RECONOCIMIENTO DE FONEMAS MEDIANTE REDES NEURONALES PREDICTIVAS

F. Freitag, E. Monte
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 predictive neural networks. The neural networks are used to predict observation vectors of speech frames. The prediction error obtained is used as distortion measure in a Viterbi decoding step. We evaluate the performance of the neural predictive networks on both the training database and the test database. Recognition results on the training database are similar to a four state HMM, results on the test database are comparable to a three state HMM. Different observation vector combinations are tested in the experiments.

1. INTRODUCCION

El reconocimiento del habla se suele realizar con sistemas basados en modelos ocultos de Markov (HMMs). Recientemente, se ha empezado investigar alternativas para modelar la señal de la voz. Se ha demostrado que resultados de reconocimiento que son comparables a sistemas basados en HMMs pueden ser obtenidos con sistemas basados en redes neuronales, en los cuales las redes neuronales se utilizan para estimar probabilidades a posteriori de fonemas [1]. Una aplicación alternativa de redes neuronales en sistemas de reconocimiento del habla es él de utilizarlas para la predicción de vectores de observación, como se ha demostrado en [2], [3] y [4]. Aplicando redes neuronales de esta manera, se utiliza el error de predicción como medida de distorsión en la decodificación con el algoritmo de Viterbi. La secuencia de modelos con el mínimo error de predicción acumulado se elige como secuencia de fonemas reconocidos.

2. IMPLEMENTACION

Para cada fonema de la base de datos se genera una red neuronal que realiza la predicción de un vector de observación utilizando uno o más vectores de observación. Así, la entrada de la red neuronal consiste en uno o más pasados vectores de observación y la salida de la red es el predicho vector de observación. Las redes neuronales se entrenan con el algoritmo de backpropagation para minimizar el error de predicción dado por

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

donde T es el número de vectores de observación disponibles en la base de entrenamiento, $x(t)$ representa el vector de observación y $\hat{x}(t)$ el vector de observación predicho por la red neuronal. De este modo, el entrenamiento de una red neuronal se consigue presentando a la red los vectores de observación de la base de entrenamiento y usando online backpropagation para minimiza el error de predicción.

Cada fonema se ha modelado con una red neuronal. Utilizando una red por fonema corresponde a una arquitectura de HMMs 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.

El tipo de red neuronal utilizado fue el perceptron multicapa con una capa oculta. El número de neuronas de la capa oculta fue 25. Las neuronas en la capa oculta tenían como función de activación la sigmoide. La función de activación de la capa de salida fue lineal.

En el algoritmo de Viterbi se utilizaba como medida de distorsión el error de predicción que se obtenía en cada trama. La secuencia de modelos con el mínimo error acumulado representaba la secuencia de fonemas reconocidos.

Las redes neuronales se entrenaban con una base de datos pre-segmentada. Después de entrenarlas con la base de datos segmentada, se las podía utilizar para segmentar una segunda base de datos. Teniendo la segunda base de datos segmentada, se podía seguir entrenando las redes iterando entre segmentación y entrenamiento.

Con el propósito de poder comparar los resultados obtenidos con las redes neuronales, se realizaron también experimentos con HMMs con la misma base de datos. Los HMMs utilizados fueron de densidad continua con un numero de 3 o 4 estados.

3. BASE DE DATOS

Para el entrenamiento de las redes neuronales se utilizó la base de datos "Valencia" segmentada. Esta base de datos consiste en 7 locutores que pronuncian 77 frases. Las frases son segmentadas y etiquetadas en 24 fonemas. En total, esta base de datos consiste en 2259 fonemas para el entrenamiento.

Se ha reservado una parte de la base de datos Eurom para un segundo entrenamiento de las redes y la otra parte de Eurom como base de test. Esta segunda parte consistía de 225 frases conteniendo un total de 12928 fonemas. Los locutores y las frases en la base de test son 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 era de 25 ms, y el movimiento de la trama era de 10 ms.

4. EXPERIMENTOS PRELIMINARES

Utilizamos en la fase de decodificación con el algoritmo de Viterbi el error de predicción de las redes neuronales como medida de distorsión, suponiendo que un modelo representado por una red neuronal entrenada tenga un error de predicción mas bajo en sus correspondientes tramas que en las tramas que pertenezcan a otros fonemas. Para evaluar la capacidad de discriminación de dicha medida de distorsión, se ha realizado un experimento preliminar en que todos los modelos se entrenaron con sus correspondientes tramas. Después del entrenamiento de los modelos se ha calculado el porcentaje de las veces con la cual los respectivos modelos obtenían el error de predicción mínimo en sus correspondientes tramas presentando las tramas también a todos los demás modelos. Este experimento se ha realizado con la base de datos "Valencia" segmentada. Las porcentajes de discriminación de los modelos en sus correspondientes tramas se presentan en Tabla 4.1. Como discriminación se considera la medida que se obtiene registrando cuantas veces el modelo correcto obtiene el mínimo error de predicción en sus correspondientes tramas respecto a los demás modelos. En este experimento se ha utilizado como entrada a cada red neuronal un vector de observación de 24 elementos consistiendo de MFCCs de orden 12 y sus correspondientes parámetros delta. La salida de la red ha sido 12 coeficientes MFCC de la siguiente trama.

modelo fonema	X	b	c	d	e	f	a	g	h	i	k	l
discriminación en %	81.81	33.82	55.12	41.53	53.40	72.61	67.70	34.89	38.04	73.26	50.00	56.99

modelo fonema	m	n	o	p	r	s	t	u	y	z	S	@
discriminación en %	57.86	62.34	57.77	40.60	37.77	77.23	35.14	58.55	38.51	48.45	52.80	40.95

Tabla 4.1: Discriminación en % de modelos en sus correspondientes tramas.

Otro experimento se ha realizado para visualizar como el error de predicción de una red neuronal varía en una frase entera. Como ejemplo se ha elegido la red neuronal que representaba el modelo del fonema "a", y que se ha entrenado con segmentos del fonema "a". El fichero f101_l1 de la base de datos "Valencia" se utilizaba para observar el error de predicción que se obtenía con la red neuronal del modelo "a" en toda la frase. En este fichero las tramas 21-27 y 140-147 corresponden a segmentos del fonema "a". En Fig. 4.1 se puede ver que el error de predicción del modelo "a" es relativamente bajo en sus correspondientes segmentos. La red neuronal utilizado ha sido la misma como en el experimento anterior.

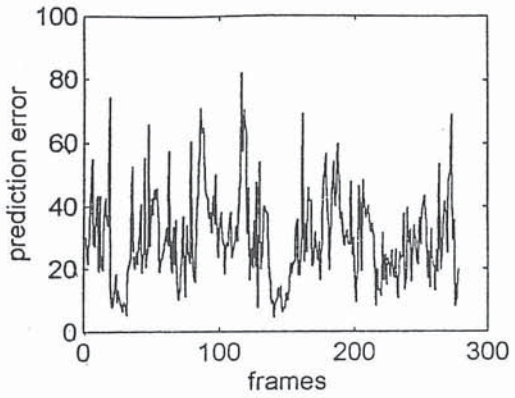

Fig. 4.1. Error de predicción de la red neuronal del fonema "a" en el fichero f101_11.

El siguiente experimento tenía como objetivo la observación de la distribución del error de predicción de un modelo en sus correspondientes tramas y la distribución del error de predicción de este modelo en tramas que pertenecen a otros fonemas. Se ha elegido para el experimento el modelo del fonema "b". En Fig. 4.2a se puede ver la distribución del error de predicción de la red neuronal del fonema "b" en todas las tramas de "b" en la base de datos "Valencia" segmentada. En la Fig. 4.2b se puede ver el error de predicción que se obtenía con el modelo de "b" en las tramas del fonema "d". Se puede observar que el error de predicción del modelo de "b" es más bajo en sus correspondientes tramas.

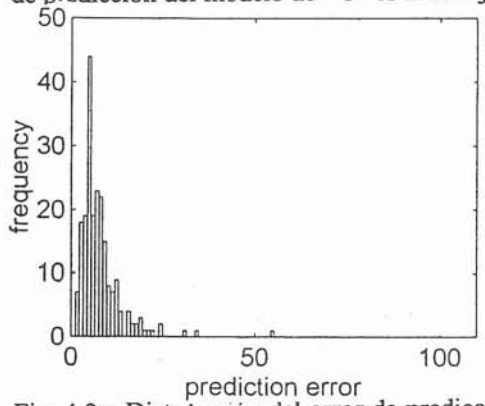

Fig. 4.2a: Distribución del error de predicción del modelo "b" en las tramas "b".

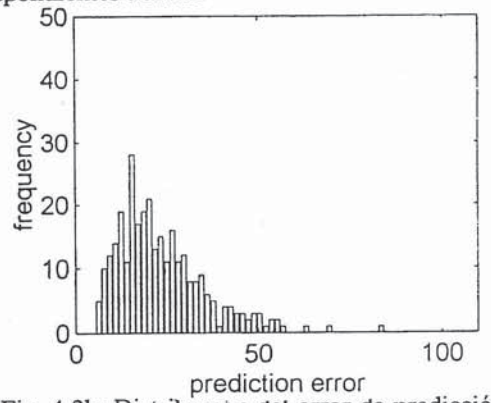

Fig. 4.2b: Distribución del error de predicción del modelo "b" en las tramas "d".

5. RESULTADOS EXPERIMENTALES

En los experimentos de reconocimiento de fonemas hemos observado que se produjeron un alto número de inserciones. Las inserciones se produjeron tanto en la base de test como en la base de entrenamiento. Estas inserciones han reducido la medida de % accuracy aunque la medida % correct era alta. Para reducir el número de inserciones hemos introducido una constante 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 un umbral que se ha determinado experimentalmente, y $E_j(t)$ es el error de predicción del modelo j registrado anteriormente.

En la Tabla 5.1 se pueden ver los resultados de reconocimiento. El tipo de red neuronal utilizado fue el perceptron multicapa. La capa oculta tenía 25 neuronas. La salida de la red eran los predichos coeficientes mel cepstrum de orden 12. La entrada de la red consistía en el primer caso de Tabla 5.1 de 2 vectores de observación, en los restantes casos se utilizaban 1 vector de observación. Las parametrizaciones

fueron coeficientes mel cepstrum, mel cepstrum más coeficientes delta, y mel cepstrum, coeficientes delta, energía y energía delta.

		base de entrenamiento	base de test Eurom
entrada 2 MFCC, salida 1MFCC	% correct	69.01	40.72
	% accuracy	64.10	32.08
entrada 1 MFCC, salida 1 MFCC	% correct	72.42	40.97
	% accuracy	69.37	33.01
entrada 1 MFCC-D, salida 1 MFCC	% correct	74.28	-
	% accuracy	67.60	-
entrada 1 MFCC-D-E, salida 1 MFCC	% correct	73.20	-
	% accuracy	67.20	-

Tabla 5.1: Resultados del reconocimiento de fonemas usando perceptrons multicapa.

HMM	3 estados	base de datos de entrenamiento	base de datas 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 5.2a: Resultados del reconocimiento de fonemas usando HMMs de densidad continua de 3 estados.

HMM	4 estados	base de datos de entrenamiento	base de datas 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 5.2b: Resultados del reconocimiento de fonemas usando HMMs de densidad continua de4 estados.

En Tabla 5.2 se pueden ver los resultados obtenidos con HMMs de densidad continua. Comparando Tablas 5.1 y 5.2 se puede comprobar que las prestaciones obtenidas con las redes neuronales predictivas en la base de entrenamiento son parecidas a las prestaciones de HMMs con 4 estados. Los resultados obtenidos en la base de test son parecidos a los resultados de HMMs de 3 estados. La diferencia de las prestaciones en la base de entrenamiento y en la base de test es mayor en las redes neuronales predictivas.

6. CONCLUSIONES

Hemos presentado un sistema para el reconocimiento de fonemas en que el error de predicción se utilizaba como medida de distorsión. Se han conseguido altas prestaciones en el reconocimiento de la base de entrenamiento, y en la base de test los resultados obtenidos son parecidos a los resultados de un HMM de densidad continuo de tres estados. Un paso que probablemente incrementará las prestaciones del sistema será modelar cada fonema por varios estados y modelar de esta forma con más exactitud la secuencia de los vectores de observación.

7. 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.