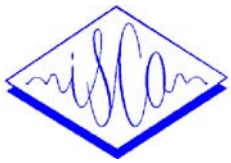

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

PREDICTIVE NEURAL NETWORKS APPLIED TO PHONEME RECOGNITION

F. Freitag, E. Monte, J. Salavedra
Polytechnic University of Catalunya
Department of Signal Theory and Communications
C/Gran Capità, s/n, E - 08034 Barcelona
E-mail: felix@gps.tsc.upc.es Fax: 34-3-4016447 Phone: 34-3-4016435

ABSTRACT

In this paper a phoneme recognition system based on predictive neural networks is proposed. Neural networks are used to predict observation vectors of speech frames. The obtained prediction error is used for phoneme recognition as 1) distortion measure on the frame level and 2) as feature, which is statistically modeled by the Rayleigh distribution. Continuous speech phoneme recognition experiments are performed different settings of the system are evaluated.

1. INTRODUCTION

An important part of continuous speech recognition systems is acoustic-phonetic decoding. The task of acoustic-phonetic decoding is to find the "best" representation of the acoustic signal in terms of a sequence of speech units. The "best" sequence is usually defined as that sequence which maximizes some similarity measure. Different approaches have been proposed for performing this task. Most often the similarity measure applied is based on statistical models. Then, the best recognized sequence is the most likely sequence. This approach based on Hidden Markov Models (HMMs) is used in most current speech recognition systems. In the HMM approach each speech unit w_i is represented by a HMM M_i. The HMM can be considered as a stochastic finite automata with symbol emission probabilities and transition probabilities associated with each state. In the recognition phase for an unknown observation O the probability of a speech unit w_i given the observation O is given according to Bayes´ rule by

$$P(w_i / O) = \frac{P(O / w_i) P(w_i)}{P(O)} . \qquad (1)$$

Being $P(w_i)$ a language dependent probability and being $P(O)$ the same for all speech units, the a posteriori probability $P(w_i / O)$ depends on the likelihood $P(O / w_i)$. For acoustic-phonetic decoding and when each speech unit w_i is represented by the model M_i, $P(O / w_i)$ is approximated by the probability $\hat{P}(O / M_i)$ of the most likely state sequence over state path Q of the model. The probability $\hat{P}(O / M_i)$ corresponds to the likelihood that the observation O is produced by the model M_i over state sequence Q. We can obtain the probability $\hat{P}(O / M_i)$ in log likelihood form as

$$\hat{P}(O / M_i) = log\, P(w_i) + \max_Q \left\{ \log(a_{Q(0)Q(1)} + \sum_{t=1}^{T} (\log(b_{Q(t)}(x(t)) + \log(a_{Q(t)Q(t+1)}) \right\} \qquad (2)$$

with

$$\log(b_{Q(t)}(x(t))) = \log \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} - \frac{1}{2}(x(t) - \mu_j)^T \Sigma^{-1} (x(t) - \mu_j)$$

being $b_{Q(t)}(x(t))$ the similarity measure on the frame level obtained by a multivariate Gaussian function and a_{ij} fixed transition probabilities between states. It can be observed in equation (2) that speech recognition results depend on 1) the accuracy of the statistical (typically Gaussian) models used as state models of the HMM for locally representing the distribution of the observation vector coefficients; 2) the discriminative capacity of the distortion measure when comparing different $\hat{P}(O / M_i)$, being $\hat{P}(O / M_i)$ essentially a sum of log probabilities obtained at the frame level; and 3) the quality of the used speech features [5]. Besides of the Gaussian models, other methods have also been presented in the past for obtaining the probability given in equation (1). Particularly successful has been the use of neural networks to estimate a posteriori probabilities [1]. In that approach recognition results comparable to state-of-the-art HMM systems have been obtained.

In this work the application of the prediction error of neural networks as distortion measure and its use for phoneme recognition is studied [2] - [4]. The neural networks are trained to map a previous observation vector into a current observation vector. This way even if static parameters without delta coefficient are used as features, the network will acquire knowledge about temporal relations between vectors due to the required mapping. It is supposed that the relation between two vectors, defined both by their location and their trajectory, is a sufficiently discriminative characteristic for being able to classify observation vector sequences into phonemes. In other words it can also be said the

characteristic of successive vector pairs is supposed to be unique for each phoneme.

2. SYSTEM IMPLEMENTATION

In the phoneme recognition system that we propose each phoneme of the database is represented by a model consisting of 3 or more states (Fig. 1). A model is given by one entry and one exit state, and at least one modeling state. Transitions between the modeling state are allowed from one state to itself or to the following state. A transition between states does not increase the value of the accumulated distortion measure. Each of the modeling states of the model consist of one neural network, which generates a distortion measure, based on its prediction of the current observation vector given one past observation vector. The obtained prediction error is the distortion measure that we use, obtained for each speech frame and for each state. For training the observation vector pairs *x(t-1)* and *x(t)* are presented to the network and the neural networks are trained with the backpropagation algorithm in order to minimize their prediction error.

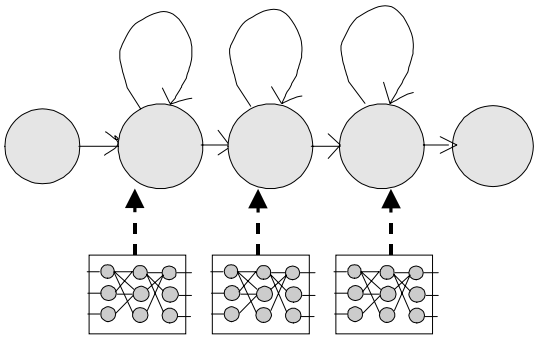

Fig. 1. 5 state phoneme model structure.

For phoneme recognition we substitute the term $log\, P(w_i)$ and the similarity measure $\log(b_{Q(t)}(x(t)))$ given in equation (2) by a phoneme dependent threshold *th*, and by the prediction error of the neural networks given as

$$E_{Q(t)}(t) = \frac{1}{N}\sum_{i=1}^{N} (x_i(t) - \hat{x}_i(t))^2 \qquad (3)$$

where N are the number of output neurons equal to the size of the observation vector, $x_i(t)$ is the *ith* coefficient of the observation vector at time *t* and $\hat{x}_i(t)$ is the *ith* predicted coefficient of the predicted observation vector. Thus, the system that we propose uses for recognition of a speech unit w_i for an unknown observation vector sequence of length *T* the discrimination measure

$$D = th + \min_{Q}\left\{\sum_{t=1}^{T} E_{Q(t)}(t)\right\}, \qquad (4)$$

where Q is the state path within a model which obtains minimum accumulated prediction error. The recognized phoneme sequence given an unknown sequence *O* is then that sequence of models M_i, which obtains the minimum accumulated prediction error.

It is known that the type of the features extracted from the speech signal can have an important influence on the recognition results [5]. Therefore, another way of using the prediction error of neural networks is also investigated, in which the prediction error of the neural networks is considered as a new feature to be modeled by an appropriate statistical model. In Fig. 2, the distribution of the average squared prediction error given in equation (3) for the phonemes “a”, “b”, “e” and “m” is shown, as well as a corresponding statistical model based on the Rayleigh distribution. It can be seen that the Rayleigh distribution can model the distribution of the average prediction error of the neural networks (Fig. 2). In order to use state models based on the Rayleigh distribution, we model the average prediction error of the neural networks with

$$p(r) = \frac{r}{\sigma^2}\exp\left(\frac{-r^2}{2\sigma^2}\right), \qquad (5)$$

where $\sigma = \dfrac{average\ \ prediction\ error}{\sqrt{\pi}/2}$

and $r = E$, with *E* given in equation (3).

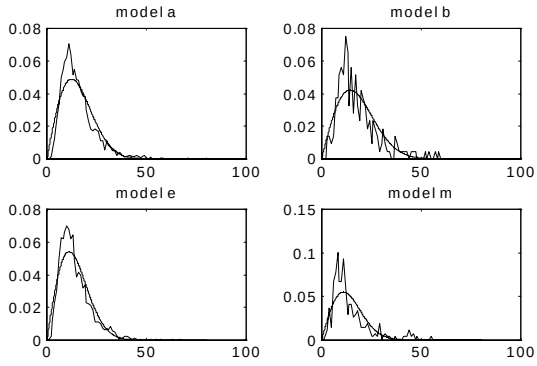

Fig. 2. Distribution of the average prediction error and the Rayleigh distribution models for the phonemes “a”, “b”, “e” and “m”.

For our experiments we used multilayer perceptrons (MLPs) with a hidden layer of 25 neurons. The activation function of the hidden layer was the sigmoid function.

The input to the network was a past observation vector at time *t-1*, the output of the network was the predicted observation vector for time *t*. The activation function of the output layer was linear. For reference purpose, a continuous density HMM system was also set up.

3. DATABASE AND TASK

For the initial training of the neural networks, the phonetically segmented part of the Spanish continuous speech database VALENCIA was used. This database consists of 7 speakers pronouncing 77 phrases. The continuous speech phrases are available with speech segmented into a total of 24 phonemes. Thus, the database which was used for training of the predictive neural networks consisted of 2259 training phonemes. A second training step was performed using a part of the database EUROM (EUROM_TR) consisting of 146 phrases containing 6397 phonemes. This part was phonetically segmented by the trained models. Then iteratively the nets were trained and the database was segmented again into phonemes. For the continuous speech phoneme recognition task another part of the database EUROM (EUROM_TE) was used as test database. This part consisted of 225 phrases containing a total of 12928 phonemes. This way, the speech material and the speakers in the test database were different to that of the training databases.

The sampled speech data in waveform of the databases was Hamming-windowed and parametrized into 12 liftered mel-frequency cepstral coefficients (MFCCs). Each parametrized speech frame represented a time window of 25 ms. A window shift of 10 ms was used.

4. PHONEME RECOGNITION RESULTS

In the first experiment the neural networks were fed with input vectors of 12 MFCCs and the distortion measure was calculated based on the predicted MFCC output vector of dimension 12. In this experiment the neural nets were trained only on the database VALENCIA. Phoneme recognition experiments were performed both on the test and training databases. For recognition both the prediction error and the Rayleigh similarity measure were tested as distortion measure on the frame level.

MFCCs MLP	**Training database VALENCIA**		**Test database EUROM_TE**	
	Prediction Error	**Rayleigh**	**Pred.Error**	**Rayleigh**
3 states				
%corr.	78.04	58.30	46.68	33.28
%acc.	71.67	52.28	26.71	29.51
4 states				
%corr.	73.79	66.31	44.28	32.37
%acc.	69.28	63.44	26.71	30.09
5 states				
%corr.	72.16	64.13	44.75	33.37
%acc.	70.47	61.84	32.08	31.21

Table 1: Phoneme recognition results of the system with MFCC input vector.

In Table 1 the results for the first experiment are given. It can be seen that when using the prediction error as distortion measure better recognition rates than for the Rayleigh distribution based measure are obtained. For the test database, however, the difference is smaller. A significant difference between the recognition results in the test and training database can also be observed. One reason for this difference might be that the test database is statistically not identical to the training database.

In the second experiment the nets were fed with input vectors of dimension 24 consisting of 12 MFCCs and delta coefficients (MFCC_D). Like in the first experiment the output vector are the predicted 12 MFCCs without deltas. As a difference to the first experiment the delta parameters as additional inputs contain temporal information, which may be useful for the net to better predict the current observation vector. Like in the first experiment, the prediction error is computed from the 12 predicted MFCCs. Delta coefficients are not predicted. In this second experiment the neural nets are trained on the database VALENCIA and on the database EUROM_TR. Phoneme recognition experiments are performed both on the test and training databases. For recognition experiments with the training database both the prediction error and the Rayleigh distortion measure are tested as distortion measure. For experiments with the test database, the prediction error is used as distortion measure.

MFCC_D	MLP	Training database VALENCIA	
		Prediction Error	Rayleigh
3 states	% correct	76.10	61,80
	% accuracy	72,07	57,77
4 states	% correct	77,38	61,22
	% accuracy	74,68	58,57
5 states	% correct	73,62	60,80
	% accuracy	71,85	58,88

Table 2: Phoneme recognition results of the system with MFCC_D input vector for the database VALENCIA.

MFCC_D	MLP	Training database EUROM_TR	Test database EUROM_TE
		Pred. Error	Pred. Error
3 states	% correct	55,90	45,36
	% accuracy	46,91	33,47
4 states	% correct	62,04	47,43
	% accuracy	51,03	35,25
5 states	% correct	56,57	47,08
	% accuracy	51,30	39,03

Table 3: Phoneme recognition results of the system with MFCC_D input vector for the database EUROM.

In Tables 2 and 3 the results of the second experiment can be seen. Better recognition rates can be observed for the training databases VALENCIA and EUROM_TR. Concerning EUROM_TR, better recognition results are obtained compared to EUROM_TE, although both belong to the same speech corpus. Therefore, a good adaptation of the neural nets to the training data can be observed. In fact, when the neural nets are trained on presented observation vectors, after a certain number of epochs they start adapting to the training material and loose the capacity of generalization. This effect can be observed during recognition. It can be observed that initially recognition results both on the training and test database improve during training, while after a certain number of epochs the results on the test database remain constant or decrease. Concerning the choice of the number of states in the models, better results for the test database EUROM_TE are obtained when the number of states was increased.

MFCC	HMM	Training database VALENCIA	Test database EUROM
3 states	% corr.	56.22	43.15
(1 mix.)	% acc.	47.50	32.18
4 states	% corr.	64.40	44.63
(1. mix.)	% acc.	57.24	38.52

Table 4: Phoneme recognition results of the HMM reference system.

In Table 4 the recognition rates obtained with a 1-Gaussian-mixture continuous density HMM system using MFCC parametrization for the same speech recognition task is given. It can be seen, that the recognition rates on the training database are below the rate obtained with the neural network based system, being on the other hand slightly better on the test database. It has to be mentioned, however, that the HMM system was trained on a significantly larger part of the database EUROM_TR. More training data was therefore used for the HMM system than for the neural network based system.

5. CONCLUSIONS

With the proposed neural network based system recognition rates comparable to simple continuous density HMM systems were obtained. The recognition results on the continuous speech task indicate that the prediction error of neural networks is a useful distortion measure, which carries discriminative information and which can be used in the Viterbi search during recognition. Considering the average prediction error as a feature, predictive neural networks can also be used to generate new features. For this feature the Rayleigh distribution models can be used on the state level to model the distribution of the average prediction error of the neural networks.

REFERENCES

[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 91*, pp. 61-64, Toronto, 1991.

[3] F. Freitag, E. Monte. “Acoustic-Phonetic Decoding based on Elman Predictive Neural Networks”. *Proc. ICSLP 96*, pp. 522-525, Philadelphia 1996.

[4] F. Freitag, E. Monte. “Statistical modeling of the prediction error of neural networks by means of the Rayleigh distribution, *Proc. NEURAP 97*, pp. 169-172, Marseille, March 1997.

[5] S. Furui. “Speaker-Independent Isolated Word Recognition using Dynamic Features of the Speech Spectrum. *IEEE ASSP*-34(1), pp. 52-59, February 1986.