ACOUSTIC-PHONETIC DECODING BASED ON ELMAN PREDICTIVE NEURAL NETWORKS[1]

F. Freitag, E. Monte

Universitat Politècnica de Catalunya
Department of Signal Theory and Communications
Barcelona
Spain

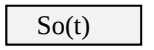
E-mail: felix@gps.tsc.upc.es

ABSTRACT

In this paper we present a phoneme recognition system based on the Elman predictive neural networks. The recurrent neural networks are used to predict the observation vectors of speech frames. Recognition of phonemes is done using the prediction error as distortion measure in the Viterbi algorithm. The performance of the neural predictive networks is evaluated on both the training database and on a speaker independent test database. The results obtained on the training database are similar to a four state continuous density HMM, results on the test database results are comparable to a three state HMM.

1. INTRODUCTION

Current speech recognition systems are usually based on Hidden Markov Models (HMMs). Recently, research has also focused on alternative ways for modeling the speech signal. It was shown that speech recognition results comparable to the results using HMMs can be obtained by approaches, in which neural networks are used to estimate posterior probabilities of phonemes [1]. Another approach is to use neural networks as predictors of observation vectors of speech frames as shown in [2], [3], and [4]. When working as predictors, neural networks map past observation vectors into a predicted observation vector, and the prediction error is used in a Viterbi decoding. In this paper we propose to use recurrent neural networks as predictors of speech frames. Unlike feed-forward networks, the recurrent architecture permits to include context information about past frames in form of previous hidden layer states in the recurrent connections. Thus, feature trajectories theoretically can be modeled in a better way. In the recognition phase the prediction error is used as distortion measure and the model sequence with the minimum accumulated prediction error is selected after a dynamic programming phase. In Fig. 1.1 the architecture of an Elman neural network is shown, where *Si(t)* indicates the states of the input layer, *Sh(t)* the states of the hidden layer and *So(t)* the states of the output layer.



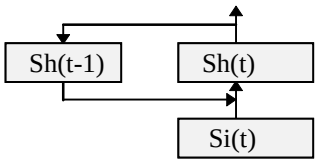


Fig. 1.1: Architecture of an Elman neural network.

2. IMPLEMENTATION

For each phoneme of the database one neural network is used to predict the current speech frame given a number of past speech frames. In this way, the input of the network is represented by the observation vectors of the past frames and the output of the neural network is the predicted observation vector of the current frame. We have chosen to use the observation vectors **x**(t-1) and **x**(t-2) as input to the network in order to predict the observation vector **x**(t). The neural networks are trained with backpropagation in order to minimize the prediction error given as

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

where T is the number of observations vectors available for training, $x(t)$ being the observation vector and $\hat{x}(t)$ being the predicted observation vector. Training of the Elman neural networks is done presenting the corresponding speech frames of the training database and using online backpropagation to minimize the prediction error.

An important point in training of neural networks is the aspect concerning the behavior of the trained network to unseen data. The capacity of generalization of the network can decrease after a number of training epochs in spite of a decrease of the prediction error on the training data set. Generally, overtraining of the neural network can be avoided if a third database for cross validation next to the training and test database is used to evaluate the

[1] This work was supported by the TIC 95-1022-C05-03 grant.

performance of the neural network continuously. The number of adjustable parameters of the neural network, which is strongly related to the size of the network, can also influence in the capacity of generalization. If the size of the network is too large for the required task or if the training pattern set is small, neurons can become inactive or are badly trained. On the other hand, a network that is too small for a task might not be able to perform the required mapping. We have chosen for the experiments a hidden layer of 25 neurons. The activation function of the hidden layer was the sigmoid function. The activation function of the output layer was linear.

In the recognition phase the prediction error centered on the mean prediction error of the training pattern set is used as distortion measure.

Using the Viterbi algorithm, the path with the smallest accumulated prediction error is selected to represent the recognized phoneme sequence.

The neural networks are initialized on a pre-segmented database. The obtained trained neural network models are then used to obtain the segmentation of a second labeled but unsegmented database. After the segmentation of this database with the initially trained models, the database can also be used for training of the neural network models.

3. DATABASE AND EXPERIMENTAL SETTINGS

Both the neural network models and the continuous density HMMs use a database containing continuous Spanish speech for initial training. This database consists of 7 speakers pronouncing 77 phrases. The phrases are available in segmented form with speech labeled into a total of 25 phonemes. Thus, the database used for training consisted of 2259 training phonemes.

For testing and additional training the continuous speech database Eurom is used. The speech material is available in labeled form, but is not segmented. The Eurom database is divided into two parts: The first part after segmentation is also used for training of the models. The second part is used for testing. The segmentation of the first part of the Eurom database, which is necessary for training of the neural network models, is obtained by the Viterbi algorithm and the initialized neural network models. With the segmentation obtained the initialized models can be further trained. The second part of the Eurom database used for testing consists of 225 phrases containing a total of 12928 phonemes. The speech material and the speakers in the test database are different to the training database.

Speech data was parametrisized into 12 liftered mel-frequency cepstral coeficients (MFCCs) without delta coeficients. The analysis window is 25ms and the window shift 10ms.

Each phoneme is modeled by one neural network. The architecture of the neural networks which is seen during recognition with the Viterbi algorithm (when the neural network models provide the prediction error as distortion measure) corresponds to a HMM with 3 states (supposing that the second state is modeling the speech signal, and the first and last state act as input and output states, respectively).

For reference purpose, experiments with continuous density HMMs were also performed on the same database. Each HMM represented one phoneme. The experiments were done with HMMs of 3 and 4 states. The observation vector used is also of 12 MFCCs without delta coeficients.

4. EXPERIMENTAL RESULTS

In a first stage, preliminary experiments were done with feed-forward networks to evaluate the capacity of generalization of the neural networks. After a certain number of training epochs simultaneous recognition of the training and test database was done. Table 4.1 illustrates that although the recognition rates on the training database improved, the recognition of the test database was rougly constant or decreased after a certain number of epochs. This behavior can indicate that after a certain number of epochs the neural networks become overtrained on the training patterns and loose capacity of generalization, which decreases the recognition rate on the test database.

It was observed in the experiments that in the recognition phase a high number of insertions was introduced by the neural networks. This fact was observed both in the training and test database. Insertions reduce the % accuracy measure of the phoneme recognition while the % correct measure is high. To reduce the number of insertions, we introduced a penalty added to the prediction error in the path obtained by the Viterbi algorithm before recording a new model. This modification has the effect that a new phoneme model is only recorded if the difference between the error of the previously recorded model and the current model is below a certain threshold. Thus, a new model i in the recognition phase is recorded at frame t if

MLP	epochs	20	30	40	50	60	70
TRAIN DB	% correct	59.06	60.63	61.73	62.19	63.53	63.98
	% accuracy	52.57	54.36	55.48	56.82	57.49	58.17
TEST DB	% correct	40.73	40.84	40.73	40.05	39.36	39.59
	% accuracy	31.97	33.33	34.81	34.58	33.11	34.58

Table 4.1: Phoneme recognition results obtained with a feed forward neural network after different training epochs.

$E_i(t) + th < E_j(t)$

where $E_i(t)$ is the prediction error of the model *i* to be recorded, *th* is an experimentally determined threshold, and $E_j(t)$ is the prediction error of the previously recorded model *j*.

Applying 4.1 in the recognition experiments has shown to reduce the number of insertions significantly.

In order to further eliminate insertions, another modification was introduced in which the average prediction error of the current and previous frame of a neural network model is used as distortion measure during recognition instead of the instantaneous prediction error. In Table 4.2 the recognition rates are shown when using the average prediction error. In Table 4.3 the recognition rates obtained using the instantaneous prediction error can be seen. Using average prediction error slightly improves the % accuracy measure. It also can be observed that there is a significant difference in performance between the recognition of the test and training data.

Elman neural network	Training database	Eurom test database
% correct	70.69	43.39
% accuracy	65.07	32.01

Table 4.2: Phoneme recognition with Elman neural network using average prediction error.

Elman neural network	Training database	Eurom test database
% correct	73.35	44.63
% accuracy	66.27	28.02

Table 4.3: Phoneme recognition with Elman neural network using the instantaneous prediction error.

A third experiment was done in which in the recognition phase the duration of each model was forced to be at least two frames. This modification can be considered to be equivalent to a 4 state topology in which, since calculated by the same neural network model, the prediction errors of the second and third states (which model the speech signal) are equal and the transitions are allowed to the following state. In Table 4.4 the recognition results are given.

Elman neural network	Training database	Eurom test database
% correct	70.03	44.35
% accuracy	63.70	29.84

Table 4.4: Phoneme recognition with Elman neural using minimal duration of 2 frames.

The results shown in Tables 4.2 - 4.4 obtained with the Elman neural networks can be compared with reference results obtained using continuous density HMMs on the same databases. In Tables 4.5 the performance of continuous density HMMs on the same test and training database are given.

HMM	3 states	Training database	Eurom test database
1 mixture	% correct	56.22	43.15
	%accuray	47.50	32.18
3 mixtures	% correct	61.81	47.07
	%accuray	58.65	37.32

Table 4.5a : Phoneme recognition using a 3-state continuous density HMM.

HMM	4 states	Training database	Eurom test database
1 mixture	% correct	64.40	44.63
	%accuray	57.24	37.52
3 mixtures	% correct	72.60	48.87
	%accuray	68.48	42.40

Table 4.5b : Phoneme recognition using a 4-state continuous density HMM.

Comparing the results obtained with the Elman neural networks shown in Table 4.2 - 4.4 and the results obtained with the continuous density HMMs, it can be seen that the performance of the Elman neural networks on the training database is similar to the performance of the 4 states 3 mixture HMM. Concerning the test database, the performance of the Elman neural network is similar to the performance of the 3 state 1 mixture HMM. The difference in performance of the HMMs in training and test database is less than the difference in performance of the neural networks on the two databases.

It has to be mentioned concerning the above given results that the HMMs, in contrast to the Elman neural networks, were also trained on the training part of the Eurom database. It was noticed, however, that in the case of the 3 state 1 mixture HMMs additional training did not change recognition results significantly. When the 3 state 1 mixture HMMs were trained only on the pre-segmented database, a recognition rate of 41.34 % correct and 32.93 % accuracy on the test database was obtained. This result is similar to the one given in Table 4.5 when the HMM is also trained on the training part of Eurom. Additional training of the Elman neural networks on the training part of the Eurom database also did not change the recognition rates significantly. We obtained 41.31 % correct and 34.41 % accuracy. Therefore, it is assumed that in order to improve the performance on the test database a better way is to provide additional features that are more robust to different speakers and speech material rather than using additional training of the models.

5. CONCLUSIONS

A phoneme recognition system using the prediction error obtained with Elman neural networks was presented. With this approach a high recognition rate on the training database was obtained. On the test database the recognition rate obtained was similar to a

three state HMM. Using the neural networks a significant difference in recognition performance between the training and test database was observed. The approach based on neural predictive networks can be further developed beyond the work presented in this paper experimenting with some modifications which probably can improve the recognition rates. To account in a more accurate way to the time-invariant nature of speech, one or more networks can be used to predict various states of the observation vector sequence of phonemes. Improvements probably can also be achieved using for discrimination in the Viterbi algorithm distortion measures different to the instantaneous prediction error. Another aspect which can be further considered is the choice of the input data vector to 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*, 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.