Reconocimiento de Voz en Español Mediante Sílabas

Dr. Sergio Suarez Guerra
Profesor del CIC-IPN
Dr. en C. José Luis Oropeza Rodríguez
Profesor del CIDETEC-IPN

Actualmente, el uso de los fonemas tiene implícitas 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 tiempo, analizando la forma en la cual el paradigma de la sílaba responde a tal labor dentro del español. Durante los experimentos realizados se examinaron 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 como la Función de Energía Total en Corto Tiempo se usaron en un corpus de dígitos en donde los resultados alcanzados usando sólo la Función de Energía, 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 del 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 emplearon Cadenas Ocultas de Markov de densidad continua con tres y cinco estados, con 3 mixturas Gaussianas por estado.

Introducción

Un Sistema de Reconocimiento Automático del Habla (SRAH) es aquel sistema automático capaz de gestionar la señal de voz emitida por un individuo. Dicha señal ha pasado 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. El entrenamiento es una de las etapas críticas dentro de estos sistemas, y gran parte del éxito de un sistema de reconocimiento de voz recae en esta parte. 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, usando 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 del idioma 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 en la que se lee son muy semejantes; esto se debe 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, entre los que 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 origen a una gran cantidad de resultados e implementación de 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 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 entre la forma en que se pronuncia y la que se escribe, puede establecerse como elemento primordial de un SRAH.

c) La separación automática de estructuras fonéticas sigue siendo 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 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 de este trabajo.

Este artículo se encuentra dividido en 6 partes, las 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 en la 6 las referencias bibliográficas.

Reconocimiento de voz por computadora

Historia

Se considera que el reconocimiento de voz por computadora es una tarea muy compleja, debido a todos sus requerimientos 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, la cantidad de aspectos a tener en cuenta es muy grande. 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 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.

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 el reporte de los experimentos de este trabajo, la frecuencia de muestreo utilizada fue de 11025 Hz. La señal de voz necesita ser analizada para extraer información relevante una vez que ha sido digitalizada.

A manera de resumen, dentro de esta tarea, se aplican 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, tanto dependientes como independientes del locutor
- Voz aislada, discontinua y continua.
- Voz aplicada a tareas o en general
- Voz leída o espontánea
- Condiciones adversas

Metodologías de evaluación

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 se listan a continuación mostrando además excepciones a las mismas.

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

Figura 1. Representación de consonantes inseparables.

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.

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 SRAH actuales son implementados usando el fonema como parte fundamental, gran parte de ellos se encuentran basados en los Hidden Markov Models (HMM), que se concatenan como se muestra en la **Figura 2** (Savage, 1995) para obtener palabras ó frases.

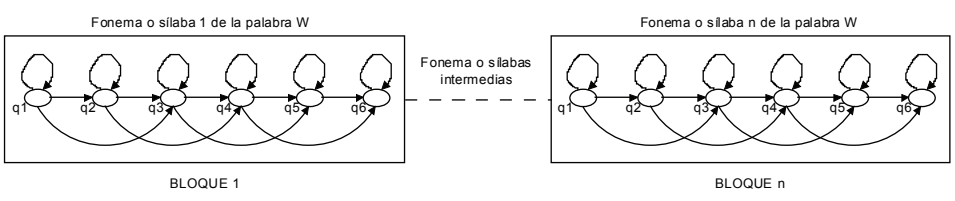

Figura 2. Esquema de representación de Cadenas Ocultas de Markov concatenadas para conformar palabras o frases continúas.

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

Por conveniencia, usamos la notación compacta:

$$\lambda = (A, B, \pi) \qquad [1]$$

para indicar el conjunto de parámetros completos del modelo. Este conjunto de parámetros, por supuesto, define 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 $\lambda = (A, B, \pi)$, ¿cómo podemos calcular eficientemente $P(O|\lambda)$, la probabilidad de la secuencia de observación producida por el modelo?

Problema 2

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

Problema 3

¿Cómo ajustar los modelos de los parámetros *l=(A, B, p)* para maximizar *P(O½l)?*. 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, se denomina secuencia de entrenamiento porque es usada para entrenar el HMM. El problema del entrenamiento es una de las aplicaciones más cruciales para el HMM, ya que 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:

$$\alpha_t(i) = P(O_1 O_2 \ldots O_t, q_t = S_i | \lambda) \qquad [2]$$

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

1. Inicialización

$$\alpha_1(i) = \pi_i b_i(O_1) \qquad 1 \le i \le N \qquad [3]$$

2. Inducción

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

3. Terminación

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i) \qquad [5]$$

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={O1, O2, ..., OT)*, como se muestra a continuación:

1. Inicialización

$$\delta_1(i) = \pi_i b_i(O_1) \qquad 1 \le i \le N \qquad [6]$$

2. Recursión

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

$$1 \le j \le N, \quad 1 \le t \le T\text{-}1 \qquad [7]$$

3. Terminación

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

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

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 cambios de los estados del modelo de Markov. Con el análisis anterior, Baum-Welch logró obtener la siguiente expresión para la implementación de su algoritmo; dicha ecuación, nos permite determinar el número de transiciones del estado s_i al estado s_j.

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

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):

$$a_{ij} = \frac{\text{número esperado de transiciones del estado } s_i \text{ al estado } s_j}{\text{número esperado de transiciones del estado } s_i} \qquad [11]$$

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

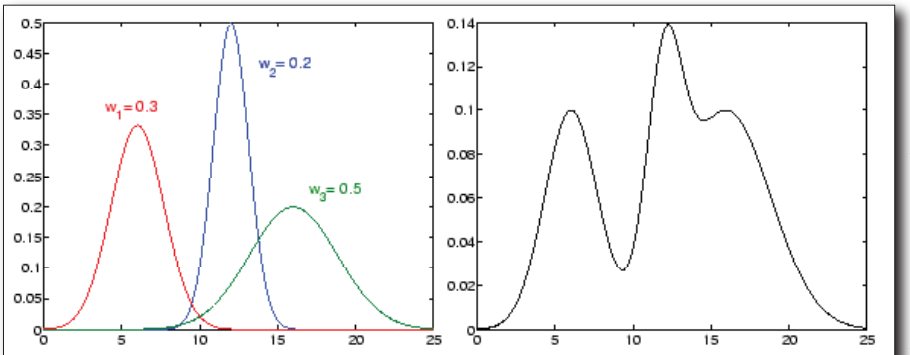

Figura 3. Representación esquemática de una tripleta de mixturas Gaussianas.

Los parámetros de la función de densidad de probabilidad (PDF, *Probability Density Function*), 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 Σ. Para encontrar estos parámetros, que de alguna forma describen a una determinada función de probabilidad de un conjunto de datos, se utiliza el algoritmo iterativo de máxima esperanza (*EM*).

Mixturas de Gaussianas

Las mixturas Gaussianas, son combinaciones de distribuciones normales o funciones de Gauss. Una mixtura de k Gaussianas puede 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 Gaussiano es de la forma:

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

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

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

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

$$\sum_{i=1}^{K} w_i = 1 \quad \forall \quad i \in \{1,.....,K\} \quad : w_i \geq 0 \qquad [15]$$

En la **Figura 3**, se muestra un ejemplo de una mixtura Gaussiana, que consiste de tres Gaussianas sencillas.

Al variar el número de Gaussianas *K*, los pesos w_i, y los parámetros de cada una de las funciones de densidad μ y Σ, las mixturas Gaussianas pueden usarse para describir algunas Funciones de Densidad de Probabilidad Complejas (FDPC).

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 del sistema.

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 aplicar las reglas pertinentes, la energía en corto tiempo de la señal y la energía en corto tiempo del parámetro RO, se procede a realizar la división en unidades silábicas de cada uno

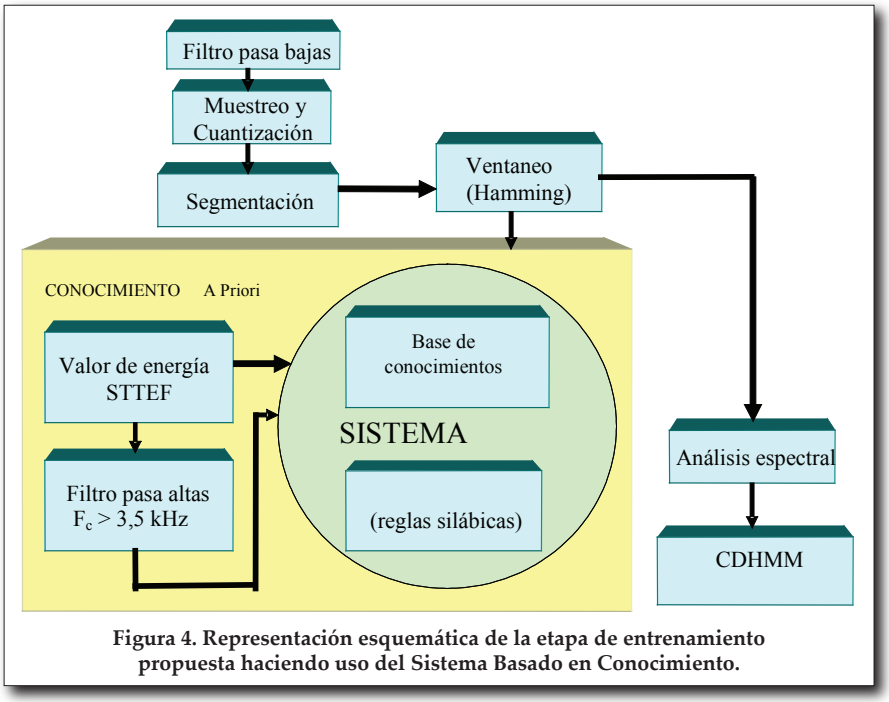

Figura 4. Representación esquemática de la etapa de entrenamiento propuesta haciendo uso del Sistema Basado en Conocimiento.

de los elementos de entrada, con lo cual se logran 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 propuesto, se puede resumir con el diagrama a bloques mostrado en la **Figura 4**, el cual demuestra la finalidad y función del mismo.

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

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, se provee la cantidad de bloques que serán usados en la etapa de etiquetado silábico de las señales de voz.

Los resultados obtenidos en la división silábica, al hacer uso de este sistema sobre sílabas independientes, frases de distintos corpus y textos escritos fue 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 creo un sistema de reconocimiento para el habla discontinua, para un determinado conjunto de sílabas, generando los resultados de reconocimiento mostrados en la **Tabla 1**. El proceso sigue la lógica de los árboles de inferencia mostrados en la **Figura 5.**

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.

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 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, pero 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 [16] y es una modificación a RO, lo que representa una contribución del presente trabajo:

$$ERO = \sum_{i=0}^{N-1} ROi^2 \qquad [16]$$

Dado lo anterior, se procedió a crear una nueva forma de segmentación automática, usando el parámetro ERO como artífice para ello.

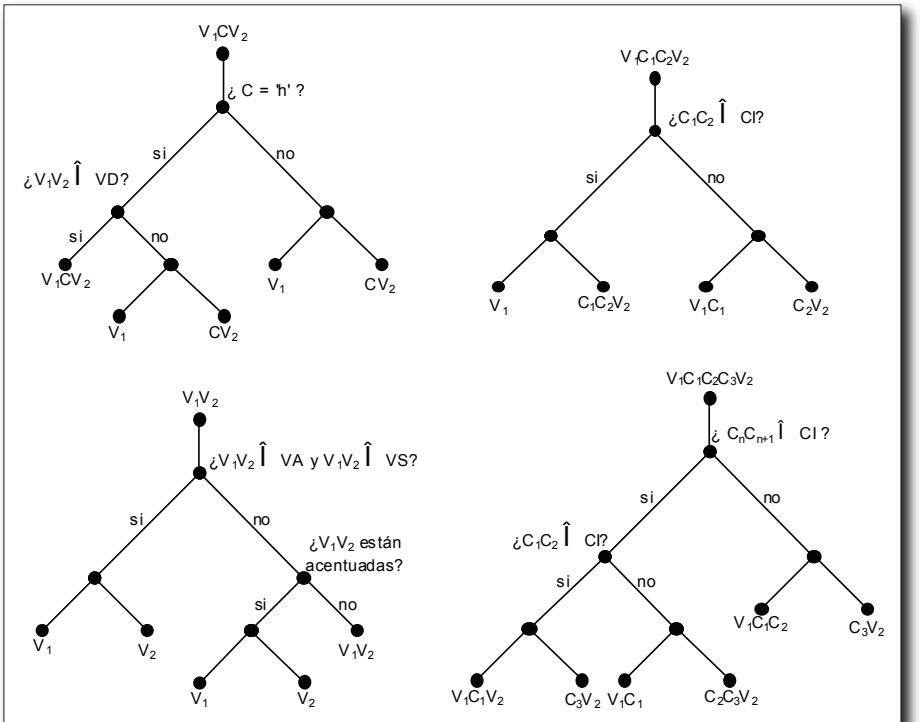

Figura 5. Árboles de inferencia inmersos dentro del Sistema Basado en Conocimiento.

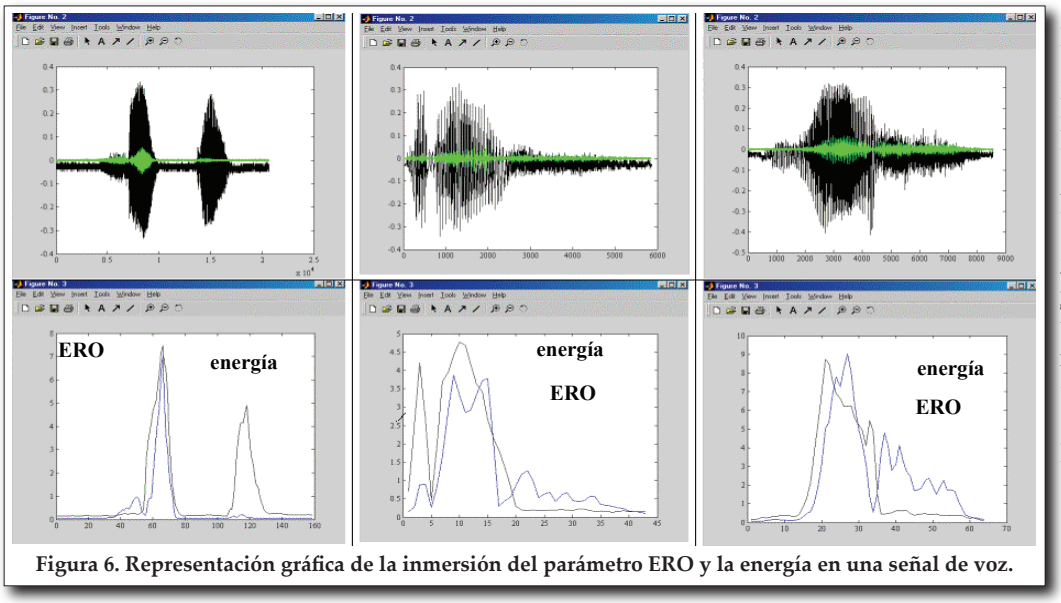

Figura 6. Representación gráfica de la inmersión del parámetro ERO y la energía en una señal de voz.

Los beneficios que conlleva este nuevo recurso radican 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.**
- 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 o 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 se realizan 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*.

La región de transición energía-parámetro ERO

Con los puntos anteriores, se obtiene una segmentación que toma la representación numérica y esquemática de la **Figura 7**. Las líneas de color negro representan las regiones de la señal de voz en donde la energía de la señal sin filtrar produce efectos; las líneas de color gris muestran las regiones en donde la señal de voz ya filtrada genera efectos, es decir, los puntos en donde las componentes de alta frecuencia se encuentran presentes; y, finalmente las líneas de color gris claro indican las *regiones de transición energía-parámetro RO*, que permiten ver las regiones en donde ninguno de los dos elementos tiene su aparición, pero que son necesarias para ligar la aparición o continuación de una sílaba.

Al usar los aspectos mencionados sobre un corpus de dígitos, se reporta un reconocimiento de sílaba individual del 94.7%, para el corpus de voz del habla continua, mientras que las figuras anteriores muestran la comparación entre reconcer con palabras completas y con la concatenación de las sílabas, la cual es más eficiente (91% y 96% respectivamente). El parámetro RO se usa para 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.

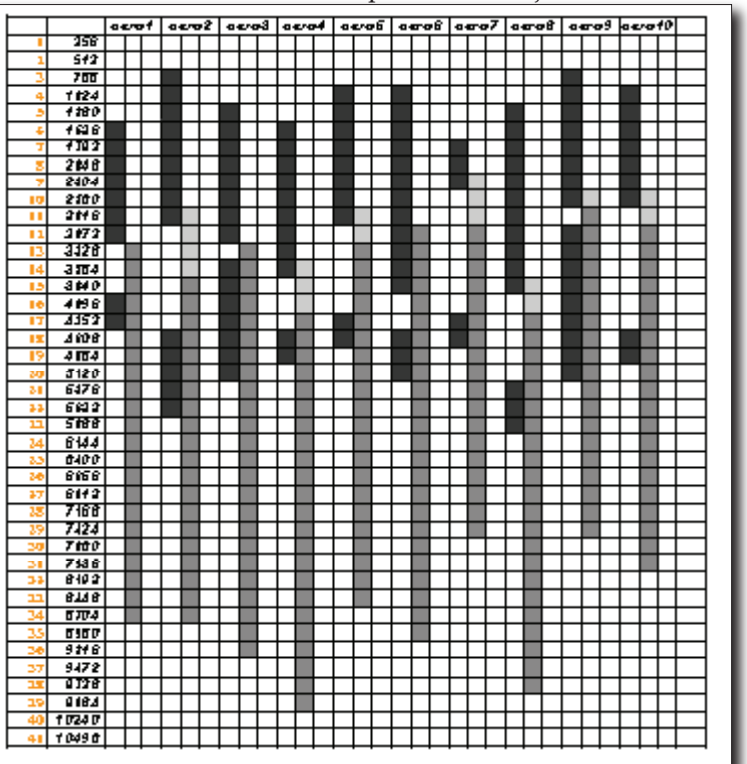

Figura 7. Regiones de manifestación de la energía, parámetro ERO y región de transición energía-parámetro ERO.

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 2. Corpus de voz sin 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

Tabla 3. Corpus de voz con confusión lingüística.

Con el fin de extender la aplicación anterior a un corpus de sílabas más extenso, se usaron las frases de las **Tabla 2** y **3** 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.

Se utilizaron Mixturas Gaussianas, 3 de ellas para cada estado, y HMM con 5 y 3 estados; se usaron 12 coeficientes CLPC como componentes de observación, generándose los Modelos independientes por sílaba y realizándose la concatenación de las mismas utilizando las probabilidades obtenidas a través del modelo bi-

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

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

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

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

gram del lenguaje, para comenzar con el entrenamiento global de la frase. Estos programas se ejecutaron sobre MATLAB, y los resultados obtenidos se muestran en la **Tabla 4** para el corpus final "1", y en la **Tabla 5** para el corpus final "2".

Se hicieron 20 repeticiones con 5 personas (3 hombres y 2 mujeres), de las cuales 50% se usaron para el entrenamiento y 50% para el 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

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

Tabla 6. Porcentajes de reconocimiento obtenidos para el corpus de voz 2 usando habla discontinua.

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

Tabla 7. Porcentajes de reconocimiento para el corpus de voz 2 usando habla continua.

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 agrego el doble de elementos. Los resultados para el corpus final "2" se muestran en las **Tablas 6 y 7**.

Para finalizar, se pretendió verificar el efecto que tiene considerar la acentuación de las palabras que conforman al corpus final "2"; el resultado generó un porcentaje de reconocimiento entre el 50 y 60%, lo cual implica que debe utilizarse un análisis de señal de voz que permita identificar la variación que provoca la acentuación. El pitch (Tono fundamental), la amplitud y filtros adaptivos pueden ser herramientas utilizadas para tratar de resolver este problema.

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; y por otro, la inmersión de técnicas de Inteligencia Artificial a los procesos de reconocimiento de voz se hace cada vez más necesaria.

2) El esquema propuesto, representa una extensión al planteado por Furui, lo cual implica un aporte importante al área de investigación del reconocimiento de voz.

3) Uno de los puntos esenciales buscados por 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. La mayor parte de las investigaciones propuestas se fundamentan en ella; más aún, los resultados obtenidos demuestran que tal aseveración es prácticamente cierta.

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 definen el espacio de búsqueda. De la misma forma, aunque este trabajo no explora los beneficios de la programación paralela, algunas de sus conclusiones 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, el uso de máquinas paralelas y concurrentes puede ser ampliamente ventajosa en la investigación del reconocimiento de voz; asimismo, la combinación de la metodología empleada en el 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 tanto, tal es el caso del idioma Español.

Referencias Bibliográficas

[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.