International Journal on
"Technical and Physical Problems of Engineering"
(IJTPE)
Published by International Organization of IOTPE

ISSN 2077-3528
IJTPE Journal
www.iotpe.com
ijtpe@iotpe.com

September 2023 Issue 56 Volume 15 Number 3 Pages 34-44

ACCURACY IMPROVING OF IDENTIFYING BIOMETRIC ARTIFACTS IN ANALYSIS OF DYNAMIC MULTIMODAL DATA FLOWS

G.S. Lebedev [1] E.Y. Linskaya [2] A.K. Lampezhev [3]

1. Department of Information and Internet Technologies, Moscow State Medical University, Moscow, Russia
gs.lebedev@inbox.ru
2. Science and Technology Park of Biomedicine, Moscow State Medical University, Moscow, Russia
linskaya.yelena@bk.ru
3. Institute of Design and Technology Informatics, Russian Academy of Sciences, Moscow, Russia, abas.lampezhev@mail.ru

Abstract- The object of this research is the identification algorithms for biometric artifacts that are used in telemedicine systems and other systems with dynamic multimodal data flows. This research aims to develop a procedure for improving the accuracy of systems for identifying biometric artifacts in terms of speech signals. Special attention is paid to the issue of various distorting factors, which cause a drop in the quality of the data of the analyzed biometric parameters. In particular, it is necessary to understand that when analyzing speech, identification systems exhibit a rather high sensitivity to various effects that are formed during processing and transmitting information, and they are also extremely sensitive to the physiological characteristics of the speaker and the acoustic properties of the environment. To date, deep learning algorithms, including convolutional neural networks, are one of the most effective approaches to automatic pattern recognition. An urgent scientific, technical and practical task in the development of digital systems for storing and processing multimodal data is the development of novel efficient neural network algorithms aimed at identifying biometric artifacts. The paper proposes a model for speaker identification, based on the preliminary processing of speech signals, which provides cleaning of phonograms from various defects. In the process of assessing neural networks, we use several indicators that can determine the percentage of correct responses regarding a particular category of information that should be clarified additionally. The proposed neural network algorithm improves the accuracy of voice activity fragment identification by about 2-3%. The existing modern methods and algorithms for the analysis of speech signals were analyzed, out and a procedure was proposed to improve the identification accuracy.

Keywords: Neural Network Algorithm; Biometric Parameters; Speech Signal; Neural Networks; Identification.

1. INTRODUCTION

Biometric personal identification systems, that is, personality recognition tools that are based on analyzing various unique behavioral and physiological properties of a person, are becoming an increasingly important element for a wide variety of everyday subject areas [1]. This thesis is confirmed by a fairly large number of research publications [2, 3]. The analysis of human biometric parameters acts as a fundamental factor for the operation of such methods of personal identification. For example, this can imply a digital image of a fingerprint [4], a face [5] or an eye retina [6]. At the same time, one should consider the fact that such a recording of a speech signal will act as a kind of digital impression of a person, forming his/her unambiguous characteristic.

Very often, such biometrics tools are used in videoconferencing sessions. Here much attention is paid to the task of proctoring. It should also be noted that various biometric methods aimed at building high-complexity access control and management systems are becoming more and more widespread. However, it should be understood that in all considered cases of using such technology, an imperfect quality of the speech signal and face image can be observed, which is caused by a rather significant number of distorting factors [7]. The thing is that such identification systems are extremely sensitive to the effects that form during processing and transmitting information, and to the acoustic properties of the environment.

It should also be noted that such personal identification systems based on the analysis of one biometric parameter are characterized by a low level of reliability, and therefore, further research is required on unimodal algorithms to improve them, and to develop personal identification systems based on two or more parameters [8]. It should be understood that an approach based on a combination of modalities will not only increase the stability and accuracy of biometric systems functioning, but also significantly improve their reliability in terms of the possibility of obtaining unauthorized access [9].

The use of deep learning algorithms acts as the most effective tool to provide automatic recognition of any patterns [10]. Currently, convolutional neural networks are one of the main tools for image analysis [11, 12]. They are most widely used when it is necessary to solve problems related to biometric identification, which is based on the analysis of a person's voice and face. The key feature of this approach is that during the implementation of the algorithm, all descriptors are automatically generated. The convolution operation will act as the main structural block in such a network. Owing to such descriptors, it is possible to achieve better results in terms of the tasks of detecting, segmenting and recognizing objects within digital images.

Even despite the popularity of voice biometrics, each system capable of successfully recognizing the speaker has some weaknesses, which include dependence on the microphone and data transmission channels, the physiological specifics of the speaker's voice, and the acoustics of the external background. The authentication algorithm may encounter a problem when users are registered in near-ideal conditions, whereas testing and operation of the device occurs in a noisy environment. The inability to control external factors and non-compliance with the rules for collecting biometric data can significantly reduce the accuracy of such a system.

As already noted in this research, the development of new neural network algorithms for identifying a person is an extremely urgent scientific and technical task, and also has direct practical interest from the perspective of the development of various digital systems for storing, processing and analyzing dynamic data flows. This research aims to develop a procedure for improving the accuracy of such personal identification systems that analyze speech signals. Achieving the research goal involves the implementation of the following list of tasks:

1) Developing an improved voice activity detector, which will surpass classical analogues in its characteristics;

2) Developing an algorithm for identifying a person in the presence of interference and noise of speech signals based on the existing artificial intelligence technologies.

2. LITERATURE REVIEW

In everyday life, a wide variety of biometric personal identification systems are increasingly encountered [13]. To begin with, it is necessary to note two main tasks for personality recognition, namely, identification and verification. Let us take a closer look at their differences. If we consider the essence of these processes from the viewpoint of user registration, for example, within some program, there will be no differences between these factors. Based on registration, a certain unified object will be formed within the database, which stores information about individual biometric parameters [14]. In this case, a certain set of parameters, such as a feature vector [15] or a digital model, will be an object. For example, when it is necessary to solve speaker recognition problems, Gaussian mixture models have gained the most popularity [16]. As soon as the of the user base is finally formed, the direct operation of the system begins.

Each potential user is verified through biometric requests that are matched against those existing within the database. Authenticity is determined using special credibility metrics. If the task of identification is considered, much attention is paid to the process of identifying a person based on a limited set of people registered in the system. Based on the results of the comparison, the optimal object is selected, which most closely matches the test object.

Facial images and a person's voice have quite large advantages over other biometric features. The absence of the need for physical contact with the recording devices is one of the main advantages here. If the procedure for recognizing a person's face is directly considered, its significant drawback lies in the extremely high dependence on the degree of illumination of the premises, and the angle of rotation of the head. An equally important role is given to the quality of the optical device. Another problem is that such algorithms are extremely sensitive to age-related changes, as a result, within the framework of such algorithms, a huge number of side factors must be considered, which will make it possible to ensure the operation of the system in practical conditions [17]. To date, the demand for such systems remains extremely high.

If we return to the consideration of systems related to voice analysis and speaker recognition, such techniques are becoming more widespread as speech technologies become popular. Particularly, there is now a rapid increase in the need for applications with which one can search for and recognize audio materials, and a variety of voice assistants. Noteworthy, such solutions will provide a qualitative leap for the entire industry of speech technologies. At the current stage, an extremely active development of various methods and algorithms is underway, which help deal with automatic speaker recognition [18]. Constant improvements to such systems have led to the fact that today they recognize voice information at almost the same level as a person. However, one should not forget about the disadvantages of such systems, which since they are extremely dependent on the effects of the information transmission channel, and the microphone used. Along with this, it is necessary to consider the physiological characteristics of the speaker and the acoustic properties of the environment. A decrease in the accuracy of the functioning of such systems is often observed in situations where there is no possibility to control external factors, and the prescribed rules for collecting biometric information are not followed.

One should not forget that the very method for recognizing a person's voice is almost identical to the methods that perform successful face identification. Figure 1 shows a generalized block diagram describing voice biometrics systems in more detail. It is noted that, if necessary, such a system includes a speaker separation block, where the technology implies the division of the input audio signal into a set of homogeneous segments that correspond to belonging to a particular speaker.

As practice shows, convolutional neural networks demonstrate high efficiency here [19]. The thing is that the CNN-based techniques and algorithms demonstrate extremely high results.

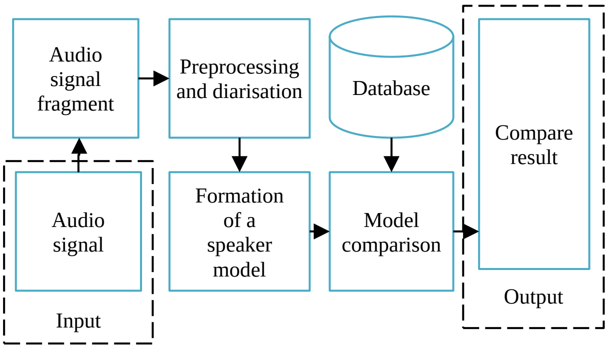


Figure 1. A generalized block diagram of speaker recognition system

They have become extremely widespread in text processing and analysis, medicine, robotics, and various biochemical studies. It is important to note that such algorithms inherently belong to the deep learning class, which is explained by the multi-connected hierarchical structure of such neural network systems, while their learning process takes an extremely long time. The very definition of "convolutional" indicates that building such networks requires to use the mathematical operation of convolution, which in its essence implies the execution of an operation based on two functions of a real argument.

The activation function acts as an essential element of the CNN CHC [20]. It is intended to determine whether the neuron will be activated as a result of the input signal. When activated, the signal will continue its own movement towards deeper layers. The kernel will determine the weight parameters of the convolutional layers. The learning process of the neural network will also imply the adaptation of weight parameters. Currently, the linear rectification block acts as one of the most common activation functions; however, today its various modifications already exist. This function has quite significant advantages over its counterparts. One of them is to quickly calculate the derivative function during training, by cutting off the negative part of the scalar value. The gradients in the system turn out to be not only large, but also consistent. At the same time, this activation function can wake up part of the neurons, which enables to make the layers sparse and reduce the computational load in the course of ongoing work. Thus, the performance increases significantly.

The subsampling operation acts as one of the most important elements of deep neural networks. The main advantage of pooling using generalization of selected features, and their compaction, provides an invariant representation of the input data. This process will obviously imply a partial loss of information, but at the same time there will be a drop in dimensionality. The dimensionality reduction operation with the choice of the maximum value is One of the special cases of discretization [21]. Such an operation will act as a non-linear feature compaction, where it is assumed that the maximum value will be selected from the property map area, the operation is repeated based on the entire feature map with a fixed step.

In addition, there are a fairly large number of varieties of pooling operations, among which one can note global and local averaging, and statistical pooling [22]. The latter approach implies the determination of the mathematical expectation, and the root-mean-square deviation of the entire feature map. From a theoretical perspective, there can be an unlimited number of convolutional layers, but as they grow, the requirements of the system for computing resources will increase. Figure 2 demonstrates in more detail the block diagram for the convolutional layer of the classical CNN.

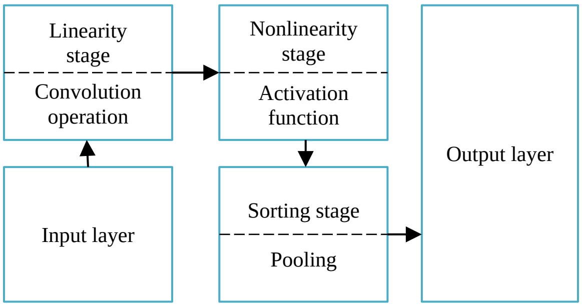


Figure 2. A structural diagram of the CNN convolutional layer

Fully connected layers are used at the output of existing classical CNNs, and several of them can exist simultaneously. The last of them contributes to the formation of an *N*-dimensional vector, which depends on the total number of objects under study. Note that at the output of the CNN, each number in the output set of parameters will act as a probability of a certain class. Their total number is equal to one. If we consider practical activities, the one with the most weight always wins. If we consider the learning process of the CNN, the layers of linear rectification and subsampling will be constant functions here. The error backpropagation method acts as a tool for adapting existing parameters. This approach should be considered as a modification to the classical gradient descent technique.

The residual network [23] is another common architecture. It should be noted that convolutional neural networks based on this architecture are capable of demonstrating extremely high results in the course of face recognition, and they are also considered promising from the viewpoint of the need to solve the problem associated with the biometric identification by voice. The main feature of this architecture is that it involves the use of a set of quick access links, which make it possible to significantly increase the effective depth of trainable networks. Such links are used within the limits of successive convolutions. The link is implemented by elementwise addition of the input and output of the difference block. Through the use of difference blocks, two problems can be successfully solved, the first of which is related to the damped gradient problem, where with the network layer deepening, the gradients decrease to update the weights, which does not allow training deep layers.

Noteworthy, a fairly large number of variations of the ResNet architecture differ from each other depending on the network depth. The theoretical increase in the accuracy of the network can be achieved by increasing the number of difference blocks, but in practice this is extremely rare. The difference in the accuracy of networks of 18 and 152 layers will differ very little, while the amounts of resources needed for training will grow rapidly. That is why a relatively simple architecture is used in this research.

Additionally, in terms of speaker recognition, one of the most effective approaches involves the use of neural networks, which are based on time delay blocks. The key idea of this concept is to analyze the frequency representation of the signal, while the analysis is carried out from the position of the time series. Context fragments are used to analyze the various hidden layers in the network. The analysis of a two-dimensional representation of a speech signal is another extremely common approach, although the approach itself can be interpreted as a variation of a digital image. If we consider the task in such a context, it will be reduced to the identification of visual images using a CNN based on two-dimensional convolution kernels. There is an extension that involves the analysis of a spectral feature map based on image analysis, for which two-dimensional convolution kernels are used.

It should also be noted that various x-vector systems have recently become more widespread, where TDNN networks act as basic structural blocks. The advantage of this architecture is that it allows achieving extremely high results in the context of the speaker recognition problem. Figure 3 shows in more detail the architecture of one of the most successful x-similar algorithms.

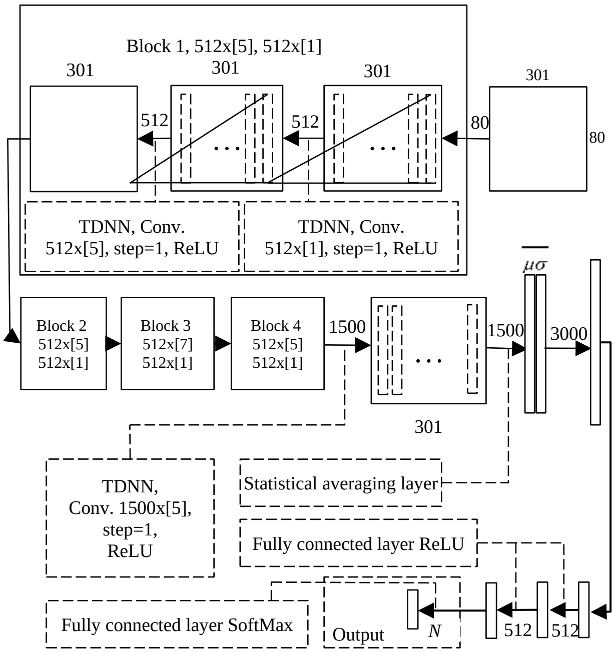

Block 1, 512x[5], 512x[1]
301
301
301
301
512
512
80
80
TDNN, Conv. 512x[5], step=1, ReLU
TDNN, Conv. 512x[1], step=1, ReLU
$\mu\sigma$
Block 2 512x[5] 512x[1]
Block 3 512x[7] 512x[1]
Block 4 512x[5] 512x[1]
1500
1500
3000
301
TDNN, Conv. 1500x[5], step=1, ReLU
Statistical averaging layer
Fully connected layer ReLU
Fully connected layer SoftMax
Output
N
512
512

Figure 3. X-vector architecture based on TDNN blocks

The creation of multimodal algorithms is one of the most significant directions in the development of biometric systems, since they will increase the reliability and stability of the personality recognition process. The key idea of this approach is to conduct a complex analysis for two or more biometric parameters. The crucial feature of the multimodal approach is the universality property, that is, one of several biometric parameters can be used to verify identity [24].

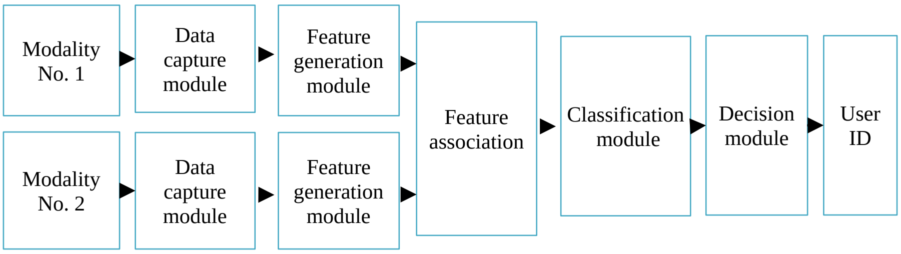


Figure 4. A scheme of combining modalities at the level of generated features

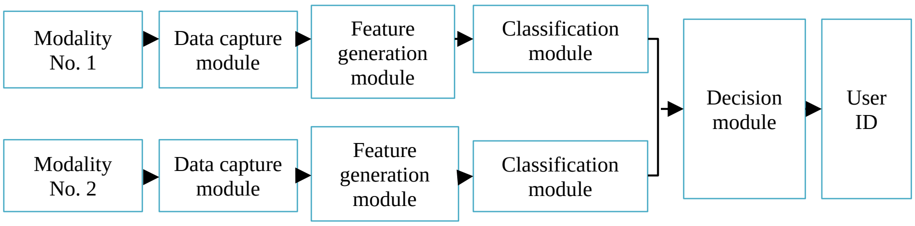


Figure 5. A scheme of combining modalities at the decision-making level

Figure 4 shows a generalized scheme, within which the main features of the multimodal approach to combining modalities at the feature level are displayed in more detail. A parallel analysis of various biometric data is conducted, and this analysis will contribute to the formation of specific features aimed at describing the modality. Then they are combined, which leads to the creation of a common set of features that can be used to form a classification and make final decisions.

Figure 5 demonstrates in more detail the logic of the approach aimed at combining modalities at the decision-making level. It is important to note that the essence of this approach is to conduct an independent analysis and classification of each specific modality.

Along with this, it should be understood that there are a fairly large number of other approaches to the combined analysis of biometric data. For example, work is underway aimed at creating multi-algorithmic systems, the idea of which is that one modality is analyzed, but different algorithms are used for this [25]. The specific operating conditions of the system will determine the specific mode of operation of the algorithms. For example, one of the algorithms can be used in daylight face analysis conditions, while another algorithm will work at night and in the evening. It is also admitted to build a system with the simultaneous operation of all algorithms, while the final decision is determined by a specific decision-making module.

It should be noted that one more approach also needs to be considered, within which the issue of developing multi-sensor biometric systems is considered. It is an alternative to the existing approach to the analysis of several modalities. The essence of this approach is that one specific biometric parameter is analyzed using two or more physical sensors, characterized by different effects. The multiple exemplar approach acts as another way to classify combinational biometric systems. Its essence is reduced to the simultaneous analysis of several biometric exemplars, which belong to the same biometric modality. When considering the problem of multimodal biometrics, the issue of using convolutional layers as feature extractors, which are used in conjunction with classical machine learning algorithms, is extremely relevant here. However, in most cases the best results are achieved through the use of solutions developed using neural network architectures.

It is important to understand here that the development of algorithms for biometric identification of a person based on neural network algorithms requires the use of an enormous data array. Moreover, the exact size and quality of such information will determine all the key characteristics of biometric systems, namely, accuracy, overall stability of operation, and the absence of strict technical requirements. In practice, two main types of systems are used to form such a data set, namely, text-dependent and text-independent ones [26]. Systems that have information about which phrase should be pronounced by the user during identification are considered text-independent. This approach implies that during identification, the user pronounces some key phrase, after which the main features that characterize the user's voice are extracted. The last stage of work will imply the formation of a new feature vector or a speaker model. When considering text-dependent systems, in most cases they use the concept of hidden Markov models, that is, random models that provide a statistical representation of sounds formed by a person. Intensity analysis is used to describe changes within a speech signal, with regard to its duration and pitch. Other techniques apply the Gaussian mixture model, which uses an approach based on the information about the voice to create a state vector that characterizes the physiological peculiarities of a particular person. Also note that the control process will imply the two-factor authentication.

Such text-independent systems do not imply the use of a priori information about the phrase that is pronounced by the user. This system would be much more flexible, as it would be able to recognize the speaker in situations where the person does not want to be identified. This is extremely effective in terms of combating telephone terrorism and mobile fraud. It should be noted that, firstly, proctoring systems are characterized by spontaneous speech of participants at the stage of passing control measures. Secondly, the use of text-independent algorithms in access control systems makes it possible to increase the reliability of identification, and expand the scope of their practical use.

A person in spontaneous speech cannot do without stops; therefore, pauses are always present in recorded voice sets. The presence of such stops in speech from the viewpoint of speaker recognition systems is a negative factor. This is related to the fact that the work of most neural network methods is based on the analysis of short speech fragments lasting 2-4 seconds. This is enough to recognize the speaker with high accuracy. Additionally, it is worth noting that a phonogram is an extremely complex representation of information, which directly depends on the quality characteristics of the recording devices, and the degree of transmission channel noise and the acoustic properties of the environment. As a result, any of the analyzed short-term fragments can completely or partially contain pauses without speech or only noise, which can significantly degrade the performance of the speaker recognition algorithms. To combat this, fragments containing only speech are selected from the phonogram at the stage of preprocessing. DGA algorithms are used for this procedure.

To date, a fairly large number of audiovisual voice data sets are freely distributed worldwide [27]. In most cases, they are focused on English speech. Summarizing the results of a detailed study of the available algorithms, techniques and a number of technologies applied in modern personal identification complexes, it can be concluded that a typical convolutional neural network is one of the key tools that contribute to the improvement of technologies for successful identification of people. An extremely high dependence of the overall level of operation accuracy on various distorting factors is one of the main problems in the development of this industry. Particularly, there is a strong dependence of such speaker identification systems on the effects that occur during data transmission, and because of the characteristics of the microphone and the speaker's physiological characteristics. The combination of these factors further confirms the relevance of the task associated with the development of combined methods of personal identification, since such approach will significantly increase the stability and accuracy of biometric systems, and provide a sufficiently high level of protection against unauthorized access.

3. MATERIALS AND METHODS

When considering the main features of the functioning of classical voice activity analysis algorithms, its basis is a fundamental understanding of the key features of the signal transmission process. Here, special attention is paid to energy, since it is a defining characteristic, which enables to most accurately describe the specific properties of the signal, and the dynamics of their changes in time and space. In addition, one should not forget that a certain type of energy can be calculated by the power integral, determined by the full cycle of the existence of the pulse under study, generally caused by speech, external noise component, and the advantages and disadvantages of recording devices.

Table 1. Classical voice activity detector algorithms

Classical voice activity detector algorithm (VAD)	The essence of the algorithm (the main stages of the operation of such a detector)
VAD1 algorithm based on the energy analysis [28]	The time domain signal is divided into windows 10-30ms long. Further, in each window, the square of the amplitude is calculated for all counts inside the window. A threshold filtering operation is performed. Thus, if the current energy value exceeds the set threshold value – θ, the processed section will remain unchanged. The specificity of this operation is determined by the fact that a number of the studied speech fragments are characterized by a significant energy potential, at the same time, each fragment containing noise and pauses will be expressed by a weak energy component. And the category of impulse noise is an exception in this case. Concurrently, algorithms based on the principles of energy analysis can be expressed by the equation: $S = \{S_1, S_2, S_3 \ldots S_j\}$, where $S_j = (s_1, s_2, s_3, \ldots, s_n)$, $E_j = \sum_{i=1}^{N} E(i) = \sum_{i=1}^{N} s^2(i)$, $V = \begin{cases} S_j, if\theta \pounds E_j \\ 0, if\theta > E_{j'} \end{cases}$ $S' = \{V_1 V_2 V_3, \ldots, V_W\}$, $\theta = k \times (E_{\max} - E_{\min})$ where, S is the initial speech impulse; S_j represents the j-th segments of the initial impulse; $s(i)$ is the value of the amplitude of the i-th recalls; E_j is the energy potential of the i-th recalls; E_j is the energy of the j-th sections of the initial impulse; Emax, Emin represent the largest and smallest values of the energies of the considered phonogram; N is the window length; V is a speech fragment according to the set threshold θ; k is an empirical coefficient enabling to adjust threshold θ; w is the number of windows containing speech; S' is the processed signal.
VAD2 algorithm based on the analysis of Teager-Kaiser energy	The algorithm is based on the fact that the signal is not divided into windows, and for each time count the energy is calculated as follows $E(i) = s^2(i) - s(i-1) \times s(i+1)$ where, $s(i)$ is the i-th count of the phonogram
VAD3 this algorithm performs a frequency study of phonograms [29]	At the first stage, the original phonograms are processed by a number of transformations proposed by Fourier, as a result of which spectrograms are constructed. Then the periodograms determined by the spectrograms are calculated. Next, the periodograms are processed by a multi-stage bandpass filter, since human speech is limited to a frequency spectrum ranging from 290 Hz to 3.45 kHz. At the next stage, it is necessary to determine the maximum range power for each fragment. Thus, if the current value exceeds the previously recorded threshold θ, the processed section is marked as a segment that carries information only about the speaker's voice

The available stops in the speech signal will be a negative factor for the signal, since this complicates the process of data identification. The solution to this problem is modern algorithms, which help select individual areas characterized by a high level of intensity. Also, such algorithms will highlight sections with a weak energy component. The description of the most efficient algorithms of this type is presented in Table 1.

Noteworthy, the algorithms presented in the VAD table are not demanding on computing resources; therefore, within the framework of this research, an attempt is made to create an algorithm that will be characterized by a higher level of complexity and efficiency. In addition, the analysis of the detector performance will be based on a poly-criteria approach implying that the assessment is based on metrics, which are presented in more detail in Table 2. The first metric involves the analysis of the percentage of correct answers. It is important to understand here that the number of speech fragments is much larger than that of the fragments containing pauses or noise; and therefore, such a set of audio signals will be unbalanced. This feature will be considered as part of the metric for assessing the proportion of correct answers. Along with this, it is possible use the harmonic mean between accuracy and recall, since these parameters will be rather effective in terms of evaluating the quality of work.

Table 2. Characteristics enabling to evaluate the performance of speech activity detectors

Indicator	The essence of the algorithm for evaluating the performance of the voice activity detector
Proportion of correct responses (*acc*)	To assess (*acc*) the proportion of correct responses the following formula is used: $acc = (\sum_{i=1}^{n} c(x_i)) / n$ where, x_i denotes the section of the phonogram corresponding to the number i, 10ms long; n is the number of all fragments 10 ms long in the set of VADSpeakersD under consideration; $c(x_i)$ is the indicator of the correctness of recognizing the i-th fragment of the speech signal. $c(x_i) = \begin{cases} 1, y_i = y_i' \\ 0, y_i{}^1 y_i' \end{cases}$ where, y_i is a target label of the fragment; y_i' is the label of the fragment that determines the result of the VAD operation.
Proportion of correct responses regarding data imbalance, (*accb*)	To assess (*accb*) the proportion of correct responses considering the fact that the number of speech fragments is significantly larger than that of the fragments containing pauses or noise, that is, for unbalanced data, the following formula is used: $accb = (acc_{y_i=1} + acc_{y_i=0}) / n$ where, $acc_{y_i=1}$ the proportion of correctly detected fragments representing the "speech" class; $acc_{y_i=0}$ is the proportion of correctly detected fragments representing the "noise/pause" class; n is the number of all fragments 10ms long in the set of VADSpeakersD under consideration.
Average harmonic in terms of accuracy and recall [30], (*F*-measure, *F*)	The average value is evaluated according to this expression: $F = x = \dfrac{2PR}{P+R}$ where P is accuracy, a metric that determines type I errors, R is recall, a metric that determines type II errors

Average harmonic in terms of accuracy and recall, regarding VADSpeakersDB data imbalance (F_{macro})	To assess the harmonic mean, with regard to the fact that the number of speech sections exceeds the number of sections that include noise, or pauses, that is, for unbalanced data, a macro averaging approach is used regarding the calculation of the metric within each class (y_i=1 is "speech", y_i=0 is "noise/pause") followed by normalization to the total number of classes: $F_{macro} = (F_{y_i=1} + F_{y_i=0}) / 2$

To improve the quality of identifying sections which contain human speech, within the framework of this research, a special voice activity detector is applied, which uses the data presented in Table 1. Along with this, the stacking method is used. Its essence is to use each VAD indicated in Table 1 as an independent detector. In this case, all conclusions are additionally combined to be analyzed using a generalizing classifier. A more detailed structural diagram of this solution is presented in Figure 6. And the results of the analysis of the effectiveness of using this approach are described in more detail in Table 3. Thus, when analyzing the information from the table, it can be concluded that the use of this algorithm increases the overall accuracy of determining speech fragments by about 2-3%.

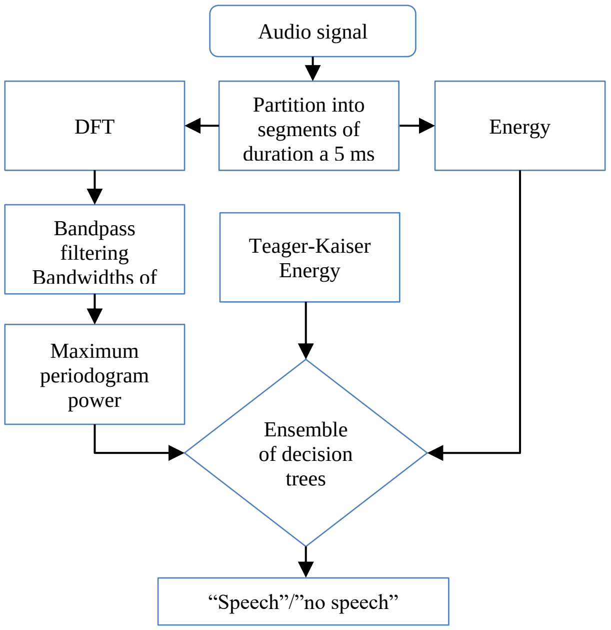

Audio signal
DFT
Partition into segments of duration a 5 ms
Energy
Bandpass filtering Bandwidths of
Teager-Kaiser Energy
Maximum periodogram power
Ensemble of decision trees
"Speech"/"no speech"

Figure 6. A structural diagram of the proposed KVAD algorithm

Further, it is proposed to dwell in more detail on the consideration of the principle of KVAD operation. In particular, the input receives fragments of a phonogram, the length of which is about 10ms. After that, each detector is engaged in the analysis of the input sample of the analyzed phonogram, on the basis of which three independent predictions are formed. As a result, the generalizing classifier will form the final decision to understand where it is required to attribute some input fragment of the phonogram. The analog model is a classifier using the regression tree method.

Table 3. Comparative analysis of voice activity detectors

Indicators	*acc*	*accb*	*F*	F_{macro}
VAD_1	0.90	0.88	0.93	0.88
VAD_2	0.89	0.88	0.92	0.88
VAD_3	0.89	0.87	0.92	0.87
KVAD	0.91	0.90	0.94	0.90

When considering the structure of the developed procedure for personal identification, the proposed combined voice activity detector will be considered as a special tool for processing a large databases of speech signals. The main goal of this process is to form sets intended for training and testing algorithms for biometric personal identification, for which the principles of voice biometrics will be used. Currently, the main problem is the absence of any publicly available sets and tasks that can consider the Russian speech, and function in conditions of significant variability of acoustic properties. In this regard, two text-independent sets of audio data were formed, which will be used in the personal identification procedure to be developed. This is a special set that has been prepared through a wide range of recording devices, and in the realities of excessive changes in the acoustic properties of the environment.

This data set is used to create a special testing and training sample, for which k-block cross-validation is used. Using this parameter, the number of equal parts of the training set will be determined, after which they will be used to train the entire model. Training is repeated as many as there are parts, which enables to get an average estimated characteristic. Figure 7 demonstrates in more detail the scheme of the developed approach related to the process of training and testing the derived KVAD algorithm. The adjustment is made depending on the total number of trees, the specific splitting criterion, and the maximum depth of the trees. Meanwhile, one should also be guided by the total number of evaluated characteristics that reflect the functioning of each node involved in decision-making. However, this technique demonstrates increased sensitivity, due to the fact that total number of fully trained sorters can be expressed in several thousands.

It is important to remember that speech is essentially an acoustic signal, which is described using a function of time. Successful solution of the task associated with speaker recognition will imply the analysis of the frequency properties of the speech signal, which will characterize the individual characteristics of the voice. That is, the solution of the problem of personal identification, based on speech signals, implies the determination of the unique acoustic properties and features of the speaker's voice. Here, special attention is paid to special frequency representations of speech signals in the form of spectrograms. This is one of the most effective and popular types of speech information presentation. The essence of this approach is that it can be used when solving the problem of identifying the speaker and the speech recognition system. The process itself will imply that when analyzing the frequency representation, the main features of the human auditory system will be analyzed. The MEL scale will link the pitch of the sound to its actually measured frequency.

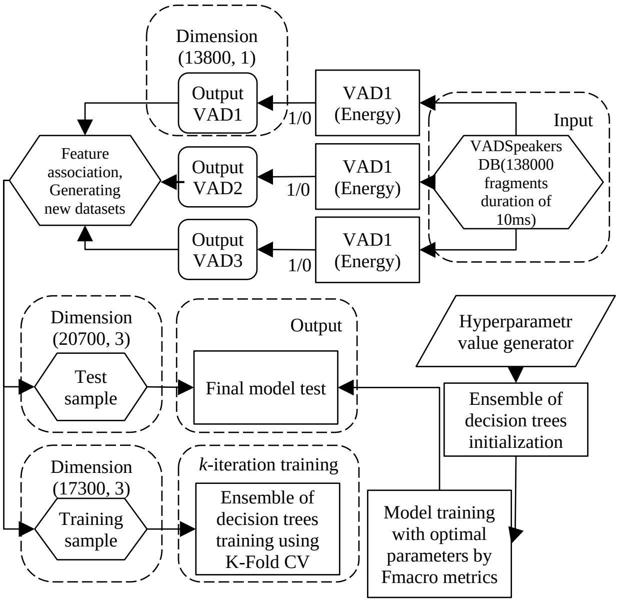

Dimension (13800, 1)
Output VAD1
1/0
VAD1 (Energy)
Input
VADSpeakers DB(138000 fragments duration of 10ms)
Feature association, Generating new datasets
Output VAD2
1/0
VAD1 (Energy)
Output VAD3
1/0
VAD1 (Energy)
Dimension (20700, 3)
Test sample
Output
Final model test
Hyperparametr value generator
Ensemble of decision trees initialization
Dimension (17300, 3)
Training sample
k-iteration training
Ensemble of decision trees training using K-Fold CV
Model training with optimal parameters by Fmacro metrics

Figure 7. A scheme of the KVAD training and testing process

An important advantage of the MEL scale is that it can be used help form a complete set of triangular filters that will be used in relation to the periodogram by rolling it along the axis. This process is presented in more detail in Figure 8.

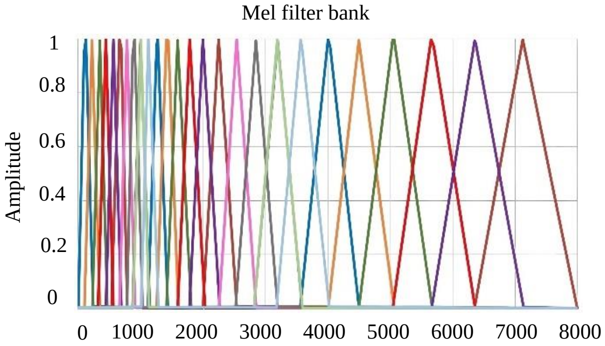


Figure 8. The bank of triangular MEL filters

Using these filters, it is possible to accumulate spectral energy in the region of central frequencies. The size of the triangular filter bank will determine the total number of frequencies. The frequency range will be the determining factor for the bandwidth of each filter. It should be understood that the filters will have a narrow band around the zero frequency. This is also one of the features of the human auditory system. Figure 9 demonstrates the block diagram of this procedure for calculating the MFCC in more detail.

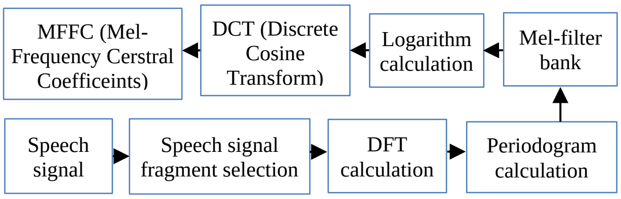

MFFC (Mel-Frequency Cerstral Coefficeints)
DCT (Discrete Cosine Transform)
Logarithm calculation
Mel-filter bank
Speech signal
Speech signal fragment selection
DFT calculation
Periodogram calculation

Figure 9. A scheme for calculating the MFCCs

In general, the main stages of the MFCC formation include the direct selection of a fragment of a speech signal having a fixed length, the determination of a spectrogram using the discrete Fourier transform, the application of a MEL filter bank to the periodogram and the logarithm calculation.

Personal identification using speech signals will be carried out using the X-speech neural network algorithm, which will be the final stage of the proposed speaker identification procedure. Such a solution will be determined by a number of reasons, namely: the use of such algorithms will provide an extremely high level of accuracy during operation. The second feature is that the topology of this network will be based on the one-dimensional convolution operation, which is characterized by low computational resources. Moreover, the most important advantage of this approach is that it involves the analysis of temporal fragments of the speech signal over the entire frequency band.

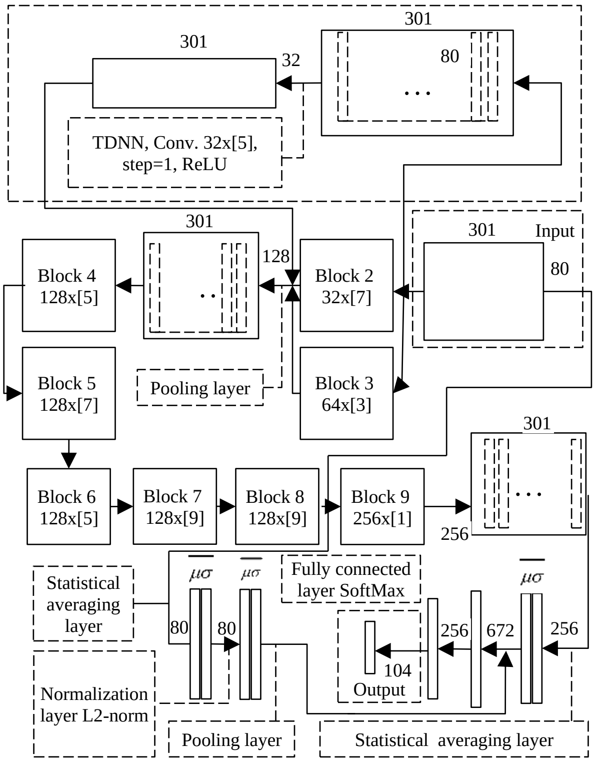


Figure 10. The X-Speech neural network architecture

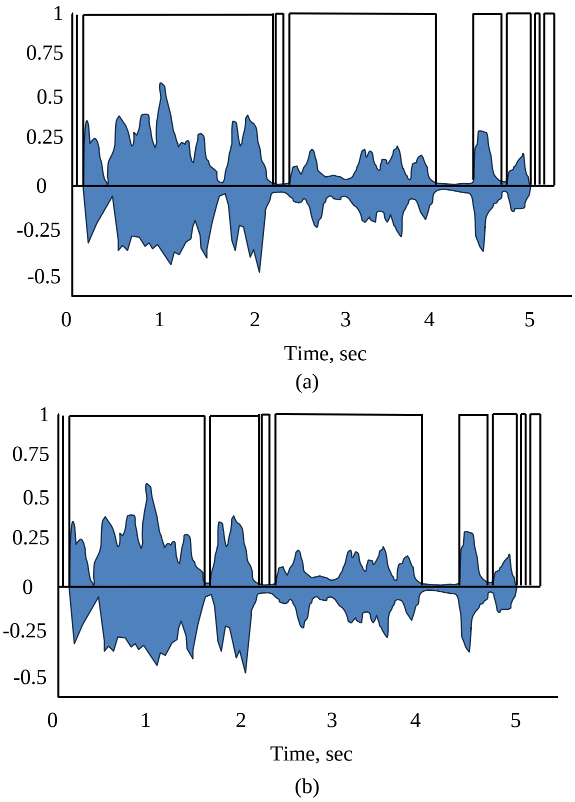
Figure 10 shows in more detail the block diagram of the developed convolutional neural network.

At the input, the network has 3 blocks of neural networks with a time delay. The first and second blocks each have 32 filters with frame sizes of 5 and 7, respectively. One frame of the spectrogram is the frequency decomposition of the speech signal within the width of the window, which is used 70 when analyzing the signal in the time domain. In this research, the window width is 512 counts, which corresponds to the time section of the initial signal with a duration of 32ms. The third block consists of 64 filters. The outputs from these blocks are combined to form a 128×301 map.

This is followed by 5 blocks in sequence, where the number of filters remains unchanged and equals 128. Next comes "Block 9", which increases the width of the feature map by using 256 filters.

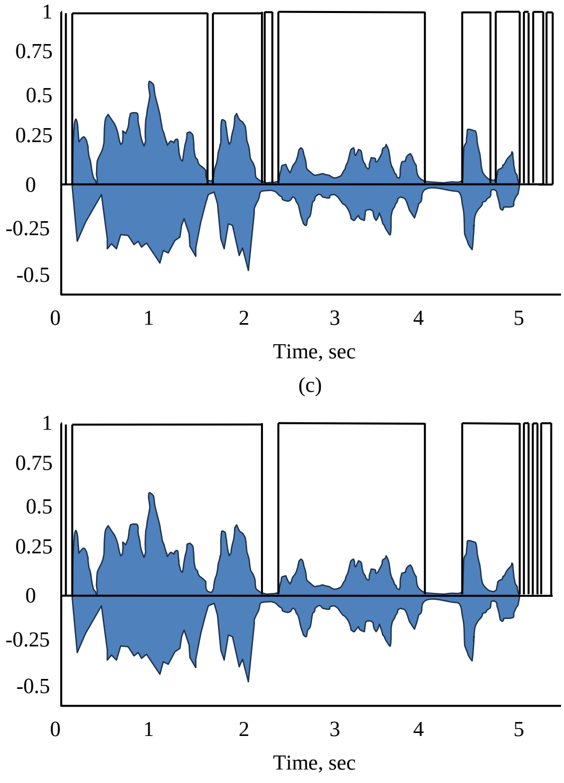
The next level uses a statistical averaging layer that calculates the mean and mean-root-square deviation for each feature. The calculation is carried out along the time axis. As a result, two 256-dimensional vectors are formed – $\overline{\mu_1}$ and $\overline{\sigma_1}$. Further, these vectors are combined with the vectors $\overline{\mu_2}$ and $\overline{\sigma_2}$. The latter are formed as a result of using one more layer of statistical averaging, to the input of which the MFCC map is directly fed. As a result of combining $\overline{\mu_1}$, $\overline{\sigma_1}$, $\overline{\mu_2}$ and $\overline{\sigma_2}$, a common feature vector having length of 672 is formed. After that, the generalized vector is sent to the fully connected layer, and then the layer with the "SoftMax" activation option is activated, in which the number of outgoing neurons will correspond to the population covering each class being evaluated (104).

All convolutional layers use ReLU as an activation function. The total number of layers is 18, of which 9 are convolutional. Thus, it can be concluded that this approach will have extremely low requirements for computing resources, and also demonstrates extremely high accuracy in identification and overall noise immunity. Figure 11 shows in more detail the results of the operation of such detectors, the visual assessment of which confirms the very fact of improvement in the identification of voice activity zones. As can be seen from the oscillograms, the proposed KVAD algorithm makes it possible to reduce the number of voice activity zones, which helps increase the speech recognition accuracy.



(a)

(b)



(c)

(d)

Figure 11. The operating results of the voice activity detectors: (a) based on energy analysis; (b) based on the Teager-Kaiser energy analysis; (c) based on the signal frequency analysis; (d) the proposed KVAD algorithm

4. CONCLUSIONS

Based on the research results, a system was developed that can successfully identify the speaker. The system uses the capabilities of combined detectors that capture voice activity. The model, that is proposed in this paper, is based on the preliminary processing of speech signals. The main advantages of neural network algorithm, of the model are:

- Proposed algorithm significantly improves the quality of the speech impulse owing to the cascade filtering of each phonogram, removing pauses, a number of inappropriate hitches and limiting the quantity of the voice activity zones. In addition, this algorithm improves the accuracy of identifying each fragment where voice activity is recorded (approximately by 1-3%). Ultimately, the creation of a neural network algorithm can be used for the automated identification of the speaker, relying upon a detailed analysis of each speech impulse.
- Developed algorithms require a relatively small amount of computing resources, hence, they can be used to create real-time biometric systems.

ACKNOWLEDGEMENTS

The presented results were obtained under the Subsidy Agreement No. 075-11-2021-050 of June 24, 2021, "Development of a multi-format information system for automated distributed monitoring and on-line quality control of information and medical services using artificial intelligence technologies" with the Ministry of Science and Higher Education of Russia.

REFERENCES

[1] S. Dargan, M. Kumar, "A Comprehensive Survey on the Biometric Recognition Systems Based on Physiological and Behavioral Modalities", Expert Systems with Applications, Vol. 143, pp. 113-114, 2020.

[2] J. Bergstra, Y. Bengio, "Random Search for Hyper-Parameter Optimization", Journal of Machine Learning Research, Vol. 13, No. 2, pp. 281-305, 2012.

[3] M.I. Jordan, T.M. Mitchell, "Machine Learning: Trends, Perspectives, and Prospects", Science, Vol. 349, No. 6245, pp. 255-260, 2015.

[4] G. Guven, U. Guz, H. Gurkan, "A Novel Biometric Identification System Based on Fingertip Electrocardiogram and Speech Signals", Digital Signal Processing, Vol. 121, p. 103306, 2022.

[5] C. Bisogni, G. Iovane, R.E. Landi, M. Nappi, "ECB2: A Novel Encryption Scheme Using Face Biometrics for Signing Blockchain Transactions", Journal of Information Security and Applications, Vol. 59, p. 102814, 2021.

[6] F. Sadikoglu, S. Uzelaltinbulat, "Biometric Retina Identification Based on Neural Network", Procedia Computer Science, Vol. 102, pp. 26-33, 2016.

[7] N. Alghamdi, S. Maddock, J. Barker, G.J. Brown, "The Impact of Automatic Exaggeration of the Visual Articulatory Features of a Talker on the Intelligibility of Spectrally Distorted Speech", Speech Communication, Vol. 95, pp. 127-136, 2017.

[8] C. Kose, C. Ikibas, "A Personal Identification System Using Retinal Vasculature in Retinal Fundus Images", Expert Systems with Applications, Vol. 38, No. 11, pp. 13670-13681, 2011.

[9] E.A. Alkeem, et al., "Robust Deep Identification using ECG and Multimodal Biometrics for Industrial Internet of Things", Ad Hoc Networks, Vol. 121, p. 102581, 2021.

[10] M. Alghamdi, P. Angelov, L.P. Alvaro, "Person Identification from Fingernails and Knuckles Images Using Deep Learning Features and the Bray-Curtis Similarity Measure", Neurocomputing, Vol. 513, pp. 83-93, 2022.

[11] W. Hu, et al., "A State-of-the-Art Survey of Artificial Neural Networks for Whole-Slide Image Analysis: From Popular Convolutional Neural Networks to Potential Visual Transformers", Computers in Biology and Medicine, Vol. 161, p. 107034, 2023.

[12] A. Dhandayuthapani, J. Lawrence, "Plant Disease Recognition Using Optimized Image Segmentation Technique", International Journal on Technical and Physical Problems of Engineering (IJTPE), Issue 50, Vol. 14, No. 1, pp. 211-218, March 2022.

[13] N. Singla, M. Kaur, S. Sofat, "Automated Latent Fingerprint Identification System: A Review", Forensic Science International, Vol. 309, p. 110187, 2020.

[14] A Olejnik, A. Kapitanov, I. Alexandrov, A. Tatarkanov, "Designing a Tool for Cold Knurling of Fins", Journal of Applied Engineering Science, Vol. 18, No. 3, pp. 305-312, 2020.

[15] X. Wang, W. Cai, M. Wang, "A Novel Approach for Biometric Recognition Based on ECG Feature Vectors", Biomedical Signal Processing and Control, Vol. 86, p. 104922, 2023.

[16] A.R. Avila, D. O'Shaughnessy, T.H. Falk, "Automatic Speaker Verification from Affective Speech Using Gaussian Mixture Model Based Estimation of Neutral Speech Characteristics", Speech Communication, Vol. 132, pp. 21-31, 2021.

[17] L. Boussaad, A. Boucetta, "Deep-Learning Based Descriptors in Application to Aging Problem in Face Recognition", Journal of King Saud University, Computer and Information Sciences, Vol. 34, No. 6, pp. 2975-2981, 2022.

[18] K.J. Devi, N.H. Singh, K. Thongam, "Automatic Speaker Recognition from Speech Signals Using Self Organizing Feature Map and Hybrid Neural Network", Microprocessors and Microsystems, vol. 79, p. 103264, 2020.

[19] Z.T. Liu, M.T. Han, B.H. Wu, A. Rehman, "Speech Emotion Recognition Based on Convolutional Neural Network with Attention-Based Bidirectional Long Short-Term Memory Network and Multi-Task Learning", Applied Acoustics, Vol. 202, p. 109178, 2023.

[20] S. Kilicarslan, C. Kozkurt, S. Bas, A. Elen, "Detection and Classification of Pneumonia Using Novel Superior Exponential (SupEx) Activation Function in Convolutional Neural Networks", Expert Systems with Applications, Vol. 217, p. 119503, 2023.

[21] M. Garcia Torres, R. Ruiz, F. Divina, "Evolutionary Feature Selection on High Dimensional Data Using a Search Space Reduction Approach", Engineering Applications of Artificial Intelligence, Vol. 117, p. 105556, 2023.

[22] G. Lin, S. Zhao, J. Shen, "Video Person Re-Identification with Global Statistic Pooling and Self-Attention Distillation", Neurocomputing, Vol. 453, pp. 777-789, 2021.

[23] V. Jain, Ap. Jain, V. Garg, Ac. Jain, M. Demirci, M.C. Taplamacioglu, "Siamese Neural Networks for Pandemic Detection Using Chest Radiographs", International Journal on Technical and Physical Problems of Engineering (IJTPE), Issue 51, Vol. 14, No. 2, pp. 104-110, June 2022.

[24] G. Fenu, M. Marras, "Demographic Fairness in Multimodal Biometrics: A Comparative Analysis on Audio-Visual Speaker Recognition Systems", Procedia Computer Science, Vol. 198, pp. 249-254, 2022.

[25] F. Alonso Fernandez, et al., "Cross Sensor Periocular Biometrics in a Global Pandemic: Comparative Benchmark and Novel Multialgorithmic Approach", Information Fusion, Vol. 83-84, pp. 110-130, 2022.

[26] S. Farsiani, H. Izadkhah, S. Lotfi, "An Optimum End-to-End Text-Independent Speaker Identification System Using Convolutional Neural Network", Computers and Electrical Engineering, Vol. 100, p. 107882, 2022.

[27] J. Liu, et al., "Audio-Video Database from Subacute Stroke Patients for Dysarthric Speech Intelligence Assessment and Preliminary Analysis", Biomedical Signal Processing and Control, Vol. 79, p. 104161, 2023.

[28] Yu.N. Matveev, “Technologies of Biometric Identification of a Person by Voice and Other Modalities”, Engineering Journal: Science and Innovation, No. 3, 2012.

[29] M.H. Moattar, M.M. Homayounpour, “A Simple but Efficient Real-Time Voice Activity Detection Algorithm”, The 17th European Signal Processing Conference, pp. 2549-2553, Glasgow, UK, 24-28 August 2009.

[30] D. Powers, “Evaluation: from Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation”, International Journal of Machine Learning Technology, Vol. 2, No. 1, pp. 37-63, 2011.

BIOGRAPHIES

Name: **Georgy**
Middle Name: **Stanislavovich**
Surname: **Lebedev**
Birthday: 19.12.1964
Birth Place: Lviv, USSR
Master: Mathematician in "Software for the Functioning of Automated Control Systems", Department of Information and Analytical Work, Faculty of Special Information Technologies, Military Engineering Institute of the Red Banner (Military Space Academy), St. Petersburg, Russia, 1987
The Last Scientific Position: Head of Department of Information and Internet Technologies, Director of Institute of Digital Medicine, Moscow State Medical University, Ministry of Health of Russia (Sechenov University), Moscow, Russia, since 2016
Research Interests: Neural Networks, IT Technologies and E-Health, Standardization of Software and Information Technologies
Scientific Publications: 142 Papers, 9 Books, 6 Patents, 6 Projects, 28 Theses

Name: **Elena**
Middle Name: **Yurievna**
Surname: **Linskaya**
Birthday: 07.07.1988
Birth Place: Moscow, Russia
Master: Mathematical Engineering in Applied Mathematics, Department of Applied and Computational Mathematics, Faculty of Informatics, Moscow State University of Instrument Engineering and Informatics, Moscow, Russia, 2020
The Last Scientific Position: Leading Engineer, Science and Technology Park of Biomedicine, Moscow State Medical University, Ministry of Health of Russia (Sechenov University), Moscow, Russia, Since 2021
Research Interests: Automated Control Systems, Informatics and Information Systems
Scientific Publications: 14 Papers, 6 Patents, 4 Theses

Name: **Abas**
Middle Name: **Khasanovich**
Surname: **Lampezhev**
Birthday: 07.07.1997
Birth Place: Murmansk, Russia
Bachelor: Navigation and Ballistic Support for the Use of Space Technology, Department of Dynamics and Flight Control of Rockets and Spacecraft, Faculty of Special Mechanical Engineering, Bauman Moscow State Technical University, Moscow, Russia, 2016
Master: Navigation and Ballistic Support for the Use of Space Technology, Department of Dynamics and Flight Control of Rockets and Spacecraft, Faculty of Special Mechanical Engineering, Bauman Moscow State Technical University, Moscow, Russia, 2020
The Last Scientific Position: Researcher, Institute of Design and Technology Informatics of RAS, Moscow, Russia, Since 2020
Research Interests: Multi-Parameter Optimization and Mathematical Modeling, Multi-Parameter Optimization and Mathematical Modeling
Scientific Publications: 12 Papers, 3 Theses