UDC 004.93, 004.032.26
DOI: 10.15587/1729-4061.2024.298687

THE DEPENDENCE OF THE EFFECTIVENESS OF NEURAL NETWORKS FOR RECOGNIZING HUMAN VOICE ON LANGUAGE

Aigul Nurlankyzy
PhD Student*
Almaty University of Power Engineering and Telecommunications
Baytursynuli str., 126/1, Almaty, Republic of Kazakhstan, 050013

Ainur Akhmediyarova
Corresponding author
PhD*
E-mail: a.akhmediyarova@satbayev.university

Ainur Zhetpisbayeva
PhD***

Timur Namazbayev
Master, Senior Lecturer
Department of Solid State Physics and Nonlinear Physics
Al-Farabi Kazakh National University
Al-Farabi ave., 71, Almaty, Republic of Kazakhstan, 050040

Asset Yskak
Master***

Nurdaulet Yerzhan
Student
Department of Cybersecurity,
Information Processing and Storage**

Bekbolat Medetov
PhD***

*Department of Software Engineering**
**Institute of Automation and Information Technology
Satbayev University
Satpaev str., 22a, Almaty, Republic of Kazakhstan, 050013
***Department of Radio Engineering,
Electronics and Telecommunications
S. Seifullin Kazakh Agro Technical Research University
Zhenis ave., 62, Astana, Republic of Kazakhstan, 010011

This study examines the effectiveness of neural network architectures (multilayer perceptron MLP, convolutional neural network CNN, recurrent neural network RNN) for human voice recognition, with an emphasis on the Kazakh language. Problems related to language, the difference between speakers, and the influence of network architecture on recognition accuracy are considered. The methodology includes extensive training and testing, studying the accuracy of recognition in different languages, and different sets of data on speakers. Using a comparative analysis, this study evaluates the performance of three architectures trained exclusively in the Kazakh language. The testing included statements in Kazakhs and other languages, while the number of speakers varied to assess its impact on recognition accuracy.

During the study, the results showed that CNN neural networks are more effective in recognizing human voice than RNN and MLP. Also, it was found that the CNN has a higher accuracy in recognizing the human voice in the Kazakh language, both for a small and for a large number of announcers. For example, for 20 speakers, the recognition error in Russian was 21.86 %, whereas in Kazakhs it was 10.6 %. A similar trend was observed for 80 speakers: 16.2 % Russians and 8.3 % Kazakhs. It can also be argued that learning one language does not guarantee high recognition accuracy in other languages. Therefore, the accuracy of human voice recognition by neural networks depends significantly on the language in which training is conducted.

In addition, this study highlights the importance of different sets of speaker data to achieve optimal results. This knowledge is crucial for advancing the development of reliable human voice recognition systems that can accurately identify different human voices in different language contexts

Keywords: Artificial intelligence, neural networks, CNN, RNN, MLP, voice activity detector, human voice recognition, the effectiveness of training, language specifics, recognition accuracy

Received date 08.12.2023
Accepted date 15.02.2024
Published date 28.02.2024

How to Cite: *Nurlankyzy, A., Akhmediyarova, A., Zhetpisbayeva, A., Namazbayev, T., Yskak, A., Yerzhan, N., Medetov, B. (2024). The dependence of the effectiveness of neural networks for recognizing human voice on language. Eastern-European Journal of Enterprise Technologies, 1 (9 (127)), 72–81. doi: https://doi.org/10.15587/1729-4061.2024.298687*

1. Introduction

Human voice recognition is of great practical importance in various economic and technological sectors. Voice Activity Detectors (VADs) are commonly employed in telecommunication systems to minimize the amount of voice data transmitted. These devices are responsible for identifying and removing non-speech elements from the audio stream. Additionally, in speech recognition, speaker identification, and other speech technologies, the task of isolating speech fragments from a continuous stream of audio data is primarily solved. It is noteworthy that in all these areas, the task of recognizing human voice is solved in different ways. For example, in traditional VAD systems, the energy and entropy of the signal are analyzed; that is, an algorithmic approach is still used. However, in reality, it is almost impossible to accurately describe the parametric features of human voice algorithmically. In modern conditions of speech technologies, various artificial neural networks (ANNs) are used to recognize human voices. It should be noted that ANN-based methods based on ANNs provide impressive results.

The use of artificial neural networks in the field of speech processing is an important area of development; however, most existing models are trained and tested on data from only one language, which limits their applicability to other languages. To overcome this problem, it is necessary to develop methods for language-independent recognition of the human voice that are capable of working with various languages and dialects. The results of these studies on the development of language-independent human voice recognition methods are relevant and can be useful in improving modern human speech recognition technologies in the field of information technology.

Therefore, research on the dependence of the effectiveness of neural networks for recognizing human voices on language is extremely important for understanding the influence of different linguistic features on speech processing. The results of such studies can be applied in practice to improve multilingual speech recognition systems and may also be useful for creating more effective training programs and technologies for teaching foreign languages. It can also lead to improved communication in multilingual environments and possibly to increased access to technology for speakers of different languages.

2. Literature review and problem statement

In [1], the results of an analysis of the accuracy and performance of VAD using a multilayer perceptron (MLP), recurrent neural network (RNNs), and convolutional neural network (CNNs) were presented, which showed that for most of the models under consideration, the best performance was achieved using CNNs. To conduct training and model testing, a TIMIT dataset of various dialects of English was used, containing 6,300 utterances uttered by 192 women and 438 men. The recognition accuracy was 97.60 %. However, it was noted that it is also necessary to use different datasets, that is, speech corpora of different languages, to evaluate the performance of VAD.

VAD models based on deep learning using the TIMIT English dataset were also proposed in another study [2]. This study highlights the importance of using different spectral and temporal characteristics to increase VAD performance. Nevertheless, it is also noted that further research is needed to overcome complex problems, such as evaluating the performance of VAD using datasets of different languages, as well as analyzing the number of speakers needed to obtain a highly efficient VAD system.

In [3], emphasis was placed on the detection of voice activity based on the self-attention attention mechanism, the task of which is to identify patterns only between input data. A single set of pure English speech data, LibriSpeech, was used as input data. To test the effectiveness of the proposed method, 6,500 training data points and 6,500 test data points were used.

In [4], the authors achieved a high VAD performance based on deep learning methods at any decision threshold. The English speech corpus LibriSpeech was used as the training data. However, this study did not consider tasks for evaluating VAD performance using datasets from other languages.

In [5], a comparative analysis of the effectiveness of the end-to-end neural network Marble Net and basic CNN was considered. The results show the high performance of the Marble Net network compared to the base model, which makes VAD more accessible under conditions of limited computing resources, such as mobile and wearable devices. However, the training and testing of the model were also carried out using only a single set of AVA-speech English language data.

In [6], a VAD learning strategy using Supervised Contrastive Learning (Supervised Contrastive Learning for Voice Activity Detection, SCLVAD) was proposed for the first time. The proposed method was used in combination with audio–specific data augmentation methods, which were trained using two common sets of English language speech data: the Google Speech Commands Dataset V2 and audio samples from the site freesound.org, and then evaluated using the third AVA-Speech English dataset. However, in this study, no studies have evaluated the performance of VAD based on sets of different languages.

In [7], a study on the effectiveness of VAD based on 8064 audio signals from the TIMIT English speech corpus was conducted. The efficiency of this VAD system averaged 91.02 %. Nevertheless, it is important to conduct reeə search based on the speech corpora of other languages.

In [8], a VAD system based on a multitasking learning U-shaped neural network (MTU-Net) was considered. The proposed VAD method demonstrated a high performance compared to existing models. The training and testing of the model were also carried out based on a single speech corpus of the English language TIMIT. Training and testing of a multilayer neural network (MTS-NN) based on the TIMIT speech corpus for detecting voice activity was performed in [9]. The proposed algorithm was trained and tested using data from the TIMIT English Speech Corpus (16 men and 8 women from eight different dialects). Nevertheless, there is a need for research using datasets of different languages, as well as an analysis of the number of speakers required to obtain a highly efficient VAD system.

A method for detecting voice activity based on a deep neural network is proposed in [10]. Training and testing of the network were performed using the Korean speech corpus. Thus, in this study, all analyses were conducted in only one language.

The paper [11] presents an application for smartphones using a convolutional neural network to determine voice activity in real time with low audio latency. The architecture of the convolutional neural network has been optimized in such a way as to ensure the processing of audio frames in real time without skipping any frames while maintaining high accuracy in determining voice activity. Speech files from the PN/NC corpus version 1.0 were used to teach and evaluate the developed CNN VAD. This corpus consists of 20 native speakers (10 men, 10 women) from two dialect regions of American English (Pacific Northwest and Northern Cities) pronouncing 180 sentences of the IEEE "Harvard" set. In total, it consists of 3,600 audio files. However, this study does not consider the tasks for evaluating VAD performance using datasets from other languages, and there are no studies devoted to analyzing the number of speakers required to obtain a highly efficient VAD system.

In [12], the VAD method, which uses statistical functions based on the linear spectrum and frequency, was presented. The experiments were conducted on a database of more than 350 h, consisting of data from various sources, such as YouTube. The accuracy of the system was 99.43 %. However, there were also no studies related to the study of the number of required speakers.

In [13], a fast and efficient uncontrolled VAD method using fractal dimension estimation was proposed. Two databases in different languages are used to evaluate the effectiveness of the proposed method. The first is the database of the Massachusetts Institute of Technology Texas Instruments (TIMIT) and the second is the database of Arabic speech at King Saud University (KSU). The language of the TIMIT speech database is English, representing 630 male and female speakers from eight dialect regions. The language of the KSU speech database is Arabic, consisting of 328 speaking men and women who recorded both pre-written and spontaneous texts. However, the training and testing of the VAD system in this study were performed separately for each dataset.

In [14], a VAD speech enhancement function based on a variational autoencoder was proposed. In this study, let's use pure utterances from the Aurora 4 database [14], which contained 7,138 continuous speech utterances for training and 330 utterances for testing. To build a 35-hour training set, all 7,138 statements from a pure training set were used. The statements in the Aurora4 corpus are short, and approximately 80 % of them are speech.

In [15], the voice activity detection (VAD) function was combined with end-to-end automatic speech recognition with an online speech interface and decryption of very long audio recordings. The Japanese language corpus CSJ was used in the first assessment. It contains approximately 650 hours of data on spontaneous Japanese speech. The TED-LIUMV2 corpus, which is a set of TED talks in English with transcriptions, was used for the second assessment. It contained approximately 200 h of speech data. Two types of network structures were investigated: a 6-layer bidirectional LSTM encoder with a 1-layer LSTM decoder and a 6-layer unidirectional LSTM encoder with a 2-layer LSTM decoder. The experimental results on non-segmented data show that the proposed method surpasses the basic methods using traditional VAD methods based on energy and neural networks. However, as in many studies, no study has been conducted on the analysis of the number of speakers.

In [16], an end-to-end segmentation model was trained, which performed a combination of three subtasks: detection of voice activity, detection of speaker change, and detection of overlapping speech. The DIHARD3 speech corpus, developed by the Linguistic Data Consortium and containing approximately 34 hours of speech data in English and Chinese, was selected for the study. In this study, the database was divided into two parts: 192 files used as a training set and the remaining 62 files used for testing. Experiments with multi-speaker diarization datasets have concluded that this model can be used with great success for both detecting speech activity and detecting overlapping speech.

In [17], a neural network was developed that combined trainable filters and recurrent layers to detect voice activity directly from the waveform. Experiments with a complex DIHARD dataset showed that the proposed end-to-end model achieved the most up-to-date indicators and surpassed the option in which the trained filters were replaced by standard cepstral coefficients. The same DIHARD dataset taken from 11 different areas was used to evaluate the two scenarios. In an intra-domain scenario, when training and testing sets cover the same domains, where it is shown that the domain-competitive approach does not reduce the performance of the proposed end-to-end model. In an off-domain scenario, where the test domain differs from the training domain, this results in a relative improvement of more than 10 %. However, the training and testing were conducted using only a single dataset.

In [18], the problem of detecting speech activity based on a convolutional neural network was considered. Experimental studies were conducted on two speech datasets: AMI, a multimodal dataset consisting of 100 hours of recordings of meetings in English, and the CHiME-6 dataset. The CHiME-6 case contained over 60 h of recordings organized into 20 sessions.

In [19], a comprehensive multitasking model for VAD with speech enhancement was proposed. The Wall Street Journal English Language Dataset (WSJ 0) was used as the source of the pure speech. It contains 12776 statements from 101 speakers for training, 1206 statements from 10 speakers for verification, and 651 statements from 8 speakers for evaluation. The experimental results demonstrate that the multitasking method is significantly superior to the single-tasking analog VAD.

In [20], a convolutional neural network (CNN) with multiple inputs and one output was proposed, which uses a new combination of functions to evaluate the VAD. The voices of 168 speakers from the TIMIT dataset, including both men and women, were used for the training and validation. Experimental results for a single-speaker scenario show that the proposed CNN can distinguish speech from blocks of nonspeech signals, thereby surpassing the basic CNN. In addition, the results show that the proposed method can be adapted to various invisible acoustic conditions and background noise.

In research on the use of methods for language-independent recognition of the human voice, there is a problem: large computing resources are required. However, the use of programmable logic integrated circuits (FPGAs) can be an effective and inexpensive solution to accelerate neural computing, which is also applicable for signal processing onboard satellites [21].

It is obvious that the methods of language-independent recognizing the human voice using ANNs are the most effective and promising; however, this task is still far from its final solution, and there are many unresolved problems and questions. As a result of studying and analyzing studies devoted to this area, it was found that there are currently few studies devoted to language-independent detection of voice activity, and no studies have analyzed the number of speakers needed to obtain a highly effective language-independent VAD system. In many of the studies conducted, the artificial neural network was trained and tested in only one language. According to the research conducted, it is not possible to assess the effectiveness of human voice recognition because speech can sound in different languages. Quite often, these issues are related to ensuring a sufficient amount of training data for ANNs. For example, when training an ANNs, one of the main requirements is to ensure the completeness, diversity, and parity of data. To do this, in most cases, a method for language-independent recognition fairly large amount of data is needed, which is not always possible to find and prepare, or their preparation requires huge human resources, time, and computing resources. In this case, the ideal variety of training data would mean that in order to teach ANNs to recognize the language-independent human voice, it would be necessary to use the voices (speech utterances) of different people in all languages of the world, ranging from children to elderly men and women. Of course, fulfilling such a requirement is quite difficult, so the question arises as to whether it is possible to limit these requirements, for example, to the number of languages and/or speakers so that the accuracy of recognition of the human voice of the ANNs remains at an acceptable level.

There are also not detected studies devoted to improving the accuracy of human voice recognition, because a variety of training data for language-independent human voice recognition would require using the voices (speech utterances) of different people in all languages of the world. Therefore, the unsolved problem lies in the study of methods and techniques that make it possible to select the optimal set of linguistic and phonetic characteristics for training neural networks to improve their accuracy and reliability in recognizing human voice.

All this allows to assert that it is expedient to conduct a study on limiting requirements, for example, to the number of languages and/or speakers, so that the accuracy of recognition of the human voice of the ANN remains at an acceptable level.

3. The aim and objectives of the study

The aim of this study шs to conduct a comprehensive assessment of the dependence of the effectiveness of neural networks on the language used in teaching, as well as the number of native speakers, on the accuracy of human voice recognition using neural networks.

The practical significance of this work lies in the fact that it has direct application value for the development and improvement of speech recognition systems in various languages. Research related to the dependence of the effectiveness of neural networks for recognizing human voices on language may lead to the development of new methods for training neural networks on multilingual data, improving speech recognition algorithms in different languages, and creating more accurate and universal models. The results obtained can be useful for the development of multilingual speech recognition systems, voice command recognition systems, and other applications working in different languages and even for translators. This is important in the context of the growing influence of a multilingual environment in a world in which there is a need to develop technologies that can work effectively in different languages and dialects. The results obtained in this work can have a significant impact on the development of speech recognition technologies and increase their effectiveness in multilingual environments.

To achieve this aim, the following objectives фre coniducted:

– to conduct a comparative analysis of MLP, CNN, and RUN neural networks to assess their effectiveness in recognizing the human voice;

– to evaluate the influence of language on the accuracy of human voice recognition using ANNs;

– to evaluate the impact of the number of speakers on the accuracy of human voice recognition using ANNs.

4. Materials and methods of research

The objects of research in this work are various artificial neural networks used to recognize human voices. Their ability to effectively recognize the human voice, regardless of language, is considered learning from a small number of speakers. The following three types of ANNs were considered as artificial neural networks used to solve the voice recognition problem: fully connected perceptron (MLP), convolutional neural network (CNN), and recurrent neural network (RNN).

The main hypothesis of this study is that despite the fact that the phonetics of different languages differ from each other, they have many common phonemes; therefore, a neural network trained in some languages should recognize human voices in other languages with the same efficiency. It was also assumed that in order to obtain acceptable accuracy of human voice recognition by neural networks, they can be trained on a limited number of speakers, approximately several hundred; however, it is also necessary to observe the parity of male and female voices. To answer these questions, the following simplifications and agreements were adopted.

– all neural networks were trained on the same datasets, representing the statements of the speakers in one language only (in Kazakh);

– the ratio of the amount of data to the number of paprameters of neural networks in all cases, which was the same and equal to 1:3;

– the required number of epochs in training neural networks is selected individually for each network based on the analysis of training and validation accuracy;

– when studying the dependence of the accuracy of neural networks on the number of speakers, the total amount of training data is a fixed constant, where the proportion of each speaker is the same and decreases proportionally with their growth;

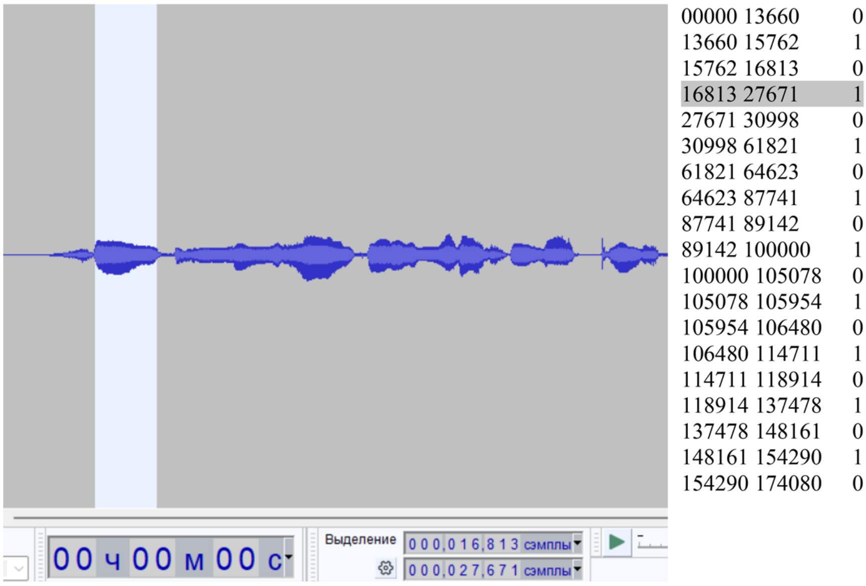
– the results of testing neural networks consider only those sections of audio data where speech is contained, that is, the ability of neural networks to correctly recognize only speech sections is analyzed.

To conduct training and testing of neural networks, datasets from the Institute of Smart Systems and Artificial Intelligence (ISSAI) of Nazarbayev University were used, namely the Kazakh speech corpus [22], Russian speech corpus [23], Turkish language corpus [24], and Uzbek language corpus [25]. One of the largest open datasets, the Common Voice Dataset [26], was also used, namely, the corpus of the Kyrgyz language, the corpus of the English language, and the corpus of the French language. From each data set, 15 male and 15 female voices were selected in a special way so that the voices were of different intonation, pitch, age, etc.

In this work, the audio files of the Kazakh speech corpus were manually marked [21] using Audacity 3.4.2 software product. The area of the audio file where the sound was present was marked as 1, and the area with missing sound was marked as 0. An example of manual markup of an audio file is shown in Fig. 1.

Next, each block of audio data was divided into fragments of 20 ms duration. Then, for each fragment, the spectral coefficients (MFCC), as well as the delta and delta-delta coefficients, were calculated. Thus, each fragment of the speech signal was represented by 36 coefficients.

When training MLP networks, 36 values were input, consisting of 12 MFCC coefficients, 12 delta coefficients, and 12 delta-delta coefficients. Accordingly, the input layer of the MLP network contains 36 neurons, and there is only one neuron at the output, which can accept either 1 if the current fragment is speech or 0 if the current fragment is not speech. The general structure of the MLP network used in this study is shown in Fig. 2.



00000 13660	0
13660 15762	1
15762 16813	0
16813 27671	1
27671 30998	0
30998 61821	1
61821 64623	0
64623 87741	1
87741 89142	0
89142 100000	1
100000 105078	0
105078 105954	1
105954 106480	0
106480 114711	1
114711 118914	0
118914 137478	1
137478 148161	0
148161 154290	1
154290 174080	0

Fig. 1. Manual partitioning of audio data into blocks

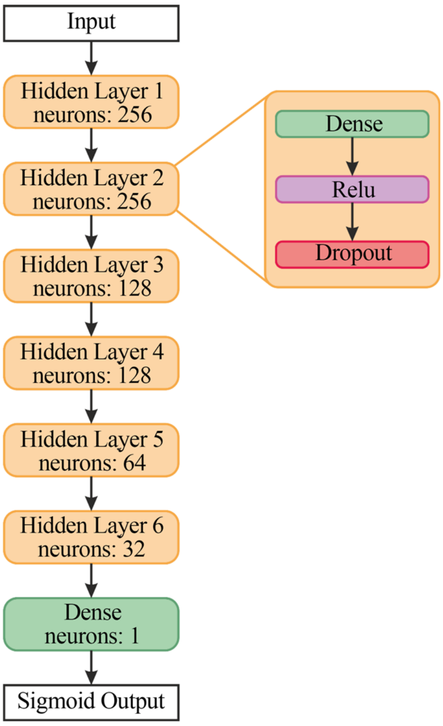


Fig. 2. Structure of multilayer perceptron networks

For CNN-type networks, only 12 MFCC coefficients were used, without the delta and delta-delta coefficients. Nevertheless, there are also 36 neurons at the input of this network; however, unlike MLP networks, the remaining 24 values are taken from subsequent neighboring fragments. Fig. 3 shows the general structure of the CNN-type networks used in this study.

The structure of the RNN-type networks used in this study is similar to that of the CNN. It also did not use delta and delta coefficients. In this case, the dynamics of the audio data were set using the MFCC coefficients of the three adjacent fragments of the signal.

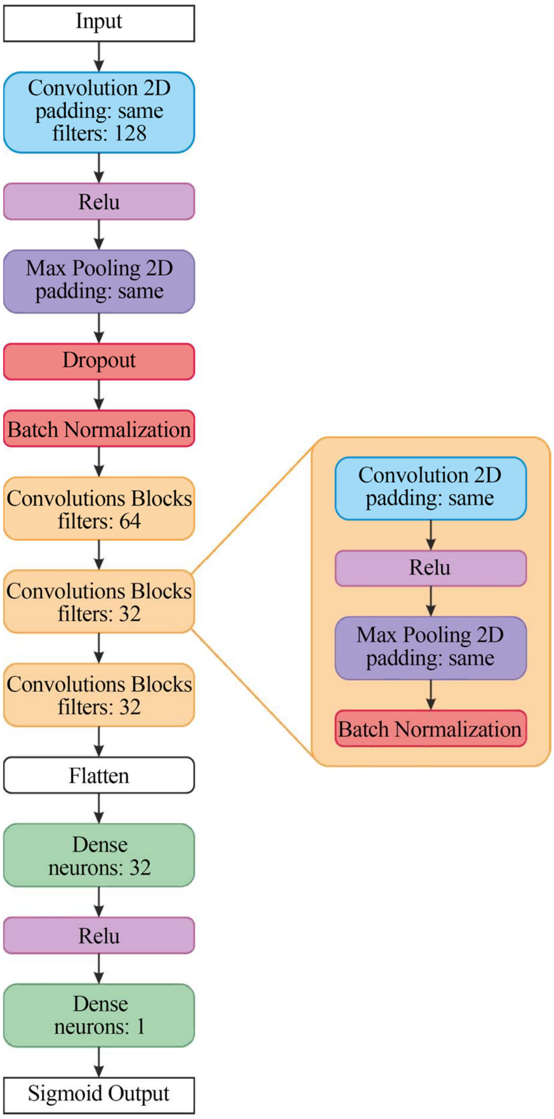


Fig. 3. Structure of multilayer perceptron networks

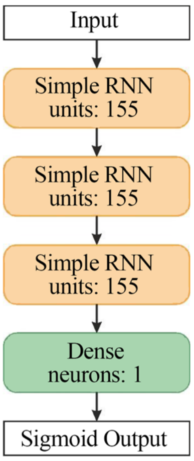


Fig. 4. The general structure of the recurrent neural network

Thus, our task was to determine how the accuracy of ANNs for recognizing the language-independent human voice depends on the number of speakers and languages used. For example, an ANNs can be trained on statements in English, but simultaneously, it is necessary to recognize the human voice from statements in German. That is, it is necessary to answer the question of whether it is possible to train a network in some languages and then use it to recognize a human voice in another language. Another important issue is determining the required number of speakers to ensure sufficient accuracy in human voice recognition. The task was to assess which type of ANNs would be most effective for the implementation of the task under consideration. Three types of INS were considered in this study: conventional multilayer perceptron (MLP), convolutional neural network (CNN), and recurrent neural network (RNN). Simultaneously, to ensure the same comparison conditions, ANNs were created so that the number of network parameters was approximately the same.

5. The result of the analysis of the effectiveness of neural networks for human voice recognition

5. 1. The result of a comparative analysis of the effectiveness of multilayer perceptron, convolutional neural network and recurrent neural networks for human voice recognition

This study demonstrated the dependence of human voice recognition errors on the number of speakers using MLP, CNN, and RNN networks. At the same time, all these networks are trained in only one Kazakh language, and their effectiveness has already been verified with the help of statements in other languages, including Kazakh.

The following figures show the results of testing various neural networks using statements from different languages. For example, Fig. 5 shows the dependence of human voice recognition error on the number of speakers for the MLP network for statements in Uzbek.

To approximate the dependence of recognition errors on the number of speakers, it is possible to consider nonlinear functions of the following types:

$$y = C \cdot x^{n}, \qquad (1)$$

$$y = C \cdot n^{x}, \qquad (2)$$

$$y = C \cdot e^{n \cdot x}, \qquad (3)$$

$$y = a + b \cdot \log(x), \qquad (4)$$

where y – recognition error, x – number of speakers, a, b, C and n – some empirical real numbers.

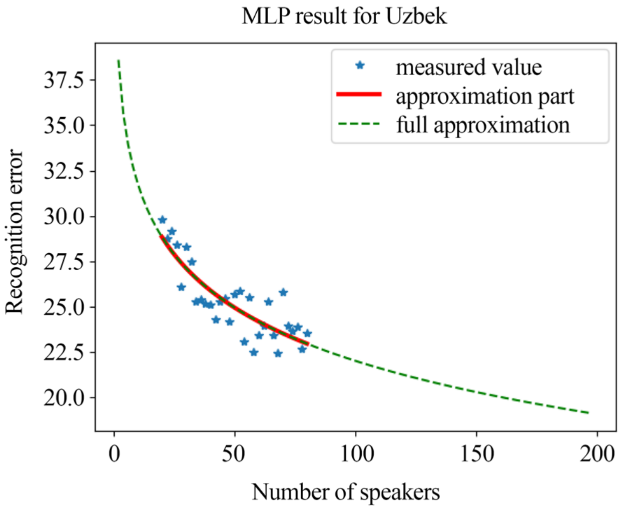


Fig. 5. Dependence of the human voice recognition error on the number of speakers for the multilayer perceptron network for statements in Uzbek

Our calculations show that the logarithmic approximation function is best suited for MLP. Accordingly, in Fig. 5, the red and green curves show the interpolation and extrapolation parts of approximation function (4).

Fig. 6 shows a graph of the dependence of the human voice recognition error on the number of speakers for the CNN network for French statements.

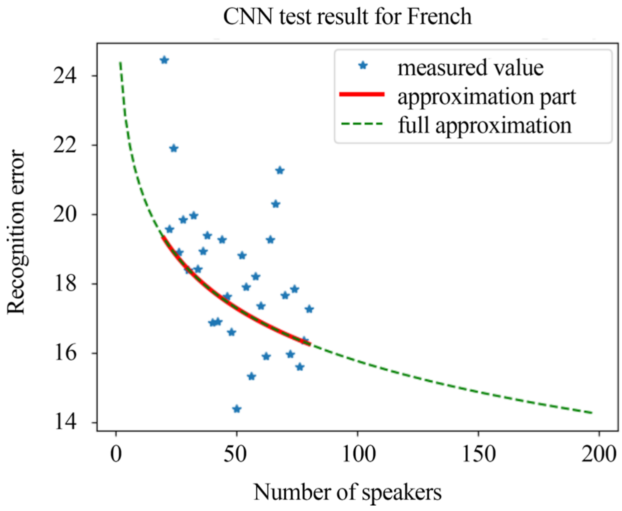


Fig. 6. Dependence of the human voice recognition error on the number of speakers for the convolutional neural network for statements in French

As an example, Fig. 7 shows the dependence of the human voice recognition error on the number of speakers for the RNN network for statements in the Kyrgyz language.

It should be noted that for the CNN and RNN networks, the dependence of the human voice recognition error on the number of speakers is difficult to approximate using the above four nonlinear functions. For CNN and RNN networks, a logarithmic approximation function was used in order to have a certain opportunity to compare the effectiveness of different neural networks among themselves. Fig. 6, 7 show the results obtained by using the logarithmic approximation function.

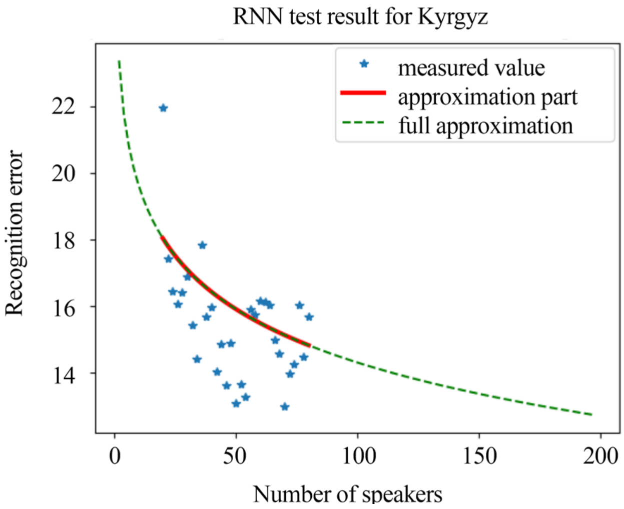


Fig. 7. Dependence of the human voice recognition error on the number of speakers for the recurrent neural network for statements in the Kyrgyz language

After selecting a general function for approximating the dependence of human voice recognition error on the number of speakers for all types of neural networks, the effectiveness of each neural network can be compared. Therefore, in Fig. 8, as an example, graphs of approximating functions for the MLP, CNN, and RNN networks are shown when testing them using statements in the Kazakh language.

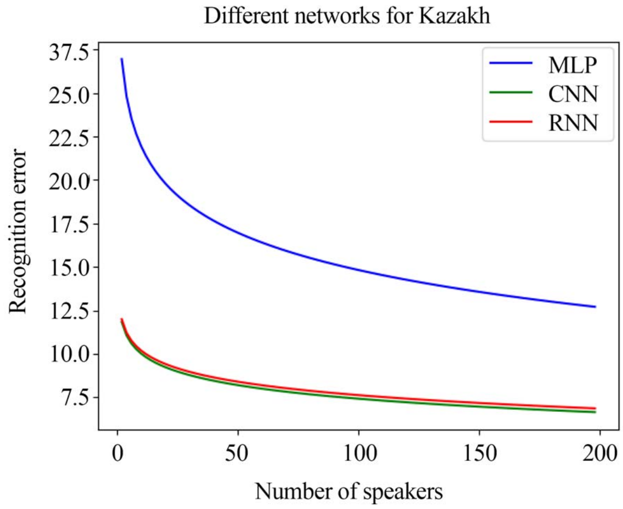


Fig. 8. Graphs of the logarithmic approximating function of the dependence of the human voice recognition error on the number of speakers for three types of neural networks multilayer perceptron, convolutional neural network and recurrent neural network

As shown in Fig. 8, the MLP neural network is much less accurate than the other two types of neural networks. It is also observed that the accuracies of the CNN and RNN neural networks are very close, with a slight advantage of the CNN. Therefore, based on these data, it is possible to assert that CNN is the most effective neural network for solving the problem of human voice recognition. Let's remind that the number of trained parameters for all three types of neural networks were set to be approximately the same.

5. 2. Assessment of the influence of language on the accuracy of human voice recognition using the artificial neural networks

To assess the impact of language on the accuracy of human voice recognition, let's use established approximation functions. Fig. 9 shows graphs of the dependence of human voice recognition error on statements in various languages on the number of speakers for the MLP neural network.

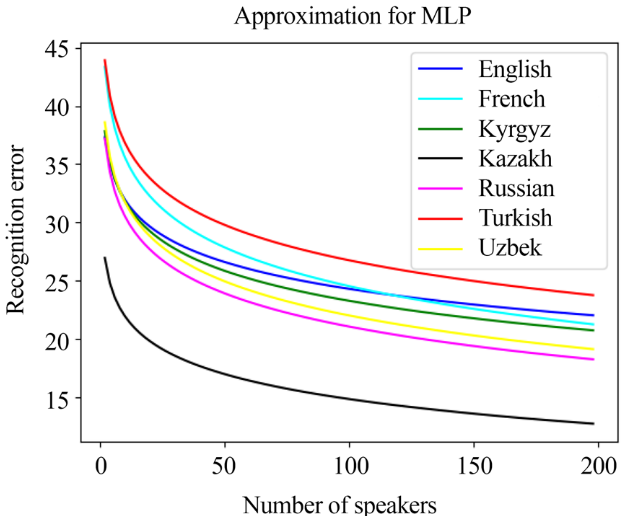


Fig. 9. Graphs of the dependence of the human voice recognition error on statements in different languages on the number of speakers for the multilayer perceptron neural network

Fig. 10 shows the same results for the CNN.

Finally, Fig. 11 shows graphs of the dependence of human voice recognition error on statements in various languages on the number of speakers for the MLP neural network.

Again, let's recall that all these neural networks were trained on a set of speech data only in the Kazakh language. The networks were tested using statements from various languages, including Kazakhs.

As shown in Fig. 9–11, the human voice is better recognized by statements in the Kazakh language by a noticeable margin than in other languages by all three neural networks. Another interesting fact is that different networks place different languages in second place. For example, for RNN, Kyrgyz is second after Kazakh, and for CNN and MLP, Russian is second. And in all cases, the Turkish language turned out to be an outsider, although it belongs to the same family of languages along with the Kazakh language.

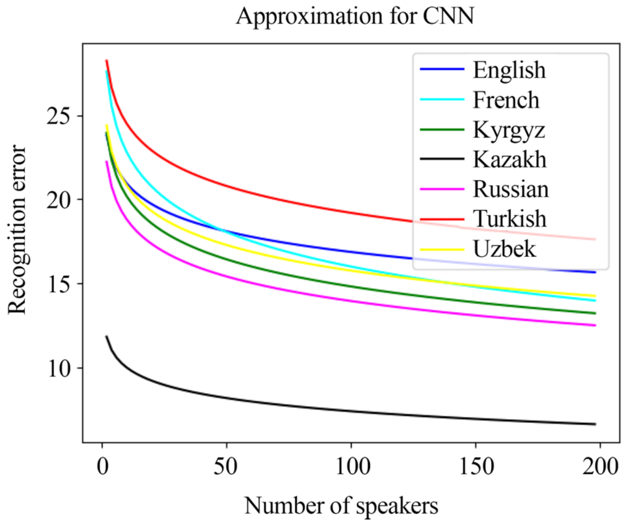


Fig. 10. Graphs of the dependence of the human voice recognition error on statements in different languages on the number of speakers for the CNN neural network

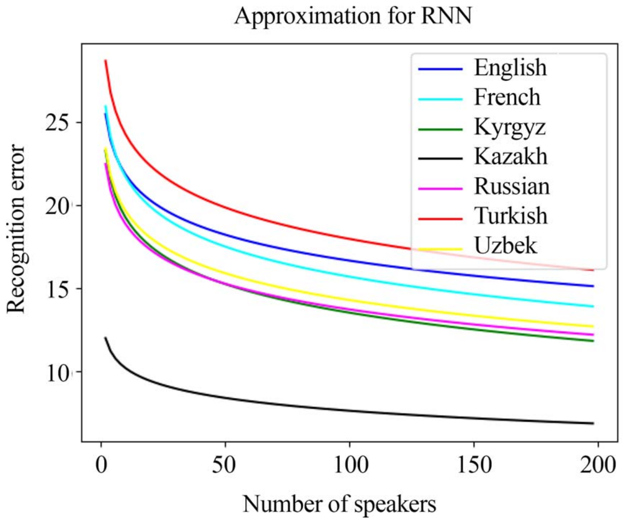


Fig. 11. Graphs of the dependence of the human voice recognition error on statements in different languages on the number of speakers for the recurrent neural network

5. 3. Assessment of the impact of the number of speakers on the accuracy of human voice recognition using the artificial neural networks

All the previous graphs shown in Fig. 5–11 indicate that there is an unambiguous dependence of the accuracy of human voice recognition by neural networks on the number of speakers used in training these networks. Obviously, with an increase in the number of speakers, the error in recognizing the human voice decreases, that is, the efficiency of the neural network increases. As a result, the dependence of the recognition error on the number of speakers for the MLP network (using the example of the Kazakh language) can be represented as follows:

$$y = 29{,}10 - 3{,}10 \cdot \log(x), \quad (5)$$

and for the CNN network, as well as for the Kazakh language, this dependence can be expressed using the following expression:

$$y = 12{,}62 - 1{,}13 \cdot \log(x), \quad (6)$$

finally, there is the following pattern for the RNN network:

$$y = 12{,}77 - 1{,}12 \cdot \log(x), \quad (7)$$

where in all formulas (5)–(7) y means the magnitude of the error in recognizing the human voice by the neural network, and x is the number of speakers whose statements were used in training neural networks.

Calculations show that if these patterns are correct, then in order to get a human voice recognition error of no more than 3 % (accuracy of at least 97 %), it is necessary to have at least 4.5 thousand voice samples for training neural networks.

6. Discussion of the results of evaluating the accuracy of human voice recognition

Initially, it was assumed that the efficiency of neural networks depends weakly on the speaker's language since all languages have a large number of similar phonemes. Thus, it is expected that by training a neural network in one language, it would be possible to detect human voices in other languages with fairly good accuracy using the same neural network. However, these expectations have not been fully fulfilled. However, there is a dependence on language. This can be seen in Fig. 9–11, where MLP, CNN, and RNN neural networks trained in Kazakh were much better at detecting voices in the same language than voices in other languages. However, at the same time, a strange result is obtained. Although Kazakh, Kyrgyz, Uzbek, and Turkish are considered related languages, that is, they belong to the same group of Turkic languages, neural networks trained in Kazakhs recognize voices in Russian better than others. This can be observed in Fig. 9, 10. Perhaps this is due to the fact that the Turkish language has a much greater phonetic difference from the Kazakh language than the same russian language. Thus, the methods and results of this study could be useful for studying phonetic similarities between different languages.

Regarding the choice of neural network architecture for optimal human voice recognition, the results in Fig. 8 show that the CNN is the best network. The correctness of this statement can be easily verified using formulas (5) and (6), where the functions for approximating errors in human voice recognition by CNN and RNN are given. For example, if 1000 speakers are used in training, calculations using formula 5 for the CNN network give an error of 4.8 %, and for RNN, according to formula (6), the recognition error is 5 %. As it is possible to see, in our case, the CNN network has a slight advantage over the RNN network. However, with an equal number of trained parameters, the MLP neural network demonstrated significantly lower efficiency in this task.

The results of the study presented in Fig. 5–7 confirm the dependence of the accuracy of the recognition of neural networks on the number of speakers used in training. An increase in the number of speakers significantly affects the accuracy of human voice recognition using networks. For example, for an MLP, such an increase is logarithmic. However, for CNN and RNN, such a functional relationship cannot be precisely established. Nevertheless, to compare different networks, let's also use the same logarithmic approximation function for CNN and RNN. The end result of this study is the answer to the question of how many different voices should be used when training a neural network to achieve high recognition accuracy. The presence of an approximation function provides an answer to this question. Calculations carried out according to the obtained approximation functions showed that, for example, in order to achieve 97 % accuracy, at least 4,500 speakers are needed.

The main problem that arises when conducting such studies is related to the fact that they require huge computing resources. In this regard, an inexpensive solution for accelerating neural computing is the use of FPGAs [27]. In general, the computing resources of FPGAs in the implementation of artificial neural networks can also be applied to the processing of other signals, especially onboard low-orbit satellites [28, 29].

In this study, in addition to the three types of artificial neural networks considered, other types of ANNs could be used, such neural networks (SNN). SNN-type networks are closest to real biological neurons and can describe time series as accurately as possible. Many properties of SNN networks can be obtained by studying their neurodynamics [30, 31]. Accordingly, let's believe that these future studies should be carried out with a large number of types of neural networks, while it is desirable to implement them on FPGAs to obtain the necessary computational performance.

7. Conclusions

1. The best architecture for human voice recognition is the CNN neural network, which has a slight advantage over RNN. For example, for the starting value of the number of speakers (20 speakers), the error of voice recognition in Kazakhs by the CNN network was 10.6 % and for the RNN network was 11.9 %. Thus, for 20 speakers, the accuracy of rechognition by the CNN was 1.3 % higher than that of the RNN. For a finite number of announcers (80 announcers), the error of voice recognition in Kazakhs by the CNN network is 8.3 %, and for RNN, it is 8.93 %, that is, CNN accuracy is 0.9 % points higher than that of RNN.

2. The accuracy of human voice recognition by a neural network strongly depends on the choice of language. In our case, neural networks were trained in the Kazakhs alone, and the effectiveness of these networks was tested using people's statements in other languages. The dependence of the accuracy of human voice recognition on language can be seen in the example of a CNN. For the Kazakh language, when 20 speakers were used, the CNN network showed an error of 10.6 %, and for the nearest result (Russian), it showed an error of 21.86 %. Among the 80 speakers, 8.3 % and 16.2 %, respectively. That is, the accuracy of voice recognition in Kazakh is at least two times higher than that in other languages. Accordingly, in order to obtain a neural network capable of recognizing human voices with high accuracy, it cannot be trained in language alone.

3. In addition to language, the accuracy of human voice recognition by neural networks is strongly influenced by the number of speakers used in training these networks. For example, for a neural network trained in a particular language to recognize human voices in the same language with an accuracy of at least 97 %, it is necessary to use different voices from more than 4.5 thousand speakers. It is quite possible that, for a more complex structure (with a large number of trainable parameters), the required number of speakers may be very different. Therefore, our estimate could not provide the final number of required speakers. Nevertheless, it shows that the accuracy of recognizing any human voice using a neural network depends quite strongly on the variety of voices used in training the corresponding neural networks.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study and the results reported in this paper.

Financing

This research has been/was/is funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP19678995) "Development of a speaker recognition method using deep neural networks for ultrashort duration of pure speech".

Data availability

All data are available in the main text of the manuscript.

Use of artificial intelligence

The authors used artificial intelligence technologies within acceptable limits to provide their own verified data, which are described in the section "Research Methodology" section.

References

1. Mihalache, S., Burileanu, D. (2022). Using Voice Activity Detection and Deep Neural Networks with Hybrid Speech Feature Extraction for Deceptive Speech Detection. Sensors, 22 (3), 1228. https://doi.org/10.3390/s22031228
2. Lee, Y., Min, J., Han, D. K., Ko, H. (2020). Spectro-Temporal Attention-Based Voice Activity Detection. IEEE Signal Processing Letters, 27, 131–135. https://doi.org/10.1109/lsp.2019.2959917
3. Sofer, A., Chazan, S. E. (2022). CNN self-attention voice activity detector. arXiv. https://doi.org/10.48550/arXiv.2203.02944
4. Zhang, X.-L., Xu, M. (2022). AUC optimization for deep learning-based voice activity detection. EURASIP Journal on Audio, Speech, and Music Processing, 2022 (1). https://doi.org/10.1186/s13636-022-00260-9
5. Jia, F., Majumdar, S., Ginsburg, B. (2021). MarbleNet: Deep 1D Time-Channel Separable Convolutional Neural Network for Voice Activity Detection. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/icassp39728.2021.9414470
6. Heo, Y., Lee, S. (2023). Supervised Contrastive Learning for Voice Activity Detection. Electronics, 12 (3), 705. https://doi.org/10.3390/electronics12030705

7. Faghani, M., Rezaee-Dehsorkh, H., Ravanshad, N., Aminzadeh, H. (2023). Ultra-Low-Power Voice Activity Detection System Using Level-Crossing Sampling. Electronics, 12 (4), 795. https://doi.org/10.3390/electronics12040795
8. Lee, G. W., Kim, H. K. (2020). Multi-Task Learning U-Net for Single-Channel Speech Enhancement and Mask-Based Voice Activity Detection. Applied Sciences, 10 (9), 3230. https://doi.org/10.3390/app10093230
9. Arslan, O., Engin, E. Z. (2019). Noise Robust Voice Activity Detection Based on Multi-Layer Feed-Forward Neural Network. Electrica, 19 (2), 91–100. https://doi.org/10.26650/electrica.2019.18042
10. Oh, Y. R., Park, K., Park, J. G. (2020). Online Speech Recognition Using Multichannel Parallel Acoustic Score Computation and Deep Neural Network (DNN)- Based Voice-Activity Detector. Applied Sciences, 10 (12), 4091. https://doi.org/10.3390/app10124091
11. Sehgal, A., Kehtarnavaz, N. (2018). A Convolutional Neural Network Smartphone App for Real-Time Voice Activity Detection. IEEE Access, 6, 9017–9026. https://doi.org/10.1109/access.2018.2800728
12. Mukherjee, H., Obaidullah, Sk. Md., Santosh, K. C., Phadikar, S., Roy, K. (2018). Line spectral frequency-based features and extreme learning machine for voice activity detection from audio signal. International Journal of Speech Technology, 21 (4), 753–760. https://doi.org/10.1007/s10772-018-9525-6
13. Ali, Z., Talha, M. (2018). Innovative Method for Unsupervised Voice Activity Detection and Classification of Audio Segments. IEEE Access, 6, 15494–15504. https://doi.org/10.1109/access.2018.2805845
14. Jung, Y., Kim, Y., Choi, Y., Kim, H. (2018). Joint Learning Using Denoising Variational Autoencoders for Voice Activity Detection. Interspeech 2018. https://doi.org/10.21437/interspeech.2018-1151
15. Yoshimura, T., Hayashi, T., Takeda, K., Watanabe, S. (2020). End-to-End Automatic Speech Recognition Integrated with CTC-Based Voice Activity Detection. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/icassp40776.2020.9054358
16. Bredin, H., Laurent, A. (2021). End-To-End Speaker Segmentation for Overlap-Aware Resegmentation. Interspeech 2021. https://doi.org/10.21437/interspeech.2021-560
17. Lavechin, M., Gill, M.-P., Bousbib, R., Bredin, H., Garcia-Perera, L. P. (2020). End-to-End Domain-Adversarial Voice Activity Detection. Interspeech 2020. https://doi.org/10.21437/interspeech.2020-2285
18. Cornell, S., Omologo, M., Squartini, S., Vincent, E. (2020). Detecting and Counting Overlapping Speakers in Distant Speech Scenarios. Interspeech 2020. https://doi.org/10.21437/interspeech.2020-2671
19. Tan, X., Zhang, X.-L. (2021). Speech Enhancement Aided End-To-End Multi-Task Learning for Voice Activity Detection. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/icassp39728.2021.9414445
20. Varzandeh, R., Adiloglu, K., Doclo, S., Hohmann, V. (2020). Exploiting Periodicity Features for Joint Detection and DOA Estimation of Speech Sources Using Convolutional Neural Networks. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/icassp40776.2020.9054754
21. Medetov, B., Kulakayeva, A., Zhetpisbayeva, A., Albanbay, N., Kabduali, T. (2023). Identifying the regularities of the signal detection method using the Kalman filter. Eastern-European Journal of Enterprise Technologies, 5 (9 (125)), 26–34. https://doi.org/10.15587/1729-4061.2023.289472
22. Mussakhojayeva, S., Khassanov, Y., Atakan Varol, H. (2022). KSC2: An Industrial-Scale Open-Source Kazakh Speech Corpus. Interspeech 2022. https://doi.org/10.21437/interspeech.2022-421
23. Mussakhojayeva, S., Khassanov, Y., Atakan Varol, H. (2021). A Study of Multilingual End-to-End Speech Recognition for Kazakh, Russian, and English. Lecture Notes in Computer Science, 448–459. https://doi.org/10.1007/978-3-030-87802-3_41
24. Mussakhojayeva, S., Dauletbek, K., Yeshpanov, R., Varol, H. A. (2023). Multilingual Speech Recognition for Turkic Languages. Information, 14 (2), 74. https://doi.org/10.3390/info14020074
25. Musaev, M., Mussakhojayeva, S., Khujayorov, I., Khassanov, Y., Ochilov, M., Atakan Varol, H. (2021). USC: An Open-Source Uzbek Speech Corpus and Initial Speech Recognition Experiments. Lecture Notes in Computer Science, 437–447. https://doi.org/10.1007/978-3-030-87802-3_40
26. Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J. et al. (2020). Common voice: A massively-multilingualspeech corpus. arXiv. https://doi.org/10.48550/arXiv.1912.06670
27. Medetov, B., Serikov, T., Tolegenova, A., Zhexebay, D., Yskak, A., Namazbayev, T., Albanbay, N. (2023). Development of a model for determining the necessary FPGA computing resource for placing a multilayer neural network on it. Eastern-European Journal of Enterprise Technologies, 4 (4 (124)), 34–45. https://doi.org/10.15587/1729-4061.2023.281731
28. Aigul, K., Altay, A., Yevgeniya, D., Bekbolat, M., Zhadyra, O. (2022). Improvement of Signal Reception Reliability at Satellite Spectrum Monitoring System. IEEE Access, 10, 101399–101407. https://doi.org/10.1109/access.2022.3206953
29. Aitmagambetov, A., Butuzov, Y., Butuzov, Y., Tikhvinskiy, V., Tikhvinskiy, V., Kulakayeva, A. et al. (2021). Energy budget and methods for determining coordinates for a radiomonitoring system based on a small spacecraft. Indonesian Journal of Electrical Engineering and Computer Science, 21 (2), 945. https://doi.org/10.11591/ijeecs.v21.i2.pp945-956
30. Albanbay, N., Medetov, B., Zaks, M. A. (2021). Exponential distribution of lifetimes for transient bursting states in coupled noisy excitable systems. Chaos: An Interdisciplinary Journal of Nonlinear Science, 31 (9). https://doi.org/10.1063/5.0059102
31. Albanbay, N., Medetov, B., Zaks, M. A. (2020). Statistics of Lifetimes for Transient Bursting States in Coupled Noisy Excitable Systems. Journal of Computational and Nonlinear Dynamics, 15 (12). https://doi.org/10.1115/1.4047867