Resultados preliminares sobre SLHMM.

J.E. Díaz Verdejo, J.C. Segura Luna, A.J. Rubio Ayuso
P. García Teodoro, Victoria Sánchez Calle
Dpto. Electrónica y Tecnología de Computadores. Facultad de Ciencias.
Universidad de Granada. 18071 - GRANADA (Spain)

Resumen

En este trabajo se propone un nuevo sistema híbrido para el reconocimiento de voz continua que integra HMM y ANN. Dicho sistema se compone de 3 clases de bloques (LVQ, SLHMM y DP), todos ellos redes neuronales, si bien los denominados SLHMM pueden ser interpretados y entrenados de acuerdo con los HMM. Un SLHMM es, básicamente, una expansión en una red con un número fijo de capas de una red neuronal recurrente con una topología conveniente. Se presentan algunos resultados experimentales preliminares que, comparados con los obtenidos a partir de un sistema basado únicamente en HMM, muestran un incremento en el rendimiento del sistema sencillamente debido a la topología utilizada.

1 Introducción

Varios trabajos recientes han puesto de manifiesto las fuertes relaciones existentes entre ambos tipos de Redes Neuronales Artificiales (ANN), concretamente las llamadas Redes Neuronales Recurrentes (RNN) y los Modelos Ocultos de Markov (HMM) [1] [2]. De hecho, estos trabajos han probado que las RNN, con una topología apropiada, son equivalentes a los HMM.

Nuestro objetivo es la utilización de este tipo de redes neuronales como núcleo de un sistema de reconocimiento de voz continua, con la finalidad de mantener la interpretabilidad según HMM. Esto puede permitirnos mejorar el rendimiento de ambos tipos de sistemas (los basados en HMM y los basados en ANN) aplicando técnicas de redes neuronales a los HMM y viceversa. Por otra parte, si se mantiene la interpretabilidad según HMM de uno de los módulos del reconocedor, es posible entrenar sus parámetros con los algoritmos usuales de los HMM, que son más simples que los correspondientes a las ANN.

Para reconocer voz continua utilizando ANN es necesario evaluar las probabilidades de la secuencia a reconocer localmente, debido a la estructura fija de la red. Este problema aparece cuando se utilizan ANN porque no es posible construir un "supermodelo" de la secuencia de voz simplemente concatenando los modelos elementales, tal como se hace en el caso de los reconocedores basados en HMM.

Para solucionar este problema hemos desarrollado un formalismo que puede evaluar el camino óptimo para el conjunto de probabilidades locales. Este formalismo permite diseñar

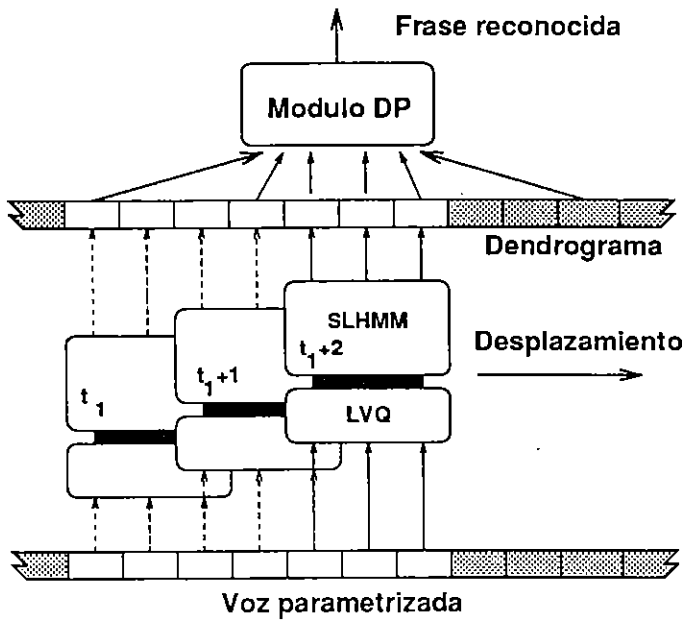

Figura 1: Diagrama de bloques del sistema propuesto.

un sistema modular compuesto por varios tipos de ANN que pueden ser entrenadas mediante los algoritmos convencionales utilizados en ANN, como p.e. retropropagación.

2 Descripción del sistema.

El sistema propuesto está compuesto por 3 tipos de módulos, de los cuales 2 son directamente ANN. El tercer módulo puede ser implementado como una ANN, aunque no es entrenable. Estos módulos son (figura 1): módulos de cuantización LVQ, módulos de estimación de probabilidades locales basado en los SLHMM, y módulo de programación dinámica (DP).

El sistema se compone de un módulo LVQ y un módulo SLHMM asociados entre sí para cada una de las unidades de reconocimiento (en nuestro caso, fonemas) y un módulo de programación dinámica. El procesamiento realizado por el sistema se describe a continuación. La señal de voz es convenientemente parametrizada, obteniéndose una serie de vectores. Esta secuencia de vectores es procesada por cada uno de los módulos SLHMM y LVQ asociados a cada uno de las unidades a reconocer. Para ello se selecciona inicialmente un segmento de longitud determinada de la secuencia de vectores. Este segmento es procesado primero por el módulo LVQ, con lo que se obtiene una versión cuantizada del vector de entrada, que es procesada por el módulo SLHMM. A la salida se obtiene un conjunto de probabilidades parciales del segmento seleccionado a la entrada. Desplazando los dos módulos sobre la secuencia a reconocer es posible obtener todas las probabilidades parciales para todos los instantes de tiempo y para todos los fonemas. El resultado es procesado por el módulo de programación dinámica, que obtendrá la secuencia óptima de fonemas mediante búsqueda sobre el conjunto de probabilidades obtenido anteriormente.

En una primera fase, el módulo LVQ ha sido suprimido debido a que no afecta al núcleo del sistema y, actualmente, nuestro interés reside en el desarrollo de los módulos SLHMM y DP.

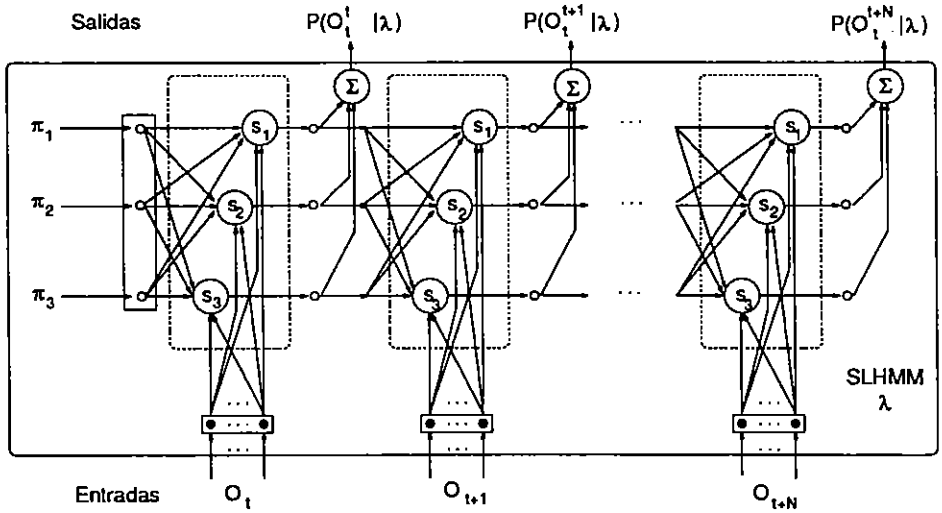

Figura 2: Esquema de un SLHMM.

3 El módulo SLHMM.

El módulo SLHMM es el núcleo del sistema. Básicamente consiste en un desarrollo de una RNN-HMM obtenido expandiendo la red para un número fijo de capas. De esta forma, cada instante temporal en la evaluación de una RNN ha sido sustituida por una capa en la estructura del SLHMM. Obviamente, esta expansión de la red impone restricciones temporales: la longitud máxima de la secuencia a evaluar es el número de capas del SLHMM.

La estructura del SLHMM se muestra en la figura 2. Sus entradas son un número fijo de elementos de la secuencia de símbolos de entrada, mientras que su salida es un vector compuesto por las probabilidades parciales de generación. De esta forma, para un SLHMM con $N+1$ capas, la entrada será la secuencia parcial de símbolos

$$O_{t_1}^{t+N} = \{O_{t_1}, O_{t_1+1}, \cdots, O_{t_1+N}\}$$

mientras que la salida serán las probabilidades de todas las subsecuencias $O_{t_1}^{t_1+N}$ con $1 \leq n \leq N$ y t_1 como instante inicial:

$$\{P(O_{t_1}^{t_1}|\lambda), P(O_{t_1}^{t_1+1}|\lambda), \cdots, P(O_{t_1}^{t_1+N}|\lambda)\}$$

Desplazando la red sobre la secuencia completa (fig. 1) se puede obtener un *dendrograma* (conjunto de todas las posibles segmentaciones) de la señal ya que se evaluarán todas las probabilidades de la forma

$$p(O_{t_1}^{t_2}|\lambda) \quad con \quad 0 \leq t_1 \leq T-N;\ t_1 \leq t_2 \leq t_1+N$$

4 El módulo DP.

El módulo DP obtendrá el camino óptimo sobre el dendrograma. Este camino óptimo viene determinado por el número de unidades que componen la secuencia, L (p.e. el número de

fonemas), la secuencia de unidades óptima, $\{\lambda_1, \lambda_2, \cdots, \lambda_L\}$ y los índices de los instantes temporales finales de dichas unidades $\{t_1, t_2, \cdots, t_{L-1}\}$.

El funcionamieto del módulo DP es diferente dependiendo de si se está entrenando el sistema o reconociendo una secuencia.

Durante el entrenamiento del sistema, las unidades que componen la frase son conocidas. Por lo tanto, el módulo DP realiza un simple alineamiento temporal. Si llamamos

$$\mathcal{L}(t,l) \equiv \log P(O_1^t|\lambda_1, \lambda_2, \cdots, \lambda_l)$$

las ecuaciones de programación dinámica son:

$$\begin{array}{ll} \textit{Comienzo} & \mathcal{L}(1,1) = \log P(O_1^1|\lambda_1) \\ \text{Recursión} & \mathcal{L}(t,l) = \max\limits_{1\leq\Delta t\leq N}\{\mathcal{L}(t-\Delta t, l-1) + \log P(O_{t-\Delta t+1}^t|\lambda_l)\} \\ \textit{Final} & \log P(O_1^T|\lambda_1, \lambda_2, \cdots, \lambda_L) = \mathcal{L}(T,L) \end{array}$$

Este alineamiento puede ser utilizado para el entrenamiento de los módulos SLHMM y LVQ.

Durante la fase de reconocimiento, el módulo DP realiza un proceso muy similar al Viterbi Beam Search debido al hecho de que no se conocen ni el número de unidades que componen la secuencia ni, obviamente, cuáles son dichas unidades. Así, las cantidades a evaluar son de la forma $\mathcal{L}(t,l,\lambda_l)$ dado que dependen del modelo usado. El procedimiento de reconocimiento es el que sigue:

- Dado un valor conocido para $\mathcal{L}(t,l,\lambda_l)$ se evalúan todos los posibles valores $\mathcal{L}(t+\Delta t, l+1, \lambda_{l+1})$ de acuerdo a una gramatica preestablecida

$$\mathcal{L}(t+\Delta t, l+1, \lambda_{l+1}) = \mathcal{L}(t,l,\lambda_l) + \log P(O_t^{t+\Delta t}|\lambda_{l+1}$$

- La inicialización de los valores se hace de la forma

$$\mathcal{L}(1,1,\lambda_i) = \log P(O_1^1|\lambda_i)$$

- Una vez que se ha alcanzado el estado final de la gramatica y el instante de final de la secuencia de entrada, T, se selecciona el valor máximo para $\mathcal{L}(T, L_i, \lambda_{final})$ y se reconstruye el camino óptimo.

5 Resultados experimentales.

El sistema propuesto está siendo evaluado actualmente. Como ya se ha comentado, el módulo LVQ ha sido suprimido en esta fase inicial. Los valores iniciales de los parámetros de todos los SLHMM han sido estimados a partir de un entrenamiento de los HMM de acuerdo con las técnicas habituales para su aplicación a voz continua, con la esperanza de que de esta forma se disminuirá el número de iteraciones necesarias para el entrenamiento del sistema. Los resultados experimentales que se muestran han sido obtenidos a partir de los mismos,

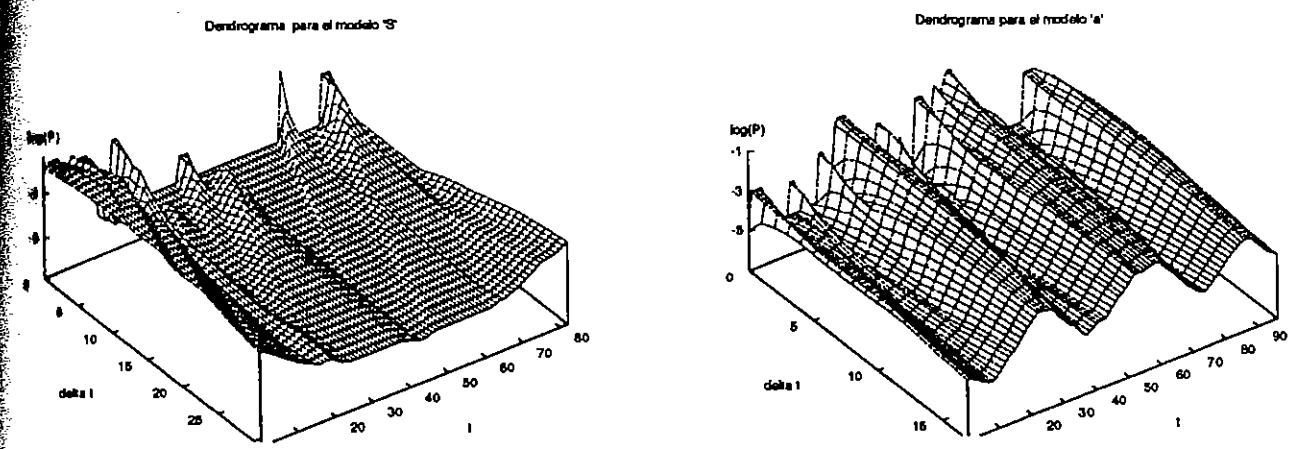

Figura 3: Ejemplos de los dendrogramas obtenidos para varios de los modelos para la frase /dameagua/.

no habiéndose realizado, por tanto, ningún entrenamiento posterior de los modelos mediante las técnicas propias de las redes neuronales.

La evaluación preliminar del sistema se ha realizado sobre un conjunto de 200 frases pronunciadas cada una por 4 locutores (800 frases en total), que han sido divididas en un conjunto de entrenamiento y otro de evaluación. El conjunto de entrenamiento está compuesto por 150 frases, mientras que el de evaluación lo está por las 50 frases restantes. Cada una de las frases ha sido muestreada a 8 kHZ y convenientemente parametrizada de acuerdo a los valores del Cepstrum, ΔCepstrum, Energía y ΔEnergía [3].

En la figura 3 se muestran las probabilidades obtenidas para el fonema /a/ y el silencio en la frase /dameagua/, siendo Δt la duración del intervalo temporal considerado, t el instante inicial del segmento y $log(P)$ el logaritmo de la probabilidad de generación del segmento. En la figura 4 se muestran los valores máximos correspondientes a cada uno de los instantes de tiempo iniciales para todos los fonemas. En dicha gráfica puede observarse el buen comportamiento del sistema, ya que no aparece ningún fonema de los no existentes en la frase con una probabilidad por encima del umbral fijado en la gráfica, y el fonema que localmente proporciona un máximo para la probabilidad se encuentra, por lo general, en la posición adecuada. Unicamente se observan dos pequeños problemas, que son fácilmente eliminados por el módulo de programación dinámica, encargado de proporcionar la secuencia final de fonemas reconocida. Estos dos problemas corresponden a una nasalización de la /d/ inicial (aparece una zona con alta probabilidad para la /m/) y al valor obtenido para la /a/ final de la frase, ya que la el modelo asociado al fonema /u/ proporciona mayor valor para la probabilidad.

Los tasas de reconocimiento obtenidas se muestran en la tabla 1 tanto para el sistema de reconocimiento de voz continua basado en HMM como para el sistema propuesto. En dicha tabla puede observarse que si no se realiza poda (100% de umbral), ambos sistemas proporcionan la misma tasa de error. Esto era de esperar, puesto que los valores de los parámetros de los modelos son los mismos en ambos casos. Sin embargo, para valores similares para el umbral de poda, los resultados sobre la tasa de error son ligeramente favorables para el sistema propuesto. Esta ventaja se hace más evidente si se considera el

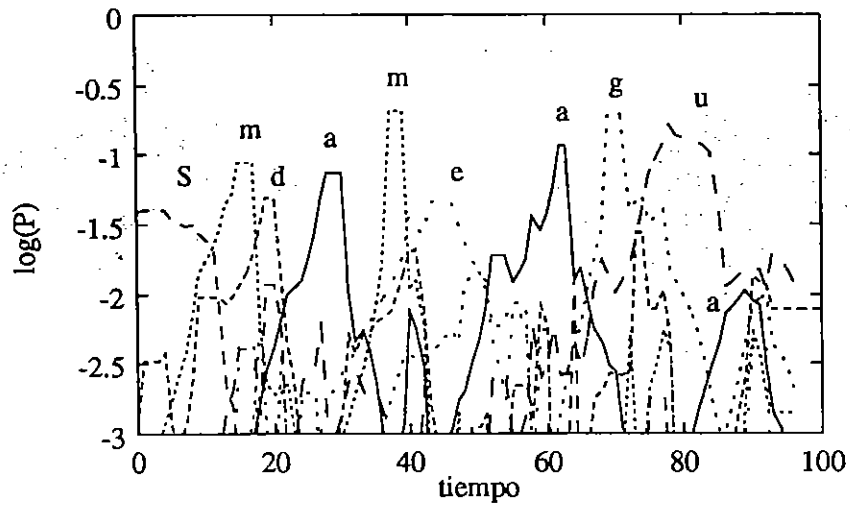

Figura 4: Valores maximos de la probabilidad para los fonemas de la frase /dameagua/.

HMM			SLHMM		
Umbral	Num. nodos	Error	Umbral	Num. nodos	Error
10	8275	34.1	12	3256	32.79
20	16920	9.16	20	13735	8.12
100	—	0.0	100	—	0.0

Tabla 1: Resultados experimentales sobre el conjunto de frases de evaluación.

número de posibles candidatos parciales evaluados (número de nodos), que es bastante más reducido en el caso del sistema propuesto. Esto es consecuencia del propio proceso de poda, ya que en el caso del reconocedor basado en HMM, la decisión de podar o no un camino se hace de forma síncrona, esto es, cada vez que se procesa un vector de entrada, se decide qué candidatos son seleccionados. Por el contrario, en el sistema basado en SLHMM esta poda se realiza de forma asíncrona: la decisión de eliminar o no un posible candidato se realiza atendiendo a las probabilidades de fonemas completos reconocidos.

6 Conclusiones.

Se ha propuesto un sistema de reconocimiento de voz continua que integra modelos ocultos de Markov y redes neuronales cuyo núcleo es el denominado SLHMM. Un SLHMM es una red neuronal que puede ser interpretada según un HMM. El sistema integra también los módulos denominados LVQ que actúan como un cuantizador, con la ventaja de que su estructura permite la utilización de un algoritmo de entrenamiento que optimiza el rendimiento global del sistema. Finalmente, se utiliza un módulo de programación dinámica que obtiene el alineamiento óptimo para la secuencia reconocida.

Un aspecto importante de los SLHMM consiste en la posibilidad que presentan de modelizar la duración de los fonemas. Si durante el entrenamiento de dicho bloque se permite que los pesos de las interconexiones entre las diferentes capas puedan ser diferentes de una capa a otra, el propio algoritmo de entrenamiento obtendrá los valores óptimos de acuerdo

n las posibles duraciones de los fonemas.

eferencias

J. Díaz, "A new neuron model for an Alphanet-Semicontinuous HMM," *Proc. ICASSP'93*, Vol.1, pp. 529–532, 1993.

J. S. Bridle, "Alpha-Nets: A Recurrent 'Neural' Network Architecture with a Hidden Markov Model Interpretation," *Speech Communications*, Vol.9, pp. 83–92, 1990.

J. C. Segura, "Variantes del Modelado Oculto de Markov para Señales de Voz," in *Monografías del Dpto. de Electrónica*, vol. 24. Dept. de Electrónica y Tecnología de Computadores. Universidad de Granada, Diciembre 1991.