jet INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2018, VOL. 64, NO. 4, PP. 425–432
Manuscript received March 24, 2018; revised October, 2018. DOI: 10.24425/123541

Determination of Input Parameters of the Neural Network Model, Intended for Phoneme Recognition of a Voice Signal in the Systems of Distance Learning

Berik Akhmetov, Igor Tereykovsky, Aliya Doszhanova, and Lyudmila Tereykovskaya

***Abstract*—The article is devoted to the problem of voice signals recognition means introduction in the system of distance learning. The results of the conducted research determine the prospects of neural network means of phoneme recognition. It is also shown that the main difficulties of creation of the neural network model, intended for recognition of phonemes in the system of distance learning, are connected with the uncertain duration of a phoneme-like element. Due to this reason for recognition of phonemes, it is impossible to use the most effective type of neural network model on the basis of a multilayered perceptron, at which the number of input parameters is a fixed value. To mitigate this shortcoming, the procedure, allowing to transform the non-stationary digitized voice signal to the fixed quantity of mel-cepstral coefficients, which are the basis for calculation of input parameters of the neural network model, is developed. In contrast to the known ones, the possibility of linear scaling of phoneme-like elements is available in the procedure. The number of computer experiments confirmed expediency of the fact that the use of the offered coding procedure of input parameters provides the acceptable accuracy of neural network recognition of phonemes under near-natural conditions of the distance learning system. Moreover, the prospects of further research in the field of development of neural network means of phoneme recognition of a voice signal in the system of distance learning is connected with an increase in admissible noise level. Besides, the adaptation of the offered procedure to various natural languages, as well as to other applied tasks, for instance, a problem of biometric authentication in the banking sector, is also of great interest.**

***Keywords*—neural networks, a phoneme, recognition of a voice signal,the system of distance learning, mel-cepstral coefficients, the spectral analysis**

I. Introduction

INTERNATIONAL experience in the development of distance learning systems demonstrates that one of the most perspective ways to increase their efficiency is the introduction of interactive training materials, based on the application of means of voice signals recognition [1, 2]. Except the proved improvement of training quality, the specified means application allows to boost the applicability of distance learning systems due to getting rid of a rigid binding to the lesson schedule, to meet the requirements of listeners with limited opportunities more efficiently, and to enhance the security of such systems at the expense of biometric authentication means introduction. At the same time, in the majority of existing distance learning systems the means of a voice signal recognition are absent, though the possibility of their introduction is confirmed by the extensive use of software applications (Google+, Microsoft Office, VoiceNavigator, Siri) with the corresponding functionality. At the same time, introduction of the known means of a voice signal recognition in the popular systems of distance learning triggers the necessity of their complicated adaptation to the variability of application conditions, connected with the development term, volume of the recognizable words dictionary, formation of sample databases, admissible size of a recognition error, acoustic factors, resource intensity of creation and functioning. One more barrier of its introduction is the high cost of the existing commercial means of recognition and lack of detailed scientific and technical documentation [2, 3]. In such formulation, the problem of models development, methods and means of recognition of voice signals, adapted to conditions of systems of distance learning, is relevant.

B. Akhmetov is Rector of the Caspian State University of Technologies and Engineering named after Sh. Yessenov, Aktau, Republic of Kazakhstan (e-mail: berik.akhmetov@kguti.kz).

I. Tereykovsky is with National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine (e-mail: terejkowski@ukr.net).

A Doszhanova is with Department IT-engineering, Almaty University of Power Engineering and Telecommunications, Almaty, Republic of Kazakhstan (e-mail: d_alia.81@mail.ru).

L Tereykovskaya is with Kyiv National University of Construction Architecture,Ukraine (e-mail: tereikovskal@ukr.net).

II. Literature Review Paper

As a rule, voice interaction in the course of distance learning should be applied, when holding lectures, seminars, consultations, laboratory and practical training [3]. Quite a natural addition to the provided list is the application of voice authentication of users. It is worth noting that the interaction between components of distance learning system, based on voice signals recognition, is to be understood as voice interaction. In many cases, such recognition has to be carried out in the automatic mode, for example, at computer testing when a student has to provide a voice answer to the question posed. Similarly, the system of recognition has to deal with two problems: 1. To determine correctness of the answer. 2. To carry out user authentications, that is to determine whether it is the stated student that interacts with the system. The solution of both tasks is based on the recognition of voice information of the user. It is the complexity of recognition that determines the emergence of the problem voice interactions in information technology of distance learning. It is worth mentioning that

generally such recognition consists in the consecutive solution of two tasks: 1. Creation of the formal description of voice information. 2. Carrying out the semantic analysis of the received formal description. Quite often, the formal description of voice information is understood as its textual representation. At the same time modern theoretical practices in the field of the semantic analysis of text information do not allow to create highly reliable tools [4, 5], and in many cases the answer and the identity of the student can be ascertained on the basis of identification of certain words in voice information. Thus, concerning information technology of distance learning, recognition of voice information is limited to the search for keywords in it and to the announcer recognition. According to [1, 6], a standard analysis algorithm of a voice signal for search of keywords and recognition of the announcer consists of seven stages. The flowchart of this algorithm is shown in fig. 1.

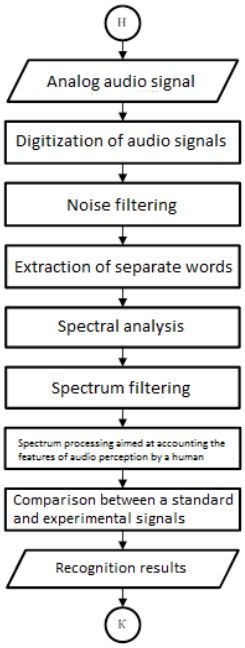

Fig. 1. Flowchart of a standard analysis algorithm of a voice signal

For determining of perspective ways of the keywords search and the announcer recognition, the analysis of separate stages of the given algorithm is carried out. The analysis was carried out in terms of application of the algorithm in information technology of distance learning. As a rule, a pre-digitized signal is being input to modern systems of recognition of audio information. Digitization is implemented by means of the standard equipment of a personal computer - the sound card and the microphone. When using the popular equipment, sampling of a signal is ranging from 8000 up to 44000 Hz. At the same time the analog input signal is being quantized, that is presented in the form of sixteen-digit or thirty-two-digit numbers. It is obvious that at application in the system of distance learning it is necessary to be guided by some standard parameters of digitization. It can represent parameters, which correspond to the weakest configuration of the sound equipment: frequency of sampling is 8000 Hz and sixteen-digit quantization. Prime filtration of noise in the sampled signal consists in overlaying of windows of different types, such as Kayser, Hamming and others, on this signal. Following noise filtration, the energy analysis of a voice signal during each 10-20 ms is applied to extract separate words from a sound stream [6]. Besides, it is possible to determine the beginning/end of a word by splash/attenuation of a signal size. Processing of the input digitized signal with a view to reduce the input data volume consists in application of various methods of the spectral analysis of data. The spectral analysis of data is implemented whether by means of the Fourier window discrete transformation, or by means of the discrete wavelet transformation. The Fourier discrete transformation provides representation of a row in the form of several processes, consisting of sinusoids and cosinusoids of various frequencies. For this purpose, Fourier needs to spread out the digitized data in a row:

$$\bar{X}(t) = a_0 + \sum_{i=1}^{q}(a_i c_i(t) + b_i s_i(t)) + e(t) \qquad (1)$$

where $c_i(t) = \cos(2\pi f_i t), s_i(t) = \sin(2\pi f_i t), a_0$ is a coefficient, a_i, bi are regression coefficients which indicate degree of correlation of functions $c_i(t) = \cos(2\pi f_i t)$ and $s_i(t) = \sin(2\pi f_i t)$ with statistical data, $f_i = \frac{i}{T} - i$-and a harmonica of the main frequency of $\frac{1}{T}$, $q = \frac{T-1}{2}$, T is a quantity of points of a row, is a casual component. The expression $2\pi f_i$ is designated as λ_i and is called circular frequency. For the solution (1) it is necessary to calculate coefficients on the basis of statistical data ai and b_i:

$$a_i = \frac{2}{T}\sum_{t=1}^{T}(\bar{X}(t)c_i(t)), i = 1, 2, ..., q, \qquad (2)$$

$$b_i = \frac{2}{T}\sum_{t=1}^{T}(\bar{X}(t)c_i(t)), i = 1, 2, ..., q. \qquad (3)$$

Based on the received results, the vectors of mel-frequency cepstral coefficients are calculated (Mel Frequency Cepstral Coefficients MFCC) [7]. The received sequence of the normalized MFCC vectors represents input information for the modern systems, intended for comparison reference and examinees of voice signals. Also it should be noted that many of these systems use neural network methodology of phoneme recognition. So, in the work [8] development of the neural network method of delimitation between phonemes is declared. The formation procedure of input parameters of the model, which represent discrete veyvlet transformation coefficients of a MFCC vector of each stationary segment (frame) of a voice signal, is shown. The computer experiment of delimitation of phonemes of some Russian words is described. During experiments, a three-layer perceptron with ten neurons in the hidden layer and one output neuron was applied. In the means of voice signals recognition, applied in the Google search engine, the Android operating system, and the Apple products, so-called deep neural networks are used [1]. Structurally deep neural network represent a multilayered perceptron with a large number of the hidden layers of neurons

and two-stage training. At the first stage, there is a preliminary control of weight coefficients. The method "without teacher", based on the use of the limited Boltzmann machine, is used. At the second stage of training, the method "with a teacher" with use of an algorithm of the return distribution of an error is implemented. The need of two-stage training is connected with low efficiency of the standard algorithms of weight coefficients correction at a large number of the hidden neurons and with formation complexity of sufficient marked educational examples. In case of a small amount of hidden layers of a neural network (2-4) and rather large number of the marked educational examples, there is no need in two-stage training. The use of deep neural networks allows to process powerful voice streams, however requires considerable computing capacities that can cause difficulties in the systems of distance learning. In the neural network system of voice signals recognition of the Microsoft Company (MSNET) the neural network models like a multilayered perceptron are used. Though it was not possible to find the detailed description of this system, according to indirect data [19, 24, 83, 88, 143, 146] and proceeding from practical experience, it is possible to note that for determination of an optimum look and parameters of a neural network model, the multicriteria approach is used; a method of training is optimized with positions of minimization of training term. Besides, at creation of the system the computing resources limitations are considered. In the work [9] for a solution of the problem of speaker-independent recognition of phonemes in spoken language, arising during creation of systems of automatic recognition of words of the discrete and conjoint speech on the basis of the phoneme-focused method, the authors offered an algorithm, based on the use of bagging ensemble of "multilayered perceptron" neural networks. On material of the TIMIT speech corpus, an advantage of bagging ensemble of neural networks both over the single neural network recognizer, and over the recognizer on the basis of the hidden Markov models, is experimentally shown. The received results 69,17% of correctly recognizable phonemes illustrate the competitiveness of the neural network algorithm, offered by the authors, and its applicability in the systems of conjoint speech recognition. In the work [12] influence of various sources of a speech internal variation (speech tempo, effort, style and a dialect or accent) on perception of the human speech is investigated. As a result of computer experiments, it is shown that the distance of spectral level between the corresponding speech segment and a range of the masking noise is a good predictor for separate phonemes recognition of high-quality. Recognition of voice commands by means of a convolutional neural network is offered in [18]. Such networks operate with two-dimensional data - neurons in each layer form the planes. The input layer is represented by one plane. Its dimension coincides with the one of input data. The following layers of network are convolute and consist of the planes of neurons (the card of signs). Each neuron of the convolute layer is connected to a small subarea of the previous layer. The last two layers of network represent a neural network with direct distribution of a signal. The example of practical application is given. Advantages of convolutional neural networks are shown: small amount of parameters of training and high precision of training. One more perspective view of the neural network model, intended for recognition of voice signals, is presented in the works [13-15]. It is about the time-delay neural network (TDNN), multi state time-delay neural networks (MS-TDNN) and also about the network of long short-term memory (LSTM). In contrast with traditional networks on the basis of a multilayered perceptron such networks have recurrent communications that, as envisioned by their creators, allows to adapt them to the analysis of numerical data ranks. It is specified that the accuracy of recognition of such networks is over 90%. At the same time in the works [1, 10, 17] insufficient study of this kind of networks is pointed out that results in their high resource intensity, instability of training and functioning. Thus, as a result of the carried-out analysis of the guide-books, devoted to application of neural networks for voice signals recognition and also on the basis of scientific research and practical works of the authors, stated in [1, 6] the prospects of introduction in the system of distance learning of neural network means of phoneme recognition of voice signals are determined. In this case, for recognition of a voice signal it is necessary to solve three various problems: 1. To divide the word selected in a voice stream into separate phonemes (a phoneme-like element) 2. On the basis of the analysis of a phoneme-like element, to distinguish a phoneme 3. On the basis of the analysis of the recognizable phonemes, to recognize a keyword Taking into account that the first and third tasks can be solved on the basis of the works [1, 8, 10, 15]. Therefore, in this article the accent to be set on the solution of the third task. Application of multilayered perceptron neural networks is implied. At the same time, the main difficulties of the neural network model creation are connected with uncertain duration of a phoneme-like element. Due to this reason, it is impossible to determine the number of input parameters of the neural network model.

III. Formulation of the Research Problems

Development of the determination procedure of the neural network model input parameters, intended for recognition of phonemes in the system of distance learning.

IV. Procedure of Determination of Input Parameters

Generally, in any natural languages there are several types of phonemes. This work is focused on the Russian language, in which five types of phonemes are distinguished: 1) vowels ([], [], [], [], [], []), acoustic basis of which is represented only by tone; 2) sonorants ([], [], [], []), acoustic basis of which is tone, noise is practically absent, by this feature they are closest to vowels; 3) voiced consonants ([], [], [], [], [], []), in which the tone prevails over noise; 4) voiceless consonants ([], [], [], [], [], [], []), in which noise prevails over tone; 5) hissing consonants ([], [], [], []), acoustic basis of which is represented only by noise. It is worth noting that 2-5 types of phonemes form a group of concordant phonemes. Taking into account that each vowel phoneme can be stressed and unstressed, and each concordant one can be hard and soft,

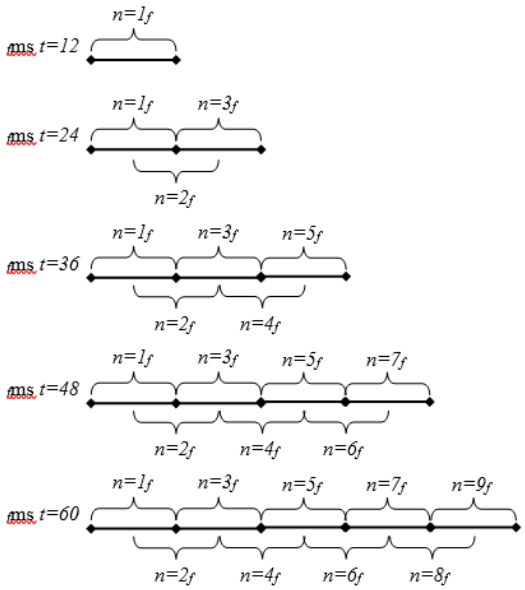

Fig. 2. Quantity of possible intervals of the analysis at recognition of phonemes

the phonetic alphabet of the Russian language consists of 48 elements.

According to [8, 10] as the first approximation at normal tempo of speech, duration of a Russian phoneme can vary from $t_{\phi}^{max} = 60ms$ to t_{ϕ}^{min} , that is equal to duration of a quasistationary fragment of a voice signal [15, 50, 91]. It is believed that duration of a quasistationary fragment of a voice signal is equal to $t_{CT} = 12ms$. subsequently the selected fragment of a voice signal, which corresponds to a separate phoneme is to be called a phoneme-like fragment. As shown in fig. 2, the analysis of a voice signal is implemented by parts, on quasistationary fragments, which have half overlapping. Therefore, the duration of a phoneme falls within the limits of:

$$t_{\phi} = \left[t_{\phi}^{min}..t_{\phi}^{max}, t_{CT}\right] = [12...60, 12]\,. \tag{4}$$

Depending on phoneme duration, the quantity of possible corresponding quasistationary fragments of a voice signal is equal to:

$$n_{\phi} = 2\frac{t_{\phi}}{t_{cm}} - 1. \tag{5}$$

According to [8-10], from 6 to 24 mel-cepstral coefficients correspond to a quasistationary fragment there. Considering (2), the number of such coefficients, which will correspond to input of the neural network model, can be expressed as follows:

$$N_x = (12..48)\frac{t_{\phi}}{t_{cm}} - (6..24). \tag{6}$$

where N_x is a quantity of the neural network model inputs. Substitution (1) in (3) shows that the number of input parameters of the neural network model is within an interval:

$$N_x \epsilon\,[54..216, 6]\,. \tag{7}$$

At the same time, the existing neural network models are not adapted for educational examples with the variable number of input parameters. Therefore, directly it is inexpedient to use mel-cepstral coefficients, which describe a voice signal as the input parameters of the neural network model. For correction of this shortcoming, the results [6, 8-10] are used, testifying to a possibility of high-quality recognition of boundaries of separate phonemes and to a basic possibility of nonlinear scaling of a voice signal fragment, which corresponds to a separate phoneme. For calculation of large-scale coefficient, results [10] are used, in which the general model of a separate phoneme duration is given:

$$\begin{gathered} T_j = \left\lfloor T_j^{(0)} - k_n(N-j)\right\rfloor \times \\ \left[1 - (1-k_T)\frac{N-j+1}{N}\right] \times \frac{T}{N}. \end{gathered} \tag{8}$$

where k_n is a compression coefficient; k_T is a speech tempo coefficient; is quantity of phonemes in a statement; is a statement duration; $T_j(0)$ is duration j-phoneme under stress; j is a number of a phoneme in a statement.After adaptation of this model to conditions of system of distance learning [1] it is expressed as follows:

$$\begin{gathered} t_{\phi j} = \lfloor t_{\phi i}e - k_n(N_{\phi} - j)\rfloor \times \\ \left[-(1 - k_T\frac{N_{\phi}-j+1}{N_{\phi}})\right] \times \frac{T}{N_{\phi}}, i\epsilon\,[1...K_{\phi}]\,. \end{gathered} \tag{9}$$

where $t_{\phi j}$ is duration of a j-phoneme-like fragment of an experimental voice signal; $t_{\phi i}e$ is standart duration i-phoneme; N_{ϕ} is a quantity of phoneme-like fragments in an experimental voice signal; T is duration of an experimental voice signal; is a number of a phoneme-like fragment in an experimental voice signal; K_{ϕ} is a quantity of phonemes, on which the system of recognition is focused. In this case, the number of input parameters of the neural network model will correspond to the quantity of mel-cepstral coefficients of a standard phoneme, and the selected fragment of a voice signal is subject to the nonlinear scaling procedure. When using a standard -phoneme the large-scale coefficient for a j-phoneme-like fragment is calculated as follows:

$$k_{i,j} = \frac{t_{\phi i}e}{t_{\phi j}}, i\epsilon\,[1...K_{\phi}] \tag{10}$$

With the minimum duration of a phoneme-like fragment and the maximum duration of a standard phoneme, the large-scale coefficient is equal to:

$$k_{max} = 5/1 = 5, \tag{11}$$

where k_{max} is a maximum size of the large-scale coefficient. With the minimum duration of a standard phoneme and the maximum duration of a phoneme-like fragment the minimum size of the large-scale coefficient is:

$$k_{min} = 1/5 = 0,2, \tag{12}$$

where k_{min} is minimum size of the large-scale coefficient. As at recognition of $t_{\phi j}$ and $t_{\phi i}$ phonemes only discrete change can take place, so the change of large-scale coefficient also has the discrete nature. Finally, the functionality of range of changes of the large-scale coefficient can be described as follows:

$$if\, t_{\phi i}e > t_{\phi i} \rightarrow k_{\phi_{i,j}}\epsilon\,[1..5] \tag{13}$$

$$if t_{\phi j} > t_{\phi i} e \rightarrow k_{\phi_{i,j}} \epsilon [0/, 2..1] \quad (14)$$

The procedure of the selected phoneme-like fragment scaling can be presented in the form of the following expression:

$$\vec{y_I} \rightarrow \vec{Y_I}, \quad (15)$$

where $\vec{y_I} = \{y_0, y_1, ...y_I\}$ is a vector, which corresponds to the initial phoneme-like fragment, set by means of the I counting; $\vec{Y_j} = \{Y_0, Y_1, ...Y_j\}$ is a vector, which corresponds to the scaled phoneme-like fragment set by means of J counting. At the same time $J = k \times I$. At $k_{\phi_j,i} > 1$ to implement (11), it is necessary to distinguish in $\vec{Y_j}$ vector the vector of basic counting $\vec{q}$ and the vector of intermediate counting $\vec{g}$. For this purpose, it is necessary to use the following expressions:

$$q_i = trunk(k \times i), i\epsilon [0, I], \quad (16)$$

$$\vec{g} = \vec{Y_j} - \vec{q}, \quad (17)$$

where q_i is -reference point; trunc is a rounding operation of a real number to the nearest whole; k is a large-scale coefficient. It is worth noting that the number of basic counts is equal to , and the number of intermediate counts is equal to

$$g = J - I. \quad (18)$$

In each basic count it is necessary to establish sizes of $\vec{Y_j}$ components, equal to the corresponding $\vec{y_I}$ components:

$$Y_{qi} = y_i, \quad (19)$$

where Y_{qi} is $\vec{Y_j}$ valuesin q_i -basic count.

By analogy with the linear interpolation of discrete function methodology, for calculation of $\vec{Y_j}$ components in any g intermediate, which is between q_i and q_{i+1} basic counting, there is an expression:

$$Y_g = Y_{qi} + \frac{Y_{qi+1} + Y_{qi}}{q_{i+1} - q_i} \times (q_{i+1} - g), \quad (20)$$

where Y_g is $\vec{Y_j}$ values in g-intermediate point. In a case $k_{\phi_{j,i}} < 1$, in respective methodology of compression of a discrete signal, $\vec{Y_j}$ components can be calculated by means of expressions:

$$Y_j = \frac{1}{k_{\phi_{j,i}}} \sum_{i=A}^{B} y_i, j = 1..J, \quad (21)$$

$$A = 1 + (j-1) \times k_{\phi_{j,i}}, \quad (22)$$

$$B = j \times k_{\phi_{j,i}}. \quad (23)$$

The use of the scaling results procedure cause the necessity of addition to input parameters of the neural network model of the parameter which corresponds to the size of a large-scale coefficient. Thus, in educational examples of the neural network model, intended for recognition of some phoneme, the number of input parameters will exceed the quantity of mel-cepstral coefficients by a unit, which characterize a standard of this phoneme. Calculation of concrete values of input parameters can be carried out, using expressions (5, 7-19). Besides, the specified research indicates a possibility to use several neural network models in the system of recognition of phonemes. The minimum quantity of such models has to be equal to a quantity of phonemes that have the standards with identical duration. The maximum quantity of models is equal to the quantity of standards of phonemes, which have to be distinguished. In addition, it should be noted that adaptation of the procedure to other natural languages, for example English, will consist only in accounting of the phonetic analysis results.

TABLE I
STRUCTURE OF EXPERIMENTAL INSTALLATION

Name	Receiving source	Functionality
Audacity	In free access, (http://audacity.sourceforge.net)	Record of a voice signal from the microphone to the file., Preliminary processing of a voice signal
Google Chrome	In free access (http://www.google.com.ua)	Recognition of separate words and phonemes
WavTest	Author's development	Extraction of separate words and phoneme-like elements.
AnalizWav	Author's development	Analysis of phoneme-like elements., Neural network, recognition of phonemes., Recognition of separate words.
Decibel Meter	In free access, (http://www.microsoft.com)	Measurement of noise level

V. PILOT STUDIES

For the purpose of verification of the offered procedure of input parameters determination, computer experiments on recognition of Russian vowel phonemes under conditions close to the ones during voice signals recognition in the system of distance learning are made [1]. The experimental installation is used with the structure, brought in tab. I.

The Windows-oriented software AnalizWav and WavTest are the main part of experimental installation. The main window of AnalizWav is shown in fig. 3. The top part of the Calculation of the parameters of sound recordings window is intended for display of parameters

The top part of the Calculation of the parameters of sound recordings window is intended for display of parameters of the current studied voice signal, recorded in the wav-file, the choice of which is implemented by pressing the Analysis of sound file button. For an example in fig. 3 parameters of the voice signal are shown. The signal is recorded in the li2.wav file. The lower part of the Comparison of sound recordings window is intended for recognition of phonemes by search of the most similar standard. Search is implemented by means of the neural network model. For ensuring service of listening of sound records in the same part of a window there are Play the first most similar entry and Play the second most similar record buttons. Identification of record is carried out according to the name of the file. Results of comparison are displayed in the field 3. For example, the record "la la1

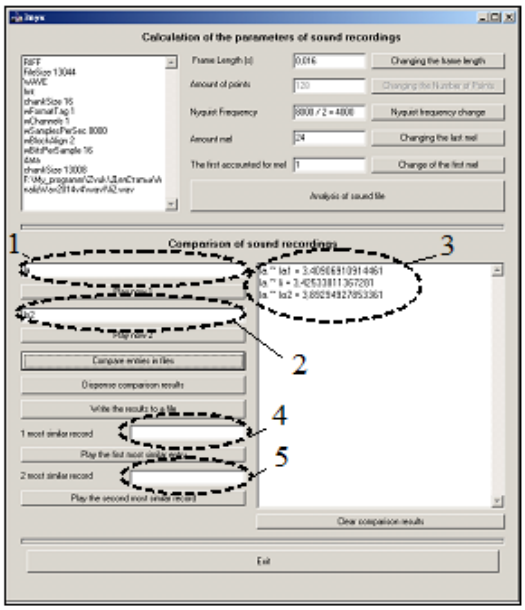

Fig. 3. Main window of the AnalizWav program

= 3,40906910914461" means that the difference between the voice signals, which are recorded in the la.wav and la1.wav, makes up 3,40906910914461. The minimum difference when voice signals coincide is equal to 0. For the recognition procedure implementation by means of AnalizWav it is necessary to execute the following operations: 1. Having pressed the Analysis of sound file button, to determine parameters of the training selection of a voice signal standards (phonemes), which have to be recognized, and parameters of an experimental voice signal (a phoneme-like element). 2. To type a name of the file into the field 1, which corresponds to the experimental phoneme-like element. 3. To consistently type the names of the files with standards of phonemes into the field 2. After each input it is necessary to press the Compare entries in files button that leads to display results of comparison in the field 3. 4. For sorting of the comparison results it is necessary to press the Dispense comparison results button. Except that in the field 3 the results of sorting are displayed, in the field 4 and 5 the names of two files with records the most similar to the experimental phoneme-like element are displayed. 5. Pressing of the Write the results to a file button leads to transferring of the results of comparison, shown in the field 3, in the text file. In the first series of computer experiments recognition of the vowel phonemes, extracted from the Russian-speaking speech corpus, were conducted. Recognition was carried out both by means of author's solutions and by means of the voice signals recognition tools, integrated into the Google Chrome browser. As words for extraction of phonemes five mere verbiages corresponding to figures from 0 to 9 were used. The words had been dictated by two female and male voices. According to the expected application conditions in the system of distance learning, the level of sound volume at record was from 40 dB to 60 dB. At the same time, the microphone was at distance of 0,2-0,3 m from the announcer, at an angle 0-20 degrees between an axis of the microphone and a source of a voice signal. The noise level measured by means of the Decibel Meter software was 10-20 dB. The Google Chrome tools were used according to public recommendations from the Google Company. When using author's development, the sequence of recognition in general corresponded to the standard flowchart shown in fig. 1. For an example in fig. 4 sonograms of [a], [o], [y] phonemes, recorded by means of Audacity, are shown.

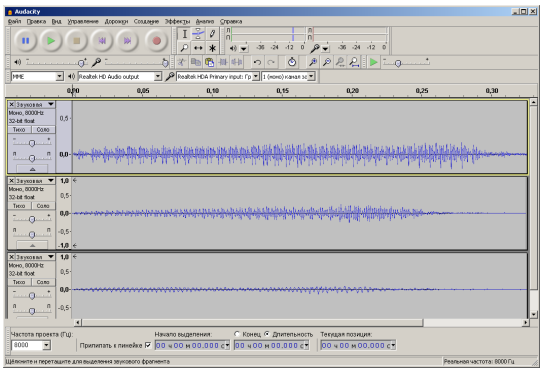

Fig. 4. Sonograms of [a], [o], [y] phonemes

Also for an example in fig. 5 the spectrogram of [a] phoneme is shown. The spectrogram is also received by means of Audacity.

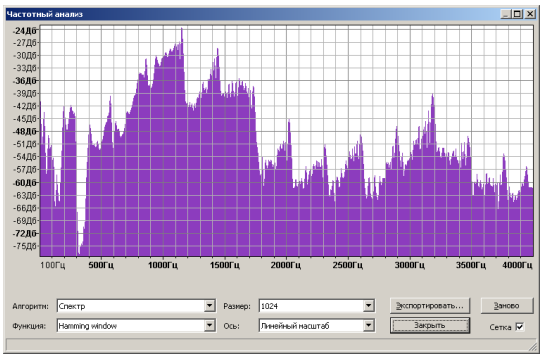

Fig. 5. Spectrogram of a phoneme [a]

It should be mentioned that spectrograms of phonemes are the basis for determination of input parameters of the neural network model. At the same time, extraction of phoneme-like elements is made by means of WavTest. On the basis of the results [8, 9] it is assumed that quantity of mel-cepstral coefficients characterizing one quasistationary fragment is equal to 24. Taking into account the need of display for input signals of the neural network model the value of large-scale coefficient, as well as the expression (4) the following is received:

$$N_x = 216 + 1 = 217. \tag{24}$$

It is worth noting that the first 216 input parameters of the neural network model correspond to the values of mel-cepstral coefficients, and the 217th input parameter corresponds to the value of large-scale coefficient (7). The two-layer perceptron neural network model with one output neuron is used for simplification of modeling. According to data [17], the amount of

TABLE II
FRAGMENT OF INPUT PARAMETERS OF THE NEURAL NETWORK MODEL

Phoneme	Values of input parameters			
	x_1	x_2	x_3	x_4
[a]	- 0,1562	- 0,0489	0,0005	- 0,0184
[]	- 0,1385	- 0,0368	- 0,0348	- 0,0103
[]	0,0465	0,0447	- 0,1265	- 0,0717
[o]	- 0,1310	- 0,0430	- 0,0432	- 0,0129
[]	- 0,1487	- 0,0518	- 0,0612	- 0,0498
[]	- 0,0539	0,0243	- 0,0074	- 0,0200

hidden neurons is accepted to be equal to 20. The fragment of the entrance data obtained by means of the offered procedure is given in tab. 2.

Results of recognition are shown in fig. 6. Also fig. 6 shows the values of average accuracy of recognition of the Russian vowel phonemes by means of the voice data input service, integrated into the Google Chrome browser. Average accuracy of

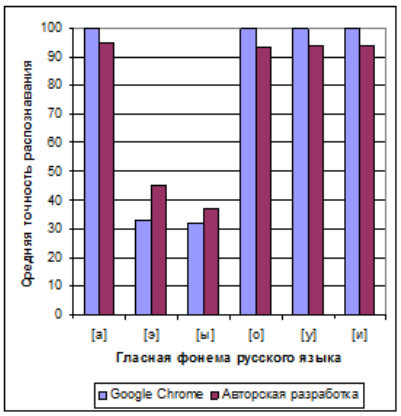

Fig. 6. Histogram of average accuracy of Russian phonemes recognition.

all vowel phonemes recognition, when using author's development, is approximately equal to 76,33According to the results of the first series of experiments, in general the accuracy of neural network recognition of vowel phonemes, implemented by means of the offered procedure of determination of input parameters, is commensurable with the recognition accuracy of innovative industrial methods. Although the author's development has some advantage over Google Chrome at recognition of [] and [] phonemes, the recognition accuracy of these phonemes remains insufficient. Thus, the applicability of the offered determination procedure of input parameters in the neural network model of phoneme recognition of a voice signal within the system of distance learning is possible to consider to be proved. Besides, being based on results of experiments and data [1, 6, 8-10, 12], it is concluded that for the acceptable accuracy of words recognition the average recognition accuracy of vowel phonemes shall not fall below 75In the second series of computer experiments recognition of the same phonemes at various noise level caused by various sources were carried out. Influence of the noise caused by a conditioner and the noise caused by a TV news broadcast was investigated. According to the expected conditions of distance learning system, noise level had been changing, ranging from 10 up to 60 dB [1]. Moreover, the top limit of noise level is received, proceeding from use of distance learning system in office rooms. The main results of experiments are shown in fig. 7 and fig. 8.

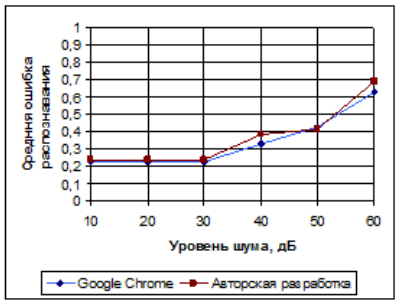

Fig. 7. The chart of dependence of an average error of vowel phonemes recognition on the noise level caused by the conditioner

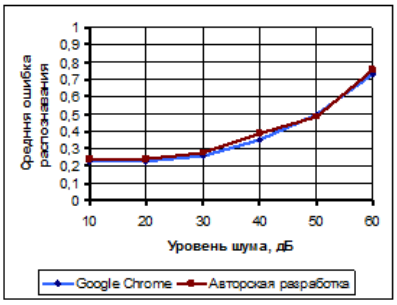

Fig. 8. The chart of dependence of an average error of vowel phonemes recognition on the noise level caused by the TV news broadcast

Analyzing the received charts it is possible to determine approximately identical resistance to noise by the system of recognition that uses author solutions, and the Google Chrome recognition systems. Also from the charts shown in fig. 7 and fig. 8 it is visible that the sufficient recognition accuracy of vowel phonemes can be provided only at the noise level to be no more than 30 dB. Therefore, the considered systems of a voice signal recognition require substantial enhancement for adaptation to the expected noise level, when using the system of distance learning in office rooms, which predetermines one of the main directions of further research. At the same time, the considered systems of recognition provide sufficient accuracy, when using the system of distance learning in premises, where a standard noise level does not exceed 30 dB.

VI. CONCLUSION

As a result of the research conducted, it is shown that the main difficulties in developing the neural network model, intended for recognition of phonemes in the system of distance learning, are connected with uncertain duration of a phoneme-like element. Due to this reason, for recognition of phonemes it is impossible to use the most effective type of neural network model on the basis of a multilayered perceptron, at which the number of input parameters is a fixed value. To mitigate this

shortcoming, the procedure, allowing to transform the non-stationary digitized voice signal to the fixed quantity of mel-cepstral coefficients, which are the basis for calculation of input parameters of the neural network model, is developed. In contrast to the known ones, the possibility of linear scaling of phoneme-like elements is available in the procedure. The conducted computer experiments confirmed expediency of the fact that use of the offered coding procedure of input parameters provides the acceptable accuracy of neural network recognition of phonemes under near-natural conditions of the distance learning system. It is also shown that prospects of further researches in the field of development of neural network means of phoneme recognition of a voice signal in the system of distance learning are connected with the necessity to increase the admissible noise level. Moreover, adaptation of the offered procedure to various natural languages and other applied tasks, for instance, a problem of biometric authentication in the banking sector, is also of great interest.

REFERENCES

[1] V. Mikhaylenko, Neural network models and methods of recognition of phonemes in a voice signal in the system of distance learning: [Monograph] / V. M. Mikhailenko, L. O. Tereykovskaya, I. A. Tereykovsky., B. B. Akhmetov. - K .: CP ”Komprint”, 2017.- 252 p.

[2] A Najib, A Basari, A Pee, M Daimon, A Rahman, L Tahir, Online performance dialogue system model (e-DP): a requirement analysis study at batu pahat district education office, Journal of Theoretical and Applied Information Technology, 31st December 2017, vol.95, no 24, p. 6699 6706.

[3] A. Mosa, M. Mahrin, R. Yusoff, A systematic literature review of technological factors for e-learning readiness in higher education, Journal of Theoretical and Applied Information Technology, 30th November 2016, vol.93, no.2, p. 500 521.

[4] I. Veritawati, I. Wasito, T. Basaruddin, Text interpretation using a modified process of the ontology and sparse clustering, Journal of Theoretical and Applied Information Technology, 15th March 2017, vol.95, no 5, p. 1019-1028.

[5] A.Kadir, A. Yauri, Automated semantic query formulation using machine learning approach, Journal of Theoretical and Applied Information Technology, 30th June 2017, vol. 95, no 12, p. 2761-2775.

[6] J. Park, J. Yoon, Y. Seo, G. Jang, Spectral energy based voice activity detection for real-time voice interface, Journal of Theoretical and Applied Information Technology, 15th September 2017, vol. 95, no17, p. 4304-4312.

[7] A. Agranovsky, D. Lednov, Theoretical aspects of algorithms for processing and classifying speech signals, - M . Radio and Communication, 2004. - Ch. 1. 164 c.

[8] L. Babenko, D. Subbotin, V. Fedorov, P. Yurkov, Definition of the borders between the fonemas by a neuroet network method, Izvestiya Southern Federal University. Technical science, 2003, no 4, t 33, pp. 321-323.

[9] T. Kartbayev, B. Akhmetov, A. Doszhanova, K. Mukapil, A. Kalizhanova, G. Nabiyeva, L. Balgabayeva, F. Malikova, Development of a computer system for identity authentication using artificial neural networks, Image Analysis & Stereology, 10.5566/ias.1612. V.36, 1, 2017.

[10] O. Fedyaev, I. Bondarenko, Neural network algorithm for speaker-independent recognition of speech phonemes, USIM, 2013, no. 4 p. 41- 50.

[11] B. Meyer , T. Jrgens, T. Wesker, T. Brand, B. Kollmeier, Human phoneme recognition depending on speech-intrinsic variability, J Acoust Soc Am, 2010 Nov;128(5):3126-41.

[12] Y. Qian, M. Bi, T. Tan, K. Yu, ”Very deep convolutional neural networks for noise robust speech recognition”, IEEE/ACM Trans. Audio Speech Language Process. , vol. 24, no. 12, pp. 2263-2276, 2016.

[13] V. Lila, E. Puchkov, Methodology of training a recurrent artificial neural network with dynamic stack memory, International magazine ”Software products and systems”, Tver, no 4, 2014 p. [on pages 132-135].

[14] Understanding LSTM Networks Posted on August 27, 2015 (http://colah.github.io/posts/2015-08-Understanding-LSTMs/) .

[15] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. Lang, Phoneme Recognition Using Time - Delay Neural Networks, IEEE Transactions On Accoustics, Speech And Signal Processing, vol. 37, 1989.

[16] M. Gusev, Methods and models of recognition of Russian speech in information systems, dis. doctors of techn. Sciences: 05.13.01 / MN Gusev - St. Petersburg, 2014. - 378 p.

[17] I. Tereykovskii, Optimization of the structure of a two-chirped perceptron, possible distribution of fertility of anomalous influences of experimental parameters of computer technology, Scientific and technical journal ”Management of branching of folded systems”, Kiev. National University of Architecture. - 2011. - vol. 5. - S. 128-131.

[18] I. Boykov, A. Ivanov, D. Kalashnikov, Algorithm of the construction of the statistic discrete-continuous description of the duration of the audio sources of the increased speech of the dictor, News of higher educational institutions. The Volga region, no 4 (36), 2015 p.64-76.