Reconocimiento automático del habla mediante redes neuronales

Enric Monte.
Departament de Teoria del Senyal i Comunicacions.UPC

I- Introducción.

En la década de los 80 se popularizaron una serie de arquitecturas de ordenadores que tienen una gran semejanza en cuanto a su estructura con las redes neuronales. Estas arquitecturas, junto con algoritmos capaces de ejecutarse en ellas, dieron lugar a la aparición de lo que se conoce como modelos conexionistas o redes neuronales artificiales. En este tipo de máquinas el cálculo se realiza de manera distribuida en un gran número de procesadores sencillos, dotados de una gran riqueza de conexiones entre ellos. Las aplicaciones en las que se han empleado los modelos conexionistas, se caracterizan por necesitar gran potencia de cálculo y por no tener una solución práctica mediante tecnologías convencionales. Las aplicaciones en donde ha sido natural el uso de las redes neuronales han sido principalmente control, imagen y voz.

Las arquitecturas conexionistas intentan heredar algunas de las propiedades del sistema nervioso. Estas propiedades abarcan la capacidad de clasificación rápida de imágenes o sonidos, o el control motor de un sistema con gran número de grados de libertad como es el caso de la mano. Además de ser capaz de realizar este tipo de tareas una arquitectura conexionista ha de ser capaz de funcionar con procesadores de gran simplicidad y extremadamente lentos. Otra propiedad que ha de tener una arquitectura conexionista, es que los algoritmos que se puedan ejecutar sobre ella, sean fácilmente paralelizables. Esta última propiedad es casi tan importante como la existencia de un hardware capaz de soportar un número elevado de procesadores altamente interconectados, y ha sido la razón detuvo la investigación en este área durante casi 20 años. En los años 50 y 60 se produjo un auge en la investigación de arquitecturas que fueran una alternativa a las arquitecturas secuenciales tipo Von Newman. La investigación en este área se detuvo tras el fracaso de las búsquedas de algoritmos que se pudieran ejecutar en arquitecturas de este tipo. Por último una propiedad que ha de tener este tipo de arquitectura es la capacidad de aprender. Este aprendizaje puede ser a partir de ejemplos presentados al sistema por un supervisor o se puede producir por autoorganización interna de las conexiones entre procesadores. Además presentan una serie de propiedades de las cuales nombramos unas cuantas a continuación:

**Paralelismo masivo: La característica que define las redes neuronales es la existencia de gran cantidad de procesadores simples, con una topología de conexiones muy complicada y rica. Esta propiedad proporciona versatilidad de las redes, en cuanto a problemas que puede resolver, velocidad de tratamiento, tolerancia a fallos.

**Aprendizaje: Otra de las características que hace atractivas las redes, es la existencia de algoritmos para entrenar las redes de manera automática a partir de ejemplos. Muchos de estos algoritmos son locales, por lo que cada «neurona», únicamente se ha de comunicar con las «neuronas» de un entorno topológico, lo que facilita las comunicaciones. Además algunos algoritmos por ser locales permiten que cada «neurona», pueda aprender de manera independiente de las otras, y por tanto se puede introducir paralelismo en la red.

**Modelado estocástico, incertidumbre y variabilidad: La propiedad de generalización de las redes neuronales permite tratar la variabilidad y el ruido de las muestras. Esta propiedad consiste en que con muestras que la red no ha visto durante el entrenamiento, la red proporciona resultados parecidos a los que daría con las muestras de entrenamiento. Otra ventaja de las redes no presuponen distribuciones de probabilidad de los datos de entrada, sino que realizan un modelado no paramétrico de los mismos.

Aprendizaje de funciones no lineales: Las redes neuronales se pueden usar como clasificadores no lineales y para aprender funciones arbitrarias de **Rsup5(N) a **Rsup5(M)**, lo que puede llevar a mejorar las prestaciones en aplicaciones de clasificación y de aprendizaje de funciones.

**Integración de información procedente de diversos orígenes: La operación básica de la mayor parte de los modelos de redes neuronales es la combinación lineal de entradas de una «neurona». Este tipo de operación, hace que para la red sea transparente el tipo de entrada que tenga. El proceso de aprendizaje, hace que los pesos de la combinación lineal sean tales que las variables de entrada queden escaladas a un mismo margen.

Para una introducción al los diversos tipos de arquitecturas neuronales se puede consultar el artículo de Lippman [1] o el libro de Hetch-Nielsen [2].

II- Relación entre las redes neuronales biológicas y electrónicas.

Para realizar un modelo de las

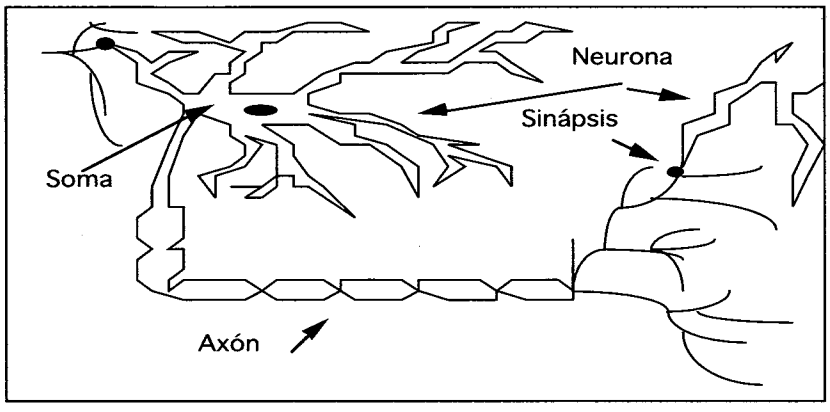
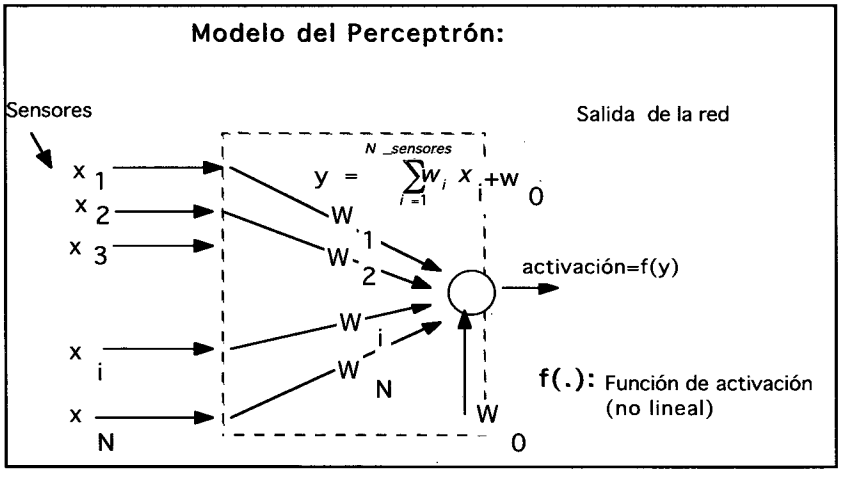


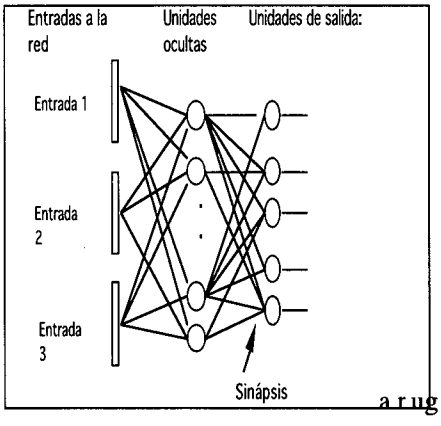
Figura 1. Dibujo en el que se esquematiza el diagrama de una neurona biológica.

redes neuronales biológicas, es necesario tener un conocimiento de su comportamiento. En la figura 1 presentamos un dibujo de una neurona biológica. En el dibujo se observa que la neurona está constituida por un cuerpo central denominado soma, en el que confluyen sinápsis procedentes de otras neuronas. En el soma se produce una integración de las señales procedentes de las sinapsis. Empíricamente se ha observado que las neuronas responden de manera no lineal a las señales procedentes de las sinapsis. Cuando la intensidad de las señales que confluyen a las sinapsis supera un cierto umbral, la neurona responde emitiendo una señal que se genera en el soma y que se propaga por el axón hasta llegar a las dendritas, donde transmiten la señal propia de la neurona a otras neuronas a través de las sinapsis.

El modelo de una neurona biológica constará de unos elementos que simularán las sinápsis. Estos elementos que simularán las sinapsis serán canales de comunicaciones entre elementos procesadores. Además el modelo que haremos de la neurona biológica ha de ser capaz de simular el efecto umbral, es decir, ha de ser capaz de activarse o dar señal a partir de unas condiciones relacionadas con las entradas de la red. De manera arbitraria en el modelo de la neurona supondremos que la neurona se dispara o genera señal a partir del momento en que la suma de las entradas procedentes de otras neuronas ponderadas por la eficacia sináptica asociada con cada conexión. En la figura 2 presentamos el modelo de una neurona, que llamaremos perceptrón. En este modelo se observa que la entrada esta ponderada por la eficacia sináptica, representada aquí como wsdo5(i), una vez se ha realizado la combinación lineal de las entradas, se simula la actividad del cuerpo de la neurona (Soma), por medio de una función no lineal arbitraria **f(.)** que incluye el efecto de umbral de la red biológica.Como se ha observado en la figura anterior, el modelo que hemos realizado de la neurona biológica es de una gran simplicidad, únicamente efectúa una combinación lineal de entradas, que pueden proceder de otras neuronas o de sensores. La manera de aumentar la capacidad de cómputo consiste en crear una red de neuronas simples altamente conectadas. Intuitivamente se puede pensar que este tipo de arquitectura puede heredar algunas de las propiedades que tiene el tejido nervioso. Existe el teorema de Kolmogorov, que dice que una red de procesadores que realizan las operaciones del modelo de neurona que hemos descrito en el párrafo anterior puede implementar cualquier función continua y arbitraria entre un cubo n-dimensional [0,1]sup5(n) y el conjunto Rsup5(m), sin imponer restricciones en la dimensionalidad de la entrada o de la salida. Este resultado es muy importante pues, demuestra que con arquitecturas del tipo mostrado en la figura 3, se pueden aprender aplicaciones arbitrarias. En particular estas aplicaciones pueden ser el control del movimiento de una mano, el reconocimiento de imágenes, el reconocimiento y síntesis del habla y en general tareas que son difíciles de resolver mediante otro tipo de técnicas.

Figura 2. Modelo matemático de una neurona biológica





a r u g i F

fig 3. Diagrama de una arquitectura neuronal.

III- El reconocimiento automático del habla mediante redes neuronales.

Una de las peculiaridades que tiene el problema del reconocimiento del habla es la naturaleza secuencial de la señal de voz. Esta peculiaridad

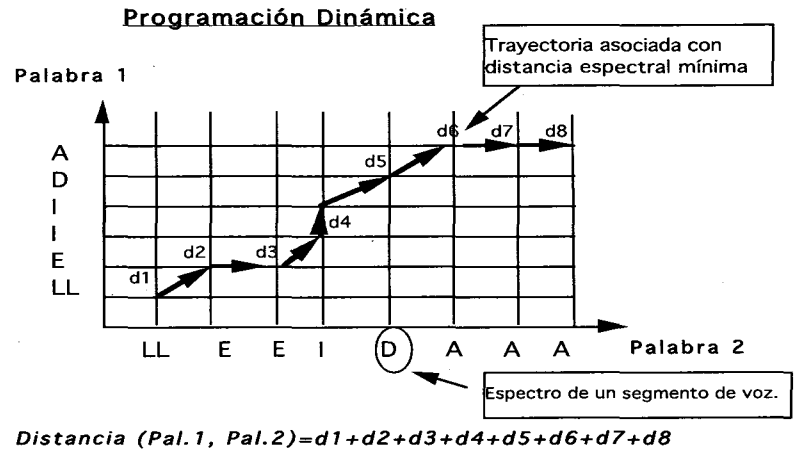


Figura 4. Descripción del funcionamiento de la programación dinámica.

dificultará la tarea de reconocimiento pues nos encontraremos que se tendrán que comparar señales de duración diferente. La herramienta que se ha usado para realizar la comparación de señales de duración diferente es la programación dinámica desarrollada por Bellman en la década de los 60. Para realizar la programación dinámica se ha de dividir la señal en segmentos de la misma duración. Si queremos comparar señales de duraciones diferentes tendremos que comparar dos conjuntos de segmentos con un número de componentes diferentes. Para poder realizar esta comparación se genera una matriz de distancias y se busca el camino de distancia mínima entre el origen y el final. En la figura 4 mostramos un ejemplo de como funciona la programación dinámica en el que se calcula la distancia entre dos realizaciones de la palabra «*LLEIDA*» .

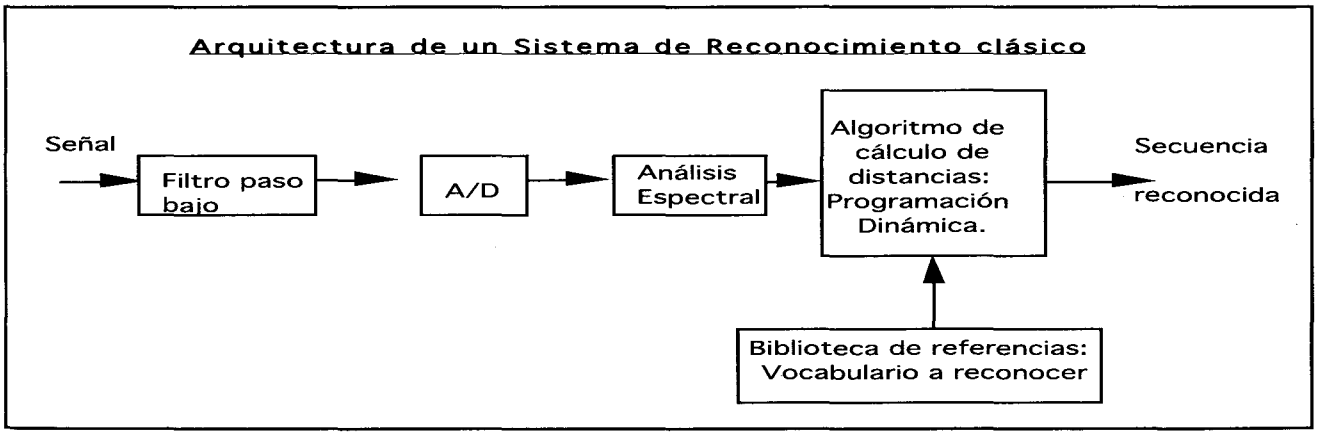
El uso de las redes neuronales en el campo del reconocimiento automático del habla, representa una innovación cualitativa respecto a las aproximación clásica basada en distancias y en la programación dinámica. En la figura 4 presentamos un diagrama de un sistema de reconocimiento clásico. El núcleo de un sistema de de reconocimiento clásico consiste en un módulo en el que comparan dos secuencias de duración diferente: la señal de entrada con la señal almacenada en la biblioteca de referencias.

Dado que las arquitecturas basadas en redes neuronales, son una herramienta reciente, se han ideado diversos sistemas para tratar el problema del reconocimiento del habla. Algunos de estos sistemas tratan el problema de manera homogénea y resuelven la secuencialidad del habla sin utilizar programación dinámica. Sin embargo existen diversos sistemas híbridos en los que se utilizan las redes neuronales para resolver los aspectos del problema de clasificación en los que éstas funcionan mejor que los sistemas tradicionales; dejando el problema del alineamiento temporal a la programación dinámica. Esto es debido a que las redes actuales no tratan el problema de la secuencialidad del habla de manera totalmente satisfactoria. A pesar de que las redes neuronales como línea de investigación tienen un futuro prometedor, las prestaciones actuales de los sistemas basados en redes neuronales son limitados. Con los sistemas actuales es difícil realizar un sistema de reconocimiento automático del habla, independiente del locutor y de gran vocabulario (>1000 palabras) basado en redes neuronales.El estudio lo llevaremos a cabo sobre el tipo de unidad fonética sobre el que trabaja cada una de las redes, ya sea a nivel fonético o nivel de palabra.

Una aproximación al reconocimiento basado en sistemas que combinan las redes neuronales con la programación dinámica, se basa en usar la programación dinámica para alinear la señal de entrada a una duración preestablecida. En este caso la red neuronal actúa como un clasificador, con un grupo de neuronas que se activa según sea la clasificación que decida la red. Tras el procesado que consiste en un análisis espectral de los segmentos de voz, tenemos la señal con una duración constante que usaremos como entrada a una red con un número de sensores fijos. La red realiza la clasificación de la señal de entrada activando la neurona de salida que corresponde a una palabra del vocabulario. En la figura 6 mostramos un diagrama del funcionamiento de este sistema de reconocimiento.

Una solución alternativa consiste en usar una red realimentada,

figura 5



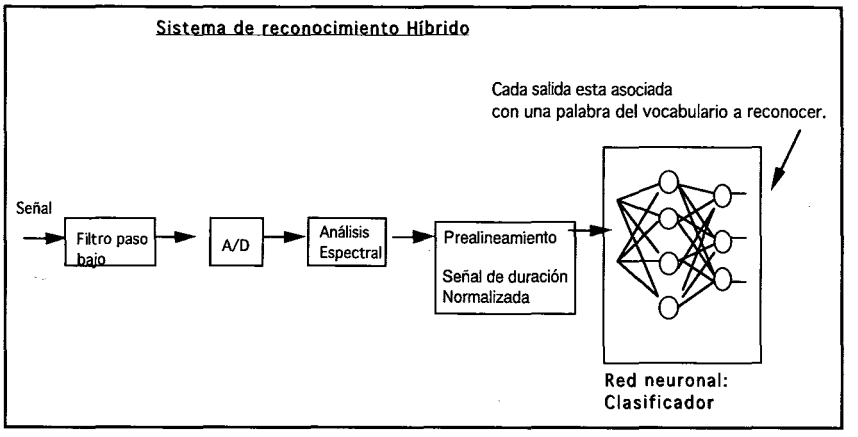


Figura 6. Diagrama de un sistema de clasificación de señales basado en la combinación de un alineamiento temporal (que soluciona el problema de la variabilidad de la duración de las palabras), con una red neuronal que trabaja como un clasificador.

con estructura semejante a un filtro IIR, en el que se realimenta ya sea la salida de la red o las activaciones de la capa oculta de la red. En este tipo de sistema de reconocimiento no es necesario alinear la señal a una longitud fija, pues la realimentación se encarga de conservar memoria sobre la historia pasada de la señal de voz que usamos como entrada a la red. La ventaja de un sistema de este tipo es que no se ha de realizar ningún tipo de alineamiento de la señal, ni ningún tipo de segmentación.

Existen otros tipo de sistemas de reconocimiento automático del habla basados en redes neuronales. En estos sistemas se le agrega memoria a la red (requisito indispensable para poder comparar señales de duración diferente), mediante lineas de desplazamiento en las neuronas (Time Delay Neural Nets) o mediante el uso de Modelos Ocultos de Markov.

ENRIC MONTE MORENO. Doctor Ingeniero de Telecomunicaciones por la UPC (1992).
Actualmente es profesor asociado del Departament de Teoria del Senyal i Comunicacions de la UPC donde imparte la asignatura "Electroacústica". Es autor de diversas publicaciones dentro de la especialidad de Procesado y Reconocimiento de Voz, así como de numerosas ponencias en congresos internacionales.

IV- Linéas futuras.

El tratamiento de la información realizado por las redes neuronales, promete mejoras espectaculares en un futuro cercano, dado que la capacidad de integrar en silicio estructuras del tipo que hemos discutido en el artículo se puede realizar con las tecnologías actuales. La aparición de hardware especializado en este tipo de arquitecturas de ordenador permitirá que en un futuro cercano se puedan realizar sistemas de reconocimiento del habla, de gran vocabulario (en la actualidad está limitado a 1000 palabras) con independencia del locutor (en la actualidad como máximo los sistemas aceptan decenas de locutores distintos)

V-Referencias.

[1] Lippmann,R.P. «An introduction to computing with neural nets».IEEE.ASSP Mag. Abril,1987.

[2] Hecht-Nielsen, R. «Neurocomputing»; Addison-Wesley,1991.

[3] Lippmann,R.P. and Gold,B. «Neural-net classifiers useful for speech recognition. IEEE Int. Conf. Neural Networks, Junio, 1987.

[4] Elman,J.L. y Zipser, D. «Learning the hidden structure of speech.» Technical Report, University of California, San Diego, Feb. 1987.

[4] Monte, E. «Reconocimiento Automático del habla mediante redes neuronales y técnicas híbridas». Tesis Doctoral, Universidad Politécnica de Catalunya. 1992.

[5] Waibel, A., Hanazawa, T., Hinton,G., Shikano, K., y Lang K. «Phoneme recognition using time-delay neural networks.» IEEE Trans. Acoust. Speech Signal Process. Marzo 1989.

[6] Wartrous,R. «Speech recognition using connectionist networks.» Ph.D. thesis,Universidad de Pensilvania. 1988.

[7] Waibel, A. «Modular construction of time-delay neural networks for speech recognition.» Neural Computation. 1989.

[8] Leung, H.C. y Zue, V.W. «Applications of error back-propagation to phonetic classification.» Advances in Neural Information Processing Systems. D. S. Touretzky (ed.) Morgan Kaufman, San Mateo. 1989.

[9] Hampshire, J., y Waibel, A. «A novel objective function for improved phoneme recognition using time delay neural networks.» IEEE Trans. Neural Networks, Junio, 1990.

[10] Peeling S., y Moore, R. «Experiments in isolated digit recognition using the multi-layer perceptron.» Technical Report 4073, Royal Speech and Radar Establishment, Diciembre 1987.

[11] Burr, D.J. «Speech recognition experiments with perceptrons.» Advances in Neural Information Processing Systems.» Morgan Kaufmann. 1988.

[12] Bottou, L. Fogelman-Soulie, F.,Blanchet, P., y Lienard, J.S. «Experiments with time-delay networks and dynamic time warping for speaker independent isolated digits recognition.» Proc. Eurospeech, Septiembre 1989.

[13] Tank,D.W. y Hopefield,J.J. «Neural computation by concentrating information in time.» Proc. Natl. Acad. Sci.USA. Abril 1987.

[14] Sakoe,H. Isotani,R.Yoshida,K, Iso,K. y Watanabe,T. «Speaker independent word recognition using dynamic programming neural networks.» IEEE Int. Conf. Acoustics Speech and Signal Processing 1990.

[15] Iso, K. y Watanabe,T. «Speaker independent word recognition using a neural prediction model.» IEEE Int. Conf. Acoustics, Speech and Signal Processing. 1990

[16] Levin, E. «Speech recognition using hidden control neural network architecture.» Proc. Int. Conf. Acoustics, Speech and Signal Processing. 1990.

[17] Miyatake, M., Sawai, H. y Shikano, K. «Integrated training for spotting Japanese phonemes using large phonemic time delay neural networks.» IEEE Int. Conf. Acoustics, Speech and Signal Processing. 1990.