Paweł POWROŹNIK, Dariusz CZERWIŃSKI[1]

Lublin University of Technology (1)

doi:10.15199/48.2016.07.08

Effectiveness comparison on an artificial neural networks to identify Polish emotional speech

***Abstract**. This article presents the issue of Polish emotional speech recognition based on Polish database prepared by Medical Electronic Division of the Lodz University of Technology. The main goal of this article was to show the differences is artificial neuron networks learning processes. Researches were conducted on the basis of the five most popular variants of the back propagation algorithm. The neuron activation function was the second analyzed issue.*

***Streszczenie.** Artykuł prezentuje możliwe zastosowanie sztucznych sieci neuronowych do identyfikacji stanów emocjonalnych mówcy. Badania zostały przeprowadzone w oparciu o pięć najpopularniejszych wariantów algorytmu wstecznej propagacji błędów. W badaniach wykorzystano polska bazę mowy emocjonalnej przygotowaną przez Zakład elektroniki Medycznej Politechniki Łódzkiej. (**Porównanie skuteczności uczenia sztucznych sieci neuronowych do identyfikacji polskiej mowy emocjonalnej**)*

Keywords: Polish emotional speech recognition, artificial neural networks, backpropagation algorithm
Słowa kluczowe: identyfikacja polskiej mowy emocjonalnej, sztuczne sieci neuronowe, algorytm wstecznej propagacji błedów

Introduction

The recognition of an emotional state of a speaker, based on speech signals analysis, is relatively new issue, although, its significance is rapidly increasing. Development of human - computer type of communication systems may be one of the reasons of such a direction of changes. The other reason is the growth of number of applications involving the processing of the speech signal. Attempt to find a universal set of parameters which clearly describe appropriate emotion seems to be another reason for increasing interest of this subject.

Carried out researches has been mostly based on databases where each speech sample has been matched with specific emotional tone of the voice [1]. Although, achieved results can be better. It might be a result of the fact that it is possible for us to identify another person's emotional state only in 60% of all cases [2].

The basic classifier applied to this type of research is the support vector machine (SVM), and the k-Nearest Neighbours algorithm (or k-NN for short).

The subject of this research is the comparison of effectiveness of artificial neural networks which were learned by different methods: basic backpropagation algorithm (GD), backpropagation with adaptive learning rate (GDA), backpropagation with momentum (GDM), backpropagation with adaptive learning rate and momentum (GDX) and scaled conjugate gradient backpropagation algorithm (SCG). The neurons were activated by three different methods: linear, sigmoidal and hyperbolic tangent. All simulation has been conducted in Matlab and Simulink systems.

Signification of emotional speech

The dynamic development of interface based on the human-computer type of interaction is leading to guaranteeing almost completely non-absorbing methods of interaction [3]. Devices which have been used in communication so far, such as the mouse, keyboard, pad, or monitor, do not suffice. It is expected that new kinds of interface will be created, that is, ones that will allow a completely intuitive approach to operating a device by using the senses of speech, sight, or touch. Along with the development of information society and a notable improvement in both data transfer speed and its quality, a distinctive tendency to use speech signals as one of the ways of human-computer interaction has been observed [4]. Such an approach might result from the fact that speech is the most intuitive manner of communication, which allows the use of cameras, microphones, or touch screens. From the perspective of speech signal processing, these impulses contain the following information [5]:

- semantic, focusing on the meaning of a statement and its content,
- emotional, allowing the identification of the speaker's emotions to some extent,
- individual, allowing the identification of the speaker as possessing a unique voice.

Emotions which are expressed in the voice constitute an essential part of a message and they are a complement for the semantic information. Comfort and intuitiveness in using a voice interface can be significantly improved by the support of voice recognition systems which are capable of detecting the emotions of the speaker. It means that a new quality in the human-computer communication is being introduced, and it is a matter of utmost importance [3]. The knowledge of qualities influencing the fact that voice is an important means of conveying emotional information lies at the very foundation of human communication.

What is fundamental for the identification of basic emotions, such as anger, joy, sadness, boredom, fear [2, 6], based on the analysis of speech signals is the paradigm of Sherer [2]. It claims that each basic emotion can be universally described. Such a description should include an unambiguous model or a set of acoustic parameters. It is one of the reasons behind the necessity of possessing an accurately extensive and varied set of recordings. This researches were conducted with usage of database of the Polish emotional speech, prepared and shared by Lodz University of Technology [7].

Research methodology

The structure of emotion recognition system has been shown on Fig. 1.

The researches focused primarily on the effectiveness of neural network which were learned with different backpropagation algorithm. The second analysed issue is learning time.

Speech signal parameters

All researches which are connected with processing of speech signal require a distinction of characteristic parameters in the voice.

Carried out researches has no results in the determination of a uniform and universal set of features so far, what forced researchers to adopt heuristic approach [8]. The main idea of this approach is to extract from speech signal as many parameters as possible. Then select those which described the researched matter best. For selection experimental methods are used as well as algorithms. This researches has been focused on following parameters: average value and energy of the speech signal, sex of the speaker (1 - woman, 0 - man), minimum and maximum sample value for a given signal.

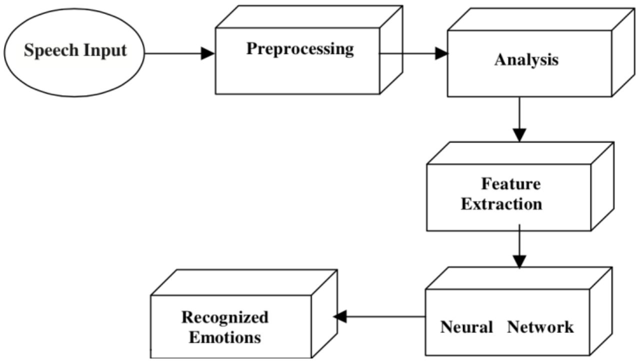


Fig.1. The structure of emotion recognition system

The average value of the whole signal is described by the following formula [9]:

$$\overline{x_N} = \lim_{n\to\infty}\frac{1}{2N+1}\sum_{n=-N}^{N} x(n)$$,

where *x(n)* – value of *n* – sample, *N* – total number of samples.

The energy of the signal is defined as the integral of the square of the signal, that is, energy emitted with unitary resistance. For digital signals it is defined in the following way [9]:

$$Ex = \sum_{n=0}^{N} x^2(n),$$

where *n* – sample number, $x^2(n)$ – square value of *n* – sample.

Before mentioned above parameters have been extracted, the values of particular samples have undergone the process of normalization.

Short description of conducted research

The Polish database of emotional speech, prepared and shared by the Medical Electronics Division of the Lodz University of Technology, has been used for research. The collection was prepared by 8 actors of both sex and contains 240 records in six different emotional states, that is: anger, boredom, fear, joy, sadness, or without any emotional tone. Each of the speakers pronounce five different sentences: 'I stop to shave from today on', 'Johnny was today at the hairdresser', 'They have bought a new car today', 'This lamp is on the desk today' and 'His girlfriend is coming here by plane'. The database contains sound files in the 'wav' format sampled with 44.1 kHz frequency and the bit rate of 16 bps [1].

All researches has been conducted based on four layers artificial neural network (ANN). The input layer contain 5 nodes: average value of signal, signal energy, sex of speaker, minimum and maximum value indicated in whole signal. The learning method was a difference. ANN has two hidden layers: the first one contains 12 hidden neurons, the second one – six. The output layer was built with six neurons one for each recognised emotional state. A structure of ANN has been shown on Fig. 2. During researches three different neuron activation methods has been used: linear, sigmoidal and hyperbolic tangent.

The main problem was to compare the effectiveness of abovementioned ANN which was learned by different learning algorithms.

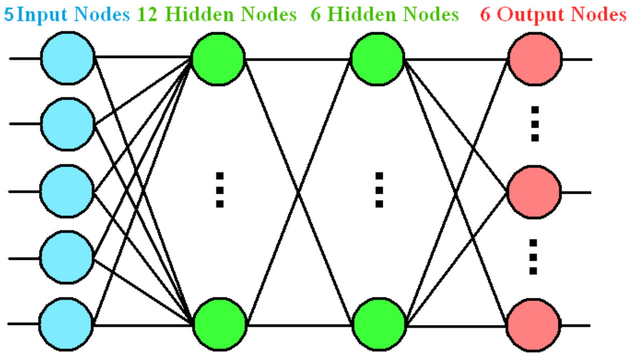


Fig.2. A structure of used ANN

The first one was commonly used basic gradient descent backpropagation algorithm (GD) which based on minimalizing the mean square error between the network's desire output and achieved output. The GD algorithm updates bias and synapses weights along the negative gradient of error energy function. Once the error achieved by ANN has decreased to the specified level, the network is consider to be trained [10].

The second was gradient descent with adaptive learning rate backpropagation algorithm. Commonly known backpropagation algorithm defines a error energy used for monitoring learning process as generalized value of all errors calculated by the least-squares formulation. Abovementioned error is represented by Mean Squared Error (MSE) as follows [10]:

$$MSE = \frac{1}{MP}\sum_{j=1}^{P}\sum_{k=1}^{M}(d_k - y_k)^2$$

where *M* is the number of neurons in output layer, *P* - number of training patterns, d_k - desired outputs and y_k - actual output.

The main difference between standard GD algorithm and GDA is learning rate parameter. In GD this factor is constant through training. Although the backpropagation algorithm is sensitive to the proper setting of this rate. This is the reason why GDA algorithm was developed. The main idea of GDA is to allow the learning rate parameter to be adaptive to keep the learning step size as large as possible while keeping learning stable. The adaptive rate value changes with the gradient's trajectory on the error surface [10].

The third used algorithm was Gradient Descent with Momentum Backpropagation (GDM). The main advantage of this learning methods is that often provides faster convergence then other based on backpropagation algorithms. During learning process momentum allows networks to respond to the local gradient as well as recent trends in error surface [10].

The GDM behaves like low-pass filter what allows ANN to ignore small features in the error surface. The main danger in typical GD is that during learning process the network can stuck in shallow local minimum, momentum makes that such minimum can be skipped [10].

The combination of GDM and GDA is Gradient Descent with Momentum and Adaptive Learning Rate Backpropagation Algorithm(GDX). The main idea was to add the momentum coefficient as an additional training parameter to mentioned above Adaptive Learning Rate Backpropagation. Thus, the weight vector has been modified based on GDM but to verify learning rate GDA algorithm has been used. To adjust weight basic back propagation algorithms use the direction in which the cost function is decreasing most rapidly (negative of the gradient). Such a procedure does not necessarily produce the fastest convergence. In Scaled Conjugate Gradient Backpropagation (SCG) to adjust neuron weights a search along conjugate directions is performed. Typical conjugate gradient algorithms start out by searching on the first iteration. The SCG training algorithm was developed to avoid this time-consuming line search [10].

Achieved results

As it was mentioned above neural network used in this research contained 4 layers. The main issue of conducted research was to compare the effectiveness of different learning methods for ANN in case of recognition of Polish emotional speech. The neurons has been activated by three functions: linear, sigmoidal and hyperbolic tangent. On Fig. 3 was shown learning time (ten first tests) for different learning methods in case when neuron activation function was sigmoidal.

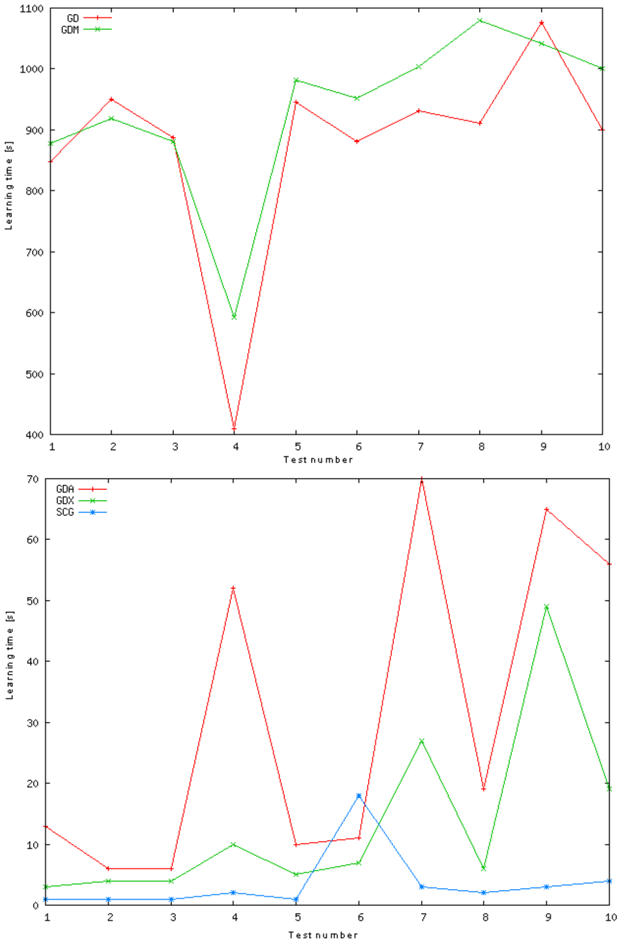


Fig. 3. Learning time comparison for sigmoidal function

The best results were achieved for artificial neural network which neurons was activated by hyperbolic tangent function and the ANN was learned using backpropagation with momentum algorithm. Results achieved by different learning methods are quite similar (the average difference in recognition is 1,2%) . In Fig. 4 achieved results for abovementioned ANN was shown. In Fig. 5 was shown confusion matrix for this network.

All achieved results during research was compared on Figures 6 to 10. Results were divided based on neural network learning function.

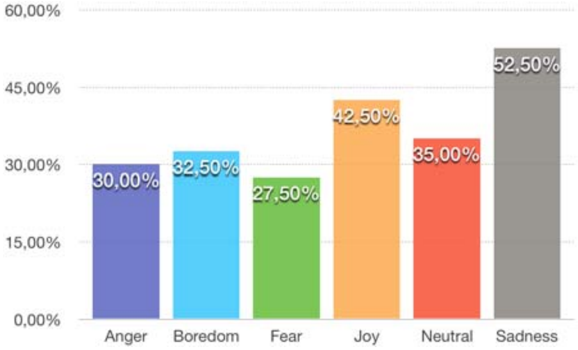


Fig. 4 The effectiveness of the recognition of individual emotional states by ANN activated by hyperbolic tangent function and learned by GDM.

	Anger	Boredom	Fear	Joy	Neutral	Sadness
Anger	12	5	10	11	1	1
Boredom	4	13	6	1	11	5
Fear	9	9	11	5	4	2
Joy	10	1	7	17	4	1
Neutral	4	9	4	1	14	8
Sadness	1	7	5	1	5	21

Fig. 5. Confusion matrix for ANN activated by hyperbolic tangent function and learned by GDM.

	Linear function	Sigmoidal function	Hyperbolic tangent
Anger	25,00%	27,50%	32,50%
Boredom	32,50%	30,00%	30,00%
Fear	27,50%	27,50%	30,00%
Joy	32,50%	37,50%	35,00%
Neutral	32,50%	32,50%	32,50%
Sadness	42,50%	47,50%	50,00%

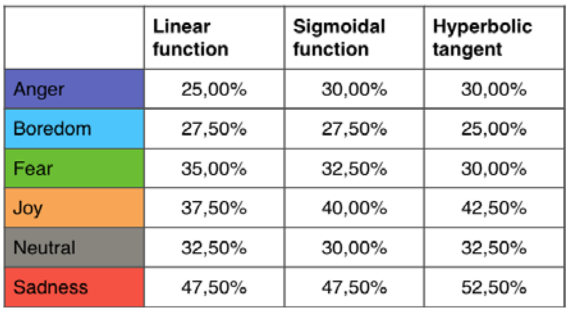
Fig. 6. The effectiveness of the recognition of individual emotional states by ANN learned by GD.

	Linear function	Sigmoidal function	Hyperbolic tangent
Anger	27,5%	27,5%	25%
Boredom	30%	30%	30%
Fear	32,5%	35%	32,5%
Joy	35%	40%	37,5%
Neutral	32,5%	35%	35%
Sadness	42,5%	42,5%	45%

Fig. 7. The effectiveness of the recognition of individual emotional states by ANN learned by GDA.

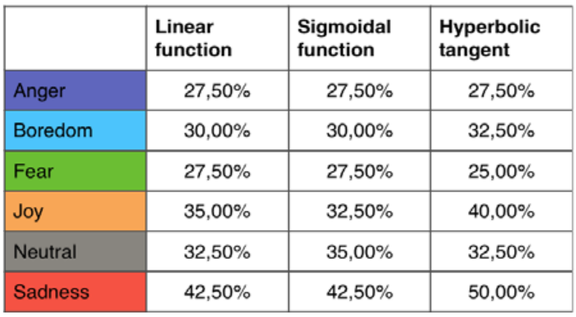
	Linear function	Sigmoidal function	Hyperbolic tangent
Anger	25,00%	27,50%	30,00%
Boredom	30,00%	32,50%	32,50%
Fear	32,50%	30,00%	27,50%
Joy	37,50%	37,50%	42,50%
Neutral	35,00%	35,00%	35,00%
Sadness	45,00%	47,50%	52,50%

Fig. 8. The effectiveness of the recognition of individual emotional states by ANN learned by GDM



	Linear function	Sigmoidal function	Hyperbolic tangent
Anger	25,00%	30,00%	30,00%
Boredom	27,50%	27,50%	25,00%
Fear	35,00%	32,50%	30,00%
Joy	37,50%	40,00%	42,50%
Neutral	32,50%	30,00%	32,50%
Sadness	47,50%	47,50%	52,50%

Fig. 9. The effectiveness of the recognition of individual emotional states by ANN learned by GDX



	Linear function	Sigmoidal function	Hyperbolic tangent
Anger	27,50%	27,50%	27,50%
Boredom	30,00%	30,00%	32,50%
Fear	27,50%	27,50%	25,00%
Joy	35,00%	32,50%	40,00%
Neutral	32,50%	35,00%	32,50%
Sadness	42,50%	42,50%	50,00%

Fig. 10. The effectiveness of the recognition of individual emotional states by ANN learned by SCG

Conclusions

The recognition of emotions in speech signal is a difficult task, and the achieved results are far from ideal. Nevertheless, it is possible to improve the achieved results [2]. But what is worthy to said that depends on neuron activation function and neural network learning methods in case of recognition of Polish emotional speech achieved results are quite spread. Carried out researches shown that it is so important to choose appropriate learning method and neuron activation. As can be observer if there is necessity to focus only on one of emotional state the GDM method may not be the best one. The effectiveness of emotion recognition can be also increased by combining a voice analysis system with semantic analysis [11]. A natural direction of development is, first and foremost, the application and testing of the suggested solutions in both the process of abstracting signal and classifier parameters.

***Authors**: dr hab. inż. Dariusz Czerwiński, prof. nadzw., Politechnika Lubelska, Instytut Informatyki, ul. Nadbystrzycka 38a, 20-618 Lublin, E-mail: d.czerwinski@pollub.pl; mgr Paweł Powroźnik, Instytut Elektrotechnki i Elektrotechnologii, ul. Nadbystrzycka 38a, 20-618 Lublin, E-mail: ppowroznik@gmail.com*.

REFERENCES

[1] Kamińska D., Pelikant A., Zastosowanie multimedialnej klasyfikacji w rozpoznawaniu stanów emocjonalnych na podstawie mowy spontanicznej, *IAPGOŚ*, 3/2012, 36 – 39

[2] Scherer K., Vocal communication of emotions: A Review of Research Paradigms in Speech Communication 40, 2003, 227 – 256

[3] Igras M., Ziółko B., Baza danych nagrań mowy emocjonalnej, *Studia Informatica*, Vol. 34, 2013, 67 – 77

[4] Ziółko B, Ziółko M., Przetwarzanie mowy, *Wydawnictwo AGH,* 2011, 16 – 18

[5] Paeschke A., Sendlmeier W. F., Prosodic Characteristics of Emotional Speech: Measurements of Fundamental Frequency Movements, *ITRW on Speech and Emotions,* Newcastle 2000

[6] Lewis M., Haviland-Jones J. M., Psychologia emocji*, Gdańskie Wydawnictwo Psychologiczne*, Gdańsk 2005,

[7] http://www.eletel.p.lodz.pl/bronakowski/med_catalog/ (18.01.2015)

[8] Polzehl T., Schmitt A., Metze F., Approaching multilingual emotion recognition from speech on language dependency of acoustic/prosodic features for anger recognition, *Proc. of Speech Prosody*, Chicago 2010

[9] Zieliński T., Przetwarzanie sygnałów cyfrowych. Od teorii do zastosowań, *WKŁ*, 2009

[10] Gopalakrishnan K., Effects of training algorithms on neural network aided pavement diagnosis, *International Journal of Engineering, Science and Technology, Vol. 2, No. 2, 2010, 83 – 92*

[11] Wang Y., Guan L., Recognizing human emotional state from audiovisual signals*, Proc. IEEE Transactions on multimedia*, Vol. 10, 2008, p. 659 – 668