Reconocimiento de Palabras Aisladas utilizando Modelos Ocultos de Markov* .

Maria Teresa Escrig Monferrer.[1]

Francisco Casacuberta Nolla.[2]

[1]*Departamento de Matemáticas e Informática.*
Universidad Jaume I de Castellón.
Campus Penyeta Roja. 12071 Castellón, España.

[2]*Departamento de Sistemas Informáticos y Computación.*
Universidad Politécnica de Valencia.
Valencia, España.

Key words: Automatic Speech Recognition, Isolated Word Recognition, Hidden Markov Models, Speaker Independent Recognition.

RESUMEN.

Actualmente son necesarias interfaces en las cuales la comunicación entre el ser humano y las computadoras sea eficiente, natural y que no requiera un gran esfuerzo de aprendizaje.En ello trabaja la línea de investigación "Reconocimiento Automático del Habla". En este trabajo se describe un sistema para el **Reconocimiento de Palabras Aisladas (RPA)** para vocabularios reducidos, independiente del locutor.y se muestran los resultados de su implementación en una estación de trabajo HP-9000, sobre la que el sistema funciona en tiempo real.

A partir de la señal vocal (en nuestro caso una palabra aislada del vocabulario), se obtiene una representación paramétrica de dicha señal en la etapa denominada **preproceso**, en el cual se realiza un tratamiento mediante **Banco de Filtros**, un cálculo del **cepstrum** y un **etiquetado**, que reducirá la cantidad de información con la que tratar y enfatizará las características más relevantes de la señal.

Señales así parametrizadas serán utilizadas para el aprendizaje de un **Modelo de Markov Oculto (MMO)** para cada palabra del vocabulario. Para el reconocimiento se tomarán muestras igualmente parametrizadas y se calculará la probabilidad de generar dicha muestra por cada uno de los modelos, eligiéndose como palabra aquella cuyo modelo de la probabilidad más alta.

El sistema permite, además, la posibilidad de generación de nuevas tareas, aprendiendo los modelos de cualquier otro vocabulario.

ABSTRACT.

Nowardays it is needed interfaces in which human-computer comunication is efficient, natural and that don't require a large effort of training. The line of investigation"Speech Automatic Recognition" is working in

* Este proyecto ha sido parcialmente subvencionado por la Comisión Internacional de Ciencia y Tecnología (CICYT): "Construcción de Sistemas de Reconocimiento de Habla" TIC 448/89.

this subjet. In this work we show a Speaker-Independent, Isolated Word Recognizer for small vocabularies, and we show the implementation results on a workstation HP-9000, in real time.

We obtain a parametric representation from a speech signal (in our case an isolated word of the vocabulary) in a preprocessing stage in which a filter-bank treatment, a cepstrum calculation and a labelling are made. This will reduce the amount of information to be treated and will emphasize the most outstanding signal features.

This parametric signal will be used for training a distinct Hidden Markov Model (HMM) for each word of the vocabulary. Recognition consists of computing the probability of generating the test word with each model and choosing the word whose model gives the highest probability.

This system allows us to generate new tasks to train models of other vocabulary.

1. INTRODUCCION.

El área de investigación en la cual está enmarcado este trabajo, Reconocimiento Automático del Habla (RAH), apunta hacia un estudio obligatorio en nuestra sociedad actual, el de las interfaces que faciliten la comunicación entre el ser humano y sus computadoras de forma eficiente, natural y sin requerir gran esfuerzo de aprendizaje.Entre la distintas disciplinas que confluyen en el RAH destacan la Teoría de la Señal, el Reconocimiento de Formas y la Inteligencia Artificial.

Debido a la enorme dificultad que entraña la comunicación oral, aún no ha sido posible diseñar un sistema de RAH general, que interprete cualquier discurso de cualquier locutor. Para simplificar el problema general hasta llevarlo a planteaminetos abordables, se introducen restricciones que hacen que los sistemas resultantes queden especializados en determinados aspectos del habla, por ejemplo Reconocimiento de Palabras Aisladas [3] o Reconocimiento del Discurso Continuo.

En este trabajo se describe un sistema para Reconocimiento de Palabras Aisladas (RPA) independiente del locutor y utilizando unos vocabularios reducidos.

A partir de la señal vocal (en nuestro caso, una palabra aislada del vocabulario), todo sistema de RAH debe obtener una representación paramétrica de dicha señal en una etapa denominada preproceso [9] [10], que reducirá la cantidad de información con la que tratar y enfatizará las características más relevantes de la señal.

Los vectores de parámetros acústicos serán clasificados en categorias microfonéticas en un proceso denominado etiquetado. La señal resultado de ese proceso servirá para el aprendizaje de una estructura o modelo que representará a cada palabra del vocabulario.

El criterio seguido en la etapa de reconocimiento, para determinar a qué palabra del vocabulario corresponde la palabra pronunciada, es el de máxima probabilidad [2] [4].

La estructura que representa cada palabra del vocabulario es un Modelo Oculto de Markov (MMO) que es la base de una técnica probabilística [3] [6] [10] [13].

2. ESTRUCTURA DEL SISTEMA.

En la figura 1 podemos ver un esquema del sistema que constituye nuestro reconocedor de palabras aisladas [6].

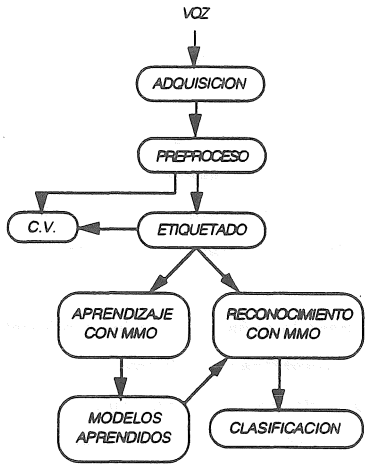

Figura 1. Sistema propuesto para nuestro Reconocedor de Palabras Aisladas.

La señal vocal podrá ser utilizada para el aprendizaje de un modelo de Markov que la representa, después del paso por las fases de adquisición, parametrización y etiquetado, denominadas en conjunto **preproceso** [6].

En la fase de **adquisición** [2], las ondas de presión producidas por el aparato fonador humano, son captadas por un micrófono que convierte esa onda sonora de presión en señal eléctrica. Posteriormente es amplificada para llevar la débil señal eléctrica obtenida hasta niveles manejables y después filtrada.

La señal analógica captada y amplificada, es digitalizada por un conversor analógico-digital; muestreando la señal analógica (midiendo la amplitud de la señal cada cierto tiempo) a 8555 Hz. y cuantificando la señal muestreada (pasando a una serie de valores numéricos) a 12 bits. Esta señal muestreada debe cumplir el teorema de muestreo o de Nyquist que indica que la máxima frecuencia del espectro no nulo de la señal a muestrear debe ser inferior a la mitad de la frecuencia de muestreo, de lo contrario se producen solapamientos de los espectros sucesivos que impiden reconstruir la señal original (efecto "aliasing") [11].

La siguiente fase por la que pasa la señal vocal es la de **parametrización** que servirá para eliminar la información redundante y enfatizar aquella que es relevante [2]. La parametrización consta de una multiplicación por ventanas de 256 puntos por la señal vocal de forma solapada. La ventana utilizada es la de HAMMING de 30 ms., por ser la más adecuada cuando la transformación aplicada a la señal vocal va dirigida a un análisis frecuencial de la señal [2], con un desplazamiento de 15 ms.

Después se realiza el cálculo de la Transformada de Fourier Dependiente del Tiempo que es procesada con la **Transformada Rápida de Fourier (FFT)** [7] [11].

Posteriormente se realiza un cálculo mediante **Banco de Filtros** de N filtros pasabanda, con sus intervalos frecuenciales distribuidos según unas escalas, que representan la división natural de las frecuencias audibles por el ser humano (bandas críticas) y su unidad es el "mel". La escala utilizada es la ESCALA DE MEL de 18 filtros que modeliza la respuesta de la membrana basilar del oido y cuyos resultados han sido comprobados en la práctica [2] [3].

Después del tratamiento de la señal mediante un banco de filtros, se realiza el cálculo de **coeficientes cepstrales** mediante la transformada inversa del coseno del logaritmo de los módulos [11].

La siguiente fase por la que pasa la señal vocal es la de **etiquetado**. Previamente a esta fase se ha tenido que efectuar una **Cuantificación Vectorial** (CV), técnica de Reconocimiento de Formas basada en una técnica de agrupamiento para el aprendizaje no supervisado ("Clustering"), que permitirá reducir la cantidad de información necesaria para representar la señal vocal. Mediante la CV creamos un conjunto de vectores prototipo ("codebook") a partir de unos vectores muestra, que definen un conjunto de clases. A cada clase se le asigna una etiqueta que la identifica, en principio de forma aleatoria [1] [5] [12].

Una vez definidas las clases y los prototipos representantes de cada clase, en la fase de etiquetado se asignará una etiqueta a cada vector de parámetros obtenido en la fase de parametrización utilizando la técnica de los k-vecinos más próximos, con k=3. La complejidad de este algoritmo es proporcional al número de prototipos por cada vector de parámetros.

Señales procesadas según lo anteriormente expuesto servirán, en una fase de entrenamiento, para el aprendizaje de los **Modelos de Markov Ocultos (MMO)** que representen cada una de las palabras del vocabulario. En una etapa de reconocimiento, se elegirá como palabra pronunciada aquella cuyo modelo ha dado la probabilidad más alta de generación de la muestra a reconocer.

3. MODELOS OCULTOS DE MARKOV.

Un MMO consta de una componente estructural, que viene dada por un conjunto de estados finito, y otra probabilística, formada por conjuntos de funciones de distribución de probabilidades [4] [10].

Transición entre estados $A = \{a_{ij}\} / a_{ij} \in [0,1]$; $1 \le i,j \le N$, que cumple que $\Sigma a_{ij} = 1$
Emisión de símbolos $B = \{b_i(k)\} / b_i(k) \in [0,1]$; $1 \le i \le N$; $1 \le k \le M$, que cumple que $\Sigma b_i(k) = 1$
Y distribución de estados iniciales $\pi = \{\pi_i\}$, que cumple que $\Sigma \pi_i = 1$

Sin embargo, debido a los buenos resultados obtenidos empíricamente [10] se utiliza una estructura de MMO con unas restricciones, dando lugar a los Modelos de Markov Ocultos de Izquierda a Derecha (MMOID) (figura 2), caracterizado por las propiedades siguientes:

1- $a_{ij} = 0$ si $j<i$.

2- $\pi_1 = 1$ y $\pi_j = 0$ para todo $j<>1$.

3- q_N es un estado final o absorbente.

A partir de los resultados experimentales obtenidos en [10] elegimos para este trabajo Modelos de Markov de Izquierda a Derecha (MMOID), sin saltos entre estados no continuos y con 10 estados, aunque, como veremos en el apartado siguiente se realizaron algunos experimentos con modelos de 20 estados.

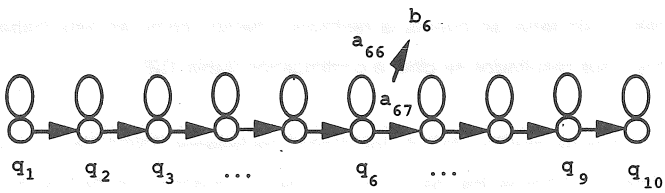

Figura 2. Modelo de Markov de Izquierda a Derecha de 10 estados y sin saltos entre estados no contiguos.

Al utilizar los MMOID se plantearon una serie de problemas que son expuestos en los siguientes párrafos.

Dada una secuencia de observaciones O, será reconocida como aquella palabra perteneciente al vocabulario ($w^* \in V$) cuyo MMO M dé la probabilidad más alta de generar la muestra [4] [6] [10].

$$w^* = \underset{1 \le v \le V}{\operatorname{argmax}} P_v(M/O)$$

El cálculo de la probabilidad de que un modelo genere una muestra O se realiza mediante los algoritmos "Hacia Delante" o "Hacia Atrás" [6] [10].

El problema de la precisión o "underflow" (probabilidades tendentes a 0) que es clásicamente resuelto introduciendo un factor de escalado [10], se soluciona en este trabajo utilizando una representación logarítmica de las probabilidades [6] [13]. Se pasa de trabajar con números reales a trabajar con números enteros. Además el cálculo de logaritmos simplifica algunas operaciones como son la multiplicación (suma de los logaritmos) y la suma que, aunque es más complejo, puede tabularse (en una tabla cuyo tamaño depende de la base logarítmica), de manera que el logaritmo de una suma se reduce a una suma, una resta, una comparación y una búsqueda en tabla. A pesar de que con los logaritmos se introducen errores, se demuestra empíricamente [13] que produce idénticos resultados a los obtenidos mediante la técnica del escalado.

La sustitución del escalado por el cálculo del logaritmo de las probabilidades supuso un ahorro sustancial del tiempo requerido para el aprendizaje y el reconocimiento utilizando los MMOID [6] [13].

4. EXPERIMENTOS Y RESULTADOS.

Los experimentos realizados con nuestro reconocedor tuvieron lugar en un ambiente ruidoso de un laboratorio de trabajo, lo cual, sin lugar a dudas, influyó en que los resultados obtenidos sean ligeramente peores a aquellos obtenidos con muestras adquiridas sin ruidos y comprobadas con un módulo de reconocimiento no integrado con el sistema de adquisición/preproceso y sin tener en cuenta la restricción tiempo como en este trabajo. Algunos de esos experimentos y sus resultados se citan a continuación (tabla 1)[6].

- Reconocimiento, en sistema multilocutor, de los dígitos castellanos (**DIG1**), en el cual, para 300 muestras de aprendizaje y 210 muestras para reconocimiento, se obtuvo un 94'3 % de aciertos, en un tiempo medio de reconocimiento de 1 s. 85 cs.

- Reconocimiento, en sistema multilocutor, de los dígitos castellanos, pero utilizando un modelo por cada dígito y cada sexo (**DIG2**), con un número similar de muestras para aprendizaje y reconocimiento, se obtuvo un 93'5 % de aciertos en reconocimiento para hombres y un 95' 7% para las mujeres. Lo cual indica que, a pesar del incremento del espacio y el tiempo necesario que supone aprender modelos por separado para los hombres y las mujeres los resultados son similares.

Tabla 1. Tabla de resultados obtenidos para diversos experimentos (ver texto), en los que se indica el número de muestras utilizado para el aprendizaje y el test, los tiempos de aprendizaje y de test medios, así como las tasas de muestras reconocidas correctamente.

	APRENDIZAJE		RECONOCIMIENTO		
	MUESTRAS	TIEMPO	MUESTRAS	TIEMPO	ACIERTOS
DIG1	300 (3h, 3m)	5m. 26s.	210 (1h, 3m)	1s. 85cs.	94'3%
DIG2	250 300	5m. 26s.	140 140	1s. 85cs.	93'5% 95'7%
DIG3	200 100	5m. 26s.	100 190	1s. 85cs.	80% 80%
NOMPER30	900 (3h, 3m)	19m. 28s.	250 (1h, 2m)	2s. 71cs.	87'2%
CIUD_20E	100	9m. 3s.	80	4s. 26cs.	99'5%

- Reconocimiento, en sistema multilocutor, de los dígitos castellanos separando modelos para hombres y mujeres, pero con un aprendizaje común inicial para ambos (**DIG3**); con un número similar de muestras, el porcentaje de aciertos en el reconocimiento disminuye hasta un 80 % tanto para hombres como para mujeres, debido a que los modelos están aprendidos con demasiadas iteraciones y llegan a "especializarse" en las muestras aprendidas.

- Los experimentos realizados con vocabularios más extensos (30 nombres de personas) **(NOMPER30)** y, con un número similar de muestras de aprendizaje, daban aún resultados peores, demostrando que se necesitan muchas muestras de aprendizaje para entrenar los modelos de vocabularios extensos y palabras parecidas.

- Se realizaron varios experimentos en un sistema monolocutor, con un vocabulario de 20 nombres de ciudades, de los que podemos extraer las siguientes conclusiones. El porcentaje de aciertos aumenta tal como vamos aprendiendo más los modelos hasta un cierto punto en el cual disminuye, por el problema antes mencionado. Si se utiliza un codebook propio para la aplicación la tasa de reconocimiento aumenta significativamente. El mejor resultado se obtuvo aprendiendo los modelos sobre un modelo de 20 estados sin saltos entre estados no contiguos **(CIUD_20E)** en vez de los modelos de 10 estados utilizados en los experimentos anteriores , 99'5%. Sin embargo, el tiempo medio de reconocimiento, aumenta en este caso significativamente.

5. CONCLUSIONES.

La construcción de modelos separados para hombres y mujeres no producen una mejora en el porcentaje de aciertos en el reconocimiento respecto a los que no hacen esa disticción. Para vocabularios relativamente extensos y con palabras parecidas, se necesitan bastantes muestras de aprendizaje para entrenar los modelos y obtener porcentajes de aciertos aceptables. En un sistema monolocutor (y con un codebook genérico) el mejor resultado se obtiene aprendiendo los modelos en dos fases, utilizando 5 muestras de cada palabra cada vez y con 10 iteraciones en el algoritmo de aprendizaje. Si los modelos están más entrenados la tasa de reconocimiento decrece. Para el mismo sistema obtenemos mejores resultados si modificamos la topología de los MMOID de 10 a 20 estados, aunque el tiempo de respuesta al usuario en este caso sea inaceptable.

Posibles mejoras al reconocedor construido serían la aplicación del algoritmo de Viterbi para el reconocimiento, lo cual podría mejorar el tiempo de respuesta; realizar una detección de bordes más fina; y distinguir las palabras pronunciadas que no pertenecen al vocabulario.

6. AGRADECIMIENTOS.

Deseo expresar mi más sincero agradecimiento a mi director de proyecto y al catedrático Enrique Vidal Ruiz, miembros del grupo de investigación "Reconocimiento Automático del Habla" de la Universidad Politécnica de Valencia, por su inestimable apoyo.

7. REFERENCIAS.

[1] Gabriela Andreu, Enrique Vidal y Francisco Casacuberta. **"An Empirical Evaluation of feature Maps and other Clustering Tecnhniques for frame labeling of speech"**. Elsevier Science Publisher B.V., 1990.

[2] Jose miguel Benedí Ruiz. **"Estudio de un Sistema de Reconocimiento Automático del Habla: TABARCA"**. Tesis doctoral. Universidad Politécnica de Valencia. 1988.

[3] Enrique Vidal Ruiz y Francisco Casacuberta Nolla. **"Reconocimiento Automático del Habla"**. MARCOMBO. BARCELONA. 1989.

[4] Francisco Casacuberta Nolla. **"Modelos de Markov Ocultos y Reconocimiento de Palabras Aisladas"**. DSIC. Universidad Politécnica de Valencia. Febrero 1991.

[5] R.O. Duda, P.E. Hart. **"Pattern Classification and Scene Analisys"**. John Wiley and Sons. 1973.

[6] María Teresa Escrig Monferrer. **"Un Reconocedor de Palabras Aisladas en tiempo real, para una estación de trabajo HP-9000 utilizando Modelos Ocultos de Markov"**. Proyecto Fin de Carrera. Universidad Politécnica de Valencia. 1991.

[7] F.J. Harris. **"On the use of windows for Harmonic Analisys with the Discrete Fourier Transform"**. Proc. IEEE, Vol 66, No 1, pp 51-84. 1978.

[8] S. Levinson. **"A Unified Theory of Composite Pattern Analisys for Automatic Speech Recongnition"**. En "Computer Speech Processing". F. Fallside y W.A. Woods. Prentice-Hall. 1985.

[9] Andrés Marzal Varó. **"Nuevos Métodos de Segmentación para la Decodificación Acústico-Fonética"**. Proyecto Fin de Carrera. Universidad Politécnica de Valencia. 1990.

[10] Begoña Más Valor. **"Reconocimiento de Palabras Aisladas y Decodificación Acústico-Fonética utilizando Modelos Ocultos de Markov"**. Proyecto Fin de Carrera. Universidad Politécnica de Valencia. 1989.

[11] L.R. Rabiner, R.W. Schafer. **"Digital Processing of Speech Signals"**. Prentice-Hall. 1978.

[12] L.R. Rabiner, S.E. Levison, M.M. Sondhi. **"On the aplication of vector Quantization and Hidden Markov Models to Speaker-independent, Isolated Word Recognition"**. The Beel System Tech. Journal. Vol. 62, No 4, pp 1075-1104. 1983.

[13] Luís Sanchez. **"Decodificación Acústico-Fonética utilizando Modelos Ocultos de Markov"**. Proyecto Fin de Carrera. Enero 1992.