DOI: 10.37943/20TTRV6747

Aigul Kulakayeva
PhD, Associate Professor, Department of Radio Engineering and Electronics
a.kulakayeva@iitu.edu.kz, orcid.org/0000-0002-0143-085X
International Information Technology University, Kazakhstan

Valery Tikhvinskiy
Candidate of Technical Sciences, Professor, Department of Radio Engineering and Electronics
vtniir@mail.ru, orcid.org/0000-0002-3443-5171
International Information Technology University, Kazakhstan

Aigul Nurlankyzy
PhD candidate, Department of Electronics, Telecommunications and Space Technologies
nurlankyzyaigulya@gmail.com, orcid.org/0000-0002-0791-8573
Satbayev University, Kazakhstan
Senior Lecturer, Department of Space Engineering
Energo University, Kazakhstan

Timur Namazbayev
Senior lecturer, Department of Solid State Physics and Nonlinear Physics
timur.namazbayev@gmail.com, orcid.org/0000-0002-2389-2262
Al-Farabi Kazakh National University, Kazakhstan

COMPARATIVE ANALYSIS OF THE EFFECTIVENESS OF NEURAL NETWORKS AT DIFFERENT VALUES OF THE SNR RATIO

Abstract: This work is devoted to a comparative analysis of the effectiveness of neural networks, CNN and RNN, at different SNR ratios. The research conducted within the framework of this work showed that CNN convolutional neural networks demonstrate higher efficiency in speech signal recognition tasks, regardless of different levels of SNR ratio and language. Thus, the CNN neural network showed stable superiority over RNN under all conditions, especially at low SNR ratios. It was revealed that with an increase in the SNR ratio, the difference in accuracy between the CNN and RNN neural networks decreases, but the CNN continues to lead, which indicates its higher adaptability and ability to learn under conditions of different noise and interference levels. It is especially important to note that the advantage of CNN becomes noticeable at low SNR values, where the accuracy of the RNN decreases more significantly. As a result, with an SNR ratio of 3 dB, the recognition accuracy using CNN was 80% for the Kazakh language, whereas RNN showed a result in the region of 75%. With an increase in the SNR ratio to 21 dB, the difference in accuracy between CNN and RNN decreased, but CNN continued to lead, reaching 88% accuracy compared to 86% for RNN. In addition, the results showed that the effectiveness of the CNN and RNN depended on the language in which they were trained. Neural networks trained in Kazakh showed the best results in recognizing Kazakh speech but also successfully coped with recognizing the Russian language. This highlights the importance of considering language features when developing and training neural networks to improve their performance in multilingual environments.

Received: 29.08.2024   Accepted: 17.10.2024   Published: 30.10.2024

Keywords: artificial neural networks (ANN); convolutional neural network (CNN); recurrent neural network (RNN); voice activity detector (VAD); signal-to-noise ratio

Introduction

Speech is the main means of human communication and plays an important role in interactions between people. In recent years, there has been growing interest in the use of speech technologies, which may prove to be more effective than traditional methods of entering information [1]. This interest has become the basis for active research in the field of automatic speech recognition (ASR). Nevertheless, the correct operation of ASR in a noisy environment remains an urgent problem because background noise and environmental distortions can significantly reduce the accuracy of speech signal recognition [2]. The problem of low speech recognition performance in a noisy environment is mainly caused by the discrepancy between the training and testing conditions of the ASR systems. One of the key components used to improve the accuracy of ASR systems in noisy environments is the Voice Activity Detector (VAD) [3]. VAD is responsible for separating speech segments from non-speech segments in the audio signal, which allows the ASR system to focus solely on speech data, ignoring noise. This is especially important in conditions of low SNR, where background noise can interfere with accurate speech signal recognition. Proper functioning of the VAD not only reduces computational costs but also improves recognition accuracy by removing unnecessary non-speech fragments.

Noise reduction and distortion removal remain the most important tasks in speech-recognition systems. Different SNR levels can significantly distort speech signals, leading to a decrease in their recognition accuracy. In this regard, the use of VAD with neural network architectures is important for solving the problem of low ASR performance under adverse conditions [4].

In [5], a procedure was proposed for accurately determining the presence of a VAD voice, which included three stages: signal preprocessing, generation of derived variables, and application of VAD using classification and smoothing techniques based on machine learning. The experiments were conducted on the basis of datasets including verbal interaction of 15 groups, 3 people each. The results demonstrated high reliability in extracting speech signals. However, it is important to note that these studies were conducted without considering the effects of noise and extraneous sounds, which may limit the overall applicability of the results.

In [6], an algorithm for noise elimination was proposed, aimed at improving the quality of speech data in the Kannada language, which was integrated into a system of oral queries in real time at the stage of acoustic feature extraction. The system uses an automatic speech recognition model with a minimum level of word errors, tested on 500 farmers/speakers from Karnataka under uncontrolled conditions. By combining a noise reduction algorithm with a time-delayed neural network, this study achieved a relative reduction in the level of errors in words of 1.59% compared to the previous version of the oral query system. However, to obtain such results, it is necessary to involve a large number of speakers, which may limit the possibility of using the algorithm in situations that lack data.

In [7], a study was conducted on the performance of the VAD algorithm under conditions of background noise, that is, environmental noise. For the analysis, 25 statements taken from the noisy AURORA sample were used, containing the speech of 13 men and 16 women. The authors also emphasize the need for further research involving more speakers. In addition, the amount of training data has a significant impact on system performance, which indicates the importance of increasing the amount of available data to improve the accuracy of the algorithm.

In [8], a system was proposed for creating speech datasets with noise for several languages. These datasets are used to train and test neural models to improve speech quality using loss functions based on self-supervised speech representation (SSSR). The results of this study showed that the audio recording language used in training had a minimal impact on system performance. However, it is emphasized separately that the amount of training data for a particular language significantly affects the effectiveness of the model, which indicates the importance of having enough data to achieve high results.

In [9], a real-time study of a VAD based on a neural network was conducted. For the analysis, a dataset consisting of a recording of a meeting with 14 speakers was used, which created conditions for overlapping speech and other complex scenarios. This study focuses on evaluating the effectiveness of the proposed approach in determining the presence of voice in such difficult acoustic conditions, which is an important task for improving ASR systems. Thus, the results of this study can contribute to further development of technologies related to real-time speech processing.

In [10], a study of VAD performance based on singing voice was conducted. The songs of 40 singers presented in the five languages were used as the dataset. Despite the vastness of the study, certain problems were related to invisible singers, the use of communication codecs, differences in languages, and specific musical contexts. Such factors have had an impact on the accuracy and reliability of the VAD operation in the context of music recordings, which underscores the need for further research and improvements in this area. Thus, the results may be useful for optimizing speech and music-processing algorithms in the future.

In [11], a method for detecting voice activity using ultra-wideband radar (UWB) was proposed. For the analysis, a dataset containing recordings of the voices of 12 announcers representing various geographical regions, including the United Kingdom, China, Pakistan, and other countries, was used. The speakers were chosen on the basis of their unique accents, physical characteristics, and pronunciation styles associated with their ethnicity. As a result of the study, high performance of the method was achieved, but the experiments were carried out under ideal recording conditions, while noise and extraneous sounds were not taken into account. These limitations indicate the need for further research to verify the stability and accuracy of this method under more complex and noisy acoustic conditions.

In [12], a speech recognition system in the Moroccan dialect was presented under various conditions of additional noise. The ten most frequently used greetings in the Moroccan dialect, extracted from telephone conversations, were selected. This corpus was recorded by 60 speakers (30 men and 30 women). Each speaker pronounced each expression three times under both natural and noisy conditions. Although automatic speech recognition systems demonstrate high performance in the absence of noise, their effectiveness is significantly reduced in the presence of noise.

As a result of the conducted research analysis, it was found that currently there are an insufficient number of studies on the effect of noise on ASR systems. Most of the existing work focuses on improving pure speech recognition algorithms, whereas the impact of various types of noise and acoustic distortion on ASR performance still requires more in-depth study. A variety of scenarios, such as urban noise and other complex acoustic conditions, are insufficiently covered, which may limit the possibility of using ASR in real conditions. Given the growing need for efficient ASR systems for various applications, it is important to conduct additional research aimed at adapting and improving ASR systems to work under conditions of interference and noise, that is, different levels of SNR.

Thus, in this paper, the objects of research are the neural networks CNN and RNN, which are widely used for analyzing speech signals. The main hypothesis of this study is that adding noise to the training data with different SNR ratios may affect the accuracy of speech signal recognition.

The scientific problem of this study is to analyze the influence of different levels of SNR ratio on the efficiency of the neural networks CNN and RNN. This study aims to identify how changing the SNR level affects the performance of these two types of neural networks to determine their abilities and limitations in conditions of different noise levels.

The purpose of this study is to determine which of the presented neural network architectures will show high noise immunity and accuracy in speech signal recognition in the presence of various noise levels. To achieve this goal, it is planned to conduct a number of experiments in which the characteristics of CNN and RNN architectures under the influence of noise of various levels are evaluated, which will reveal their advantages and disadvantages in processing and recognizing speech signals.

The following tasks were set up in this study: to analyze the effect of different levels of SNR on the accuracy of speech signal recognition using the neural networks CNN and RNN; to compare the adaptability and efficiency of CNN and RNN neural network architectures under different noise conditions in different languages; to conduct a comparative analysis and study the impact of different levels of SNR on the performance of CNN and RNN; evaluate how language features and the number of speakers affect the performance of trained neural networks in different languages.

Methods and Materials

This study is a continuation of [13], where it was hypothesized that neural networks trained in a particular language can effectively recognize human voices in other languages due to the presence of common phonemes. It was also noted that to achieve acceptable accuracy of speech signal recognition, a limited number of speakers can be used while maintaining parity between male and female voices. The study was conducted on datasets representing the statements of speakers in the Kazakh language, taking into account certain requirements, such as an equal ratio of the amount of data and the parameters of neural networks. In addition, the calculations given in [13] showed that to achieve a human voice recognition error of no more than 3% (accuracy of at least 97%), it is necessary to use at least 4.5 thousand speakers during the training of neural networks. As a result, CNN demonstrated better speech signal recognition quality than RNN. For example, when learning on 20 speakers, the recognition error for CNN was 10.6%, whereas for RNN, it was 11.9%. Using 80 announcers, the results were 8.3% for the CNN and 8.93% for the RNN.

In this study, the same speech corpus was used as in [13]. Datasets from the Institute of Smart Systems and Artificial Intelligence (ISSAI) of Nazarbayev University were used to conduct training and testing of neural networks, namely the Kazakh speech corpus [14], Russian speech corpus [15], Turkish language corpus [16], and Uzbek language corpus language [17]. One of the largest open datasets, the Common Voice Dataset [18], was also used, namely, the corpus of the Kyrgyz language, corpus of the English language, and corpus of the French language. From each dataset, 20 male and 20 female voices were selected in a special way so that the voices were of different intonation, pitch, age, etc.

The speakers of the Kazakh language corpus were in the sound studio , "Nazarbayev University Research and Innovation System". The studio had a high degree of sound insulation and was equipped with the necessary sound equipment. The digitization of speech was carried out using an external LEXICON I-ONIX U82S sound card, whose highly efficient DACs/ADCs guarantee a pure 24-bit/96 kHz sound to transmit every smallest nuance of speech. To maintain a balance between recording quality and hard disk space, the following recording characteristics were used: sampling rate – 16 kHz, sample size, 16 bits; and compression algorithm, linear

PCM. The average recording time for each speaker was 40-45 minutes, and the average number of recording sessions per week was 14-15 sessions. The acoustic data were grouped into a separate kazspeech database with a total volume of 8 GB or approximately 25 h of speech.

One of the largest databases of audio fragments, samples, recordings, and audio signals was used as the noise source, FREESOUND.ORG [19], released under Creative Commons licenses. 10 types of noise were used, such as the sounds of electrical appliances, rain, birdsong, autobahns, airplanes and stadiums.

In this work, manual markup of audio files of the Kazakh speech corpus was carried out [14] using the Audacity 3.4.2 software product. The audio recording area is shown in Figure 1. file, where 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.

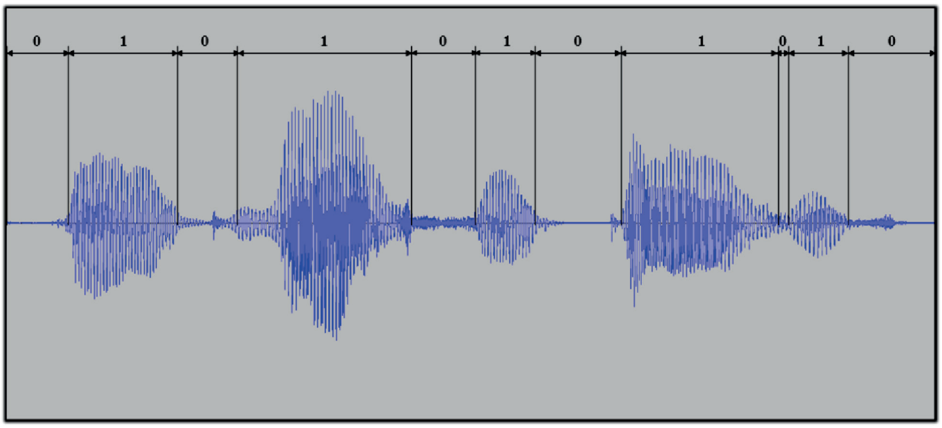

Figure 1. Manual markup of audio data

Twelve MFCC coefficients obtained from audio with noise addition were used as the input data. These three coefficients were combined by 3 and submitted to the input in the form of a two-dimensional array. The task was a binary classification: the allocation of areas with and without speech, where the labels were values 1 and 0, indicating the presence and absence of speech, respectively. An Accuracy metric was selected to evaluate the effectiveness of the model. This choice was justified because the task focused on binary classification, and in our case, the classes were balanced, which makes the accuracy sufficient to assess the quality of classification.

Metrics such as Speed Factor (SF) and Word Error Rate (BER) were not used because the goal was not to isolate individual words from speech but to determine the presence or absence of speech fragments. This makes the use of these metrics impractical in our task.

The *Adam* optimizer was used to train the CNN, and the losses were calculated using the *Binary Crossentropy* function. The *Mean squared error* loss function was used for the RNN. During training, functions were used to stop learning early (*EarlyStopping*) and save the best model (*ModelCheckpoint*). EarlyStopping stopped learning when the value of “val_accuracy” did not improve for 10 epochs. ModelCheckpoint resaves the maximum value of “val_accuracy” every time. Fig. 2 and Fig. 3 show the algorithms of the neural networks, CNN and RNN.

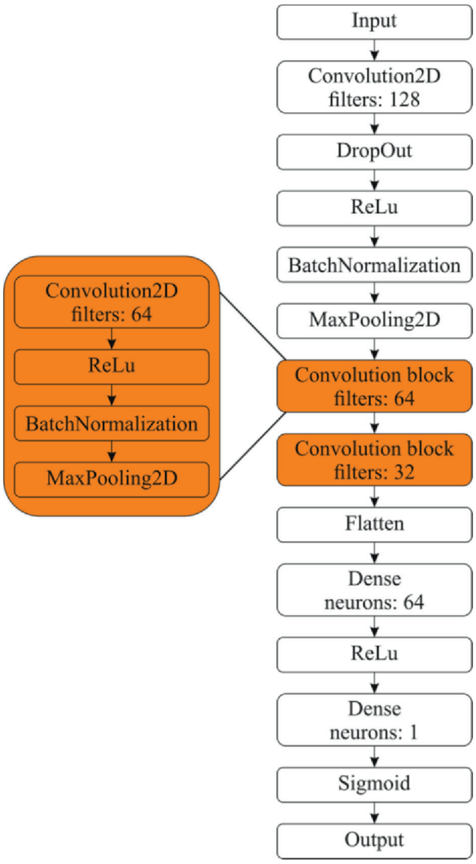

Figure 2. Structure of CNN networks

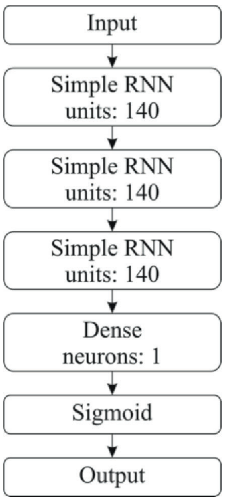

Figure 3. Structure of RNN networks.

The main difference between this study and the previous one [13] is the analysis of the effect of different noise levels on the accuracy of the speech signal recognition. Within the framework of this study, experiments were conducted with different levels of SNR ratio (from 3 dB to 21 dB), which is allowed to identify how the performance of CNN and RNN neural networks changes under background noise conditions. This methodology allows us to study the stability of neural network architectures of CNN and RNN in real environmental conditions, when noise can significantly distort speech signals.

Thus, in this study, the focus is on studying the effect of different levels of SNR ratio on the performance of the neural networks CNN and RNN. In a previous study [13], the main task was to compare the accuracy of speech signal recognition based on the number of speakers and languages; in this work, a new parameter was added – noise load, that is, different levels of background noise.

Thus, the objects of research in this study are artificial neural networks CNN and RNN. The main hypothesis of the study is how adding noise to training data with different levels of SNR ratio affects the accuracy of speech signal recognition by CNN and RNN neural networks and to determine which neural network architecture is more noise-resistant.

Results

This study, which, as previously noted, is a logical continuation of the study [13], where the analysis was expanded by including various levels of the SNR ratio in order to study the effect of noise on the performance of neural networks CNN and RNN. As a result, 40 speakers were used in this study to train and test neural networks, such as CNN and RNN. The recordings of the male and female announcers were used for training. Each speaker had its own characteristics, such as intonation, pitch, age, and other speech characteristics, which made the dataset for training the neural networks diverse. CNN and RNN neural networks were trained exclusively in the Kazakhs, which allowed them to focus on the specifics of this language and its features in speech recognition. The effectiveness of the trained models was evaluated in other languages such as Russian, Uzbek, Kyrgyz, Turkish, English, and French. As part of this study, comprehensive experiments were conducted to evaluate the effectiveness of various CNN and RNN architectures. The main purpose of this study was to determine how differences in the architecture of neural networks and different levels of signal-to-noise ratio (SNR) affect the accuracy of speech recognition. The results obtained will allow for a deeper understanding of the impact of these factors on the performance of speech recognition systems in a multilingual environment.

Results for SNR ratio = 3 dB. With an SNR ratio of 3 dB, the signal transmission conditions are extremely unfavorable, which significantly complicates the task of speech recognition. At this noise level, the CNN demonstrated a clear advantage over the RNN in all languages. For example, for the Kazakh language, the recognition accuracy using CNN was approximately 80%, whereas RNN showed a result of approximately 75%. A similar pattern was observed in Russia, where CNN reached 77%, while RNN reached about 73-74%. These data are shown in Fig. 4, which shows the accuracy of the speech recognition at SNR = 3 dB.

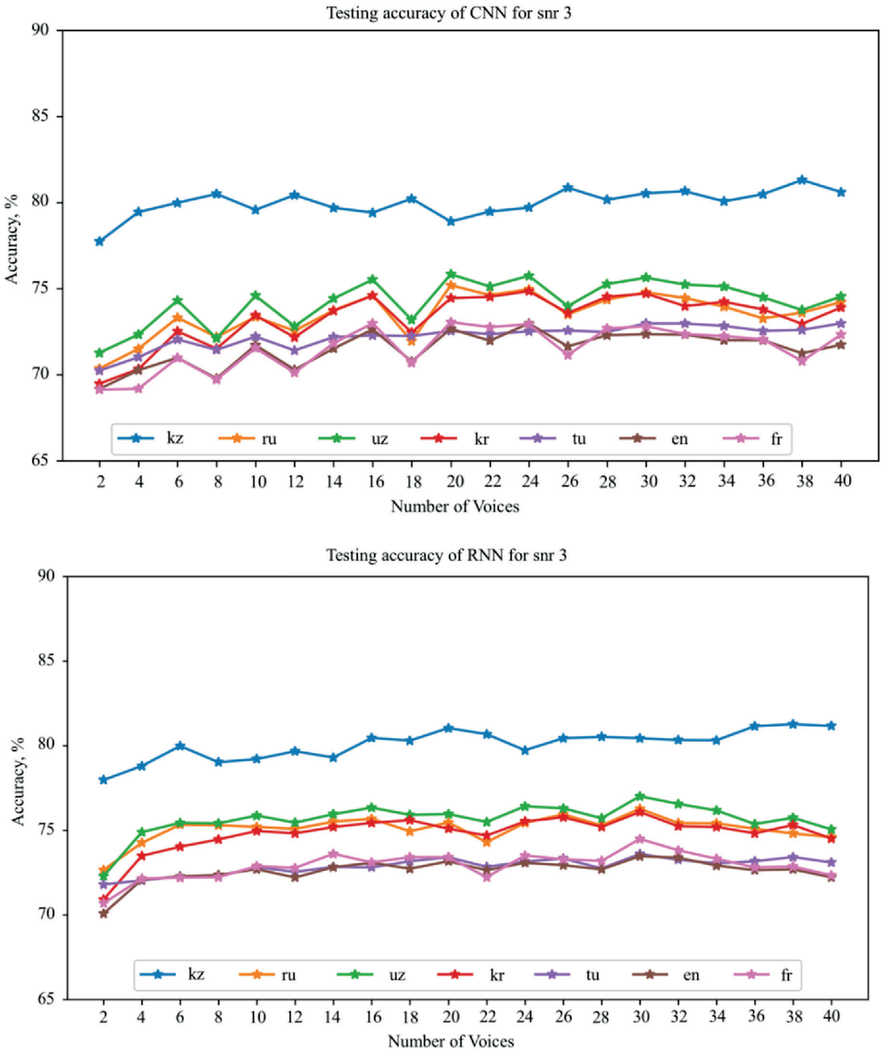

Figure 4. The accuracy of speech recognition by neural networks CNN and RNN at a ratio of SNR = 3 dB depends on the number of speakers in different languages.

It is also interesting to note that in Uzbek and Kyrgyz, the gap between the results of the CNN and RNN was more noticeable. Thus, the accuracy of CNN recognition was 4-5% higher than that of RNN. In French, the difference was less significant, but CNN still showed better recognition (approximately 70than RNN to (67%). Turkish, despite its linguistic proximity to Kazakhs, proved to be the most difficult for both networks, where CNN also showed a small but stable advantage.

Results for SNR ratio = 9 dB. Increasing the SNR to 9 dB makes the signal transmission conditions less noisy, which improves the efficiency of the neural networks. As a result, for the Kazakh language, the recognition accuracy when using CNN increased to 83%, whereas RNN reached 80%. CNN also performed better in Russian and Uzbek, outperforming RNN by 2-3%. The Kyrgyz language, as before, showed lower results for the RNN, which was 3-4% inferior to the CNN%. These data are shown in Fig. 5, which shows the accuracy of the speech recognition at SNR = 9 dB.

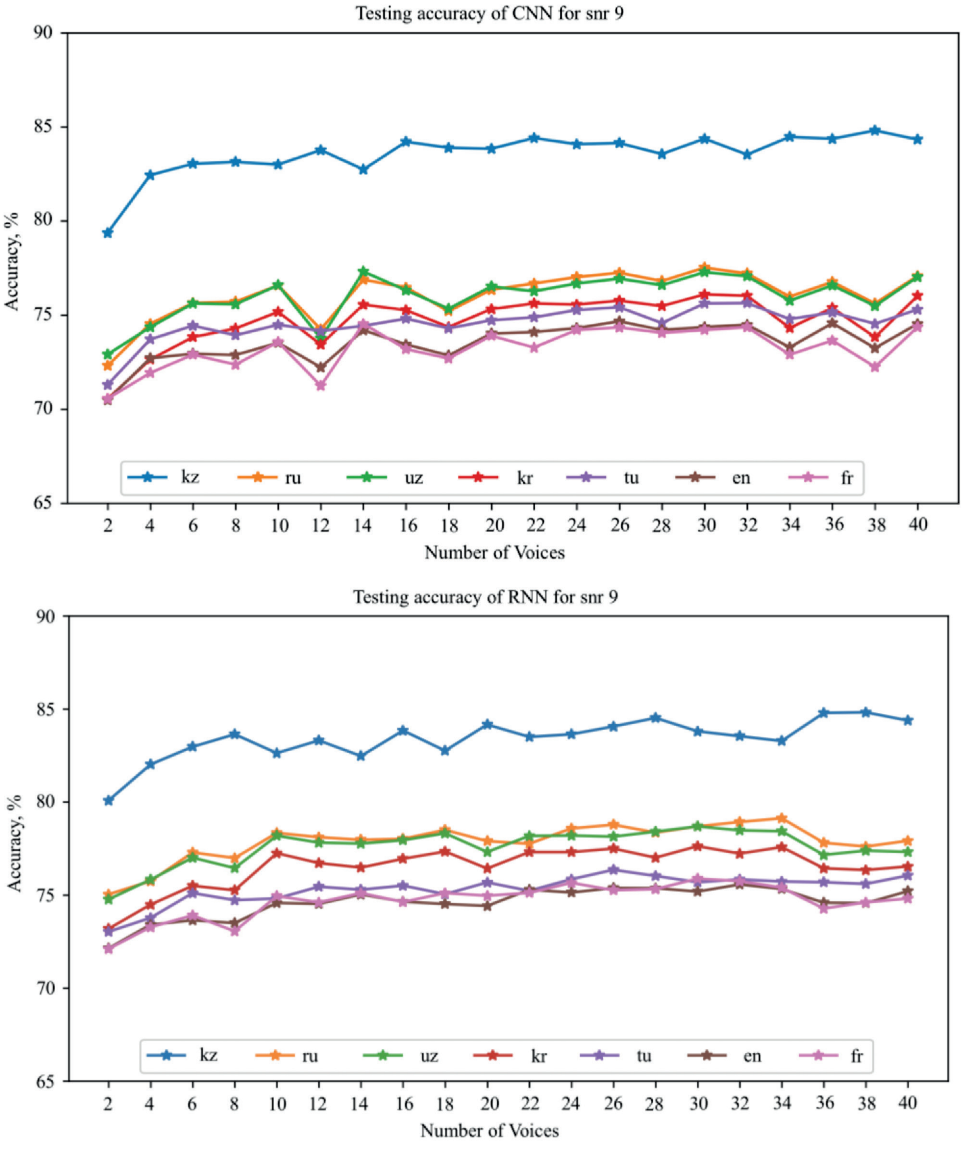

Figure 5. The accuracy of speech recognition by neural networks CNN and RNN at a ratio of SNR = 9 dB depends on the number of speakers in different languages.

In France, the difference between the results of CNN and RNN became less pronounced, but CNN continued to lead, which confirms its resistance to medium-noise conditions.

Results for SNR ratio = 15 dB. With a further increase in the SNR to 15 dB, the conditions for speech recognition become even more favorable. With this SNR ratio, the results in Kazakh for CNN reached an accuracy of 86%, whereas RNN showed a result of approximately 84%. In Russia, CNN showed an accuracy of approximately 81%, which is 2% higher than that of RNN. In Turkish, the gap between the networks has been reduced to 1-2%, which indicates the complexity of this task for both networks, even under conditions of high signal strength. However, the CNN continued to show better recognition than the RNN. These data are shown in Fig. 6, which shows the accuracy of the speech recognition at SNR = 15 dB.

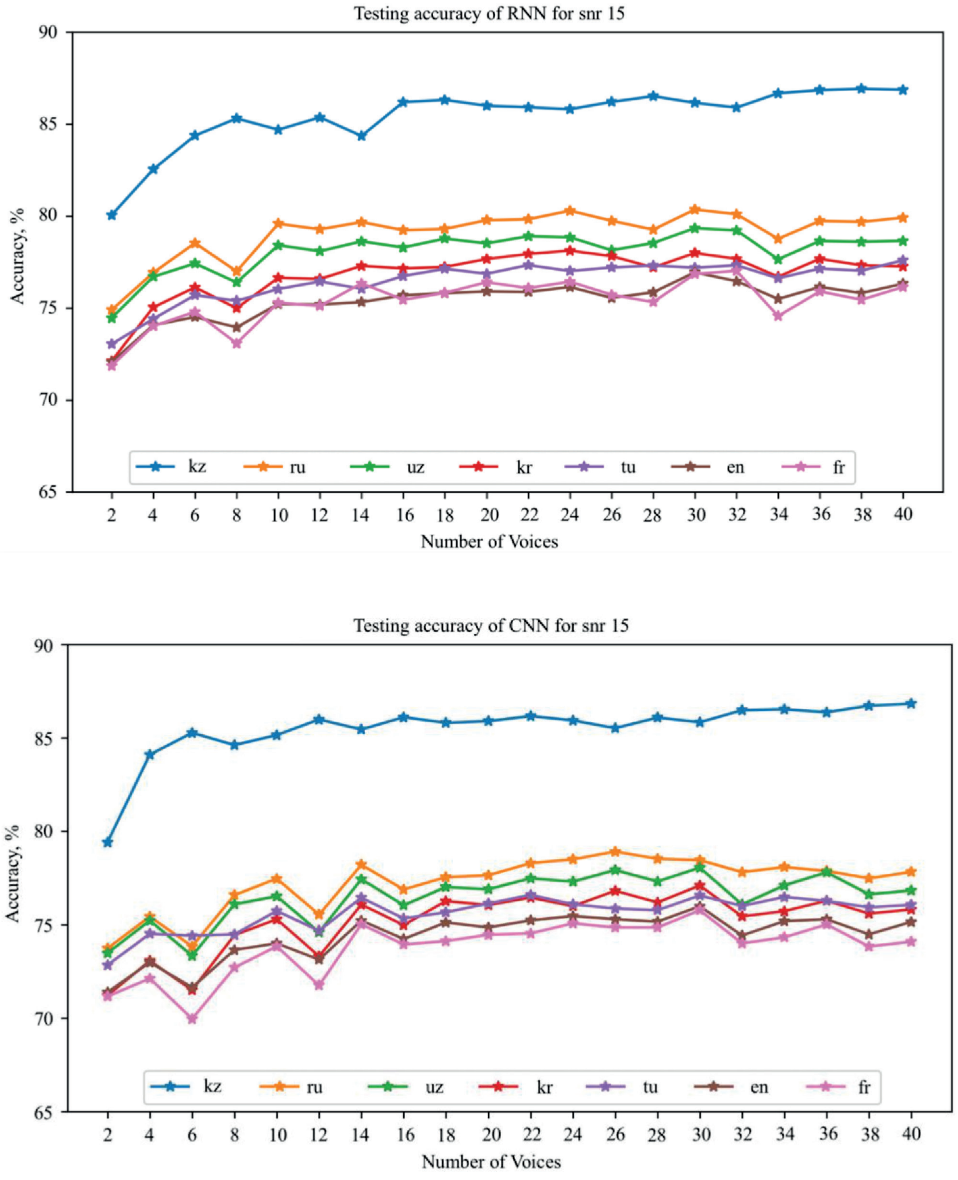

Figure 6. The accuracy of speech recognition by neural networks CNN and RNN at a ratio of SNR = 15 dB depends on the number of speakers in different languages.

Results for SNR ratio = 21 dB. With a further increase in the SNR ratio to 21 dB, the signal became almost pure. Using this ratio, the CNN and RNN neural networks demonstrated consistently high results. The CNN achieved a maximum accuracy of 88% for the Kazakh language, while the RNN showed a result in the region of 86%. In Russia and Uzbek, the difference between the networks was less than 2%, but the CNN still showed the best results. These data are shown in Fig. 7, which shows the accuracy of the speech recognition at SNR = 21 dB.

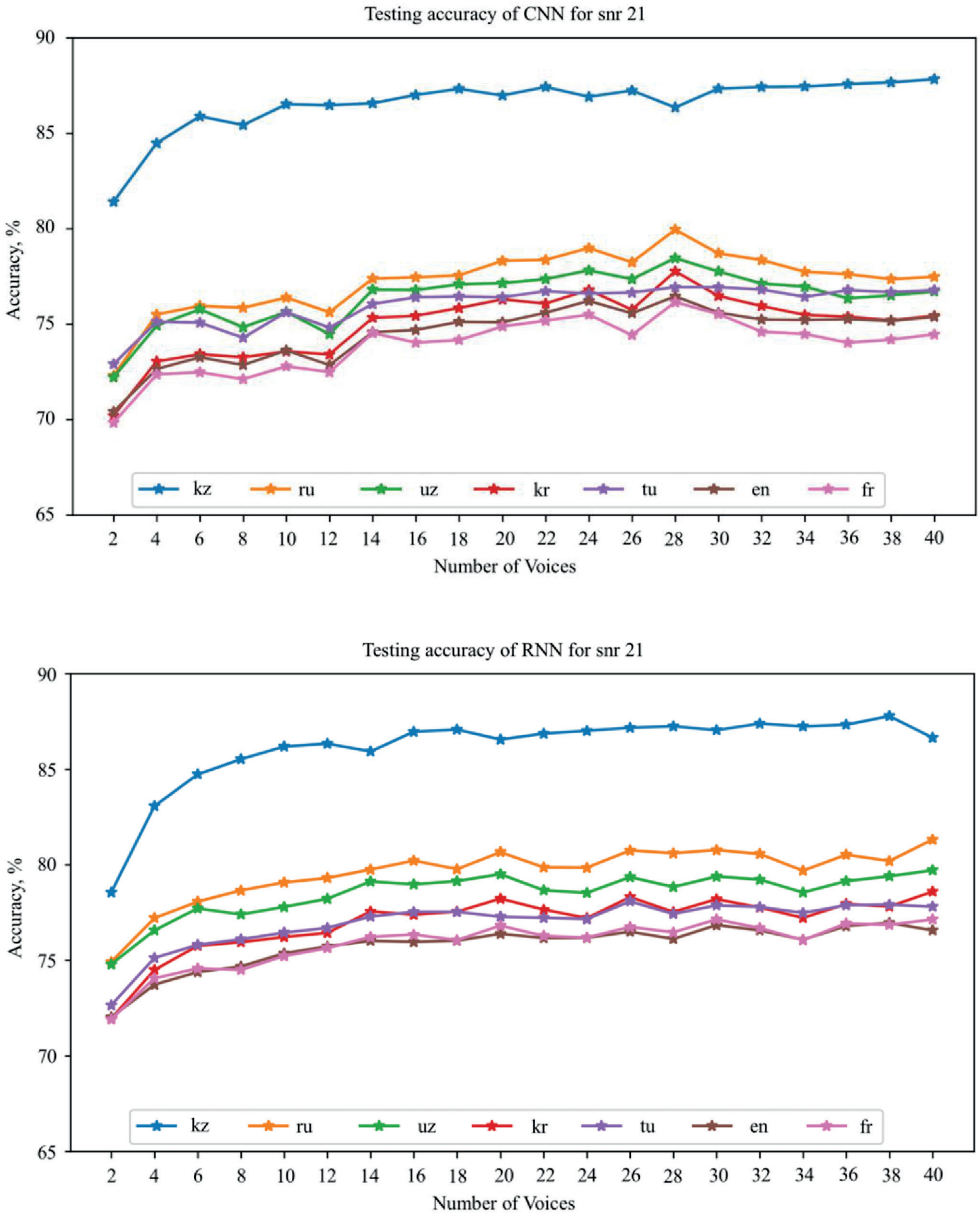

Figure 7. The accuracy of speech recognition by neural networks CNN and RNN at a ratio of SNR = 21 dB depends on the number of speakers in different languages.

It is also worth noting that in French, the gap between the networks became insignificant, but CNN continued to show better results, which confirms its higher stability and efficiency. In Turkish, both networks showed similar results, which is probably due to the peculiarities of the phonetics of this language, which do not provide a clear advantage to one architecture over the other.

In addition, a comparative analysis of the performance of CNN and RNN at different SNR values were allowed to identify important differences in their effectiveness. The results obtained during testing and recognition of speech signals in various languages are presented in Table 1.

Table 1. Comparative analysis of CNN and RNN performance at different SNR levels

Language	SNR=3 dB		SNR=9 dB		SNR=15 dB		SNR=21 dB	
	CNN	RNN	CNN	RNN	CNN	RNN	CNN	RNN
Kazakh	85%	80%	83%	80%	86%	84%	88%	86%
Russian	77%	73%	85%	82%	85%	83%	87%	85%
Uzbek	75%	70%	78%	76%	82%	80%	84%	82%
Kyrgyz	74%	70%	76%	74%	78%	76%	80%	78%
Turkish	67%	62%	70%	68%	72%	70%	74%	72%
English	69%	65%	71%	69%	73%	71%	75%	73%
French	70%	67%	72%	70%	74%	72%	76%	74%

Table 2 shows the difference in the performance of the CNN and RNN neural networks at different SNR ratios for different languages.

Table 2. Difference in performance of CNN and RNN networks

Language	Δ at SNR 3dB	Δ at SNR 9dB	Δ at SNR 15dB	Δ at SNR 21dB
Kazakh	5%	3%	2%	2%
Russian	4%	3%	2%	2%
Uzbek	5%	2%	2%	2%
Kyrgyz	4%	2%	2%	2%
Turkish	5%	2%	2%	2%
English	4%	2%	2%	2%
French	3%	2%	2%	2%

The difference in percentage points between the accuracy of CNN and RNN for each language and each SNR value was calculated using the following formula:

$$\Delta = CNN - RNN \tag{1}$$

Thus, it can be argued that the CNN neural network shows higher accuracy than the RNN at all levels of SNR and for all the languages studied. A particularly noticeable advantage of CNN is manifested at low SNR levels, which indicates the higher resistance of this architecture to noise and interference. It can also be argued that the performance of neural networks depends on the language of the instruction and testing.

Discussion

The analysis of the results showed that the architecture of the CNN provides higher accuracy of speech signal recognition compared to the RNN for all values of the SNR ratio and for all tested languages. This is particularly noticeable at a low SNR ratio when the signal transmission conditions are the least favorable. Thus, with an SNR ratio of 3 dB, the difference in recognition accuracy between the CNN and RNN reached approximately 5%, which indicates a better ability of the CNN to adapt to adverse conditions. This result is explained by the fact that the CNN neural network is able to extract more stable features from noisy data, which makes it more effective in conditions of a low SNR ratio.

As previously noted, with an increase in the SNR ratio, both networks showed good results. With a ratio of SNR = 9 dB and 15 dB in Kazakhs and Russians, respectively, CNN showed an accuracy exceeding RNN by approximately 3%. This indicates that the CNN copes more successfully with the task of speech signal recognition with an increase in the amount of training data and improved signal transmission conditions.

In addition, with an SNR ratio of 21 dB, where the conditions for speech signal recognition become as favorable as possible, the gap between the networks narrowed, but the CNN continued to show better results. This confirms its high adaptability and ability to learn, even under conditions of high signal strength. Moreover, the difference in performance decreased with an increase in the SNR from 5% to 2% for different SNR levels.

Within the framework of this study, it was found that different languages demonstrated different results at different levels of SNR. For example, despite the kinship between the Kazakh and Kyrgyz languages, the CNN was more successful in recognizing the Russian language. This may indicate a greater similarity in phonetic features between Kazakh and Russian languages than between Kazakh and Kyrgyz. This result requires further detailed research and analysis to identify phonetic features that affect the accuracy of speech signal recognition.

In addition, in this study, it was found that the use of a limited number of speakers (in our case, 40) allowed us to achieve good results in the accuracy of speech signal recognition with different SNR ratios compared to other studies where up to 500 speakers participated in the learning process [20]. This underlines the effectiveness of the chosen approach for the training and optimization of neural networks, as presented in [13].

Conclusion

Based on the research conducted, the following conclusions can be drawn. The results showed that an increase in the SNR ratio led to a significant increase in the accuracy of automatic speech recognition systems using the CNN and RNN architectures of neural networks. High SNR values, such as 21 dB, show a significant improvement in recognition accuracy compared with lower SNR values, such as 3 dB. It was found that the architecture of the CNN neural network was more effective for speech signal recognition, regardless of the SNR ratio and language. In addition, the use of a limited number of speakers for training has demonstrated the high efficiency and adaptability of CNN to various speech conditions. It was revealed that with an increase in the SNR ratio, the difference in accuracy between the CNN and RNN decreased, but CNN continued to show better results. This indicates its higher adaptability and ability to learn under different noise levels, which makes it versatile for use in various conditions. The effectiveness of neural networks depends on the language in which they are trained and the number of speakers. Networks trained in Kazakh showed better results in recognizing Kazakh speech but also successfully coped with recognizing Russians. This indicates the need to consider language features when learning and using neural networks to recognize speech signals.

References

[1] Kumar S., Rani R., Chaudhari U. Real-time sign language detection: Empowering the disabled community (2024) MethodsX, 13, art. no. 102901. DOI: 10.1016/j.mex.2024.102901

[2] Wang J., Saleem N., Gunawan T.S. Towards Efficient Recurrent Architectures: A Deep LSTM Neural Network Applied to Speech Enhancement and Recognition (2024) Cognitive Computation, 16 (3), pp. 1221 – 1236. DOI: 10.1007/s12559-024-10288-y

[3] Tan Y.W., Ding X.F. Heterogeneous Convolutional Recurrent Neural Network with Attention Mechanism and Feature Aggregation for Voice Activity Detection (2024) APSIPA Transactions on Signal and Information Processing, 13 (1), art. no. e6. DOI: 10.1561/116.00000158

[4] Janani S., Akhil Hassan G., Madhankumar S., Kumar M.A. Speech Enhancement Algorithm Analysis for a Reliable Speech Recognition System using Artificial Intelligence Methods (2023) 1st Interna-

tional Conference on Emerging Research in Computational Science, ICERCS 2023 – Proceedings. DOI: 10.1109/ICERCS57948.2023.10434226

[5] D. Lim, H. Kang, B. Choi, W. Hong and J. Lee, "An Interpersonal Dynamics Analysis Procedure With Accurate Voice Activity Detection Using Low-Cost Recording Sensors," in *IEEE Access*, vol. 12, pp. 68427-68440, 2024, doi: 10.1109/ACCESS.2024.3387279.

[6] G. T.Y., B.G. N., Jayanna H.S. Development of noise robust real time automatic speech recognition system for Kannada language/dialects (2024) Engineering Applications of Artificial Intelligence, 135, art. no. 108693. DOI: 10.1016/j.engappai.2024.108693

[7] Pavani C. Mumtaz B.M. CNN-based noise reduction for multi-channel speech enhancement system with discrete wavelet transform (DWT) preprocessing (2024) PeerJ Comput. Sci., DOI 10.7717/peerj-cs.1901

[8] George C., Thomas H., Stefan G. The effect of spoken language on speech enhancement using self-supervised speech representation loss functions. 2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. October 22-25, 2023, New Paltz, NY.

[9] Gurvich et al., "A Real-Time Active Speaker Detection System Integrating an Audio-Visual Signal with a Spatial Querying Mechanism," *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Republic of, 2024, pp. 8781-8785, doi: 10.1109/ICASSP48485.2024.10446169.

[10] Y. Zang, Y. Zhang, M. Heydari and Z. Duan, "SingFake: Singing Voice Deepfake Detection," *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Republic of, 2024, pp. 12156-12160, doi: 10.1109/ICASSP48485.2024.10448184.

[11] Li, H., et al.: Speaker identification using Ultra-Wideband measurement of voice. IET Radar Sonar Navig. 18(2), 266–276 (2024). https://doi.org/10.1049/rsn2.12525

[12] Ouisaadane A., Safi S., Frikel M. An experiment of Moroccan dialect speech recognition in noisy environments using PocketSphinx (2024) International Journal of Speech Technology, 27 (2), pp. 329 – 339. DOI: 10.1007/s10772-024-10103-x

[13] Nurlankyzy A., Akhmediyarova A., Zhetpisbayeva A., Namazbayev T., Yskak A., Yerzhan N., Medetov B. THE DEPENDENCE OF THE EFFECTIVENESS OF NEURAL NETWORKS FOR RECOGNIZING HUMAN VOICE ON LANGUAGE (2024) Eastern-European Journal of Enterprise Technologies, 1 (9(127)), pp. 72 – 81 DOI: 10.15587/1729-4061.2024.298687

[14] Mussakhojayeva, S., Khassanov, Y. , Varol, H.A.: KSC2: An Industrial-Scale Open-Source Kazakh Speech Corpus. In: Proceedings of the 23rd INTERSPEECH Conference: pp. 1367-1371. 2022.

[15] Mussakhojayeva S., Khassanov Y., Atakan Varol H. (2021) A Study of Multilingual End-to-End Speech Recognition for Kazakh, Russian, and English. In: Karpov A., Potapova R. (eds) Speech and Computer. SPECOM 2021. Lecture Notes in Computer Science, vol 12997. Springer, Cham. https://doi.org/10.1007/978-3-030-87802-3_41

[16] Mussakhojayeva, S.; Dauletbek, K.; Yeshpanov, R.; Varol, H.A. Multilingual Speech Recognition for Turkic Languages. Information 2023, 14, 74.

[17] 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. In: Karpov A., Potapova R. (eds) Speech and Computer. SPECOM 2021. Lecture Notes in Computer Science, vol 12997. Springer, Cham. https://doi.org/10.1007/978-3-030-87802-340

[18] Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F.M., Weber, G.: Common voice: A massively-multilingualspeech corpus. In: LREC. pp. 4218–4222. ELRA (2020)

[19] Font, F.; Roma, G.; Serra, X. Freesound Technical Demo. In Proceedings of the 21st ACM International Conference on Multimedia MM '13, Barcelona, Spain, 21 October 2013; Association for Computing Machinery: New York, NY, USA, 2013; pp. 411-412.

[20] S. Yadav, P.A.D. Legaspi, M.S.O. Alink, A.B.J. Kokkeler and B. Nauta, "Hardware Implementations for Voice Activity Detection: Trends, Challenges and Outlook," in *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 70, no. 3, pp. 1083-1096, March 2023, https://doi.org/10.1109/TCSI.2022.3225717