Economía, XXIII, 14 (1998), pp. 7-20

La estadística como herramienta para el desarrollo de sistemas automaticos reconocedores del habla

Statistics as a tool for the development of speech recognition automatic systems

José Luciano Maldonado*

Resumen

Desde hace mucho tiempo el mundo científico viene haciendo grandes esfuerzos en la construcción de máquinas que sean capaces de intercambiar información en forma hablada con sus operadores humanos. Como consecuencia directa de esos esfuerzos se han alcanzado logros muy importantes;sin embargo, queda mucho por hacer. En ese sentido, se está iniciando el desarrollo de un sistema prototipo que tenga la capacidad de reconocer pronunciaciones del español hablado en Venezuela. Una actividad destacable que se ha cumplido hasta ahora, como parte de las etapas que hay que superar para lograr el objetivo, es la revisión de técnicas y algoritmos que se han utilizado a nivel mundial en la construcción de sistemas semejantes a los que se propone implementar. Producto de esa revisión se destaca el rol de la herramienta de la estadística en esta área de investigación. En ese sentido se muestra la importancia de la aplicación de la estadística en los progresos que ha tenido la tecnología del habla desde la década de los años setenta hasta nuestros días.

1. Introducción

Es bien conocido en el mundo científico que el hombre desde hace mucho tiempo ha deseado comunicarse con las máquinas a través de la voz. Por este motivo, los investigadores del habla (Ingenieros del Habla) han tratado de construir máquinas que sean capaces de recibir órdenes y/o mensajes por medio del habla, interpretar esos mensajes, realizar las actividades solicitadas y eventualmente presentar resultados, también en

* Universidad de Los Andes, Facultad de Ciencias Económicas y Sociales, Instituto de Estadística Aplicada y Computación

forma hablada. Este objetivo ha resultado muy difícil de alcanzar, y hasta existen investigadores que sostienen que ésta es una tarea imposible de lograr y que a lo más, se podrán concebir máquinas que puedan manejar sólo pequeños componentes del habla. Otros investigadores sostienen que el desarrollo de un sistema de tal capacidad, sólo se logrará si éste tiene una estructura modular e incremental. Con base en esta última idea se han orientado, a nivel mundial, proyectos a pequeños módulos que manejan precisamente algunos componentes básicos de la voz. Vale resaltar que a este nivel se han desarrollado sistemas que cumplen funciones específicas muy importantes, entre los cuales se encuentran los siguientes: sistemas de alarmas, sistemas traductores de texto escrito en un idioma a otro idioma, máquinas reconocedoras del hablante, sistemas que guían a su operador, etc.

Como el objetivo de este artículo es presentar el rol de la estadística en el desarrollo de sistemas reconocedores del habla, en la sección 1 se da una breve explicación de lo que es la tecnología del habla, destacando sus objetivos principales, luego en la sección 2 se explica el funcionamiento de los sistemas reconocedores, en la sección 3, se muestran sus componentes principales y las funciones de cada uno de éstos y finalmente en la sección 4 se señalan las principales herramientas estadísticas que se han venido usando en el desarrollo de tales sistemas.

2. Tecnología del habla

Es un área de investigación cuyo objetivo principal es construir máquinas capaces de interactuar con el hombre en forma hablada y en lenguaje natural. Sus retos están enmarcados en dos campos bien definidos: el reconocimiento automático del habla y la síntesis del habla. El primer campo tiene como objetivo específico desarrollar técnicas y algoritmos que lleven a crear sistemas con la capacidad de escuchar e interpretar mensajes dados a través de la voz, mientras que el segundo campo tiene como objetivo desarrollar técnicas y algoritmos que lleven a crear sistemas con la capacidad de producir voz, es decir, con el don de hablar. La integración de esos subsistemas en uno solo, constituye el

gran reto de esta interesante área de investigación. En estos dos campos se han venido usando herramientas estadísticas, con mayor énfasis en el primero de ellos.

Como en la actualidad, se desarrolla un sistema prototipo para el reconocimiento automático de pronunciaciones del español hablado en Venezuela, entonces, vamos a referirnos a partir de ahora al uso de la estadística en el reconocimiento automático del habla.

3. Funcionamiento de los sistemas reconocedores del habla

Estos sistemas tienen un ciclo de vida que comprende dos etapas. Una primera etapa llamada entrenamiento y una segunda llamada de reconocimiento o de identificación.

Durante la etapa de entrenamiento se le presentan al sistema una cantidad de pronunciaciones (elementos del habla: unidades básicas de las palabras, palabras, frases, oraciones, etc.) que se desea que éste "memorice" y durante la etapa de reconocimiento (superada la etapa de entrenamiento) se le pide que identifique una pronunciación particular dada, como alguna de las que ya conoce o parecida a las que conoce o simplemente como desconocida. Esto significa que la pronunciación a reconocer no tiene que ser, necesariamente, una de las que se usan en la etapa de entrenamiento.

Es importante señalar que la información almacenada o retenida por el reconocedor está constituida por propiedades extraídas de todas las pronunciaciones de entrenamiento, es decir, no se almacenan las pronunciaciones, sino propiedades de ese conjunto. Esto se hace con el fin de evitar en lo posible, almacenar datos redundantes y con ello darle al sistema la propiedad de responder en forma rápida, a cualquier solicitud de identificación de alguna señal de entrada. Lo ideal es que los sistemas respondan en tiempo real, sin embargo, éste es el gran problema de los sistemas actuales, que con el propósito de manejar grandes vocabularios de palabras, grandes conjuntos de frases, modelar gramáticas y lenguajes asociados a esas gramáticas, almacenan muchos datos y por lo tanto su velocidad de respuesta es muy baja. Este problema también constituye

una de las razones por las cuales se han venido construyendo sistemas de propósitos específicos.

4. Estructura general de un sistema de reconocimiento automático del habla

Como se muestra en la figura 1, un sistema reconocedor, presenta una serie de componentes que vamos a describir a continuación:

1. *Módulo de adquisición de datos*. Este es un subsistema cuya función es la de hacer la conversión de la señal sonora (sonidos del habla) a señal eléctrica (esto se hace a través de micrófonos y amplificadores electrónicos); luego esa señal eléctrica es convertida a una secuencia de valores numéricos, que es lo que maneja el resto del sistema. En otras palabras, el módulo se encarga de hacer la conversión analógica a digital de las pronunciaciones y de almacenar los datos producto de esa conversión.

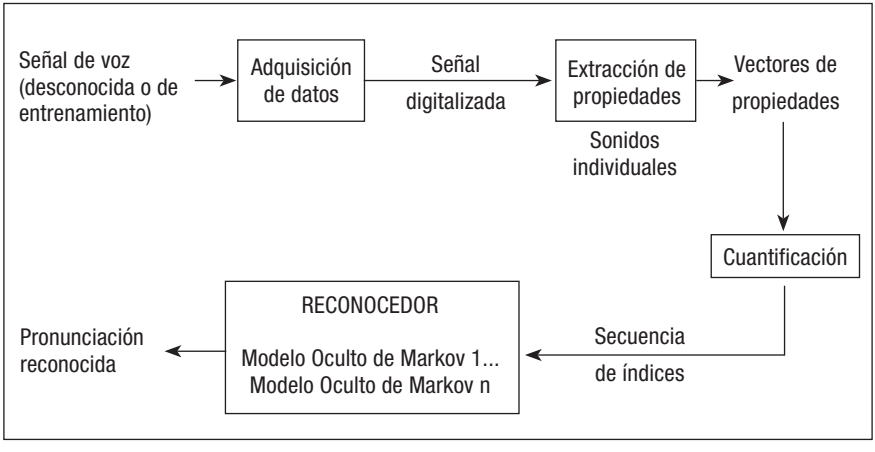

Figura 1. Esquema general de un sistema de reconocimiento

2. *Módulo de extracción de propiedades de la señal de voz*. Este subsistema se encarga de obtener propiedades de la señal (energía espectral, tono, formantes, donde empieza el sonido, donde termina el sonido, etc.) correspondiente a una pronunciación, por ello se encarga de dividir la secuencia de valores obtenida por el subsistema de adquisición de datos, en segmentos correspondientes a una duración de entre 10 y 35 milisegundos (toda pronunciación tiene una cierta duración). La razón para realizar esa separación es que se ha determinado que la duración de todos los sonidos del habla, está en ese rango. Aquí lo que realmente se realiza es una compresión de los datos para obtener un vector de propiedades de cada segmento y de cada sonido de la pronunciación, esto implica el uso de técnicas espectrales, FFT, modelos autoregresivos (ARMA) y regresivos (MA), Modelos de Predicción Lineal (LPC), Análisis Cepstral, filtrajes, etc. La salida de este módulo comprende una secuencia de vectores de propiedades de los segmentos.

3. *Módulo de cuantificación de los sonidos*. Este subsistema se encarga de identificar los distintos sonidos que están presentes en la pronunciación, para ello utiliza cada vector de la secuencia de vectores de propiedades obtenida en el módulo anterior. Cada vector está asociado a un sonido del habla, luego la salida de este módulo es una secuencia de valores, donde cada valor representa el sonido con el que está asociado un vector de propiedades. Está claro que un mismo valor y por lo tanto un mismo sonido puede aparecer varias veces en esta secuencia de salida.

4. *Módulo reconocedor propiamente dicho*. Este es el subsistema que finalmente va a identificar a una pronunciación dada, como conocida, parecida a una conocida o como desconocida. Para ello recibe desde el módulo de cuantificación la secuencia de valores que corresponde a una mezcla de los sonidos que puede tratar el sistema; estos sonidos individualmente corresponden a un segmento de la señal de la voz pero en conjunto y en la secuencia constituyen la señal completa de la pronunciación que se desea reconocer o

memorizar. La complejidad de este módulo depende del tipo de identificación que se requiera. Por ejemplo, un reconocedor de gramáticas será más complejo que un reconocedor de palabras y uno de palabras será más complejo que uno de letras, fonemas y fonos.

5. Algunas herramientas estadísticas usadas en el desarrollo de los sistemas de reconocimiento

En todos los sistemas reconocedores que se han construido hasta nuestros tiempos, la estadística ha intervenido de diversas maneras mediante la aplicación de sus técnicas, entre las cuales destacan de manera muy sobresaliente los modelos ocultos del Markov y las técnicas de grupo.

5.1. Los modelos ocultos de Markov (MOM) (Hidden Markov Models)

Son autómatas de estados finitos estocásticos, máquinas abstractas que permiten modelar procesos estocásticos. En el campo de la Tecnología del Habla se están usando para modelar las pronunciaciones, dada la gran variabilidad de dichas señales. La teoría de modelos ocultos de Markov tiene su origen en la idea que tuvieron unos estadísticos, en la década de los 50, para caracterizar procesos estocásticos para los cuales no se contaba con muchas observaciones. La idea básicamente consistía en modelar un proceso estocástico "doble", donde se asumía que los datos observados eran producto de hacer pasar el proceso real (oculto) a través de un medio cuyo resultado era el proceso observado (Deller y otros, 1993). De allí, surgió el algoritmo de identificación conocido como el algoritmo de Máxima Estimación (ME).

Para la aplicación de esta teoría al procesamiento de la voz, Baum y Welch hicieron una modificación al algoritmo mencionado y lo llamaron Baum-Welch, posteriormente surgió el algoritmo Viterbi. Estos dos algoritmos se vinieron usando casi sin competencia y con la misma efectividad hasta la década actual en la que aparecen los modelos de redes neurales artificiales para hacer ese tipo de entrenamiento e

identificación. A continuación se hace referencia a algún término de esta técnica.

a) *Topología de los modelos ocultos de Markov*: Su topología está determinada por el número de estados y las transiciones permitidas entre esos estados. A manera de ejemplo, en la figura 2 se muestra una topología típica.
Un Modelo Oculto de Markov de cuatro estados en el que se admite transiciones desde un estado hacia cualquier otro. Los MOM son entrenados para que generen una secuencia de sonidos correspondientes a una pronunciación determinada (una secuencia de observaciones). Un MOM genera secuencias de sonidos (observaciones) por la emisión de uno de éstos al pasar de un estado a otro. En otras palabras, por cada segmento del que se extrae un sonido (observación) de la señal pronunciada que se desea reconocer, ocurre una transición de estado en el modelo.

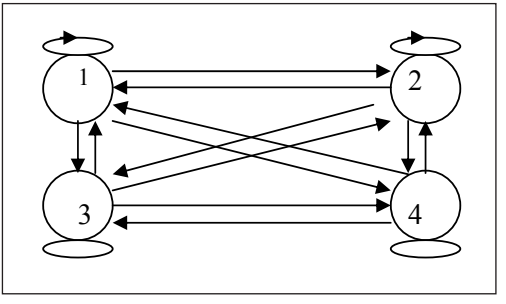

Figura 2. El Modelo Oculto de Markov más general

b) *Procesos aleatorios asociados con un MOM*: Un MOM tiene asociados dos procesos aleatorios: el proceso aleatorio de los estados (el proceso oculto) y el proceso aleatorio de los sonidos u observaciones (proceso observado). Estos dos procesos se describen a continuación:

i) *Proceso aleatorio de los estados*: A continuación se define matemáticamente el proceso aleatorio de los estados a través del cual se genera la secuencia de sonidos.

$\underline{x}$: proceso aleatorio de estados.
$\underline{x}(t)$: variables aleatorias asociadas al proceso aleatorio.
$a(i\backslash j)=P(\underline{x}(t) = i|)\ \underline{x}(t-1) = j)$: La probabilidad de transición desde el estado j hacia el estado i para un segmento t arbitrario.

A[a(i\j)]: Matriz de probabilidades de transición de estado.

$\prod$ (t) = [p(x(t) =i]: Vector de probabilidades de estado en el segmento t, i=1,2,...,S.

Donde S es el total de estados del modelo.
En cualquier segmento t se cumple: $\prod$ (t)=A$\prod$(t – 1) y por consiguiente, $\prod$ (t)=A^{t-1} $\prod$ (1)

$\prod$ (1): Vector de probabilidades de los estados en el segmento inicial.

ii) *Proceso aleatorio de las observaciones o sonidos*: La secuencia de observaciones (sonidos) se modela también como un proceso aleatorio discreto en tiempo, $\underline{y}$ con variable aleatoria $\underline{y}(t)$ Al entrar a un estado i en un instante o segmento t, se genera una observación y(t). La generación de dicha observación está gobernada por una función de densidad de probabilidad $f_{y(t)/x(t)}(\varepsilon / i)$

Un MOM que produce un conjunto finito de K valores u observaciones distintas, se ha usado por mucho tiempo en aplicaciones de tratamiento de señales de voz, ese es un tipo de MOM conocido como de observaciones discretas. En este artículo se describe este tipo de MOM, recordando que también existen MOM de observaciones continuas.

b (k\i)=P($\underline{y}(t)$ = k\ $\underline{x}(t)$ =i) : Es la probabilidad de que ocurra el sonido k (k es uno de los de los K sonidos) en el estado i, o b(y(t)\ i)=p($\underline{y}(t)$ = $\underline{y}$(t)\$\underline{x}(t)$ =i)

B [b(k\i)] : Matriz de probabilidades de los sonidos.
gP (t) = [P($\underline{y}(t)$ =k)] : Vector de probabilidades de los sonidos.
P (t)=B$\prod$ (t): Ecuaciones de las observaciones (sonido), o
P (t)=B A^{t-1} $\prod$ (1).

c) *Estructura matemática de los modelos ocultos de Markov descritos*: Sobre la base de las definiciones anteriores encontramos la siguiente estructura para un MOM m:

$$m = \{S, \prod(1), A, B, \{y_k, 1 \leq k \leq K\}\}$$

y_k representa uno de K sonidos u observaciones distintas que puede modelar un MOM m.

En la figura 3, se muestra un MOM entrenado donde se puede apreciar que en cada estado puede ocurrir o se puede generar un sonido (observación) correspondiente a un segmento de una señal pronunciada.
El entrenamiento de un MOM consiste en ajustar sucesivamente los parámetros $\prod(1)$, A,B hasta lograr que el modelo genere una secuencia deseada de sonidos. El entrenamiento comienza con valores arbitrarios para esos parámetros y una secuencia de sonidos de una pronunciación conocida que se obtiene en el módulo de cuantificación, esa secuencia es lo que se quiere modelar. Cada vez que se hace un ajuste, se verifica con qué probabilidad el modelo puede generar la secuencia y dicha probabilidad debe mejorar con cada ajuste, de ocurrir lo contrario el proceso de entrenamiento finalizará. El modelo con los últimos mejores parámetros es el que se acepta como el que genera o representa la secuencia de sonidos. En los MOM ese ajuste se hace con los algoritmos Baum-Welch y Viterbi que se basan en la estimación por Máxima Verosimilitud.

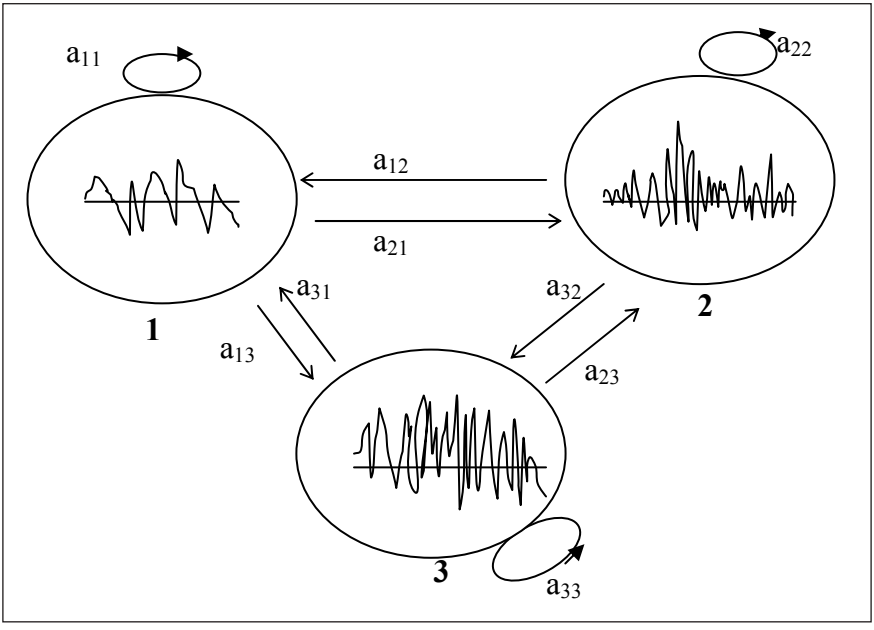

Figura 3. Un modelo oculto de Markov en cuyos estados se modela un segmento de señal de voz

Un sistema reconocedor contiene al menos tantos modelos ocultos de Markov como pronunciaciones diferentes puede identificar. Un sistema robusto, tendrá varios niveles de redes de MOM.
En la etapa de identificación, la pronunciación reconocida es aquella representada por el MOM que resulta con la mayor probabilidad de generarla.

5.2. Las técnicas de agrupamiento

Para hacer la identificación de los todos los sonidos diferentes que puede manipular un sistema reconocedor, en el módulo de cuantificación también se hace un entrenamiento. Este entrenamiento realmente consiste en un proceso de agrupamiento (aplicación de la teoría de *clustering*), de los vectores de observaciones "más parecidos" en el sentido de una métrica o medida de distorsión (las más usadas son las métricas *Euclidia* y la de *Itakura-Saito*, 1993). Cada grupo está asociado a un

sonido distinto y habrá tantos grupos como sonidos pueda manejar el reconocedor.

A continuación se explica el proceso de agrupamiento y cuantificación con el apoyo de la figura 4. Supongan que se tiene un espacio de vectores de observaciones, donde cada vector contiene propiedades extraídas de un segmento de la señal de una pronunciación. Suponga también que desea construir un sistema que maneje un número K de sonidos diferentes y que sea capaz de reconocer pronunciaciones de distintas personas que hablen un mismo idioma.

Como las pronunciaciones varían de persona a persona (por influencia del ambiente, el ruido, estado de ánimo, estructura física del aparato fonador humano, etc.), es de esperar que los sonidos individuales de cada persona no sean exactamente iguales a los de otras, pero si "parecidos". Debido a esto, nuestro espacio vectorial de observaciones lo podemos dividir en K grupos, donde cada grupo contiene aquellos vectores que corresponden a un sonido distinto. Esta transformación del espacio vectorial de observaciones original a K grupos, se realiza usando la teoría de *clustering* y específicamente a través de una de sus herramientas, la conocida como cuantificación vectorial. La cuantificación vectorial trabaja con base del siguiente algoritmo:

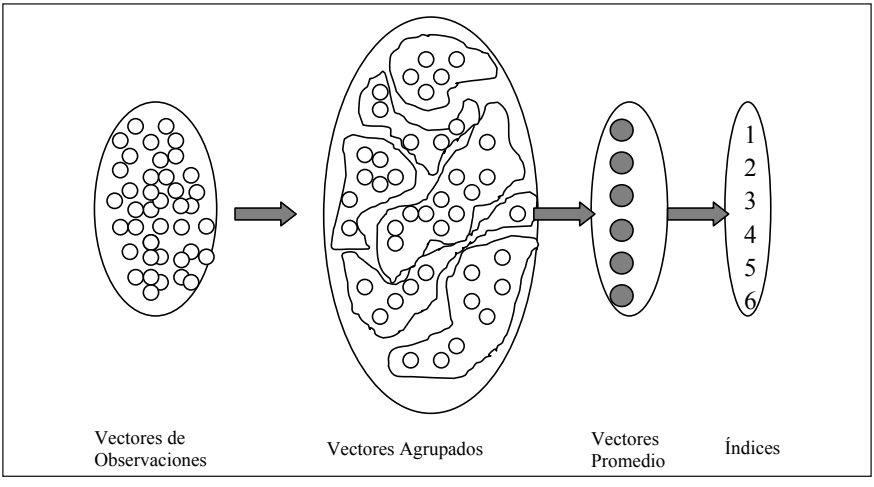

Figura 4. Ejemplo de cuantificación vertical

1. Se seleccionan arbitrariamente K vectores del espacio vectorial de observaciones, esos vectores constituyen el llamado código.
2. A cada vector del espacio de observaciones, lo asociamos con aquel vector de los K del Código, con el que más se identifique en el sentido de una medida de distorsión (con aquel cuya distorsión resulte más pequeña).
3. Calculamos la distorsión total producto de la asociación hecha en el paso 2. Si esa distorsión es suficientemente pequeña, paramos el proceso.
4. Por cada grupo que se forma, se calcula su vector promedio. Los nuevos vectores promedios de los grupos constituyen el nuevo código.
5. Se vuelve al paso 2.

El algoritmo que implementa la cuantificación vectorial de esta manera (en el contexto de la teoría de clustering), es el llamado algoritmo LBG en honor a sus creadores LLoyd, Buzo y Gray y es una extensión del algoritmo de las K-Medias. En la actualidad, también se hace cuantificación vectorial en el contexto de las Redes Neurales Artificiales a través del algoritmo Learning Vector Quantizer, LVQ.

Como resultado de esta cuantificación, el espacio de observaciones original fue transformado a un espacio de grupo de vectores "parecidos" y luego a un espacio constituido sólo por los vectores centroides (los promedios) de los grupos. Además, como se puede apreciar en la figura 4, cada centroide se puede representar por un valor o índice.

El proceso de cuantificación descrito constituye la etapa de entrenamiento del llamado módulo de cuantificación, mientras que en el proceso de identificación de ese módulo, lo que hace es una comparación (en el sentido de una medida de distorsión) entre cada vector que le entra y los respectivos centroides que tiene almacenados, se asocia ese vector con aquel centroide con el cual la distorsión es más pequeña y se marca o identifica a través del índice del grupo.

Es claro que el proceso de cuantificación comprende una compresión de datos, lo que reduce la carga computacional de los sistemas y por lo tanto, el tiempo de respuesta es considerablemente más rápido al empleado, si no se hiciera este tipo de manejo de propiedades.

6. Conclusiones

Debido a que los logros alcanzados por la Tecnología del Habla todavía no cumplen el objetivo planteado de manera satisfactoria, la búsqueda de nuevas técnicas y la modificación de las que existen continuará, por lo tanto, la Estadística seguirá con su contribución a través del uso y la evolución de sus herramientas actuales.

7. Referencias

Casacuberta, Francisco y Enrique Vidal (s.f.). "Reconocimiento automático del habla". Marcombo, Boixareu Editores, Barcelona-México.

Dayhoff, Judith (1990). *Neural Networks Arquitectures*. Editorial Van Nostrand Reinhold, New York.

Deller, John, Proakis John y Jansen Jhon (1993). *Discrete-Time Processing of Speech Signals*. New York: Macmilllan Publishing Company.

Haykin, Simon (s.f.). *Neural Networks, A comprehensive Foundation*. MAC MILLAN, IEEE PRESS.

Juang, B. y L. Rabiner (s.f.). "Issues Using Hidden Markov Models for Speech Recognition". Speech Research Department, AT&T Bell Laboratories.

Maldonado, José Luciano (1994). "Una aplicación de las redes neurales artificiales al reconocimiento de las vocales expresadas en español". Tesis de Grado para optar al título de Magister Scientiae en Ingeniería de Control. Facultad de Ingeniería, ULA.

____. (1997). "El estado del arte de la Tecnología del Habla". Charla dictada en el ciclo de Seminarios del postgrado en Ingeniería de Control, Facultad de Ingeniería, ULA, julio.

____. (1998). "Algoritmos para el reconocimiento automático del Habla". Charla dictada en el ciclo de Seminarios del postgrado en Ingeniería de Control, Facultad de Ingeniería, ULA, marzo.

Neural Networks (1993). Electronics World + Wireless World. August.

PAUL, D.B. (1990). "Speech Recognition Using Hidden Markov Models". *The Lincoln Laboratory Journal*, Volume 3, Number 1.

Picone, Joseph (1990). Continuous Speech Recognition Using Hidden Markov Models. July. *IEEE ASSP Magazine*.

Rabiner, L. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of The IEEE*, Vol. 77, No. 2, February.

Torres, I., and F. Casacuberta (s.f.). *Spanish Phone Recognition Using Semicontinuos Hidden Models*. Universidad del País Vasco y Universidad Politécnica de Valencia, España.

Waibel, Alex y John Hampshire (1989). *Building Blocks for Speech*. BYTE, August.

Zhao, Yunxin (1993). A Speaker-Indepent Continuous Speech Recognition System Using Continuos Mixture Gaussian Density HMM of Phoneme-Siuzed Units. *IEEE Transactions On Speech And Audio Processing*, Vol. 1, No. 3, July.