Computing and Informatics, Vol. 30, 2011, 321–334

MODULAR ECHO STATE NEURAL NETWORKS IN TIME SERIES PREDICTION

Štefan Babinec, Jiří Pospíchal

Department of Mathematics
Faculty of Chemical and Food Technology
Slovak University of Technology
812 37 Bratislava, Slovakia
&
Institute of Applied Informatics
Faculty of Informatics and Information Technologies
Slovak University of Technology
842 16 Bratislava, Slovakia
e-mail: stefan.babinec@stuba.sk, pospichal@fiit.stuba.sk

Manuscript received 30 March 2010; revised 30 November 2010
Communicated by Vladimír Kvasnička

Abstract. Echo State neural networks (ESN), which are a special case of recurrent neural networks, are studied from the viewpoint of their learning ability, with a goal to achieve their greater predictive ability. In this paper we study the influence of the memory length on predictive abilities of Echo State neural networks. The conclusion is that Echo State neural networks with fixed memory length can have troubles with adaptation of its intrinsic dynamics to dynamics of the prediction task. Therefore, we have tried to create complex prediction system as a combination of the local expert Echo State neural networks with different memory length and one special gating Echo State neural network. This approach was tested in laser fluctuations and turbojet gas temperature prediction. The prediction error achieved by this approach was substantially smaller in comparison with prediction error achieved by standard Echo State neural networks.

Keywords: Echo State neural networks, recurrent neural networks, time series prediction, gating of artificial neural networks

Mathematics Subject Classification 2000: 68T05

1 INTRODUCTION

From the point of information transfer during processing, neural networks can be divided into two types: feed-forward neural networks and recurrent neural networks [4]. Unlike the feed forward networks, recurrent neural networks contain at least one cyclical path, where the same input information repeatedly influences the activity of the neurons in a cyclical path. The advantage of such networks is their close correspondence to biological neural networks, but there are many theoretical and practical difficulties connected with their adaptation. The common problem of all such networks is the lack of an effective supervised training algorithm. The problem of difficult training was largely overcome by Echo State neural networks, where only weights to output neurons are trained. That can be achieved in one step [2, 3]. On one hand their application bypasses a problem of efficient training, but on the other hand, by imposing an echo-state property we restrict the ESN recurrent dynamics to contractions, making it less general (unlike fully trained recurrent neural networks, ESN cannot learn, e.g. the context-free grammar [13]). A very fast algorithm is used in these networks consisting of a calculation of one pseudo-inverse matrix, which is a standard numerical task.

In this paper we studied the influence of the memory length on predictive abilities of these special neural networks. We have found that Echo State neural networks (ESN) with fixed memory length can have troubles with adaptation of its intrinsic dynamics to dynamics of the prediction task. Therefore we suggest to create complex prediction system as a combination of the local expert Echo State neural networks with different memory length and one special gating ESN. The increase of number of the Echo State neural networks in the whole prediction system does not mean any substantial increase in computational demands due to very fast training algorithm. The advantage of this approach is that we get higher flexibility and better quality of prediction.

Connection between "liquid state" computing, related to echo states, and backpropagation was mentioned previously in [8, 12]. In our previous work [9, 10, 11] we explored a possibility to improve "one-step" learning by evolutionary approaches and Anti-Oja's learning. Mixtures of local expert neural networks and its application in time series prediction can be found in [1, 5, 6].

2 ECHO STATE NEURAL NETWORK

Echo State neural networks are atypical in architecture and training of recurrent neural networks (RNN). This new approach leads to a fast, simple and constructive supervised learning algorithm for the RNN. The basic idea of ESN is an application of a huge reservoir, as a source of dynamic behavior of a neural network, from which neural activities are combined into the required output.

The activity of hidden layer neurons in an RNN is further denoted as $\mathbf{x}(n) = (x_1(n), x_2(n), \ldots, x_N(n))$, where $x_i(n)$ is the output of the i^{th} hidden neuron in time n, and N is the number of hidden neurons. Under certain conditions, each $x_i(n)$

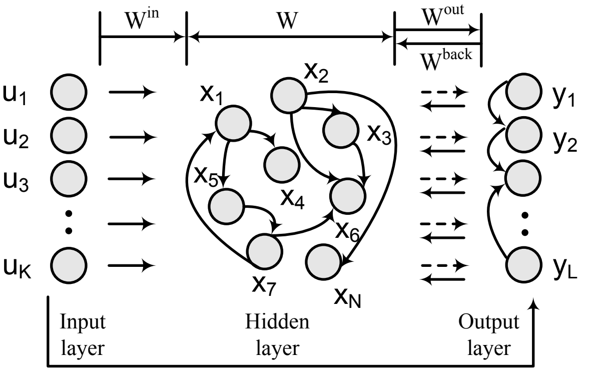

Fig. 1. The typical architecture of Echo State neural networks. The only weights which are trained in this architecture are the weights from the hidden to the output layer (dashed arrows).

is a function of the networks previous inputs $\mathbf{u}(n), \mathbf{u}(n-1), \ldots$, processed by the network. The input vector is denoted as $\mathbf{u}(n) = (u_1(n), u_2(n), \ldots, u_K(n))$, where $u_i(n)$ is the input of the i^{th} input neuron at the time n and K is the number of input neurons. Therefore, there exists such a function, E, so that:

$$\mathbf{x}(n) = E(\mathbf{u}(n), \mathbf{u}(n-1), \ldots). \tag{1}$$

Metaphorically speaking, the state of the neural network $\mathbf{x}(n)$ can be considered as an "echo", or in other words, a reflection of its previous inputs.

2.1 Description of the Neural Network

Neural network consists of K input, N hidden and L output neurons. The state of the neurons in the input layer at the time n is characterized by the vector

$$\mathbf{u}(n) = (u_1(n), u_2(n), \ldots, u_K(n)),$$

in the output layer by the vector

$$\mathbf{y}(n) = (y_1(n), y_2(n), \ldots, y_L(n)),$$

and in the hidden layer by the vector

$$\mathbf{x}(n) = (x_1(n), x_2(n), \ldots, x_N(n)).$$

The values of all the synaptic weights will be stored in matrices. An input weight matrix will be created: $\mathbf{W}^{in} = (w_{ij}^{in})$ of size $N \times K$, a weight matrix between hidden neurons: $\mathbf{W} = (w_{ij})$ of size $N \times N$, a matrix of output weights: $\mathbf{W}^{out} = (w_{ij}^{out})$ size of $L \times (K + N + L)$, and a matrix of weights from the output back to the reservoir: $\mathbf{W}^{back} = (w_{ij}^{back})$ size of $N \times L$. It is notable that, in this type of network, both direct input-output weights, as well as the recurrent weights between output neurons are allowed.

The structure and topology of ESN can be adjusted according to their current task. It is not necessary, for example, to use sigmoid output neurons, back weights from the output layer to the reservoir may or may not exist (they were not used), and even the input neurons may not be used (they were used). The output neurons used in this application were with no loops and with $f = \tanh$ (hyperbolic tangent) activation function. We can find detailed description of the learning algorithm in [2, 7].

We will just introduce computation of activities of the internal and output neurons. The states of hidden neurons in "dynamical reservoir" are calculated by the formula

$$\mathbf{x}(n+1) = f(\mathbf{W}^{in}\mathbf{u}(n) + \mathbf{W}\mathbf{x}(n) + \mathbf{W}^{back}\mathbf{d}(n)), \tag{2}$$

where f is the activation function of hidden neurons. The states of output neurons are calculated by the formula

$$\mathbf{y}(n+1) = f^{out}(\mathbf{W}^{out}(\mathbf{u}(n+1), \mathbf{x}(n+1), \mathbf{y}(n)), \tag{3}$$

where f^{out} is the activation function of output neurons.

3 TESTING DATA

The goal of this paper is to compare the results achieved by original "one-step" learning algorithm with our new approach. We have used two different data sets.

The first data set represents exact laboratory measurements. This data set was composed of a time sequence of 1 000 samples of laser fluctuations data, and the quality of prediction was measured by an error of prediction in the next 100 steps.

The second data set was composed of a time sequence of 555 samples of turbojet gas temperature, and the quality of prediction was measured by an error in the next 30 samples. In addition there were also available turbine rotation and air temperature values.

A mean absolute percentage error (MAPE)[1] was used to measure the quality of prediction on these testing sets, where test values P_i^{real} and predicted values P_i^{calc} are used, and N is the number of couples of values (the length of the predicted time series):

$$MAPE = \frac{\sum_{i=1}^{N} \left| \frac{P_i^{real} - P_i^{calc}}{P_i^{real}} \right|}{N} \times 100. \tag{4}$$

[1] The MAPE error was used in order to allow a comparison with results of prediction on the predicted data achieved by other methods [9, 11], which also used MAPE error.

A mean squared error (MSE) and standard deviation (SD) were also used for the final evaluation of results:

$$MSE = \frac{\sum_{i=1}^{N}(P_i^{real} - P_i^{calc})^2}{N}. \tag{5}$$

$$SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(P_i^{calc} - y')^2}, \tag{6}$$

where y' is the mean (average) of predicted values.

4 MOTIVATION AND METHOD

One of the most important parameters influencing the predictive ability of Echo State neural networks is spectral radius (α) of synaptic weights matrix W. The appropriate choice of the spectral radius has crucial importance for the eventual success of the ESN training. This is because α is closely connected to the intrinsic timescale of the dynamics of the dynamic reservoir (DR). Small value of α means that we have a fast DR, large value of α (close to 1) means that we have a slow DR. The intrinsic timescale of the task should match the DR timescale. Standard settings of α lies in a range between 0.6 and 0.99 (based on empirical observations). Proper setting of this parameter is crucial for the prediction quality of the resulting ESN. It should be small for the fast training data dynamics and large for the slow training data dynamics. Typically, α needs to be hand-tuned by trying out several settings. The DR timescale depends exponentially on $1 - \alpha$, so e.g. settings of $\alpha = 0.99, 0.98, 0.97$ will yield an exponential speedup of DR timescale, not a linear one [2]. From the other point of view α is influencing the memory of the ESN. If the value of α is less than 1, we can use the response of DR in more complex ESN architecture. On the other hand, the neural network has the tendency to gradually forget information from the previous time steps. In other words, ESN has memory with exponential decrease of information.

With the change of the α parameter we are also changing the number of the previous inputs, which will have the influence on the current state of the DR. We have made the following simple experiments. Different Echo State neural networks were trained and used for the time series prediction. The only difference among individual ESN networks was in the α parameter value. In the end, every ESN network had different prediction accuracy of the whole testing set; but what is more important, not every sample from the testing set was predicted most accurately by the winning ESN. In other words, Echo State neural networks with different memory length differed in the prediction accuracy for the individual time steps of the whole testing set.

In Figure 2 we can see the comparison of a part of the laser fluctuations testing set for two Echo State neural networks with different α parameter. This part represents a quick change in the data set, which is very difficult for neural network to adapt to. As we can see, the Echo State neural network with α equal to 0.67 (faster dynamics) can better adapt to a sudden change in the data, but still the better prediction accuracy at the whole testing set was achieved with the Echo State neural network with α equal to 0.8.

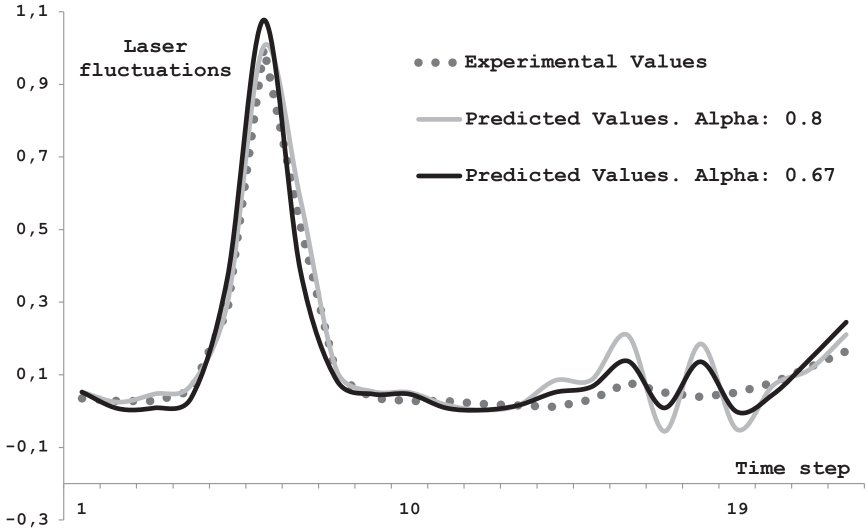

Fig. 2. Comparison of prediction accuracy of a part of the laser fluctuations testing set for two Echo State neural networks with different α parameter. The overall prediction accuracy was 29.52 % for the ESN with α 0.8 and 36.7 % for α equal to 0.67.

Thus our goal was to create a complex prediction system as a combination of the local expert Echo State neural networks with different α parameter and one special gating ESN. Local expert Echo State neural networks were used for the testing set prediction, but the task of the gating ESN was different. This neural network should determine which local expert ESN has the best prediction accuracy for the given time step of the testing set. The increase of number of the Echo State neural networks in the whole prediction system does not mean any substantial increase in computational demands, because the whole ESN training process is only computation of one pseudo inverse matrix. We can see the scheme of the whole prediction system in Figure 3.

5 EXPERIMENTS

The whole learning process was divided into three parts. The task of the first part was to find parameters of local expert ESN networks, which would be optimal for the quality of prediction on the laser fluctuations and turbojet gas temperature

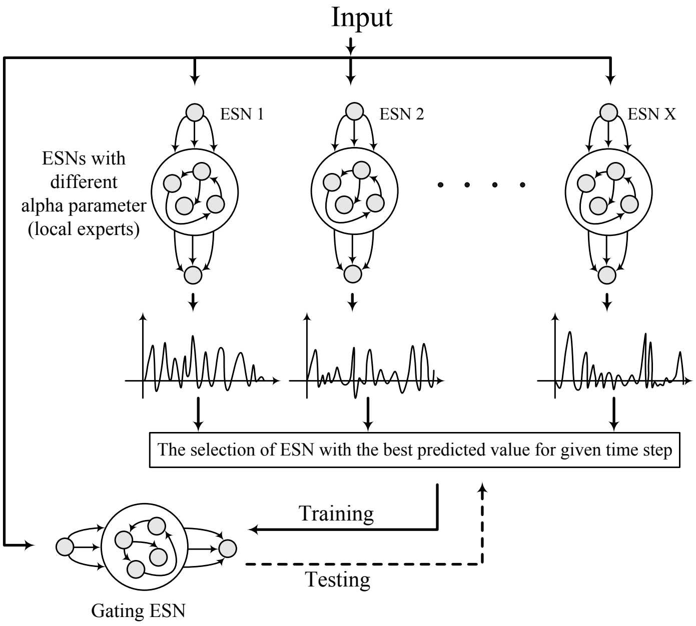
Input
ESN 1
ESN 2
ESN X
ESNs with different alpha parameter (local experts)
The selection of ESN with the best predicted value for given time step
Training
Testing
Gating ESN

Fig. 3. The scheme of the whole prediction system

sets. The achieved results should also serve for comparison with results achieved in the third part. The second part of experiments was aimed at finding the best parameters of gating Echo State neural network and prepares the whole system for the final third part. The third part of experiments was focused on the evaluation of prediction results, where the above mentioned gating ESN was used as global expert with simple task – finding the local ESN with the best prediction accuracy for a given time step.

5.1 The First Part of Experiments

As mentioned in Section 5, the first part involves two tasks: Finding best parameters of local expert ESN networks and obtaining prediction results. These results will be used later for comparison with the ones achieved in the third part.

The ESN consists of dynamic reservoir with one output neuron and one input neuron for laser data and three input neurons for turbojet gas temperature data. In the case of turbojet gas temperature, the first input represents gas temperature outgoing from the turbojet, the second input represents air temperature incoming to the turbojet and the third input represents rotation of the turbine.

The weight matrix between hidden neurons (**W**) should be sparse, to encourage rich variety of dynamics in dynamical reservoir. For that reason, only 2 % of all connections in dynamical reservoir were created. The network's hidden units are standard sigmoid units, with a transfer function $f = tanh$ (hyperbolic tangent) and the synaptic weights were initialized from interval $[-1, 1]$ with uniform distribution.

A considerable number of experiments was carried out, the representative results of which can be seen in Table 1 for laser data and in Table 2 for turbojet gas temperature data.

Index	*DR*	α	*Average MAPE*	*The best MAPE*
1	200	0.8	33.86 %	31.24 %
2	250	0.8	34.23 %	29.52 %
3	250	0.7	36.29 %	31.34 %
4	300	0.7	35.94 %	32.85 %

Table 1. Results of representative experiments in the first part: laser time series

Index	*DR*	α	*Average MAPE*	*The best MAPE*
1	250	0.8	2.93 %	1.81 %
2	300	0.7	2.70 %	1.78 %
3	300	0.8	2.64 %	1.50 %
4	350	0.8	3.58 %	2.33 %

Table 2. Results of representative experiments in the first part: turbojet gas temperature series

In these tables, DR represents the dynamic reservoir; α is the spectral radius of the weight matrix W, which is influencing the ability of the neural network to exhibit echo states. These DR and α values were chosen in accordance with the proposal used by Jaeger [2]. We have changed the number of neurons in DR from 50 to 1 000 with step equal to 25. The α value was changed from 0.59 to 0.99 with step 0.1. Experiments were carried out in the following way. For each value of DR and the parameter α, the values of synaptic weights in DR were randomly generated 1 000 times. This number was estimated to be large enough for statistical evaluation of prediction error on a testing set and for each initialization of weights the error for the testing set was calculated. Further, an average error of all 1 000 trials is presented in the columns *Average MAPE* (Table 1 and 2). Also, the smallest achieved error was recorded in the *Best MAPE* in the same tables.

A clear correlation between Best and Average value columns is apparent from Tables 1 and 2. When a better *Average MAPE* was achieved, there is also a better *Best MAPE.* This way we have found best parameters (number of neurons in DR and α) for each individual local expert. The best results for both prediction sets were achieved with α 0.8 and with DR consisting of 250 neurons for laser data and of 300 neurons for turbojet gas temperature data. We can see graphical representation of the best error for laser testing set in Figure 4 and for air temperature set in Figure 5.

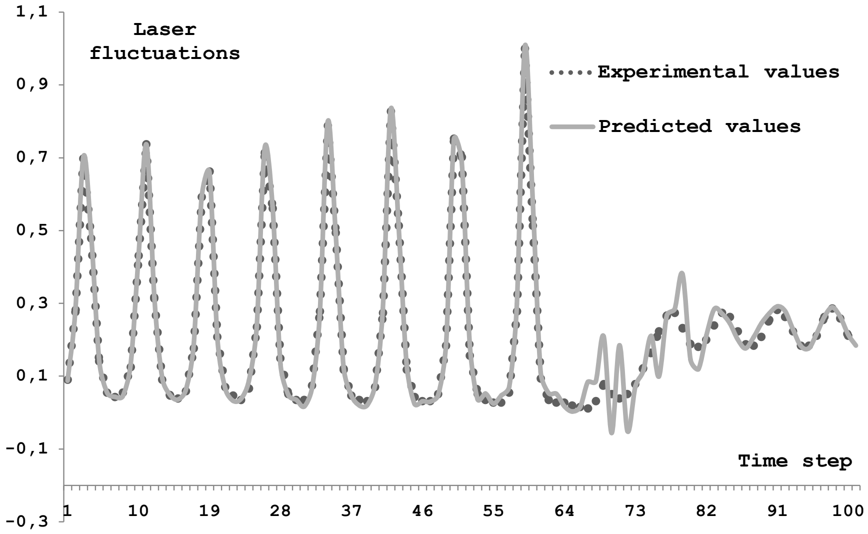

Fig. 4. Testing data: 100 records of laser fluctuations and 100 values predicted by standard Echo State neural network (MAPE 29.52 %).

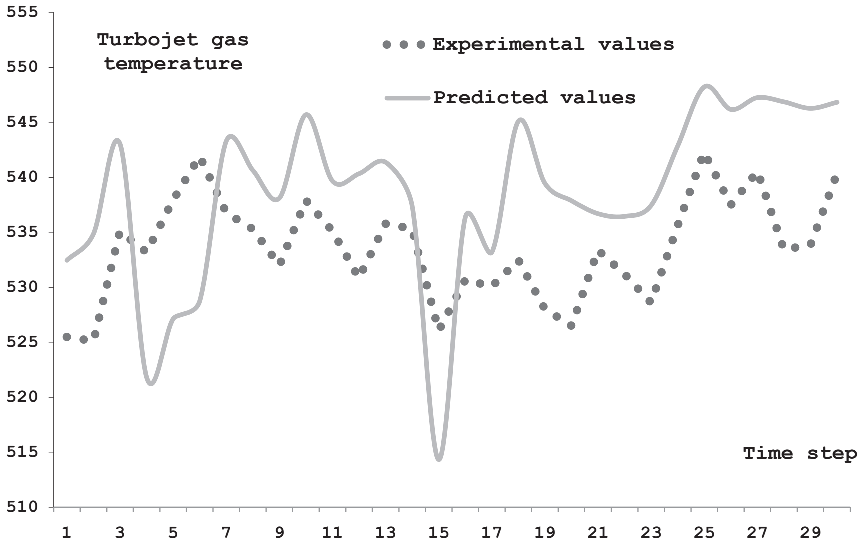

Fig. 5. Testing data: 30 records of turbojet gas temperature values and 30 values predicted by standard Echo State neural network (MAPE 1.5 %)

5.2 The Second Part of Experiments

In this part of experiments we have optimal parameters of the local expert ESN networks, in regard to the used laser fluctuations and turbojet gas temperature sets. These networks are already trained and ready to use in this second part.

Now we have to find the best parameters of the gating Echo State neural network. First, we had to create the training set for this special neural network. Therefore we had trained 40 ESN networks with different values of the α parameter. The rest of the parameter values and initialization of the synaptic weights of these neural networks were chosen based on the best results from the first phase of the experiments. The α parameter value of individual ESN ranged between 0.59 and 0.99 with sampling step 0.1.

The training of ESN is based on linear regression and one-step learning algorithm. Therefore, we have no possibility to catch the error signal during this learning process. So the local expert ESN networks were used for self prediction of the whole laser fluctuations and turbojet gas temperature sets and the time series obtained in that way reflects the quality of adaptation of each local expert to the training sets.

Afterwards, we were able to select the local expert ESN networks with the best prediction accuracy for the individual time steps of the training sets. The architecture and input of the gating ESN was the same as the one used for local expert ESN networks for individual testing sets. The only difference was in the desired output. In the case of gating ESN, this output was represented by the sequence number of the local expert ESN network with the best prediction accuracy for the following timestep. Thus the gating ESN had simple classification task.

A considerable number of experiments was carried out; the representative results are in the following Table 3 for laser and in Table 4 for turbojet gas temperature time series.

Index	*Size of DR*	*Alpha*	*Classification accuracy*
1	250	0.75	76.4 %
2	300	0.76	81.6 %
3	350	0.84	72.2 %

Table 3. Results of representative experiments in the second part: laser time series

Index	*Size of DR*	*Alpha*	*Classification accuracy*
1	200	0.74	63.6 %
2	270	0.81	76.4 %
3	300	0.86	67.8 %

Table 4. Results of representative experiments in the second part: turbojet gas temperature time series

Experiments in this part and description of attributes from Tables 3 and 4 are equivalent to experiments and description of attributes from the first part. Therefore, these descriptions will not be repeated again.

5.3 The Third Part of Experiments

The main experiments were carried out in the third phase with already trained local expert ESN networks and the gating ESN. All 40 local expert ESN networks were used for the prediction of the next 100 values for laser time series and 30 values for turbojet gas temperature series, which were not a part of the training sets (these values were used for the prediction quality measurements). After this process, for each training set we have created one time series as an average of all 40 time series predicted by local expert ESN networks. This time series served as an input into the gating ESN. The gating ESN was consequently used for the determination of the local expert ESN with the best prediction accuracy for individual time steps of the testing sets.

In the following Table 5 we can see the comparison of best achieved errors on testing data sets for standard Echo State neural network and our new approach – Modular Echo State neural networks. We can see graphical representation of these two approaches in Figures 6 and 7. It is clear from this table and figures that the modular approach can increase the quality of prediction considerably in comparison with standard Echo State neural network.

Testing data	*Original Approach* MAPE/MSE/SD	*Modular ESN approach* MAPE/MSE/SD
Laser fluctuations	29.42 %/1.18e–3/0.221	16.62 %/0.29e–4/0.218
Turbojet gas temperature	1.50 %/74.86/7.852	0.70%/19.04/7.120

Table 5. Results of experiments in the third part

6 CONCLUSIONS

Echo State neural networks are relatively new in the domain of neural networks. Their advantage is a closer connection with biological models inherent to recurrent neural networks and in their usage of the reservoir of dynamic behavior without adjusting the weights within the hidden layer. Echo State neural networks have a substantial advantage over other types of recurrent networks in their "one-step" learning ability.

However, there is no incremental approach further improving performance of an Echo State neural network. We decided to improve the performance by using a set of Echo State neural networks, each trained with different value of alpha parameter. This parameter controls the length of time, during which the old inputs can influence the output of the Echo State neural network, metaphorically "the

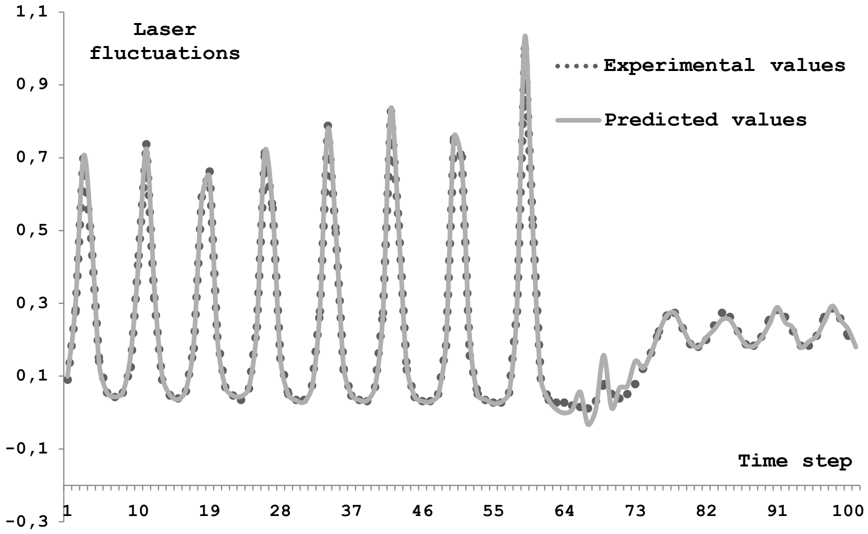

Fig. 6. Testing data: 100 records of laser fluctuations and 100 values predicted by Modular Echo State neural networks approach (MAPE 16.62 %)

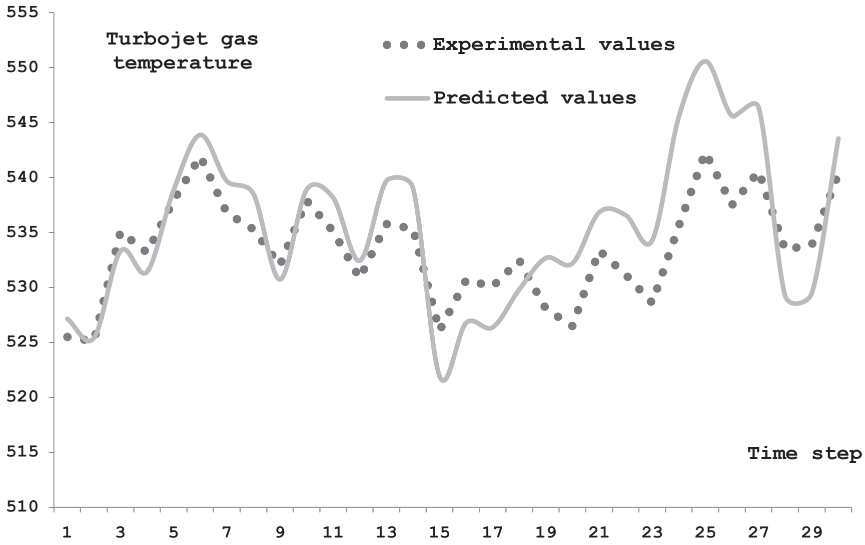

Fig. 7. Testing data: 30 records of turbojet gas temperature values and 30 values predicted by Modular Echo State neural networks approach (MAPE 0.70 %)

length of the Echo". Gating Echo State neural network was trained to decide, in which part of the dynamic phase the modeled system is. In each time step of the test, this gating network selected from the set the best Echo State neural network trained with suitable alpha parameter. The answer of this selected network was used as a response in the current time step.

We have chosen laser fluctuations and turbojet gas temperature as the testing data. Our aim was to find out whether this approach is able to increase prediction quality in comparison with original Echo State neural networks. From the results shown in the paper, it is clear that this aim has been accomplished. Modular Echo State neural networks approach can increase the quality of the network's prediction.

Acknowledgement

This work was supported by Scientific Grant Agency Vega of Slovak Republic under grants 1/0141/10 and 1/0804/08.

REFERENCES

[1] Jacobs, R.—Jordan, M.—Nowlan, S.—Hinton, G.: Adaptive Mixtures of Local Experts. Neural Computation, Vol. 3, 1991, pp. 79–87.

[2] Jaeger, H.: The Echo State Approach to Analysing and Training Recurrent Neural Networks. German National Research Center for Information Technology, GMD report 148, 2001.

[3] Lukosevicius, M.—Jaeger, H.: Reservoir Computing Approaches to Recurrent Neural Network Training. Computer Science Review, Vol. 3, 2009, No. 3, pp. 127–149.

[4] Haykin, S.: Neural Networks – A Comprehensive Foundation. Macmillan Publishing, New York 1994.

[5] Jordan, M.—Jacobs, R.: Hierarchical Mixtures of Experts and the EM Algorithm. Neural Computation, Vol. 6, 1994, pp. 181–214.

[6] Carvalho, A.—Tanner, M.: Mixtures-Of-Experts of Autoregressive Time Series: Asymptotic Normality and Model Specification. IEEE Transactions on Neural Networks, Vol. 16, 2005, No. 1, pp. 39–56.

[7] Jaeger, H.: Short Term Memory in Echo State Networks. German National Research Center for Information Technology, GMD report 152, 2002.

[8] Natschlager, T.—Maass, W.—Markram, H.: The "Liquid Computer": A Novel Strategy for Real-Time Computing on Time Series. Special Issue on Foundations of Information Processing of TELEMATIK, Vol. 8, 2002, No. 1, pp. 39–43.

[9] Babinec, Š.—Pospíchal, J.: Optimization in Echo State Neural Networks by Metropolis Algorithm. In R. Matousek, P. Osmera (Eds.): Proceedings of the 10th International Conference on Soft Copmputing, Mendel 2004, VUT Brno Publishing 2004, pp. 155–160.

[10] Babinec, Š.—Pospíchal, J.: Two Approaches to Optimize Echo State Neural Networks. In R. Matousek, P. Osmera (Eds.): Proceedings of the 11th International Conference on Soft Computing, Mendel 2005. VUT Brno Publishing 2005, pp. 39–44.

[11] Babinec, Š.—Pospíchal, J.: Improving the Prediction Accuracy of Echo State Neural Networks by Anti-Oja's Learning. Lecture Notes in Computer Science, Vol. 4668, 2007, pp. 19–28.

[12] Goldenholz, D.: Liquid Computing: A Real Effect. Technical report, Boston University Department of Biomedical Engineering, 2002.

[13] Čerňanský, M.—Tiňo, P.: Comparison of Echo State Networks with Simple Recurrent Networks and Variable-Length Markov Models on Symbolic Sequences. Lecture Notes in Computer Science, Vol. 4668, 2007, pp. 618–627.

Štefan Babinec received his M. Sc. degree in artificial intelligence in June 2003 from Technical University of Košice, and the Ph. D. degree in applied informatics from Faculty of Chemical and Food Technologies at Slovak University of Technology, Bratislava in 2007. Now he is working at the Department of Mathematics at the same university as Assistant Professor. He has co-authored 26 works in journals and conferences related to artificial neural networks and evolutionary algorithms.

Jiří Pospíchal received a diploma degree in physical chemistry from University of Jan Evangelista Purkyně in Brno, Czech Republic in 1984, and the Ph. D. degree in chemistry from Faculty of Chemical and Food Technologies at Slovak University of Technology, Bratislava, in 1990. From 1990 till now, he held positions at Slovak University of Technology, Bratislava, as Assistant Professor and Associate Professor of Applied Informatics at Department of Mathematics, Faculty of Chemical and Food Technologies, and as Professor of Applied Informatics at Institute of Applied Informatics at Faculty of Informatics and Information Technology. His research interests are evolutionary algorithms, artificial intelligence, cognitive science, neural networks, mathematical chemistry and graph theory.