitud finita $x[n]$, que es cero fuera del intervalo $0 \le n \le N-1$, la transformada discreta de Fourier (TDF) de $x[n]$ está dada por:

$$X[k] = \sum_{n=0}^{N-1} x[n]\, e^{-j\frac{2\pi}{N}kn} \qquad (4)$$

donde $X[k]$ es la TDF de la secuencia finita $x[n]$, y corresponde a un conjunto de muestras igualmente espaciadas de la transformada de Fourier de $x[n]$, v.g.,

$$X[k] = X[e^{j\omega}]\Big|_{\omega=\frac{2\pi k}{N}} \tag{5}$$

Esta representación de Fourier es útil para obtener una representación de $x[n]$ en el dominio de la frecuencia, bajo la suposición de que las propiedades de la señal no varían con el tiempo; sin embargo, en algunas situaciones prácticas ésto no se cumple. Un ejemplo de tales situaciones son las señales de voz, donde una simple estimación de la TDF no es suficiente para describirlas, y hace necesario introducir el concepto de **transformada de Fourier dependiente del tiempo (TFDT)**, también conocida como la **transformada de Fourier de tiempo corto (TFTC)** (el hecho de que las señales de voz sigan un modelo no-estacionario será analizado en el capítulo III).

La TFDT de $x[n]$ está definida como:

$$X[n,\omega) = \sum_{m=-\infty}^{\infty} x[n+m]\, w[m]\, e^{-j\omega m} \tag{6}$$

donde $w[m]$ es una secuencia ventana. La notación $X[n,\omega)$ se utiliza para remarcar que n es una variable discreta y ω una variable continua.

En la ecuación (6) la señal unidimensional $x[n]$, que es función de una variable discreta, es convertida en una función bidimensional de la variable temporal n y la variable frecuencial ω, siendo esta última continua. Debe notarse que la TFDT es periódica con respecto a ω con periodo 2π.

Sin embargo, para fines computacionales el cálculo de $X[n,\omega)$ sólo puede hacerse para un número finito de valores de ω, correspondiente a un muestreo de la TFDT en el dominio de la frecuencia. Por ejemplo, supóngase que la ventana tiene longitud L con las muestras comenzando en m=0, v.g.,

$$w[m] = 0 \quad \textit{fuera del intervalo} \quad 0 \le m \le L-1 \tag{7}$$

Si se muestrea $X[n,\omega)$ a N frecuencias igualmente espaciadas $\omega_k = 2\pi k / N$ con $N \ge L$, entonces es posible recuperar la secuencia original a partir de la TFDT muestreada.

Específicamente, si se define $X[n,k]$ como:

$$X[n,k] = X[n, 2\pi k / N] = \sum_{m=0}^{L-1} x[n+m]\, w[m]\, e^{-j(2\pi/N)km}\ ,\ 0 \le k \le N-1 \qquad (8)$$

entonces $X[n,k]$ es la TDF de la secuencia $x[n+m]w[m]$. Usando la TDF inversa,

$$x[n+m]w[m] = \frac{1}{N}\sum_{k=0}^{N-1} X[n,k]\, e^{j(2\pi/N)km}\ ,\ 0 \le m \le L-1 \qquad (9)$$

y ya que se ha supuesto que la ventana $w[m] \neq 0$ para $0 \le m \le L-1$, los valores de la secuencia pueden ser recuperados en el intervalo comprendido desde n hasta (n+L-1), usando

$$x[n+m] = \frac{1}{Nw[m]}\sum_{k=0}^{N-1} X[n,k]\, e^{j(2\pi/N)km}\ ,\ 0 \le m \le L-1 \qquad (10)$$

El punto importante es que la ventana tiene longitud finita y que se han tomado al menos tantas muestras en la dimensión ω, como muestras hay en la ventana, v.g., $N \ge L$. En general, para la región de soporte de la ventana de la ecuación (7), se puede definir la TFDT muestreada como:

$$X[rR,k] = X[rR, 2\pi k / N) = \sum_{m=0}^{L-1} x[rR+m]\, w[m]\, e^{-j(2\pi/N)km} \qquad (11)$$

dónde r y k son enteros tales que $-\infty < r < \infty$ y $0 \le k \le N-1$, y R especifica el paso de recorrimiento, medido en muestras de la señal $x[n]$ a través de la ventana $w[m]$.

La TFDT es bastante útil en los modelos no estacionarios que siguen las señales de voz, ya que ésta es producida por la excitación de un tubo acústico llamado **tracto vocal**, terminado en un extremo por los labios y en otro por la glotis, que es donde se encuentran las cuerdas vocales.

En voz normal, la forma del tracto vocal cambia de una manera relativamente lenta con el tiempo, a medida que los labios y la lengua hacen los gestos de la voz; así, el tracto vocal puede ser modelado como un un filtro variante en el tiempo que impone sus propiedades de respuesta a la frecuencia sobre el espectro de la señal de excitación

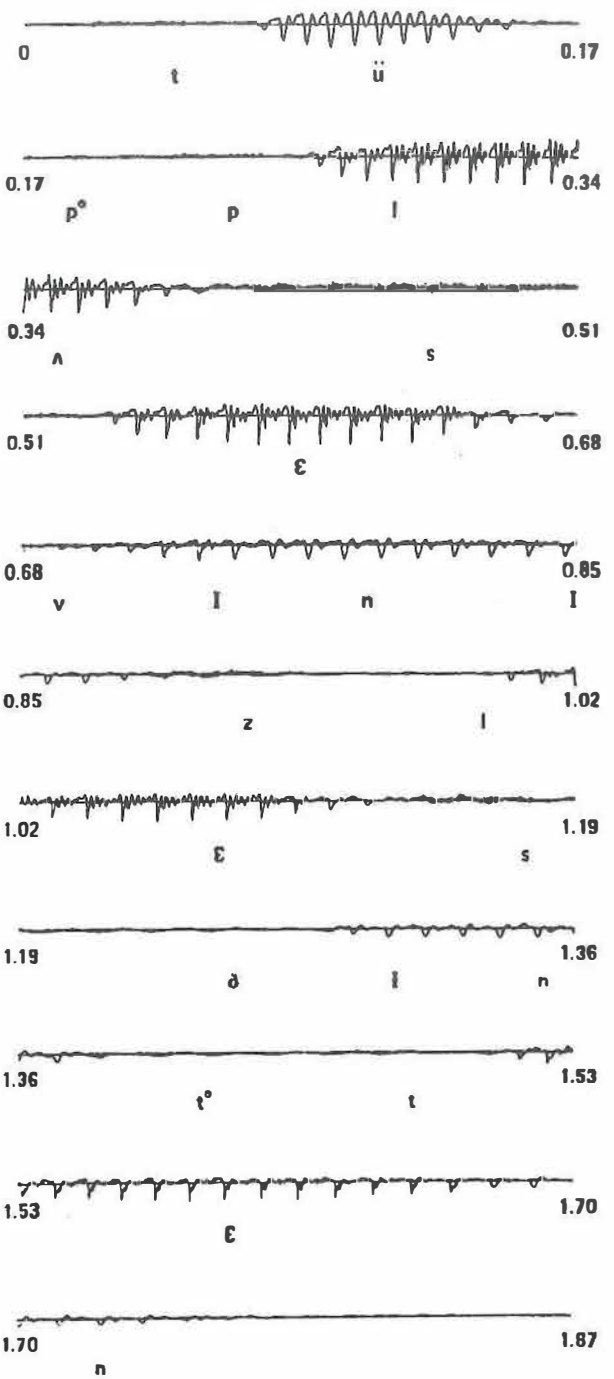


Figura 4. Forma de onda de la frase "Two plus seven is less than ten" (escala en segundos). Debajo de la forma de onda se muestra la transcripción fonética alineada temporalmente [Oppenheim y Schafer, 1989].

De la breve descripción anterior del proceso de producción de la voz, la cual será ampliada en el siguiente capítulo, se puede ver que la voz es una señal no-estacionaria; sin

embargo, observando la figura 4 se puede suponer que las características de la señal de voz permanecen constantes sobre intervalos de tiempo del orden de 30 a 40 ms.

La figura 4 es la forma de onda de la frase "Two plus seven is less than ten", e ilustra el hecho de que ésta consiste de una secuencia de sonidos cuasiperiódicos, interpuestos con segmentos que parecen ruido, de manera que si la longitud L de la ventana no es demasiado grande, las propiedades de la señal no cambiarán dentro de la ventana.

Así, la TDF de un segmento de voz ponderado por una ventana, despliega las propiedades frecuenciales de la señal dentro de la ventana. Por ejemplo, si la longitud de la ventana es suficientemente larga para incluir varios ciclos de un segmento cuasiperiódico, la TDF de ese segmento debe mostrar una serie de picos a múltiplos enteros de la frecuencia fundamental de la señal en ése intervalo; si la ventana es muy corta, no podrán apreciarse las armónicas, pero la forma espectral general permanecerá evidente. La selección de la longitud de la ventana es un compromiso entre resolución temporal y resolución frecuencial, si la ventana es muy larga las propiedades de la señal pueden cambiar demasiado a través de la ventana, y si ésta es muy corta se perderá la resolución de componentes de banda angosta.

La figura 5 muestra el despliegue de un par de **espectrogramas** de la forma de onda de la figura 4, éstos fueron obtenidos calculando la TFDT de la misma señal. En la escala horizontal de cada espectrograma se encuentra el eje del tiempo n, y en la vertical el eje de la frecuencia f; la magnitud de la TFDT se representa por las marcas oscuras, a mayor oscuridad corresponde mayor magnitud, debajo de cada espectrograma se presenta también la forma de onda.

La figura 5a en particular es un **espectrograma de banda ancha**, el cual resulta de una ventana relativamente corta y se caracteriza por su pobre resolución frecuencial, y una buena resolución temporal. Ya que la tasa de muestreo fue 16 KHz, $\omega = \pi$ corresponde a 8 KHz; el espectrograma fue calculado usando una ventana Hamming

$$w[m] = 0.54 - 0.46\cos\left[\frac{2\pi m}{L-1}\right] \qquad (12)$$

con una duración de 6.7 ms, lo cual corresponde a L=108; el valor de R fue 16, lo que representa incrementos de 1 ms.

Las barras oscuras horizontales a través del espectrograma corresponden a resonancias del tracto vocal, el cual como se puede ver cambia con el tiempo. La apariencia verticalmente estriada del espectrograma se debe a la naturaleza cuasiperiódica de las porciones sonoras de la forma de onda , que se hacen evidentes al comparar las variaciones de ésta con las variaciones del espectrograma, es decir, ya que la longitud de la ventana de análisis es del orden de la longitud de un periodo de la forma de onda, a medida que la señal se desliza por la ventana, ésta cubre alternativamente segmentos de alta y baja energía, produciendo las estrías verticales durante los segmentos sonoros.

La figura 5b muestra un **espectrograma de banda angosta**, donde la longitud de la ventana $w[m]$ es mayor, lo cual provee un aumento en la resolución frecuencial, pero un decremento en la resolución temporal. En este caso la longitud de la ventana Hamming fue 45 ms que corresponde a L=720; el valor de R fue 16. Si bien las resonancias o frecuencias formantes siguen evidentes, no se observan las estrías verticales en las porciones sonoras, y en cambio se pone en evidencia la frecuencia fundamental y sus armónicas. Las regiones no sonoras se caracterizan por la ausencia de periodicidad en el dominio de la frecuencia

El espectrograma fue introducido en 1962 por Kersta como un medio de identificación personal (ver [Doddington, 1985]), y fue en su reporte dónde surgió el concepto de **voz impresa**. Los experimentos de Kersta consistieron en identificar visualmente la voz impresa (espectrogramas) de 12 locutores de referencia; la eficiencia observada por Kersta fué de un 99%. Desafortunadamente esa alta eficiencia no ha sido observada en experimentos posteriores bajo un ambiente real, y el problema es que diferentes muestras de un mismo locutor suenan, y se ven espectrográficamente diferentes

(muy diferentes con mucha frecuencia); este fenómeno es llamado **varianza intralocutor**.

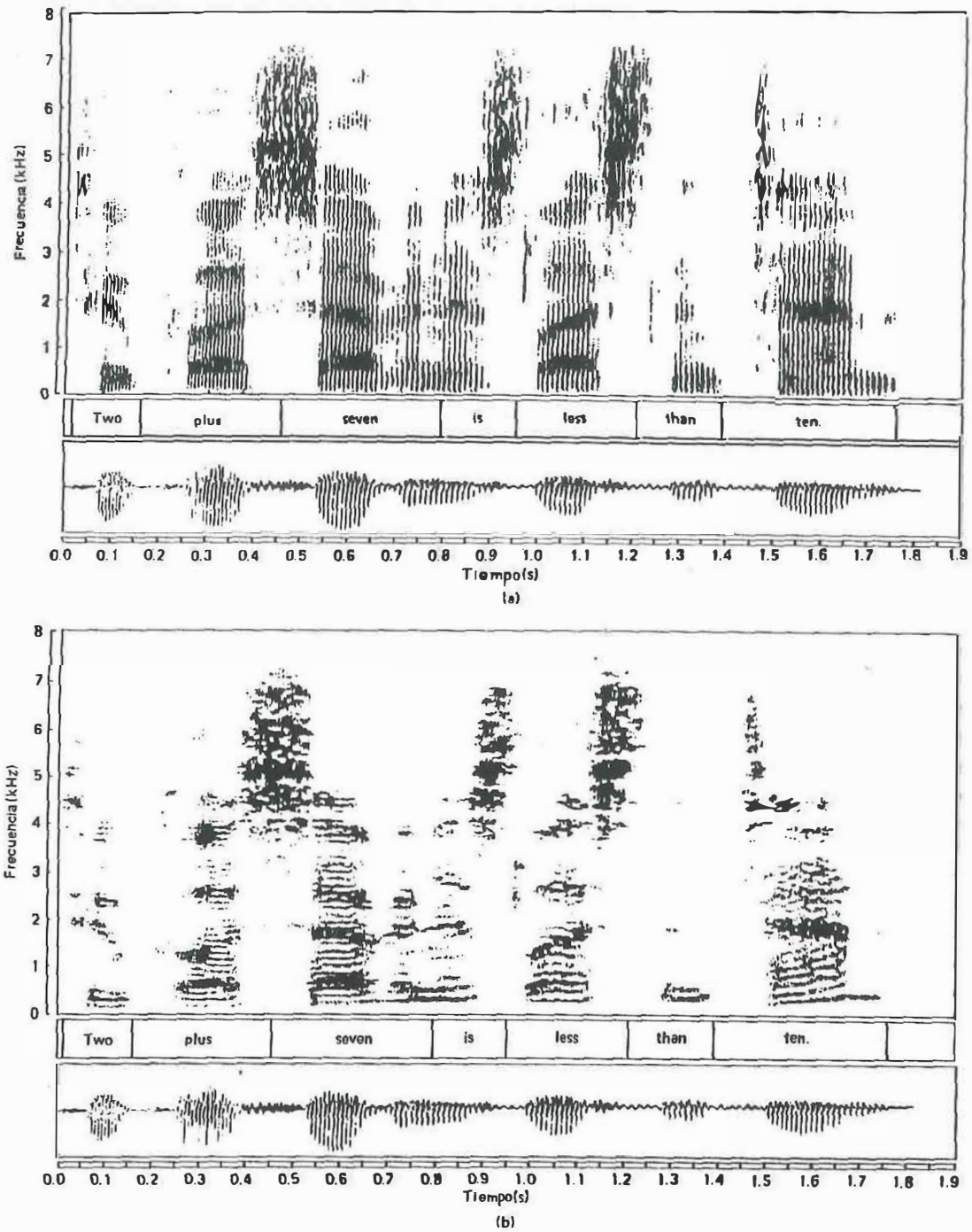


Figura 5. (a) Espectrograma de banda ancha de la forma de onda de la figura 5.
(b) Espectrograma de banda angosta. [Oppenheim y Schafer, 1989]

La figura 5b muestra un **espectrograma de banda angosta**, donde la longitud de la

ventana $w[m]$ es mayor, lo cual provee un aumento en la resolución frecuencial, pero un decremento en la resolución temporal. En este caso la longitud de la ventana Hamming fue 45 ms que corresponde a L=720; el valor de R fue 16. Si bien las resonancias o frecuencias formantes siguen evidentes, no se observan las estrías verticales en las porciones sonoras, y en cambio se pone en evidencia la frecuencia fundamental y sus armónicas. Las regiones no sonoras se caracterizan por la ausencia de periodicidad en el dominio de la frecuencia

El espectrograma fue introducido en 1962 por Kersta como un medio de identificación personal (ver [Doddington, 1985]), y fue en su reporte dónde surgió el concepto de **voz impresa**. Los experimentos de Kersta consistieron en identificar visualmente la voz impresa (espectrogramas) de 12 locutores de referencia; la eficiencia observada por Kersta fué de un 99%. Desafortunadamente esa alta eficiencia no ha sido observada en experimentos posteriores bajo un ambiente real, y el problema es que diferentes muestras de un mismo locutor suenan, y se ven espectrográficamente diferentes (muy diferentes con mucha frecuencia); este fenómeno es llamado **varianza intralocutor**.

Los especialistas en procesamiento de voz tienden a ser muy críticos con la técnica de reconocimiento de locutor por medio de **voz impresa** por las siguientes razones:

a) La técnica no es un procedimiento bien definido, lo cual es un arte contrario a la ciencia.
b) La eficiencia de la identificación está fuertemente influenciada por condiciones específicas, que sin un modelo subyacente, no pueden ser predeterminadas.
c) Varias evaluaciones de la exactitud del método han sido equívocas.

Así, la confiabilidad de la técnica de voz impresa es bastante dudosa ya que está sujeta a muchas condiciones acústicas, ambientales y del locutor, además de que su uso es bastante cuestionable si se toma en cuenta que se puede obtener una eficiencia más alta a través del juicio auditivo de los escuchas.

CAPÍTULO III MODELOS Y PARÁMETROS DE LA SEÑAL DE VOZ UTILIZADOS EN RECONOCIMIENTO AUTOMÁTICO DE LOCUTOR (RAL).

III.1 El proceso de producción de la voz.

La figura 6 muestra un esquema del sistema vocal humano. El tracto vocal comienza en la apertura de las cuerdas vocales (glotis) y termina en los labios, y está formado por dos partes principales: la faringe (conexión del esófago a la boca) y la boca o cavidad oral; en el hombre la longitud total promedio del tracto vocal es 17 cm. El área transversal del tracto vocal, determinada por la posición de la lengua, labios, quijada y vélum, varía desde cero (completamente cerrado) hasta $20\ cm^2$ aproximadamente. El tracto nasal comienza en el vélum y termina en los orificios nasales; cuando el vélum se encuentra abierto (v.g., se encuentra abajo) el tracto nasal está acústicamente acoplado con el tracto vocal para producir los sonidos nasales de la voz.

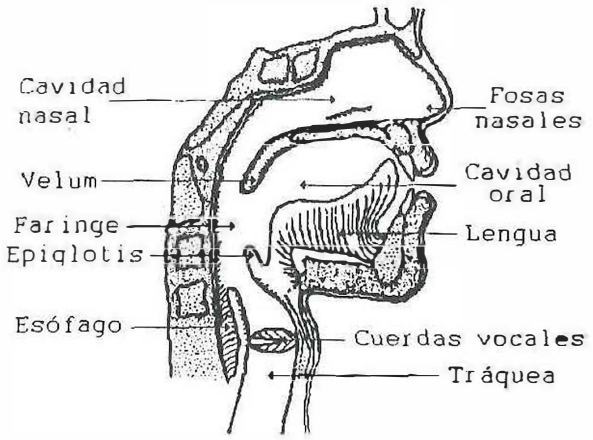


Figura 6. Esquema del sistema vocal humano.

El sistema productor de la voz puede ser dividido en tres grupos [Álvarez Constantino, 1977] :

a) Órganos de la respiración: pulmones, bronquios y tráquea.

b) Órganos de la fonación: laringe, cuerdas vocales y glotis.

c) Órganos de la articulación: labios, mandibula, dientes, vélum,

faringe, cavidad oral y cavidad nasal.

Previa inhalación de aire en los pulmones, la voz se produce generalmente cuando el aire es exhalado; al expanderse el diafragma, los órganos de la respiración proporcionan un flujo de aire a los órganos de la fonación, y es ahí donde el sonido adquiere sus características primarias. A éstas, se agregarán características adicionales impuestas por los órganos de la articulación al restringir el paso del aire en algún punto del tracto vocal.

Los sonidos de la voz pueden clasificarse desde el punto de vista de la excitación en tres grupos [Rabiner y Schafer, 1978]:

a) **Sonidos sonoros**, los cuales son producidos al forzar el aire a través de la glotis con la tensión de las cuerdas vocales ajustada de manera tal que se produce una oscilación relajada de las mismas. Ejemplos de tales sonidos son las vocales del español /a, e, i, o, u/.

b) **Sonidos fricativos**, que se producen al restringir el paso del aire en algún punto del tracto vocal (usualmente cerca de la boca), de tal manera que el aire fluye a una velocidad suficiente para producir una turbulencia; ésto crea una fuente de ruido de gran ancho de banda para excitar el tracto vocal. Ejemplos de tales sonidos se encuentar en /s, f/.

c) **Sonidos oclusivos**, que resultan de cerrar completamente el tracto vocal en algún punto (usualmente al frente de éste), creando una alta presión que al ser liberada por la reapertura del tracto vocal produce el sonido. Ejemplos de tales sonidos se hallan en /p, t, k/.

En el contexto de la producción de la voz las resonancias del tracto vocal son llamadas *frecuencias formantes o formantes* simplemente. Las *frecuencias formantes* dependen de la forma y dimensiones del tracto vocal; cada forma está caracterizada por un conjunto particular de frecuencias formantes. La dependencia del área transversal con la

distancia a lo largo del tracto vocal se designa como **función área**.

Para formar diferentes sonidos es necesario variar la forma del tracto vocal, y así, las propiedades espectrales de la señal de voz varían con el tiempo a medida que la forma del tracto vocal cambia. Las frecuencias formantes pueden ser identificadas como los máximos de la TFDT de la señal de voz, y se denotan como F1, F2, F3, etc., partiendo de la más baja frecuencia formante. El símbolo F0 es utilizado para denotar la frecuencia de oscilación de las cuerdas vocales en sonidos sonoros, también conocida como frecuencia fundamental (del inglés *pitch*). Debe notarse que las frecuencias formantes F1, F2, F3, etc., son múltiplos enteros de F0 ya que son armónicas de la frecuencia fundamental.

III.2 Modelo digital para la señal de voz.

El apéndice A de esta tesis contiene un breve análisis del modelado del tracto vocal utilizando tubos uniformes sin pérdidas, sin embargo, en aplicaciones que requieran del uso de estos modelos es necesario contar con los valores de los parámetros de la excitación y del tracto vocal con el fin de producir una señal de voz. En la práctica, la obtención de esos parámetros es bastante complicada, por lo cual se hace necesario utilizar modelos cuyos parámetros sean más simples de calcular.

El apéndice A ha sido incluido debido a la estrecha relación que existe entre los parámetros que caracterizan al modelo de tubos concatenados sin pérdidas y el modelo que se analizará a continuación, además de que constituye un buen antecedente para este último.

III.2.1 Modelo análogo a la salida.

La mayoría de las aplicaciones de voz se concentran en el modelado de los aspectos acústicos de la voz en vez de la dinámica del tracto vocal que la produce [O'Shaughnessy, 1987]; estos **modelos análogos a la salida** (del ingés *terminal analog models*) sirven para representar el proceso de producción de la voz en términos de su salida (del inglés

terminal). El modelo es así equivalente al modelo físico en su salida, pero su estructura interna no simila la física de la producción de voz.

La idea que se está manejando se encuentra representada en la figura 7, donde el modelo de tubos concatenados sin pérdidas del apéndice A, representado por los coeficientes de reflexión, es sustituído por un modelo análogo a la salida; este último modelo está representado por un sistema lineal cuya salida tiene las propiedades de la señal de voz, y es controlado por un conjunto de parámetros relacionados de alguna manera con el proceso de producción de la voz.

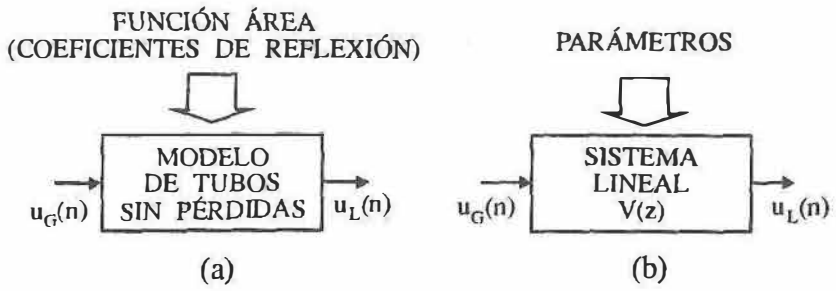


Figura 7. (a) Representación del modelo de tubos sin pérdidas;
(b) Modelo análogo a la salida.

Para producir una señal de voz con este modelo los parámetros y la excitación deben de cambiar con el tiempo, sin embargo, ya se ha señalado que las características de la señal de voz cambian de una manera relativamente lenta con el tiempo, y es razonable considerar que *V(z)* es un sistema lineal variando lentamente en el tiempo, y más aún, se puede considerar que las propiedades de la excitación y el tracto vocal permanecen fijas por períodos de 10-20 ms [Rabiner y Schafer, 1978].

Considérese que en el modelo de la figura 7b *V(z)* está dada como

$$V(z) = \frac{G}{1 - \sum_{k=1}^{N} \alpha_k z^{-k}} \qquad (13)$$

obtenida a partir de la ecuación (168) del apéndice A mediante la agrupación de los elementos del numerador en el factor *G*, y la eliminación del retardo fijo. Este modelo, que tiene únicamente polos y no tiene ceros, es útil para representar los efectos del tracto vocal en la mayoría de los sonidos de la voz, excepto nasales, los cuales requieren de resonancias (polos) y también antirresonancias (ceros).

Sin embargo, ya que los ceros de la función de transferencia del tracto vocal caen dentro del círculo unitario en el plano *z*, cada cero puede ser representado por múltiples polos [Atal y Hanauer, 1971]; esto equivale a demostrar que si $|a|<1$

$$1 - az^{-1} = \frac{1}{\sum_{n=0}^{\infty} a^n z^{-n}} \tag{14}$$

y por otra parte, la localización de un polo es considerablemente más importante perceptualmente que la localización de un cero, los cuales en la mayoría de los casos sólo contribuyen al balance espectral [Atal y Hanauer, 1971].

Ya que los coeficientes α_k en la ecuación (13) son reales, los polos de *V(z)* son reales o complejos conjugados; en el caso particular de los formantes, éstos están especificados por un par de polos complejos conjugados, los cuales se denotan como

$$s_k, s_k^* = -\sigma_k \pm j2\pi F_k \tag{15}$$

que al ser trasladados al plano z se convierten en

$$z_k, z_k^* = e^{-\sigma_k T \pm j2\pi F_k T} \tag{16}$$

dónde el valor del formante *k* es F_k (Hz), y su ancho de banda es σ_k / π (Hz) [Rabiner y Schafer, 1978].

La especificación completa del *modelo análogo a la salida* requiere, además del tracto vocal, considerar la representación de una excitación cambiante y el efecto de radiación en los labios. Considerando primero la radiación en los labios, hay que notar que el

modelo de las ecuaciones (168) [del apéndice A] y (13) relaciona la velocidad del volumen en los labios con la velocidad del volumen en la glotis. Sin embargo, usualmente se desea un modelo para la presión en los labios, para ello hay que hacer uso de la ecuación (151) del apéndice A, que en caso de un modelo discreto es

$$P_L(z) = R(z)\,U_L(z) \tag{17}$$

La impedancia de radiación *R(z)* es usualmente modelada como una primera diferencia [Rabiner y Schafer, 1978][O'Shaughnessy, 1987], esto es

$$R(z) = R_0(1 - z^{-1}) \tag{18}$$

Para el caso de la excitación hay que recordar que la mayoría de los sonidos de la voz pueden ser clasificados como sonoros o sordos; bajo esta consideración, la generación de sonidos sonoros se realiza excitando un modelo *G(z)* de la glotis con un tren de impulsos unitarios separados un intervalo de tiempo igual al período fundamental *T0*. Los elementos para la producción de sonidos sonoros pueden observarse en la parte superior de la figura 8, dónde se ha incluído un factor *As* para controlar la amplitud de los pulsos de excitación.

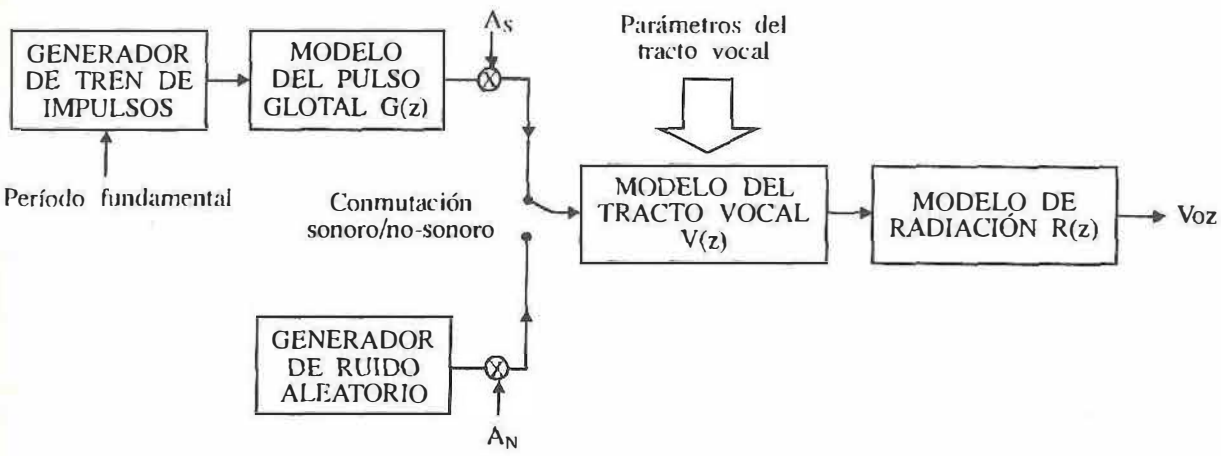


Figura 8. Modelo típico análogo a la salida para la producción de voz.

Para el caso de los sonidos sordos la excitación se obtiene conmutando hacia un generador de ruido aleatorio con espectro de frecuencia plano [O'Shaughnessy, 1987], el

cual es aplicado directamente al modelo del tracto vocal *V(z)*, previa multiplicación por el factor de ganancia A_N, según se observa en la parte inferior de la figura 8. La distribución de probabilidad de las muestras de ruido no parece ser crítica [Rabiner y Schafer, 1978].

Algunas de las limitaciones de este modelo son: no es posible la generación de sonidos fricativos sonoros debido a que las excitaciones son excluyentes, además de que la generación de sonidos transientes, tales como los oclusivos, no es demasiado buena pero sigue siendo aceptable [Rabiner y Schafer 1978]. Sin embargo, la simplicidad del modelo justifica su uso tan difundido [O'Shaughnessy, 1987].

III.2.2 Cálculo de los parámetros por medio de codificación predictiva lineal (CPL).

El modelo presentado en la figura 8 puede ser sustituído por el de la figura 9, donde los efectos combinados de *G(z), V(z) y R(z)* han sido agrupados en *H(z)*; y el parámetro *G* está relacionado con la ganancia A_S o A_N según sea el caso. Bajo estas consideraciones *H(z)* es un filtro digital variando lentamente en el tiempo, cuya función de transferencia es

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{19}$$

dónde *s(n)* es la salida del filtro y *u(n)* es la excitación, esta última puede establecerse en forma de impulsos periódicos o ruido aleatorio. Un aspecto importante de este modelo es que la excitación puede considerarse espectralmente plana, de tal manera que los detalles espectrales más importantes de *s(n)* estarán dados por *H(z)*. A partir de la figura 9 puede establecerse que el modelo está caracterizado por una decisión *sonoro/sordo*, el período fundamental *T0*, el parámetro de ganancia *G*, y los *p* coeficientes a_k del filtro digital.

Una cuestión importante que puede ser analizada ahora es el orden *p* del predictor. Si se desea una correspondencia exacta entre *H(z)* y *S(z)* el orden del predictor debe ser

infinito (ya que la excitación se considera de espectro plano), obviamente este resultado es imposible de obtener debido a limitaciones de memoria y tiempo de cómputo. En la práctica el orden de predicción es dependiente de la tasa de muestreo, y se ha demostrado que con el fin de modelar adecuadamente las resonancias del tracto vocal, la memoria del predictor debe ser de al menos dos veces el tiempo requerido para que las ondas sonoras viajen desde la glotis hasta
los labios (o fosas nasales) [Atal y Hanauer, 1971].

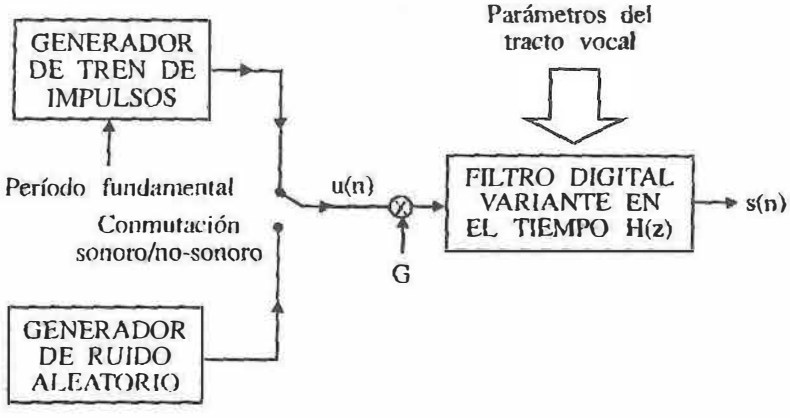


Figura 9. Diagrama simplificado del modelo para la producción de voz.

Si se supone que la longitud desde la glotis hasta los labios es 17 cm, el tiempo t que tarda la onda en viajar esa distancia es

$$t = \frac{l}{c} = \frac{0.17\ m}{340\ m/s} = 0.5\ ms \tag{20}$$

de tal manera que el orden p está dado por

$$p = 2\frac{0.5\ ms}{T} = 1x10^{-3}\ f_m \tag{21}$$

dónde f_m es la frecuencia de muestreo. Sin embargo, es necesario agregar 2-4 polos para aproximar posibles ceros en el espectro y la forma general de éste; de esta manera, una

norma para muestreo a 8 kHz es p=10 [O'Shaughnessy, 1987].

Con el fin de modelar los formantes de alta frecuencia, que generalmente son de menor amplitud, la señal de voz es *preenfatizada* muy frecuentemente antes de realizar el análisis de predicción lineal. Esto es, la señal de entrada al analizador de CPL es

$$y(n)=s(n)-\alpha s(n-1) \tag{22}$$

donde α controla el grado de *preénfasis*; el valor de α es establecido usualmente alrededor de 0.94. El *preénfasis* contribuye a "aplanar" el espectro de la señal de voz en el intervalo de tiempo que se está analizando y que es considerando estacionario. En sistemas que hacen uso del predictor lineal para la síntesis la etapa final debe contener un filtro de *deénfasis* con el fin de cancelar el efecto de *preénfasis*.

La idea fundamental detrás del análisis de predicción lineal es que una muestra de voz puede ser aproximada como una combinación lineal de muestras pasadas. Al minimizar la suma de las diferencias al cuadrado entre las muestras reales y las muestras predeterminadas sobre un intervalo finito, es posible obtener un conjunto de coeficientes del predictor. Existen varios métodos para el cómputo de esos coeficientes (v.g., autocorrelación, malla, covarianza, etc.); el análisis que se hará enseguida se concentrará únicamente en el método de la autocorrelación, el cual presenta las siguientes ventajas [Rabiner y Schafer, 1978]:

a) Los requerimientos de memoria son de los más bajos, comparables al método de la covarianza.

b) Los requerimientos de cómputo son los más bajos comparados contra los métodos de la covarianza y de la malla (del inglés *lattice*).

c) La estabilidad de $H(z)$ está garantizada, lo cual no ocurre con el método de la covarianza, pero sí con el de la malla (del inglés *lattice*).

d) El método proporciona parámetros intermedios que han sido utilizados en RAL.

Método de la autocorrelación

De la ecuación (19) la ecuación de diferencias que relaciona s(n) y u(n) es

$$s(n) = -\sum_{k=1}^{p} a_k s(n-k) + Gu(n) \tag{23}$$

y la salida de un predictor lineal con coeficientes α_k es

$$\hat{s}(n) = -\sum_{k=1}^{p} \alpha_k s(n-k) \tag{24}$$

Si se define el error de predicción como

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} \alpha_k s(n-k) \tag{25}$$

entonces la secuencia *error de predicción* es la salida de un sistema con función de transferencia

$$A(z) = 1 + \sum_{k=1}^{p} \alpha_k z^{-k} \tag{26}$$

Si la señal de voz sigue el modelo de la ecuación (23) y si $\alpha_k = a_k$, entonces $e(n) = Gu(n)$ y $A(z)$ será un filtro inverso de $H(z)$, esto es, de la ecuación (19)

$$H(z) = \frac{G}{A(z)} \tag{27}$$

El problema fundamental del análisis de predicción es encontrar el conjunto de coeficientes $\{\alpha_k\}$ a partir de la señal de voz, de tal manera que se tenga una buena estimación de las propiedades espectrales de ésta al usar la ecuación (27). Debido a la naturaleza no estacionaria de la señal de voz los coeficientes α_k deben ser estimados a partir de segmentos cortos de la forma de onda de voz (20-30 ms, como ya se ha mencionado).

Considérese $x(n)$ la secuencia obtenida al ponderar una trama de N muestras de $s(n)$ con una ventana $w(n)$ tal que

$$x(n) = \begin{cases} w(n)s(n) & 0 \le n \le N-1 \\ 0 & de\ otra\ manera \end{cases} \tag{28}$$

El error de predicción se define como

$$E = \sum_{n=-\infty}^{\infty} e_x^2(n) = \sum_{n=-\infty}^{\infty} \left[x(n) + \sum_{k=1}^{p} \alpha_k x(n-k) \right]^2 \tag{29}$$

donde $e_x(n)$ es la secuencia *error de predicción* de *x(n)*. Los valores de α_k que minimizan el error se encuentran estableciendo $\partial E / \partial \alpha_i = 0$ para *i=1, 2, ...,p*, lo cual conduce al conjunto de ecuaciones

$$\sum_{k=1}^{p} \alpha_k \sum_{n=-\infty}^{\infty} x(n-k)x(n-i) = -\sum_{n=-\infty}^{\infty} x(n)x(n-i) \quad , \quad i = 1, 2, \ldots, p \tag{30}$$

con p incógnitas α_k. Tomando en cuenta que la secuencia *x(n)* tiene longitud finita se puede demostrar que el segundo término de la igualdad es igual a la función autocorrelación de orden *i*, esto es

$$R(i) = \sum_{n=i}^{N-1} x(n)x(n-i) \quad , \quad i = 1, 2, \ldots, p \tag{31}$$

y tomando en cuenta que R(i) es una función par, al sustituir (31) en (30) se obtiene

$$\sum_{k=1}^{p} \alpha_k R(|i-k|) = -R(i) \quad , \quad i = 1, 2, \ldots, p \tag{32}$$

y se puede demostrar que el error de predicción E_p para un modelo de p polos es [Bristow, 1984]:

$$E_p = R(0) + \sum_{k=1}^{p} \alpha_k R(k) \tag{33}$$

El conjunto de ecuaciones descrito por (32) puede ser representado en forma matricial como

$$\begin{bmatrix} R(0) & R(1) & R(2) & \dots & R(p-1) \\ R(1) & R(0) & R(1) & \dots & R(p-2) \\ R(2) & R(1) & R(0) & \dots & R(p-3) \\ \dots & \dots & \dots & \dots & \dots \\ R(p-1) & R(p-2) & R(p-3) & \dots & R(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \dots \\ \alpha_p \end{bmatrix} = - \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \dots \\ R(p) \end{bmatrix} \tag{34}$$

El sistema de ecuaciones representado por (34) puede ser resuelto por alguno de los métodos convencionales (v.g., Gauss-Jordan, inversión de matriz, etc.), sin embargo, la matriz del sistema de ecuaciones es Toeplitz (v.g., los elementos en cualquiera de las diagonales son todos iguales) y simétrica; para la solución de este tipo de sistemas de ecuaciones existe un método recursivo optimizado conocido como el *algoritmo de Levinson-Durbin*, el cual se sintetiza en las siguientes ecuaciones [Bristow, 1984]:

$$E^{(0)} = R(0) \tag{35}$$

$$K_m = -\left[R(m) + \sum_{k=1}^{m-1} \alpha_k^{(m-1)} R(m-k) \right] \Big/ E^{(m-1)} \tag{36}$$

$$\alpha_m^{(m)} = K_m \tag{37}$$

$$\alpha_k^{(m)} = \alpha_k^{m-1} + K_m \alpha_{m-k}^{(m-1)} \qquad , \qquad 1 \le k \le m-1 \tag{38}$$

$$E^{(m)} = (1 - K_m^2) E^{(m-1)} \tag{39}$$

donde las ecuaciones (36)-(39) se resuelven recursivamente para m=1, 2, ..., p, y la solución final está dada por

$$\alpha_k = \alpha_k^{(p)} \qquad k = 1, 2, \dots, p \tag{40}$$

Debe notarse que en el proceso para obtener la solución para un predictor de orden p está involucrada la solución de los p-1 predictores de orden menor. La ecuación (39) se puede reescribir como

$$E^{(m)} = R(0) \prod_{i=1}^{m} (1 - K_m^2) \qquad m = 1, 2, \dots, p \tag{41}$$

Esta variable es el error de predicción y debe decrecer a medida que el orden del predictor se incrementa, de lo contrario no tendría sentido continuar con el proceso iterativo.

La tabla I muestra los resultados de una simulación donde una secuencia de 360 datos, cuyo valor es 1.01, es ponderada por una ventana Hamming de igual longitud. En el proceso para la obtención de los coeficientes de un predictor lineal de orden 10 puede observarse como el error de predicción decrece a medida que el orden se incrementa. Bajo esta consideración, a partir de la ecuación (41) se debe cumplir que

$$|K_m| < 1 \qquad 1 \leq m \leq p \tag{42}$$

lo cual se puede verificar en la tabla I.

Tabla I. Resultados de una simulación para la obtención de los coeficientes de predicción lineal para 360 datos de valor 1.01.

Índice	Error de predicción	Coeficiente de reflexión	Coeficiente de predicción
m	$E^{(m)}$	K_m	α_m
1	0.024919	0.999914	1.092568
2	0.019273	-0.0475986	0.006055
3	0.017222	-0.326249	-0.002448
4	0.016197	-0.243898	0.002936
5	0.015561	-0.198185	0.0012087
6	0.015130	-0.166421	-0.000400
7	0.014823	-0.142447	0.002473
8	0.014583	-0.127213	-0.003529
9	0.014407	-0.110050	0.003132
10	0.014255	-0.102533	-0.102533

Se puede demostrar [Bristow, 1984] que la condición establecida en la ecuación (42) es necesaria y suficiente para garantizar la estabilidad del sistema *H(z)* (v.g., todas las raíces de *A(z)* se encuentran dentro del círculo unitario). Las variables K_m del *algoritmo de Levinson-Durbin* se conocen como **coeficientes de reflexión**, y sus negativos como **coeficientes de correlación parcial (CORPAR)**.

Los parámetros K_m pueden ser interpretados como el negativo de la correlación parcial entre *x(n)* y *x(n+m)* manteniendo *x(n+1)*, ..., *x(n+m-1)* fijos [Makhoul, 1975]. A través del método de la malla es posible demostrar que la función de transferencia *H(z)* tiene la misma forma que la función de transferencia de un modelo de *p* tubos sin pérdidas concatenados [Rabiner y Schafer, 1978], tal que

$$r_m = K_m \qquad m = 1, 2, \ldots, p \tag{43}$$

de ahí el nombre de *coeficientes de reflexión* para los parámetros K_m.

Haciendo uso de la ecuación (158) del apéndice A se puede ver que

$$K_m = \frac{A_{m+1} - A_m}{A_{m+1} + A_m} \qquad m = 1, 2, \ldots, p \tag{44}$$

esto es, K_m representa la cantidad reflejada en la unión de los tubos *m* y *m*+1. Para el caso *m*=*p* se considera que el *(p+1)-ésimo* tubo es de longitud infinita, de tal manera que no hay onda sonora viajando desde el tubo en sentido inverso. De la ecuación (44) se puede ver que

$$\frac{A_{m+1}}{A_m} = \frac{(1 + K_m)}{(1 - K_m)} \qquad m = 1, 2, \ldots, p \tag{45}$$

la cual representa el cociente de áreas de tubos adyacentes. Es conveniente señalar que las áreas no pueden ser determinadas absolutamente, sin embargo, cualquier normalización sobre su valor produce la misma función de transferencia.

Una representación importante para los *coeficientes de reflexión* K_m que expande la región cercana a $|K_m| = 1$ es

$$g_m = \log\left|\frac{A_{m+1}}{A_m}\right| = \log\left|\frac{1 + K_m}{1 - K_m}\right| \qquad m = 1, 2, \ldots, p \tag{46}$$

donde los parámetros g_m son llamados **logaritmo del cociente de las áreas**, y son utilizados como parámetros cuantificables uniformemente de baja sensitividad espectral.

La respuesta al impulso *h(n)* del sistema con la función de transferencia de la ecuación (27) puede obtenerse recursivamente a partir de los coeficientes de predicción lineal como

$$h(n) = -\sum_{k=1}^{p} \alpha_k h(n-k) + G\delta(n) \qquad n \geq 0 \qquad (47)$$

donde *h(n)* se define 0 para $n<0$.

III.3 Cepstrum

En los 60's fueron desarrolladas dos técnicas de procesamiento llamadas **análisis cepstrum** y **deconvolución homomórfica.** La palabra *cepstrum* se obtiene intercambiando la posición de las letras en la palabra *spectrum*, y es adjudicado a Bogert *et al* quien afirmó "... *en general nos encontrábamos trabajando en el dominio de la frecuencia con métodos que se acostumbran en el dominio del tiempo y viceversa* ..." [Oppenheim y Schafer, 1989].

El término *sistemas homomórficos* fue propuesto por Oppenheim para una clase de sistemas que satisfacen una generalización del principio de superposición. Que en su sentido clásico establece: *si una señal de entrada a un sistema lineal está compuesta de una combinación lineal de dos señales (v.g., suma), la salida del sistema es una combinación lineal de las salidas correspondientes a cada una de ellas.* Este concepto puede ser representado por el siguiente par de ecuaciones

$$L[x(n)] = L[x_1(n) + x_2(n)] = L[x_1(n)] + L[x_2(n)] = y_1(n) + y_2(n) = y(n) \qquad (48)$$

$$L[a\,x(n)] = a\,L[x(n)] = a\,y(n) \qquad (49)$$

donde L representa un operador lineal.

Por analogía con el principio de superposición para sistemas lineales, se pueden definir una clase de sistemas que siguen un principio generalizado de superposición, donde la suma puede ser reemplazada, por ejemplo, por convolución. Ésto es

$$H[x(n)] = H[x_1(n) * x_2(n)] = H[x_1(n)] * H[x_2(n)] = y_1(n) * y_2(n) = y(n) \qquad (50)$$

Los sistemas que tienen la propiedad expresada en la ecuación (50) son llamados sistemas homomórficos para la convolución, y es posible demostrar que cualquier sistema homomórfico puede ser descompuesto en tres sistemas homomórficos en cascada [Oppenheim y Schafer, 1989].

La transformación de una señal en su *cepstrum* es una transformación homomórfica para la convolución; esto es, el *cepstrum* es un caso particular del análisis homomórfico, y es la primera de las tres transformaciones homomórficas necesarias para realizar una deconvolución. Ya que el concepto de *sistemas homomórficos* es bastante amplio, no será expuesto aquí; y más aún, no se hará un análisis del *cepstrum complejo*. El análisis de esta sección se limitará al análisis del *cepstrum* como parámetro útil en el reconocimiento automático de locutor (RAL), y en la obtención del periodo fundamental *T0*.

En el análisis de la producción de la señal de voz se ha considerado que la señal de voz *s(n)* es producto de la respuesta del tracto vocal *h(n)*, que es excitado por una señal *u(n)* la cual está formada ya sea por pulsos cuasiperiódicos o por ruido aleatorio. La señal observable *s(n)* es producto de la convolución de la excitación *u(n)* con la respuesta al impulso del tracto vocal *h(n)*; esto es

$$s(n) = u(n) * h(n) \tag{51}$$

La separación o deconvolución de las componentes de *s(n)* podría ser de gran utilidad; sin embargo, la deconvolución de dos señales convolucionadas es prácticamente imposible, a menos que las señales tengan características frecuenciales muy diferentes, como se considera que ocurre con la señal de voz [O'Shaughnessy, 1987].

Considérese nuevamente que las propiedades de la señal de voz *s(n)* no varían en un intervalo de tiempo correspondiente a N muestras; por ejemplo, la señal *x(n)* se obtiene al ponderar un segmento de *s(n)* con una ventana *w(n)*, esto es

$$x(n) = w(n)s(n) \qquad 0 \le n \le N-1 \tag{52}$$

y de acuerdo con la ecuación (51) se obtiene que

$$x(n) = w(n)[u(n) * h(n)] \qquad 0 \le n \le N-1 \qquad (53)$$

La mayoría de las investigaciones hechas con el *cepstrum* han supuesto que la ecuación (53) se puede aproximar por

$$x(n) \approx [u(n)w(n)] * h(n) \qquad (54)$$

lo cual implica que *u(n)* varía más lentamente *h(n)* [Oppenheim y Schafer, 1989]. Para el caso de voz sonora *u(n)* es un tren de impulsos de la forma

$$u(n) = \sum_{k=0}^{M-1} \delta(n - kN_0) \qquad (55)$$

por lo que

$$p(n) = u(n)w(n) = \sum_{k=0}^{M-1} w(kN_0)\delta(n - kN_0) \qquad (56)$$

donde se ha supuesto que el período fundamental es N_0 y que la ventana *w(n)* cubre *M* períodos. Al hacer esas consideraciones se tiene

$$x(n) = p(n) * h(n) \qquad (57)$$

Figura 10. Realización práctica para la obtención del *cepstrum*.

El cepstrum *c(n)* de *x(n)* puede obtenerse usando el esquema de la figura 10, donde se cumplen las siguientes ecuaciones

$$X(k) = \sum_{n=0}^{N-1} x(n)\, e^{-j\frac{2\pi}{N}kn} \qquad 0 \le k \le N-1 \qquad (58)$$

$$C(k) = \log|X(k)| \qquad 0 \le k \le N-1 \qquad (59)$$

$$c(n) = \frac{1}{N}\sum_{k=0}^{N-1} \log|X(k)|\, e^{j\frac{2\pi}{N}kn} \qquad 0 \le n \le N-1 \tag{60}$$

dónde *c(n)* representa la combinación de las componentes de x(n), pero ya no por medio de la convolución sino en forma de suma, ya que de las ecuaciones (57) y (60) se obtiene

$$C(k) = \log|X(k)| = \log|P(k)V(k)| = \log|P(k)| + \log|V(k)| \tag{61}$$

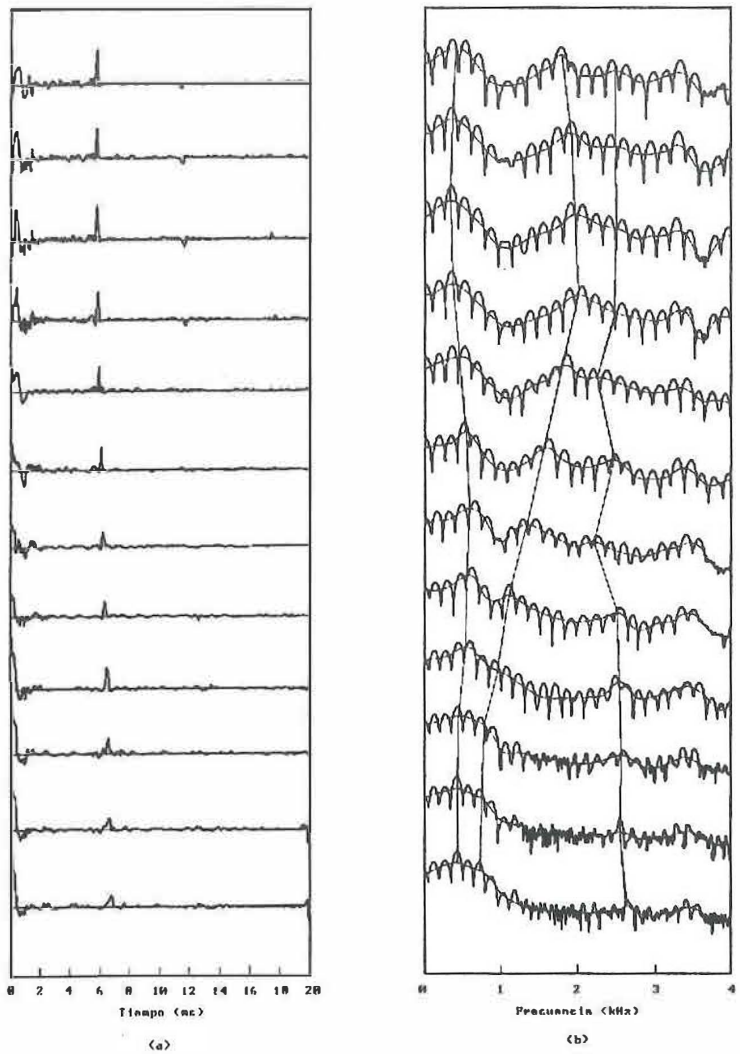


Figura 11. (a) cepstrum y (b) logaritmo del espectro de magnitud para segmentos secuenciales de voz sonora [Oppenheim, 1989].

Es importante establecer que no es posible obtener la secuencia *x(n)* a partir de *c(n)*,

ya que esta última secuencia fue obtenida a partir de la magnitud de *X(k)*. Para que el proceso sea invertible es necesario utilizar el cepstrum complejo $\hat{x}(n)$, en el cual está incluída la información de fase.

La figura 11 muestra una serie de secuencias *cepstrum* y logaritmo de magnitud para 12 tramas sucesivas de duración 20 ms de una señal sonora. Los formantes *F1, F2* y *F3* se encuentran unidos por líneas en la figura 11b, y es posible observar la manera en la que éstos cambian de trama a trama. La presencia de un pico prominente alrededor de 7 ms en la figura 11a es indicativo de la periodicidad de la señal de voz, y su posición indica el valor del período fundamental *T0* en cada trama analizada.

En el caso del *cepstrum* de señales que representan sonidos sordos el pico mencionado no existe o es de magnitud muy pequeña, en base a esta observación, un método para realizar una decisión *sonoro/sordo* sugerido por [Rabiner y Schafer, 1978] establece como primer paso la obtención del *cepstrum*, para luego realizar la detección de un pico alrededor del período fundamental esperado. Si la magnitud del pico es superior a un umbral pre-establecido se considera que la trama es de voz sonora, y que su *período fundamental T0* es igual a la posición del pico en el eje del tiempo, en caso contrario se considera que la trama de voz no corresponde a un sonido sonoro.

La obtención del *cepstrum* puede hacerse también usando el modelo del filtro digital todo-polos, cuyos coeficientes α_k pueden ser obtenidos por análisis de predicción lineal. Considérese el modelo de la ecuación (27) con G=1, esto es

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{k=1}^{p} \alpha_k z^{-k}} \tag{62}$$

la relación entre los coeficientes *cepstrum* c_n y los coeficientes de predicción lineal α_k está dada por la siguiente relación recursiva [Furui, 1981]:

$$c_n = -\alpha_n - \sum_{k=1}^{n-1}(1-k/n)\alpha_k c_{n-k} \qquad 1 \le n \le p \tag{63}$$

Si se desea una relación que calcule un número menor de divisiones, de la ecuación (63) se puede obtener

$$c_n = -\alpha_n - \frac{1}{n}\sum_{k=1}^{n-1} k c_k \alpha_{n-1} \tag{64}$$

Aún cuando la obtención del *cepstrum* se puede hacer a través de la transformación de Fourier o a partir de los coeficientes de predicción lineal, la eficiencia de los coeficientes obtenidos por ambos métodos al ser aplicados en RAL es similar. Sin embargo, el cálculo por el último método es aproximadamente dos veces más rápido [O'Shaughnessy, 1986].

III.4 Efectividad de los parámetros de la señal de voz en reconocimiento automático de locutor (RAL).

No es fácil encontrar en la literatura un estudio que incorpore los diferentes parámetros de la señal de voz con el fin de determinar cuales son más efectivos en RAL, y por otra parte, no es posible realizar una comparación entre estudios experimentales que hayan utilizado diferentes parámetros, a menos que hayan utilizado la misma base de datos (locutores), bajo las mismas condiciones experimentales y el mismo criterio de efectividad.

En el desarrollo de este trabajo se examinaron dos reportes de estudios comparativos de la efectividad de los parámetros de la señal de voz en RAL. En el primero de ellos [Sambur, 1975] se utilizó una base de datos generada en 3.5 años, y que estaba compuesta de las siguientes frases habladas en el idioma inglés:

1) Cool shirts please me.

2) Pay the man first please.

3) I cannot remember it.

4) Papa needs two singers.

5) A few boys bought them.

6) Cash this bond please.

Esta base de datos se considera muy rica fonéticamente y fue adquirida de 21 adultos masculinos. En el estudio se consideraron 92 parámetros de la señal de voz que incluyen: anchos de banda de formantes, localización de polos, formantes de las vocales, duración de algunos eventos acústicos, y frecuencia fundamental.

Los experimentos realizados fueron de identificación de locutor de umbral abierto (v.g., sólo se identificaron "clientes"), y el criterio de clasificación se basó en eliminar subsecuentemente la característica menos efectiva, de tal manera que al final se obtuvo como resultado la siguiente lista de parámetros ordenada por efectividad:

1) F2 de /n/ en las palabras /needs/ y /man/.

2) F3 de /u/ en la palabra /two/.

3) F2 de /I/ en la palabra /this/.

4) Duración de /k/ en la palabra /cash/.

5) F3 de /m/ en la palabras /remember/ y /man/.

En una evaluación con este conjunto de 5 parámetros se tuvo solamente un error en 320 identificaciones de locutor. Desafortunadamente, las condiciones experimentales estuvieron muy lejanas de aplicaciones en el mundo real, y entre ellas se pueden citar las siguientes:

a) Todas las grabaciones se hicieron en una cámara anecoica.

b) Los locutores del experimento fueron entrenados para mantener la cadencia en la pronunciación de las frases.

c) Se tuvo un conocimiento pleno de la estructura fonética de las frases habladas, lo cual hizo posible su explotación visual (v.g., segmentación y/o detección de fonemas).

El segundo reporte analizado [Atal, 1974] se centra únicamente en la comparación entre parámetros obtenidos de la predicción lineal, y se señalan como ventajas que éstos pueden obtenerse fácilmente a partir de la señal de voz, y que no es necesario hacer ninguna suposición *a priori* acerca de cuales parámetros podrían ser más efectivos; los coeficientes de predicción lineal representan la información combinada de los formantes y sus anchos de banda, frecuencia fundamental, y forma de onda glotal, entre otros.

En este experimento Atal recolectó un total de 60 frases obtenidas de seis repeticiones de la frase "May we all learn a yellow lion roar" hablada por 10 locutores del sexo femenino. Las frases fueron grabadas en una cámara anecoica en un intervalo de 27 días con una frecuencia de muestreo de 10 kHz. Los silencios fueron eliminados de las grabaciones calculando la energía sobre intervalos de 10 ms, y las frase comprimidas fueron divididas en 40 tramas de igual duración, siendo ésta proporcional a la duración de la frase.

En cada trama se realizó un análisis de predicción lineal de orden 12 usando el método de la autocorrelación. Además de utilizar los coeficientes de predicción lineal, también se usaron los siguientes parámetros:

a) Respuesta al impulso $h(n)$ $1 \leq n \leq 12$

b) Función autocorrelación $R(n)$ $1 \leq n \leq 12$.

c) Función área (coeficientes de reflexión) K_n $1 \leq n \leq 12$.

d) Coeficientes *cepstrum* c_n $1 \leq n \leq 12$

Los experimentos realizados fueron de identificación de locutor de umbral abierto. En el entrenamiento se consideraron de manera independiente cada una de las tramas, y los parámetros que las representaban fueron promediados sobre 5 frases; la frase restante fue usada para prueba. Esto se realizó de manera rotatoria, es decir, intercambiando las frases entre los conjuntos de entrenamiento y de prueba. La identificación de locutor se realizó obteniendo el vector representativo de la trama con la menor distancia al vector de prueba.

Los resultados del experimento se muestran en la figura 12. Como puede verse la eficiencia más alta la tuvieron los coeficientes *cepstrum*.

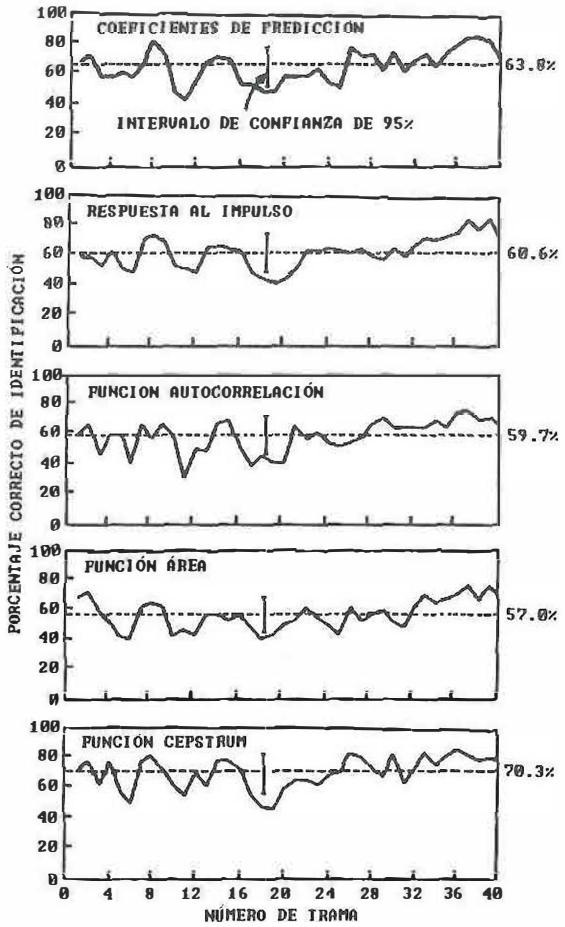


Figura 12. Porcentaje de exactitud en la identificación de locutor locutor como función de la trama utilizada [Atal, 1974].

CAPÍTULO IV TÉCNICAS DE RECONOCIMIENTO AUTOMÁTICO DE LOCUTOR (RAL).

IV.1 Generalidades.

En este capítulo se presentará un análisis de las técnicas que se han utilizado con mayor frecuencia en la tarea de RAL de acuerdo con la literatura. El análisis se restringe a tres técnicas que han producido resultados aceptables, si bien, el problema de RAL sigue sin solución completa dada la amplia gama de consideraciones bajo las cuales éste puede ser analizado; por ejemplo:

a) Funcionamiento en un ambiente real o con un nivel de ruido acústico ambiental muy alto.

b) Baja tasa de error aún si los "clientes" presentan alguna enfermedad o estado de ánimo que cambie las características de su voz

c) Evitar que el sistema pueda ser "burlado" por medio de alguna grabación magnetofónica.

d) Mantener baja la tasa de error a pesar de la variación temporal de las características de la voz (v.g., la mayoría de los sistemas de RAL presentan una tasa de error mayor a medida que el tiempo entre entrenamiento y prueba aumenta).

Es necesario mencionar, por una parte, que no fueron analizadas técnicas que involucran el uso de redes neuronales artificiales (RNA), las cuales han cobrado un gran auge; mientras que por otra parte, las técnicas que han sido aplicadas en RAL son similares a las aplicadas previamente en RAH, donde los resultados satisfactorios obtenidos motivaron su aplicación en reconocimiento de locutor. Ésto es consecuencia del hecho de que ha habido una cantidad mayor de trabajos relacionados con RAH que con RAL.

De acuerdo con el modo en que generan y manejan los modelos o patrones de locutor, las técnicas de reconocimiento automático de locutor (RAL) pueden ser clasificadas

en dos categorías [O'Shaughnessy, 1987]:

a) Técnicas que obtienen un modelo explotando el análisis de sonidos específicos.

b) Técnicas que obtienen un modelo analizando estadísticamente muestras de gran duración temporal de la voz del locutor (v.g., secuencias de palabras).

La primera categoría es usada con mucha frecuencia en sistemas *dependientes del vocabulario*, mientras que la segunda en sistemas *independientes del vocabulario.*

Por razones de eficiencia computacional es conveniente representar los modelos o patrones de locutor con una cantidad mínima de información; de aquí que todas las técnicas realicen una reducción de la información por medio de una parametrización y extracción de características concernientes al locutor.

IV.2 Alineamiento dinámico en el tiempo (ADT).

La técnica de **alineamiento dinámico en el tiempo (ADT)** pertenece a la categoría de técnicas que explotan el análisis de sonidos específicos y es eminentemente *dependiente del vocabulario.* Como punto de partida para analizar esta técnica se puede mencionar que en el método utilizado por Atal mencionado en la última sección del capítulo anterior, se realizó una comparación temporal trama a trama bajo la suposición de que la duración temporal de todas las frases habladas era la misma, y por lo tanto se dividió la frase en 40 tramas que se suponían de duración igual. A reserva de conocer a detalle la manera en la que fue obtenida la base de datos (ya que la referencia no profundiza demasiado en este aspecto), es necesario señalar que este método no considera las variaciones temporales que se presentan en la emisión de una palabra o frase; es decir, la velocidad a la que son emitidas las muestras de voz no solamente varía de locutor a locutor, sino que también varía entre las muestras de un mismo locutor; y más aún, la velocidad con que son emitidos los mensajes no es constante, sino que varía de manera no lineal a través de la frase entera.

Lo anterior trae como consecuencia que para un mismo mensaje hablado, la duración

temporal interlocutor e intralocutor varíe de muestra a muestra. Además, ocasiona que los eventos acústicos subyacentes en las muestras, que pueden estar representados por información temporal o frecuencial, ocurran a diferentes escalas temporales cuando estas últimas son comparadas. Este hecho ocasiona que la comparación "punto a punto" de patrones de locutor que considera que las velocidades de habla son constantes dentro de cada muestra, y por lo tanto proporcionales de muestra a muestra, presente una tasa de error más alta comparada con la de los sistemas que toman en cuenta el comportamiento no lineal de éstas, y en consecuencia realizan una normalización.

De acuerdo con lo expuesto anteriormente, antes de realizar una comparación de patrones de locutor es necesario realizar una normalización no lineal en las escalas temporales bajo las cuales fueron adquiridas las muestras de voz, esto es, se deben alinear temporalmente los eventos acústicos presentes.

Considérense dos muestras de voz X e Y, que pueden o no pertenecer al mismo locutor, representadas por las secuencias $(x_1, x_2, \ldots, x_N)$ y $(y_1, y_2, \ldots, y_M)$ respectivamente, donde x_n e y_m son vectores de parámetros representativos de las características acústicas de tiempo corto (v.g., sobre una trama); los índices N y M no necesariamente deben ser iguales, y de hecho, es de esperarse que sean diferentes. La disimilitud entre X e Y puede ser definida considerando alguna función distorsión espectral de tiempo corto $d(x_n, y_m)$, la cual será denotada por simplicidad como $d(n, m)$, donde n=1, 2, ..., N y m=1, 2, ..., M. Ya que el orden secuencial de los sonidos es crítico en la composición de las palabras y frases, es necesario que los índices de los vectores comparados cumplan ciertas restricciones (v.g., es poco probable que un evento acústico en el inicio de una muestra corresponda con el evento final de otra).

Quizá la solución más simple para el problema de alineamiento temporal y normalización sea una técnica de normalización lineal temporal. En normalización lineal

temporal, la disimilitud entre X e Y se define como

$$D(X,Y) = \sum_{n=1}^{N} d(n,m) \tag{65}$$

donde n y m satisfacen

$$m = \frac{M}{N} n \tag{66}$$

y $d(n,m)$ es la medida de distorsión de tiempo corto. Ya que n y m son enteros, es necesario hacer alguna clase de redondeo en la ecuación (66). La sumatoria en la ecuación (65) puede ser tomada también desde m=1 hasta m=M, dependiendo de la dirección deseada para la normalización.

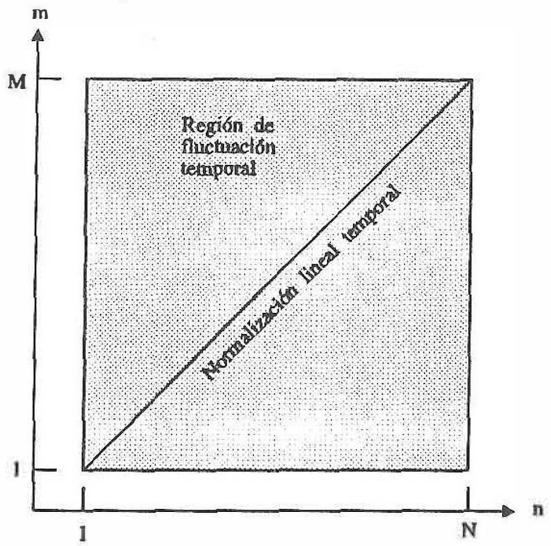


Figura 13. Normalización lineal temporal para dos secuencias de diferente longitud.

Este método de normalización se encuentra representado en la figura 13, donde se ha distribuido cada una de las secuencias de vectores en el eje horizontal y vertical respectivamente. A partir de la figura puede observarse que la evaluación de las distorsiones trama a trama, y por lo tanto la disimilitud entre las muestras X e Y, tiene lugar a lo largo de la trayectoria diagonal que une las tramas iniciales y finales en el plano (n,m).

Obviamente, esta restricción rígida de las fluctuaciones de la velocidad de habla no modela adecuadamente la situación real para muestras reales de voz, de aquí que se haga necesaria una manera más realista de alineamiento y normalización temporal de un par arbitrario de patrones de voz.

En la literatura se han propuesto varios métodos para resolver el problema de alineamiento temporal de patrones [Rabiner *et al*, 1978]; por conveniencia se analizará únicamente un método bastante general, el cual es una variante de uno de los propuestos por Rabiner *et al* [Furui,1981]. Este método es conocido como una técnica de **Mínimo Local No-restringida a los Límites (MLNL)** (del inglés *Unconstrained Endpoints, Local Minimum UELM*).

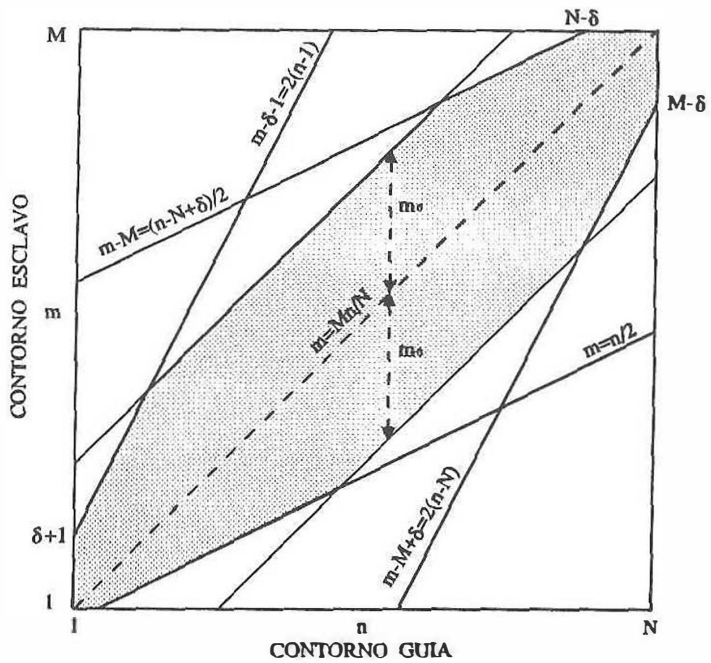


Figura 14. Región permisible de la ruta función alineamiento.

El método que se va a analizar a continuación forma parte de un conjunto de técnicas que tienen por nombre genérico **Alineamiento Dinámico en el Tiempo (ADT),** (del inglés *Dynamic Time Warping DTW*) y se encuentra representado en la figura 14. El objetivo del método es encontrar una función alineamiento $w(n)$ entre dos secuencias de parámetros

(v.g., secuencias de vectores) *X e Y* conocidas como *contornos*. La secuencia distribuida a lo largo del eje horizontal se designa como *contorno guía*, mientras que aquella distribuida a lo largo del eje vertical se le llama *contorno esclavo*. La función alineamiento *w(n)* entre *n* y *m* se denota como

$$m = w(n) \tag{67}$$

esto es, el índice del *contorno esclavo* es reasignado con los valores de *w(n)*.

La función alineamiento *w(n)* debe cumplir además un conjunto de restricciones en los límites de las muestras de voz y en la forma que ésta sigue. El conjunto de restricciones está representado por las siguientes ecuaciones:

$$w(n+1) - w(n) = 0, 1, 2 \quad para \quad w(n) \neq w(n-1) \tag{68}$$

$$w(n+1) - w(n) = 1, 2 \quad para \quad w(n) = w(n-1) \tag{69}$$

$$1 \leq w(1) \leq \delta + 1 \tag{70}$$

$$M - \delta \leq w(N) \leq M \tag{71}$$

$$\max w(n) = M \qquad para\ N - \delta \leq n \leq N \tag{72}$$

$$\frac{M}{N} n - m_0 \leq w(n) \leq \frac{M}{N} n + m_0 \tag{73}$$

las ecuaciones (68) y (69) ocasionan que *w(n)* sea monotónicamente creciente con una pendiente máxima de 2 y una pendiente mínima de 1/2. La restricción de mínima pendiente es consecuencia de la prohibición de dos pasos consecutivos de pendiente cero. Las restricciones de pendiente son muy importantes; por una parte, si se permite que la función de alineamiento tenga una gran pendiente, un evento acústico de corta duración en el *contorno guía* corresponderia a un evento acústico de gran duración en el *contorno esclavo*, además de que una gran cantidad de eventos de este último contorno serían eliminados del alineamiento. Por otra parte, si se permiten demasiados pasos con pendiente cero, una evento acústico de gran duración en el *contorno guía* sería alineado con un evento acústico de corta duración en el *contorno esclavo*. Concretamente, si las restricciones en la pendiente

son muy severas (v.g., si existe una ruta única de alineamiento) el método no podría funcionar adecuadamente, y si éstas son muy laxas, la discriminación de contornos de parámetros para diferentes locutores se degradaría.

Las ecuaciones (70)-(73) definen el área permitida de alineamiento para la función *w(n)* en el plano *(n,m)*, la cual se encuentra sombreada en la figura 14. La ecuación (73) impide que la ruta se aproxime a los cruces de las rectas que limitan el área de alineamiento, lo cual reduce el número de posibles rutas. Ésto se justifica, ya que es bastante improbable que dos muestras pertenecientes a un mismo locutor puedan ser alineadas por una función que se aproxime a estos puntos.

En las ecuaciones (70)-(72) δ representa el máximo intervalo de error en la detección de los límites temporales del mensaje hablado por el locutor (v.g., puntos inicial y final de la palabra o mensaje hablado por el locutor), y se mide en tramas. Ésta es la razón por la que este método es clasificado como de Mínimo Local No-restringido a los Límites (MLNL), ya que presupone que puede haber un error en la detección automática de los límites del mensaje hablado, y permite que la trama inicial del *contorno guía* pueda ser alineada con cualquiera de las primeras δ tramas del *contorno esclavo*; y por otra parte, permite mayor flexibilidad en cuanto al punto terminal de la función *w(n)*, ya que éste puede estar localizado en cualquiera de los conjuntos de puntos $\{n,M \mid N-\delta \le n \le N\}$ y $\{N,m \mid M-\delta \le m \le M\}$. En las primeras versiones de los algoritmos de alineamiento δ valía cero, de ahí que las tramas iniciales y finales de los contornos estuvieran restringidas a ser alineadas bajo la función alineamiento *w(n)*; ésas versiones son conocidas como **Restringidas a Límites de frase, de Intervalo de Pendiente 2 a 1 (RLIP2-1)** (del inglés *Constrained Endpoints, 2 to 1 range of slope CE2-1* [Rabiner *et al*, 1978].

Debe notarse que la función alineamiento *w(n)* puede alcanzar la trama final del *contorno esclavo* antes que la trama final del *contorno guía*, en cuyo caso no tendría sentido

continuar la ruta. Esto es

$$w(n)=M \qquad para \qquad n<N \qquad (74)$$

Una pregunta que puede surgir ahora es: ¿cual de los contornos debe ser el guía y cual el esclavo ?. En experimentos de reconocimiento automático del habla (RAH) realizados con esta técnica [Rabiner *et al*, 1978] no se encontró que éste fuera un factor importante en la tasa de error, mientras que en experimentos de reconocimiento automático de locutor (RAL) [Furui, 1981] se sugiere que el más corto de los contornos sea asignado como guía y el más largo como esclavo; esto a partir de la observación de que si la variación intralocutor de la duración de las muestras habladas no es grande, hay una gran probabilidad de que la ruta óptima de alineamiento comience en la primera trama del más corto de los contornos.

Con el fin de encontrar la función ruta óptima que alinea los contornos, es necesario definir una medida de similitud punto a punto, ésta podría ser una función distancia *d(n.m)*. Dada esta medida de similitud puntual, se selecciona la función ruta óptima *w(n)* que minimiza la distancia total acumulada D a lo largo de la ruta de alineamiento, esto es:

$$D = \min_{\{w(n)\}} \sum_{n=1}^{N} d(n, w(n)) \qquad (75)$$

Si la función alineamiento *w(n)* alcanza el final del contorno esclavo antes de alcanzar el final del contorno guía, la distancia D puede ser normalizada por el factor (N / N_s), donde N_s es el índice de la trama en el contorno guía a la cual la ecuación (74) se satisface.

El problema de encontrar la ruta óptima *w(n)* que alinea los contornos esclavo y guía puede ser resuelto por un método de "fuerza bruta" tal y como calcular todas las rutas posibles dentro del area permitida y elegir la ruta con la mínima distancia acumulada D. Sin embargo, tal método es computacionalmente ineficiente ya que consume demasiado tiempo y memoria; una alternativa es un método de programación dinámica (optimización) los cuales han probado ser una buena solución para este tipo de problemas [Dreyfus y Law,

1977]; de aquí el nombre del algoritmo Alineamiento Dinámico en el Tiempo (ADT).

Usando este tipo de algoritmos, la distancia acumulada en cualquier punto *(n,m)* de la rejilla en la figura 14 puede ser determinada recursivamente como

$$D(n,m) = d(n,m) + \min[D(n-1,m)g(n-1,m), D(n-1,m-1), D(n-1,m-2)] \quad (76)$$

donde *g(n,m)* es un peso de la forma

$$g(n,m) = \begin{cases} 1 & si \ \ w(n) \neq w(n-1) \\ \infty & si \ \ w(n) = w(n-1) \end{cases} \quad (77)$$

Este algoritmo se basa en el hecho de que la ruta óptima que minimiza la distancia acumulada, puede ser obtenida al minimizar localmente la distancia puntual; esto es, la distancia acumulada desde cualquier punto *(n,m)* de la ruta óptima *w(n)* hasta el inicio o el fin de los límites de alineamiento (v.g., las condiciones de frontera) es también la óptima.

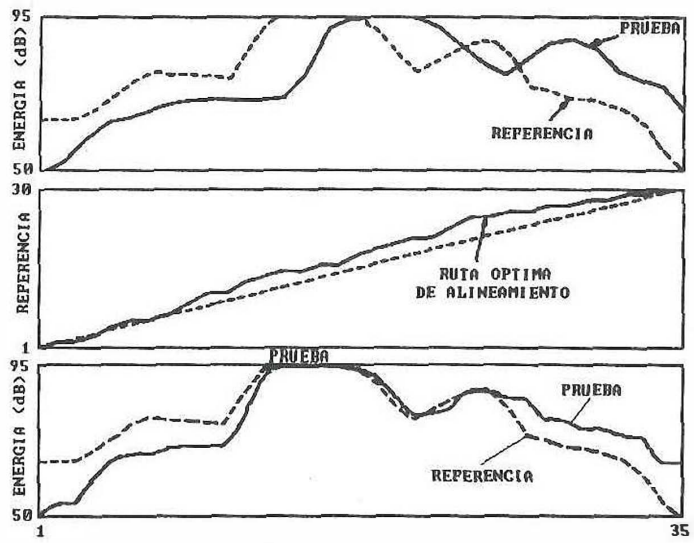


Figura 15. Ilustración de la efectividad del alineamiento dinámico en el tiempo para el contorno de energía de dos versiones de la palabra inglesa */seven/*.

La figura 15 muestra el efecto del algoritmo de alineamiento dinámico en el tiempo (ADT) para el contorno de energía de dos versiones de la palabra inglesa */seven/*. En la parte superior se muestran ambos contornos sobrepuestos sin alineamiento y en una escala lineal, ahí puede observarse la necesidad de un alineamiento no lineal; en la parte media se muestra

la función ruta óptima de alineamiento *w(n)* necesaria para alinear ambos contornos, y en la parte baja el efecto del alineamiento sobre el contorno de prueba. Puede observarse que los picos de energía fueron alineados sin modificar substancialmente los índices de las primeras tramas de ambos contornos.

El método de alineamiento dinámico en el tiempo ha sido aplicado en experimentos de reconocimiento automático de locutor dependiente del vocabulario (RAL-DV), utilizando como parámetros los coeficientes *cepstrum*, obtenidos sobre tramas de tiempo de la señal de voz. En los resultados de un experimento de verificación de locutor [Velius, 1988] la tasa de equierror (TEE) estuvo por debajo del 6% para órdenes de predicción mayores a 14 (v.g., más de 14 coeficientes *cepstrum* por trama) a partir de una base de datos de 20 locutores que contenía por igual palabras o frases, y que fueron capturadas a través de una red telefónica local. Para este tipo de técnicas no hay evidencia de que la TEE se reduzca cuando se incrementa la longitud de la muestra de entrenamiento (v.g., cambiar de una palabra a una frase); sin embargo, el uso de contornos promediados de varias muestras del mismo mensaje para obtener un modelo de referencia de locutor,sí contribuye a disminuir la tasa de equierror (TEE), tal y como se verá en el capítulo de experimentos y resultados.

IV.3 Cuantificación Vectorial (CV).

IV.3.1 Resumen de Cuantificación Vectorial (CV).

Cuantificación vectorial (CV) es el nombre de una técnica para codificación fuente desarrollada originalmente en teoría de la información para comunicaciones, pero que ha extendido su aplicación hacia RAH y RAL. Codificación fuente se refiere al conjunto de técnicas que son utilizadas para convertir una señal de información en un conjunto de códigos relacionados con la fuente que genera esa información (v.g., vectores de coeficientes de predicción lineal), con el fin de trasmitirlos y reproducir la señal original en un lugar e instante de tiempo diferentes, con un nivel aceptable de distorsión.

Específicamente, el objetivo de codificación fuente es lograr la mínima distorsión posible para una tasa de bits de transmisión predeterminada, o bien, lograr un nivel preestablecido de distorsión a la menor tasa posible de bits.

En el contexto de reconocimiento automático de locutor (RAL), la técnica de cuantificación vectorial (CV) ha sido utilizada en sistemas *independientes del vocabulario* [Soong *et al*, 1987], al igual que en sistemas que dependen del mensaje hablado [Buck *et al*, 1985] [Rosenberg y Soong, 1986] con tasas de error bastante bajas en ambos casos.

Para analizar el método de cuantificación vectorial (CV) considérese que se dispone de una secuencia $\{X_1, X_2, \ldots, X_I\}$, donde X_i representa vectores de K componentes. Estos vectores son obtenidos generalmente al segmentar señales de voz en tramas temporales y realizar algún procedimiento de parametrización en cada una de ellas (v.g., predicción lineal de orden K).

Un cuantificador vectorial de N entradas y K dimensiones es una función q tal que asigna a cada vector de entrada $X_i = (x_0, x_1, \ldots, x_{K-1})$ un vector $\hat{x}_n = q(X_i)$ de un conjunto finito $\hat{\mathrm{A}} = \{\hat{x}_n : n = 1, 2, \ldots, N\}$. El cuantificador q está completamente descrito por el conjunto $\hat{\mathrm{A}}$ y la partición $S = \{s_n : n = 1, 2, \ldots, N\}$, donde $s_n = \{X | q(X) = \hat{x}_n\}$. El conjunto $\hat{\mathrm{A}}$ es conocido como **libro de códigos**, y los vectores $\hat{x}_n$ son los centroides de los subespacios s_n. Una medida de la eficiencia del cuantificador q al cuantificar una secuencia de I vectores $\{X_1, X_2, \ldots, X_I\}$ es la distorsión promedio producida en la cuantificación

$$D = \frac{1}{I} \sum_{i=1}^{I} d(X_i, q(X_i)) \tag{78}$$

donde D es la distorsión promedio y d es una medida de distorsión espectral (v.g., distancia). En el sentido más general el cuantificador se considera óptimo si minimiza la distorsión promedio al cuantificar una secuencia de vectores; sin embargo, en la aplicación de CV al RAL el concepto de *optimalidad* tiene otra connotación, como se verá más adelante.

Con el fin de emplear cuantificación vectorial (CV) se deben seguir dos etapas: diseño del cuantificador (v.g., obtención del libro de códigos $\hat{A}$), y clasificación o codificación de las secuencias de vectores; estas dos tareas adquieren el nombre de *entrenamiento* y *prueba* al utilizar CV en el reconocimiento de locutores.

El diseño del cuantificador vectorial se realiza con una gran cantidad de vectores $\{v_1, v_2, \ldots, v_L\}$ que forman un conjunto de *entrenamiento*. El conjunto de entrenamiento es utilizado para generar el conjunto óptimo de vectores del *libro de códigos* que represente la variabilidad espectral observada en el conjunto de entrenamiento. Por razones de algorítmica y de máximo aprovechamiento del número de bits para etiquetar cada uno de los códigos del cuantificador, el tamaño N (v.g., número de entradas al cuantificador) del *libro de códigos* es usualmente una potencia de 2; esto es $N = 2^B$, donde B es llamada la *tasa del cuantificador*, o bien se estára hablando de un cuantificador vectorial de B bits. De esta manera, se requiere que el tamaño L del conjunto de entrenamiento sea mucho mayor que el tamaño N del *libro de códigos*, con el fin de encontrar los mejores N vectores de una manera robusta.

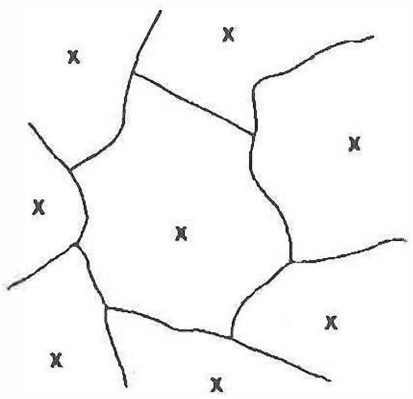


Figura 16. Partición de un espacio vectorial en subespacios, donde cada subespacio es representado por un centroide.

También dentro de la etapa de entrenamiento del cuantificador es necesario definir la

medida de similitud o distancia d de la ecuación (78), y un procedimiento para la obtención de los centroides, el cual trabaja sobre la base de que se realiza una partición del espacio de L vectores de entrenamiento, para agruparlos en N subconjuntos. Los componentes del *libro de códigos* son seleccionados como los centroides de cada uno de los N subespacios, tal y como se muestra en la figura 16 para un espacio de dos dimensiones (v.g., K=2).

Finalmente, para la utilización del cuantificador vectorial es necesario definir un proceso que clasifique vectores arbitrarios de análisis espectral de la voz, tal que elija el vector del *libro de códigos* más cercano a el vector de entrada y use el índice (de B bits) como su representación espectral resultante. Este procedimiento es frecuentemente referido como **Regla del Vecino Más Cercano (RVMC) o Procedimiento de Codificación Óptima (PCO).** El procedimiento de clasificación es escencialmente un cuantificador que acepta como entrada un vector espectral de voz, y proveé como salida el índice del vector del *libro de códigos* que más se le parece, en el sentido de la medida de similitud d.

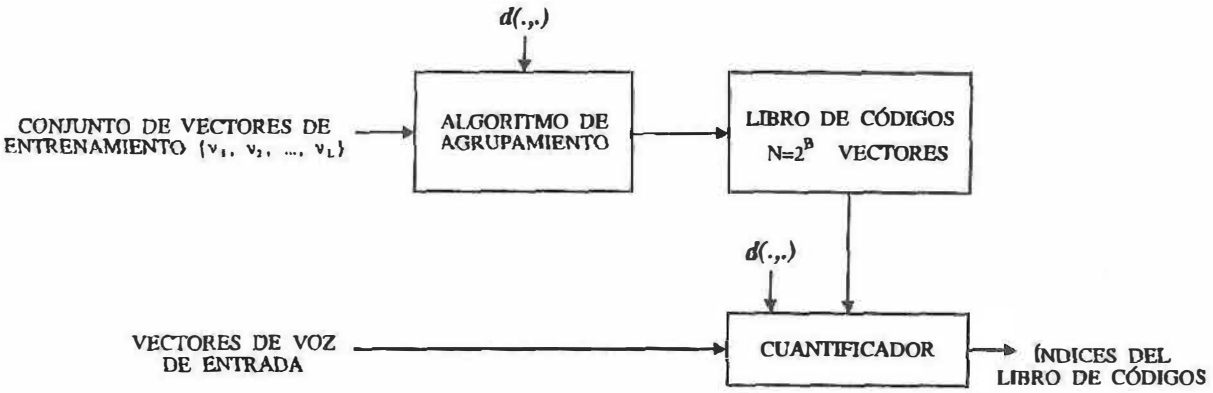


Figura 17. Diagrama a bloques de la estructura de un cuantificador vectorial.

La figura 17 muestra un diagrama a bloques de la estructura de un cuantificador vectorial que resume todo lo que se ha anotado; la parte más compleja en el diseño de un cuantificador es el algoritmo de agrupamiento, de los cuales existen varios reportes en la literatura. A continuación se describirán brevemente dos de los algoritmos más utilizados en

los sistemas de reconocimiento automático de locutor (RAL). Es conveniente establecer que el desarrollo de algoritmos de agrupamiento es un tema nada trivial que incluye aspectos de estadística y optimización que salen del alcance de esta tesis.

Algoritmo Generalizado de Lloyd (AGL).

El algoritmo generalizado de Lloyd (GLA en inglés) es uno de los más conocidos en la literatura. De acuerdo con este algoritmo, la manera en la que un conjunto de L vectores de entrenamiento pueden ser agrupados en un *libro de códigos* de N entradas es la siguiente [Rabiner y Juang, 1993]:

1.-) **Inicialización**: Elegir arbitrariamente N vectores (inicialmente fuera del conjunto de entrenamiento de L vectores).

2.-) **Búsqueda del vecino más cercano**: Para cada vector de entrenamiento encontrar el vector dentro del *libro de códigos* que esté más cerca (en términos de distancia espectral), y asignar ese vector al conjunto asociado al centroide correspondiente.

3.-) **Actualización del centroide**: Actualizar el *libro de códigos* calculando el centroide de cada uno de los conjuntos de vectores de entrenamiento formados.

4.-) **Iteración:** Repetir los pasos 2 y 3 hasta que la distorsión promedio D sea menor que un umbral preestablecido.

Algoritmo de división binaria.

Si bien el algoritmo mencionado anteriormente puede funcionar adecuadamente, se ha encontrado que es más conveniente diseñar un cuantificador de N entradas en etapas [Rabiner y Juang, 1993], esto es, se diseña primero un cuantificador vectorial con una sola entrada (v.g., sólo un código), se usa entonces una técnica para dividir el código único en dos, y el proceso continúa iterativamente hasta que se obtiene el *libro de códigos* de N entradas deseado. El algoritmo puede ser realizado siguiendo los siguientes pasos:

1.-) Diseñar un *libro de códigos* de un vector, el cual puede ser elegido como el centroide de todo el conjunto de vectores de entrenamiento.

2.-) Duplicar el tamaño del libro de códigos dividiendo cada código $\hat{x}_n$ de acuerdo a la regla

$$\begin{aligned} \hat{x}_n^+ &= \hat{x}_n(1+\varepsilon) \\ \hat{x}_n^- &= \hat{x}_n(1-\varepsilon) \end{aligned} \tag{79}$$

donde n varía desde 1 hasta el tamaño actual del *libro de códigos*, y ε es un parámetro de división que se establece usualmente entre 0.01 y 0.05).

3.-) Usar el algoritmo generalizado de Lloyd (AGL) para obtener el mejor conjunto de centroides del nuevo *libro de códigos* (v.g., el *libro de códigos* de tamaño duplicado).

4.-) Repetir los pasos 2 y 3 hasta que se obtenga un *libro de códigos* del tamaño N deseado.

IV.3.2 Aplicación de Cuantificación Vectorial (CV) en Reconocimiento Automático de Locutor (RAL).

La idea detrás de la aplicación de cuantificación vectorial en reconocimiento automático de locutor (RAL) es que es posible obtener un *libro de códigos* para cada locutor, a partir de secuencias de vectores que representan algún tipo de información espectral. Como podrá entreverse, esas secuencias de vectores son obtenidas al segmentar en tramas las muestras de voz de cada locutor, y realizar algún tipo de análisis con el fin de extraer un conjunto de parámetros. Únicamente las secuencias de vectores de cada locutor son utilizadas para generar su *libro de códigos*; tal y como se muestra en la figura 18 para el caso de 2 parámetros por vector. El procedimiento de obtener los *libros de códigos* para cada locutor equivale a la función de *entrenamiento* en el contexto de RAL.

Como se mencionó anteriormente, esta técnica ha sido aplicada en experimentos de reconocimiento automático de locutor independiente del vocabulario RAL. Si bien la técnica

de cuantificación vectorial ha sido aplicada con algunas variantes en diferentes experimentos de reconocimiento automático de locutor, uno de los experimentos más representativos con esta técnica fue realizado por Soong *et al* [Soong *et al*, 1987], a quienes se debe el esquema de la figura 19, donde se muestra un procedimiento esquematizado para reconocer locutores utilizando los coeficientes de predicción lineal (CPL) obtenidos sobre tiempo corto.

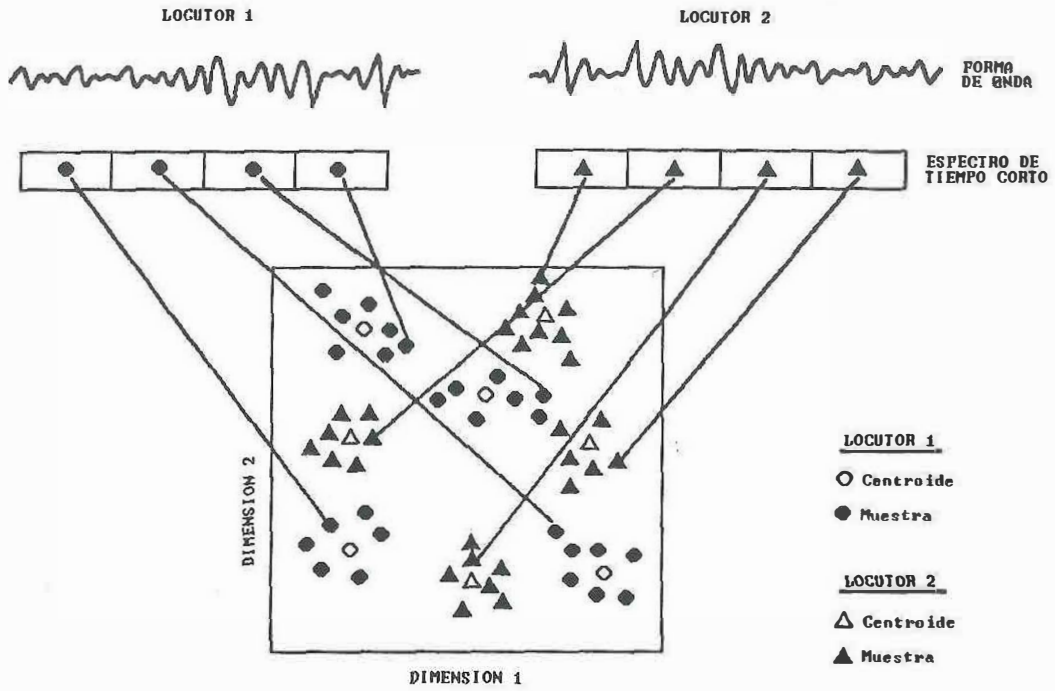


Figura 18. Diagrama conceptual ilustrando la obtención del *libro de códigos*.

Previa obtención del *libro de códigos* para cada uno de los M locutores, el procedimiento para identificar locutores (con umbral abierto) consiste en cuantificar la secuencia de vectores, obtenidos de la voz del locutor de prueba, con cada uno de los *libros de códigos*, y elegir como locutor identificado al dueño del *libros de códigos* que generó la menor distorsión promedio. Esto es, si se proporcionan L vectores de prueba $\{X_1, X_2, \ldots, X_L\}$, la distorsión producida por el i-ésimo *libro de códigos* es

$$D_i = \frac{1}{L}\sum_{\ell=1}^{L} \min_{1\le n\le N} d(X_\ell, \hat{x}_n) \qquad (80)$$

y la decisión del sistema está dada por

$$i^* = \underset{1 \le m \le M}{argmin} D_m \tag{81}$$

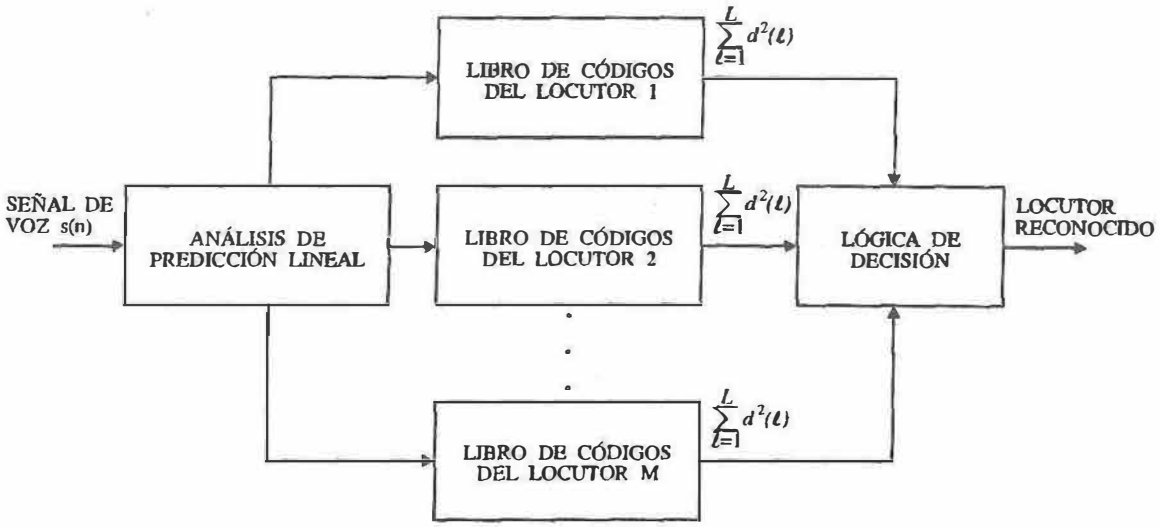


Figura 19. Diagrama a bloques del sistema de reconocimiento automático de locutor basado en cuantificación vectorial (CV) [Soong *et al*, 1987].

El esquema de verificación de locutor tiene una estructura similar al de la figura 19, pero en este caso únicamente se codifica la secuencia de vectores de prueba con el *libro de códigos* cuyo presunto dueño es el locutor. La distorsión obtenida es comparada contra un umbral preestablecido para *rechazar* o *aceptar* al locutor desconocido.

En sus experimentos, Soong *et al* recolectaron muestras de los 10 dígitos (hablados en inglés) de 100 locutores, 50 hombres y 50 mujeres; cada uno de los locutores proporcionó 20 repeticiones de cada dígito a través de una línea telefónica local en 5 sesiones de grabación.

Los 200 dígitos aislados fueron divididos en dos grupos, los primeros 100 fueron usados para *entrenamiento* (generación del *libro códigos*) y los últimos 100 para pruebas (*reconocimiento*). Los parámetros utilizados fueron los coeficientes de predicción lineal obtenidos de la señal de voz, de la cual se conservó sólo la banda de 200 Hz a 3200 Hz, y

que fue muestreada a 6.67 kHz. El orden de predicción utilizado fue 8, sobre tramas de longitud 45 ms traslapadas 30 ms (v.g., 15 ms de corrimiento entre trama contiguas).

La figura 20 muestra el porcentaje de error en identificación de locutor de *umbral abierto* con secuencias de dígitos diferentes al variar la longitud de la secuencia de prueba y la tasa del *libro de códigos*. Como puede observarse la tasa de error disminuye al aumentar el número de dígitos de prueba, así como también el tamaño del libro de códigos, hasta un mínimo de 1.5 % para 10 dígitos de prueba y un *libro de códigos* de 64 entradas.

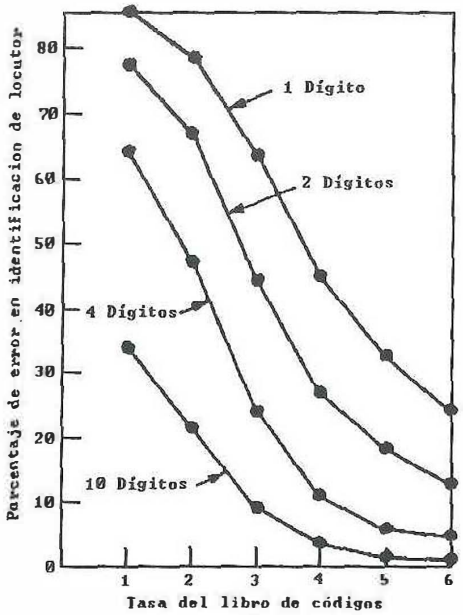


Figura 20. Tasa de error en identificación de locutor contra tasa de *libro de códigos* para secuencias de 1, 2, 4 y 10 dígitos.

En la figura 21 se muestran los resultados para pruebas de identificación de locutor de *umbral abierto* con 10 dígitos repetidos y una tasa de *libro de códigos* de 6. Puede observarse que la tasa de error más baja la produce el dígito /nine/, esto puede ser asociado al hecho de que los sonidos nasales han mostrado buenos resultados en el reconocimiento de locutor; sin embargo, estos resultados no superan a los mejores obtenidos en el experimento anterior (10 dígitos diferentes y tasa de *libro de códigos* 6) que se muestran en la línea punteada de esta misma figura.

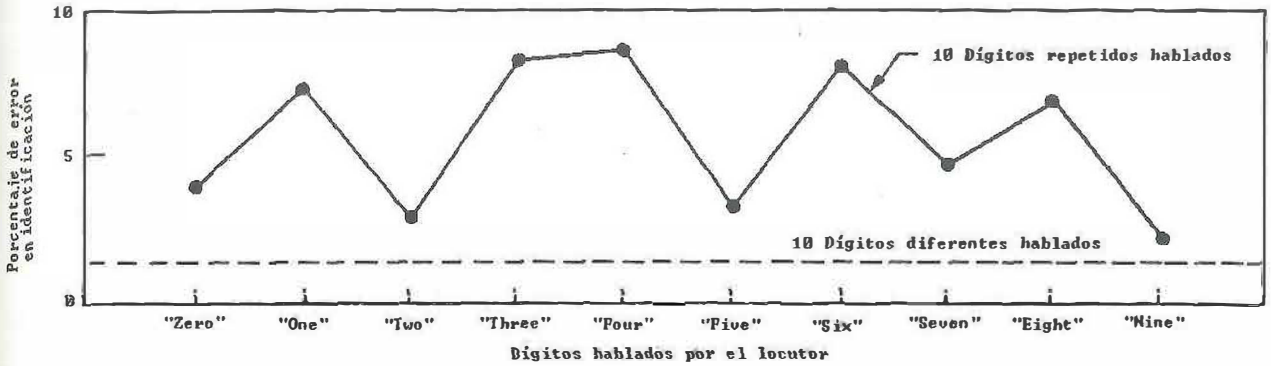


Figura 21. Porcentaje de error de identificación contra dígitos hablados por el locutor, repetidos y diferentes (Soong *et al*, 1987).

Finalmente, en la figura 22 se muestran los resultados obtenidos por Soong *et al* en verificación de locutor. Los experimentos fueron realizados utilizando 100 dígitos del locutor dueño del *libro de códigos* y 100 dígitos adicionales tomados de los demás locutores para que actuaran como "impostores". El tamaño del *libro de códigos* fue 64 y el umbral de decisión fue elegido *a posteriori* como aquel que igualaba las dos clases de error (*falsa aceptación y falso rechazo*); cabe mencionar que no son muy conocidos los métodos para establecer umbrales *a priori*.

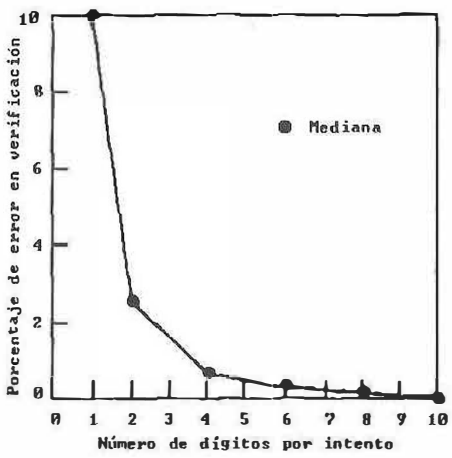


Figura 22. Mediana de la tasa de equierror (TEE) contra número de dígitos de prueba en experimentos de verificacion de locutor [Soong *et al*, 1987].

La figura 22 corresponde a la mediana de las (TEE) de cada uno de los locutores, y

de acuerdo con otros resultados presentados por el mismo autor correspondientes a la media de las TEE, ésta fue ligeramente menor al 2% al verificar locutores con una cadena de 10 dígitos. La gran diferencia entre la media y la mediana se debe a que un pequeño grupo de locutores presentó unas TEE bastante altas, mientras que para la mayoría de los locutores fue muy cercana a cero.

IV.4 Modelos Ocultos de Markov (MOM).

Ninguno de los dos métodos que se han planteado para RAL (v.g., ADT y CV) puede considerarse basado en modelado estadístico en un sentido estricto. Si bien se han usado técnicas estadísticas para realizar el agrupamiento (v.g., obtención del *libro de códigos*) en la técnica de cuantificación vectorial (CV), esta técnica podría ser clasificada como un método simplificado no paramétrico, en el cual una cierta cantidad de muestras del habla son utilizadas para caracterizar a diferentes locutores.

Modelos Ocultos de Markov (MOM) es el nombre de una técnica que se basa en el modelado estocástico, y que ha sido utilizada con bastante frecuencia en los últimos años en las tareas de RAL y RAH. Ni la teoría de los *Modelos Ocultos de Markov(MOM)* ni su aplicación en Reconocimiento Automático de Voz (RAV) son nuevos, la teoría básica fue publicada a finales de los 60's y principios de los 70's por Baum *et al*, y fue implantada para RAH en los 70's por Baker en la Universidad de Carnegie-Mellon y Jelinek *et al* en IBM. Sin embargo, su comprensión y amplia difusión ha ocurrido durante los últimos años [Rabiner, 1989].

En esta sección se hará un breve análisis primero de las *cadenas de Markov*, luego de su extensión hacia *Modelos Ocultos de Markov (MOM)*, para finalizar con su aplicación en RAL. Cabe mencionar que un análisis profundo de cualquiera de los tres puntos mencionados queda fuera del contexto de este trabajo; sin embargo, se presentarán a continuación las bases mínimas necesarias para explicar su aplicación en RAL.

IV.4.1 Cadenas de Markov.

Considere el proceso de una sucesión de pruebas realizadas a intervalos regulares de tiempo, cuyos resultados pertenecen a un conjunto finito de estados $\{S_1, S_2, \ldots, S_N\}$ llamado espacio de estados; tal y como se ilustra en la figura 23 para el caso N=5.

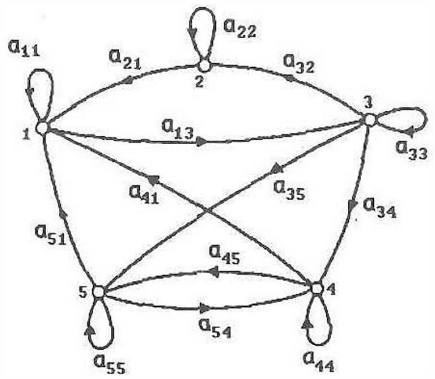


Figura 23. Cadena de Markov de 5 estados con transiciones de estado seleccionadas.

En cada una de las pruebas que se realizan a espacios discretos de tiempo regularmente espaciados ocurre un cambio de estado (o posiblemente se regresa al mismo) de acuerdo con un conjunto de probabilidades asociadas con el estado. Los instantes en los cuales ocurre un cambio de estado se denotan como t= 1, 2, ...; y el estado correspondiente al tiempo t es denotado como q_t . Una descripción probabilística completa del proceso podría requerir la especificación del estado actual q_t , así como también de los estados anteriores. Para el caso especial de una cadena de Markov de primer orden, esta descripción probabilística está truncada a únicamente el estado actual y el estado anterior; esto es

$$P[q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, \ldots] = P[q_t = S_j | q_{t-1} = S_i] \qquad (82)$$

donde $P[q_t = S_j | q_{t-1} = S_i]$ es la probabilidad de transitar hacia el estado S_j en el instante t, dado que el sistema se encontraba en el estado S_i en el instante t-1.

Se considerará además que en el proceso $P[q_t = S_j | q_{t-1} = S_i]$ es independiente del tiempo, conduciendo a un conjunto de probabilidades de transición de estado a_{ij} de la forma

$$a_{ij} = P[q_t = S_j | q_{t-1} = S_i] \qquad 1 \le i, j \le N \qquad (83)$$

donde los coeficientes probabilidad de transición de estado tienen las propiedades

$$a_{ij} \geq 0 \tag{84}$$

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{85}$$

y pueden ser arreglados en una matriz estocástica llamada *matriz de probabilidad de transición de estados*

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{bmatrix} \tag{86}$$

El proceso estocástico anterior puede ser llamado un *Modelo Observable de Markov o Modelo Evidente de Markov (MEM)*, ya que la salida del mismo pertenece al conjunto de estados, donde cada estado corresponde a un evento fisico observable. Para clarificar ideas puede revisarse el ejemplo del apéndice C.

IV.4.2 Extensión a Modelos Ocultos de Markov (MOM).

En general los *Modelos Evidentes de Markov (MEM)* describen procesos en los cuales cada estado corresponde a un evento determinísticamente observable, es decir, la salida en cualquiera de los estados no es aleatoria. Si bien éstos pueden ser aplicados en muchos problemas de interés, en la aplicación relacionada con esta tesis es necesario extender el concepto de *Modelos de Markov* hacia situaciones en las cuales la observación es una función probabilística del estado; esto es, el modelo resultante llamado *Modelo Oculto de Markov (MOM)* es un proceso estocástico doblemente empotrado, con un proceso estocástico subyacente que no es observable directamente (v.g., se encuentra *oculto*), sino a través de otro conjunto de procesos estocásticos que producen la secuencia

de observaciones.

Con el fin de introducir los *Modelos Ocultos de Markov (MOM)* considérese el siguiente escenario:

En un cuarto con una barrera (v.g., una cortina) se encuentran dos hombres, uno a cada lado de la barrera, que no pueden verse uno al otro, sólo escucharse. El sujeto 1 se encuentra realizando experimentos de lanzamientos de dado (o dados), y éste no le dirá nada acerca de lo que está haciendo al sujeto 2, excepto el resultado de los lanzamientos.

En este escenario se está realizando una secuencia de lanzamientos de dados ocultos, que dan como resultado una secuencia de observaciones tomadas del conjunto {1, 2, 3, 4, 5, 6}; por ejemplo una secuencia observación típica sería $O = o_1, o_2, \ldots, o_7 = 3, 1, 3, 6, 4, 2$.

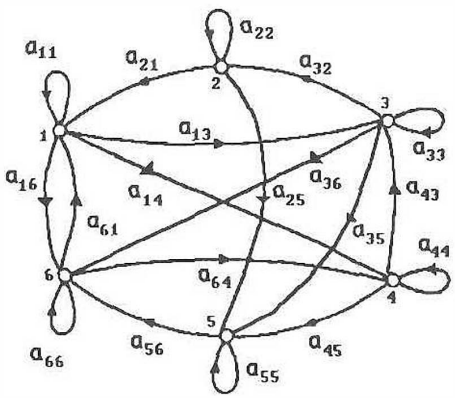


Figura 24. Cadena de Markov de 6 estados para el modelo de un dado.

Dada esta situación, el problema consiste en formular un modelo para explicar las secuencias de dígitos observados. Si se decide emplear un MOM, el primer problema es decidir a que corresponden los estados y cuantos deben ser éstos. A partir de las observaciones podría inferirse que se trata de lanzamiento de dados y suponer que se trata de solamente uno, lo cual podría ser analizado bajo un *Modelo Evidente de Markov (MEM)* de 6 estados como el mostrado en la figura 24, donde cada estado corresponde a una de las caras del dado y el único problema es encontrar los valores de los coeficientes de la matriz

de transición de estados.

Otra forma de plantear el modelo es utilizando un MOM, suponiendo que la secuencia de observaciones proviene del lanzamiento de dos dados, tal y como se muestra en la figura 25. En este caso hay dos estados en el modelo, y cada estado corresponde a un dado diferente que es lanzado. Cada estado está caracterizado por una distribución de probabilidad de los dígitos {1, 2, 3, 4, 5, 6}, y las transiciones entre estados (dados) están determinadas por una matriz de transición de estados.

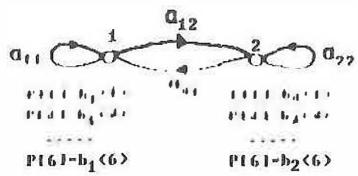


Figura 25. *Modelo Oculto de Markov (MOM)* de 2 estados para dos dados.

El modelo de la figura 25 representa una buena idea acerca de lo que es un MOM y como puede ser aplicado en algunas situaciones. Con el fin de establecer formalmente un MOM es conveniente enumerar los elementos que lo componen. Un MOM está caracterizado por [Rabiner, 1989]:

1) N el número de estados en el modelo, donde los estados individuales son denotados como $S = \{S_1, S_2, \ldots, S_N\}$ y el estado en el tiempo t es q_t.

2) M el número de símbolos distintos de observación en cada estado, los cuales corresponden a la salida física del sistema que está siendo modelado.

3) La *matriz de probabilidad de transición de estados* $A = \{a_{ij}\}$, donde

$$a_{ij} = P[q_{t+1} = S_j | q_t = S_i] \qquad 1 \le i, j \le N \qquad (87)$$

4) La distribución de probabilidad de los símbolos en el estado j, $B = \{b_j(k)\}$, donde

$$b_j(k) = P[v_k \text{ en } t | q_t = S_j], \quad 1 \le j \le N, \quad 1 \le k \le M \qquad (88)$$

5) La distribución de probabilidad inicial de estados $\Pi = \{\pi_i\}$, donde

$$\pi_i = P[q_1 = S_i], \qquad 1 \le i \le N \tag{89}$$

Dados los valores apropiados de *N, M, A, B* y Π, el MOM puede ser utilizado como un generador para dar la secuencia de observaciones

$$O = o_1, o_2, \ldots, o_T \tag{90}$$

En el resto de este capítulo se utilizará por conveniencia la notación compacta

$$\lambda = (A, B, \prod) \tag{91}$$

para referirse al conjunto completo de parámetros de un MOM.

En la aplicación de MOM existen tres problemas básicos [Rabiner, 1989]:

Problema 1: Dada la secuencia de observaciones $O = o_1, o_2, \ldots, o_T$ y un modelo $\lambda = (A, B, \prod)$, ¿Cómo se calcula eficientemente $P[O, \lambda]$, la probabilidad de producir una secuencia de observaciones a partir del modelo ?

Problema 2: Dada la secuencia de observaciónes $O = o_1, o_2, \ldots, o_T$ y el modelo λ, ¿ Cómo se puede elegir la secuencia de estados correspondiente $Q = q_1, q_2, \ldots, q_T$ que es óptima en algún sentido (v.g., que mejor "explica" las observaciones) ?

Problema 3: ¿ Cómo se pueden ajustar los parámetros del modelo $\lambda = (A, B, \prod)$ para maximizar $P[O, \lambda]$?

El problema 1, puede ser visto también como la evaluación de ¿Que tan bien ajusta el modelo la secuencia de observaciones dada ?; este punto de vista es bastante conveniente ya que se puede aplicar en el caso de que estemos tratando de elegir, entre varios modelos, áquel que se ajusta mejor a la secuencia de observaciones. Si bien existe un procedimiento de "fuerza bruta" para resolver este problema, también existe un procedimiento optimizado llamado el *procedimiento de adelanto-atraso* [Rabiner, 1989].

En el problema 2 se está tratando de "descubrir" la parte oculta del modelo; esto es tratando de encontrar la secuencia "correcta" de estados. La solución de este problema permite aprender acerca de la estructura del modelo oculto, u obtener promedios

estadísticos de estados individuales. Sin embargo, tal secuencia "correcta" no existe, y en situaciones prácticas usualmente se utiliza un criterio de optimización para resolver el problema de la mejor manera posible; por ejemplo, un criterio podría ser elegir la secuencia de estados que es *individualmente* más probable, lo cual maximizaría el número de estados "correctos" individualmente. Otro criterio podría ser obtener la mejor secuencia de estados que maximice $P[Q| O,\lambda]$, lo cual es equivalente a maximizar $P[Q, O| \lambda]$; bajo este criterio existe una técnica basada en programación dinámica, y que es conocida como el *algoritmo de Viterbi* [Rabiner, 1989].

El problema 3 es por mucho el más complejo de los tres, esto es, ajustar los parámetros del modelo $\lambda = (A, B, \prod)$ para satisfacer un criterio de optimización. No existe manera analítica de encontrar el conjunto de parámetros del modelo que maximizan la probabilidad de generar una secuencia de observaciones en una manera cerrada. Sin embargo, es posible elegir los parámetros del modelo $\lambda = (A, B, \prod)$ tal que la probabilidad $P[O,\lambda]$ pueda ser maximizada localmente usando un procedimiento iterativo conocido como el algoritmo de Baum-Welch [Rabiner, 1989].

La descripción detallada de los métodos para la solución de cada uno de los tres problemas expuestos está fuera del alcance de esta tesis, y más aún su aplicación. Para obtener más información acerca de esos procedimientos se pueden revisar las referencias [Levinson *et al*, 1983] y [Rabiner, 1989].

IV.4.3 Aplicación de Modelos Ocultos de Markov (MOM) en Reconocimiento Automático de Locutor (RAL).

Como podría entreverse a partir del análisis planteado anteriormente, la idea en la aplicación de MOM en RAL es fundamentalmente la de generar un MOM para cada "cliente" del sistema (v.g., entrenamiento) y a partir de cada modelo almacenado y dada una secuencia de vectores de observación, realizar el reconocimiento del locutor (v.g.,

identificación o verificación).

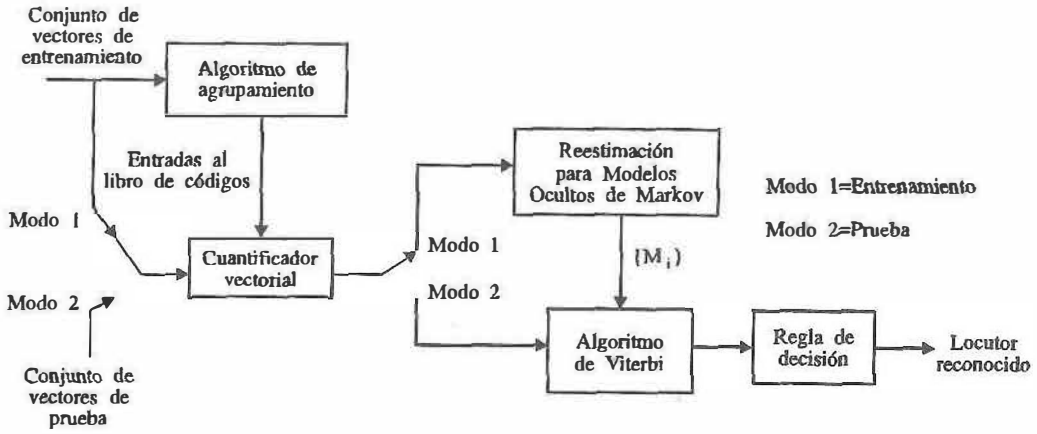


Figura 26. Diagrama a bloques del sistema de RAL basado en MOM.

La figura 26 muestra el diagrama a bloques completo de un sistema para Reconocimiento Automático de Locutor (RAL); El sistema funciona en dos modos: entrenamiento y prueba. Durante el entrenamiento, a partir de todas las muestras de voz de todos los locutores, que han sido transformadas en secuencias de vectores de parámetros, se obtiene un cuantificador vectorial con información relativa a todos los locutores. Posteriormente los vectores correspondientes a cada uno de los locutores son codificados a través del cuantificador vectorial para producir una secuencia de observaciones; esta secuencia ha de servir para la obtención de un MOM que maximice la probabilidad de generarla (v.g., solución al problema 3 de la sección anterior).

De esta manera se obtiene un MOM para cada locutor.

En el modo de prueba, un conjunto de vectores de parámetros del locutor desconocido son codificados a través del cuantificador vectorial, lo cual genera una secuencia de observaciones de prueba. Esta secuencia de observaciones es aplicada a todos o uno de los MOM, según se trate de identificación o verificación de locutor respectivamente, y se calcula la probabilidad de generar la secuencia de observaciones con

los MOM almacenados (esto se hace usualmente a través del *algoritmo de viterbi* [Rabiner *et al*, 1983]). En el caso de IAL de umbral abierto (v.g., sólo "clientes") el sistema entrega como salida al "dueño" del MOM con la probabilidad más alta de generar la secuencia de observaciones de entre todos los MOM; si se trata de IAL de umbral cerrado (v.g., "clientes" e "impostores") esa máxima probabilidad es comparada con un umbral de decisión. Si se trata de VAL la probabilidad de generar la secuencia de observaciones con el MOM dado como presunto locutor es también comparada con un umbral, si la probabilidad es mayor que el umbral el locutor es "aceptado" de otra manera es "rechazado".

En experimentos realizados con el esquema de la figura 26 [Zheng y Yuan, 1988] en un sistema *dependiente del vocabulario* y utilizando una clase especial de cadenas de Markov llamadas *circulares*, se generaron MOM para 10 locutores (6 hombres y 4 mujeres) utilizando los coeficientes de predicción lineal de orden 12 en cada trama y obteniendo un cuantificador vectorial de tasa 6 (v.g., 64 entradas) usando el *algoritmo de división binaria*. Se generaron MOM para cada locutor con una matriz probabilidad de transición de estados A de orden 6 y una matriz de probabilidad de observaciones de 6x64. Al entrenar y evaluar el sistema con 7 frases chinas independientes en la modalidad de IAL de umbral abierto se obtuvo una tasa de error del 6.3%. Al cambiar de cadenas circulares de Markov hacia cadenas de Markov serie de izquierda-a-derecha la tasa de error aumentó al 11.3%.

Si bien se han planteado los MOM a partir de observaciones discretas; es decir, los vectores de entrada son codificados por un cuantificador vectorial y de esta manera las densidades de probabilidad asignadas a cada estado son discretas, también es posible utilizar densidades de probabilidad contínuas, que tienen la ventaja de eliminar la distorsión causada por la cuantificación vectorial. El análisis de tales clases de MOM queda fuera del alcance de esta tesis, y sólo se mencionará que en experimentos realizados con MOM de densidades contínuas e independencia del vocabulario (en realidad la misma base de datos que se utilizó

en los experimentos analizados en la sección de cuantificación vectorial (CV)), se obtuvo una tasa de equierror (TEE) ligeramente inferior al 3% con una secuencia de prueba de 4 dígitos [Tishby, 1991]. El autor concluye que la ligera mejoría obtenida sobre la técnica de CV (con la misma base de datos) no justifica la complejidad y esfuerzo requerido para implantar los MOM particulares que utilizó, sin embargo, es su opinión que la manera correcta de adaptar MOM para RAL es un problema abierto.

CAPÍTULO V REALIZACIÓN DEL SISTEMA EXPERIMENTAL DE RECONOCIMIENTO AUTOMÁTICO DE LOCUTOR (SERAL).

V.1 Discusión.

El sistema realizado es dependiente del vocabulario y se basa en una técnica de alineamiento dinámico en el tiempo (ADT). La técnica particular implantada es la descrita en la sección IV.2, y los parámetros seleccionados para caracterizar al locutor son los coeficientes *cepstrum*, los cuales son obtenidos a partir de los coeficientes de predicción lineal (CPL); éstos a su vez son obtenidos por medio del método de la autocorrelación y el *algoritmo de Levinson-Durbin* sobre tramas de tiempo cuya longitud y corrimiento puede ser establecida, al igual que el orden de predicción (v.g., el número de parámetros por trama). Finalmente, antes de describir el sistema es necesario establecer la razón del adjetivo "EXPERIMENTAL" dentro del nombre SERAL; ésto se hará bajo las siguientes consideraciones:

a) El sistema que se describirá a continuación es completamente configurable, teniendo como única limitante la cantidad de memoria que puede ser manejada convencionalmente por una computadora personal, administrada por el sistema operativo MS-DOS, y la cantidad de memoria disponible en la tarjeta de adquisición, procesamiento y síntesis de señales (TAPS). Esta característica de configurabilidad lo hace acreedor al calificativo "EXPERIMENTAL", ya que es posible realizar experimentos con el sistema bajo diferentes parámetros, locutores, modalidades, etc.; tal y como se verá más adelante.

b) La eficiencia del sistema estará altamente influenciada por todas los factores que interfieren con la tarea de RAL, y como tal no debe de considerarse un sistema óptimo. Ésta es otra razón para considerarlo "EXPERIMENTAL".

V.2 Infraestructura física utilizada.

La componente física de SERAL está formada principalmente por una tarjeta desarrollada *exprofeso* llamada "Tarjeta de Adquisición, Procesamiento y Síntesis de señales (TAPS)", y una computadora personal compatible con los modelos de IBM; además de equipo externo tal como micrófono, grabadora/reproductora de audiocintas, amplificador de potencia de salida, y bocina.

V.2.1 Tarjeta de adquisición, procesamiento y síntesis de señales (TAPS).

La figura 27 muestra un diagrama a bloques de la TAPS. El núcleo de la tarjeta lo constituye el circuito procesador digital de señales DSP32 de la compañía AT&T, el cual funciona con un reloj de 25 MHz y es capaz de ejecutar 12.5 MOPFS (millones de operaciones en punto flotante por segundo) [Martínez y Félix, 1993]. El DSP32 puede manejar un total de 64 koctetos; 6 koctetos se encuentran integrados internamente como memoria de lectura/escritura (MLE), al igual que 2 koctetos de memoria de lectura exclusiva (MLEX) que contienen rutinas de uso general pregrabadas por el fabricante.

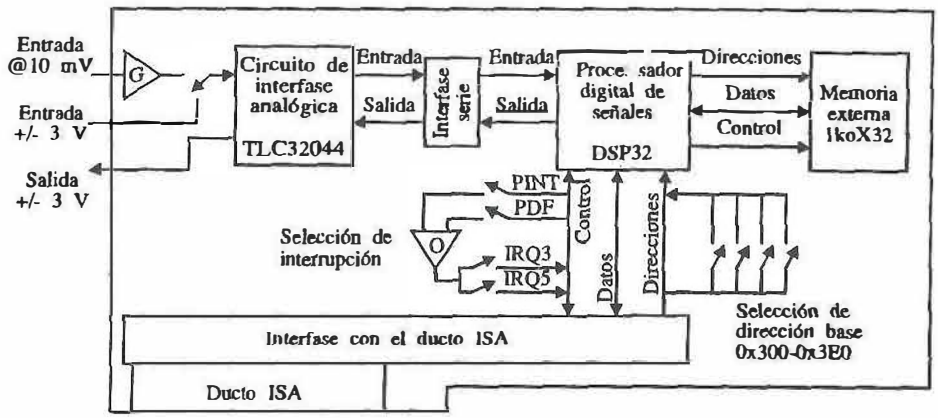


Figura 27. Diagrama a bloques de la tarjeta de adquisición, procesamiento y síntesis de señales (TAPS).

El procesador digital de señales DSP32 y su memoria externa de 32 koctetos se encuentran integrados en un módulo de procesamiento que es una réplica del desarrollado

por Martínez y Félix en 1992 [Martínez y Félix, 1993]. La TAPS fue desarrollada para ser albergada y controlada por una computadora personal compatible con los modelos de IBM; en particular, esta tarjeta fue insertada en una de las ranuras de expansión de una computadora personal con el microprocesador 80386 funcionando con un reloj de 33 MHz.

La comunicación entre la computadora y la TAPS se realiza a través del ducto de aquella, mediante una interfaz adecuada; la cual le permite accesar cada uno de los registros internos del DSP32. El direccionamiento de los registros internos se hace a partir de una dirección base (la cual corresponde a un registro particular y por omisión corresponde a 0x300).

Existen dos líneas físicas del DSP32 que pueden ocasionar una interrupción de circuitería por medio del ducto ISA, el nombre de esas líneas es PINT y PDF. En la TAPS se han proveido mecanismos para que cualquiera de esas dos líneas (o ambas) puedan establecer una interrupción a través de las líneas del ducto IRQ3 y/o IRQ5.

Se dispone en la TAPS de dos canales de adquisición de señales analógicas, uno para señales de bajo nivel (@10 mVrms) y otro para señales de alto nivel (+/- 3 V max), además de un canal de salida (+/- 3 V max). La señal de bajo nivel es amplificada por un preamplificador de ganancia variable, cuyo intervalo dinámico va desde 0 dB hasta 66.6 dB.

La adquisición y síntesis de señales analógicas se realiza a través del circuito de interfase analógica (CIA) TLC32044 de la compañía Texas Instruments. Las muestras para adquisición y síntesis son enviadas en un formato de 14 bits mediante un protocolo de comunicación serie entre el TLC32044 y el DSP32. Si bien el protocolo de comunicación del TLC32044 es compatible con varios miembros de la familia de PDS de Texas Instruments, éste no era el caso con el DSP32, por lo que fue necesario diseñar una interfase de comunicación serie entre ambos circuitos.

Se seleccionó el circuito TLC32044 debido a que ofrece la ventaja de integrar en una

simple pastilla los siguientes componentes (figura 28):

a) Filtros pasabajas y pasaaltas anti-tralslape de entrada de capacitores conmutados de orden 8 y 4 respectivamente.

b) Convertidor A/D de 14 bits en formato de magnitud y signo.

c) Convertidor D/A de 14 bits en formato de magnitud y signo.

d) Filtros pasabajas de reconstrucción y de corrección [(*sen x*)/*x*] de salida,de capacitores conmutados de orden 8 y 2 respectivamente.

Después del encendido, el circuito de interfase analógica toma por omisión una configuración tal que la señal analógica es filtrada pasabanda desde 100 Hz hasta 3.7 kHz y muestreada a una frecuencia de 8 kHz. Esta configuración puede ser alterada con el fin de cambiar esos y otros parámetros, enviando mandos de configuración desde el DSP32 hacia el TLC32044 vía serie siguiendo un protocolo regido por este último. La configuración por omisión fue alterada para que la adquisición se hiciera a escala completa (v.g., 14 bits); en vez de la media escala (v.g., 13 bits) que utilizaba originalmente.

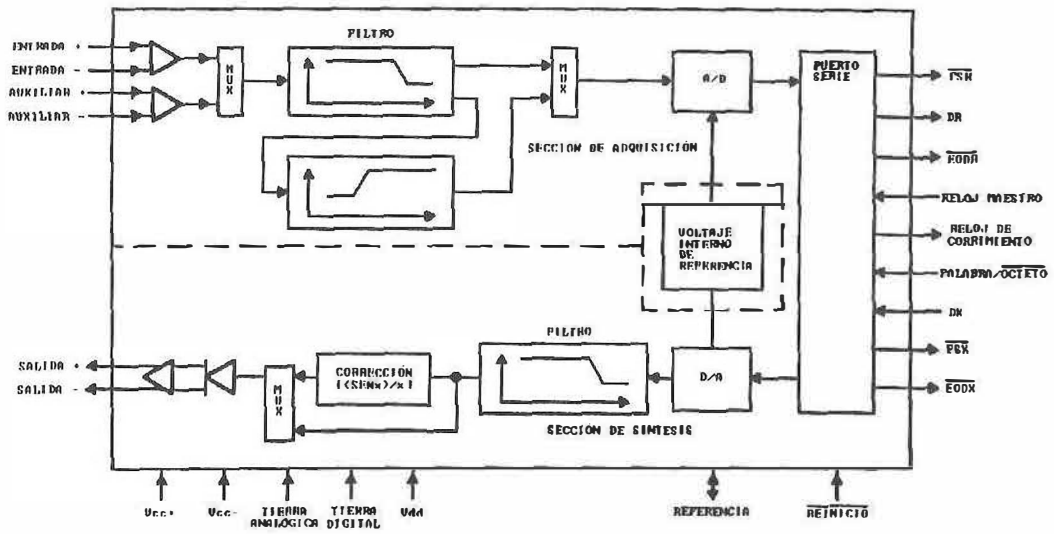


Figura 28. Diagrama a bloques del circuito de interfase analógica (CIA) TLC32044.

V.2.2 Sistema físico integrado.

Idealmente, el único elemento adicional para el sistema, además de la TAPS y la computadora, sería un micrófono que iría conectado al preamplificador de ganancia variable; sin embargo, al hacer pruebas de adquisición de señales de voz bajo esta configuración se encontró que había una cantidad considerable de ruido eléctrico que se agregaba a la señal a través de la línea de tierra del preamplificador, la cual fué tomada de la tierra analógica del ducto de la computadora.

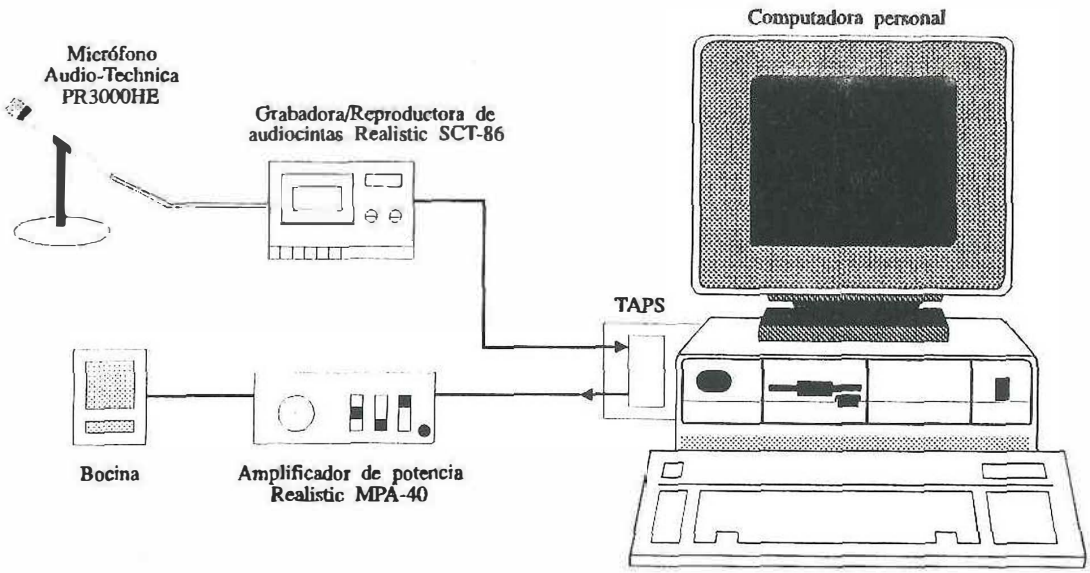


Figura 29. Diagrama esquemático del sistema físico integrado.

La alternativa final fue utilizar un amplificador externo para la señal proveniente del micrófono y adquirirla a través del canal de señal de alto nivel de la TAPS, tal y como se observa en la figura 29. La amplificación externa se hizo a través de la grabadora /reproductora de audiocintas *Realistic SCT-86*, y la cantidad de ruido agregado disminuyó considerablemente.

Si bien para los propósitos de este trabajo no era necesario disponer de reproducción (i.e, síntesis) de voz, esta alternativa fue agregada como una herramienta de validación a

nivel circuitería y programación, de tal manera que durante la adquisición y reproducción es posible "escuchar" las señales de voz conectando la salida de la TAPS a un amplificador de potencia y de éste a una bocina, o directamente a esta última, ya que el TLC32044 puede proveer potencia para una carga de baja impedancia. En este caso se decidió utilizar el amplificador de potencia *Realistic MPA-40*, ya que se dispone de la ventaja de controlar el volumen manualmente así como la igualación de frecuencias que tiene éste disponible.

La figura 29 muestra el sistema físico integrado. El micrófono utilizado fue un *Audio-technica PR3000HE*, el cual tiene como principales características: un patrón *hipercardiode*, una bobina móvil como transductor, y una planicidad frecuencial aceptable.

V.3 Infraestructura programática desarrollada.

El sistema completo es controlado y administrado por un programa principal llamado "SERAL" que fue desarrollado en lenguaje de alto nivel "C" utilizando el compilador de Borland *TURBOC 2.0*. Además de controlar al sistema, el programa SERAL sirve como interfase con el usuario y tiene una cantidad considerable de capacidades gráficas y analíticas, entre las cuales destacan:

a) Capacidad de adquisición de secuencias de hasta 8 segundos de longitud (v.g., 64,000 muestras) para ser almacenadas en el disco duro.

b) Capacidad de reproducción de secuencias almacenadas en el disco duro.

c) Pseudo-osciloscopio integrado con capacidad de cambiar ambas escalas vertical y horizontal para verificar el canal de adquisición.

d) Capacidades gráficas para visualizar secuencias adquiridas, *contornos* de energía (ver siguiente sección), *contornos* de parámetros, alineamiento y comparación de *contornos*; entre otras.

e) Modo manual de entrenamiento para generar patrones de referencia para cada loctutor ("cliente") de la base de datos.

f) Identificación y verificación de locutor en modo directo (v.g., tiempo real: inmediatamente después de la adquisición) o en modo archivo (i.e, a partir de una secuencia almacenada en el disco duro).

g) Identificación y verificación de locutor con ambos umbrales: individual (v.g., un umbral para cada "cliente") y general (v.g., el mismo umbral para todos los "clientes").

Con el fin de desarrollar no solamente las tareas descritas anteriormente, sino todas sus capacidades, el programa SERAL se auxilia de 5 programas desarrollados en lenguaje ensamblador del DSP32 y compilados bajo un programa compilador de la compañía AT&T. Un programa desarrollado en "C" se encarga de cambiar el formato de uno de los archivos generados por el compilador (*nombre*.L) a un formato propio de SERAL (*nombre*.COD), el cual es un formato ASCII de la forma:

Dirección{tabulador}*Dato o instrucción*{Retorno de carro}

donde *Dirección* es una palabra ASCII que representa un número de 4 dígitos hexadecimales y *Dato o instrucción* puede variar desde 2 hasta 8 dígitos hexadecimales. Ésto se hizo con la intención de disponer en el programa SERAL de una rutina para *cargar* los programas *nombre*.COD. A continuación de describirán brevemente los programas desarrollados para el DSP32, de los cuales tres pueden ser considerados algoritmos implantados.

V.3.1 Adquisición y reproducción de señales.

La adquisición de señales se realiza cuando la computadora carga y permite la ejecución en el DSP32 del programa GRABA.COD. En el arribo de cada muestra al DSP32, el orden de los bits de ésta es invertido y se cambia de formato *magnitud y signo* al formato *complemento a 2*. El DSP32 genera luego una interrupción a la computadora para la lectura de la muestra [ver el apéndice E].

El modo en el cual se realiza la adquisición/ síntesis es síncrono, es decir, al mismo tiempo que el TLC32044 está escribiendo una muestra en la muelle serie de entrada del DSP32, está también leyendo la muelle serie de salida.

La secuencia adquirida es almacenada en el disco duro de la computadora bajo un nombre (con extensión VOZ por omisión; v.g., *nombre*.VOZ), para su posterior procesamiento, o es pasada directamente como un parámetro a un proceso de identificación/verificación de locutor.

El proceso de reproducción de señales (síntesis) se realiza cuando la computadora carga y permite la ejecución en el DSP32 del programa REPROD.COD. La secuencia a reproducir es leída del disco duro y enviada por la computadora muestra a muestra a un registro interno del DSP32, previa verificación de una bandera de lectura [ver apéndice E].

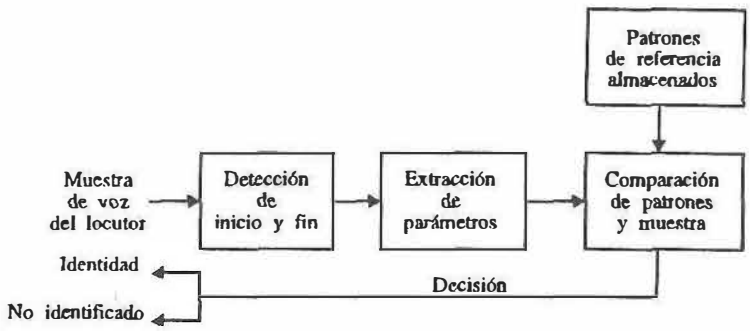


Figura 30. Diagrama esquemático del proceso de *identificación de locutor*.

El programa REPROD.COD que se encuentra corriendo en el DSP32 sincroniza la lectura del registro donde se escribió la muestra con el arribo de una nueva muestra en la muelle de entrada; la muestra leida es cambiada de formato *complemento a dos* a *magnitud y signo*, el orden de sus bits es invertido, y ésta es escrita en la muelle de salida.

V.3.2 Algoritmos implantados.

Considérese que el esquema de las figuras 2 y 3 para *identificación y verificación de*

locutor ha sido modificado para incluir un procedimiento de *deteccción de inicio y fin*, dando como resultado las figuras 30 y 31 respectivamente. Los tres bloques principales de estas figuras:

a) Detección de inicio y fin.

b) Extracción de parámetros.

c) Comparación de patrón(es) y muestra.

corresponden con los tres algoritmos que fueron implantados, y que serán analizados a continuación.

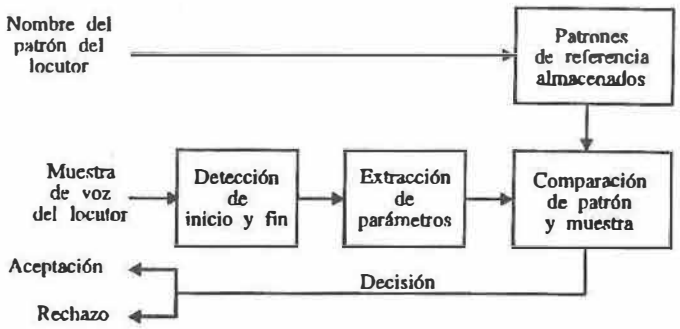


Figura 31. Diagrama esquemático del proceso de *verificación de locutor*.

V.3.2.1 Detección de inicio y fin de palabras aisladas.

El procedimiento de *detección de inicio y fin* es incluido como un algoritmo que permita discriminar entre los segmentos de las secuencia *s(n)* que contienen información dada por el locutor (hablada), y los demás eventos que ocurren durante la grabación; los cuales pueden ser agrupados bajo el nombre de *ruido de fondo*.

Este procedimiento no es trivial, excepto en aquellos casos en los que el ambiente de grabación permite una alta relación señal a ruido (cámaras anecoicas, cuartos a prueba de ruido) y se pueden realizar grabaciones de alta calidad. En esas situaciones el nivel de los sonidos débiles de la voz es mayor que el nivel del ruido de fondo, y una medición del nivel

de energía de la señal es suficiente para discriminar entre voz y ruido de fondo.

Sin embargo, tales situaciones no ocurren en aplicaciones en el mundo real, y existe muy poco material publicado al respecto, debido a que la mayoría de los sistemas de laboratorio usan grabaciones considerablemente limpias que les permiten realizar la detección de inicio-fin directamente en el contorno de energía de la señal, mientras que los fabricantes de productos comerciales, que tienen que preocuparse acerca de las condiciones de un ambiente real, se muestran renuentes a publicar sus algoritmos robustos.

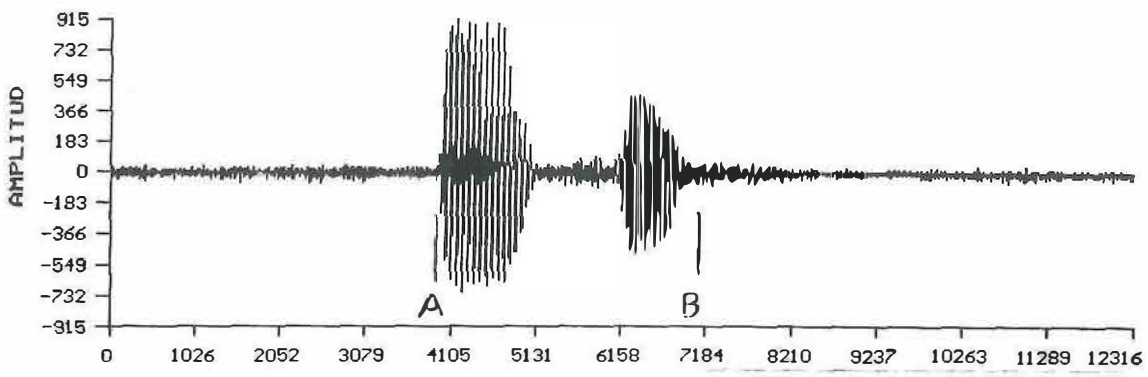


Figura 32. Forma de onda de la palabra /fases/.

El procedimiento de *detección de inicio y fin* (del inglés *endpoint detection* o *word boundaries detection*) recibe ese nombre porque su propósito es obtener un estimado de los índices de la secuencia *s(n)* en los cuales *inician* y *terminan* las palabras habladas durante la adquisición. La mayoría de los algoritmos propuestos para la *detección de inicio y fin* llevan a cabo esta tarea segmentando la secuencia *s(n)* en tramas y obteniendo un estimado de algunos parámetros en cada una de ellas. El algoritmo implantado está basado en uno sugerido por Lamel *et al* [Lamel *et al*, 1981], y tiene la restricción de que debe ser evaluado en un ambiente con una relación señal a ruido (S/R) de al menos 25 dB.

La efectividad de la *detección de inicio-fin* de palabras aisladas está altamente relacionada con el tipo de sonidos que las limitan. Uno de los problemas más grandes que se

puede encontrar ocurre cuando las palabras inician y/o terminan con sonidos fricativos. La naturaleza de los sonidos fricativos ocasiona que estos puedan ser confundidos con ruido, de tal manera que es difícil determinar los límites de una palabra acotada con tales sonidos aún "a simple vista."

La figura 32 es un ejemplo claro de ésto, ahí se muestra la forma de onda de la palabra /fases/, esta palabra inicia y termina con los fricativos /f/ y /s/, respectivamente, además de contener un fricativo más en la parte intermedia. Sin ningún conocimiento *a priori* acerca de la palabra con que se está tratando, sería bastante lógico suponer que el inicio y fin de la palabra corresponden a los puntos marcados A y B, respectivamente. Sin embargo, estas suposiciones son erróneas ya que se están omitiendo por completo a los sonidos de los fricativos /f/ y /s/ que limitan a la palabra; tal y como se verá más adelante.

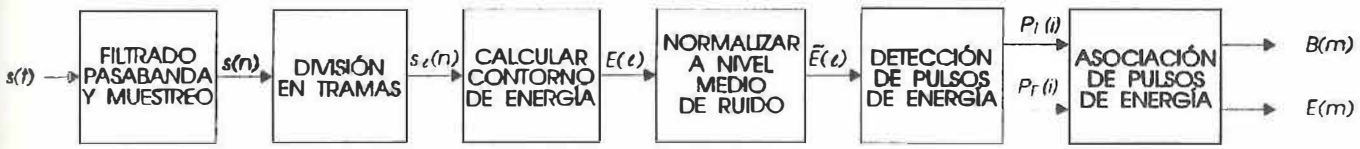


Figura 33. Diagrama a bloques del algoritmo de *detección de inicio y fin.*

Otro de los factores que complican la *detección de inicio y fin*, es que algunas palabras suelen venir acompañadas de una inhalación y exhalación de aire, en el principio y fin de la palabra, respectivamente. Esto se manifiesta generalmente como un pulso corto de baja energía que no debe ser considerado parte de la palabra. A continuación se presenta una lista de diversas categorías de problemas encontrados en la *detección de inicio y fin* de palabras aisladas [Rabiner y Schafer, 1978]:

a) Sonidos fricativos débiles (/f/, /d/, /j/) al inicio o fin.

b) Ráfagas de oclusivos débiles (/p/, /t/, /k/) al inicio o fin.

c) Nasales (/m/, /n/) al final.

d) Fricativos sonoros que llegan a ser no sonoros al final de las palabras.

e) Sonidos vocales disminuidos al final de una palabra.

La figura 33 muestra un diagrama a bloques del algoritmo. La señal de entrada *s(t)* es primero filtrada por medio de un filtro pasabanda, y luego muestreada, este proceso es realizado automáticamente al hacer una adquisición. La señal digitalizada *s(n)* es entonces preenfatizada, usando un filtro digital de primer orden cuya transformada z es:

$$H(z) = 1 - 0.95z^{-1} \tag{92}$$

La señal preenfatizada es después segmentada en tramas de N muestras, con un recorrimiento entre tramas de L muestras; los valores de N y L fueron 360 y 120 (45 ms y 15 ms) respectivamente. Cada trama es ponderada por una ventana Hamming de la forma:

$$Ham(n) = 0.54 - 0.46\cos\left[\frac{2\pi n}{N-1}\right] \qquad 0 \le n \le N-1 \tag{93}$$

Se denota la ℓ-ésima trama ponderada de la señal como $s_\ell(n)$ definida para $0 \le n \le N-1$. Después del procesamiento anterior se obtiene el contorno de energía de la señal calculando la suma de los cuadrados de los componentes de cada trama, lo cual equivale a calcular el coeficiente de autocorrelación de orden 0. El contorno de energía debe ser expresado en dB, esto es:

$$E(\ell) = 10\log_{10} R_\ell(0), \qquad \ell = 1,2,\ldots,NT \tag{94}$$

donde NT es el número total de tramas en el intervalo de grabación y $R_\ell(0)$ es el coeficiente de autocorrelación de orden 0 para la trama ℓ, que se calcula:

$$R_\ell(0) = \sum_{k=0}^{N-1}[s_\ell(k)]^2 \tag{95}$$

El siguiente procesamiento realizado sobre el contorno de energía es la normalización de nivel adaptable, que tiene como propósito normalizar el contorno de energía al nivel medio de ruido de fondo. Primero se calcula $E_{\min}$ como:

$$E_{\min} = \min_{1\le\ell\le NT}[E(\ell)] \tag{96}$$

el valor de E_{min} es entonces restado de $E(\ell)$:

$$\hat{E}(\ell) = E(\ell) - E_{min} \qquad \ell = 1, 2, \ldots, NT \qquad (97)$$

Enseguida se calcula un histograma del contorno de energía $\hat{E}(\ell)$ para todos los valores no mayores de 10 dB; se aplica entonces un suavizador de mediana de tres puntos [Rabiner *et al*, 1975] sobre el histograma calculado y se obtiene la moda de éste, el cual es el valor que se presentó con mayor frecuencia. Finalmente se calcula el contorno de energía normalizado al nivel medio de ruido $\tilde{E}(\ell)$, $\ell = 1, 2, \ldots, NT$:

$$\tilde{E}(\ell) = \hat{E}(\ell) - \text{moda} \qquad (98)$$

Este contorno normalizado tiene la propiedad de que en las tramas de silencio (v.g., ruido de fondo) el valor de la energía fluctúa alrededor de 0 dB, mientras que en las tramas de voz su valor es mucho más alto.

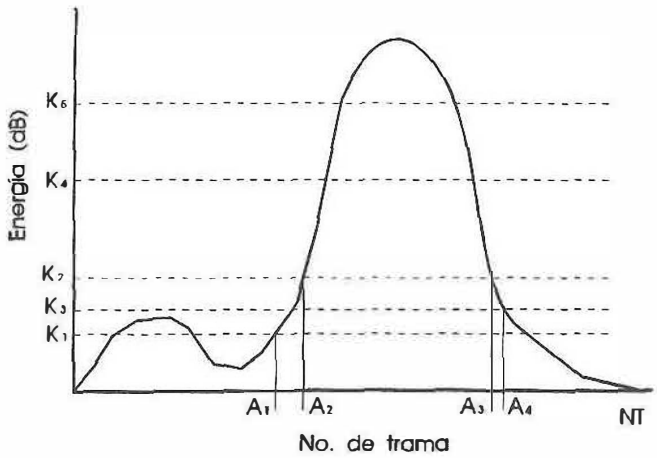


Figura 34. Diagrama ilustrativo de la detección de pulsos de energía.

El siguiente paso es la detección de pulsos de energía sobre $\tilde{E}(\ell)$. El algoritmo de *detección de inicio y fin*, se basa en la suposición de que cada una de las palabras está formada por uno o más pulsos de energía; el problema radica en encontrar esos pulsos y determinar cuales de ellos pertenecen a las palabras habladas y cuales al *ruido de fondo*.

Basándose en los valores del contorno normalizado $\hat{E}(\ell)$, se definen 5 niveles de

energía, tal y como se ilustra en la figura 34. El detector de pulsos de energía procede de izquierda a derecha, los valores de $\hat{E}(\ell)$ son monitoreados variando ℓ y cuando $\hat{E}(\ell)$ excede el primer umbral K_1, el número de trama A_1 es registrado como un posible inicio de pulso; si $\hat{E}(\ell)$ excede el valor de K_2 sin caer antes por debajo del valor de K_1, se detecta un inicio de pulso. El inicio de pulso es almacenado como A_1, a menos que la diferencia entre A_1 y A_2 sea demasiado grande, esto es, mayor que T_1, en cuyo caso el inicio del pulso es almacenado como A_2; esta situación indica generalmente inhalación al inicio de la palabra

La detección de fin de pulso se realiza de manera similar, en este caso si la diferencia entre A_3 y A_4 es muy grande (mayor que T_2 tramas) la trama A_3 es tomada como fin de pulso, ya que esto indica exhalación al final de la palabra, de lo contrario A_4 es almacenado como fin de la palabra.

Se realizan dos mediciones posteriores sobre cada pulso detectado. Se obtiene la energía pico en el pulso, y si ésta es menor que el umbral K_4 el pulso es deshechado, también se mide la duración del pulso y si éste es menor que T_3 tramas, el pulso es rechazado. El propósito del umbral K_5 es deshechar todos aquellos contornos normalizados $\hat{E}(\ell)$ cuyo valor pico sea menor que este nivel de energía.

El valor de los parámetros K_1, K_2 y K_3 determinado experimentalmente fue 5, 9 y 5 dB, respectivamente. Todos los pulsos con un nivel pico de energía inferior a K_4=12 dB fueron rechazados, al igual que todas las grabaciones con una energía pico inferior a K_5=25 dB. Los parámetros T_1 y T_2 fueron establecidos en 5 tramas, al igual que T_3, mientras que T_4 fue establecido en 10 tramas.

La salida del detector de pulsos de energía es un conjunto de pares de índices indicando el inicio $P_I(m)$ y el fin $P_F(m)$ de cada pulso para m=1,2,...,M, donde M es el número de pulsos detectados en el intervalo de grabación.

En el último bloque de la figura 33, el propósito de la asociación de pulsos de energía

es determinar que pulsos de energía están suficientemente cerca como para formar parte de una palabra multipulso. Es frecuente que palabras con sonidos plosivos intermedios presenten una disminución considerable de energía en el instante en que ocurrió la oclusión del tracto vocal, ocasionando que ésta sea detectada como varios pulsos de energía.

Bajo esta consideración, los pulsos de energía son agrupados en combinaciones de pares de inicio-fin para formar un sólo par inicio-fin de palabras. La figura 35 muestra tres pulsos de energía con separación D_1 y D_2 tramas, si ambos D_1 y D_2 son menores que T_4 tramas, los puntos A y F son marcados como inicio y fin de la palabra. Si D_1 es menor que T_4 tramas, y D_2 no lo es, se detectan dos palabras con pares inicio-fin A-D y E-F, respectivamente. De manera similar se resuelve para D_1 no menor que T_4 tramas, y D_2 menor que T_4 tramas.

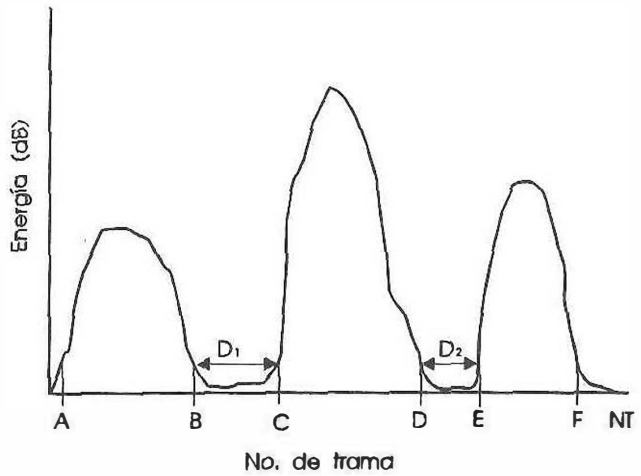


Figura 35. Diagrama ilustrativo de la asociación de pulsos de energía.

El diagrama de flujo interactivo computadora-DSP32 para el algoritmo de *detección de inicio y fin* puede verse en el apéndice E. La figura 36 muestra un ejemplo de la aplicación de este algoritmo sobre la secuencia adquirida de la palabra /fases/ de la figura 32. Esta figura corresponde a un despliegue gráfico obtenido por medio de un submenú del programa SERAL. En la parte superior se muestra el contorno de energía $\tilde{E}(\ell)$, mientras

que en la parte baja se encuentra la secuencia adquirida. Las líneas verticales indican los índices de trama detectados como inicio y fin de la palabra.

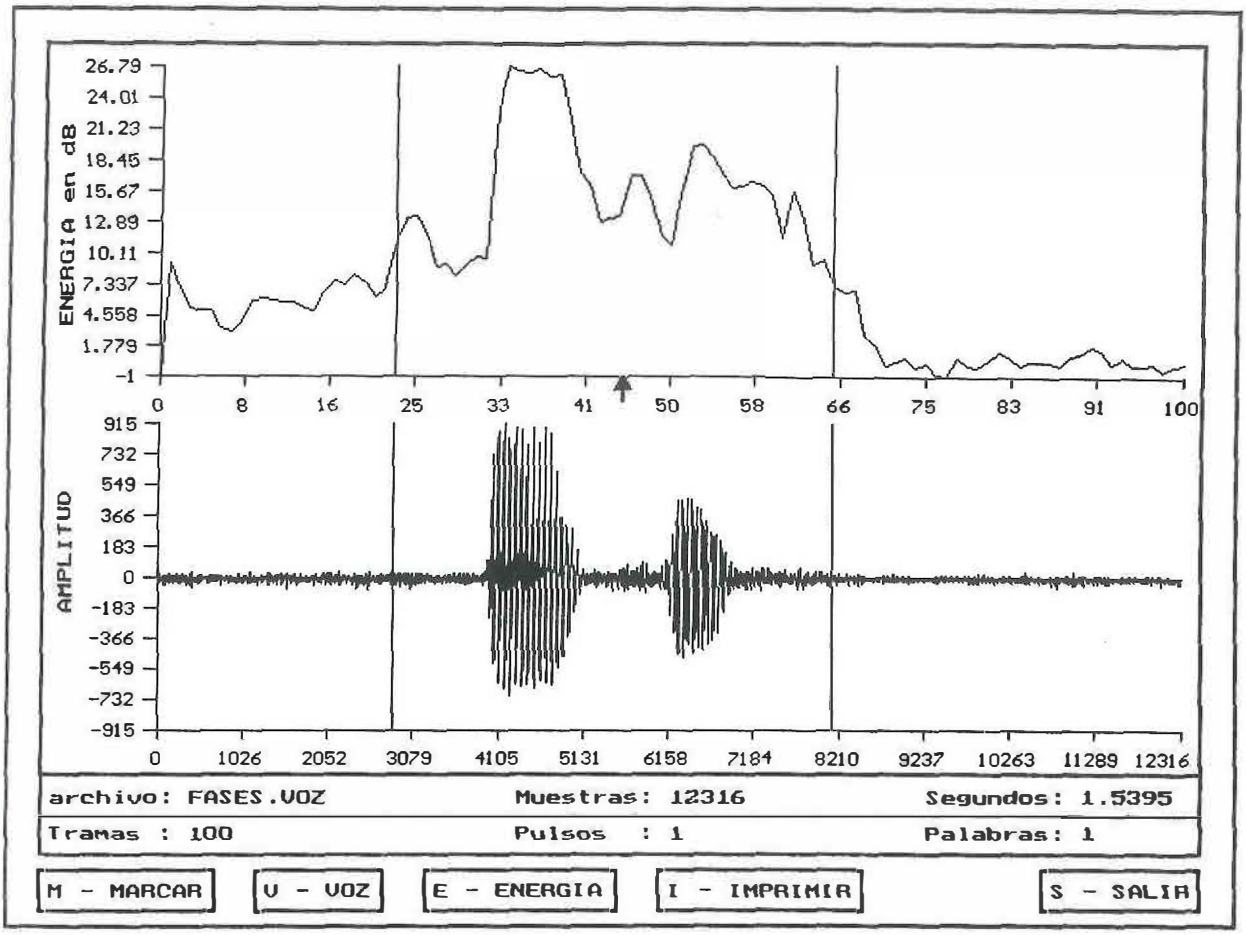


Figura 36. Despliegue gráfico de SERAL ilustrando la *detección de inicio y fin* para la palabra /fases/ de la figura 32.

Al pie de las gráficas se ha impreso información relativa al archivo *fuente* sobre el que se realizó la *detección de inicio y fin*, además de información relacionada con el algoritmo tal como: número de tramas *NT* en la secuencia, número de pulsos detectados, y número de palabras detectadas.

En realidad el despliegue gráfico de SERAL para la *detección de inicio y fin*, marca primero los índices inicio y fin de pulsos con líneas segmentadas, y aquellos índices que

corresponden a inicio y fin de palabras son marcados con líneas continuas.

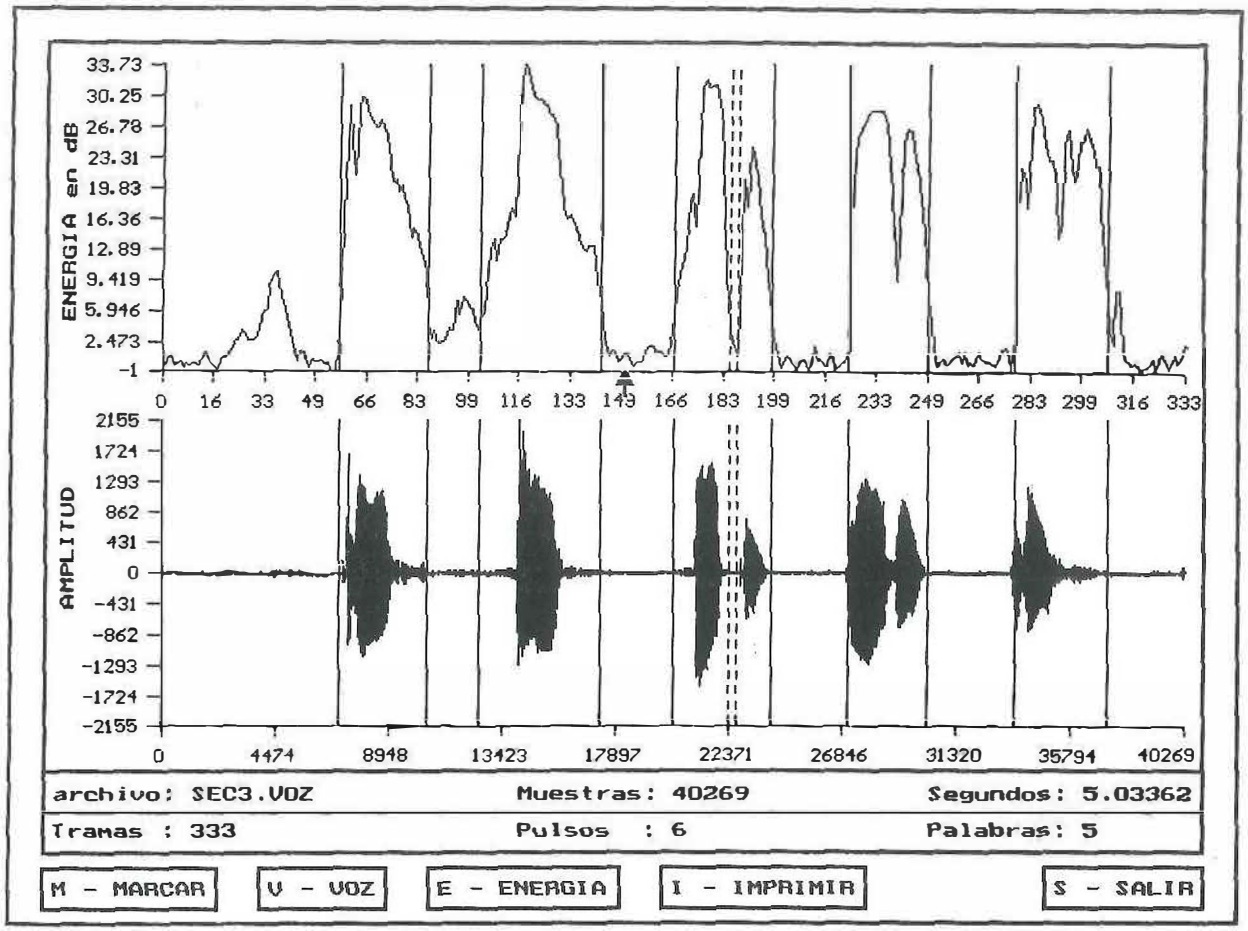


Figura 37. Despliegue gráfico de la *detección de inicio y fin* para la secuencia de palabras /tres seis siete nueve dos/.

Otro buen ejemplo de la aplicación del algoritmo se puede obtener de la figura 37. Ahi se muestran los resultados de la aplicación del algoritmo a la secuencia de palabras aisladas (separadas) /tres seis siete nueve dos/. Puede verse que debido al oclusivo intermedio /t/ de la palabra /siete/, dos pulsos de energía la conforman; ya que la energía detectada durante la oclusión es muy baja. Sin embargo, el algoritmo asoció estos pulsos como parte de una misma palabra debido a su cercanía.

V.3.2.2 Extracción de parámetros.

El siguiente bloque en el proceso de *identificación/verificación de locutor* de las figuras 30 y 31 es la extracción de parámetros en los segmentos de voz detectados (palabras o pulsos). Como ya se mencionó anteriormente los parámetros seleccionados son los coeficientes *cepstrum* ya que han mostrado una eficiencia mayor en la tarea de RAL, además de que existen métodos bastante conocidos para su cómputo.

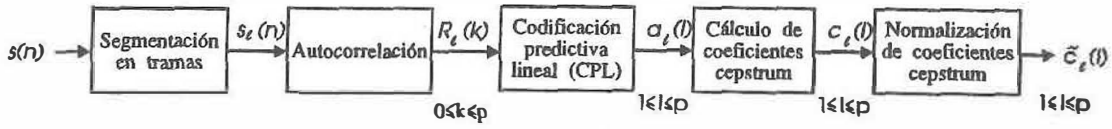


Figura 38. Diagrama a bloques del procedimiento de extracción de parámetros.

Con el fin de explicar la obtención de los coeficientes *cepstrum* considérese el esquema de la figura 38. La secuencia *s(n)* debe ser considerada como la salida del procedimiento de *detección de inicio y fin*, tal que la primera muestra de *s(n)* corresponde a la primera muestra de la trama cuyo índice fue detectado como *inicio* de la primera palabra en la secuencia original adquirida; y la última muestra de *s(n)* corresponde a la última muestra de la trama cuyo índice fue detectado como *fin* de la última palabra en la secuencia adquirida. Esto es, *s(n)* comienza donde se ha detectado el inicio del habla en la secuencia procesada por el *detector de inicio y fin*, y termina donde se ha detectado el fin del habla en la misma secuencia.

De esta manera, *s(n)* puede contener varias palabras mezcladas con posibles silencios entre palabras, o bien únicamente una palabra. En experimentos de RAL con la técnica de ADT [Velius, 1988] se ha encontrado que la eficiencia al utilizar frases (secuencias de palabras) no es claramente mejor que al utilizar únicamente una palabra.

En el primer bloque de la figura 38 la **señal de voz** *s(n)* es preenfatizada usando el filtro digital de la ecuación (92), y luego segmentada en tramas de longitud N muestras con

un recorrimiento entre tramas subsecuentes de L muestras. Cada trama es ponderada por la ventana Hamming de la ecuación (93). La ℓ-ésima trama ponderada de ***s(n)*** se denota como $s_\ell(n)$ definida para $0 \le n \le N-1$. En cada trama de voz se realiza un análisis de predicción lineal de orden p usando el método de la autocorrelación y el *algoritmo de Levinson-Durbin* descritos en la sección III.2.2. Los coeficientes de predicción lineal son transformados en coeficientes *cepstrum* del mismo orden usando la ecuación recursiva (64) de la sección III.3.

En el último bloque de la figura 38 los coeficientes *cepstrum* son normalizados restando el promedio de cada uno de ellos, el cual es obtenido sobre toda la secuencia ***s(n)***. Esto es, si la **secuencia de voz *s(n)*** está formada de NT tramas, y p es el orden de predicción; los coeficientes *cepstrum* normalizados son obtenidos como

$$\tilde{c}_\ell(i) = c_\ell(i) - \frac{\sum_{k=0}^{NT-1} c_k(i)}{NT} \qquad 1 \le i \le p\,,\ 0 \le \ell \le NT \qquad (99)$$

Esta normalización ha probado ser conveniente para reducir la variabilidad intralocutor en secuencias de larga duración [Furui, 1981].

A pesar de que los coeficientes *cepstrum* son obtenidos sobre tramas de tiempo, en algunas referencias éstos son considerados como *información instantánea*, y se ha tratado de mejorar la eficiencia de los sistemas de RAL incorporando *información transicional.*

En las técnicas de cuantificación vectorial (CV) esta información ha sido incorporada agregando un *libro de códigos* con este tipo de información, para cada locutor [Soong y Rosenberg, 1986], mientras que en las técnicas de ADT esta información se agrega trama a trama [Furui, 1981].

En ambos casos la *información transicional* se obtiene haciendo una aproximación polinomial de cada uno de los p coeficientes *cepstrum* sobre un número predeterminado de T tramas contiguas; ésto es, para la ℓ-ésima trama representada por los p coeficientes *cepstrum* se hace la aproximación desde la trama $\ell - K$ hasta la trama $\ell + K$; de aquí T=2K+1.

La aproximacón polinomial de los coeficientes *cepstrum* puede hacerse utilizando un conjunto de polinomios ortogonales, lo cual facilitaría el cálculo de los coeficientes de cada uno de ellos, ya que los valores disponibles de la *ordenada* (v.g., los coeficientes *cepstrum*) son obtenidos a intervalos espaciados de la *abscisa* (v.g., trama a trama).

Un conjunto de polinomios ortogonales que pueden ser utilizados es [Johnson y Leone, 1964]:

$$P_0(z) = 1 \tag{100}$$

$$P_1(z) = z - \frac{T+1}{2} \tag{101}$$

$$P_2(z) = z^2 - (T+1)z + \frac{T^2+3T+2}{6} \tag{102}$$

donde la aproximación polinomial está dada por:

$$P(z) = a\,P_0(z) + b\,P_1(z) + c\,P_2(z) + \ldots \tag{103}$$

Los coeficientes *a, b, y c* pueden ser obtenidos de las ecuaciones:

$$a = \frac{\sum_{z=1}^{T} x_z}{T} \tag{104}$$

$$b = \frac{\sum_{z=1}^{T} x_z\,P_1(z)}{\sum_{z=1}^{T} P_1^2(z)} \tag{105}$$

$$c = \frac{\sum_{z=1}^{T} x_z\,P_2(z)}{\sum_{z=1}^{T} P_2^2(z)} \tag{106}$$

Los coeficientes *a, b, y c* representan valor promedio, pendiente y curvatura respectivamente. En experimentos realizados por Furui [Furui, 1981] se ha demostrado que

el coeficiente b (v.g., la pendiente) es más util para RAL, ya que presenta una mayor tasa *distancia interlocutor/distancia intralocutor* [ver siguiente capítulo] que los coeficientes a y c. A partir de esto, en este trabajo se utiliza únicamente el coeficiente b como información transicional.

Dado que $T=2K+1$ es el número total de tramas sobre el cual se realiza la aproximación polinomial, para cualquier K se tiene un T impar; por lo tanto la ecuación (105) puede ser transformada en:

$$\Delta\tilde{c}_\ell(i) = b = \frac{\sum_{k=-K}^{K} k\tilde{c}_{\ell+k}(i)}{\sum_{k=-K}^{K} k^2} \qquad para\ 1 \le i \le p,\ \ 0 \le \ell \le NT-1 \qquad (107)$$

En síntesis, la ecuación (107) permite obtener la pendiente para cada cada coeficiente *cepstrum*, de cada trama de voz, sobre un intervalo de T tramas contiguas. Debe notarse que pueden ocurrir problemas al no disponer de información suficiente para los índices $\ell=0$ y $\ell=NT-1$, a menos que se consideren K tramas anteriores al inicio de la(s) palabra(s) detectadas y K tramas posteriores al fin detectado; esta solución fue utilizada en la implantación del algoritmo.

Por otra parte, cabe mencionar que el agregar *información transicional* aumenta la cantidad de información utilizada para caracterizar al locutor; en el caso de la técnica de CV se requieren dos *libros de códigos* para cada locutor, mientras que en SERAL, con la técnica de ADT, cada trama está representada por $2p$ parámetros. Esto es, los p coeficientes *cepstrum* y las p pendientes correspondientes.En el apéndice E puede encontrarse el diagrama de flujo interactivo computadora-DSP32 para la extracción de parámetros.

La figura 39 muestra el contorno de parámetros de la palabra /fases/ de la figura 32; en este caso el contorno se ha obtenido de toda la señal adquirida, no solamente de los segmentos de voz o de la palabra detectada. Al pié de las gráficas puede verse información

relativa a la extracción de parámetros tal como: duración de las tramas (45 ms), recorrimiento entre tramas subsecuentes (15 ms), número de tramas (95), orden de predicción (15), y número de tramas para *información transicional* (T=5 o K=2). Todos los parámetros anteriores pueden ser establecidos en SERAL.

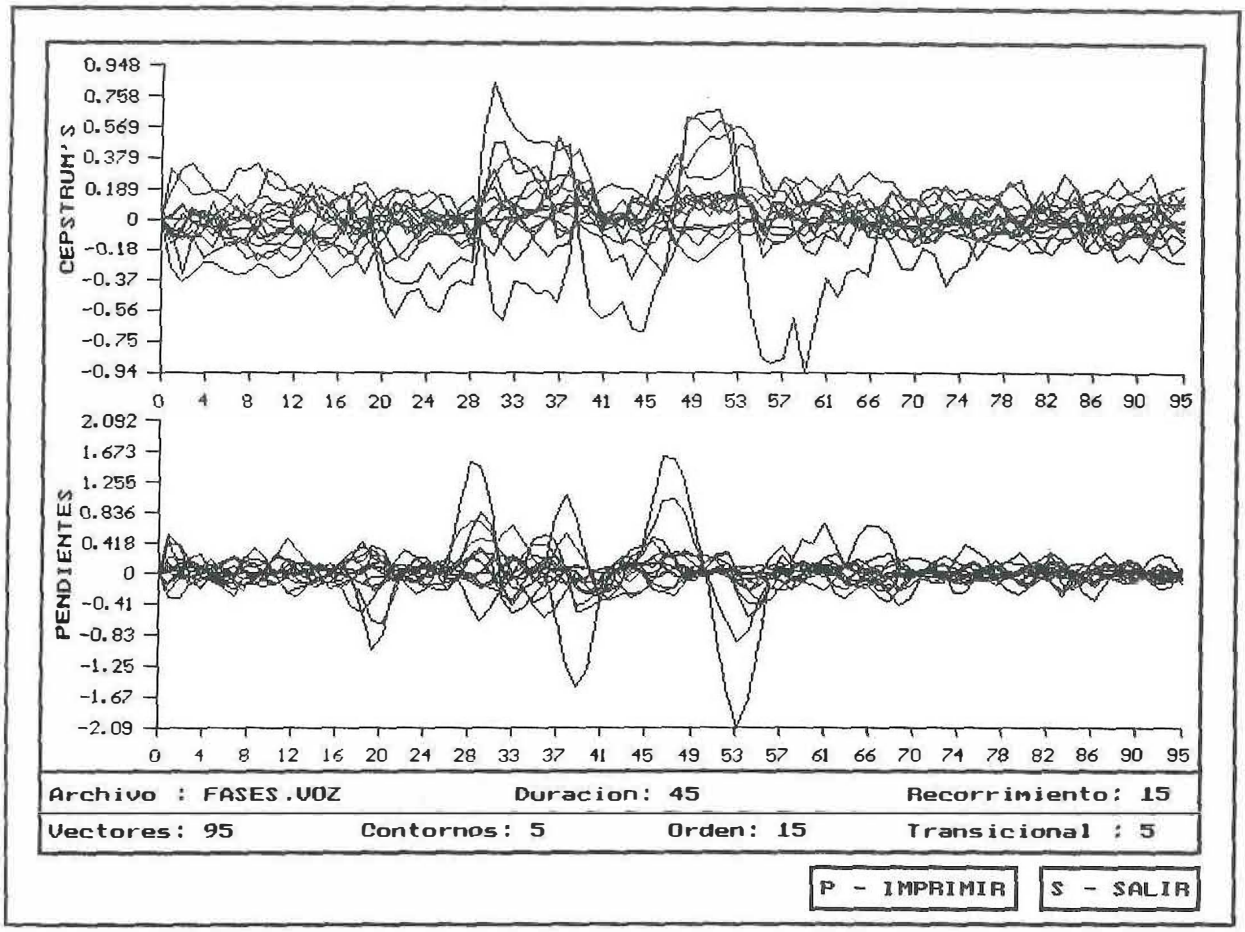


Figura 39. Despliegue gráfico de SERAL mostrando el *contorno de coeficientes cepstrum y el contorno de pendientes* para la palabra /fases/ de la figura 32.

V.3.2.3 Alineamiento, generación y comparación de patrones de locutor.

El alineamiento, generación y comparación de patrones de locutor se realiza bajo la técnica de alineamiento dinámico en el tiempo (ADT) expuesta en el capítulo anterior. Si bien debe aclararse que no se realiza comparación entre patrones de locutor propiamente,

sino entre muestras de voz provenientes de un locutor desconocido que son transformadas en un contorno de parámetros (v.g., *cepstrum* y pendientes), y uno o todos los patrones de locutor almacenados en la base de datos.

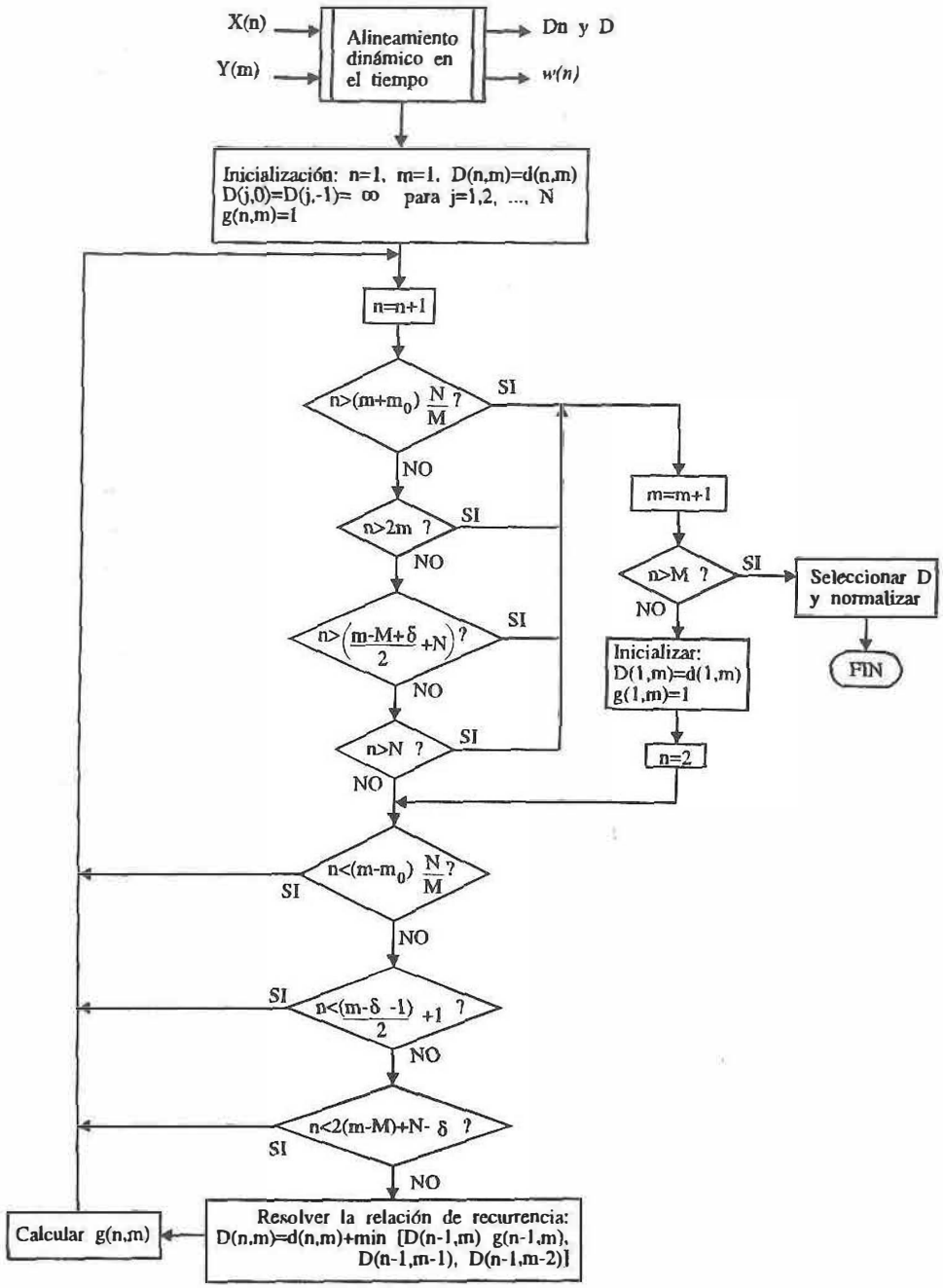


Figura 40 Diagrama de flujo del procedimiento para el ADT de dos *contornos*.

La generación de patrones de locutor se realiza a partir de una o varias muestras del mensaje de entrenamiento (y prueba) del sistema. En particular, se sigue un método aplicado por Furui [Furui, 1981] en cual, si se dispone de varias muestras de entrenamiento, una de ellas es elegida como referencia inicial; esta primera muestra es alineada bajo el algoritmo de ADT con otra de las muestras de entrenamiento y ambas muestras (transformadas en *contornos* de parámetros) son promediadas bajo la función ruta de alineamiento *w(n)*. El procedimiento es repetido iterativamente hasta que se agotan todas las muestras de entrenamiento disponibles para cada locutor.

La implantación del algoritmo de ADT se realizó utilizando el diagrama de flujo de la figura 40. El algoritmo recibe como entradas dos *contornos* de parámetros y entrega como salida la función ruta de alineamiento *w(n)*, la distancia acumulada, y la distancia normalizada entre ambos ambos contornos al ser alineados bajo la función *w(n)*. La medida de distorsión espectral utilizada en la técnica de ADT fue una distancia euclidiana pesada al cuadrado de la forma

$$d(n,m) = d(x_n, y_m) = \sum_{i=1}^{v} g_i^{\,2}[x_n(i) - y_m(i)]^2 \qquad (108)$$

donde x_n *e* y_m son los n-ésimo y m-ésimo vectores de los contornos guía y esclavo respectivamente, v es el número de parámetros por trama (v.g., *2p o p* según se utilice o no *información transicional*), y g_i es el peso del i-ésimo parámetro.

El procedimiento de ADT consiste en establecer las condiciones iniciales para el alineamiento (v.g., condiciones de frontera), para luego hacer un *barrido* de la zona permitida de alineamiento calculando la distancia mínima acumulada desde cualquiera de los puntos permitidos como inicio hasta el punto (n,m). El procedimiento se termina cuando se ha calculado *D(N,M)*, y la distancia mínima acumulada puede ser seleccionada de cualquiera de los puntos permitidos como fin de la función de alineamiento *w(n)*. Esta distancia mínima acumulada es normalizada por el factor (N / N_s) para compensar el hecho de que la función

ruta *w(n)* no alcance el final del contorno guía. Para información relacionada con la implantación interactiva (v.g., computadora-DSP32) del algoritmo se puede revisar el apéndice E.

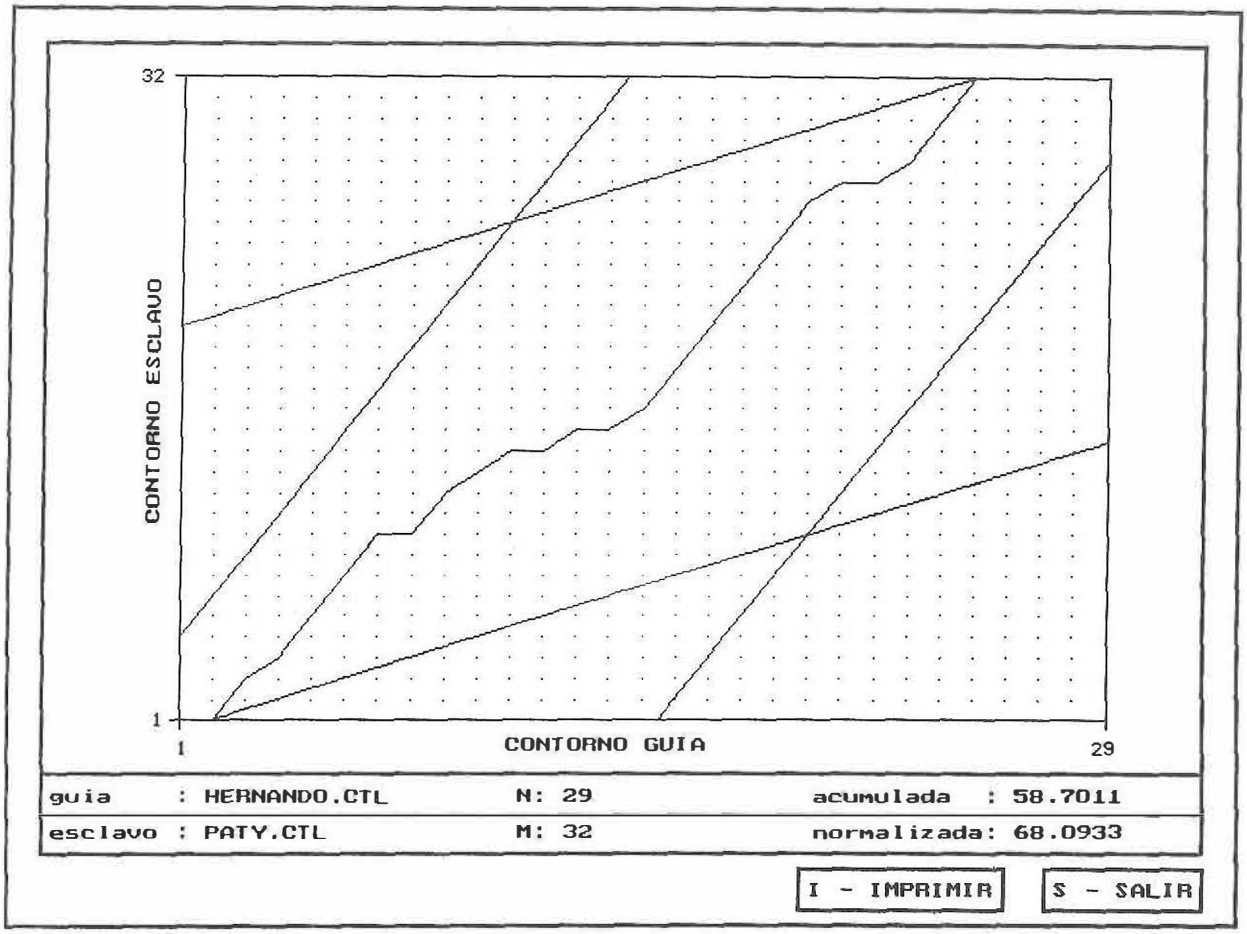


Figura 41. Despliegue gráfico de SERAL mostrando el alineamiento y la comparación de los patrones de dos locutores al utilizar la palabra /NÚMERO/.

La distancia normalizada es el único parámetro utilizado para hacer la decisión, además del umbral. La función ruta de alineamiento *w(n)* se utiliza en algunas herramientas gráficas de SERAL tal como la mostrada en la figura 41, donde se puede observar la función ruta de alineamiento *w(n)* obtenida al alinear y comparar los patrones generados para dos lócutores que utilizaron la palabra /NÚMERO/. Al pié de la figura se puede observar el

nombre de cada locutor, el número de vectores de información de cada uno de ellos y ambas distancias acumulada y normalizada.

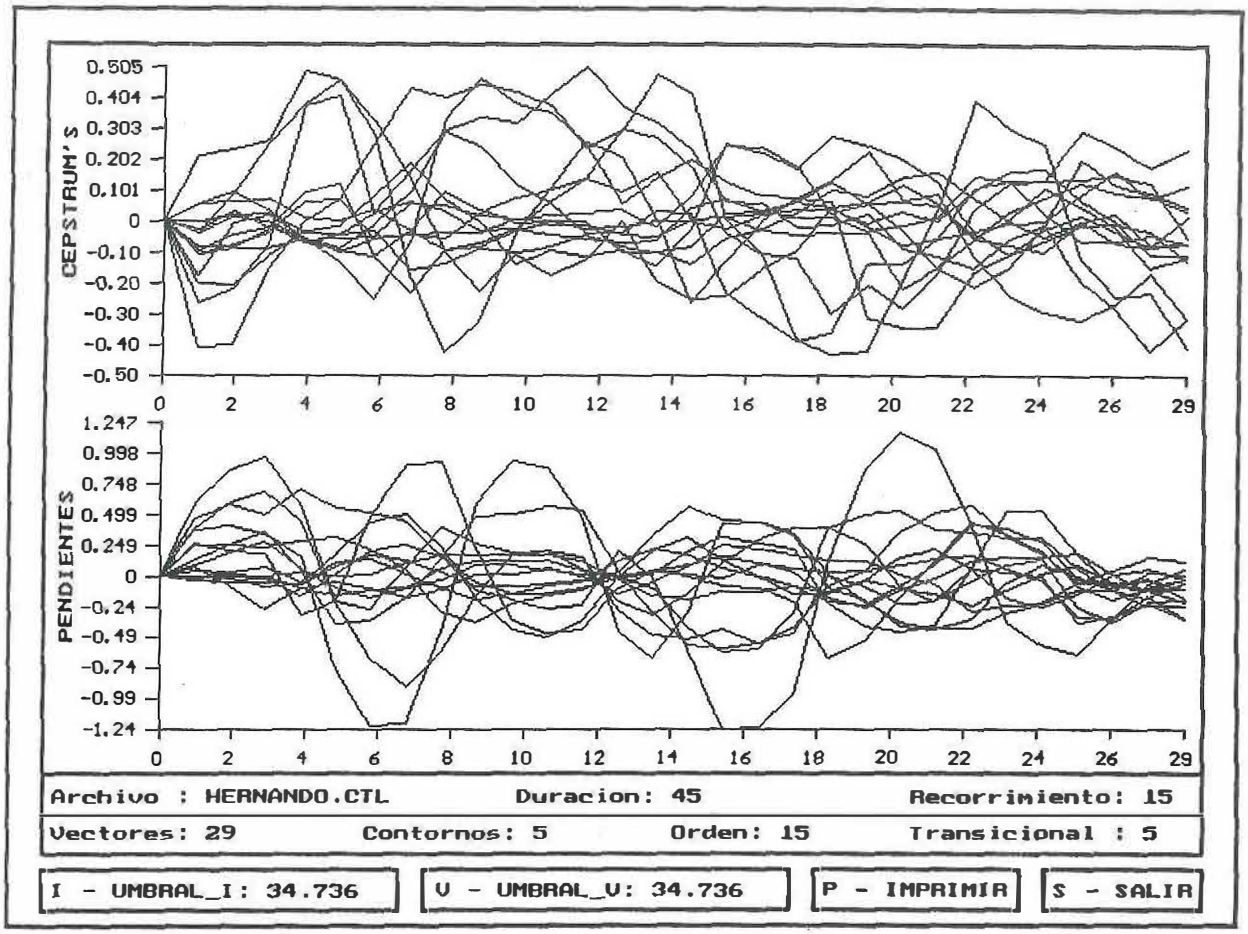


Figura 42. Despliegue gráfico de SERAL mostrando la información contenida en un patrón típico de locutor.

La figura 42 muestra la información almacenada en un patrón típico de locutor de SERAL. Éste se encuentra almacenado como un archivo bajo el nombre del locutor en el disco duro, y contiene información relacionada con los parámetros bajo los cuales fue generado el patrón del locutor. Entre esos parámetros se encuentra el número de muestras utilizadas para generar el patrón, y los umbrales individuales asignados a cada locutor en particular, tanto para *verificación* como para *identificación.* El tema de la estimación de los

umbrales será abordado en el siguiente capítulo.

Es importante establecer que si bien los patrones de locutor se encuentran almacenados en el disco duro, éstos son cargados de inmediato en la memoria de la computadora al ejecutar el programa administrador SERAL.EXE; de esta manera, los patrones se encuentran disponibles para las tareas de *identificación/verificación*, lo cual permite acelerar el proceso de emitir la decisión.

V.3.3 Archivos auxiliares y de configuración.

Para el despliegue gráfico de la información en SERAL, es necesario disponer de varios archivos auxiliares que contienen los tipos de letra utilizados en el desarrollo de las interfases gráficas, además del manejador de modo gráfico. Estos archivos son:

a) EGAVGA.BGI

b) LITT.CHR

c) SANS.CHR

La configuración de SERAL está contenida en un archivo de formato ASCII llamado PARAM.DAT, el cual puede ser modificado con un editor de texto que trabaje ése formato, o bien desde el mismo programa principal SERAL. La información contenida en el archivo y la forma en que está ordenada es la siguiente:

a) Dirección base de TAPS en hexadecimal (v.g., 0x300).

b) Número de interrupción utilizada por TAPS (v.g., 3 o 5).

c) Longitud de las tramas de tiempo en milisegundos.

d) Recorrimiento de las tramas de tiempo en milisegundos.

e) Orden de predicción.

f) Parámetro K para el cálculo de información transicional.

g) Parámetro δ del algoritmo de ADT.

h) Parámetro m_0 del algoritmo de ADT.

i) Número de contornos necesarios para crear un patrón de referencia.

j) Número de locutores ("clientes") en en el sistema.

k) Umbral general de *identificación.*

l) Umbral general de *verificación.*

m) Bandera para conmutar entre umbrales generales o individuales.

n) Bandera para actualizar los patrones de locutor si se da una decisión positiva.

ñ) Bandera para conmutar entre impresión en impresora láser o de puntos.

o) Bandera para seleccionar entre impresión de media hoja u hoja completa.

Además de archivo de configuración descrito anteriormente se requiere de un archivo en formato ASCII que contenga los pesos asignados para cada parámetro. El nombre del archivo es PESOS.DAT, y el número de datos que debe contener es igual al orden de predicción si no se usa *información transicional*, o dos veces el orden de predicción en el caso de que ésta se use.

CAPÍTULO VI EXPERIMENTOS Y RESULTADOS.

VI.1 Base de datos.

Para llevar a cabo los experimentos se adquirió una base de datos de 10 locutores, 5 hombres y 5 mujeres; todos adultos. Cada locutor proporcionó 15 muestras aisladas de las palabras "HOLA" y "NÚMERO". Las muestras fueron adquiridas en dos sesiones con una semana de diferencia, en la primera sesión se tomaron alrededor de 2 ó 3 muestras por locutor para pruebas preliminares, y en la segunda sesión las muestras restantes. El tiempo promedio entre muestras sucesivas adquiridas de un mismo locutor promedió 10 segundos.

La adquisición de la base de datos se llevó a cabo en un laboratorio de computadoras bajo las siguientes condiciones:

a) Sólo se encontraba en funcionamiento la computadora de SERAL.

b) La computadora utilizada fue colocada bajo la mesa de trabajo y cubierta con un material con propiedades de atenuación de ondas acústicas. Lo anterior se hizo con la intención de atenuar el ruido acústico causado por la computadora, así como también evitar que la vibración mecánica se trasmitiera al micrófono.

c) Sólo se permitió la presencia un locutor a la vez en la sala de grabación.

d) Luego de la adquisición de las muestras de cada locutor, el nivel pico de energía fue verificado para permitir que pudieran ser procesadas por el *detector de inicio y fin*; en el caso de que una o varias de las muestras no tuvieran el nivel mínimo de energía requerido se le pidió al locutor la repetición de éstas.

e) El laboratorio donde se realizaron las grabaciones no tenía protección especial contra ruido diseñada *exprofeso* para este tipo de experimentos.

f) Ninguno de los locutores fue compelido a mantener cierta cadencia o

entonación en las muestras; si bien se le indicó cual debería ser el volumen adecuado y la distancia de la boca al micrófono más conveniente. Esto con el único fin de evitar un sobreflujo en el convertidor A/D o una energía pico muy baja.

La base de datos adquirida fue separada en dos grupos. El primer grupo se formó de las primeras cinco muestras (de cada palabra) adquiridas de cada locutor, este grupo fue seleccionado para el entrenamiento (v.g., generación de patrones de referencia) del sistema; el segundo grupo se formó con las diez muestras restantes y fue utilizado para pruebas y caracterización.

VI.2 Estimación de umbrales.

La intención de los experimentos fue evaluar la tasa de error del SERAL bajo un conjunto de "clientes" dado, y bajo diferentes parámetros de operación. La tasa de error fue evaluada para las modalidades de:

a) *Identificación automática de locutor (IAL) de umbral abierto.*

b) *Identificación automática de locutor (IAL) de umbral cerrado.*

c) *Verificación automática de locutor (VAL).*

En las modalidades b) y c) es necesaria la estimación de un umbral de decisión; ya que SERAL puede trabajar tanto con umbrales generales (v.g., el mismo umbral para todos los locutores) como con umbrales individuales (v.g., un umbral para cada locutor), estas dos clases de umbrales fueron estimados. Los umbrales estimados fueron establecidos por igual para VAL como para IAL.

Las clases de errores que puede producir cada una de las modalidades de RAL fueron mencionadas en el capítulo II, para propósitos de notación considérense las siguientes clases de errores que pueden producirse en las modalidades a), b), y c) para ambos tipos de

umbrales individual y general.

a) Falsas identificaciones de umbral abierto(FIUA).

b) Falsas identificaciones de umbral individual (FIUI).

Falsas no-identificaciones de umbral individual (FNUI).

Falsas identificaciones de umbral general (FIUG).

Falsas no-identificaciones de umbral general (FNUG).

c) Falsas aceptaciones de umbral individual (FAUI).

Falsos rechazos de umbral individual (FRUI).

Falsas aceptaciones de umbral general (FAUG).

Falsos rechazos para umbral general (FRUG).

La estimación de los umbrales se realizó sobre la suposición de que tanto el conjunto de *distancias intralocutor*, como el de *distancias interlocutor*, podían ser aproximados por distribuciones paramétricas de probabilidad; tales como las distribuciones normales mostradas en la figura 43.

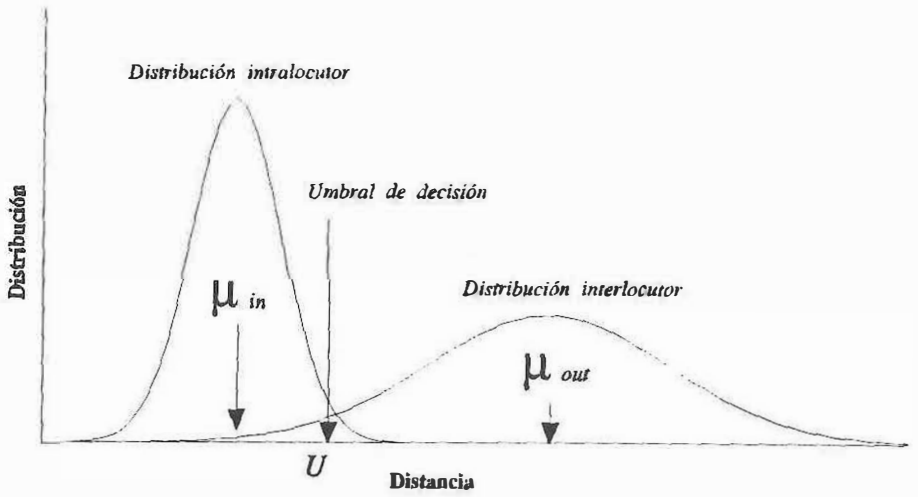


Figura 43. Distribuciones típicas de *distancias intralocutor e interlocutor*.

El umbral de decisión fue seleccionado *a posteriori*, como áquel que igualaba la

probabilidad de error; produciendo de aquí una tasa de equierror (TEE). En la figura 43, lo anterior equivale a seleccionar como umbral, el punto U en el eje de la distancia tal que el área bajo la curva de la *distribución interlocutor* desde menos infinito a U es igual al área bajo la curva de la *distribución intralocutor* desde U hasta infinito. Los procedimientos para la obtención de umbrales *a priori* no se encuentran bien definidos, y si bien ha habido algunas propuestas para su cómputo [Furui, 1981], la mayoría de los experimentos de VAL han calculado el umbral *a posteriori* (v.g., [Soong *et al*, 1987] [Tishby, 1991]).

Las *distancias intralocutor* son obtenidas a partir de la medición de las distancias entre los patrones de locutor generados y las muestras pertenecientes (o habladas) por el locutor "dueño" de cada patrón, esto es, sólo se comparan muestras y patrones de un mismo locutor; mientras que las *distancias interlocutor* son calculadas como las distancias entre los patrones de locutor y las muestras *no habladas* por el dueño de cada patrón, es decir, sólo se comparan muestras y patrones que pertenecen a un locutor diferente.

Las distribuciones de probabilidad de la figura 43 siguen la ecuación [Spiegel, 1970]:

$$P(D) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(D-\mu)^2}{2\sigma^2}} \tag{109}$$

donde μ es la media y σ es la desviación estándar. Para seleccionar el umbral U que produce la TEE, éste debe ser elegido un número igual de desviaciones estándar σ_{in} *y* σ_{out} alejado de cada media μ_{in} *y* μ_{out}, esto es, de la figura 43

$$U - \mu_{in} = n\sigma_{in} \tag{110}$$

$$\mu_{out} - U = n\sigma_{out} \tag{111}$$

Despejando e igualando n de las ecuaciones (110) y (111) se puede despejar el umbral U dando como resultado:

$$U = \frac{\sigma_{in}\mu_{out} + \sigma_{out}\mu_{in}}{\sigma_{in} + \sigma_{out}} \tag{112}$$

y la ecuación (112) puede ser utilizada para calcular el umbral general si se definen:

μ_{in} y σ_{in}=Media y desviación estándar de las distancias intralocutor para toda la población de locutores (100 mediciones posibles).

μ_{out} y σ_{out}=Media y desviación estándar de las distancias interlocutor para toda la población de locutores (900 mediciones posibles).

Para el caso de los 10 umbrales individuales, éstos se obtienen utilizando:

$$U_i = \frac{\mu_i^{in}\sigma_i^{out} + \mu_i^{out}\sigma_i^{in}}{\sigma_i^{in} + \sigma_i^{out}} \qquad i = 1,2,\ldots,10 \qquad (113)$$

donde se define:

μ_i^{in} y σ_i^{in}=Media y desviación estándar de las distancias intralocutor para el i-ésimo locutor (10 mediciones posibles).

μ_i^{out} y σ_i^{out}=Media y desviación estándar de las distancias interlocutor para el i-ésimo locutor (90 mediciones posibles).

La TEE puede ser obtenida para todos los umbrales (individuales y general) acudiendo a una tabla de áreas bajo la curva normal estandarizada (v.g., media cero y desviación estándar unitaria) [Lipschutz, 1971]. La TEE puede considerarse como un parámetro suficiente para caracterizar al sistema de RAL en los experimentos; sin embargo, con la intención de verificar *grosso modo* la bondad de ajuste bajo la distribución normal, así como de contabilizar los errores, en sus diferentes modalidades, se introdujo el esquema experimental de cómputo de la figura 44.

En este esquema las variables i, j, k representan:

i=Índice del patrón de referencia del i-ésimo locutor tal que $1 \leq i \leq 10$.

j=Índice del j-ésimo locutor tal que $1 \leq j \leq 10$.

k=Índice de la k-ésima muestra del j-ésimo locutor tal que $1 \leq k \leq 10$.

De esta manera, $D(i,j,k)$ denota la distancia normalizada (obtenida por ADT) entre el

i-ésimo patrón (v.g., modelo del *i*-ésimo locutor) y la *k*-ésima muestra del *j*-ésimo locutor.

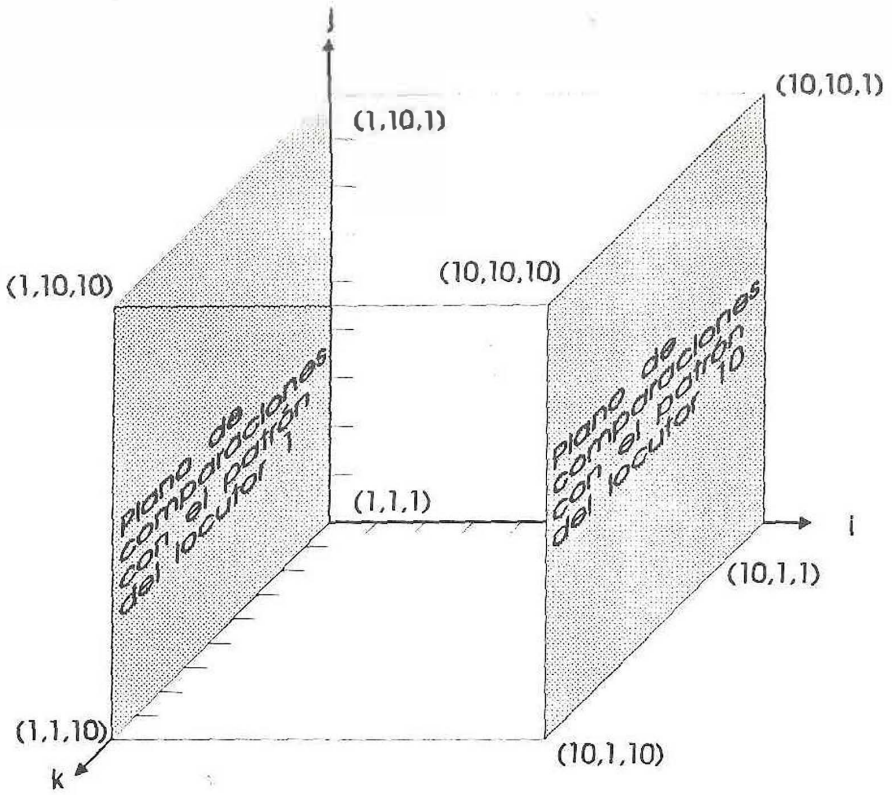


Figura 44. Esquema experimental de cómputo para estimación de las TEE.

De acuerdo con el esquema de la figura 44 y la definición de los índices, la cantidad de errores en cada modalidad de RAL y con ambos tipos de umbrales puede ser definida como:

$$FIUA = \#\{D(i,i,k) \mid D(i,i,k) > \min_{1 \le j \le 10,\ j \ne i} [D(j,i,k)]\} \quad \forall i,k \tag{114}$$

$$FIUI = \#\left\{ D(i,i,k) \left| \left(D(i,i,k) > \left(\ell = \min_{1 \le j \le 10,\ j \ne i} [D(j,i,k)] \right) \right) y \left(\ell < U_m \left| D(m,i,k) = \ell \right. \right) \right. \right\} \quad \forall i,k \tag{115}$$

$$FNUI = \#\left\{ D(i,i,k) \left| \left(\ell = \min_{1 \le j \le 10,\ j \ne i} [D(j,i,k)] \right) y \left(\ell \ge U_m \left| D(m,i,k) = \ell \right. \right) \right. \right\} \quad \forall i,k \tag{116}$$

$$FIUG = \#\left\{ D(i,i,k) \middle| \left(D(i,i,k) > \left(\ell = \min_{1 \le j \le 10,\ j \ne i} [D(j,i,k)] \right) \right) y \left(\ell < U \right) \right\} \quad \forall i,k \qquad (117)$$

$$FNUG = \#\left\{ D(i,i,k) \middle| \left(\ell = \min_{1 \le j \le 10,\ j \ne i} [D(j,i,k)] \right) y \left(\ell \ge U \right) \right\} \quad \forall i,k \qquad (118)$$

$$FAUI = \#\left\{ D(i,j,k) \middle| D(i,j,k) < U_i \right\} \quad \vee \quad 1 \le i,j,k \le 10,\ j \ne i \qquad (119)$$

$$FRUI = \#\left\{ D(i,i,k) \middle| D(i,i,k) \ge U_i \right\} \quad \forall\ i,k \qquad (120)$$

$$FAUG = \#\left\{ D(i,j,k) \middle| D(i,j,k) < U \right\} \quad \vee \quad 1 \le i,j,k \le 10,\ j \ne i \qquad (121)$$

$$FRUG = \#\left\{ D(i,i,k) \middle| D(i,i,k) \ge U \right\} \quad \forall\ i,k \qquad (122)$$

donde # denota *cardinalidad.*

Debe notarse que para todos los tipos de errores de *identificación,* de las ecuaciones (114)-(118), existe una ocurrencia máxima de 100 (v.g., igual al número de muestras de prueba); además de que no se están utilizando "impostores" para los experimentos. De aquí que las *falsas identificaciones* se contabilicen como el número de ocasiones en las cuales el sistema responde con el nombre del *locutor equivocado*, mientras que las *falsas no-identificaciones* se contabilizan simplemente como el número de ocasiones en que el sistema responde *No-identificado.* En el caso de los errores de *verificación* hay una ocurrencia máxima de 100 para los *falsos rechazos*, y una ocurrencia máxima de 900 para las *falsas aceptaciones*.

VL3 Experimentos y resultados.

Aprovechando la característica de configurabilidad de SERAL se llevaron a cabo una serie de experimentos cambiando el valor de algunos de los parámetros y obteniendo las tasas de error bajo cada configuración particular.

El sistema fue establecido en una configuración inicial, y sólo se varió un parámetro a la vez sobre intervalos encontrados en la literatura [Soong *et al*, 1987] [Furui, 1981]. Se seleccionó un valor como óptimo a partir de un compromiso entre TEE, memoria y tiempo

de cómputo, y se procedió a variar el siguiente parámetro.Ésta metodología tiene la obvia desventaja de despreciar una posible correlación entre los valores de los parámetros, la cual podría producir mejores resultados; sin embargo, analizar esa posible correlación tomaría demasiado tiempo, además de que la intención de los experimentos es monitorear la tasa de error sobre condiciones más bien generales.

La tabla II muestra los parámetros que se variaron en los experimentos, así como el orden en el cual se llevó a cabo esa variación. La primera columna muestra los valores de los parámetros en la configuración inicial del sistema, la segunda columna el parámetro variado, y en la última columna se observan los valores dados a cada uno de ellos.

Durante todos los experimentos el valor del parámetro m_0 fue establecido en 100 tramas, con lo cual no tenía ningún efecto sobre el algoritmo de ADT. Los valores de los pesos en la distancia euclidiana pesada al cuadrado fueron todos establecidos a un valor unitario, lo cual equivale a utilizar una distancia euclidiana al cuadrado.

Tabla II. Parámetros variados en los experimentos y sus valores.

Valor inicial	Parámetro	Valores experimentados
10	Orden de predicción	5, 10, 15, 20 y 25.
45 ms	Longitud de trama	25, 35, 45, 55 y 65 ms.
15 ms	Recorrimiento de trama	5, 10, 15 y 20 ms.
0	Tramas para información transicional	3, 5, 7, 9 y 11.
0	Parámetro δ del algoritmo de ADT	2, 4, 6 y 8.
3	Muestras para entrenamiento	1, 2, 3, 4 y 5.
"HOLA"	Vocabulario	"HOLA" y "NÚMERO"

Los valores de la configuración inicial del sistema fueron tomados de algunos experimentos reportados en la literatura. A partir de la tabla II puede notarse lo siguiente: no se está utilizando información transicional, se supone hay un buen alineamiento en los límites (v.g., inicio y fin) de los contornos guía y esclavo, se están tres de las muestras disponibles

para entrenamiento, y se ha tomado como vocablo inicial la palabra "HOLA".

Los experimentos realizados generaron una gran cantidad de información relativa a tasas de error (v.g., tablas III, IV, V, VI, VII y VIII), sin embargo, utilizar la totalidad de esa información complicaría bastante el establecimiento de un criterio de optimalidad del valor del parámetro. A partir de esta consideración se tomaron en cuenta únicamente los valores de la TEE para el umbral general y el promedio de las TEE para los umbrales individuales, los cuales se encuentran en las dos primeras líneas de las tablas de datos.

El resto de los datos en las tablas de resultados contiene información relativa al porcentaje real de errores en cada modalidad de RAL experimentada, y puede utilizarse para verificar la bondad de ajuste de las curvas de la distribución normal de la figura 43. Si se desea observar los resultados cualitativamente más que cuantitativamente, se pueden revisar las gráficas de la figura 45 al final de la siguiente sección, donde se han resumido gráficamente los resultados de las TEE de todos los experimentos.

Tabla III. Resultados variando el orden de predicción.

Tasa de error \ Valor	**5**	**10**	**15**	**20**	**25**
Umbral general	12.1	5.82	5.26	5.59	5.71
Umbral individual	9.63	3.98	3.40	3.02	3.14
%FAUI	7.33	3.56	2.88	2.11	2.44
%FRUI	7.00	4.00	4.00	2.00	3.00
%FAUG	8.78	4.33	3.89	2.89	3.33
%FRUG	15.00	7.00	4.00	7.00	5.00
%FIUA	14.00	6.00	6.00	6.00	4.00
%FIUI	8.00	3.00	4.00	2.00	1.00
%FNUI	10.00	6.00	5.00	5.00	5.00
%FIUG	9.00	6.00	5.00	4.00	2.00
%FNUG	13.00	3.00	3.00	6.00	5.00

La tabla III muestra los resultados de los experimentos al variar el orden de

predicción. Puede observarse que en general, al aumentar el orden de predicción la tasa de error disminuye, si bien hay un incremento muy ligero en las TEE para los órdenes de predicción más altos. Se puede considerar que para obtener buenos resultados se debe utilizar un orden de predicción de al menos 10. El parámetro seleccionado de aquí es 15, ya que presenta la ganancia neta más alta en términos de las TEE, si bien requiere una carga computacional relativamente grande.

Tabla IV. Resultados variando la longitud de trama.

Tasa de error \ Valor	25 ms	35 ms	45 ms	55 ms	65 ms
Umbral general	6.06	5.59	5.26	5.59	39.36
Umbral individual	4.07	3.44	3.40	3.68	32.81
%FAUI	4.66	3.33	2.88	2.78	31.78
%FRUI	4.00	4.00	4.00	3.00	33.00
%FAUG	4.33	3.56	3.89	3.67	42.89
%FRUG	3.00	6.00	4.00	7.00	39.00
%FIUA	3.00	3.00	6.00	5.00	82.00
%FIUI	3.00	1.00	4.00	2.00	39.00
%FNUI	3.00	5.00	5.00	5.00	47.00
%FIUG	2.00	1.00	5.00	3.00	81.00
%FNUG	3.00	6.00	3.00	6.00	1.00

La tabla IV resume los resultados obtenidos al variar la longitud de trama. Las longitudes experimentadas no son realmente cortas y se puede considerar que las TEE se mantienen constantes para los valores más bajos experimentados, si bien para la longitud más grande experimentada se observa un incremento considerable. Los resultados de las tablas III y IV coinciden bastante con los reportados por Velius para un sistema de similares características [Velius, 1988].

Los resultados de los experimentos al variar el recorrimiento entre tramas subsecuentes se muestran en la tabla V. Puede notarse que las tasas de error para los

recorrimientos más bajos son en general menores que para los recorrimientos más altos; ésto se justifica, ya que para recorrimientos más pequeños se tiene una mayor cantidad de información almacenada (v.g., vectores de información). Se puede ver que la ganancia neta más alta se obtiene para un recorrimiento de 10 ms, sin embargo, un recorrimiento de 15 ms no desmerece demasiado y reduce la cantidad de información en un tercio. Valores de longitud de trama de 45 ms y recorrimiento de 15 ms han sido utilizados en una gran cantidad de experimentos de RAL (v.g., [Soong *et al*, 1987][Velius, 1988], todos ligados a Rabiner). El valor elegido para continuar los experimentos es 15 ms.

Tabla V. Resultados variando el recorrimiento de trama.

Tasa de error \ Valor	**5 ms**	**10 ms**	**15 ms**	**20 ms**
Umbral general	5.48	5.16	5.26	9.51
Umbral individual	3.41	3.13	3.40	6.59
%FAUI	3.00	2.67	2.88	5.11
%FRUI	4.00	4.00	4.00	3.00
%FAUG	3.67	3.11	3.89	7.22
%FRUG	8.00	5.00	4.00	4.00
%FIUA	6.00	5.00	6.00	7.00
%FIUI	3.00	2.00	4.00	5.00
%FNUI	5.00	6.00	5.00	5.00
%FIUG	5.00	4.00	5.00	6.00
%FNUG	6.00	4.00	3.00	3.00

En la tabla VI se muestran los resultados al agregar información transicional para diferente número de tramas. No existe un valor que muestre claramente una mejoría con respecto a los demás valores, si bien la ganancia neta más alta es para 5 tramas.

Si se comparan las TEE para 5 tramas de la tabla VI y el caso elegido como óptimo de la tabla V (v.g., recorrimiento igual a 15 ms) puede notarse que al agregar la información transicional la TEE para el umbral general disminuyó en más de un punto porcentual,

recorrimientos más bajos son en general menores que para los recorrimientos más altos; ésto se justifica, ya que para recorrimientos más pequeños se tiene una mayor cantidad de información almacenada (v.g., vectores de información). Se puede ver que la ganancia neta más alta se obtiene para un recorrimiento de 10 ms, sin embargo, un recorrimiento de 15 ms no desmerece demasiado y reduce la cantidad de información en un tercio. Valores de longitud de trama de 45 ms y recorrimiento de 15 ms han sido utilizados en una gran cantidad de experimentos de RAL (v.g., [Soong *et al*, 1987][Velius, 1988], todos ligados a Rabiner). El valor elegido para continuar los experimentos es 15 ms.

Tabla V. Resultados variando el recorrimiento de trama.

Tasa de error \ Valor	5 ms	10 ms	15 ms	20 ms
Umbral general	5.48	5.16	5.26	9.51
Umbral individual	3.41	3.13	3.40	6.59
%FAUI	3.00	2.67	2.88	5.11
%FRUI	4.00	4.00	4.00	3.00
%FAUG	3.67	3.11	3.89	7.22
%FRUG	8.00	5.00	4.00	4.00
%FIUA	6.00	5.00	6.00	7.00
%FIUI	3.00	2.00	4.00	5.00
%FNUI	5.00	6.00	5.00	5.00
%FIUG	5.00	4.00	5.00	6.00
%FNUG	6.00	4.00	3.00	3.00

En la tabla VI se muestran los resultados al agregar información transicional para diferente número de tramas. No existe un valor que muestre claramente una mejoría con respecto a los demás valores, si bien la ganancia neta más alta es para 5 tramas.

Si se comparan las TEE para 5 tramas de la tabla VI y el caso elegido como óptimo de la tabla V (v.g., recorrimiento igual a 15 ms) puede notarse que al agregar la información transicional la TEE para el umbral general disminuyó en más de un punto porcentual,

mientras que la TEE promedio para los umbrales individuales se redujo en un tercio de punto porcentual. El valor seleccionado es 5 tramas para información transicional.

Tabla VI. Resultados al agregar *información transicional*.

Tasa de error \ Valor	3	5	7	9	11
Umbral general	4.65	4.01	4.36	4.85	5.26
Umbral individual	2.95	3.08	3.47	3.48	3.59
%FAUI	3.00	3.67	3.44	3.44	3.22
%FRUI	3.00	4.00	4.00	4.00	5.00
%FAUG	3.44	3.89	4.00	4.11	3.66
%FRUG	5.00	4.00	4.00	6.00	6.00
%FIUA	6.00	3.00	5.00	5.00	5.00
%FIUI	5.00	3.00	3.00	3.00	2.00
%FNUI	3.00	3.00	5.00	5.00	7.00
%FIUG	5.00	3.00	4.00	4.00	4.00
%FNUG	4.00	3.00	4.00	4.00	4.00

Tabla VII. Resultados al variar el parámetro δ del algoritmo de ADT.

Tasa de error \ Valor	2	4	6	8
Umbral general	4.20	4.09	4.36	4.85
Umbral individual	3.11	2.97	3.28	3.80
%FAUI	3.44	3.22	3.56	3.44
%FRUI	4.00	4.00	2.00	4.00
%FAUG	4.11	4.44	4.44	4.33
%FRUG	2.00	2.00	4.00	5.00
%FIUA	7.00	6.00	6.00	6.00
%FIUI	4.00	3.00	3.00	3.00
%FNUI	6.00	7.00	5.00	7.00
%FIUG	7.00	6.00	6.00	6.00
%FNUG	1.00	2.00	4.00	4.00

En la tabla VII se muestran los resultados al variar el valor del parámetro δ del

algoritmo de ADT. En todos lo anteriores experimentos tal parámetro estaba establecido en cero, por lo que se estaban alineando la primera y última trama de los contornos guía y esclavo respectivamente. De los resultados observados se puede establecer que no hay una mejoría significativa, y por el contrario si se permite un δ demasiado grande las TEE aumentan, lo cual puede ser atribuido a una degradación en el proceso de ADT.

Los resultados de la tabla VII sugieren que el algoritmo de *detección de inicio y fin* es bastante eficiente y no requiere aprovechar la propiedad de *no restricción* de los límites de alineamiento del algoritmo de ADT implantado; sin embargo, tal afirmación debe ser verificada con un vocabulario mucho más completo del que se está utilizando aquí. El valor seleccionado de este experimento es δ=2, basándose en la ligera mejoría obtenida.

Tabla VIII. Resultados para varias muestras de entrenamiento.

Tasa de error \ Valor	**1**	**2**	**3**	**4**	**5**
Umbral general	5.48	4.36	4.09	3.51	3.01
Umbral individual	4.49	2.42	2.97	2.39	2.57
%FAUI	4.78	2.11	3.22	2.00	2.67
%FRUI	3.00	3.00	4.00	4.00	3.00
%FAUG	5.33	4.11	4.44	3.11	2.89
%FRUG	5.00	3.00	2.00	6.00	2.00
%FIUA	12.00	9.00	6.00	3.00	1.00
%FIUI	10.00	2.00	3.00	1.00	1.00
%FNUI	5.00	10.00	7.00	5.00	3.00
%FIUG	12.00	9.00	6.00	3.00	1.00
%FNUG	1.00	0.00	2.00	4.00	2.00

La tabla VIII muestra los resultados de utiizar varias muestras para entrenamiento (de entre las cinco seleccionadas inicialmente). El análisis de la TEE muestra una tendencia a decrecer a medida que se utilizan más muestras para generar los patrones de referencia en el caso del umbral general; sin embargo, la TEE promedio para los umbrales individuales

muestra un comportamiento *zig-zag* al aumentar el número de muestras de entrenamiento.

El análisis de los porcentajes reales de error podría ayudar a dilucidar una conclusión con respecto a este parámetro, y de hecho se puede observar que en ninguna de las tablas de resultados anteriores el decremento de las tasas reales de error es tan evidente como lo es en la tabla VIII. Una conclusión definitiva (y esperada) acerca de este experimento es que al aumentar el número de muestras para entrenamiento la tasa de error disminuye.

En el último de los experimentos se cambió el vocabulario de la palabra "HOLA" a la palabra "NÚMERO", para ambos entrenamiento y prueba. Se utilizaron todas las muestras disponibles de entrenamiento (v.g., 5) para generar el patrón de referencia.

Tabla IX. Resultados óptimos para las palabras "HOLA" y "NÚMERO".

Tasa de error \ Palabra	**"HOLA"**	**"NÚMERO"**
Umbral general	3.01	3.59
Umbral individual	2.57	1.51
%FAUI	2.67	0.44
%FRUI	3.00	1.00
%FAUG	2.89	1.67
%FRUG	2.00	6.00
%FIUA	1.00	2.00
%FIUI	1.00	0.00
%FNUI	3.00	2.00
%FIUG	1.00	0.00
%FNUG	2.00	6.00

La tabla IX muestra los resultados óptimos obtenidos para la palabra "HOLA" y la palabra "NÚMERO". Como puede observarse ocurrió un gran decremento en las tasas de error relacionadas con los umbrales individuales, si bien ocurrió un pequeño incremento en las tasas de error relacionadas con el umbral general.

En general puede considerarse que con esta base de datos los mejores resultados se

obtienen al emplear umbrales individuales con la palabra "NÚMERO". La palabra "NÚMERO" puede considerarse una palabra nasal (v.g., con sonidos nasales), y ésta es la razón que se puede aducir para justificar una mayor eficiencia, ya que en experimentos reportados en la literatura se ha encontrado que los sonidos nasales presentan una buena eficiencia en RA:L.

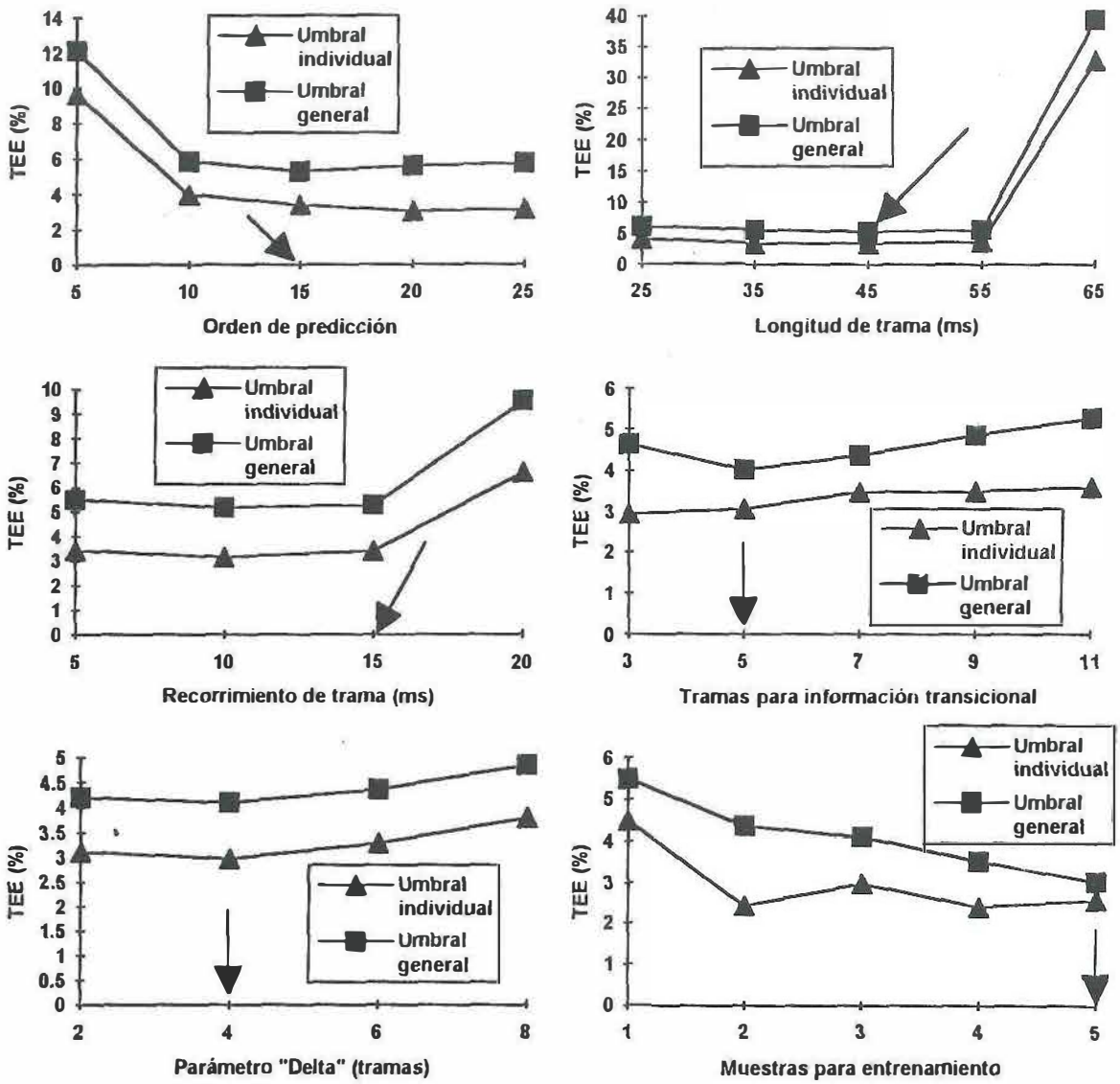


Figura 45. Resultados experimentales resumidos gráficamente para la TEE del umbral general y el promedio de las TEE de los umbrales individuales.

Los resultados generales de los experimentos pueden ser resumidos en la figura 45, donde se han graficado la TEE para el umbral general y el promedio de las TEE para los umbrales individuales. Las flechas indican los valores óptimos elegidos en los experimentos.

CAPÍTULO VII CONCLUSIONES Y RECOMENDACIONES.

CONCLUSIONES

Este trabajo proveé la información necesaria para analizar el problema de RAL en sus diferentes modalidades, así como los parámetros y modelos de la señal de voz que han sido utilizados para enfrentarlo.

Las técnicas de RAL analizadas han sido aplicadas previamente en RAH, de donde se extendió su uso, ésto se debe a la ausencia de técnicas explícitamente enfocadas a RAL; el éxito logrado con RAH ha sido mucho mayor que en RAL.

El sistema experimental desarrollado (SERAL) es *autocontenido* en el sentido de que proveé todas las herramientas necesarías para generar y experimentar con una nueva base de datos (locutores). Es de esperarse que si se utiliza una base de datos más "limpia" (v.g., con una cantidad menor de ruido acústico ambiental) la tasa de error se reducirá; amén de considerar el vocabulario.

Además de lo mencionado anteriormente, la parte programática de bajo nivel de SERAL (v.g., programas para el DSP32) es completamente modular. Esta característica permite la mejora, modificación, y/ó reutilización de los algoritmos implantados. Por ejemplo:

a) El programa GRABA.COD permite la adquisición de cualquier señal bajo las mismas características de configuración establecidas para el TLC32044.

b) Por una parte, el algoritmo de *detección de inicio y fin* puede ser modificado para ajustarse estrictamente al algoritmo *híbrido* planteado por Lamel, y por otra parte, la utilización de SERAL en un ambiente con un mayor o menor nivel de *ruido acústico ambiental* requerirá un reajuste de los parámetros de detección de pulsos de energía.

c) Ya que los coeficientes *cepstrum* son obtenidos como un subproducto de la

predicción lineal, el programa PARAM.COD puede ser modificado fácilmente (v.g., mediante la eliminación de algunas líneas) para utilizar parámetros tales como: coeficientes de predicción lineal, coeficientes de reflexión o coeficientes de autocorrealción

d) El algoritmo de ADT es completamente transparente al tipo de parámetros, así como al número de éstos por trama. De esta manera, no se requiere realizar modificación alguna para utilizar este algoritmo con otro tipo de parámetros, y de hecho, en SERAL existe una opción que permite alinear contornos de energía, la cual utiliza el algoritmo de ADT implantado.

e) Es posible la realización de un sistema de RAH a partir de los cinco programas desarrollados para el DSP32, todo lo que se requiere es diseñar el programa de control y administración para la computadora personal.

En los experimentos realizados con SERAL la tasa de error observada es similar a la reportada en la literatura para este tipo de técnicas, e incluso menor que en algunos experimentos de RAL con técnicas *dependientes del vocabulario* (v.g., [Velius, 1988], [Zheng y Yuan, 1988]. En particular, en todos los experimentos la tasa de error fue menor al utilizar umbrales individuales que al utilizar un umbral general, y por otra parte, la tasa de error fue ligeramente menor para la palabra nasalizada "NÚMERO" que para la palabra vocal "HOLA".

En el trabajo inconcluso de esta tesis se encuentra la realización de experimentos aprovechando la capacidad de SERAL de trabajar con distancias euclidianas pesadas al cuadrado, así como la actualización de patrones de locutor cuando se da una respuesta afirmativa al locutor (v.g., aceptación o identidad). Esto ha sido sugerido como una herramienta de reentrenamiento para considerar la variabilidad temporal de las características del locutor [Furui, 1981] [O'Shaughnessy, 1986].

El reconocimiento automático de locutor (RAL) está lejos de ser un problema resuelto, y como prueba de ello se puede notar que hay una notable ausencia de productos comerciales de RAL comparado con RAH. El conocimiento que se tiene acerca de como el humano explota las características espectrales para reconocer sonidos es mucho mayor que el de como reconoce locutores.

Las técnicas *dependientes del vocabulario* han producido los mejores resultados bajo condiciones favorables, pero la eficiencia se degrada bajo condiciones adversas que pueden encontrarse en aplicaciones prácticas; por ejemplo: distorsiones en líneas telefónicas, locutores no cooperativos y variabilidad temporal.

Es importante establecer que no será posible (al menos a corto plazo) conseguir cualquier nivel de eficiencia deseado en RAL, y esperar que éste sólo esté limitado por los parámetros y las técnicas empleadas. Un papel importante lo juegan el locutor y el control sobre el ambiente de desempeño. Se puede conseguir una alta eficiencia con locutores cooperativos y bajo condiciones controladas, sin embargo, si el locutor no es cooperativo y no se tiene control sobre las condiciones de desempeño, será imposible conseguir cualquier nivel de eficiencia deseado.

RECOMENDACIONES

La explotación de características fonéticas tales como *frecuencias formantes* de ciertos sonidos podría ayudar a mejorar la eficiencia de RAL, sin embargo, deben desarrollarse primero métodos automáticos confiables para la detección de tales sonidos. Se puede preveer que a la par que se desarrollen métodos más confiables para RAH, su aplicación será extendida hacia RAL.

Es de esperarse que las aplicaciones prácticas de RAL en el futuro cercano, apliquen técnicas *dependientes del vocabulario*, y ocurran en VAL; esto debido a que el incremento

del tamaño de la población de locutores dificulta la tarea de IAL, mientras que en VAL no es un factor de peso.

Es necesario también establecer bases de datos de registro (del inglés *benchmark database*) con el fin de evaluar objetivamente la eficiencia de los parámetros y métodos propuestos; de la discusión realizada en este trabajo debe ser obvio que la eficiencia de la tarea de RAL es altamente susceptible a la base de datos empleada y las condiciones en que ésta fue adquirida.

Finalmente, para desarrollar sistemas comerciales es necesario enfocar los esfuerzos hacia el RAL a través de líneas telefónicas, el cual podría ser el medio de comunicación más conveniente desde el punto de vista de la aplicación (v.g., transacciones y acceso a bancos de información por vía telefónica, etc.).

REFERENCIAS

Álvarez Constantino, J., 1977. Gramática funcional del español. Tomo I: Fonología y ortografía. Editorial Avante. México, DF.205 p.

Atal, B. S., 1974. Efectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. Journal of the acoustical society of américa, vol. 55, no. 6, pp. 1304-1312.

Atal B. S., 1976. Automatic recognition of speakers from their voices. Proceedings of the IEEE, vol. 64, no. 4, pp. 460-475.

Atal B. S. y Hanauer S. L., 1971. Speech analysis and synthesis by linear prediction of the speech wave. Journal of the acoustical society of américa, vol. 50, no. 2 (parte 2), pp. 637-655.

Bristow G., 1986. Electronic speech recognition. McGraw-Hill book company, USA, 395 p.

Bristow G., 1984. Electronic speech synthesis . Mc Graw-Hill book company, USA, 346 p.

Buck J., Burton D., y Shore J., 1985. Text dependent speaker recognition using vector quantization. ICASSP, pp. 399-402.

Doddington G. R., 1985. Speaker recognition-Identifying people by their voices. Proceedings of the IEEE, vol. 73, no. 11, pp. 1651-1664.

Dreyfus S. E. y Law A. M., 1977. The art and theory of dynamic programming. vol. 130 en Mathematics in science and engineering. Academic press, Inc., Orlando, Florida. 284 p.

Furui S., 1981. Cepstral analysis technique for automatic speaker verification. IEEE transactions on ASSP, vol. 29, no. 2, pp. 254-272.

Johnson N. L. y Leone F. C., 1964. Statistical and experimental design in engineering and the physical sciences. vol. 1, John Wiley & Sons, Inc., New York, USA, 523 p.

Lamel L. F., Rabiner L. R., Rosenberg A. E. y Wilpon J. G., 1981. An improved endpoint detector for isolated word recognition. IEEE transactions on ASSP, vol 29, no. 4, pp.

777-785.

Levinson S. E., Rabiner L. R. y Sondhi M. M., 1983. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. BSTJ, vol. 62, no. 4, parte 1, pp. 1035-1074.

Lipschutz S., 1971. Probabilidad. McGraw-Hill de México. México, DF, 151 p.

Makhoul J., 1975. Linear prediction: a tutorial review. Proceedings of the IEEE, vol. 63, no. 4, pp. 561-580.

Martínez García-Moreno C. A. y Félix Acuña R., 1993. Sistema mínimo de procesamiento digital de señales. Informe técnico. Comunicaciones académicas, serie electrónica y telecomunicaciones, CICESE. 24 pp. CTETT9302.

O'Shaughnessy D., 1987. Speech communication: human and machine. Adison-Wesley publishing company, USA, 568 p.

O'Shaughnessy D., 1986. Speaker recognition. IEEE ASSP magazine (october), USA, pp. 4-17.

Oppenheim A. V. y Schafer R. W., 1989. Discrete-time signal processing. Prentice Hall, Englewood Cliffs, New Jersey, USA, 879 p.

Rabiner L. R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286.

Rabiner L. R., Rosenberg A. E. y Levinson S. E., 1978. Considerations in dynamic time warping algorithms for discrete word recognition. IEEE transactions on ASSP, vol. 26, no. 6, pp. 575-582.

Rabiner L. R., Levinson S. E. y Sondhi M. M., 1983. On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition. Bell system technical journal, vol. 62, no. 4, pp 1075-1105.

Rabiner L. R., Sambur M. R. y Schmidt C. E., 1975. Applications of a nonlinear smoothing algorithm to speech processing. IEEE transactions on ASSP, vol. 23, no. 6, pp. 552-557.

Rabiner L. R. y Juang B. H, 1993. Fundamentals of speech recognition. PTR Prentice Hall, Englewood Cliffs, New Jersey, USA, 507 p.

Rabiner L. R. y Schafer R. W., 1978. Digital processing of speech signals. Prentice Hall, Inc., Englewood Cliffs, New Jersey, USA, 512.

Rosenberg A. E. y Soong F. K., 1986. Evaluation of a vector quantization talker recognition system in text independent and text dependent modes. ICASSP'86 Tokio, pp. 873-876.

Sambur M. R., 1975. Selection of acoustic features for speaker identification. IEEE transactions on ASSP, vol. 23, no. 2, pp. 176-182.

Soong F. K., Rosenberg A. E. y Juang B. H., 1987. A vector quantization approach to speaker recognition. AT&T technical journal, vol. 66, issue 2, pp. 14-26.

Soong F. K. y Rosenberg A. E., 1986. On the use of instantaneous and transitional spectral information in speaker recognition. ICASSP'86 Tokio, pp. 877-880.

Spiegel M. R., 1970. Estadística. McGraw-Hill de México, México, D.F., pp. 357.

Tishby N. Z., 1991. On the application of mixture AR hidden Markov models to text independent speaker recognition. IEEE transactions on signal processing, vol. 39, no. 3, pp. 563-570.

Velius G., 1988. Variants of cepstrum based speaker identity verification. IEEE CH2561-9/88/0000-0583, pp. 583-586.

Zheng Y. C. y Yuan B. Z., 1988. Text-dependent speaker identification using circular hidden Markov models. IEEE CH2561-9/88/0000-0580, pp. 580-582.

apéndice A - Teoría acústica de la producción de la voz y modelos de tubos sin pérdidas.

A.1 Teoría acústica de la producción de la voz.

Ya que las ondas sonoras son creadas por vibración y son propagadas en el aire o en otro medio por vibración de las partículas en ese medio, las leyes de la física son la base para describir la generación y propagación de sonido en el sistema vocal. En particular se requiere del uso de las leyes de conservación de masa, conservación de momento, y conservación de la energía; además de los principios de la termodinámica y mecánica de fluidos, que son todas aplicables a un fluido compresible de baja viscosidad como lo es el aire, el medio para la propagación de la voz. Una teoría acústica detallada de la producción de la voz debe considerar los siguientes efectos [Rabiner y Schafer, 1978]:

a) Variación temporal de la forma del tracto vocal.

b) Pérdidas por conducción de calor y fricción viscosa en las paredes del tracto vocal.

c) Maleabilidad de las paredes del tracto vocal.

d) Radiación de sonido en los labios.

e) Acoplamiento nasal.

f) Excitación de sonido en el tracto vocal.

Sin embargo, la discusión de una teoría que incorpore todos los efectos mencionados está fuera del alcance de este trabajo, y aún actualmente es motivo de investigación y estudio. El análisis que se presentará enseguida no pretende ser exhaustivo, la intención de este análisis es establecer la relación que existe entre los modelos del tracto vocal y los parámetros que lo caracterizan y que han sido utilizados en RAL.

A.1.1 Acústica básica de la propagación del sonido.

En la figura 46 se muestra una de las configuraciones físicas más simples que pueden

utilizarse para analizar el proceso de producción de la voz, dónde el tracto vocal es modelado como un tubo no uniforme de sección transversal variante en el tiempo. En general, en el espacio libre las ondas sonoras viajan desde la fuente en todas direcciones, sin embargo, a la mayoria de las frecuencias de interés en la voz (v.g., debajo de 4 kHz) la longitud de onda λ se puede considerar mucho mayor que el diámetro del tracto vocal; y es razonable suponer que se está tratando con la propagación de una onda plana a lo largo del tubo [O'Shaughnessy, 1987]. Por ejemplo, para una frecuencia de 4 kHz

$$\lambda = \frac{c}{f} = \frac{340\ m/s}{4000/s} = 8.5\ cm \tag{123}$$

la cual es mucho mayor que un diámetro promedio de 2 *cm* para el tracto vocal. Bajo esa consideración, y con la suposición de que no hay pérdidas por viscosidad o conducción térmica, ya sea en el volumen del fluido o en las paredes del tubo, se ha demostrado que las ondas sonoras en el tubo satisfacen el siguiente par de ecuaciones [Rabiner y Schafer, 1978]:

$$-\frac{\partial p}{\partial x} = \rho \frac{\partial (u\ A)}{\partial t} \tag{124}$$

$$-\frac{\partial u}{\partial x} = \frac{1}{\rho c^2}\frac{\partial (pA)}{\partial t} + \frac{\partial A}{\partial t} \tag{125}$$

dónde:

$p = p(x,t)$ es la presión de sonido en la posición x y el tiempo t.

$u = u(x,t)$ es la velocidad del volumen de aire en la posición x y el tiempo t.

ρ es la densidad de aire en el tubo (aproximadamente 1.2 mg/cc).

c es la velocidad del sonido.

$A=A(x,t)$ es la función área del tubo como función del tiempo.

La solución completa de (124) y (125) requiere que se calcule la presión y velocidad del volumen en las regiones limitadas por la glotis y los labios, para eso es necesario conocer

las condiciones de frontera en los límites del tubo. En los labios, las condiciones de frontera deben tomar en cuenta la radiación de sonido, mientras que en la glotis, la condición de frontera es impuesta por la naturaleza de la excitación. Además de lo anterior debe conocerse también la función área $A(x,t)$, la cual puede considerarse invariante en el tiempo para sonidos continuantes pero no para sonidos no continuantes.

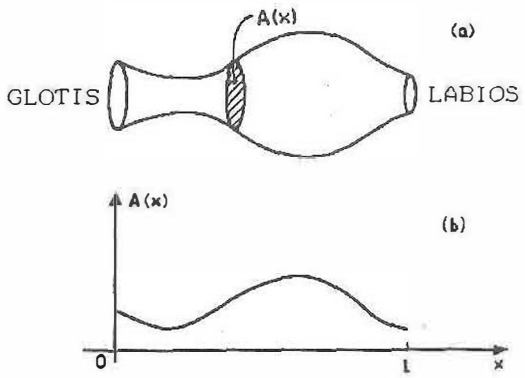


Figura 46. (a) Tracto vocal como un tubo no uniforme.
(b) Función área [Rabiner y Schafer, 1978].

La solución completa de (124) y (125) es muy complicada aún si se determina $A(x,t)$ de manera exacta; afortunadamente, bajo las condiciones más generales no es necesaria la solución de tales ecuaciones para obtener una visión de la naturaleza de las señales de voz.

A.1.2 Tubo uniforme sin pérdidas.

El modelo presentado anteriormente puede ser simplificado drásticamente si se considera que la función área del tracto vocal es constante en x y en t (v.g., un tubo invariante en el tiempo con sección transversal uniforme). Si bien este modelo es bastante simple, puede servir como base de modelos más reales que se examinarán más adelante.

La figura 47 muestra un tubo de sección transversal uniforme que es excitado por una fuente ideal de velocidad de volumen de aire, la cual es representada por un pistón que se puede mover en cualquier modo deseado independientemente de las variaciones de

presión en el tubo. Una simplificación adicional es que en la terminal abierta del tubo no hay variaciones en la presión del aire, sólo en la velocidad del volumen de aire. Esas fuertes simplificaciones son imposibles de conseguir en la práctica, sin embargo, la solución para este modelo tiene mucho en común con modelos más reales, los cuales pueden ser obtenidos concatenando tubos uniformes, como se verá más adelante.

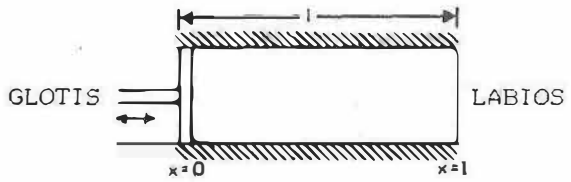


Figura 47. Tubo uniforme sin pérdidas con terminaciones ideales [Rabiner y Schafer, 1978].

Si $A(x,t) = A$ es una constante, las ecuaciones (124) y (125) se reducen a:

$$-\frac{\partial p}{\partial x} = \frac{\rho}{A}\frac{\partial u}{\partial t} \tag{126}$$

$$-\frac{\partial u}{\partial x} = \frac{A}{\rho c^2}\frac{\partial p}{\partial t} \tag{127}$$

y se puede demostrar (v.g., por sustitución) que la solución de (126) y (127) tiene la forma:

$$u(x,t) = [u^+(t - x/c) - u^-(t + x/c)] \tag{128}$$

$$p(x,t) = \rho\frac{c}{A}[u^+(t - x/c) + u^-(t + x/c)] \tag{129}$$

dónde $u^+(t - x/c)$ y $u^-(t + x/c)$ pueden interpretarse como ondas viajando en las direcciones positiva y negativa respectivamente. La relación entre esas ondas está determinada por la condiciones de frontera.

Para obtener una representación del comportamiento espectral del tubo supóngase que éste es excitado por una velocidad de volumen senoidal $u_G(t)$ aplicada en la glotis.

Esto es

$$u(0,t) = u_G(t) = U_G(\Omega)\, e^{j\Omega t} \tag{130}$$

dónde Ω es la frecuencia de la excitación en radianes y $U_G(\Omega)$ es su amplitud. Ya que las ecuaciones (126) y (127) son lineales, la solución $u^+(t-x/c)$ y $u^-(t+x/c)$ debe ser de la forma:

$$u^+(t-x/c) = k^+ e^{j\Omega(t-x/c)} \tag{131}$$

$$u^-(t+x/c) = k^- e^{j\Omega(t+x/c)} \tag{132}$$

Sustutiyendo (131) y (132) en (128) y (129), y aplicando la condición de frontera

$$p(l,t)=0 \tag{133}$$

en la terminal labial del tubo y la ecuación (130) en la terminal glotal se puede resolver para las constantes k^+ y k^-. Las ecuaciones resultantes para $p(x,t)$ y $u(x,t)$ son:

$$p(x,t) = jZ_0 \frac{sen[\Omega(l-x)/c]}{cos[\Omega l/c]} U_G(\Omega)\, e^{j\Omega t} \tag{134}$$

$$u(x,t) = \frac{cos[\Omega(l-x)/c]}{cos[\Omega l/c]} U_G(\Omega)\, e^{j\Omega t} \tag{135}$$

dónde $$Z_0 = \frac{\rho c}{A} \tag{136}$$

es llamada la *impedancia acústica característica* del tubo.

Las ecuaciones (134) y (135) expresan la relación entre la velocidad del volumen de aire de la excitación, y la presión y velocidad del volumen en cualquier punto del tubo. En particular, si se considera la relación entre la velocidad del volumen en los labios y la velocidad del volumen en la excitación se obtiene de la ecuación (135)

$$u(l,t) = U(l,\Omega)\, e^{j\Omega t} = \frac{1}{cos(\Omega l/c)} U_G(\Omega) e^{j\Omega t} \tag{137}$$

El cociente

$$\frac{U(l,\Omega)}{U_G(\Omega)} = V_a(\Omega) = \frac{1}{\cos(\Omega l/c)} \tag{138}$$

es la respuesta en frecuencia relacionando la velocidad de los volúmenes de entrada y salida. La figura 48a muestra esta relación para valores c=35,000 *cm/s* y l=17.5 *cm*, el cual es un valor típico para el tracto vocal de un hombre adulto. Los valores de l para mujeres y niños son 13 cm y 10 cm respectivamente [O'Shaughnessy, 1987].

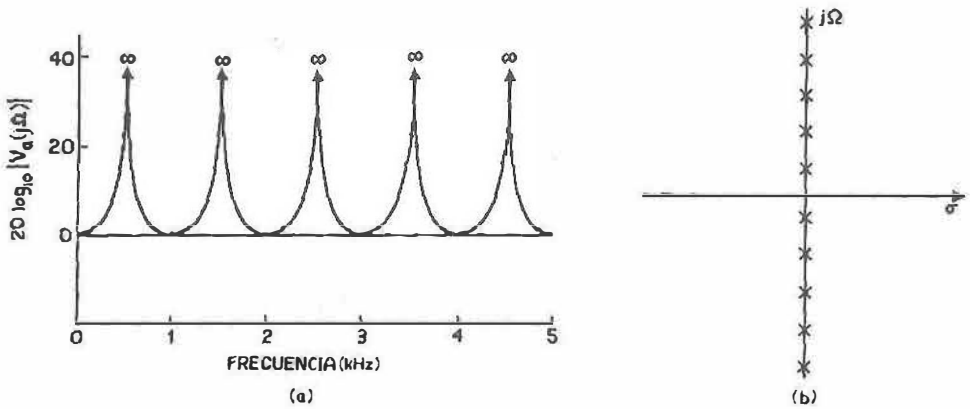


Figura 48. (a) Respuesta en frecuencia. (b) Localización de los polos para un tubo uniforme sin pérdidas [Rabiner y Schafer, 1978].

Si se reemplaza Ω por s/j en la ecuación (138) se obtiene la transformada de Laplace o función de transferencia

$$V_a(s) = \frac{2\,e^{-sl/c}}{1+e^{-s2l/c}} \tag{139}$$

la cual tiene un número infinito de polos igualmente espaciados en el eje jΩ en

$$s_n = \pm j\left[\frac{(2n+1)\pi c}{2l}\right] \qquad n = 0, \pm 1, \pm 2, \ldots \tag{140}$$

La localización de estos polos se muestra en la figura 48b; estos polos son las *frecuencias formantes* del sistema y para este caso particular su ancho de banda es cero.

Las ecuaciones obtenidas en esta sección se basan en la suposición de que no hay pérdidas de energía en en el tubo. Sin embargo, estas pérdidas existen y pueden ser ocasionadas ya sea por fricción viscosa entre el aire y las paredes del tubo, o por conducción

de calor y vibración de las mismas. Analizar el efecto de esas pérdidas queda fuera del contexto de este trabajo. Estos efectos han sido analizados por [Rabiner, 1978] y [O'Shaughnessy, 1987].

A.2 Modelo de concatenación de tubos sin pérdidas.

Un modelo que se ha usado ampliamente en el procesamiento de voz se basa en la suposición de que el tracto vocal puede ser modelado como una concatenación de tubos acústicos sin pérdidas como el mostrado en la figura 49.

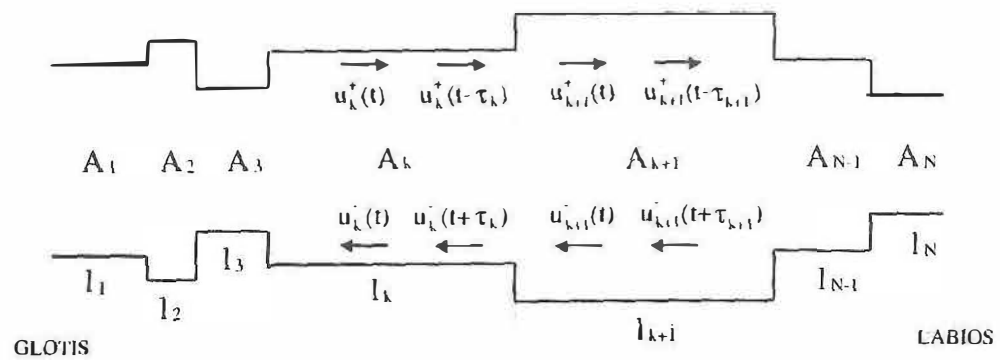


Figura 49. Modelo del tracto vocal como una concatenación de tubos acústicos sin pérdidas.

Las áreas transversales constantes $\{A_k\}$ de los tubos se eligen de tal manera que aproximen la función área $A(x)$ del tracto vocal. Si se elige un número grande de tubos de longitud corta se puede esperar que los *formantes* tengan características muy similares a los de un tubo con la función área del tracto vocal. Sin embargo, este modelo desprecia las pérdidas por fricción, conducción de calor y vibración de las paredes; y es de esperarse que los anchos de banda de los formantes difieran de aquellos modelos que si consideran estas pérdidas; una alternativa es atribuir estas pérdidas a la glotis y los labios, de manera tal que el modelo pueda representar adecuadamente las propiedades de resonancia del tracto vocal.

La importancia de este modelo radica en que provee una transición bastante conveniente entre modelos de tiempo contínuo y modelos de tiempo discreto, cuyos

parámetros se encuentran estrechamente relacionados y han sido aplicados en RAL.

A.2.1 Propagación de ondas en tubos concatenados sin pérdidas.

Ya que cada tubo en la figura 49 se supone sin pérdidas, la propagación del sonido en cada uno de ellos cumple las ecuaciones (128) y (129). En particular, si se considera el *k-ésimo* tubo con área transversal A_k, la presión y velocidad del volumen para este tubo tiene la forma

$$p_k(x,t) = \frac{\rho c}{A_k}\left[u_k^+(t - x/c) + (u_k^-/t + x/c)\right] \tag{141}$$

$$u_k(x,t) = u_k^+(t - x/c) - u_k^-(t + x/c) \tag{142}$$

dónde x es la distancia medida desde la terminación izquierda del *k-ésimo* tubo $(0 \le x \le l_k)$ y $u_k^+(t)$ y $u_k^-(t)$ son ondas viajando en sentido positivo y negativo en el *k-ésimo* tubo.

La relación entre las ondas viajando en tubos adyacentes está dada por el principio fisico que establece que la presión y velocidad del volumen son continuos en el tiempo y en el espacio en cualquier punto del sistema; y así, considerando los *k-ésimo* y *(k+1)-ésimo* tubos se deben cumplir

$$p_k(l_k,t) = p_{k+1}(0,t) \tag{143}$$

$$u_k(l_k,t) = u_{k+1}(0,t) \tag{144}$$

y sustituyendo las ecuaciones (141) y (142), en (143) y (144) se obtiene

$$\frac{A_{k+1}}{A_k}\left[u_k^+(t - \tau_k) + u_k^-(t + \tau_k)\right] = u_{k+1}^+(t) + u_{k+1}^-(t) \tag{145}$$

$$u_k^+(t - \tau_k) - u_k^-(t + \tau_k) = u_{k+1}^+(t) - u_{k+1}^-(t) \tag{146}$$

dónde $\tau_k = l_k / c$ es el tiempo que tarda la onda sonora en viajar a través del *k-ésimo* tubo.

De la figura 49 se puede inferir que una parte de la onda que viaja en sentido positivo es propagada hacia la derecha al alcanzar la juntura, mientras que otra parte es

reflejada. De manera similar, una parte de la onda que viaja en sentido negativo es propagada hacia la izquierda, mientras que otra es reflejada hacia la derecha. La velocidad neta del volumen que viaja tanto en sentido positivo como negativo $u_{k+1}^{+}(t)$ y $u_{k}^{-}(t+\tau_k)$, puede calcularse despejando $u_{k}^{-}(t+\tau_k)$ de la ecuación (146) y sustituyendo en (145)

$$u_{k+1}^{+}(t) = (1+r_k)\,u_{k}^{+}(t-\tau_k) + r_k\,u_{k+1}^{-}(t) \tag{147}$$

y restando la ecuación (146) de la ecuación (145) se obtiene

$$u_{k}^{-}(t+\tau_k) = -r_k\,u_{k}^{+}(t-\tau_k) + (1-r_k)\,u_{k+1}^{-}(t) \tag{148}$$

dónde

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} \tag{149}$$

es la cantidad de velocidad de volumen que es reflejada en sentido positivo y negativo respectivamente, y es llamado **coeficiente de reflexión** para la *k-ésima* juntura. Se puede demostrar que ya que todas las áreas son positivas

$$-1 \le r_k \le 1 \tag{150}$$

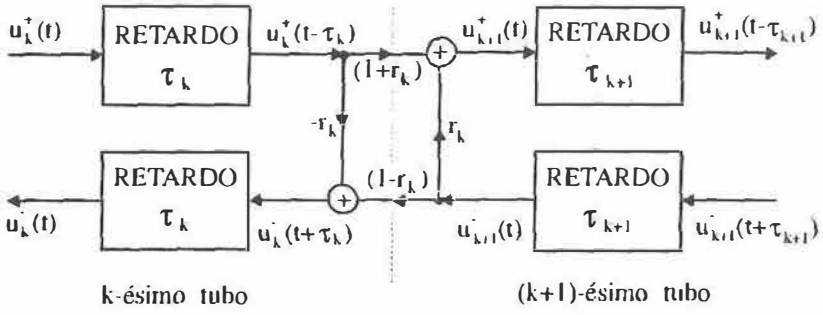


Figura 50. Flujo de señal en la juntura entre dos tubos sin pérdidas.

Las ecuaciones (147) y (148) pueden ser representadas en una gráfica de flujo como la que se muestra en la figura 50, donde el tiempo de propagación τ_k se convierte en un elemento de retardo; las flechas etiquetadas denotan multiplicaciones y los nodos circulares sumas. Un modelo de N tubos tendría N pares de elementos de retardo y N-1 junturas, cada

una de ellas caracterizada por un *coeficiente de reflexión.*

A.2.2 Condiciones de frontera.

En la unión de los labios la presión $p_N(l_N,t)$ y la velocidad del volumen $u_N(l_N,t)$, a la salida del *N-ésimo* tubo, están relacionados por la ecuación

$$P_N(l_N,\Omega) = Z_L\, U_N(l_N,\Omega) \tag{151}$$

Si se supone por simplicidad que Z_L es real entonces se obtiene la relación en el dominio del tiempo

$$\frac{\rho c}{A_N}\left[u_N^+(t-\tau_N)+u_N^-(t+\tau_N)\right] = Z_L\left[u_N^+(t-\tau_N)-u_N^-(t+\tau_N)\right] \tag{152}$$

y despejando de esta relación $u_N^-(t+\tau_N)$ se obtiene

$$u_N^-(t+\tau_N) = -r_L\, u_N^+(t-\tau_N) \tag{153}$$

donde el coeficiente de reflexión en los labios es

$$r_L = \left[\frac{\rho c / A_N - Z_L}{\rho c / A_N + Z_L}\right] \tag{154}$$

y la velocidad del volumen de salida en los labios es

$$u_N(l_N,t) = u_N^+(t-\tau_N) - u_N^-(t+\tau_N) = (1+r_L)\, u_N^+(t-\tau_N) \tag{155}$$

El efecto de esta terminación se puede apreciar en el extremo derecho de la figura 51.

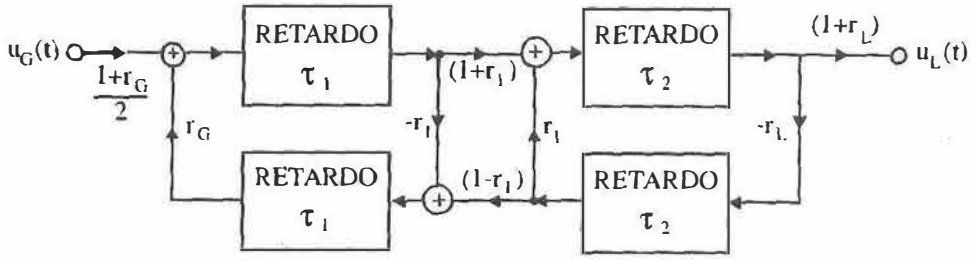


Figura 51. Gráfica de flujo de señal para un modelo completo de dos tubos.

Algunos modelos para la excitación glotal consideran que ésta puede ser modelada como una fuente de velocidad de volumen $u_G(t)$ en paralelo con una impedancia acústica

Z_G. Para el primer tubo esta relación está dada por

$$U_1(0,\Omega) = U_G(\Omega) - P_1(0,\Omega) / Z_G \tag{156}$$

Suponiendo nuevamente que Z_G es real, la velocidad neta de volumen en el primer tubo es

$$u_1^+(t) - u_1^-(t) = u_G(t) - \frac{\rho c}{A_1}\left[\frac{u_1^+(t) + u_1^-(t)}{Z_G}\right] \tag{157}$$

y despejando $u_1^+(t)$ se obtiene

$$u_1^+(t) = \frac{1+r_G}{2}\, u_G(t) + r_G\, u_1^-(t) \tag{158}$$

donde el coeficiente de reflección glotal es

$$r_G = \left[\frac{Z_G - \frac{\rho c}{A_1}}{Z_G + \frac{\rho c}{A_1}}\right] \tag{159}$$

El gráfico de flujo de la ecuación (158) se encuentra representado en el extremo izquierdo de la figura 51.

A.2.3 Interpretación digital.

Ya que el modelo de concatenación de tubos sin pérdidas de la figura 51 contiene únicamente sumas, multiplicaciones y retardos, éste puede transformarse en un modelo digital; para ello considérese un sistema compuesto de N tubos concatenados sin pérdidas de longitud $\Delta x = l/N$, dónde l es la longitud total del tracto vocal, tal como se muestra en la figura 52. El retardo para cada uno de los tubos es

$$\tau = \frac{\Delta x}{c} \tag{160}$$

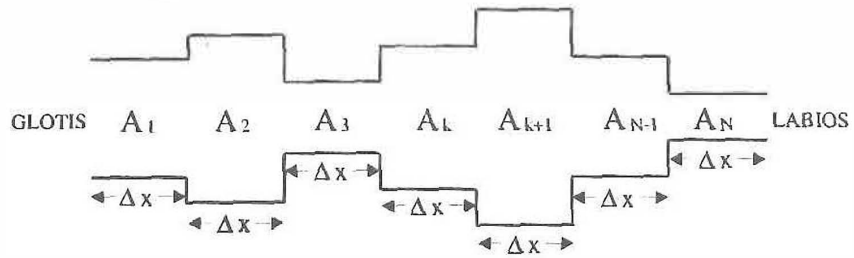


Figura 52. Concatenación de N tubos sin pérdidas de longitud igual.

Suponiendo que se excita este sistema con un impulso (v.g., $u_G(t) = \delta(t)$), se puede considerar que el impulso es parcialmente reflejado y parcialmente propagado en las junturas. Un análisis detallado del proceso confirmará que la respuesta al impulso (v.g., la velocidad de volumen en los labios debida a un impulso en la glotis) tiene la forma

$$v_a(t) = \alpha_0 \delta(t - N\tau) + \sum_{k=1}^{\infty} \alpha_k \delta(t - N\tau - 2k\tau) \quad , \quad \alpha_k \in \Re \tag{161}$$

de donde puede verse claramente que el tiempo mínimo para que un impulso alcance la salida es $N\tau$ segundos, con impulsos sucesivos debidos a las reflexiones que alcanzan la salida a múltiplos de 2τ segundos más tarde. La cantidad 2τ es el tiempo requerido para que una onda se propague en ambos sentidos en un tubo.

La función de transferencia de este sistema es de la forma

$$V_a(s) = \sum_{k=0}^{\infty} \alpha_k \, e^{-s(N+2k)\tau} = e^{-sN\tau} \sum_{k=0}^{\infty} \alpha_k e^{-s2\tau k} \tag{162}$$

donde el factor $e^{-sN\tau}$ es el retardo temporal necesario para una propagación en las N secciones, mientras que la cantidad

$$\hat{V}_a(s) = \sum_{k=0}^{\infty} \alpha_k e^{-sk2\tau} \tag{163}$$

es la función de transferencia de un sistema lineal cuya respuesta al impulso es $\hat{v}_a(t) = v_a(t + N\tau)$. Ya que la respuesta en frecuencia $\hat{V}_a(\Omega)$ es

$$\hat{V}_a(\Omega) = \sum_{k=0}^{\infty} \alpha_k e^{-j\Omega 2k\tau} \tag{164}$$

se puede demostrar que

$$\hat{V}_a(\Omega + \frac{2\pi}{2\tau}) = \hat{V}_a(\Omega) \tag{165}$$

Lo establecido en la ecuación (165) es una característica de los sistemas discretos, y de hecho, si la señal de excitación es limitada a frecuencias menores que $\pi/(2\tau)$ es posible muestrear la entrada con período $T=2\tau$ y filtrar la señal muestreada con un filtro digital cuya respuesta al impulso es

$$\hat{v}_a(n) = \begin{cases} \alpha_n & n \geq 0 \\ 0 & n < 0 \end{cases} \tag{166}$$

Para un período de muestreo $T=2\tau$, el retardo de $N\tau$ corresponde a un corrimiento de $N/2$ muestras. Hay que notar que si N es par, $N/2$ es un entero y el retardo puede ser implementado simplemente recorriendo la secuencia de salida $\hat{v}_a(n)$; sin embargo, si N es impar es necesario hacer una interpolación de la secuencia $\hat{v}_a(n)$ para obtener la secuencia de salida. En la práctica este retardo puede ser ignorado en la mayoría de los casos, sin consecuencias en la mayoría de las aplicaciones de voz [Rabiner y Schafer, 1978].

La transformada z de $\hat{v}_a(n)$ se obtiene sustituyendo e^{sT} por z en la ecuación (163), y de esta manera

$$\hat{v}_a(z) = \sum_{k=0}^{\infty} \alpha_k z^{-k} \tag{167}$$

La figura 53a muestra la gráfica de flujo de señal para un modelo del tracto vocal como tres tubos concatenados sin pérdidas. En la figura 53b se muestra la versión de la figura 53a como un sistema discreto donde los retardos han sido sustituídos por el operador $z^{-1/2}$, lo cual implica un retardo de media muestra ya que $\tau=T/2$. El retardo de media muestra de la figura 53b requiere una interpolación entre dos muestras consecutivas; una

configuración en la que no se requiere la interpolación se muestra en la figura 53c, que se obtiene al mover los retardos de la parte inferior hacia la parte superior. Puede notarse que los retardos en cualquier ruta cerrada de las figuras 53b y 53c se conservan, sin embargo, el retardo total desde la entrada hasta la salida no se mantiene, lo cual es de una significancia menor en la práctica y puede ser compensado teóricamente con la cantidad adecuada de adelanto $z^{N/2}$ [O'Shaughnessy, 1987].

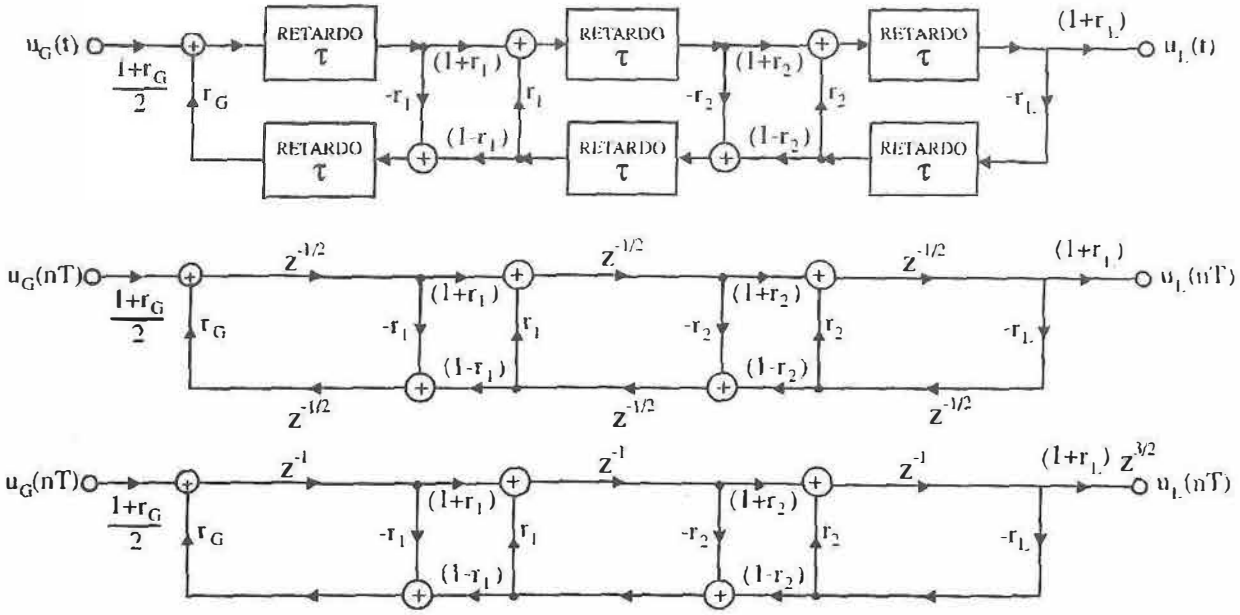


Figura 53. (a) Gráfica de flujo se señal para el modelo del tracto vocal como tres tubos sin pérdidas concatenados; (b) Sistema equivalente de tiempo discreto; (c) Sistema equivalente de tiempo discreto usando únicamente retardos completos.

Se puede demostrar [Atal y Hanauer, 1971][Rabiner y Schafer, 1978] que la función de transferencia para un modelo de N tubos concatenados, sin pérdidas y sin considerar la radiación en los labios, está dada por

$$V(z) = \frac{U_L(z)}{U_G(z)} = \frac{1+r_G}{2} \frac{\prod_{k=1}^{N}(1+r_k)z^{-N/2}}{1-\sum_{k=1}^{N}\alpha_k z^{-k}} \tag{168}$$

dónde α_k son coeficientes que dependen de r_k. Puede observarse en la ecuación (168) que *V(z)* tiene un retardo correspondiente al número de secciones del modelo, y que no tiene ceros finitos, únicamente polos; estos polos definen las *frecuencias formantes* del modelo.

apéndice B Glosario de términos y acrónimos español-inglés.

ADT	Alineamiento dinámico en el tiempo.
DTW	Dynamic time warping.
AGL	Algoritmo generalizado de Lloyd.
GLA	Generalized Lloyd's algorithm.
CIA	Circuito de interfase analógica.
AIC	Analog interface circuit.
CORPAR	Correlación parcial.
PARCOR	Partial correlation.
CPL	Codificación predictiva lineal.
LPC	Linear predictive coding.
CV	Cuantificación vectorial.
VQ	Vector quantization.
IAL	Identificación automática de locutor.
ASI	Automatic speaker identification.
MEM	Modelos evidentes de Markov.
OMM	Observable Markov model.
MLNL	Mínimo local no restringido a límites de palabra.
UELM	Unconstrained endpoints, local minimum.
MOM	Modelo oculto de Markov.
HMM	Hidden Markov model.
PCO	Procedimiento de codificación óptima.
RAH	Reconocimiento automático del habla.
ASR	Automatic speech recognition.
RAL	Reconocimiento automático de locutor.
ASR	Automatic speaker recognition.
RAL-DV	Reconocimiento automático de locutor dependiente del vocabulario.

RAL-IV	Reconocimiento automático de locutor independiente del vocabulario.
RAV	Reconocimiento automático de voz.
RLIP2-1 CE2-1	Restringido a límites de palabra , de intervalo de pendiente 2 a 1. Constrained endpoints, 2 to 1 range of slope.
RNA ANN	Red neuronal artificial. Artificial neural network.
RVMC	Regla del vecino más cercano.
SERAL	Sistema experimental de reconocimiento automático de locutor.
TAPS	Tarjeta de adquisición, procesamiento y síntesis de señales.
TDF DFT	Transformada discreta de Fourier Discrete Fourier transform
TEE EER	Tasa de equierror. Equal error rate.
TFDT TDFT	Transformada de Fourier dependiente del tiempo. Time dependent Fourier transform.
TFTC STFT	Transformada de Fourier de tiempo corto. Short time Fourier transform.
VAL ASV	Verificación automática de locutor. Automatic speaker verification.

apéndice C Ejemplo de modelos evidentes de Markov (MEM).

EJEMPLO: Un hombre maneja su auto o toma el tren para ir a trabajar cada día, pero nunca toma el tren dos días seguidos; si conduce su auto para ir a trabajar existe igual probabilidad de que al día siguiente maneje o tome el tren.

Éste proceso puede ser representado por un *Modelo Evidente de Markov* de dos estados

$$A = \begin{matrix} \\ tren \\ auto \\ \\ tren \quad auto \end{matrix} \begin{bmatrix} 0 & 1 \\ 0.5 & 0.5 \end{bmatrix}$$

dónde el estado S_1 representa el evento de que el hombre tome el tren, mientras que el estado S_2 representa el evento de que el hombre conduzca su auto.

RESOLVER: Si este hombre ha conducido su auto hoy (v.g., $q_t = S_2$), ¿ Cuál es la probabilidad de que durante los próximos 5 días decida utilizar "auto, tren, auto, auto, tren" ?.

SOLUCIÓN:

Ésto equivale a determinar la probabilidad de la secuencia de observaciones $O = \{S_2, S_2, S_1, S_2, S_2, S_1\}$ correspondiente a t=1, 2, ..., 6 dado el modelo. Esta probabilidad puede ser evaluada y expresada como

$$\begin{aligned} P[O|Modelo] &= P[S_2, S_2, S_1, S_2, S_2, S_1|Modelo] \\ &= P[S_2]\, P[S_2|S_2]\, P[S_1|S_2]\, P[S_2|S_1]\, P[S_2|S_2]\, P[S_1|S_2] \\ &= \pi_2\, a_{22}\, a_{12}\, a_{21}\, a_{22}\, a_{12} \\ &= 1\,(0.5)\,(1)\,(0.5)\,(0.5)\,(1) = 0.125 \end{aligned}$$

dónde $\prod = [\pi_1, \pi_2, \ldots, \pi_N]$ es la distribución inicial de probabilidad de estados.

apéndice D Diagramas esquemáticos de la TAPS.

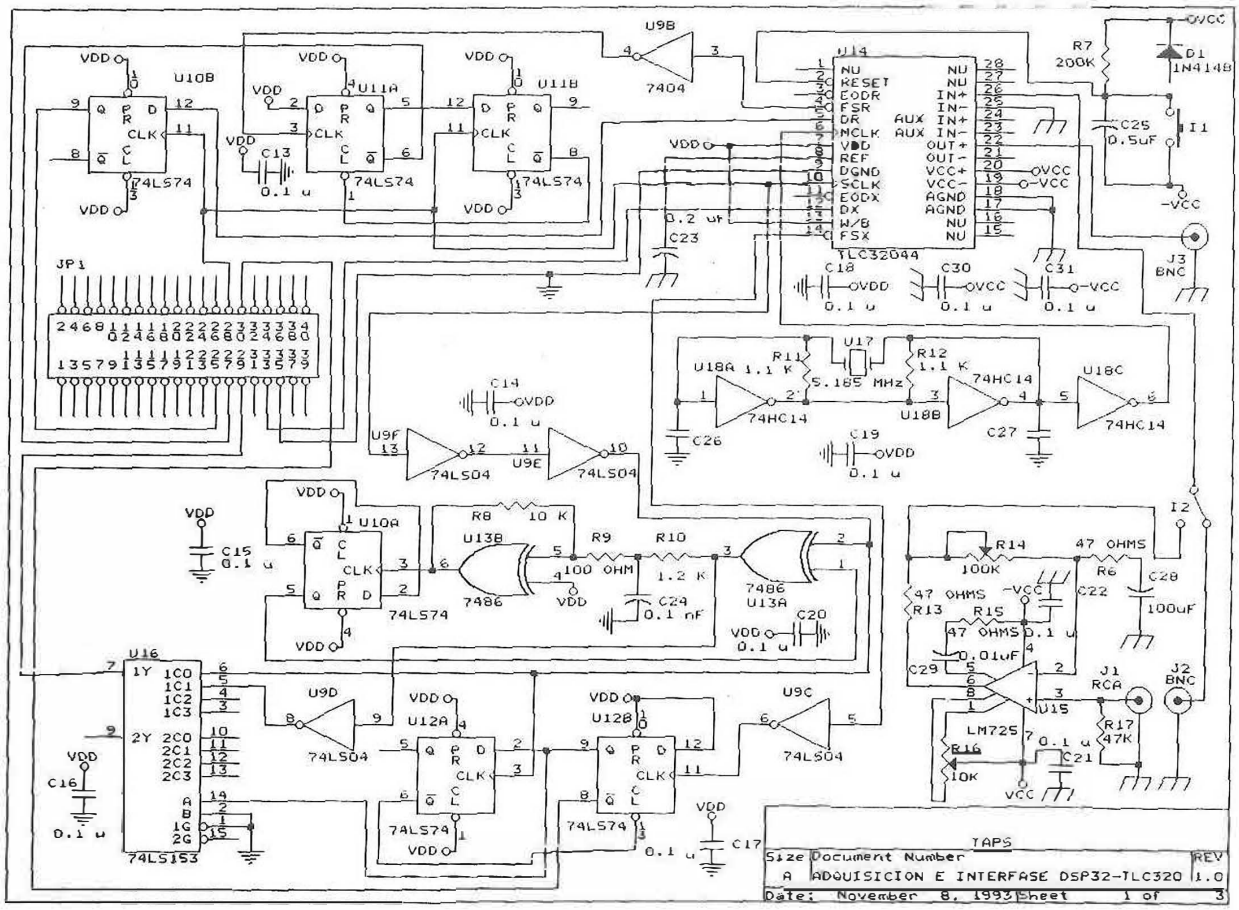


Figura 54. Diagrama esquemático de la sección de adquisición, ésta incluye la interfase DSP32-TLC32044.

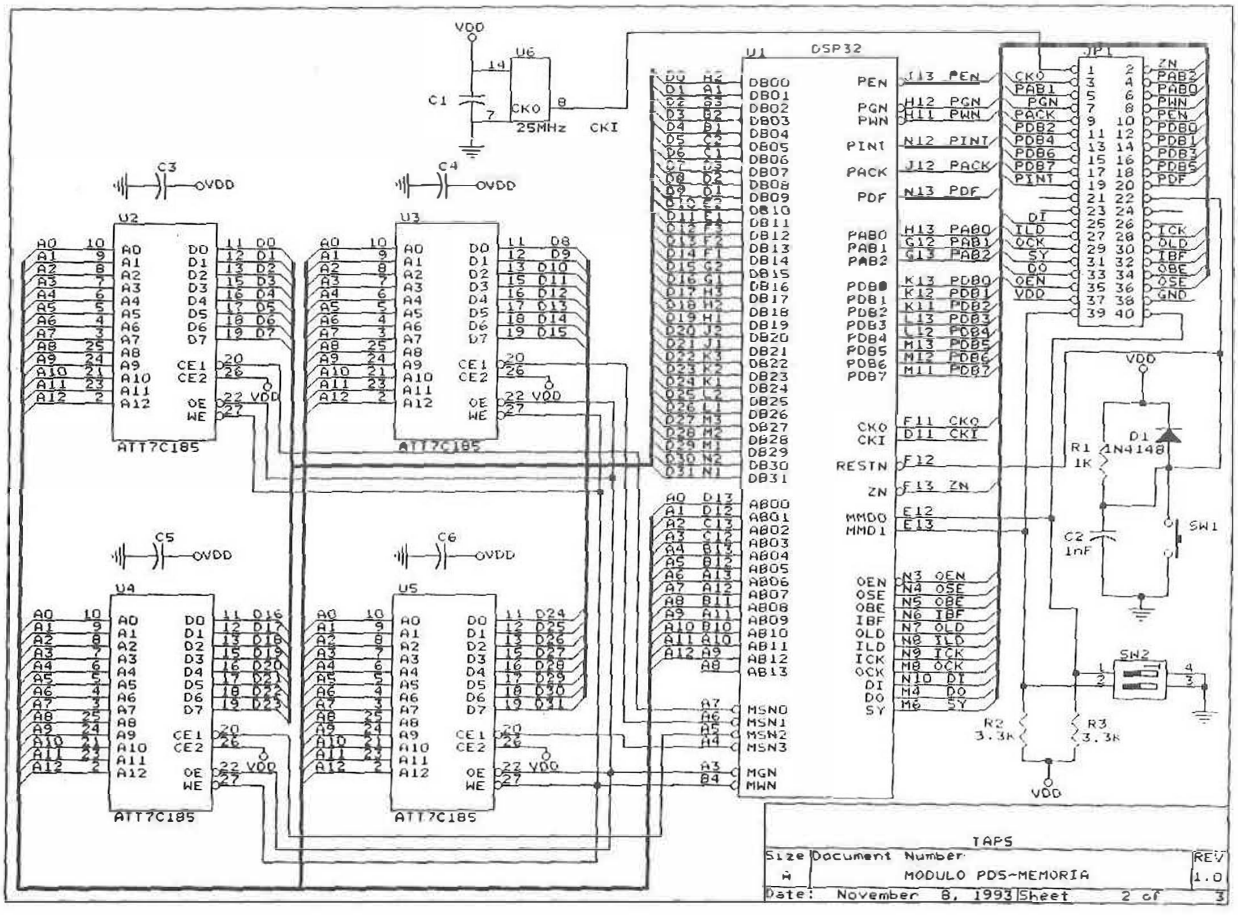


Figura 55. Diagrama esquemático del módulo de procesamiento, éste incluye el DSP32 y su memoria externa [Martínez y Félix, 1993].

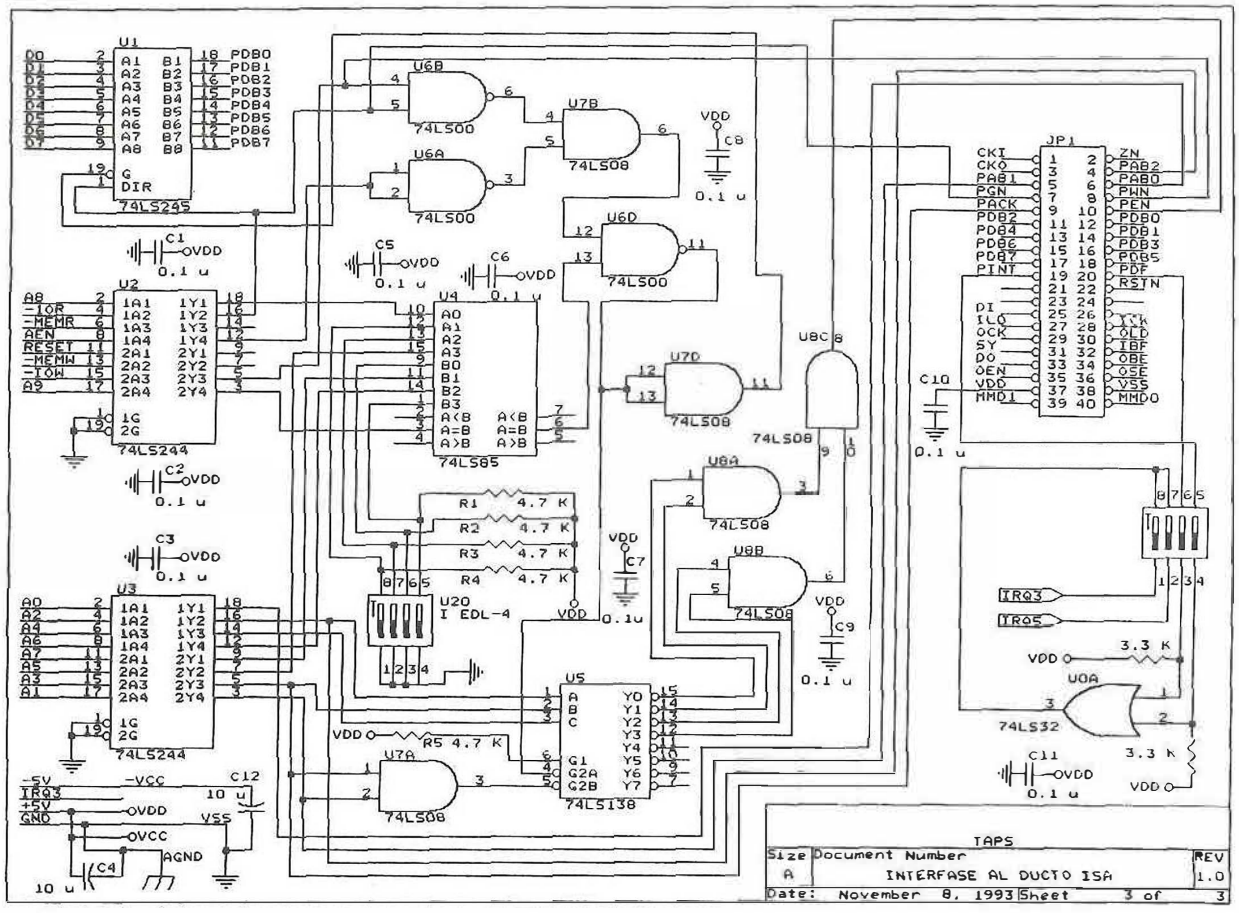


Figura 56. Diagrama esquemático de la sección de interfase al ducto ISA.

apéndice E Diagramas de flujo interactivos computadora-DSP32.

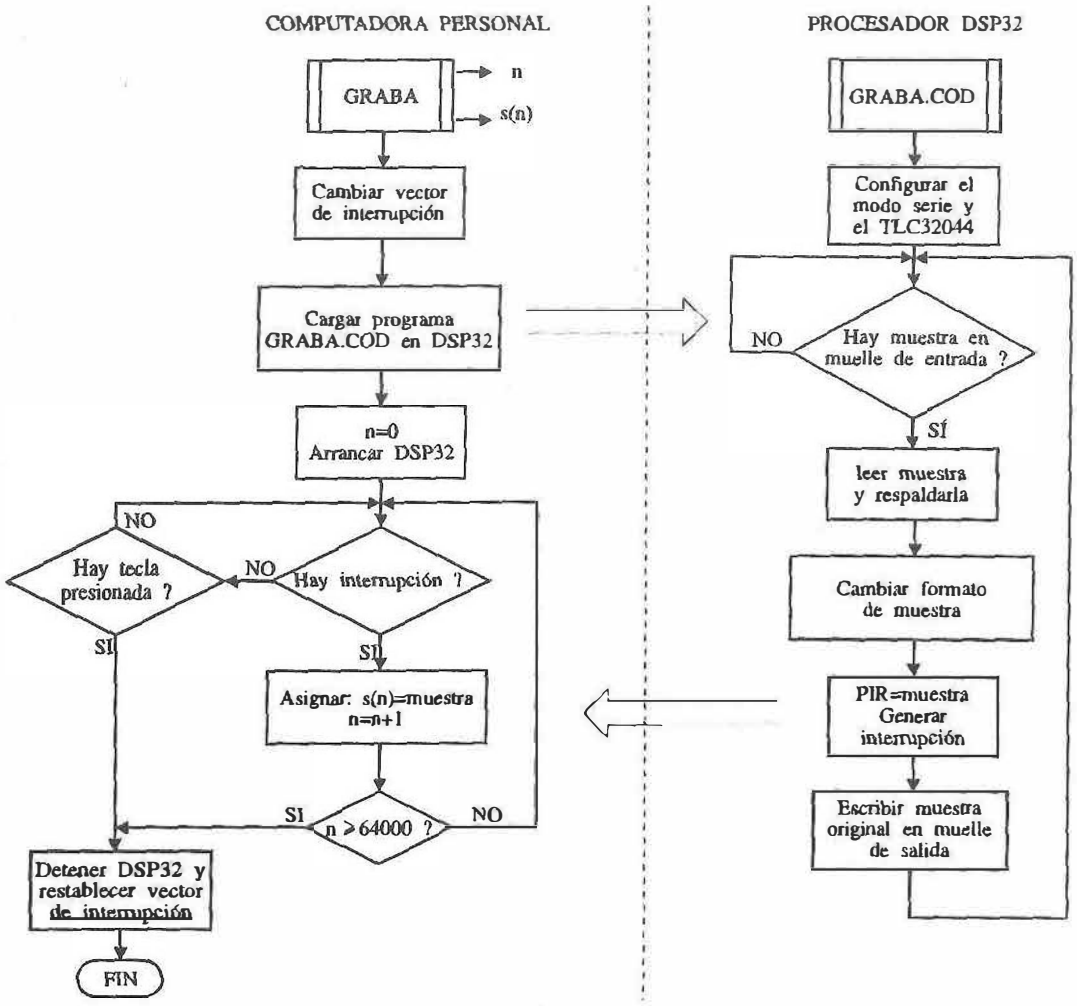


Figura 57. Diagrama de flujo interactivo computadora-DSP32 para la adquisición de señales.

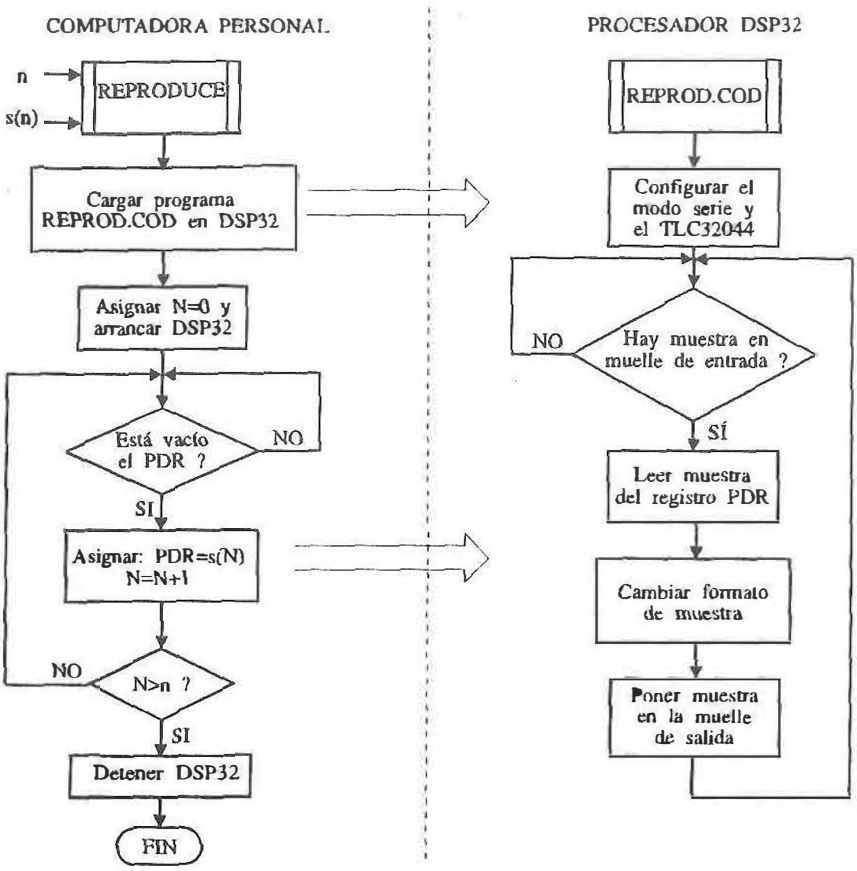


Figura 58. Diagrama de flujo interactivo computadora-DSP32 para la síntesis de señales.

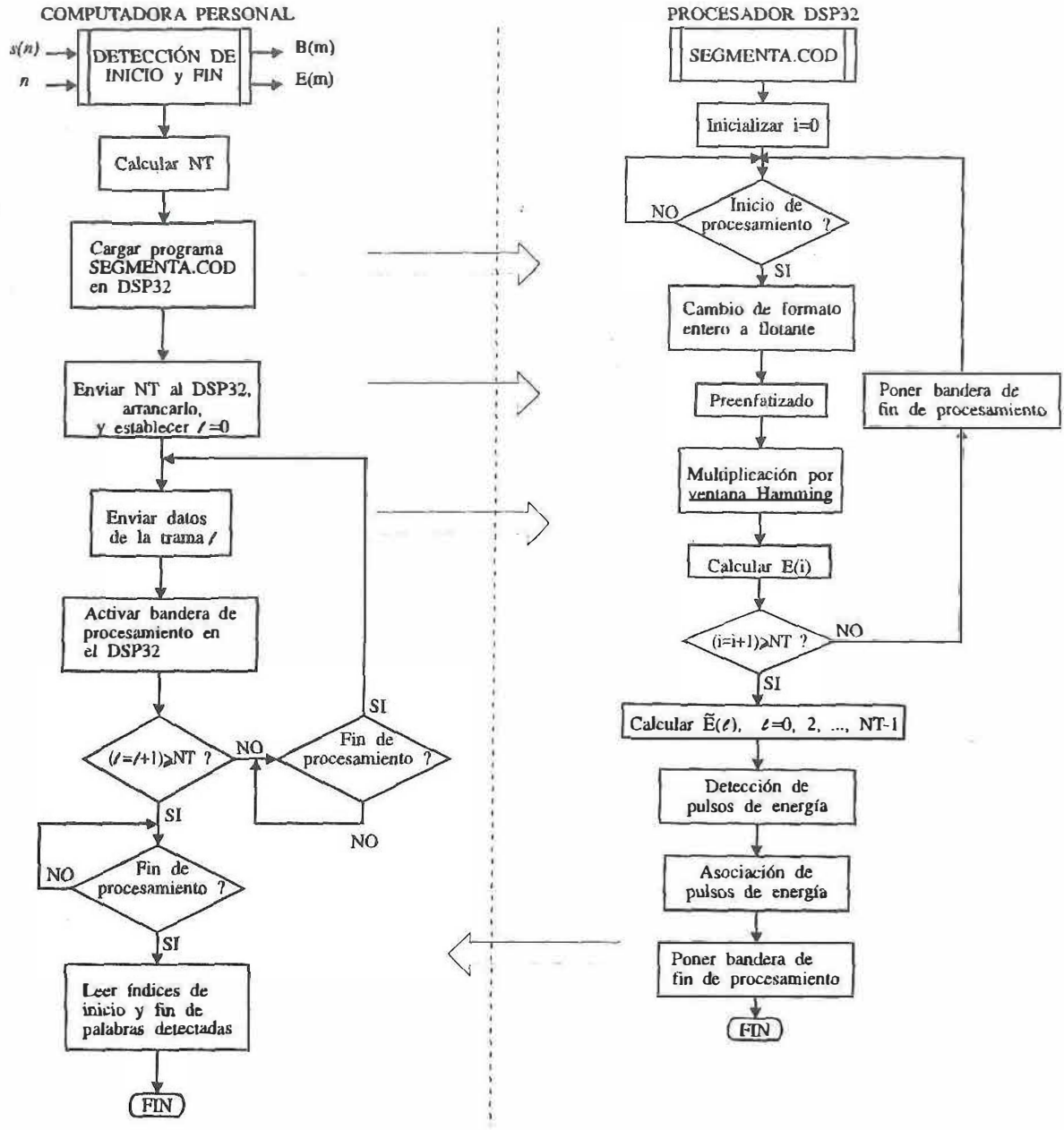


Figura 59. Diagrama de flujo interactivo computadora-DSP32 para la detección de inicio y fin de palabras aisladas.

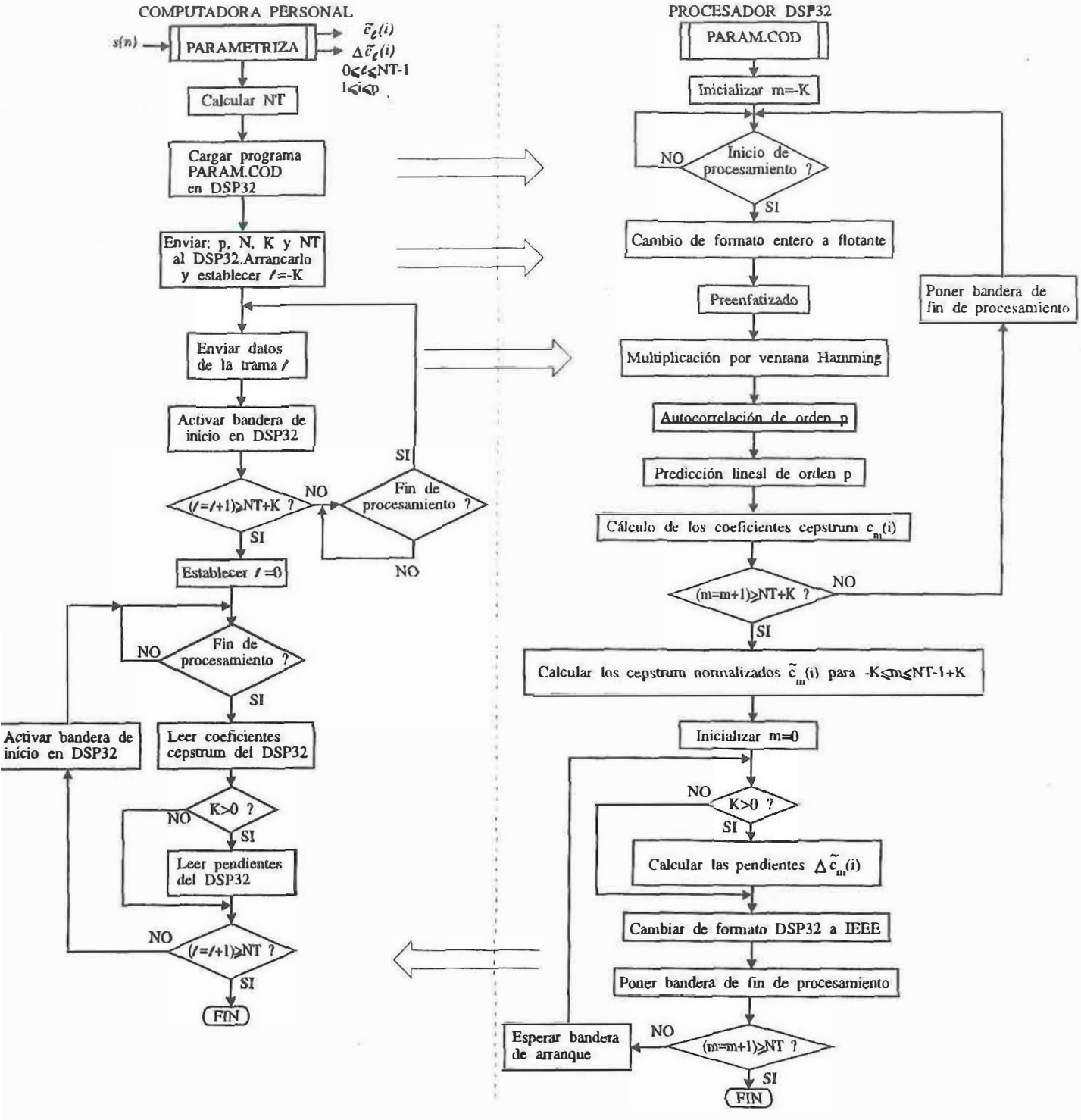


Figura 60. Diagrama de flujo interactivo computadora-DSP32 para la extracción de parámetros.

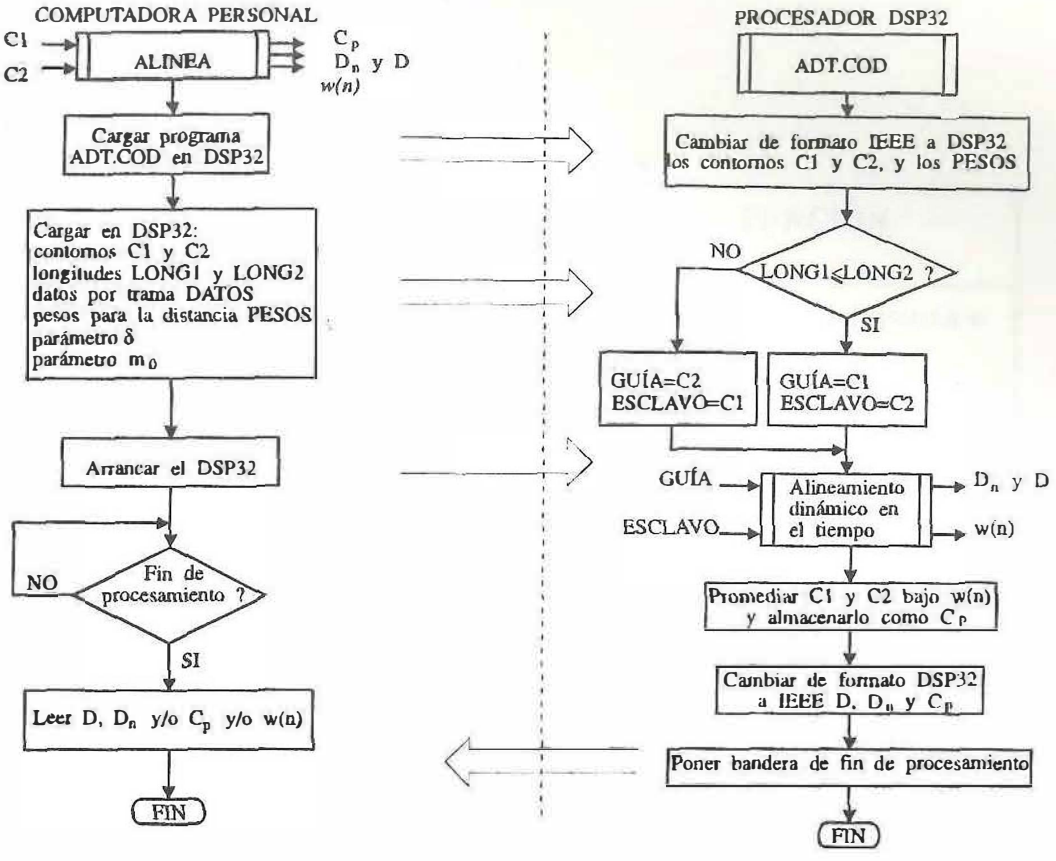


Figura 61. Diagrama de flujo interactivo computadora-DSP32 para el alineamiento, comparación y promediación de contornos.

apéndice F Lista de programas de SERAL.

PROGRAMAS

PROGRAMA FUENTE	PROGRAMA EJECUTABLE	FUNCIÓN
SERAL.C SERAL1.C	SERAL.EXE	Administración del sistema e interfase con el usuario.
GRABA.S	GRABA.COD	Adquisición de señales.
REPROD.S	REPROD.COD	Síntesis de señales.
SEGMENTA.S	SEGMENTA.COD	Detección de inicio y fin de palabras aisladas.
PARAM.S	PARAM.COD	Extracción de parámetros
ADT.S	ADT.COD	Alineamiento, comparación y promediación de contornos

ARCHIVOS AUXILIARES Y DE CONFIGURACIÓN.

ARCHIVO	CONTENIDO
PARAM.DAT	Parámetros de configuración del sistema.
PESOS.DAT	Pesos para la distancia euclidiana pesada al cuadrado.
EGAVGA.BGI	Manejador para las interfases gráficas de SERAL
LITT.CHR SANS.CHR	Tipos de letra utilizados en SERAL.

UTILERÍAS

PROGRAMA FUENTE	PROGRAMA EJECUTABLE	FUNCIÓN
DEPURA.C	DEPURA.EXE	Generar un archivo *nombre*.COD a partir de un archivo *nombre*.L, el cual fué a su vez generado por la compilación de *nombre*.S.
DSP-SS1.C	DSP-SS1.EXE	Cambia el formato de un archivo *nombre*.VOZ de SERAL al formato del programa de despliegue espectrográfico SS1 de la compañía ARIEL. El archivo de salida se llama *nombre*.DAT.
SS1-DSP.C	SS1-DSP.EXE	Cambia el formato de un archivo *nombre*.DAT de SS1 al formato de SERAL. El archivo de salida se llama *nombre*.VOZ