ОБРАБОТКА СИГНАЛОВ, ИЗОБРАЖЕНИЙ, РЕЧИ, ТЕКСТА И РАСПОЗНАВАНИЕ ОБРАЗОВ

SIGNAL, IMAGE, SPEECH, TEXT PROCESSING AND PATTERN RECOGNITION

УДК 004.94 + 534.77
https://doi.org/10.37661/1816-0301-2026-23-1-69-87

Поступила в редакцию | Received 26.01.2026
Подписана в печать | Accepted 23.02.2026
Опубликована | Published 31.03.2026

Распознавание эмоций по речи на основе LSTM-сетей с мультивекторным механизмом внимания

Д. В. Краснопрошин✉, М. И. Вашкевич
✉E-mail: daniil.krasnoproshin@gmail.com

Белорусский государственный университет информатики и радиоэлектроники, ул. П. Бровки, 6, Минск, 220013, Беларусь

Аннотация

Цели. Целью исследования является повышение точности распознавания эмоций по речевому сигналу с помощью моделей на основе рекуррентных нейронных сетей (РНС) с долгой краткосрочной памятью.

Методы. В работе предложен мультивекторный механизм внимания для РНС на основе ячеек LSTM. Данный механизм представляет собой обобщение классического мягкого внимания и позволяет модели одновременно анализировать различные аспекты временны́х зависимостей. Предложенные архитектуры РНС применены к задаче распознавания эмоций по речевому сигналу. В качестве входных данных использовались последовательности мел-частотных кепстральных коэффициентов, отражающих частотно-временную структуру речевого сигнала. Эксперименты проводились на общедоступном наборе данных RAVDESS. Для автоматизированного подбора оптимальных гиперпараметров моделей использовался метод байесовской оптимизации.

Результаты. Результаты экспериментов с LSTM-сетями, имеющими различную размерность скрытого состояния (64, 96, 128), показывают, что применение мультивекторного механизма внимания приводит к статистически значимому улучшению среднего значения точности на величину от 0,88 до 1,56 %.

Заключение. Полученные результаты подтверждают целесообразность использования предложенного механизма мультивекторного внимания в архитектурах LSTM-сетей для задачи классификации эмоций в речи.

Ключевые слова: обработка речи, распознавание эмоций, глубокое обучение, рекуррентные нейронные сети, механизм внимания

Для цитирования. Краснопрошин, Д. В. Распознавание эмоций по речи на основе LSTM-сетей с мультивекторным механизмом внимания / Д. В. Краснопрошин, М. И. Вашкевич // Информатика. – 2026. − Т. 23, № 1. – С. 69–87. – https://doi.org/10.37661/1816-0301-2026-23-1-69-87.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Speech emotion recognition based on LSTM networks with multi-vector attention

Daniil V. Krasnoproshin✉, Maxim I. Vashkevich
✉E-mail: daniil.krasnoproshin@gmail.com

Belarusian State University of Informatics and Radioelectronics,
st. Brovki, 6, Minsk, 220013, Belarus

Abstract

Objectives. Improvement of speech emotion recognition accuracy using Long Short-Term Memory (LSTM) recurrent neural network (RNN) models.

Methods. The paper proposes a multi-vector attention mechanism for LSTM-based RNNs. This mechanism generalizes the classical soft attention and allows the model to simultaneously analyze different aspects of temporal dependencies. The proposed RNN architectures were applied to the task of speech emotion recognition. Input data consisted of sequences of mel-frequency cepstral coefficients (MFCCs), which reflect the time-frequency structure of the speech signal. Experiments were conducted on the publicly available RAVDESS dataset. Bayesian optimization was employed for automated hyperparameter tuning of the models.

Results. The experimental results with LSTM networks having different hidden state dimensions (64, 96, 128) demonstrate that the application of the multi-vector attention mechanism leads to a statistically significant improvement in the average accuracy metric (UAR) by 0.88 to 1.56 %.

Conclusion. The obtained results confirm the effectiveness of using the proposed multi-vector attention mechanism in LSTM-based architectures for speech emotion classification.

Keywords: speech processing, emotion recognition, deep learning, recurrent neural networks, attention mechanism

For citation. Krasnoproshin D. V., Vashkevich M. I. *Speech emotion recognition based on LSTM networks with multi-vector attention*. Informatika [*Informatics*], 2026, vol. 23, no. 1, pp. 69–87 (In Russ.). https://doi.org/10.37661/1816-0301-2026-23-1-69-87.

Conflict of interests. The authors declare of no conflict of interest.

Введение

Распознавание эмоций по речи является одной из важных задач в области обработки естественного языка и разработки человеко-машинных интерфейсов [1]. Развитие этой области обусловлено широким спектром практических приложений – от интеллектуальных голосовых ассистентов до анализа поведения пользователей и оценки эмоционального состояния в системах дистанционного обучения и медицинского обслуживания. Несмотря на существенный прогресс в области обработки естественного языка, достигнутый благодаря применению глубоких нейронных сетей, задача точного и устойчивого распознавания эмоций по речи остается сложной из-за высокой вариативности речевого сигнала, индивидуальных особенностей дикторов и контекстной зависимости эмоциональных проявлений [1–3].

Традиционные подходы к распознаванию эмоций, основанные на экспертных признаках (например, мел-частотных кепстральных коэффициентах, спектральных моментах и просодических характеристиках), демонстрируют приемлемую точность на небольших и «чистых» наборах данных [4]. Однако их обобщающая способность резко

снижается при работе с большими объемами зашумленных данных, что связано с ограниченной гибкостью параметризации и высокой чувствительностью к вариациям акустических условий [5].

С развитием глубокого обучения в задаче распознавания эмоций активно применяются сверточные нейронные сети (СНС), работающие с визуальным представлением речи – спектрограммами. Такие модели эффективно извлекают спектрально-временные признаки, достигая высоких результатов [6, 7]. Вместе с тем использование двумерных представлений сигнала приводит к частичной потере тонкой временно́й динамики, так как последовательная природа речи искусственно сводится к пространственной структуре, характерной для изображений. Данное ограничение особенно критично для задач, где временны́е зависимости (например, интонационные переходы) играют важную роль.

Современные исследования фокусируются на применении для задачи распознавания эмоций моделей с архитектурой трансформер, предобученных на аудиоданных [2, 8]. Благодаря способности моделировать долгосрочные зависимости трансформеры позволяют достигать высокой точности. Однако их практическое внедрение сдерживается высокой вычислительной сложностью: модели с десятками и сотнями миллионов параметров требуют значительных ресурсов для обучения и исполнения (англ. inference), что делает их непригодными для реального времени и устройств с ограниченным вычислительным ресурсом.

Учитывая указанные ограничения, перспективным направлением для разработки практичных систем распознавания эмоций являются РНС. Различные архитектуры РНС изначально ориентированы на обработку последовательных данных, что позволяет эффективно учитывать временну́ю динамику речевого сигнала [9]. Параметризируя аудиопоток в виде последовательности векторов, РНС сохраняют информацию о структуре и динамике речевого сигнала. Ключевое преимущество данного подхода – значительно меньшее количество параметров по сравнению с трансформерами при сохранении способности к моделированию контекстных зависимостей. Это делает РНС оптимальным выбором для сценариев, где важны скорость обработки и возможность запуска модели на мобильном или портативном устройстве.

В данной работе исследуется система распознавания эмоций на основе РНС с долгой краткосрочной памятью (англ. LSTM, Long Short-Term Memory) [10]. LSTM-сеть представляет собой ячейку памяти, которая посредством механизмом управляющих затворов (англ. gate) изменяет во времени свое состояние в зависимости от текущего входного значения и предыдущего состояния. Наличие в структуре нескольких управляющих затворов позволяет LSTM-сети моделировать долгосрочные зависимости, продуктивно решая проблему «исчезающего» градиента, которая характерна для классических РНС [9]. Тем не менее, несмотря на способность LSTM-сетей моделировать временны́е зависимости, такие архитектуры не всегда эффективно фокусируются на наиболее значимых участках сигнала, что важно для правильного распознавания эмоций. Механизмы внимания (англ. attention mechanism) позволяют решать эту проблему, предоставляя модели возможность адаптивно выделять релевантные части входной последовательности. В частности, механизм мягкого внимания (англ. soft attention) показал свою эффективность в задаче классификации эмоций [11]. В настоящей работе предлагается новый ме-

ханизм мультивекторного внимания, который используется для формирования улучшенного вектора контекста из последовательности выходов LSTM-сети. Механизм мультивекторного внимания является обобщением механизма мягкого внимания, который также ранее использовался при решении задачи классификации эмоций с применением РНС [3, 11].

Целью исследования являются разработка и количественная оценка мультивекторного механизма внимания для LSTM-сетей в контексте задачи распознавания эмоций в речи. Предполагается, что увеличение числа векторов внимания позволит модели формировать улучшенное представление эмоционального содержания аудиосигнала, что приведет к повышению точности классификации эмоциональных состояний при сохранении умеренной вычислительной сложности.

Извлечение признаков

Для параметризации речевого сигнала в работе используются мел-частотные кепстральные коэффициенты (МЧКК). Основная идея метода расчета МЧКК заключается в приближенном моделировании особенностей восприятия звука в слуховой системе человека. Данный метод относится к кратковременному спектральному анализу речевых сигналов, при котором исходный сигнал разбивается на фреймы длительностью 10–30 мс. Такой временной интервал выбран исходя из предположения о квазистационарности речевого сигнала на указанных отрезках, что позволяет применять спектральные методы анализа. Преобразования речевого сигнала в процессе расчета МЧКК показаны на рис. 1.

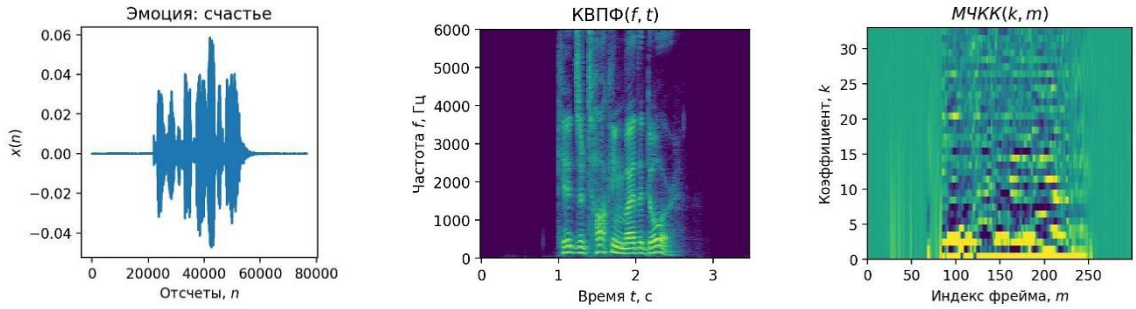

Рис. 1. Преобразования речевого сигнала в последовательность векторов с использованием МЧКК
Fig. 1. Transformations of speech signal into sequence of vectors using MFCC

Согласно рис. 1 процесс извлечения МЧКК включает следующие шаги [12]:

1) сигнал разделяется на короткие фреймы длины L с перекрытием h_{size};

2) для каждого фрейма вычисляется кратковременное преобразование Фурье (КВПФ) и находится квадрат модуля КВПФ;

3) выполняется переход от КВПФ к мел-спектрограмме (энергия сигнала из шкалы герц переводится в мел-шкалу, отражающую свойства человеческого слуха);

4) рассчитывается логарифм от энергии сигнала в мел-частотных полосах;

5) применяется декоррелирующее преобразование, в качестве которого выступает дискретное косинусное преобразование II типа (ДКП-II).

В результате параметризации речевого сигнала при помощи МЧКК образуется компактное представление речевого сигнала в последовательность векторов:

$$\mathbf{X} = [\mathbf{x}_0, \quad \mathbf{x}_1, \quad \ldots \quad \mathbf{x}_{T-1}], \mathbf{x}_t \in \mathbb{R}^n, \tag{1}$$

где T – длина последовательности (количество фреймов сигнала).

В данной работе обрабатывались речевые сигналы с частотой дискретизации 48 кГц, размер фрейма выбирался равным $L = 1024$, а перекрытие – $h_{size} = 512$ отсчетам. Размерность вектора МЧКК выбиралась равной $n = 34$.

LSTM-сеть с механизмом внимания

В настоящем исследовании разрабатывался мультивекторный механизм внимания, который основывается на способности LSTM-сети моделировать временны́е зависимости в последовательных данных, таких как речь. LSTM является улучшенной разновидностью РНС, которая решает проблему исчезающего градиента. LSTM-сети моделируют временны́е зависимости в последовательности $\mathbf{X}$ с помощью скрытого состояния $\mathbf{h}_t$, которое обновляется на каждом временно́м шаге и передается для вычисления следующего состояния, что позволяет сохранять информацию о предыдущих элементах последовательности. Математически LSTM-ячейку можно описать набором уравнений для вычисления состояния и выхода на каждом временно́м шаге t:

$$\begin{aligned}
\mathbf{f}_t &= \sigma(\mathbf{W}_f \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f), \\
\mathbf{i}_t &= \sigma(\mathbf{W}_i \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i), \\
\hat{\mathbf{C}}_t &= \tanh(\mathbf{W}_C \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_C), \\
\mathbf{C}_t &= \mathbf{f}_t \odot \mathbf{C}_{t-1} + \mathbf{i}_t \odot \hat{\mathbf{C}}_t, \\
\mathbf{o}_t &= \sigma(\mathbf{W}_o \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o), \\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{C}_t),
\end{aligned} \tag{2}$$

где $\mathbf{f}_t$, $\mathbf{i}_t$, $\mathbf{o}_t \in \mathbb{R}^n$ – забывающий, входной и выходной затворы соответственно; $\mathbf{C}_t \in \mathbb{R}^n$ – внутреннее состояние LSTM-ячейки; $[\mathbf{h}_{t-1}, \mathbf{x}_t]$ – конкатенация предыдущего вектора скрытого состояния ячейки на шаге $t-1$ и входного вектора на шаге t; $\tanh(\cdot)$ – функция активации гиперболический тангенс; $\sigma(\cdot)$ – функция активации логистического сигмоида; $\mathbf{W}_f$, $\mathbf{W}_i$, $\mathbf{W}_C$, $\mathbf{W}_o \in \mathbb{R}^{n \times 2n}$ – обучаемые матрицы весов; $\mathbf{b}_f$, $\mathbf{b}_i$, $\mathbf{b}_C$, $\mathbf{b}_t \in \mathbb{R}^n$ – обучаемые векторы смещений; $\odot$ – операция поточечного умножения векторов.

Рассмотрим кратко функционирование модели (2). Ее сердцевиной является вектор состояния ячейки $\mathbf{C}_t$ – долгосрочная память, которая хранит данные на протяжении всех временны́х шагов. Вектор внутреннего состояния $\mathbf{h}_t$ – это текущая (или краткосрочная) память, значение которой подается на выход модели. Таким образом, $\mathbf{h}_t$ хранит информацию, которая актуальна для текущего шага. В структуре LSTM-ячейки важную роль играет понятие затворов. Под затворами понимаются векторные переменные, которые определяют важность или, наоборот, неважность информации, хранимой в векторе-состоянии ячейки $\mathbf{C}_t$. Например, затвор $\mathbf{f}_t$ определяет, какую часть информации своего

предыдущего состояния LSTM-ячейка должна «забыть», а затвор $\mathbf{o}_t$ определяет, какая часть информации долгосрочной памяти важна на текущем шаге.

Таким образом, классическая LSTM-сеть принимает на вход последовательность входных векторов $\mathbf{x}_t$ и генерирует на выходе последовательность выходных векторов $\mathbf{h}_t$. Требуется доопределить ее структуру, чтобы сеть была способна решать задачу классификации эмоций. С точки зрения классификации задач [9], которые решаются при помощи РНС, распознавание эмоций может быть отнесено к классу «один ко многим» (рис. 2).

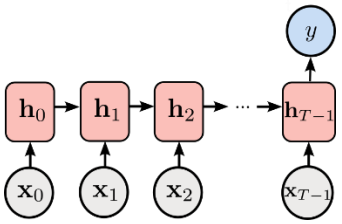

Рис. 2. Классификация эмоций при помощи РНС
Fig. 2. Emotion classification by RNN using many-to-one scheme

Диаграмма на рис. 2 показывает, что метка эмоции может быть получена путем классификации с использованием последнего скрытого состояния, полученного в LSTM-сети. Для этого достаточно вектор $\mathbf{h}_{T-1}$ подать на полносвязный слой с функцией классификации $\mathrm{softmax}(\cdot)$. Ошибки, полученные в данном слое, затем распространяются обратно к началу последовательности $\mathbf{x}_t$. В литературе данный подход известен как обучение по последнему фрейму (англ. final-frame training) [2]. Его основная идея заключается в предположении, что последнее скрытое состояние РНС содержит достаточную для корректной классификации эмоций информацию, извлеченную из всей речевой последовательности. При таком подходе РНС рассматривается в качестве адаптивного средства, преобразующего переменную по длине последовательность векторов $\mathbf{X}$ в вектор фиксированной размерности $\mathbf{h}_{T-1}$.

Однако, как показано в работе [3], практическая эффективность данного подхода оказывается ограниченной. Это объясняется тем, что, хотя LSTM-архитектура и обладает улучшенной способностью к сохранению долгосрочных зависимостей по сравнению с обычными РНС, механизм забывания все же приводит к постепенной потере информации. В результате в последнем скрытом состоянии данные, относящиеся к начальным сегментам последовательности, могут быть представлены недостаточно полно. Для решения этой проблемы был предложен метод временно́го усреднения выходов LSTM-сети [11, 13], который позволяет агрегировать информацию со всех временны́х шагов. Под агрегацией в данном случае понимается глобальное усреднение всех выходных состояний LSTM-сети.

Таким образом, можно сделать вывод, что LSTM-сети обладают способностью учитывать долговременные зависимости в последовательных данных, однако их внутренняя структура не предполагает механизма приоритизации отдельных элементов временно́го контекста [3, 11].

Улучшить работу LSTM-сети можно за счет внедрения в ее структуру механизма внимания. Его добавление позволяет модели вычислять веса «важности» для различных временны́х шагов, что способствует получению более информативного выходного вектора, на основании которого выполняется классификация эмоции.

В настоящем исследовании в качестве базовой модели рассматривается LSTM-сеть с механизмом мягкого внимания, предложенная в работе [3]. В данной модели на основании последовательности выходных состояний $\mathbf{h}_t$ рассчитывается взвешенная сумма скрытых состояний, которая называется вектором контекста:

$$\mathbf{h}_{wp} = \sum_{t=0}^{T-1} \alpha_t \mathbf{h}_t , \tag{3}$$

где α_t – весовые коэффициенты, отражающие значимость вектора состояния $\mathbf{h}_t$ в формировании вектора контекста. Обозначение *wp* берется от англ. weighted pooling – взвешенное усреднение.

На основании вектора контекста $\mathbf{h}_{wp}$ выполняется классификация эмоции, т. е. $\mathbf{h}_{wp}$ подается на полносвязный слой с активационной функцией softmax. Общая схема описанного подхода изображена на рис. 3.

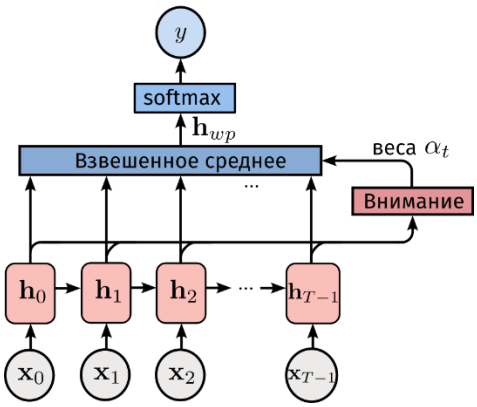

Рис. 3. Классификация эмоций при помощи РНС с механизмом внимания
Fig. 3. Emotion classification using RNN with attention mechanism

В оригинальной работе [3] для формирования весов использовался механизм мягкого внимания (англ. soft attention):

$$\alpha_t = \text{softmax}(e_t) = \frac{\exp(e_t)}{\sum_{k=0}^{T-1} \exp(e_k)} , \tag{4}$$

где $e_t = \mathbf{u}^T \mathbf{h}_t$ – оценка внимания (англ. attention score), $\mathbf{u}$ – вектор внимания (обучаемый параметр).

Добавление механизма мягкого внимания существенно улучшает работу LSTM-сети за счет того, что вектор контекста $\mathbf{h}_{wp}$ адаптивно выделяет из выходной последовательности состояний те компоненты, которые наиболее важны для задачи распознавания эмоций.

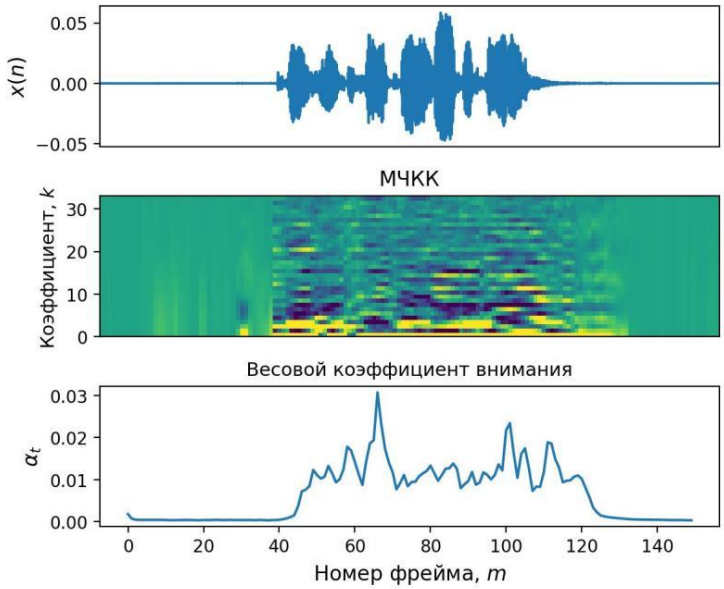

Рис. 4. Пример вычисления весов внимания
Fig. 4. Example of local attention weights calculation

На рис. 4 показаны речевой сигнал, последовательность коэффициентов МЧКК, а также веса внимания, рассчитанные обученной LSTM-сетью. Видно, что после обучения механизм внимания приобретает ожидаемую селективность. Например, фреймам, которые относятся к паузам, присваиваются очень низкие значения весов внимания, а фреймам, несущим эмоциональную окраску, – высокие значения.

Мультивекторный механизм внимания

Рассмотренный выше механизм мягкого внимания можно интерпретировать следующим образом. Механизм в неявном виде предполагает, что в векторном пространстве, к которому принадлежат векторы состояний $\mathbf{h}_t$, имеется некоторое направление, соответствующее вектору внимания $\mathbf{u}$. При этом значимость вектора $\mathbf{h}_t$ тем больше, чем больше его проекция на вектор внимания $\mathbf{u}$. Таким образом, в процессе обучения модель стремится «вытянуть» все релевантные векторы состояний $\mathbf{h}_t$ вдоль направления вектора внимания $\mathbf{u}$.

В настоящей работе предлагается мультивекторный механизм мягкого внимания, согласно которому оценка внимания рассчитывается по формуле

$$e_t = \max\left(\begin{pmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \vdots \\ \mathbf{u}_{N_{att}}^T \end{pmatrix} \mathbf{h}_t\right), \quad (5)$$

где N_{att} – число векторов внимания; $\mathbf{u}_i$ – набор векторов внимания, каждый из которых отвечает за независимое направление в пространстве векторов скрытых состояний.

Следует отметить, что мультивекторный механизм внимания включает в себя обычный механизм мягкого внимания как частный случай, когда $N_{att}=1$.

Предполагается, что различные векторы внимания $\mathbf{u}_i$ могут отвечать за различные проявления эмоций в последовательности $\mathbf{h}_t$. Таким образом, мультивекторный механизм внимания более эффективно использует пространство векторов скрытых состояний $\mathbf{h}_t$, формируя в нем не одно, а несколько базовых направлений, которые способны повлиять на получение более информативного контекстного вектора $\mathbf{h}_{wp}$.

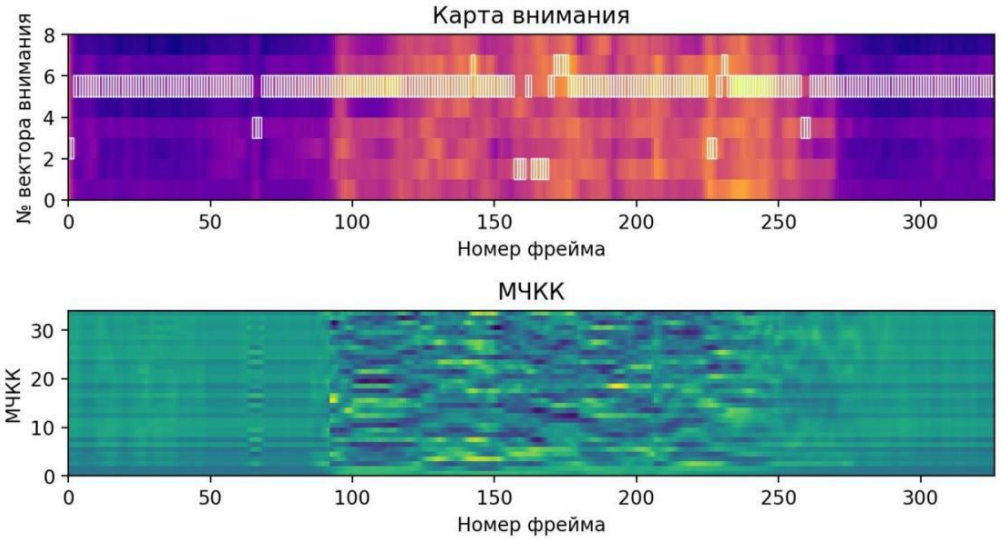

Рис. 5. Иллюстрация работы механизма мультивекторного внимания
Fig. 5. Illustration of the multi-head attention mechanism

На рис. 5 показана работа обученного механизма мультивекторного внимания. На верхней панели рис. 5 изображена карта внимания, вычисленная для модели с восемью векторами внимания. Белым контуром выделены оценки внимания, которые имеют максимальное значение на текущем временно́м шаге. Рис. 5 демонстрирует, что чаще всего максимальное значение оценки внимания дает проекция на вектор $\mathbf{u}_6$. Однако в отдельные моменты времени максимальные оценки дают проекции $\mathbf{h}_t$ на другие векторы внимания. Таким образом, применение механизма мультивекторного внимания позволяет LSTM-сети формировать более информативно насыщенное представление об эмоции, содержащейся в речевом сигнале.

Описание базы данных

При проведении исследования использовался набор данных Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). RAVDESS содержит записи 24 актеров (12 мужчин и 12 женщин). Для каждого актера имеется 104 различных сообщения (60 речевых высказываний и 44 песенных). В рамках данной работы использована только часть базы RAVDESS, содержащая речевые высказывания. В этой части содержатся 1440 файлов в формате wav (16 бит, 48 кГц) – 60 записей на каждого из 24 актеров. Речевые эмоции включают: нейтральность, спокойствие, счастье, грусть, гнев, страх,

удивление и отвращение. Все эмоциональные состояния, кроме нейтрального, озвучивались на двух уровнях эмоциональной громкости (нормальная и повышенная). Актеры повторяли каждую вокализацию дважды.

Постановка эксперимента

Для экспериментальной оценки мультивекторного механизма внимания была реализована однослойная РНС на основе LSTM-ячеек. В качестве входных данных использовались нормализованные МЧКК размерностью $n = 34$. Исследование проводилось для различных размерностей скрытого состояния РНС: $h_{size} \in \{64, 96, 128\}$. Количество векторов внимания варьировалось в дискретном множестве значений $N_{att} \in \{1, 2, 4, 8, 16, 32, 64\}$. Для модели со скрытой размерностью 128 дополнительно были протестированы конфигурации с 128 и 256 векторами внимания.

Разработанные LSTM-сети с механизмом внимания обучались с помощью метода оптимизации Adam [9]. В качестве функции потерь использовалась перекрестная энтропия, которая оптимальна для задач многоклассовой классификации и обеспечивает эффективное вычисление градиентов при работе с вероятностными распределениями:

$$CE = \sum_{n=0}^{N-1} \sum_{i=0}^{N_c-1} y_{true}^{(i,n)} \cdot \ln(y_{pred}^{(i,n)}), \quad (6)$$

где N – число обучающих примеров, N_c – количество классов, y_{true} – истинное распределение меток, y_{pred} – предсказанные вероятности классов.

Инициализация весов модели выполнялась с учетом архитектурных особенностей LSTM-сети: веса инициализировались методом Ксавье с нормальным распределением [9], что обеспечивает сохранение дисперсии сигналов при прямом и обратном распространении через слой; смещения всех затворов LSTM-ячейки инициализировались нулевыми значениями, за исключением затвора забывания, смещения которого устанавливались равными единице для обеспечения начального сохранения информации в ячейке памяти.

Оптимизация гиперпараметров проводилась с использованием фреймворка Optuna, который обеспечивает эффективный автоматизированный поиск конфигураций модели [14]. Этот подход динамически формирует пространство поиска в процессе оптимизации. В качестве алгоритма сэмплирования значений гиперпараметров Optuna использует метод древовидных оценок Парзена (англ. TPE, Tree-Structured Parzen Estimator). TPE – это байесовский метод оптимизации, который моделирует плотности вероятности высокоэффективных и низкоэффективных конфигураций гиперпараметров по отдельности [15]. Сравнивая эти распределения, TPE интеллектуально направляет поиск в перспективные области пространства гиперпараметров, значительно ускоряя сходимость по сравнению со случайными методами или методами поиска по сетке.

Оптимизация гиперпараметров включала пять основных параметров: скорость обучения (выборка в логарифмическом масштабе в диапазоне $\eta = [3 \cdot 10^{-5}, \ 2 \cdot 10^{-4}]$), коэффициент затухания весов (выборка в логарифмическом масштабе в диапазоне $\lambda = [2 \cdot 10^{-5}, \ 2 \cdot 10^{-2}]$), отсечение (англ. dropout) в полносвязных слоях (равномерная выборка в диапазоне $p_{drop} = [10^{-1}, \ 5 \cdot 10^{-1}]$), количество циклов отжига в планировщике ко-

синусного отжига $T_0 \in \{1, 2, 3, 5, 10\}$ и размер пакета $batch\_size \in \{16, 32, 64\}$. Каждая модель обучалась в течение 200 эпох, поскольку точность на валидационном наборе после этого практически не улучшалась.

Оптимизация выполнялась для базовой модели с одним вектором внимания. Полученные оптимальные параметры использовались далее для всех экспериментов с различным числом векторов внимания. Для каждой конфигурации модели обучение повторялось 10 раз с разными начальными значениями весов. Для итоговой оценки качества модели вычисляли среднее невзвешенное значение полноты (англ. unweighted average recall, UAR):

$$\mathrm{UAR} = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{A_{i,i}}{\sum_{j=1}^{N_c} A_{i,j}}, \tag{7}$$

где A – матрица ошибок (англ. confusion matrix), N_c – количество классов.

Метрика UAR применяется для оценки общей эффективности моделей многоклассовой классификации. В UAR полнота, полученная для каждого класса, вносит равный вклад в итоговую оценку независимо от распределения данных. Значения UAR нормированы в диапазоне [0, 1], где значения, приближающиеся к единице, соответствуют более высокой производительности классификатора. Для статистически надежной оценки модели применялась процедура перекрестной проверки по пяти блокам (англ. 5-fold cross-validation). В исследовании использовалась схема разбиения дикторов на блоки, предложенная в работе [2], для обеспечения воспроизводимости результатов и возможности их сопоставления с предыдущими исследованиями [4, 5, 12].

Результаты экспериментов

В данной работе для обозначения исследуемых моделей принята нотация LSTM-hX-vY, где X соответствует размерности скрытого состояния, а Y – количеству векторов механизма внимания.

Результаты, полученные в ходе экспериментов для моделей LSTM-h64, приведены в табл. 1.

Таблица 1
Результаты экспериментального исследования LSTM-сети со скрытым слоем размерности 64

Table 1
Experimental study results for LSTM network with hidden layer dimension of 64

Модель *Model*	Число параметров *Number of parameters*	Точность *UAR*	Пиковое значение точности UAR_{max}
LSTM-h64-v1	26 312	0,5510 ± 0,0071	0,5625
LSTM-h64-v2	26 376	0,5566 ± 0,0104	0,5742
LSTM-h64-v4	26 504	0,5548 ± 0,0065	0,5618
LSTM-h64-v8	26 760	0,5593 ± 0,0079	0,5710
LSTM-h64-v16	27 272	0,5586 ± 0,0112	**0,5775**
LSTM-h64-v32	28 296	0,5581 ± 0,0109	0,5729
LSTM-h64-v64	30 344	**0,5598 ± 0,0083**	0,5742

Анализ табл. 1 позволяет сделать вывод, что добавление механизма мультивекторного внимания повышает точность модели. В частности, среднее значение метрики UAR для модели с 64 векторами внимания на 0,88 % превышает аналогичный показатель модели с одним вектором внимания. Полученные наборы значений UAR были проанализированы с использованием парного t-критерия Стьюдента. Статистический тест показал, что различие в средних значениях между моделью с одним и 64 векторами внимания не достигает общепринятого уровня статистической значимости ($p = 0,052$). Однако, несмотря на формальное отсутствие статистической значимости на уровне $\alpha = 0,05$, наблюдаемые различия могут иметь практическую важность. В частности, из полученных данных следует, что добавление дополнительных векторов внимания приводит к повышению максимальной достигнутой точности модели в серии экспериментов. Например, пиковое значение UAR для модели с 16 векторами внимания составляет 0,5775, что на 1,5 % выше аналогичного показателя базовой модели с одним вектором вниманием.

Результаты, полученные в ходе экспериментов для моделей LSTM-h96, приведены в табл. 2.

Таблица 2
Результаты экспериментального исследования LSTM-сети со скрытым слоем размерности 96

Table 2
Experimental study results for LSTM network with hidden layer dimension of 96

Модель *Model*	Число параметров *Number of parameters*	Точность *UAR*	Пиковое значение точности UAR_{max}
LSTM-h96-v1	51 752	0,5485 ± 0,0059	0,5573
LSTM-h96-v2	51 848	0,5515 ± 0,0068	0,5605
LSTM-h96-v4	52 040	0,5508 ± 0,0150	0,5729
LSTM-h96-v8	52 424	0,5538 ± 0,0093	0,5716
LSTM-h96-v16	53 192	0,5515 ± 0,0079	0,5684
LSTM-h96-v32	54 728	0,5566 ± 0,0091	0,5788
LSTM-h96-v64	57 800	**0,5641 ± 0,0112**	**0,5859**

Полученные результаты также показывают, что увеличение количества векторов внимания улучшает качество классификации. Так, при помощи парного t-критерия была выявлена статистически значимая разница между моделью с одним и 64 векторами внимания ($p<0,02$). При этом среднее значение UAR для модели с 64 векторами внимания превышает аналогичный показатель модели с одним вектором внимания на 1,56 %. Отметим также, что пиковое значение UAR для модели LSTM-h96-v64 составляет 0,5859, что на 2,86 % выше, чем у базовой модели с одним вектором внимания.

Результаты, полученные в ходе экспериментов для моделей LSTM-h128, приведены в табл. 3. Они демонстрируют устойчивую тенденцию, наблюдаемую в экспериментах: увеличение количества векторов внимания положительно влияет на качество классификации модели. Во-первых, анализ средних значений метрики UAR показывает прогрессирующий рост точности. Во-вторых, при помощи парного t-критерия была вы-

явлена еще большая статистически значимая разница между моделью с одним и 256 векторами внимания ($p<0,0008$). Можно отметить, что средний UAR для модели с 256 векторами внимания превышает показатель базовой модели (с одним вектором) на 1,45 %. Это указывает на то, что мультивекторное внимание позволяет модели более эффективно извлекать и использовать релевантную информацию из последовательностей данных.

Таблица 3
Результаты экспериментального исследования LSTM-сети со скрытым слоем размерности 128

Table 3
Experimental study results for LSTM network with hidden layer dimension of 128

Модель *Model*	Число параметров *Number of parameters*	Точность *UAR*	Пиковое значение точности UAR_{max}
LSTM-h128-v1	85 384	0,5600 ± 0,0097	0,5794
LSTM-h128-v2	85 512	0,5636 ± 0,0078	0,5801
LSTM-h128-v4	85 768	0,5727 ± 0,0117	0,5840
LSTM-h128-v8	86 280	0,5708 ± 0,0111	0,5859
LSTM-h128-v16	87 304	0,5658 ± 0,0062	0,5768
LSTM-h128-v32	89 352	0,5716 ± 0,0126	0,5964
LSTM-h128-v64	93 448	0,5754 ± 0,0092	0,5859
LSTM-h128-v128	101 640	0,5766 ± 0,0117	**0,5996**
LSTM-h128-v256	118 024	**0,5781 ± 0,0069**	0,5892

Как и в предыдущих сериях экспериментов, максимальная точность модели также значительно возрастает. Например, пиковое значение UAR для конфигурации с 128 векторами внимания достигает 0,5996, что на 1,95 % выше, чем у базового варианта с одним вектором внимания.

На рис. 6 изображена матрица ошибок, полученная для модели LSTM-h128-v128, у которой UAR=0,5996.

Анализ матрицы ошибок показывает, что модель для распознавания эмоций демонстрирует наилучшие результаты для классов «отвращение» (73 %), «страх» (69 %) и «удивление» (67 %). Это свидетельствует о наличии у них уникальных и легко выделяемых закономерностей, отражаемых в данных. Наибольшую сложность представляет идентификация «нейтральности» (44 %) и «счастья» (46 %), которые часто путают как друг с другом, так и с другими эмоциями, что указывает на размытость границ этих состояний в признаковом пространстве. Взаимные ошибки часто возникают при классификации «счастья» и «грусти» (счастье принимается за грусть в 12,5 %, а грусть за счастье в 10,4 %), что, возможно, связано с неяркой выраженностью этих эмоций в наборе данных RAVDESS. Обращает на себя также внимание тот факт, что 20,8 % нейтральных образцов классифицированы как «счастье» – это достаточно распространенная проблема в системах распознавания эмоций, поскольку спокойная речь может восприниматься за слабовыраженную радость.

Таким образом, модель демонстрирует бóльшую эффективность при различении эмоций по степени их интенсивности и яркости проявления (отвращение, страх, удивление), однако меньшую эффективность при точной идентификации их положительной или отрицательной окраски, что особенно заметно для состояний со слабой выраженностью (грусть, счастье).

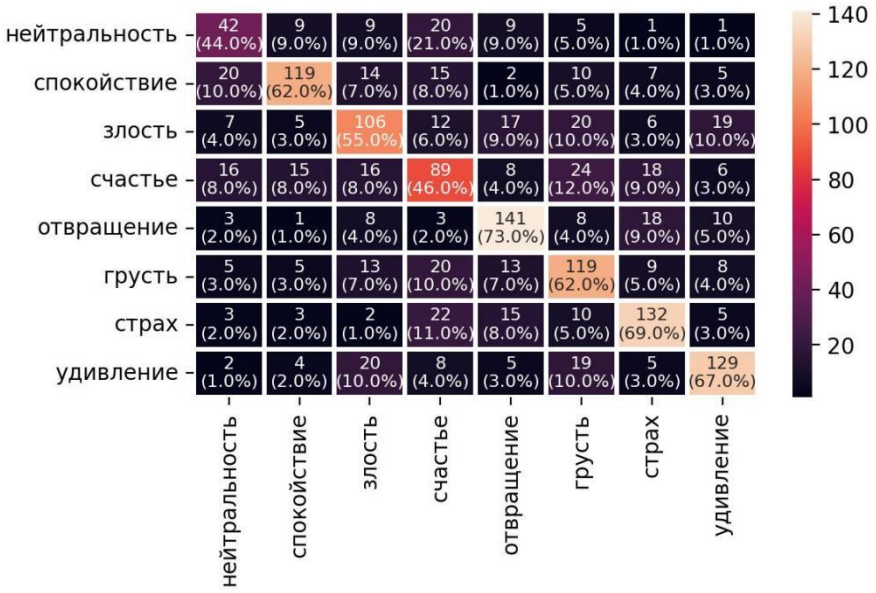

Рис. 6. Матрица ошибок для модели LSTM-h128-v128
Fig. 6. Confusion matrix for LSTM-h128-v128

Из вышесказанного можно сделать вывод, что результаты для различных конфигураций модели на основе архитектуры LSTM согласуются с гипотезой, проверяемой в рамках данного исследования: переход от мягкого внимания с одним вектором к мультивекторному вниманию является эффективным методом, статистически значимо повышающим как среднюю, так и пиковую точность модели в задаче классификации.

На рис. 7 показаны сглаженные распределения значений UAR для моделей с 64, 96 и 128 векторами внимания. Представленные графики позволяют сделать вывод, что распределение значений UAR для моделей с несколькими векторами внимания сдвинуто вправо относительно распределения модели с одним вектором. Полученные распределения наглядно демонстрируют, что при одних и тех же условиях модель с несколькими векторами внимания показывает более высокие показатели качества, чем модель с одним вектором внимания.

В целом рост точности можно объяснить тем, что каждый вектор внимания учится фокусироваться на различных аспектах эмоциональных признаков – интонации, тембре, спектральной энергии и т. д. Таким образом, использование нескольких таких векторов создает более богатое представление входного сигнала, повышая способность модели выделять эмоционально релевантные сегменты речи.

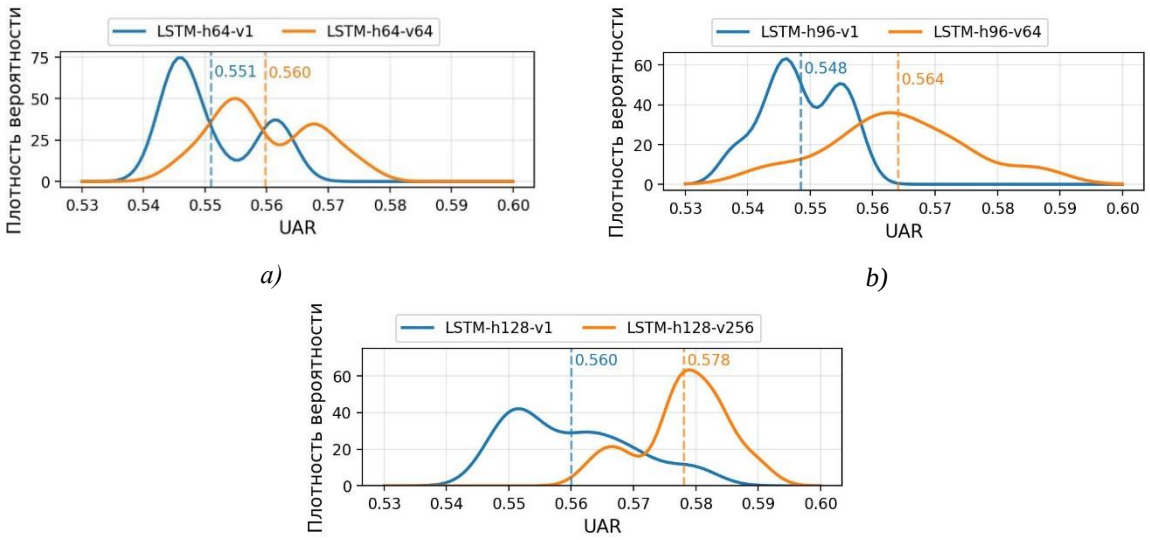

Рис. 7. Сглаженные распределения точностей (UAR) моделей: *a)* LSTM-h64-v1 и LSTM-h64-v64; *b)* LSTM-h96-v1 и LSTM-h96-v64; *c)* LSTM-h128-v1 и LSTM-h128-v256

Fig. 7. Smoothed distributions of model UAR: a) LSTM-h64-v1 and LSTM-h64-v64; b) LSTM-h96-v1 and LSTM-h96-v64; c) LSTM-h128-v1 and LSTM-h128-v256

Результаты, полученные в рамках настоящего исследования, были сопоставлены с известными из литературы методами классификации эмоций по речи (табл. 4). Поскольку в опубликованных работах, как правило, приводятся точечные оценки производительности моделей, а не интервальные, в сводной табл. 4 для моделей указаны их максимальные достигнутые значения метрики UAR.

Таблица 4
Сравнение производительности различных моделей классификации эмоций с использованием набора данных RAVDESS

Table 4
Comparison of the performance of various speech emotion classification models using the RAVDESS database

Модель *Model*	Число параметров *Number of parameters*	Точность *UAR*
SVM с 4096-мерным вектором признаков, полученных из модели AlexNet [2]	61 000 тыс.	0,4580
SVM с 304-мерным вектором надсегментных признаков [12]	–	0,4820
LDA со 190-мерным вектором признаков [4]	–	0,5380
GResNet+S [16]	–	0,5970
LSTM-h128-v128 [предлагаемая]	101,6 тыс.	0,5996
AlexNet [2]	61 000 тыс.	0,6167
CNN14 [2]	81 000 тыс.	0,7658
Модифицированная LARGE xlsr-Wav2Vec2.0 [17]	317 000 тыс.	0,8182

Анализ табл. 4 позволяет выделить некоторые тенденции в эволюции методов распознавания эмоций по речи. Традиционные методы машинного обучения показывают ограниченную эффективность: SVM (англ. support vector machine) с признаками, извлеченными предобученной AlexNet, имеет точность UAR = 0,4580, а его модификация из работы [12] достигает UAR = 0,4820. Примечателен результат классификатора на основе линейного дискриминантного анализа (англ. LDA, linear discriminant analysis) [4] с UAR = 0,5380, учитывая простоту данного метода.

Более высокие показатели демонстрируют специализированные архитектуры нейронных сетей. Модель GResNet+S [16], основанная на блоках со стробируемыми остаточными связями (англ. gated residual networks blocks), имеет достаточно высокую точность (UAR = 0,5970) и извлекает признаки непосредственно из спектрограмм. Предобученные модели, дообученные (англ. fine tuned) на целевом наборе данных, показывают прирост точности. Так, в работе [2] предобученная модель AlexNet дообучалась на базе RAVDESS, в результате достигнута точность UAR = 0,6167. Хотя эта точность превосходит результаты, которые достигаются при помощи рекуррентных и сверточных сетей, ее можно считать ограниченной, учитывая большой размер модели – порядка 61 млн параметров. Авторы статьи [2] связывают такой результат с тем, что сеть была предобучена не на речевых данных, а на базе изображений ImageNet.

Переход к моделям, предобученным непосредственно на аудиоданных, обеспечивает качественный скачок. CNN14 [8], дообученная в работе [2], демонстрирует UAR = = 0,7658. Наивысший результат (UAR = 0,8182) достигается при использовании крупной трансформерной модели xlsr-Wav2Vec2.0 [17], хотя ее сложность (317 млн параметров) ограничивает практическое применение.

Таким образом, использование больших предобученных моделей подтверждает возможность построения высокоточных систем распознавания эмоций. Однако для практических задач необходимы архитектуры с умеренной вычислительной сложностью. В этом контексте предложенные модели с мультивекторным механизмом внимания представляют перспективное направление, сочетающее приемлемую точность с вычислительной эффективностью.

Заключение

В работе проведено исследование влияния количества векторов внимания на качество распознавания эмоций в речи с использованием LSTM-сетей. Основное внимание уделялось разработке и экспериментальной оценке мультивекторного механизма внимания, предназначенного для более гибкого выделения информативных фрагментов акустического сигнала.

Эксперименты, выполненные на наборе данных RAVDESS, показали, что увеличение числа векторов внимания приводит к статистически значимому росту метрики точности UAR по сравнению с базовой моделью, использующей один вектор внимания. Это свидетельствует о том, что предложенный механизм способствует более точному моделированию эмоциональных состояний в речи. Несмотря на умеренное абсолютное улучшение метрики, статистическая значимость и воспроизводимость указывают на устойчивое положительное влияние разработанного подхода.

Перспективными направлениями дальнейших исследований представляются: интеграция разнородных акустических признаков, включая мел-спектрограммы и хромограммы, для формирования более полного описания речевого сигнала; использование сверточных слоев для автоматического извлечения иерархических признаков непосредственно из спектральных представлений аудиосигнала [7].

Вклад авторов. *Д. В. Краснопрошин* занимался решением задач исследования, анализом полученных результатов и формированием структуры статьи; *М. И. Вашкевич* определил цель и задачи, которые необходимо было решить в ходе проведения исследования, руководил исследованием и принял участие в разработке метода интерпретации результатов эксперимента, редактировании текста статьи и подготовке графического материала.

Список использованных источников

1. A review of affective computing: From unimodal analysis to multimodal fusion / S. Poria, E. Cambria, R. Bajpai, A. Hussain // Information Fusion. – 2017. – Vol. 37. – P. 98–125.

2. Multimodal emotion recognition on RAVDESS dataset using transfer learning / C. Luna-Jiménez, D. Griol, Z. Callejas [et al.] // Sensors. – 2021. – Vol. 21. – P. 1–29.

3. Mirsamadi, S. Automatic speech emotion recognition using recurrent neural networks with local attention / S. Mirsamadi, E. Barsoum, C. Zhang // Proc. of IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 05–09 Mar. 2017. – New Orleans, 2017. – P. 2227–2231.

4. Краснопрошин, Д. В. Отбор признаков на основе техники переноса обучения для классификации эмоций в речи с помощью полносвязной нейронной сети прямого распространения / Д. В. Краснопрошин, М. И. Вашкевич // Системный анализ и прикладная информатика. – 2025. – № 1. – С. 38–43.

5. Краснопрошин, Д. В. Анализ подходов к построению систем распознавания эмоций по речи с использованием методов глубокого обучения / Д. В. Краснопрошин, М. И. Вашкевич // Big Data and Advanced Analytics : сб. науч. ст. XI Междунар. науч.-практ. конф., Минск, 23–24 апр. 2025 г. – Мн., 2025. – С. 343–353.

6. Dal Rí, F. A. Speech emotion recognition and deep learning: an extensive validation using convolutional neural networks / F. A. Dal Rí, F. C. Ciardi, N. Conci // IEEE Access. – 2023. – Vol. 11. – P. 116638–116649.

7. Waleed, G. T. Speech emotion recognition on MELD and RAVDESS datasets using CNN / G. T. Waleed, S. H. Shaker // Information. – 2025. – Vol. 16, no. 7. – P. 518.

8. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition / Q. Kong, Y. Cao, T. Iqbal [et al.] // IEEE/ACM Transactions on Audio, Speech, and Language Processing. – 2020. – Vol. 28. – P. 2880–2894.

9. Николенко, С. Глубокое обучение / С. Николенко, А. Кадурин, Е. Архангельская. – СПб. : Питер, 2019. – 480 с.

10. Hochreiter, S. Long short-term memory / S. Hochreiter, J. Schmidhuber // Neural Computation. – 1997. – Vol. 9, no. 8. – P. 1735–1780.

11. Context-aware attention mechanism for speech emotion recognition / G. Ramet, P. N. Garner, M. Baeriswyl, A. Lazaridis // 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 Dec. 2018. – Athens, 2018. – P. 126–131.

12. Краснопрошин, Д. В. Метод распознавания эмоций в речевом сигнале с использованием машины опорных векторов и надсегментных акустических признаков / Д. В. Краснопрошин, М. И. Вашкевич // Доклады БГУИР. – 2024. – Т. 22, № 3. – С. 93–100.

13. Bahdanau, D. Neural machine translation by jointly learning to align and translate / D. Bahdanau, K. Cho, Y. Bengio // 3rd Intern. Conf. on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015. – San Diego, 2015. – URL: https://arxiv.org/abs/1409.0473 (date of access: 13.11.2025).

14. Optuna: A next-generation hyperparameter optimization framework / T. Akiba, S. Sano, T. Yanase [et al.] // Proc. of the 25th ACM SIGKDD Intern. Conf. on Knowledge Discovery & Data Mining (KDD'19), Anchorage, AK, USA, 4–8 Aug. 2019. – Anchorage, 2019. – P. 2623–2631.

15. Algorithms for hyper-parameter optimization / J. S. Bergstra, R. Bardenet, Y. Bengio, B. Kégl // NIPS'11: Proc. of the 25th Intern. Conf. on Neural Information Processing Systems, Granada, Spain, 12–15 Dec. 2011. – Granada, 2011. – P. 2546–2554.

16. Spectrogram based multi-task audio classification / Y. Zeng, H. Mao, D. Peng, Z. Yi // Multimedia Tools and Applications. – 2019. – Vol. 78, no. 3. – P. 3705–3722.

17. A proposal for multimodal emotion recognition using aural transformers and action units on RAVDESS dataset / C. Luna-Jiménez, R. Kleinlein, D. Griol [et al.] // Applied Sciences. – 2022. – Vol. 12, no. 1. – P. 1–23.

References

1. Poria S., Cambria E., Bajpai R., Hussain A. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 2017, vol. 37, pp. 98–125.

2. Luna-Jiménez C., Griol D., Callejas Z., Kleinlein R., Montero J. M., Fernández-Martínez F. Multimodal emotion recognition on RAVDESS dataset using transfer learning. *Sensors*, 2021, vol. 21, pp. 1–29.

3. Mirsamadi S., Barsoum E., Zhang C. Automatic speech emotion recognition using recurrent neural networks with local attention. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 05–09 March 2017*. New Orleans, 2017, pp. 2227–2231.

4. Krasnoproshin D. V., Vashkevich M. I. *Transfer learning based feature selection for feedforward neural network for speech emotion classifier*. Sistemnyj analiz i prikladnaja informatika [*System Analysis and Applied Information Science*], 2025, no. 1, pp. 38–43 (In Russ.).

5. Krasnoproshin D. V., Vashkevich M. I. *Analysis of approaches to building speech emotion recognition systems using deep learning methods*. Big Data and Advanced Analytics: sbornik nauchnyh statej XI Mezhdunarodnoj nauchno-prakticheskoj konferencii, Minsk, 23–24 aprelja 2025 g. [*Big Data and Advanced Analytics: Collection of Scientific Articles of the XI International Scientific and Practical Conference, Minsk, 23–24 April 2025*]. Minsk, 2025, pp. 343–353 (In Russ.).

6. Dal Rí F. A., Ciardi F. C., Conci N. Speech emotion recognition and deep learning: an extensive validation using convolutional neural networks. *IEEE Access*, 2023, vol. 11, pp. 116638–116649.

7. Waleed G. T., Shaker S. H. Speech emotion recognition on MELD and RAVDESS datasets using CNN. *Information*, 2025, vol. 16, no. 7, p. 518.

8. Kong Q., Cao Y., Iqbal T., Wang Y., Wang W., Plumbley M. D. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020, vol. 28, pp. 2880–2894.

9. Nikolenko S., Kadurin A., Archangelskaya E. Glubokoe obuchenie. *Deep Learning*. Saint Petersburg, Piter, 2019, 480 p. (In Russ.).

10. Hochreiter S., Schmidhuber J. Long short-term memory. *Neural Computation*, 1997, vol. 9, no. 8, pp. 1735–1780.

11. Ramet G., Garner P. N., Baeriswyl M., Lazaridis A. Context-aware attention mechanism for speech emotion recognition. *2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018*. Athens, 2018, pp. 126–131.

12. Krasnoproshin D. V., Vashkevich M. I. *Speech emotion recognition method based on support vector machine and suprasegmental acoustic features*. Doklady BGUIR [*BGUIR Proceedings*], 2024, vol. 22, no. 3, pp. 93–100 (In Russ.).

13. Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015*. San Diego, 2015. Available at: https://arxiv.org/abs/1409.0473 (accessed 13.11.2025).

14. Akiba T., Sano S., Yanase T., Ohta T., Koyama M. Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'19), Anchorage, AK, USA, 4–8 August 2019*. Anchorage, 2019, pp. 2623–2631.

15. Bergstra J. S., Bardenet R., Bengio Y., Kégl B. Algorithms for hyper-parameter optimization. *NIPS'11: Proceedings of the 25th International Conference on Neural Information Processing Systems, Granada, Spain, 12–15 December 2011*. Granada, 2011, pp. 2546–2554.

16. Zeng Y., Mao H., Peng D., Yi Z. Spectrogram based multi-task audio classification. *Multimedia Tools and Applications*, 2019, vol. 78, no. 3, pp. 3705–3722.

17. Luna-Jiménez C., Kleinlein R., Griol D., Callejas Z., Montero J. M., Fernández-Martínez F. A proposal for multimodal emotion recognition using aural transformers and action units on RAVDESS dataset. *Applied Sciences*, 2022, vol. 12, no. 1, pp. 1–23.

Информация об авторах

Краснопрошин Даниил Вадимович, магистр технических наук, аспирант кафедры электронных вычислительных средств, Белорусский государственный университет информатики и радиоэлектроники.
E-mail: daniil.krasnoproshin@gmail.com

Вашкевич Максим Иосифович, доктор технических наук, доцент, профессор кафедры электронных вычислительных средств, Белорусский государственный университет информатики и радиоэлектроники.
E-mail: vashkevich@bsuir.by

Information about the authors

Daniil V. Krasnoproshin, M. Sci. (Eng.), Postgraduate Student of Computer Engineering Department, Belarusian State University of Informatics and Radioelectronics.
E-mail: daniil.krasnoproshin@gmail.com

Maxim I. Vashkevich, Dr. Sci. (Eng.), Assoc. Prof., Prof. of Computer Engineering Department, Belarusian State University of Informatics and Radioelectronics.
E-mail: vashkevich@bsuir.by