CONCEPTUAL PRINCIPLES OF NEURONETWORK RECOGNITION OF PHONEMES IN THE VOISE SIGNAL OF DISTANCE LEARNING SYSTEM MEMBERS

Liudmyla Tereikovska[1]
Ihor Tereikovskyi[2]

DOI: https://doi.org/10.30525/978-9934-26-364-4-5

Abstract. One of the most promising ways to improve the effectiveness of distance learning systems is the introduction of interactive educational materials based on the use of voice recognition tools. The introduction of known voice recognition tools into domestic distance learning systems is associated with significant financial costs and is complicated by the need for complex adaptation to the variability of application conditions, which explains the urgency of the task of developing models, methods and tools for recognizing voice signals adapted to the conditions of the distance learning system. *The subject of the research* are models, methods and means of neural network recognition of voice signals of members of the distance learning system. *The purpose of the work* is to develop the conceptual foundations of neural network recognition of phonemes in the voice signal of members of the distance learning system. *The research methodology* is based on the methods of digital signal processing, the theory of neural networks, the theory of voice recognition and involves the analysis of actual problems of recognizing members' voice signals in the distance learning system, the analysis of modern neural network solutions in the field of voice recognition, the development of a conceptual model and the principles of effective application of neural networks for recognizing phonemes in the voice signal of distance learning system members. *As a result of the conducted research*, it was determined that the main functions of the

[1] Doctor of Technical Sciences, Associate Professor,
Professor of the Department of Information Technologies Design and Applied Mathematics,
Kyiv National University of Construction and Architecture, Ukraine
[2] Doctor of Technical Sciences, Professor,
Professor of the Department of System Programming and Specialized Computer Systems,
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine

module for recognizing the voice signals of members of the distance learning system are identification at the entrance, determination of the voice response in the process of computer testing, and determination of the voice command when using the services. It is substantiated that the functioning of such a module can be implemented on the basis of neural network means of phoneme recognition. At the same time, the conditions for the implementation of voice recognition tools are characterized by limitations on the development period, the involvement of labor resources, and restrictions on access to the audio recording databases necessary for the training of the neural network model, and lead to the need to forecast the load on the server of the distance learning system and take into account the requirements for the voice transmission channel signals, microphone, room acoustics, hardware and software. Inadequate adaptation of known neural network means of recognizing voice signals to the conditions of domestic distance learning systems has been proven. It is substantiated that effective recognition can be implemented on the basis of the neural network method of recognizing selected phonemes, which involves the implementation of basic procedures related to the selection of the type and parameters of the neural network model, adaptation of the learning method, effective coding of the initial parameters, formation of an effective training sample, and forecasting sufficient computing resources. A conceptual model of neural network recognition of phonemes in the voice signal of members of the distance learning system has been developed, which ensures the formalization of the process of building neural network tools for phoneme recognition. The principles of the use of neural networks for phoneme recognition have been formed, which should be used as a basis for creating models of the processes of using neural network means of phoneme recognition in the voice signal of members of the distance learning system. The proposed list of basic procedures for building effective neural network tools for phoneme recognition, the developed conceptual model of neural network recognition of phonemes in the voice signal of members of the distance learning system, and the formed principles of using neural networks for phoneme recognition form the conceptual foundations of neural network recognition of phonemes in the voice signal of members of the distance learning system.

1. Introduction

The world experience in the development of distance learning systems (DLS) shows that one of the most promising ways to increase their effectiveness is the introduction of interactive educational materials based on the use of voice signal (VS) recognition tools. In addition to the proven increase in the quality of education, the use of the specified tools allows to increase the convenience of using the DLS due to the lack of strict binding to the class schedule, to better meet the needs of students with disabilities, and to increase the security of the DLS due to the introduction of biometric authentication tools. At the same time, in most of the known DLS, there are no means of recognition of VS, although the possibility of their implementation is confirmed by the wide use of software applications (Google+, Microsoft Office, VoiceNavigator) with the appropriate functionality. At the same time, the introduction of well-known means of recognition of VS in domestic DLSs necessitates their complex adaptation to the variability of application conditions related to the term of development, the volume of the dictionary, the formation of educational databases, the permissible value of the recognition error, acoustic factors, resource-intensive creation and operation. Also, the disadvantages of common means of recognition of VS are high cost and lack of detailed scientific and technical documentation. In such a setting, the task of developing models, methods and means of recognition of VS, adapted to the conditions of DLS, is urgent.

According to [7, p. 125], the most complex stage of the VS recognition system (VSRS) is the implementation of the recognition procedure, the result of which is the determination of a standard corresponding to an unknown VS. The complexity of the recognition procedure is explained by the non-linear change in the pace of speaking words and the different duration of pauses at the beginning and end of the word. Therefore, the recognition procedure is divided into several stages. The input VS is divided into elements – phonemes, allophones, diphones, triphones, syllables. There are standards for the specified elements, and with the help of the standards of the elements, there are standards of individual words. The division of VS into separate elements is performed on the basis of the analysis of energy components of VS.

The methods of recognizing individual words based on element standards are also considered sufficiently tested and reliable. At the same time, the

problem of finding standards of individual elements is far from being solved. The results [19, p. 346] allow us to assert the perspective of using phonemes as individual elements, which is explained by their relative small number compared to the number of syllables, allophones, diphones, and triphones. The process of determining the limits of individual phonemes in the VS is described in [20, p. 8]. The analysis showed that the vast majority of tested VSRS are built on the basis of dynamic programming methods, neural networks (NN) and hidden Markov models. The advantages of the dynamic programming method are the ease of establishing the time correspondence between the input VS and the reference one, and the disadvantages are high computational complexity and diction dependence.

The application of NNs is based on their ability to classify VSs, given using coefficients that correspond to the calculated vector of VS features [21, p. 19; 23, p. 32]. Advantages of neural network methods: proven effectiveness in solving difficult-to-formalize problems, resistance to noise in input data, high speed of calculations and low needs of computing resources when making a decision, resistance to partial failures in the hardware implementation of neural network models (NNM). Disadvantages include the difficulty of adapting the NNM to a non-stationary input signal, the problems of choosing the parameters of the NNM.

The use of hidden Markov models is based on the postulate that the VS can be represented using a hidden Markov chain. The advantages of the method are the simplicity of its application. The main disadvantages of hidden Markov models include the complexity of forming a database of multivariate reference elements of words and the high computational complexity of calculating the parameters of the Markov model. These shortcomings are somewhat compensated by the combined use of Markov models and NNM, which in turn negatively affects the complexity of the model. In addition, there are known attempts to use dynamic Bayesian networks, support vector machines, and the theory of non-force interaction in VSRS. The widespread use of these methods is hindered by low reliability and the need for adaptation to the practical aspects of application in VSRS. Thus, one of the promising ways to solve the scientific and practical problem of recognition of the VS of DLS members is the development of neural network methods for recognizing phonemes extracted from this VS. Since the basis of the development of such methods is not sufficiently

covered in the available scientific and practical sources, the purpose of this scientific study is to develop the conceptual foundations of neural network recognition of phonemes in the voice signal of members of the distance learning system. To achieve the goal of the research, the following tasks should be solved:

– To characterize the functions of the VS recognition module in DLS.

– To characterize modern neural network means (NNM) of VS recognition from the point of view of the possibility of their application in DLS.

– To develop a conceptual model for ensuring the effectiveness of neural network phoneme recognition in the VS of a member of DLS.

– To form the principles of the application of NN for phoneme recognition in the VS of a member of DLS.

In order to solve the stated tasks, the research methodology provides for the analysis of the current problems of recognition of the VS of DLS members, the analysis of modern NNM and neural network methods of recognition of VS, the development of a conceptual model and the principles of effective application of NN for the recognition of phonemes in the VS of a DLS member.

2. Actual problems of VS recognition in DLS

It is generally accepted that DLS is a complex of educational services that are provided to students using a specialized information and educational environment, which is based on the means of remote exchange of educational information [3, p. 12]. Most DLSs are web-oriented and built on a client-server architecture. DLS uses a thin client, so most computing operations are performed by the web server.

One of the promising ways to improve the quality of education of students of the DLS is the use of interactive educational materials, which are based on the use of means of automating voice interaction [22, p. 93]. The use of these tools makes it possible to increase the ease of use of DLS and meet the needs of members with disabilities. The results of the analysis of DLS and similar ISs for supporting the educational process indicate that they lack means of automating voice interaction. It should be noted that the list of considered DLS included: Lotus (USA), Blackboard (USA), ATutor (Canada), Claroline (Belgium), Dokeos (Belgium), LAMS (Netherlands), Moodle (Australia), OLAT (Switzerland), OpenACS (Germany), Sakai

(USA), Chamilio (Belgium), DocebolLMS (USA), Dokeos (France), ILIAS (Germany), Open Elms (USA), SharePointLMS (EU), Adobe Connect Training (USA), Collaborator (Ukraine).

In general, the task of VS recognition consists in its automatic processing in order to determine the sequence of words [24, p. 125]. The complexity of solving such a problem is explained by the need to take into account the variability of the VS, the type of language input, the size of the dictionary, and the level of ambient noise. To solve the problem of VS recognition, VSRS are created, the creation of which is complicated by the need to take into account various factors, for example, the location of the microphone. Modern VSRS, as a rule, have a hierarchical structure. At the first acoustic level, preliminary processing and selection of acoustic features that characterize the VS is performed. The next level of VSRS is the linguistic level, which includes the procedure for searching for VS in reference dictionaries. In addition, VSRS can include phonetic, phonological, morphological, lexical, syntactic and semantic levels. A typical sequence of VSRS functioning is shown in Figure 1.

In general, the development of VSRS in DLS is a complex scientific and practical task. At the same time, the results [20, p. 6] indicate that the actual practical tasks for DLS are the development of means of voice identification of the user when entering the DLS, means of determining the voice response in the computer process computer testing and means of determining the voice command of the DLS user. From the point of view of the theory of VS recognition, the development of such tools boils down to solving the problem of recognizing isolated words in VS, which, according to the results of clause 1.1, can be implemented due to phoneme recognition using NN. At the same time, due to the requirements for the characteristics of the VS transmission channel, the microphone, the acoustics of the premises, and the hardware and software, it is possible to significantly reduce the number of factors that affect the recognition efficiency.

In addition, the conditions for the implementation of the NNM of VS recognition are characterized by limitations on the development period, the involvement of labor resources, and restrictions on access to the databases of audio recordings necessary for the training of the NNM. Also, based on the client-server architecture, it is possible to conclude that a new server recognition module should be added to the DLS. The functioning of such

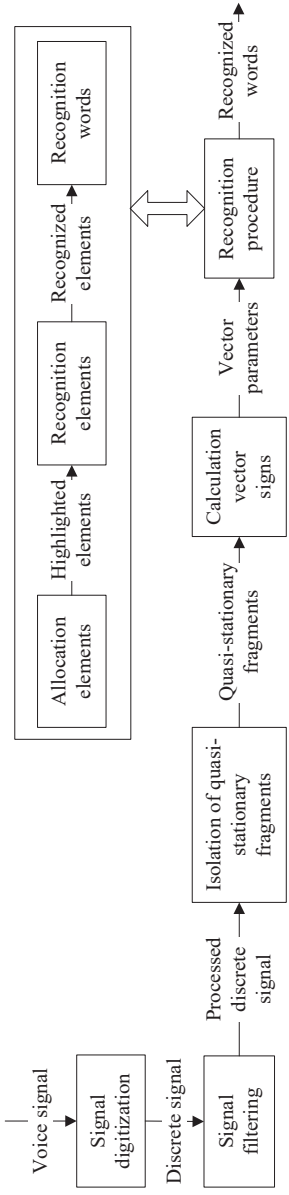

Figure 1. Typical sequence of functioning of the voice signal recognition system

a module can lead to the use of additional computing resources of the server, the amount of which is fixed and quite limited in domestic DLSs. Therefore, the application of VSRS in DLS leads to the need to predict the sufficiency of the amount of computing resources of the web server.

3. Analysis of neural network solutions for VS recognition

In the process of analysis of works devoted to the use of NN for recognition of VS, it was found that most of them are characterized by some inconsistency in the name and description of the given development. For example, the title of the work indicates the development of a neural network system (NNS), although the actual development consists in defining an algorithm for processing input parameters for NNM. Therefore, the analysis of these works was carried out from the single point of view of defining the main characteristics of neural network methods and models. The obtained data is presented below.

The algorithm for automatic syllable recognition based on linear autoregressive NN and fuzzy set theory is presented in [18, p. 59]. The paper proposes a definition of a phoneme as a vague set of minimal language units. An algorithm for syllable recognition is proposed. The input of the algorithm is a set of reference minimal language units.

The task of the algorithm is to assign an unknown language unit to one of the reference ones. Examples of application of the algorithm are shown. The given results indicate insufficient accuracy of recognition, which is in the range of 3-35%.

In [6, p. 151], a method for recognizing voice commands using the Kohonen map is given. An example of designing a command-based announcer-dependent system is shown. The mathematical support for constructing the Kohonen map topology is described. In order to train such a network, it is proposed to prepare a training sample from various variants of spoken specified commands.

A neural network with a time delay for recognizing the consonant sounds "B", "D" and "G" is presented in [27, p. 330]. The input signals are 16 normalized scaled spectral coefficients. Scaled coefficients are calculated from the signal power spectrum. The method of inverse error propagation was used in the neural network with a time delay. It is proposed to adjust the weight coefficients for time-shifted connections, namely to update each weight of the corresponding connection to the average value of all time-delayed weight changes. Such a computational procedure is quite complex, which is explained by the large number of iterations that are necessary for memorizing a complex multidimensional space of weighting coefficients and training images.

In [20, p. 10], an approach to learning a neural network using a genetic algorithm is proposed. An example of recognizing ten words is given. The word recognition mechanism provides for the association of each individual word with a separate NN. An unknown example is applied to the inputs of all NNs. After that, the NN with the maximum output signal is determined. A word associated with a given network is assumed to match an unknown example. The training of the NN was performed by sequentially presenting the training examples and adjusting the connection weights until the training error over the entire set reached an acceptably low level.

Recognition of voice commands using convolutional neural networks is proposed in [13, p. 20]. Such networks work with two-dimensional data – neurons in each layer form planes. The input layer is represented by one plane, the dimension of which coincides with the dimension of the input data. The next layers of the network are convolutional and consist of neuron planes (feature maps). Each neuron of the collapsed layer is connected to

a small subregion of the previous layer. The last two layers of the network represent the NN of direct propagation. An example of practical application is given. The advantages of convolutional neural networks are shown: a small number of training parameters and high training accuracy.

In [14, p. 657], an algorithm for recognizing isolated words was developed using the RBF network, which was used to select the most informative features of standards. The RBF network was trained using the methods of cluster analysis and gradient descent. Also presented is the algorithm for configuring VSRS for a new announcer. It is proven that the use of the RBF network allows to significantly increase the frequency of the correct result of recognizing problematic (acoustically similar) words and to reduce the volume of the training sample for the procedure of setting up the recognition system for a new speaker.

The method of word recognition based on two-dimensional vectors is shown in [20, p. 11]. The classification of vectors is carried out with the help of NN, the input of which is low-frequency two-dimensional wavelet transformation of intervals of spectrograms of individual words. The spectrogram is scaled to a square form and a two-dimensional wavelet transform is performed. It is shown that the main advantage of the method is the concentration of the most noise-resistant frequency-time information in a compact frequency-time vector. An example of recognizing a voice stream with a limited number of words is given. The results allow us to determine the insufficient accuracy of speech recognition (30-35%).

The method of automatic recognition of isolated words is described in [25, p. 139]. When building the architecture of the NN and choosing the number of input neurons, the actual aspects of human language perception are taken into account. Therefore, the NN consists of 25 neurons of the upper layer, three neurons of the inner layer and one output neuron, the activation function is sigmoidal. In the experiment, ten numbers were spoken, as well as the names of 40 cities. The number of announcers is 17 women and 15 men. The mel-cepstral analysis algorithm was used for phoneme selection. The given results indicate insufficient accuracy of speech recognition, which is in the range of 13-23%.

The method of phoneme recognition in the voice signal is described in [20, p. 12]. It is shown that the use of phoneme recognition methods requires preliminary segmentation of the VS to obtain the vector of properties of

individual phonemes. The shortcomings associated with the predetermined type and number of conversion factors, which can lead to duplication and redundant information in the VS, are pointed out. The proposed structure of the phoneme recognition system in VS. The input signal enters the input of the delay line, which forms the input vector NN. Next, the signal enters the neural correlation layer. Then – to the input of the analysis layer. The outputs of the analysis layer are features of a separate phoneme. The advantages of the system include the compactness of its structure.

In the neural network system for recognition of VS of the company Microsoft, NNMs of the type with direct signal propagation are used. Although it was not possible to find a detailed description of this system, but based on indirect data [5, p. 125; 20, p. 12] and based on practical experience, it is possible to note that a multi-criteria approach was used to determine the optimal type and parameters of the NNM, the training method was optimized with positions to minimize the training period. In addition, when building the system, limitations related to computing resources are taken into account.

The approach of the combined use of multilayer NNMs and hidden Markov models is considered in [26, p. 127]. The considered advantages of NN are the possibility of learning with the help of a small acoustic database with phonetic transcription, while the parameters of the Markov chain are estimated only on the basis of a large lexical database. It is indicated that the disadvantage of NNM is its poor adaptability to work with time series. It is also shown that the compatible approach makes it possible to use conditional probabilities at the stage of linguistic modeling and thereby narrow the search circle.

The neural network training method without word segmentation is proposed in [17, p. 57]. The use of NN with direct signal propagation, which is intended for the recognition of individual phonemes in VS, is considered. As input parameters, 16 energy characteristics of VS obtained in 15 time windows were used. Each of the output neurons of the NN corresponds to a separate phoneme. In the process of learning on each individual word, it is provided not to adjust the weight coefficients of connections leading to output neurons for which there are no corresponding phonemes in this word. Due to this, the possibility of refusing the procedure of preliminary segmentation of the VS is declared.

In [4, p. 10], the principle of increasing the speed of learning and generalizing capabilities of NN due to the use of wavelet transformations is proposed. Based on this approach, a new modification of NN – with a module of inverse wavelet decomposition of the output signal was developed. It is proposed to use the specified approach and modified NN for recognition of VS. The results of comparative experiments on the recognition of the isolated sound "a" using conventional and modified NN are presented. A significant advantage of the latter is noted, which is expressed in a tenfold reduction in the training time and in a reduction in the relative recognition error from 11.57% to 0.003%.

The neural network algorithm for speech recognition is described in [16, p. 96]. The recognition algorithm consists of the following stages: entering an acoustic signal into the computer and selecting word boundaries, selecting parameters that characterize the spectrum of the signal, using NN to assess the proximity of acoustic parameters, and comparing with standards in the dictionary. As 39 input signals of NN, the results of spectral analysis of VS were used. The given results prove that the accuracy of neural network recognition is not inferior to the accuracy of traditional acoustic-phonetic methods.

In [2, p. 23], the clustering of the space of features of speech signals using the Kohonen map is described. The possibility of using the Kohonen map for phoneme clustering and word clustering is shown. Various modifications of the methods of teaching NN have been proposed. It is proved that the accuracy of recognition of VS is within 96-99%.

In [15, p. 102], the rationale for the selection of linguistic features for learning NN is given. It is proved that phonemes can be the initial data for learning NN. It is shown that it is possible to apply the results of: wavelet transform, windowed Fourier transform, or Hilbert-Huang transform to process the energy indicators of VS to form input data of NN. The given results indicate the same effectiveness of the indicated transformations.

The method of fuzzy matching of speech images in the neural network basis for recognizing isolated words of the Russian language is described in [8, p. 115]. In the method, the speech signal is presented in the form of a two-dimensional spectral-temporal image obtained using a windowed Fourier transform. The unknown image is considered as a clear relation between a set of frequencies and a set of time intervals. It is indicated that when using this method, the accuracy of recognition reaches only 76%.

In [12, p. 59], a neural network approach to the integrated representation and processing of information in intelligent systems was developed. Methods and algorithms of neural network processing of language information were developed and keyword recognition NNS was created. It is proposed to use dynamic associative memory devices based on recurrent NNs. The parameters characterizing the sound wave were used as input information. In general, the recognition system represents a hierarchical structure that analyzes the speech flow at the acoustic-phonetic, morphological, lexical and semantic levels. For the analysis at the acoustic-phonetic level, the Kohonen map was used, the input data of which was obtained by applying the perceptual linear prediction method to the VS.

In [9, p. 45], a set of neural network models for phoneme recognition was considered in VS when controlling a text editor. A two-layer perceptron with a 20-10-1 connection structure was used, the number of inputs of which corresponds to the period of the main tone of a specific announcer.

In [20, p. 15], a system of automatic speech recognition based on neural network technology is described. The system has a hierarchical structure that corresponds to the traditional levels of linguistic information processing – acoustic-phonetic, lexical and phrasal. A specific structure of the NNM, which is based on the use of the RBF network, has been developed. The peculiarities of setting up the recognition system for a specific announcer are considered. General approaches to the use of recurrent NNs at the upper levels of analysis are also described.

In [10, p. 512], the method of determining the most informative features of the speech signal is considered. The method is designed to determine the input parameters of the NN, which is used for the selection of vocalized segments and the classification of vowel phonemes. It has been proven that the correlation parameter between readings is highly informative in the task of isolating vocalized segments. It is proposed to use parameters that take into account the structure of vowel sounds to search for vowel formants.

In [11, p. 3], a method for building a speech recognition device based on a hybrid neuro-Markov model is proposed. It has been proven that NNs allow creating acoustic models that are more compact compared to models based on the theory of hidden Markov models. It is declared possible to create such a model based on NN with direct and reverse connections. Two cases of combining NNM and Markov model are described. In the first

case, NN is used only at the stage of acoustic analysis, and at the stage of lexical analysis, a hidden Markov chain is used. In the second case, NN is used to determine the transition probabilities of the hidden Markov chain. It was noted that the difficulties of creating the NNM are related to the non-stationary nature of the VS.

The article [9, p. 47] describes the method of sliding phonetic analysis and the structure of a multi-level system of automatic speech recognition based on NN. The method is based on the postulate that vocalized VS consists of stable intervals that characterize phonemes and unstable intervals that refer to interphoneme transitions. To ensure insensitivity to changes in the duration of phoneme sound, the method uses the idea of approximating the stationary intervals of the NNM of reference elements of speech. To describe the standard, it is proposed to use NNM with the structure 80-10-10-1. The given results of phoneme recognition confirm the promisingness of the proposed method.

A neural network model for fusion speech recognition is proposed in [28, p. 1678]. NNMFSR is based on the concept of dividing VS into separate phonemes and transitions between them. The model consists of three modules representing NNs capable of learning using Hebb's rule. The input of the model is the mel-cepstral coefficients of the VS. In the first module, the VS is segmented into intervals of different lengths. Each of the intervals corresponds to various variants of phonemes and transitions between them. In the second module, phonemes are recognized, and in the third – individual words. The given results indicate that the recognition accuracy is about 90%.

So-called deep NNs are used in the VS recognition tools used in the Google search system and the Android operating system [13, p. 21]. At the first stage, the weighting coefficients are pre-adjusted. The method "without a teacher" is used, which is based on the application of a restricted Boltzmann machine. At the second stage of training, the method "with the teacher" is implemented using the backpropagation algorithm. The use of DNN allows processing powerful voice streams, but requires significant computing power. We should also note that a complete detailed description of the methods of construction of DNN has not been found.

In [1, p. 2778], the automatic phoneme recognition system in VS is described. The system provides a two-stage recognition method. At the first

stage, VS is divided into phonemes. For this, the analysis of the frequency characteristics of the VS is used. Phoneme recognition occurs at the second stage using NNM or hidden Markov models. The structure of APRS is presented, the preparation of its input parameters and the segmentation algorithm are described.

As a result of the analysis, taking into account the defined conditions of the tasks of recognition of VS in DLS, by analogy with [24, p. 126], it was determined that the effectiveness of NNM recognition of VS in DLS is associated with the implementation in them a set of basic procedures ***G***: single-criterial and multi-criteria selection of the type of NNM (G_1, G_2), single-criteria and multi-criterial selection of NNM parameters (G_3, G_4), adaptation to the training method (G_5), effective coding of output parameters (G_6), selection of permissible types of NNM (G_7), formation of effective training set (G_8), predicting the sufficiency of computational resources (G_9). It is also possible to argue about the insufficient adaptation of known NNMs to the conditions of domestic DLSs. As a result, the conducted analysis indicates that effective recognition of VS in DLS can be implemented on the basis of the neural network method of recognizing selected phonemes, which involves the implementation of certain basic procedures. The development of the specified method leads to the need to improve the methodological principles of the application of the NNM of phoneme recognition in the DLS VS and predict the adequacy of computing resources of the DLS web server.

4. Conceptual model of neural network phoneme recognition

A peculiarity of the task of neural network recognition of phonemes in the VS of DLS members is the need for theoretical substantiation of the characteristics of neural network models and methods adapted to the conditions of DLS both in the case of announcer-dependent and in the case of announcer-independent recognition. The indicated conditions include the permissible development period, the possibility of attracting labor resources, the availability of access to the databases of audio recordings necessary for the training of NNM, the peculiarities of the acoustic parameters of the VS and the available amount of computing resources of the DLS web server. Solving the given scientific task will allow solving such practical problems as voice response recognition in the process of computer testing, voice

command recognition and voice authentication of DLS users by recognizing the secret word (password) spoken by them. In the case of announcer-dependent recognition, the specified practical tasks take into account the features of a specific user's VS, and in the case of announcer-independent recognition, these features are not taken into account. At the same time, the tasks of filtering the VS, selecting phonemes from the VS, and forming individual words from the recognized phonemes are considered solved.

According to recommendations [19, p. 347], the starting point for solving the formulated task was the development of a conceptual model for ensuring the effectiveness of neural network phoneme recognition. In general, a conceptual model is a model of a subject area consisting of a list of interrelated concepts that are used to describe this area together with properties and characteristics, classification of these concepts by types, situations, signs in this area, and the laws of flow in it processes. A conceptual model is a reflection of a concept, the concept of which is understood as a certain method of judgment, interpretation of certain phenomena, a basic point of view, a guiding idea for their systematic coverage. We note that the development of a conceptual model is a generally accepted starting point for the development of a methodological base, which is a system of principles and methods of organizing and building theoretical and practical activities, as well as teaching about this system. Since the practical result of the dissertation work involves the creation of software and hardware for phoneme recognition, to determine the effectiveness of the process of neural network phoneme recognition, it is planned to use definitions from the field of computer and software engineering. According to international standards in this field, efficiency is a set of attributes that determine the relationship between the levels of execution of the software system, the use of resources (tools, equipment, materials, etc.) and services performed by full-time service personnel, etc. The performance characteristics of the software system include:

– responsiveness – an attribute that indicates the time of response, processing and performance of functions;

– resource intensity – an attribute that determines the amount and duration of resources used when performing functions of the software system;

– compliance – an attribute that indicates the compliance of this attribute with given standards, rules and prescriptions.

In accordance with the above definitions, at the first stage of creating a conceptual model, the terminology used in the field of application of NN for recognition of VS was harmonized. Harmonization is carried out from the standpoint of reflecting the current state of science and practice and supports the solution of the problems of the dissertation work. As a result, the following terms are defined:

– VS is a complex acoustic signal, the source of which is the human voice. In the context of this dissertation, a synonym for the term VS is a speech signal, although in general there are certain differences between these terms.

– Phoneme-like element – a fragment selected in the VS, the parameters of which correspond to a separate phoneme.

– Phoneme is a minimal structural and functional sound unit of language, which serves to find differences and identify significant units of language.

– NN is a network of artificial neurons interconnected by synaptic (weighted) connections.

– NNM – a model of NN, characterized by the method of learning, the method of signal propagation, the structure of connections and the type of artificial neuron. The specified parameters and their combinations determine the type of NNM. A synonym of the concept of the type of NNM is the architecture of NNM. Derived from the term NNM are neural network methods, NNS and NNM, that is, these are methods, systems and tools that are based on NN. Since in the general case the concept of a tool is understood as a tool (object, device, set of devices), the concept of NNM is collective for NNM and NNS, which are used to recognize phonemes in VS DLS. The hardware and software implementation of such devices will be called instrumental NNM. It is also determined that in relation to the task of this dissertation research, the conceptual model is intended for the formalization of cause-and-effect relationships, which are inherent in the process of phoneme recognition in VS, determined by the need to increase the efficiency of DLS. In addition, the conceptual model should take into account:

– operating conditions of the NNM of phoneme recognition, determined by the nature of the interaction of its individual parts and components of the DLS;

– the need to implement the effective application of NNM for phoneme recognition and the direction of improving its functioning;

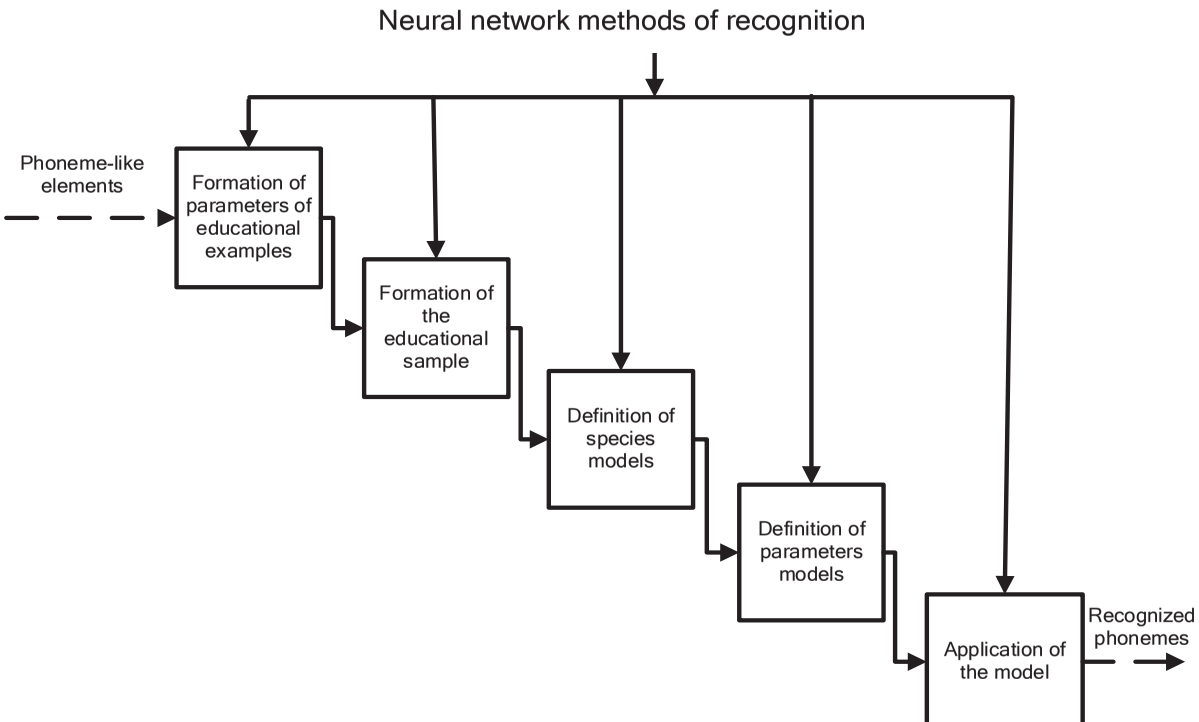

Figure 2. Decomposition diagram of neural network phoneme recognition

– possibilities of managing the NNM and defining its controlled variables.

Taking into account the generally accepted technology of the application of the NNM, constructed shown in Figure 2. Decomposition diagram of neural network phoneme recognition. The purpose of the component parts of this diagram is as follows:

– Formation of parameters of educational examples – determination of a set of input and output parameters for each phoneme and the method of their coding to a form suitable for use in NNM.

– Formation of the training sample – determination of such a set of training examples that correspond to the standards of phonemes. The number, quality, and nomenclature of examples should be sufficient for teaching NNM.

– Determination of the type and parameters of NNM – the choice for the application of this type of NNM, with such parameters that most fully meet the conditions of the task of phoneme recognition in the VS of a specific DLS.

– Application of NNM – phoneme recognition in VS. It should be taken into account that the use of NNM leads to a load on the DLS web server and may lead to the exhaustion of its computing resources.

The next stage of creating a conceptual model was the development shown in Figure 3 schemes of NNM phoneme recognition components.

The scheme takes into account the peculiarities of the implementation of the NNM in the conditions of the DLS:

– imperfection of the methods of forming the parameters of training examples for NNM, which are intended for phoneme recognition;

– a long period of training sample formation for NNM in the case of limited access to the phoneme database;

– difficulty of accessing existing phoneme databases;

– additional load on the DLS web server due to the functioning of recognition tools.

Therefore, the scheme provides for the possibility of forming the parameters of training examples and the training sample with the help of expert data.

Analysis of the data shown in Figure 2 and Figure 3 allows us to state that the effectiveness of neural network phoneme recognition is influenced by a number of factors shown in Figure 4.

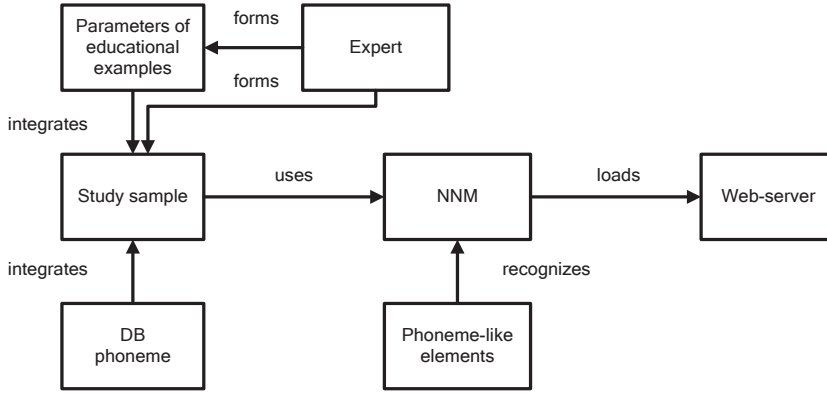

Figure 3. Scheme of interaction of NNM components of phoneme recognition in VS DLS

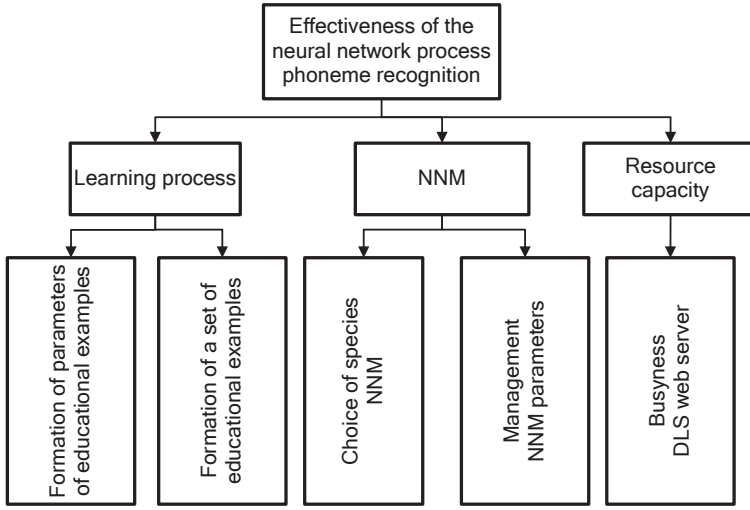

Figure 4. Factors affecting recognition efficiency

In addition, it can be argued that the effectiveness of neural network recognition should be evaluated from the point of view of the effectiveness of the NNM application process and from the point of view of NNM training. Performance indicators should reflect the duration, resource intensity and accuracy of the specified processes. Thus, the data shown in Figure 5 indicators for evaluating the effectiveness of neural network phoneme recognition.

As a result, it was determined that in an analytical form, the conceptual model of ensuring the effectiveness of the neural network phoneme recognition process can be displayed using expressions (1-3).

$$E_{\Sigma} = f\left(E_{NN}, E_{DS}\right), \tag{1}$$

$$E_{NN} = f\left(e_1, e_2\right), \tag{2}$$

$$E_{DS} = f\left(e_3, e_4, e_5\right), \tag{3}$$

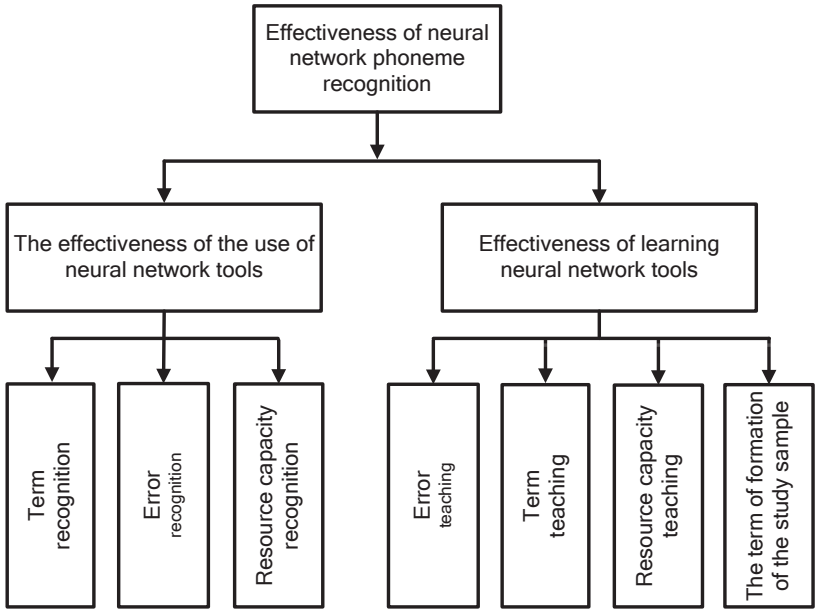

Figure 5. Indicators for evaluating the effectiveness of neural network recognition

where E_{Σ} is the integral efficiency of the process; E_{NN} – efficiency of creation and application of NNM; E_{DS} – effectiveness of creating a training sample; e_1 – definition of effective types of NNM; e_2 – determination of NNM parameters; e_3 – resource intensity of NNM application; e_4 – determination of parameters of educational examples; e_5 – training sample formation.

The analysis of the developed conceptual model allows us to state that for the effective application of NNM for phoneme recognition in the VS DLS, it is necessary to supplement the methodological base with the following principles: the admissibility of the application of the type of NNM, the determination of the set of effective types of NNM, the evaluation of the effectiveness of the NNM type, the determination of the expected output signal for phoneme standards, the forecast use of phoneme recognition NNM of computing resources of the DLS web server, assessment of NNM efficiency and use of expert knowledge to form a training sample.

5. Principles of application of neural networks

The principle of determining the set of effective types of NNM for phoneme recognition in VS DLS. According to the conclusions of [21, p. 78], the determination of the set of types of NNM, which ensure effective recognition of phonemes in the VS DLS, is presented as a procedure of the following type:

$$\boldsymbol{M}_{\boldsymbol{d}} \rightarrow \boldsymbol{M}_{\boldsymbol{d}} \rightarrow \boldsymbol{M}_{\boldsymbol{e}}, \tag{4}$$

where $\boldsymbol{M}_{\boldsymbol{a}}$ is the set of available types of NNM; $\boldsymbol{M}_{\boldsymbol{d}}$ is a set of permissible types of NNM; $\boldsymbol{M}_{\boldsymbol{e}}$ is the set of effective types of NNM.

The principle of admissibility of the application of the NNM type for phoneme recognition in the VS DLS. As evidenced by the results of the conducted research, the main factor that affects the formation of a set of permissible types of NNM used to recognize phonemes in VS in DLS is the provision of effective training of NNM. To do this, it is necessary to perform the following procedures in an acceptable time: determine and code the input and output parameters of the NNM, create a training sample and implement the training process. The first procedure is carried out at the preparatory stage of the development of the NNM. Therefore, attention is paid to the implementation of the second and third procedures. The acceptable term for the creation of the training sample and the training of NN is determined based on the requirements for the creation of DLS resources, i.e.:

$$t_g \le t_d , \tag{5}$$

where t_g is the general term of learning NN to recognize phonemes in VS; t_d is an acceptable term for the creation of the NNM of phoneme recognition in VS.

Thus, the principle of admissibility of using the *i*-th type of NNM for phoneme recognition in VS DLS is given by the following rule:

$$If\ t_g\left(m_i\right) \le t_d \rightarrow m_i \in \boldsymbol{M_d} , \tag{6}$$

where m_i is *i*-th type of NNM; $\boldsymbol{M_d}$ is the set of admissible types of NNM.

The principle of evaluating the effectiveness of the type of NNM intended for phoneme recognition in the VS DLS. By analogy with [19, p. 350], it is considered that among the set of admissible i-th type of NNM is the most effective if the efficiency function for it takes the maximum value:

$$\max_{V_i} = \left\{V_1, V_2, \ldots V_I\right\} , \tag{7}$$

where *I* is the number of NNM species; V_i is the efficiency function of the *i*-th type of NNM.

The calculation of the value of the efficiency function is performed as follows:

$$V_i = \sum_{k=1}^{K} \alpha_k R_k\left(m_i\right), m_i \in \boldsymbol{M_d}, i = 1, \ldots I \tag{8}$$

where α_k is the weighting factor of the *k*-th efficiency criterion; m_i – *i*-th type of NNM; $\boldsymbol{M_d}$ is the set of permissible types of NNM; *K* is the number of efficiency criteria; $R_k\left(m_i\right)$ is the value of the *k*-th criterion for the NNM of the *i*-th type.

According to the results of [19, p. 350], the criteria for choosing the most effective type of NNM should reflect the extent of its adaptability to the given applied problem. Thus, by the *k*-th criterion for determining the most effective type of NNM, we will understand the degree of provision in the NNM of the *k*-th requirement of the problem of phoneme recognition in the VS DLS. An example of such a criterion is the measure of ensuring the requirement for non-iterative learning in NNM.

The principle of determining the expected output signal of NNM for phoneme standards. The value of the output signal should reflect the similarity of the initial examples. Otherwise, learning NNM can become difficult. Therefore, the output signal for phoneme standards is proposed to be described by the expression:

$$\boldsymbol{Y}_{\boldsymbol{\phi}} = f(d_{\phi}), \tag{9}$$

where $\boldsymbol{Y}_{\boldsymbol{\phi}}$ is the expected output signals of NNM for phoneme standards $\boldsymbol{\Phi}$; d_{ϕ} is a set of measures of similarity between the components of $\boldsymbol{\Phi}$.

Since phonemes are supposed to be recognized based on the analysis of their acoustic characteristics, it is proposed that these characteristics be reflected in the degree of similarity between phonemes.

The principle of forecasting the use of NNM phoneme recognition of computing resources of the DLS web server. The load on the DLS web server due to the use of phoneme recognition NNM directly depends on the number of calls to this system. Similar processes are effectively described using the theory of wavelet transforms. Therefore, it is proposed to predict the load of the web server using a wavelet model of the change in the number of requests:

$$Q_{web} = f\left(W\right), \tag{10}$$

where Q_{web} is the load of the DLS web server; W is a wavelet model of changes in the number of calls to the DLS web server.

The principle of evaluating the effectiveness of NNM phoneme recognition. By analogy with [20, p. 37], it is proposed to evaluate the effectiveness of NNM based on a set of parameters that indicate the degree of ensuring that they perform the procedures that are considered necessary for effective phoneme recognition in the VS DLS:

$$E_s = f(\boldsymbol{\Pi}), \tag{11}$$

where E_s is the efficiency of NNM; $\boldsymbol{\Pi}$ is a set of proposed parameters.

The principle of using expert knowledge to form an educational sample. This principle assumes that training examples can be formed on the basis of expert knowledge about phonemes in the form of production rules of the form:

$$If\ x_1 \in \left[X_1^{min}, X_1^{max}\right] \wedge ... \wedge x_K \in \left[X_K^{min}, X_K^{max}\right] \rightarrow Y, \tag{12}$$

where $x_1,\ldots x_K$ are phoneme identifying parameters; $\left[X_1^{min}, X_1^{max}\right],\ldots\left[X_K^{min}, X_K^{max}\right]$ – given ranges $x_1,\ldots x_K$; K is the number of identifying parameters; Y is the result of the production rule (expected phoneme).

It is expedient to use the developed principles described by expressions (4-12) as a basis for creating models of the processes of using NNM for phoneme recognition.

6. Conclusions

It was determined that the main functions of the module for recognizing the voice of DLS members are their identification when entering the DLS, determining the voice response in the process of computer testing and determining the voice command when using DLS services. It is substantiated that the functioning of such a module can be implemented on the basis of the NNM of phoneme recognition in VS. At the same time, the conditions for the implementation of the NNM of VS recognition are characterized by limitations on the development period, the involvement of labor resources, and restrictions on access to the audio recording databases necessary for the training of the NNM, and lead to the need to forecast the load on the DLS server and take into account the requirements for the VS transmission channel, microphone, acoustics premises, hardware and software.

Inadequate adaptation of known NNMs of VS recognition to the conditions of domestic DLSs has been proven. It is substantiated that effective recognition of VS in DLS can be implemented on the basis of a neural network method for recognizing selected phonemes, which involves the implementation of basic procedures: single-criteria and multi-criteria selection of the type of NNM, single-criteria and multi-criteria selection of NNM parameters, adaptation of the learning method, effective coding of initial parameters, definition of admissible types of NNM, formation of an effective training sample, prediction of the sufficiency of computing resources.

A conceptual model of neural network recognition of phonemes in the VS of the DLS member has been developed, which ensures the formalization of the process of building the NNM of phoneme recognition in the VS of the DLS member.

The principles of the application of NM for recognizing phonemes in the VS of the DLS member were formed: the principle of determining the set of effective types of NNM for recognizing phonemes in the VS of the DLS member; the principle of admissibility of using the NNM type for phoneme recognition in the VS DLS; the principle of evaluating the effectiveness of the type of NNM intended for phoneme recognition in the VS of the DLS; the principle of determining the expected output signal of the NNM for phoneme standards; the principle of forecasting the use of NMS phoneme recognition computing resources of the DLS web server; the principle of evaluating the effectiveness of NNM phoneme recognition; the principle of

using expert knowledge to form an educational sample. It is expedient to use the specified principles as a basis for creating models of the processes of using the NNM of phoneme recognition in the VS of a member of DLS.

Thus, as a result of the conducted research, the conceptual principles of neural network recognition of phonemes in the VS of DLS members were developed.

References:

1. Ali A. (1998). An acoustic-phonetic feature-based system for automatic phoneme recognition in continuous speech. *Journal of the Acoustical Society of America*, vol. 103(5), pp. 2777–2778.

2. Ajinkya N., Nagaraj J., Dharwadkar V. (2018). A Speaker Recognition System Using Gaussian Mixture Model, EM Algorithm and K-Means Clustering. *International Journal of Modern Education and Computer Science*, vol. 10, no. 11, pp. 19–28.

3. Bykov V., Burov O., Gurzhii A., Zhaldak M., Leshchenko M., Lytvynova S., Lugovoi V., Oliynyk V., Spirin O., Shishkina M. (2019). *Theoretical and methodological principles of informatization of education and practical implementation of information and communication technologies in the educational sphere of Ukraine*. Kyiv: Comprint, 214 p. (in Ukrainian)

4. Chervyakov N., Astapov A. (2009). Using wavelets to improve the parameters of neural networks in speech recognition problems. *Infocommunication technologies*, vol. 4, pp. 9–12.

5. Chernyshev D., Mihaylenko V,, Tereikovska L. (2020). Development and research of tools for neural network analysis of the voice of listeners of the distance learning system. *Management of Development of Complex Systems*, vol. 43, pp. 123–130.

6. Du X.-P., He P.-L. (2006). The clustering solution of speech recognition models with SOM, *Proc. IEEE International Symposium on Neural Networks*, vol. 3972, pp. 150–157.

7. Dychka I.A., Tereikovskyi I.A., Didus A.V., Tereikovska L.O., Boyarinova Yu.Ye. (2023). Assessment of the efficiency of keyword spotting in voice signals. *Academic notes of TNU named after V.I. Vernadskyi. Series: technical sciences*, vol. 34 (73), no. 3, is. 1, pp. 123–129.

8. Fedyaev O. I., Bondarenko I. Yu. (2006). Analysis of the efficiency of the fuzzy pattern matching method for recognizing isolated words. *VI Mizhnarodna naukovo-praktichna konferenciya. Intellectual analysis of information IAI-2006*. May 16-19, 2006. Kyiv, pp. 112–117.

9. Fedyaev O., Bondarenko I. (2013). Neural network algorithm for speaker-independent recognition of speech phonemes, *USIM*, vol. 4, pp. 41–50.

10. Hu Z., Tereikovskyi I., Korystin O., Mihaylenko V., Tereikovska L. (2021). Two-Layer Perceptron for Voice Recognition of Speaker's Identity. *Advances in Intelligent Systems and Computing*. Springer, Cham, vol. 1247, pp. 508–517.

11. Ivanov A., Petrovsky A. (2006). First-order Markov property of the auditory spiking neuron model response. *14th European Signal Processing Conference*. Florence, Italy, pp. 1–5.

12. Kharlamov A., Raevsky V. (2004). Networks constructed of neuroid elements capable cf temporal summation of signals. *Neural Information Processing Research and Development*. Springer-Verlag, vol. 3, pp. 56–76.

13. Kotomin A. (2012). Recognition of voice command with the use of convolutional neuron networks. *Knowledge – intensive information technologies*, vol. 1, pp. 17–28.

14. Kushnir D. (2004). Automatic speech recognition system based on neural network technology. *Artificial intelligence: a scientific and theoretical journal*, vol. 3, pp. 654–659.

15. Medvedev M., Schukov S. (2016). The visualization system of patrol squad coordinates with a voice user interface. *IOP Conference Series: Materials Science and Engineering*, vol. 537, is. 4, pp. 101–104.

16. Misyurev A. (2003). Using an artificial neural network to assess the proximity of vectors of acoustic parameters. *Intelligent technologies for input and processing of information*, vol. 1, pp. 94–98.

17. Ovchinnikov P., Semin Y. (2006). Training a perceptron without segmenting words from the training set in the problem of sound recognition. *Neuroinformatics*, vol. 3, pp. 212–216.

18. Savchenko L. (2013). Automatic recognition of syllables based on a linear autoregressive neural network and the theory of fuzzy sets. *Neuroinformatics*, vol. 1, pp. 52–62.

19. Seilova N., Tereykovskaya L., Nadgi A. (2016). Conceptual model to ensure the efficiency of neural network recognition of phonrmes in distance learning. *Vestnik KazNRTU*, vol. 2 (114), pp. 345–351.

20. Tereikovskyi I., Tereikovska L. (2022). *Digital processing of signals and images: recognition of phonemes in the voice signal using neural networks*. Kyiv: KPI named after Igor Sikorsky, 120 p.

21. Tereikovskyi I., Bushuev D., Tereikovska L. (2022). *Artificial neural networks: basic principles*. Kyiv: KPI named after Igor Sikorsky, 123 p.

22. Tereikovska L. (2020). Architecture of a neural network analyzer of biometric parameters of listeners of the distance learning system. *Management of Development of Complex Systems*, vol. 44, pp. 91–99.

23. Tereikovska L. (2020). The method of neural network analysis of the voice signal. *Cyber security: education, science, technology*, vol. 3 (7), pp. 31–42.

24. Tereikovska L., Tereikovskyi I. (2020). Application of a convolutional neural network for the analysis of biometric parameters. *Academic notes of TNU named after V.I. Vernadskyi. Series: technical sciences*, vol. 31 (70), no. 5, pp. 124–128.

25. Titov Y., Kilgour K., Stüker S., Waibel A. (2011). Kit Quaero Speech-to-Text System. *Proc. 14 Int. Conf. "Speech and Computer"*, pp. 136–143.

26. Verenich I., Fedyaev O. (2007). Analysis of methods for constructing speech recognition systems based on neural network and hidden Markov

models. *III International Scientific and Technical Conference of Young Scientists and Students "Informatics and Computer Technologies"*, DonNTU, December 11-13, 2007, pp. 126–129.

27. Waibel A., Hanazawa T., Hinton G. (1989). Phoneme Recognition Using Time-Delay Neural Networks. *Transaction on acoustics, speech and signal processing*, vol. 37, pp. 328–339.

28. Yu H., Oh Y. (1995). A Neural Network using Non-Uniform Unit for Continuous Speech Recognition. *EU-Rospeech'95*, Madrid, Spain, September, vol. 3, pp. 1677–1680.