

# The recognition of speech defects using convolutional neural network

Olha Pronina<sup>1</sup>, Olena Piatykop<sup>1</sup>

<sup>1</sup>*Pryazovskyi State Technical University, 19 Dmytro Yavornytskyi Ave., Dnipro, 49005, Ukraine*

## Abstract

The paper proposes a solution to improve the efficiency of recognition of speech defects in children by processing the sound data of the spectrogram based on convolutional neural network models. For a successful existence in society, a person needs the most important skill – the ability to communicate with other people. The main part of the information a person transmits through speech. The normal development of children necessarily includes the mastery of coherent speech. Speech is not an innate skill for people, and children learn it on their own. Speech defects can cause the development of complexes in a child. Therefore, it is very important to eliminate them at an early age. So, the problem of determining speech defects in children today is a very urgent problem for parents, speech therapists and psychologists. Modern information technologies can help in solving this problem. The paper provides an analysis of the literature, which showed that models of CNN can be successfully used for this. But the results that are available today have not been applied to speech in Ukrainian. Therefore, it is important to develop and study models and methods of convolutional neural networks to identify violations in the speech of children. The paper describes a mathematical model of oral speech disorders in children, the structure of a convolutional neural network and the results of experiments. The results obtained in the work allow to establish one of the speech defects: dyslexia, stuttering, difsonia or dyslalia with recognition results of 77-79%.

## Keywords

speech defects, smart data processing, CNN, model of a convolutional neural network, Deep Learning

## 1. Introduction

The development of speech is the main skill that allows a person to communicate qualitatively in society. The developmental stage of human speech begins in early childhood. That is, in the interval from one to six years, the foundation is laid, thanks to which the child will build his entire social and communication life. Speech defects can cause the development of complexes in a person. Therefore, it is very important to eliminate them at an early age [1].

The due to the violation of oral speech, there may be difficulties in schooling. The sound pronunciation is important for children in order to write and read correctly, not to skip letters when writing, not to make mistakes in the analysis of sounds and letters. Also the child may find it difficult to connect with their peers and teachers. There may be problems with the

---

*CoSinE 2022: 10th Illia O. Tepytskyi Workshop on Computer Simulation in Education, co-located with the ACNS Conference on Cloud and Immersive Technologies in Education (CITeD 2022), December 22, 2022, Kyiv, Ukraine*


✉ [pronina.lelka@gmail.com](mailto:pronina.lelka@gmail.com) (O. Pronina); [pyatikopalena@gmail.com](mailto:pyatikopalena@gmail.com) (O. Piatykop)

🌐 [https://kn.pstu.edu/?page\\_id=7038](https://kn.pstu.edu/?page_id=7038) (O. Pronina); [http://kn.pstu.edu/?page\\_id=7038](http://kn.pstu.edu/?page_id=7038) (O. Piatykop)

🆔 0000-0001-7085-8027 (O. Pronina); 0000-0002-7731-3051 (O. Piatykop)

© 2023 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

child's psyche and lagging behind the school curriculum. In addition to psychological phobias associated with speech, for example, a phobia of communicating with people, a child may have problems in a team, peers may react negatively due to problems in speech. Such a reaction can give rise to new phobias and mental problems in a child.

The causes of poor diction in children can be such combinations of factors [2]:

- biological – these include speech disorders caused by pathogenic factors that acted on the body during fetal development and in childhood, for example, severe infections, injuries. This includes disorders caused by genetic predisposition [3]. Manifestations of biological causes – stuttering, nasal voice, speech retardation, swallowing individual sounds;
- social – social causes that led to a violation of diction include problems in communicating with others, personal complexes and self-doubt. A child who doubts his abilities speaks quietly and indistinctly, “swallows” sounds, trying to end the conversation quickly and not stand out.

The formation of a child's speech occurs under the direct influence of adult speech and related exercises and training. Intensive speech practice, proper upbringing and training, timely and accurate detection of speech disorders, correct speech of surrounding adults are important conditions for the normal speech development of a child.

Currently, there is no single classification of speech disorders. The ineffectiveness of the classification of speech disorders is explained by the fact that a person does not have specific organs for performing speech functions. The generation of speech and voice is carried out by adapted organs and systems that initially perform other physiological functions.

In this case, it is possible to perform a general division of speech disorders into classes, and the following defects are more common:

- dyslexia – a selective specific violation of the reading process, difficult formation of reading skills;
- dyslalia – a violation in which it is difficult for a child to pronounce hissing sounds;
- rhinosalia – during such a pathology, the child's voice sounds vile;
- stuttering – this pathology is associated with a violation of the rhythm of speech, stretching of individual sounds;
- dysphonia – a defect associated with disturbances in the vocal apparatus, characterized by the absence of phonation or a partial loss of pitch, strength and timbre of the voice, in extreme cases, a complete loss of speech.

Loss of speech is one of the most complex manifestations of speech disorders, this violation can have several variations:

- aphasia is a speech disorder that affects the function of speech at the level of its processing;
- alalia – the absence or underdevelopment of speech;
- alogia – the limited use of speech in mental disorders.

Motor alalia is characterized by the fact that the child understands what is addressed to him, but does not speak. Sensory alalia is characterized by a speech understanding

disorder while maintaining elementary hearing. If you do not develop the child's speech, then he may completely miss the opportunity to talk. Speech activity in children is directly related to the formation of the psyche and affects the intellectual, sensory, emotional and volitional spheres of life.

An important step is the timely diagnosis of speech disorders in a child. For this, it is advisable to use modern information technologies. The possibilities of smart data processing will allow you to recognize and determine the type of speech defect. The development and use of such software will enable a speech therapist to carry out high-quality diagnostics. Also, such a program would be useful for parents. This will make it possible to conduct an initial diagnosis at home and contact a specialist in a timely manner. To develop such software, it is first necessary to develop and investigate a model for speech defect recognition.

## 2. Literature review

To date, there are many works on speech recognition for its further transcription. The question of speech recognition for detecting defects suggests a slightly different mechanism. In addition, it is worth paying attention to the fact that the specificity of each language has its own pronunciation features, and hence the detection of speech defects. Most often, neural networks are used for speech recognition [4, 5, 6, 7, 8, 9]. This is due to the specifics of the tasks for which they are used, as well as the quality of the final result.

In [5] an overview of modern deep learning architectures of neural networks, algorithms and systems for speech programs is presented. Due to the growing amount of sensor data and cloud computing for processing and training deep neural networks. And with increasing complexity in mobile technology, intelligent systems are poised to revolutionize to adapt to the task at hand. This article reviews some of the most successful deep learning models for intelligent vision and speech systems to date. From this work, it can be concluded that recent advances in deep learning of artificial neural networks are driving rapid innovation and development in intelligent vision and language systems.

The work of Bahuleyan [8] contains a comparison of the performance of two classes of models. The first is a deep learning approach where a CNN model is trained end-to-end to predict the genre label of an audio signal solely using its spectrogram. The second approach uses hand-crafted features in both the time and frequency domains. The author trains four traditional machine learning classifiers with these features and compares their performance. Also, the work defines the functions that are most conducive to solving the problem of multiclass classification. Experiments are carried out with the audio dataset.

Kourkounakis et al. [7] highlighted the problem of identifying and classifying various forms of stuttering. Unlike most of the existing work identifying stuttering with speech models, their work proposes a model based solely on acoustic features, allowing the identification of several variants of stuttering dysfunctions without the need for speech recognition. This model uses a deep residual network and bidirectional short-term long-term memory layers to classify different kinds of stutters and achieves an average miss rate.

Latif et al. [10] discusses the problems and key characteristics of models of teaching representation and discussion. It also considers possible future trends. Speech processing research

has traditionally been treated as a task of developing hand-crafted acoustic characteristics as a separate problem. This task differs from the task of developing efficient machine learning models for making prediction and classification decisions. Although the general structure can be traced. There are two main disadvantages of this approach: firstly, this is an engineering function – a manual that is cumbersome and requires human knowledge; secondly, the functions developed may not be the best for achieving the goals.

Work by Purwins et al. [6] is devoted to the current state of deep learning methods for processing audio signals and deep learning models, including convolutional neural networks. Speech, music, and environmental sound processing are considered parallel. To point out the similarities and differences between the two, the author reviews common methods, issues, key references, and the potential for cross-training.

Chlasta et al. [11] proposes a new approach to automated detection of depression in people in speech using a convolutional neural network (CNN) and multi-party interactive learning. Early detection and treatment of depression is essential for recovery, prevention of relapse and reduction of the severity of the disease. The experiment applied data to residual CNNs in the form of image spectrograms automatically generated from audio samples.

Work by Sheikh et al. [12] is devoted to a comprehensive examination of audio and the identification of features, methods for classifying stuttering/disagreement based on statistical and deep learning. The identification of stuttering is an interesting interdisciplinary research problem involving pathology, psychology, acoustics, and signal processing, which in turn makes it difficult to identify the problem.

Kourkounakis et al. [13] proposed an end-to-end deep neural network FluentNet, capable of identifying a number of different types of disagreements. FluentNet consists of a residual convolutional neural network that facilitates the study of strong frame-level spectral representations. Frames are followed by a set of bi-directional long short-term memory layers to aid in learning effective timing. In addition, FluentNet uses the attention mechanism to focus on important parts of speech in order to get better performance.

In [14] the MCLNN performance was evaluated using the Ballroom and Homburg music genre datasets. Conditional Neural Networks (CLNN) and Conditional Masked Neural Networks (MCLNN) exploit the nature of multidimensional time signals. CLNN captures the effect of conditional time between frames in a window. And the mask in MCLNN provides a systematic sparsity that follows a pattern similar to a filter bank. The mask induces the network to learn about the frequency representation of time in bands, allows the network to tolerate frequency shifts. In addition, the mask in MCLNN usually automates the exploration of a number of feature combinations made through exhaustive manual searches.

Wang and Chen [15] formulated a newer approach to speech separation as a supervised learning problem. Where discriminatory patterns of speech, speakers and background noise are obtained from the training data. This article provides a comprehensive overview of research based on deep learning in broadcast separation control over the past few years and presents a formulation of supervised separation and a discussion of the three main components of supervised learning: machine learning, learning goals, and acoustic performance.

Thus, most researchers use neural networks [5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17]. However, most of them are not focused on the Ukrainian language, so ready-made solutions are not suitable for the specifics of the Ukrainian language and children. It was decided to use convolutional

neural networks to recognize the voice messages of children when they were tested to identify deviations in speech. Which in turn is the object of research when detecting speech deviations.

### 3. Mathematical model of speech disorders in children

The construction of a mathematical model is carried out in order to mathematically represent a variant of the model as a system.

The input data for the model are audio signals with labels for the type of speech defect established by the speech therapist. To describe the violation in the speech of children, a mathematical model was developed. Speech disorders are presented as a tuple:

$$M = \{A, B, C, F, G\}, \quad (1)$$

where  $M$  is the set of violations of the child's oral speech;  $A$  – violation of the physical condition;  $B$  – mental problems;  $C$  – loss of speech;  $F$  – difficulties in schooling;  $G$  – mental retardation of the child.

In turn, the violation of the physical state  $A$  consists of the signs given in formula (2):

$$A = \{Q, W, E, R\}, \quad (2)$$

where  $A$  is a set of physical condition disorders;  $Q$  – nasal voice;  $W$  – speech retardation;  $E$  – swallowing individual sounds;  $R$  – stuttering.

Mental problems, that is, sign  $B$  is detailed in formula (3):

$$B = \{Y, U\}, \quad (3)$$

where  $B$  is the set of violations of the physical state;  $Y$  – “swallowing” sounds;  $U$  – quiet and slurred speech.

Sign  $C$ , characterized by loss of speech, in turn consists of the signs given in formula (4):

$$C = \{I, O, S\}, \quad (4)$$

where  $C$  is loss of speech;  $I$  – aphasia;  $O$  – alalia;  $S$  – alogia.

Alalia also consists of two features given in formula (5):

$$O = \{K, L\}, \quad (5)$$

where  $O$  is alalia;  $K$  is motor alalia;  $L$  – sensory alalia.

Learning problems consist of the features presented below in formula (6):

$$F = \{J, Z\}, \quad (6)$$

where  $F$  – difficulties in schooling;  $J$  – lagging behind the training program;  $Z$  – misunderstanding with peers and teachers.

The mental retardation of the child – sign  $G$  – appears due to the fact that oral speech is one of the main methods of teaching and improving the personality. The speech activity of the child

is associated with different areas of development that are responsible for the mental abilities of the child. Sign G includes the following signs, which are contained in formula (7):

$$G = \{D, X, V\}, \quad (7)$$

where G is the mental retardation of the child; D – autism; X – mutism; V – general underdevelopment of speech.

To determine which defect the current state of the child refers to, it was decided to use the probability of determining the disease. Thus, this will allow not only to indicate the disease, but also to understand how accurately the deviation was detected. This is due to the fact that there are situations when a child has several defects, and then, in addition to testing with the help of the system, one should additionally contact a speech therapist or defectologist.

#### **4. Modeling a convolutional neural network for speech defects recognition**

Based on the features of the task and analysis of the literature, a convolutional neural network was chosen as a neural network. Since it has a number of advantages for solving the problem of speech recognition. When using a convolutional neural network, you can significantly reduce the number of training parameters and get a high quality classification. Another reason why a convolutional neural network was chosen is that it provides partial resistance to scale changes, displacements, rotations, angle changes, and other distortions. Based on the advantages, the structure of a convolutional neural network was developed, which is used to train the model to solve the problem of determining the type of speech impairment. Its graphical representation is shown in figure 1.

The zero layer of a convolutional neural network is the layer with the input data.

The first layer of a convolutional neural network is the Convolution (Conv2D) layer. It performs the convolution process. Its main input parameters are the number of filters, the size of the convolution window, and an array with the value of the size of the input data. The value for the number of filters was set to 32, the size of the convolution window was set to 3 by 3. The value for the size of the input was set to 64 by 64.

The second layer is to reduce the data dimension and get important information using the AveragePooling2D() function. AveragePooling2D() in turn acquires the size of the pooling window, the value for it was set to 2 by 2.

The third layer is the activated layer. The relu activation function was chosen as the main activation function. Its main advantage is that it avoids and fixes the disappearing gradient problem and is less computationally intensive.

The fourth layer is the Conv2D convolution layer. It had the following values: 64 for filters, 3 by 3 convolution window size and offset type – same. If you select the “same” offset, the image size remains the same.

The fifth layer consists of the AveragePooling2D() function, which reduces the data. The pool window size is two.

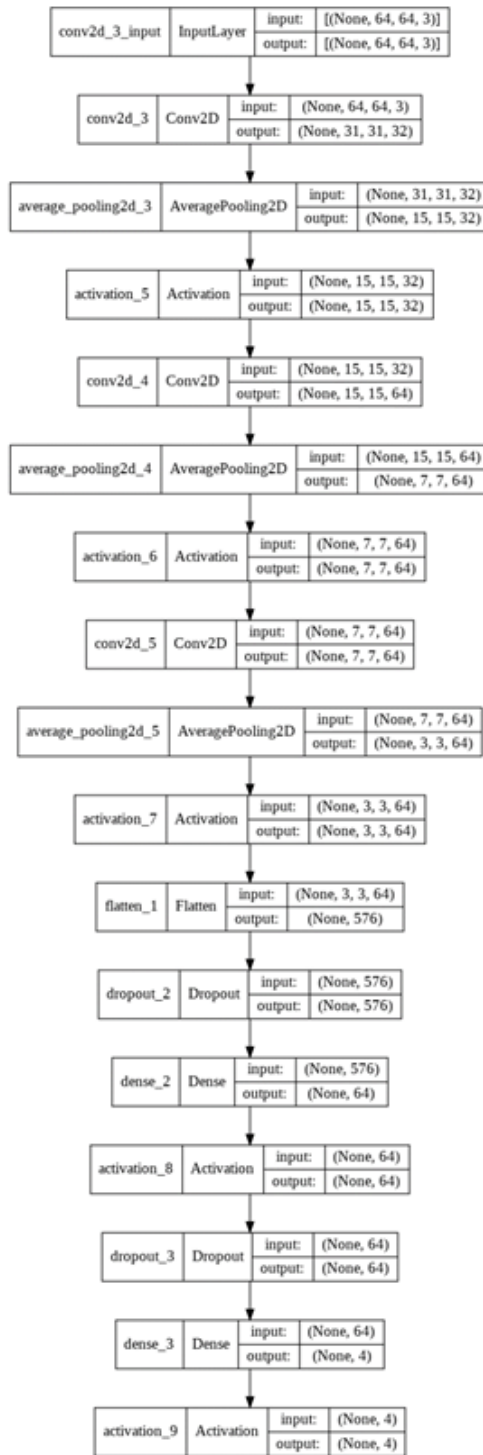


Figure 1: Model of a convolutional neural network.

The sixth layer is the activated layer. The ReLU activation function is chosen for this activation layer. This function returns 0 if it accepts a negative argument, if it accepts a positive argument, it returns it.

The seventh layer is convolutional. It has the same input values and parameters as the fourth layer.

The eighth layer serves to reduce the dimension of the data. It consists of the function AveragePooling2D and takes values such as the size of the decrease window, which has been set to 2 by 2.

The ninth layer is the activation layer, like the sixth layer, the main activation function is ReLU.

The tenth layer is the pull-out layer. The Flatten function performs a tensor flattening operation, changing the shape of the tensor so that it has a shape equal to the number of elements contained in the tensor, regardless of the batch size. The value was left at the default.

The eleventh layer is the model adjustment layer. Using the Dropout() function, we indicate that the layer is regulating. The parameter to be passed to Dropout is a value indicating how likely the neurons will be disabled. It was set to 0.5, which means that half of the neuron will be turned off.

The twelfth layer is a fully connected Dense layer. The main function of this layer is to receive information from all nodes of the previous layer. The input value is the dimension of the output space, for this layer the dimension has been set to 64.

The thirteenth layer is the active layer, like the ninth layer, it is the activation layer with ReLU activation function. This function was chosen because there are no special requirements for the output value of the neuron within the framework of the task.

The fourteenth layer is the learning adjustment layer. Like the eleventh layer, it takes only one parameter and it is equal to 0.5, which means that half of the neurons will not be involved in training.

The fifteenth layer is a fully connected layer that receives information from all other layers. The dimension of this method depends on the number of classes in which the model is trained. The number of training classes was chosen to be 4, so the dimension value was set to 4.

The last layer is the activation layer. The Softmax activation function was installed for it. Softmax turns a vector of real numbers into a vector of probabilities for the selected classes. The deviation turns them into values between 0 and 1 so that they can be interpreted as probabilities. However, if one of the inputs is significantly small or negative, the Softmax function turns it into a low probability.

Conversely, if the input is large, Softmax turns it into a large probability. In this case, the values will always remain between 0 and 1, which is suitable for the problem statement for the selected subject area. The activation function is presented below in formula:

$$s(x_i) = \frac{e^{x_j}}{\sum_{j=1}^n e^{x_j}}, \quad (8)$$

where  $x$  – the input vector,  $e^{x_j}$  – the standard exponential function,  $\sum_{j=1}^n e^{x_j}$  – the sum of all initial values;  $n$  – the number of class.



Thus, the Softmax function transforms the feature vector obtained in the neural network for each image, a vector consisting of numbers, which in turn are in the range from 0 to 1, and the sum of all elements of each such vector is equal to one. That is, each element of this vector can be estimated as the probability that the image in question belongs to the corresponding element of the class.

The process of identifying speech defects includes several stages. In order to develop a functioning model, all the signs of a speech disorder were considered. This was done in order to select video materials for training the developed neural network. All video materials were taken with speech disorders of children who have not yet started working with a speech therapist. Each video is approximately ten seconds long. The input data for the model are audio signals that were selected and processed from those found in the public domain, with labels for the type of speech defect established by the speech therapist.

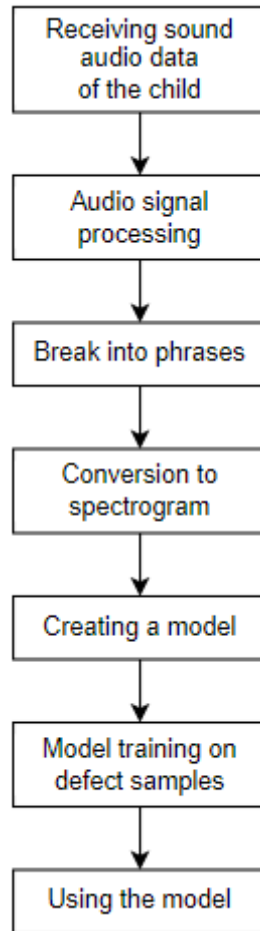
The signals of these video materials are processed and cut into phrases and words. The splitting of the speech audio file is done with the help of python libraries, namely pydup, as well as Librosa. That allows you to both break the replicas into words and build a spectrogram for further analysis. These audio signals are then converted into a spectrogram. A spectrogram is a visual representation of the spectrum at the frequencies of a signal and how it changes over time. The sound spectrogram is used for phonetic identification of spoken words and for their analysis. It is rendered as a heat map, that is, an image with intensity displayed by changing color or brightness. After that, the spectrogram data were fed to the neural network as a training sample. After training the convolutional neural network, it was tested on data taken with and without speech disorders.

The algorithm for general data processing and stages of working with a model for detecting speech defects consists of several stages. The general principle of defect handling is shown in figure 2.

At the first stage, the data with which the system will work is prepared. This data can be obtained from the global network, namely a video in which speech therapists talk or work with children who have speech impairments. At the second stage, after processing the audio signals and splitting them into phrases and words, these audio files are converted into a spectrogram. Frequencies are shown on the vertical axis and time on the horizontal axis. And the intensity of the color increases with increasing density. The final stage of the algorithm is the creation and training of the model. After that, the training and training of the model takes place. With high-quality training and confirmation of performance, its further use occurs.

## **5. Experimental research**

In order to analyze the trained model, we used recordings of conversations of children with speech impairments, which were taken from a YouTube video hosting, where children talk to a doctor before they begin to undergo treatment. The entire data sample was divided according to the principle of 70 to 30. The model was trained on a test set made independently from video and audio files available in the public domain. Namely, the conversation of children with speech therapists. A date set was made from audio data with labels of belonging to the defect class. After training, a dataset not previously used in training was taken and testing was carried out.



**Figure 2:** The general process of recognizing speech defects.

A total of 32 experiments were conducted, 8 experiments were run for each type of speech disorder. The minimum accuracy was 0.7293 and the maximum was 0.8143. The results of the experiment are presented in table 1. The “Pm” column is the class to which the model assigned the loaded audio recording, and the “Oc” column indicates the possibility that the audio recording is the class that the Pm model defined.

The division into speech disorders is shown in figure 3.

The chart in figure 3 shows data for each type of speech impairment, i.e. mean value from eight experiments. This chart displays several speech disorders as bars growing in a given direction from the baseline. It can be seen that the highest probability of determining dyslalia is 78.78%, the next violation is dyslexia – 78.74%, followed by stuttering with 77.55%, and the last violation is dysphonia with 76.25% probability.

During the experiment, audio recordings of children with speech impairments were collected for each class. The age of the children in the audio files ranges from 5 to 14 years old. To analyze the satisfactory training of the trained convolutional neural network model, each class was

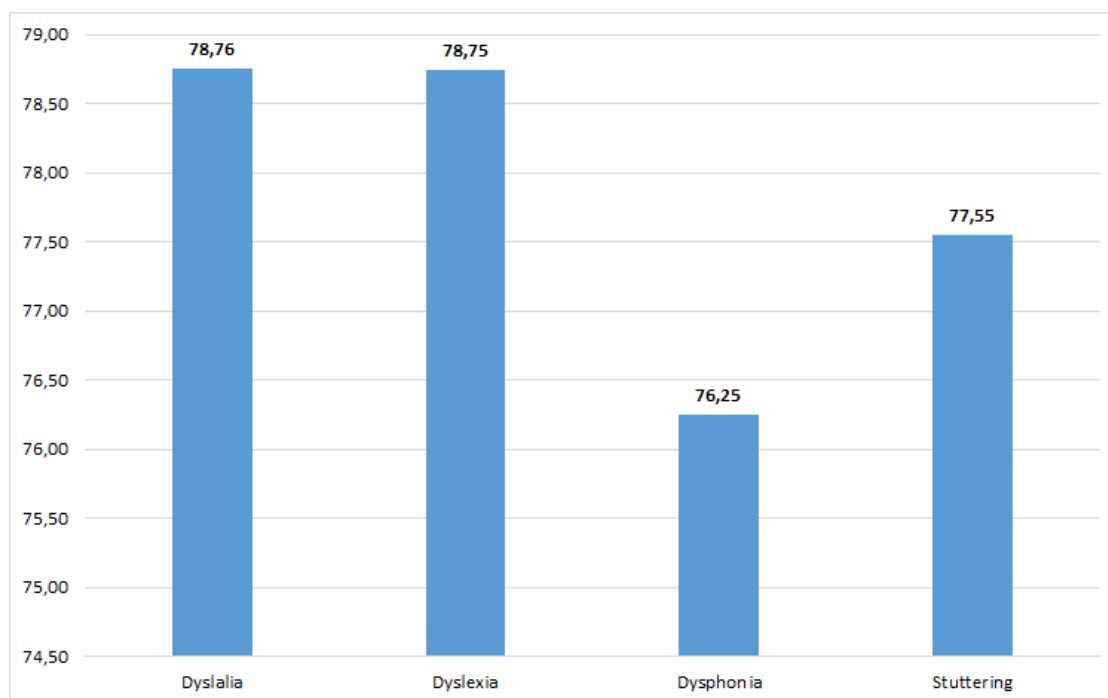
**Table 1**

The result of the experiment.

Experiment number	Oc, %	Pm
1	78,55	Dyslalia
2	80,22	Dyslalia
3	80,06	Dyslalia
4	79,76	Dyslalia
5	76,44	Dyslalia
6	79,05	Dyslalia
7	78,65	Dyslalia
8	77,54	Dyslalia
9	78,54	Dyslexia
10	81,1	Dyslexia
11	77,65	Dyslexia
12	79,08	Dyslexia
13	78,77	Dyslexia
14	78,51	Dyslexia
15	78,12	Dyslexia
16	78,19	Dyslexia
17	75,92	Dysphonia
18	77,1	Dysphonia
19	75,92	Dysphonia
20	76,91	Dysphonia
21	75,03	Dysphonia
22	77,12	Dysphonia
23	75,5	Dysphonia
24	76,51	Dysphonia
25	76,43	Stuttering
26	76,49	Stuttering
27	76,74	Stuttering
28	81,43	Stuttering
29	75,23	Stuttering
30	77,81	Stuttering
31	76,87	Stuttering
32	79,4	Stuttering

carried out on a test set, the number of audio tracks is different. The result of the experiment is presented in table 2.

As a result of this experiment, it was concluded that the successful identification of trained classes: dyslexia, stuttering, difsonia and dyslalia is 77-79%. This is a correct result and allows you to accurately determine the likely violation in the child's speech.



**Figure 3:** The values for each speech defects.

**Table 2**

The result of the experiment.

	Dyslexia	Stuttering	Dysphonia	Dyslalia
Number of audio files	22	19	17	18
Number of recognitions	17	15	13	14
Percentage of successful recognition	77.25	78.95	76.45	77.80

## 6. Conclusions

The subject of the study is the models and methods of convolutional neural networks for detecting speech defects in children. Classes of disturbances in speech were preliminarily investigated. On the basis of these data, a general mathematical model of speech impairment in children was built.

The paper presents the results of the analysis of studies on the use of convolutional neural networks for the recognition of sound information. The analysis showed that convolutional neural network models can be successfully used to classify speech defects. But the results that are available today have not been applied to speech in Ukrainian. Therefore, an architectural model and structure of a convolutional neural network for speech defect recognition was developed. Experiments have been conducted that have shown a successful definition of classes of defects: dyslexia, stuttering, difsonia and dyslalia is 77-79%.

The next step will be to expand to cover a wider range of speech deficiencies and deviations.

## References

- [1] I. V. Martynenko, Psychological principals for communication activity development in senior preschool age children with system speech disorders, The thesis for obtaining the Scientific Degree of the Doctor of Psychological Sciences in speciality 19.00.08. – Special Psychology, M. P. Dragomanov National Pedagogical University, Kyiv, 2017. URL: [https://npu.edu.ua/images/file/vidil\\_aspirant/dicer/%D0%94\\_26.053.23/Martynenko.pdf](https://npu.edu.ua/images/file/vidil_aspirant/dicer/%D0%94_26.053.23/Martynenko.pdf).
- [2] O. V. Boryak, The specificity of a cognitive component in the speech activity at mental retardation, Aktualni pytannia korektsiinoi osvity. Pedagogichni nauky 7(1) (2016) 38–49. URL: [http://nbuv.gov.ua/UJRN/apko\\_2016\\_7%281%29\\_\\_6](http://nbuv.gov.ua/UJRN/apko_2016_7%281%29__6).
- [3] T. H. Kolomoiets, D. A. Kassim, Using the Augmented Reality to Teach of Global Reading of Preschoolers with Autism Spectrum Disorders, in: A. E. Kiv, V. N. Soloviev (Eds.), Proceedings of the 1st International Workshop on Augmented Reality in Education, Kryvyi Rih, Ukraine, October 2, 2018, volume 2257 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 237–246. URL: <http://ceur-ws.org/Vol-2257/paper24.pdf>.
- [4] S. Semerikov, I. O. Teplytskyi, Y. V. Yechkalo, O. M. Markova, V. N. Soloviev, A. Kiv, Computer Simulation of Neural Networks Using Spreadsheets: Dr. Anderson, Welcome Back, in: V. Ermolayev, F. Mallet, V. Yakovyna, V. S. Kharchenko, V. Kobets, A. Kornilowicz, H. Kravtsov, M. S. Nikitchenko, S. Semerikov, A. Spivakovsky (Eds.), Proceedings of the 15th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. Volume II: Workshops, Kherson, Ukraine, June 12-15, 2019, volume 2393 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 833–848. URL: [http://ceur-ws.org/Vol-2393/paper\\_348.pdf](http://ceur-ws.org/Vol-2393/paper_348.pdf).
- [5] M. Alam, M. Samad, L. Vidyaratne, A. Glandon, K. Iftekharuddin, Survey on Deep Neural Networks in Speech and Vision Systems, *Neurocomputing* 417 (2020) 302–321. doi:10.1016/j.neucom.2020.07.053.
- [6] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, T. Sainath, Deep Learning for Audio Signal Processing, *IEEE Journal of Selected Topics in Signal Processing* 13 (2019) 206–219. doi:10.1109/JSTSP.2019.2908700.
- [7] T. Kourkounakis, A. Hajavi, A. Etemad, Detecting Multiple Speech Disfluencies Using a Deep Residual Network with Bidirectional Long Short-Term Memory, in: ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6089–6093. doi:10.1109/ICASSP40776.2020.9053893.
- [8] H. Bahuleyan, Music Genre Classification using Machine Learning Techniques, 2018. doi:10.48550/ARXIV.1804.01149.
- [9] M. Dawodi, J. A. Baktash, T. Wada, N. Alam, M. Z. Joya, Dari speech classification using deep convolutional neural network, in: 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2020, pp. 1–4. doi:10.1109/IEMTRONICS51293.2020.9216370.
- [10] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, B. W. Schuller, Survey of Deep Representation

- Learning for Speech Emotion Recognition, *IEEE Transactions on Affective Computing* (2021) 1–1. doi:10.1109/TAFFC.2021.3114365.
- [11] K. Chlasta, K. Wołk, I. Krejtz, Automated speech-based screening of depression using deep convolutional neural networks, *Procedia Computer Science* 164 (2019) 618–628. doi:10.1016/j.procs.2019.12.228, cENTERIS 2019 - International Conference on ENTERprise Information Systems / ProjMAN 2019 - International Conference on Project MANagement / HCist 2019 - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2019.
- [12] S. A. Sheikh, M. Sahidullah, F. Hirsch, S. Ouni, Machine learning for stuttering identification: Review, challenges and future directions, *Neurocomputing* 514 (2022) 385–402. doi:10.1016/j.neucom.2022.10.015.
- [13] T. Kourkounakis, A. Hajavi, A. Etemad, Fluentnet: End-to-end detection of stuttered speech disfluencies with deep learning, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021) 2986–2999. doi:10.1109/TASLP.2021.3110146.
- [14] F. Medhat, D. Chesmore, J. Robinson, Music genre classification using masked conditional neural networks, in: D. Liu, S. Xie, Y. Li, D. Zhao, E. M. El-Alfy (Eds.), *Neural Information Processing - 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part II, volume 10635 of Lecture Notes in Computer Science*, Springer, 2017, pp. 470–481. doi:10.1007/978-3-319-70096-0\_49.
- [15] D. Wang, J. Chen, Supervised Speech Separation Based on Deep Learning: An Overview, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26 (2018) 1702–1726. doi:10.1109/taslp.2018.2842159.
- [16] A. Sokoliuk, G. Kondratenko, I. Sidenko, Y. Kondratenko, A. Khomchenko, I. Atamanyuk, Machine Learning Algorithms for Binary Classification of Liver Disease, in: *2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T)*, 2020, pp. 417–421. doi:10.1109/PICST51311.2020.9468051.
- [17] Y. Kondratenko, I. Atamanyuk, I. Sidenko, G. Kondratenko, S. Sichevskiy, Machine Learning Techniques for Increasing Efficiency of the Robot’s Sensor and Control Information Processing, *Sensors* 22 (2022) 1062. doi:10.3390/s22031062.