Machine learning-based voice assistant: optimizing the efficiency of speech conversion for people with speech disorders

M.H. Antor [1], N.V. Chudinovskikh [1], M.V. Bachurin [1], A.A. Shurpikov [1], N.A. Khlebnikov [1], B.A. Bredikhin [1]
[1] Ural Federal University named after the first President of Russia B. N. Yeltsin, Ekaterinburg, Russia, 620002, Mira street, 19

Abstract

An automatic speech recognition system has the possibility of enhancing the standard of living for persons with disabilities by solving issues such as dysarthria, stuttering, and other speech defects. In this paper, we introduce a voice assistant using hyperkinetic dysarthria (HD) defect speeches. It contains the data preprocessing steps and the development of a novel convolutional recurrent network (CRN) model that is built depending on the convolutional neural networks and recurrent neural networks. We implemented data preprocessing methods, including filtering, down-sampling, and splitting, to prevent overfitting and decrease processing power as well as time. In addition, the technique of Mel Frequency Cepstral Coefficients (MFCC) has been utilized to extract speech characteristics. The proposed model is trained to recognize HD speech disorders using a dataset including 2000 Russian speeches. The experimental results demonstrate that the proposed method obtains a character error rate (CER) of 14.76 %. It indicates that approximately 85 % of characters are able to correctly recognize on the test dataset. We have created a telegram bot that utilizes our trained model to help people with hyperkinetic dysarthria speech disorder. This bot is capable of providing assistance independently, without the need for any third-party assistance.

Keywords: natural language processing, hyperkinetic dysarthria, speech recognition, feature extraction, optimization.

Citation: Antor MH, Chudinovskikh NV, Bachurin MV, Shurpikov AA, Khlebnikov NA, Bredikhin BA. Machine learning-based voice assistant: optimizing the efficiency of speech conversion for people with speech disorders. Computer Optics 2025; 49(1): 124-131. DOI: 10.18287/2412-6179-CO-1482.

Introduction

Hyperkinetic dysarthria is highlighted through abnormal spontaneous actions that have an impact on the articulatory, respiratory, and pronator systems, which eventually impact speech and deglutition [1]. According to the Department of Neurology at Mayo Clinic in the United States, the number of patients recognized with motor speech disorders over 1993-2008 was estimated with average 57 %. Among these group of people, around 20 % were recognized with hyperkinetic dysarthria [2]. People with the hyperkinetic dysarthria problem face several difficulties in social and professional interactions. The human-computer interaction (HCI) technique is an approach to develop interactive systems; it generates designs that are user-friendly by focusing on the requirements and needs of the users [3]. Although the recently introduced entirely neural speech recognizers have shown positive performance [4], Hidden Markov Models (HMM) remain the fundamental component of effective speech recognition systems [5]. There are many methods that have been developed for the recognition of speech and other applications, including convolutional neural networks [6], gaussian mixture models [7], and hybrid systems that combine artificial neural networks with HMMs [8]. However, there has been very little research on people with hyperkinetic dysarthria.

Motivated by the considerable advancement obtained through neural networks in the field of natural language processing (NLP) [9], the implementation of combined neural networks for the hyperkinetic dysarthria speech recognition has become an attractive area [10]. According to the outcomes of previous investigations, the application of neural networks has not only decreased the necessity for speech preprocessing but also increased the accuracy of recognition. This research introduces a novel methodology for evaluating hyperkinetic dysarthria speech recognition. The neural networks based methodology encounters difficulties in training model due to insufficient speech dataset and defining most effective network model structures. Here is a summary of the major contributions:

- To overcome the insufficient disorder speeches, this work proposed a Russian hyperkinetic dysarthria disorder (RHDD) speech dataset, which can improve the neural network based model adaptability and ensure against the overfitting. The initial process refers to the collection of speeches by real hyperkinetic dysarthria disorder patient, which are then put through digital speech processing techniques, including data filtering, down-sampling, and splitting techniques.
- In order to enhance the efficiency and maintain numerical stability of feature extraction, this study proposes a convolutional recurrent network (CRN)

model. This model combines a convolutional neural network (CNNs) and a recurrent neural network (RNNs) for the analysis of RHDD speech. This model includes several modifications, including adjustments to the kernel size, several 2D convolution and linear blocks, and integrating an extra recurrent neural network block.

The findings of the experiment illustrate that the proposed CRN model performed better than the other baseline models, with a CER score of 14.76 %. The recognition rate increases with using a dataset of 2000 RHDD speeches, which presents more effective generalization and robustness. The rest of the manuscript is divided into several sections. Initially, it introduces the related work. Afterwards, the methods employed are briefly described. The next section presented the results and discussion. The manuscript concludes with a summary of our research and a projection of future endeavors.

1. Related works

The evaluation of speech disorders has garnered significant focus from researchers because for the abilities to enhance the quality of life for people with disabilities. Consequently, multiple automatic speech recognition systems have been explored to minimize speech disorder problems using publicly available databases [11 – 13]. Takashima et al. [14] proposed a convolutional neural network-based feature extraction to deal with the small local fluctuations of the speech uttered by a person with an articulation disorder. They evaluated on a word recognition task for one male with recorded 216 words included in the ATR Japanese speech database. The convolutional restricted Boltzmann machine parameters were obtained from training data and used as the initial value for the convolution layer for the convolutive bottleneck networks, achieving a maximum recognition accuracy of 86.11 %.

Passricha & Aggarwal [15] proposed a deep architecture based on convolutional neural networks and support vector machines together. The inter connected architecture called convolutional support vector machine (CSVM) which replaces the SoftMax layer of CNN with SVM for better classification. The model was tested with TIMIT dataset and results shows that there was 13.21 % improvement. Also, CSVM achieves 16.9 % and 17.1 % phone error rate (PER) on TIMIT and WSJ datasets respectively. Hinton et al. [16] and Abdel-Hamid et al. [17] successfully used convolutional neural networks (CNNs) for speech recognition tasks. It can perform any sequence-to-sequence mapping with great accuracy. Wang & Xiao [18] proposed a novel approach to speech recognition using a duration-distribution-based hidden markov model (DDBHMM). For the purpose of isolated word recognition, they utilized 1,254 Chinese words, and for continuous voice recognition, they utilized 48,000 sentences. The average recognition rate was 90.68 %, an increase of 0.85 % in recognition accuracy, and a decrease of 8.36 % in error rate compared to the traditional hidden markov model (HMM).

Muhammad et al. [19] utilized the conventional characteristics and categorization of automatic speech recognition approach in order to recognize the speech of patients who were suffering from voice disorder. The research investigated samples of Arabic speech from 62 patients with dysphonia and six vocal abnormalities, along with 50 participants who were considered as normal. The Arabic digits that were pronounced by normal people were recognized with a recognition accuracy of 100 percent. The authors [20] highlight the efficacy of multiple supervised learning techniques and an evaluation of performance has been conducted between end-to-end systems and hybrid systems utilizing deep neural network-hidden markov model (DNN-HMM).

The studies mentioned above highlight the extensive application of neural networks in the field of speech recognition, generating advantageous results. However, it is important to note that those studies mainly utilize CNN based models for the purpose of speech recognition, without introducing any additional improvements. In this study, we present a CRN model to recognition RHDD speech using CNNs and RNNs model to enhance the feature extraction. Additionally, this model is specifically engineered to handle audio data in real-time.

2. Methodology

The suggested procedure refers to the utilization of hyperkinetic dysarthria speech recognition technology for converting speech to text. The input contains mainly audio speeches, which are subsequently converted into MFCC coefficients. These are then fed into the CRN model in order to generate text. The generated text is after that transmitted to the telegram bot.

2.1. Dataset

The accessibility of datasets for Russian disordered speech is extremely limited. The dataset used in this study was collected at Ural Federal University, Russian Federation as part of a larger investigation into dysarthric Russian speech. The proposed dataset consists of 2000 RHDD sentences with a sampling rate of 48 kHz [21]. The audio samples have been collected from a person with dysarthria named Boris Andreevich Bredikhin, who is 24 years old and male. He has symptoms of a certain type of dysarthria called hypokinetic. During the data acquisition process, dysarthric patients were asked to read 2000 short Russian sentences. The defect speeches were recorded using a Realme C2 mobile with a 4× Cortex-A53 1.8 GHz processor and 4GB of RAM. Table 1 provides a brief description of the speech recording parameters used in this experiment.

The proposed HyperDysarthria-RusSpeechData dataset strength lies in its specificity and quality. It

contains 2000 sentences from a single patient with hypokinetic dysarthria, providing rich, specific data. The recordings are high-quality, ensuring reliable research. On the other hand, the current EasyCall Corpus dataset [22] does not mention the dysarthria subtype and has noise due to different recording conditions.

Tab. 1. Recording speech parameters

Parameters	Values
Recording format	wav
Duration of recordings	1.25 –12 seconds
Recording frequency	48000 Hz
Bitrate	128 kbps
Channels	2

2.2. Dataset Filtering and Down-sampling

The audio speech dataset goes through data filtering method to eliminate noise caused by different ambient conditions throughout the recording process.

$$Noisy\ Defect\ Speech \xrightarrow{\text{Removal of Speech Noise}} Clean\ Speech\ . \quad (1)$$

The signal is down-sampling to a sampling rate of 24000 Hz and filtered using a Band-pass filter with a frequency range of 20 to 2000 Hz.

$$Original_{48000\ Hz} \xrightarrow{Down-sampling} Compressed_{24,000\ Hz}\ , \quad (2)$$

where, original frequency is 48000 Hz, and the compressed frequency is 24000 Hz. This shows the compressed audio after the sampling rate has been decreased in by half.

2.3. Dataset Splitting

It essential to split the dataset into testing and training sets in order to minimize the overfitting. The dataset has been split using the most efficient K-Fold Cross Validation (k-FCV) technique [23]. The data has been divided into k-fold subsets, where one subset is identified as the test set and the remaining subsets are utilized for training. The training and testing subsections are allocated 85 % and 15 % of the total data, respectively. After splitting the dataset, there are 1700 samples in the training set and 300 samples in the testing set.

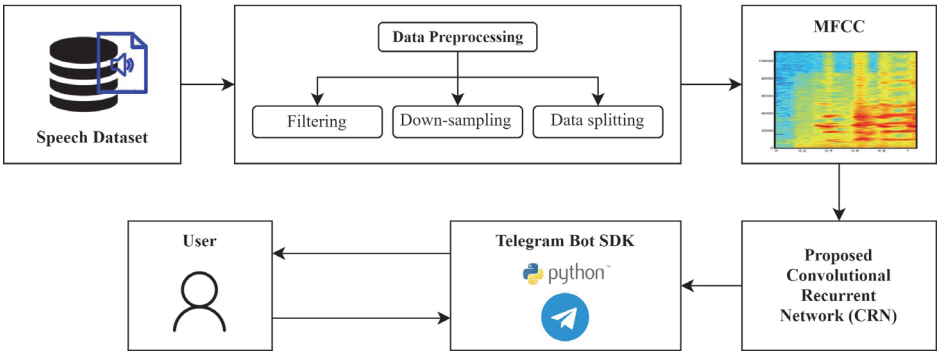

Fig. 1. Block diagram of speech recognition based hyperkinetic dysarthria speeches

2.4. MFCC coefficients

The MFCC are a commonly employed technique in speech recognition [24], specifically targeting features such as the frequency spectrum for the vocal tract. The features collected from the MFCC coefficients are provided to the proposed model instead of the raw signal directly. The MFCC extraction processes includes window function, Fourier transform, Mel filter, logarithm, and discrete cosine transform. The following is the sequential procedure for converting MFCC coefficients:

- The use of the hamming window function smoothly minimizes spectral leakage, while each audio observe frame is 40 milliseconds in duration.

$$W[n] = 0.54 - 0.46\cos(2\pi n / N - 1);\ 0 \le n \le N\text{-}1\ ,$$

where, N is the length of the window, and n is the sample number.

- The frequency spectrum is gathered through the integration of a Fourier transform to each and every frame.

$$X[k] = \sum x[n] \times e^{(-j2\pi kn/N)};\ 0 \le k \le N - 1\ ,$$

where, $x[n]$ is the *nth* sample of the signal, N is the total number of samples, and $X[k]$ is the *kth* sample of the Fourier transform.

- The spectrum has been determined from the earlier Fourier transform and transformed to the Mel scale. The mapping has been conducted through triangular overlapping windows.

$$M(f) = 2595\ \log 10(1 + f / 700)\ ,$$

where, f is the frequency in Hz.

- A logarithmic scale has been constructed from magnitude numbers that represent the logarithm of a power at each Mel frequency.

$$Y = \log(X)\ ,$$

where, X is the output of the Mel filter.

- After that, the discrete cosine transform was implemented to modify the sequence of Mel log powers. The outcome is a set of coefficients known as the MFCC coefficients.

$$C[m] = \Sigma\ (1/N) \times Y[n] \times \cos[\pi(m)(2n+1)/2N];$$
$$0 \le m \le N - 1\ ,$$

where, $Y[n]$ is the *nth* log Mel power, N is the total number of log Mel powers, and $C[m]$ is the *mth* MFCC.

2.5. Proposed Model

An convolutional recurrent network (CRN) model has been proposed for defect speech recognition based on CNNs [25] and RNNs [26]. The proposed model consists of 2D convolution (conv2d), linear, and recurrent blocks. The 2D convolution and linear layers is employed for image recognition, and recurrent layers are utilized for text prediction. The input MFCC coefficients image size is 32×32×3 pixel, batch size 5 for every epoch and the optimizer used is AdamW [27] with connectionist temporal classification (CTC) loss function [28] for 35 epochs. The optimal rate of learning has been determined with a value of 0.0005.

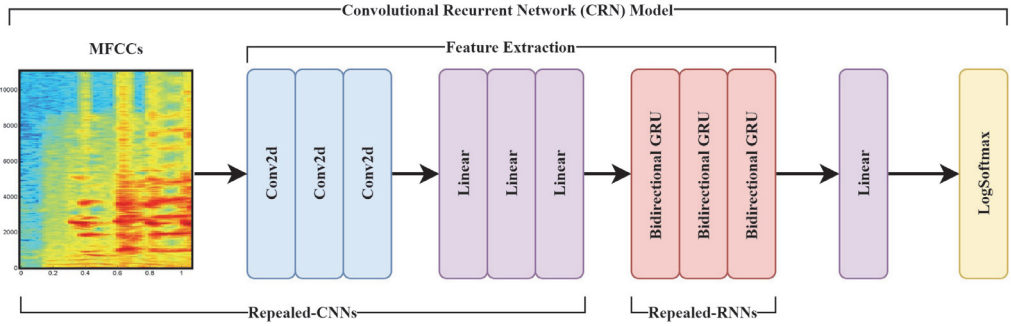

Fig. 2. The structure of proposed CRN model

Initially, a design block called Repealed-CNN is created, which is based on the typical CNN structure. A smaller kernel size leads to faster estimation and training times respectively, resulting in 32 kernels in the first conv2d layer with dimensions of 3×3. This differs from standard CNNs, that typically include a one-dimensional convolutional layer with dimensions of 7×7. The noise is filtered by the second conv2d layer using 64 kernels of dimensions 5×5. The third conv2d layer consists of 32 kernels, each with dimensions of 3×3. Gaussian Error Linear Unit (GELU) activation function has been used to overcome the vanishing gradient problem and BatchNorm2d normalization layer is subsequently implemented with each convolutional layer. This configuration is designed to enhance the network's capability to extract small features and faster estimation. The spectrogram image was then transformed into three linear blocks with LayerNorm and GELU activation function. It is responsible for transforming processed images into a linear structure. In this case, the data involves a transformation from its original lower-level features to a new set of higher-level features, resulting in an improvement in the speech's quality.

Following the development of the Repealed-CNN, the Repealed-RNN architecture is constructed. The architectural design encompasses a configuration comprising of three Bidirectional GRU (BiGRU) elements. The BiGRU layer receive a linear data structure as input from last linear layer. It enables the evaluation of linear data structures, which is essential for the recognition of word and letter sequences. According to different audio file lengths, BiGRU layers could improve recognition in situations where context is crucial. The last linear layer receives input from the BiGRU layer that includes the outcome of the processing of the data structures. In order to enhance efficiency and ensure numerical stability, the last layer is built as a LogSoftmax layer.

3. Results and discussion

This section contains the outcomes of the research study conducted based on the proposed methodology. This study is carried out on a personal computer platform (ASUS Vivo Book 15 M513UA, Processor AMD Ryzen 5 5500U with Radeon Graphics 2.10 GHz, Installed RAM 16GB DDR4), The Kaggle platform provides access to 2 types of video cards - NVIDIA TESLA P100 GPU, NVIDIA T4 (×2). The efficiency of the CRN model was evaluated using the character error rate (CER) and word error rate (WER) metrics [29]. As the Russian language is complex and has many suffixes, prefixes, endings, etc., we ignore WER as it evaluates too strictly and considers the whole word wrong for one wrong letter. CER is a more absorbent metric and looks at letters, not words. The levenshtein distance technique [30] was used to determine the incorrect characters and words. The CER and WER metrics equations are:

$$CER = Wrong\ Characters / Total\ Characters, \quad (3)$$

$$WER = Wrong\ Words / Total\ Words. \quad (4)$$

The results highlight that the CER achieves 14.76 % while the WER achieves 62.13 %. The CER is stable at approximately 0.15, meaning that nearly 85 % of characters are accurately recognized from speech with defects. Furthermore, it has been noticed that the WER remains constant at approximately 0.62 – 0.65. This indicates that around 35 – 38 % of words are accurately recognized from speech with defects. Comparatively, the proposed solution outperforms competitors in terms of learning speed and character error rate metrics.

We compared the proposed CRN model performance with CSVM [15], and DDBHMM [18] models using training dataset in terms of train character error rate. We can observe that the CRN model is capable of an accuracy of up to 90.11 % for training, whereas the

CSVM has a maximum accuracy 81.83 % and DDBHMM has maximum accuracy 83.87 %. The error rate in the CSVM and DDBHMM models stopped decreasing by 3 points. The error rate in the proposed model continued decreasing and reached a low point of 9.89. The results show that CRN model does better than the CSVM and DDBHMM model on proposed dataset, when the same training conditions are used.

Tab. 2. Proposed CRN model configuration, Number of epochs: 35; Batch size: 5; learning rate = 0.0005; loss functions: CTC; optimizer: AdamW

Model Configuration					
Layer	Kernel Size	Input Neuron	Output Neuron	Stride	Padding
Convolutional Blocks					
Conv2d	(4, 4)	1	32	(3, 3)	(2, 2)
BatchNorm2d	-	32	32	-	-
GELU	-	-	-	-	-
Conv2d	(3, 3)	32	64	(1, 1)	(1, 1)
BatchNorm2d	-	64	64	-	-
GELU	-	-	-	-	-
Conv2d	(3, 3)	64	32	(1, 1)	(1, 1)
BatchNorm2d	-	32	32	-	-
GELU	-	-	-	-	-
Linear Blocks					
Linear	-	224	270	-	-
LayerNorm		270	270	-	-
GELU	-	-	-	-	-
Linear	-	270	270	-	-
LayerNorm	-	270	270	-	-
GELU	-	-	-	-	-
Linear	-	270	270	-	-
Recurrent Blocks					
Bidirectional GRU	-	270	270	-	-
Bidirectional GRU	-	540	270	-	-
Bidirectional GRU	-	540	270	-	-
Linear	-	540	34	-	-
LogSoftmax	-	-	-	-	-

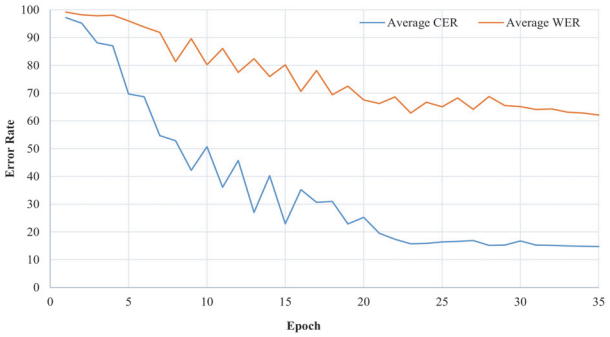

Fig. 3. Error Rate (CER and WER) comparison curves for test dataset using proposed CRN model

A telegram bot has been developed to implement the proposed CRN model, that utilizes the model we have trained to assist people with hyperkinetic dysarthria speech impairment. The first part of this diagram illustrates the interface created using an input. Another part of the diagram displays the outcomes recognized by the proposed model. The function enables the reception of voice messages, which are then converted into text using a trained CRN model, and subsequently output a response message. We have developed a fundamental feature that allows users to listen to their message and respond accordingly.

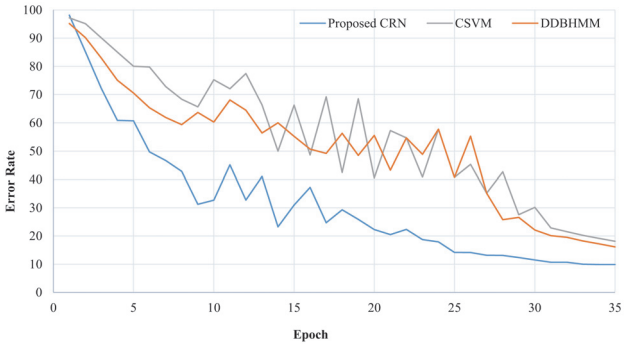

Fig. 4. CER Error on train dataset using existing CSVM, and DDBHMM model with proposed CRN model

Tab. 3. Comparison between proposed CRN model and existing CSVM and DDBHMM model based on character error rate using train dataset

Models	CER
Proposed CRN model	9.89%
CSVM	18.17%
DDBHMM	16.13%

Conclusion

The main objectives of this research work were to find out the solutions for voice conversion with speech defects. A comprehensive dataset consisting of 2000 RHDD speeches has been suggested to ensure an adequate quantity of defected speech. Furthermore, a CRN model was developed based on CNNs and RNNs. The obtained findings are considered satisfactory, as the proposed model exhibits a higher overall testing CER metric is 14.76 % compared to other models. Also, a voice assistant in the form of a telegram bot has been develop using proposed model that can help people.

The suggested approach could be utilized as a classroom assistance that provides speech-to-text features to assist students with hyperkinetic dysarthria. This would enable them to actively participate in discussions and successfully perform their assigned responsibilities. Another possible use could be in the healthcare sector, specifically in speech therapy treatments. It can be utilized to enhance the pronunciation and intelligibility of people with speech disorders. Adding more hyperkinetic dysarthria speeches from different person to the existing dataset is one way to further developing our research. Additionally, we can expand proposed dataset to include other speech disorders in order to build better models for the future. Additionally, the combination of a predictive typing system into an existing application can enhance the accuracy of transforming entire messages, compared to individual characters.

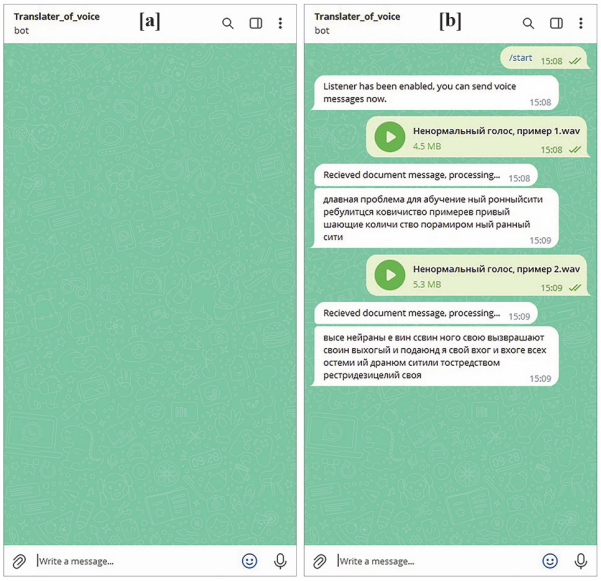

Fig. 5. Screenshot of developed Telegram bot: (a) interface; (b) recognized result

References

The main objectives of this research work were to find out the solutions for voice conversion with speech defects. A comprehensive dataset consisting of 2000 RHDD speeches has been suggested to ensure an adequate quantity of defected speech. Furthermore, a CRN model was developed based on CNNs and RNNs. The obtained findings are considered satisfactory, as the proposed model exhibits a higher overall testing CER metric is 14.76% compared to other models. Also, a voice assistant in the form of a telegram bot has been develop using proposed model that can help people.

The suggested approach could be utilized as a classroom assistance that provides speech-to-text features to assist students with hyperkinetic dysarthria. This would enable them to actively participate in discussions and successfully perform their assigned responsibilities. Another possible use could be in the healthcare sector, specifically in speech therapy treatments. It can be utilized to enhance the pronunciation and intelligibility of people with speech disorders. Adding more hyperkinetic dysarthria speeches from different person to the existing dataset is one way to further developing our research. Additionally, we can expand proposed dataset to include other speech disorders in order to build better models for the future. Additionally, the combination of a predictive typing system into an existing application can enhance the accuracy of transforming entire messages, compared to individual characters.

References

[1] Darley FL, Aronson AE, Brown JR. Clusters of deviant speech dimensions in the dysarthrias. J Speech Hear Res 1969; 12(3): 462-96. DOI: 10.1044/jshr.1203.462.

[2] Barkmeier-Kraemer JM, Clark HM. Speech-language pathology evaluation and management of hyperkinetic disorders affecting speech and swallowing function. Tremor Other Hyperkinet Mov 2017; 7: 489. DOI: 10.5334/tohm.381.

[3] Sadeghi Milani A, Cecil-Xavier A, Gupta A, Cecil J, Kennison S. A systematic review of Human-Computer Interaction (HCI) research in medical and other engineering fields. Int J Human–Computer Interact 2024; 40(3): 515-536. DOI: 10.1080/10447318.2022.2116530.

[4] Nassif AB, Shahin I, Attili I, Azzeh M, Shaalan K. Speech recognition using deep neural networks: A systematic review. IEEE Access 2019; 7: 19143-19165. DOI: 10.1109/ACCESS.2019.2896880.

[5] Saon G, Sercu T, Rennie S, Kuo H-KJ. The IBM 2016 english conversational telephone speech recognition system. arXiv Preview. 2016. Source: <https://arxiv.org/abs/1604.08242>. DOI: 10.48550/arXiv.1604.08242.

[6] Hashan AM, Al-Saeedi Adnan Adhab K, Islam RMRU, Avinash K, Dey S. Automated human facial emotion recognition system using depthwise separable convolutional neural network. 2023 IEEE Int Conf on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT) 2023: 113-117. DOI: 10.1109/IAICT59002.2023.10205785.

[7] Yu D, Deng L. Gaussian mixture models. In Book: Yu D, Deng L. Automatic speech recognition. A deep learning approach. London: Springer-Verlag; 2015: 13-21. DOI: 10.1007/978-1-4471-5779-3_2.

[8] Palaz D, Magimai-Doss M, Collobert R. End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition. Speech Commun 2019; 108: 15-32. DOI: 10.1016/j.specom.2019.01.004.

[9] Wang W, Gang J. Application of convolutional neural network in natural language processing. 2018 Int Conf on Information Systems and Computer Aided Education (ICISCAE) 2018: 64-70. DOI: 10.1109/ICISCAE.2018.8666928.

[10] Kukharchik P, Martynov D, Kheidorov I, Kotov O. Vocal fold pathology detection using modified wavelet-like features and support vector machines. 2007 15th European Signal Processing Conf 2007: 2214-2218.

[11] Kim H, Hasegawa-Johnson M, Perlman A, et al. Dysarthric speech database for universal access research. 9th Annual Conf of the International Speech Communication Association (INTERSPEECH 2008) 2008: 1741-1744. DOI: 10.21437/Interspeech.2008-480.

[12] Joy NM, Umesh S. Improving acoustic models in TORGO dysarthric speech database. IEEE Trans Neural Syst Rehabil Eng 2018; 26(3): 637-645. DOI: 10.1109/TNSRE.2018.2802914.

[13] Hashan AM, Dmitrievich CR, Valerievich MA, Vasilyevich DD, Alexandrovich KN, Bredikhin BA. Hyperkinetic Dysarthria voice abnormalities: a neural network solution for text translation. Int J Speech Technol 2024; 27(1): 255-265. DOI: 10.1007/s10772-024-10098-5.

[14] Takashima Y, Nakashika T, Takiguchi T, Ariki Y. Feature extraction using pre-trained convolutive bottleneck nets for dysarthric speech recognition. 2015 23rd European Signal Processing Conf (EUSIPCO) 2015: 1411-1415. DOI: 10.1109/EUSIPCO.2015.7362616.

[15] Passricha V, Aggarwal RK. Convolutional support vector machines for speech recognition. Int J Speech Technol 2019; 22(3): 601-609. DOI: 10.1007/s10772-018-09584-4.

[16] Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR. Improving neural networks by preventing co-adaptation of feature detectors. arXiv Preview. 2012. Source: <https://arxiv.org/abs/1207.0580>. DOI: 10.48550/arXiv.1207.0580.

[17] Abdel-Hamid O, Mohamed A, Jiang H, Deng L, Penn G, Yu D. Convolutional neural networks for speech recognition. IEEE/ACM Trans Audio Speech Lang Process 2014; 22(10): 1533-1545. DOI: 10.1109/TASLP.2014.2339736.

[18] Wang Z, Xiao X. Duration-distribution-based HMM for speech recognition. Front Electr Electron Eng 2006; 1(1): 26-30. DOI: 10.1007/s11460-005-0010-z.

[19] Muhammad G, Mesallam TA, Malki KH, Farahat M, Alsulaiman M, Bukhari M. Formant analysis in dysphonic patients and automatic Arabic digit speech recognition. Biomed Eng OnLine 2011; 10: 41. DOI: 10.1186/1475-925X-10-41.
[20] Gurunath Shivakumar P, Narayanan S. End-to-end neural systems for automatic children speech recognition: An empirical study. Comput Speech Lang 2022; 72: 101289. DOI: 10.1016/j.csl.2021.101289.
[21] Hashan AM, Chaganov RD, Melnikov AV, Dorokh DV, Khlebnikov NA, Bredikhin BA. Hyperkinetic Dysarthria voice abnormalities: a neural network solution for text translation. Int J Speech Technol 2024; 27(1): 255-265. DOI: 10.1007/s10772-024-10098-5.
[22] Turrisi R, Braccia A, Emanuele M, et al. EasyCall corpus: a dysarthric speech dataset. arXiv Preview. 2021. Source: <https://arxiv.org/abs/2104.02542>. DOI: 10.48550/arXiv.2104.02542.
[23] Yadav S, Shukla S. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. 2016 IEEE 6th Int Conf on Advanced Computing (IACC) 2016: 78-83. DOI: 10.1109/IACC.2016.25.
[24] Ittichaichareon C, Suksri S, Yingthawornsuk T. Speech recognition using MFCC. Int Conf on Computer Graphics, Simulation and Modeling (ICGSM'2012) 2012: 135-138.
[25] Wu Y, Feng J. Development and application of artificial neural network. Wirel Pers Commun 2018; 102: 1645-1656. DOI: 10.1007/s11277-017-5224-x.
[26] Salehinejad H, Sankar S, Barfett J, Colak E, Valaee S. Recent advances in recurrent neural networks. arXiv Preview. 2018. Source: <https://arxiv.org/abs/1801.01078>. DOI: 10.48550/arXiv.1801.01078.
[27] Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv Preview. 2017. Source: <https://arxiv.org/abs/1711.05101>. DOI: 10.48550/arXiv.1711.05101.
[28] Ariesta MC, Wiryana F, Suharjito, Zahra A. Sentence level Indonesian sign language recognition using 3D convolutional neural network and bidirectional recurrent neural network. 2018 Indonesian Association for Pattern Recognition Int Conf (INAPR) 2018: 16-22. DOI: 10.1109/INAPR.2018.8627016.
[29] Habeeb IQ, Al-Zaydi ZQ, Abdulkhudhur HN. Selection technique for multiple outputs of optical character recognition. Eurasian J Math Comput Appl 2020; 8(2): 41-51. DOI: 10.32523/2306-6172-2020-8-2-41-51.
[30] Sugiarto, Diyasa IGSM, Diana IN. Levenshtein distance algorithm analysis on enrollment and disposition of letters application. 2020 6th Information Technology International Seminar (ITIS) 2020: 198-202. DOI: 10.1109/ITIS50118.2020.9321030.

Authors' information

Mahamudul Hashan Antor, (b. 1995), received the M.S.c. degree in Innovative Software Systems: Design, Development & Application from National University of Science and Technology "MISIS", Moscow, Russian Federation in 2020. Currently he is a graduate student (Ph.D.) and working as a Assistant in the Institute of Fundamental Education at Ural Federal University, Yekaterinburg, Russian Federation. His research interests include digital image processing, computer vision, signal analysis, pattern recognition, and speech recognition.
E-mail: *hashan.antor@gmail.com* ORCID: 0000-0001-7926-9245.

Nikolai Vitalievich Chudinovskikh (b. 2003) studies at the Ural Federal University, Yekaterinburg, Russian Federation, Institute of Radio Electronics and Information Technologies - RTF in the 3rd year of a bachelor's degree in the specialty "Applied Informatics". Research interests include machine learning, natural language processing, and data analysis. E-mail: *kolyan.chudinovskix@mail.ru* ORCID: 0009-0006-4593-2401.

Matvey Vladimirovich Bachurin (b. 2003) is studying at the Ural Federal University, Yekaterinburg, Russian Federation, Institute of Radio Electronics and Information Technologies - RTF in the 3rd year of a bachelor's degree in the specialty "Applied Informatics". Research interests include neural networks, natural language processing, and speech recognition. E-mail: *matvey_1703@mail.ru* ORCID: 0000-0002-1707-2380.

Alexey Alexandrovich Shurpikov (b. 2003) studies at the Ural Federal University, Yekaterinburg, Russian Federation, Institute of Radio Electronics and Information Technologies - RTF in the 3rd year of a bachelor's degree in the specialty "Applied Informatics". Area of scientific interests is natural language processing, machine learning, and speech analysis. E-mail: *vasav7001@gmail.com* ORCID: 0009-0002-6217-5351.

Nikolai Alexandrovich Khlebnikov (b. 1981), graduated from the Faculty of Physics and Technology of UPI named after CM. Kirov, specialty "Information systems in materials science". Associate Professor, Director in the Institute of Fundamental Education (InFO), Ural Federal University, Yekaterinburg, Russian Federation, 620002.
E-mail: *na.khlebnikov@urfu.ru* ORCID: 0000-0003-3662-1039.

Boris Andreevich Bredikhin (b. 2000), graduated from Ural Federal University in 2018, major in Software Engineering. Currently he is master student at Ural Federal University, specialty "Information Systems and Technologies". Research interests are machine learning in cybersecurity, habilitation and rehabilitation tools using machine learning. E-mail: *boris.bredikhin@urfu.me* ORCID: 0009-0005-7370-9947.

Code of State Categories Scientific and Technical Information (in Russian – GRNTI)): 16.31.21
Received December 12, 2023. The final version – April 04, 2024.