CC BY

http://dx.doi.org/10.35596/1729-7648-2026-24-1-75-82

УДК 004.934

# НЕЙРОННАЯ СЕТЬ НА ОСНОВЕ СВЕРТОЧНЫХ, РЕКУРРЕНТНЫХ СЛОЕВ И МЕХАНИЗМА ВНИМАНИЯ ДЛЯ ВИЗУАЛЬНОГО РАСПОЗНАВАНИЯ РЕЧИ

Д. А. МАКАР, М. И. ВАШКЕВИЧ

*Белорусский государственный университет информатики и радиоэлектроники (Минск, Республика Беларусь)*

**Аннотация.** Визуальное распознавание речи представляет собой задачу классификации произносимых слов или букв по видеопотоку, фиксирующему движения губ. В статье представлены синтез и исследование нейросетевой архитектуры для визуального распознавания речи на основе комбинации сверточных и рекуррентных нейронных сетей с механизмом внимания. Обучение и оценка модели проводились на базе данных AVLetters2 в наиболее сложном дикторонезависимом режиме. Архитектура модели включает кодировщик на основе сверточных слоев для извлечения пространственных признаков, рекуррентные слои на основе блоков GRU для моделирования временных зависимостей и механизм внимания для выделения информативных фрагментов речевой последовательности. Для оценки точности модели проведена пятикратная перекрестная проверка. Подбор гиперпараметров модели осуществлялся на основе байесовской оптимизации, позволившей определить оптимальную конфигурацию параметров модели и процесса обучения. В результате проведенных экспериментов достигнута средняя точность распознавания 14,3 %. Анализ результатов выявил значительную вариативность качества распознавания в зависимости от характеристик дикторов (точность составила от 3,9 до 31,9 %), что указывает на необходимость дальнейшего повышения инвариантности модели к междикторским различиям.

**Ключевые слова:** визуальное распознавание речи, AVLetters2, сверточная нейронная сеть, рекуррентная нейронная сеть, механизм внимания.

**Конфликт интересов.** Авторы заявляют об отсутствии конфликта интересов.

**Для цитирования.** Макар, Д. А. Нейронная сеть на основе сверточных, рекуррентных слоев и механизма внимания для визуального распознавания речи / Д. А. Макар, М. И. Вашкевич // Доклады БГУИР. 2026. Т. 24, № 1. С. 75–82. http://dx.doi.org/10.35596/1729-7648-2026-24-1-75-82.

# NEURAL NETWORK BASED ON CONVOLUTIONAL, RECURRENT LAYERS AND AN ATTENTION MECHANISM FOR VISUAL SPEECH RECOGNITION

DARYA MAKAR, MAXIM VASHKEVICH

*Belarusian State University of Informatics and Radioelectronics (Minsk, Republic of Belarus)*

**Abstract.** Visual speech recognition is the task of classifying spoken words or letters from a video stream capturing lip movements. This paper presents the synthesis and study of a neural network architecture for visual speech recognition based on a combination of convolutional and recurrent neural networks with an attention mechanism. The model was trained and evaluated on the AVLetters2 dataset in the most challenging speaker-independent mode. The model architecture includes an encoder based on convolutional layers for extracting spatial features, recurrent layers based on GRU units for modeling temporal dependencies, and an attention mechanism for highlighting informative fragments of the speech sequence. To assess the accuracy of the model, five-fold cross-validation was performed. Model hyperparameters were selected using Bayesian optimization, which allowed us to determine the optimal configuration of the model parameters and the training process. As a result of the experiments, an average recognition accuracy of 14.3 % was achieved. Analysis of the results revealed significant variability in recognition quality depending on the characteristics of the speakers (accuracy ranged from 3.9 to 31.9 %), which indicates the need to further improve the invariance of the model to inter-speaker differences.

**Keywords:** visual speech recognition, AVLetters2, convolutional neural network, recurrent neural network, attention mechanism.

**Conflict of interests.** The authors declare that there is no conflict of interests.

**For citation.** Makar D., Vashkevich M. (2026) Neural Network Based on Convolutional, Recurrent Layers and an Attention Mechanism for Visual Speech Recognition. *Doklady BGUIR*. 24 (1), 75–82. http://dx.doi.org/10.35596/1729-7648-2026-24-1-75-82 (in Russian).

## Введение

Визуальное распознавание речи – это задача классификации произносимых слов или букв по видеопотоку, отображающему движение губ. Данная область актуальна для разработки бесшумных интерфейсов, помощи людям с нарушением речи (приобретенной потерей речи) и для мультимодальных систем распознавания речи. Важной подзадачей в этой области является классификация произносимых на видеоизображении букв исключительно на основе анализа артикуляции губ. Традиционный подход к решению такой задачи основывался на инженерном проектировании визуальных признаков (контуры губ и проч.), извлекаемых из последовательности кадров на видеоизображении [1, 2]. Однако с развитием методов глубокого обучения [3] произошел сдвиг парадигм в сторону сквозного (англ. end-to-end) обучения [4], при котором модель автоматически учится извлекать релевантные признаки из исходных данных и затем выполнять классификацию.

Основная сложность в задаче визуальных классификаций произносимых букв – необходимость одновременного учета двух аспектов:

– пространственной информации (каждый отдельный кадр содержит статистическую информацию о форме, об открытии и о конфигурации губ);

– временной информации (процесс артикуляции и жестикуляции речевого аппарата представляет собой последовательность плавно изменяющихся положений губ).

Основная информация для различения букв часто содержится именно в динамике их произношения.

В статье рассмотрена задача разработки и сквозного обучения нейронной сети на основе сверточных и рекуррентных слоев для распознавания букв, произносимых на видеоизображениях. Сверточные нейронные сети (СНС) эффективны в извлечении пространственных признаков из изображений, а рекуррентные нейронные сети (РНС) – в моделировании и классификации последовательностей. Таким образом, в модели классификации букв, произносимых на видеоизображении, СНС будет отвечать за извлечение набора визуальных признаков, которые затем будут подаваться на РНС, отвечающую за классификацию видеоизображения.

## Описание базы данных

В исследовании использовался набор данных AVLetters2 [1]. Это набор коротких видео, где пять дикторов на камеру произносят одну из 26 букв английского алфавита от A до Z (каждый диктор произносит каждую букву семь раз). Общее число произношений – 910, видеозаписи имели разрешение 1920×1080 пикселей. На рис. 1 показан пример первых восьми фреймов видеозаписи, на которой диктор произносит букву A.

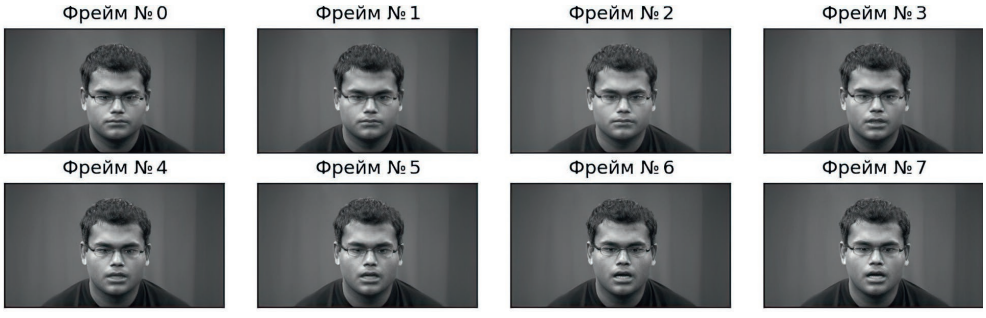


**Рис. 1.** Пример кадров видеоизображения из базы AVLetters2 (буква A)
**Fig. 1.** Example of video frames from the AVLetters2 database (letter A)

### Структура системы чтения по губам

Рассмотрим автоматизированную систему чтения по губам, обработку информации в которой можно разбить на четыре этапа:

1) предобработка видео (детектирование региона губ);
2) извлечение визуальных признаков;
3) моделирование временных зависимостей;
4) классификация.

На первом этапе предобработки видео происходит обнаружение лица и области губ. Выполнение этапов обработки со второго по четвертый предлагается объединить и выполнять в рамках одной нейросетевой модели. Для извлечения признаков из видеокадров рекомендуется использовать СНС, поскольку сети данного типа успешно применяются для детектирования объектов на изображениях [3]. А для моделирования временных зависимостей движения губ следует использовать РНС типа GRU (англ. Gated Recurrent Unit). Модели на основе GRU способны запоминать информацию из предыдущих кадров и формировать информативный вектор контекста, позволяющий выполнить классификацию слова, произносимого диктором.

### Кодировщик на основе сверточной нейронной сети

Входом нейросетевой модели являлось изображение области губ, выделенное на каждом кадре видеопотока. Далее изображение масштабировалось до размеров 64×96 пикселей. Для выделения релевантных визуальных признаков использовалась СНС, структура которой представлена на рис. 2.

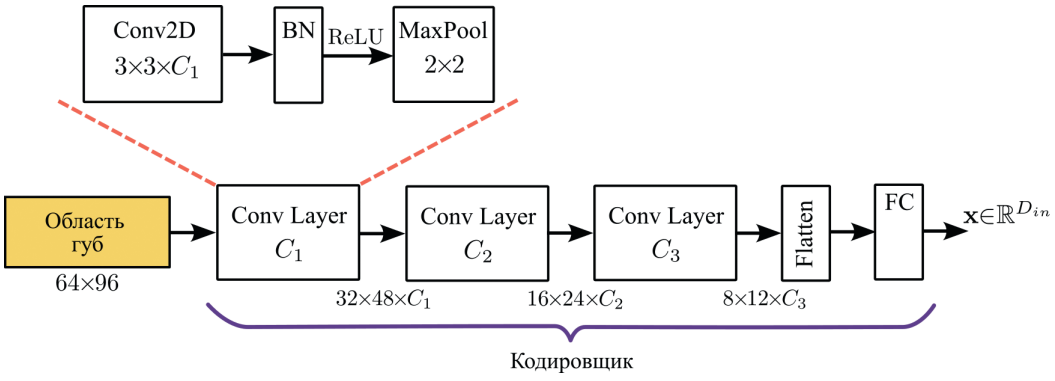


**Рис. 2.** Структура кодировщика на основе сверточной нейронной сети
**Fig. 2.** Structure of an encoder based on a convolutional neural network

Модель на рис. 2 называют кодировщиком, поскольку она выполняет роль формирования векторного представления (англ. embedding) изображения области губ с уменьшенной, по сравнению с начальной, размерностью. Кодировщик состоит из трех сверочных блоков, обозначенных на схеме как Conv Layer. Каждый блок включает вычисление двумерной свертки, нормализацию по минибатчам (англ. batch normalization, BN), вычисление активационной функции ReLU и субдискретизацию (англ. maxpooling). Сверточный блок также имеет настраиваемый гиперпараметр $C$, который определяет число ядер свертки (т. е. число обучаемых двумерных фильтров). В дальнейшем данный гиперпараметр выбирался в процессе оптимизации. После прохождения входного изображения через сверточные блоки получившийся тензор размерами $8\times12\times C_3$ «вытягивается» (англ. flatten) в одномерный вектор и подается на полносвязный слой (англ. fullyconnected layer, $FC$), который выполняет его проецирование в векторное пространство пониженной размерности $D_{in}$.

### Рекуррентная нейронная сеть для классификации

Кодировщик преобразует входную последовательность кадров в последовательность векторных представлений $\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_{N-1}$, которые затем подаются на двунаправленную рекуррентную сеть типа GRU. Структура ячейки GRU представлена на рис. 3, она является усовершенствованной архитектурой простой РНС и предназначена для эффективной обработки последовательных данных [3].

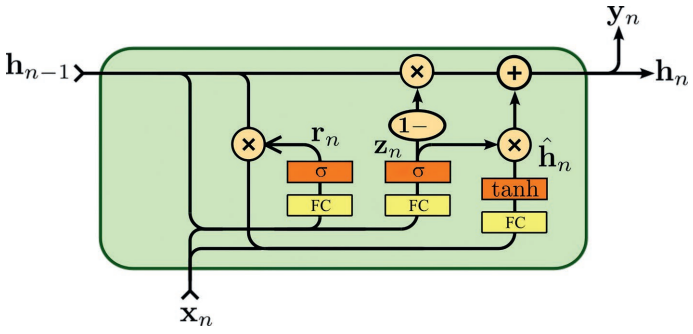


**Рис. 3.** Структура ячейки рекуррентной сети типа GRU
**Fig. 3.** The structure of the GRU cell

Входными данными для ячейки являются текущий входной вектор последовательности $\mathbf{x}_n$ и предыдущее скрытое состояние $\mathbf{h}_{n-1}$. Внутри блока GRU формируется несколько векторных переменных, которые имеют общее название «гейт» (от англ. gate). Гейты определяют процессы обработки информации (забывание и обновление), происходящие внутри ячейки.

Гейт забывания (англ. reset gate) $\mathbf{r}_n$ определяет, какая часть информации о предыдущем состоянии должна быть «забыта»:

$$\mathbf{r}_n = \sigma\left(\mathbf{W}_r\begin{bmatrix}\mathbf{h}_{n-1}\\ \mathbf{x}_n\end{bmatrix} + \mathbf{b}_r\right), \tag{1}$$

где $\mathbf{W}_r$ – матрица весов; $\mathbf{b}_r$ – вектор смещений; σ() – функция логистического сигмоида.

Гейт обновления (англ. update gate) определяет часть информации, которая должна сохраниться о предыдущем состоянии:

$$\mathbf{z}_n = \sigma\left(\mathbf{W}_z\begin{bmatrix}\mathbf{h}_{n-1}\\ \mathbf{x}_n\end{bmatrix} + \mathbf{b}_z\right), \tag{2}$$

где $\mathbf{W}_z$ – матрица весов; $\mathbf{b}_z$ – вектор смещений.

Внутри ячейки также происходит вычисление вектора-кандидата на новое состояние

$$\hat{\mathbf{h}}_n = \tanh\left(\mathbf{W}_h\begin{bmatrix}\mathbf{r}_n \odot \mathbf{h}_{n-1}\\ \mathbf{x}_n\end{bmatrix} + \mathbf{b}_h\right). \tag{3}$$

где $\mathbf{W}_h$ – матрица весов; $\mathbf{b}_h$ – вектор смещений; $\odot$ – операция поэлементного перемножения векторов.

На заключительном этапе формируется выходное текущее скрытое состояние как комбинация вектора-кандидата и предыдущего состояния

$$\mathbf{h}_n = (\mathbf{1} - \mathbf{z}_n) \odot \mathbf{h}_{n-1} + \mathbf{z}_n \odot \hat{\mathbf{h}}_n. \tag{4}$$

Для того чтобы нейронная сеть лучше формировала контекст для каждого обрабатываемого кадра, используется двунаправленная рекуррентная сеть GRU. При этом один блок GRU пропускает последовательность в прямом порядке, а второй – в обратном. Получающиеся на каждом временном шаге $n$ выходные векторы блоков конкатенируются и подаются на блок внутреннего внимания.

В [4] указывалось, что без использования механизма внимания модель очень быстро «забывает» входную последовательность и в результате имеет очень низкую производительность. Механизм внимания принимает на вход последовательность векторов $\mathbf{h}_0, \mathbf{h}_1, ..., \mathbf{h}_{N-1}$, сформированных рекуррентной сетью, и вычисляет для каждого из них весовые коэффициенты $\alpha_0, \alpha_1, ..., \alpha_{N-1}$, которые затем используются для формирования единого вектора контекста $\mathbf{c}$, имеющего фиксированную длину. Вектор контекста далее подается на классифицирующий полносвязный слой с активационной функцией softmax. Общая структура описанной модели представлена на рис. 4.

Математическая модель нейронной сети имеет следующее описание:

$$\mathbf{h}_n = \text{biGRU}\left(\mathbf{x}_n\right), \tag{5}$$

где $\mathbf{h}_n$ – выходной вектор состояния двунаправленного блока GRU, получается в результате конкатенации двух векторов, полученных от блоков GRU, выполняющих прямой и обратный проход по последовательности.

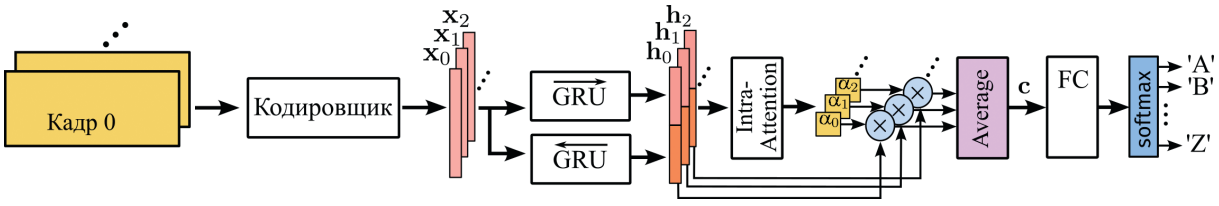


**Рис. 4.** Структура нейронной сети с механизмом внутреннего внимания для классификации произносимых букв

**Fig. 4.** The structure of a neural network with an intra-attention mechanism for classifying pronounced letters

На следующем этапе происходит вычисление вектора коэффициентов внимания

$$\left[\alpha_0, \alpha_1, ..., \alpha_{N-1}\right] = \text{IntraAttention}\left(\left[\mathbf{h}_0, \mathbf{h}_1, ..., \mathbf{h}_{N-1}\right]\right). \tag{6}$$

Полученные коэффициенты используются в модели для формирования вектора контекста как взвешенной суммы выходов двунаправленного блока GRU

$$\mathbf{c} = \frac{1}{N}\sum_{n=0}^{N-1}\alpha_n \mathbf{h}_n. \tag{7}$$

Распределение вероятности для выходной буквы генерируется при помощи полносвязного слоя с функцией активации softmax

$$\mathbf{y} = \text{softmax}(\text{FC}(\mathbf{c})), \tag{8}$$

где $\text{FC}()$ – функция, описывающая работу полносвязного слоя.

Механизм внимания в предлагаемой модели реализовывался как внутреннее внимание, выполненное по принципу, предложенному в [5]. На первом этапе вычисляется оценка (англ. score) внимания при помощи следующего выражения:

$$s_n = \mathbf{v}^{\mathrm{T}}\text{GeLU}\left(\mathbf{W}_{att}\mathbf{h}_n\right), \tag{9}$$

где $\mathbf{v} \in \mathbb{R}^{D_{in}}$ – обучаемый вектор параметров блока внимания; $\text{GeLU}(\cdot)$ – функция активации; $\mathbf{W}_{att} \in \mathbb{R}^{D_{in} \times 2D_{in}}$ – матрица линейного преобразования, выполняющая проецирование векторов $\mathbf{h}_i$ в пространство меньшей размерности.

На втором этапе выполняется нормализация оценок внимания при помощи функции softmax

$$\alpha_n = \text{softmax}\left(s_n\right). \tag{10}$$

Получаемые в (10) весовые коэффициенты $\alpha_n$ применяются в (7) для формирования вектора контекста. Коэффициенты $\alpha_n$ имеют большую величину для тех векторов $\mathbf{h}_n$, у которых большее значение с позиции последующей классификации.

### Экспериментальные исследования

Для оценки обобщающей способности модели в условиях, приближенных к реальным, был применен наиболее сложный режим тестирования – дикторонезависимый (англ. speaker-independent). В этом режиме модель обучается на одних дикторах и тестируется на других, что максимально приближено к практическому сценарию, когда система должна распознавать речь ранее не встречавшихся людей.

База AVLetters2 содержит записи пяти дикторов, поэтому для объективной оценки применялась схема пятикратной перекрестной проверки (англ. 5-fold cross-validation), где на каждой итерации данные трех дикторов использовались для обучения, данные одного диктора – для валидации, данные оставшегося диктора – для тестирования. Такая организация эксперимента обеспечивает проверку способности модели к обобщению, поскольку тестирование всегда проводится на дикторах, не представленных в обучающей выборке. Это исключает возможность «запоминания» моделью индивидуальных особенностей артикуляции конкретных дикторов и гарантиру-

ет оценку истинной способности модели распознавать визуальные паттерны речи. Описанный подход является наиболее сложным для систем визуального распознавания речи, так как требует от модели абстрагирования от индивидуальных особенностей дикторов и выделения инвариантных признаков артикуляции.

Для количественной оценки эффективности модели использовался показатель правильности (англ. accuracy), рассчитываемый как процент верно классифицированных образцов. В рамках пятикратной перекрестной проверки правильность вычислялась отдельно для каждого тестового блока данных, соответствующего разным дикторам. Результирующая метрика качества модели вычислялась как среднее значение правильности по всем тестовым блокам данных. Для оценки стабильности работы модели и вариативности результатов относительно разных дикторов рассчитывалось стандартное отклонение точности.

Поскольку задачей, которую решает модель, является классификация букв, произносимых на видеоизображении, в качестве функции потерь использовался отрицательный логарифм функции правдоподобия (англ. NLLLoss – negative log-likelihood loss). Для оптимизации параметров модели применялся метод Adam – один из самых эффективных алгоритмов градиентного спуска с адаптивным шагом [3]. С целью предотвращения переобучения и улучшения обобщающей способности модели была применена L2-регуляризация, которая заключается в добавлении к функции потерь члена $\lambda \sum_i w_i^2$, где $w_i$ – веса модели; $\lambda$ – коэффициент регуляризации (англ. weight decay). Значение $\lambda$ подбиралось в процессе оптимизации гиперпараметров.

Для регулировки скорости обучения $\eta$ модели в процессе градинетного спуска использовался планировщик скорости обучения по методу косинусного отжига с перезапуском (англ. cosine anneling with warm up restart). В данном планировщике предполагается, что скорость обучения плавно изменяется от значения $\eta_{max}$ по косинусному закону до $\eta_{min}$ в течение периода $T_0$, после чего процесс повторяется. Общее число циклов повторения определяется как $N_{epochs}/T_0$, где $N_{epochs}$ – общее число эпох обучения модели. В процессе обучения число эпох $N_{epochs}$ выбиралось равным 180, а параметр $\eta_{min} = 10^{-6}$. Показатели $T_0$ и $\eta_{max}$ существенным образом влияют на качество обучения модели, поэтому их значения выбирались в процессе оптимизации гипперпараметров.

Поиск гипперпараметров осуществлялся методом TPE (англ. tree-structured Parzen estimator), который является вариантом байесовской оптимизации, реализованным в библиотке Optuna языка Python. Особенность метода TPE в том, что в нем выполняется вероятностное моделирование пространства гипперпараметров модели и происходит целенаправленный поиск точек данного пространства, для которых ожидается максимальное значение метрики качества модели. При помощи метода TPE были найдены оптимальные значения гиперпараметров: $\eta_{max} = 4{,}63 \cdot 10^{-4}$, $\lambda = 4{,}12 \cdot 10^{-4}$, $T_0 = 10$, $C_1 = C_2 = C_3 = 16$ (число ядер сверток в слоях кодировщика); $D_{in} = 96$ (размерность выхода кодировщика); $D_h = 192$ (размерность скрытого состояния блока GRU); $p_{drop} = 0{,}40$ (дропаут-регуляризация для полносвязных слоев модели); размер батча – 32.

В результате проведения кросс-валидации было получено, что средняя точность работы модели составляет 14,3 %. Результаты по отдельным тестовым блокам данных представлены на рис. 5.

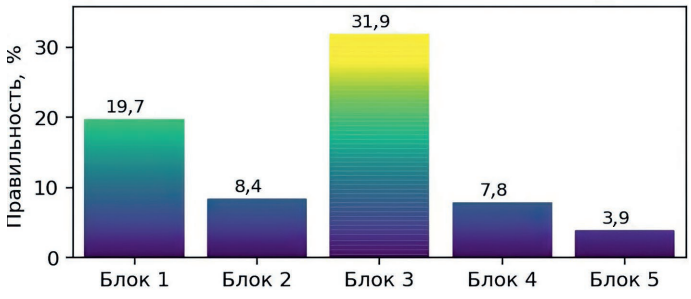


**Рис. 5.** Результат перекрестной проверки по тестовым блокам данных
**Fig. 5.** Cross-validation result on test data blocs

Минимальная точность модели была зафиксирована на пятом тестовом блоке данных и составила 3,9 %, что статистически не отличается от случайного выбора при 26 классах (ожидаемая точность случайного классификатора ≈3,85 %). Детальный анализ показал, что низкая точность на последнем блоке данных связана с тем, что дикторы в обучающем наборе имели светлый тип

кожи, а диктор в тестовом наборе был темнокожим. Таким образом, плохой результат объясняется тем, что модель не смогла корректно выделить признаки для диктора с цветом кожи, отличающимся от дикторов в тренировочной выборке.

Тем не менее предложенный подход демонстрирует значительное улучшение по сравнению с методом, представленным в [1]. При использовании активной модели внешнего вида (англ. Active Appearance Model, AMM) для извлечения признаков и скрытых марковских моделей для временного моделирования в [1] была достигнута точность 8,0 %, в то время как разработанная в данном исследовании модель показывает среднюю точность 14,3 %, что соответствует относительному улучшению на 6,3 %.

Достижение высокой точности распознавания на базе данных AVLetters2 в условиях дикторонезависимого тестирования представляет значительную сложность, что подтверждается ограниченным количеством публикаций с соответствующими результатами. Показательно, что в исследовании [6], посвященном сравнительному анализу методов визуального распознавания речи, авторы приводят оценки точности для дикторозависимого сценария, но не включают результаты для дикторонезависимого режима тестирования.

Также существуют подходы, которые демонстрируют более высокую точность распознавания для базы AVLetters2, чем полученную в настоящем исследовании. Так, в [7] предложена модель на основе полносвязных слоев и двунаправленных рекуррентных сетей типа LSTM, которая позволила получить среднюю точность, равную 36,8 %. Особенностью модели [7] является то, что в нее явным образом подаются как текущий кадр видеоизображения, так и разностный кадр. Поэтому можно сделать вывод, что предварительная обработка видеокадров может иметь существенное влияние на результативность работы модели.

**Заключение**

1. Получена и исследована архитектура нейронной сети для визуального распознавания речи, объединяющая сверточные слои для извлечения пространственных признаков из области рта, рекуррентные слои типа GRU для моделирования временных зависимостей и механизм внимания для выделения информативных фрагментов в речевых последовательностях.

2. Экспериментальная оценка модели проводилась на базе данных AVLetters2 в наиболее сложном дикторонезависимом режиме с использованием пятикратной перекрестной проверки. Проведенный анализ показал, что предложенная архитектура обеспечивает среднюю точность распознавания 14,3 %. Исследование выявило значительную вариативность качества распознавания в зависимости от характеристик диктора (от 3,9 до 31,9 %), что указывает на чувствительность модели к индивидуальным особенностям артикуляции и внешнего вида. Такой почти восьмикратный разброс результатов свидетельствует о том, что модель хорошо адаптируется к «легким» дикторам, но терпит затруднения на «сложных».

3. Оптимизация гиперпараметров с использованием байесовского подхода позволила определить оптимальную конфигурацию модели, включая размерность признакового пространства, параметры регуляризации и архитектурные характеристики. Несмотря на это, абсолютное значение точности остается низким для практического применения, тем не менее является сопоставимым с результатами, достигаемыми на ранних этапах разработки в данной области.

4. Полученные результаты демонстрируют перспективность использования комбинированных архитектур «сверточные нейронные сети – рекуррентные нейронные сети» с механизмами внимания для задач визуального распознавания речи, однако также подчеркивают необходимость дальнейшего повышения инвариантности моделей к междикторской вариативности для обеспечения устойчивой работы в реальных условиях. В качестве перспективных направлений для повышения устойчивости модели можно выделить: аугментацию данных, направленную на симуляцию различных артикуляционных паттернов; применение методов адаптации к диктору; исследование более сложных механизмов внимания, способных выделять и игнорировать второстепенные и индивидуальные особенности; использование предобученных моделей для извлечения визуальных признаков.

**Список литературы**

1. The Challenge of Multispeaker Lip-Reading / S. Cox [et al.] // International Conference on Auditory-Visual Speech Processing. 2008. P. 179–184.

2. Extraction of Visual Features for Lipreading / I. Matthews [et al.] // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002. Vol. 24, No 2. P. 198–213.
3. Николенко, С. Глубокое обучение. Погружение в мир нейронных сетей / С. Николенко, А. Кадурин, Е. Архангельская. СПб.: Питер, 2020.
4. Lip Reading Sentences in the Wild / S. J. Chung [et al.] // Conference on Computer Vision and Pattern Recognition. 2017. https://doi.org/10.48550/arXiv.1611.05358.
5. Cheng, J. Long Short-Term Memory-Networks for Machine Reading / J. Cheng, L. Dong, M. Lapata // EMNLP 2016 Conference. https://doi.org/10.48550/arXiv.1601.06733.
6. Pei, Y. Unsupervised Random Forest Manifold Alignment for Lipreading / Y. Pei, T.-K. Kim, H. Zha // IEEE International Conference on Computer Vision. 2013. P. 129–136.
7. End-to-End Visual Speech Recognition for Small-Scale Datasets / S. Petridis [et al.] // Pattern Recognition Letters. 2020. P. 131, 421–427. https://doi.org/10.48550/arXiv.1904.01954.

Поступила 12.10.2025 Принята в печать 23.12.2025

**References**

1. Cox S., Harvey R., Lan Y., Newman J. L., Theobald B.-J. (2008) The Challenge of Multispeaker Lip-Reading. *International Conference on Auditory-Visual Speech Processing.* 179–184.
2. Matthews I., Cootes T. F., Bangham J. A., Cox S., Harvey R. (2002) Extraction of Visual Features for Lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 24 (2). 198–213.
3. Nikolenko S., Kadurin A., Arkhangelskaya E. (2020) *Deep Learning: A Dive into the World of Neural Networks*. St. Petersburg, Piter Publ. (in Russian).
4. Chung S. J., Senior A., Vinyals O., Zisserman A. (2017) Lip Reading Sentences in the Wild. *Computer Vision and Pattern Recognition.* https://doi.org/10.48550/arXiv.1611.05358.
5. Cheng J., Dong L., Lapata M. (2016) Long Short-Term Memory-Networks for Machine Reading. *EMNLP 2016 Conference*. https://doi.org/10.48550/arXiv.1601.06733.
6. Pei Y., Kim T.-K., Zha H. (2013) Unsupervised Random Forest Manifold Alignment for Lipreading. *IEEE International Conference on Computer Vision*. 129–136.
7. Petridis S., Wang Y., Ma P., Li Z., Pantic M. (2020) End-to-End Visual Speech Recognition for Small-Scale Datasets. *Pattern Recognition Letters*. 131, 421–427. https://doi.org/10.48550/arXiv.1904.01954.

Received: 12 October 2025 Accepted: 23 December 2025

**Вклад авторов**

Макар Д. А. выполнила анализ результатов и подготовила рукопись статьи.

Вашкевич М. И. поставил научную задачу, определил методологию исследования, курировал проектирование архитектуры модели, участвовал в интерпретации результатов и подготовке текста статьи.

**Authors' contribution**

Makar D. performed the analysis of the results and prepared the manuscript of the article.

Vashkevich M. formulated the scientific problem, defined the research methodology, supervised the model architecture design, participated in the interpretation of the results, and the preparation of the text of the article.

**Сведения об авторах**

**Макар Д. А.,** асп. каф. электронных вычислительных средств, Белорусский государственный университет информатики и радиоэлектроники

**Вашкевич М. И.,** д-р техн. наук, проф. каф. электронных вычислительных средств, Белорусский государственный университет информатики и радиоэлектроники

**Адрес для корреспонденции**

220013, Республика Беларусь,
Минск, ул. П. Бровки, 6
Белорусский государственный университет
информатики и радиоэлектроники
Тел.: +375 17 293-84-20
E-mail: vashkevich@bsuir.by
Вашкевич Максим Иосифович

**Information about the authors**

**Makar D.,** Postgraduate of the Electronic Computing Facilities Department, Belarusian State University of Informatics and Radioelectronics

**Vashkevich M.,** Dr. Sci. (Tech.), Professor at the Electronic Computing Facilities Department, Belarusian State University of Informatics and Radioelectronics

**Address for correspondence**

220013, Republic of Belarus,
Minsk, P. Brovki St., 6
Belarusian State University
of Informatics and Radioelectronics
Tel.: +375 17 293-84-20
E-mail: vashkevich@bsuir.by
Vashkevich Maxim