Computación y Sistemas Vol. 9 Núm. 3, pp. 270-286

Algoritmos y Métodos para el Reconocimiento de Voz en Español Mediante Sílabas

Algorithms and Methods for the Automatic Speech Recognition in Spanish Language using Syllables

Graduated: José Luis Oropeza Rodríguez
Centro de Investigación en Computación-IPN
Av. Juan de Dios Bátiz s/n esq. Miguel Othón Mendizábal C. P. 07738 México D. F.
j_orope@yahoo.com.mx
Graduado en diciembre 15, 2006

Advisor: Sergio Suárez Guerra
Centro de Investigación en Computación-IPN
Av. Juan de Dios Bátiz s/n esq. Miguel Othón Mendizábal C. P. 07738 México D. F.
ssuare@cic,ipn,mx

Resumen

Actualmente el uso de los fonemas tiene implícita varias dificultades debido a que la identificación de las fronteras entre ellos por lo regular es difícil de encontrar en representaciones acústicas de voz. 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. 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. 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.
Palabras clave: Reconocimiento de voz, reconocimiento de sílabas, sistemas expertos, procesamiento de voz.

Abstract

This work 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.
Keywords: Speech recognition, Syllables recognition, Expert System, Speech processing.

I Introducción

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. Un Sistema de Reconocimiento Automático del Habla (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. 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.
- Sistema de reconocimiento híbrido, etc.

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.

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 {*pla}* 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 en la elaboración del presente trabajo.

El presente trabajo se encuentra dividido en 6 partes, los cuales tienen la siguiente estructura:

a) La parte 2, presenta el conjunto de historia y antecedentes de los sistemas de reconocimiento de voz.
b) La parte 3, presenta las metodologías de evaluación.
c) En la sección 4, se muestran los resultados alcanzados con la metodología propuesta.
d) La sección 5 presenta las conclusiones y la 6, referencias bibliográficas.

2 Reconocimiento de voz por computadora

2.1 Historia

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 tener en cuenta. 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 Antecedentes

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 compararon 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.

2.3 Generalidades

El reconocimiento de voz por computadora es una tarea compleja de reconocimiento de patrones y de los sistemas biométricos. Por lo regular, la señal de voz se muestrea en un rango entre los 8 y 16 KHz. En lo reportado de los experimentos de este trabajo, la frecuencia de muestreo 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.
- Codificación Predictiva Lineal.
- Análisis de los coeficientes Cepstrales.
- Predicción Lineal Perceptiva.

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
- SISTEMAS DEPENDIENTES E INDEPENDIENTES DEL LOCUTOR
- VOZ AISLADA, DISCONTINUA Y CONTINUA.
- VOZ APLICADA A TAREAS O EN GENERAL
- VOZ LEÍDA O ESPONTÁNEA
- CONDICIONES ADVERSAS

3 Metodologías de evaluación

3.1 Las reglas de la sílaba

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. 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 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. 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 las mismas.

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 1, no puede ser separado al dividir una palabra en sílabas.

br, bl, cr, cl, dr, fr, fl, gr, gl, kr, ll, pr, pl, tr, rr, ch

Fig. 1. Representación 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 forma 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.

REGLA 8. La h entre dos vocales, no destruye un diptongo.

REGLA 9. La acentuación sobre la vocal cerrada de un diptongo provoca su destrucción.

REGLA 10. La unión de tres vocales forma un triptongo.

Los SARH actuales son implementados usando el fonema como parte fundamental, gran parte de ellos se encuentran basados en los Hidden Harkov Models (HMM's) que se concatenan como se muestran en la figura 2 (Savage, 1995) para obtener palabras ó frases:

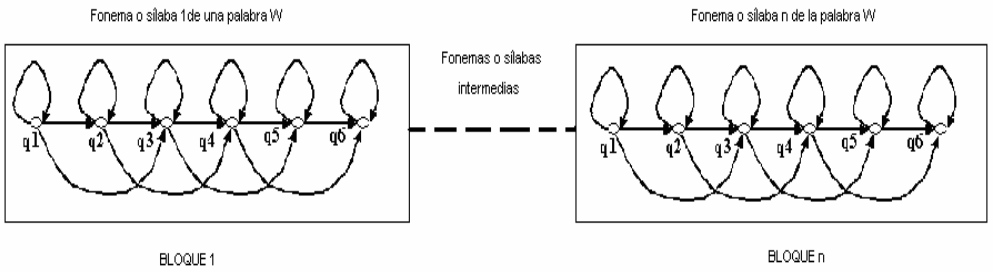

Fig. 2. Esquema de representación de Cadenas Ocultas de Markov concatenadas para conformar palabras o frases continúas.

3.2 Los tres problemas básicos de las Cadenas Ocultas de Markov (COM)

Por conveniencia, usamos la notación compacta:

$$\lambda = (A, B, \pi) \tag{1}$$

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

LOS TRES PROBLEMAS BÁSICOS DE LAS COM

Dados los datos de la HMM anterior, 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, ..., O_T), y un modelo λ = (A, B, π), como podemos calcular eficientemente P(O $|$ λ), la probabilidad de la secuencia de observación producida por el modelo.

Problema 2

Dada la secuencia de observación $O = (O_1, O_2, ..., O_T)$, y el modelo λ, como seleccionamos una secuencia de estados correspondiente $Q = (q_1, q_2, ..., q_T)$ que sea óptima en algún sentido.

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 el 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
Procedimiento hacia adelante

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

$$\alpha_t(i) = P(O_1 O_2 ... O_t, q_t = S_i | \lambda) \quad \textbf{(2)}$$

Esto es, la probabilidad de la secuencia de observación parcial, $O_1O_2...O_t$, (mientras el tiempo *t*) y analizada en el estado i, dado el modelo λ. En (Zhang, 1999) se menciona que podemos resolver $\alpha_t(i)$ inductivamente, como se muestra en las ecuaciones siguientes:

1. Inicialización

$$\alpha_1(i) = \pi_i b_i(O_1) \qquad 1 \le i \le N \quad \textbf{(3)}$$

2. Inducción

$$\alpha_{t+1}(j) = b_j(O_{t+1}) \sum_{i=1}^{N} \alpha_t(i) a_{ij} \qquad 1 \le j \le N, \quad 1 \le t \le T\text{-}1 \quad \textbf{(4)}$$

3. Terminación

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i) \quad \textbf{(5)}$$

Solución al problema 2

El *algoritmo de Viterbi* encuentra la mejor de las secuencias de estados, *Q={q1, q2, ..., qT}*, para una secuencia de observaciones dada $O=\{O_1, O_2, ..., O_T\}$, como se muestra en las ecuaciones siguientes.

1. Inicialización

$$\delta_1(i) = \pi_i b_i(O_1) \qquad 1 \le i \le N \quad \textbf{(6)}$$

2. Recursión

$$\delta_{t+1}(j) = b_j(O_{t+1}) \left[\max_{1 \le i \le N} \delta_t(i) a_{ij} \right] \qquad 1 \le j \le N, \quad 1 \le t \le T\text{-}1 \quad \textbf{(7)}$$

3. Terminación

$$p^* = \max[\delta_T(i)] \qquad 1 \le i \le N \quad \textbf{(8)}$$

$$q^* = \arg\max[\delta_T(i)] \qquad 1 \le i \le N \quad \textbf{(9)}$$

Solución al problema 3

Para poder dar solución al problema 3, se hace uso del *algoritmo de Baum-Welch*, que al igual que los anteriores, 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. 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.

$$\varepsilon_t(i, j) = \frac{\alpha_t(i) a_{ij} \beta_{t+1}(j) b_j(o_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} \beta_{t+1}(j) b_j(o_{t+1})} \quad \textbf{(10)}$$

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):

$$a_{ij} = \frac{\text{número esperado de transiciones del estado } s_i \text{ al estado } s_j}{\text{número esperado de transiciones del estado } s_i} \quad \textbf{(11)}$$

$$b_j(k) = \frac{\text{número esperado de veces que estando en j aparece el símbolo } v_k}{\text{número esperado de veces que se analiza el estado j.}} \quad \textbf{(12)}$$

3.3 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 k 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:

$$g(\mu,\Sigma)(x) = \frac{1}{\sqrt{2\pi^d}\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \quad \textbf{(13)}$$

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

$$gm(x) = \sum_{k=1}^{K} w_k * g(\mu_k, \Sigma_k)(x) \quad \textbf{(14)}$$

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

$$\sum_{i=1}^{K} w_i = 1 \qquad \forall \quad i \in \{1,\ldots\ldots,K\} \quad : w_i \geq 0 \quad \textbf{(15)}$$

En la figura 3 se muestra un ejemplo de una mixtura Gaussiana, que consiste de tres Gaussianas sencillas:

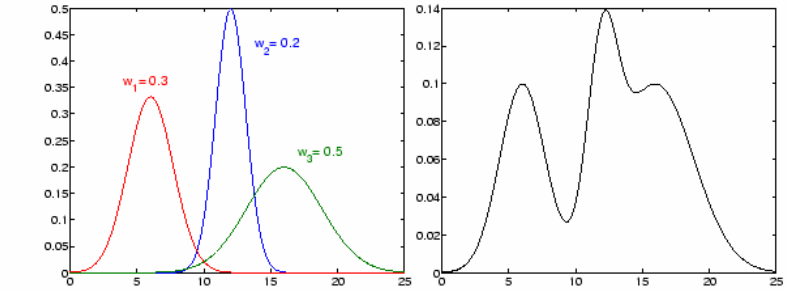

Fig. 3. Representación esquemática de una tripleta de mixturas Gaussianas

Al variar el número de Gaussianas *K*, los pesos w_i, y los parámetros de cada de las funciones de densidad μ y Σ, las mixturas 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 μ y la matriz de covarianza $\sum$.

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*) es utilizado.

4 Implantación del Sistema Basado en Conocimientos

En nuestro caso la base de conocimientos se encuentra constituida por todas las reglas de clasificación de sílabas del lenguaje español, 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 de la figura 4, el cual demuestra la finalidad y función del mismo:

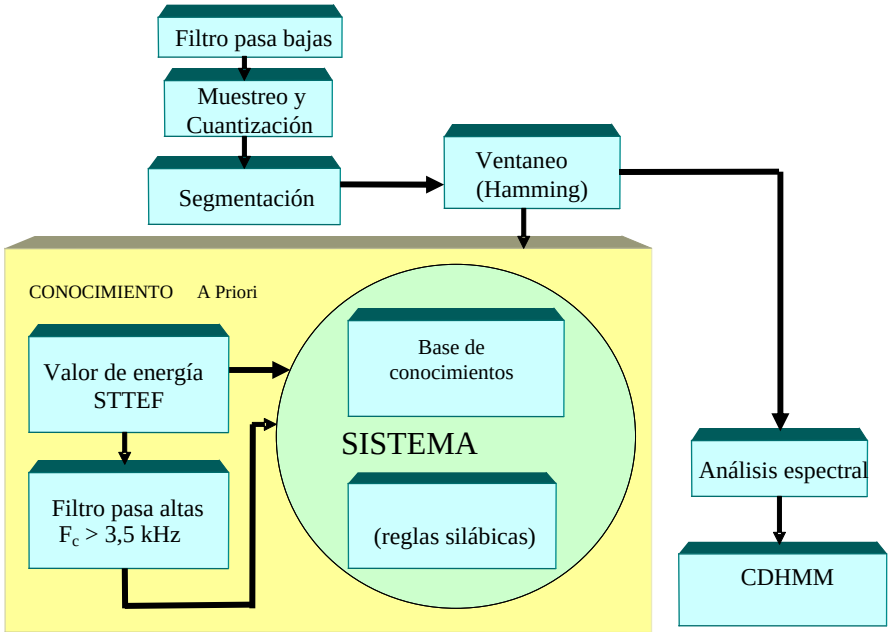

Fig. 4. Representación esquemática de la etapa de entrenamiento propuesta haciendo uso del Sistema Basado en Conocimiento.

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 habérsele 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 SRAH 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.

Los resultados obtenidos en la división silábica al hacer uso de este sistema sobre sílabas independientes, frases de distintos corpus y textos escritos fue demasiado óptima Uno de los puntos importantes al hacer uso de sílabas es que existe gran preponderancia de los grupos V, CV, VC, CCV, CVC, sobre otras representaciones (VVV, CVVC, CVVVC, 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 mostrados en la tabla 1. El proceso sigue la lógica de los siguientes árboles de inferencia de la figura 5:

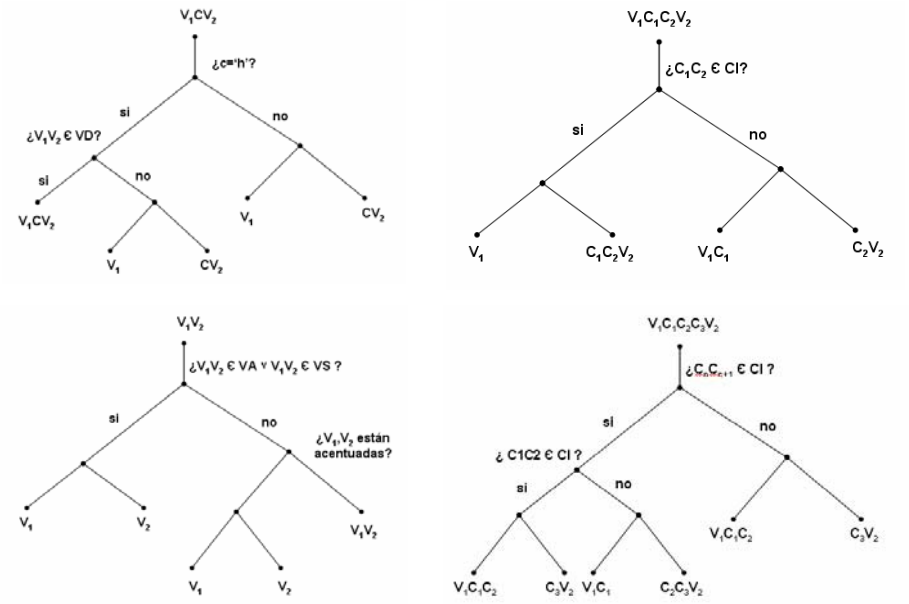

Fig. 5. Árboles de inferencia inmersos dentro del Sistema Basado en Conocimiento.

Tabla 1. Porcentajes de reconocimiento para el caso de sílabas aisladas.

	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%

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.

4.1 Efecto del parámetro ERO en una señal de voz

Para poder incrementar la taza de reconocimiento tanto a corpus de dígitos analizados y al corpus que se empleó al final 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 [7] y es una modificación a RO lo que representa una contribución del presente trabajo:

$$ERO = \sum_{i=0}^{N-1} ROi^2 \qquad \textbf{(16)}$$

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 bando 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 muestra la forma en la cual, una vez aplicado al filtro a la señal original, se comporta la gráfica de energía de dicha señal. Observe en la figura 6 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.

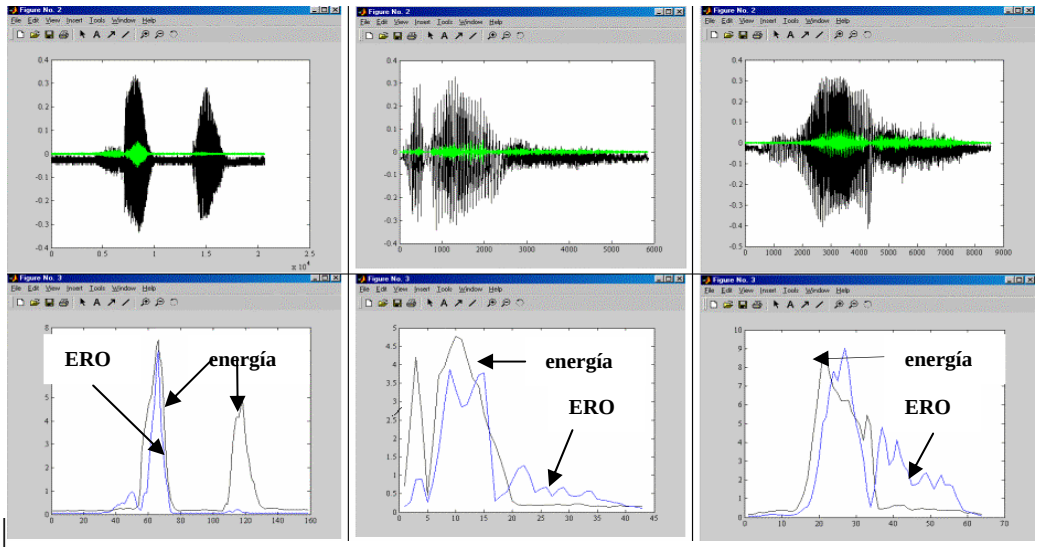

Fig. 6. Representación gráfica de la inmersión del parámetro ERO y la energía en una señal de voz.

Dado lo anterior, se procedió a crear una nueva forma de realizar la segmentación automática usando el parámetro ERO como artífice 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 7.
- 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.

4.2 La región de transición energía-parámetro ERO

Con los puntos anteriores se obtiene una segmentación que toma la representación numérica y esquemática de la siguiente figura 7. 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.

Al hacer uso de los aspectos antes mencionados sobre un corpus de dígitos se reporta un reconocimiento de sílaba individual del 94.70%, para el corpus de voz del habla continua. Mientras que las figuras anteriores muestran la comparación entre el reconocimiento hecho con palabras completas y con la concatenación de las sílabas, que resulta más eficiente (91% y 96% respectivamente). 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.

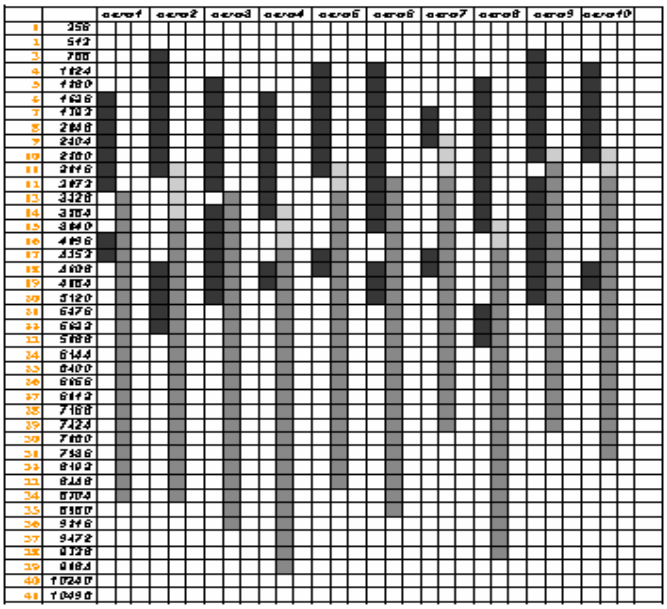

Fig. 7. Regiones de manifestación de la energía, parámetro ERO y región de transición energía-parámetro ERO.

Con el fin de extender la aplicación anterior a un corpus de sílabas más extenso. Se usaron las siguientes frases para el sistema de reconocimiento, la primera tabla muestra un corpus que carece de elementos que se denominan de confusión lingüística y la segunda es un corpus con un alto índice de confusión lingüística:

Tabla 2. Corpus de voz sin confusión lingüística.

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**

Tabla 3. Corpus de voz con confusión lingüística.

1	**A la mejor ésta es el ala**
2	**El ala del ave está mejor**
3	**Alá es el mejor del mejor**
4	**A la mejor no está**
5	**Es mejor Alá**
6	**Es Alá el mejor**
7	**El ala no la cubre Alá**
8	**A la mejor, Alá está allá**
9	**El ala de Alá está allá**
10	**A la mejor es el ala de Alá**

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":

Tabla 4. Porcentajes de reconocimiento obtenidos para el corpus de voz 1 usando habla discontinua.

Segmentación	**Modelos**	**de Harkov**
	3 estados	**5 estados**
energía	89.5%	95.5%
ERO	95%	97.5%

Tabla 5. Porcentajes de reconocimiento para el corpus de voz 1 usando habla continua.

Segmentación	**Modelos**	**de Markov**
	3 estados	**5 estados**
energía	77.5%	75.5%
ERO	79%	80.5%

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:

Tabla 4. Porcentajes de reconocimiento obtenidos para el corpus de voz 2 usando habla discontinua.

Segmentación	**Modelos**	**de Markov**
	3 estados	**5 estados**
energía	85.5%	86.3%
ERO	90%	90.8%

Tabla 5. Porcentajes de reconocimiento para el corpus de voz 2 usando habla continua.

Segmentación	**Modelos**	**de Markov**
	3 estados	**5 estados**
energía	65.5%	64.5%
ERO	72%	74.2%

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 (Tono fundamental), la amplitud y filtros adaptivos pueden ser herramientas utilizadas para intentar resolver este problema.

5 Conclusión

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 del 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.

A manera de resumen se tiene que:

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, 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.

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. 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. Hay idiomas donde la sílaba es una muy buena alternativa, en otros no, tal es el caso del idioma Español.

6 Referencias

1. 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.
2. 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.
3. Giarratano and Riley (2001). Giarratano Joseph y Riley Gary, International Thompson Editores, Sistemas expertos, principios y programación 2001.
4. Hauenstein (1996). Hauenstein A., “The syllable Re-revisited”, Technical Report, Siemens AG, Corporate Research and Development, München Alemania, 1996.
5. 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
6. 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.
7. 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.
8. Kirschning (1998). Kirschning Albers Ingrid, "Automatic Speech Recognition with the Parallel Cascade Neural Network", PhD Thesis, Tokyo Japan, March 1998.
9. Kosko (1992). Kosko B., "Neural Networks for Signal Processing", Prentice Hall, U.S.A., 1992.

10. Meneido et al. (1999). Meneido Hugo, Neto Joâo 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, Hunagry, September 5-9, 1999.
11. 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.
12. Mermelstein (1975). Mermelstein Paul "Automatic Segmentation of Speech into Syllabic Units". Haskins Laboratories, New Haven, Connecticut 06510, pp. 880-883,58 (4), June 1975.
13. Oropeza (2000). Oropeza Rodríguez José Luis, "Reconocimiento de Comandos Verbales usando HMM". Tesis de maestría, Centro de Investigación en Computación, Noviembre 2000.
14. Rabiner and Biing-Hwang (1993). Lawrence Rabiner and Biing-Hwang Juang, "Fundamentals of Speech Recognition", Prentice Hall, 1993.
15. 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, November 15, 2001.
16. 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, November 15, 2001.
17. Russell and Norvig (1996). Russell Stuart and Norvig Peter, Inteligencia Artificial un enfoque moderno, Prentice Hall, 1996.
18. Savage (1995). Savage Carmona Jesus, "A Hybrid Systems with Symbolic AI and Statistical Methods for Speech Recognition". PhD Thesis, University of Washington, 1995.
19. Suárez (2005). Suárez Guerra Sergio, ¿100% de reconocimiento de voz?. Trabajo inédito, no publicado.
20. 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.
21. Weber (2000). Weber K., "Multiple Timescale Feature Combination Towards Robust Speech Recognition". Konferenz zur Verarbeitung natürlicher Sprache KOVENS2000, Ilmenau, Alemania, 2000.
22. Wu (1998). Wu, S., "Incorporating information from syllable-length time scales into automatic speech recognition", PhD Thesis, Berkeley University, 1998.
23. 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.
24. Zhang (1999). Zhang Jialu, "On the syllable structures of Chinese relating to speech recognition", Institute of Acoustics, Academia Sinica Beijing, China, 1999.

Jose Luis Oropeza Rodriguez. Graduated in Telecommunication Engineering in ESIME CU (Culhuacan Unit), Mexico City, 1994. M. Sc. in Computer Engineering in the Computing Research Center, National Polytechnic Institute, Mexico City, since 2000. Area of interest: Speech Recognition.

Sergio Suarez Guerra. Graduated in Electrical Engineering in the Superior Polytechnic Institute José Antonio Echeverria, Havana, Cuba, in 1972. PhD developed in the Academic of Sciences of the Soviet Union, Moscow, 1979. Since 1998 he is Professor and Senior Researcher in the National Polytechnic Institute, Mexico. Areas of interest: Digital Signal Processing, Speech Processing and Speech Recognition.