This paper considers the efficiency of neural networks for human voice recognition. The objects of the study are artificial neural networks used for human voice recognition. Their ability to effectively recognize a human voice regardless of language, trained on a small number of speakers in noisy conditions, has been considered. The task being solved is to enhance the accuracy of speech activity detection, which plays a significant role in improving the functioning of automatic speech recognition systems, especially under conditions of a low signal-to-noise ratio.

The findings showed that the accuracy of human voice recognition in languages of different phonetic proximity could vary greatly. As a result of the study, it was found that the recurrent neural network (RNN) demonstrates high accuracy in voice recognition – 95 %, which exceeds the results of the convolutional neural network (CNN), reaching an accuracy of 94 %. Special features of the results are the adaptation of neural networks to multilingual features, which made it possible to increase the efficiency of their work. An important conclusion was that training neural networks on data with different languages and types of speakers significantly improves recognition accuracy. The study confirmed that training neural networks on different languages and speaker types could significantly affect recognition accuracy. The results are an important contribution to the development of speech recognition technologies and have the potential for application in various fields where high accuracy in human voice recognition is required

Keywords: convolutional neural network, recurrent neural network, voice activity detector

UDC 004.032.26

DOI: 10.15587/1729-4061.2025.321659

EVALUATING THE EFFECTIVENESS OF A VOICE ACTIVITY DETECTOR BASED ON VARIOUS NEURAL NETWORKS

Bekbolat Medetov
PhD*

Ainur Zhetpisbayeva
Corresponding author
PhD, Associate Professor*
E-mail: aigulji@mail.ru

Ainur Akhmediyarova
PhD
Department of Software Engineering***

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

Timur Namazbayev
Master, Senior Lecturer
Al-Farabi Kazakh National University
Al-Farabi ave., 71, Almaty, Republic of Kazakhstan, 050040

Aigul Kulakayeva
PhD
Department of Radio Engineering, Electronics and Telecommunications
International Information Technology University
Manasa str., 34 a, Almaty, Republic of Kazakhstan, 050000

Nurtay Albanbay
PhD
Department of Cybersecurity, Information Processing and Storage***

Mussa Turdalyuly
PhD, Associate Professor
Department of Software Engineering***

Asset Yskak
Lead Design Engineer
Department of Software Development
Ghalam LLP
Turan ave., 89, Astana, Republic of Kazakhstan, 010000

Gulzhazira Uristimbek
Master's Student
Department of Algebra and Geometry**
*Department of Radio Engineering, Electronics and Telecommunications**
**L.N. Gumilyov Eurasian National University
Satbaev str., 2, Astana, Republic of Kazakhstan, 010000
***Satbayev University
Satbayev str., 22, Almaty, Republic of Kazakhstan, 050000

Received 23.10.2024
Received in revised form 13.12.2024
Accepted date 30.12.2024
Published date 28.02.2025

How to Cite: *Medetov, B., Zhetpisbayeva, A., Akhmediyarova, A., Nurlankyzy, A., Namazbayev, T., Kulakayeva, A., Albanbay, N., Turdalyuly, M., Yskak, A., Uristimbek, G. (2025). Evaluating the effectiveness of a voice activity detector based on various neural networks. Eastern-European Journal of Enterprise Technologies, 1 (5 (133)), 19–28. https://doi.org/10.15587/1729-4061.2025.321659*

1. Introduction

Spoken language, as a fundamental tool of human communication, has a unique ability to convey complex ideas and emotional states. In the era of globalization and internationalization, adequate speech recognition is of strategic importance. Automatic Speech Recognition (ASR) technologies transform spoken language into text, finding application in various domains of life.

Improving the accuracy and reliability of ASR systems is critical in situations where noise interference, accents, and other destabilizing factors significantly affect the quality of recognition. The problem of effective recognition in a multilingual and multi-accent environment remains insufficiently solved, despite significant advances in recent decades. Voice Activity Detection (VAD) systems, which help isolate speech fragments in audio streams, play a significant role in improving the functioning of ASR but their effectiveness under conditions of low signal-to-noise ratio remains a subject for further improvement.

Therefore, ongoing research aimed at optimizing speech recognition and analysis technologies is becoming highly relevant. It specifically address the key challenges of improving the accessibility of information technologies, ensuring increased levels of safety and comfort in everyday life, and contributing to the development of innovative methods of machine learning and artificial intelligence. The research results are essential for progress in the field of human-machine interaction, highlighting the critical need for continued scientific developments in this area.

Thus, ongoing scientific research in the field of automatic speech recognition and VAD systems remains highly relevant as it contributes to solving a number of applied and theoretical tasks, thereby confirming the need for further advances in this area.

2. Literature review and problem statement

Modern ASR systems have made significant progress in recent decades. Despite the widespread use and diversity of systems in use, full speech recognition remains an unsolved problem [1]. This task is caused by many factors, such as noise, transmission distortions, accents, as well as differences in the rate and manner of speech. These factors can significantly affect the accuracy and reliability of speech recognition systems, creating significant challenges for developers and researchers in this field. To reduce the impact of non-speech fragments on speech signal processing systems, it is assumed that the input audio signal contains only human speech. For this purpose, VAD, also known as speech activity detection or speech recognition, is used. The VAD system is an important component of the ASR speech recognition system. Correct and error-free operation of the ASR system is an important factor for improving the quality and accuracy of speech signal recognition and detection under real-world conditions.

In [2], it was proposed to use VAD based on deep learning (DNN) with any decision threshold. The only English language database LibriSpeech ASR, containing 1000 hours of read English speech, was used to train and test this method. A large sound effects library containing over 20,000 sound effects and the NOISEX-92 database containing 9 noise scenarios were also used. However, despite the effectiveness of the designed VAD system, there are still unresolved issues related to evaluating the performance of VAD based on multiple datasets.

In [3], the TIMIT English speech corpus was used for training, which included 4620 sentences and 1680 clean utterances in both the training and test sets. The training dataset was supplemented with the NOISEX-921 dataset, including eight types of noise. However, the work notes that further research is needed to overcome complex issues, such as evaluating the performance of VAD in a multilingual environment.

In [4], a study was conducted on an experiment on noise reduction to improve speech in noisy environmental situations with a limited number of speakers (13 male voices, 16 emale voices). The AURORA and LibriSpeech English speech datasets were used to train deep learning models. However, despite the fact that this work attempted to achieve high VAD performance on a limited number of speakers, issues related to recognition in a multilingual environment remained unresolved.

The task of deploying deep learning models for ecological speech recognition was considered in [5]. The Libri-Speech dataset was also used in the work to detect vocal activity. The balance between male and female voices was 1:1. The ESC-50 and BirdClef 2017 datasets were used as a noise source. However, in the work, as in many papers, the training and testing of the model was carried out using only one single English language dataset, i. e., the problem of multilingual recognition remains unsolved.

To achieve high performance in voice activity detection, a heterogeneous convolutional recurrent neural network (HCRNN) with an attention mechanism was proposed in [6]. The AI SHELL corpus was used to evaluate the effectiveness of the proposed method. The model was trained using data from 340 speakers (120,098 utterances, a total of 340 hours). However, the paper notes the importance of further research to achieve the same high performance with a smaller number of speakers.

In [7], a deep neural network (DNN) system was proposed for automatic speech detection in audio signals. Several types of DNN were investigated, including multilayer perceptrons (MLP), recurrent neural networks (RNN), and convolutional neural networks (CNN). The systems were trained and tested on several subsets of data from the Japanese CENSREC-1-C database, considering different simulated ambient noise conditions. For each noise type, 52 female and 52 male voices were used. However, the authors of the paper note the need to conduct research on the performance of VAD using a dataset of other languages.

In [8], the process of automating the design of the neural network architecture was used for the VAD task. The TIMIT dataset, which contains about 5.4 hours of speeches in English, was used as a source of pure voice. The SoundIdeas dataset was applied as noise sources. However, the training and testing of the model was also carried out on the basis of a single English speech corpus.

This paper differs from our previous work [9] on the study of language-independent human voice recognition. The main objective of this study is to find the type of function that most accurately describes the dependence of the human voice recognition error on the number of speakers for CNN, RNN, and MLP neural networks.

Based on our analysis, it can be argued that the development of an effective and reliable VAD algorithm is a relevant area, which is due to several factors. Firstly, 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 makes them limited in applicability to other languages. Secondly, there are currently few studies aimed at investigating the required number of speakers for effective human voice recognition. In addition, combining traditional VAD methods with neural networks allows us to achieve the most effective VAD system for studying phonetic proximity between different languages.

Thus, the development of an effective and reliable VAD algorithm is an important factor in improving the quality of the speech signal under real operating conditions in automatic speech recognition systems.

3. The aim and objectives of the study

The aim of our study is to evaluate the efficiency of using CNN, RNN, and MLP neural networks in speech signal recognition tasks under noise conditions with a S/N ratio of 20 dB, taking into account the influence of language features and the number of speakers on the recognition accuracy. This could make it possible to achieve high accuracy of human voice recognition in a multilingual and multi-accent environment on a limited data set due to the use of a training strategy aimed at minimizing only one error component.

To achieve the goal, the following tasks were set:

– to determine a suitable function for describing dependences in CNN, RNN, and MLP neural networks;

– to perform a comparative analysis of the efficiency of CNN, RNN, and MLP neural networks in human voice recognition tasks using function approximation methods;

– to investigate how language features and the number of speakers affect the accuracy of voice recognition in different types of neural networks.

4. The study materials and methods

The study objects of our work are various artificial neural networks used for human voice recognition. Their ability to effectively recognize the human voice regardless of language, trained on a small number of speakers under noisy conditions, has been considered.

The main hypothesis of the study assumes 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 one language should recognize human voices in other languages with the same efficiency. It was also assumed that in order to achieve acceptable accuracy of human voice recognition by neural networks, they can be trained on a limited number of speakers, approximately several hundred, but it is necessary to maintain parity between male and female voices.

To train and test neural networks, datasets from the Institute of Smart Systems and Artificial Intelligence (ISSAI) at Nazarbayev University were used, namely the corpus of Kazakh speech, the corpus of Russian speech, the corpus of the Turkish language, and the corpus of the Uzbek language. One of the largest open datasets, Common Voice Dataset, was also used, namely the Kyrgyz language corpus and the English language corpus, the French language corpus. From each dataset, 20 male and 20 female voices were chosen and selected in a special way so that the voices had different intonation, pitch, age, etc.

As a noise source, one of the largest databases of audio fragments, samples, recordings, sound signals FREESOUND.ORG, released under Creative Commons licenses, was used. Also, 10 types of noise were applied, such as sounds of electrical appliances, sounds of rain, birds singing, sounds of the highway, airplane, stadium, etc.

In this work, manual marking of audio files of the Kazakh speech corpus was carried out using the Audacity 3.4.2 software. The area of the audio file where the sound is present was marked as 1, the area with no sound was marked as 0. An example of manual marking of an audio file is shown in Fig. 1.

The main focus of the study is on determining the functional dependence of the accuracy of human voice recognition by neural networks on the number of speakers used to create the training data. It is obvious that the accuracy of the trained neural networks depends not only on the number of speakers but also on many other factors. In general, it can be assumed that the accuracy of neural networks recognition is a function of many variables of the following type:

$$y = f(a, b, c, \ldots), \tag{1}$$

where *a* is the number of announcers, *b* is the signal-to-noise ratio, *c* is the number of layers of the neural network, etc. It is clear that if during training and testing of neural networks all parameters are constant, except for the number of announcers, a one-dimensional function can be considered. Thus, the function specified in (1) takes the following form:

$$y = f(a). \tag{2}$$

In order to achieve the dependence of the accuracy of human voice recognition by neural networks only on the number of announcers, the data for training the neural network were formed as follows:

– a long audio file was created, consisting only of different noises;

– an audio file was created, consisting of *K* voices of the Kazakh language. *K* varies from 2 to 40 with a step of 2 (1 male and 1 female voice);

– if the length of the audio file of *K* voices was less than the length of the audio file of noise, then *K* voices were duplicated and added to the file until the length of the two files was the same;

– the resulting two audio files were mixed with each other with an SNR value of 20 dB;

– 36 *MFCC* coefficients (taking into account delta and delta-delta) are calculated from the mixed audio file, which serve as a set of training input data.

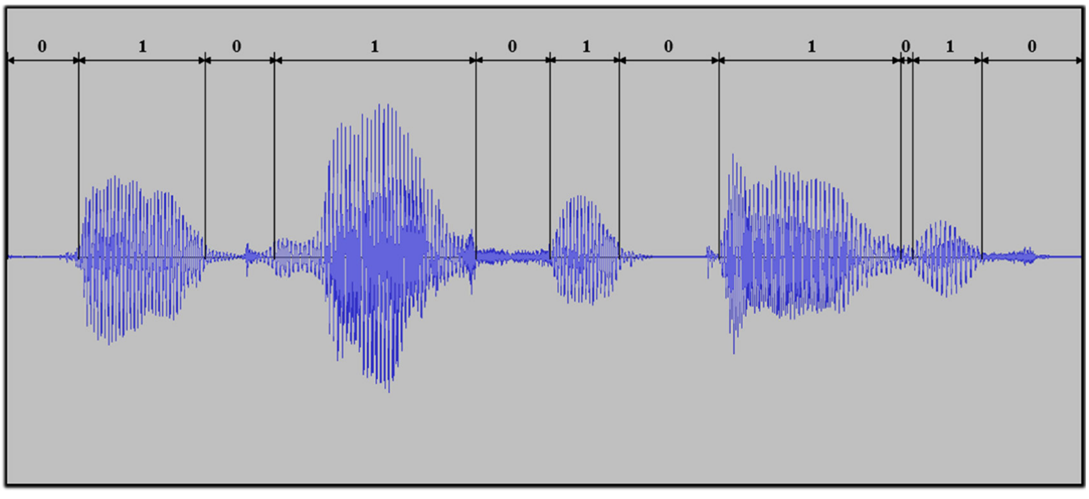

Fig. 1. Manual audio data tagging

The generated files, consisting of *MFCC* coefficients, were used to train neural networks. Depending on the type of neural network, *MFCC*s were arranged differently. For the MLP neural network, the training data was supplied as a vector of length 36. Namely, all 36 *MFCC* coefficients were used.

When working with CNN and RNN neural networks, only the first 12 *MFCC* coefficients were used. And the input data of these neural networks was a matrix in the form of 12×3, where the *MFCC* coefficients were sequentially combined.

Table 1 gives information on the types of neural networks used and the number of their parameters.

Table 1

The number of neural network parameters and training data

Neural network	Total amount of trainable data	Number of neural network parameters
CNN	1 393 100	98 657
MLP	1 507 403	97 985
RNN	1 393 103	98 981

Modeling of the neural network structure, their configuration and training, as well as testing were carried out using the Python programming language version 3.9 and the TensorFlow library version 2.10. The Adam optimizer was used to train all neural networks, and losses were calculated using the Binary Crossentropy function. Only for the RNN neural network, the Mean squared error loss function was used.

The structures of the CNN RNN MLP neural networks are shown in Fig. 2–4.

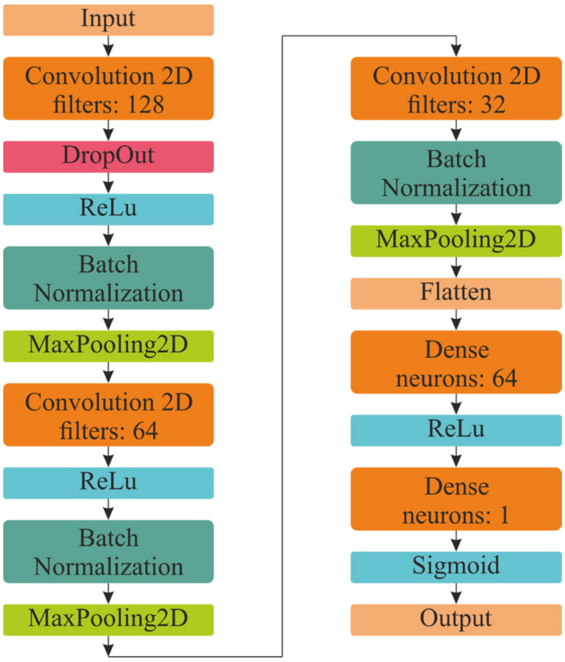

Fig. 2. Structure of a convolutional neural network (CNN)

During training, the functions for early stopping of training (EarlyStopping) and saving the best model (ModelCheckpoint) were used. EarlyStopping stopped training when the "val_accuracy" value did not improve for 10 epochs in a row. And ModelCheckpoint saved the maximum "val_accuracy" value each time. To improve the performance of operations during training and testing of the created neural networks, an NVIDIA RTX4090 video card with 24 GB of memory was used.

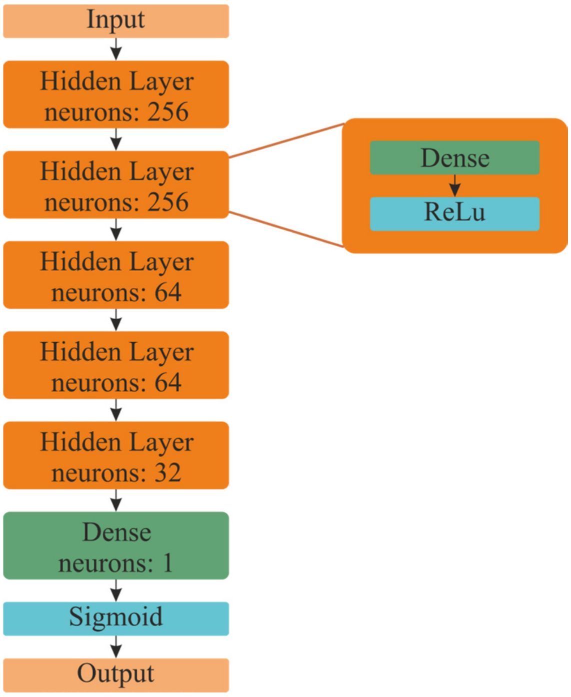

Fig. 3. Structure of a Multilayer Perceptron (MLP)

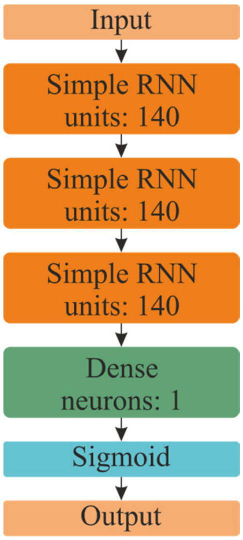

Fig. 4. Structure of a recurrent neural network (RNN)

5. Results of analyzing the efficiency of neural networks for human voice recognition

5. 1. Determining a suitable function for describing dependences in neural networks of the CNN, RNN, and MLP types

A function was determined that most accurately describes the dependence of the speech signal recognition error on the number of announcers for different types of neural networks. For each type of neural networks (in this case, CNN, RNN, and MLP), it was taken into account that the generalized error includes two components:

1) erroneous recognition of a section that does not contain speech data as a human voice (False positive);

2) erroneous recognition of a section containing speech data as a non-human voice (False negative).

Fig. 5 shows the result of testing the generalized accuracy of human voice recognition by a CNN neural network (the structure of this network is shown in Fig. 2) depending on the number of announcers.

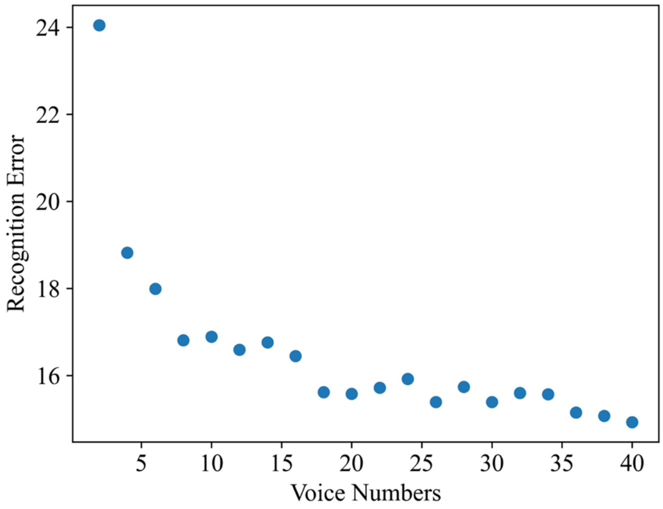

Fig. 5. Experimentally measured dependence of the total error of human voice recognition using CNN on the number of announcers

It is clearly seen from Fig. 5 that the overall error of human voice recognition using CNN, depending on the number of announcers used in the formation of training data, is constantly decreasing. It can also be noted that this dependence is clearly nonlinear. In this case, to describe the functional form of this dependence, shown in Fig. 5, various nonlinear functions of the following types are considered:

$$y = C \cdot x^{n}, \tag{3}$$

$$y = C \cdot x^{x}, \tag{4}$$

$$y = C \cdot e^{nx}, \tag{5}$$

$$y = a + b \cdot \log(x), \tag{6}$$

where y is the recognition error, x is the number of announcers, a, b, C, and n are some empirical real numbers that are determined using regression analysis of the experimental data. Thus, using formulas (3) to (6), it is possible to determine the function that most accurately characterizes the dependence shown in Fig. 5. The most suitable function is selected based on the calculation of the error in determining the significant coefficients of the approximating functions using the experimental data, for example, for functions (3) to (5), the significant parameter is n, and for function (6), the parameter b. Table 2 gives errors in determining the significant parameters of the approximating functions for three types of neural networks.

As the data from Table 2 show, the best function that most accurately describes the dependence of the human voice recognition error on the number of announcers for all types of neural networks considered is the power function of form (3).

Fig. 6 shows an example of a plot of the approximating function (3) for a CNN network trained only in the Kazakh language and tested on statements also in the Kazakh language.

Table 2

Value of the error in determining the significant parameters of approximating functions

Network, function	$C{\cdot}x^n$	$C{\cdot}n^x$	$C{\cdot}e^{n\cdot x}$	$a+b{\cdot}\log(x)$
CNN	16 %	33 %	34 %	19 %
RNN	15 %	34 %	34 %	19 %
MLP	14 %	31 %	31 %	15 %

Table 3 gives the type of approximating function for the dependence of the accuracy of human voice recognition on the number of announcers for different types of neural networks.

Fig. 7, 8 show plots of approximating functions for RNN and MLP neural networks, constructed according to the data in Table 2.

In Fig. 6–8, the experimental data are indicated by dots, the red line shows the section of the approximation function corresponding to the range of experimental data (from 2 to 40 announcers), the dotted green part corresponds to the predicted values of the approximation function. And in Table 3, as an example, all the approximation functions found for all three types of neural networks when testing on statements in the Kazakh language are given.

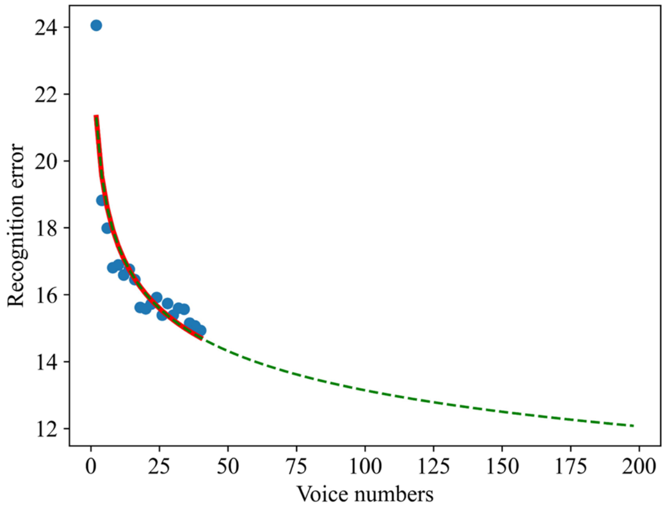

Fig. 6. Dependence of the total error of human voice recognition by the CNN network on the number of announcers

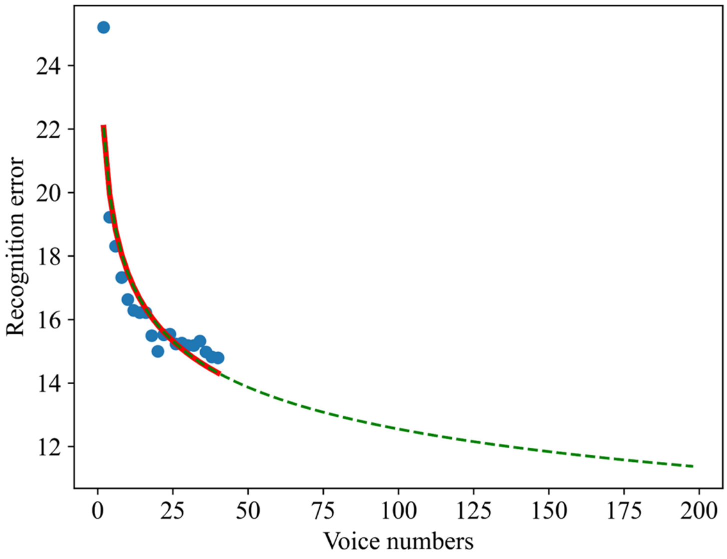

Fig. 7. Dependence of the total recognition error of the RNN network of human voice on the number of announcers

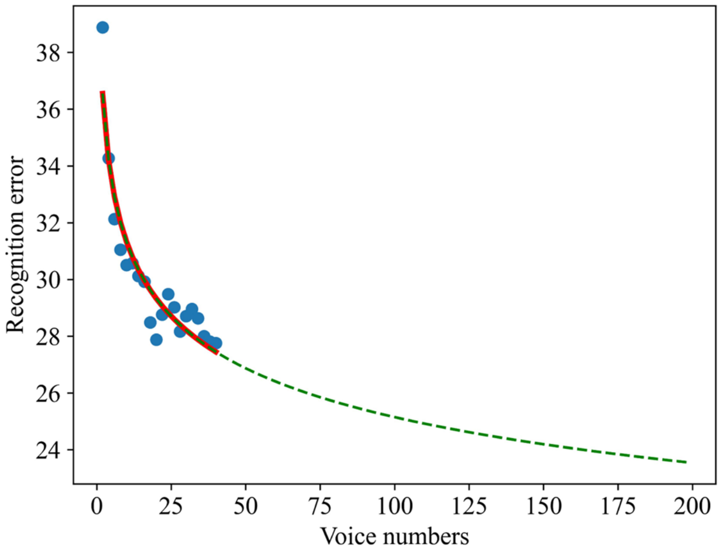

Fig. 8. Dependence of the total recognition error of the MLP network of human voice on the number of announcers

Table 3

Approximation function of the dependence of the accuracy of human voice recognition on the number of announcers for different types of neural networks

No.	Neural network	Approximation function
1	CNN	$y=23.2\cdot x^{-0.123}$
2	RNN	$y=24.36\cdot x^{-0.144}$
3	MLP	$y=39.05\cdot x^{-0.096}$

5. 2. Conducting a comparative analysis of the effectiveness of neural networks CNN, RNN, and MLP

A comparative analysis of the neural networks CNN, RNN, and MLP in the task of recognizing the human voice using approximating functions is shown in Fig. 9.

Fig. 9 provides an answer to the question of which of the neural networks CNN, RNN, or MLP gives the best results in recognizing the human voice. As we can see from Fig. 9, the best results are provided by the RNN neural network. Also, using the types of approximating functions from Table 3, we can approximately calculate how many different announcers should be used when training the neural network to achieve a certain accuracy of human voice recognition. For this purpose, we can use the following formula:

$$x = e^{\frac{\ln(y)-\ln(c)}{n}}, \tag{7}$$

where x is the required number of announcers, y is the required network recognition accuracy, c and n are the parameters of the approximating function. For example, for a network operation accuracy of 95 %, the value is y=5. And the values of the parameters c and n are taken from Table 3. Then, for the RNN network, if the required human voice recognition accuracy is at least 95 %, then according to the calculation using formula (7), it is necessary to use about 59,668 different announcers when training this network. The calculations also show that when using the same number of announcers during training, equal to 59,668, the CNN neural network gives a recognition accuracy of 94 %, i.e., slightly worse than the RNN network.

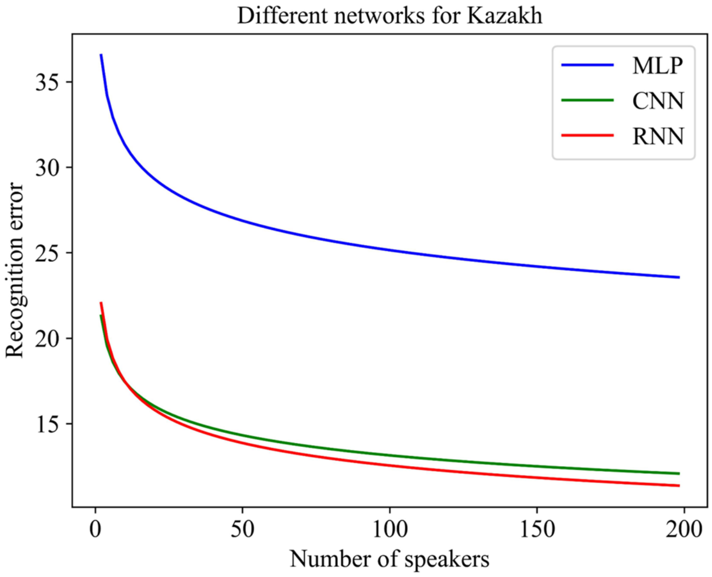

Fig. 9. Dependence of the total recognition error of the MLP network of human voice on the number of announcers

5. 3. Evaluation of the influence of language and the number of announcers on the accuracy of human voice recognition

After finding out that, all other things being equal, the human voice is best recognized by an RNN-type network, it is planned to consider the following question. Will a neural network trained on data from one language be able to recognize a human voice from data in another language with the same efficiency? To answer this question, neural networks trained only on data in the Kazakh language were tested using data in other languages. The result of this testing on an RNN-type network is shown in Fig. 10.

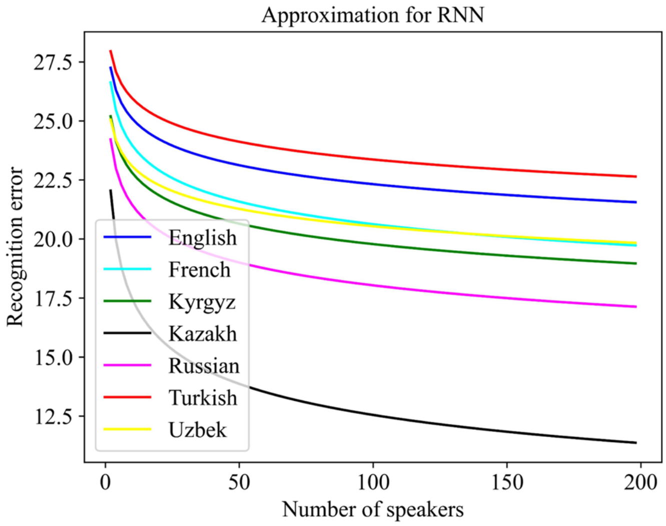

Fig. 10. Dependence of the accuracy of human voice recognition by a RNN network on the number of announcers, when training the network with data in the Kazakh language, but when testing in other languages

It should be noted that the dependence of the human voice recognition error in other languages, as in the case of the Kazakh language, has a power form. For example, this dependence for the Russian language takes the following form:

$$y = 25.53 \cdot x^{-0.075}. \tag{8}$$

As can be seen from Fig. 10, the neural network trained only on data in the Kazakh language recognizes human voices in other languages quite poorly. For example, the accuracy of human voice recognition in Russian for the RNN neural network trained using 59,668 announcers, according to formula (8), would have a value of 89 %. This is quite worse compared to 95 % accuracy for the Kazakh language, but not much. The accuracy values of human voice recognition by the RNN neural network for all other languages are given in Table 4.

Table 4

Estimated accuracy of human voice recognition in different languages by an RNN trained on the Kazakh language for 59,668 announcers

No.	Language	Accuracy of recognition
1	Kazakh	95.0 %
2	Russian	88.85 %
3	Kyrgyz	86.68 %
4	French	86.41 %
5	Uzbek	85.15 %
6	English	83.89 %
7	Turkish	82.57 %

As can be seen from Table 4, the accuracy of human voice recognition still significantly depends on the language if the network is trained in one language and tested on data from another language.

However, one big problem arises here. It is related to the fact that in order to achieve an accuracy of over 95 % for human voice recognition using an RNN, it is necessary to use voice samples of almost 60 thousand announcers. This will constitute too large a set of training data, the preparation of which requires enormous resources. But this problem can be significantly alleviated if we consider not the overall error of the network but only one of its parts, associated with the erroneous recognition of the section containing speech data as not a human voice (False Negative).

Indeed, for VAD systems, a sufficient and necessary condition is not to lose speech data when analyzing audio data. Then the neural network can be trained so that only the False Negative error is minimized. Fig. 11 shows the dependence of the False Negative error of RNN on the number of announcers when testing on Kazakh language data.

The dependence in Fig. 11, as in the previous cases, has a power-law character, determined from the following formula:

$$y = 34 \cdot x^{-0.244}. \tag{9}$$

In this case, in order to achieve 95 % accuracy of correct recognition of speech segments only, only 2582 samples of human voice are required. It can be seen that this is a much smaller number than is required when trying to provide 95 % accuracy for the total error. In Fig. 12, one plot shows the dependence of recognition errors for the total error and the False Negative error of the RNN network on the number of announcers when testing on Kazakh language data.

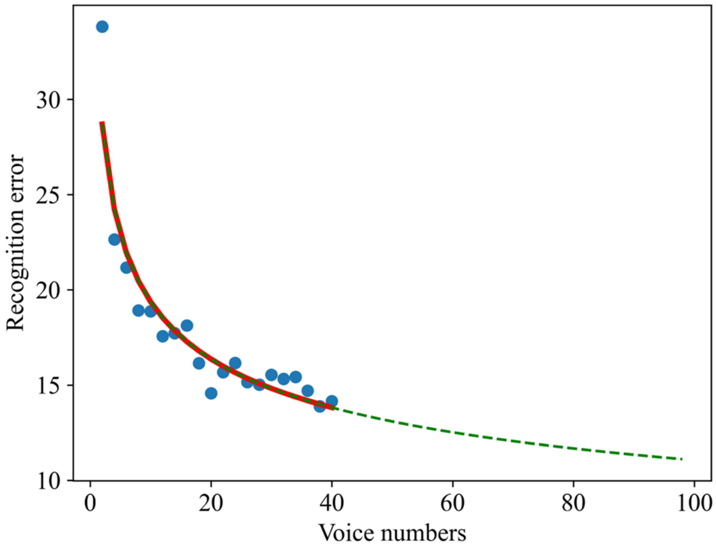

Fig. 11. Dependence of the False Negative recognition error type of the RNN network on the number of announcers when testing on Kazakh language data

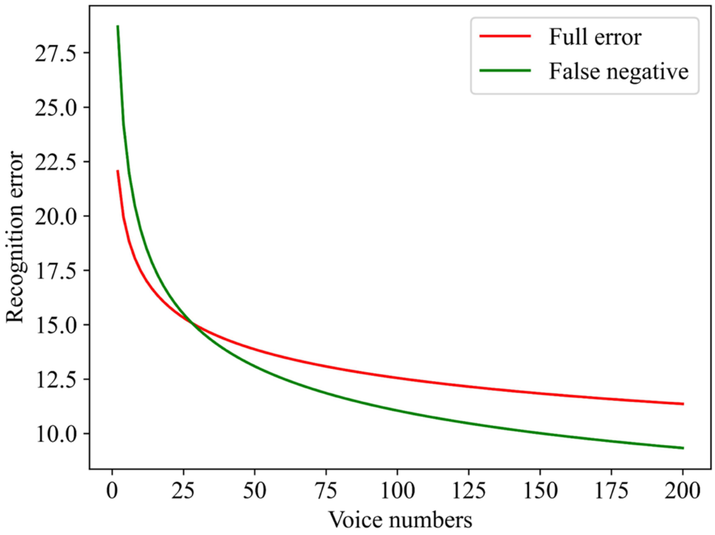

Fig. 12. Dependences of the total error and the False Negative error type of the RNN network on the number of announcers (test data in Kazakh)

The dependence plots in Fig. 12 are constructed using the corresponding approximation functions. As can be seen from Fig. 12, the False Negative error function decreases faster with an increase in the number of announcers used to generate the training data. This is understandable since with an increase in the number of announcers, sections of the audio signal containing human speech will be recognized more accurately. However, an increase in the number of announcers has a very weak effect on the accuracy of recognizing non-speech sections. For this reason, the total error containing both types of errors decreases more slowly with an increase in the number of announcers used in training. It should be noted that a neural network trained to minimize the total error can operate as an independent VAD system since it takes into account all types of errors. But a network focused on minimizing only the False Negative error cannot independently perform the functions of VAD systems in full. This is due to the fact that it does not actually control the errors associated with recognizing non-voice sections as voiced. In general, a neural network oriented towards minimizing the False Negative error can be used in combination with traditional VAD systems operating on the basis of analyzing the energy and spectral characteristics of the signal. The diagram of the possible operation of such a hybrid system is shown in Fig. 13.

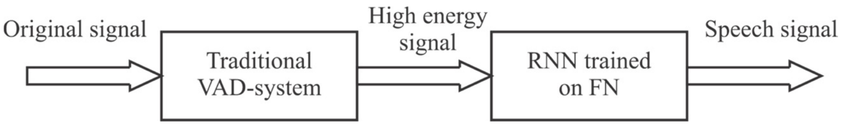

Fig. 13. Diagram of possible operation of a combined VAD system

According to the scheme shown in Fig. 13, a neural network of the RNN type, trained to minimize only one type of error, False Negative, is used as an additional element that improves the performance of traditional VAD systems.

6. Discussion of results based on the efficiency of neural networks for the task of human voice recognition

Unlike works [2–4], in which various neural network algorithms for implementing VAD, built within the framework of only one language model, were considered, the results obtained in this work include an assessment of the influence of the training language of neural networks on the accuracy of human voice recognition in other languages. The results, shown in Fig. 10 and Table 4, allow us to conclude that a neural network trained on data from only one language does not recognize human voices in other languages very effectively. For example, a neural network of the RNN type, trained only in the Kazakh language, provides recognition accuracy in the same language of 95 %, while the accuracy of recognizing the Turkish language is about 82 %. The difference is quite noticeable and in this case is equal to 13 % points.

In [6], it is proposed to use a hybrid convolutional-recurrent neural network to create a VAD system that is capable of providing very high accuracy of human voice recognition. However, in the same work, the authors emphasize the importance of further research to achieve similar results using a smaller number of announcers. In this work, we partially solved this problem by identifying empirical dependences of the accuracy of human voice recognition on the number of announcers, which made it possible to significantly reduce their number without losing the recognition quality. These found dependences are given in formula (8), in Table 3, and plots in Fig. 6–8. Calculations using these approximation functions showed that the best accuracy of human voice recognition is provided by an RNN type network, but not much better than CNN. For example, when using about 60 thousand announcers during training, the RNN network provides an accuracy of 95 %, and the CNN network 94 %, i.e., only 1 % higher. Calculations also showed that to obtain a purely neural network VAD system with an accuracy of over 95 %, training data using voice samples of about 60 thousand announcers is required. This is a lot. In this regard, unlike works [7–9], we considered a neural network that minimizes only one error associated with the erroneous recognition of a speech section as a non-human voice (False Negative). In this case, according to the plot in Fig. 12, to achieve an accuracy of over 95 % for False Negative, only 2582 announcers are required, which is almost 24 times less than when considering the total network error. It should be taken into account that this neural network does not focus on False Positive errors, but which can be cut off quite effectively by a traditional VAD system. Therefore, the solution may be the scheme shown in Fig. 12, which proposes to combine traditional VAD systems with a neural network solution that minimizes only the False Negative error. In this case, it is possible to achieve the same high accuracy of human voice recognition but with a much smaller number of announcers and volume of training data.

It is quite possible that when using other neural network structures than those considered in this paper, the types of approximating functions may be different. This may also be greatly influenced by the data language used to train these neural networks. Therefore, it cannot be said that the proposed solutions are universal. But they can provide guidance in designing VAD systems based on neural networks.

7. Conclusions

1. A suitable activation function for describing dependences in neural networks of the CNN, RNN, and MLP types has been determined. The activation function is selected based on the calculation of the error in determining the significant coefficients of the approximating functions based on experimental data, where the significant parameter is some empirical real numbers and the signal-to-noise ratio value. Thus, by determining the activation function, not the overall network error is considered but only one of its parts associated with the erroneous recognition of the area containing speech data (False Negative).

2. Based on our calculations, a comparative analysis of the effectiveness of CNN, RNN, and MLP neural networks in human voice recognition tasks was carried out using function approximation methods. The study showed that the best results are provided by the RNN neural network. Based on the approximating functions, the required number of announcers was calculated to achieve a certain accuracy of human voice recognition. Calculations have shown that to achieve at least 95 % accuracy in human voice recognition, it is necessary to use approximately 59,668 different announcers when training this network.

3. The impact of language and the number of announcers on the accuracy of human voice recognition has been assessed. It was found that a neural network trained on data from only one language is not very effective in recognizing human voices in other languages. For example, if an RNN-type neural network trained on only the Kazakh language provides 95 % accuracy when recognizing a human voice in Kazakh, then the same network provides about 82 % accuracy for the Turkish language. The difference is quite significant and in this case is equal to 13 percentage points. This value can be used as a measure of the phonetic proximity of two languages. The smaller this difference, the more similar these languages sound at the phonemic level.

Also, despite the fact that the Kyrgyz, Uzbek, and Turkish languages, together with the Kazakh language, belong to the same Turkic group, the closest to the Kazakh language in phonetics (in sound) is the Russian language, which does not belong to the Turkic group of languages. This is quite possibly due to the fact that the Russian language is very widespread in the Republic of Kazakhstan and virtually all the invited announcers also speak Russian fluently. Apparently, this creates a somewhat similar phonetic background for the Kazakh and Russian languages in the Republic of Kazakhstan. For such neural networks, the meaning of what is said is not important, but what is most important is how the voices sound phonetically. This approach, when the neural network is trained in one language, but tested in another language, could be used to study the phonetic proximity of different languages.

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, as well as the results reported in this paper.

Funding

This research was 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

The data will be provided upon reasonable request.

Use of artificial intelligence

The authors used artificial intelligence technologies within acceptable limits to conduct their own verified research, which are described in the "Research Methodology" chapter.

References

1. Dhouib, A., Othman, A., El Ghoul, O., Khribi, M. K., Al Sinani, A. (2022). Arabic Automatic Speech Recognition: A Systematic Literature Review. Applied Sciences, 12 (17), 8898. https://doi.org/10.3390/app12178898
2. 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
3. Tang, M., Huang, H., Zhang, W., He, L. (2024). Phase Continuity-Aware Self-Attentive Recurrent Network with Adaptive Feature Selection for Robust VAD. ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 11506–11510. https://doi.org/10.1109/icassp48485.2024.10446084
4. Cherukuru, P., Mustafa, M. B. (2024). CNN-based noise reduction for multi-channel speech enhancement system with discrete wavelet transform (DWT) preprocessing. PeerJ Computer Science, 10, e1901. https://doi.org/10.7717/peerj-cs.1901
5. Priebe, D., Ghani, B., Stowell, D. (2024). Efficient Speech Detection in Environmental Audio Using Acoustic Recognition and Knowledge Distillation. Sensors, 24 (7), 2046. https://doi.org/10.3390/s24072046
6. Tan, Y., Ding, X. (2024). Heterogeneous Convolutional Recurrent Neural Network with Attention Mechanism and Feature Aggregation for Voice Activity Detection. APSIPA Transactions on Signal and Information Processing, 13 (1). https://doi.org/10.1561/116.00000158
7. 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
8. Rho, D., Park, J., Ko, J. H. (2022). NAS-VAD: Neural Architecture Search for Voice Activity Detection. Interspeech 2022, 3754–3758. https://doi.org/10.21437/interspeech.2022-975
9. 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. https://doi.org/10.15587/1729-4061.2024.298687
10. Liu, P., Wang, Z. (2004). Voice activity detection using visual information. 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1, I-609–612. https://doi.org/10.1109/icassp.2004.1326059