# Reconocedor de dígitos

*ITZIAR GORETTI ALONSO GONZÁLEZ*


## *RESUMEN*

En los últimos años, el reconocimiento de voz se ha integrado en aplicaciones para facilitar la comunicación entre hombre-máquina. Los reconocedores existentes en el mercado se caracterizan por ser simples y seguros en su tarea, pero no hay que dudar que se trata de un sistema que está en creciente evolución y desarrollo, mejorando lo existente y dando nuevos servicios a nuevas aplicaciones.
Aquí proponemos una mera introducción para el desarrollo de un reconocedor discreto y dependiente de locutor. Se ha separado en tres etapas: análisis de la voz, clasificación de patrones y, por último, la implementación del reconocedor y resultados. Todo aquel que quiera trabajar en reconocimiento de voz debería usar como cuadernos de campo: la estadística (Modelos ocultos de Markov) y procesado digital de señales.

## *ABSTRACT*

### *Digital recognition*

*In recent years, applications designed to promote communication between man and machines have included voice recognition techniques. Existing technology in this field is straightforward and performs its function satisfactorily, but this is obviously a growing field in which constant innovations offer improved services to new applications.*
*This study comprises an introduction to the development of a discreet voice recognition system, dependent on the speaker. The process is divided into three stages: voice analysis, pattern classification and, lastly, implementation of recognition and results. Anyone wanting to work in voice recognition should use the following as field notebooks: statistics (Markov's Hidden Models) and the digital processing of signals.*


## INTRODUCCIÓN

El reconocimiento de voz es fundamentalmente una tarea de clasificación de patrones. El objetivo es tomar un patrón de entrada, que en este caso es la señal de voz, y clasificarla como cero, uno, dos, etc. Los patrones de entrada, señal de voz, pueden ser tratados como palabras, sílabas o fonemas. Si estos patrones fuesen invariantes el problema sería trivial, es decir, simplemente se compararía la secuencia de patrones de entrada con patrones previamente almacenados y encontrar el más similar. Pero esto no es así, la principal dificultad del reconocimiento es que la señal de voz es muy variable, debido a la gran variedad de locutores (hombres, mujeres, niños), diferentes velocidades a la hora de hablar, ambientes, distintas condiciones acústicas, e incluso el esta-

do anímico del locutor (si está enfadado, agresivo, etc.).

El estudio del reconocimiento discreto de voz está basado en tres principios:

- El primero, que la información en la señal de voz puede ser representada por el conjunto de segmentos espectrales de la señal. Esto quiere decir que la señal de voz correspondiente a un dígito puede ser dividida (en el tiempo) en varios segmentos, extrayendo parámetros espectrales de cada uno de ellos y trabajar con estos parámetros en etapas posteriores.

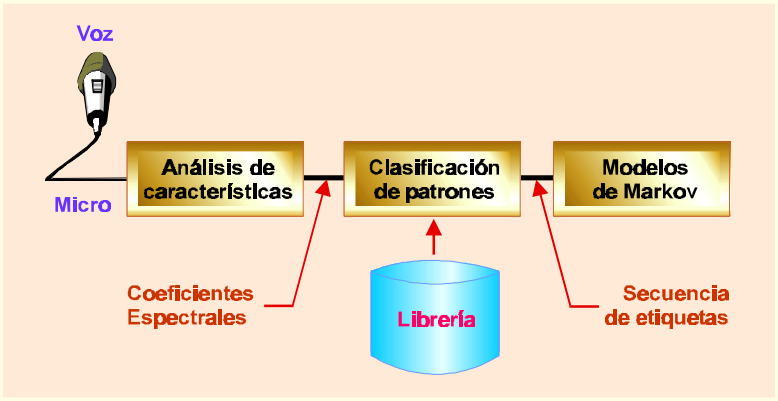
- El segundo, que el contenido de la señal de voz puede ser expresado como una secuencia de símbolos fonéticos (etiquetas). Esto se deriva del punto anterior y significa que al extraer parámetros de cada segmento, la voz quedaría representada como una secuencia de vectores-parámetros que posteriormente éstos se transformarán en etiquetas mediante la utilización de un Cuantificador Vectorial, que comentaremos más adelante.

- El tercero, que el reconocimiento es un proceso estadístico basado en los modelos ocultos de Markov.

Se podría incluir una siguiente etapa, la cual no deja de ser importante, en la que se tuviesen en cuenta la gramática, semántica y estructura del lenguaje. Esta etapa se incluiría en sistemas de reconocimiento de vocabularios más amplios y más generales. En nuestro estudio no lo hemos añadido ya que sólo reconocemos dígitos.

FIGURA 1

Diagrama de bloques del proceso de reconocimiento




En la figura 1, aparece un esquema del diagrama de bloques de un posible sistema de reconocimiento. A continuación se analizará por separado cada uno de los bloques de esta figura.

# ANÁLISIS DE LA SEÑAL DE VOZ

## Conversión A/D

El primer paso es la captura de la señal de voz. Para ello se utiliza un micrófono que convierte la señal acústica en una eléctrica, con la que podemos trabajar. La salida del micrófono es una señal analógica y lo que procede es digitalizar la señal mediante un conversor analógico digital, CAD.

Este conversor lo que hace es tomar muestras de la señal analógica cada cierto tiempo, Ts, tiempo de muestreo (ver figura 2), es decir, una señal analógica se convierte en una secuencia de valores discretos. En realidad no se habla de tiempo de muestreo sino de frecuencia de muestreo, que es 1/Ts. Es importante recordar que para muestrear una señal se ha de cumplir los criterios de Nysquist. [5]

FIGURA 2

Señal analógica muestreada cada Ts.

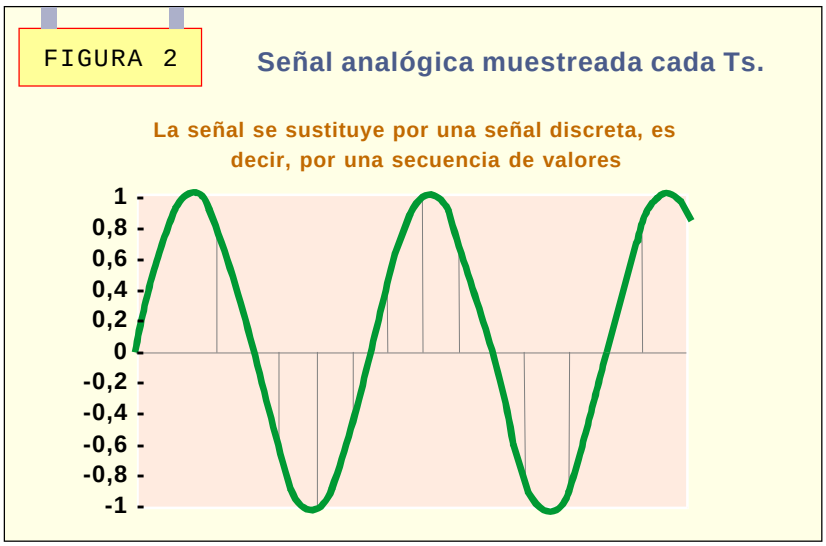
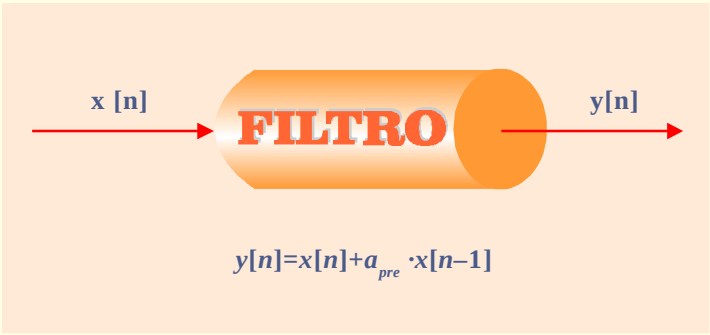


FIGURA 3

Filtro de preénfasis




A causa del limitado ancho de banda de los canales telefónicos y el uso generalizado de la frecuencia de muestreo de 8 KHz en la telefonía digital, se ha utilizado este valor como frecuencia de muestreo. Sin embargo, con la salida reciente de redes digitales de banda ancha, podremos ver pronto nuevas aplicaciones de telecomunicaciones que utilizan calidad más alta en entrada de audio, y esto supondrá utilizar una frecuencia de muestreo más alta.

Hay que tener en cuenta que tanto el micrófono utilizado así como el proceso de conversión analógico/digital introducen efectos indeseados, ruido de la línea, pérdidas a altas y bajas frecuencias. Esto hace que el sistema completo sea altamente dependiente del convertidor CAD y del micrófono.

El propósito principal del proceso de digitalización es producir una representación de los datos muestreados de la señal de voz con una relación señal a ruido lo más alta posible manteniéndose por encima de los 30 dB.

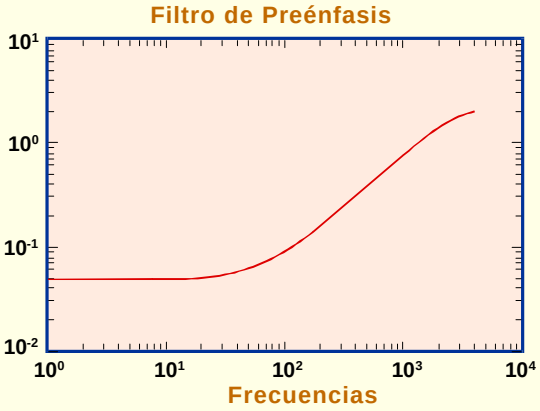
El siguiente paso una vez digitalizada la señal es el filtrado. Se utiliza un filtro digital FIR (Finite Impulse Response) cuya ecuación viene dada en la ecuación

$$H_{pre}(z)=1+a_{pre}z^{-1}$$

Se trata de un filtro de un solo coeficiente denominado filtro preénfasis. Un valor típico de **a** es 0.95. El filtro de preénfasis se destina para alzar el espectro de la señal aproximadamente 20 dB por década. Hay dos explicaciones que justifican su utilización: primero, los segmentos de voz sonoros tienen una pendiente espectral negativa (aproximadamente 20 dB por década), este filtro tiende a contrarrestar esta pendiente mejorándose la eficiencia de las etapas posteriores; y, segundo, es que la audición es más sensible por encima de 1 KHz en la región del espectro. Este filtro amplifica esta zona del espectro ayudando a las etapas posteriores de análisis a modelar los aspectos más importantes del espectro de la voz.

FIGURA 4

Respuesta en frecuencia del Filtro de Preénfasis




## Análisis Espectral

Una vez que la señal se ha digitalizado y filtrado con el filtro preénfasis, la señal de voz se segmenta en tramas de 20 ó 30 mseg. con un desplazamiento cuyo valor típico es 10 mseg. Por ejemplo, imaginemos que tenemos la señal de la figura 5. Las rayas verticales de la gráfica se corresponden con 20 mseg de tiempo. Para analizar este trozo de voz se procederá de la siguiente manera, la primera trama de voz es la indicada en la figura 5. La segunda trama no comenzará en el siguiente segmento indicado por el trazo vertical, sino que estará desplazado 10 mseg. respecto del comienzo de la trama

anterior, y así sucesivamente. En la figura 5, la punta de la flecha y su longitud indica comienzo y duración de la trama. Es por esto que se habla de duración de las tramas y desplazamiento.

Cada trama de 20 mseg. se procesa de la siguiente forma:

1. Después de la segmentación se aplica una ventana Hamming [1], la cual elimina los problemas causados por los cambios rápidos de la señal en los extremos de cada trama de voz. Es por eso por lo que se utiliza la segmentación con un desplazamiento para conseguir transiciones suaves entre tramas. En la práctica es deseable normalizar la ventana para que la potencia de la señal sea aproximadamente igual a la potencia de la señal antes del enventanado.

La teoría de la ventana fue un tema activo de investigación en el procesado digital de señal, hay muchos tipos de ventana: rectangular, Hamming, Hanning, Blackman, Bartlett, y Kaiser. Hoy en día, en reconocimiento de voz, se utiliza exclusivamente la ventana Hamming, que es un caso específico de la Hanning.

Una ventana Hamming se define como:

$$w(n) = 0.54 - 0.46 \cdot \cos\left(\frac{2\pi n}{N_s - 1}\right)$$

Ns es el número de muestras de la trama.

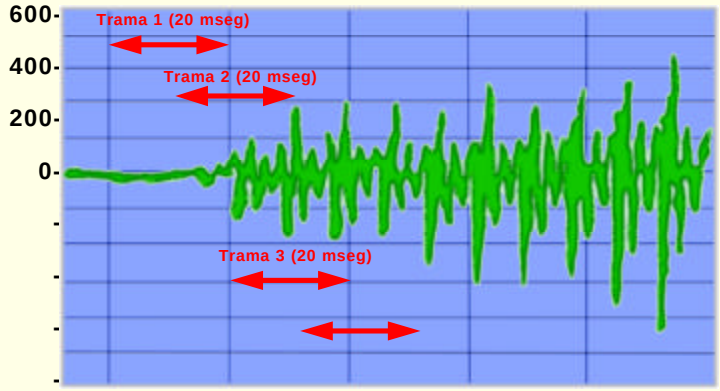
Y n está 0£n£Ns

2. De cada segmento de voz se obtienen los llamados coeficientes LPC (Linear Prediction Coefficients) [1]. Estos coeficientes LPC se convierten a otros coeficientes denominados Cepstra [1], éstos representan la transformada de Fourier de la magnitud logarítmica del espectro.

Una mejora en cuanto a representación puede ser utilizar, además, la información de la derivada temporal de los coeficientes *cepstra*, serían los llamados *delta-cepstra*. Incluso se podría utilizar la segunda derivada.

FIGURA 5

**Segmentación de la señal de voz**




CUADRO 1

**Cuadro-resumen del preprocesado de señal**

**Para concretar este apartado, conviene recordar que por cada dígito pronunciado, se deben realizar los siguientes pasos:**

- Digitalizar la señal de voz
- Segmentar en tramas de 20 mseg y desplazamiento 10 mseg. El número total de tramas obtenidas sería:

$$N_{tramas} = Tiempo_{total}(dígito) / Tiempo_{Desplazamiento}$$

- De cada trama se aplica una ventana Hamming y se calculan los coeficientes LPC, seguidamente los *cepstra* y *delta-cepstra*. El número de coeficientes *cepstra* utilizados comúnmente en reconocimiento es 13, por tanto, tenemos un total de 26 coeficientes, 13 *cepstra* y 13 *delta-cepstra*.

Todos estos coeficientes se almacenan en una matriz cuyas dimensiones son:

$$N_{trama} x N_{coeficientes}$$

# CLASIFICACIÓN DE PATRONES

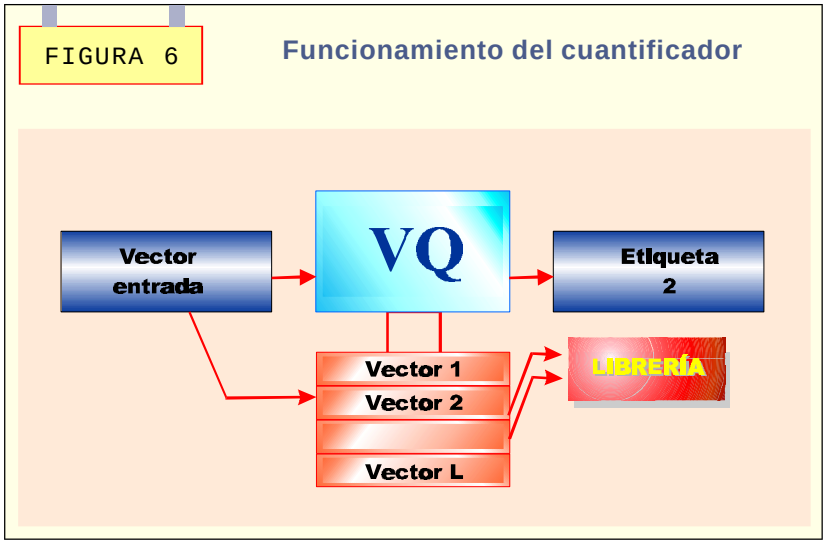
El siguiente paso en reconocimiento de voz sería el proceso de clasificación de patrones, que será estudiado en dos partes: cuantificación vectorial y los modelos ocultos de Markov, HMM (Hidden Markov Models).

A su vez tenemos que separar ambas etapas en dos fases: *entrenamiento y test.* La fase de entrenamiento consiste en tomar un conjunto de voces de referencia y a partir de éstos ajustar los parámetros para que el sistema funcione correctamente. Una vez que se ha entrenado, se pasa a la fase de test, que consiste en verificar el entrenamiento. Si es correcto, el sistema está preparado para trabajar, si no lo está se vuelve a entrenar.

## Cuantificador Vectorial

¿Qué es lo que hace el cuantificador vectorial? El cuantificador vectorial, en lo sucesivo VQ, lo que hace es simplificar la variedad de un conjunto muy amplio de vectores (en este caso de coeficientes o parámetros) de entrada a un conjunto finito de vectores almacenados en una librería. La elección de los vectores de la librería se obtiene mediante la aplicación de algoritmos como el de las K-medias o LBG, [4]. Cada vector de coeficientes se substituye bajo un criterio de medidas de distancia (ejemplo: distancia Euclidea) por un vector de la librería. Esto facilita enormemente el funcionamiento de etapas posteriores, ya que sólo trabajan con un número finito de vectores, siempre conocidos.

FIGURA 6 Funcionamiento del cuantificador




### *Entrenamiento y Test del VQ*

En la primera se diseña una librería de vectores de dimensión o longitud L (se considera que sean potencia de 2). Estos vectores son representativos del conjunto de entrenamiento. La segunda fase no es de test propiamente dicho, sino de funcionamiento. Un vector de entrada se sustituye por uno de la librería (ver figura 6).

La figura 6 intenta representar el funcionamiento del VQ, que es el siguiente: el vector de entrada es codificado por el VQ con el vector 2 de la librería. La salida del VQ es la posición que ocupa dentro de la librería, en este caso, el 2. Es decir el vector de entrada es codificado con la etiqueta 2.

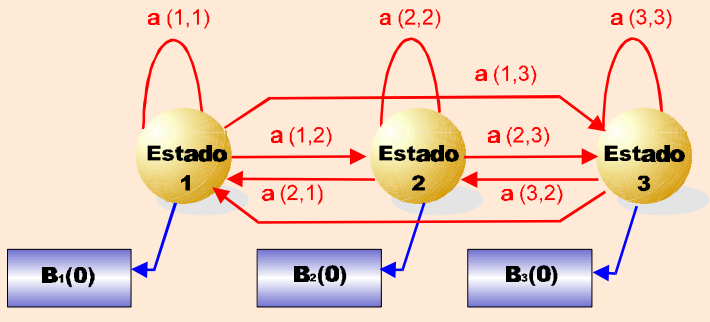
¿Qué se consigue con el VQ? Primero, reducir la información a procesar en la siguiente etapa, es decir, hemos pasado de trabajar con un vector de 26 coeficientes a una etiqueta representada por un número entero [1,L]. Segundo, se reduce además el cálculo computacional

en etapas posteriores, y tercero, tener una representación discreta de la voz.

Las desventajas de utilizar un cuantificador vectorial es la distorsión que se produce al sustituir un vector por otro. La librería elegida tiene un número finito de vectores, y el proceso de elegir la mejor representación del vector conduce a un error de cuantificación.

FIGURA 7

**Diagrama de Estados de un modelo de Markov sencillo**




## Modelos Ocultos de Markov

Los sistemas de Markov, en lo sucesivo nos referiremos a ellos como los modelos HMM, es el algoritmo actualmente más eficaz y más utilizado en reconocimiento. La principal ventaja de los HMM es que este sistema retiene información estadística acerca de los patrones de voz.

La clave del procedimiento estadístico de los HMM en reconocimiento de voz es que la voz puede ser modelada estadísticamente durante un proceso automático. Se crea un modelo estadístico por dígito.

En la formulación de los HMM [2 y 3], la voz se asume como dos procesos probabilísticos. En el primero, la voz se modela como una secuencia de transiciones de estados a(i,j) (ver figura 7); y en el segundo, en cada estado se produce un evento observable $B_i(O)$.

En la figura 7 se representa un posible modelo de Markov. Las transiciones entre estados vienen determinadas por la matriz de probabilidades de transiciones a(i,j). El índice i representa el estado inicial y j el estado final. La matriz $B_i(O)$, es la matriz de probabilidades de la secuencia de observación de cada estado.

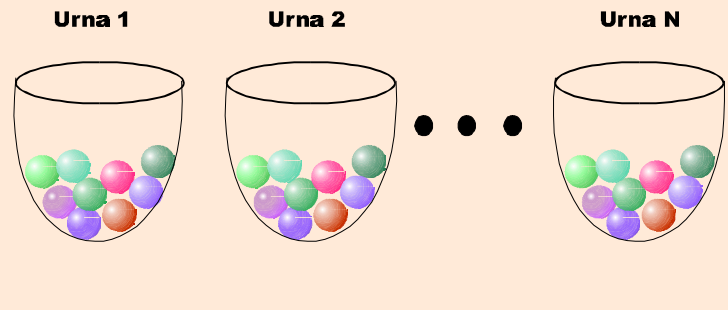
Se dice que el modelo de Markov es oculto porque no se puede observar directamente en qué estado se encuentra el modelo, únicamente se puede observar las observaciones generadas en cada estado. Un ejemplo que se expone claramente en [3] y que ayudará a entender los HMM es el de las urnas y bolas:

*"Consideremos las urnas de la figura 8, imaginemos que hay N urnas y cada una de ellas con bolas de colores. Hay M colores diferentes. Imaginemos el siguiente proceso aleatorio, una niña elige una urna y saca una bola y anota el color. La bola es devuelta a la urna y una nueva urna es elegida y se selecciona una nueva bola. Este proceso se repite un número finito de veces, T. Por tanto, se ha generado una secuencia de observaciones $B_i(O)$ de colores y los estados serían las urnas".*

En reconocimiento, la secuencia de observación $B_i(O)$, (secuencia de colores en el ejemplo anterior), es el conjunto de

FIGURA 8

**Un modelo de N urnas-estados y bolas ilustran el caso general de un modelo discreto de HMM**



CUADRO 2

### Problemas matemáticos que pueden resolverse con los HMM

**Problema 1**

*D*ado varios modelos y una secuencia de observación, ¿cómo calculamos la probabilidad de que la secuencia de observación fue producida por un modelo determinado? La solución a este problema nos permite encontrar un modelo adaptado al conjunto de observaciones. Para solucionarlo se utiliza el algoritmo de Baum. Aquel modelo que haya sido entrenado para reconocer esa secuencia de observación dará una probabilidad mayor que el resto de los modelos.

**Problema 2**

¿Cómo encontrar la secuencia correcta de estados? Se soluciona aplicando el algoritmo de Viterbi.

**Problema 3**

Dado una secuencia de observación, ¿cómo ajustar el modelo? Esto es lo que se conoce como entrenamiento y se aplica el algoritmo de Baum-Welch. Es decir, dado un conjunto de secuencias de observación de entrenamiento, el modelo se adapta para reconocerlo como si hubiese sido generado por él. Solucionando esta etapa se adaptan los parámetros a(i,j) y $B_i(O)$.

vectores-etiquetas de la librería del VQ.

En el cuadro 2 se recogen los tres problemas básicos de interés que deben resolverse para que el modelo pueda solucionar problemas del mundo real y vienen tratados en profundidad en [3].

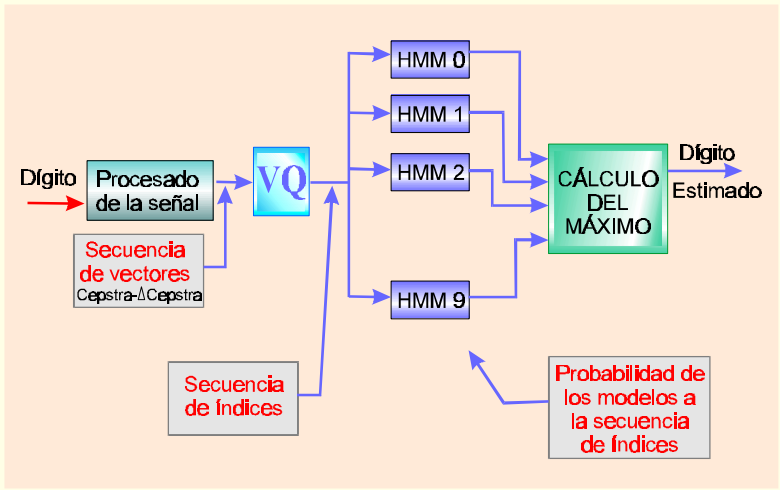
La aplicación de estos problemas en reconocimiento es el siguiente:

*Problema 1:* Se definen 10 modelos, uno por dígito. La entrada a estos modelos es la secuencia de etiquetas, procedentes del VQ, y que constituyen la secuencia de observación. El modelo que reconozca esta secuencia de etiquetas como suya dará una probabilidad alta a su salida. El modelo con la probabilidad más alta, es el que gana (ver figura 9).

*Problemas 2 y 3*: se aplican en la fase de entrenamiento de los modelos. Cada modelo se entrena para que reconozca un dígito determinado. Por ejemplo, el modelo del cero se entrena sólo con secuencias de etiquetas (observación) procedentes del cero. El modelo, una vez entrenado, responderá con una probabilidad alta cuando la entrada se corresponda con el dígito cero, y baja frente a otros dígitos.

FIGURA 9

### Diagrama de Bloques de un reconocedor discreto




## IMPLEMENTACIÓN DE UN RECONOCEDOR DE VOZ

En este apartado comentaremos cada uno de los pasos para realizar el reconocedor discreto de dígitos. En la figura 9 se representa un esquema del diagrama de bloques realizado, donde el cuantificador Vectorial (VQ) es el calculado para cada locutor. El funcionamiento del sistema es el siguiente, se pronuncia el dígito, se analiza espectralmente, se calculan los coeficientes *cesptra* y *delta-cepstra*, se cuantifican con los de la librería del VQ y posteriormente se pasan a todos los modelos de Markov. Aquel modelo que dé como resultado una probabilidad mayor con las observaciones o vectores de entrada es el que gana.

## Base de Datos

Lo primero que hay que tener es una base de datos de dígitos, para ello se eligieron 5 locutores: 3 varones y 2 mujeres. Está constituida por 660 grabaciones repartidas de la siguiente manera: 5 locutores y 132 grabaciones por locutor. Cada grabación contiene los dígitos del 0 al 9.

Las voces fueron grabadas con una tarjeta de Sound Blaster 16 ASP. El formato de grabación fue el siguiente: frecuencia de muestreo de 8000 Hz, 16 bits, mono y formato WAVE. A continuación se convertía todos los ficheros WAVE[1] a formato MATLAB[2]. La señal fue filtrada con un filtro paso alto de frecuencia 60 Hz para eliminar el tono de 50 Hz acoplado al sistema de grabación. La etapa siguiente fue la de aislar los dígitos.

## Preprocesado de Voz

El preprocesado de voz es el comentado en el apartado *Análisis de la señal de voz*. Se filtra la señal con un filtro de preénfasis cuyo coeficiente es 0.97. Se segmenta la señal y enventana con una ventana Hamming, de longitud 20 mseg y desplazamiento 10 mseg. De cada segmento de voz se calculan los coeficientes *cesptra* y *delta-cepstra*.

TABLA1

**Tasa de reconocimiento (en %) de los cinco locutores**

| Locutor 1 | Locutor 2 | Locutor 3 | Locutor 4 | Locutor 5 |
|---|---|---|---|---|
| 92.7 | 97.9 | 98.0 | 97.7 | 99.3 |

## Cuantificación Vectorial

En un principio para entrenar el VQ, se tomó toda la base de datos de dígitos, es decir, 6600 dígitos, lo que supone contar con más de 350.000 vectores con sus 26 coeficientes. La memoria necesaria era de 62 Mbytes. Debido al volumen de memoria necesaria se hacía implanteable con los medios de los que actualmente se disponen realizar semejante entrenamiento pues tardaría meses. Se optó realizar un cuantificador vectorial por cada locutor. La longitud de la librería es de 128.

## HMM ( Modelos Ocultos de Markov)

Los modelos ocultos de Markov han sido desarrollados a partir del artículo clásico de Rabiner [3]. El desarrollo de estos modelos de Markov ha supuesto implementar toda una serie de algoritmos, que se explican en [3], extensamente. Dichos algoritmos se han desarrollado para trabajar en el entorno MATLAB. Las características de los modelos son las siguientes: número de estados, 6, número de observaciones por estado es igual a la longitud de la librería de cuantificación, en este caso 128.

## Resultados obtenidos

De las 132 realizaciones que se tienen por locutor, se dejaron las 100 primeras para el test de los modelos y 32 para el entrenamiento. Para estos datos se ha obtenido una tasa de reconocimiento media es 97 %. (Ver Tabla 1).

En la actualidad seguimos trabajando en la mejora de este reconocedor: tasa de reconocimiento, cálculo computacional, aplicación de nuevos algoritmos de entrenamiento, así como la posibilidad de trabajar en tiempo real.

## GLOSARIO

**WAVE:** Se trata de un formato de grabación de sonidos, voces, música, etc. La extención de cualquier fichero que contenga este tipo de información es WAV.

**MATLAB:** Software de simulación muy útil en aplicaciones de procesado de señal. Es muy recomendable cuando se trabaja con vectores y matrices.

## BIBLIOGRAFÍA


1. **Joseph W. Picone:** "Signal Modeling Techniques in Speech Recognition", *Proceedings of IEEE*, vol. 81. Núm. 9, Septiembre 1993.

2. **Lawrence R. Rabiner, Biing-Hwang Juang** (1993): *Fundamentals of speech recognition*. Englewood Cliffs, NJ: Prentice-Hall.

3. **Lawrence R. Rabiner:** "A tutorial on hidden Markov Models and Seleted Aplications in Speech Recognition", *Proceedings of IEEE*, vol. 77. Núm. 2, Febrero 1989.

4. **John Markoul, Salim Roucos and Herbert Gish:** "Vector Quantitation in Speech Coding", *Proceedings of the IEEE*, vol. 73, núm.11, Noviembre 1985.

5. **Alan V. Oppenheim, Ronald W. Schafer** (1989): *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall.

6. **L.Rabiner and R.W.Schafer (1978):** *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall.

7. **Sadaoki Furui** (1989): *Digital Speech Processing, Synthesis, and Recognition*. New York, Marcel Dekker.


## BIOGRAFÍA


### ITZIAR GORETTI ALONSO GONZÁLEZ

Ingeniero Técnico de Telecomunicación por la Universidad de Las Palmas de Gran Canaria en 1991, e Ingeniero de Telecomunicación por la misma Universidad en 1993. Es Profesora asociada de la ULPGC desde 1994.

**Dirección:**
Escuela Universitaria Ingeniería Técnica de Telecomunicación
Edificio Pabellón A, Campus de Tafira
CP:35017, Tafira
Tlf: 45 12 50 Fax: 45 12 43
e-mail: ialonso@cic.teleco.ulpgc.es



*Este trabajo ha sido patrocinado por:*

***UNIÓN ELÉCTRICA DE CANARIAS, S.A. (UNELCO)***