TECNIACUSTICA
BARCELONA • 1996

# Sistema de Descodificación Acústico-Fonética basada en el uso de redes neuronales

*F. Freitag, E. Monte, P. Pachès-Leal*

*Departamento de Teoría de la Señal y Comunicaciones*
*Universidad Politécnica de Catalunya*
*C/Gran Capità, s/n, 08034 Barcelona*
*E-mail felix(gps.tsc.upc.es*

**ABSTRACT**

In this paper we present a phoneme recognition system based on neural networks. The neural networks are used to predict observation vectors of speech frames. For this task we have applied both feed-forward and recurrent neural networks. The prediction error is used as distortion measure in a Viterbi decoding step. The performance of the system is evaluated on both the training database and the test database. Experimental results on the training database are similar to a four state HMM, results on the test database are comparable to a three state HMM.

**INTRODUCCION**

La mayoría de los sistemas para el reconocimiento del habla se basa en el uso de modelos ocultos de Markov (HMMs). Una de las ventajas de esta técnica es que su estructura permite tanto el modelado local del habla como el alineamiento temporal de la señal. Por otra parte, se han obtenido resultados de reconocimiento comparables a los de sistemas basados en HMMs con sistemas, en los cuales redes neuronales se utilizaban para estimar probabilidades a posteriori de fonemas [1].

Una forma alternativa de aplicar redes neuronales para el reconocimiento del habla es de usarlas para la predicción de vectores de observación, como se ha propuesto en [2], [3] y [4]. En este trabajo utilizamos redes neuronales de tipo feed-forward y redes neuronales recurrentes (red Elman) para la predicción de vectores de observación. En comparación con las redes feed-forward, las redes neuronales de tipo Elman incluyen información sobre tramas pasadas en sus conexiones recurrentes, lo cual les puede permitir modelar mejor la trayectoria de los vectores de observación. En la figura 1 se ve la arquitectura de una red

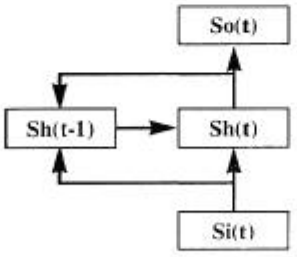


*Fig. 1: Arquitectura de una red Elman.*

Elman, donde *Si(t)* indica los estados de la capa de entrada, *Sh(t)* los estados de la capa oculta, y *So(t)* los estados de la capa de salida.

En la fase de descodificación acústico-fonética se utiliza el error de predicción de la red neuronal en cada trama como medida de distorsión. Aplicando el algoritmo de Viterbi se elige la secuencia de fonemas con el mínimo error acumulado como secuencia reconocida.

## IMPLEMENTACION

Se ha generado para cada fonema de la base de datos una red neuronal. Cada red realiza la predicción del vector de observación actual basándose en pasados vectores de observación. De esta manera, la entrada de una red neuronal consiste en uno ó más pasados vectores de observación, y la salida de la red representa el vector de observación predicho.

En la fase de entrenamiento las redes neuronales se entrenan con el algoritmo de backpropagation con el objetivo de minimizar el error de predicción

$$E = \sum_{t=1}^{T} (x(t) - \hat{x}(t))^2 \qquad 2.1$$

donde $T$ es el número de vectores de observación disponible para el entrenamiento, $x(t)$ representa el vector de observación actual y $\hat{x}(t)$ el vector de observación predicho por la red neuronal.

Realizando la predicción del vector de observación con una sola red neuronal por fonema corresponde a la estructura de un HMM de 3 estados, asumiendo que el primer y tercer estado actúan como entrada y salida, respectivamente, y el segundo estado modela la señal de voz. Se han utilizado redes neuronales de tipo feed-forward y de Elman. Las redes poseían una capa oculta con 25 neuronas. La función de activación de la capa oculta fue la sigmoide, la capa de salida tenía una función de activación lineal.

En la fase de reconocimiento se utilizaba el error de predicción de la red neuronal en cada trama como medida de distorsión. Utilizando el algoritmo de Viterbi, se eligió como secuencia de fonemas reconocidos la secuencia con el mínimo error acumulado.

Para el entrenamiento de las redes neuronales se utilizó una base de datos segmentada. La descodificación acústico-fonética se evaluó tanto con una base de test independiente como con la base de entrenamiento. Con el propósito de poder comparar los resultados obtenidos con el sistema que proponemos, se realizaron experimentos con un sistema basado en HMMs utilizando la misma base de datos. Los HMMs fueron de densidad continua con un número de 3 o 4 estados.

## BASE DE DATOS

La base de datos con la cual se realizó el entrenamiento fue la "Valencia segmentada". Esta base de datos consistía en 77 frases pronunciadas por 7 locutores distintos. Las frases fueron disponibles de forma segmentada en un total de 24 fonemas. De esta manera, un total de 2259 fonemas fueron utilizados para el entrenamiento de las redes neuronales.

Para el reconocimiento de fonemas se ha utilizado una parte de la base de datos Eurom como base de test. Esta parte consistía de 225 frases con 12928 fonemas. Los locutores y las frases en la base de test eran distintos a los de la base de entrenamiento.

La señal de la voz se ha parametrizado en coeficientes mel cepstrum (MFCCs) de orden 12. Una trama del señal era de 25 ms, y el movimiento de la trama era de 10 ms.

## RESULTADOS EXPERIMENTALES

En los experimentos de reconocimiento de fonemas hemos observado que se producía un alto número de inserciones de fonemas en la secuencia reconocida. Estas inserciones se produjeron tanto en la base de test como en la base de entrenamiento. Como consecuencia de las inserciones se ha reducido la medida de %accuracy aunque la medida %correct fue alta. Para disminuir el número de inserciones de fonemas en la secuencia reconocida hemos introducido una constante $th$ que se ha añadido al error de predicción de una red neuronal. De esta manera, un modelo $i$ se registra en la fase del reconocimiento en la trama $t$ si

$$E_i(t) + th < E_j(t) \qquad 4.1$$

donde $E_i(t)$ es el error de predicción del modelo $i$, $th$ es la constante que se ha determinado experimentalmente, y $E_j(t)$ es el error de predicción del modelo $j$ registrado anteriormente. La aplicación de 4.1 en los experimentos de reconocimiento ha reducido significativamente el número de inserciones.

En la tabla 4.1 se pueden ver los resultados de reconocimiento utilizando redes neuronales de tipo feed-forward. En el primer experimento de tabla 4.1 se han utilizado los vectores de observación ***x(t-1)*** y ***x(t-2)*** como entrada a la red neuronal. En el segundo experimento la entrada consistía únicamente del vector de observación ***x(t-1)***, lo cual puede significar una reducción de información para la red. Los resultados de los

dos experimentos, sin embargo, no indican una diferencia clara entre la primera y segunda elección de la entrada a la red. Por otra parte, se puede observar en la tabla 4.1 una diferencia importante entre las tasas de reconocimiento de la base de test y la de entrenamiento.

| FEED-FORWARD | | base de entrenamiento | base de test Eurom |
|---|---|---|---|
| 1. entrada 2 MFCC, salida 1MFCC | % correct | 69.01 | 40.72 |
| | % accuracy | 64.10 | 32.08 |
| 2. entrada 1 MFCC, salida 1 MFCC | % correct | 72.42 | 40.97 |
| | % accuracy | 69.37 | 33.01 |

**Tabla 4.1:** Resultados del reconocimiento de fonemas usando redes neuronales feed-forward.

La tabla 4.2 recoge los resultados de reconocimiento obtenidos con redes de tipo Elman. El primer experimento de tabla 4.2 se realizó en condiciones idénticas al primer experimento de tabla 4.1. Una comparación de los resultados obtenidos en estos experimentos indica prestaciones parecidas en los dos tipos de redes neuronales en la tarea dada. En el segundo experimento de tabla 4.2, en vez del error instantáneo para cada trama, se ha utilizado el promedio del error de predicción en dos tramas seguidas como medida de distorsión. Comparando con el primer experimento de tabla 4.2 se obtenía una mejora en el % accuracy de la base de test. Los resultados de reconocimiento obtenidos con la base de entrenamiento son - igual a los resultados obtenidos con el sistema basado en redes feed-forward - significativamente mejor que los resultados obtenidos con la base de test.

| ELMAN | | base de entrenamiento | base de test Eurom |
|---|---|---|---|
| 1. entrada 2 MFCC, salida 1MFCC | % correct | 73.35 | 44.63 |
| | % accuracy | 66.27 | 28.02 |
| 2. entrada 2 MFCC, salida 1 MFCC, | % correct | 70.69 | 43.39 |
| promedio de error | % accuracy | 65.07 | 32.01 |

**Tabla 4.2:** Resultados del reconocimiento de fonemas usando redes neuronales de tipo Elman.

En las tablas 4.3 se han recogido los resultados obtenidos con HMMs de densidad continua. Se puede comprobar que respecto a la base de entrenamiento se han obtenido con el sistema basado en redes neuronales resultados comparables a un sistema de HMMs de 4 estados. Los resultados del sistema basado en redes neuronales con la base de test, por otra parte, son parecidos a un sistema de HMMs de 3 estados. Como demuestran las tablas 4.1 y 4.2, se ha obtenido una diferencia importante de las prestaciones en la base de test y en la de entrenamiento utilizando el sistema basado en redes neuronales. Según los resultados experimentales en tabla 4.3, en el sistema basado en HMMs que hemos utilizado esta diferencia de prestaciones es menor.

| HMM | 3 estados | base de entrenamiento | base de test Eurom |
|---|---|---|---|
| 1 mezcla | % correct | 56.22 | 43.15 |
| | %accuray | 47.50 | 32.18 |
| 3 mezcla | % correct | 61.81 | 47.07 |
| | %accuray | 58.65 | 37.32 |

**Tabla 4.3a:** Resultados del reconocimiento de fonemas usando HMMs de densidad continua de 3 estados.

| HMM | 4 estados | base de entrenamiento | base de test Eurom |
|---|---|---|---|
| 1 mezcla | % correct | 64.40 | 44.63 |
| | %accuray | 57.24 | 37.52 |
| 3 mezcla | % correct | 72.60 | 48.87 |
| | %accuray | 68.48 | 42.40 |

**Tabla 4.3b:** Resultados del reconocimiento de fonemas usando HMMs de densidad continua de 4 estados.

## CONCLUSIONES

Hemos presentado un sistema basado en redes neuronales para la descodificación acústico-fonética, en el cual la medida de distorsión esta representada por el error de predicción. En los experimentos realizados no se ha observado una diferencia clara en prestaciones de los diferentes tipos de redes neuronales utilizadas. Se han obtenido en la base de entrenamiento resultados de reconocimiento comparables a un HMM de 4 estados, los resultados obtenidos en la base de test fueron parecidos a un HMM de 3 estados. Un próximo paso que probablemente incrementará las prestaciones del sistema puede incluir el modelado de cada fonema por varios estados. Así, la trayectoria de los vectores de observación puede representarse con más exactitud.

## REFERENCIAS


[1] N. Morgan, H. Boulard. *"Neural Networks for Statistical Recognition of Continuous Speech"*, Proc. of the IEEE, pp. 742-770, vol. 83, no. 5, May 1995.

[2] J. Tebelskis, A. Waigel, B. Petek, O. Schmidbauer, *"Continuous speech recognition using Linked Predictive Neural Networks"*, Proc. ICASSP, pp. 61-64, 1991.

[3] K. Na, J. Ryu, D. Chang, S.Chae, S. Ann, *"Recurrent neural prediction models for speech recognition"*, Proc. Europ. Conf. on Speech Communication and Technology, pp. 2213-2216, Madrid, September 1995.

[4] M. Paping, H. Marti, M. Renfer, *"Predictive connectionist speech recognition with a new discriminant learning algorithm"*, Proc. Europ. Conf. on Speech Communication and Technology, pp. 2193-2196, Madrid, September 1995.