
**МАТЕМАТИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ,
КОМПЛЕКСОВ И КОМПЬЮТЕРНЫХ СЕТЕЙ/MATHEMATICAL SOFTWARE FOR COMPUTERS,
COMPLEXES AND COMPUTER NETWORKS**

DOI: <https://doi.org/10.60797/IRJ.2026.172.90> EDN: BZUNHE

ИССЛЕДОВАНИЕ ГЛУБИНЫ НЕЙРОННЫХ СЕТЕЙ В ЗАДАЧЕ РАСПОЗНАВАНИЯ ГОЛОСОВЫХ КОМАНД

Научная статья

Данько Е.В.^{1,*}

¹ ORCID : 0000-0002-8651-5296;

¹ Алтайский государственный университет, Барнаул, Российская Федерация

* Корреспондирующий автор (evdanko88[at]gmail.com)

Предложена: 30.06.2026; Принята: 02.10.2026; Опубликовано: 16.10.2026

Аннотация

В настоящее время искусственные нейронные сети активно используются для задач, исходные данные для которых представлены в виде аудиосигнала. Примерами таких задач являются: классификация жанров музыки, выявление эмоций человека по голосу, улучшение качества звукового сигнала (удаление шума), идентификация голосовых команд при управлении различными устройствами.

В данном исследовании рассматривается задача распознавания англоязычных голосовых команд с помощью нейронных сетей, создаваемых с использованием библиотек Keras и TensorFlow языка Python. При этом для решения поставленной задачи используются только сверточные нейронные сети. Целью проводимого эксперимента является определение оптимального количества свёрточных блоков нейронной сети для качественного распознавания команд.

Ключевые слова: сверточные нейронные сети, распознавание голосовых команд, обучение нейронной сети, распознавание речи, слой Conv2D, вычисление мел-частотных кепстральных коэффициентов.

RESEARCH ON THE DEPTH OF NEURAL NETWORKS IN THE TASK OF VOICE COMMAND RECOGNITION

Research article

Danko E.V.^{1,*}

¹ ORCID : 0000-0002-8651-5296;

¹ Altai State University, Barnaul, Russian Federation

* Corresponding author (evdanko88[at]gmail.com)

Suggested: 30.06.2026; Accepted: 02.10.2026; Published: 16.10.2026

Abstract

Currently artificial neural networks are actively used for tasks where the initial data is presented in the form of an audio signal. Examples of such tasks include: classifying music genres, identifying human emotions based on voice, improving the quality of audio signals (noise removal) and identifying voice commands for controlling various devices.

This study examines the task of recognizing voice commands in English using neural networks created with Keras and TensorFlow in Python. In this experiment only convolutional neural networks are used to solve the problem. The purpose of this experiment is to determine the optimal number of convolutional blocks in a neural network for high-quality command recognition.

Keywords: convolutional neural networks, voice commands recognition, neural network training, speech recognition, Conv2D layer, calculation of mel-frequency cepstral coefficients.

Введение

На современном этапе во многих сферах применяются искусственные нейронные сети. В связи с этим исследования особенностей архитектуры и структуры таких моделей становятся довольно актуальными. Для многих задач с исходными данными в форме графической информации применяются сверточные нейронные сети, основы которых подробно освещены во многих работах: [1], [3], [4]. Кроме того, сверточные сети также используют и при решении задач детекции объектов на изображениях [2], что показывает универсальность моделей, реализующих операцию свёртки. В зависимости от особенностей решаемой задачи, параметры и количество сверточных слоев сети может существенно различаться, более подробно об этом отмечено в работах [6], [7]. Помимо работы с графической информацией, сверточные нейронные сети часто применяются для задач с исходной информацией в звуковой форме, подобные задачи рассмотрены в [8], [10]. В работе [5] проведено сравнение архитектур сверточной нейронной сети, гибридной архитектуры Conformer, свёрточно-рекуррентной сети и сетей с остаточными связями (residual networks) для задач выявления эмоций и распознавания голоса. Авторами сделан вывод о предпочтительном использовании сверточных нейронных сетей в случае ограниченности вычислительных ресурсов. В исследовании [9] предложена нейронная сеть, включающая одномерные и двумерные сверточные слои для увеличения качества распознавания речи человека. В исследовании [11] авторы представили гибридную модель, содержащую помимо сверточных слоев блоки внимания, характерные для сетей-трансформеров, включение которых положительно влияет на качество решения

задачи автоматического распознавания речи (ASR). В работе [12] предложена свёрточно-рекуррентная нейронная сеть, включающая блоки внимания для более качественного распознавания эмоций человека по голосу. В исследовании [13] описан эксперимент по классификации 30 голосовых команд с подбором гиперпараметров свёрточной нейронной сети. В данном исследовании в качестве предобработки данных использовано вычисление мел-частотных кепстральных коэффициентов.

Таким образом изучение особенностей моделей сверточных нейронных сетей при работе со звуком является актуальной современной задачей, определившей направление нашего эксперимента.

Целью текущего исследования является выбор оптимального количества блоков сверточных слоев сети в задаче распознавания голосовых команд. В рамках исследования используется датасет “Mini speech commands”, содержащий 8 голосовых англоязычных команд: up, down, yes, no, stop, go, left, right. Мы обучим и протестируем набор сверточных нейронных сетей с разным количеством сверточных блоков слоев для выбора оптимальной модели. Актуальность данного эксперимента подтверждается активным применением архитектуры свёрточных сетей для задач распознавания речи, классификации голосовых команд, выявления эмоций и других подобных задач, описанных выше. В соответствии с поставленной целью исследования необходимо решить следующие задачи:

- создать несколько базовых моделей,
- провести обучение каждой модели,
- протестировать обученные сети на отложенной выборке,
- повторить эксперимент несколько раз,
- сравнить полученные результаты и сделать вывод об эффективности обученных моделей.

Научная новизна исследования состоит в изучении влияния глубины свёрточной нейронной сети на качество классификации голосовых команд. Так как модели на основе свёрточных сетей особо актуальны для устройств с ограниченными вычислительными ресурсами (например, мобильных устройств), дополнительное изучение особенностей структуры таких моделей позволит повысить качество решения данной задачи, актуальной для массового потребления.

Практическая значимость исследования состоит в повышении качества решения задачи распознавания голосовых команд, что активно используется во многих мобильных приложениях.

Таким образом, результаты нашего эксперимента позволят более осознанно подходить к формированию структуры свёрточных сетей для решения поставленной задачи с максимальным качеством.

Методы исследования

Датасет “Mini speech commands”, содержащий 8000 файлов с расширением wav, был скачан с сайта storage.googleapis.com.

Так как в этом датасете для каждой голосовой команды имеется 1000 файлов, дополнительные шаги по корректровке баланса классов не требовались. Весь набор данных был разбит на 3 части: обучающая выборка (6400 файлов), проверочная и тестовая выборки (по 800 файлов каждая). В качестве среды для работы была использована среда Google Colab, все вычисления выполнены с использованием графического процессора T4 GPU.

В качестве предварительной обработки исходных данных использовано вычисление мел-частотных кепстральных коэффициентов (MFCC). Этот подход часто используется для предобработки звуковой информации, более подробно о его особенностях отмечено в работе [10]. Альтернативные подходы к предобработке звука включают построение спектрограмм, вычисление спектрального центроида, спектрального спада и других параметров, что также используется во многих исследованиях: [8] и [9]. Дополнительно отметим, что алгоритм вычисления MFCC может немного различаться, в зависимости от особенностей используемых инструментов. В нашем исследовании алгоритм вычисления MFCC включает следующие шаги:

1. Выполнение кратковременного преобразования Фурье исходного аудиосигнала
2. Получение спектрограммы сигналов (абсолютных значений результатов преобразования Фурье)
3. Перевод спектрограммы в масштаб мел
4. Вычисление логарифмов по всем значениям мел-спектрограмм
5. Выполнение дискретного косинусного преобразования списка полученных логарифмов мел-спектрограмм.

Оценка качества работы обученных моделей проводилась с использованием метрики assigasy (доли правильных ответов классификатора), а также функции `classification_report` библиотеки Sklearn языка Python. Процесс обучения всех нейронных сетей проведен с использованием сокращения скорости обучения и ранней остановки процесса. Скорость обучения сокращалась в 4 раза при отсутствии улучшений метрики assigasy на проверочном множестве в течение 6 подряд эпох обучения, ранняя остановка обучения проводилась в случае, если метрика assigasy на проверочном множестве не улучшалась в течение 12 подряд итераций обучения.

Далее отметим особенности архитектуры обученных моделей.

```
[23]
✓ 0
чек.

inp = layers.Input(shape=input_shape)

x = layers.Conv2D(32, (3,3), padding='same')(inp)
x = layers.MaxPooling2D((2,2), padding='same')(x)

x = layers.Conv2D(64, (3,3), padding='same')(x)
x = layers.MaxPooling2D((2,2), padding='same')(x)

x = layers.Flatten()(x)
x = layers.Dropout(0.5)(x)
x = layers.Dense(256, activation='relu')(x)
x = layers.Dropout(0.5)(x)
out = layers.Dense(num_labels, activation='softmax')(x)

mod_1 = models.Model(inputs=inp, outputs=out)
```

Рисунок 1 - Определение первой нейронной сети
DOI: <https://doi.org/10.60797/IRJ.2026.172.90.1>

На рисунке 1 приведено создание первой модели сверточной сети с использованием функционального подхода. Здесь мы включили 2 сверточных блока, каждый из которых содержит слои Conv2D и MaxPooling2D со стандартными размерами ядер свёртки и окон пулинга. Число свёрточных фильтров увеличивается с 32 (в первом блоке) до 64 (во втором блоке). Далее следует слой Flatten для сокращения размерности тензоров, после чего идут Dropout (сокращение переобучения модели), промежуточный полносвязный слой Dense, снова Dropout и выходной слой для классификации аудиофайлов.

```
[31]
✓ 0
чек.
▼
```

mod_1.summary()

Model: "functional"

Layer (type)	Output Shape	Param #
input_layer	(None, 98, 40, 1)	0
conv2d	(None, 98, 40, 32)	320
max_pooling2d	(None, 49, 20, 32)	0
conv2d_1	(None, 49, 20, 64)	18,496
max_pooling2d_1	(None, 25, 10, 64)	0
flatten	(None, 16000)	0
dropout	(None, 16000)	0
dense	(None, 256)	4,096,256
dropout_1	(None, 256)	0
dense_1	(None, 8)	2,056

Total params: 4,117,128 (15.71 MB)
Trainable params: 4,117,128 (15.71 MB)
Non-trainable params: 0 (0.00 B)

Рисунок 2 - Общая информация о базовой модели
DOI: <https://doi.org/10.60797/IRJ.2026.172.90.2>

На рисунке 2 приведена таблица, показывающая полный набор использованных слоев в базовой модели. Здесь мы также можем увидеть информацию о размерности выходных тензоров после каждого имеющегося слоя, отметим, что общее количество настраиваемых параметров модели составляет более 4,1 млн. шт.

Помимо рассмотренной базовой модели, были также определены:

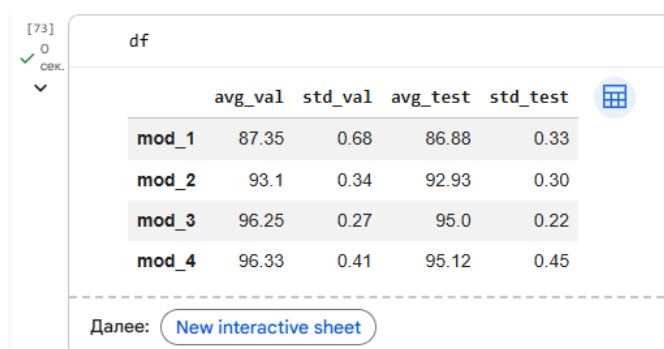
- mod_2 (нейронная сеть, включающая 3 сверточных блока: 32, 64 и 128 фильтров свертки в каждом). Включает более 2,2 млн. параметров.
- mod_3 (модель, содержащая 4 сверточных блока: 32, 64, 128 и 256 фильтров свертки в каждом). Всего — более 1,7 млн. параметров.
- mod_4 (нейронная сеть, включающая 5 сверточных блоков: 32, 64, 128, 256 и 512 фильтров свертки в каждом). Содержит более 2,6 млн. параметров.

Все отмеченные выше модели отличаются между собой только количеством сверточных блоков, при этом размеры ядер свертки, подвыборки и использованные полносвязные слои нейронных сетей mod_1, mod_2, mod_3 и mod_4 — одинаковы.

При обучении всех моделей использовался метод оптимизации Adam и функция потерь “sparse_categorical_crossentropy”, так как выходные тензоры представляли собой метки классов в виде целых чисел.

Основные результаты

Для надежности получаемых выводов эксперимент с созданием и обучением нейронных сетей был повторен 5 раз. Для анализа полученных результатов создан объект DataFrame, где отмечены достигнутые средние значения метрики accuracy на проверочных и тестовых данных для всех нейронных сетей.



	avg_val	std_val	avg_test	std_test
mod_1	87.35	0.68	86.88	0.33
mod_2	93.1	0.34	92.93	0.30
mod_3	96.25	0.27	95.0	0.22
mod_4	96.33	0.41	95.12	0.45

Рисунок 3 - Усредненные значения метрики accuracy и стандартные отклонения

DOI: <https://doi.org/10.60797/IRJ.2026.172.90.3>

На рисунке 3 представлена таблица с усредненными долями правильных ответов всех обученных классификаторов. Столбцы “avg_val” и “avg_test” — это усредненные значения accuracy на проверочном и тестовом множествах соответственно, а столбцы “std_val” и “std_test” — это стандартные отклонения значений метрики в проведенном эксперименте.

Можем отметить, что максимальное усредненное значение доли правильных ответов на проверочном множестве (96,33%) было достигнуто моделью mod_4, при этом максимальное значение accuracy на тестовых данных для mod_4 составило 95,12%. Чуть ниже средние значения accuracy для модели mod_3 - 96,25% и 95,0% соответственно. Однако при этом, значения accuracy для mod_3 характеризуются меньшим разбросом, чем соответствующие значения метрики модели mod_4, что хорошо заметно по значениям стандартных отклонений accuracy, полученных при повторении эксперимента для mod_3 и mod_4.

Вполне очевидно, что нейронные сети с 4-мя и 5-ю блоками сверточных слоев (mod_3 и mod_4 соответственно) работают в условиях текущей задачи значительно лучше, чем модели mod_1 и mod_2. При этом разница между сетями mod_3 и mod_4 не столь велика, что говорит о том, что дальнейшее увеличение сверточных блоков для моделей уже не приносит большой пользы.

Обсуждение

По результатам проведенного эксперимента можем отметить, что лучшим выбором количества сверточных блоков в рамках рассмотренной архитектуры является 4 или 5 блоков. Разница по усредненной доле правильных ответов между 4 или 5 блоками совсем незначительна в пользу модели с 5-ю блоками сверточных слоев (mod_4), однако сеть с четырьмя блоками (mod_3) имеет более устойчивые значения accuracy при повторении эксперимента. Также следует отметить, что модель mod_3 имеет меньшее число настраиваемых параметров, чем mod_4 (1,7 млн. шт. вместо 2,6 млн. шт.), поэтому использует меньший объем вычислительных ресурсов. В связи с этим, mod_3 может работать и обучаться быстрее, чем mod_4.

Без проведения дополнительного исследования сложно сделать однозначный выбор между сетями с 4-ми и 5-ю блоками. Однако мы можем отметить, что эффект переобучения нейронных сетей mod_3 и mod_4 гораздо менее выражен по сравнению с моделями mod_1 и mod_2. Для более глубокого анализа результатов исследования, для всех обученных моделей также были получены отчеты по классификации тестовых аудиофайлов. Вполне очевидно, что лучшие значения метрик precision, recall и f1-score по каждой голосовой команде имеют модели mod_3 и mod_4.

Таким образом, если скорость работы и меньший объем вычислительных ресурсов имеют наиболее важное значение, то можно рекомендовать к использованию модель mod_3. В случае, когда важно максимальное качество распознавания голосовых команд, требуется проведение дополнительного эксперимента для однозначного выбора между моделями mod_3 и mod_4. Здесь следует напомнить, что путем изменения других гиперпараметров процесса обучения (начальной скорости обучения, метода оптимизации, размера батча и др.) возможно дополнительное повышение качества работы для любой из обученных моделей.

Таким образом, при использовании двумерных сверточных слоев (Conv2D) для классификации команд по датасету “Mini speech commands” рекомендуется выбирать 4 или 5 блоков для максимизации качества работы модели. Полученный вывод вполне согласован с результатами современных исследований, например в работе [5] отмечена аналогичная зависимость общего количества параметров модели от качества ее работы. В нашем эксперименте

структура с наименьшим количеством настраиваемых параметров (mod_3) демонстрирует близкую к максимальным значениям долю правильных ответов. Это указывает на некий оптимальный баланс между общим размером модели и качеством её работы, так как слишком простые модели будут тяготеть к недообучению, а слишком сложные — к переобучению. Также отмеченные выше моменты подтверждаются результатами исследования [9], где авторы отмечают склонность к недообучению модели, включающей слои Conv1D (меньше настраиваемых параметров) по сравнению с сетью, использующей слои Conv2D (больше настраиваемых параметров) в задаче распознавания речи.

Заключение

В соответствии со сформулированной целью эксперимента, можем отметить, что при использовании датасета “Mini speech commands”, оптимальным количеством двумерных сверточных блоков является 4 или 5 блоков. Меньшее количество сверточных блоков в условиях рассмотренной задачи является менее эффективным, так как такие модели выполняют классификацию команд хуже, чем их более глубокие аналоги. Проведенный эксперимент показывает важность правильного выбора количества сверточных блоков в задаче распознавания голосовых команд и заметное влияние этого параметра на качество классификации команд.

В целом особенности выбора оптимальных параметров сверточных сетей для решения задач классификации звуковых данных мало освещены в современных исследованиях, поэтому результаты проведенного эксперимента имеют высокую практическую значимость. Особенности выбора гиперпараметров сверточных сетей сегодня все ещё остаются малоизученными, в связи с этим проведенный эксперимент полезен для более глубокого понимания зависимости структуры модели и качества её работы.

Дополнительно отметим, что полученные выводы справедливы только в рамках выбранной конфигурации нейронной сети и процесса обучения, так как в проведенном эксперименте варьировалось только число сверточных блоков. Одним из вариантов углубления проведенного эксперимента является подбор других гиперпараметров обучения сетей с четырьмя и пятью сверточными блоками с целью максимизации доли правильных ответов моделей на новых данных.

Конфликт интересов

Не указан.

Рецензия

Кацко С.Ю., Сибирский государственный университет геосистем и технологий, Новосибирск Российская Федерация
DOI: <https://doi.org/10.60797/IRJ.2026.172.90.4>

Conflict of Interest

None declared.

Review

Katsko S.Y., Siberian State University of Geosystems and Technologies, Novosibirsk Russian Federation
DOI: <https://doi.org/10.60797/IRJ.2026.172.90.4>

Список литературы / References

1. Лыкасов А.А. Свёрточные нейронные сети / А.А. Лыкасов // Современные научные исследования и инновации. — 2025. — № 4. — URL: <https://web.snauka.ru/issues/2025/04/103236> (дата обращения: 29.06.2026).
2. Зарипова Д.Н. Применение свёрточных нейронных сетей (CNN) для детекции объектов на аэроснимках / Д.Н. Зарипова, М.П. Шлеймович // Вестник науки. — 2025. — Т. 2. — № 5 (86). — С. 808–815. — ISSN 2712-8849. — URL: <https://www.vestnik-nauki.rf/article/22927> (дата обращения: 29.06.2026).
3. Трубин А.Е. Построение оптимальной модели сверточной нейронной сети для решения задач распознавания сложных символов / А.Е. Трубин, А.В. Батищев, А.Н. Алексахин [и др.] // Информационные технологии. — 2023. — Т. 29. — № 2. — С. 84–90. — DOI: 10.17587/it.29.84-90.
4. Krizhevsky A. ImageNet classification with deep convolutional neural networks / A. Krizhevsky, I. Sutskever, G.E. Hinton // Advances in Neural Information Processing Systems. — 2012. — № 25 (2). — P. 84–90. — DOI: 10.1145/3065386.
5. Kapu Nirmal Joshua. Towards Advanced Speech Signal Processing: A Statistical Perspective on Convolution-Based Architectures and its Applications / Kapu Nirmal Joshua, Raghav Karan // TechRxiv. — 19 November 2024. — DOI: 10.36227/techrxiv.173203269.96743718/v1
6. Zhou D.X. Theory of deep convolutional neural networks: Downsampling / D.X. Zhou // Neural Networks. — 2020. — Vol. 124. — P. 319–327.
7. Ронкин М.В. Глубокое обучение систем компьютерного зрения : учебное пособие / М.В. Ронкин, А.Ю. Долганов ; М-во науки и высш. образования РФ. — Екатеринбург: Изд-во Урал. ун-та, 2025. — 124 с.
8. Кошкарёва Ю.И. Выбор оптимальной системы распознавания речи в задаче анализа новостной информации / Ю.И. Кошкарёва, В.И. Терехов // Современные научные исследования и инновации. — 2016. — № 5. — URL: <https://web.snauka.ru/issues/2016/05/67547> (дата обращения: 29.06.2026).
9. Diep Q.B. Crossmixed convolutional neural network for digital speech recognition / Q.B. Diep, H.Y. Phan, T.C. Truong // PLoS One. — 2024. — Vol. 19. — № 4. — P. e0302394. — DOI: 10.1371/journal.pone.0302394.
10. Муртазин Р.А. Алгоритм выявления синтезированного голоса на основе кепстральных коэффициентов и сверточной нейронной сети / Р.А. Муртазин, А.Ю. Кузнецов, Е.А. Фёдоров [и др.] // Научно-технический вестник информационных технологий, механики и оптики. — 2021. — Т. 21. — № 4. — С. 545–552.
11. Deng F. CCE-Net: Causal Convolution Embedding Network for Streaming Automatic Speech Recognition / F. Deng, Y. Ming, B. Lyu // International Journal of Network Dynamics and Intelligence. — 2024. — Vol. 3. — № 3. — P. 100019.
12. Zhang H. Pre-trained Deep Convolution Neural Network Model with Attention for Speech Emotion Recognition / H. Zhang, R. Gou, J. Shang [et al.] // Frontiers in Physiology. — 2021. — Vol. 12. — P. 643202. — DOI: 10.3389/fphys.2021.643202.



13. Mahmood Arzo. Speech recognition based on Convolutional neural networks and MFCC algorithm / Mahmood Arzo, Köse Utku // *Advances in Artificial Intelligence Research*. — 2021. — Vol. 1. — № 1. — P. 6–12.

Список литературы на английском языке / References in English

1. Lykasov A.A. Svjortochnye nejronnye seti [Convolutional neural networks] / A.A. Lykasov // *Sovremennye nauchnye issledovaniya i innovacii* [Modern Scientific Research and Innovations]. — 2025. — № 4. — URL: <https://web.snauka.ru/issues/2025/04/103236> (accessed: 29.06.2026). [in Russian]
2. Zaripova D.N. Primenenie svjortochnyh nejronnyh setej (CNN) dlja detekcii ob#ektov na ajerosnimkah [Application of convolutional neural networks (CNN) for object detection in aerial images] / D.N. Zaripova, M.P. Shlejmovich // *Vestnik nauki* [Bulletin of Science]. — 2025. — Vol. 2. — № 5 (86). — P. 808–815. — ISSN 2712-8849. — URL: <https://www.вестник-науки.рф/article/22927> (accessed: 29.06.2026). [in Russian]
3. Trubin A.E. Postroenie optimal'noj modeli svertochnoj nejronnoj seti dlja reshenija zadach raspoznavaniya slozhnyh simvolov [Construction of an optimal convolutional neural network model for solving complex character recognition problems] / A.E. Trubin, A.V. Batishhev, A.N. Aleksahin [et al.] // *Informacionnye tehnologii* [Information Technologies]. — 2023. — Vol. 29. — № 2. — P. 84–90. — DOI: 10.17587/it.29.84-90. [in Russian]
4. Krizhevsky A. ImageNet classification with deep convolutional neural networks / A. Krizhevsky, I. Sutskever, G.E. Hinton // *Advances in Neural Information Processing Systems*. — 2012. — № 25 (2). — P. 84–90. — DOI: 10.1145/3065386.
5. Kapu Nirmal Joshua. Towards Advanced Speech Signal Processing: A Statistical Perspective on Convolution-Based Architectures and its Applications / Kapu Nirmal Joshua, Raghav Karan // *TechRxiv*. — 19 November 2024. — DOI: 10.36227/techrxiv.173203269.96743718/v1
6. Zhou D.X. Theory of deep convolutional neural networks: Downsampling / D.X. Zhou // *Neural Networks*. — 2020. — Vol. 124. — P. 319–327.
7. Ronkin M.V. Glubokoe obuchenie sistem komp'juternogo zrenija [Deep learning of computer vision systems] : textbook / M.V. Ronkin, A.Ju. Dolganov ; Ministry of Science and Higher Education of the Russian Federation. — Yekaterinburg: Ural University Publishing House, 2025. — 124 p. [in Russian]
8. Koshkareva Ju.I. Vybor optimal'noj sistemy raspoznavaniya rechi v zadache analiza novostnoj informacii [Choosing the optimal speech recognition system for the task of news information analysis] / Ju.I. Koshkareva, V.I. Terehov // *Sovremennye nauchnye issledovaniya i innovacii* [Modern Scientific Research and Innovations]. — 2016. — № 5. — URL: <https://web.snauka.ru/issues/2016/05/67547> (accessed: 29.06.2026). [in Russian]
9. Diep Q.B. Crossmixed convolutional neural network for digital speech recognition / Q.B. Diep, H.Y. Phan, T.C. Truong // *PLoS One*. — 2024. — Vol. 19. — № 4. — P. e0302394. — DOI: 10.1371/journal.pone.0302394.
10. Murtazin R.A. Algoritm vyjavlenija sintezirovannogo golosa na osnove kepstral'nyh ko#efficientov i svertochnoj nejronnoj seti [Algorithm for detecting synthesized voice based on cepstral coefficients and a convolutional neural network] / R.A. Murtazin, A.Ju. Kuznecov, E.A. Fjodorov [et al.] // *Nauchno-tehnicheskij vestnik informacionnyh tehnologij, mehaniki i optiki* [Scientific and Technical Bulletin of Information Technologies, Mechanics and Optics]. — 2021. — Vol. 21. — № 4. — P. 545–552. [in Russian]
11. Deng F. CCE-Net: Causal Convolution Embedding Network for Streaming Automatic Speech Recognition / F. Deng, Y. Ming, B. Lyu // *International Journal of Network Dynamics and Intelligence*. — 2024. — Vol. 3. — № 3. — P. 100019.
12. Zhang H. Pre-trained Deep Convolution Neural Network Model with Attention for Speech Emotion Recognition / H. Zhang, R. Gou, J. Shang [et al.] // *Frontiers in Physiology*. — 2021. — Vol. 12. — P. 643202. — DOI: 10.3389/fphys.2021.643202.
13. Mahmood Arzo. Speech recognition based on Convolutional neural networks and MFCC algorithm / Mahmood Arzo, Köse Utku // *Advances in Artificial Intelligence Research*. — 2021. — Vol. 1. — № 1. — P. 6–12.