

**INSTITUTO POLITÉCNICO NACIONAL**

**CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN**

**CIC-IPN**



*Algoritmos y métodos para el  
reconocimiento de voz en español  
mediante sílabas*

**TESIS**

que para obtener el grado de

*Doctor en Ciencias de la  
Computación*

*Presenta:*

José Luis Oropeza Rodríguez

*Asesor:*

*Dr. Sergio Suárez Guerra*

*México D. F., diciembre de 2005*

# AGRADECIMIENTOS

*Una vez terminado el presente trabajo quisiera agradecer enormemente al Consejo Nacional de Ciencia y Tecnología (CONACyT) el apoyo brindado en el desarrollo del presente trabajo; asimismo al Centro de Investigación en Computación por la labor que realiza en pos de la investigación de este país.*

*De la misma manera quisiera agradecer enormemente al Dr. Sergio Suárez Guerra, director de esta tesis, por el tiempo, espacio y ayuda brindada para que los trabajos realizados hayan llegado a culminarse de manera adecuada. A su vez y muy en especial al Dr. Edgardo Manuel Felipe Riverón por sus acertados consejos con relación a gran parte del presente escrito, al Dr. John Goddard Close por sus pertinentes comentarios en los resultados del presente trabajo. De la misma manera agradecer a todos y cada uno de los restantes miembros del jurado Dr. Oleksiy Pogrebnyak, Dr. Jesús Guillermo Figueroa Nazuno y Dr. Héctor Manuel Pérez Meana, que han tenido a bien por bastante tiempo brindarme los consejos y aclaraciones necesarias para que esto termine de forma adecuada.*

*Finalmente, quisiera agradecer a las personas del laboratorio de Tiempo Real y de Digitales, al M. en C. Ricardo Barrón Fernández, por haberme guiado en los principios del área de reconocimiento de voz, al M. en C. Osvaldo Espinosa Sosa, por sus valiosos consejos y amistad; así como también, al M. en C. Marco Antonio Ramírez Salinas por haberme permitido compartir el lugar de estancia en el Centro de Investigación en Computación.*

# DEDICATORIAS

*En primera instancia dedico el presente trabajo a mi madre Guadalupe Rodríguez Primero, por todo lo que ha significado a lo largo de mi vida, ahora si es el fin, gracias por la paciencia y por todo.*

*A mi abuelita Dominga Primero López con todo mi cariño y respeto.*

## CONTENIDO

---

CONTENIDO	Página
GLOSARIO DE TÉRMINOS	I
GLOSARIO DE ABREVIATURAS	IV
ÍNDICE DE FIGURAS	IX
ÍNDICE DE TABLAS	XI
RESUMEN	XV
ABSTRACT	XVIII
	XIX

CAPÍTULO 1 INTRODUCCIÓN	1
1.1 Motivación	2
1.2 Descripción del problema	3
1.3 Objetivos de la tesis	7
1.4 Organización de la tesis	9
1.5 Resumen del capítulo	9
CAPÍTULO 2 ESTADO DEL ARTE	10
2.1 Antecedentes históricos	11
2.2 Estado del arte	12
2.3 Fundamentos del reconocimiento de voz	20
2.4 Tópicos del reconocimiento de voz	21
2.5 Adecuación del proceso de reconocimiento	22
2.6 Métodos de recuperación del error	25
2.7 Resumen del capítulo	25
CAPÍTULO 3 LA SÍLABA, SU ESTRUCTURA Y SU INMERSIÓN EN LOS SRAH	26
3.1 El paradigma de la sílaba	27
3.2 Las sílabas en el reconocimiento de voz humana	27
3.2.1 Las sílabas como unidades básicas	28
3.2.2 Identificación de la sílaba	30
3.2.3 Identificación de la sílaba (caso práctico)	30
3.2.4 Las sílabas en un acceso léxico	32
3.2.5 Definición de la sílaba	32
3.3 Las reglas de la sílaba	33

3.4	Número de sílabas	37
3.5	Las sílabas en conversaciones de voz	42
3.6	Las sílabas en los SRAH	43
3.7	Resumen del capítulo	45
CAPÍTULO 4 HERRAMIENTAS MATEMÁTICAS		46
4.1	Codificación predictiva lineal	47
4.2	Caracterización de los parámetros de la codificación predictiva lineal	47
4.3	Cálculo de los perfiles espectrales	48
4.4	Algoritmo de Levinson-Durbin	49
4.5	Comparación y semejanza de perfiles espectrales	50
4.6	Detección del vecino más próximo	52
4.7	Cálculo del centroide	53
4.8	Cálculo de la distorsión total	53
4.9	Construcción del libro código	54
4.10	Iteración de Lloyd	54
4.11	Algoritmo de bipartición	55
4.12	Cadenas Ocultas de Markov (COM)	56
4.12.1	Los elementos de una cadena oculta de Markov	57
4.13	Los tres problemas básicos de las COM	58
4.14	Mixturas Gaussianas	62
4.15	Reconocimiento de voz estadístico	64
4.15.1	La aproximación determinística	64
4.15.2	El plano estocástico	65
4.15.3	Simplificaciones al modelo estocástico	67
4.15.4	Simplificaciones de $P(W   \Theta)$	67
4.15.5	Simplificaciones de $P(S   W, \Theta)$	68
4.15.6	Simplificaciones de $P(X   W, S, \Theta)$	68
4.16	Resumen del capítulo	69
CAPÍTULO 5 INTEGRACIÓN DE LA SÍLABA EN LOS SRAH		70
5.1	Implantación del Sistema Experto	71
5.2	Detección de las fronteras de las sílabas y ejemplos de reconocimiento	75
5.3	Resumen del capítulo	86
CAPÍTULO 6 MÉTODO DE RECONOCIMIENTO E INTEGRACIÓN DE LAS SÍLABAS EN SRAH DEL HABLA CONTINUA		88
6.1	Análisis de un algoritmo de reconocimiento de voz del habla discontinua	89

6.1.1	Entrenamiento	89
6.1.2	Reconocimiento	89
6.2	Evaluación de la efectividad	92
6.2.1	Distribución en regiones	92
6.3	El reconocimiento de voz continua usando Mixturas Gaussianas y Cadenas Ocultas de Markov	103
6.3.1	Definición del parámetro ERO	109
6.3.2	Efecto del parámetro ERO en una señal de voz	111
6.3.3	La región de transición energía-parámetro ERO	113
6.3.4	Análisis de un corpus de voz	115
6.3.5	Análisis de perplejidad	118
6.3.6	Resultados obtenidos	121
6.4	Resumen del capítulo	122
CAPÍTULO 7 CONCLUSIONES		124
7.1	Remembranza	125
7.2	Discusión	125
7.2.1	Implicaciones del uso de las sílabas para los SRAH	127
7.3	Contribuciones de la tesis	128
7.4	Trabajos futuros	130
7.5	Reflexiones sobre el futuro de las investigaciones de los SRAH	131
7.6	Conclusión final	132

Referencias bibliográficas

Glosario de términos

Glosario de abreviaturas

Apéndice A. Resultados obtenidos

## ***GLOSARIO DE TÉRMINOS***

**Acentuación.** f. Acción de acentuar, es decir, dar acento prosódico a las palabras: acentuar bien al hablar. Poner el acento ortográfico: acentuar una vocal. FONÉTICA. Colocación del acento ortográfico sobre una vocal: *el examen tenía muchas faltas de acentuación.*

**Acepciones.** f. Sentido en el que se toma una palabra. LING. Cada uno de los significados que puede adquirir una palabra o frase según el contexto: *el verbo "abrir" tiene muchas acepciones.*

**Alófono.** m. FON. Cada una de las variantes o realizaciones fonéticas de un fonema dentro de la cadena hablada, según los sonidos contiguos y su posición en la palabra: la "d" fricativa de "nada" y la "d" oclusiva de "fonda" son alófonos del fonema español /d/.

**Autómata.** m. Máquina que imita los movimientos de un ser animado. INFORM. Dispositivo o conjunto de reglas que realizan un encadenamiento automático y continuo de operaciones capaces de procesar una información de entrada para producir otra de salida.

**Bigram.** Utiliza la información acerca de dos y sólo dos categorías o representaciones anteriores para calcular las probabilidades de secuencia de información.

**Centroide.** m. En geometría, el centroide o baricentro de un objeto perteneciente a un espacio dimensional, es la intersección de todos los hiperplanos que dividen en dos partes de igual cantidad de movimiento con respecto al hiperplano. Informalmente, es el promedio de todos los puntos de un conjunto de vectores o matrices.

**Cóclea.** f. Parte del oído interno con forma de caracol que contiene estructuras celulares delicadas (células ciliadas) vital para el proceso de audición. La cóclea normal cambia los sonidos de vibraciones en pulsos eléctricos que pasan al nervio acústico y de ahí al cerebro.

**Coherencia.** f. Conexión, relación de unas cosas con otras: *no hay coherencia entre lo que dices y lo que haces.*

**Corpus.** m. Conjunto de muestras de voz que son utilizadas para los procesos de entrenamiento y reconocimiento en un SRAH.

**Correlación.** f. Correspondencia o relación recíproca entre dos o más cosas, ideas, personas, etc.: *correlación entre calidad y precio.*

**Dependencia de contexto.** LING. Entorno lingüístico, pragmático y social del que depende el significado de una palabra o un enunciado: [deduje el significado de la palabra por el contexto.](#)

**Entonación.** f. Acción de entonar, es decir, ajustarse al tono al cantar. Dar cierto tono: voz mal entonada. Modulación de la voz que acompaña a la secuencia de sonidos del habla, y que puede reflejar diferencias de sentido, de intención, de emoción y de origen del hablante: [por su entonación me pareció chino.](#)

**Entramado.** m. Arq. Estructura de madera que sirve para establecer una pared. En este caso se refiere a un conjunto de muestras de voz que se separan del resto de la información para su análisis, es decir, la señal de voz se coloca en tramas.

**Etiquetado.** m. Acción o efecto de etiquetar, es decir, colocar un rótulo o inscripción.

**Filtro digital.** m. Un filtro es un sistema que discrimina lo que pasa a su través de acuerdo a algunos parámetros. Los filtros digitales tienen como entrada una señal analógica o digital y a su salida tienen otra señal analógica o digital, pudiendo haber cambiado en amplitud y/o fase dependiendo de las características del filtro.

**Gestionar.** tr. Hacer los trámites o diligencias necesarios para resolver un asunto: [estoy gestionando la venta de varios pisos.](#) Dirigir o administrar una empresa o negocio.

**Gramática.** f. Conjunto de reglas del arte de enseñar a hablar y escribir correctamente.

**Hardware.** m. Conjunto de elementos materiales que componen un ordenador. En dicho conjunto se incluyen los dispositivos electrónicos y electromecánicos, circuitos, cables, tarjetas, armarios o cajas, periféricos de todo tipo y otros elementos físicos.

**Léxico.** adj. Perteneciente al vocabulario de una lengua o región. LING. Vocabulario, conjunto de palabras de una lengua, de una región, de un colectivo, una actividad, etc.: [léxico latino, aragonés, técnico.](#)

**Lineamientos.** m. Delineación de un cuerpo. Orientación, directriz. Conjunto de líneas que forman el dibujo de un cuerpo, por el cual se distingue y conoce su figura.

**Mecanismos de inferencia.** m. Estrategia de solución de un Sistema Experto haciendo uso de elementos de inferencia.



**Modelo.** m. Representación en pequeña escala: modelo de una máquina, estereotipo o elemento con patrones de referencia a seguir. Arquetipo digno de ser imitado que se toma como pauta a seguir: *la antigüedad clásica se convirtió en el modelo artístico del Renacimiento.*

**Modelo oculto de Markov.** m. Es la composición de dos procesos estocásticos, en las que la secuencia de unidades de reconocimiento subyacente y las observaciones acústicas están modeladas como procesos de Markov.

**Modelo oculto de Markov de densidad continua.** m. Define distribuciones de probabilidad de las observaciones en un espacio continuo donde es necesario aplicar ciertas restricciones para limitar la complejidad de los procesos de estimación y cálculo de las probabilidades ahocicadas. La forma más usual caracteriza cada modelo como una mezcla de funciones del mismo tipo (generalmente Gaussianas).

**Monofónico.** m. Emisión de señal acústica por un solo canal o manifestación de un único fonema.

**Muestreo.** f. Resultado obtenido al tomar información en espacios de tiempo equitativos a una señal definida en el dominio del tiempo. En estadística se le conoce como el estudio de la variación de una característica determinada en función de las muestras escogidas para una encuesta. Selección de las personas que se van a someter a una encuesta por medio de un sondeo para obtener un resultado representativo.

**Oído interno.** m. Parte interna del oído humano en donde se encuentra el caracol, que es un órgano capaz de filtrar las frecuencias de forma natural.

**Paradigma.** m. Un paradigma está constituido por los supuestos teóricos generales, las leyes y las técnicas para su aplicación que adoptan los miembros de una determinada comunidad científica.

**Perfil espectral.** m. Determina el estado de resonancia que guarda el tracto vocal del sistema productor de voz en un momento dado.

**Perplejidad.** f. Irresolución, duda. En este caso se refiere a un parámetro de medición de la calidad del uso del modelo del lenguaje bigram.

**Pitch.** m. Tono de frecuencia fundamental emitido por el tracto vocal al emitir una vocal o sonido fónico.

**Premisa.** f. (Del latín praemisa, puesta o colocada delante). Log. Cada una de las dos primeras proposiciones de un silogismo: de las premisas se saca la conclusión. Señal o indicio por el que se deduce o conoce algo.

**Prosodia.** f. Parte de la gramática que estudia las reglas de la pronunciación y acentuación. Conjunto de las reglas relativas a la cantidad de las vocales. GRAM. Parte de la gramática que enseña la correcta pronunciación y acentuación: la prosodia da las normas para lograr una cuidada pronunciación. LING. Parte de la fonología dedicada al estudio de los rasgos fónicos que afectan a unidades inferiores o superiores al fonema: la prosodia se encarga del estudio de la entonación.

**Psicoacústica.** f. La audición humana es un proceso extraordinariamente complejo, que apenas está comenzando cuando el sonido golpea el tímpano y es convertido de variaciones en la presión del aire a impulsos nerviosos. De ahí en adelante, es asunto de la mente, y la psicología se convierte en factor importante para estudiar y analizar los sonidos, así como las reacciones de las personas ante éstos.

La psicoacústica puede ser definida simplemente como el estudio psicológico de la audición. El objetivo de la investigación psicoacústica es averiguar cómo funciona la audición. En otras palabras, el objetivo es descubrir cómo los sonidos que entran al oído son procesados por éste y el cerebro, con el fin de dar a la persona que escucha información útil acerca del mundo exterior.

**Psicolingüística.** f. Ciencia que estudia el lenguaje y la expresión verbal en relación con los mecanismos psicológicos que la hacen posible: la psicolingüística permite ver los procesos de aprendizaje de una lengua.

**Redes neuronales.** f. Referidas habitualmente de forma más sencilla como redes de neuronas o redes neuronales, las redes de neuronas artificiales (RNA) son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Consiste en simular las propiedades observadas en los sistemas neuronales biológicos a través de modelos matemáticos recreados mediante mecanismos artificiales (como un circuito integrado, un ordenador o un conjunto de válvulas).

**Sílaba.** f. Sonido o conjunto de sonidos articulados que constituyen un solo núcleo sónico entre dos depresiones sucesivas de la emisión de la voz. (Diccionario Real Academia Española). Sonido o sonidos articulados que constituyen un solo núcleo fónico entre dos depresiones sucesivas de la emisión de voz: **las palabras esdrújulas tienen tres o más sílabas.**

**Síntesis.** f. Método que procede de lo simple a lo compuesto, de los elementos al todo, de la causa a los efectos, del principio a las consecuencias. Para este caso, se refiere a la capacidad de realizar síntesis de voz, es decir, a partir de texto producir una señal acústica por un medio computacional que resulta entendible.

**Software.** m. Es un conjunto de programas, documentos, procedimientos, y rutinas asociados con la operación de un sistema de cómputo. Distinguiéndose de los componentes físicos llamados hardware.

**Tramas.** f. Conjunto continuo de elementos que se siguen uno tras otro bajo una secuencia lógica. Conjunto de líneas que integran la imagen.

**Traslapar.** m. Cubrir una cosa a otra de un modo más o menos completo.

**Umbral.** m. Del latín umbratilis, que está a la sombra. Entrada o paso de alguna cosa, en este caso se refiere a rebasar una determinada cantidad numérica establecida como referencia o valor útil de información. Valor a partir del cual empiezan a ser perceptibles los efectos de un agente físico: [umbral luminoso](#).

## **GLOSARIO DE ABREVIATURAS**

ASR	Automatic Speech Recognition
BD-PUBLIC	Base de datos PUBLIC
Bell Labs.	hoy AT&T
BPF	Band Pass Filter
CDHMM	Continuous Density Hidden Markov Models
CLPCs	Cepstrum Linear Predictive Coefficients
CMU	Carnegie Mellon University
COMDC	Cadenas Ocultas de Markov de Densidad Continua
COM	Cadenas Ocultas de Markov
CPL	Codificación Predictiva Lineal
DARPA	Agencia de Investigación de Proyectos Avanzados de Defensa
DTW	Dynamic Time Warping
EE. UU.	Estados Unidos de América
EM	Estimation Maximum
ERO	Energía del parámetro RO
ESPRESSO	Proyecto de investigación usando sílabas
FDPC	Función de Densidad de Probabilidad Continua
FFT	Fast Fourier Transform
FIR	Filter Impulse Response
HAL 9000	Heuristically programmed ALgorithmic computer
HMM	Hidden Markov Models
HMM-NN	Hidden Markov Models-Neural Networks
Hz	Hertz
IBM	International Business Machinery
log	Logaritmo
LPC	Linear Predictive Coding
LPF	Low Pass Filter
ms	milisegundos
NEC labs	Nippon Electric Corporation Laboratories
OGI	Oregon Graduate Institute
PhonDatII	Base de datos PhontDatII
PLP	Perceptual Linear Prediction
PLP-cepstrales	Perceptual Linear Predictive
pdf	probability density function
RAH	Reconocimiento Automático del Habla
RCA Labs	The Sarnoff Corporation is the successor organization to the David Sarnoff Research Center and the RCA Laboratories in Princeton, New Jersey. General Order S-56 of the Radio Corporation of America, issued March 5, 1941, mandated that "All research, original development and patent and licensing

	activities of the Corporation and its Associated Companies will be consolidated in RCA Laboratories which will be responsible for all such work in the future."
RO	Parámetro RO, filtro digital con frecuencias de corte por encima de los 4 KHz., utilizado para analizar los cruces por cero en tales frecuencias.
SBC	Sistema Basado en Conocimiento
SRAH	Sistema de Reconocimiento Automático del Habla
SPHINX	speaker-independent large vocabulary continuous speech recognizer under Berkeley's style license. It is also a collection of open source tools and resources that allows researchers and developers to build speech recognition system.
STTEF	Short Term Total Energy Function
SWITCHBOARD	Public Domain Large Vocabulary Continuous Speech Recognition Software
TIMIT	corpus of read speech
TRF	Transformada Rápida de Fourier
Verbmobil	project at Institut für Maschinelle Sprachverarbeitung
VoiceXML	Lenguaje de Marcado eXtendido para voz
WER	Word Error Rate

# ÍNDICE DE FIGURAS

	Página
<b>CAPÍTULO 1 INTRODUCCIÓN</b>	
Fig. 1.1 Adquisición de información por medio de una tarjeta de sonido	4
Fig. 1.2 Diferentes herramientas ocupadas en el reconocimiento de voz	5
Fig. 1.3 Elementos fonéticos usados en el reconocimiento de voz	6
<b>CAPÍTULO 2 ESTADO DEL ARTE</b>	
Fig. 2.1 Implantación de bancos de filtros del tipo Cepstral	20
<b>CAPÍTULO 3 LA SÍLABA, SU ESTRUCTURA Y SU INMERSIÓN EN SISTEMAS AUTOMÁTICOS DE RECONOCIMIENTO DEL HABLA (SRAH)</b>	
Fig. 3.1 Gráfica de tiempos de duración de sílabas en un corpus de voz	30
Fig. 3.2 Gráfica de tiempos de duración de sílabas en un corpus de voz (continuación)	31
Fig. 3.3 Gráfica de tiempos de duración de las sílabas en un corpus de voz (continuación)	31
Fig. 3.4 Gráfica del comportamiento de la señal de voz y su energía	33
Fig. 3.5 Esquema de clasificación de letras del alfabeto del español	34
Fig. 3.6 Grupo de consonantes inseparables	35
Fig. 3.7 Gráfica del estudio de las estructuras silábicas en textos científicos	40
Fig. 3.8 Análisis de estructuras silábicas en un locutor del Latino-40	41
Fig. 3.9 Gráfica del comportamiento de las sílabas en general	42
Fig. 3.10 Gráfica de Cadenas Ocultas de Markov aplicadas a sílabas y fonemas	43
Fig. 3.11 Gráfica del sistema de reconocimiento de voz aislado usando sílabas	44
<b>CAPÍTULO 4 HERRAMIENTAS MATEMÁTICAS</b>	
Fig. 4.1 El modelo de la urna y la pelota	56
Fig. 4.2 Paso de inducción para encontrar una optimización de los parámetros de la Cadena Oculta de Markov	62
Fig. 4.3 Representación esquemática de 3 Gaussianas en una gráfica	63
Fig. 4.4 Ejemplo de las Mixturas Gaussianas en una Cadena de Markov	64

concatenada	
Fig. 4.5 Esquema de funcionamiento de un simple sistema de reconocimiento de voz	65
CAPÍTULO 5 INTEGRACIÓN DE LAS SÍLABAS EN LOS SRAH	
Fig. 5.1 Diagrama a bloques del sistema de reconocimiento propuesto, haciendo uso de un Sistema Experto	71
Fig. 5.2 Árboles de inferencia del Sistema Experto	74
Fig. 5.3 Gráficas de energía y espectrograma de sílabas CV	76
Fig. 5.4 Gráficas de energía y espectrograma de diferentes estructuras de sílabas	77
Fig. 5.5 Gráficas de sílabas del corpus de dígitos (sílabas u {V}, sílabas tres {CCVC})	79
Fig. 5.6 Gráficas de sílabas segmentadas del corpus de dígitos	80
Fig. 5.7 Gráficas del tiempo de duración de las sílabas del corpus de dígitos	81
Fig. 5.8 Gráficas del promedio de duración de las sílabas del corpus de dígitos	81
Fig. 5.9 Gráficas de segmentación en unidades silábicas de una palabra de forma manual	83
Fig. 5.10 Gráficas del proceso de segmentación y cálculo de la energía del corpus de dígitos	83
Fig. 5.11 Gráficas de la segmentación silábica de las sílabas del corpus de dígitos	84
CAPÍTULO 6 MÉTODO DE RECONOCIMIENTO E INTEGRACIÓN DE LAS SÍLABAS A LOS SRAH DEL HABLA CONTINUA	
Fig. 6.1 Gráficas de señal de voz y energía de palabra 'negro'	90
Fig. 6.2 Gráficas de las sílabas 'ne'-'gro'	90
Fig. 6.3 Caja de diálogo de vectores de autocorrelación	91
Fig. 6.4 Caja de diálogo de coeficientes LPC	91
Fig. 6.5 Datos resultantes de un modelo oculto de Markov optimizado	92
Fig. 6.6 Gráfica de la Distribución en regiones de la sílaba "tres" para 4 Regiones	93
Fig. 6.7 Gráfica de la Distribución en regiones de la sílaba "tres" para 16 Regiones	93
Fig. 6.8 Gráfica de distribución de las regiones de un corpus experimental	94
Fig. 6.9 Diagrama de Errores de optimización para 32 regiones con 200 muestras usando sílabas	95
Fig. 6.10 Diagrama de Errores de optimización para 64 regiones con 200 muestras usando sílabas	95

<i>Fig. 6.11</i> Diagrama de Errores de optimización para 128 regiones con 200 muestras usando sílabas	96
<i>Fig. 6.12</i> Diagrama de Errores de optimización para 32 regiones con 400 muestras usando sílabas	97
<i>Fig. 6.13</i> Diagrama de Errores de optimización para 64 regiones con 400 muestras usando sílabas	97
<i>Fig. 6.14</i> Diagrama de Errores de optimización para 128 regiones con 400 muestras usando sílabas	98
<i>Fig. 6.15</i> Diagrama de Errores de optimización para 32 regiones con 200 muestras usando palabras	98
<i>Fig. 6.16</i> Diagrama de Errores de optimización para 64 regiones con 200 muestras usando palabras	99
<i>Fig. 6.17</i> Diagrama de Errores de optimización para 128 regiones con 200 muestras usando palabras	99
<i>Fig. 6.18</i> Diagrama de Errores de optimización para 32 regiones con 400 muestras usando palabras	100
<i>Fig. 6.19</i> Diagrama de Errores de optimización para 64 regiones con 400 muestras usando palabras	100
<i>Fig. 6.20</i> Diagrama de errores de optimización para 128 regiones con 400 muestras usando palabras	101
<i>Fig. 6.21</i> Representación esquemática de Mixturas Gaussianas por cada estado de la Cadena Oculta de Markov	104
<i>Fig. 6.22</i> Esquematización de las Mixturas Gaussianas por Modelo Oculto de Markov después de la segunda iteración del algoritmo propuesto	104
<i>Fig. 6.23</i> Esquematización en el dominio del tiempo de la palabra 'cero'	105
<i>Fig. 6.24</i> Esquematización en el dominio del tiempo de la palabra 'cero' y su energía correspondiente	105
<i>Fig. 6.25</i> Esquematización en el dominio del tiempo y de la frecuencia de la palabra 'cero'	106
<i>Fig. 6.26</i> Gráfica de reconocimiento del corpus de dígitos usando segmentación silábica y modelos ocultos de Markov	109
<i>Fig. 6.27</i> Esquematización del análisis lingüístico del corpus de dígitos	107
<i>Fig. 6.28</i> Esquematización de un filtro digital y la señal en el dominio del tiempo ya filtrada	111
<i>Fig. 6.29 (a)</i> Esquematización de las señales 'cero' y 'tres' antes y después de haberles aplicado un filtro digital	111
<i>Fig. 6.29 (b)</i> Esquematización de las señales 'cinco', 'tres' y 'diez' antes y después de haberles aplicado el filtro digital	112
<i>Fig. 6.30</i> Esquematización de las regiones de transición energía-parámetro RO para el caso de la palabra 'ce'-'ro'	113
<i>Fig. 6.31</i> Gráficas comparativas del reconocimiento haciendo uso de palabras completas y concatenación de sílabas del corpus de dígitos	114
<i>Fig. 6.32</i> Segundo corpus final de prueba	119



<i>Fig. 6.33</i> Probabilidades de transición del segundo corpus final de prueba	119
<i>Fig. 6.34</i> Probabilidades de transición de las sílabas del corpus de prueba que se utilizan para inicializar las probabilidades de transición de los Modelos Ocultos de Markov globales	121

# ÍNDICE DE TABLAS

	Página
<b>CAPÍTULO 3 LA SÍLABA, SU ESTRUCTURA Y SU INMERSIÓN EN SISTEMAS AUTOMÁTICOS DE RECONOCIMIENTO DEL HABLA (SRAH)</b>	
Tabla 3.1 Notación utilizada para expresar las reglas de división silábica	34
Tabla 3.2 Frecuencia de aparición de fonemas en textos en español	37
Tabla 3.3 Frecuencia de pronunciación de las sílabas	39
<b>CAPÍTULO 5 INTEGRACIÓN DE LAS SÍLABAS EN LOS SRAH</b>	
Tabla 5.1 Porcentaje de efectividad en la aplicación del Sistema Experto a diferentes corpus de voces	74
Tabla 5.2 Índices de reconocimiento para determinados tipos de sílabas en el español	77
Tabla 5.3 Estructura silábica para un corpus de dígitos	80
Tabla 5.4 Porcentajes de reconocimiento para un corpus de dígitos	82
Tabla 5.5 Valores obtenidos de una señal de voz y proceso de segmentación en sílabas de forma automática	83
Tabla 5.6 Tabla de confusión para el caso del corpus de dígitos	84
Tabla 5.7 Porcentaje de reconocimiento usando una segmentación basada en energía	84
Tabla 5.8 Frecuencia de aparición de palabras con N sílabas en el diccionario y corpus del Latino40	85
Tabla 5.9 Frecuencia de aparición de 10 monosílabas más usadas dentro del Latino40	86
Tabla 5.10 Análisis a nivel de sílabas del corpus de voz Latino40	86
<b>CAPÍTULO 6 MÉTODO DE RECONOCIMIENTO E INTEGRACIÓN DE LAS SÍLABAS A LOS SRAH DEL HABLA CONTINUA</b>	
Tabla 6.1 Reconocimiento de palabras de un diccionario experimental	102
Tabla 6.2 Reconocimiento de palabras con las que no fue entrenado el sistema	103
Tabla 6.3 Estructuras silábicas de un corpus de dígitos	107
Tabla 6.4 Tabla de confusión para el caso del corpus de dígitos	108
Tabla 6.5 Tabla de confusión para el caso del corpus de dígitos utilizando sílabas concatenadas y STTEF	108

Tabla 6.6 Tablas de confusión para el caso del corpus de dígitos usando regiones de transición energía-parámetro RO, sílabas independientes y sílabas concatenadas	114
Tabla 6.7 División de un corpus experimental en sílabas	115
Tabla 6.8 Análisis del número de palabras que conforman al corpus de prueba	116
Tabla 6.9 Matriz de transición de palabras en el corpus analizado	117
Tabla 6.10 Valores logarítmicos del corpus de prueba	118
Tabla 6.11 Valores logarítmicos del conjunto de prueba final	120
Tabla 6.12 Porcentajes de reconocimiento usando Cadenas Ocultas de Markov con 3 y 5 estados respectivamente para el habla discontinua del corpus final "1"	121
Tabla 6.13 Porcentajes de reconocimiento usando Cadenas Ocultas de Markov con 3 y 5 estados respectivamente para el habla continua del corpus final "1"	121
Tabla 6.14 Porcentajes de reconocimiento usando Cadenas Ocultas de Markov con 3 y 5 estados respectivamente para el habla discontinua del corpus final "2"	122
Tabla 6.15 Porcentajes de reconocimiento usando Cadenas Ocultas de Markov con 3 y 5 estados respectivamente para el habla continua del corpus final "2"	122
APÉNDICE A RESULTADOS OBTENIDOS	
Tabla A.1 Ejemplos de coeficientes LPC para los experimentos de las sílabas sa, se, si, so y su	A2
Tabla A.2 Ejemplos de coeficientes LPC para la sílaba si en un corpus de sílabas sa, se, si, so y su	A2
Tabla A.3 Ejemplos de coeficientes LPC para la sílaba so en un corpus de sílabas sa, se, si, so y su	A2
Tabla A.4 Ejemplos de coeficientes LPC para la sílaba su en un corpus de sílabas sa, se, si so y su	A2
Tabla A.5 Algunos datos representativos del libro código global	A2
Tabla A.6 Probabilidades de transición de estados para el modelo U-NO	A2
Tabla A.7 Probabilidades de transición de estados para el modelo DOS	A3
Tabla A.8 Probabilidades de transición de estados para el modelo TRES	A3
Tabla A.9 Probabilidades de transición de estados para el modelo CUATRO	A3
Tabla A.10 Probabilidades de transición de estados para el modelo CIN-CO	A3
Tabla A.11 Probabilidades de transición de estados para el modelo SEIS	A3
Tabla A.12 Probabilidades de transición de estados para el modelo SIE-TE	A3
Tabla A.13 Probabilidades de transición de estados para el modelo O-CHO	A3
Tabla A.14 Probabilidades de transición de la matriz B para el modelo O-	A6

CHO	
Tabla A.15 Tabla de reconocimiento para el modelo O-CHO	A6
Tabla A.16 Análisis de la duración de las palabras del diccionario	A8
Tabla A.17 Análisis de duración de las sílabas que componen el diccionario	A9
Tabla A.18 Distribución de vectores en Regiones	A12
Tabla A.19 Distribución en Regiones de la sílaba "tres"	A13
Tabla A.20 Errores por optimización para 32 regiones con 200 muestras para sílabas	A15
Tabla A.21 Errores por optimización para 64 regiones con 200 muestras para sílabas	A15
Tabla A.22 Errores por optimización para 128 regiones con 200 muestras para sílabas	A16
Tabla A.23 Errores por optimización para 32 regiones con 400 muestras para sílabas	A16
Tabla A.24 Errores por optimización para 64 regiones con 400 muestras para sílabas	A16
Tabla A.25 Errores por optimización para 128 regiones con 400 muestras para sílabas	A16
Tabla A.26 Errores por optimización para 32 regiones con 200 muestras para palabras	A17
Tabla A.27 Errores por optimización para 64 regiones con 200 muestras para palabras	A17
Tabla A.28 Errores por optimización para 128 regiones con 200 muestras para palabras	A17
Tabla A.29 Errores por optimización para 32 regiones con 400 muestras para palabras	A18
Tabla A.30 Errores por optimización para 64 regiones con 400 muestras para palabras	A18
Tabla A.31 Errores por optimización para 128 regiones con 400 muestras para palabras	A18

# RESUMEN

Esta tesis muestra los resultados de la incorporación de las unidades silábicas en sistemas de reconocimiento de voz para el idioma español. Actualmente el uso de los fonemas representa varias dificultades debido a que la identificación de las fronteras entre ellos por lo regular es difícil de encontrar en representaciones acústicas de voz, dado lo anterior, el empleo de los fonemas no ha alcanzado los resultados que se desean.

Durante los experimentos realizados fueron examinados para la tarea de segmentación tres elementos esenciales: a) la Función de Energía Total en Corto Tiempo, b) la Función de Energía de altas frecuencias Cepstrales (conocida como Energía del parámetro RO), y c) un Sistema Basado en Conocimiento; todas las cuales representan el conjunto de las aportaciones esenciales de la presente tesis. Al ser aplicados en corpus continuos y discontinuos de voces creados en laboratorio, mostraron buenos resultados.

Tanto el Sistema Basado en Conocimiento y la Función de Energía Total en Corto Tiempo fueron usados en un corpus de dígitos en donde los resultados alcanzados usando sólo la Función de Energía Total en Corto Tiempo, fueron de 90.58%. Cuando se utilizaron los parámetros Función de Energía Total en Corto Tiempo y la Energía del parámetro RO se obtuvo un 94.70% de razón de reconocimiento. Lo cual causa un incremento del 5% con relación al uso de palabras completas en un corpus de voz dependiente de contexto.

Por otro lado, cuando se utilizó un corpus de laboratorio del habla continua al usar la Función de Energía Total en Corto Tiempo y el Sistema Basado en Conocimiento, se alcanzó un 78.5% de razón de reconocimiento y un 80.5% de reconocimiento al usar los tres parámetros anteriores. El modelo del lenguaje utilizado para este caso fue el bigram y se utilizaron Cadenas Ocultas de Markov de densidad continua con tres y cinco estados, con 3 mixturas Gaussianas por estado.

La inclusión de un número mayor de filtros digitales y técnicas de Inteligencia Artificial en la etapa de entrenamiento y reconocimiento, respectivamente, incrementaron de forma significativa los resultados obtenidos. Esta investigación demostró el potencial del paradigma de la unidad silábica en sistemas de reconocimiento de voz tanto del habla continua como discontinua. Finalmente, se crearon las reglas de inferencia dentro del Sistema Basado en Conocimiento tomando como base la segmentación en sílabas de palabras del idioma español.

# ABSTRACT

This thesis examines the results of incorporating into Automatic Speech Recognition the syllable units for the Spanish language. Because of the boundaries between phonemes-like units its often difficult to elicit them; the use of these has not reached a good performance in Automatic Speech Recognition.

In the course of the developing the experiments three approaches for the segmentation task were examined: a) the using of the Short Term Total Energy Function, b) the Energy Function of the Cepstral High Frequency (named ERO parameter), and c) a Knowledge Based System. They represent the most important contributions of this work; they showed good results for the Continuous and Discontinuous speech corpus developed in laboratory.

The Knowledge Based System and Short Term Total Energy Function were used in a digit corpus where the results achieved using Short Term Total Energy Function alone reached 90.58% recognition rate. When Short Term Total Energy Function and RO parameters were used a 94.70% recognition rate was achieved.

Otherwise, in the continuous speech corpus created in the laboratory the results achieved a 78.5% recognition rate using Short Term Total Energy Function and Knowledge Based System, and 80.5% recognition rate using the three approaches mentioned above. The bigram model language and Continuous Density Hidden Markov Models with three and five states incorporating three Gaussian Mixtures for state were implemented.

By further including a major number of digital filters and Artificial Intelligent techniques in the training and recognition stages respectively the results can be improved even more. This research showed the potential of the syllabic unit paradigm for the Automatic Speech Recognition for the Spanish language. Finally, the inference rules in the Knowledge Based System associated with rules for splitting words in syllables in the cited language were created.

# CAPÍTULO 1

---

## Introducción

En este capítulo se presentan las diferentes premisas que permiten incorporar en la tarea de reconocimiento de voz, el uso de las sílabas. Sobre la base de estos lineamientos, se plantean los objetivos que persigue la presente investigación, los cuales tienen como parte fundamental indagar la importancia del uso de la sílaba como elemento de reconocimiento en el español, el que como es sabido, es un idioma altamente dependiente del contexto.

Para la realización de la tarea de reconocimiento, se hace uso de las técnicas que hasta el día de hoy tienen mayor repercusión en ésta; las Cadenas Ocultas de Markov, tanto para corpus del habla continua y discontinua.

Una vez establecidos los elementos anteriores, se procede a bosquejar el contenido del presente documento.

## **1.1 MOTIVACIÓN**

Como es conocido, la tarea de investigación reúne una gran cantidad de actividades que permiten complementar su desarrollo y el logro de los objetivos que se intentan alcanzar con la misma. Dentro del área del Reconocimiento Automático del Habla (RAH o ASR – Automatic Speech Recognition - por sus siglas en inglés), la tarea se incrementa en gran medida debido al uso de diferentes parámetros a considerar dentro de ella (fluctuaciones de la voz, el medio ambiente, los medios de captación, etcétera).

El estudio del reconocimiento de voz data ya desde un poco más de 50 años; sin embargo, los resultados obtenidos distan aún, de ser los deseados. Por un momento y debido a sus características, se pensó que el fonema sería el parámetro de apoyo sobre el cual todos los problemas de reconocimiento recaerían; pero al paso del tiempo se ha observado que no es así, y de hecho, la inclusión de nuevos parámetros de estudio se hace más que necesario, indispensable.

Dichas consecuencias que se analizan dentro del presente documento, han dado cabida a generar vertientes de investigación totalmente nuevas para este campo. Para el caso específico de este trabajo, la sílaba es un parámetro de gran relevancia por las cuestiones que le son inherentes.

Los SRAH (Sistemas de Reconocimiento Automáticos del Habla) tienen hoy en día una vasta aplicación en la industria, el hogar, la oficina, etc. Su advenimiento es apenas el comienzo de una nueva era donde la tecnología forma parte de nuestras vidas. Esto quizás no sea raro, pues al fin y al cabo la señal de voz es parte del ser humano, se nace y se vive con ella, aunque es una de las últimas manifestaciones del individuo en su capacidad de acoplamiento con el entorno. Muchos de los elementos que permiten una vida estable al ser humano son gracias al equilibrio que logra a lo largo de su edad temprana, en lo que se conoce como etapa de sondeo, que se manifiesta en la coordinación entre los sonidos emitidos por el tracto vocal y los que percibe el oído interno.

A lo largo del tiempo estas características se van modificando gracias en parte a la región, costumbres y lugar de residencia del individuo. Con lo que la entonación y acentuación de lo que se pronuncia cambia, y es diferente para los individuos no sólo a lo largo del mundo, sino también, a lo largo de una región como un país.

Sin embargo, la maquinaria humana es tan fascinante y tiene un alto rendimiento en cuestiones de reconocimiento, que incluso en las situaciones más adversas logra realizar un reconocimiento altamente confiable. Además es capaz de modificar sus estructuras internas con el fin de realizar un acoplamiento en tiempo real, del sentido de las frases y de los elementos que conllevan todo tipo de



información. Esto es factible siempre y cuando la relación señal a ruido sea lo menos significativa posible, o bien la atenuación de la señal de información no sea demasiado grande.

Para el caso específico del reconocimiento de voz por computadora, esto no es una realidad por desgracia. El crear un sistema de reconocimiento para la computadora aún dista mucho de ser el sistema perfecto que posee el ser humano, sin embargo, se pretende aproximarlos en gran medida.

El presente trabajo plantea una alternativa a la forma en la que el reconocimiento de voz se ha estado implementando desde hace ya bastante tiempo, analizando la forma en la cual el paradigma de la sílaba responde a tal labor dentro del español. A su vez, da lugar a ser un eslabón más en el campo del desarrollo tecnológico del reconocimiento de voz, al considerar a las sílabas como unidades elementales de reconocimiento.

La inmersión del estudio basado en sílabas o de cualquier otra unidad básica pretende por un lado, incrementar los índices de reconocimiento que actualmente se tienen dentro de los sistemas actuales, y por otro, analizar el resultado de incrustar esta unidad para este caso al idioma español.

Tras un análisis que se efectúa sobre los componentes del problema del reconocimiento basado en sílabas, se anexan las tareas realizadas en pos de alcanzar los objetivos del trabajo. Se aplica como técnica principal de reconocimiento las Cadenas Ocultas de Markov, por ser un algoritmo altamente empleado en la actualidad dentro de los sistemas de Reconocimiento.

## **1.2 DESCRIPCIÓN DEL PROBLEMA**

Un SRAH es aquel sistema automático que es capaz de gestionar la señal de voz emitida por un individuo. Dicha señal ha sido pasada por un proceso de digitalización para obtener elementos de medición (muestras), las cuales permiten denotar su comportamiento e implementar procesos de tratamiento de la señal, enfocados al reconocimiento.

Bajo este esquema, la señal de voz se ve inmersa en dos bloques importantes: entrenamiento y reconocimiento. Una forma gráfica de representar dicho proceso se muestra en la figura 1.1:

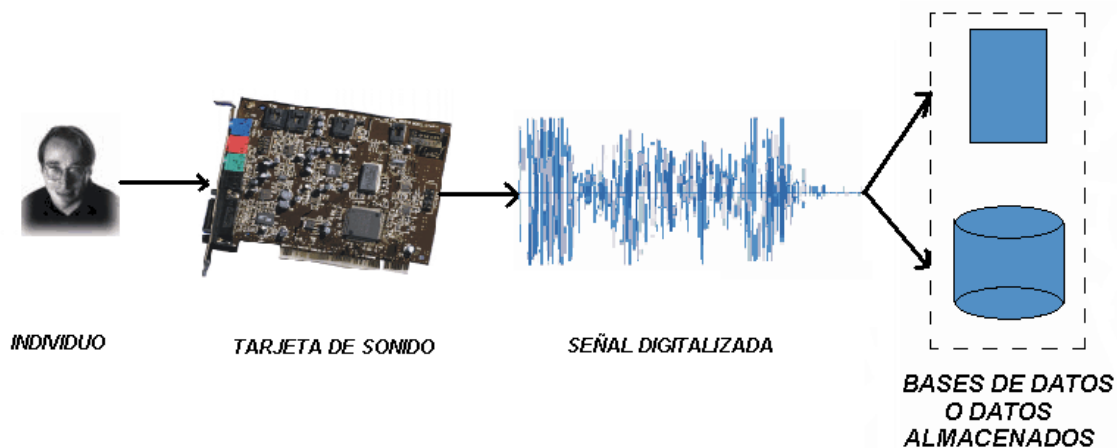


Fig. 1.1. Adquisición de información por medio de una tarjeta de sonido.

La cantidad de muestras depende de la configuración de la tarjeta de adquisición de datos que se esté usando. Una vez que se obtiene la señal, se utiliza para llevar a cabo la etapa de entrenamiento. Los micrófonos del mercado poseen en la actualidad alta eficiencia en la captura de información.

Dicho entrenamiento, es una de las etapas más críticas dentro de estos sistemas y gran parte del éxito de un sistema de reconocimiento de voz recae en esta etapa. Como un referente esencial de lo anterior, el presente trabajo presenta la incrustación de un bloque destinado al refuerzo de la obtención de los datos a ser procesados que hace uso de un Sistema Basado en Conocimiento (SBC al cual se le denominará también Sistema Experto), capaz de realizar la clasificación de la señal de entrada en unidades silábicas, por medio de la aplicación de un conjunto de reglas lingüísticas que prevalecen en el español.

La razón por la cual se pensó en un Sistema Basado en Conocimiento es debido a que en el español en contraparte del inglés por ejemplo, la forma en la que se escriben los textos y la que se lee no dista mucho de ser semejante. Esto es debido a que el español es altamente dependiente del contexto y de la prosodia. Los elementos anteriores justifican la aplicación del experto en esta parte del sistema.

La etapa de entrenamiento puede llevarse a cabo por varios métodos, dentro de los cuales destacan:

- ❖ Bancos de Filtros.
- ❖ Codificación Predictiva Lineal.
- ❖ Modelos Ocultos de Markov.
- ❖ Redes Neuronales Artificiales.
- ❖ Lógica Difusa.
- ❖ Sistemas de reconocimiento híbrido, etc.

Los algoritmos y las unidades fonéticas usadas se manifiestan como los recursos que dan cabida a soluciones expresadas para este campo. En el presente trabajo la sílaba y los modelos ocultos de Markov son la materia prima a usarse para alcanzar los objetivos que se plantean. La figura 1.2 muestra algunas de las técnicas empleadas en sistemas de reconocimiento del habla (Savage, 1995) y (Kershaw, 1996).

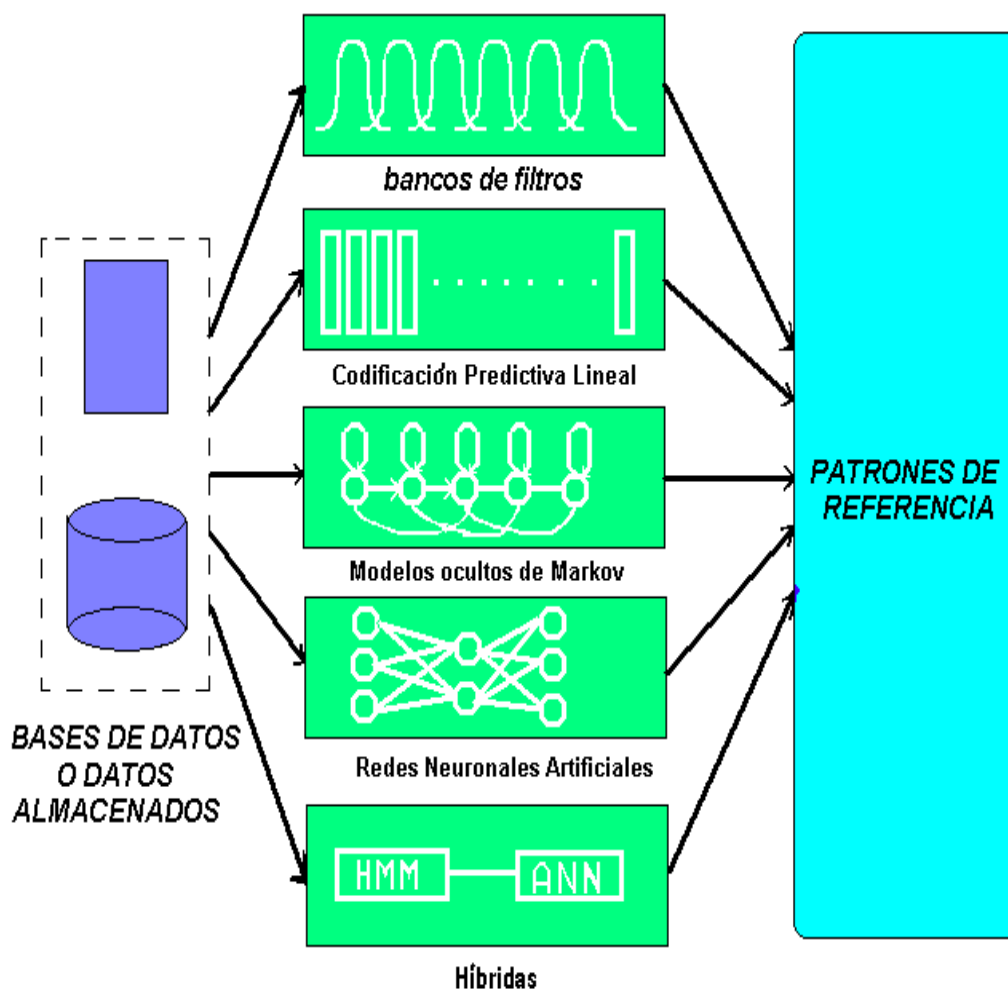
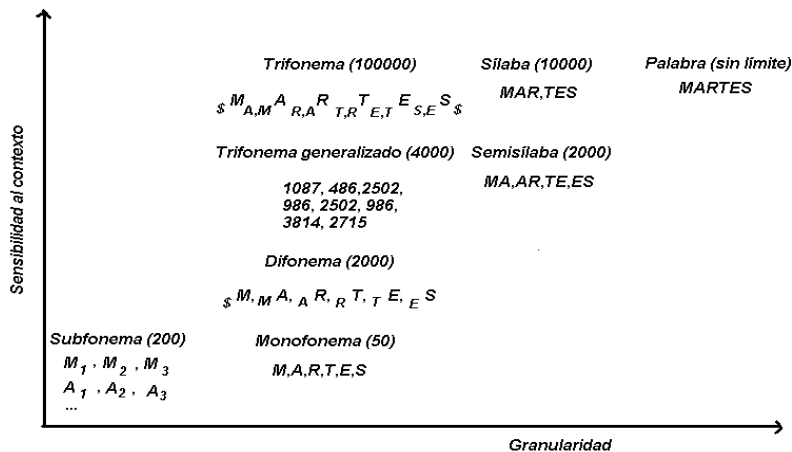


Fig. 1.2. Diferentes herramientas ocupadas en el reconocimiento de voz.

En los sistemas de reconocimiento de voz se ha usado la estructura fonética como elemento primordial. La figura 1.3 muestra algunas de las distintas estructuras que se han utilizado además del fonema (Wu, 1998).



Granularidad vs. Sensibilidad al contexto, para la palabra MARTES.

Fig. 1.3. Elementos fonéticos usados en el reconocimiento de voz.

Como se observa, la señal de voz ha sido analizada desde distintos puntos de vista. Se han hecho análisis desde fonemas, hasta la palabra misma. Esto ha dado como origen una gran cantidad de resultados e implementación de algunas técnicas relacionadas. El presente trabajo se enfoca al área de la sílaba y se analiza su alta sensibilidad al contexto, tal y como se muestra en la figura 1.3.

Una vez que un sistema de reconocimiento ha pasado por la etapa de entrenamiento y se han creado los patrones respectivos, las aplicaciones pueden ser vastas. Desde el acceso a bases de datos en INTERNET (Tu and Loizou, 1999) hasta la apertura o cierre de un elemento electrónico.

Dentro de todo el conjunto de lenguajes existentes en el mundo, se pueden encontrar una gran variedad de estructuras fonético-lingüísticas que los caracterizan. Las manifestaciones de los lenguajes basados en sílabas se ponen de manifiesto en el Oriente donde países como China, Japón, Hong Kong, etc., poseen una estructura alfabética apegada en gran medida a las sílabas.

Cabe mencionar algunas lenguas propias de aquellas regiones Cantones (Zhang 1999), (Chang 2000) y (Peskin et al. 1991), Mandarín (Chang 2000), etc. Como se mencionó con anterioridad, la sílaba para el presente trabajo es considerada como el medio que tiene mayor peso en el aprendizaje del español. Esto es debido a las siguientes características:

- a) En una señal de voz, la sílaba es una estructura que es independiente para cualquier idioma que se ponga de ejemplo, pues no es posible encontrar errores de coarticulación en su estructura interna como sucede en el caso de los fonemas.

Considere la sílaba {p/a} de las palabras plazo y plato, si no se realiza una división de sus elementos fonéticos y se estudian sus características, se concluye que es exactamente igual en cualquier caso, aunque se use en dos palabras distintas.

Ahora bien, considere al fonema /f/ de las palabras foca, y fofa, en el primer y segundo caso al fonema /f/ le prosiguen dos fonemas totalmente diferentes /o/ y /a/ de las sílabas {fo} y {fa}. Aquí, el problema es que el fonema pierde sus características propias al tener adyacentes dos fonemas totalmente diferentes.

A esto se le conoce como el problema de la coarticulación y es la fuente de las grandes dificultades que manifiestan los sistemas de reconocimiento actuales.

b) La sílaba en el caso del español al contener cierta semejanza a la forma en que se pronuncia con la que se escribe, puede establecerse como elemento primordial de un SRAH.

c) La separación automática de estructuras fonéticas, sigue siendo aún hasta nuestros días, un problema que no se ha podido resolver. Tal es el caso de que un sistema de reconocimiento de voz que basa sus principios en este esquema, tiene que realizarla en ocasiones, de manera semiautomática para poder incrementar las tasas de reconocimiento.

Obviamente esto no quiere decir que en la sílaba no pueda suceder, pero por sus propias características intrínsecas pueden permitir una mejora en los esquemas de segmentación.

Las razones expuestas en los incisos anteriores y las que se presentan en (Wu 1998) son un punto de apoyo para la elaboración del presente trabajo.

### **1.3 OBJETIVOS DE LA TESIS**

#### **OBJETIVO GENERAL**

Desarrollar el estudio y análisis del efecto que tiene introducir unidades silábicas en esquemas de reconocimiento de voz dentro del español, apoyándose en la técnica de las Cadenas Ocultas de Markov, con el fin de lograr incrementar los índices de reconocimiento que prevalecen en la actualidad y analizar su comportamiento.

#### **OBJETIVOS PARTICULARES**

Para lograr alcanzar el objetivo antes señalado se plantean las siguientes actividades a alcanzar:

- Realizar un estudio de las características que tienen las sílabas dentro del español, analizando su estructura interna, sus elementos y la relación que guarda con otros idiomas, donde la sílaba resulta bastante útil en esquemas de reconocimiento ya implantados.

Con el fin de entender el comportamiento de las estructuras silábicas que conforman al idioma en cuestión y, basándose en ello, crear las herramientas necesarias que permitan contener dichas características.

- Demostrar la factibilidad que tiene la sílaba como elemento básico de reconocimiento, en esquemas regulados bajo la dependencia del contexto.

Con el fin de asegurarnos de la factibilidad del uso de la sílaba para las tareas de reconocimiento de voz por computadora, se hará un análisis de sus propiedades internas dentro de textos y en corpus de voces existentes.

- Realizar el estudio de la inmersión de un Sistema Experto para la tarea de segmentación automática en los SRAH.

Lo que permitirá indagar el comportamiento que tiene el agregar conocimiento a priori en la etapa de entrenamiento y su repercusión en el reconocimiento.

- Aplicar la técnica de Cadenas Ocultas de Markov para demostrar la utilidad de la sílaba en los sistemas de reconocimiento de voz dentro del español, mediante estrategias de caracterización basadas en esta unidad del habla.

Lo que permitirá implantar de forma práctica los resultados obtenidos en los puntos anteriores, descansando sobre un algoritmo de gran uso en la actualidad, que por su flexibilidad resulta óptimo para los fines del presente trabajo.

- Crear variantes de investigación dentro del español, con lo que el estado del arte actual en esta área se verá beneficiado y la contribución del presente trabajo servirá como una vista alternante a los elementos actuales.

Una vez alcanzadas las metas anteriores, se pretende generar la ruta de inicio para el uso del paradigma de la sílaba en los SRAH de voz para el español.

## **HIPÓTESIS**

La inclusión de sílabas en los esquemas de reconocimiento de voz dentro del español puede permitir el incremento en los índices de reconocimiento que se manejan en la actualidad, sobre todo para cuestiones donde las aplicaciones son totalmente dependientes del contexto.

Esto puede ser posible dado que los mecanismos perceptivos del ser humano se realizan basándose en el uso extensivo de la información temporal, que son muy similares a los que caracterizan a las sílabas.

#### **1.4 ORGANIZACIÓN DE LA TESIS**

El presente trabajo se encuentra dividido en 7 capítulos, los cuales tienen la siguiente estructura:

- a) El capítulo 2, realiza un estudio referente a los conceptos internos que los sistemas de reconocimiento del habla manifiestan. Estableciéndose los conceptos básicos que conforman a esta área de la computación.
- b) El capítulo 3, muestra un estudio de las características inherentes de las sílabas dentro del idioma español, sus elementos esenciales y sus características, así como también su comportamiento en textos y corpus del español.
- c) El capítulo 4, realiza un estudio de los diferentes algoritmos utilizados en el reconocimiento de voz por computadora, logrando con ello establecer las bases matemáticas para los experimentos a realizar.
- d) El capítulo 5, realiza un análisis exhaustivo del comportamiento de las Cadenas Ocultas de Markov en un corpus de palabras determinado.
- e) El capítulo 6, realiza el estudio del comportamiento de las sílabas en un sistema de reconocimiento del habla continua.
- f) El capítulo 7, termina con el análisis de los resultados obtenidos y las conclusiones del presente trabajo.

#### **1.5 RESUMEN DEL CAPÍTULO**

En el presente capítulo se analizaron los diferentes elementos que constituyen un Sistema de Reconocimiento Automático del Habla (SRAH), tomando en cuenta los componentes computacionales que se requieren, así como las técnicas empleadas para tal fin. Se presentaron además un conjunto de elementos primordiales, para considerar a la sílaba como una unidad de reconocimiento que puede permitir incrementar los índices de reconocimiento actuales.

Se analizó la problemática que tienen los sistemas de reconocimiento basados en fonemas y la solución planteada al usar sílabas. Se considera como parte fundamental de este trabajo el análisis de las características de la sílaba, tanto en el habla continua como en la discontinua.

# CAPÍTULO 2

---

## Estado del arte

Este capítulo trata sobre el estado del arte actual del Reconocimiento de Voz por Computadora, así como la descripción de los principios y fundamentos de cada uno de los elementos que componen esta área de la Inteligencia Artificial.

Se aborda, además, la problemática actual del fonema como unidad básica de reconocimiento, lo que indirectamente ha dado lugar a la presente investigación, haciendo esencial hincapié en el uso del paradigma de la sílaba como unidad alternativa de reconocimiento para el español.

Dentro de los sistemas de reconocimiento de voz se encuentra una amplia gama de elementos que imponen su estudio exhaustivo. A su vez, el procesamiento de la señal de voz es en gran medida una tarea que, como ya se sabe, depende del desarrollo de varios elementos para lograr una calidad óptima.

Dentro de este campo de investigación existe una diversidad de fundamentos y conceptos básicos relacionados. Sin embargo, cabe hacer la aclaración de que al ser el reconocimiento de voz un campo bastante amplio, es prácticamente imposible abarcarlos todos durante su descripción.

Aunado a lo anterior, existe un conjunto de factores que los sistemas de reconocimiento de voz aún no pueden controlar, los cuales son causas de errores o salidas ambiguas.



## 2.1 | ANTECEDENTES HISTÓRICOS

Se considera que el reconocimiento de voz por computadora es una tarea muy compleja debido a todos los requerimientos que le son implícitos (Suárez, 2005). Además del alto orden de los conocimientos que en ella se conjugan, deben tenerse nociones de los factores inmersos que propician un evento de análisis individual (estados de ánimo, salud, etc.). Por tanto, en los SARH, ya sea para tareas específicas o generales, es inmensa la cantidad de aspectos a realizar.

La historia esencial de los sistemas de reconocimiento de voz se puede resumir con las siguientes premisas (<http://www.gtc.cps.unizar.es>):

- Los inicios: años 50's
  - Bell Labs. Reconocimiento de dígitos aislados monolocutor.
  - RCA Labs. Reconocimiento de 10 sílabas monolocutor.
  - University College in England. Reconocedor fonético.
  - MIT Lincoln Lab. Reconocedor de vocales independiente del hablante.
- Los fundamentos: años 60's – Comienzo en Japón (NEC labs)
  - Dynamic Time Warping (DTW – Alineación Dinámica en Tiempo -). Vintsyuk (Soviet Union).
  - CMU (Carnegie Mellon University). Reconocimiento del Habla Continua. HAL 9000.
- Las primeras soluciones: años 70's - El mundo probabilístico.
  - Reconocimiento de palabras aisladas.
  - IBM: desarrollo de proyectos de reconocimiento de grandes vocabularios.
  - Gran inversión en los EE. UU.: proyectos DARPA.
  - Sistema HARPY (CMU), primer sistema con éxito.
- Reconocimiento del Habla Continua: años 80's - Expansión, algoritmos para el habla continua y grandes vocabularios
  - Explosión de los métodos estadísticos: Modelos Ocultos de Markov.
  - Introducción de las redes neuronales en el reconocimiento de voz.
  - Sistema SPHINX.
- Empieza el negocio: años 90's - Primeras aplicaciones: ordenadores y procesadores baratos y rápidos.
  - Sistemas de dictado.
  - Integración entre reconocimiento de voz y procesamiento del lenguaje natural.
- Una realidad : años 00's - Integración en el Sistema Operativo

- Integración de aplicaciones por teléfono y sitios de Internet dedicados a la gestión de reconocimiento de voz (Voice Web Browsers).
- Aparece el estándar VoiceXML.

## **2.2 ESTADO DEL ARTE**

A lo largo del tiempo, el estudio de las sílabas como base para definir modelos de lenguaje ha arrojado algunos resultados beneficiosos. Hu et al. utilizaron las sílabas en un experimento que permitía el reconocimiento de sílabas pertenecientes a un vocabulario de los meses del año en inglés (Hu et al., 1996). Encontraron un total de 29 unidades silábicas del corpus con un 84.4% de eficiencia. Boulard Dupont (1996) mencionó el uso de las sílabas en trabajos bastante similares para el alemán.

Haustein, al comparar el rendimiento en un SRAH híbrido (HMM-NN Hidden Markov Models-Neural Networks –Cadenas Ocultas de Markov y redes neuronales–) utilizando sílabas y fonemas como unidades básicas para el modelo, encuentra que ambos sistemas presentan ventajas que se pueden aprovechar de manera combinada (Hauenstein, 1996). Wu et al. propusieron la integración de información al nivel de sílabas dentro de los reconocedores automáticos del habla para mejorar el rendimiento y aumentar la robustez (Wu, 1998) y (Wu et al., 1997). La razón de error alcanzada fue del 10% para un corpus de voz de dígitos del corpus de OGI (Oregon Graduate Institute). En (Wu, 1998), se reportan resultados del orden del 6.8% para un corpus de dígitos proveniente de conversaciones telefónicas, haciendo uso de un sistema híbrido fonema-sílaba.

Jones et al. (1999) experimentaron con los modelos ocultos de Markov (HMM - Hidden Markov Models) para obtener las representaciones de las unidades al nivel de sílaba, encontrando que se puede mejorar substancialmente los rendimientos del SRAH en una base de datos de tamaño mediano al compararlos con modelos monofónicos (Jones et al., 1999). Logrando un 60% de reconocimiento que lo comparan con un 35% que se obtiene al utilizar monofonemas, dejando en claro que las aplicaciones prácticas deben de conformarse por un sistema híbrido. Fosler et al. encontraron que una gran cantidad de fenómenos fonéticos en el habla espontánea son de carácter silábico y presentaron un modelo de pronunciación que utiliza ventanas fonéticas contextuales mayores a las utilizadas en los SRAH basados en fonemas (Fosler et al., 1999).

Weber, al experimentar con marcos (“tramas”) de diferentes duraciones, encuentra una mejora en las tasas de error de reconocimiento de palabra (WER Word Error Rate –razón de error de palabra–) cuando se introduce ruido a la señal de voz (Weber, 2000). El hecho de que sus ventanas abarquen hasta los cientos de milisegundos sugiere, aunque de manera indirecta, que una unidad conveniente de modelado lo serían las sílabas.

Meneido y Neto utilizan la información al nivel de sílaba para analizar los sistemas de segmentación automática que, aplicado al portugués, mejora la WER (Meneido and Neto, 2000). Comienza a surgir la idea de diseñar un sistema híbrido de modelado de lenguaje para aplicarlo al SRAH. El trabajo de Meneido en portugués da una pauta de cómo funcionaría la incorporación de la información al nivel de sílaba en español, ya que ambos idiomas comparten las características de tener sílabas bien definidas (Meneido et al., 1999) y (Meneido and Neto, 2000). El trabajo de Meneido (Meneido et al., 1999), reporta un 93% de detección de inicios de la sílaba, además de considerar que ventanas de entrada de contexto amplias (260 ms), son las más apropiadas.

En 1996, en el Centro de Lenguaje y Procesamiento de Voz de la Universidad John Hopkins, se celebró un taller de verano de reconocimiento del habla continua de vocabulario grande, donde se demostró que el uso de la información al nivel de sílaba y de marcas de "presión interna del individuo", puede reducir la WER de sistemas basados en trifenemas (Wu, 1998). En ese mismo año (Hauenstein 1996), realiza la inmersión de unidades silábicas a sistemas de reconocimiento del habla continua, realizando la comparación con unidades fonéticas. Los experimentos demostraron que los sistemas basados en fonemas presentan una mejor efectividad en sistemas del habla continua, y que las sílabas mejoran los índices de reconocimiento para el habla discontinua, alcanzando un 17.7% de razón de error como máximo.

Cinco años después la idea de incorporar la información al nivel de sílaba, con la que se obtiene un nivel similar al del fonema, fue retomada por Ganapathiraju et al. (2001), quienes observaron que existe un número considerable de trifenemas que tienen poca información para ser modelados. Además, notaron que en los trifenemas se presenta una mínima integración de las dependencias espectrales y temporales debido a la corta duración de los marcos, y que el hecho de que el mapeo uno a uno de las palabras con sus fonemas, introduce una gran cantidad de categorías diferentes para un mismo sonido.

Proponen un sistema donde integran un inventario de modelos de sílabas y una mezcla de modelos acústicos que van desde las palabras monosílabas hasta fonemas dependientes del contexto. Su sistema reduce en un 1% el WER con respecto a un sistema basado en trifenemas en la base de datos Switchboard. Reportan un porcentaje del 11.1% de razón de error al hacer uso de la sílaba independiente del contexto dentro del corpus anteriormente mencionado.

En 1997 se generó en Edimburgo el proyecto "Mejora del reconocimiento del habla utilizando estructura silábica" (Proyecto ESPRESSO), motivado por el hecho de que los SRAH convencionales basados en HMM se habían estancado, en el sentido de que su rendimiento no había podido ser sustancialmente mejorado, y de la

conveniencia de encontrar modelos más efectivos que los modelos de fonemas dependientes del contexto.

Se buscó que el modelo capturara la dependencia del contexto sin necesidad de llegar al número excesivo de parámetros requeridos para su representación. La primera fase terminó en 1999 en que se investigó el uso de las sílabas fonéticamente caracterizadas para el reconocimiento de voz (King, 2000), mientras que en la segunda fase se examinaron nuevos modelos para el reconocimiento de voz que tomaran en cuenta la naturaleza continua de los esquemas encontrados en la fase anterior. Esta fase se encuentra aún en proceso de realización.

Para el español se han desarrollado reconocedores de voz de propósitos específicos. Córdoba y colaboradores trabajaron en un sistema SRAH para dígitos con aplicaciones en telefonía (Córdoba et al., 1995). También se ha abordado el problema de la variabilidad del estado emocional del hablante (Montero, 1999). En México se ha trabajado sobre los SRAH, con una versión para el español que se habla en nuestro país basado en una plataforma desarrollada originalmente para el inglés (Munive et al., 1998) y (Fanty, 1996).

Con respecto al estudio de las sílabas, Feal menciona la idea de utilizarlas como unidad para la síntesis del español, y presenta el listado de las sílabas que se utilizan en este idioma tomando una base de datos creada a partir de libros de literatura española (Feal, 2000). Asimismo, propone un algoritmo para la segmentación de sílabas en texto continuo a partir de la transcripción fonética de las palabras.

El estado del arte en lo que toca a los SRAH recae sobre el uso de los fonemas. Los resultados obtenidos hasta estos momentos en lo que cabe a vocabularios medianos ha sido satisfactorio, pero cuando el vocabulario comienza a crecer (mayor que 50000 palabras) los resultados no son los deseados.

La tarea principal de un SRAH ha quedado en un punto sin avance, en donde los nuevos trabajos no han podido hacer que los índices de reconocimiento mejoren (Delaney, 2000). Debido a esto, en años recientes se ha hecho hincapié en la introducción de nuevas técnicas que permitan alcanzar tal objetivo.

Para el idioma inglés, uno de los trabajos pioneros reportado en (Hu et al., 1996) muestra la integración de la unidad silábica como unidad de reconocimiento. Se menciona que entre las principales motivaciones para usar la sílaba para tal efecto, es la de que son unidades que realizan un modelado de la señal de voz; además, este método ofrece un mejor entramado para la integración de técnicas de modelado dinámico en los sistemas de reconocimiento de voz.

Como una parte fundamental que confirma la efectividad del uso de la sílaba, consideran que las unidades silábicas pueden captar no solamente la dinámica dentro de las fronteras de los fonemas, sino también la dinámica en la región de transición entre ellos. Todo esto debido a que, por ejemplo, las transiciones entre los fonemas (vocal-vocal) son difíciles de detectar con los algoritmos de segmentación actuales, los cuales de acuerdo al corpus en el que se trabaje suelen agruparse en un conjunto especial denominado unidades semejantes a las sílabas.

Para ello utilizaron un corpus de los nombres de los 12 meses del año perteneciente a un subconjunto del conjunto de datos Census que se mantiene en OGI (Oregon Graduate Institute). También utilizaron 8 coeficientes PLP (Perceptual Linear Prediction) como vectores característicos por cada estado y 10 estados, por lo que el vector característico fue en total de 81 componentes, ya que agregaron el logaritmo de la duración de la señal. Utilizaron redes neuronales para el reconocimiento.

Uno de los trabajos de mayor relación con la importancia de la sílaba es el referido en (Lizana et al., 2000), en donde se indica que no existe una definición exacta para esta unidad lingüística. Este es uno de los trabajos pioneros que utiliza el parámetro de energía de la sílaba para realizar su segmentación. Analizan el problema que existe al intentar segmentar la sílaba y hacen hincapié en que el 80% del idioma inglés, se encuentra regido por expresiones monosilábicas (estructuras CV - Consonante Vocal -, V - Vocal -, VC - Vocal Consonante -, CVC - Consonante Vocal Consonante -, etc.).

En este trabajo se hace referencia y una comparación con el algoritmo de división en unidades silábicas para el inglés propuesto por Mermelstein (1975).

El algoritmo propuesto en este trabajo para realizar la segmentación silábica es parecido al empleado en el trabajo antes mencionado, pero orientado hacia el idioma español. El algoritmo de Mermelstein utiliza un filtro Butterworth pasa banda en el rango 500-4000 Hz. Utilizan los cruces por cero como referencia. La propuesta añadió un pico de energía y utilizaron un valor de 72 niveles de presión sonora como umbral de corte de los segmentos de sílaba.

Para el experimento utilizaron 12 frases balanceadas fonéticamente con un total de 74 sílabas y compararon sus salidas con un etiquetado realizado a mano con anterioridad.

Como otro de los trabajos sobresalientes en esta área se encuentra el de Jones et al. (1999), en el que se usó una base de datos telefónica y un conjunto de modelos ocultos de Markov para representar a las unidades silábicas del corpus Subscriber que contiene un total de 1313 palabras distintas. Su trabajo es sobre 3063 frases, de las cuales 1810 fueron para las pruebas. Existieron 750 locutores

en el entrenamiento y 500 en el reconocimiento. Una parte importante que refieren es la dificultad de delimitar las fronteras de las sílabas dentro del corpus. Puntualizan que la utilización de unidades silábicas permite la construcción de gramáticas de palabras completas. La sílaba más común dentro del corpus se manifestó 3290 veces. Alcanzando un reconocimiento óptimo del 60%.

Un aspecto importante del análisis realizado radica en que el 95% de las sílabas con pobre calidad de reconocimiento aparece al menos 30 veces. Concluyen mencionando los altos costos computacionales de la unidad silábica, pero sin embargo demuestran la aplicación del reconocimiento de voz en un corpus de tamaño medio usando a la sílaba como unidad elemental.

Dentro de la literatura referente a resolver el problema de la segmentación en unidades silábicas de las palabras del idioma inglés, en (Hartmut et al., 1998) se propone un algoritmo basado en la detección del núcleo de la estructura silábica. Dicho núcleo, por lo regular, se ha caracterizado por estar constituido de una estructura vocálica, rodeada del inicio de la sílaba ("onset") y por la parte final de la misma ("coda"). Ambas partes alrededor del núcleo son comúnmente consonantes del idioma.

En este experimento utilizaron las bases de datos PhonDatII y Verbmobil etiquetadas por 3 fonetistas, cuya labor sirvió para ajustar los parámetros del algoritmo y evaluar su proporción de error. Del corpus PhonDatII utilizaron 200 frases de 16 locutores con un total de 3177 sílabas; del Verbmobil se usaron 960 frases con un total de 14048 sílabas.

Para dialectos como el cantonés, que se hablan en países como China y que por lo regular se basan en unidades silábicas, también se han hecho un conjunto de experimentos para analizar los resultados que un sistema de reconocimiento del habla basado en sílabas (Peskin et al., 1991 y Lee and Ching, 1998). En (Lee and Ching, 1998), emplearon un sistema de reconocimiento dependiente del locutor con un vocabulario de entre 40 y 200 sílabas. En el caso de las 200 sílabas, se logró un 81.8% de reconocimiento adecuado.

El dialecto cantonés tiene como característica ser monosilábico y basado en tonos, por lo que el sistema propuesto en (Peskin et al., 1991) se compone de dos partes esenciales: un detector de tonos y un reconocedor de sílabas básico. Utilizaron para ello una red neuronal del tipo perceptrón multicapa. El tamaño del vocabulario era de 40 a 200 sílabas, que representan el 6.8 y 34.4 por ciento del total de las sílabas de dicho dialecto. El número de locutores fue de 10, de los cuales 5 eran hombres y 5 mujeres.

Otro dialecto que ha sido objeto de estudio es el mandarín (Chang, 2000), que al igual que el anterior se encuentra basado en tonos y sílabas, pero que a diferencia

del cantonés tiene reforzamiento del tono al final de la sílaba. Dentro de las principales características de éste está que inserta al tono fundamental ("pitch") como parámetro de reconocimiento. La base de datos comprendió un total de 500 locutores y 1000 frases; el algoritmo que obtiene el "pitch" corrió en tiempo real y además agregaron componentes delta y doble delta. Obtuvieron una razón de error de 7.32% cuando ningún procesamiento tonal se realizó.

Cuando al modelo final, le agregaron el análisis de tonos se alcanzó una razón de error de 6.43%. Cuando la información del pitch y el conjunto de sílabas se combinaron, la razón de error fue del 6.03% (con lo que lograron una reducción de error acumulada del 17.6%). Estos resultados denotan que las fuentes de información tales como la energía y la duración pueden contribuir al problema de la ambigüedad de diferentes tonos.

La referencia más cercana a los trabajos del idioma español son los trabajos que Meneido y Neto (Meneido et al., 1999) y (Meneido and Neto, 2000) hicieron para el portugués. En (Meneido et al., 1999) autores manifiestan que desarrollos recientes han permitido observar que la sílaba puede ser utilizada como una unidad de reconocimiento para el portugués, idioma altamente silábico, debido a que las fronteras de las sílabas están mejor definidas que en el caso de los fonemas. Dentro de sus principales trabajos se encuentra el poder realizar la segmentación de la señal de voz en unidades silábicas por medio de la extracción de características orientadas a la percepción.

Dichas características fueron post-procesadas por medio de un mecanismo simple de umbral o por medio de una red neuronal, basados en las estimaciones de las fronteras de las sílabas. Los resultados obtenidos mostraron que con una ventana de aproximadamente 260 ms se alcanzó un 93% de detección de los inicios, con una razón de inserción del 15%.

La incursión de los trifenemas a la par del estudio de la sílaba es otro aspecto relevante que mencionan en el artículo (Meneido and Neto, 2000). Como preámbulo a su investigación, consideraron que un conocimiento exacto de los inicios de la sílaba puede ser útil para incrementar los índices de reconocimiento que logran actualmente. La base de datos que utilizaron para la prueba fue la BD-PUBLIC, de la cual obtuvieron una segmentación adecuada de la misma del orden del 72%.

En su trabajo mencionan cuatro métodos para segmentar la señal en unidades silábicas. Bajo la base de datos analizada extrajeron un corpus de un total de 750 frases, con un total de 3408 palabras, de las cuales 1314 eran diferentes con un total de 616 sílabas distintas.

Cabe destacar que el trabajo de Meneido (Meneido and Neto, 2000) se ve fuertemente influenciado por el trabajo de Villing Rudi et al. (2004), en el cual se menciona la utilidad que tiene la unidad silábica a partir de un conjunto de experimentos de alfa dígitos, mismos que surgieron a partir de su aplicación a cierta parte de la base de datos SWITCHBOARD. En tal experimento demostraron que las sílabas tienen mejor ejecución que los trifenemas en un orden del 1.1%.

Asimismo, se presenta el análisis de un algoritmo utilizado para la segmentación de unidades silábicas, que se comparan con los resultados de los algoritmos de Mermelstein y Howitt utilizados para realizar la misma actividad. Se presenta un porcentaje de reconocimiento de 93.1% mayor al 76.1% del algoritmo de Mermelstein y del 78.9% del de Howitt.

En otra aportación de Meneido que se reporta en (Meneido and Neto, 2000) se analizó el rendimiento que tiene la utilización de las Cadenas Ocultas de Markov y las redes neuronales en el reconocimiento de frases del portugués por medio de un nuevo método de segmentación en unidades silábicas. En (Hartmut et al., 1996), se presenta un algoritmo para la segmentación de la palabra en unidades silábicas que contiene una razón de error del 12.87% y del 21.03% para el habla discontinua y el habla espontánea respectivamente. Dicho algoritmo empleado utiliza una gran cantidad de filtros digitales para realizar la segmentación correspondiente.

Una parte importante que menciona en este trabajo es que a lo largo de los años, los investigadores han encontrado numerosas formas de sugerir que las sílabas pueden ser desde el punto de vista perceptual muy importantes, debido a que los mecanismos de percepción humana hacen un uso extensivo de la información temporal propia de éstas.

Al igual que en el experimento referido en (Meneido et al., 1999), utilizan 12 coeficientes PLP-cepstrales y el logaritmo de la energía cada 10 ms aplicados a tramas de 20 ms. Con este nuevo método alcanzaron un 79.7% en la calidad de la segmentación. Los resultados obtenidos demuestran una reducción de la proporción de error de palabra (WER) muy significativa con relación a los reconocedores basados en fonemas.

Uno de los trabajos más sobresalientes en cuanto al reconocimiento automático del habla basado en sílabas es el descrito en (Hauenstein, 1996), donde se compara un trabajo realizado con unidades silábicas para ampliar el estado del arte con relación a los fonemas. En él emplearon Cadenas Ocultas de Markov con redes neuronales, y el reconocimiento de voz con un corpus pequeño y mediano.

Tras realizar la comparación de ambos sistemas mediante la proporción de error de palabra, el sistema basado en sílabas resultó ser superior en el reconocimiento



de palabras aisladas, mientras que el sistema basado en fonemas resultó tener mayores ventajas en el reconocimiento del habla continua.

El corpus del experimento fueron los dígitos, extraídos de una base de datos de OGI. La base de datos contiene señales de voz pronunciadas de forma espontánea y muestreadas a 8 kHz. La mayor parte de las palabras son de habla continua, pero existen algunas palabras aisladas. El vocabulario comprende 92 palabras diferentes: dígitos (0-9), números cardinales, números ordinales y palabras que no son números.

Utilizaron como modelo del lenguaje el bigram, con una razón de perplejidad de 12.97; el conjunto de entrenamiento contiene una perplejidad de 4.71 y el conjunto de prueba una perplejidad de 2.23. El corpus antes mencionado contiene un total de 96 sílabas diferentes. Los resultados encontrados en este trabajo muestran una diferencia de 2% en el caso del reconocimiento de sílabas de dígitos aislados sin ruido y de 4.4% para el caso de la señal con ruido en comparación con los fonema. Sin embargo, cuando se realiza un reconocimiento con habla continua usando fonemas, éstos superaron el uso de las sílabas en un 4.7%.

Como parte final de este panorama de la aplicación del uso de la sílaba en sistemas de reconocimiento del habla continua, se tiene el proyecto ESPRESSO, el cual trabaja sobre las bases de datos SWITCHBOARD y TIMIT (dos bases de datos destinadas al reconocimiento del habla para el inglés). Como se dijo, este proyecto se encuentra actualmente en desarrollo.

Los resultados anteriores nos demuestran la utilidad de la sílaba en su aplicación a los sistemas de reconocimiento de voz. Sin embargo, su aplicación como se menciona está destinada al idioma inglés e incluso a vocablos de dialectos chino y japonés. En lo que respecta al idioma español, los trabajos desarrollados sobre este tenor resultan ser efímeros. De hecho, en nuestro país existe en la actualidad la inquietud reciente de utilizar tal unidad del lenguaje para realizar el reconocimiento de la voz. El presente trabajo representa un esfuerzo más entre los que las instituciones de investigación de nuestro país están realizando.

Finalmente, el presente trabajo basa su desarrollo en las Cadenas Ocultas de Markov de densidad continua para el caso del reconocimiento de voz continuo. La principal inserción o modificación a los esquemas preestablecidos, radica en la etapa de entrenamiento, en la cual se realizó la inserción de un sistema basado en conocimiento que fundamentado en los comentarios hechos por Hauenstein en su trabajo referido en (Hauenstein, 1996). Además, este proceso refuerza la etapa de segmentación de la unidad silábica la cual utiliza parte de los esfuerzos de Meneido al agregar filtros digitales para la segmentación de la señal de voz en la etapa de entrenamiento. El parámetro de la energía de RO (ERO), lo que constituye una de las aportaciones del presente trabajo, muestra la utilidad de los filtros digitales en la segmentación de las unidades silábicas de una frase o frases del corpus.

## 2.3 FUNDAMENTOS DEL RECONOCIMIENTO DE VOZ

El reconocimiento de voz por computadora es una tarea compleja de reconocimiento de patrones. Por lo regular, la señal de voz se muestrea en un rango entre los 8 y 16 KHz. En el desarrollo de los experimentos de esta tesis, la frecuencia de muestreo más utilizada fue de 11025 Hz. La señal de voz necesita ser analizada para extraerle información relevante una vez que ha sido digitalizada.

A manera de resumen, dentro de esta tarea existen las siguientes técnicas de extracción de parámetros característicos de la señal (Kirschning, 1998), (Jackson, 1986), (Kosko, 1992) y (Sydral et al., 1995):

- **Análisis de Fourier.** Como se puede observar en la figura 2.1, consiste en aplicar la Transformada Rápida de Fourier, TRF (FFT - Fast Fourier Transform) al conjunto de muestras que en ese momento se está analizando. Regularmente, dicha representación en el dominio de la frecuencia se distribuye por medio de la conocida escala de MEL, en donde las frecuencias menores a 1KHz se analizan de forma lineal y las superiores de forma logarítmica (Bernal et al., 2000), con el fin de crear una analogía con la cóclea interna del oído que de manera natural trabaja como un divisor de frecuencias. Esto se puede observar en la expresión de la ecuación 2.1.

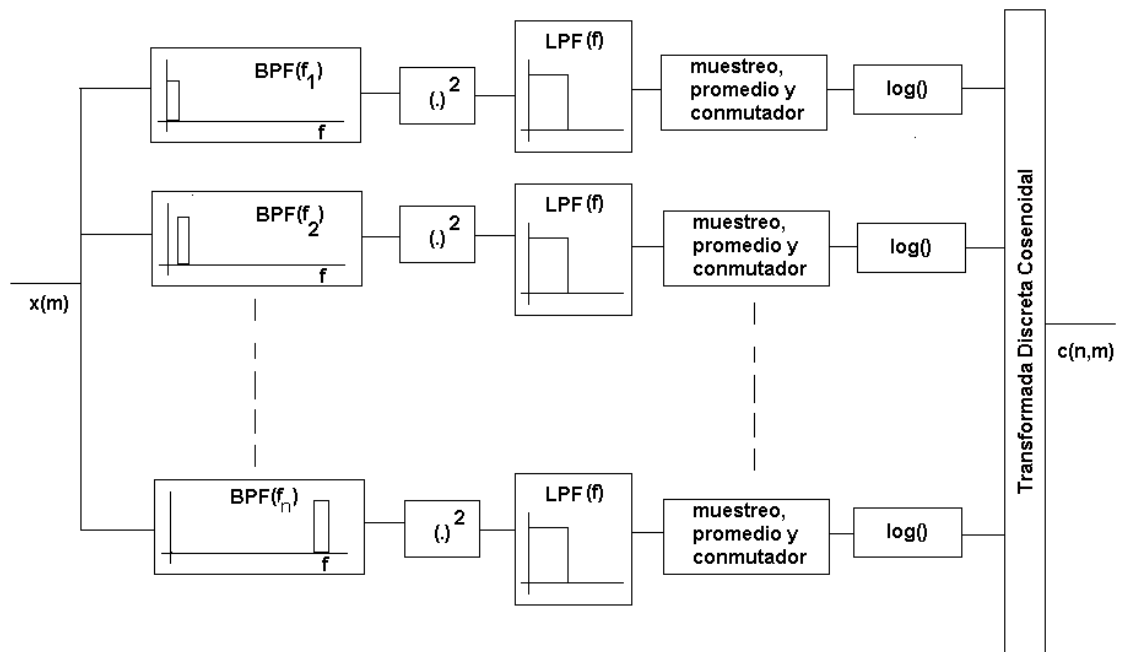


Fig. 2.1. Implantación de bancos de filtros del tipo Cepstral (Suk and Flanagan, 1999).

$$Mel(f) = b \log_{10} \left( 1 + \frac{f}{c} \right), \quad (2.1)$$

- **Codificación Predictiva Lineal**, CPL (LPC –Linear Predictive Coding-). La finalidad del método es encontrar un conjunto de vectores representativos denominados en conjunto vectores código, que forma una matriz de vectores representativos que a su vez conforma lo que se denomina Libro Código,. Basa su hipótesis en el hecho de que la muestra  $x_t$  de una señal puede ser predeterminada a partir de las k muestras anteriores, si conocemos el peso por el cual cada una de ellas está afectada por todas las  $x_{t-k}, x_{t-k-1}, x_{t-k-2},$  etc., muestras anteriores.
- **Análisis de los coeficientes Cepstrales**. Representan la transformada inversa de Fourier del logaritmo del espectro de potencia de la señal (Rabiner and Biing-Hwang, 1993).
- **Predicción Lineal Perceptiva** (PLP| -Perceptual Linear Prediction -). Resultante de las características fisiológicas, pero no puede ser representada gráficamente (Kirschning, 1998).

## 2.4 TÓPICOS DEL RECONOCIMIENTO DE VOZ

La característica esencial de la señal de voz es su excesiva variación en función del tiempo. En la actualidad, el reconocimiento de voz se ve desde dos diferentes vertientes: el nivel acústico y el nivel temporal (Tebelskis, 1995). Desde el punto de vista acústico, se analizan aspectos tales como: el acento, la pronunciación, la frecuencia fundamental de resonancia del tracto vocal ("pitch" en inglés), volumen, entre otros. En el caso de la variación temporal se analizan las diferentes duraciones que manifiesta un conjunto de muestras de voz. Por lo regular, los dos aspectos anteriores no son independientes; sin embargo, para fines esenciales del reconocimiento de voz, se toman como si lo fuesen.

De los dos aspectos antes señalados, la variación en el tiempo es más fácil de gestionar. En principio, se utilizó un tipo de deformación lineal de una señal de voz desconocida para compararla con una señal tipo muestra; el resultado no fue óptimo. Posteriormente se utilizó una deformación del tipo no lineal, la cual dio como consecuencia la aparición del algoritmo DTW (Dynamic Time Warping) (Rabiner and Levinson, 1990). En la actualidad tal algoritmo ha dejado de usarse en gran medida. La variación acústica es más difícil de modelar, debido a su naturaleza heterogénea. Por lo tanto, el estudio del reconocimiento de voz ha ampliado su campo en este aspecto. Las diferentes perspectivas por donde se analiza el reconocimiento de voz se reducen a las siguientes:

- a) Modelos de referencia o plantillas,
- b) Conocimiento,
- c) Modelos estocásticos o estadísticos,
- d) Redes neuronales artificiales,
- e) Métodos híbridos.

El método más popular que se ha venido utilizando es el método estadístico que hace uso de las Cadenas Ocultas de Markov (Kita et al., 1993).

Varios de los sistemas desarrollados con tales métodos se presentan en dominios específicos y llegan a ser limitados cuando sus límites se expanden. Por ello surge la necesidad de la investigación de nuevas vertientes.

## **2.5 ADECUACIÓN DEL PROCESO DE RECONOCIMIENTO**

Al ser la señal de voz variable en el tiempo, se hace indispensable su digitalización para ser tratada por los recursos computacionales con los que se haría posible su reconocimiento.

Una de las características esenciales a definir en el proceso de captura de la señal de voz es la frecuencia de muestreo. Este factor es muy importante, pues es la limitante y posible causante de diferenciar entre una buena calidad de señal y los problemas que se pueden presentar si no se respetan las reglas que el procesamiento digital de señales enmarca. Específicamente hablando y suponiendo que el problema anterior se ha resuelto, quedan aún muchos factores a analizar dentro de los cuales se encuentran los siguientes (Kirschning, 1998):

### ➤ TAMAÑO DEL VOCABULARIO Y CONFUSIÓN

Los sistemas, conforme más palabras se desea que reconozcan, tienden a incrementar los índices de error. Se tienen reportes de un aceptable porcentaje de reconocimiento cuando se trabaja con números de palabras menores a 1000, pero el problema se agrava cuando este número se incrementa. Esto da origen a que el porcentaje de reconocimiento se vea gravemente afectado, pues con frecuencia el sistema tiende a caer en inestabilidades y por ende a perder características de confiabilidad.

### ➤ SISTEMAS DEPENDIENTES E INDEPENDIENTES DEL LOCUTOR

La gran controversia dentro de los sistemas de reconocimiento de voz se ve plasmada en los sistemas de reconocimiento dependientes y no dependientes del locutor. A lo largo de la historia que tiene el advenimiento de los SRAH, se

han gestado los sistemas dependientes del locutor como los que han hecho una realidad al alcanzar un alto índice de reconocimiento.

Sin embargo, y en contraparte, se encuentran los sistemas independientes del locutor, donde se hace evidente la necesidad de implementar mecanismos cada vez más sofisticados que representan un problema aún no resuelto en nuestros días. A pesar de esto, los avances no se han hecho esperar y aunque actualmente podemos hablar de reconocimiento de voz para un grupo determinado de personas, es una realidad que resulta difícil la extensión de estos esquemas para que cubran a toda una población.

➤ VOZ AISLADA, DISCONTINUA Y CONTINUA.

Gran parte del desarrollo de este trabajo se ve enfocado en estos tópicos. Se entiende por voz aislada aquella que podemos percibir como unidades del habla que forman un núcleo elemental de entendimiento dentro de la estructura lingüística en donde se gesten.

Cabe hacer la mención de que este hecho es importante porque la sílaba y el fonema pueden entrar dentro de esta clasificación. Sin embargo, existe un problema con el fonema: como pieza independiente carece de sentido, siendo totalmente abstracto y sin relevancia cuando se manifiesta de manera independiente. En contraparte, la sílaba es totalmente autónoma sin necesidad de compartir espacios temporales con algún otro medio lingüístico. Por tal motivo, podemos decir que su contenido de información es vasto y enorme.

La voz discontinua es una manifestación en donde las palabras o secuencias sonoras se encuentran entrelazadas por una pauta que no permite que haya continuidad entre una estructura anterior y la siguiente, sino más bien es simplemente el intermedio entre lo continuo y lo individual. En este esquema, los reconocimientos que se realizan son de gran importancia para la investigación.

Finalmente, el habla continua es por naturaleza la forma que los seres humanos tienen para comunicarse entre ellos. Es importante observar la forma en la que los elementos anteriores se presentan en la vida. El ser humano en su proceso de adaptación permite que los sonidos ayuden al equilibrio óptimo de las funciones básicas del cerebro que lo acompañaran durante toda la vida. Al pasar el tiempo, el individuo comienza a coordinar sus estructuras sonoras de tal forma que las acopla al medio que le rodea; esto es, no importa si haya nacido en México, por ejemplo, o si es trasladado a otra región del mundo. Este tiende a aprender y a acoplarse al medio donde se encuentre independientemente de las nacionalidades.

De lo anterior y de la forma de aprender del individuo, se asevera que las primeras manifestaciones de coordinación real del lenguaje entendible se dan por estructuras más complejas al fonema, las que bien pueden ser las sílabas.

#### ➤ VOZ APLICADA A TAREAS O EN GENERAL

Los sistemas de reconocimiento de voz se encuentran altamente ligados a la aplicación que se esté llevando a cabo durante su implementación, es decir, muchos sistemas se encuentran limitados en contexto por la tarea que tienen que realizar, mientras que otros quedan completamente abiertos. Piénsese en un SRAH destinado a gestionar conversaciones telefónicas, reservaciones de vuelos aéreos, etc. Como es de suponerse, la cantidad de elementos que tiene este vocabulario se encuentra limitada a unas cuantas palabras.

Uno de los principales objetivos de este trabajo es precisamente demostrar que las sílabas se acoplan de manera satisfactoria a sistemas destinados a ciertas aplicaciones.

#### ➤ VOZ LEÍDA O ESPONTÁNEA

Los SRAH existentes hasta estos momentos se manifiestan en dos grandes vertientes, sobre todo cuando se habla de bases de datos destinadas para tal fin. Los corpus de voces almacenados para estudio se diferencian en el hecho de que sus grabaciones se encuentran hechas por personas que pronuncian las frases cuando las leen, o cuando se encuentran en una charla normal.

TIMIT es una base de datos que demuestra este hecho. Gran parte de las muestras de voz que en ella se encuentran almacenadas son realizadas por personas que se encontraban en charlas de oficina o en lugares concurridos, en donde la voz que se percibe es totalmente continua y espontánea. Esto es, que no existió un esquema de conversación preestablecido.

Caso contrario sucede con las muestras de voces leídas, en donde la muestra de voz es obtenida de una secuencia de frases preestablecidas (lecturas, formatos, etc.) y por ende, el hablante pone más cuidado en lo que está diciendo y la claridad se nota en gran parte del texto.

#### ➤ CONDICIONES ADVERSAS

Este tema se trató ya con anterioridad y se refiere específicamente a las perturbaciones que una señal de voz puede recibir por causas del medio ambiente.

Cabe recalcar que es importante tener en cuenta este hecho, pues si se desea tener un sistema de reconocimiento de voz que opere bajo ciertas características (lugar de trabajo, condiciones atmosféricas, etc.), se deberán de tener las muestras de voz extraídas bajo las mismas condiciones en que funcionará el sistema.

## **2.6 MÉTODOS DE RECUPERACIÓN DEL ERROR**

Debido a que el alcance de los sistemas de reconocimiento de voz es más general, otros tópicos ganan terreno en esta área; tal es el caso de los métodos de recuperación del error.

Los métodos de recuperación del error permiten discernir ante un resultado erróneo y en ocasiones realizan la reestructuración del mismo. Este tipo de tarea puede ser realizada en varios niveles, los cuales involucran diferentes tipos de información. Dependiendo de su complejidad, el sistema puede contar con más o menos detalles de la información que está tratando o saber más acerca de las muestras analizadas. En el presente trabajo el dominio se encuentra limitado, pues el contexto es conocido.

Por lo general, podemos mencionar cuatro tipos principales de métodos de corrección del error, a saber:

- a) Modificación por medio del diálogo interactivo,
- b) Modificación por la repetición del diálogo,
- c) Modificación según el contexto,
- d) Modificación por medio de la reparación interna.

## **2.7 RESUMEN DEL CAPÍTULO.**

En el presente capítulo se analizaron los elementos primordiales que conforman el estado del arte en lo que a los sistemas de reconocimiento de voz se refiere, sobre todo aquellos basados en sílabas. Se expusieron las limitantes que hasta estos momentos existen dentro de esta área de investigación. Se hizo un especial énfasis en el uso de las sílabas como nueva unidad de reconocimiento en otros idiomas en sistemas desarrollados por numerosos grupos de investigación. Dichas investigaciones y proyectos demuestran la posibilidad de desarrollo y el interés de realizar investigaciones en esta rama del reconocimiento de voz.

# CAPÍTULO 3

---

## La sílaba, su estructura y su inmersión en los SRAH

Dentro de las diferentes definiciones y acepciones que ha sufrido la sílaba se encuentran que representa los picos de sonoridad, pulsos altos de energía, unidades necesarias en la organización mental y producción de la voz, un grupo de movimientos de la voz y como una unidad de percepción de la voz (Wu, 1998).

El presente capítulo está destinado a analizar las características que las sílabas guardan dentro del contexto lingüístico y su relación con los sistemas de reconocimiento planteados con anterioridad, cabe hacer la aclaración que dentro de los esquemas planteados dentro de los SRAH, una unidad básica se debe de considerar como una forma intermedia de la información de voz alrededor de la cual la mayor parte del proceso de reconocimiento es organizado ya sea por las máquinas o bien por los seres humanos.



### **3.1 EL PARADIGMA DE LA SÍLABA**

Dentro del esquema de los sistemas de reconocimiento de voz, las unidades básicas son idealmente una salida del procesamiento fonético acústico y una etapa del procesamiento de etapas del léxico. Es sabido que una unidad básica, debe de ser lo suficientemente pequeña como para expresar gran variedad de elementos en la manifestación de la voz, ésta debe ser eficientemente computacional y poseer propiedades adecuadas tales que permitan una fácil interpretación de los mismos.

Investigaciones actualmente realizadas en las áreas de psicoacústica y psicolingüística, sugieren que la sílaba debe de ser una unidad básica en el proceso de la percepción de la voz humana (Wu, 1998). Los investigadores han creado la hipótesis de que la sílaba, o una unidad de equivalencia relativa, debe ser la clave en el proceso de reconocimiento en los seres humanos e integrar la información de la señal de voz.

Desde un punto de vista de ingeniería, la sílaba debe ser una eficiente y útil unidad intermedia de voz que puede potencialmente reducir el cálculo redundante y su almacenamiento en el reconocimiento automático de la voz. El aprendizaje en altos niveles del idioma hablado puede ser expresado de forma más natural y compacta en términos de las sílabas; esto se debe a que las sílabas son relativamente pequeñas en comparación a las palabras y tienen características propias muy marcadas.

Existen, sin embargo, muchas preguntas aún sin responder acerca del comportamiento real de las sílabas en el idioma humano, y muchas dificultades prácticas en el uso de sílabas como unidades básicas en los sistemas de reconocimiento de voz.

### **3.2 LAS SÍLABAS EN EL RECONOCIMIENTO DE VOZ HUMANA**

Una de las tareas prioritarias que existen en los sistemas de reconocimiento es que deben de emular las actividades realizadas por el cerebro humano y de esta manera activar ciertos esquemas de reconocimiento que son propios del sistema, esta tarea sin embargo, aún es una gran utopía por lo que dentro de estos esquemas han existido grandes debates.

Con relación a la inclusión de la sílaba como unidad básica de reconocimiento dentro de la literatura enfocada a ella, podemos considerar que existe un análisis amplio. Su estudio es extensivo y a su vez es beneficioso, el presente capítulo muestra de forma descriptiva y un poco analítica todas sus características e implementaciones.

### **3.2.1 LAS SÍLABAS COMO UNIDADES BÁSICAS**

Una de las cuestiones que deben quedar definitivamente claras dentro del contexto del procesamiento de señales de voz es que ésta por sus características, es no lineal; sin embargo, la percepción de la voz es un proceso altamente dependiente del contexto.

Greenberg indicó que el mecanismo de reconocimiento de voz es un proceso multicapas con docenas de representaciones amplias, que se combinan de una forma no lineal pero con la finalidad de efectuar una robusta y eficiente capacidad de reconocimiento de voz en los seres humanos (Wu et al. 1997).

La mayoría de las personas consideran que ninguna unidad básica de las existentes (basada en fonemas, sílabas o palabras), han llegado a ser una unidad básica para todas las condiciones de audición. En contraparte, los investigadores generalmente creen que unas cuantas representaciones dominan las unidades organizacionales en el sistema de percepción de la voz humana (Wu, 1998) y (Wu et al., 1997).

Los dos mayores competidores para convertirse en estas unidades básicas lo son la sílaba y el fonema, aunque los trifenemas, son otro paradigma que está analizándose actualmente (Rabiner and Biing-Hwang, 1993) y (Rabiner, 1989). A su vez, también se ha llegado a concluir que no existe una unidad integrada que representa de manera total a las palabras, se habla entonces de modelos conjugados donde la sílaba y el fonema pueden aparecer en conjunto (Jones et al., 1999).

Pero existen sin embargo, bastantes pruebas que indican que las sílabas son más representativas que los fonemas. Lo que ha desembocado en el hecho de que muchas propiedades prosódicas tales como el pitch, el acento y el stress, son expresadas de forma más natural por medio de las sílabas. Algunos investigadores consideran que la sílaba es la unidad primaria de segmentación de la voz y la unidad básica del léxico dentro del cerebro (Arai and Greenberg, 1997), (Fujimura, 1975a), (Fujimura, 1975b) y (Lee and Ching, 1998).

Un punto de vista a favor de las sílabas con respecto a las estructuras fonéticas es que son altamente perceptibles dentro del contexto más que los fonemas, por receptores no entrenados.

En este momento deseo hacer un especial énfasis; cuando una persona desea o tiene la necesidad de aprender algún idioma, varias de las formas de lograrlo es por medio de la repetición, o bien del entrenamiento mutuo. Sin embargo, no comienza con aprender un fonema para después conformar una sílaba, después palabra y posteriormente una frase, sino que inconscientemente crea dentro de sí

estructuras más complejas dentro de las cuales, si es usted observador, son las sílabas, pues éstas son el punto intermedio entre los abstractos y nunca considerados en este aprendizaje fonemas y las aliviantes y extensas palabras, lo que se trata, es destacar el aspecto de que la sílaba juega un papel preponderantemente importante en este hecho y es por ello la razón de la necesidad de su estudio, asimismo, la cuestión gramatical interviene también.

En concordancia con lo anterior se ha encontrado que de acuerdo con estudios psicológicos, los niños con problemas de lectura tienden a aminorar este problema, cuando se les presentan los caracteres chinos, esto se debe básicamente al hecho de un mapeo de caracteres a un nivel mayor que el de los fonemas, con lo que se determina que la sílaba ayuda en demasía en las cuestiones de lectura.

A su vez, los niños son capaces de identificar segmentos silábicos desde una edad temprana a diferencia de los fonemas. Mehler, Dommergues y Feauenfelder, denotaron que los niños comienzan a entender al fonema sólo cuando son capaces de reconocer los símbolos del alfabeto correspondiente generando en ellos la asimilación de modelos abstractos, pero antes de ello, ya estos han establecido conversaciones que los involucran y sin necesidad de percatarse de que los están utilizando (Mehler, 1981).

Para el estudio de las sílabas se establece que su importancia radica en los siguientes aspectos:

- El sistema auditivo humano integra duraciones de aproximadamente 200 ms para captar la señal de voz, lo cual tiene una adecuada correlación con la duración de las sílabas. Esto presupone que una percepción humana robusta puede ser modelada adecuadamente si se hace uso de las sílabas en lugar de los fonemas.
- La duración relativa de las sílabas depende en menor medida de las variaciones de la pronunciación que en la duración de los fonemas.
- Se considera que el tiempo de análisis de 250 ms es adecuado para los métodos de extracción de vectores cepstrales. Bajo esta consideración la señal de ruido estacionaria puede ser discriminada con ventajas de la señal de voz no estacionaria. Tiempos de captura más cortos (de 100 ms, por ejemplo) pueden capturar la parte estacionaria de una vocal. Esto muestra la ventaja potencial de la utilización de la sílaba, al usar ventanas entre 200 y 250 ms para cada unidad de clasificación.

Los argumentos antes planteados nos permiten suponer que los sistemas de reconocimiento automáticos del habla basados en sílabas, pueden ser más robustos que los basados en fonemas, especialmente cuando se trabaja con señales de voz espontáneas.

### 3.2.2 IDENTIFICACIÓN DE LA SÍLABA

Uno de los métodos más populares referentes al estudio de unidades de identificación de la voz en los seres humanos, es el paradigma de "reacción en tiempo". Esta orden de experimentos comprende una correlación entre que tan rápido un humano puede reconocer y responder a los estímulos de la voz. Dentro de la gran cantidad de estudios relativos a la polémica de la sílaba y fonema, Massaro encontró que la sílaba era preponderante para este caso (Massaro, 1972). Resultado de estos estudios existe un conjunto de palabras monosilábicas tales como pan, zar, cal, etc.

### 3.2.3 IDENTIFICACIÓN DE LA SÍLABA (CASO PRÁCTICO)

Estudios realizados con anterioridad han demostrado que los seres humanos son sensibles en intervalos de tiempo de 300 msecs, dichos intervalos se relacionan de manera adecuada con el límite superior de la duración de las sílabas. Para el caso específico del español, esto es una variante bastante fuerte, los siguientes datos demuestran la enorme variación que existe al analizar un corpus de voz comprendido por 32 sílabas y 16 palabras, se presenta el análisis de sílabas para un locutor y 320 muestras de las sílabas comprendidas en el siguiente corpus.

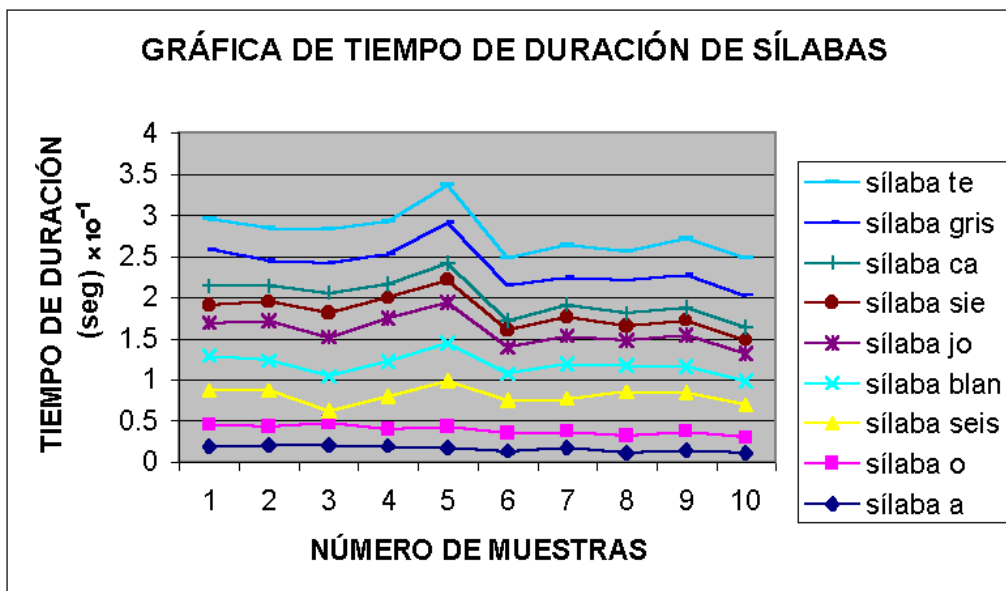


Fig. 3.1 Gráfica de tiempos de duración de sílabas en un corpus de voz.

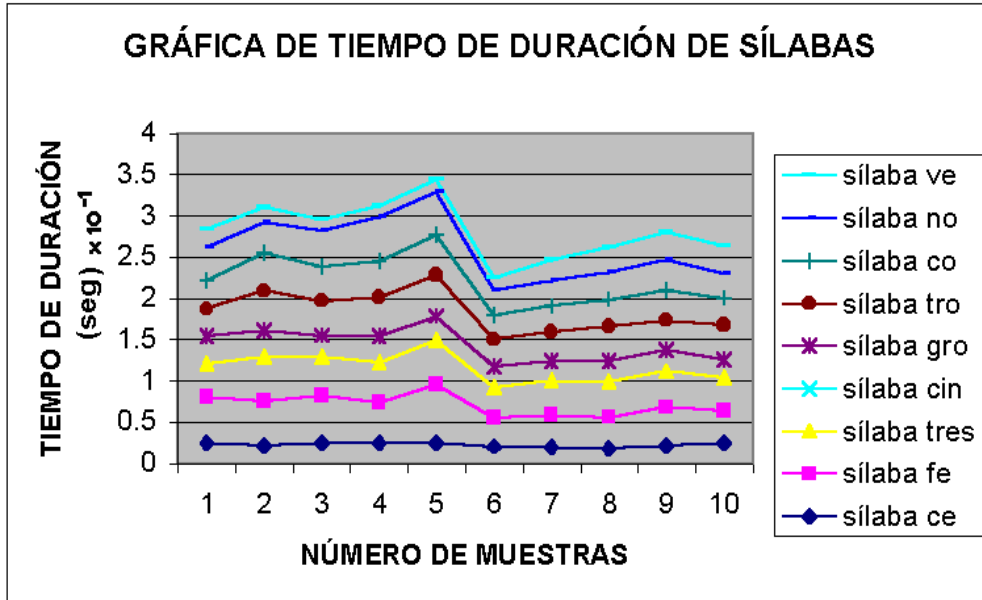


Fig. 3.2 Gráfica de tiempo de duración de sílabas en un corpus de voz (continuación).

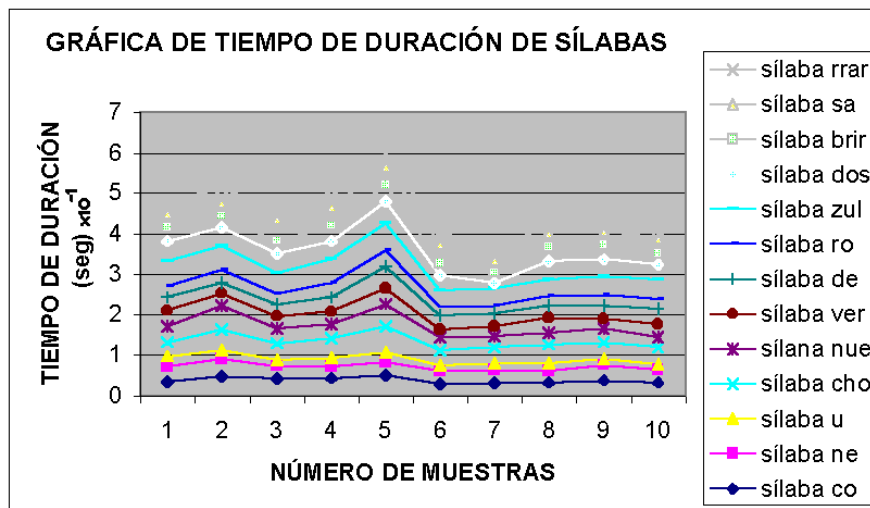


Fig. 3.3 Gráfica de tiempos de duración de sílabas en un corpus de voz (continuación).

Además de los experimentos expuestos con anterioridad se ha demostrado que la duración crítica para que una información sonora pueda ser inteligible es muy similar a la longitud de una sílaba. Toda la serie de experimentos relacionados a la longitud de la sílaba implican que ésta juega un papel muy importante. De los resultados anteriores, se encontró que el promedio de la sílaba resulta ser bastante óptimo para aplicaciones de reconocimiento. Los valores de las gráficas están colocados con el fin de ofrecer claridad al lector evitando su traslape, por lo que no debe confundirse el hecho de que la muestra cinco haya sido siempre la de mayor longitud, sino que por claridad se colocó en esa posición.

### **3.2.4 LAS SÍLABAS EN UN ACCESO LÉXICO**

A lo largo del tiempo los humanos asimilan las palabras como una secuencia de sílabas en lugar de fonemas. Ladefoged (1993) menciona que en la historia de la escritura, muchos idiomas han sido generados a partir de símbolos que representan sílabas, tal es el caso del Japonés. Cabe en estos momentos, sin embargo, hacer la siguiente reflexión, la sílaba es una unidad de voz la cual no tiene una definición satisfactoria. No existe una frontera en la que la sílaba se pueda observar, pero cada una de ellas se encuentra separada y es distinta a las demás.

Las estadísticas anteriores demuestran este hecho, la integración de sílabas en estructuras reafirma este aspecto, pues es obvio que el tiempo de duración de las sílabas de la estructura V (vocal - a,e,i,o,u -) es menor a la de las estructuras (CCVC Consonante Consonante Vocal Consonante - pros, pres, fres, etc. -)

### **3.2.5 DEFINICIÓN DE LA SÍLABA**

La sílaba es un sonido o conjunto de sonidos articulados que constituyen un solo núcleo sónico entre dos depresiones sucesivas de la emisión de la voz. (Diccionario Real Academia Española).

Se ha mencionado durante bastante tiempo que la sílaba se construye alrededor de un núcleo que por lo general es el componente de mayor intensidad dentro de la estructura y generalmente el de única obligatoriedad.

Dentro del estudio lingüístico se considera que la sílaba tiene como elemento principal a una vocal y en sus extremos a algunas consonantes, esto no siempre sucede así, pues en ocasiones las sílabas pueden tener como núcleo una consonante determinada y a los lados una consonante y una vocal. Asimismo, se hace referencia al hecho de que las sílabas representan los picos de sonoridad lo que es análogo a las regiones de mayor energía de sonido y son pensadas para corresponder al núcleo de la sílaba.

En general las personas entienden intuitivamente el concepto de sílaba y pueden además identificar gran parte de sus características dentro de una palabra, pues se pueden identificar una gran cantidad de sílabas y sus fronteras.

La figura 3.4 demuestra el comportamiento de la energía para una muestra de voz determinada, observe el comportamiento de la energía cuando se manifiestan los picos de alta sonoridad de la señal de voz.

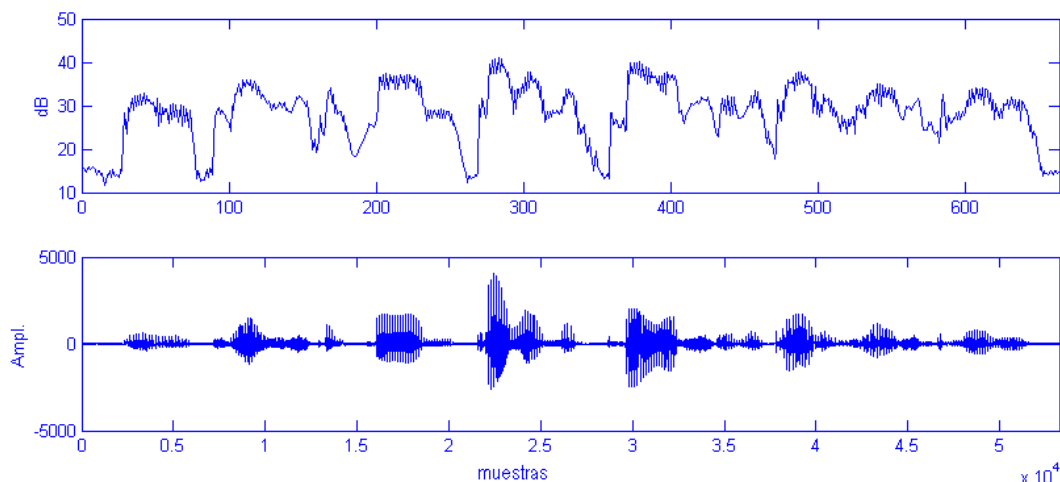


Fig. 3.4 Gráfica del comportamiento de la señal de voz y su energía.

Desde un punto de vista abstracto, una sílaba forzosamente se encuentra constituida por un conjunto de fonemas y tiene manifestaciones acústicas. Una sílaba puede ser analizada en términos de sus propiedades y sonidos que la constituyen, o en términos de su manifestación cuando emerge del hablante. Dentro del español se encuentra una clasificación definida para las sílabas que en él coexisten, tal clasificación a diferencia de otros idiomas se encuentra bien enmarcada por estudiosos de la lengua.

### 3.3 LAS REGLAS DE LA SÍLABA

Palabras monosilábicas:

Son las que están formadas por una sílaba: luz, mar.

Palabras bisilábicas: Son las que están formadas por dos sílabas: silla, mesa.

Palabras trisilábicas: Son las que están formadas por tres sílabas: ventana, cabeza.

Palabras polisilábicas: Son las que están formadas por cuatro o más sílabas: Argentina, Polideportivo.

En (Feal, 2000) se menciona que en el español existen 27 letras, las cuales están clasificadas de acuerdo a su pronunciación en dos grupos: vocales y consonantes. En la figura 3.5 se muestra un esquema de la clasificación de las letras para una mejor comprensión. El grupo de las vocales está formado por cinco, su pronunciación no dificulta la salida del aire. La boca actúa como una caja de resonancia abierta en menor o mayor grado y de acuerdo a esto, las vocales se clasifican en abiertas, semiabiertas y cerradas (Oropeza, 2000) y (Rabiner and Biing-Hwang, 1993).

El otro grupo de letras, las consonantes, está formado por veintidós letras, con las cuales se forman tres son consonantes compuestas, llamadas así, por ser letras simples en su pronunciación y dobles en su escritura. Las letras restantes son llamadas consonantes simples, por ser simples en su pronunciación y en su escritura.

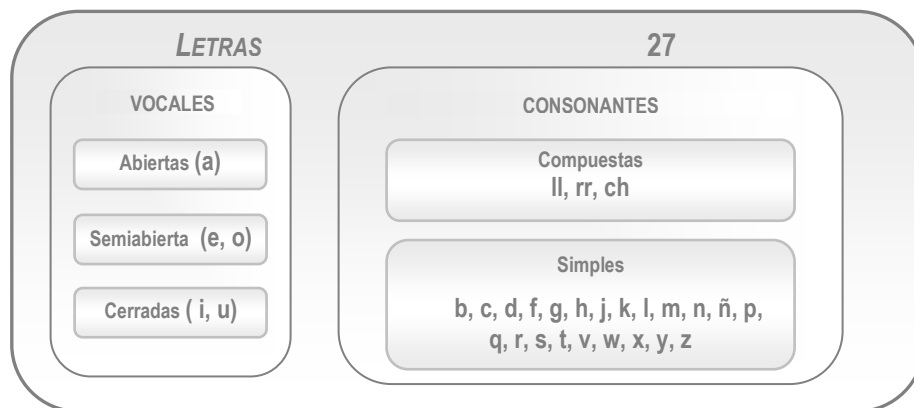


Fig. 3.5 Esquema de clasificación de letras del alfabeto del español.

### 3.3.1 Reglas del idioma español para la formación de sílabas

Para expresar las reglas con claridad se utiliza la notación mostrada en la tabla 3.1.

Símbolo	Descripción
+	Utilizado para concatenar sílabas
( )	Utilizada para agrupación
-	Indican las letras donde se aplicará la regla en cuestión
	Utilizada para establecer posibilidades alternativas

Tabla 3.1. Notación utilizada para expresar las reglas de división silábica.

En el idioma español existen diez reglas, las cuales determinan la separación de las sílabas de una palabra. Estas reglas son listadas a continuación mostrando enseguida excepciones a la misma.

#### REGLA 1

En las sílabas, por lo menos, siempre tiene que haber una vocal. Sin vocal no hay sílaba.



Excepción. Esta regla no se cumple cuando se presenta la “y”.

## REGLA 2

Cada elemento del grupo de consonantes inseparables, mostrado en la figura 3.6, no puede ser separado al dividir una palabra en sílabas.

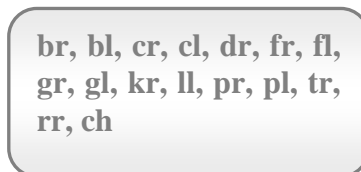


Fig. 3.6. Grupo de consonantes inseparables.

## REGLA 3

Cuando una consonante se encuentra entre dos vocales, se une a la segunda vocal.

## REGLA 4

Cuando hay dos consonantes entre dos vocales, cada vocal se une a una consonante.

Excepción: Esto no ocurre en el grupo de consonantes inseparables (regla 2).

## REGLA 5

Si son tres las consonantes colocadas entre dos vocales, las dos primeras consonantes se asociarán con la primera vocal y la tercera consonante con la segunda vocal.

Excepción. Esta regla no se cumple cuando la segunda y tercera consonante forman parte del grupo de consonantes inseparables.

## REGLA 6

Las palabras que contienen una h precedida o seguida de otra consonante, se dividen separando ambas letras.

## REGLA 7

El diptongo es la unión inseparable de dos vocales. Se pueden presentar tres tipos de diptongos posibles:

- Una vocal abierta + Una vocal cerrada.
- Una vocal cerrada + Una vocal abierta.
- Una vocal cerrada + Una vocal cerrada.

Son diptongos sólo las siguientes parejas de vocales: ai, au, ei, eu, io, ou, ia, ua, ie, ue, oi, uo, ui, iu, ay, ey, oy .

La unión de dos vocales abiertas o semiabiertas no forma diptongo, es decir, deben separarse en la segmentación silábica. Pueden quedar solas o unidas a una consonante.

#### REGLA 8

La h entre dos vocales, no destruye un diptongo.

*Ejemplo:*

Palabra	sílabas
ahuyentar ↑↑↑	ahu + yen + tar

#### REGLA 9

La acentuación sobre la vocal cerrada de un diptongo provoca su destrucción.

Palabra	sílabas
María ↑↑	Ma + rí + a

#### REGLA 10

La unión de tres vocales forma un triptongo. La única disposición posible para la formación de triptongos es la siguiente:

Vocal cerrada + (vocal abierta | vocal semiabierta) + vocal cerrada

Sólo las siguientes combinaciones de vocales, forman un triptongo: iai,iei, uai, uei, uau, iau, uay, uey.

### 3.4 NÚMERO DE SÍLABAS

Dentro de la lingüística destinada al español se ha realizado un análisis que corresponde a la forma en la cual se presentan los diferentes elementos que lo constituyen, para el caso específico de los fonemas y su consecuente aparición en las frases de este idioma se muestra la tabla 3.2 que muestra la manera en la cual se comporta este elemento. Cabe destacar que este análisis demuestra la ponderación que tiene algunos fonemas en el idioma (Lizana et al., 2000).

LETRA	FREC. REL.	CANTIDAD DE INF.	PARA EL PROMEDIO
A	0.111	3.171	0.352
B	0.017	5.878	0.100
C	0.060	4.059	0.244
D	0.053	4.238	0.225
E	0.149	2.747	0.409
F	0.004	7.966	0.032
G	0.011	6.509	0.072
H	0.004	7.966	0.032
I	0.072	3.796	0.273
J	0.002	8.966	0.018
L	0.052	4.265	0.222
LL	0.002	8.966	0.018
M	0.033	4.921	0.162
N	0.059	4.083	0.241
Ñ	0.001	9.966	0.010
O	0.095	3.396	0.323
P	0.029	5.108	0.148
Q	0.012	6.381	0.077
R	0.054	4.211	0.227
S	0.074	3.756	0.278
T	0.020	5.644	0.113
U	0.030	5.059	0.152
V	0.015	6.059	0.091
X	0.002	8.966	0.018
Y	0.007	7.158	0.050
Z	0.006	7.381	0.044
	ENTROPIA		3.827

Tabla 3.2. Frecuencia de aparición de fonemas en textos en español.

Para el caso que nos interesa en este estudio se cuenta con el siguiente análisis que identifica la manera en la que determinada sílaba se presenta en el español (Lizana et al., 2000).

SILABA	FR	SILABA	FR	SILABA	FR	SILABA	FR
A	0.0191	CRE	0.0006	GAR	0.0012	ME	0.0049
AC	0.0074	CRI	0.0012	GE	0.0012	MEN	0.0093
AL	0.0037	CU	0.0012	GEN	0.0012	MER	0.0012
AM	0.0012	CUA	0.0012	GER	0.0006	MI	0.0080
AN	0.0006	CUAL	0.0006	GI	0.0012	MIS	0.0031
AR	0.0012	CUAN	0.0006	GIR	0.0006	MO	0.0105
AS	0.0012	CUEN	0.0006	GO	0.0043	MOS	0.0068
AU	0.0006	CUES	0.0012	GRA	0.0006	MU	0.0012
BA	0.0031	CUNS	0.0006	GRAN	0.0006	MUL	0.0016
BAR	0.0012	DA	0.0068	GUIR	0.0006	MUN	0.0012
BE	0.0037	DAD	0.0037	GUN	0.0019	MUY	0.0006
BEN	0.0006	DAN	0.0037	HA	0.0019	NA	0.0123
BI	0.0031	DAS	0.0019	HAY	0.0006	NAL	0.0012
BIEN	0.0012	DE	0.0530	HE	0.0019	MUY	0.0006
BLA	0.0019	DEL	0.0012	HI	0.0006	MUN	0.0012
BLAN	0.0006	DEN	0.0025	HIS	0.0006	NAS	0.0025
BLE	0.0025	DES	0.0019	HO	0.0006	NE	0.0068
BLES	0.0012	DI	0.0049	HOM	0.0012	NEN	0.0006
BO	0.0006	DIA	0.0006	HU	0.0012	NER	0.0019
BON	0.0006	DIAL	0.0006	I	0.0012	NES	0.0111
BOOM	0.0006	DIEN	0.0012	IM	0.0006	NI	0.0031
BRA	0.0006	DIO	0.0019	IN	0.0093	NO	0.0111
BRE	0.0037	DIR	0.0069	JA	0.0019	NOS	0.0037
BRES	0.0006	DO	0.0086	JAR	0.0006	NU	0.0006
BRIN	0.0006	DOS	0.0037	JE	0.0006	NUE	0.0019
BRIR	0.0006	DRA	0.0006	JEM	0.0006	O	0.0080
C	0.0012	DRAS	0.0006	JER	0.0006	OB	0.0012
CA	0.0210	DRI	0.0006	JO	0.0006	OR	0.0006
CAR	0.006	DRO	0.0006	JUI	0.0006	OX	0.0006
CAS	0.0019	DU	0.0012	JUN	0.0006	PA	0.0062
CE	0.0037	DUC	0.0043	JUS	0.0056	PAR	0.0006
CEN	0.0006	E	0.0234	LA	0.0308	PE	0.0019
CEP	0.0025	EL	0.0130	LAS	0.0130	PEC	0.0019
CER	0.0012	EM	0.0025	LAR	0.0006	PEN	0.0012
CES	0.0019	EN	0.0117	LE	0.0019	PER	0.0025
CHO	0.0006	ES	0.0148	LEC	0.0006	PERS	0.0006
CI	0.0049	EX	0.0031	LES	0.0025	PIE	0.0006
CIA	0.0019	FA	0.0019	LI	0.0062	PIO	0.0006
CIAL	0.0006	FE	0.0012	LIO	0.0006	PLE	0.0012
CIAS	0.0031	FEC	0.0043	LIS	0.0006	PLI	0.0012
CIE	0.0031	FI	0.0068	LLA	0.0012	PLO	0.0006

CIEN	0.0031	FIAN	0.0019	LLAS	0.0012	PO	0.0105
CIO	0.0130	FIN	0.0006	LLO	0.0025	POR	0.0080
CION	0.0099	FO	0.0006	LO	0.0074	POS	0.0012
CIOS	0.0006	FOR	0.0056	LOS	0.0105	PRAC	0.0037
CIR	0.0006	FRAC	0.0006	LU	0.0012	PRE	0.0031
CO	0.0154	FRE	0.0006	LUE	0.0006	PRI	0.0012
COM	0.0006	FU	0.0006	MA	0.0099	PRO	0.0086
CON	0.0142	FUE	0.0019	MAL	0.0012	PU	0.0006
COR	0.0006	FUN	0.0006	MAN	0.0012	PUE	0.0019
COS	0.0006	GA	0.0012	MAS	0.0068	PUN	0.0006

Tabla 3.3. Frecuencia de pronunciación de las sílabas.

Dentro de los elementos necesarios para alcanzar los objetivos del presente trabajo, se vio la necesidad de crear un Sistema Experto capaz de determinar el número de sílabas, su tipo y características de un texto determinado y del conjunto de corpus que se analizaron, cuyas características serán descritas en el capítulo 5.

Dicho Sistema Experto es alimentado por el conjunto de frases, palabras o sílabas del corpus en cuestión y se encarga de aplicar las reglas antes mencionadas, para con ello, determinar las características referentes a sílabas de dicho corpus.

El estudio más importante relacionado a las sílabas dentro del español es el de Feal Pinto(2000) de la Universidad de Valladolid, en donde se propone utilizarlas como una unidad de síntesis para el español.

Se utilizó el método de análisis de textos en su trabajo, que consiste de cuatro etapas: obtención de los textos, extracción de palabras, transcripción fonética y división en sílabas.

- Los textos se obtuvieron del Proyecto Gutenberg: <http://www.promo.net/pg/>, de las Obras de Miguel de Cervantes Saavedra: <http://cervantes.alcala.es/obras.htm>, etc.
- Se efectuó la extracción de palabras y la transcripción fonética se realizó usando el alfabeto fonético SAMPA.
- La división en sílabas la realizó un autómata.
- Las conclusiones fueron la obtención de un total de 1456067 sílabas, de las cuales 1894 eran distintas. Algunas no se podrían dar en el español pues procedían de palabras del inglés. Al final se detectaron un total de 1641 sílabas distintas.

A continuación, se muestra el caso de estudio realizado por el Sistema Experto a un conjunto de dos textos de índole científico, los textos fueron extraídos del Congreso Internacional de Instrumentación y Sistemas Digitales del año 2002.

La gráfica resalta la aparición de las estructuras silábicas por su importancia, haciendo especial hincapié en el hecho de que las estructuras CV preponderan en mayor grado que las otras, el mismo caso se presentó al analizar otros corpus diferentes.

## PROPORCIÓN DE ESTRUCTURAS SILÁBICAS EMPLEADAS EN 10 ARTÍCULOS CIENTÍFICOS DEL CIGC INDI 2002

V	VC	VV	VCC	C	CV	CVC	CVCC	CVV	CVVC	CVVV	CVVVC	CCV	CCVC	CCVCC	CCVV	CCVVC	T
217	459	1	2	62	2475	936	11	182	270	2	9	243	69	2	1	3	4944

V	VC	VV	VCC	C	CV	CVC	CVCC	CVV	CVVC	CVVV	CVVVC	CCV	CCVC	CCVCC	CCVV	CCVVC	T
202	309	3	0	41	1739	939	7	140	142	0	1	207	38	0	5	1	3774

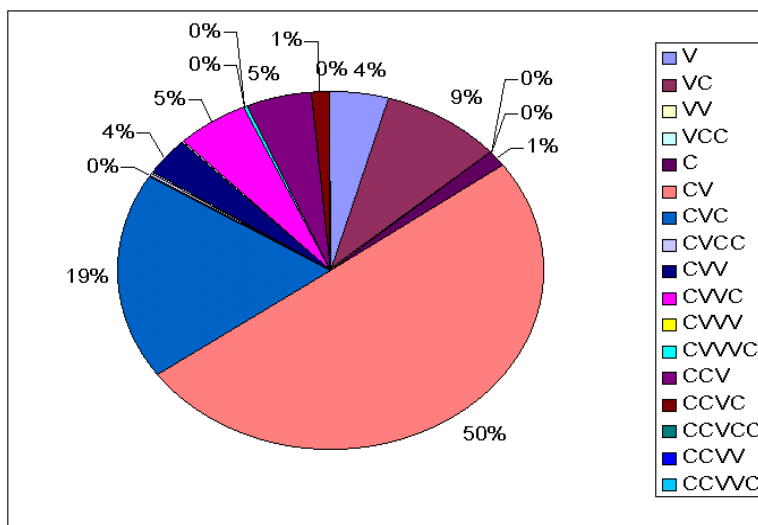


Fig. 3.7 Gráfica del estudio de las estructuras silábicas en textos científicos.

Se obtuvieron los siguientes resultados, tras de la aplicación de dicho sistema, para algunos locutores del corpus del Latino-40 (Martínez, 2002) y (Bernstein, 1994).

# ANÁLISIS DEL LATINO-40, UN CORPUS MÁS GRANDE

V	VC	VV	VCC	C	CV	CVC	CVCC	CVV	CVVC	CVVVC	CCV	CCVC	CCVCC	CCVV
52	12	2	0	0	282	46	0	32	2	0	8	6	0	0

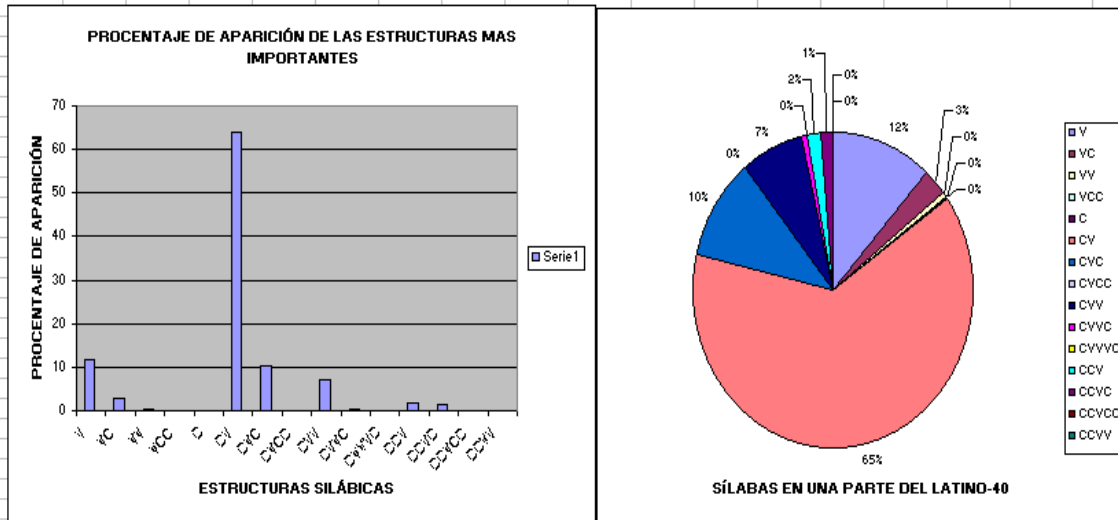


Fig. 3.8 Análisis de estructuras silábicas en un locutor del Latino-40.

Las definiciones de la frontera de las sílabas son tan difíciles de encontrar como la definición de la sílaba misma. Es claro que todo idioma posee una gran cantidad de sílabas, en el caso específico del español se cuentan con alrededor de 1500 de estas estructuras (esenciales), aunque el número dinámico de las mismas se encuentra oscilando entre las 8000, de esas 1500 alrededor de 1300 son usadas en un 92% en las conversaciones diarias y por un individuo normal. De lo anterior, se infiere que el número de sílabas es mucho más pequeño que el número de palabras pero mucho más grande que el de los fonemas (Feal, 2000).

De acuerdo a (Wu, 1998) se realizó un trabajo en donde el número de unidades silábicas es mucho más grande que el número de fonemas independientes del contexto con buenos resultados, lo que demostró que se pueden utilizar en sistemas de reconocimiento de dígitos e incluso en sistemas de tareas de conversación de voz.

Históricamente la cantidad enorme de sílabas existentes ha sido la oposición a su estudio más amplio e implementación en sistemas de reconocimiento automático de voz. Dichos sistemas que se implementan para tal fin hacen uso de los llamados trifonemas, los cuales son tan numerosos como las sílabas. Los sistemas basados en trifonemas tienen muchos cientos de modelos, por ejemplo el sistema HTK de la Universidad de Cambridge para el Wall Street Journal y el sistema de Dragon para Switchboard (Peskin, 1997) y (Hauenstein, 1996).

En la juventud se tiene un sistema muy dependiente de los trifenemas. Young menciona que se requieren de un total de 60000 de ellos para completar el total del vocabulario empleado en este nivel (Young, 1995).

De todas las especulaciones mencionadas, el número de sílabas dentro de cierto rango de elementos esperado puede variar en distintos aspectos, costumbres, área de desarrollo, extroversión, introversión, etc.

Además de los aspectos antes mencionados, una frase o conjunto de palabras usadas en un lugar determinado son propios de ese entorno por lo que dependiendo de los resultados y de a quién vaya dirigido el sistema creará la cantidad de elementos que debe de tener.

El análisis del número de sílabas para el español es de 1500 máximo y 800 mínimo para los trabajos de los ASR. Además, a lo largo de los análisis desarrollados, la duración promedio de una sílaba en el español es de aproximadamente de entre 200 y 350 milisegundos.

### 3.5 LAS SÍLABAS EN CONVERSACIONES DE VOZ

Como es sabido, las bases de datos que conforman los corpus de voces son herramientas para realizar estudios de las diferentes frases que las componen.

Para el caso específico de la sílaba, existen reportes que el comportamiento que generan dichos sistemas con relación a la cantidad de sílabas expuestas y el total del vocabulario toman una forma aproximadamente exponencial como se expone en la figura 3.9 (Wu, 1998).

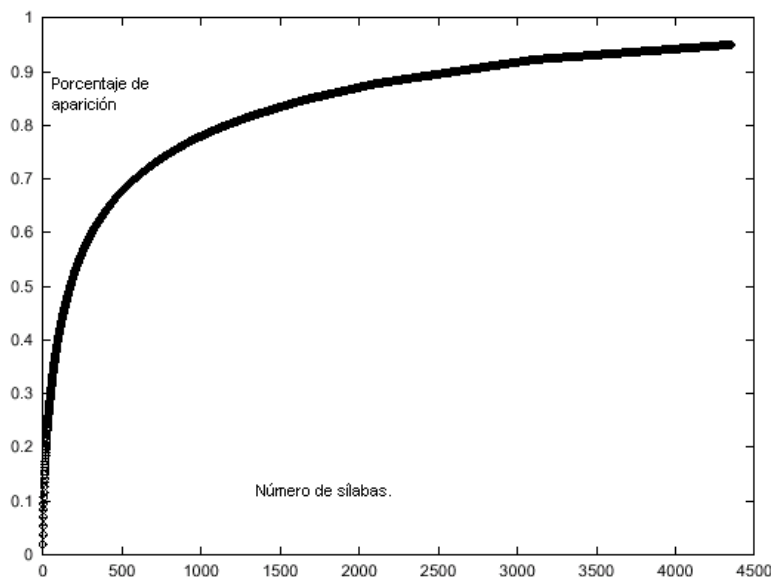


Fig. 3.9 Gráfica del comportamiento de las sílabas en general.



Como se puede observar en la figura 3.9, cuando se trabaja en una aplicación específica el número de sílabas se reduce y se alcanza mucho más rápido el vocabulario que se está analizando. De hecho en la figura 3.8 se observa que con pocas sílabas, se alcanza un alto porcentaje de los elementos existentes en la aplicación.

Gran parte de los estudios enfocados en las sílabas recaen en estos parámetros, algunos otros estudios realizados por Kirchoff en el idioma Alemán, Arai y Greenberg en el Japonés (Arai and Greenberg, 1997), (Fujimura, 1975a), (Fujimura, 1975b) y (Lee and Ching, 1998) así lo demuestran.

### 3.6 LAS SÍLABAS EN LOS SRAH

Los ASR's actuales son implementados usando el fonema como parte fundamental, gran parte de ellos se encuentran basados en los HMM's que se concatenan como se muestran en la figura 3.10 (Savage, 1995) para obtener palabras ó frases:

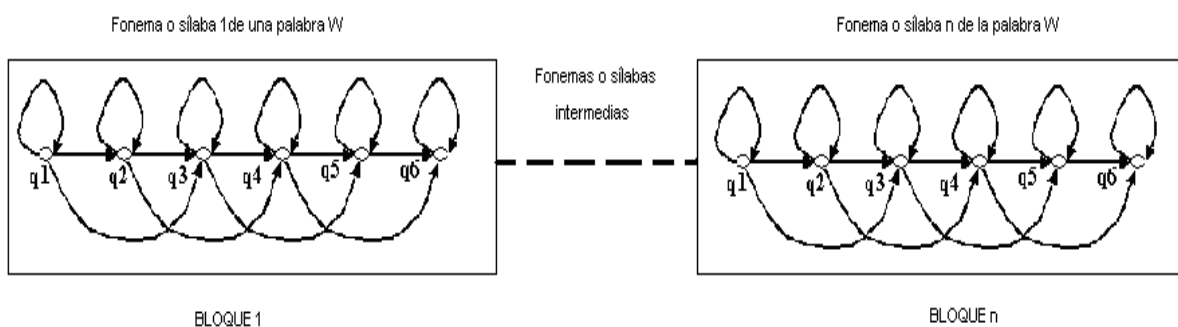


Fig. 3.10 Gráfica de Cadenas Ocultas de Markov aplicadas a sílabas y fonemas.

Los requerimientos actuales dentro del campo del reconocimiento de voz convergen en incrementar la fiabilidad del tipo de sistema que se utilice, en cuanto a los elementos existentes actualmente para realizar un reconocimiento de manera óptima, estos han encontrado en el fonema la base de su modelación, sobretodo en sistemas de reconocimiento con un vocabulario extenso y por ende aplicados a las necesidades del sistema del habla continua.

Sin embargo, dados los intereses creados y las pruebas a las que actualmente se ha llegado, se deduce que un sistema de reconocimiento del habla basado en fonemas, aún no reúne los elementos de reconocimiento total que se requieren, de ahí la propuesta del uso de las estructuras como las sílabas en un intento de mejorar y conseguir lo anterior.

A continuación se muestra la propuesta del esquema básico de un SARH del habla discontinua en etapa de entrenamiento basado en estos elementos de reconocimiento y usando vectores de codificación predictiva lineal:

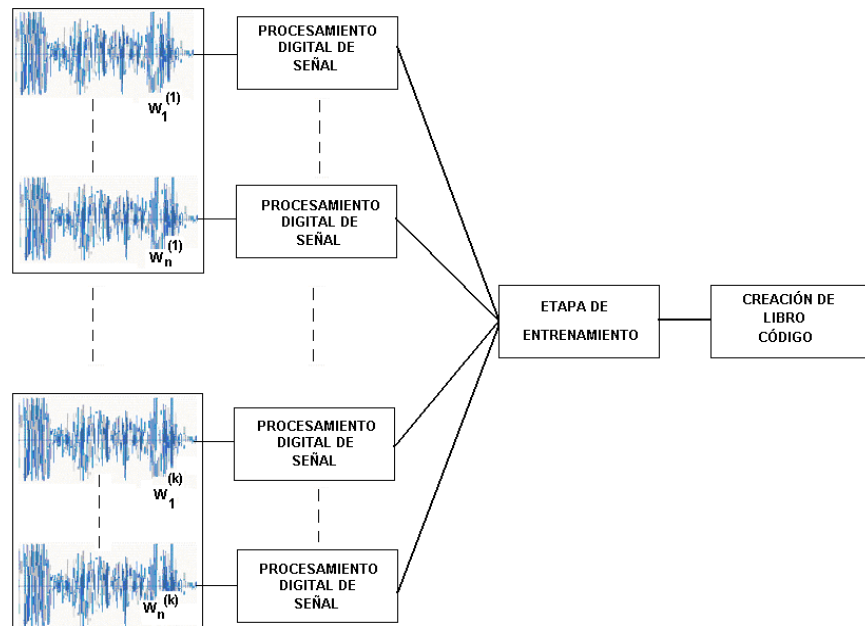


Fig. 3.11 Gráfica del sistema de reconocimiento aislado usando sílabas.

De la figura 3.11 y para fines de estudio se deduce que debe de existir una base de datos que contenga las unidades básicas del idioma a reconocer, es aquí en donde se empleará la sílaba como parte elemental dentro del español. El número de sílabas máximo como se mencionó ya con anterioridad es de 3000 como límite adecuado.

Dentro de la lingüística se tienen 2 casos especiales de sílabas dentro de las cuales se denotan las siguientes:

- a)- Sílaba tónica.- que es en donde recae la fuerza de la pronunciación.
- b)- Sílaba atónica.- que es la de mayor intensidad pero siempre acompañada de una sílaba tónica.

El sistema de reconocimiento propuesto contendrá como elementos cruciales, el bloque de composición de sílabas, encargado básicamente de lo siguiente.

- Comparar los vectores espectrales con los modelos de unidades básicas que para este caso son las sílabas, éste se encarga básicamente de adjudicar a cual de los elementos dentro del vocabulario pertenece la señal que se está analizando. Una vez clasificada dicha unidad, se procede a asignar cierta coherencia en la estructura gramatical de la cual estamos deseando construir nuestros elementos de referencia, de acuerdo a la figura 3.11, se tiene la necesidad de primeramente extraer del léxico la pronunciación de cada palabra en términos de sus unidades básicas y posteriormente se

concatenan los modelos de estas unidades para formar los modelos de las palabras pertinentes (Savage, 1995).

Una de las ventajas de usar los fonemas y su razón de existir, es que estos son los responsables de que la palabra "paz" y "pan" se escuchen de diferente forma debido a la aparición de los fonemas /z/ y /n/.

### **3.7 RESUMEN DEL CAPÍTULO**

A lo largo del presente capítulo se describieron las características esenciales que las sílabas poseen dentro del español, se analizaron los tiempos de duración de las mismas, poniendo de manifiesto que las diferentes sílabas analizadas poseen diferentes tiempos de aparición.

El análisis de las reglas de las sílabas en el español, generó el interés de realizar la creación de un Sistema Experto destinado a realizar las tareas de segmentación y estudio de las diferentes estructuras analizadas.

Finalmente, se hace una recopilación de los diferentes resultados obtenidos a lo largo del tiempo en lo referente a las aplicaciones de las sílabas en los sistemas de reconocimiento del habla.

Se encontró que la cantidad máxima de sílabas a utilizar para los trabajos de reconocimiento de voz en español es en promedio de 3000 unidades.

En este capítulo se analizaron los lineamientos y aspectos esenciales de la integración de la sílaba a los ASR's, así como también, sus beneficios y perjuicios de éstos.

# CAPÍTULO 4

---

## Herramientas matemáticas.

Uno de los elementos esenciales en la tarea de reconocimiento de voz es el proceso matemático que se aplica a la señal. Para los fines específicos del presente trabajo, se utilizan dos diferentes herramientas de reconocimiento, en el caso del reconocimiento del habla no continua, se comprueba el estudio de las sílabas con los métodos de Codificación Predictiva Lineal (CPL) y Cadenas Ocultas de Markov (COM). Para el análisis del reconocimiento de voz del habla continua, se hace uso de las Cadenas Ocultas de Markov de Densidad Continua (COMDC o CDHMM por sus siglas en inglés).

En este capítulo se muestran los fundamentos de cada una de éstas herramientas para posteriormente ser utilizadas en los procesos de reconocimiento de los capítulos posteriores.

Posteriormente se desarrollará el análisis de los parámetros para dos aplicaciones: 1) encontrando los parámetros de una mixtura de densidad de Gaussianas, 2) encontrando los parámetros de un Modelo Oculto de Markov (HMM) (dados por el algoritmo de Baum-Welch) para tanto modelos de observación discretos y de mixturas de Gaussianas. Se derivará la actualización de las ecuaciones de una forma detallada pero sin realizar pruebas sobre las propiedades de convergencia.

## 4.1 CODIFICACIÓN PREDICTIVA LINEAL

Un perfil espectral determina el estado de resonancia que guarda el tracto vocal del sistema productor de voz en un momento dado (Rabiner and Biing-Hwang, 1993), podríamos asegurar que son las características de un determinado fenómeno que cambia de manera abrupta con el tiempo, pero que es posible descomponerlo en instantes breves, donde sus características permanecen estáticas. Haciendo la analogía con la reproducción de una cinta fílmica, ésta consta de un conjunto de pequeños trozos de imágenes captadas en breves instantes de tiempo, éstas son estáticas, sin embargo la presentación continua de estas imágenes ante los ojos del ser humano a una velocidad determinada, provocan que éste los vea en total y completo movimiento (Oropeza, 2000).

## 4.2 CARACTERIZACIÓN DE LOS PARÁMETROS DE LA CODIFICACIÓN PREDICTIVA LINEAL

Dicho estado de resonancia está completamente caracterizado por los parámetros LPC que intervienen en la función de transferencia del tracto vocal. Si la función de transferencia del tracto vocal está dada por  $G/A(z)$ , entonces los ceros de  $A(z)$  indican las frecuencias de resonancia. Los parámetros LPC son los coeficientes de  $A(z)$  y usando el principio de mínima acción se pueden caracterizar como sigue:

De todas las posibilidades para coeficientes de  $A(z)$  se va a tomar la que utilice una excitación  $U(z)$  de energía mínima para producir la señal de voz  $X(z)$  (Barrón, 1998).

$$X(z) = \left[ \frac{1}{A(z)} \right] U(z) \quad (4.1)$$

De acuerdo al Teorema de Parseval, la energía de  $U(z)$  está dada por:

$$E = E[U(z)] = \frac{1}{2\pi} \int_{-\pi}^{\pi} |U(e^{j\omega})|^2 d\omega = \sum_{n=-\infty}^{\infty} |u(n)|^2 \quad (4.2)$$

Es de notar que la mayoría de las leyes de la física, si no es que todas, se pueden obtener usando este principio de economía de la naturaleza (Barrón, 1998).

En términos fisiológicos, la forma del tracto vocal es la que determina sus propiedades de resonancia y esta forma de acuerdo al principio de mínima acción es la que produce una señal de voz dada con una excitación de energía mínima.

Otra forma de caracterizar los parámetros LPC es como sigue:  
Despejando la excitación  $U(z)$  de (1) se obtiene:

$$U(z) = A(z)X(z) \quad (4.3)$$

con

$$A(z) = \sum_{k=0}^m a_k z^{-k} \quad \text{y} \quad a_0=1$$

entonces  $U(z)$  es el error predictivo hacia adelante y la energía de  $U(z)$  es el error cuadrático total del predictor.

$$u(n) = \sum_{k=0}^m a_k x(n-k) = x(n) - \left[ -\sum_{k=1}^m a_k x(n-k) \right] \quad (4.4)$$

donde:

$$-\sum_{k=1}^m a_k x(n-k) \quad \text{es un predictor lineal de } x(n)$$

La condición de optimización para los parámetros del predictor lineal es que el error cuadrático total sea mínimo.

Por lo tanto, si hacemos  $a_0 = 1$  los coeficientes del polinomio en el denominador de la función de transferencia del tracto vocal son los parámetros del predictor lineal óptimo (Parámetros LPC).

### 4.3 CÁLCULO DE LOS PERFILES ESPECTRALES

Como se vio anteriormente, para encontrar los parámetros que determinan los perfiles espectrales, basta con encontrar los coeficientes del predictor lineal óptimo de la señal de voz. El cálculo de los coeficientes predictivos se facilita si el problema se plantea en un contexto matemático adecuado (Barrón, 1998).

Considérese el error cuadrático total de un predictor lineal:

$$\sum_{n=-\infty}^{\infty} e^2(n), \quad e(n) = x(n) - \text{predictor}[x(n)],$$

$$\text{predictor}[x(n)] = -\sum_{k=1}^m a_k x(n-k)$$

tomando  $a_0=1$ ,  $e(n)$  se puede escribir como:

$$e(n) = \sum_{k=0}^m a_k x(n-k)$$

(4.5)

por lo que:

$$E = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{n=-\infty}^{\infty} \left( \sum_{i=0}^m a_i x(n-i) \right) \left( \sum_{j=0}^m a_j x(n-j) \right)$$

$$= \sum_{i=0}^m a_i \sum_{j=0}^m a_j \sum_{n=-\infty}^{\infty} x(n-i)x(n-j) = \sum_{i=0}^m a_i \sum_{j=0}^m a_j r_x(|i-j|)$$

donde

$$r_x(|i-j|) = \sum_{n=-\infty}^{\infty} x(n-i)x(n-j)$$

Es la autocorrelación de  $x(n)$  evaluada en  $|i-j|$  esto en forma matricial se puede escribir como:

$$E = a^t R_x a \quad (4.6)$$

con  $a^t = (a_0, a_1, \dots, a_m)$ ,  $R_x$  es la matriz de autocorrelación  $r(|i-j|)$   $0 \leq i, j \leq m$  y  $E$  es el error cuadrático total predictivo.

#### 4.4 ALGORITMO DE LEVINSON-DURBIN

Una manera de encontrar de forma práctica los parámetros LPC es usando el algoritmo de Levinson-Durbin. Este algoritmo se basa en una condición de ortogonalidad que deben cumplir dichos coeficientes.

En (Bernal et al., 2000) se encontró que entre las ventajas que presenta el método de predicción lineal se encuentran:

Los parámetros obtenidos mediante predicción lineal muestran un espectro suavizado que proporciona la información más representativa de la voz.

LPC proporciona un modelo adecuado de la señal de voz y sus parámetros se ajustan a las características del tracto vocal, especialmente en los sonidos sonoros del habla cuyas propiedades se aproximan más a la señal estacionaria que en los sonidos sordos.

LPC es un método preciso, muy adecuado para computación, tanto por su sencillez como por la rapidez de ejecución.

Como se menciona en (Rabiner and Biing-Hwang, 1993) la matriz de autocorrelación es una matriz Toeplitz, la cual puede ser resuelta por el algoritmo de Levinson-Durbin. De esta forma se logra obtener la conversión de coeficientes de autocorrelación a un conjunto de parámetros LPC. El algoritmo se presenta de la ecuación 4.7 a la 4.11.

$$E^{(0)} = r(0) \quad (4.7)$$

$$k_i = \left\{ r(i) - \sum_{j=1}^{L-1} a_j^{(i-1)} r(|i-j|) \right\} / E^{(i-1)}, \quad 1 \leq i \leq k \quad (4.8)$$

$$a_i^{(i)} = k_i \quad (4.9)$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)} \quad (4.10)$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)} \quad (4.11)$$

El conjunto de ecuaciones anteriores se resuelve de manera recursiva para  $i=1,2,\dots,k$ , los índices entre paréntesis indican la iteración actual o la anterior, la solución final esta dada por:

$$a_m = \text{Coeficientes LPC} = a_m^{(k)}, \quad 1 \leq m \leq k$$

Los coeficientes LPC obtenidos son un conjunto de vectores que nos proporcionan un modelo apropiado de la señal de voz porque sus parámetros se ajustan a las características del tracto vocal.

#### 4.5 COMPARACIÓN Y SEMEJANZA DE PERFILES ESPECTRALES

Cuando se tienen dos perfiles espectrales diferentes, muchas veces es necesario compararlos entre sí o medir su parecido respecto a un tercero, para esto es necesario contar con una medida de distorsión o semejanza, que efectivamente dados dos perfiles espectrales nos proporcione un valor numérico proporcional a su parecido, respecto a los atributos que nos interesa medir (Barrón, 1998). Una vez identificado el o los atributos a medir y propuesta una función que los compare, dicha función de semejanza debe tener las siguientes propiedades.

- a) Los valores que arroja deben tener correspondencia con el parecido real de los objetos comparados.
- b) Se debe poder evaluar en términos prácticos, es decir para empezar debe ser realizable y de preferencia en el menor tiempo posible.



Para el caso de la comparación de los perfiles espectrales se propone una medida de Itakura-Saito modificada, esta medida de distorsión está basada en el producto interior entre dos perfiles espectrales ya visto anteriormente y se define como:

$$d(\bar{a}, a) = \frac{\bar{a}^t R_x \bar{a}}{a^t R_x a} \quad (4.12)$$

donde,  $R_x$  es la matriz de autocorrelación de  $x(n)$  y  $a$  es su perfil espectral óptimo como  $\bar{a}$  no es óptimo para  $x(n)$  se tiene:

$$a^t R_x a \leq \bar{a}^t R_x \bar{a} \quad \text{por lo que} \quad d(\bar{a}, a) = \frac{\bar{a}^t R_x \bar{a}}{a^t R_x a} \geq 1$$

Si tomamos el logaritmo del lado derecho tendríamos:

$$\log d(\bar{a}, a) = \log(\bar{a}^t R_x \bar{a}) - \log(a^t R_x a) \quad (4.13)$$

Lo que es más compatible con nuestro concepto de comparación. Una primera observación es que esta medida no es simétrica por lo que hay que tener cuidado con el orden en que aparecen los perfiles a comparar. Sin embargo la matriz de autocorrelación es positiva definida, por lo que la medida de distorsión siempre es positiva (Barrón, 1998):

$$\begin{aligned} d(\bar{a}, a) &> 0 \\ d(\bar{a}, a) &\neq d(a, \bar{a}) \end{aligned}$$

para dos perfiles espectrales arbitrarios  $a$  y  $\bar{a}$ .

Para los fines que nos interesan en este trabajo ni siquiera es necesario el cálculo del denominador en (4.12); por lo que se tomará como la medida de distorsión a:

$$d(\bar{a}, a) = \bar{a}^t R_x \bar{a} \quad (4.14)$$

Esta medida de distorsión en realidad lo que mide es la eficiencia del perfil espectral respecto a la señal  $x(n)$  de la que  $a$  es el perfil óptimo. Entre más pequeño es el valor de la medida de distorsión es más alta la eficiencia del perfil espectral, en el sentido que se necesita una excitación de menor energía para generar la señal  $x(n)$ .

Esta eficiencia por supuesto está acotada por el mismo perfil, que es contra el que se compara  $\bar{a}$  y el orden más pequeño de la distancia  $a^t R_x a$  se obtiene cuando se compara  $a$  contra  $a$  es decir  $d(a,a)=a^t R_x a$ .

El cálculo de la medida de distorsión involucra la evaluación de una forma cuadrática de la forma  $a^t R_x a$  con  $a^t=(a_0, a_1, \dots, a_m)$

$$a^t R_x a = \sum_{i=0}^m \sum_{j=0}^m a_i r(|i-j|) a_j \quad (4.15)$$

Esta es una suma de  $a_i r(|i-j|) a_j$  sobre un arreglo rectangular de  $(m+1)*(m+1)$ . La suma en la ecuación 4.15 se puede hacer normalmente por renglones o por columnas pero en este caso como la matriz de autocorrelación es constante sobre sus diagonales y simétrica, es mejor hacer la suma sobre los diagonales obteniéndose (Barrón, 1998):

$\sum_{i=0}^m \sum_{j=0}^m a_i r_x(|i-j|) a_j =$  suma sobre el diagonal principal más dos veces la suma sobre los diagonales debajo de la principal

$$= \sum_{k=0}^m a_k^2 r_x(0) + 2 \sum_{i=1}^m \sum_{j=0}^{m-i} a_j a_{j+i} r_x(i) = r_x(0) \sum_{k=0}^m a_k^2 + 2 \sum_{i=1}^m r_x(i) \sum_{j=0}^{m-i} a_j a_{j+i}$$

pero:  $\sum_{j=0}^{m-i} a_j a_{j+i}$

es la autocorrelación del perfil espectral  $a$ , es decir:

$$r_a(n) = \sum_{k=0}^{m-n} a_k a_{k+n}$$

por lo tanto:

$$a^t R_x a = r_x(0) r_a(0) + 2 \sum_{i=1}^m r_x(i) r_a(i) \quad (4.16)$$

#### 4.6 DETECCIÓN DEL VECINO MÁS PRÓXIMO

Dado un libro de códigos  $C$  formado por  $N$  vectores espectrales  $(a_k)_{k=1}^N$  y una secuencia  $x(n)$  el problema es encontrar el perfil espectral  $a_i$  más cercano al perfil óptimo  $a$  de  $x(n)$ .

Para resolver este problema lo primero que se tiene que hacer es calcular el vector de autocorrelación de  $x(n)$  y después calcular  $d(a_k, a) = a_k^t R_x a_k$   $1 \leq k \leq N$  y elegir el más próximo como:

$$V = \operatorname{argmin} d(a_k, a) = \operatorname{argmin} a_k^t R_x a_k \quad (4.17)$$

#### 4.7 CÁLCULO DEL CENTROIDE

Dado un libro de códigos  $C = (a_k)_{k=1}^N$  se puede partir un espacio muestra  $X$  en  $N$  regiones disjuntas  $R_k$ ,  $1 \leq k \leq N$  y cada una de estas regiones se les pueden asociar un centroide  $C_k$ , que en caso óptimo debería de corresponder con el vector código asociado a la región. El centroide viene a ser el mejor representante de la región que se obtiene, es el que en promedio difiere menos de cada uno de los vectores de la región (Barrón, 1998).

Una región está definida por todos aquellos vectores del espacio muestra que tiene un mismo vector código como vecino más próximo:

Con  $x$  en el espacio muestra  $X$  y  $a_i, a_j$  en el libro de códigos  $C$ ,  $x$  es una señal, pero se entiende que el cálculo de la distorsión  $d(a_i, x)$  se hace contra su perfil espectral óptimo correspondiente. Una vez que se tiene definida una región el centroide se calcula como sigue:

$$R_i = \left\{ x : d(a_i, x) < d(a_j, x) \right\}, i \neq j$$

#### 4.8 CÁLCULO DE LA DISTORSIÓN TOTAL

Una manera de evaluar la calidad de un grupo de centroides como representantes de una población  $(X_k)_{k=1}^M$  es calculando la dispersión o distorsión total promedio  $\Delta$ .

$$\Delta = \frac{1}{M} \sum_{k=1}^M d(C_k, a_k) \quad \text{donde } C_k \text{ es el vecino más próximo de } a_k \text{ en el libro de códigos}$$

Entre otras cosas la distorsión total nos permite elegir entre varios modelos dados por libros código, cuál es el más apegado a la familia  $(X_k)_{k=1}^M$  que pueden representar segmentos de una palabra, y en esto precisamente está basado el esquema propuesto de reconocimiento (Barrón, 1998).

#### 4.9 CONSTRUCCIÓN DEL LIBRO CÓDIGO

Son dos las condiciones que debe cumplir un libro de códigos óptimo (Barrón, 1998):

- a) Debe tener asociadas regiones bien definidas a cada uno de sus vectores código.
- b) Cada vector código debe ser el centroide de su región correspondiente.

Si estas dos condiciones se cumplen entonces los vectores del libro código conforman un mínimo local de la función de distorsión total.

Suponiendo que se tiene un espacio muestra o de entrenamiento  $X$  y se quiere construir un libro código de  $N$  vectores, entonces aplicando sucesivamente la iteración de Lloyd se llega a un libro código que cumpla con las dos condiciones anteriores.

#### 4.10 ITERACIÓN DE LLOYD

La iteración de Lloyd permite encontrar dado un conjunto de vectores código, otro conjunto de vectores código que sean mejores representantes, es decir que tengan una distorsión total menor.

$$\text{Sean } A = (a_k)_{k=1}^N$$

Un conjunto de vectores código y sean las regiones asociadas en base al criterio del vecino más próximo las siguientes:

$$R_i = \{x: d(a_i, x) < d(a_j, x)\}, i \neq j$$

y sean

$$B = (b_i)_{i=1}^N$$

los centroides de cada región  $R_i$  si  $a_i = b_i, 1 \leq i \leq N$ .

#### 4.11 ALGORITMO DE BIPARTICIÓN

Supongamos que tenemos un espacio de entrenamiento  $X$  y que deseamos construir un libro código de orden  $N=2^n$ , el procedimiento es el siguiente (Barrón, 1998):

- a) Se construye un libro código de tamaño 1, para esto basta calcular el centroide  $X$ .
- b) En general, si se tiene un libro código óptimo de orden  $k$ , se construye uno óptimo de tamaño  $2^k$ . Primero perturbando cada vector código del libro óptimo, para obtener por cada vector dos vectores perturbados  $y^+ = (1+\varepsilon)y$ ,  $y^- = (1-\varepsilon)y$ , después, una vez que se tiene un libro código no óptimo de orden  $2k$ , aplicar el algoritmo correspondiente para optimizarlo.
- c) El paso b se repite hasta que se obtiene el libro código óptimo del tamaño deseado.

Como comentario del método antes citado, se utiliza principalmente para tareas de reconocimiento de voz aislada, en aplicaciones de reconocimiento de comandos se fundamenta su aplicación.

Debido a que como parte fundamental del desarrollo del presente trabajo, está el hecho de encontrar la respuesta que tiene un corpus de voz. Se procederá a encontrar la respuesta que tiene un sistema creado en laboratorio donde se hará uso del método de los LPC y haciendo uso de las sílabas como unidades básicas de reconocimiento. Todo esto haciendo uso de un corpus aplicado a cuestiones de la pronunciación de comandos.

Se procederá a analizar la respuesta que tiene el sistema a cada una de las unidades de segmentación y la forma en la cual el uso de las sílabas ayuda a la tarea de reconocimiento en el caso de ser comandos los utilizados.

Lo anterior permitirá establecer el rango de aplicación que pueden tener las sílabas para el caso de ser utilizado en aplicaciones de tipo comandos. La utilidad de esta representación nos permitirá conocer los aspectos necesarios para considerar si el uso de la sílaba en este tipo de aplicaciones causa buenos rendimientos en comparación a por ejemplo el uso de las palabras u otra unidad utilizada para este fin.

Lo anterior permitirá crear los fundamentos que se necesitan para trasladar estas aplicaciones al uso del reconocimiento del habla continua.

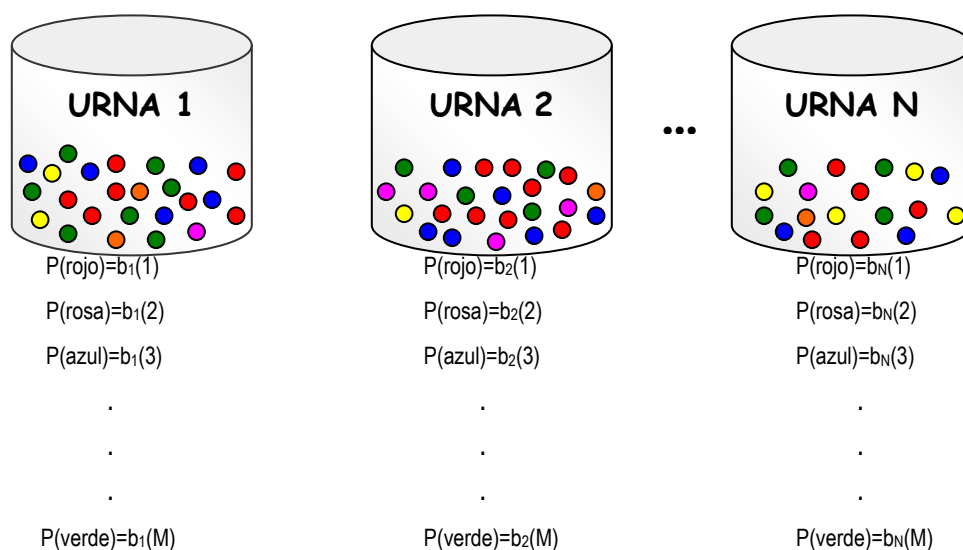
## 4.12 CADENAS OCULTAS DE MARKOV (COM)

Las Cadenas Ocultas de Markov (HMM) se encargan de encontrar un modelo óptimo para cada palabra que pertenezca a un vocabulario definido. A continuación se explica el típico ejemplo de las urnas para comprender su funcionamiento.

### El modelo de la urna y la pelota

En (Rabiner, 1989) se menciona el ejemplo de considerar que hay  $N$  urnas de vidrio en un cuarto como se muestra en la figura 4.1, dentro de cada una existen una gran cantidad de pelotas de  $M$  distintos colores, el proceso físico para obtener las observaciones es como sigue:

Una persona está en un cuarto, de acuerdo a un procedimiento aleatorio, elige una urna inicial; de esta urna, una pelota es seleccionada al azar, y su color es registrado como una observación. La pelota entonces se reemplaza en la urna de la que fue seleccionada. Se escoge una nueva urna de acuerdo al proceso aleatorio, y el proceso de selección de pelota se repite, de esta forma se genera una secuencia de observaciones finitas de colores, la cual podríamos modelar como la salida observable de una HMM.



$O = \{\text{verde, rosa, azul, rojo, amarillo, rojo,....., azul}\}$

Figura 4.1. El modelo de la urna y la pelota.

Debería ser obvio que el HMM más simple en el proceso de la urna y la pelota, es en el que cada estado pertenece a una urna específica y para la cual la probabilidad de una pelota de color está definida por cada estado. La elección de

las urnas está dictada por la matriz de los estados de transición del HMM. Se debe de tomar en cuenta que las pelotas de colores en cada urna deben ser las mismas y la distinción sobre varias urnas está en el modo de la colección de pelotas de colores que la componen. Por lo tanto, una observación aislada de una pelota de color en particular no nos dice inmediatamente de cual urna proviene.

#### **4.12.1 LOS ELEMENTOS DE UNA CADENA OCULTA DE MARKOV**

En (Rabiner, 1989) y (Resch, 2001b) se menciona que una Cadena Oculta de Markov para observaciones de símbolos discretos tales como un modelo de urnas y pelotas se caracteriza por lo siguiente:

##### Número de estados en el modelo (N)

Debido a que los estados se encuentran ocultos, para muchas aplicaciones prácticas existe algún significado relacionado a los estados o conjuntos de estados del modelo. En el modelo de la urna y las pelotas, cada estado corresponde a las urnas. Generalmente, los estados están interconectados de tal forma que cualquier estado puede alcanzar a otro; como veremos, existen una gran cantidad de interconexiones entre estados de interés y esto se puede trasladar a aplicaciones de reconocimiento de voz. Se etiquetan los estados individuales como  $\{1, 2, \dots, N\}$ , y denotan el estado al tiempo  $t$  como  $q_t$ .

##### NÚMERO DE DIFERENTES OBSERVACIONES DE CADA SÍMBOLO POR ESTADO (M)

Los símbolos de observación corresponden a la salida física del sistema que está siendo modelado. Para el modelo de la urna con las pelotas son los colores de las pelotas seleccionadas de las urnas. Se denotan los símbolos individuales como  $V = \{v_1, v_2, \dots, v_M\}$ .

##### Distribución de probabilidad de cada estado de transición (A)

La distribución de probabilidad de cada estado de transición es  $A = \{a_{ij}\}$ , donde  $a_{ij}$  se define como se muestra en la ecuación 4.18.

$$a_{ij} = P[q_{t+1} = j | q_t = i], \quad 1 \leq i, j \leq N \quad (4.18)$$

para el caso especial en el cual un estado puede alcanzar a otro estado en un paso sencillo, tenemos  $a_{ij} > 0$  para todos los  $i, j$ . Para otros tipos de HMM, podríamos tener  $a_{ii} = 0$  para uno o más pares de  $(i, j)$ .

### Observación de los símbolos de la distribución de probabilidad ( $B$ )

La observación de los símbolos de la distribución de probabilidad,  $B = \{b_j(k)\}$ , en la cual  $b_j(k)$  se define por la ecuación 4.19, donde se definen los símbolos de distribución en el estado  $j$ ,  $j=1,2,\dots,N$ .

$$b_j(k) = P[o_t = v_k | q_t = j], \quad 1 \leq k \leq M, \quad (4.19)$$

### Distribución del estado inicial ( $\pi$ )

La distribución del estado inicial  $\pi = \{\pi_i\}$ , está definida por la ecuación 4.20.

$$\pi_i = P[q_1 = i], \quad 1 \leq i \leq N \quad (4.20)$$

se puede observar que un modelo de un HMM requiere la especificación de dos parámetros para él número de estados ( $N$ ) y número de diferentes observaciones de cada símbolo por estado ( $M$ ), la especificación de los símbolos de observación, y la especificación de los tres estados de probabilidad medidos  $A$ ,  $B$  y  $\pi$ . Por conveniencia, usamos la notación compacta:

$$\lambda = (A, B, \pi)$$

para indicar el conjunto de parámetros completos del modelo. Este conjunto de parámetros, por supuesto, definen una medida de probabilidad para  $O$ , por ejemplo,  $P(O | \lambda)$ .

## **4.13 LOS TRES PROBLEMAS BÁSICOS DE LAS COM**

Dados los datos de la HMM anterior y como se expuso antes, los tres problemas básicos que deben ser resueltos por el modelo son los siguientes:

### Problema 1

Dada una secuencia de observación  $O = (o_1, o_2, \dots, o_T)$ , y un modelo  $\lambda = (A, B, \pi)$ , como podemos calcular eficientemente  $P(O | \lambda)$ , la probabilidad de la secuencia de observación producida por el modelo.

### Problema 2

Dada la secuencia de observación  $O = (o_1, o_2, \dots, o_T)$ , y el modelo  $\lambda$ , como seleccionamos una secuencia de estados correspondiente  $Q = (q_1, q_2, \dots, q_T)$  que sea óptima en algún sentido.



En este problema intentamos encubrir la parte oculta del modelo, esto es, encontrar la secuencia de estados correcta. Esto debe ser lo suficientemente claro para que en todo los casos de modelos generados, no haya una secuencia de estados correcta a ser encontrada.

### Problema 3

Como ajustar los modelos de los parámetros  $\lambda=(A, B, \pi)$  para maximizar  $P(O/\lambda)$ .

En este problema intentamos optimizar los parámetros del modelo para describir de mejor forma como se construye una secuencia de observación dada. La secuencia de observación usada para ajustar los parámetros del modelo es llamada la secuencia de entrenamiento porque es usada para entrenar la HMM. El problema del entrenamiento es una de las aplicaciones más cruciales para el HMM, porque nos permite adaptar de manera óptima los parámetros del modelo que son observados durante el entrenamiento de datos.

### SOLUCIÓN AL PROBLEMA 1

Se desea calcular la probabilidad de la secuencia de observaciones,  $O=(o_1, o_2, \dots, o_T)$ , dado el modelo  $\lambda$ , por ejemplo,  $P(O/\lambda)$ . La forma mas adecuada de hacer esto es a través de la numeración de cada posible secuencia de estados de longitud  $T$  (número de observaciones). Considere una de las siguientes secuencias mezcladas.

$$Q=(q_1, q_2, \dots, q_T)$$

en donde  $q_1$  es el estado inicial, la probabilidad de la secuencia de observación dado la secuencia de estados anterior es la ecuación siguiente.

$$P(O | Q, \lambda) = \prod_{t=1}^T P(O_t | q_t, \lambda) \quad (4.21)$$

en donde hemos asumido la independencia estadística de observaciones, por lo que obtenemos la ecuación siguiente:

$$P(O/q, \lambda) = bq_1(O_1) \cdot bq_2(O_2) \dots bq_T(O_T)$$

La probabilidad de cada secuencia estática  $Q$  puede ser escrita como la siguiente expresión:

$$P(Q|\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \dots a_{q_{T-1} q_T}$$

La probabilidad que  $O$  y  $Q$  ocurran simultáneamente se muestra en la siguiente expresión, la cual es simplemente el producto de las dos ecuaciones anteriores.

$$P(O, Q|\lambda) = P(O|Q, \lambda) P(Q|\lambda)$$

La probabilidad de  $O$  (dado el modelo) es obtenida como se muestra en la siguiente expresión, pero sumando este tipo de probabilidad a través de todas las posibles secuencias  $q$  dadas,

$$\begin{aligned} P(O, Q|\lambda) &= \sum_{\text{todas } Q} P(O|Q, \lambda) P(Q|\lambda) \\ &= \sum_{q_1, q_2, \dots, q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \dots a_{q_{T-1} q_T} b_{q_T}(O_T) \end{aligned}$$

La interpretación del cálculo en la ecuación anterior es la siguiente: inicialmente (al tiempo  $t=1$ ) estamos en el estado  $q_1$  con una probabilidad igual a  $\pi_{q_1}$ , y genera los símbolos  $O_1$  (en este estado) con la probabilidad  $b_{q_1}(O_1)$ . El reloj cambia del tiempo  $t$  a  $t+1$  (tiempo=2) y realizamos una transición al estado  $q_2$  desde el estado  $q_1$  con la probabilidad  $a_{q_1 q_2}$ , que genera el símbolo  $O_2$  con la probabilidad  $b_{q_2}(O_2)$ . Este proceso continúa de esta manera, mientras que se hace, se lleva a cabo la traslación anterior (al tiempo  $T$ ) del estado  $q_{T-1}$  al estado  $q_T$  con probabilidad  $a_{q_{T-1} q_T}$  y genera el símbolo  $O_T$  con probabilidad  $b_{q_T}(O_T)$ .

Un pequeño análisis permite verificar que el cálculo de  $P(O|\lambda)$ , de acuerdo a su definición en la ecuación involucra el orden de  $2T \cdot N^T$  cálculos, dado que en cada  $t=1, 2, \dots, T$ , hay  $N$  posibles secuencias de estados, y para cada secuencia de estado aproximadamente  $2T$  cálculos son requeridos por cada término en la suma de la ecuación. Para ser precisos, se necesitarían  $(2T-1)N^T$  multiplicaciones, y  $N^T-1$  sumas, este cálculo es computacionalmente irrealizable, dado que para valores pequeños de  $N$  y  $T$ , por ejemplo  $N=5$ ,  $T=100$  (observaciones), hay alrededor de 1072 cálculos. Claramente un procedimiento más eficiente se requiere para resolver el problema 1, afortunadamente tal procedimiento existe y se llama el *procedimiento hacia adelante*.

## **PROCEDIMIENTO HACIA ADELANTE**

Considere la variable regresiva  $\alpha_t(i)$  definida como se muestra en la ecuación 4.22

$$\alpha_t(i) = P(O_1 O_2 \dots O_t, q_t = S_i | \lambda) \quad (4.22)$$

esto es, la probabilidad de la secuencia de observación parcial,  $O_1 O_2 \dots O_t$  (mientras el tiempo  $t$ ) y el estado  $i$  al tiempo  $t$ , dado el modelo  $\lambda$ . En (Zhang, 1999) se menciona que podemos resolver para  $\alpha_t(i)$  inductivamente, como se muestra en las ecuaciones siguientes:

1. Inicialización

$$\alpha_1(i) = \pi_i b_i(O_1) \quad 1 \leq i \leq N$$

2. Inducción

$$\alpha_{t+1}(j) = b_j(O_{t+1}) \sum_{i=1}^N \alpha_t(i) a_{ij} \quad 1 \leq j \leq N, \quad 1 \leq t \leq T-1$$

3. Terminación

$$P(O | \lambda) = \sum_{i=1}^N \alpha_T(i)$$

## SOLUCIÓN AL PROBLEMA 2

El *algoritmo de Viterbi* encuentra la mejor de las secuencias de estados,  $Q = \{q_1, q_2, \dots, q_T\}$ , para una secuencia de observaciones dada  $O = \{O_1, O_2, \dots, O_T\}$ , como se muestra en las ecuaciones siguientes.

1. Inicialización

$$\delta_1(i) = \pi_i b_i(O_1) \quad 1 \leq i \leq N$$

2. Recursión

$$\delta_{t+1}(j) = b_j(O_{t+1}) \left[ \max_{1 \leq i \leq N} \delta_t(i) a_{ij} \right] \quad 1 \leq j \leq N, \quad 1 \leq t \leq T-1$$

3. Terminación

$$p^* = \max[\delta_T(i)] \quad 1 \leq i \leq N$$

$$q^* = \arg \max[\delta_T(i)] \quad 1 \leq i \leq N$$

## SOLUCIÓN AL PROBLEMA 3

Para poder dar solución al problema 3, se hace uso del *algoritmo de Baum-Welch*, que al igual que los otros realiza por medio de inducción la determinación de valores que optimicen las probabilidades de transición en la malla de posibles transiciones de los estados del modelo de Markov. A manera de ejemplo del cálculo de inducción de las trayectorias se tiene en la figura 4.2:

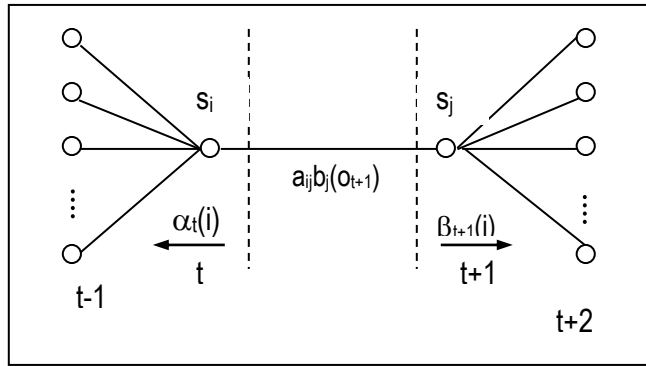


Figura 4.2. Paso de inducción para encontrar una optimización de los parámetros de la Cadena Oculta de Markov

Con el análisis anterior, Baum-Welch logró obtener la siguiente expresión para la implementación de su algoritmo. La ecuación 4.23, nos permite determinar el número de transiciones del estado  $s_i$  al estado  $s_j$ .

$$\varepsilon_t(i, j) = \frac{\alpha_t(i) a_{ij} \beta_{t+1}(j) b_j(o_{t+1})}{\sum_{i=1}^N \sum_{j=1}^N \alpha_t(i) a_{ij} \beta_{t+1}(j) b_j(o_{t+1})} \quad (4.23)$$

Una manera eficiente de optimizar los valores de las matrices de transición en el algoritmo de Baum-Welch, es de la forma siguiente (Oropeza, 2000) y (Rabiner and Biing-Hwang, 1993):

$$a_{ij} = \frac{\text{número esperado de transiciones del estado } s_i \text{ al estado } s_j}{\text{número esperado de transiciones del estado } s_i}$$

$$b_j(k) = \frac{\text{número esperado de veces que estando en } j \text{ aparece el símbolo } v_k}{\text{número esperado de veces que se analiza el estado } j.}$$

#### 4.14 MIXTURAS DE GAUSSIANAS

Las mixturas de Gaussianas como se ha comentado con anterioridad son combinaciones de distribuciones normales o funciones de Gauss. Una mixtura de Gaussianas puede por lo tanto ser escrita o ser vista como una suma de densidades de Gaussianas (Resch, 2001a), (Resch, 2001b), (Kamakshi et al., 2002) y (Mermelstein, 1975).

Como es sabido, la función de densidad de probabilidad del tipo Gaussiana es de la forma:

$$g(\mu, \Sigma)(x) = \frac{1}{\sqrt{2\pi^d} \sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \quad (4.24)$$

Una mezcla de Gaussianas con ciertos valores de pesos se ve de la forma:

$$gm(x) = \sum_{k=1}^K w_k * g(\mu_k, \Sigma_k)(x) \quad (4.25)$$

en donde los pesos son todos positivos y la suma de los mismos es igual a 1:

$$\sum_{i=1}^K w_i = 1 \quad \forall \quad i \in \{1, \dots, K\} : w_i \geq 0 \quad (4.26)$$

En la figura 4.3 se muestra un ejemplo de una mezcla Gaussiana, que consiste de tres Gaussianas sencillas:

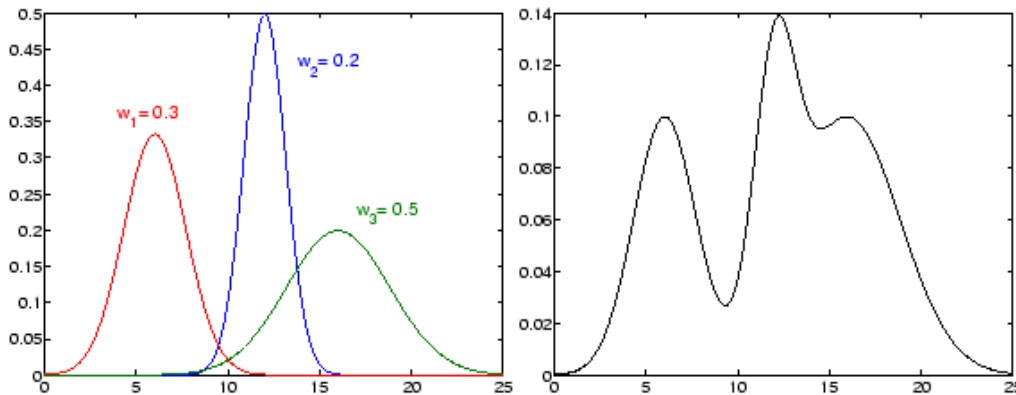


Figura 4.3. Representación esquemática de 3 Gaussianas en una gráfica.

Al variar el número de Gaussianas  $K$ , los pesos  $w_i$  y los parámetros de cada de las funciones de densidad  $\mu$  y  $\Sigma$ , las mezclas de Gaussianas pueden ser usadas para describir algunas Funciones de Densidad de Probabilidad Complejas (FDPC).

Los parámetros de la función de densidad de probabilidad (pdf) son el número de Gaussianas, sus factores de peso, y los parámetros de cada función de Gaussiana tales como la media  $\mu$  y la matriz de covarianza  $\Sigma$ .

Para encontrar estos parámetros que de alguna forma describen a una determinada función de probabilidad de un conjunto de datos un algoritmo iterativo, el de máxima esperanza (*EM*) será utilizado.

La siguiente figura 4.4 muestra la aplicación de tales elementos a la cadena de Markov (Peskin et al., 1991).

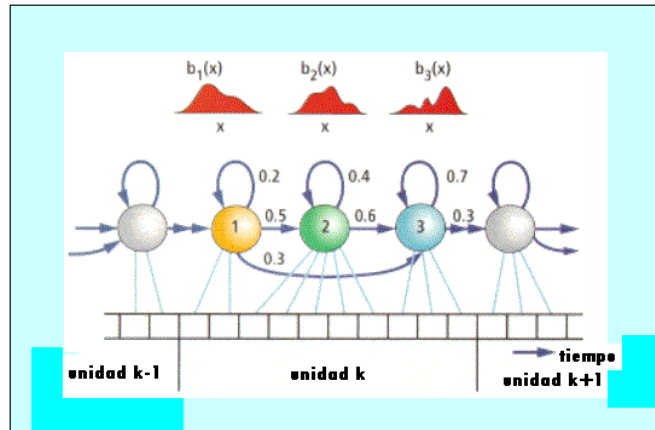


Figura 4.4. Ejemplo de las Mixturas Gaussianas en una Cadena de Markov concatenada (Resch, 2001b).

## 4.15 RECONOCIMIENTO DE VOZ ESTADÍSTICO

La señal de voz es representada por secuencias de palabras que se describen de la forma  $W = (w_1, w_2, \dots, w_t)$  en donde  $w_t$  es la palabra genérica pronunciada en el tiempo discreto " $t$ ". Las secuencias de palabras son conectadas a través de la voz que es una secuencia de sonidos acústicos  $\chi$ , de acuerdo a Becchetti and Prina (1999).

En términos matemáticos, el reconocimiento es un operador " $f$ " que mapea un  $X$  conjunto de datos que pertenece al conjunto completo de secuencias acústicas  $\chi$  a un conjunto  $W$  que está incluido en el conjunto  $\varpi$ , de acuerdo a Becchetti and Prina (1999):

$$f: X \rightarrow W, \quad X \in \chi \quad y \quad W \in \varpi \quad (4.27)$$

### 4.15.1 La aproximación determinística

El problema del reconocimiento de voz, como muchos otros, puede ser resuelto pero relacionando a un modelo  $\Theta$  con el fenómeno. La estrategia consiste en construir un modelo  $\Theta$  que regrese todas las posibles secuencias  $\chi$  a todos los eventos admisibles  $W$ . En el caso de la voz,  $h(W, \Theta)$  regresa todas las posibles secuencias  $\chi$  que puedan ser asociadas a la  $W$  dada. El reconocimiento es

realizado pero encontrando la secuencia de palabras que de acuerdo a  $\Theta$  regresan una secuencia  $\chi$  que se relaciona de mejor forma. En esencia, el reconocimiento se realiza pero haciendo uso de un conocimiento a priori de todo el mapeo de secuencias de palabras acústicas. Este conocimiento se encuentra almacenado en  $\Theta$ , de acuerdo a Becchetti and Prina (1999).

De forma matemática, la expresión  $d(X', X'')$  define una distancia entre dos secuencias  $X'$  y  $X''$ , la secuencia de secuencia de palabras  $W^*$  asociada a un  $\chi$  que esta dada por:

$$W^* = \text{ArgMin}_{W \in \Theta} d(h(W, \Theta), X) \quad (4.28)$$

El procedimiento general se esquematiza en la figura 4.4, el cual consiste de 2 pasos, de acuerdo a Becchetti and Prina (1999):

- Entrenamiento: el modelo es construido tomando un gran número de diferentes correspondencias  $(X', W')$ . Este es el mismo procedimiento de entrenamiento de un ser humano en su edad temprana: entre más grande es el número de acoplamientos  $(X', W')$ , más grande es el reconocimiento adecuado.
- Reconocimiento: todas las secuencias posibles de palabras  $W$  son probadas para encontrar la  $W^*$  en donde la secuencia acústica  $X = h(W^*, \Theta)$  mejor se relacionan a uno dado.

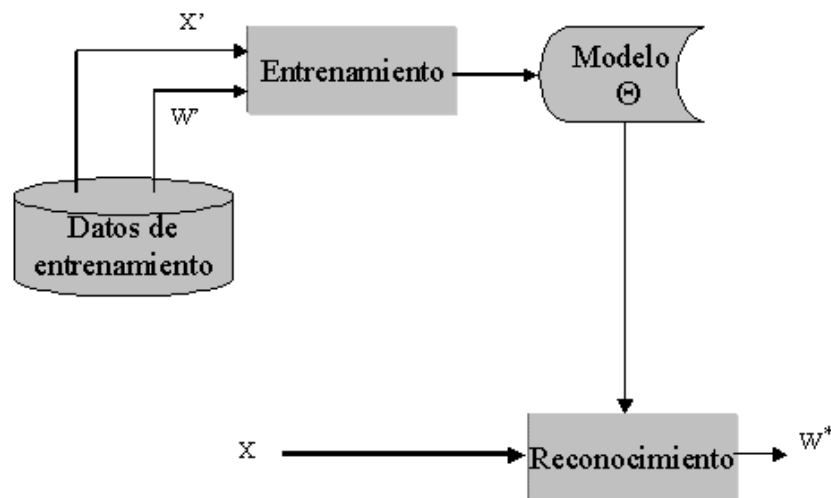


Figura 4.5. Esquema de funcionamiento de un simple sistema de reconocimiento de voz.

### 4.15.2 El plano estocástico

Se considera que el reconocimiento estadístico de un conjunto de secuencias de palabras, puede ser obtenido como la maximización de la probabilidad dada por la

secuencia acústica  $\chi$  actual. Conocida con el nombre de "observación", la secuencia acústica  $\chi$  e introduciendo la probabilidad a posteriori  $P(W|X)$  se tiene la  $W$  condicionada al hecho de que la  $X$  es observada. La secuencia reconocida de palabras  $W^*$  es:

$$W^* = \text{ArgMax}_{W \in \mathcal{W}} P(W | X) \quad (4.29)$$

Consideraremos que la probabilidad condicional  $P(x|w)$  tiene una variable estocástica  $x$  condicionada a un valor particular de una segunda variable  $w$ , lo cual es relacionada a la probabilidad conjunta  $P(x, w)$  por  $P(x|w) = P(x, w) / P(w)$ , pero haciendo uso de la fórmula de Bayes se tiene, de acuerdo a Becchetti and Prina (1999):

$$P(W | X) = \frac{P(X, W)}{P(X)} = \frac{P(X, W)P(W)}{P(X)} \quad (4.30)$$

La fórmula 4.29 se convierte en:

$$W^* = \text{ArgMax}_{W \in \mathcal{W}} P(X, W) = \text{ArgMax}_{W \in \mathcal{W}} P(X | W)P(W) \quad (4.31)$$

donde el término  $P(X)$  es eliminado dado que no cambia el resultado de la maximización, de acuerdo a Becchetti and Prina (1999).

La fórmula (4.31) permite a un modelo ser introducido, usando conceptos que son similares a aquellos delineados en la sección anterior. De la expresión (4.31) tenemos de acuerdo a Becchetti and Prina (1999):

$$W^* = \text{ArgMax}_{W \in \mathcal{W}} P(X | W, \Theta)P(W | \Theta) \quad (4.32)$$

De forma similar (4.31) y (4.32) tienen diferentes significados. (4.31) expresa que, siguiendo el criterio de (Máximo a posteriori, MAP -Maximum a Posteriori-) la secuencia de reconocimiento puede ser encontrada, pero maximizando la probabilidad a posteriori. Las probabilidades deben de ser encontradas de alguna forma, de acuerdo a Becchetti and Prina (1999).

La fórmula (4.32) sugiere que un modelo  $\Theta$  puede ser seleccionado de dos conjuntos de probabilidades que deben de ser estimadas:

- Las probabilidades  $P(X | W, \Theta)$  que el modelo  $\Theta$  produce la secuencia acústica  $X$  cuando la comunicación de la secuencia de palabras  $W$ .
- La probabilidad  $P(W, \Theta)$  que el modelo comunica a las palabras  $W$ .



### 4.15.3 Simplificaciones al modelo estocástico

El procedimiento de reconocimiento puede ser simplificado pero considerando que una palabra pronunciada está compuesta básicamente de secuencias de estados acústicos  $S$  que pertenecen a un conjunto  $\delta$ . La  $X$  asociada a palabras existentes puede ser modelada como secuencias de tales estados, que por ejemplo pueden corresponder al conjunto de datos acústicos similares de cada fonema, de acuerdo a Becchetti and Prina (1999).

Por ejemplo, las vocales pueden ser modeladas por tres estados acústicamente similares: el de inicio, el central y el final (el presente trabajo hace uso de este mismo esquema para referenciar a las sílabas). Diferentes secuencias de unidades pueden estar asociadas a una palabra determinada. La principal simplificación es que el posible número de unidades se encuentra limitado. Un total de 138 unidades (46 fonemas x 3 estados) son generalmente suficientes para obtener buenos resultados en la mayoría de los idiomas (Becchetti and Prina, 1999). De forma matemática, la secuencia esta dada por:

$$\begin{aligned}
 W^* &= \text{ArgMax}_{W \in \varpi, S \in \delta} P(X, W, S | \Theta) = \\
 &= \text{ArgMax}_{W \in \varpi, S \in \delta} P(W, S | \Theta) P(X | W, S, \Theta) = \\
 &= \text{ArgMax}_{W \in \varpi, S \in \delta} P(W | \Theta) P(S | W, \Theta) P(X | W, S, \Theta)
 \end{aligned} \tag{4.33}$$

Los últimos términos de (4.33) es un producto de tres probabilidades. Estas probabilidades pueden ser simplificadas para realizar un reconocimiento factible como se considera a continuación.

### 4.15.4 Simplificaciones de $P(W | \Theta)$

La probabilidad de secuencias  $P(W | \Theta)$  puede ser simplificada considerando que la palabra al momento " $t$ ",  $w_t$  depende estadísticamente sólo de un número limitado  $L$  de palabras pronunciada con anterioridad  $w_{t-1}, w_{t-2}, \dots, w_{t-L}$  por lo que el proceso no depende del tiempo. Considerando  $L=2$  se tiene, de acuerdo a Becchetti and Prina (1999):

$$P(W = w_1, w_2, \dots, w_T | \Theta) \equiv P(w = w_1 | \Theta) \prod_{t=2, \dots, T} P(w_t | w_{t-1}, \Theta) \tag{4.34}$$

La fórmula anterior establece que la probabilidad de una secuencia de palabras  $P(W = w_1, w_2, \dots, w_T)$  puede ser calculada por la probabilidad  $P(w | \Theta)$  de que una palabra en particular sea emitida y la probabilidad de que una palabra  $w_t$  sea emitida después de que la pronunciación de la palabra  $w_{t-1}$ .

#### 4.15.5 Simplificaciones de $P(S|W, \Theta)$

Se pueden realizar simplificaciones similares a la expresión  $P(S|W, \Theta)$ . Se considera que la probabilidad de emitir una unidad  $s_t$  depende solamente de la palabra actual y es independiente del tiempo en que es pronunciada. Además, la probabilidad de una unidad puede ser considerada como independiente de las características de las unidades pronunciadas de forma futura y depende solamente de la unidad pronunciada previamente. Por tanto,  $P(S|W, \Theta)$  toma la forma de acuerdo a Becchetti and Prina (1999):

$$P(S = s_1, s_2, \dots, s_T | W = w_1, w_2, \dots, w_T, \Theta) = P(s_1 | w_1, \Theta) \prod_{t=2, \dots, T} P(s_t | s_{t-1}, w_t, \Theta) \quad (4.35)$$

#### 4.15.6 Simplificaciones de $P(X|W, S, \Theta)$

Los últimos términos de (4.35) requieren de mayores simplificaciones. La primera simplificación se deriva del hecho de que la secuencia de muestras acústicas  $X$  no son las mejores observaciones para estimar  $W$ . Por lo que una función estadística  $g(X)$  debe ser encontrada. Desde el punto de vista heurístico, un conjunto de datos:  $Y = y_1, y_2, \dots, y_T = g(X)$  que contiene toda la información de  $X$  que es relevante para la estimación de  $W$ . Satisfactoriamente  $Y$  debería de contener menos datos que  $X$  en el sentido de que  $g(X)$  debería de descartar toda la información menos necesaria para estimar  $W$ . Por ejemplo, el pitch y el volumen no son particularmente significativos para el reconocimiento y pueden ser descartados en la información llevada por  $Y$ . Al introducir a  $g(X)$ , la probabilidad  $P(X|W, S, \Theta)$  en (4.29) es reemplazada por  $P(g(X)|W, S, \Theta) = P(Y|W, S, \Theta)$ , de acuerdo a Becchetti and Prina (1999).

Otras consideraciones pueden ser definidas por el proceso  $Y$  para simplificar el cálculo de  $P(Y|W, S, \Theta)$ . En particular, se considera que:

- $Y$  está estadísticamente independiente de  $W$
- $Y_t$  depende solamente de los estados actuales  $s_t$
- El proceso condicional  $y_t/s_t$  es independiente y estadísticamente distribuido.

Estas consideraciones implican que la probabilidad que  $y_t$  es emitida cuando es la unidad acústica  $s_t$  que debe ser calculada de la forma:

$$P(Y|W, S, \Theta) = \prod_{t=1, \dots, T} P(y_t | s_t, \Theta) \quad (4.36)$$

Como en las expresiones (4.34), (4.35) y (4.36) se permite a la probabilidad  $P(Y|W, S, \Theta)$  ser calculada a través de otra probabilidad estimable sencilla.

Recolectando las expresiones anteriores, una fórmula más simple puede obtenerse para obtener la secuencia  $W^*$  de la forma:

$$W^* = \text{ArgMax}_{W \in \mathcal{W}, S \in \mathcal{S}} \left\{ P(w = w_1 | \Theta) \prod_{t=2}^T P(w_t | w_{t-1}, \Theta) P(s_1 | w_1, \Theta) \prod_{t=1}^T P(s_t | s_{t-1}, w_t, \Theta) \prod_{t=1}^T P(y_t | s_t, \Theta) \right\}$$

#### 4.16 RESUMEN DEL CAPÍTULO

El presente capítulo mostró las diferentes herramientas utilizadas actualmente en lo que al campo del reconocimiento de voz por computadora se refiere, los análisis dan una explicación desde el punto de vista matemático de dichas herramientas. Con lo que se pretende generar los fundamentos teóricos que se plantean en el presente trabajo.

Asimismo, se hace un análisis de los resultados obtenidos con tales herramientas en los diferentes sistemas de reconocimiento existentes actualmente. Como elementos claves de todas las herramientas se encuentran la utilización de una buena elección de los parámetros iniciales de los datos a analizar.

Las diferentes técnicas planteadas a lo largo del presente capítulo presentan beneficios y aspectos diferentes al ser utilizadas en el reconocimiento de voz por computadora. Cuando se realice la implantación de tales herramientas en los sistemas de reconocimiento diseñados para tal fin, se comentarán en base a los resultados su ventaja y desventaja con relación a su utilización con las sílabas.

Por el momento, cabe destacar que el modelo oculto de Markov de densidad continua, es la herramienta más ampliamente utilizada y que las redes neuronales.

# CAPÍTULO

# 5

---

## Integración de la sílaba en los SRAH

Una de las primeras premisas dentro del reconocimiento de voz basado en sílabas es tener una adecuada estimación del comienzo y final de las mismas, esto conllevará a controlar el número de repeticiones y con esto mejorar las características del reconocimiento de voz no continua.

Dentro del análisis del reconocimiento de voz de este tipo, se tienen que considerar en gran medida los aspectos de detección adecuada de la frontera en la sílaba, así como la información acústica del elemento en cuestión, sin dejar a un lado los aspectos de coordinación acústica y representaciones léxicas de la voz

La conformación de un conjunto determinado de elementos permite desarrollar sistemas de reconocimiento del habla que hacen uso de características esenciales del idioma. A lo largo del tiempo, se ha hecho uso del fonema, sin embargo, se ha comprobado que este recurso resulta insuficiente para las tareas actuales.

Dado lo anterior, se hace uso de elementos alternativos como el trifenema y la sílaba que es el caso de estudio que nos interesa en este trabajo. Se presenta pues el análisis del uso de la sílaba para el caso del español, analizando cada una de las características esenciales de su inmersión en los sistemas de reconocimiento del habla no continua.

## 5.1 IMPLANTACIÓN DEL SISTEMA EXPERTO

En nuestro caso la base de conocimientos se encuentra constituida por todas las reglas de clasificación de sílabas del lenguaje español que se analizaron en el capítulo 3, la tarea es entonces entender y poner en el lenguaje de programación apropiado tales reglas para cumplir de forma satisfactoria los requerimientos que el sistema requiere.

La implantación del Sistema Experto para el presente trabajo tiene como entrada el conjunto de frases o palabras que conforman un vocabulario determinado a reconocer (corpus de voz). Tras la aplicación de las reglas pertinentes, la aplicación de la energía en corto tiempo de la señal y de la energía en corto tiempo del parámetro RO, se procede a realizar la división en unidades silábicas de cada uno de los elementos de entrada, con lo cual se logra establecer los inicios y finales de las sílabas (Russell and Norvig, 1996) y (Giarratano and Riley, 2001).

La incorporación de un experto a la fase de entrenamiento para el caso que se propone, lo podemos resumir con el siguiente diagrama a bloques, el cual demuestra la finalidad y función del mismo:

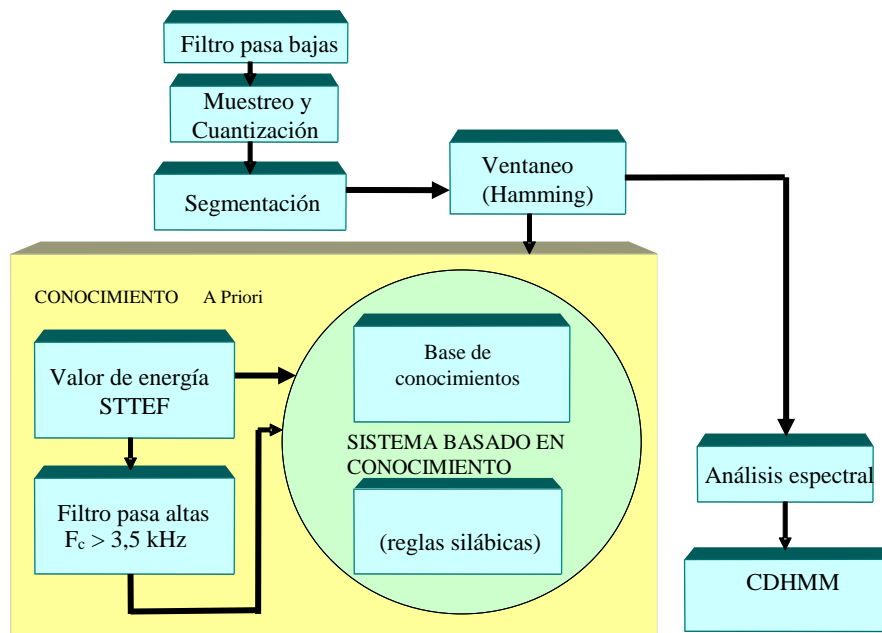


Fig. 5.1. Diagrama a bloques del sistema de reconocimiento propuesto, haciendo uso de un Sistema Experto.

Lo anterior cumple con el objetivo de que en el entrenamiento se extraigan la cantidad y tipos de sílabas que conforman el corpus a estudiar; asimismo, provee la cantidad de bloques que serán usados en la etapa de etiquetado silábico de las

señales de voz. Todo lo anterior permite encontrar e indicar un parámetro de referencia en los siguientes puntos:

- Se procede a concordar el número de sílabas obtenidas por el trabajo del experto, con el número de segmentos obtenidos tras la determinación de la energía en corto tiempo de la señal de voz y la energía extraída a la señal de voz después de habersele aplicado un filtro digital del tipo de respuesta al impulso finito (energía del parámetro RO). En caso de no coincidir se desecha la muestra analizada y se retoma una nueva.
- Para el caso del habla continua, el experto tiene la misma relevancia en la parte del entrenamiento, pues permite extraer las unidades silábicas conforme al diccionario en cuestión. Una vez obtenidas las unidades representativas se procede a realizar el entrenamiento, haciendo uso de la concatenación de los modelos obtenidos de las palabras y del modelo del lenguaje del corpus empleado.

Básicamente y debido a que la finalidad es crear el Sistema Experto que se acople a las necesidades del tipo del SRA que estamos tratando de realizar, se procedió a realizar el diseño y la programación del mismo usando lenguaje C. Las principales características de tal programa se describen a continuación:

- Posee una toma de decisión de tipo cualitativo y cuantitativo del tipo de sílabas que comprenden la entrada que se le está otorgando. Almacenando los resultados en la base de datos del experto (se usó SQL Server para ello).
- La entrada al sistema se produce a través del software diseñado para las aplicaciones referentes a la presente tesis.
- En caso de ser frases, el sistema se encarga de realizar su separación en palabras para después separarlas por sílabas.
- La base de conocimiento es usada para poder realizar tal tarea de segmentación, se tiene como referente las reglas del español para la obtención de sílabas. El uso de las sentencias if proposición then, son el estándar usado en esta capa del sistema.
- La base de programación está realizada en estructuras dinámicas de datos del tipo lista enlazada. Las cuales nos permiten ir accediendo a cada uno de los elementos de la palabra para posteriormente realizar la segmentación adecuada de la misma según las reglas silábicas del idioma.

A continuación se muestran dos segmentos del código del Sistema Experto en cuestión:

```
void CDivide_elemento::inicializa (void)
{
    posiciones_corte[0]=elemento;
```

```

elemento++;
for (i=0;i<20;i++) reglas[i]=false;
cons_insepa="brblcrcldrfrflgrglkrllprpltrrrchtl";
vocal_diptongo="aiaueieuiouiauaieueioiuiuyeyoy";
vocal_abierta="íaíoíéíúúaúéúóúíoaoeaoeaoeooe";
vocales = "aeiouíúáéó";
}

```

Este método se encarga de inicializar los elementos necesarios para analizar dentro de una palabra la existencia de elementos tales como: las consonantes inseparables, las vocales diptongo, las vocales abiertas, las vocales cerradas, etcétera.

*/\*aquí se comienza a analizar la regla 2, después de ella se identificará si existen consonantes inseparables en la cadena de texto que se está analizando\*/*

```

r=inicio; //r apunta al inicio de la lista para iniciar el recorrido de la misma
do
{
ptr1=*r->fonema; //se asigna a ptr1 el primer elemento de la lista
if (ptr1=="C") //si ese elemento es "C" entonces
{
p2=r->next; //se usa p2 para acceder al siguiente
ptr2=*p2->fonema; //fonema de la lista y se guarda en ptr2
if(ptr1==ptr2) //se comparan ambos para saber si son 2 consonantes
{ //si es que si se buscará la satisfacción de la regla 2
for (i=0;i<int(strlen(cons_insepa));i+=2)
{
if(cons_insepa[i]==palabra[contador]&&
cons_insepa[i+1]==palabra[contador+1])
{
conta_conso_insepa[cuenta_insepa]=contador;
cuenta_insepa++;
reglas[1]=true;
if (verifica_existencia(posiciones_corte,elemento,contador))
{
posiciones_corte[elemento]=contador;
elemento++;
}
}
}
}
}
}
}
}
r=r->next; //se toma el siguiente elemento de la lista

```

```

contador+=1;
} while (r->next!=NULL); //se verifica si el apuntador no es igual a NULL

```

Las sentencias anteriores permiten observar la forma en la cual se trabaja sobre un conjunto determinado de elementos de entrada para poder realizar la toma de decisión adecuada. El conjunto de banderas (que representan las reglas que se activan) conforman los resultados de la aplicación de los mecanismos de inferencia y son la base de la toma de decisión del experto. Cuando estas reglas satisfechas se contrarrestan, se procede a analizarlas posteriormente en otra etapa posterior del mismo sistema, para al final entregar los resultados adecuados. El proceso sigue la lógica de los siguientes árboles de inferencia:

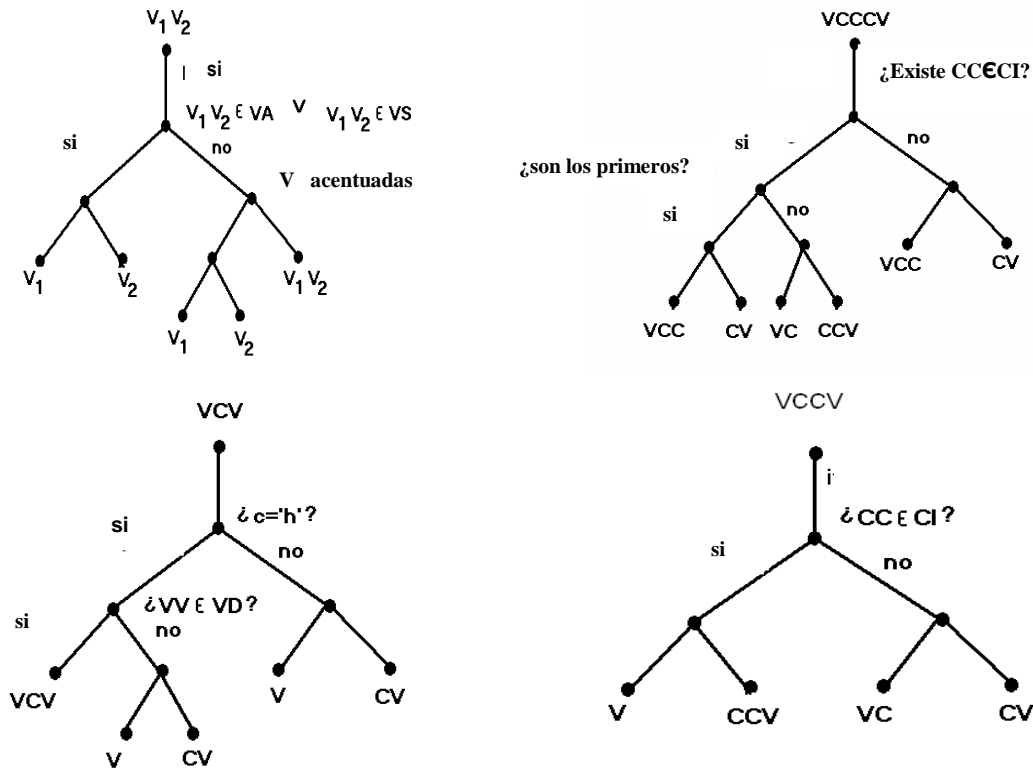


Fig. 5.2. Árboles de inferencia del Sistema Experto.

Los resultados obtenidos en la división silábica al hacer uso de este sistema sobre sílabas independientes, frases de distintos corpus y textos escritos se resumen en la tabla 5.1, manifestándose en la misma el porcentaje de efectividad alcanzado.

TIPO DE CORPUS	PORCENTAJE DE EFECTIVIDAD
Sílabas independientes	100%
Dígitos	100%
Corpus de palabras aisladas	100%
Corpus de Latino40	100%
Corpus de frases simples	100%

Tabla 5.1. Porcentaje de efectividad en la aplicación del Sistema Experto a diferentes corpus de voces.



Como podemos observar en la tabla anterior, la aplicación de las reglas antes mencionadas a cualquier corpus de los señalados otorga un alto porcentaje de segmentación silábica. En comparación con lo reportado en (Giarratano and Riley, 2001), se consideran las siguientes diferencias:

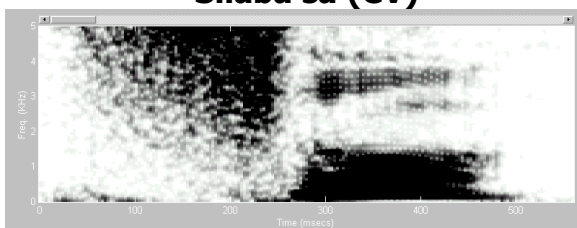
- a) La función del Sistema Experto aquí planteada es para tareas de reconocimiento del habla, mientras que en el otro caso es para síntesis.
- b) El experto aquí programado se probó para tareas de textos científicos y corpus de voz con un alto grado de efectividad.
- c) Para los fines prácticos que a este trabajo competen, las unidades silábicas que contienen acentos ortográficos son consideradas como elementos con división tal y como lo demuestran las reglas establecidas.

## 5.2 DETECCIÓN DE LAS FRONTERAS DE LAS SÍLABAS Y EJEMPLOS DE RECONOCIMIENTO

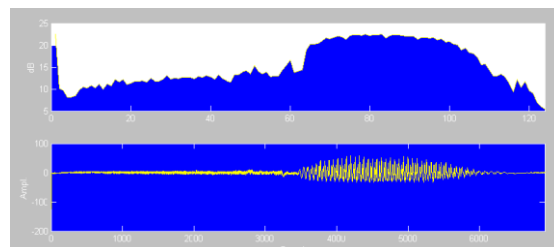
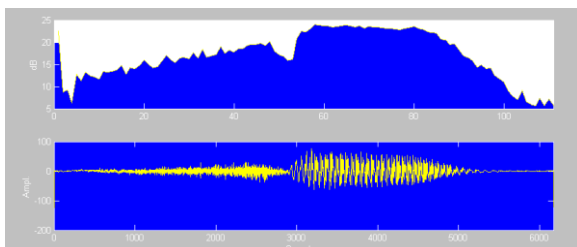
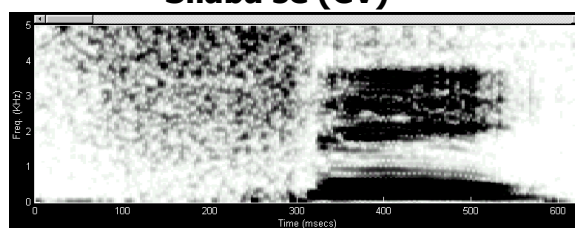
Como se ha comentado con anterioridad, las sílabas contienen un alto beneficio en su uso debido a su alta dependencia al contexto, con relación a las estructuras fonéticas. Su estructura regular sugiere que las fronteras de las sílabas son de mayor relevancia a las de los fonemas, y por ello, pueden ser definidas de mejor forma tanto en las ondas de voz en el dominio del tiempo, así como también en un espectrograma.

Realizando un análisis de las características de las sílabas individuales para reconocer sus características, se procedió a obtener las gráficas tanto en el dominio del tiempo como de la frecuencia de algunas de ellas, observándose lo siguiente:

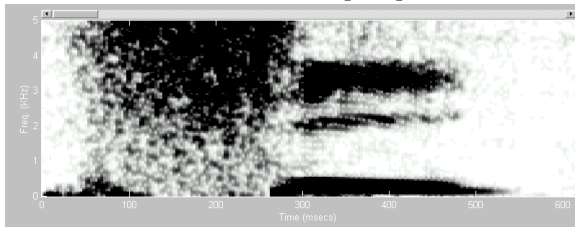
Sílaba sa (CV)



Sílaba se (CV)



**Sílaba si (CV)**



**Sílaba so (CV)**

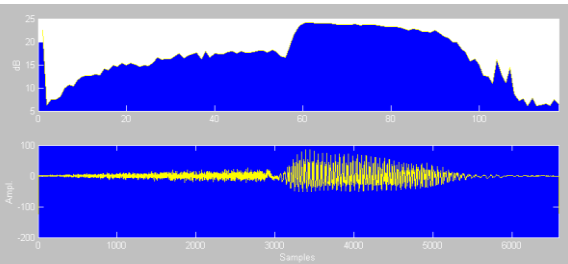
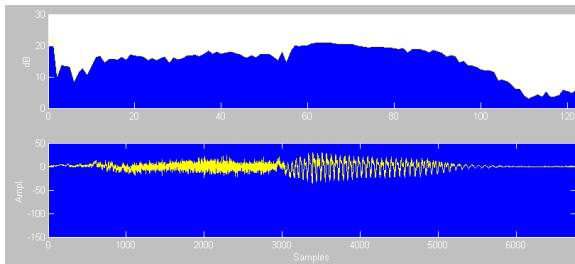
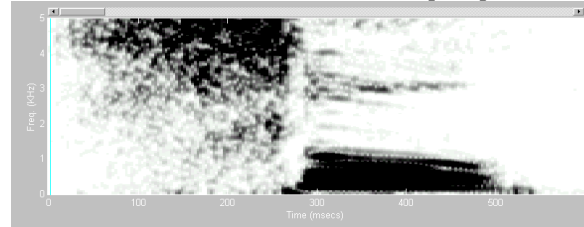
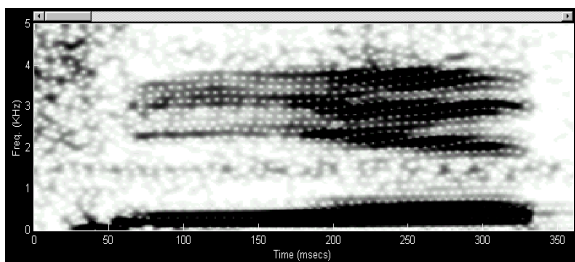
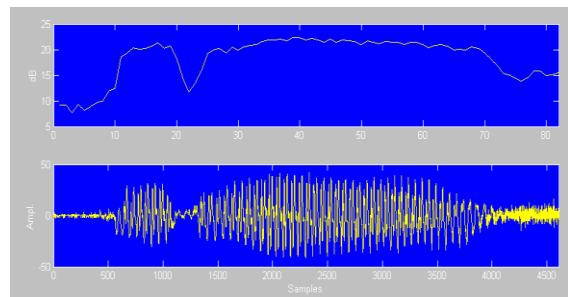
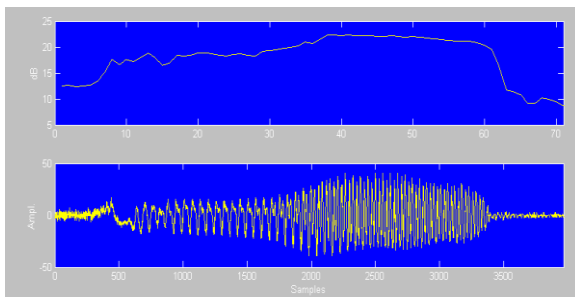
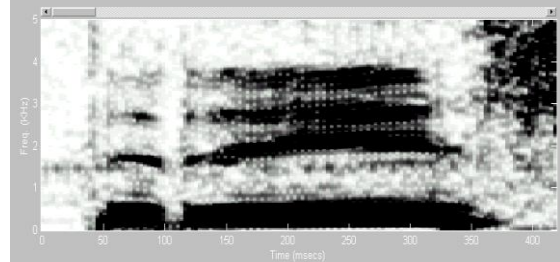


Fig. 5.3 Gráficas de energía y espectrograma de sílabas CV.

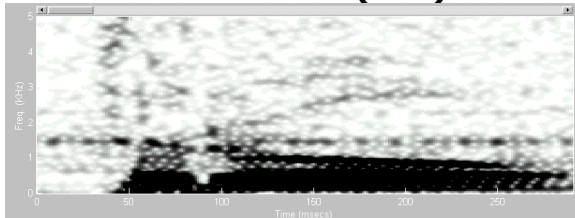
**sílaba sie (CVV)**



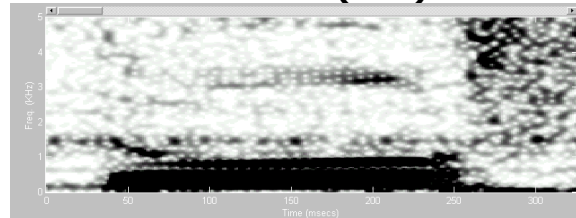
**sílaba tres (CCVC)**



**sílaba tro (CCV)**



**sílaba dos (CVC)**



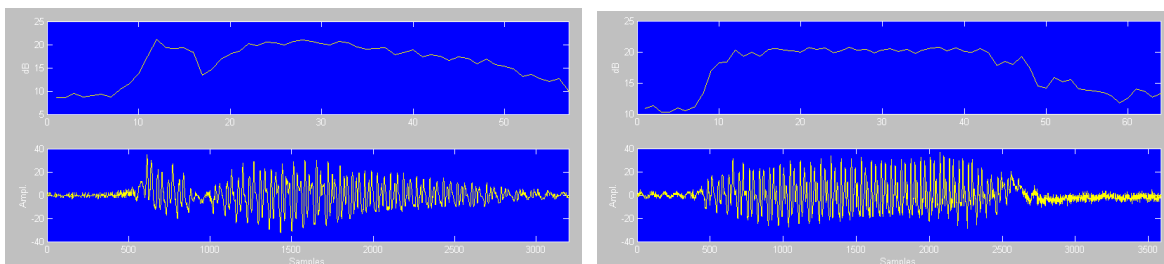


Fig. 5.4 Gráficas de energía y espectrograma de diferentes estructuras de sílabas.

Las figuras anteriores fueron extraídas para fines de análisis de la sílaba individual. Como se puede apreciar, la gráfica de la energía contiene altos niveles cuando se presenta una vocal a diferencia de cuando se manifiesta una consonante.

Asimismo, se puede observar el alto contenido de saturación en el dominio de la frecuencia de la señal al presentarse una consonante. Posteriormente a lo largo del presente trabajo, se analizarán diferentes formas de poder extraer esa información, la cual será relevante en el momento de realizar la segmentación de la señal de voz en sílabas.

Uno de los puntos importantes observados en el capítulo 3, es que existe gran preponderancia de las sílabas V, CV, VC, CCV, CVC, sobre las otras representaciones (VVV, CVVC, CVVC, etc.). Una vez realizadas las grabaciones correspondientes, se realizó la creación de un sistema de reconocimiento para el habla discontinua, para un determinado conjunto de sílabas mencionadas anteriormente. Generando los siguientes resultados de reconocimiento:

	TÉCNICA EMPLEADA	# DE UNIDADES DEL CORPUS	% DE RECONOCIMIENTO	% DE RECONOCIMIENTO ACUMULADO
V	LPC	5	98%	98%
VC-GI	LPC	5	96%	97%
VC-GII	LPC	9	96.66%	96.83%
CV-GI	LPC	5	96%	96.41%
CV-GII	LPC	5	100%	98.20%
CV-GIII	LPC	5	100%	99.10%
VC-GII	MARKOV	8	97.5%	98.30%

Tabla 5.2. Índices de reconocimiento para determinados tipos de sílabas en el español.

Donde a continuación se muestran las características de cada uno de los grupos analizados:

GRUPO V (vocal). a, e, i, o, u.

GRUPO VC-GI. el, es, ir, os, un.  
GRUPO VC-GII. al, am, el, em, es, in, ir, os, un.  
GRUPO CV-GI. la, le, li, lo, lu.  
GRUPO CV-GII. sa, se, si, so, su.  
GRUPO CV-GIII. ba, be, bi, bo, bu.

Los resultados obtenidos anteriormente demuestran el reconocimiento de sílabas individuales, y por tanto, su posible aplicación en los esquemas de reconocimiento de voz. Sin embargo, al realizar una comparación con el uso de los fonemas en las mismas aplicaciones (reconocimiento de fonemas individuales), éstos responden con mayor eficiencia, pues el reconocimiento alcanzó el 99.98% en promedio. No obstante, el porcentaje mostrado en la tabla anterior no es despreciable y comprueba la utilidad de la sílaba para los SRAH.

Las características del experimento se describen a continuación, haciendo especial énfasis en que los resultados obtenidos al usar las Cadenas Ocultas de Markov que mejoran el rendimiento del reconocimiento.

El filtro de predicción lineal de orden  $p$  determina el número de resonancias que serán encontradas, en donde cada resonancia requiere 2 raíces conjugadas complejas de  $A(z)$ . Una regla es permitir 2 polos por kHz de rango de frecuencia, más 3 o 4 polos por efectos de radiación y de la glotis (libro de Barrón).

Si la frecuencia de trabajo es de 5 kHz para ser muestreada a 10 kHz, se tomarán 10 polos más 3 o 4, lo que hace un total de 13 o 14 coeficientes.

Cabe destacar también que el grupo de VC-GII se escogió de acuerdo al conjunto de elementos de mayor preponderancia en un corpus de dígitos. Las demás estructuras, fueron analizadas por ser las de mayor aparición en los corpus analizados.

- ❖ La matriz de transiciones de Markov tuvo 6 estados, es decir, una matriz de 6x6.
- ❖ La matriz  $B$  de orden 32x12.
- ❖ El libro código para el caso de LPC tuvo un tamaño de 32x12.
- ❖ Cada sílaba se pronunció en intervalos de 5 veces.
- ❖ La ventana de Hamming fue de 256 puntos.
- ❖ El coeficiente del filtro de preénfasis fue de 0.98.
- ❖ Traslape entre tramas fue del 50%.
- ❖ Se usó la ecuación de Levinson-Durbin para encontrar los coeficientes LPC.

Estos datos son similares a experimentos donde se usó tanto LPC como patrón de reconocimiento. Tales valores fueron utilizados en (Oropeza, 2000) para el caso de palabras.

De la tabla 5.2 obsérvese el alto índice de reconocimiento que se obtiene al usar las sílabas y los coeficientes LPC como unidades de reconocimiento para los grupos CV-GI y CV-GII. Como resultado alentador se supera en estos casos en un .301875% al resultado obtenido al usar estructuras fonéticas.

Como medida de la eficiencia del uso de la sílaba para tareas de reconocimiento de voz, se construyó un corpus de dígitos (del 1 al 10), que comprende un total de 16 unidades silábicas, el cual fue probado usando tanto el patrón de reconocimiento LPC como de MARKOV.

Las siguiente figura 5.5 muestra el comportamiento en el dominio del tiempo y de la señal de energía en corto tiempo de una de las muestras del corpus antes mencionado:

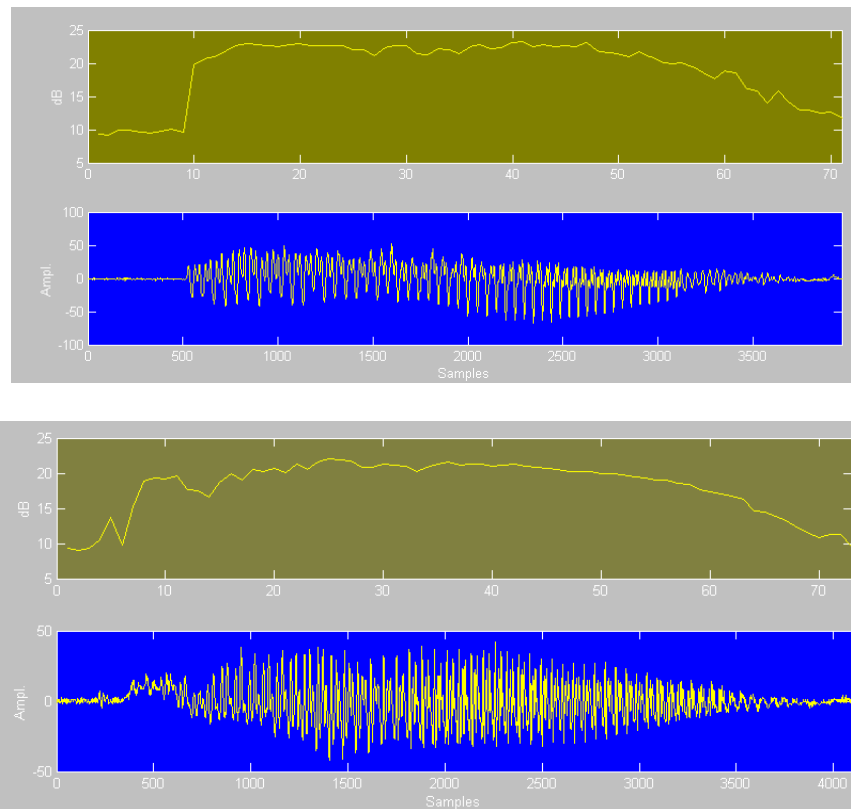


Fig. 5.5 Gráficas de sílabas del corpus de dígitos (sílabas u {V}, sílabas tres {CCVC}).

Como parámetro de segmentación de la señal de voz se utilizó la energía de la señal de voz. En la figura 5.6 se muestra una señal de voz antes y después de ser segmentada al usar este método de la energía.

La tabla 5.3 muestra el espacio de tiempo de cada una de las sílabas del corpus de dígitos usados para este caso. Observe los tiempos promedios encontrados para cada una de ellas.

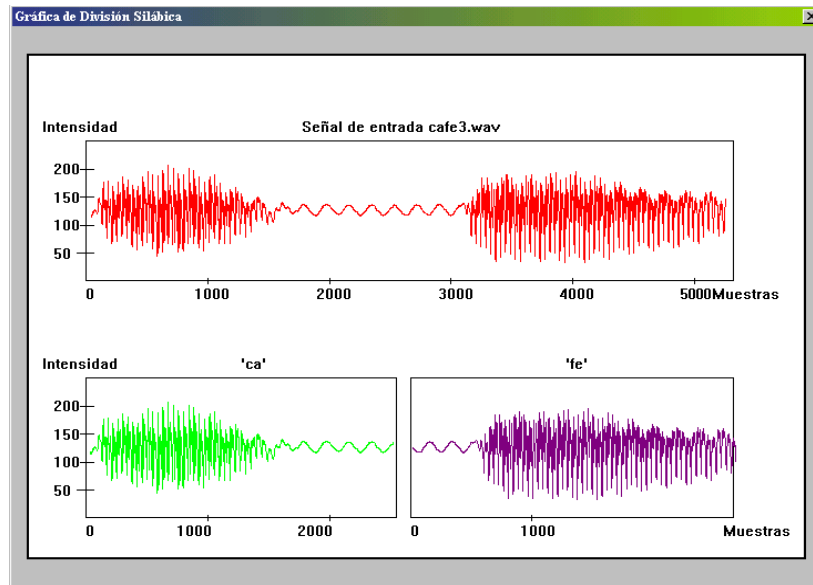


Fig. 5.6 Gráficas de sílabas segmentadas del corpus de dígitos.

<b>Sílaba</b>	<b>Estructura</b>	<b>Tiempo promedio (mseg)</b>
<b>u</b>	<b>V</b>	<b>3809</b>
<b>no</b>	<b>CV</b>	<b>493.6</b>
<b>dos</b>	<b>CVC</b>	<b>971.1</b>
<b>tres</b>	<b>CCVC</b>	<b>875.9</b>
<b>cua</b>	<b>CVV</b>	<b>479.2</b>
<b>tro</b>	<b>CCV</b>	<b>411.5</b>
<b>cin</b>	<b>CVC</b>	<b>764</b>
<b>co</b>	<b>CV</b>	<b>442.1</b>
<b>seis</b>	<b>CVVC</b>	<b>1181.1</b>
<b>sie</b>	<b>CVV</b>	<b>861.6</b>
<b>te</b>	<b>CV</b>	<b>406.7</b>
<b>o</b>	<b>V</b>	<b>437.6</b>
<b>cho</b>	<b>CCV</b>	<b>489.9</b>
<b>nue</b>	<b>CVV</b>	<b>631.7</b>
<b>ve</b>	<b>CV</b>	<b>429.8</b>
<b>diez</b>	<b>CVVC</b>	<b>865</b>

Tabla 5.3. Estructura silábica para un corpus de dígitos.

Las respuestas de tiempo pueden ser utilizadas como parámetros útiles para el reconocimiento que utiliza por ejemplo redes neuronales ya que pueden constituir un elemento de entrada a tales estructuras de reconocimiento de patrones, la información total obtenida se muestra en la siguiente figura 5.7:

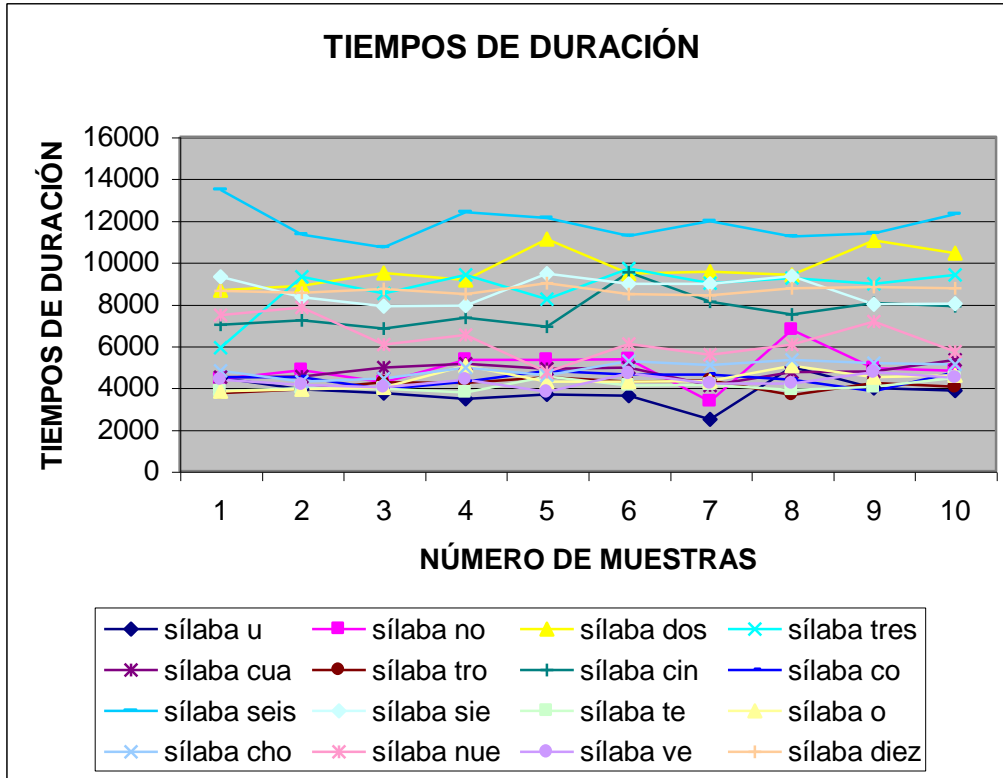


Fig. 5.7 Gráficas del tiempo de duración de las sílabas del corpus de dígitos.

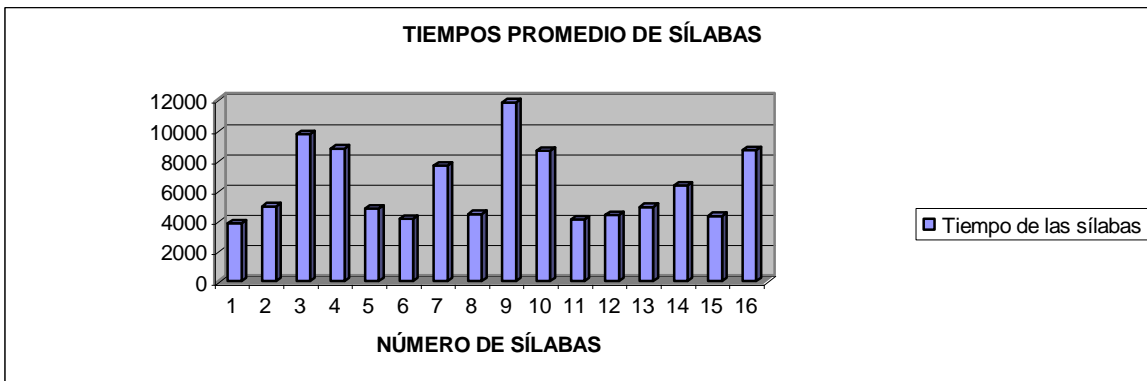


Fig. 5.8 Gráficas del promedio de duración de las sílabas del corpus de dígitos.

El proceso de reconocimiento se realizó para el habla no continua, creándose un libro código de 128 vectores LPC para el caso del uso de las Cadenas Ocultas de Markov. Del cual, se obtuvo un modelo por sílaba a reconocer, mismos que fueron comparados con las secuencias de señal de voz de entrada para el reconocimiento (Oropeza, 2000), lo cual presenta buenos resultados de acuerdo al trabajo realizado en anteriores sistemas de reconocimiento. En este caso la tarea de segmentación fue manual tanto en la etapa de entrenamiento como en la de reconocimiento.

Además:

- ❖ Los resultados obtenidos fueron para un solo locutor.
- ❖ Las cuestiones de umbral, cortes y demás, se realizaron de forma manual. Se consideró un valor por debajo del 90% del valor de los picos de energía superior encontrada a diferentes puntos de la muestra de voz.
- ❖ Cada palabra del corpus fue pronunciada de manera individual.

Encontrándose los siguientes resultados:

	<b>MARKOV</b>	<b>LPC</b>
<b>CORPUS DE DÍGITOS</b>	96.35%	95%

Tabla 5.4. Porcentajes de reconocimiento para un corpus de dígitos.

El siguiente paso fue segmentar las señales de voz usando un algoritmo especial para ello (basado en el contenido energético de la señal de voz pero de forma automática). Para estos experimentos se tuvieron las siguientes condiciones:

- ❖ Tanto el entrenamiento como el reconocimiento son de respuesta rápida, sin ayuda de ninguna herramienta más.
- ❖ En el caso de los valores de energía obtenidos, se seleccionó un umbral de 0.5 para los cortes, a partir del cual se tomaron las 4 muestras antecesoras y posteriores.
- ❖ La energía se calculó en tramas con un contenido de 128 muestras.
- ❖ Los cálculos se hicieron usando MATLAB y los resultados fueron para un locutor

A continuación se muestran tablas y gráficas que simbolizan la segmentación manual y los resultados obtenidos del reconocimiento.

#### MATRIZ DE VALORES DE ENERGIA PARA LA PALABRA CERO.

<b>ENERGÍA OBTENIDA POR SEGMENTOS DE 128 PUNTOS DE LA SEÑAL DE VOZ</b>									
0.1438	.01396	.01473	.01648	0.1500	0.1247	0.0936	0.0539	0.0797	0.2409
0.3254	0.2404	0.2286	0.2704	0.2291	0.1638	0.1607	0.1774	0.1698	0.1735
0.1677	0.1656	0.1740	0.1958	0.2440	0.2595	0.2228	0.2122	0.2158	0.2025
0.2056	0.6649	1.4025	3.1734	4.1051	5.5392	7.9105	8.8134	9.5190	10.2475
11.9152	13.4525	15.0421	16.4577	17.5969	17.9182	17.2774	15.7152	13.2223	11.4707
9.0478	6.3210	4.4554	2.5601	1.3398	0.8280	.03524	0.3288	0.5000	0.5591
0.4745	0.3623	0.2623	0.2239	0.1999	0.1939	0.1752	0.1451	0.1867	0.2479
0.2399	0.2063	0.2065	0.1915	0.1948	0.2310	0.2146	0.1925	0.2012	0.1870
0.1747	0.1706	0.1659	0.1671	0.1765	0.1749	0.1699	0.1492	0.6112	0.7737
0.6923	1.6484	3.6975	5.3254	6.9727	9.1572	11.9152	12.3934	12.9217	13.3000
14.2120	13.1026	11.9152	11.1012	11.4090	11.2646	8.8199	5.8689	4.0442	2.7961
2.1538	1.7145	1.1036	0.5715	0.2503	0.3454	0.3578	0.2634	0.2116	0.2157
0.2241	0.2420	0.2478	0.2285	0.2085	0.2050	0.1915	0.1869	0.1955	0.2013
0.1810	0.1616	0.1568	0.1479	0.1559	0.1767	0.1705	0.1540	0.1574	0.1762



<b>0.1702</b>	<b>0.1586</b>	<b>0.1610</b>	<b>0.1540</b>	<b>0.1440</b>	<b>0.1389</b>	<b>0.1294</b>	<b>0.1305</b>	<b>0.1312</b>	<b>0.1265</b>
<b>0.1340</b>	<b>0.1372</b>	<b>0.1377</b>	<b>0.1370</b>	<b>0.1314</b>	<b>0.1331</b>	<b>0.1365</b>	<b>0.1353</b>	<b>0.1302</b>	<b>0.1243</b>
<b>0.1212</b>	<b>0.1291</b>	<b>0.1457</b>	<b>0.1457</b>						

Tabla 5.5. Valores obtenidos de una señal de voz y el proceso de segmentación en sílabas de forma manual y automática.

En este caso, cada elemento (celda) representa un segmento de 128 valores de la señal de voz al usar una ventana de 256 muestras, pues cada segmento se encuentra traslapado. De la tabla 5.5 se encuentra que el segmento 33 sobrepasa el umbral de 0,5, por lo que se toma 4 segmentos anteriores. Es decir, en el 29 que tiene el valor de 0.2122.

El punto final está en el segmento 61, por lo que se toma el 65, que tiene el valor de 0.2239. Por lo que los valores iniciales están en la muestra 3712 y el final de la sílaba {ce} está en 8320. La siguiente figura 5.9 muestra las dos formas de segmentación en unidades silábicas para el caso de los dos experimentos antes expuestos. Se usó el software de Creative Sound Blaster para la segmentación manual y MATLAB para la segmentación automática, tal y como lo ilustra las figuras 5.9 y 5.10:

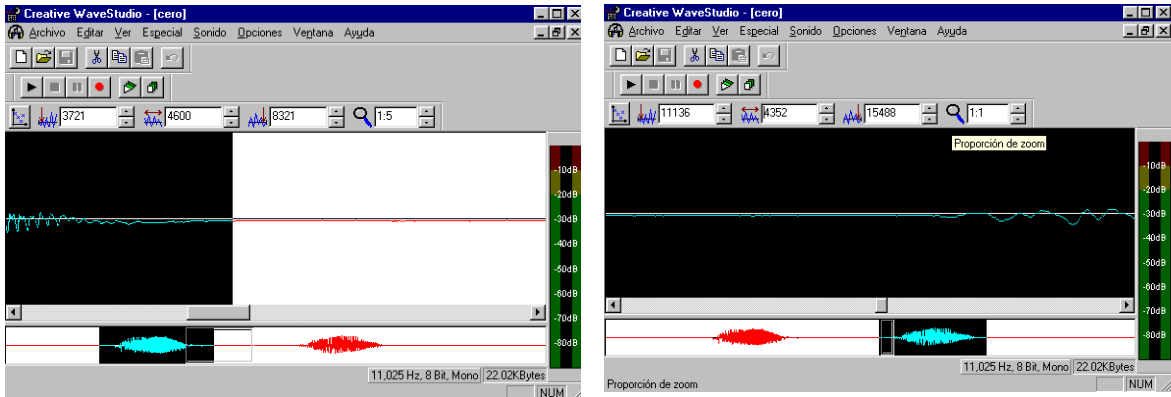


Fig. 5.9 Gráficas de segmentación en unidades silábicas de una palabra de forma manual.

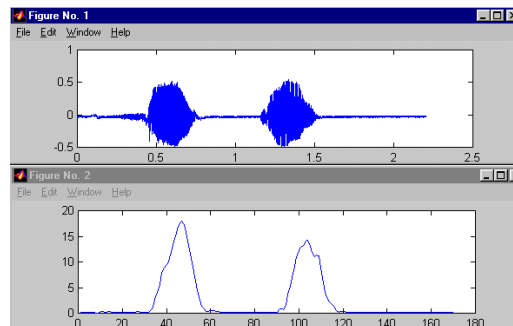


Fig. 5.10 Gráficas del proceso de segmentación y cálculo de la energía del corpus de dígitos en forma automática.

Como ejemplo de la segmentación manual se tiene la siguiente figura 5.11, en donde la sílaba a segmentar se encuentra en el intervalo de 87 (11136) hasta el 121 (15488):

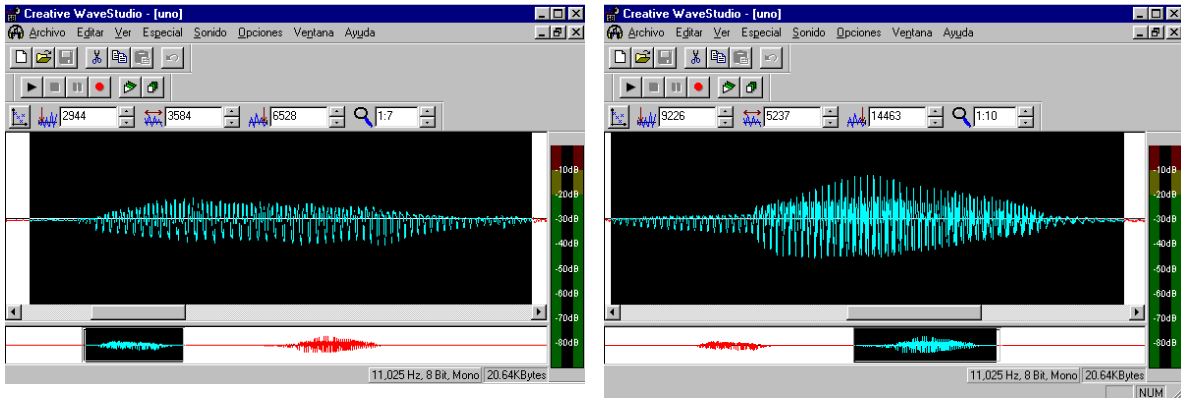


Fig. 5.11 Gráficas de la segmentación silábica manual de una sílaba del corpus de dígitos.

Una vez realizados los cortes a las señales de voz se procedió a realizar el entrenamiento, obteniéndose los modelos correspondientes a cada una de las sílabas que comprenden los dígitos del 1 al 10, con lo que se obtuvieron los siguientes resultados experimentales:

	u	no	dos	tres	cua	tro	cin	Co	seis	sie	te	o	cho	nue	ve	diez
u	10															
no		10														
dos			10													
tres				10												
cua					10											
tro						10										
cin							10									
co								10								
seis									9							1
sie										9						1
te											10					
o												10				
cho													10			
nue														10		
ve															10	
diez																10

Tabla 5.6. Tabla de confusión para el caso del corpus de dígitos.

Al hacer uso de las Cadenas Ocultas de Markov obtuvo el 99.21% de reconocimiento, incrementándose con relación al uso de LPC y del experimento anterior. Los valores de experimentación fueron utilizados en (Oropeza, 2000).

Haciendo uso de LPC se llegó a los siguientes resultados:

<b>LPC</b>	98%
<b>HMM</b>	99.21%

Tabla 5.7. Porcentaje de reconocimiento usando una segmentación basada en energía.

El incremento en el índice de reconocimiento conlleva a suponer, que al usar las Cadenas Ocultas de Markov y un algoritmo más sofisticado de segmentación, el índice se incrementa radicalmente.

Como parte complementaria del análisis de la señal de voz, se realizó un análisis de la estructura del Latino40, con la finalidad de realizar un estudio de su comportamiento a la cuestión del reconocimiento basado en sílabas.

Las siguientes son las consideraciones tomadas en cuenta:

- ❖ No se tomaron en cuenta los símbolos ortográficos.
- ❖ Se consideran como palabras los nombres propios y de lugares.
- ❖ Se respetaron las 11 reglas silábicas.

La siguiente tabla presenta el porcentaje de frecuencia del vocabulario como de la base de datos de palabras en función del número de sílabas que contienen (Martínez, 2002):

<b>N</b>	<b>% vocabulario</b>	<b>% acumulado del vocabulario</b>	<b>% base de datos</b>	<b>% acumulado de la base de datos</b>
1	1.93	1.93	39.07	39.07
2	23.01	24.94	26.09	65.16
3	38.10	63.04	21.65	86.61
4	26.03	89.07	10.04	96.85
5	8.94	98.01	2.56	99.41
6	1.71	99.72	0.54	99.95
7	0.22	99.94	0.04	99.99
8	0.03	99.97	0.005	99.995

Tabla 5.8. Frecuencia de aparición de palabras con N sílabas en el diccionario y corpus del Latino40.

De la tabla anterior se observa que las palabras monosílabas representan más de la tercera parte de las palabras de la base de datos. Aunque sólo ocupen el 2% del total del vocabulario. A su vez, las palabras con 5 sílabas representan sólo el 2% de las palabras del vocabulario.

Analizando dichas palabras monosilábicas se muestra la siguiente tabla:

<b>Palabra</b>	<b>Configuración de sílaba</b>	<b>No. ejemplos</b>	<b>% vocabulario</b>
De	<b>Oclusiva sorda +vocal</b>	1760	11.15
La	<b>Líquida+vocal</b>	1481	9.38
El	<b>Vocal+líquida</b>	1396	8.85

En	<b>Vocal+nasal</b>	1061	6.72
No	<b>Nasal+vocal</b>	1000	6.33
Se	<b>Fricativa + vocal</b>	915	5.80
Que	<b>Oclusiva sorda + vocal</b>	891	5.64
A	<b>Vocal</b>	784	4.97
Los	<b>Líquida + vocal + fricativa</b>	580	3.67
Es	<b>Vocal + fricativa</b>	498	3.15

Tabla 5.9. Frecuencia de aparición de 10 monosílabas más usadas dentro del Latino40.

Como se puede observar, los artículos (la, el, los) y preposiciones (a, de, en, que), conforman el 65.66% del total de monosílabas del corpus. Su estructura es simple CV o VC. Estos datos refuerzan el por qué de la importancia de haber realizado con anterioridad los estudios de reconocimiento de estas unidades básicas.

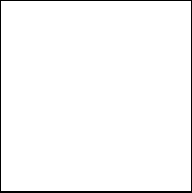
La siguiente tabla presenta el porcentaje de aparición de las estructuras silábicas de dicho corpus, del cual se tiene el reporte de 19035 sílabas. Como comparación con respecto al inglés, las dos estructuras silábicas básicas del Español abarcan el 75% del total de las sílabas utilizadas, mientras que en el inglés, éstas abarcan sólo el 30% del total (según datos obtenidos por Wu, 1998).

<b>Tipo de estructura</b>	<b>% vocabulario</b>	<b>% acumulado vocabulario</b>
CV	50.72	50.72
CVC	23.67	74.39
V	5.81	80.2
CCV	5.13	85.33
VC	4.81	90.14
CVV	4.57	94.71
CVVC	1.09	95.8

Tabla 5.10. Análisis a nivel de sílabas del corpus de voz Latino40.

### 5.3 RESUMEN DEL CAPÍTULO

A lo largo del presente capítulo se describió la creación de un Sistema Experto desarrollado para fines de integrar las reglas de las sílabas del español en la tarea de reconocimiento de voz por computadora. Se analizó su integración a tales sistemas llegando a demostrar su utilidad en los sistemas de reconocimiento del habla discontinua. Se procedió posteriormente a comprobar la utilidad de tal reconocimiento cuando se toman unidades básicas, como lo son las sílabas agrupadas en diferentes divisiones.



Para finalizar se procedió a ver la respuesta que tenía el usar tales unidades en un corpus de dígitos, el cual nos demostró que existe un gran problema con el aspecto de la segmentación al hacer uso de la energía para tal fin.

Finalmente, los resultados obtenidos nos dan pauta a buscar nuevos esquemas de segmentación y algoritmos más óptimos para tal labor.

# CAPÍTULO 6

---

## Método de reconocimiento e integración de las sílabas a los SRAH del habla continua.

Una de las tareas importantes dentro del reconocimiento de voz es el poder realizar reconocimiento del habla continua. A lo largo del tiempo esta tarea ha sido llevada a cabo haciendo uso de los fonemas y de los modelos ocultos de Markov.

Se presenta en primera instancia el análisis de la introducción del paradigma de la sílaba a sistemas del habla discontinua y después al habla continua.

En el presente capítulo se estudian las características de la introducción de las sílabas a dichos esquemas de reconocimiento. Se hace uso de las Cadenas Ocultas de Markov Discontinuas y de Densidad Continua para realizar las tareas de reconocimiento. Se comprueba que las unidades silábicas representan una alternativa interesante para ser tomada en cuenta como elemento de referencia.

Se considera el uso extenso de la sílaba como elemento de referencia para aplicaciones que tienen una alta dependencia del contexto, las aplicaciones mostradas a continuación denotan el hecho de que es posible incrustar tales aplicaciones a corpus más extensos.

## 6.1 ANÁLISIS DE UN ALGORITMO DE RECONOCIMIENTO DE VOZ DEL HABLA DISCONTINUA

A continuación se describe el desarrollo de un algoritmo creado para fines de reconocimiento de voz del habla no continua, mismo que se creó con esta finalidad. Cabe destacar que este algoritmo representa una alternativa en su implantación a los empleados con anterioridad aunque está basado en la siguiente expresión:

$$a_{ij} = \frac{\text{número esperado de transiciones del estado } s_i \text{ al estado } s_j}{\text{número esperado de transiciones del estado } s_i}$$

$$b_j(k) = \frac{\text{número esperado de veces que se analiza el estado } j \text{ y se observa el símbolo } v_k}{\text{número esperado de veces que se analiza el estado } j.}$$

En donde como se observa, no se estima el valor de las probabilidades iniciales del modelo  $\pi$  ya que estos valores por lo regular no son estimados nuevamente. En lo consecuente se muestra el análisis y repercusión que tiene la incrustación de este algoritmo a un corpus de voz, analizado desde el punto de vista de la segmentación silábica basada en energía.

El siguiente análisis muestra el comportamiento de este algoritmo para diferentes estados del corpus a analizar.

### 6.1.1 ENTRENAMIENTO

Para realizar el entrenamiento, el cual consiste en generar el libro código global y los modelos ocultos de Markov para cada elemento, se usó el mismo tipo de sistema utilizado en los experimentos descritos en el capítulo 5. Con lo cual, se obtuvieron cada uno de los modelos ocultos de Markov necesarios para poder realizar la tarea de reconocimiento.

### 6.1.2 RECONOCIMIENTO

La siguiente figura 6.1 muestra una señal de voz y su gráfica de energía correspondiente a uno de los elementos del corpus elegido.

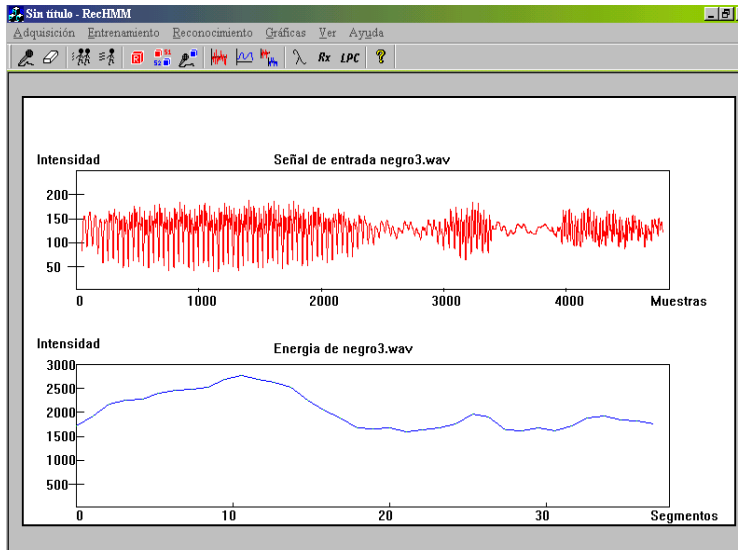


Figura 6.1. Gráficas de señal de voz y energía de la palabra 'negro'.

La siguiente gráfica 6.2 muestra la forma en la que se divide el comando a reconocer en dos diferentes partes a partir de la energía extraída de la señal original.

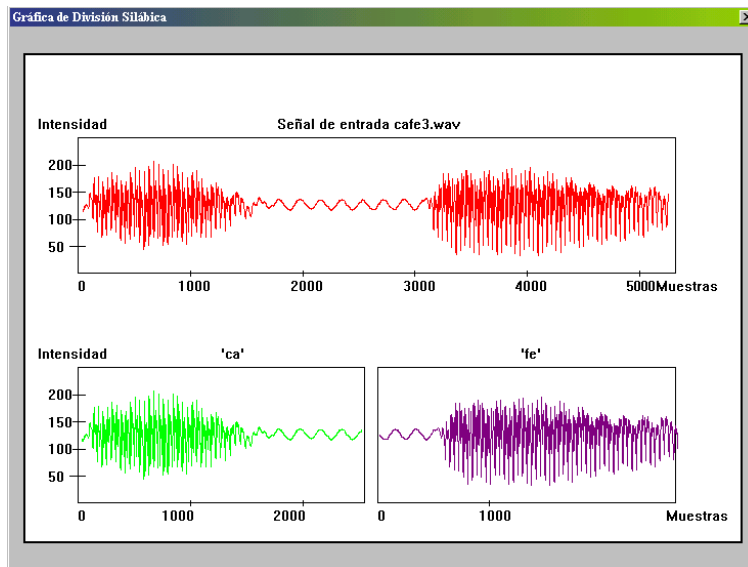


Figura 6.2. Gráficas de las sílabas 'ne'-'gro'.

La figura 6.3 muestra la información sobre el contenido de los vectores de autocorrelación obtenidos de uno de los elementos del corpus a tratar.



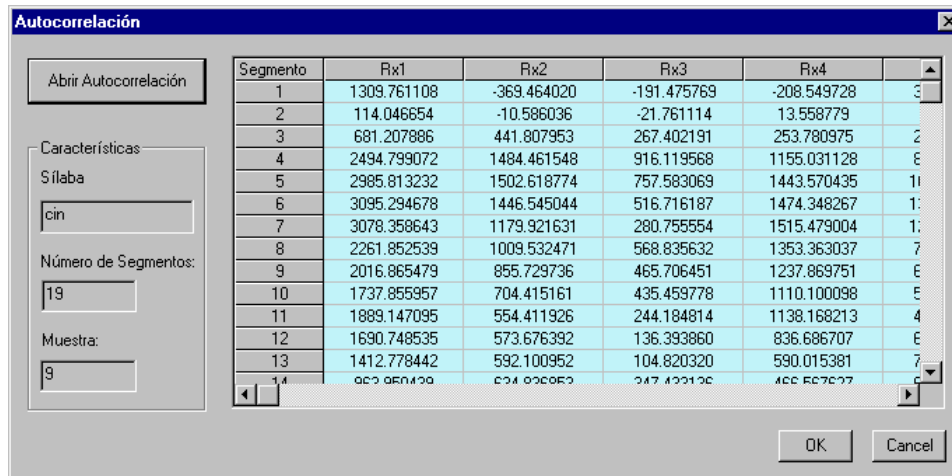


Figura 6.3. Caja de diálogo de vectores de autocorrelación.

En la figura 6.4, se listan algunos coeficientes LPC pertenecientes al libro código global con los que se generaron los modelos ocultos de Markov por cada sílaba.

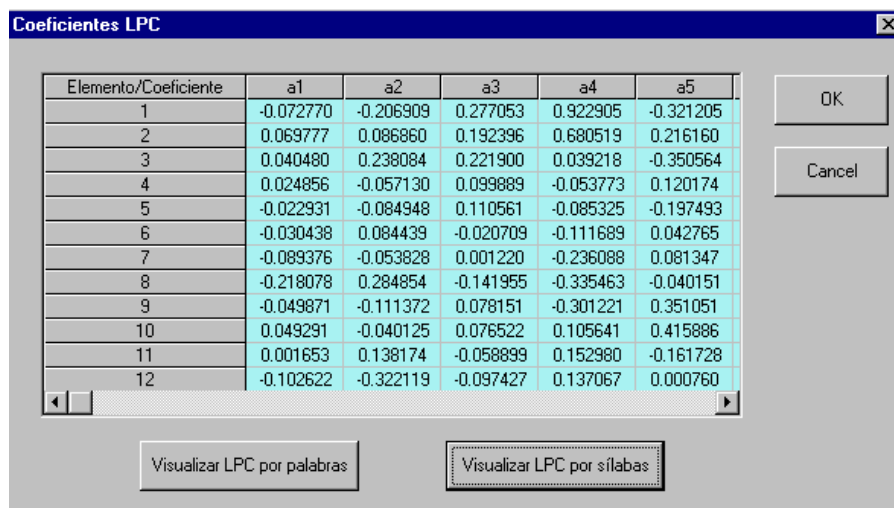


Figura 6.4. Caja de diálogo de coeficientes LPC.

La siguiente figura 6.5 muestra los parámetros del modelo oculto de Markov para una de las sílabas del corpus analizado.

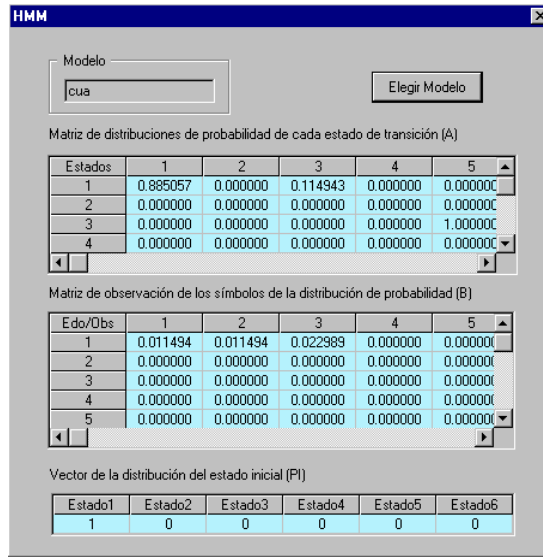


Figura 6.5. Datos resultantes de un modelo oculto de Markov optimizado.

## 6.2 EVALUACIÓN DE LA EFECTIVIDAD

Este apartado está dirigido a evaluar si el funcionamiento del sistema es el esperado. Se comienza mostrando como se realiza la distribución de regiones en el sistema, posteriormente se hace un análisis del índice de error que presenta en cada una de las optimizaciones del libro código, así como la comparación entre el reconocimiento por sílabas y el reconocimiento por palabras, finalmente se determina la tasa de error del reconocimiento.

### 6.2.1 DISTRIBUCIÓN EN REGIONES

Sabemos que dado un libro código se puede partir un espacio muestra  $X$  en  $N$  regiones disjuntas y a cada una de éstas se les asocia un centroide. Este se optimiza para convertirlo en el mejor representante de la región, el que en promedio difiere menos de cada uno de los vectores de la región.

Podemos observar la distribución de vectores en la figura 6.6, los cuales en conjunto forman 4 regiones, y su centroide fue optimizado 7 veces, a medida que los vectores cambiaban de región, se obtenía un nuevo centroide, y había que recalcularlo, por ello observamos variaciones al principio de la gráfica, pero también observamos que se vuelve estable, pues el centroide ya no cambia en las últimas 3 optimizaciones y por consiguiente, ya no hay intercambios de vectores hacia otras regiones.

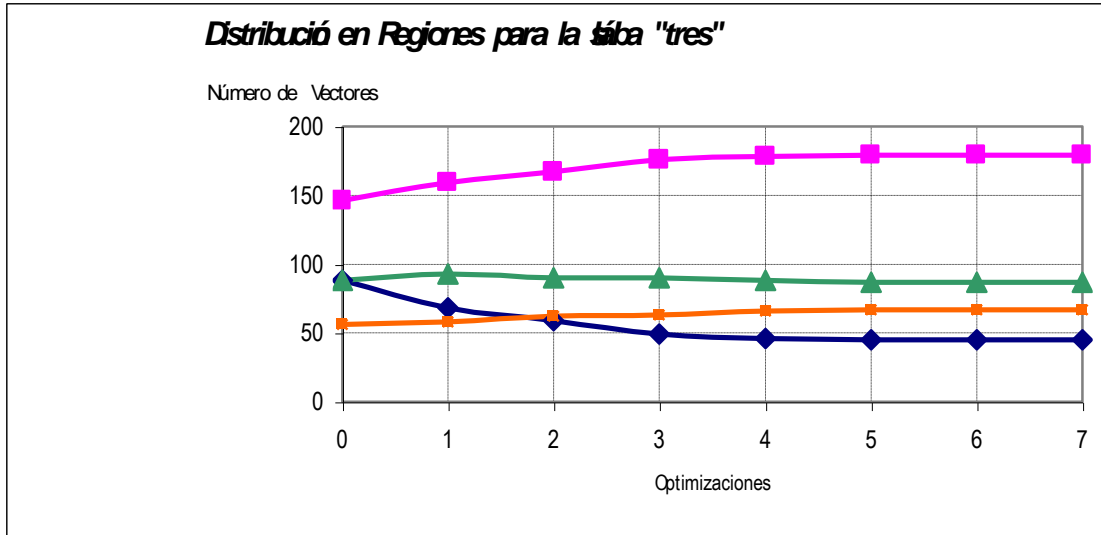


Figura 6.6. Gráfica de la Distribución en regiones de la sílaba "tres" para 4 Regiones.

En la figura 6.7 se observa el mismo proceso, con la diferencia que ahora se están modelando 16 regiones y la estabilidad se logró hasta la optimización número 12.

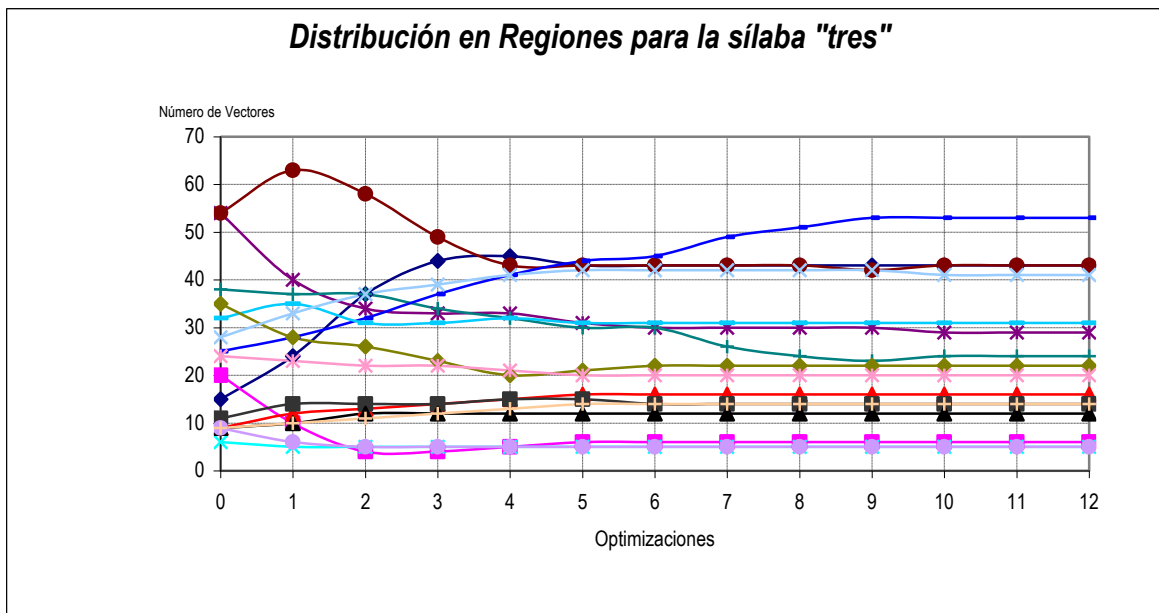


Figura 6.7. Gráfica de la Distribución en regiones de la sílaba "tres" para 16 Regiones.

Los valores con los que fueron generadas las gráficas se encuentran en el Anexo 1. En las gráficas se puede observar que el desempeño del sistema fue el esperado porque fue distribuyendo los datos hasta lograr el centroide óptimo, sin la adecuada distribución no se podrían obtener los coeficientes LPC deseados para lograr el reconocimiento. Cabe destacar que el proceso se realizó hasta 128

regiones, pero no se muestra debido a la cantidad de datos que se está manejando, por esta razón sólo mostramos la distribución de regiones final, la cual se muestra en la figura 6.8.

Cabe hacer la aclaración de que este proceso representa un gran coste computacional en el caso de tener que agregar un nuevo elemento al corpus, es decir una nueva palabra, dado que se tendría que generar de nueva cuenta el libro código global y realizar todo el proceso de la reestimación de los parámetros dados por la expresión propuesta.

También es importante recalcar que aunque tedioso este proceso se realiza en un mínimo de tiempo debido a la velocidad de procesamiento de los equipos de cómputo actuales.

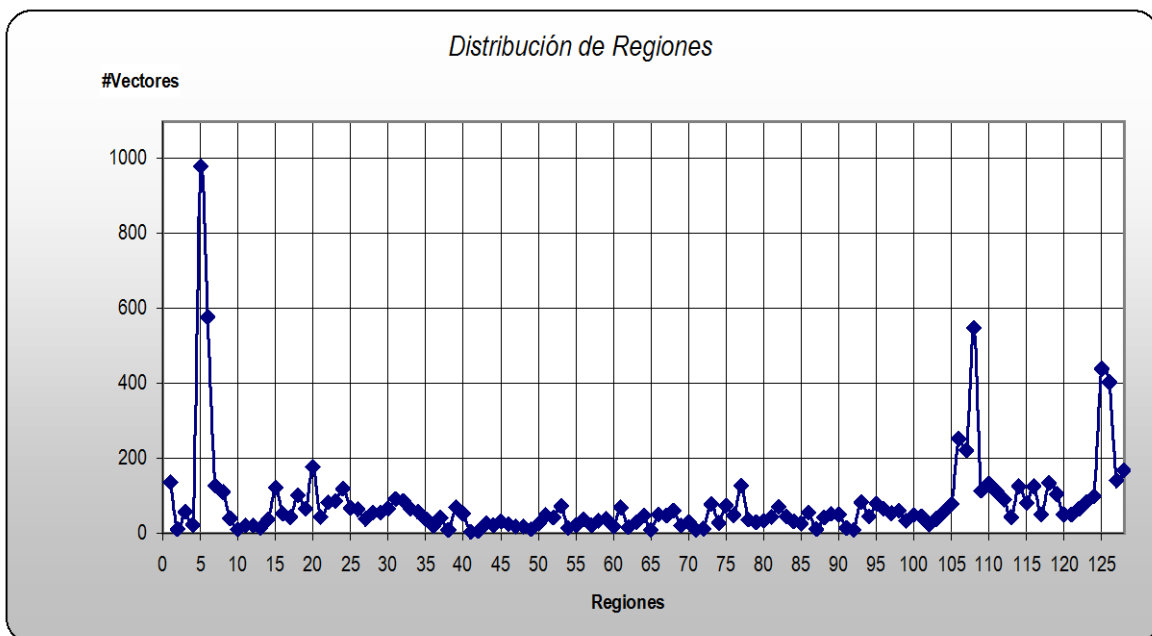


Figura 6.8. Gráfica de distribución de las regiones de un corpus experimental .

En la generación de los modelos ocultos de Markov para sílabas se determinó el número de optimizaciones con los que trabaja mejor el sistema, se estuvo evaluando para 32 regiones, con 200 muestras con las cuales fue entrenado el sistema, 2 locutores, y como se muestra en la gráfica de la figura 6.9 el número de errores está por encima de 45.

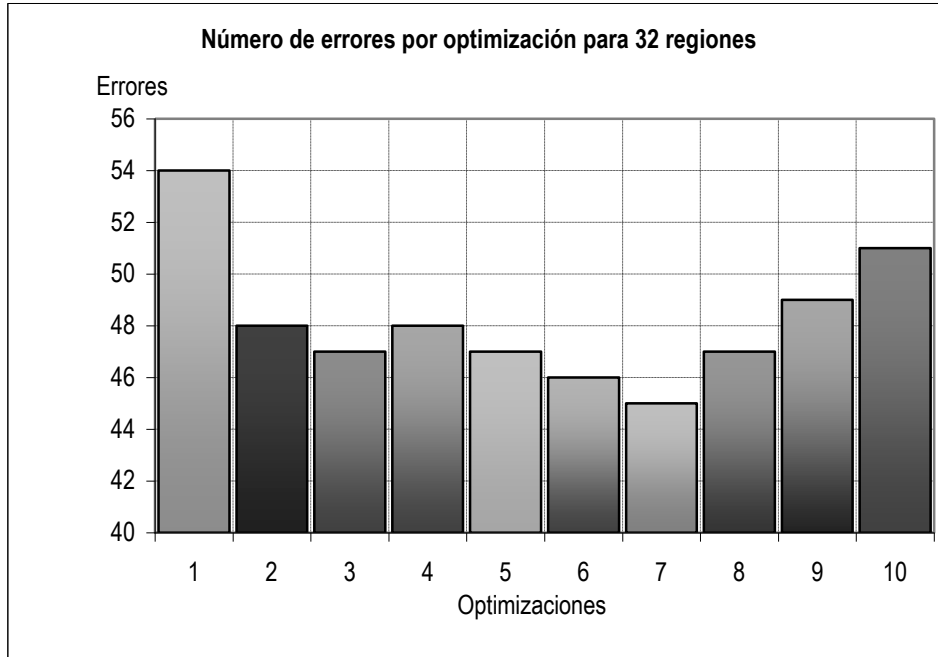


Figura 6.9. Diagrama de errores de optimización para 32 regiones con 200 muestras usando sílabas.

En la figura 6.10 se muestra el desempeño del sistema para 64 regiones las mismas muestras, y como se puede observar, el número de errores ha disminuido, ahora se encuentra por encima de 14.

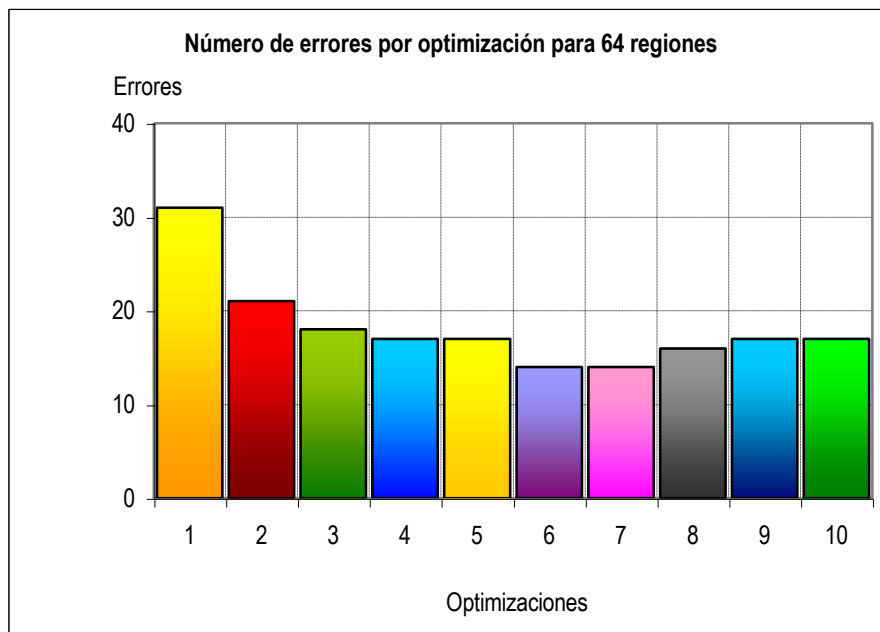


Figura 6.10. Diagrama de errores de optimización para 64 regiones con 200 muestras usando sílabas.

Por último, la figura 6.11 muestra la gráfica del número de regiones elegido, el cual es de 128 pues como se puede observar, el desempeño es mucho mejor que en los casos anteriores, no continuamos incrementando el número de regiones debido al tamaño del diccionario utilizado en el sistema, otra de las razones fue que representaría un incremento innecesario en el tiempo de cómputo.

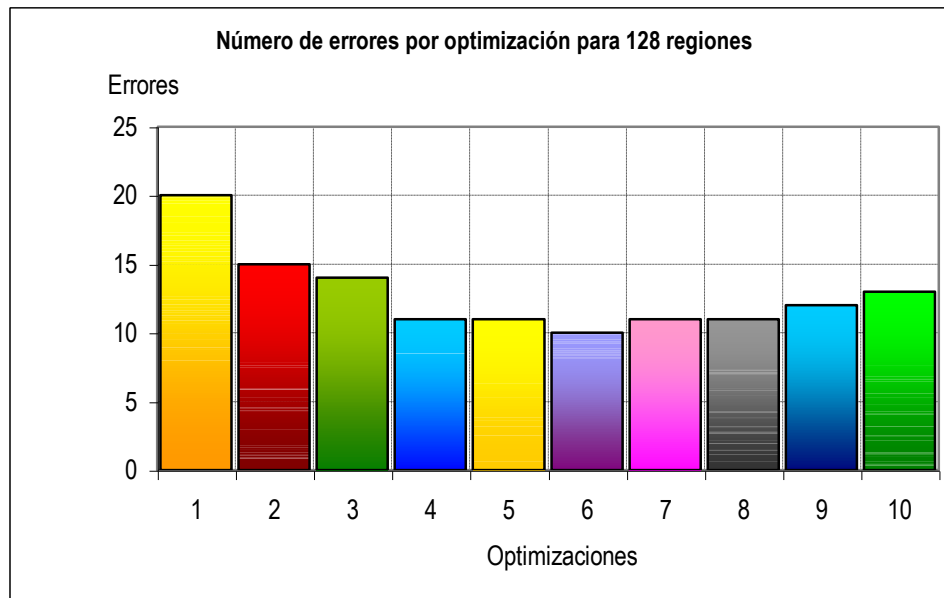


Figura 6.11. Diagrama de errores de optimización para 128 regiones con 200 muestras usando sílabas.

Como se puede observar en las gráficas del número de errores por optimización, existe un mejor desempeño en el intervalo de optimizaciones de 5 a 7, porque después de este intervalo el número de errores comienza a incrementarse. Por esta razón se tomó que el número de optimizaciones en la generación de los modelos ocultos de Markov sea de 6.

Se hizo el análisis anterior pero con más locutores para determinar de qué manera influye la variabilidad de las distintas señales de voz pues ahora se realizó con 4 locutores y con 400 muestras. En la figura 6.12 se muestra la gráfica del número de errores por optimización para 32 regiones, como se puede observar el menor error se encuentra en la optimización 6.

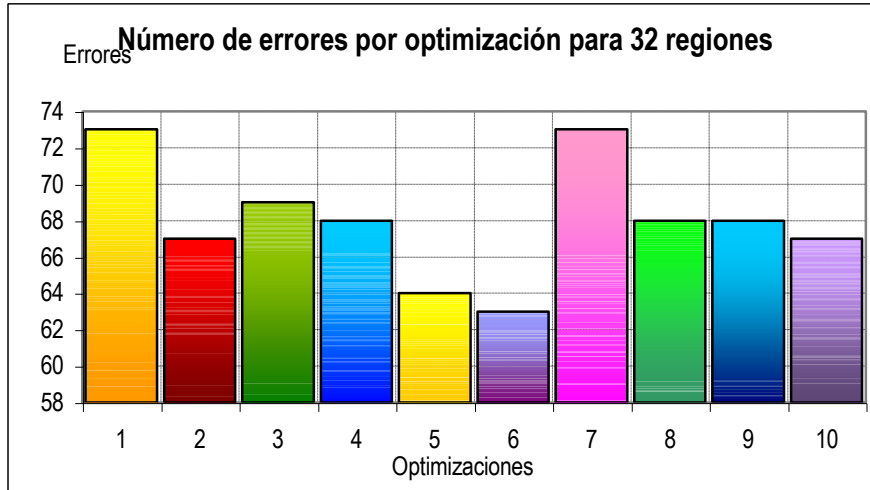


Figura 6.12. Diagrama de errores de optimización para 32 regiones con 400 muestras usando sílabas.

En la figura 6.13 se muestra el desempeño del sistema para 64 regiones con las mismas muestras, y como se puede observar, el número de errores ha disminuido, ahora está por encima de 39.

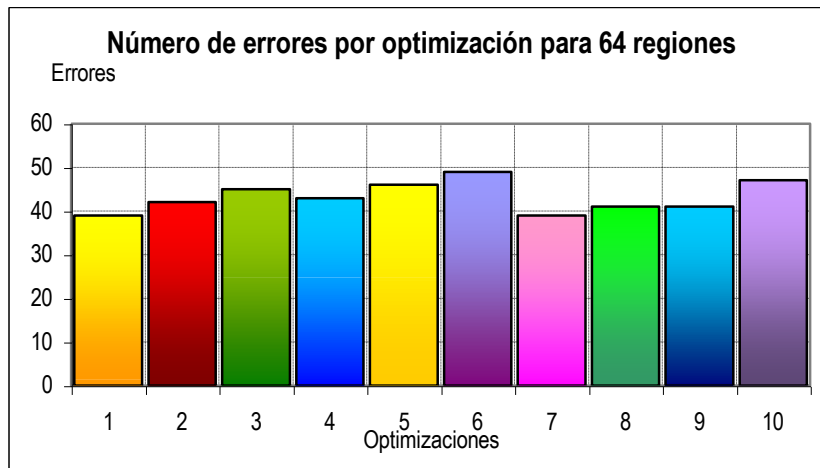


Figura 6.13. Diagrama de errores de optimización para 64 regiones con 400 muestras usando sílabas.

En la figura 6.14 se muestra la gráfica con 128 regiones y como se puede observar, el desempeño es mucho mejor que en los casos anteriores, pues con las mismas muestras los errores está por encima de 20.

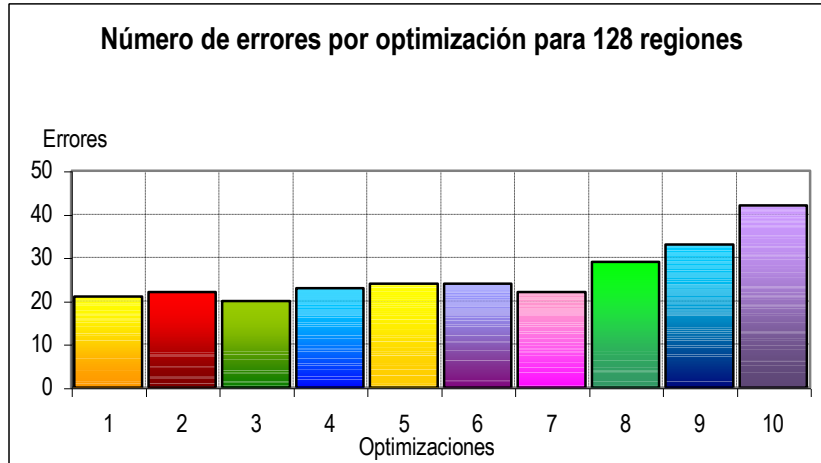


Figura 6.14. Diagrama de errores de optimización para 128 regiones con 400 muestras usando sílabas.

A continuación se mostrarán el número de errores por optimización en la generación de los modelos ocultos de Markov para el caso de palabras, en la figura 6.15 se muestra la gráfica para 200 muestras con las cuales fue entrenado el sistema, 2 locutores y 32 regiones, como se puede observar el menor error se encuentra en la optimización 2.

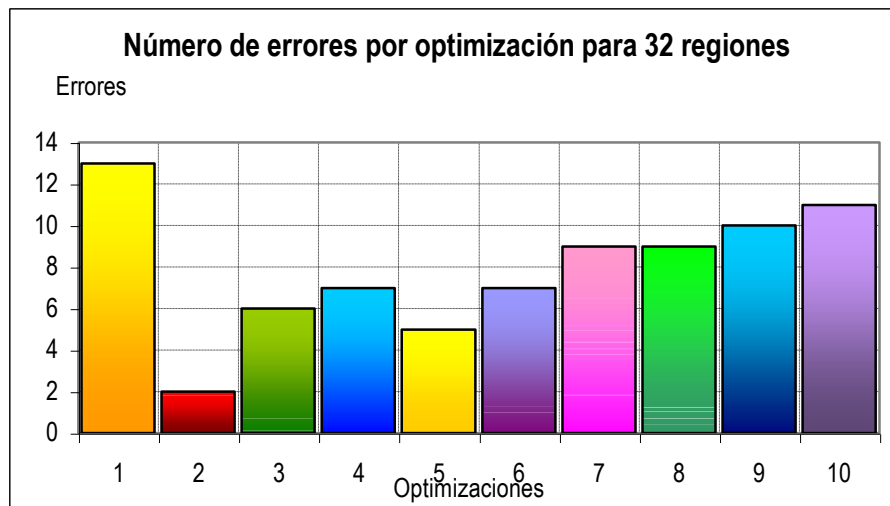


Figura 6.15. Diagrama de errores de optimización para 32 regiones con 200 muestras usando palabras.

En la figura 6.16 se muestra la gráfica para 64 regiones, el menor error está en la optimización 2.



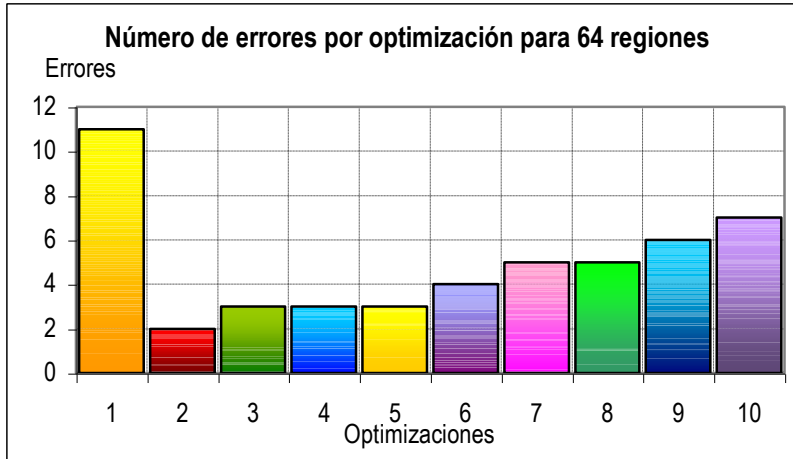


Figura 6.16. Diagrama de errores de optimización para 64 regiones con 200 muestras usando palabras.

En la figura 6.17 se muestra la gráfica para 128 regiones y como se puede observar, el desempeño es mucho mejor que para 32 o para 64 regiones, pues con las mismas muestras el número de errores que se logran es cero para la optimización 2 y 3, y en las otras optimizaciones el error es muy pequeño.

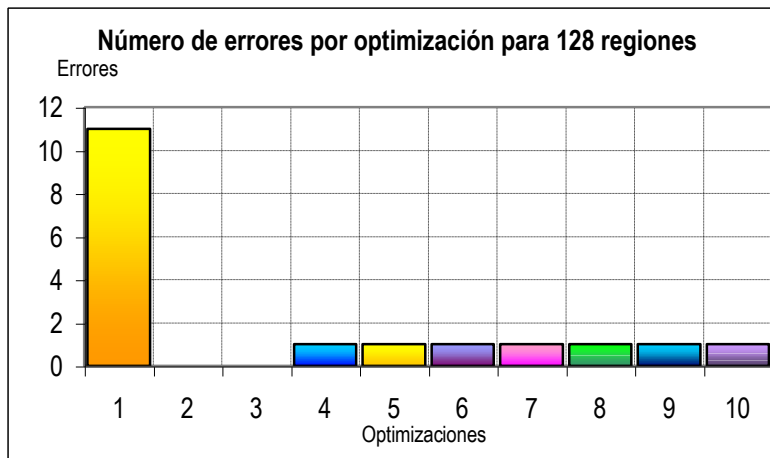


Figura 6.17. Diagrama de errores de optimización para 128 regiones con 200 muestras usando palabras.

El análisis anterior se realizó nuevamente pero con más locutores y como consecuencia con un número mayor de muestras, ahora se realizó con 4 locutores, 400 muestras con las cuales fue entrenado el sistema, para 32 regiones, en la figura 6.18 se muestra la gráfica del número de errores por optimización, el menor error se encuentra en las últimas optimizaciones.

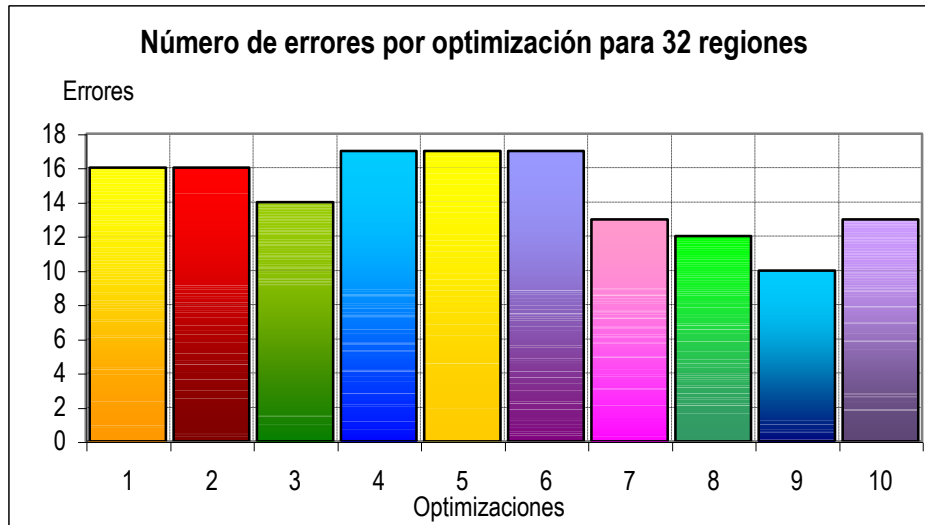


Figura 6.18. Diagrama de errores de optimización para 32 regiones con 400 muestras usando palabras.

En la figura 6.19 se muestra la gráfica para 64 regiones, el menor error está en la optimización 7.

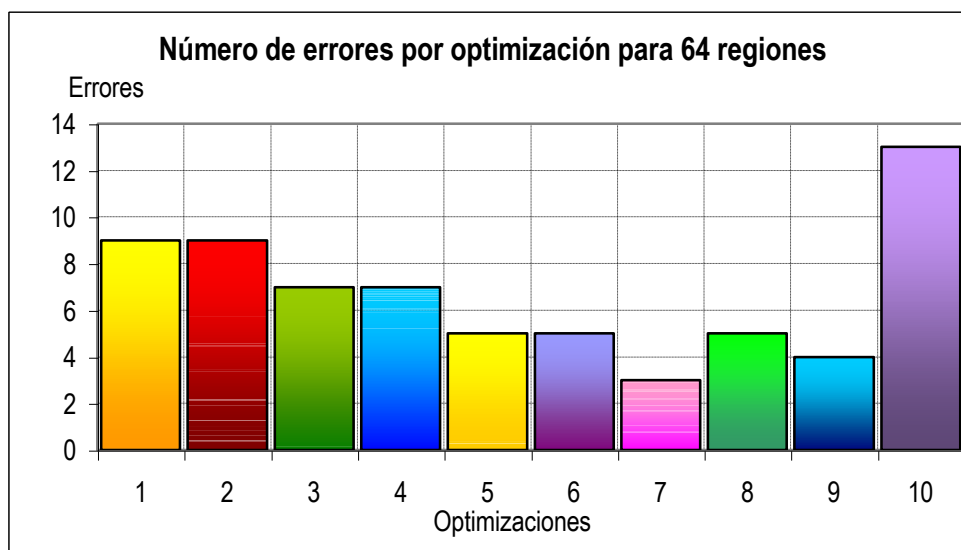


Figura 6.19. Diagrama de errores de optimización para 64 regiones con 400 muestras usando palabras.

En la figura 6.20 se muestra la gráfica con 128 regiones y como se puede observar, el desempeño es mucho mejor que para 32 o para 64 regiones, pues con el mismo número de muestras el número de errores está por debajo de 4, en la optimización 7 se logra que sólo exista un error.

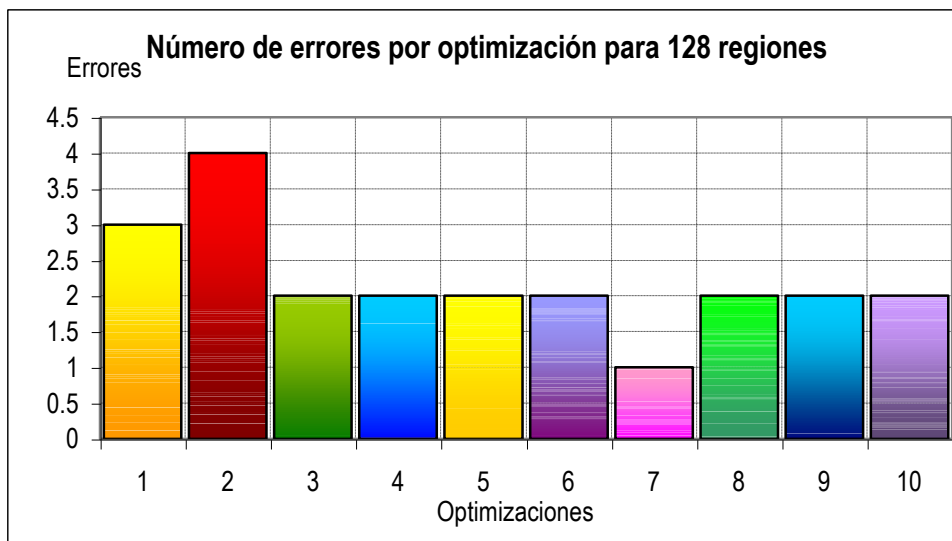


Figura 6.20. Diagrama de errores de optimización para 128 regiones con 400 muestras usando palabras.

Podemos determinar sobre la base de los resultados anteriores que existe un mejor desempeño utilizando reconocimiento por palabras, esto se debe a la división silábica que se le realizó a las muestras, ya que en el caso del blanco y el cinco existe un silencio dentro de la primera sílaba que provoca un corte no deseado. Además el tiempo de entrenamiento con sílabas es mayor que para el caso de palabras.

Aunque el rendimiento del sistema no es óptimo pudimos encontrar un punto en donde el índice de reconocimiento fue satisfactorio en un tiempo de cómputo razonable, se realizó una prueba para determinar el número de optimizaciones en donde se alcanza la estabilidad del sistema, la cual consiste en optimizar el sistema hasta que la probabilidad de una observación dado un modelo (Problema 1 de Markov, algoritmo hacia adelante) no cambie, en dicha prueba se observó que el número de optimizaciones oscila entre 700 y 800 para alcanzar su estabilidad, pero el inconveniente es el tiempo de cómputo ya que se encuentra entre 6 y 7 horas.

Los resultados obtenidos con los parámetros elegidos (128 regiones, 6 optimizaciones) y para las 200 muestras utilizadas en el entrenamiento, se pueden observar en la tabla 6.1. Los valores de experimentación fueron utilizados en (Oropeza 2000).

<b>Palabra a reconocer</b>	<b>Número de veces que reconoció la sílaba1</b>	<b>Número de veces que reconoció la sílaba2</b>	<b>Número de veces que reconoció la palabra</b>
<b>cero</b>	<b>10</b>	<b>8</b>	<b>8</b>
<b>uno</b>	<b>10</b>	<b>10</b>	<b>10</b>
<b>dos</b>	<b>10</b>	<b>-</b>	<b>10</b>
<b>tres</b>	<b>10</b>	<b>-</b>	<b>10</b>
<b>cuatro</b>	<b>10</b>	<b>10</b>	<b>10</b>
<b>cinco</b>	<b>9</b>	<b>10</b>	<b>9</b>
<b>seis</b>	<b>10</b>	<b>-</b>	<b>10</b>
<b>siete</b>	<b>8</b>	<b>8</b>	<b>8</b>
<b>ocho</b>	<b>10</b>	<b>10</b>	<b>10</b>
<b>nueve</b>	<b>10</b>	<b>10</b>	<b>10</b>
<b>azul</b>	<b>9</b>	<b>8</b>	<b>8</b>
<b>blanco</b>	<b>10</b>	<b>8</b>	<b>8</b>
<b>café</b>	<b>10</b>	<b>10</b>	<b>10</b>
<b>gris</b>	<b>10</b>	<b>-</b>	<b>10</b>
<b>negro</b>	<b>10</b>	<b>10</b>	<b>10</b>
<b>rojo</b>	<b>10</b>	<b>10</b>	<b>10</b>
<b>rosa</b>	<b>10</b>	<b>10</b>	<b>10</b>
<b>verde</b>	<b>10</b>	<b>9</b>	<b>9</b>
<b>abrir</b>	<b>10</b>	<b>10</b>	<b>10</b>
<b>cerrar</b>	<b>10</b>	<b>10</b>	<b>10</b>

*Tabla 6.1. Reconocimiento de palabras de diccionario experimental.*

Se realizaron pruebas para determinar el índice de reconocimiento con muestras con las que no fue entrenado el sistema, se realizó con 4 locutores y los resultados se muestran en la tabla 6.2.

<b>Palabra a reconocer</b>	<b>Número de veces que reconoció la sílaba1</b>	<b>Número de veces que reconoció la sílaba2</b>	<b>Número de veces que reconoció la palabra</b>
<b>cero</b>	<b>10</b>	<b>8</b>	<b>8</b>
<b>uno</b>	<b>10</b>	<b>10</b>	<b>10</b>
<b>dos</b>	<b>10</b>	<b>-</b>	<b>10</b>
<b>tres</b>	<b>2</b>	<b>-</b>	<b>2</b>
<b>cuatro</b>	<b>10</b>	<b>10</b>	<b>10</b>
<b>cinco</b>	<b>7</b>	<b>10</b>	<b>7</b>
<b>seis</b>	<b>10</b>	<b>-</b>	<b>10</b>

<b>siete</b>	<b>7</b>	<b>8</b>	<b>7</b>
<b>ocho</b>	<b>10</b>	<b>10</b>	<b>10</b>
<b>nueve</b>	<b>10</b>	<b>10</b>	<b>10</b>
<b>azul</b>	<b>8</b>	<b>7</b>	<b>7</b>
<b>blanco</b>	<b>10</b>	<b>8</b>	<b>8</b>
<b>café</b>	<b>10</b>	<b>10</b>	<b>10</b>
<b>gris</b>	<b>10</b>	<b>-</b>	<b>10</b>
<b>negro</b>	<b>10</b>	<b>8</b>	<b>8</b>
<b>rojo</b>	<b>10</b>	<b>10</b>	<b>10</b>
<b>rosa</b>	<b>10</b>	<b>10</b>	<b>10</b>
<b>verde</b>	<b>10</b>	<b>7</b>	<b>7</b>
<b>abrir</b>	<b>10</b>	<b>2</b>	<b>8</b>
<b>cerrar</b>	<b>10</b>	<b>10</b>	<b>10</b>

Tabla 6.2. Reconocimiento de palabras con las que no fue entrenado el sistema.

La tasa de reconocimiento en nuestro sistema es del 92% para la sílaba 1, del 69% para la sílaba 2 y del 86% por palabra. Esta tasa fue calculada por medio de la ecuación 6.1, al agregar la aplicación del Sistema Experto, de ERO y de la STTEF resultó en un incremento en el reconocimiento del 98% para las sílabas y del 95% en palabras:

$$Tasa\ de\ reconocimiento = \frac{Número\ de\ aciertos}{Número\ total\ de\ muestras} \quad (6.1)$$

Los resultados obtenidos anteriormente al ser comparados con los expuestos en (Hu et al.,1996) denotan que la reducción de la razón de error no fue significativa. Sin embargo, para fines prácticos resulta útil utilizar este reconocedor expuesto anteriormente, siempre y cuando los parámetros se encuentren bien controlados, de hecho una mejora al mismo se presenta en los siguientes apartados.

### **6.3 EL RECONOCIMIENTO DE VOZ CONTINUO USANDO MIXTURAS GAUSSIANAS Y CADENAS OCULTAS DE MARKOV**

Las Mixturas Gaussianas son utilizadas para modelar las funciones de densidad de probabilidad en los modelos ocultos de Markov.

Para entrenar estos modelos el algoritmo de la máxima esperanza (EM) es utilizado. En este caso, no sólo los parámetros de cada mixtura Gaussiana de cada estado en la HMM tienen que ser estimados, sino también el resto de los parámetros de la HMM, la matriz de transición y las probabilidades a priori tienen que ser estimadas en cada paso del algoritmo EM.

La forma de llevar a cabo esas estimaciones fue tratada en el capítulo 4. En esta parte nos interesa observar la respuesta que se tuvo al aplicar tal algoritmo a los modelos de las diferentes sílabas tratadas en este punto. Las siguientes gráficas muestran el comportamiento que presentan las Mixturas Gaussianas y los modelos de Markov después de las iteraciones que se realizan hasta alcanzar la estabilidad del modelo:

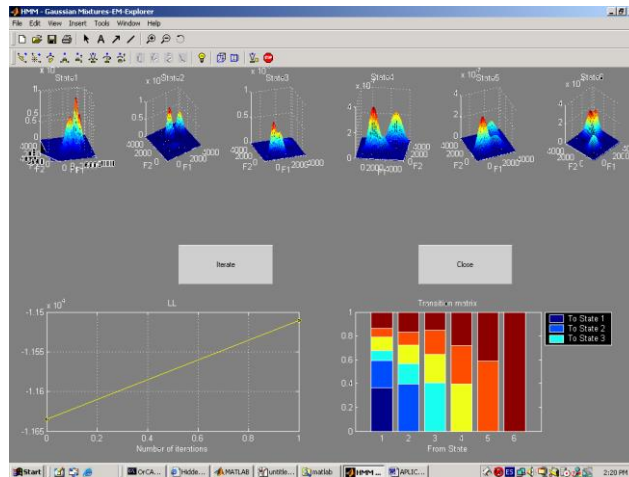


Figura 6.21. Representación esquemática de Mixturas Gaussianas por cada estado de la Cadena Oculta de Markov.

En la segunda iteración:

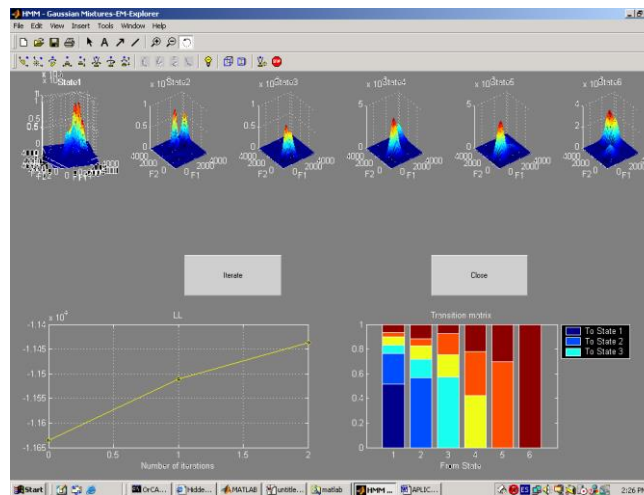


Figura 6.22. Esquemización de las Mixturas Gaussianas por Modelo Oculto de Markov después de la segunda iteración del algoritmo propuesto.

Para los fines de estos experimentos se utilizó a la energía en corto tiempo STTEF como el parámetro que permitiera la segmentación.

A continuación se analizan las características de las señales de voz de un corpus de dígitos del 0-9, para demostrar el beneficio de utilizar las sílabas como elemento

de reconocimiento. Posteriormente, se utilizó un vocabulario que incrementa el número de elementos utilizados para este fin.

La siguiente figura muestra los resultados obtenidos después de aplicar el cálculo de la energía a una muestra de voz.

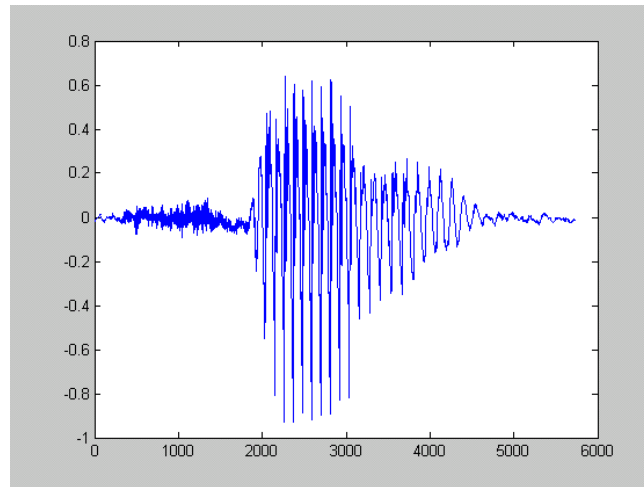


Figura 6.23. Esquematización en el dominio del tiempo de la palabra 'cero'.

Una vez obtenidas las muestras de voz, se calcula el valor de la energía por segmento, dándonos como resultado lo siguiente:

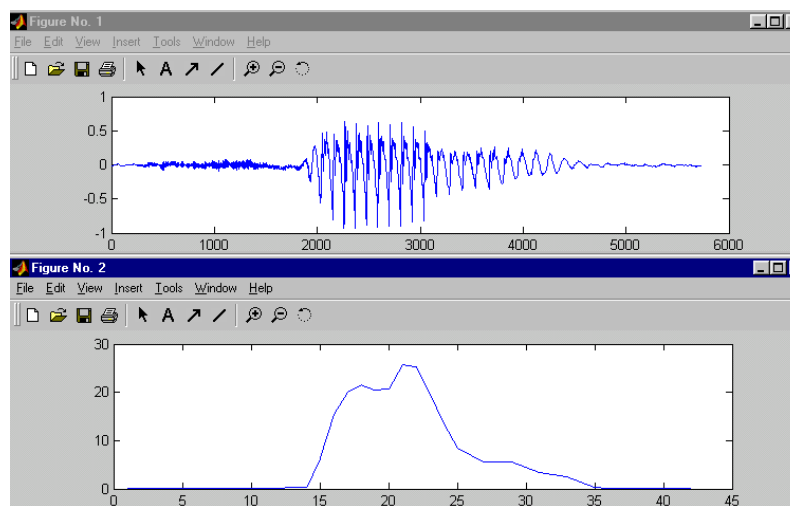


Figura 6.24. Esquematización en el dominio del tiempo de la palabra 'cero' y su energía correspondiente.

Como se puede observar, la energía extraída de la señal de voz se acumula de gran forma en donde se presentan las vocales, la palabra aquí representada es 'cero', observe la gran acumulación de energía al inicio de la vocal 'e' y la vocal 'o', aunque difícil de realizar una división bajo estos lineamientos, se consideran los siguientes criterios:

- Por visualización de la forma de onda en el dominio del tiempo se percibe el inicio del fonema 'r', al realizar pruebas de audición, se ratificó tal parámetro de segmentación.
- La representación de la señal de voz en el dominio de la frecuencia, también permitió realizar una segmentación de las sílabas de dicha palabra, la siguiente figura demuestra estos parámetros:

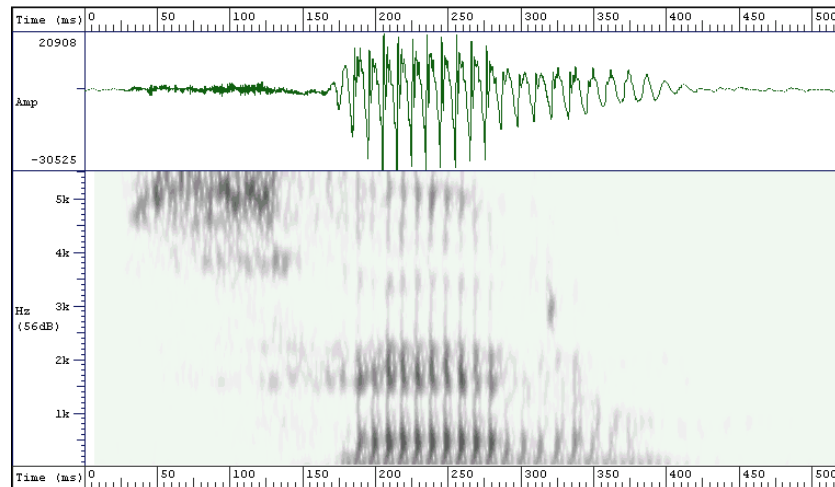


Figura 6.25. Esquematización en el dominio del tiempo y de la frecuencia de la palabra 'cero'.

La gran concentración de energía a frecuencias altas permite denotar la presencia del alófono [θ], que acompaña a la sílaba 'ce', la presencia de los formantes de la vocal 'e' se distinguen de forma automática, lo cual como se observa, termina en el tiempo en el que hace acto de presencia la sílaba 'ro', la cual posee la característica de tener presencia a muy bajas frecuencias y energía media.

Lo anterior permite obtener una forma de segmentación de la señal en sus partes elementales. La forma de hacerlo varía dependiendo del sistema a analizar. Para este caso y con la ayuda del Sistema Experto se introduce la muestra, y el sistema, se encarga en primer lugar de acuerdo a su diccionario establecer el número de sílabas que contiene la palabra, en este ejemplo 2; una vez obtenido lo anterior, la estimación de la energía busca la forma de segmentar esta señal de acuerdo a los criterios anteriormente plasmados.

Una vez realizado lo anterior se procedió a encontrar las características del vocabulario en cuestión para clasificarlo de acuerdo a la teoría de la lingüística, del análisis del corpus en cuestión se extrae la siguiente información:

Sílaba	Estructura
<b>ce</b>	<b>CV</b>
<b>ro</b>	<b>CV</b>
<b>u</b>	<b>V</b>



<b>no</b>	<b>CV</b>
<b>dos</b>	<b>CVC</b>
<b>tres</b>	<b>CCVC</b>
<b>cua</b>	<b>CVV</b>
<b>tro</b>	<b>CCV</b>
<b>cin</b>	<b>CVC</b>
<b>co</b>	<b>CV</b>
<b>seis</b>	<b>CVVC</b>
<b>sie</b>	<b>CVV</b>
<b>te</b>	<b>CV</b>
<b>o</b>	<b>V</b>
<b>cho</b>	<b>CCV</b>
<b>nue</b>	<b>CVV</b>
<b>ve</b>	<b>CV</b>

Tabla 6.3. Elementos silábicas de un corpus de dígitos.

Usando la propiedad del bigram aplicado al caso del reconocimiento de voz basado en sílabas y como una extensión del mismo se tiene:

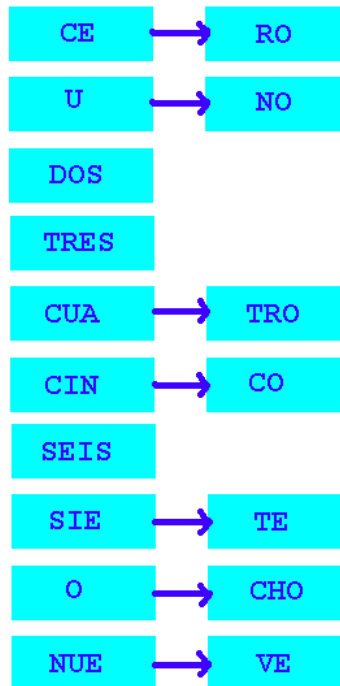


Figura 6.26. Esquematización del análisis lingüístico del corpus de dígitos.

La figura anterior nos permite observar que de acuerdo a la expresión:

$$\underline{W} = \arg \max_W P(O | W)P(W)$$

La probabilidad  $P(W)$  de la expresión anterior se extrae del análisis de la figura 6.27, la cual nos da como resultado una probabilidad de '1/10' para cada elemento de la palabra de entrada al sistema de reconocimiento (ya sea palabra o sílaba). Lo anterior es razonable pues sólo existe una representación silábica de las 17 posibles en el corpus.

Del razonamiento anterior queda como parte final la obtención de resultados a la  $P(O/W)$ , que no es más que la probabilidad obtenida del modelo oculto de Markov. Con los datos anteriores se procedió a crear el sistema de reconocimiento dando los siguientes resultados:

	ce	ro	u	no	dos	tres	cua	tro	cin	co	seis	sie	te	o	cho	nue	ve
ce	10																
ro		10															
u			6											4			
no				8			2										
dos					10												
tres						8					2						
cua					3		7										
tro								10									
cin									10								
co										9							
seis											10						
sie					1							9					
te				3									7				
o														10			
cho															10		
nue					2											8	
ve					2												8

Tabla 6.4. Tabla de confusión para el caso del corpus de dígitos.

La tabla anterior nos da como resultado un porcentaje de reconocimiento de sílaba individual del 93.42%. El cual al compararlo con los experimentos de (Hu et al., 1996) para el caso del inglés y un corpus de tamaño similar, resulta en un incremento del 8.62%, pues se reporta una razón de error de 15.2% y para este ejemplo es del 6.58%.

La siguiente tabla 6.5 y la figura 6.27 muestran el resultado del sistema de reconocimiento después de realizar la concatenación de las sílabas para el corpus de dígitos. Obteniéndose un 87% de eficiencia.

	cero	uno	dos	tres	cuatro	cinco	seis	siete	ocho	nueve
cero	10									
uno		9			1					
dos			10							
tres				7			2	1		
cuatro	1				7					2
cinco						10				
seis							8	2		
siete			1				1	8		
ocho									10	
nueve	1			1						8

Tabla 6.5. Tabla de confusión para el caso del corpus de dígitos utilizando sílabas concatenadas y SITTEF.

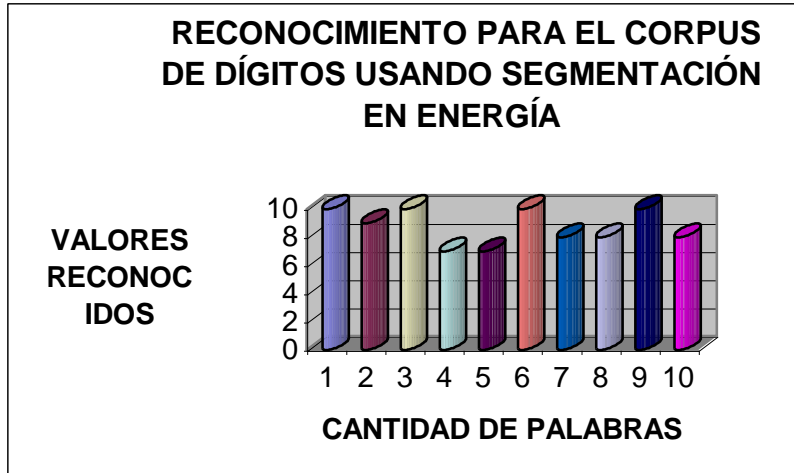


Figura 6.27. Gráfica de reconocimiento del corpus de dígitos usando segmentación silábica y modelos ocultos de Markov.

### 6.3.1 DEFINICIÓN DEL PARÁMETRO ERO

Para poder incrementar dicha tarea de reconocimiento se recurrió a utilizar una variante, el parámetro RO (parámetro que permite obtener la respuesta en frecuencia de una señal de voz por encima de los 3,500 Hz), que se obtiene tras la aplicación de un filtro digital a la señal. Cabe hacer la aclaración de que el parámetro RO ha sido utilizado en el programa para la extracción y análisis de parámetros de la voz EXPARAM 2.2 con número de registro 03-2004-052510360400-01 del registro público de derechos de autor a nombre del Dr. Sergio Suárez Guerra. En nuestro caso utilizamos el mismo algoritmo de la energía, sólo que aplicado a la señal de salida resultante del filtro digital, a lo que se ha denotado como la energía en corto tiempo del parámetro RO, ERO, que tiene la representación matemática, mostrada en 6.2 es una modificación a RO y es contribución del presente trabajo:

$$ERO = \sum_{i=0}^{N-1} ROi^2 \tag{6.2}$$

Como es sabido, la respuesta en frecuencia de un filtro digital es periódica. Del análisis de las series de Fourier cualquier función periódica puede expresarse como una combinación lineal de exponenciales complejas. Por lo tanto, la respuesta deseada de un filtro digital FIR puede ser expresada por las series de Fourier.

El truncamiento de las series de Fourier provoca los filtros digitales de respuesta al impulso finito (filtros FIR Finite Impulse Response por sus siglas en inglés) con oscilaciones indeseables en la banda de paso y en la banda de rechazo. Para reducir estas oscilaciones, una clase particular de funciones son usadas para modificar los coeficientes de Fourier (respuestas al impulso) éstas son llamadas

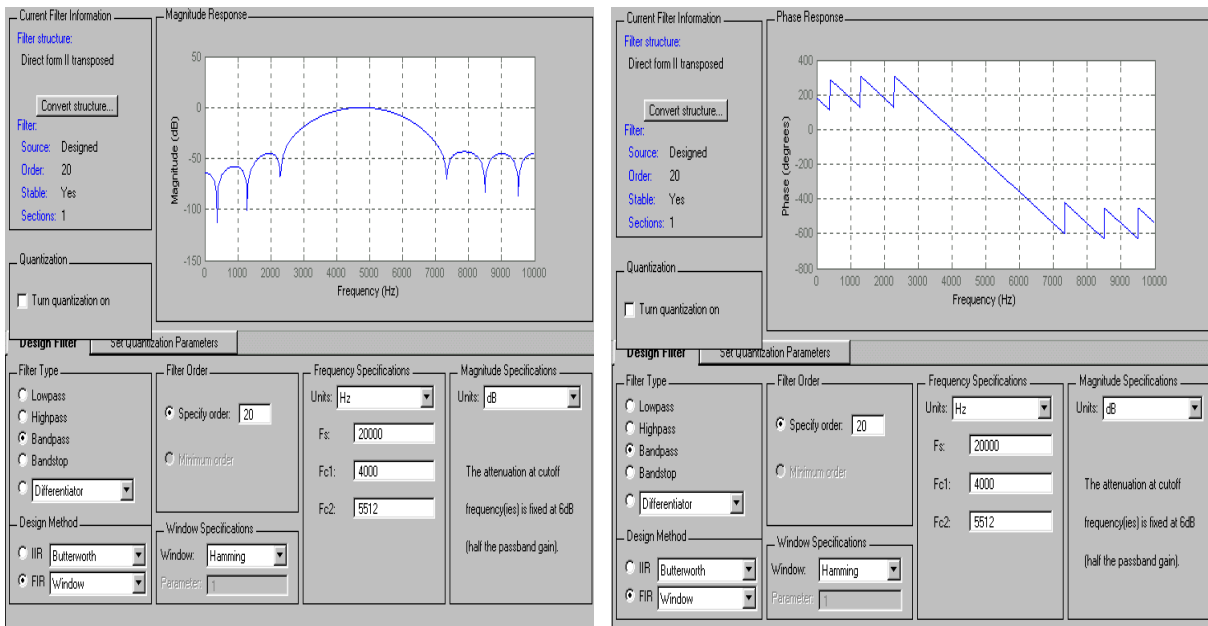
ventanas. El truncamiento de las series infinitas de Fourier es equivalente a la multiplicación de los coeficientes con la función ventana.

Las ventanas más comunes son: rectangular, Hamming, Hanning, Kaiser, Blackman, etc. Para fines de este trabajo se probó con un filtro digital pasa banda de 20 puntos y con una ventana de Hamming. Debido a que contiene una atenuación promedio entre la mayor parte de los filtros, además de ser un método de diseño ampliamente usado para este tipo de filtros y aplicaciones.

Asimismo y dado que la ventana de Hamming ha sido programada para otras secciones del presente trabajo se utilizó para este caso. En este caso se prefirió utilizar este tipo de filtros para analizar y comparar la respuesta que presentan con relación a los utilizados en (Hartmut et al., 1998). Otra de las razones del uso de esta ventana es la uniformidad que presentan sus lóbulos laterales en su representación del dominio de la frecuencia.

El número de coeficientes se extrajo de pruebas experimentales quedando en este valor para fines prácticos.

La siguiente figura 6.27 muestra las características del filtro aplicado a cada una de las señales de voz. La parte inferior muestra los valores de los coeficientes de este filtro.



<b>0.0019</b>	<b>0.2802</b>	<b>0.0019</b>
<b>-0.0037</b>	<b>-0.2412</b>	
<b>0.0039</b>	<b>0.1465</b>	
<b>0.0031</b>	<b>-0.0463</b>	
<b>-0.0194</b>	<b>-0.0168</b>	
<b>0.0324</b>	<b>0.0324</b>	
<b>-0.0168</b>	<b>-0.0194</b>	
<b>-0.0463</b>	<b>0.0031</b>	
<b>0.1465</b>	<b>0.0039</b>	
<b>-0.2412</b>	<b>-0.0037</b>	

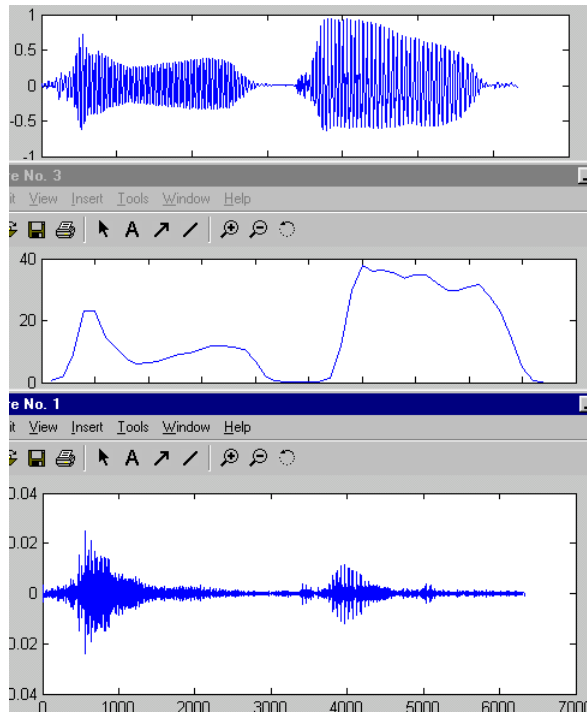


Figura 6.28. Esquemización de un filtro digital y la señal en el dominio del tiempo ya filtrada.

### 6.3.2 EFECTO DEL PARÁMETRO ERO EN UNA SEÑAL DE VOZ

La siguiente figura muestra la forma en la cual, una vez aplicado al filtro a la señal original, se comporta la gráfica de energía de dicha señal.

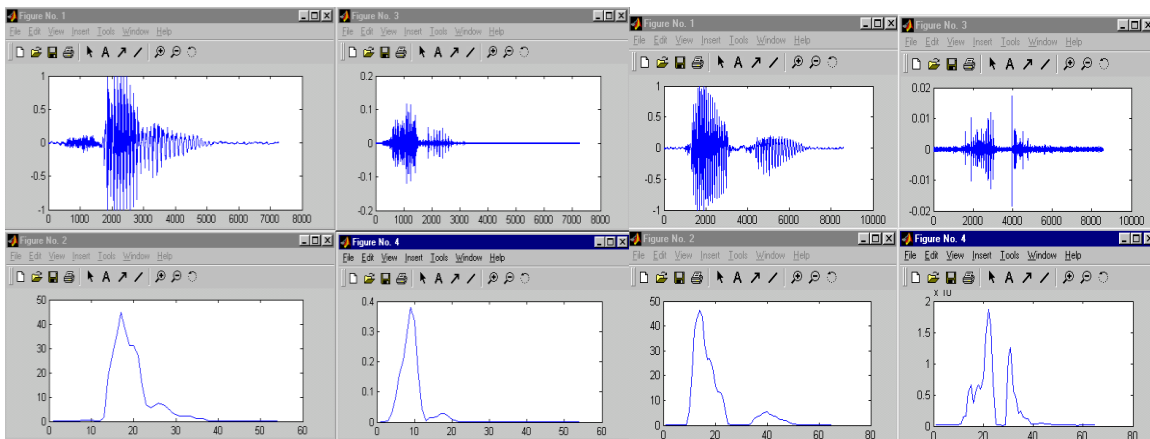


Figura 6.29 (a). Esquemización de las señales 'cero' y 'tres' antes y después de haberles aplicado el filtro digital.

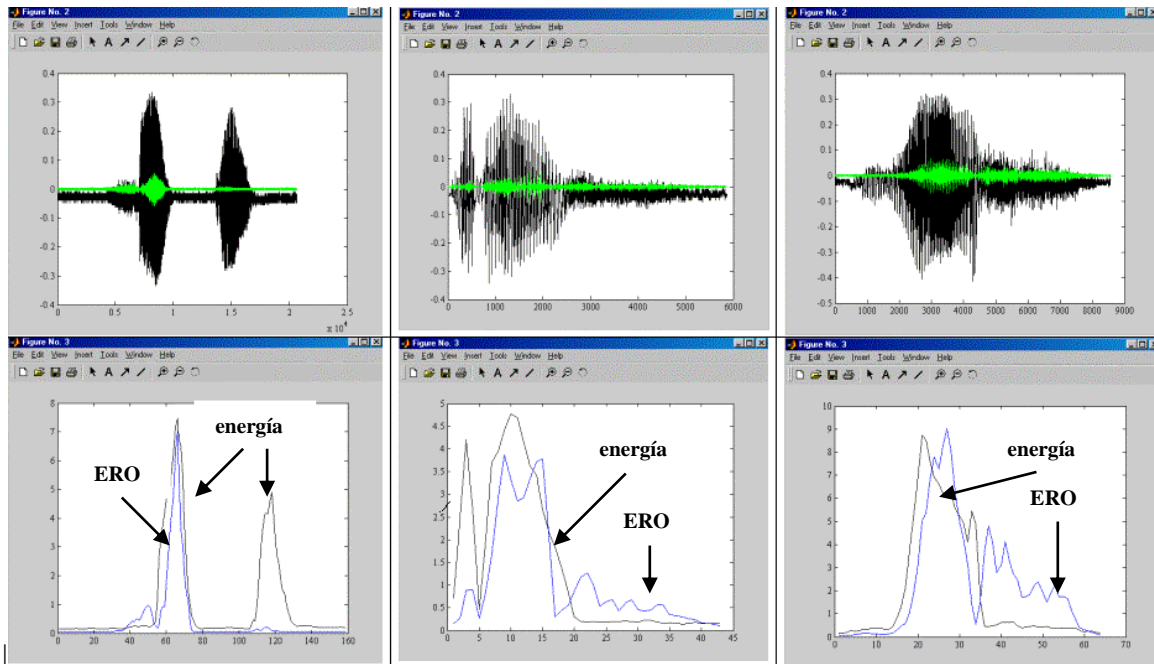


Figura 6.29 (b). Esquematación de las señales 'cinco', 'tres' y 'diez' antes y después de haberles aplicado el filtro digital, observe las regiones donde la energía con y sin parámetro ERO preponderan.

Observe de las figuras anteriores el comportamiento de la energía para los casos de cuando se ha aplicado el filtro digital y cuando no. La aplicación del parámetro ERO permite resaltar la presencia de las altas frecuencias, las imágenes demuestran el incremento de la energía de la señal filtrada a la señal no filtrada, en aquellos puntos en donde las componentes de alta frecuencia hacen sentir sus efectos.

Dado lo anterior, se procedió a crear una nueva forma de realizar la segmentación automática usando el parámetro ERO como artifice para ello. El beneficio que contrae este nuevo recurso radica básicamente en los siguientes puntos:

- ◆ Las palabras que comprendan una componente de alta energía se verán beneficiadas, pues el filtro está diseñado para dejarlas pasar, esto se puede observar en las señales de la figura 6.28b.
- ◆ Con ayuda del Sistema Experto se identifica el número de sílabas que conforman a las palabras del corpus, posteriormente se obtienen los parámetros de energía para ambos casos (aplicación y no del filtro), con lo cual se identifican tales elementos.
- ◆ Dado que el Sistema Experto analiza las componentes de los elementos del corpus, se puede deducir el número de sílabas y como están conformadas. Tales tareas aún son realizadas de forma manual y automática para fines de comparación.
- ◆ El uso de estos dos parámetros conlleva a encontrar regiones de duración de las señales de voz con y sin presencia de tales elementos.

◆ Con estos parámetros incrustados, se puede verificar que las señales de voz poseen regiones de transición energía-parámetro ERO, las cuales se comentarán posteriormente.

### 6.3.3 LA REGIÓN DE TRANSICIÓN ENERGÍA-PARÁMETRO ERO

Con los puntos anteriores se obtiene una segmentación que toma la siguiente representación numérica y esquemática:

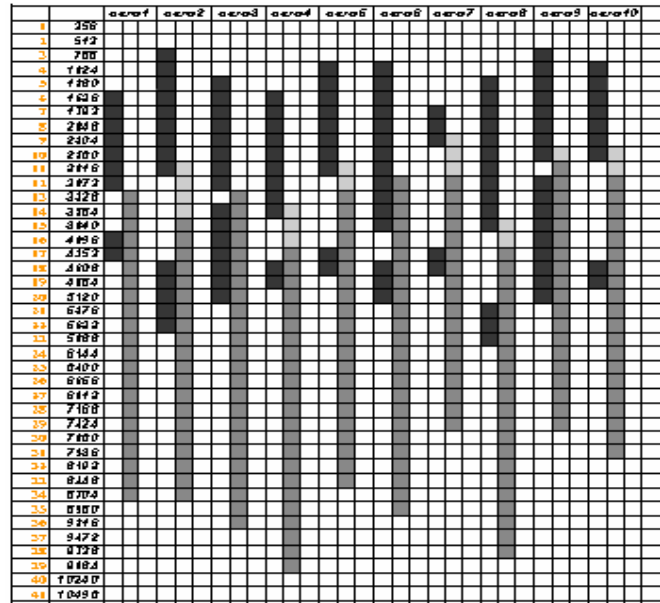


Figura 6.30. Esquemática de las regiones de transición energía-parámetro RO para el caso de la palabra 'ce'-'ro'.

Las líneas de color negro representan las regiones de la señal de voz en donde la energía de la señal sin filtrar deja sentir sus efectos, las líneas de color gris representan las regiones en donde la señal de voz ya filtrada deja sentir sus efectos, es decir; los puntos en donde las componentes de alta frecuencia se encuentran presentes y finalmente las líneas de color gris claro, representan las regiones de transición energía-parámetro RO, que permiten visualizar las regiones en donde ninguno de los dos elementos tiene su aparición, pero que es necesario para poder ligar la aparición o continuación de una sílaba.

	ce	ro	u	no	dos	tres	cua	tro	cin	co	seis	sie	te	o	cho	nue	ve
ce	10																
ro		10															
u			9											1			
no				9				1									
dos					10												
tres						10											
cua				1			9										
tro					2			8									
cin									10								
co										10							
seis											10						
sie												10					
te						1							9				
o				2										8			
cho															10		
nue																9	
ve																	9

	cero	uno	dos	tres	cuatro	cinco	seis	siete	ocho	nueve
cero	10									
uno		10								
dos			10							
tres				8			2			
cuatro					10					
cinco						10				
seis			1	1			8			
siete							1	9		
ocho									10	
nueve										10

Tabla 6.6. Tablas de confusión para el caso del corpus de dígitos usando regiones de transición energía-parámetro RO, sílabas independientes y sílabas concatenadas.

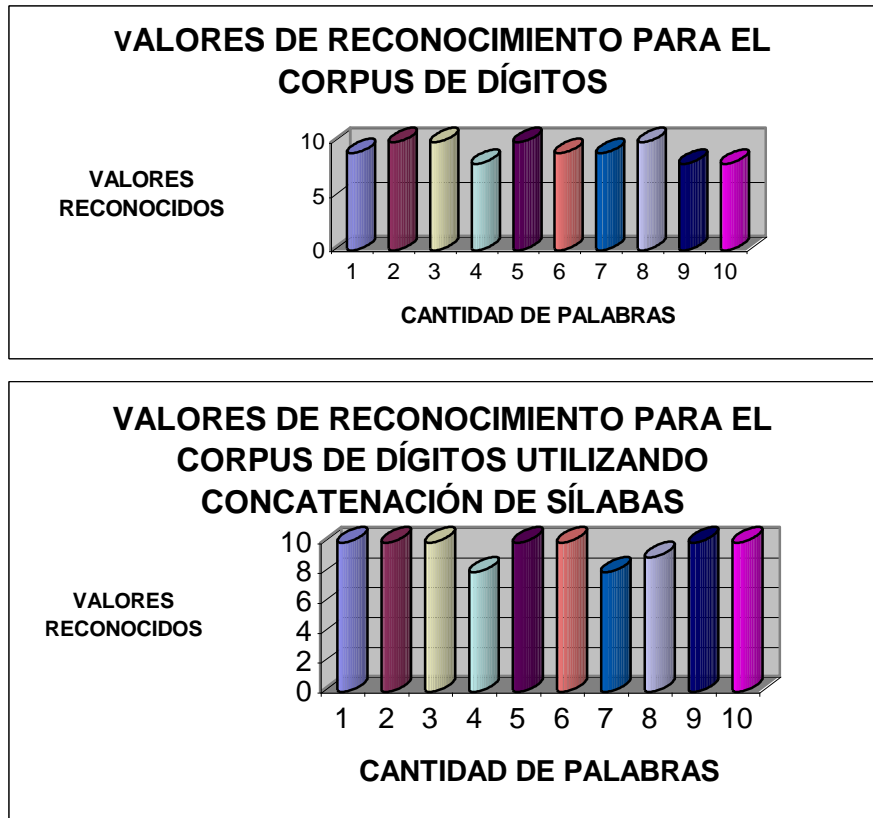


Fig. 6.31. Gráficas comparativas del reconocimiento haciendo uso de palabras completas y concatenación de sílabas del corpus de dígitos.

A esta representación le he dado por llamar *gráfica de representación energía-parámetro RO*. Dados los puntos anteriores, se hizo posible una segmentación tomando en consideración los puntos anteriores, lo cual dio como resultado:

La tabla 6.6 da como resultado un porcentaje de reconocimiento de sílaba individual del 94.70%, para el corpus de voz del habla continua. Mientras que la figura 6.31 muestra la comparación entre el reconocimiento hecho con palabras completas y con la concatenación de las sílabas, que resulta más eficiente (91% y 96% respectivamente). Cabe hacer la aclaración de que todos los diccionarios creados en los experimentos del presente trabajo son extraídos tras realizar la



tarea de segmentación a las palabras o frases que conforman las tareas de reconocimiento.

El parámetro RO es utilizado dentro del área de análisis de señales de voz en nuestro laboratorio, para fines de estudio de su aplicación en sistemas de reconocimiento de voz se muestran los resultados analizados en el presente trabajo.

Con el fin de extender la aplicación anterior a un corpus de sílabas más extenso, se procedió a analizar el siguiente corpus obteniéndose los siguientes resultados:

### 6.3.4 ANÁLISIS DE UN CORPUS DE VOZ

Se usaron las siguientes frases para el sistema de reconocimiento:

- 1 De Puebla a México
- 2 Cuauhtémoc y Cuautla
- 3 Cuautla Morelos
- 4 Espacio aéreo
- 5 Ahumado
- 6 Croacia está en Europa
- 7 Protozoarios biológicos
- 8 El trueque marítimo
- 9 Ella es seria
- 10 Sería posible desistir

Usando una herramienta de software (Sistema Experto) programada en C++, se obtuvieron los mismos resultados sílaba por sílaba del corpus:

Sílaba	#items	Sílaba	#items	Sílaba	#items
De	2	es	3	zo	1
Pue	1	pa	2	rios	1
Bla	1	cio	1	bio	1
A	5	e	2	lo	1
Me	1	o	1	gi	1
Xi	1	ahu	1	cos	1
Co	1	ma	2	el	1
Cuauh	1	do	1	true	1
Te	1	cro	1	que	1
Moc	1	cia	1	ri	2
Y	1	ta	1	ti	1
Cuau	2	en	1	lla	1
Tla	2	eu	1	se	2
Mo	2	ro	1	ria	1
Re	2	pro	1	po	1
Los	1	to	1	si	1
Ble	1	sis	1	tir	1

Tabla 6.7. División de un corpus experimental en sílabas.

Uno de los aspectos importantes a considerar cuando se trabaja con un SARH es el uso del modelo del lenguaje, para ello el presente trabajo hace uso del *bigram*, la siguiente tabla 6.7 muestra la correspondencia entre las palabras del corpus que se analiza para este caso. Observe que la primera columna es una enumeración progresiva de las frases analizadas, la columna de descripción de la frase muestra las frases utilizadas para este experimento, mientras que la última columna nos presenta el número de palabras que componen a cada una de las frases.

El número total de palabras que conforman al corpus final "1" se indican en la parte inferior de la columna del conteo de número de palabras, que para este caso es de 27 palabras. Observe además que a cada una de las frases les antecede y procede las siglas *sil*, que hacen referencia a que existe una emisión de silencio.

# frase	DESCRIPCIÓN DE LA FRASE						# palabras
1	sil	DE	PUEBLA	A	MÉXICO	sil	4
		A	B	C	D		
2	sil	CUAUHTÉMOC		Y	CUAUTLA	sil	3
			E	F	G		
3	sil	CUAUTLA	MORELOS		sil		2
			G	H			
4	sil	ESPACIO		AÉREO	sil		2
			I	J			
5	sil	AHUMADO		sil			1
			K				
6	sil	CROACIA	ESTÁ		EN	EUROPA	sil
			L	M	N	O	
7	sil	PROTOZOARIOS			BIOLÓGICOS		sil
			P	Q			2
8	sil	EL	TRUEQUE	MARÍTIMO		sil	
		R	S	T			3
9	sil	ELLA	ES	SERIA		sil	
		U	V	W			3
10	sil	SERÍA	POSIBLE		DESISTIR	sil	
		W	X	Y			3
							<b>27</b>

Tabla 6.8. Análisis del número de palabras que conforman al corpus de prueba.

Uno de los aspectos importantes y cruciales en la etapa de entrenamiento de un SARH es la concatenación de los modelos de las unidades lingüísticas seleccionadas. Para ello, se realiza un análisis desde el punto de vista de modelo del lenguaje, lo que permite conocer las probabilidades de transición que serán adjudicadas entre los distintos Modelos Ocultos de Markov determinados con las muestras de la señal de voz, y que deben de relacionarse con las frases que conforman al corpus. El acoplamiento es esencial y para este caso se ha elegido el

método de *Smoothing en la versión de Good-Turing aunado al modelo interpolado no lineal* para determinar la perplejidad del corpus (Becchetti and Prina, 1999).

La siguiente tabla 6.8 muestra las posibles transiciones que se presentan entre las palabras que conforman al corpus propuesto. La primera columna de la parte del conteo final que prosigue a la columna de la palabra *Y*, muestra el número de palabras sin tomar en cuenta el silencio correspondiente, mientras que la que le prodigue si lo hace. La columna donde se encuentran los valores del '0' y '1', muestran el número de tales valores que se presentan en cada uno de los renglones de la tabla, que corresponden con cada una de las palabras. La última columna muestra el total de elementos por cada renglón considerado. En la parte inferior de tales columnas se muestran los totales del conteo realizado en este caso.

PALABRAS	CONTEO	CONTEO FINAL																															
		sil	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	0	1				
sil		0	1	0	0	0	1	0	1	0	1	0	0	1	0	0	0	1	0	1	0	0	1	0	1	0	0	9	9	17	9	26	
A	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	25	1	26	
B	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	26	0	26	
C	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	26	0	26	
D	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	25	1	26	
E	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	26		26	
F	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	25	1	26	
G	2	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	24	2	26	
H	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	25	1	26	
I	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	26		26	
J	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	26		26	
K	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	25	1	26	
L	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	26		26	
M	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	26		26	
N	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	26		26	
O	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	25	1	26	
P	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	26		26	
Q	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	25	1	26	
R	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	26		26	
S	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	26		26	
T	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	25	1	26	
U	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	25	1	26	
V	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	26		26	
W	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	2	24	2	26	
X	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	26		26
Y	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	25	1	26	
																												26	36	653	23	676	

Tabla 6.9. Matriz de transición de palabras en el corpus analizado.

### 6.3.5 ANÁLISIS DE PERPLEJIDAD

Se muestra ahora el cálculo de la perplejidad para este corpus, cabe hacer la aclaración que debido a lo extenso de los cálculos que se requieren, sólo se muestra el análisis final. Los cálculos y las tablas correspondientes se anexan al reporte en el disco correspondiente. Se considera como ejemplo esencial tomar una de las frases para encontrar el valor que se busca. La frase que se está poniendo de prueba es la primera de la tabla 6.7.

Los datos considerados son obtenidos del ejemplo anterior con un conjunto de prueba de la forma:

{ A B C D }

		lp(A)	-3.29583687
(sil,A)	P(sil/A)	0.03340279	-3.39911582
(A,B)	P(B/A)	0.16493023	0.071428571
(B,C)	P(C/B)	0.17062974	-1.76825936
(C,D)	P(D/C)	0.17673726	-1.73309104
		0.54570002	2.0249749

Tabla 6.10. Valores logarítmicos del corpus de prueba.

Usando estos valores, la perplejidad logarítmica es de la forma:

$$\ln PP = -\frac{1}{N} \sum_{i=1}^n \ln [p(w_i | w_{i-1})] =$$

$$= -\frac{1}{7} [lp(A) + lp(SIL | A) + lp(B | A) + lp(C | B) + lp(D | C)]$$

**PP**                      **7.5759**

El valor obtenido nos demuestra un valor adecuado para los análisis de las señales de voz de acuerdo a (Hauenstein, 1996), considerándose además que el corpus es sustentable como prueba.

Para verificar el potencial que tiene la aplicación de los métodos antes señalados se aplica el mismo análisis ahora al corpus que se muestra en la siguiente figura 6.30.

El motivo de este análisis es verificar la eficiencia que tiene la utilización del método de segmentación anteriormente planteado.

La característica del corpus es que contiene una gran cantidad de palabras que de acuerdo a su lugar en donde se manifiestan representan dificultad en cuanto al sentido que se le está dando a la frase.

A	B	C	D	E	F	G	H	I	J	K	L	M	N
4	4	0	0	0	0	0	0	0	0	0	4	14	1
5	0	4	0	0	0	0	0	0	1	0	0	13	1
9	0	0	4	0	1	1	2	0	1	0	0	10	3
5	0	0	0	1	0	0	0	0	1	2	1	0	0
4	2	0	0	1	1	0	0	0	0	0	0	4	12
7	3	0	0	0	2	0	0	0	2	0	0	7	12
5	0	0	0	0	0	5	0	0	0	0	0	0	5
2	0	0	0	1	0	0	0	1	0	0	0	0	2
1	0	0	0	0	0	0	0	1	0	0	0	0	1
7	1	0	0	2	0	1	0	0	0	1	0	2	7
2	0	0	0	1	0	0	1	0	0	0	0	2	13
1	0	0	1	0	0	0	0	0	0	0	0	1	14
2	0	0	0	0	2	0	0	0	0	0	0	2	14
2	0	0	0	0	0	2	0	0	0	0	0	2	14
56	178	17	9	1	3	1							

Figura 6.32. Segundo Corpus final de prueba.

La figura 6.31 muestra las probabilidades de transición entre cada una de las palabras que conforman el corpus, extraídas del análisis del modelo interpolado no lineal. Dichas probabilidades como se mencionó con anterioridad, se utilizan para enlazar los Modelos Ocultos de Markov independientes y con ello conformar las frases del corpus, aunque las palabras están acentuadas, el análisis se realizó sin tomar en cuenta este factor.

CÁLCULO DE PROBABILIDADES CONDICIONALES. P(w(i),w(0))													
A	B	C	D	E	F	G	H	I	J	K	L	M	N
0.024310521	0.024310521	0.024310521	0.024310521	0.024310521	0.024310521	0.024310521	0.024310521	0.024310521	0.024310521	0.024310521	0.024310521	0.024310521	0.024310521
0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417
0.010804676	0.010804676	0.010804676	0.010804676	0.010804676	0.010804676	0.010804676	0.010804676	0.010804676	0.010804676	0.010804676	0.010804676	0.010804676	0.010804676
0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417
0.024310521	0.024310521	0.263518717	0.263518717	0.024310521	0.024310521	0.024310521	0.024310521	0.024310521	0.024310521	0.024310521	0.024310521	0.024310521	0.024310521
0.013891726	0.013891726	0.013891726	0.013891726	0.013891726	0.013891726	0.013891726	0.013891726	0.013891726	0.013891726	0.013891726	0.013891726	0.013891726	0.013891726
0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417	0.019448417
0.048621042	0.048621042	0.539037433	0.048621042	0.048621042	0.048621042	0.539037433	0.048621042	0.048621042	0.048621042	0.048621042	0.048621042	0.048621042	0.048621042
0.037242084	0.037242084	0.037242084	0.037242084	0.037242084	0.037242084	0.037242084	0.037242084	0.037242084	0.037242084	0.037242084	0.037242084	0.037242084	0.037242084
0.013891726	0.013891726	0.048621042	0.013891726	0.154010635	0.013891726	0.013891726	0.013891726	0.013891726	0.013891726	0.013891726	0.013891726	0.154010635	0.013891726
0.048621042	0.048621042	0.539037433	0.048621042	0.048621042	0.048621042	0.539037433	0.048621042	0.048621042	0.048621042	0.048621042	0.048621042	0.048621042	0.048621042
0.037242084	0.037242084	0.037242084	0.037242084	0.037242084	0.037242084	0.037242084	0.037242084	0.037242084	0.037242084	0.037242084	0.037242084	0.037242084	0.037242084
0.048621042	0.048621042	0.048621042	0.048621042	0.16369697	0.048621042	0.048621042	0.048621042	0.048621042	0.048621042	0.048621042	0.048621042	0.048621042	0.048621042
0.048621042	0.048621042	0.048621042	0.048621042	0.048621042	0.048621042	0.16369697	0.048621042	0.048621042	0.048621042	0.048621042	0.048621042	0.048621042	0.048621042
<b>0.3460347</b>	<b>1.2635143</b>	<b>1.028912</b>	<b>0.9692079</b>	<b>0.830764</b>	<b>0.26367</b>	<b>1.271011</b>	<b>1.661527</b>	<b>2.342222</b>	<b>0.543988</b>	<b>1.6615274</b>	<b>2.342222</b>	<b>0.801771</b>	<b>0.801771</b>
CÁLCULO DE PROBABILIDADES CONDICIONALES. P(w(i),w(0))													
A	B	C	D	E	F	G	H	I	J	K	L	M	N
0.071428571	0.071428571	0.071428571	0.071428571	0.071428571	0.071428571	0.071428571	0.071428571	0.071428571	0.071428571	0.071428571	0.071428571	0.071428571	0.071428571
0.644682748	0.01539199	0.01539199	0.01539199	0.01539199	0.01539199	0.01539199	0.01539199	0.01539199	0.01539199	0.01539199	0.01539199	0.01539199	0.01539199
0.010501072	0.4389896	0.010501072	0.116420188	0.116420188	0.116420188	0.010501072	0.116420188	0.010501072	0.010501072	0.010501072	0.010501072	0.010501072	0.010501072
0.021390506	0.021390506	0.237145959	0.021390506	0.021390506	0.021390506	0.021390506	0.021390506	0.021390506	0.021390506	0.021390506	0.021390506	0.021390506	0.021390506
0.029262859	0.029262859	0.324422843	0.029262859	0.029262859	0.029262859	0.029262859	0.029262859	0.029262859	0.029262859	0.029262859	0.029262859	0.029262859	0.029262859
0.05268595	0.05268595	0.05268595	0.05268595	0.05268595	0.05268595	0.05268595	0.05268595	0.05268595	0.05268595	0.05268595	0.05268595	0.05268595	0.05268595
0.01530153	0.01530153	0.01530153	0.01530153	0.01530153	0.01530153	0.01530153	0.01530153	0.01530153	0.01530153	0.01530153	0.01530153	0.01530153	0.01530153
0.029262859	0.029262859	0.324422843	0.029262859	0.029262859	0.029262859	0.029262859	0.029262859	0.029262859	0.029262859	0.029262859	0.029262859	0.029262859	0.029262859
0.041517023	0.041517023	0.041517023	0.041517023	0.041517023	0.041517023	0.041517023	0.460278696	0.041517023	0.041517023	0.041517023	0.041517023	0.041517023	0.041517023
0.025540564	0.025540564	0.089141578	0.025540564	0.2831556	0.025540564	0.025540564	0.025540564	0.025540564	0.025540564	0.025540564	0.025540564	0.025540564	0.089141578
0.029262859	0.029262859	0.324422843	0.029262859	0.029262859	0.029262859	0.029262859	0.029262859	0.029262859	0.029262859	0.029262859	0.029262859	0.029262859	0.029262859
0.041517023	0.460278696	0.041517023	0.041517023	0.041517023	0.041517023	0.041517023	0.460278696	0.041517023	0.041517023	0.041517023	0.041517023	0.041517023	0.041517023
0.060642093	0.060642093	0.060642093	0.211852794	0.060642093	0.060642093	0.060642093	0.060642093	0.060642093	0.060642093	0.060642093	0.060642093	0.060642093	0.060642093
0.060642093	0.060642093	0.060642093	0.060642093	0.060642093	0.060642093	0.060642093	0.060642093	0.060642093	0.060642093	0.060642093	0.060642093	0.060642093	0.060642093

Figura 6.33. Probabilidades de transición del segundo corpus final de prueba.

Los datos considerados son obtenidos del ejemplo anterior con un conjunto de prueba de la forma: **{ A B C E F G N J }**

		lp(A)	-2.63905733
(A,B)	P(B/A)	0.07142857	-2.63905733
(B,C)	P(C/B)	0.01539199	-4.173908049
(C,E)	P(E/C)	0.11642019	-2.15054932
(E,F)	P(F/E)	0.02926286	-3.531436162
(F,G)	P(G/F)	0.05268595	-2.943406455
(G,N)	P(N/G)	0.01530153	-4.179802446
(N,J)	P(J/N)	0.06064209	-2.802766027
			<b>3.13249789</b>

Tabla 6.11. Valores logarítmicos del conjunto de prueba final.

Usando estos valores, la perplejidad logarítmica es de la forma:

$$\ln PP = -\frac{1}{N} \sum_{i=1}^n \ln [p(w_i | w_{i-1})] =$$

$$= -\frac{1}{7} [lp(A) + lp(D | A) + lp(C | D) + lp(A | C) + lp(F | A) + lp(L | F) + lp(H | L)]$$

$$PP \quad \underline{\underline{22.931118}}$$

Ahora bien, se muestran los valores de las probabilidades condicionales de las sílabas que conforman al corpus y que también son importantes, pues representan las probabilidades de transición entre los elementos lingüísticos básicos de este trabajo, las sílabas.

1	sil	a	la	me	je	r	á	la	es	ta	a	llá	sil	10	
		A	B	C	D	A	B	E	F	A	G				
2	sil	a	la	me	je	r	es	el	a	la	de	á	la	sil	11
		A	B	C	D	E	H	A	B	I	A	B			
3	sil	a	la	me	je	r	és	ta	es	el	á	la	sil	10	
		A	B	C	D	E	F	E	H	A	B				
4	sil	a	la	me	je	r	no	es	ta	sil				7	
		A	B	C	D	J	E	F							
5	sil	á	la	el	me	je	r	del	me	je	r	sil	8		
		A	B	H	C	D	K	C	D						
6	sil	el	a	la	de	á	la	es	ta	a	llá	sil	10		
		H	A	B	I	A	B	E	F	A	G				
7	sil	el	a	la	del	a	ve	es	ta	me	je	r	sil	10	
		H	A	B	K	A	la	E	F	C	D				
8	sil	el	a	la	no	la	cu	bre	á	la	sil	9			
		H	A	B	J	B	M	N	A	B					
9	sil	es	á	la	el	me	je	r	sil					6	
		E	A	B	H	C	D								
10	sil	es	me	je	r	á	la	sil					5		
		E	C	D	A	B									
															86

PALABRAS	CONTEO	SIL	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
A	19	5	0	16	0	0	0	0	2	0	0	0	0	1	0	0	0
B	17	0	0	0	4	0	2	0	0	2	2	1	1	0	1	0	4
C	9	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0
D	9	0	2	0	0	2	0	0	0	0	1	1	0	0	0	0	3
E	9	2	1	0	1	0	0	5	0	2	0	0	0	0	0	0	0
F	5	0	2	0	1	0	1	0	0	0	0	0	0	0	0	0	0
G	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	7	3	5	0	2	0	0	0	0	0	0	0	0	0	0	0	0
I	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
J	2	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
K	2	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
L	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
M	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
N	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O	10	0	5	0	0	2	0	0	3	0	0	0	0	0	0	0	0
	96																86

Figura 6.34. Probabilidades de transición de las sílabas del corpus de prueba, mismas que se utilizan para inicializar las probabilidades de transición de los Modelos Ocultos de Markov.

### 6.3.6 RESULTADOS OBTENIDOS

Se utilizaron Mixturas Gaussianas con 3 de ellas para cada estado y HMM con 5 y 3 estados, se usaron 12 coeficientes CLPC's como componentes de observación, generándose los Modelos independientes por sílaba, realizándose la concatenación de las mismas utilizando las probabilidades obtenidas a través del modelo bigram del lenguaje, para comenzar con el entrenamiento global de la frase. Estos programas corrieron sobre MATLAB y los resultados obtenidos se muestran a continuación primero para el corpus final "1" y posteriormente para el corpus final "2":

Segmentación	Modelos 3 estados	de Markov 5 estados
<b>energía</b>	89.5%	95.5%
<b>ERO</b>	95%	97.5%

Tabla 6.12. Porcentajes de reconocimiento usando Cadenas Ocultas de Markov con 3 y 5 estados respectivamente para el habla discontinua del corpus final "1".

Segmentación	Modelos 3 estados	de Markov 5 estados
<b>energía</b>	77.5%	75.5%
<b>ERO</b>	79%	80.5%

Tabla 6.13. Porcentajes de reconocimiento usando Cadenas Ocultas de Markov con 3 y 5 estados respectivamente para el habla continua del corpus final "1".

Se utilizaron un total de 20 repeticiones con un total de 5 personas (3 hombre y 2 mujeres) de las cuales 50% se usaron para el entrenamiento y 50% reconocimiento. Los resultados de reconocimiento mostrados presentan el acumulado del análisis de las 1000 frases del experimento. Para las sílabas con un orden de aparición de "uno", como es el caso de "Pue", el número de muestras en el entrenamiento es de 100, sin embargo, las sílabas como "ti", presentaron complicaciones por el número de muestras tan corto, al ser distribuidas por los estados de la Mixtura Gaussiana. Para evitar esto en el proceso de inicialización se agregaron el doble de elementos). Los resultados para el corpus final "2" se muestran a continuación:

<b>Segmentación</b>	<b>Modelos 3 estados</b>	<b>de Markov 5 estados</b>
<b>energía</b>	85.5%	86.3%
<b>ERO</b>	90%	90.8%

Tabla 6.14. Porcentajes de reconocimiento usando Cadenas Ocultas de Markov con 3 y 5 estados respectivamente para el habla discontinua del corpus final "2".

<b>Segmentación</b>	<b>Modelos 3 estados</b>	<b>de Markov 5 estados</b>
<b>energía</b>	65.5%	64.5%
<b>ERO</b>	72%	74.2%

Tabla 6.15. Porcentajes de reconocimiento usando Cadenas Ocultas de Markov con 3 y 5 estados respectivamente para el habla continua del corpus final "2".

Para finalizar, se procedió a intentar verificar el efecto que tiene considerar la acentuación de las palabras que conforman al corpus final "2" antes citado. El resultado generó un porcentaje de reconocimiento entre el 50 y 60%, lo cual implica que debe de utilizarse análisis de señal de voz que permita identificar la variación que provoca la acentuación tanto en los fonemas que conforman a las sílabas y las palabras del español. El pitch, la amplitud y filtros adaptivos pueden ser herramientas utilizadas para intentar resolver este problema.

## 6.4 RESUMEN DEL CAPÍTULO

En el presente capítulo se analizó el comportamiento de un algoritmo de reconocimiento de voz discontinuo propuesto, los resultados obtenidos dejan entrever que la forma de llevar a cabo este análisis, representa una nueva alternativa a las aplicaciones de reconocimiento de voz para aplicaciones de reconocimiento de comandos.

El análisis aplicado a tal algoritmo demuestra de forma fehaciente su gran utilidad para las aplicaciones antes mencionadas, además el porcentaje de reconocimiento



obtenido tras la aplicación de las herramientas del parámetro ERO y del Sistema Experto permitieron incrementar tal tasa de reconocimiento.

Posteriormente, se analizaron las implicaciones que tendría la inmersión de tales elementos (Sistema Experto y parámetro ERO) al caso del habla continua. Demostrándose que existe un sustancial incremento del rendimiento de los parámetros del sistema tras su integración.

Los puntos anteriores permiten demostrar la utilidad de la sílaba en las dos áreas del reconocimiento de voz: el habla discontinua y la continua. Tales experimentos demuestran que el paradigma de la sílaba aplicada al español representa una alternativa al fonema, pues los resultados obtenidos son significativos, aunque al aplicarse a corpus en donde existen ambigüedades en las frases y considerando la acentuación de las palabras los resultados no fueron los esperados.

# CAPÍTULO 7

---

## CONCLUSIONES

El funcionamiento del cerebro con relación al procesamiento de la señal de voz humana aún no es del todo entendido. Existe alguna evidencia, de que la percepción de la voz incorpora información relacionada a las propiedades temporales de la sílaba.

El presente capítulo resume los aportes obtenidos con la investigación sobre la integración de la sílaba dentro del reconocimiento de voz en el idioma español. Las características esenciales de los SRAH basados en sílabas: son la adecuada segmentación de estas unidades y su incorporación a las herramientas de reconocimiento existentes.

Con los resultados obtenidos en los capítulos anteriores se concluye que la sílaba es una herramienta adecuada para las tareas de reconocimiento por computadora. La integración de tales elementos del lenguaje son adecuados sobre todo para tareas dependientes de contexto.

La integración de un sistema basado en conocimiento que incorpore las reglas del idioma es parte fundamental del desarrollo del sistema de reconocimiento. Asimismo, el uso de los parámetros de energía y ERO son la base sobre la que descansan los resultados de este trabajo.

## 7.1 REMEMBRANZA

Con formato

Esta tesis comenzó con una discusión e implantación de las reglas de la sílaba en la identificación y segmentación de las señales de voz. Una breve revisión describió algunas de las funciones en el léxico de la sílaba en los sistemas de percepción humana. La investigación en la literatura reveló que las propiedades silábicas de la voz son ampliamente conversacionales. Un estudio de la sílaba usada en conversaciones y textos indicó que una muestra representativa de la voz casual (de tamaño considerable) es relativamente simple para ser descrita con sílabas, y que las sílabas, son una representación efectiva y sustentable como elemento de reconocimiento.

Lo anterior permitió reafirmar la idea de la integración de tales unidades a las tareas de reconocimiento de voz. La exposición continuó con una discusión integra del sistema de reconocimiento en la tarea de números, la base principal de todos los experimentos en esta tesis.

La metodología usada fue desarrollada con base a la prueba y error de los experimentos propuestos. En la mayoría de los experimentos (donde se usaron segmentaciones manuales) se demostró que instancias de la segmentación silábica de la voz de entrada originan, substancialmente, pocos beneficios en la razón de error en el SRAH para el habla discontinua; esto fue debido a que los inicios fueron determinados con un modesto grado de eficiencia.

Los inicios estimados acústicamente, basados en características de energía, fueron incorporados en el SRAH resultando en un relevante porcentaje de aumento en la tarea de reconocimiento de voz. Un método para incorporar distinción de sílabas haciendo uso del parámetro  $ERO_z$  reportó una mejora substancial que al juntarlo con la energía y el sistema basado en conocimiento lograron un beneficioso decremento de la razón de error, tanto para el habla continua como la discontinua.

Estos experimentos indicaron el beneficio de utilizar la información basada en la sílaba en diferentes corpus. Dicha investigación involucró el desarrollo de sistemas experimentales basados en sílabas para la tarea de reconocimiento.

## 7.2 DISCUSIÓN

Con formato

Con formato

La investigación en este trabajo fue conducida bajo el tenor de usar la información basada en sílaba en los SRAH. El resultado final fue un incremento en el reconocimiento conforme el conjunto de experimentos se fueron realizando, consiguiendo alcanzar un sistema altamente útil en la parte final de los experimentos.

Los resultados extraídos de investigaciones han demostrado que incluso los fonetistas y lingüistas han sido atraídos por las características inherentes de la sílaba (Fosler et al., 1999).

El gran problema de usar sílabas como unidad fundamental de reconocimiento es la gran cantidad que existe de ellas, lo cual se demostró al analizar los diferentes textos y corpus a lo largo de este trabajo. Sin embargo se encontró que una gran cantidad de esos elementos esenciales se presentaban en varias ocasiones a lo largo de los textos y corpus analizados. La tarea era pues demostrar que el incremento de tales unidades era un factor que repercutía en un grado menor de porcentaje con relación al reconocimiento.

Desde los resultados reportados en (Wu, 1998), (Hu et al., 1996), (Boulard, 1996), (Hauenstein, 1996), (Hamaker, 1998) y (Wu et al., 1997) se ha considerado a la unidad silábica como una alternativa al fonema en su integración a los reconocedores de voz, los análisis desarrollados para el idioma inglés han originado el incremento del interés de su utilización. Asimismo, el uso de diferentes técnicas empleadas en el reconocimiento de la señal de voz han permitido incrementar la tasa de reconocimiento.

En la actualidad el interés por el estudio de las unidades silábicas abarca proyectos de gran extensión, tal como el expuesto en Eidemburgo. En lo que respecta al español las tareas de inserción de la sílaba en los sistemas de reconocimiento del habla no han sido estudiados en su totalidad. Por el momento la mayor cantidad de trabajos se han enfocado en los fonemas o en su defecto cuando se utiliza la sílaba para idiomas distintos al español.

Como parte esencial de los resultados obtenidos y en comparación con los trabajos realizados anteriormente se considera que la integración a los sistemas de reconocimiento del habla de una unidad especializada de conocimiento lingüístico (sistema basado en conocimiento) en la etapa de entrenamiento, refuerza la idea de conocimiento a priori que se ve inmersa en la filosofía de la inteligencia artificial. En este caso, no sólo nos interesó obtener un conocimiento de la forma en la cual se puede analizar la señal de voz, sino también, un conocimiento de las estructuras esenciales bajo las cuales en su forma básica se entrena al cerebro humano para su posterior adecuación al medio externo.

En el caso específico de los fonemas tal y como se ha expuesto a lo largo del presente trabajo y sus antecesores, se le considera como una unidad abstracta de cualquier tipo de idioma. Esto es, su sola presentación carece de sentido para quién la escucha, sin embargo, la gran ventaja de ser elementos atómicos del lenguaje, ha llevado a que sean utilizados para fines de investigación por muchos años. Empero, las investigaciones actuales han inferido que resulta cada vez más estrecho el avance que se pueda dar en cuestión de descenso de la razón de error al seguir haciendo uso de esta unidad básica.

Por tal motivo, unidades como la sílaba, el trifenema y otras más, han sido consideradas como elementos de investigación y estudio. Como referencia muy especial, se encuentra la labor que en estos tiempos realiza Lawrence Rabiner por analizar y estudiar el comportamiento de los trifenemas en el idioma inglés en las tareas de reconocimiento de voz.

Ahora bien, los estudios expuestos en (Hartmut et al., 1996) en donde para la tarea de segmentación en unidades silábicas de una señal de voz se hace uso de filtros digitales de respuesta al impulso finito pasa banda en frecuencias comprendidas entre los 200, 360, 520 y 680 en el límite inferior y de 1650, 2210, 2770 y 3330 para los límites superiores, han demostrado ser útiles para detectar el núcleo de la sílaba, sin embargo y de acuerdo con la definición de la sílaba misma, hacia falta encontrar los puntos iniciales y finales de la misma. El parámetro ERO expuesto en el presente trabajo, demuestra que regiones de frecuencia no analizadas con anterioridad, también entregan resultados útiles para las tareas de reconocimiento de voz.

Finalmente, la integración de todos los elementos antes expuestos a la tarea de reconocimiento del habla continua nos permiten ver que la sílaba para el caso específico del español, resulta ser una unidad interesante de análisis. Por lo tanto, y bajo los resultados obtenidos a lo largo de la presente investigación, se considera que la sílaba y sus características deben ser tomadas en cuenta para tareas de investigación posteriores.

### **7.2.1 IMPLICACIONES DEL USO DE LA SÍLABA PARA LOS SRAH**

Las investigaciones generalmente han considerado que hay una unidad básica para el reconocimiento de voz. Los argumentos son por lo regular enfatizados en términos de que "el fonema es lo adecuado y la sílaba es lo erróneo" o viceversa. Estos experimentos y estudios exploratorios encontraron que el uso de las sílabas, y la información correspondiente, es una alternativa a los fonemas que generan resultados satisfactorios.

El paradigma de la sílaba usado en el reconocimiento de voz por computadora puede ser interpretado en términos de una asociación dinámica entre unidades particulares e intervalos de voz idealizados. En los experimentos con la sílaba, el intervalo de voz adjudicado a la misma fue determinado por la función de energía y por los referentes lingüísticos que se implantaron en el Sistema Basado en Conocimiento, mismos que permitieron definir la cantidad de sílabas inmersas en una frase determinada.

Uno de los puntos importantes extraídos del presente trabajo es el comportamiento que las sílabas presentan en lo relacionado al tiempo de duración de las mismas. De acuerdo a los resultados analizados, los intervalos de tiempo de duración van en dependencia directa del tipo de estructura silábica que se analiza. De los resultados obtenidos, se encontró que el promedio de la sílaba resultó ser de 0.33503913 segundos, los cuales coinciden con los análisis reportados en (Hartmut et al., 1996).

Dicho tiempo de duración es cambiante y va a depender de la estructura silábica que se analiza, y por lo regular, no hay una forma de definir esas estructuras en cuanto al tiempo de duración.

De acuerdo al análisis efectuado se considera que las estructuras silábicas de mayor aparición son las CV, las cuales en la mayoría de los textos y corpus de voz abarcan entre el 50 y 65% de los mismos. Este análisis da cabida a considerar la posibilidad de generar estructuras de reconocimiento dependientes de las estructuras silábicas, de tal forma que pueda existir la posibilidad de una división de elementos de reconocimiento usando este tipo de estructuras.

Los experimentos establecieron los intervalos de la señal de voz a priori en una trama de 250-350 ms y varias repeticiones, respectivamente. Tras el uso de la segmentación basada en energía se observó la posibilidad de la segmentación dinámica de la señal de voz en el ámbito silábico con un mejor rendimiento al realizado de forma manual.

El capítulo 3 también mostró un número de ventajas y desventajas de usar las sílabas en el reconocimiento de voz. Estos experimentos han tocado una porción de las ventajas citadas a favor de las sílabas en un SRAH. Los resultados positivos reportados en este trabajo pueden ser tomados para indicar el latente beneficio de incorporar propiedades adicionales de las sílabas.

La naturaleza ambigua de las fronteras de las sílabas es probablemente el factor más importante de limitación en la implementación debido a la búsqueda del inicio de las sílabas. Sin embargo, Cook y Robinson propusieron usar el mismo esquema de inicio de sílaba, con efectos positivos en una tarea de gran vocabulario, esto demostró que los beneficios son consistentes y los métodos escalables. Lo que respalda el hecho de que aunque los corpus manifestados en el presente trabajo son reducidos en cuanto al número de elementos a reconocer, la parte fundamental se conserva y se manifiesta como una parte experimental para ponerse a prueba en sistemas de mayor cantidad de elementos de reconocimiento.

### 7.3 CONTRIBUCIONES DE LA TESIS

Con formato

Los humanos pueden entender repeticiones de corpus de números con una gran mejora en condiciones de reverberaciones limpias y moderadas. Claramente, existe mucho trabajo permanente que realizar en cuestión del reconocimiento de voz pues no es una tarea simple.

Este trabajo contribuye al avance de la ciencia de la computación al presentar un método viable para realizar reconocimiento de voz por computadora haciendo uso de las estructuras silábicas en el español. La comunidad de los SRAH parece estar generalmente inclinada en contra de usar la sílaba para varios idiomas debido a cuestiones lingüísticas no resueltas y el considerable uso del fonema. Sin embargo, las lenguas romance como también se les conoce, presentan ciertas características favorables a las sílabas que otros lenguajes estructurados como el inglés no poseen. Esto es respaldado por que en el idioma español la forma en que se escribe es la forma en la que se lee y se pronuncia.

Los resultados de este trabajo contribuyen a mostrar evidencias que los métodos de combinación de características de segmentación, tienen un potencial significativo para alcanzar un incremento en la tarea de reconocimiento de voz, tal es el caso de permitir realizar la tarea de segmentación tanto desde el punto de vista de la energía como del parámetro ERO y del SBC.

Esta investigación ha incorporado muchas ideas derivadas de las teorías de la percepción acústica humana, incluyendo el uso de inicios de las sílabas y los intervalos de longitud de la sílaba y la combinación de elementos de procesamiento acordes. Los resultados experimentales demuestran la utilidad del uso de las sílabas en sistemas del habla continua y discontinua.

Una nota personal, el trabajo desarrollado, extendió los conceptos y sugerencias originalmente compartidas por los profesores Morgan y Greenberg (Arai and Greenberg, 1997). Una gran parte del tiempo fue enfocada en desarrollar la infraestructura esencial de la etapa de entrenamiento en los experimentos para demostrar que la información basada en sílaba puede mejorar el reconocimiento.

Las investigaciones en reconocimiento de voz son conducidas en su mayor parte a través de experimentos que llevan una gran cantidad de tiempo por la cantidad de información que se tiene que procesar. Para trabajar con el reconocimiento orientado a la sílaba, se implementó la unidad de representación (CDHMM) y se entrenaron los sistemas de reconocimiento basados en sílabas. Se realizaron diversos experimentos con esos sistemas individualmente y en combinación para examinar su eficiencia.

Este trabajo permitió la implantación de software para la identificación de los inicios de las sílabas y el reconocimiento basado en ellas. Dicho software puede ser usado para propósitos referentes al análisis y combinación de sistemas arbitrarios.

Los aportes del presente trabajo se enumeran a continuación:

- ❖ La inmersión del Sistema Experto a la tarea de entrenamiento representa uno de los puntos medulares de aportación del presente trabajo. El estudio del paradigma de la sílaba por sí solo representa en el idioma español una aportación, pues el reconocimiento de voz basado en tal paradigma se encuentra en estudio, lo cual representa, una alternativa de la ideología atómica que en varias áreas de la ciencia se plantean.
- ❖ El sistema basado en conocimiento tiene su razón de aplicarse tras la investigación realizada de la sílaba, pues independientemente del punto de vista lingüístico o textual que se analice a las frases del lenguaje, las reglas silábicas existen, por lo menos para el idioma español. Al extraer

estas reglas silábicas y tomarlas como referente en las etapas de entrenamiento representa un aporte.

- ❖ En busca de la mejora del reconocimiento obtenido se hizo uso del parámetro ERO, el cual es parte esencial de la tarea de segmentación, pues permitió encontrar regiones que en conjunto con la Función de Energía en Corto Tiempo no fue posible analizar. Esto mejoró los resultados obtenidos en trabajos realizados para el Portugués (Meneido et al., 1999) de forma significativa para la misma unidad empleada.
- ❖ La incrustación de un algoritmo sencillo de ajuste de parámetros de los modelos ocultos de Markov estáticos y su análisis, representa otra de las aportaciones del presente trabajo. Su análisis desemboca en que se puede crear un sistema de reconocimiento del habla discontinua haciendo uso de este algoritmo, que en realidad la expresión no es nueva pero sí la forma en la que se realiza, y tras lo experimentos realizados demuestra su grado de efectividad.

#### **7.4 TRABAJOS FUTUROS**

Este trabajo describió el uso de la información silábica para los SRAH en dos fases: la incorporación de estimaciones de inicio de sílabas en el habla discontinua y la combinación de estos tipos de estructuras en el habla continua. Dado que los sistemas actuales trabajan con la información de los sistemas basados en fonemas y la vertiente del presente trabajo se relaciona con sistemas basados en sílabas, una nueva dirección para trabajos futuros que puede emerger es la combinación futura de estos dos tipos de información enfocados a los SRAH en el idioma español.

Este trabajo abre un campo de estudio en la inmersión de las unidades silábicas que puede ser cristalizado con una visión del desarrollo de los algoritmos de reconocimiento para un microprocesador vectorial. La sílaba tiene muchas propiedades que son deseables para la computación vectorial: 1) los modelos basados en sílabas pueden ser conducidos a remover las ramificaciones durante la ejecución y 2) los modelos basados en sílabas son una unidad de organización natural para reducir la computación redundante y define el espacio de búsqueda.

De la misma forma aunque este trabajo no explora los beneficios de la programación paralela, algunas de las conclusiones de este trabajo son aplicables al procesamiento concurrente. A saber, la combinación de información de múltiples cadenas de Markov es una operación obviamente concurrente. El decodificador de dos niveles de Fosler-Lussier puede ser mapeado cuidadosamente en una máquina de procesador múltiple, dado que las probabilidades de diferentes palabras son calculadas independientemente.



Como se mencionó en el capítulo 3, algunos avances recientes en la tecnología de reconocimiento de voz han sido atribuidos a mejoras en el funcionamiento del hardware. Si este es el caso, usando máquinas paralelas y concurrentes puede ser ampliamente ventajosa la investigación del reconocimiento de voz.

Asimismo, la combinación de la metodología empleada en el presente trabajo al unirse con la basada en fonemas abre un campo de estudio relevante.

Un punto importante que puede incrementar el camino de la investigación en lo que a la inmersión de las sílabas a los sistemas de reconocimiento se refiere, es el hecho de introducir un conjunto de filtros que permitan determinar de manera adecuada las manifestaciones de fonemas de mayor ocurrencia en un corpus de voz que conforman a las sílabas.

Además, la particularidad de mejorar el problema de la entonación logrará incrementar el alcance que la sílaba tiene dentro del idioma español.

Finalmente, los trifonemas pueden ser analizados como unidades de reconocimiento y comparar los resultados que se obtengan con los expuestos en este trabajo, procurando establecer una alternativa de utilización de ambas unidades esenciales.

## **7.5 REFLEXIONES SOBRE EL FUTURO DE LA INVESTIGACION DE LOS SRAH (ASR's)**

El campo de los SRAH está en plena etapa de maduración, como resultado de ello, el paradigma usado en este trabajo.

Un proceso de evolución similar ocurrió en el campo del diseño de los microprocesadores. No hace mucho, los microprocesadores tenían sólo cientos de transistores y los grupos de investigación diseñaron chips de propósito especial. Hoy, los microprocesadores llegan cómodamente al consumidor, las tendencias en la creación de nuevos chips son manejadas por cuestiones del mercado. El estado del arte en el campo de los microprocesadores es altamente complejo con millones de transistores y se encuentra continuamente en expansión.

Los investigadores en la academia se enfocan a investigaciones específicas, tales como la operación de baja potencia y prototipos construidos de forma simplificada. Los chips por lo regular, son fabricados por grandes corporaciones. Se menciona esto debido a que en conversaciones con gente de Intel, se analizó la posibilidad de poder incrustar a las arquitecturas de los microprocesadores algunas características del reconocimiento que permitieran optimizar los resultados obtenidos en esta área de la ciencia.

En los SRAH los sistemas de reconocimiento de academia pueden libremente competir con los de la industria. Los sistemas de las universidades, tales como

el HTK de la Universidad de Cambridge, y el grupo de la Universidad de Carnegie Mellon, son ejemplo claro de esto.

El trabajo en colaboración es una tarea difícil por sí misma, como se ha demostrado por los esfuerzos de los ingenieros de VERBMOVIL. El proyecto VERBMOVIL donde se encuentra implicada Alemania, es un proyecto que reúne 29 lugares separados con 150 investigadores e ingenieros. La integración de los esfuerzos ha sido una enorme tarea que consume tiempo y todo con el objeto de avanzar en el estudio del reconocimiento de voz.

A pesar de estas desventajas, la colaboración se convierte en una ocurrencia común en la investigación del reconocimiento de voz.

## **7.6 CONCLUSIÓN FINAL**

En este trabajo se ha demostrado que la incorporación de la sílaba en un sistema de reconocimiento de voz aplicado a corpus pequeños y medianos, genera buenos resultados en sistemas tanto de habla continua como discontinua, lo cual resulta prometedor para aplicaciones de gran robustez. El reconocimiento orientado en sílabas representa un paradigma diferente al orientado en fonemas, sobre todo, cuando se aplica al idioma español. Dicho paradigma conduce a un rendimiento estable y sostenido, los experimentos demuestran tal hecho.

Los resultados demuestran que uno de los puntos importantes para abordar el uso de las sílabas en tales sistemas es analizar tales estructuras que conforman al idioma.

El idioma español a diferencia de otros lenguajes, posee un conjunto de reglas que describen la formación de las sílabas que se presentan en dicho idioma. El estudio de tal situación, se manifiesta tanto en corpus de voces como en textos literarios, esta característica fue aprovechada al implantar dichas reglas en el Sistema Basado en Conocimiento que ayudó en gran parte de los experimentos de este trabajo, en la parte de entrenamiento del sistema.

En cuanto a los resultados obtenidos, se deduce que el uso de las sílabas repercute en una alternativa al uso de los fonemas, como parte esencial que diferencia a uno del otro, las sílabas se conforman en estructuras bien definidas dentro del lenguaje (las reglas), mientras que los fonemas, carecen de tal elemento.

A lo largo de la presente investigación se hizo uso de la Energía del parámetro RO, la cual permitió incrementar el reconocimiento de voz tanto en corpus de habla continua como discontinua, al ser utilizada en conjunto con la energía en corto tiempo de la señal de voz en el dominio del tiempo. Una de las ventajas de utilizar la energía del parámetro RO es encontrar espacios de tiempo en donde la energía en corto tiempo de la señal de voz, es prácticamente nula y

por tanto, el trabajo en conjunto de estos elementos permitió la realización de una segmentación adecuada para los problemas que esta representa.

A su vez, el Sistema Basado en Conocimiento sirvió de elemento de verificación de la segmentación silábica automática realizada por los parámetros considerados con anterioridad. Los puntos anteriores denotan dos aspectos importantes:

- 1) La introducción de un Sistema Basado en Conocimiento permite por un lado, agregar conocimiento a priori a la etapa de segmentación, la cual es fundamental en el esquema propuesto. Por otro, la inmersión de técnicas de Inteligencia Artificial a los procesos de reconocimiento de voz se hacen cada vez más necesarios.
- 2) El esquema propuesto representa una extensión al propuesto por Furui en uno de sus libros, lo cual representa un aporte importante al área de investigación del reconocimiento de voz y;
- 3) Uno de los puntos esenciales que la comunidad científica dedicada al reconocimiento de voz por computadora es el incremento en los índices de reconocimiento. La presente investigación basa sus principios en el hecho de que para poder alcanzar un índice de reconocimiento alto, la etapa de segmentación debe de ser cuidadosamente guiada y realizada. Tras lo cual, la mayor parte de las investigaciones propuestas se fundamentan en ello, mas aún, los resultados obtenidos demuestran que tal aseveración es prácticamente cierta, lo cual incrementa la veracidad de lo antes escrito.

Indudablemente el reconocimiento de voz por computadora es un paradigma de la ciencia de la computación aún difícil de corresponderse con las cuestiones de idealismo planteadas en sus inicios, como sucede con muchos otros paradigmas relacionados, sin embargo, se considera que al coadyuvar técnicas tanto de unidades de reconocimiento, técnicas de reconocimiento y elementos extraídos de la Inteligencia Artificial que permitan incrementar los índices obtenidos actualmente, el reconocimiento de voz por computadora toma un rumbo adecuado.

Aunque los resultados no fueron muy alentadores al analizar un corpus de voz abstracto, se considera que tales resultados pueden ser superados si se introducen esquemas de análisis de prosodia y entonación, mismos que podrán ser agregados de alguna forma a la etapa de reconocimiento y al mismo Sistema Basado en Conocimiento propuesto en el presente trabajo.

Introducir el análisis del modelo del lenguaje para ejemplos como el trigram e incluso mezclar las ventajas del uso de las sílabas y otras unidades de reconocimiento también tienden a crear una nueva expectativa de investigación dentro del fascinante mundo del Reconocimiento Automático del Habla por computadora.

Uno de los aspectos desafiantes que el presente trabajo toma como punto de partida para otras investigaciones es la inclusión de las características de acentuación en la prosodia del lenguaje, este aspecto al tratar de ser introducido al sistema de reconocimiento proporcionó un error significativo, que tras la introducción de elementos como los expuestos en el capítulo 6 pueden ser analizados.

Para terminar, el reconocimiento de voz es un producto en demanda, existe una motivación considerable por resolver los problemas que contienen los sistemas que implementan esta técnica en el desarrollo universal. El trabajo en esta tesis usó la unidad de la sílaba y una combinación de métodos que contribuyen a un pequeño paso al final de la meta.

# Apéndice A

## MEDICIONES

VALOR DE LOS PRIMEROS 4 VECTORES CÓDIGO LPC PARA LA SÍLABA 'SE'. CADA LPC TIENE 12 COEFICIENTES												
VECTOR	a <sub>1</sub>	a <sub>2</sub>	a <sub>3</sub>	a <sub>4</sub>	a <sub>5</sub>	a <sub>6</sub>	a <sub>7</sub>	a <sub>8</sub>	a <sub>9</sub>	a <sub>10</sub>	a <sub>11</sub>	a <sub>12</sub>
1	0.079366	0.070605	0.066394	0.077170	0.066353	0.070070	0.052146	0.058739	0.049248	0.056445	0.052804	0.060309
2	0.079525	0.070746	0.066527	0.077324	0.066486	0.070210	0.052250	0.058856	0.049347	0.056558	0.052910	0.060430
3	0.797215	0.117602	0.081414	0.048322	0.034641	0.012271	0.008481	0.059069	0.077765	0.032905	0.016808	0.015897
4	0.790525	0.186353	0.031358	0.001378	0.016360	0.051041	0.038939	0.106714	0.032126	0.029867	0.065006	0.032603

Tabla A.1 Ejemplos de coeficientes LPC para los experimentos de las sílabas sa, se, si y so.

VALOR DE LOS PRIMEROS 4 VECTORES CÓDIGO LPC PARA LA SÍLABA 'SE'. CADA LPC TIENE 12 COEFICIENTES												
VECTOR	a <sub>1</sub>	a <sub>2</sub>	a <sub>3</sub>	a <sub>4</sub>	a <sub>5</sub>	a <sub>6</sub>	a <sub>7</sub>	a <sub>8</sub>	a <sub>9</sub>	a <sub>10</sub>	a <sub>11</sub>	a <sub>12</sub>
1	0.082729	0.077888	0.065369	0.070049	0.066662	0.056316	0.070115	0.059587	0.056548	0.054676	0.052854	0.051343
2	0.086017	0.073388	0.061725	0.082475	0.069900	0.058643	0.056526	0.058234	0.052843	0.055057	0.057267	0.051711
3	0.703504	0.261604	0.000948	0.006781	0.020338	0.003139	0.022059	0.027674	0.033781	0.002712	0.001538	0.021032
4	0.805635	0.152865	0.009247	0.004471	0.011203	0.041428	0.026564	0.020551	0.004050	0.036336	0.031818	0.012566

Tabla A.2 Ejemplos de coeficientes LPC para la sílaba si en un corpus de sílabas sa, se, si y so.

VALOR DE LOS PRIMEROS 4 VECTORES CÓDIGO LPC PARA LA SÍLABA 'SI'. CADA LPC TIENE 12 COEFICIENTES												
VECTOR	a <sub>1</sub>	a <sub>2</sub>	a <sub>3</sub>	a <sub>4</sub>	a <sub>5</sub>	a <sub>6</sub>	a <sub>7</sub>	a <sub>8</sub>	a <sub>9</sub>	a <sub>10</sub>	a <sub>11</sub>	a <sub>12</sub>
1	0.048149	0.084074	0.077306	0.087515	0.043228	0.087755	0.016584	0.061534	0.092858	0.030596	0.065524	0.064961
2	0.048198	0.084158	0.077384	0.087603	0.043271	0.087843	0.016600	0.061596	0.092951	0.030627	0.065590	0.065025
3	0.804172	0.010506	0.082399	0.067343	0.059956	0.021511	0.003520	0.039749	0.041429	0.007009	0.024260	0.005619
4	0.793166	0.206055	0.051315	0.027939	0.030075	0.017175	0.064449	0.054694	0.040056	0.044605	0.041858	0.024950

Tabla A.3 Ejemplos de coeficientes LPC para la sílaba so en un corpus de sílabas sa, se, si y so.

VALOR DE LOS PRIMEROS 4 VECTORES CÓDIGO LPC PARA LA SÍLABA 'SO'. CADA LPC TIENE 12 COEFICIENTES												
VECTOR	a <sub>1</sub>	a <sub>2</sub>	a <sub>3</sub>	a <sub>4</sub>	a <sub>5</sub>	a <sub>6</sub>	a <sub>7</sub>	a <sub>8</sub>	a <sub>9</sub>	a <sub>10</sub>	a <sub>11</sub>	a <sub>12</sub>
1	0.082634	0.078126	0.080780	0.068373	0.065119	0.062036	0.051466	0.050431	0.056741	0.047469	0.061342	0.059191
2	0.082304	0.078064	0.073224	0.077692	0.065336	0.062579	0.059764	0.049841	0.057222	0.054553	0.053053	0.051418
3	0.780530	0.171189	0.014461	0.044818	0.039639	0.093347	0.014180	0.018366	0.026334	0.014236	0.038050	0.006851
4	0.811665	0.119061	0.058244	0.050387	0.053720	0.031458	0.089727	0.031176	0.012735	0.029464	0.017280	0.000633

Tabla A.4 Ejemplos de coeficientes LPC para la sílaba su en un corpus de sílabas sa, se, si y so.

A continuación se muestran los resultados obtenidos para el caso de usar las Cadenas Ocultas de Markov para un corpus pequeño:

VALOR DE LOS PRIMEROS 9 VECTORES CÓDIGO LPC DEL LIBRO CÓDIGO GLOBAL PARA LAS SÍLABAS 'SA', 'SE', 'SO' Y 'SU'												
VECTOR	a <sub>1</sub>	a <sub>2</sub>	a <sub>3</sub>	a <sub>4</sub>	a <sub>5</sub>	a <sub>6</sub>	a <sub>7</sub>	a <sub>8</sub>	a <sub>9</sub>	a <sub>10</sub>	a <sub>11</sub>	a <sub>12</sub>
1	0.055240	0.083682	0.064483	0.087761	0.050022	0.075784	0.061193	0.045027	0.065094	0.072367	0.037214	0.061180
2	0.035073	0.061373	0.077515	0.061330	0.062453	0.087064	0.085617	0.043820	0.062725	0.091861	0.012542	0.109398
3	0.091669	0.077330	0.053711	0.068623	0.067559	0.065167	0.069676	0.064786	0.055912	0.054070	0.047517	0.044904
4	0.084408	0.079448	0.057611	0.074394	0.070131	0.065099	0.058886	0.058335	0.054557	0.051401	0.053018	0.056376
5	0.078932	0.077197	0.068983	0.067924	0.064948	0.069740	0.067221	0.065636	0.055948	0.048144	0.052233	0.046069
6	0.080150	0.074052	0.078806	0.071450	0.066302	0.061154	0.058761	0.056692	0.055833	0.056212	0.053879	0.051924
7	0.072275	0.069093	0.110488	0.053319	0.078975	0.053319	0.049725	0.063689	0.065305	0.042709	0.055084	0.039915
8	0.801051	0.153750	0.073895	0.057929	0.036753	0.084185	0.117092	0.028249	0.035376	0.066894	0.010210	0.021153
9	0.829547	0.161625	0.094727	0.034956	0.066129	0.071405	0.035264	0.075745	0.046439	0.078245	0.056593	0.055249

\*Valores negativos.

Tabla A.5 Algunos datos representativos del libro código global.

MATRIZ DE TRANSICIÓN DEL MODELO OCULTO DE MARKOV PARA LA SÍLABA 'NO' DE LA PALABRA U-NO						
ESTADO	q <sub>1</sub>	q <sub>2</sub>	q <sub>3</sub>	q <sub>4</sub>	q <sub>5</sub>	q <sub>6</sub>
q <sub>1</sub>	0.994390	0	0	0	0.5	0.5
q <sub>2</sub>	0	0	0	0	1	0
q <sub>3</sub>	0	0	0	0	0.5	0.5
q <sub>4</sub>	0	0	0	0	0	1
q <sub>5</sub>	0	0	0	0	0	1
q <sub>6</sub>	0					

Tabla A.6 Probabilidades de transición de estados para el modelo NO

MATRIZ DE TRANSICIÓN DEL MODELO OCULTO DE MARKOV PARA LA SÍLABA 'DOS' DE LA PALABRA DOS						
ESTADO	q <sub>1</sub>	q <sub>2</sub>	q <sub>3</sub>	q <sub>4</sub>	q <sub>5</sub>	q <sub>6</sub>
q <sub>1</sub>	0.826923	0	0.173077	0	0	0
q <sub>2</sub>	0	0	0	0	0.5	0.5
q <sub>3</sub>	0	0	0	0	1	0
q <sub>4</sub>	0	0	0	0	0.5	0.5
q <sub>5</sub>	0	0	0	0	0	1
q <sub>6</sub>	0	0	0	0	0	1

Tabla A.7 Probabilidades de transición de estados para el modelo DOS.

MATRIZ DE TRANSICIÓN DEL MODELO OCULTO DE MARKOV PARA LA SÍLABA 'TRES' DE LA PALABRA TRES						
ESTADO	q <sub>1</sub>	q <sub>2</sub>	q <sub>3</sub>	q <sub>4</sub>	q <sub>5</sub>	q <sub>6</sub>
q <sub>1</sub>	0.785714	0	0.214286	0	0	0
q <sub>2</sub>	0	0	0	0	0.5	0.5
q <sub>3</sub>	0	0	0	0	1	0
q <sub>4</sub>	0	0	0	0	0.5	0.5
q <sub>5</sub>	0	0	0	0	0	1
q <sub>6</sub>	0	0	0	0	0	1

Tabla A.8 Probabilidades de transición de estados para el modelo TRES.

MATRIZ DE TRANSICIÓN DEL MODELO OCULTO DE MARKOV PARA LA SÍLABA 'CUA' DE LA PALABRA CUA-TRO						
ESTADO	q <sub>1</sub>	q <sub>2</sub>	q <sub>3</sub>	q <sub>4</sub>	q <sub>5</sub>	q <sub>6</sub>
q <sub>1</sub>	0.987368	0	0.012632	0	0	0
q <sub>2</sub>	0	0	0	0	0.5	0.5
q <sub>3</sub>	0	0	0	0	1	0
q <sub>4</sub>	0	0	0	0	0.5	0.5
q <sub>5</sub>	0	0	0	0	0	1
q <sub>6</sub>	0	0	0	0	0	1

Tabla A.9 Probabilidades de transición de estados para el modelo CUA-TRO.

MATRIZ DE TRANSICIÓN DEL MODELO OCULTO DE MARKOV PARA LA SÍLABA 'CIN' DE LA PALABRA CIN-CO						
ESTADO	q <sub>1</sub>	q <sub>2</sub>	q <sub>3</sub>	q <sub>4</sub>	q <sub>5</sub>	q <sub>6</sub>
q <sub>1</sub>	0.964286	0	0.035714	0	0	0
q <sub>2</sub>	0	0	0	0	0.5	0.5
q <sub>3</sub>	0	0	0	0	1	0
q <sub>4</sub>	0	0	0	0	0.5	0.5
q <sub>5</sub>	0	0	0	0	0	1
q <sub>6</sub>	0	0	0	0	0	1

Tabla A.10 Probabilidades de transición de estados para el modelo CIN-CO.

MATRIZ DE TRANSICIÓN DEL MODELO OCULTO DE MARKOV PARA LA SÍLABA 'SEIS' DE LA PALABRA SEIS						
ESTADO	q <sub>1</sub>	q <sub>2</sub>	q <sub>3</sub>	q <sub>4</sub>	q <sub>5</sub>	q <sub>6</sub>
q <sub>1</sub>	0.957672	0	0.042328	0	0	0
q <sub>2</sub>	0	0	0	0	0.5	0.5
q <sub>3</sub>	0	0	0	0	1	0
q <sub>4</sub>	0	0	0	0	0.5	0.5
q <sub>5</sub>	0	0	0	0	0	1
q <sub>6</sub>	0	0	0	0	0	1

Tabla A.11 Probabilidades de transición de estados para el modelo SEIS

MATRIZ DE TRANSICIÓN DEL MODELO OCULTO DE MARKOV PARA LA SÍLABA 'SIE' DE LA PALABRA SIE-TE						
ESTADO	q <sub>1</sub>	q <sub>2</sub>	q <sub>3</sub>	q <sub>4</sub>	q <sub>5</sub>	q <sub>6</sub>
q <sub>1</sub>	0.971014	0	0.028986	0	0	0
q <sub>2</sub>	0	0	0	0	0.5	0.5
q <sub>3</sub>	0	0	0	0	1	0
q <sub>4</sub>	0	0	0	0	0.5	0.5
q <sub>5</sub>	0	0	0	0	0	1
q <sub>6</sub>	0	0	0	0	0	1

Tabla A.12 Probabilidades de transición de estados para el modelo SIE-TE.

MATRIZ DE TRANSICIÓN DEL MODELO OCULTO DE MARKOV PARA LA SÍLABA 'CHO' DE LA PALABRA O-CHO						
ESTADO	q1	q2	q3	q4	q5	q6
q1	0.980447	0	0.019553	0	0	0
q2	0	0	0	0	0.5	0.5
q3	0	0	0	0	1	0
q4	0	0	0	0	0.5	0.5
q5	0	0	0	0	0	1
q6	0	0	0	0	0	1

Tabla A.13 Probabilidades de transición de estados para el modelo O-CHO.

MATRIZ DE TRANSICIÓN DEL MODELO OCULTO DE MARKOV PARA LA SÍLABA 'CHO' DE LA PALABRA O-CHO						
ESTADO	b1	b2	b3	b4	b5	b6
O1	0.072222	0	0	0	0	0
O2	0	0	0	0	0	0
O3	0	0	0	0	0	0
O4	0.008333	0	0	0	0	0
O5	0.002778	0	0	0	0	0
O6	0.019444	0	0	0	0	0
O7	0	0	0	0	0	0
O8	0	0	0	0	0	0
O9	0.005556	0	0	0	0	.006693
O10	0	0	0	0	0	0
O11	0	0	0	0	0	0
O12	0	0	0	0	0	0.001339
O13	0	0	0	0	0	0.001339
O14	0.005556	0	0	0	0	0.001339
O15	0	0	0	0	0	0
O16	0	0	0	0	0	0.001339
O17	0	0	0	0	0	0
O18	0	0	0	0	0	0
O19	0	0	0	0	0	0
O20	0	0	0	0	0	0
O21	0	0	0	0	0	0
O22	0	0	0	0	0	0.001339
O23	0	0	0	0	0	0
O24	0	0	0	0	0	0.001339
O25	0	0	0	0	0	0.008032
O26	0.005556	0	0	0	0	0.004016
O27	0.002778	0	0	0	0	0.009371
O28	0	0	0	0	0	0.002677
O29	0	0	0	0	0	0.006693
O30	0.002778	0	0	0	0	0.004016
O31	0.002778	0	0	0	0	0.002677
O32	0.005556	0	0	0	0	0.001339
O33	0	0	0	0	0	0
O34	0	0	0	0	0	0
O35	0	0	0	0	0	0
O36	0	0	0	0	0	0
O37	0	0	0	0	0	0
O38	0	0	0	0	0	0
O39	0	0	0	0	0	0
O40	0	0	0	0	0	0
O41	0	0	0	0	0	0
O42	0	0	0	0	0	0
O43	0	0	0	0	0	0
O44	0	0	0	0	0	0
O45	0	0	0	0	0	0
O46	0	0	0	0	0	0
O47	0	0	0	0	0	0
O48	0	0	0	0	0	0
O49	0	0	0	0	0	0
O50	0.002778	0	0	0	0	0
O51	0	0	0	0	0	0
O52	0	0	0	0	0	0
O53	0	0	0	0	0	0
O54	0	0	0	0	0	0
O55	0	0	0	0	0	0
O56	0	0	0	0	0	0
O57	0	0	0	0	0	0
O58	0	0	0	0	0	0
O59	0	0	0	0	0	0
O60	0	0	0	0	0	0
O61	0	0	0	0	0	0



O <sub>62</sub>	0	0	0	0	0	0
O <sub>63</sub>	0	0	0	0	0	0
O <sub>64</sub>	0	0	0	0	0	0
O <sub>65</sub>	0	0	0	0	0	0.001339
O <sub>66</sub>	0	0	0	0	0	0
O <sub>67</sub>	0	0	0	0	0	0
O <sub>68</sub>	0	0	0	0	0	0
O <sub>69</sub>	0	0	0	0	0	0
O <sub>70</sub>	0	0	0	0	0	0
O <sub>71</sub>	0	0	0	0	0	0
O <sub>72</sub>	0	0	0	0	0	0
O <sub>73</sub>	0.013889	0	0	0	0	0.017403
O <sub>74</sub>	0.005556	0	0	0	0	0.010710
O <sub>75</sub>	0.002778	0	0	0	0	0
O <sub>76</sub>	0	0	0	0	0	0.008032
O <sub>77</sub>	0.013889	0	0	0	0	0.022758
O <sub>78</sub>	0.036111	0	0	0	0	0.041499
O <sub>79</sub>	0.016667	0	0	0	0	0.013387
O <sub>80</sub>	0.041667	0	0	0	0	0.057564
O <sub>81</sub>	0.011111	0	0	0	0	0.005355
O <sub>82</sub>	0.013889	0	0	0	0	0.006693
O <sub>83</sub>	0	0	0	0	0	0.008032
O <sub>84</sub>	0	0	0	0	0	0.010710
O <sub>85</sub>	0.002778	0	0	0	0	0.001339
O <sub>86</sub>	0	0	0	0	0	0
O <sub>87</sub>	0	0	0	0	0	0
O <sub>88</sub>	0	0	0	0	0	0
O <sub>89</sub>	0	0	0	0	0	0
O <sub>90</sub>	0	0	0	0	0	0
O <sub>91</sub>	0	0	0	0	0	0.002677
O <sub>92</sub>	0.027778	0	0	0	0	0.004016
O <sub>93</sub>	0	0	0	0	0	0
O <sub>94</sub>	0	0	0	0	0	0
O <sub>95</sub>	0.022222	0	0	0	0	0.024096
O <sub>96</sub>	0.036111	0	0	0	0	0.026774
O <sub>97</sub>	0	0	0	0	0	0.004016
O <sub>98</sub>	0	0	0	0	0	0
O <sub>99</sub>	0.005556	0	0	0	0	0.001339
O <sub>100</sub>	0	0	0	0	0	0.001339
O <sub>101</sub>	0	0	0	0	0	0
O <sub>102</sub>	0	0	0	0	0	0.004016
O <sub>103</sub>	0	0	0	0	0	0
O <sub>104</sub>	0	0	0	0	0	0
O <sub>105</sub>	0	0	0	0	0	0
O <sub>106</sub>	0.002778	0	0	0	0	0.002677
O <sub>107</sub>	0	0	0	0	0	0
O <sub>108</sub>	0	0	0	0	0	0
O <sub>109</sub>	0	0	0	0	0	0
O <sub>110</sub>	0	0	0	0	0	0
O <sub>111</sub>	0	0	0	0	0	0
O <sub>112</sub>	0	0	0	0	0	0
O <sub>113</sub>	0	0	0	0	0	0.001339
O <sub>114</sub>	0	0	0	0	0	0
O <sub>115</sub>	0	0	0	0	0	0
O <sub>116</sub>	0	0	0	0	0	0
O <sub>117</sub>	0	0	0	0	0	0.004016
O <sub>118</sub>	0.002778	0	0	0	0	0.004016
O <sub>119</sub>	0.005556	0	0	0	0	0.002677
O <sub>120</sub>	0	0	0	0	0	0
O <sub>121</sub>	0	0	0	0	0	0
O <sub>122</sub>	0	0	0	0	0	0.008032
O <sub>123</sub>	0.005556	0	0	0	0	0.020080
O <sub>124</sub>	0.011111	0	0	0	0	0.151272
O <sub>125</sub>	0.152778	0	0	0	0.142857	0.178046
O <sub>126</sub>	0.158333	0	0	0	0	0.175368
O <sub>127</sub>	0.116667	0	0.571429	0	0.571429	0.125837
O <sub>128</sub>	0.158333	0	0.428571	0	0.285714	0

Tabla A.14 Probabilidades de transición de la matriz B para el modelo O-CHO.

	Uno	Dos	Tres	Cuatro	Cinco	Seis	Siete	Ocho
Uno	9							1
Dos		10						

Tres			10					
Cuatro				10				
Cinco			2		8			
Seis						10		
Siete							10	
Ocho								10

\*Obteniéndose un 96.25%

Tabla A.15 Tabla de reconocimiento por el O-CHO.

### ANÁLISIS DE ESTRUCTURAS SILÁBICAS

Se realizó un análisis en diferentes textos científicos sobre sus estructuras silábicas, obteniéndose los siguientes resultados.

El número de palabras contenidas en los documentos fue mayor que el número de sílabas que lo componen. Las combinaciones más frecuentes en orden decreciente fueron: consonante-vocal (CV), consonante-vocal-consonante (CVC) y vocal-consonante (VC).

Las sílabas con el mayor número de repeticiones fueron:

1. de
2. la
3. a
4. te
5. que
6. es
7. se
8. en
9. do
10. ca

Se realizó el mismo análisis para las siguientes palabras, las cuales son las que forman el diccionario del presente Trabajo Terminal: Abrir, Cerrar, Rojo, Rosa, Negro, Blanco, Verde, Café, Azul, Gris, Cero, Uno, Dos, Tres, Cuatro, Cinco, Seis, Siete, Ocho, Nueve.

A-brir, Ce-rrar, Ro-jo, Ro-sa, Ne-gro, Blan-co, Ver-de, Ca-fé, A-zul, Gris, Ce-ro, U-no, Dos, Tres, Cua-tro, Cin-co, Seis, Sie-te, O-cho, Nue-ve.

a	2	cho	1	no	1	te	1
blan	1	de	1	nue	1	tres	1
brir	1	dos	1	o	1	tro	1
ca	1	fé	1	rrar	1	u	1
ce	2	gris	1	ro	3	ve	1
cin	1	gro	1	sa	1	ver	1
co	2	jo	1	seis	1	zul	1
cua	1	ne	1	sie	1		

20 Palabras, 36 Sílabas pero sólo 31 diferentes.

Observaciones:

CV	16	V	4	CWVC	1
CVC	4	CCVC	5		
CCV	3	CVV	3		

Tabla A.16 Análisis de la duración de las palabras del diccionario.

Palabra	Tiempo	# Muestras	Palabra	Tiempo	# Muestras	Palabra	Tiempo	# Muestras
verde1	0.7	7826	negro1	0.7	7848	cero1	0.51	5750
verde2	0.57	6390	negro2	0.78	8656	cero2	0.53	5974
verde3	0.6	6750	negro3	0.62	6884	cero3	0.52	5870
verde4	0.66	7394	negro4	0.65	7206	cero4	0.63	7020
verde5	0.93	10374	negro5	0.76	8516	cero5	0.65	7210
verde6	0.54	6026	negro6	0.59	6556	cero6	0.41	4594
verde7	0.56	6216	negro7	0.58	6446	cero7	0.38	4224
verde8	0.67	7450	negro8	0.63	7084	cero8	0.42	4734
verde9	0.57	6394	negro9	0.74	8194	cero9	0.49	5438
verde10	0.69	7730	negro10	0.58	6510	cero10	0.49	5532
<b>Promedio</b>	<b>0.649</b>	<b>7255</b>	<b>Promedio</b>	<b>0.663</b>	<b>7390</b>	<b>Promedio</b>	<b>0.503</b>	<b>5634.6</b>
azul1	0.81	9020	nueve1	0.61	6852	cinco1	0.66	7366
azul2	0.79	8840	nueve2	0.76	8512	cinco2	0.71	7878
azul3	0.69	7750	nueve3	0.5	5536	cinco3	0.68	7638
azul4	0.77	8586	nueve4	0.5	5588	cinco4	0.68	7550
azul5	0.85	9502	nueve5	0.7	7794	cinco5	0.82	9124
azul6	0.55	6112	nueve6	0.45	5084	cinco6	0.51	5760
azul7	0.59	6636	nueve7	0.54	6082	cinco7	0.55	6100
azul8	0.52	5808	nueve8	0.59	6552	cinco8	0.62	6974
azul9	0.61	6776	nueve9	0.66	7380	cinco9	0.58	6430
azul10	0.59	6640	nueve10	0.61	6828	cinco10	0.59	6546
<b>Promedio</b>	<b>0.677</b>	<b>7567</b>	<b>Promedio</b>	<b>0.592</b>	<b>6620.8</b>	<b>Promedio</b>	<b>0.64</b>	<b>7136.6</b>
blanco1	0.77	8624	ocho1	0.62	6912	cuatro1	0.63	6980
blanco2	0.83	9244	ocho2	0.73	8112	cuatro2	0.84	9322
blanco3	0.84	9386	ocho3	0.68	7586	cuatro3	0.66	7418
blanco4	0.86	9600	ocho4	0.69	7722	cuatro4	0.72	8066
blanco5	0.95	10580	ocho5	0.9	10052	cuatro5	0.83	9260
blanco6	0.62	6900	ocho6	0.58	6528	cuatro6	0.63	7086
blanco7	0.73	8124	ocho7	0.61	6822	cuatro7	0.58	6458
blanco8	0.65	7216	ocho8	0.68	7564	cuatro8	0.65	7266
blanco9	0.69	7672	ocho9	0.65	7242	cuatro9	0.58	6484
blanco10	0.62	6872	ocho10	0.61	6784	cuatro10	0.65	7254
<b>Promedio</b>	<b>0.756</b>	<b>8421.8</b>	<b>Promedio</b>	<b>0.675</b>	<b>7532.4</b>	<b>Promedio</b>	<b>0.677</b>	<b>7559.4</b>
cafe1	0.8	8880	rojo1	0.66	7362	dos1	0.49	5502
cafe2	0.71	7938	rojo2	0.71	7956	dos2	0.46	5164
cafe3	0.81	9030	rojo3	0.71	7894	dos3	0.49	5470
cafe4	0.65	7260	rojo4	0.78	8740	dos4	0.45	5100
cafe5	0.9	9980	rojo5	0.8	8921	dos5	0.52	5800
cafe6	0.46	5144	rojo6	0.53	5942	dos6	0.37	4134
cafe7	0.53	5956	rojo7	0.57	6382	dos7	0.31	3466
cafe8	0.5	5566	rojo8	0.57	6392	dos8	0.47	5226
cafe9	0.61	6760	rojo9	0.7	7758	dos9	0.41	4576
cafe10	0.55	6110	rojo10	0.59	6542	dos10	0.36	4036
<b>Promedio</b>	<b>0.652</b>	<b>7262.4</b>	<b>Promedio</b>	<b>0.662</b>	<b>7388.9</b>	<b>Promedio</b>	<b>0.433</b>	<b>4847.4</b>

Palabra	Tiempo	# Muestras	Palabra	Tiempo	# Muestras	Palabra	Tiempo	# Muestras
gris1	0.44	4964	rosa1	0.58	6528	siete1	0.58	6482
gris2	0.31	3558	rosa2	0.85	9482	siete2	0.64	7128
gris3	0.36	4062	rosa3	0.71	7866	siete3	0.72	8030
gris4	0.37	4194	rosa4	0.75	8410	siete4	0.64	7156
gris5	0.49	5534	rosa5	0.8	8960	siete5	0.74	8240
gris6	0.43	4828	rosa6	0.56	6280	siete6	0.53	5956
gris7	0.33	3728	rosa7	0.58	6522	siete7	0.62	6934
gris8	0.4	4534	rosa8	0.57	6406	siete8	0.54	6080
gris9	0.4	4448	rosa9	0.61	6798	siete9	0.62	6902
gris10	0.39	4350	rosa10	0.58	6436	siete10	0.63	6994
<b>Promedio</b>	<b>0.392</b>	<b>4420</b>	<b>Promedio</b>	<b>0.659</b>	<b>7368.8</b>	<b>Promedio</b>	<b>0.626</b>	<b>6990.2</b>
uno1	0.67	7436	seis1	0.41	4578	tres1	0.4	4542
uno2	0.6	6702	seis2	0.43	4814	tres2	0.54	6058
uno3	0.6	6708	seis3	0.41	4626	tres3	0.47	5244
uno4	0.74	8302	seis4	0.39	4360	tres4	0.49	5522
uno5	0.74	8294	seis5	0.56	6318	tres5	0.54	6028
uno6	0.45	5088	seis6	0.4	4542	tres6	0.37	4190
uno7	0.46	5194	seis7	0.39	4336	tres7	0.42	4726
uno8	0.5	5566	seis8	0.54	6000	tres8	0.42	4710
uno9	0.53	5896	seis9	0.47	5266	tres9	0.43	4854
uno10	0.43	4836	seis10	0.4	4538	tres10	0.41	4596
<b>Promedio</b>	<b>0.572</b>	<b>6402.2</b>	<b>Promedio</b>	<b>0.44</b>	<b>4937.8</b>	<b>Promedio</b>	<b>0.449</b>	<b>5047</b>
abrir1	0.81	9020	cerrar1	0.64	7520			
abrir2	0.79	9040	cerrar2	0.84	9322			
abrir3	0.73	8005	cerrar3	0.58	7624			
abrir4	0.77	8521	cerrar4	0.72	8235			
abrir5	0.85	9502	cerrar5	0.83	9260			
abrir6	0.69	7800	cerrar6	0.63	9510			
abrir7	0.71	9005	cerrar7	0.8	8362			
abrir8	0.71	6280	cerrar8	0.65	7266			
abrir9	0.74	7500	cerrar9	0.63	9451			
abrir10	0.74	8500	cerrar10	0.65	7254			
<b>Promedio</b>	<b>0.754</b>	<b>8312.8</b>	<b>Promedio</b>	<b>0.73</b>	<b>8048.3</b>			

Tabla A.17 Análisis de duración de las sílabas que componen el diccionario

Sílaba	Tiempo	# Muestras	Sílaba	Tiempo	# Muestras	Sílaba	Tiempo	# Muestras
a1	0.18	1986	o1	0.27	2982	seis1	0.41	4578
a2	0.2	2242	o2	0.23	2522	seis2	0.43	4813
a3	0.2	2258	o3	0.27	2962	seis3	0.14	4626
a4	0.19	2098	o4	0.21	2342	seis4	0.39	4360
a5	0.17	1826	o5	0.25	2762	seis5	0.56	6318
a6	0.13	1426	o6	0.21	2282	seis6	0.4	4542
a7	0.17	1834	o7	0.2	2222	seis7	0.39	4336
a8	0.11	1258	o8	0.2	2202	seis8	0.54	6000
a9	0.14	1490	o9	0.23	2562	seis9	0.47	5266

a10	0.11	1218	o10	0.18	2002	seis10	0.4	4538
<b>Promedio</b>	<b>0.16</b>	<b>1763.6</b>	<b>Promedio</b>	<b>0.225</b>	<b>2484</b>	<b>Promedio</b>	<b>0.413</b>	<b>4937.7</b>
blan1	0.43	4754	jo1	0.4	4452	sie1	0.21	2282
blan2	0.37	4034	jo2	0.48	5266	sie2	0.24	2682
blan3	0.43	4786	jo3	0.47	5226	sie3	0.3	3342
blan4	0.43	4738	jo4	0.52	5692	sie4	0.25	2722
blan5	0.46	5058	jo5	0.5	5556	sie5	0.27	2942
blan6	0.33	3602	jo6	0.32	3512	sie6	0.21	2282
blan7	0.42	4602	jo7	0.35	3872	sie7	0.23	2502
blan8	0.32	3554	jo8	0.3	3344	sie8	0.18	2002
blan9	0.32	3482	jo9	0.38	4190	sie9	0.18	2002
blan10	0.29	3218	jo10	0.33	3692	sie10	0.17	1902
<b>Promedio</b>	<b>0.38</b>	<b>4182.8</b>	<b>Promedio</b>	<b>0.405</b>	<b>4480.2</b>	<b>Promedio</b>	<b>0.224</b>	<b>2466</b>
ca1	0.24	2690	gris1	0.44	4964	te1	0.37	4114
ca2	0.19	2042	gris2	0.31	3558	te2	0.4	4358
ca3	0.24	2610	gris3	0.36	4062	te3	0.42	4602
ca4	0.17	1842	gris4	0.37	4194	te4	0.39	4348
ca5	0.2	2162	gris5	0.49	5534	te5	0.47	5212
ca6	0.11	1218	gris6	0.43	4829	te6	0.33	3586
ca7	0.15	1602	gris7	0.33	3728	te7	0.39	4344
ca8	0.15	1706	gris8	0.4	4534	te8	0.36	3992
ca9	0.15	1610	gris9	0.4	4448	te9	0.44	4812
ca10	0.15	1690	gris10	0.39	4350	te10	0.45	5006
<b>Promedio</b>	<b>0.175</b>	<b>1917.2</b>	<b>Promedio</b>	<b>0.392</b>	<b>4420.1</b>	<b>Promedio</b>	<b>0.402</b>	<b>4437.4</b>
ce1	0.25	2722	fe1	0.55	6104	tres1	0.4	4542
ce2	0.22	2466	fe2	0.53	5810	tres2	0.54	6058
ce3	0.25	2706	fe3	0.57	6332	tres3	0.47	5244
ce4	0.25	2810	fe4	0.48	5330	tres4	0.49	5522
ce5	0.25	2746	fe5	0.7	7730	tres5	0.54	6028
ce6	0.2	2178	fe6	0.35	384	tres6	0.37	4190
ce7	0.19	2066	fe7	0.39	4268	tres7	0.42	4726
ce8	0.18	2018	fe8	0.38	4164	tres8	0.42	4710
ce9	0.22	2474	fe9	0.46	5064	tres9	0.43	4854
ce10	0.24	2642	fe10	0.39	4332	tres10	0.41	4596
<b>Promedio</b>	<b>0.225</b>	<b>2482.8</b>	<b>Promedio</b>	<b>0.48</b>	<b>4951.8</b>	<b>Promedio</b>	<b>0.449</b>	<b>5047</b>
<b>Sílaba</b>	<b>Tiempo</b>	<b># Muestras</b>	<b>Sílaba</b>	<b>Tiempo</b>	<b># Muestras</b>	<b>Sílaba</b>	<b>Tiempo</b>	<b># Muestras</b>
cin1	0.34	3714	gro1	0.34	3720	tro1	0.33	3596
cin2	0.32	3506	gro2	0.28	3136	tro2	0.48	5295
cin3	0.26	2834	gro3	0.31	3370	tro3	0.42	4632
cin4	0.32	3530	gro4	0.37	4046	tro4	0.48	5250
cin5	0.29	3154	gro5	0.42	1682	tro5	0.5	5540
cin6	0.25	2762	gro6	0.29	3154	tro6	0.33	3684
cin7	0.24	2658	gro7	0.26	2872	tro7	0.35	3850
cin8	0.26	2834	gro8	0.34	3748	tro8	0.42	4598
cin9	0.26	2898	gro9	0.36	3922	tro9	0.36	3974
cin10	0.22	2450	gro10	0.28	3054	tro10	0.42	4651
<b>Promedio</b>	<b>0.276</b>	<b>3034</b>	<b>Promedio</b>	<b>0.325</b>	<b>3270.4</b>	<b>Promedio</b>	<b>0.409</b>	<b>4507</b>

co1	0.34	3784	ne1	0.37	4042	u1	0.26	2842
co2	0.46	5124	ne2	0.44	4834	u2	0.22	2402
co3	0.41	4514	ne3	0.31	3426	u3	0.16	1782
co4	0.43	4776	ne4	0.28	3074	u4	0.22	2402
co5	0.49	5436	ne5	0.34	3746	u5	0.23	2482
co6	0.29	3212	ne6	0.3	3314	u6	0.15	1602
co7	0.31	3436	ne7	0.32	3482	u7	0.15	1662
co8	0.32	3574	ne8	0.29	3250	u8	0.17	1822
co9	0.37	4102	ne9	0.38	4186	u9	0.15	1702
co10	0.32	3568	ne10	0.31	3370	u10	0.14	1562
<b>Promedio</b>	<b>0.374</b>	<b>4152.6</b>	<b>Promedio</b>	<b>0.334</b>	<b>3672.4</b>	<b>Promedio</b>	<b>0.185</b>	<b>2026</b>
cua1	0.3	3298	no1	0.41	4508	ve1	0.21	2314
cua2	0.36	3938	no2	0.38	4212	ve2	0.17	1864
cua3	0.24	2698	no3	0.44	4854	ve3	0.14	1570
cua4	0.25	2730	no4	0.53	5812	ve4	0.15	1690
cua5	0.33	3634	no5	0.52	5726	ve5	0.15	1704
cua6	0.3	3314	no6	0.31	3400	ve6	0.15	1613
cua7	0.23	2522	no7	0.31	3446	ve7	0.26	2854
cua8	0.23	2582	no8	0.33	3658	ve8	0.3	3272
cua9	0.22	2422	no9	0.37	4106	ve9	0.33	3652
cua10	0.23	2514	no10	0.29	3186	ve10	0.35	3842
<b>Promedio</b>	<b>0.269</b>	<b>2965.2</b>	<b>Promedio</b>	<b>0.389</b>	<b>4290.8</b>	<b>Promedio</b>	<b>0.221</b>	<b>2437.5</b>
cho1	0.35	3844	nue1	0.4	4450	ver1	0.38	4162
cho2	0.5	5502	nue2	0.6	6562	ver2	0.31	3458
cho3	0.41	4538	nue3	0.36	3930	ver3	0.31	3426
cho4	0.48	5292	nue4	0.35	3810	ver4	0.32	3538
cho5	0.65	7204	nue5	0.54	6002	ver5	0.39	4290
cho6	0.38	4158	nue6	0.31	3382	ver6	0.2	2214
cho7	0.41	4512	nue7	0.28	3142	ver7	0.24	2594
cho8	0.48	5274	nue8	0.29	3194	ver8	0.37	4098
cho9	0.42	4594	nue9	0.33	3642	ver9	0.24	2626
cho10	0.42	4676	nue10	0.26	2898	ver10	0.31	3466
<b>Promedio</b>	<b>0.45</b>	<b>4959.4</b>	<b>Promedio</b>	<b>0.372</b>	<b>4101.2</b>	<b>Promedio</b>	<b>0.307</b>	<b>3387.2</b>

Sílaba	Tiempo	# Muestras	Sílaba	Tiempo	# Muestras	Sílaba	Tiempo	# Muestras
de1	0.32	3572	ro1	0.27	2942	zul1	0.63	6946
de2	0.26	2846	ro2	0.31	3422	zul2	0.59	6510
de3	0.29	3236	ro3	0.28	3076	zul3	0.49	5404
de4	0.34	3770	ro4	0.37	4124	zul4	0.58	6402
de5	0.54	5996	ro5	0.4	4378	zul5	0.69	7590
de6	0.34	3724	ro6	0.21	2328	zul6	0.42	4598
de7	0.32	3536	ro7	0.19	2072	zul7	0.43	4716
de8	0.3	3264	ro8	0.24	2628	zul8	0.4	4464
de9	0.33	3680	ro9	0.26	2878	zul9	0.47	5198
de10	0.38	4178	ro10	0.25	2804	zul10	0.48	5336
<b>Promedio</b>	<b>0.342</b>	<b>3780.2</b>	<b>Promedio</b>	<b>0.278</b>	<b>3065.2</b>	<b>Promedio</b>	<b>0.518</b>	<b>5716.4</b>
dos1	0.49	5502	brir1	0.34	3720	sa1	0.35	3820

dos2	0.46	5164	brir2	0.28	3136	sa2	0.33	5914
dos3	0.49	5470	brir3	0.31	3370	sa3	0.54	5238
dos4	0.45	5100	brir4	0.37	4046	sa4	0.48	4962
dos5	0.52	5800	brir5	0.42	1682	sa5	0.45	5312
dos6	0.37	4134	brir6	0.29	3154	sa6	0.48	3446
dos7	0.13	3466	brir7	0.26	2872	sa7	0.31	3734
dos8	0.47	5226	brir8	0.34	3748	sa8	0.34	3716
dos9	0.41	4576	brir9	0.36	3922	sa9	0.34	4128
dos10	0.36	4036	brir10	0.28	3054	sa10	0.37	3688
<b>Promedio</b>	<b>0.415</b>	<b>4847.4</b>	<b>Promedio</b>	<b>0.325</b>	<b>3270.4</b>	<b>Promedio</b>	<b>0.399</b>	<b>4395.8</b>
			rrar1	0.3	3298			
			rrar2	0.36	3938			
			rrar3	0.24	2698			
			rrar4	0.25	2730			
			rrar5	0.33	3634			
			rrar6	0.3	3314			
			rrar7	0.23	2522			
			rrar8	0.23	2582			
			rrar9	0.22	2422			
			rrar10	0.23	2514			
			<b>Promedio</b>	<b>0.269</b>	<b>2965.2</b>			

Tabla A.18 Distribución de vectores en Regiones

# Región	#Elementos	# Región	#Elementos	# Región	#Elementos	# Región	#Elementos
1	137	33	67	65	11	97	54
2	12	34	59	66	52	98	62
3	58	35	39	67	48	99	35
4	24	36	21	68	61	100	49
5	981	37	42	69	23	101	47
6	579	38	11	70	30	102	24
7	128	39	70	71	10	103	39
8	111	40	53	72	13	104	58
9	41	41	5	73	78	105	79
10	12	42	6	74	29	106	253
11	23	43	28	75	76	107	223
12	23	44	22	76	50	108	549
13	16	45	33	77	128	109	114
14	37	46	26	78	38	110	134
15	123	47	19	79	30	111	113
16	53	48	19	80	34	112	90
17	45	49	12	81	44	113	44
18	103	50	27	82	72	114	126

19	66	51	49	83	46	115	82
20	178	52	43	84	32	116	127
21	45	53	74	85	28	117	52
22	83	54	15	86	56	118	136
23	88	55	23	87	12	119	106
24	119	56	38	88	42	120	51
25	69	57	23	89	53	121	51
26	65	58	35	90	51	122	66
27	39	59	40	91	15	123	84
28	57	60	18	92	10	124	100
29	56	61	70	93	83	125	440
30	67	62	17	94	46	126	403
31	93	63	30	95	80	127	142
32	88	64	48	96	66	128	169

**Número Total de Rx 10250**

Tabla A.19 Distribución en Regiones de la sílaba "tres"

# Región	0	1	2	3	4
1	233	235	234	234	234
2	145	143	144	144	144
	378	378	378	378	378

# Región	0	1	2	3	4	5	6	7
1	88	68	59	49	46	45	45	45
2	146	159	167	176	178	179	179	179
3	88	93	90	90	88	87	87	87
4	56	58	62	63	66	67	67	67
	378	378	378	378	378	378	378	378

# Región	0	1	2	3	4	5	6	7	8
1	28	35	38	37	36	35	35	35	35
2	17	13	12	13	14	15	15	15	15
3	104	107	97	113	114	109	108	108	108
4	75	65	72	59	58	62	63	63	63
5	48	60	62	66	66	67	67	67	67
6	39	24	20	20	20	20	20	20	20



7	36	49	54	50	51	52	52	52	52
8	31	25	23	20	19	18	18	18	18
	378	378	378	378	378	378	378	378	378

# Región	0	1	2	3	4	5	6	7	8	9	10	11	12
1	15	24	37	44	45	43	43	43	43	43	43	43	43
2	20	10	4	4	5	6	6	6	6	6	6	6	6
3	9	12	13	14	15	16	16	16	16	16	16	16	16
4	6	5	5	5	5	5	5	5	5	5	5	5	5
5	54	40	34	33	33	31	30	30	30	30	29	29	29
6	54	63	58	49	43	43	43	43	43	42	43	43	43
7	38	37	37	34	32	30	30	26	24	23	24	24	24
8	25	28	32	37	41	44	45	49	51	53	53	53	53
9	32	35	31	31	32	31	31	31	31	31	31	31	31
10	35	28	26	23	20	21	22	22	22	22	22	22	22
11	11	14	14	14	15	15	14	14	14	14	14	14	14
12	9	10	12	12	12	12	12	12	12	12	12	12	12
13	28	33	37	39	41	42	42	42	42	42	41	41	41
14	24	23	22	22	21	20	20	20	20	20	20	20	20
15	9	6	5	5	5	5	5	5	5	5	5	5	5
16	9	10	11	12	13	14	14	14	14	14	14	14	14
	378	378	378	378	378	378	378	378	378	378	378	378	378

# Región	0	1	2	3	4	5	6	7	8	9	10	11	12
1	21	12	11	11	11	11	11	11	11	11	12	12	12
2	22	30	29	26	25	25	26	27	29	29	29	29	29
3	4	13	16	19	19	18	17	18	18	18	18	18	18
4	2	2	2	2	2	2	2	2	2	2	2	2	2
5	10	6	7	7	7	7	7	7	7	7	7	7	7
6	6	4	3	3	3	3	3	3	3	3	3	3	3
7	2	1	1	1	1	1	1	1	1	1	1	1	1
8	3	8	8	8	9	10	11	11	11	11	11	11	11
9	14	14	11	17	18	17	15	15	15	15	15	15	15
10	15	16	18	11	11	11	11	11	11	11	11	11	11
11	22	19	19	21	23	26	28	29	29	29	29	29	29
12	21	21	21	23	23	22	21	20	20	20	20	20	20
13	10	11	13	14	13	12	12	12	12	12	12	12	12
14	14	14	13	11	11	12	13	13	13	13	13	13	13
15	29	31	34	37	38	39	38	38	38	38	38	38	38
16	24	21	18	14	12	11	11	11	11	11	11	11	11
17	13	7	6	5	5	5	5	5	5	5	5	5	5
18	18	25	26	27	26	26	26	26	26	26	25	25	25
19	12	11	11	11	11	11	11	11	11	11	11	11	11
20	10	11	11	11	11	11	11	11	11	11	11	11	11
21	8	6	6	6	6	6	6	6	6	6	6	6	6

22	6	10	11	10	11	11	11	11	11	11	11	11	11
23	6	5	4	5	5	5	5	5	5	5	5	5	5
24	6	4	4	4	4	4	4	4	4	4	4	4	4
25	23	16	17	18	18	18	18	17	16	17	17	17	17
26	18	20	19	17	16	15	15	14	13	12	12	12	12
27	11	11	10	10	10	10	10	10	10	10	10	10	10
28	9	5	4	4	4	4	4	4	4	4	4	4	4
29	3	4	4	4	4	4	4	4	4	4	4	4	4
30	2	2	2	2	2	2	2	2	2	2	2	2	2
31	9	14	15	15	15	15	15	15	15	15	15	15	15
32	5	4	4	4	4	4	4	4	4	4	4	4	4
	378	378	378	378	378	378	378	378	378	378	378	378	378

**ANÁLISIS DEL NÚMERO DE ERRORES POR OPTIMIZACIÓN PARA 32, 64 Y 128 REGIONES.**

Tabla A.20 Errores por optimización para 32 regiones con 200 muestras para sílabas.

Optimización	Errores
1	54
2	48
3	47
4	48
5	47
6	46
7	45
8	47
9	49
10	51

Tabla A.21 Errores por optimización para 64 regiones con 200 muestras para sílabas.

Optimización	Errores
1	31
2	21
3	18
4	17
5	17
6	14
7	14
8	16
9	17
10	17

Tabla A.22 Errores por optimización para 128 regiones con 200 muestras para sílabas.

Optimización	Errores
1	20
2	15
3	14
4	11
5	11
6	10
7	11
8	11
9	12
10	13

Tabla A.23 Errores por optimización para 32 regiones con 400 muestras para sílabas.

Optimización	Errores
1	21
2	22
3	20
4	23
5	22
6	22
7	23
8	23
9	24
10	24

Tabla A.24 Errores por optimización para 64 regiones con 400 muestras para sílabas.

Optimización	Errores
1	39
2	42
3	45
4	43
5	46
6	49
7	39
8	41
9	41
10	47

Tabla A.25 Errores por optimización para 128 regiones con 400 muestras para sílabas.

Optimización	Errores
1	73
2	67
3	69
4	68
5	64

6	63
7	73
8	68
9	68
10	67

Tabla A.26 Errores por optimización para 32 regiones con 200 muestras para palabras.

Optimización	Errores
1	11
2	0
3	0
4	1
5	1
6	1
7	1
8	1
9	1
10	1

Tabla A.27 Errores por optimización para 64 regiones con 200 muestras para palabras.

Optimización	Errores
1	11
2	2
3	3
4	3
5	3
6	4
7	5
8	5
9	6
10	7

Tabla A.28 Errores por optimización para 128 regiones con 200 muestras para palabras.

Optimización	Errores
1	13
2	2
3	6
4	7
5	5
6	7
7	9
8	9
9	10
10	11

Tabla A.29 Errores por optimización para 32 regiones con 400 muestras para palabras.

Optimización	Errores
1	3
2	4
3	2
4	2
5	2
6	2
7	1
8	2
9	2
10	2

Tabla A.30 Errores por optimización para 64 regiones con 400 muestras para palabras.

Optimización	Errores
1	9
2	9
3	7
4	7
5	5
6	5
7	3
8	5
9	4
10	13

Tabla A.31 Errores por optimización para 128 regiones con 400 muestras para palabras.

Optimización	Errores
1	16
2	16
3	14
4	17
5	17
6	17
7	13
8	12
9	10
10	13

## REFERENCIAS BIBLIOGRÁFICAS

---

Arai and Greenberg (1997). Arai Takayuki and Greenberg Steven, "The temporal properties of spoken Japanese are similar to those of English". In proceedings of Eurospeech'97, Rhodes, Greece, September 1997. ESCA.

Barrón (1998). Barrón Fernández Ricardo, "Reconocimiento de palabras aisladas usando cuantización vectorial", tesis de maestría, Centro de Investigación en Computación, Octubre 1998.

Becchetti and Prina (1999). Becchetti Claudio and Prina Ricotti Lucio, "Speech Recognition Theory and C++ Implementation", Fondazione Ugo Bordón, Rome, Italy, John Wiley and Sons, Ltd, 1999.

Bernal et al. (2000). Bernal J., Bobadilla J., Gómez P., Reconocimiento de Voz y Fonética Acústica, Alfaomega, México D. F., 2000.

Bernstein (1994). Bernstein, Jared, et al, The Latino40 Speech Database. Entropic Research Laboratory, Washington, DC. 1994.

Boulard (1996). Boulard Dupont S. "A new ASR Approach Based On Independent Processing and Recombination of Frequency Bands". Proceedings of ICSLP, Vol 1, pp. 426-429, 1996.

Chang (2000). Chang Eric, Zhou J., Di S., Huang C. and Lee K. "Large Vocabulary Mandarin Speech Recognition with Different Approaches in Modeling Tones". In proceedings of the 2000 International Conference on Spoken Language Processing, (Beijing, Oct.) pp. 983-986, 2000.

Córdoba et al. (1995). Córdoba R., Menéndez-Pidal X., Macías Guasara J., "Development and Improvement of a Real-Time ASR System for Isolated Digits in Spanish over the Telephone Line". In Proceedings Eurospeech'95, pp. 1537-1540, Madrid 1995.

Delaney (2000). Delaney Brian W. "Voice User Interface for Wireless Internetworking". A Qualifying Examination Report, Georgia Institute of Technology, School of Electrical and Computer Engineering Atlanta Georgia. November 6, 2000.

Fanty (1996). Fanty M., "Overview of the CSLU Toolkit", Center for Spoken Language Understanding, Oregon Graduate Institute of Science & Technology, USA 1996.

Feal (2000). Feal L., "Sobre el uso de la sílaba como unidad de síntesis en el español", Informe Técnico, Departamento de Informática, Universidad de Valladolid, 2000.

Fosler et al. (1999). Fosler-Lussier E., Greenberg S., Morgan N., "Incorporating Contextual Phonetics into Automatic Speech Recognition". XIV International Congress of Phonetic Sciences, pp. 611-614, San Francisco, 1999.

Fujimura (1975a). Fujimura O., "Transactions on Acoustics Speech and Signal processing", ASSP-23(1):82-87, February 1975.

Fujimura (1975b). Fujimura O., "Syllable as a Unit of Speech Recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP Vol. 23, no. 1, pp. 82-87, February 1975.

Ganapathiraju et al. (2001). Ganapathiraju A., Hamaker J., Picone J., Doddington G., "Syllable-Based Large Vocabulary Continuous Speech Recognition". IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 4, Mayo 2001, pp. 358-366.

Giarratano and Riley (2001). Giarratano Joseph y Riley Gary, International Thompson Editores, Sistemas expertos, principios y programación 2001.

Hamaker et al. (1998). Hamaker J., Ganapathiraju A., Picone J., Godfrey J., "Advances in Alphanumeric Recognition using syllables", submitted to the IEEE International Conference on Acoustics, Speech and Signal Processing, Seattle, Washington, USA, May 1998.

Hartmut et al. (1996). Hartmut R. Pfitzinger, Susanne Burger, Sebastian Heid, "Syllable Detection in Read and Spontaneous Speech", Institut für Phonetik und Sprachliche Kommunikation, University of Munich, Germany, proceedings of 4<sup>th</sup> International Conference on Spoken Language Processing, Philadelphia PA, USA October 3-6, 1996.

Hauenstein (1996). Hauenstein A., "The syllable Re-revisited", Technical Report, Siemens AG, Corporate Research and Development, München Alemania, 1996.

Hu et al. (1996). Hu Z., Schalkwyk J., Barnard E., Cole R., "Speech Recognition using Syllable-like Units", Proceedings of ICSLP'96, Vol. 2, pp. 1117-1120, 1996.

Jackson (1986). Jackson L. B. "Digital Filters and Signal Processing". Kluwer Academic Publishers. University of Louisville, Department of Electrical and Computer Engineering, U.S.A., 1986

Jones et al. (1999). Jones R., Downey S., Mason J., "Continuous Speech Recognition Using Syllables", Proceedings of Eurospeech, Vol. 3, pp. 1171-1174, Rhodes, Grecia 1999.

Kamakshi et al. (2002). Kamakshi V. Prasad, Nagarajan T. and Murthy Hema A.. "Continuous Speech Recognition Using Automatically Segmented Data at Syllabic Units". Department of Computer Science and Engineering. Indian Institute of Technology, Madras, Chennai 600-036. 2002.

Kershaw (1996). Kershaw Daniel Jeremy, "Phonetic Context-Dependency In a Hybrid ANN/HMM Speech Recognition System". PhD thesis, St. John's College, University of Cambridge, September 1996.

King (2000). King S., Taylor P., Frankel J. and Richmond K., "Speech Recognition via Phonetically Featured Syllables". In PHONUS, volume 5, pages 15-34, Institute of Phonetics, University of the Saarland, 2000.

Kirschning (1998). Kirschning Albers Ingrid, "Automatic Speech Recognition with the Parallel Cascade Neural Network", PhD Thesis, Tokyo Japan, March 1998.

Kita et al. (1993). Kita K., Morimoto T. and Sagayama S., "LR parsing with a category reachability test applied to speech recognition". IEICE Trans. Information and Systems, vol. E76-D, no.1, pp. 23-28, 1993.

Kosko (1992). Kosko B., "Neural Networks for Signal Processing", Prentice Hall, U.S.A., 1992.

Ladefoged (1993). Ladefoged Peter. "A course in Phonetics". Harcourt Brace College Pulishers, New York, Third Edition, pp. 141-147, 1993.

Lee and Ching (1998). Tan Lee, P. C. Ching, "A Neural Network Based Speech Recognition System For Isolated Cantonese Syllables". The Chinese University of Hong Kong, proceedings of Neural Networks for Speech Processing (Lecture) NSP2L.10 Vol. 4, pp. 3269, 1998.

Lizana et al. (2000). Lizana Paulin Pablo R., Andrade García Armando, Beristáin Sergio, "Desarrollo de un traductor de texto en español a lenguaje oral aplicando la síntesis de voz por sílabas". HAVOL 2000, 1er. Taller internacional de tratamiento del habla, procesamiento de voz y el lenguaje, 2000.

Martínez (2002). Martínez Licon Fabiola Margarita, "Análisis y evaluación de sílabas en español para su utilización en sistemas de reconocimiento automático



del habla". Informe para examen predoctoral, Universidad Autónoma Metropolitana, UAM-Iztapalapa, 2002.

Massaro (1972). Massaro Dominic W.. "Perceptual Images, processing time and perceptual units in auditory perception". *Psychological review*, 79(2):124-145, 1972.

Mehler (1981). Mehler Jacques, Yves Dommergues Jeans, and Frauenfelder Uli. "The syllable's role in speech segmentation". *Journal of Verbal Learning and Verbal Behavior*, 20:298-305, 1981.

Mermelstein (1975). Mermelstein Paul "Automatic Segmentation of Speech into Syllabic Units". Haskins Laboratories, New Haven, Connecticut 06510, pp. 880-883,58 (4), June 1975.

Meneido et al. (1999). Meneido Hugo, Neto João P. and Almeida Luís B., INESC-IST, "Syllable Onset Detection Applied to the Portuguese Language". Sixth European Conference on Speech Communication and Technology (EUROSPEECH'99) Budapest, Hungary, September 5-9, 1999.

Meneido and Neto (2000). Meneido H., Neto J., "Combination of Acoustic Models in Continuous Speech Recognition Hybrid Systems". INESC, Rua Alves Redol, 9, 1000-029 Lisbon, Portugal, 2000.

Montero (1999). Montero J., Gutiérrez Arreola J., Colás J., et al. "Development of an emotional Speech Synthesiser in Spanish". In *Proceedings of Eurospeech'99*, pp. 2099-2102, Budapest 1999.

Munive et al. ( ). Munive N., Vargas A., Serridge B., Cervantes O., Kirschnning I. "Entrenamiento de un Reconocedor Fonético de Dígitos para el español mexicano usando el CSLU Toolkit". *Revista de Computación y Sistemas* Vol.3 Num. 2.

Oropeza (2000). Oropeza Rodríguez José Luis, "Reconocimiento de Comandos Verbales usando HMM". Tesis de maestría, Centro de Investigación en Computación, Noviembre 2000.

Peskin (1997). Peskin Barbara, Gillick Larry, Liberman Natalie, Newman Mike, van Mulbregt Paul and Wegmann Steven. "Progress in recognizing conversational telephone speech". In *proceedings of ICASSP'97*, volume 3, pages 1811-1814, Munich, Germany, April 1997, IEEE.

Peskin et al. (1991). Peskin Barbara, Gillick Larry, Liberman Natalie, "Progress in Recognizing Conversational Telephone Speech". *ICASSP*, Vol. 3, pp. 1811-1814, Toronto Canada, Mayo 1991. IEEE.

Rabiner (1989). Rabiner L., "A Tutorial On Hidden Markov Models And Selected Applications in Speech Recognition", Proceedings of IEEE, Vol. 77, no. 2, pp. 257-286, 1989.

Rabiner and Levinson (1990). Rabiner L. R. and Levinson S. E., "Isolated and connected word recognition-theory and selected applications". Readings in Speech Recognition, edited by A. Waibel, K. Lee, Morgan Kaufman Publishers, U.S.A., 1990, pp. 115-153

Rabiner and Biing-Hwang (1993). Lawrence Rabiner and Biing-Hwang Juang, "Fundamentals of Speech Recognition", Prentice Hall, 1993.

Resch (2001a). Resch Barbara. "Gaussian Statistics and Unsupervised Learning". A tutorial for the Course Computational Intelligence Signal Processing and Speech Communication Laboratory. [www.igi.turgaz.at/lehre/CI](http://www.igi.turgaz.at/lehre/CI), November 15, 2001.

Resch (2001b). Resch Barbara. "Hidden Markov Models". A Tutorial for the Course Computational Laboratory. Signal Processing and Speech Communication Laboratory. [www.igi.turgaz.at/lehre/CI](http://www.igi.turgaz.at/lehre/CI), November 15, 2001.

Russell and Norvig (1996). Russell Stuart and Norvig Peter, Inteligencia Artificial un enfoque moderno, Prentice Hall, 1996.

Savage (1995). Savage Carmona Jesus, "A Hybrid Systems with Symbolic AI and Statistical Methods for Speech Recognition". PhD Thesis, University of Washington, 1995.

Suárez (2005). Suárez Guerra Sergio, ¿100% de reconocimiento de voz?. Trabajo inédito, no publicado.

Suk and Flanagan (1999). Dong-Suk Yuk and James Flanagan, "Telephone Speech Recognition using Neural Networks and Hidden Markov Models". IEEE Conference on Acoustics, Speech and Signal Processing, Phoenix Arizona, pp. 157-160, 15-19 May, 1999.

Sydral et al. (1995). Sydral A., Bennet R., Greenspan S., "Applied Speech Technology", Eds (1995). CRC Press, ISBN 0-8493-9456-2, U.S.A., 1995.

Tebelskis (1995). Tebelskis Joe, "Speech Recognition using Neural Networks", PhD Thesis, School of Computer Science, Carnegie Mellon University, 1995.

Tu and Loizou. (1999). Tu Zhemin and Loizou Philipos C., "Speech Recognition Over The Internet Using Java", University of Arkansas at Little Rock, proceedings of ICASSP'99.

Villing et al. (2004). Villing Rudi, Timoney Joseph, Ward Thomas and Costello John, "Automatic blind syllable segmentation for continuous speech". ISSC 2004, Belfast, June 30-July 2, Department of electronic engineering NUI Maynooth, Maynooth Co. Kildare

Weber (2000). Weber K., "Multiple Timescale Feature Combination Towards Robust Speech Recognition". Konferenz zur Verarbeitung natürlicher Sprache KOVENS2000, Ilmenau, Alemania, 2000.

Wu (1998). Wu, S., "Incorporating information from syllable-length time scales into automatic speech recognition", PhD Thesis, Berkeley University, 1998.

Wu et al. (1997). Wu S., Shire M., Greenberg S., Morgan N., "Integrating Syllable Boundary Information into Automatic Speech Recognition ". ICASSP-97, Vol. 1, Munich Germany, vol.2 pp. 987-990, 1997.

Young (1995). Young Steve "Large Vocabulary in Speech Recognition: A review". In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, pages 3-28, Snowbird, Utah, December 1995, IEEE.

[63] Libro de Ricardo Barrón

Zhang (1999). Zhang Jialu, "On the syllable structures of Chinese relating to speech recognition", Institute of Acoustics, Academia Sinica Beijing, China, 1999.

## PUBLICACIONES

- 1 Sergio Suárez Guerra, Ricardo Barrón Fernández, José Luis Oropeza Rodríguez. Informe Técnico “Reconocimiento de Comandos Verbales utilizando Cuantización Vectorial y Redes Neuronales”, ISBN 970-18-2673-6 C.I.C.-I.P.N., México, serie ROJA No. 40, Marzo de 1999.
- 2 Sergio Suárez Guerra, Ricardo Barrón Fernández, José Luis Oropeza Rodríguez. Informe Técnico “Desarrollo de Herramientas para el procesamiento Digital de Señales. Aplicaciones al Reconocimiento de Voz en Tiempo Real”, ISBN 970-18-4211-1 C.I.C.-I.P.N., México, serie ROJA No. 75, Enero del 2000.
- 3 Humberto Sossa Azuela, Ricardo Barrón Fernández, José Luis Oropeza Rodríguez “Real-valued Pattern Recall by Associative Memory”, ENC 2004 Colima.
- 4 Humberto Sossa Azuela, Ricardo Barrón Fernández, José Luis Oropeza Rodríguez “Real-valued Pattern Recall by Associative Memory, design and evolution”, IBERAMIA 2004, Puebla.
- 5 José Luis Oropeza Rodríguez, Sergio Suárez Guerra, “Continuous Speech Recognition Using Loudness and RO parameter segmentation in syllabic units”, CIC 2004, ISBN
- 6 Sergio Suárez Guerra, Ricardo Barrón Fernández, José Luis Oropeza Rodríguez “Informe Técnico” comité editorial CIC-IPN. Título del artículo: “Reconocimiento de voz por computadora” Fecha de publicación: 2000.  
ISBN:
- 7 Sergio Suárez Guerra, Ricardo Barrón Fernández, José Luis Oropeza Rodríguez. “Informe Técnico” comité editorial CIC-IPN. Título del artículo: “Reconocimiento de dígitos por computadora”. Fecha de publicación: 2000.  
ISBN:
- 8 José Luis Oropeza Rodríguez, “Un procedimiento de segmentación automática basado en elementos subbanda para sílabas de un vocabulario pequeño del español mexicano”, CIC 2002.
- 9 José Luis Oropeza Rodríguez, “Reconocimiento de voz usando Cadenas Ocultas de Markov”, Taller Internacional de Tratamiento del habla, procesamiento de voz y el lenguaje 2000.
- 10 José Luis Oropeza Rodríguez, “Una plataforma para el desarrollo de aplicaciones de voz”, CIC 2000.

11 José Luis Oropeza Rodríguez, “Reconocimiento de voz, codificación predictiva lineal y Cadenas Ocultas de Markov”, Expo Condel 2000.

12 José Luis Oropeza Rodríguez et al, “Pruebas y validación de un sistema de reconocimiento del habla basado en sílabas con un vocabulario pequeño”, CIC2003

13 Báez Hernández Orlando, José Luis Oropeza Rodríguez, “Transmisión de datos auditivos para su reconocimiento por la red eléctrica”, RVP-AI/2005, IEEE, Sección México. Aplicaciones Industriales y Exposición Industrial.

14 Sergio Suárez Guerra, José Luis Oropeza Rodríguez, Edgardo Felipe Riverón, “Speech Recognition Using Energy Parameters to Classify Syllables in the Spanish Language”, LNCS, XCIARP 2005, La Habana Cuba.

#### EN PROCESO

15 Sergio Suárez Guerra, José Luis Oropeza Rodríguez, Edgardo Felipe Riverón, Jesús Figueroa Nazuno, “Speech Recognition Using Syllabic Units in the Spanish Language”, Pattern Recognition Letter, en proceso de arbitraje.