linear prediction, LPC, neural networks, Kohonen, covariance, nonfluency, speech, detection, perceptron, radial

Adam KOBUS[1], Wiesława KUNISZYK-JÓŹKOWIAK[1], Elżbieta SMOŁKA[1], Ireneusz CODELLO[1]

SPEECH NONFLUENCY DETECTION AND CLASSIFICATION BASED ON LINEAR PREDICTION COEFFICIENTS AND NEURAL NETWORKS

The goal of the paper is to present a speech nonfluency detection method based on linear prediction coefficients obtained by using the covariance method. The application "Dabar" was created for research. It implements three different methods of LP with the ability to send coefficients computed by them into the input of Kohonen networks. Neural networks were used to classify utterances in categories of fluent and nonfluent. The first one was Kohonen network (SOM), used to reduce LP coefficients representation of each window, which were used as input data to SOM input layer, to a vector of winning neurons of SOM output layer. Radial Basis Function (RBF) networks, linear networks and Multi-Layer Perceptrons were used as classifiers. The research was based on 55 fluent samples and 54 samples with blockades on plosives (p, b, d, t, k, g). The examination was finished with the outcome of 76% classifying.

1. INTRODUCTION

The basic idea of linear prediction is based on the fact that consecutive samples of voice signal do not change rapidly [4]. The distance between two adjoining samples is quite small, thus the next sample can be approximated with the previous p samples. This idea is expressed by the following equation [6]:

1.1. BASIC EQUATION OF LINEAR PREDICTION

$$\tilde{s}(n) = \sum_{k=1}^{p} \alpha_k s(m-k) \tag{1}$$

where $\tilde{s}(m)$ denotes a m'th value of predicted voice sample, $s(m)$ - m'th value input voice sample, p - prediction order and α_k are obtained coefficients characteristic for a given signal.

When we assume that the signal is a unitary impulse, the current sample is a linear combination of previous samples. Minimizing error of sample predicted from preceding samples and real current sample is a linear prediction goal.

The application "Dabar" was created to compute these linear prediction coefficients. It is able to compute LP coefficients with the use of the Levinson-Durbin, covariance and Burg methods, saving them to a file, by sending them into the input of Kohonen network or visualizing them.

Operating on neural networks is necessary for future reflections. Some neural networks were selected and research was based on them.

Kohonen networks are Self-Organizing Maps. They have no given response pattern, so their aim is to classify patterns without a teacher, e.g. detecting concentrations. In such networks each neuron is connected with all the elements of the normalized input data vector. In each network learning epoch, such a neuron is chosen from the neurons in the output layer whose values of weight are nearest to input data elements. Then in the neighbourhood of the winning neuron, adaptation with the use of the Kohonen rule (2) takes place[3].

[1] Institute of Computer Science, Marie Curie-Skłodowska University, Pl. M. Curie-Skłodowskiej 1, 20-031 Lublin, Poland. Corresponding author: e-mail address: kobus.adam@gmail.com.

1.2. KOHONEN RULE

$$\boldsymbol{w}_i^{(n+1)} = \boldsymbol{w}_i^{(n)} + \eta_i^{(n)}(\boldsymbol{a} - \boldsymbol{w}_i^{(n)}) \tag{2}$$

where $\boldsymbol{w}_i^{(n)}$ denotes the weight on connection with the i'th element in the n'th epoch, $\boldsymbol{a}$ is the input data vector and η_i is the neighbourhood coefficient decreasing with the distance from the winning neuron.

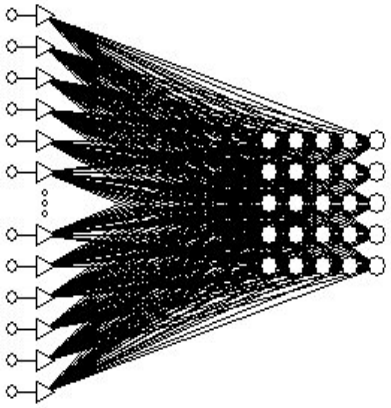

Fig. 1. Example of Kohonen network.

Multi-Layer Perceptrons (MLP) are one-direction neural networks with more than one neuron in each layer and required connections only between each neuron from one layer and each neuron from neighbour layer. Such networks are constructed with the aim of searching the best approximation of any function. The most common learning method for such networks is the Back Propagation method. Its goal is to minimize the mean-square error between the expected values and output values of the network by weight modifications[8].

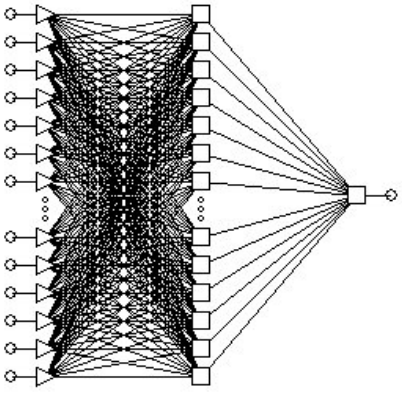

Fig. 2. Multi-Layer Perceptron network example.

The Radial Basis Function network (RBF) has a structure similar to MLP but only one input, one hidden and one output layer. Another difference is the activation functions which are radial, mainly Gaussian functions:

1.3. GAUSSIAN FUNCTION

$$\varphi_i = e^{\frac{|\boldsymbol{a}-\boldsymbol{c}_i|^2}{2\sigma_i^2}} \tag{3}$$

where $\boldsymbol{a}$ denotes the input vector, $\boldsymbol{c_i}$ denotes the centre of function and σ_i – dispersion. The weights on connections between hidden layer and input layer are constant and equal 1. There are a lot of applications of these networks. They can classify, approximate functions and learning the algorithm is simple [2].

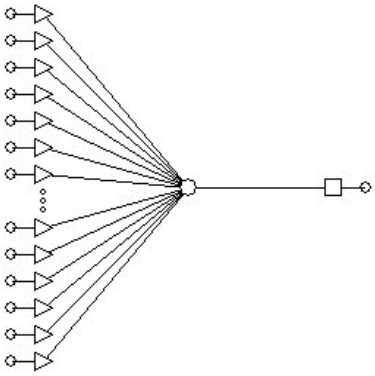

Fig. 3. Radial Basis Function network example.

Linear networks have the simplest structure of all the neural networks. They are built on two neuron layers: input and output. They are fast and they have good results for simple dependencies, e.g. white noise detecting. However, they fail when facing more complex problems [8].

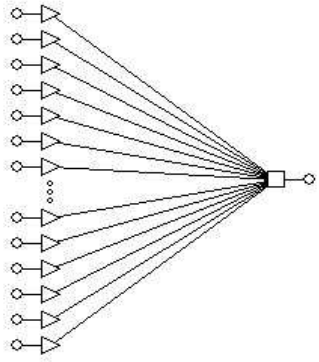

Fig. 4. Linear network example.

2. METHODOLOGY

2.1. RESEARCH CONCEPT

The aim of the research was to test the possibilities of automatic speech detection in 4sec utterances. Each utterance, saved in a *wave* format, was split over non-overlapping windows with 512 samples (ca 0.23s). Each window was multiplied by Hann function in order to improve the precision of

future computations of the linear prediction coefficients. Another step was to represent each window by 15 linear prediction coefficients obtained with the use of the covariance method.

The next step was a reduction of these coefficients to one number. In such a way, the output of the reduction from 4sec utterances was a 169-length vector of numbers, where one number represented one 0.23s window. For that purpose, a Kohonen network was build which gets 15 LP coefficients as input data for each of the 169 windows and returns a matrix of neurons with one winning neuron which represents that window.

The last step was to establish 169 values of winning neurons for an utterance as input of neural network such as the perceptron or Radial Basis Function, and to obtain information on the output concerning whether that utterance is fluent or nonfluent (it has blockades). Three Kohonen networks were examined as reductors and ten Radial Basis Function networks, linear networks and Multi-Layer Perceptrons were examined as classifiers. The research was finished by determining the best network type to solve that problem.

2.2. MATERIALS AND PROCEDURE

The classification of speech in categories of fluent and nonfluent was based on three basic blocks: speech signal acquisition and initial processing, signal parameterization and classification[5]. It is the same process as in speech recognition systems.

In the analysis, four-second utterances were used in a *wave* file format. Eight nonfluent and nine fluent speakers took part in the experiment. Nonfluent recordings contained blockades on plosives p, b, t, d, k and g at the beginning of the word. One woman, four men and three children – one girl and two boys – were examined. Blockade lengths were between 200 to 3000 ms.

The fluent recordings contained the same phrases as the nonfluent recordings. Four women and five men participated in this part.

The material was recorded by Creative Wave Studio based on SoundBlaster card with 22050Hz frequency on 16-bits for sample. Four-second samples were cut out with the same tools.

Each utterance was split into windows with 512 samples and multiplied by Hann function[1]:

2.3. HANN FUNCTION

$$w(n) = 0.5\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right) \tag{4}$$

where N denotes the size of a window, thus it is 512 samples. This function was selected as a result of the analysis of the frequency spectrum obtained from the computed linear prediction coefficients. The results of the application of that function were the best, thus, the values of coefficients were the most accurate.

In the next stage, 15 linear prediction coefficients were computed for each window. Then, each 512-sample window was represented by 15 linear prediction coefficients. The covariance method was used for these computations[6].

All the vectors of coefficients were used as input data vectors for the Kohonen network. The aim was to reduce each of these vectors possibly to one number. Three Kohonen networks with quadratic output layers were used for that purpose. The sizes of these output layers were: 5x5, 6x6 and 7x7. The aim was to test the performance of nonfluency detection dependent on the range of representation[7].

The Kohonen networks learnt in 100 epochs. The neighbour coefficient was set to 1 and the learning rate was set to 0.1 in each epoch. Increasing the number of epochs did not improve the quality of utterance modelling. The network was initialized uniformly within a range of 0 to 1. The conclusion of this observation was that each of the examined networks allows to model fluent utterances and to show the parts with blockades.

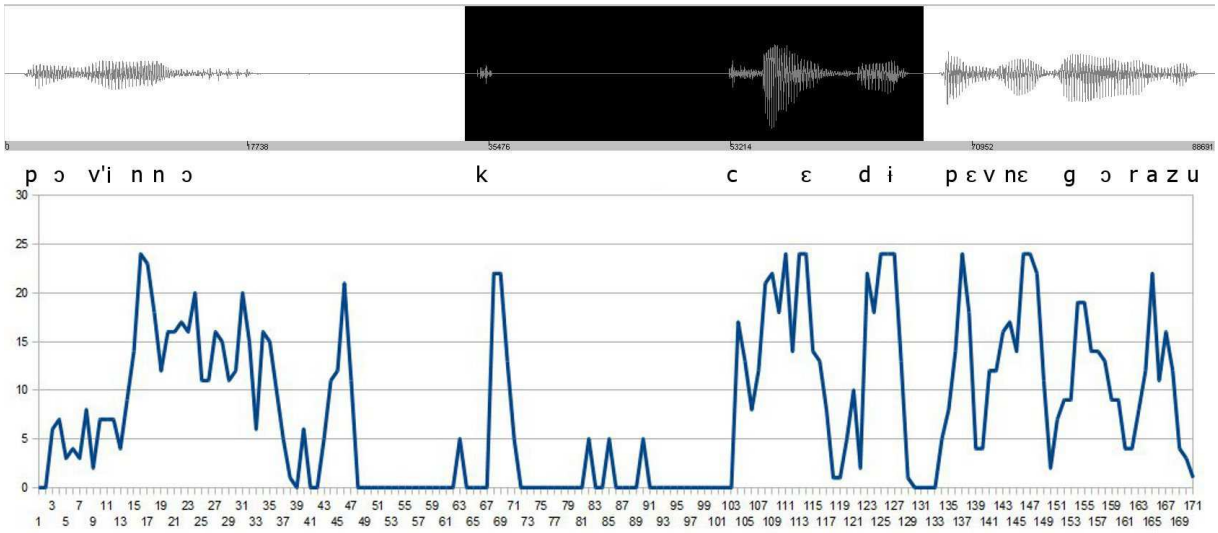

Fig. 5. Nonfluent utterance with marked word with blockade on start with winning neurons graph for Kohonen network 5x5.

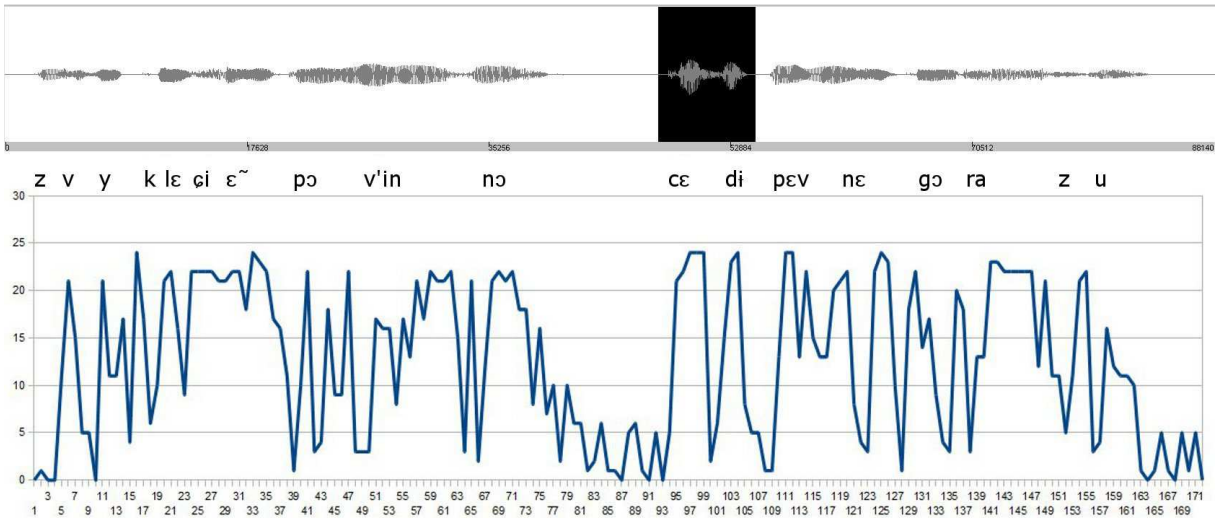

Fig. 6. Fluent utterance with the word marked with the winning neurons graph for Kohonen network 5x5.

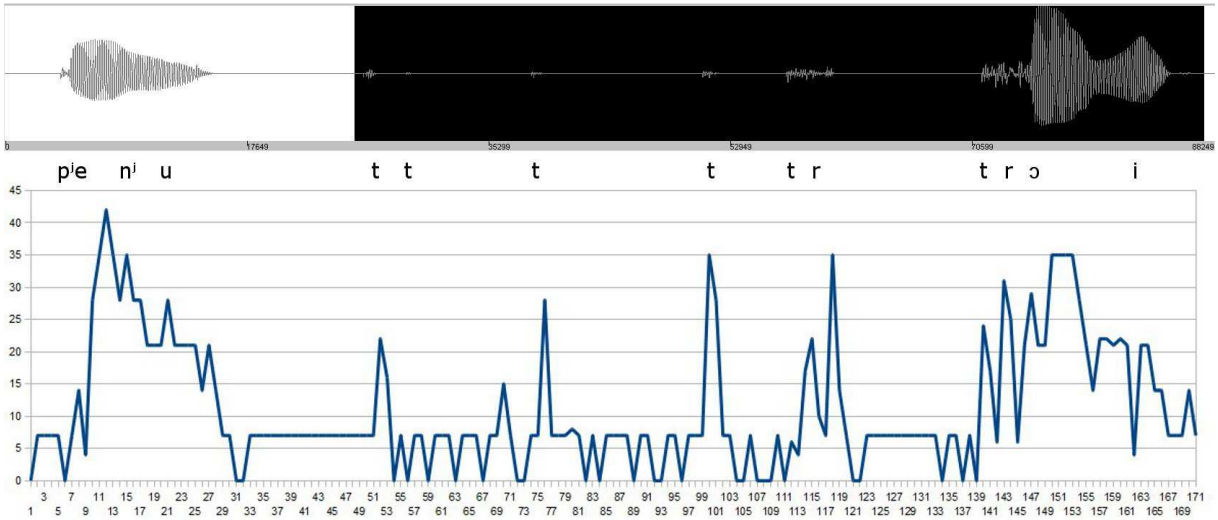

Fig. 7. Nonfluent utterance with the word with the blockade on start marked with the winning neurons graph for Kohonen network 7x7.

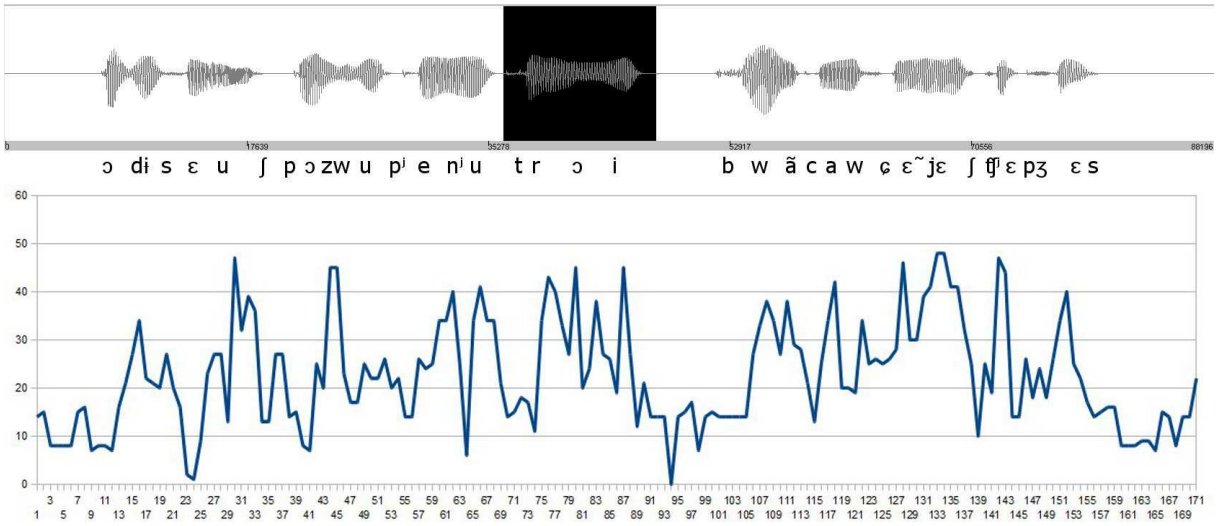

Fig. 8. Fluent utterance with the word marked with the winning neurons graph for Kohonen network 7x7.

After multiplying each window by Hann function to obtain the winning neurons of the Kohonen network for each of them, an analysis of recordings was performed with help of the author's application "Dabar". The best Kohonen network were selected for the sake of that analysis and the best classifying network was chosen with the use of Statistica© 5.5 and Statistica© 7.1 applications.

In the last stage of the speech nonfluency detection, it was examined whether it is possible to construct a learning neural network, which network achieves the best results in nonfluency detection and whether they are satisfactory.

The analysis was performed with the use of Statistica©'s 5.5 tool Solver, which allows the user to test a few network types and at the same time allows the user to assess the quality of the functions' decisions concerning the examined nonfluency. Three types of neural networks were examined: Radial Basis Function, linear and Multi-Layer Perceptron. Each of them had 169 winning neurons as the input for each utterance.

Independent on network structure, there is one neuron on the output layer. Its value, in dependency of the decision threshold, determines whether the utterance is fluent or nonfluent speech.

3. CLASSIFICATION RESULTS

The results were divided into three parts, depending on the size of the Kohonen network output layer. Each of the two charts (1. and 2, 3. and 4, 5. and 6.) contains a list of the tested networks with their structure characteristics and fluency/nonfluency detection correctness percentage.

Table 1. Chart of the best classifying networks for winning neurons taken from Kohonen network 5x5.

No.	***Type***	***Error***	***Hidden layer size***	***Learning algorithm***	***Performance***
1	MLP	0.189	53	BP50,CG10b	1
2	MLP	0.184	53	BP50,CG12b	0.9629
3	MLP	0.194	35	BP50,CG17b	0.9629
4	MLP	0.214	11	BP50,CG50b	0.9629
5	MLP	0.219	3	BP50,CG52b	0.9629
6	RBF	0.267	4	KM,KN,PI	0.8888
7	RBF	0.285	8	KM,KN,PI	0.8888
8	RBF	0.268	2	KM,KN,PI	0.8888
9	RBF	0.454	1	KM,KN,PI	0.8518
10	Linear	0.369	-	PI	0.8148

The learning algorithm provides information concerning the way the network was trained. BP50 means that the network was trained with the Back Propagation for 50 epochs. CG10b means that the conjugate gradient algorithm was used for 10 epochs. KM means K-Means algorithm, KN – K-Nearest neighbour and PI – Pseudo-Invert algorithm.

In the chart below the statistics of classification were gathered. The utterances were divided into three groups: training, verification and test. Each group was examined separately and each group contained about a half of fluent utterances and half of nonfluent ones. The best networks in each type were examined. The percentage was obtained by dividing the correctly classified utterances by all the utterances in a subgroup.

Table 2. Classification quality chart for the best network of each type among the best classifying networks for the winning neurons taken from Kohonen network 5x5 (F – fluent, N – nonfluent utterances).

	Training		***Verification***		***Test***	
	F	***N***	***F***	***N***	***F***	***N***
All	28	27	14	13	13	14
Correctly classified (MLP)	28	27	14	13	10	11
Incorrectly classified (MLP)	0	0	0	0	3	3
Unknown (MLP)	0	0	0	0	0	0
Correctly classified (RBF)	26	26	13	11	11	12
Incorrectly classified (RBF)	2	1	1	2	2	2
Unknown (RBF)	0	0	0	0	0	0
Correctly classified (linear)	28	27	11	11	9	9
Incorrectly classified (linear)	0	0	3	2	4	5
Unknown (linear)	0	0	0	0	0	0
Correctness (MLP)	**100,00%**	**100,00%**	**100,00%**	**100,00%**	**76,92%**	**78,57%**
Correctness (RBF)	92,86%	96,30%	92,86%	84,62%	84,62%	85,71%
Correctness (linear)	100,00%	100,00%	78,57%	84,62%	69,23%	64,29%

Table 3. Chart of the best classifying networks for the winning neurons taken from Kohonen network 6x6.

No.	***Type***	***Error***	***Hidden layer size***	***Learning algorithm***	***Performance***
1	MLP	0.356	7	BP50,CG56b	0.8518
2	MLP	0.316	35	BP50,CG12b	0.8518
3	RBF	0.339	4	KM,KN,PI	0.8518
4	MLP	0.321	53	BP50,CG11b	0.8518
5	RBF	0.566	1	KM,KN,PI	0.8148
6	RBF	0.334	2	KM,KN,PI	0.8148
7	RBF	0.337	3	KM,KN,PI	0.8148
8	MLP	0.365	1	BP50,CG51b	0.7777
9	MLP	0.358	5	BP50,CG50b	0.7777
10	Linear	2.357	-	PI	0.5555

Table 4. Classification quality chart for the best network of each type among the best classifying networks for the winning neurons taken from Kohonen network 6x6 (F – fluent, N – nonfluent utterances).

	Training		***Verification***		***Test***	
	F	***N***	***F***	***N***	***F***	***N***
All	28	27	14	13	13	14
Correctly classified (MLP)	28	27	12	11	11	12
Incorrectly classified (MLP)	0	0	2	2	2	2
Unknown (MLP)	0	0	0	0	0	0
Correctly classified (RBF)	26	24	12	11	12	14
Incorrectly classified (RBF)	2	3	2	2	1	0
Unknown (RBF)	0	0	0	0	0	0
Correctly classified (linear)	28	27	6	9	9	6
Incorrectly classified (linear)	0	0	8	4	4	8
Unknown (linear)	0	0	0	0	0	0
Correctness (MLP)	**100,00%**	**100,00%**	**85,71%**	**84,62%**	**84,62%**	**85,71%**
Correctness (RBF)	92,86%	88,89%	85,71%	84,62%	**92,31%**	**100,00%**
Correctness (linear)	100,00%	100,00%	42,86%	69,23%	69,23%	42,86%

Table 5. Chart of the best classifying networks for the winning neurons taken from Kohonen network 7x7.

No.	***Type***	***Error***	***Hidden layer size***	***Learning algorithm***	***Performance***
1	MLP	0.393	1	BP50,CG52b	0.8518
2	MLP	0.379	35	BP50,CG7b	0.8518
3	MLP	0.402	53	BP40b	0.8148
4	MLP	0.379	35	BP50,CG6b	0.8148
5	MLP	0.369	35	BP23b	0.8148
6	MLP	0.382	35	BP50,CG4b	0.7777
7	MLP	0.401	35	BP23b	0.7777
8	RBF	0.447	2	KM,KN,PI	0.6666
9	RBF	0.437	1	KM,KN,PI	0.6666
10	Linear	0.588	-	PI	0.6296

Table 6. Classification quality chart for the best network of each type among the best classifying networks for the winning neurons taken from Kohonen network 7x7 (F – fluent, N – nonfluent utterances).

	Training		***Verification***		***Test***	
	F	***N***	***F***	***N***	***F***	***N***
All	28	27	14	13	13	14
Correctly classified (MLP)	28	27	11	12	13	12
Incorrectly classified (MLP)	0	0	3	1	0	2
Unknown (MLP)	0	0	0	0	0	0
Correctly classified (RBF)	21	19	9	9	8	9
Incorrectly classified (RBF)	7	8	5	4	5	5
Unknown (RBF)	0	0	0	0	0	0

	Training		***Verification***		***Test***	
Correctly classified (linear)	28	27	8	9	11	10
Incorrectly classified (linear)	0	0	6	4	2	4
Unknown (linear)	0	0	0	0	0	0
Correctness (MLP)	**100,00%**	**100,00%**	**78,57%**	**92,31%**	**100,00%**	**85,71%**
Correctness (RBF)	75,00%	70,37%	64,29%	69,23%	61,54%	64,29%
Correctness (linear)	100,00%	100,00%	57,14%	69,23%	84,62%	71,43%

CONCLUSIONS

The aim of the described analysis was to examine whether it is possible to detect speech nonfluency detection with the use of neural networks with linear prediction coefficients as an input.

The results of the analysis can lead to the conclusion that such detection is possible and satisfactory, and its precision is more than 76%. The second conclusion is that the best reductor of the input data dimension is the Kohonen network with a 5x5 output layer. Input data are vectors of linear prediction coefficients for non-overlapping windows of signal. The verification results with increasing the size of the Kohonen network for any network which detects nonfluency are worse. It must be mentioned that in tests, larger Kohonen networks with Multi-Layer Perceptron classifier detect nonfluency better than those with 5x5 Kohonen network, although the training and verification were worse.

It must also be stated that the best classifying network to detect speech nonfluency with the winning neuron base obtained from the Kohonen network was the Multi-Layer Perceptron. Radial Basis Function networks give quite good results, especially in tests for smaller Kohonen networks. Linear networks were the worst in that examination.

ACKNOWLEDGEMENTS

The scientific work co-financed from the means of the European Social Fund and national budget within the framework of the Human Capital Operational Programme. Measure 4.1. “Strengthening and development of didactic potential of universities and increasing the number of graduates from faculties of key importance for knowledge-based economy”. Sub measure 4.1.1: “Strengthening and development of didactic potential of universities”. Project “Programmatic and structural teaching system reform on Faculty of Mathematics, Physics and Computer Science”.

The authors thank Natalia Fedan for language corrections.

BIBLIOGRAPHY

[1] BLACKMAN R.B., TUKEY J.W., Particular Pairs of Windows. The measurement of power spectra from the point of view of communications engineering, Bell System Tech., New York: Dover, 1959, pp. 95–101.
[2] BUHMANN M.D., Radial Basis Functions: Theory and Implementations, Cambridge University, 2003, pp. 2–5.
[3] CHANDZLIK S., KOPICERA K., The method of neuron weight vector initial values selection in Kohonen network, Journal of Medical Informatics & Technologies, Vol. 10, 2006, pp. 189–198.
[4] CODELLO I., KUNISZYK–JÓŹKOWIAK W., Digital signals analysis with the LPC method, Annales UMCS Informatica. Vol. 5, Lublin, 2006, pp. 315–323.
[5] PROKSA R., Visualization of stages of determining cepstral factors in speech recognition systems, Journal of Medical Informatics & Technologies, Vol. 13, 2009, pp. 121–128.
[6] RABINER L.R., SCHAFER R.W., Digital Processing of Speech Signals, Prentice Hall, New Jersey, 1978, pp. 396–461.
[7] SZCZUROWSKA I., KUNISZYK–JÓŹKOWIAK W., SMOŁKA E., Speech nonfluency detection using Kohonen networks, Neural Computing & Applications, Springer, Vol. 18, Number 7, London, 2009, pp. 677–687.
[8] TEBELSKIS J., Speech Recognition using Neural Networks, Ph. D. Dissertation, Carnegie Mellon University, Pittsburgh, 1995, pp. 101–146.