НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
июль–август 2023 Том 23 № 4 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
July–August 2023 Vol. 23 No 4 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2023-23-4-767-775
УДК 004.855.5

Нейросетевой метод визуального распознавания голосовых команд водителя с использованием механизма внимания

Александр Александрович Аксёнов[1]✉, Елена Витальевна Рюмина[2], Дмитрий Александрович Рюмин[3], Денис Викторович Иванько[4], Алексей Анатольевич Карпов[5]

[1,2,3,4,5] Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, 199178, Российская Федерация

[1] axyonov.a@iias.spb.su✉, https://orcid.org/0000-0002-7479-2851
[2] ryumina.e@iias.spb.su, https://orcid.org/0000-0002-4135-6949
[3] ryumin.d@iias.spb.su, https://orcid.org/0000-0002-7935-0569
[4] ivanko.d@iias.spb.su, https://orcid.org/0000-0003-0412-7765
[5] karpov@iias.spb.su, https://orcid.org/0000-0003-3424-652X

Аннотация

Введение. Визуальное распознавание речи или автоматическое чтение речи по губам все чаще применяется для преобразования речи в текст. Видеоданные доказывают свою необходимость в системах мультимодального распознавания речи, особенно когда использование акустических данных затруднено в виду сильных аудиошумов или недоступно. Основная цель исследования заключается в повышении эффективности распознавания команд водителя путем анализа визуальной информации для снижения тактильного взаимодействия с различными автомобильными системами (мультимедийными и навигационными, телефонными звонками и др.) во время вождения. **Метод.** Предложен метод автоматического чтения речи водителя по губам в процессе управления транспортным средством на основе глубокой нейронной сети архитектуры 3DResNet18. Выполнен анализ динамической информации о движении губ диктора с помощью 3D-сверточных слоев нейросети. Использование нейросетевой архитектуры с двунаправленной моделью Long Short-Term Memory и механизмом внимания позволяет добиться более высокой точности распознавания при незначительном снижении скорости работы. **Основные результаты.** Предложены и исследованы два варианта нейросетевых архитектур для визуального распознавания речи. При использовании первой нейросетевой архитектуры результат распознавания голосовых команд водителя составил 77,68 %, что ниже на 5,78 %, по сравнению со второй. Скорость работы системы определена показателем реального времени (Real-Time Factor, RTF), значение которого для первой нейросетевой архитектуры равен 0,076, а второй — 0,183, что выше более чем в два раза. Предложенный метод апробирован на данных дикторов многомодального корпуса RUSAVIC, записанных в автомобиле. **Обсуждение.** Результаты исследования могут найти применение в системах аудиовизуального распознавания речи. Подобные системы могут быть рекомендованы для применения в сильно зашумленных условиях, например, в процессе управления транспортным средством. Проведенный анализ позволил выбрать оптимальную нейросетевую модель визуального распознавания речи для последующего встраивания в ассистивную систему на базе мобильного устройства.

Ключевые слова

голосовые команды водителя, визуальное распознавание речи, автоматическое чтение речи по губам, машинное обучение, CNN, LSTM, механизм внимания

Благодарности

Исследование выполнено при поддержке РФФИ (проект № 19-29-09081-мк), ведущей научной школы Российской Федерации (грант № НШ-17.2022.1.6) и за счет средств государственного финансирования, тема FFZF-2022-0005.

Ссылка для цитирования: Аксёнов А.А., Рюмина Е.В., Рюмин Д.А., Иванько Д.В., Карпов А.А. Нейросетевой метод визуального распознавания голосовых команд водителя с использованием механизма внимания // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23, № 4. С. 767–775. doi: 10.17586/2226-1494-2023-23-4-767-775

Neural network-based method for visual recognition of driver's voice commands using attention mechanism

Alexandr A. Axyonov[1]✉, Elena V. Ryumina[2], Dmitry A. Ryumin[3], Denis V. Ivanko[4], Alexey A. Karpov[5]

St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation

[1] axyonov.a@iias.spb.su✉, https://orcid.org/0000-0002-7479-2851
[2] ryumina.e@iias.spb.su, https://orcid.org/0000-0002-4135-6949
[3] ryumin.d@iias.spb.su, https://orcid.org/0000-0002-7935-0569
[4] ivanko.d@iias.spb.su, https://orcid.org/0000-0003-0412-7765
[5] karpov@iias.spb.su, https://orcid.org/0000-0003-3424-652X

Abstract

Visual speech recognition or automated lip-reading systems actively apply to speech-to-text translation. Video data proves to be useful in multimodal speech recognition systems, particularly when using acoustic data is difficult or not available at all. The main purpose of this study is to improve driver command recognition by analyzing visual information to reduce touch interaction with various vehicle systems (multimedia and navigation systems, phone calls, etc.) while driving. We propose a method of automated lip-reading the driver's speech while driving based on a deep neural network of 3DResNet18 architecture. Using neural network architecture with bi-directional LSTM model and attention mechanism allows achieving higher recognition accuracy with a slight decrease in performance. Two different variants of neural network architectures for visual speech recognition are proposed and investigated. When using the first neural network architecture, the result of voice recognition of the driver was 77.68 %, which was lower by 5.78 % than when using the second one the accuracy of which was 83.46 %. Performance of the system which is determined by a real-time indicator RTF in the case of the first neural network architecture is equal to 0.076, and the second — RTF is 0.183 which is more than two times higher. The proposed method was tested on the data of multimodal corpus RUSAVIC recorded in the car. Results of the study can be used in systems of audio-visual speech recognition which is recommended in high noise conditions, for example, when driving a vehicle. In addition, the analysis performed allows us to choose the optimal neural network model of visual speech recognition for subsequent incorporation into the assistive system based on a mobile device.

Keywords

driver's voice commands, visual speech recognition, automatic lip reading, machine learning, CNN, LSTM, attention mechanisms

Acknowledgements

The study was supported by the Russian Foundation for Basic Research (project no. 19-29-09081-mk), the leading scientific school of the Russian Federation (grant no. NSh-17.2022.1.6) and at the expense of state funding, topic FFZF-2022-0005.

For citation: Axyonov A.A., Ryumina E.V., Ryumin D.A., Ivanko D.V., Karpov A.A. Neural network-based method for visual recognition of driver's voice commands using attention mechanism. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 4, pp. 767–775 (in Russian). doi: 10.17586/2226-1494-2023-23-4-767-775

Введение

На сегодняшний день не существует надежных систем автоматического распознавания речи, устойчивых к динамическим акустическим шумам, которые могли бы использоваться в реальных условиях вождения транспортного средства. Управление навигационной системой, кондиционером/смартфоном с применением сенсорного интерфейса может привести к отвлечению внимания водителя и стать причиной дорожно-транспортных происшествий (ДТП). Практически акустический шум является актуальной проблемой в данной области [1]. Фоновый шум оказывает влияние не только на микрофон, но и вынуждает говорящего повышать громкость голоса для того, чтобы компенсировать уровень шума в ушах (так называемый «эффект Ломбарда»). В реальных условиях применение изменения речевой активности, вызываемое шумовым воздействием на органы слуха, может повлиять на эффективность работы системы распознавания больше, чем акустический шум [2].

Современные технологии автоматического визуального распознавания речи (Visual Speech Recognition, VSR) позволяют распознавать речь людей, анализируя движения губ и лица. Системы на основе этих технологий имеют множество применений, таких как идентификация говорящего, преобразование речи в текст, а также голосовое управление техническими устройствами. Применение технологий автоматического «чтения речи по губам» в автомобильной отрасли может способствовать развитию ассистивных систем помощи водителю [3].

В последние годы для повышения эффективности распознавания речи некоторые исследователи используют визуальную информацию и анализируют, как чтение речи по губам может способствовать повышению эффективности в распознавании речи на основе аудиоинформации [4]. Визуальные сигналы содержат достаточный объем информации, позволяющий распознавать произносимые фразы [5, 6], и совместное использование этих модальностей является шагом к созданию робастной системы распознавания речи в сложных акустических условиях [7].

В связи с этим использование систем на основе анализа визуальной информации позволяет взаимодействовать с автомобильными информационно-развлекательными системами в режиме «свободных рук», снижая когнитивную нагрузку и повышая безопасность дорожного движения [8, 9].

Отметим, что внедрение систем анализа визуальной информации в автомобильной отрасли сопряжено с определенными трудностями из-за вариативности уровня освещения и расположения камер, которые, в свою очередь, влияют на точность обнаружения и отслеживания лица и рта говорящего [10]. Несмотря на эти проблемы, недавние исследования демонстрируют перспективность применения таких систем в условиях вождения [11]. Подобные системы могут быть эффективным инструментом для распознавания и отслеживания состояний сонливости и усталости водителя, что может способствовать снижению числа ДТП на дорогах [12, 13].

Основная цель настоящей работы — повышение эффективности распознавания команд водителя путем анализа визуальной информации для снижения физического взаимодействия с различными автомобильными системами (мультимедийными и навигационными системами, звонками и др.) во время вождения.

Краткий обзор предметной области

В последние годы системы VSR используют технологии глубокого обучения для извлечения информативных признаков и машинной классификации. Различные типы нейросетей, такие как сверточные нейронные сети (Convolutional Neural Network, CNN), сети прямого распространения и автоэнкодеры, используются в системах чтения речи по губам [14, 15]. Одним из перспективных подходов в VSR являются 3D CNN, которые применяются на входе системы, поскольку они отлично справляются с анализом пространственной и временной информации, а также извлечением информативных признаков [16–18]. Некоторые исследователи в дополнение к этому интегрируют механизмы пространственного внимания для обработки извлеченных признаков [19, 20]. Механизм пространственного внимания позволяет модели фокусироваться на наиболее информативных областях во входных видеокадрах.

Для распознавания слитной речи (словосочетания или фразы) также используются нейросети обработки последовательностей — рекуррентные нейронные сети (Recurrent Neural Network, RNN) [21]. В последнее время в качестве альтернативы RNN для классификации стали применять модели трансформеров с использованием механизмов внимания и временные сверточные сети [22, 23]. Однако для обучения таких моделей требуются большие вычислительные мощности, а также значительный объем обучающих данных. Потому одним из популярных подходов в таких случаях является трансферное обучение [24] — метод использования предварительно обученной модели для улучшения прогнозов в рамках другой, но схожей задачи, позволяющий сократить время обучения и уменьшить потребность в данных, а также повысить производительность нейронной сети [25].

Существуют подходы, в которых распознавание речи основывается на моделях глубоких машин Больцмана (Deep Boltzmann Mashine, DBM) [26]. DBM используются для извлечения признаков из визуальной модальности, которые потом объединяются с вектором признаков дискретного косинусного преобразования [27]. Далее проводится линейный дискриминантный анализ [28], используемый для анализа корреляции признакового пространства и уменьшения его размерности. Для моделирования и распознавания речи применяются модели смеси гауссовских распределений (Gaussian Mixture Model, GMM) [29] и скрытые марковские модели.

Кроме одномодального визуального распознавания речи широко используется многомодальное распознавание речи — аудиовизуальное распознавание речи (Audio Visual Speech Recognition, AVSR). В последние годы такие подходы демонстрируют многообещающие результаты [30, 31]. В работах [32, 33] предложено использовать иерархическую многомодальную нейросетевую архитектуру для AVSR. В работе [34] рассмотрен метод, основанный на нейросетевой архитектуре типа «последовательность-последовательность» (sequence-to-sequence) для автоматического обучения объединенному представлению из аудио- и видеомодальностей на основе его информативности, тем самым повышая устойчивость к шумам при распознавании речи за счет использования дополнительной визуальной информации.

Методы, предложенные в перечисленных работах, демонстрируют потенциал систем визуального распознавания речи, которые также могут найти применение в автомобильной сфере.

Многомодальный корпус аудиовизуальной русской речи RUSAVIC

В настоящей работе использован собственный корпус данных Russian Audio-Visual Speech in Cars (RUSAVIC) [35], созданный по разработанной методологии [36]. Корпус данных RUSAVIC разработан специально для решения задач распознавания речи на основе наиболее часто встречающихся голосовых команд водителей и предназначен для использования в системах помощи водителю. Основные характеристики многодикторного мультимодального корпуса представлены в табл. 1.

Каждый диктор произнес 62 управляющие голосовые команды (например, «Найти заправку», «Набрать номер», «Карта», «Включить свет в салоне») не менее 10 раз в течение нескольких сеансов записи. В корпусе содержатся данные, записанные в реальных условиях вождения, а также в транспортном средстве, припаркованном около оживленного перекрестка.

Корпус данных RUSAVIC обрабатывал информацию с помощью извлечения аудиосигнала из мультимедийной информации, далее применялся детектор голосовой активности (Voice Activity Detection, VAD). В работе использована специально обученная модель

Таблица 1. Основные характеристики корпуса данных RUSAVIC

Table 1. RUSAVIC corpus parameters

Характеристика	Значение
Количество дикторов	20
Разрешение видеоданных, пиксел	1920 × 1080
Частота кадров, кадр/с	60
Количество команд в словаре	62
Количество сеансов записи	10 (для каждого диктора)
Общее число фраз	около 12 400

обнаружения голосовой активности Vosk[1], которая способна достаточно точно обнаружить границы речи даже в зашумленных условиях. Результат применения VAD — получение мультимедийного файла с произнесенной фразой из словаря, в котором отсутствует лишняя информация (тишина до начала речи и после).

Метод визуального распознавания голосовых команд водителя

Метод визуального распознавания голосовых команд водителя включал в себя два этапа (рис. 1). На этапе 1 выполнена сегментация входного видеосигнала. На каждом кадре из входной последовательности определена графическая область интереса (Region of Interest, ROI), которой является область рта диктора. Детектирование ROI выполнено с помощью программной системы MediaPipe FaceMesh [37], которая определила 468 трехмерных лицевых ориентиров-маркеров. Извлеченные области губ подвергались преобразованию в градации серого, нормализации до изображения 88 × 88 пикселов. Далее области губ разбивались на сегменты, состоящие из 120 кадров с частичным перекрытием в 48 кадров (40 %). При недостатке кадров в конце последовательности недостающие кадры заполнялись последним кадром последовательности.

На этапе 2 проведено распознавание голосовой команды с помощью предварительно обученной нейросетевой модели. Визуальная речь на выходе нейросети декодирована в понятный формат и сопоставлена с метками в словаре. Распознанная голосовая команда водителя определена максимальным значением в выходном векторе вероятностных значений.

Обучение нейросетевой модели

Процесс обучения нейросетевой модели содержал параметры, которые могут влиять на эффективность распознавания обученной модели. В процессе обучения после каждой эпохи выполняется определение точности на обучающем и валидационном наборах данных. Часто точность распознавания на обучающем наборе может достигать высоких значений, вплоть до 100 %. Это значит, что модель переобучается и ее эффективность на тестовом и валидационном наборах будет низкой. Для генерации дополнительного набора обучающих данных применяются различные методы аугментации данных.

В настоящей работе на 40 % изображений и меток, которые выбраны случайным образом, применен метод аугментации данных MixUP [38]. Коэффициент слияния двух изображений и их меток (классов команд водителя) варьировался от 30 до 70 % (при сумме равной 100 % соблюдается условие нулевой прозрачности для генерируемого изображения). Для оставшихся 60 % бинарных векторов меток после применения MixUP использован метод сглаживания Label Smoothing [39].

Рассмотрим две наиболее эффективные для задачи исследования архитектуры нейросетевых моделей (рис. 2).

Архитектура 3DResNet18 + SA. Сегментированные последовательности поступают на вход обучаемой нейросетевой модели (рис. 2, а), далее слой 3D Conv свертывает их в пространстве и времени. Для извлечения наиболее значимых признаков из последовательностей использован слой подвыборки MaxPooling 3D, способный уменьшать размерности входного тензора. Отметим, что слой MaxPooling 3D позволяет ускорить обучение, а также уменьшить количество параметров модели, при этом сохраняя ее эффективность. Модифицированные остаточные блоки Residual Blocks модели ResNet18 [40] извлекают карты признаков размерностью 4 × 3 × 3 × 512 для всей входной последовательности. В последнем остаточном блоке присутствует механизм внимания (Squeeze-and-Attention, SA) [41], использование которого улучшает результаты распознавания. В SA сначала выполняется операция «сжатия» (squeeze), которая сокращает размерность последовательности данных, а операция «внимания» (attention), позволяет модели сосредоточиться на наиболее важных аспектах данных, игнорируя несущественные детали.

Признаки, поступающие с выхода сверточных слоев, преобразуются в вектор фиксированной длины за счет использования слоя Global Average Pooling 3D, суть которого заключается в том, чтобы усреднить значения признаков по всему объему 3D-пространства, полученные с последнего сверточного слоя в сети. Результатом является один признак для каждого фильтра сверточного слоя. Эти признаки затем используются для классификации. В архитектуре также применяется техника регуляризации нейросетей Dropout, которая помогает предотвратить переобучение модели.

Полученные одномерные вектора размерностью 512 признаков поступают в полносвязный слой с функцией активации ReLU, выполняющей нелинейное преобразование входных данных. Размерность этого слоя увеличивается до 1024, тем самым улучшая представление признаков и повышая точность модели. В конце этот слой соединяется с последним полносвязным слоем (Fully connected, FC), содержащим 62 нейрона с функцией активации Softmax. Это означает, что на выходе формируется вероятность принадлежности входных данных к 62 классам, которые интерпретируются

[1] Vosk Api [Электронный ресурс]. Режим доступа: https://github.com/alphacep/vosk-api, свободный (дата обращения: 23.04.2023).

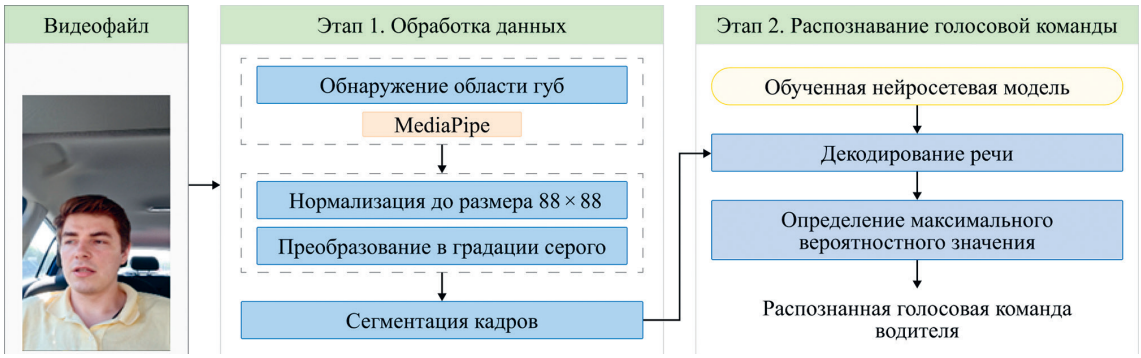

Рис. 1. Метод визуального распознавания команд водителя

Fig. 1. Method for visual recognition a driver's commands

как гипотезы предсказания модели для каждого из этих классов (62 голосовые команды).

Функция активации Softmax — одна из возможных функций активации, широко используемая в выходном слое нейронных сетей для многоклассовой классификации. Функция Softmax преобразует линейный выход полносвязного слоя в вероятность принадлежности каждого входа к определенному классу.

Архитектура 3DResNet18 + SA + BiLSTM. Вторая нейросетевая архитектура отличается от первой тем, что в ней 3D Conv обрабатывает только пространственное представление признаков. Для обработки временного пространства признаков использовано два слоя двунаправленных Long short-term memory (LSTM) (рис. 2, *b*). Длина последовательности изображений в 120 кадров поступают в 3D сверточный слой и подвергаются операции подвыборки MaxPooling 3D. Из каждого изображения входной последовательности извлекаются карты признаков размерностью 120 × 3 × 3 × 512 благодаря модифицированным остаточным блокам. Слой подвыборки Average Pooling преобразует полученные карты признаков в одномерные вектора размерностью 120 × 512, которые поступают на последующих два слоя двунаправленной LSTM. Первый слой является sequence-to-sequence, на входе и выходе которого 120 векторов признаков, а второй sequence-to-one — на выходе один вектор признаков размерностью 512. В конце полносвязный слой FC

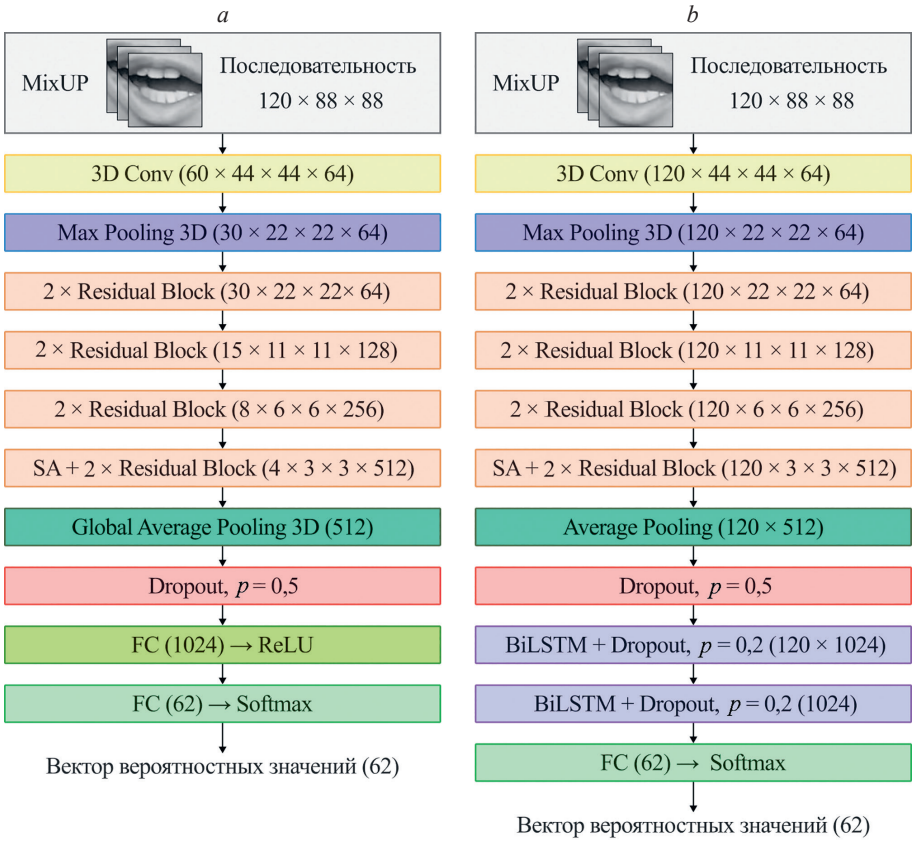

Рис. 2. Нейросетевые архитектуры: 3DResNet18 + SA (*a*); 3DResNet18 + SA+ BiLSTM (*b*)

Fig. 2. Neural network architectures: 3DResNet18 + SA (*a*); 3DResNet18 + SA+ BiLSTM (*b*)

Таблица 2. Сравнение эффективности архитектур нейросетевых моделей
Table 2. Comparison of neural network architectures performance

Архитектура нейросетевой модели	Время инициализации модели, с	Точность распознавания команд, %		RTF
		Val	Test	
3DResNet18	2,3	86,53	76,74	0,071
3DResNet18 + SA (рис. 2, *a*)	2,7	88,66	77,68	0,076
3DResNet18 + BiLSTM	7,9	82,94	75,89	0,153
3DResNet18 + SA+ BiLSTM (рис. 2, *b*)	8,2	85,55	83,46	0,183

с количеством нейронов 62 также формирует вектор вероятностных значений.

Экспериментальные результаты

Приведем результаты экспериментальных исследований с различными архитектурами нейросетей. Апробация предложенного метода выполнена на корпусе данных RUSAVIC. Набор данных для обучения и оценивания нейросетевых моделей разбивался следующим образом: обучающий набор (train) включал 14 389 видеозаписей 13 дикторов, тестовый набор (test) — 4241 видеозапись трех дикторов и валидационный (val) — 2480 видеозаписей двух дикторов. В качестве планировщика скорости обучения применялась техника косинусного отжига (Cosine Annealing), значения которого варьировались от 0,0001 до 0,001. Изначальное количество эпох устанавливалось равным 100. После каждой эпохи выполняется промежуточное оценивание модели на валидационном наборе данных. В случае если в течение 10 эпох обучаемая модель не прогрессировала, то процесс обучения останавливался, а лучшие веса модели определялись по максимальной точности, достигнутой на валидационном наборе. Валидационная выборка использована для оптимизации модели, тогда как экспериментальные результаты представлены для тестовой выборки. В качестве оптимизатора применен Adam.

Программная реализация системы визуального распознавания голосовых команд выполнена на языке программирования Python v.3.9. Для машинного обучения моделей использован фреймворк с открытым исходным кодом TensorFlow v.2.8.0 в связке с расширением модуля TensorFlow-GPU v.2.8.0. Экспериментальные исследования осуществлены на персональном компьютере под управлением операционной системы Microsoft Windows 10 Pro со следующими техническими характеристиками: CPU — AMD Ryzen 9 5950X; GPU — GeForce RTX 3090 TI 24Gb; хранилище данных — твердотельный накопитель SSD M.2.

Проведено сравнение архитектур нейросетевых моделей, основанных на архитектуре ResNet18. Выбранные модели анализируют пространственно-временные зависимости последовательности входных изображений. В табл. 2 представлены две принципиально разные архитектуры и их модернизированные версии за счет добавления механизмов внимания (рис. 2).

Параметры обучения нейросетевых архитектур, представленных в сравнении, подбирались эмпирическим путем и являются для каждой архитектуры оптимальными. Неизменные параметры: набор обучающих данных, длина последовательности и разрешение изображений, которые подаются на вход.

Исследуемые архитектуры сравнивались не только по показателю точности распознавания команд водителя, но и по таким показателям как: среднее время активации модели, размер модели и скорость распознавания по показателю реального времени (Real-Time Factor, RTF). Сравнение проводилось в одинаковых условиях. Для определения скорости работы системе предлагалось распознать 15 случайно выбранных фраз различной длительности из корпуса данных RUSAVIC. Для расчета времени инициализации модели и усредненного значения RTF описанная процедура производилась 10 раз.

По результатам анализа табл. 2 можно сделать вывод, что нейросетевая архитектура 3DResNet18 + SA + + BiLSTM оказалась точнее архитектуры без добавления слоев BiLSTM. Однако эта архитектура немного проигрывает в скорости работы. Из результатов сравнения также следует, что при добавлении механизма внимания SA независимо от типа архитектуры точность распознавания голосовых команд повышается. Добавление SA в архитектуры нейросетевых моделей оказывает незначительное влияние на скорость работы системы. Также добавление различных методов аугментации, таких как Cosine WR, MixUP и Label Smoothig, дает прирост точности [42].

Заключение

Предложен метод чтения речи по губам водителя во время управления транспортным средством с использованием модифицированной нейросетевой архитектуры ResNet18. Метод может быть встроен в системы распознавания речи, функционирующие в условиях сильного акустического шума, вызванного различными факторами, такими как скорость движения, покрытие дороги, степень открытия окон, наличие источников звука и др.

Результаты экспериментов показали, что метод позволяет эффективно распознавать произносимые водителем команды из словаря, состоящего из 62 русскоязычных фраз с точностью до 83,46 % и со скоростью распознавания RTF порядка 0,183. Метод можно использовать в режиме реального времени.

Для улучшения качества распознавания речи в дальнейшем планируется исследовать другие архитектуры глубоких нейросетей, а также расширять обучающую

базу данных. Предложенный метод визуального распознавания планируется использовать в рамках разработки системы аудиовизуального распознавания речи, что может значительно повысить точность и устойчивость к шумам существующих систем распознавания речи.

Литература

1. Lin S.C., Hsu C.H., Talamonti W., Zhang Y., Oney S., Mars J., Tang L. Adasa: A conversational in-vehicle digital assistant for advanced driver assistance features // Proc. of the 31st Annual ACM Symposium on User Interface Software and Technology. 2018. P. 531–542. https://doi.org/10.1145/3242587.3242593
2. Lee B., Hasegawa-Johnson M., Goudeseune C., Kamdar S., Borys S., Liu M., Huang T. AVICAR: Audio-visual speech corpus in a car environment // Proc. of the 8th International Conference on Spoken Language Processing. 2004. P. 2489–2492. https://doi.org/10.21437/Interspeech.2004-424
3. Ivanko D., Ryumin D., Kashevnik A., Axyonov A., Karpov A. Visual speech recognition in a driver assistance system // Proc. of the 30th European Signal Processing Conference (EUSIPCO). 2022. P. 1131–1135. https://doi.org/10.23919/EUSIPCO55093.2022.9909819
4. Xu B., Wang J., Lu C., Guo Y. Watch to listen clearly: Visual speech enhancement driven multi-modality speech recognition // Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2020. P. 1637–1646. https://doi.org/10.1109/wacv45572.2020.9093314
5. Afouras T., Chung, J.S., Senior A., Vinyals O., Zisserman A. Deep audio-visual speech recognition // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022. V. 44. N 12. P. 8717–8727. https://doi.org/10.1109/TPAMI.2018.2889052
6. Кухарев Г.А., Матвеев Ю.Н., Олейник А.Л. Алгоритмы взаимной трансформации изображений для систем обработки и поиска визуальной информации // Научно-технический вестник информационных технологий, механики и оптики. 2017. Т. 17. № 1. С. 62–74. https://doi.org/10.17586/2226-1494-2017-17-1-62-74
7. Shi B., Hsu W.N., Mohamed A. Robust self-supervised audio-visual speech recognition // Proc. of the International Conference INTERSPEECH. 2022. P. 2118–2122. https://doi.org/10.21437/interspeech.2022-99
8. Chand H.V., Karthikeyan J. CNN based driver drowsiness detection system using emotion analysis // Intelligent Automation & Soft Computing. 2022. V. 31. N 2. P. 717–728. https://doi.org/10.32604/iasc.2022.020008
9. Ivanko D., Kashevnik A., Ryumin D., Kitenko A., Axyonov A., Lashkov I., Karpov A. MIDriveSafely: Multimodal interaction for drive safely // Proc. of the 2022 International Conference on Multimodal Interaction (ICMI). 2022. P. 733–735. https://doi.org/10.1145/3536221.3557037
10. Biswas A., Sahu P.K., Chandra M. Multiple cameras audio visual speech recognition using active appearance model visual features in car environment // International Journal of Speech Technology. 2016. V. 19. N 1. P. 159–171. https://doi.org/10.1007/s10772-016-9332-x
11. Nambi A.U., Bannur S., Mehta I., Kalra H., Virmani A., Padmanabhan V.N., Bhandari R., Raman B. HAMS: Driver and driving monitoring using a smartphone // Proc. of the 24th Annual International Conference on Mobile Computing and Networking. 2018. P. 840–842. https://doi.org/10.1145/3241539.3267723
12. Kashevnik A., Lashkov I., Gurtov A. Methodology and mobile application for driver behavior analysis and accident prevention // IEEE Transactions on Intelligent Transportation Systems. 2020. V. 21. N 6. P. 2427–2436. https://doi.org/10.1109/TITS.2019.2918328
13. Jang S.W., Ahn B. Implementation of detection system for drowsy driving prevention using image recognition and IoT // Sustainability. 2020. V. 12. N 7. P. 3037. https://doi.org/10.3390/su12073037
14. Mishra R.K., Urolagin S., Jothi J.A.A., Gaur P. Deep hybrid learning for facial expression binary classifications and predictions // Image and Vision Computing. 2022. V. 128. P. 104573. https://doi.org/10.1016/j.imavis.2022.104573
15. Sunitha G., Geetha K., Neelakandan S., Pundir A.K.S., Hemalatha S., Kumar V. Intelligent deep learning based ethnicity recognition and classification using facial images // Image and Vision Computing. 2022. V. 121. P. 104404. https://doi.org/10.1016/j.imavis.2022.104404
16. Yuan Y., Tian C., Lu X. Auxiliary loss multimodal GRU model in audio-visual speech recognition // IEEE Access. 2018. V. 6. P. 5573–5583. https://doi.org/10.1109/ACCESS.2018.2796118

References

1. Lin S.C., Hsu C.H., Talamonti W., Zhang Y., Oney S., Mars J., Tang L. Adasa: A conversational in-vehicle digital assistant for advanced driver assistance features. *Proc. of the 31st Annual ACM Symposium on User Interface Software and Technology*, 2018, pp. 531–542. https://doi.org/10.1145/3242587.3242593
2. Lee B., Hasegawa-Johnson M., Goudeseune C., Kamdar S., Borys S., Liu M., Huang T. AVICAR: Audio-visual speech corpus in a car environment. *Proc. of the 8th International Conference on Spoken Language Processing*, 2004, pp. 2489–2492. https://doi.org/10.21437/Interspeech.2004-424
3. Ivanko D., Ryumin D., Kashevnik A., Axyonov A., Karpov A. Visual speech recognition in a driver assistance system. *Proc. of the 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 1131–1135. https://doi.org/10.23919/EUSIPCO55093.2022.9909819
4. Xu B., Wang J., Lu C., Guo Y. Watch to listen clearly: Visual speech enhancement driven multi-modality speech recognition. *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 1637–1646. https://doi.org/10.1109/wacv45572.2020.9093314
5. Afouras T., Chung, J.S., Senior A., Vinyals O., Zisserman A. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, vol. 44, no. 12, pp. 8717–8727. https://doi.org/10.1109/TPAMI.2018.2889052
6. Kukharev G.A., Matveev Yu.N., Oleinik A.L. Mutual image transformation algorithms for visual information processing and retrieval. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2017, vol. 17, no. 1, pp. 62–74. (in Russian). https://doi.org/10.17586/2226-1494-2017-17-1-62-74
7. Shi B., Hsu W.N., Mohamed A. Robust self-supervised audio-visual speech recognition. *Proc. of the International Conference INTERSPEECH*, 2022, pp. 2118–2122. https://doi.org/10.21437/interspeech.2022-99
8. Chand H.V., Karthikeyan J. CNN based driver drowsiness detection system using emotion analysis. *Intelligent Automation & Soft Computing*, 2022, vol. 31, no. 2, pp. 717–728. https://doi.org/10.32604/iasc.2022.020008
9. Ivanko D., Kashevnik A., Ryumin D., Kitenko A., Axyonov A., Lashkov I., Karpov A. MIDriveSafely: Multimodal interaction for drive safely. *Proc. of the 2022 International Conference on Multimodal Interaction (ICMI)*, 2022, pp. 733–735. https://doi.org/10.1145/3536221.3557037
10. Biswas A., Sahu P.K., Chandra M. Multiple cameras audio visual speech recognition using active appearance model visual features in car environment. *International Journal of Speech Technology*, 2016, vol. 19, no. 1, pp. 159–171. https://doi.org/10.1007/s10772-016-9332-x
11. Nambi A.U., Bannur S., Mehta I., Kalra H., Virmani A., Padmanabhan V.N., Bhandari R., Raman B. HAMS: Driver and driving monitoring using a smartphone. *Proc. of the 24th Annual International Conference on Mobile Computing and Networking*, 2018, pp. 840–842. https://doi.org/10.1145/3241539.3267723
12. Kashevnik A., Lashkov I., Gurtov A. Methodology and mobile application for driver behavior analysis and accident prevention. *IEEE Transactions on Intelligent Transportation Systems*, 2020, vol. 21, no. 6, pp. 2427–2436. https://doi.org/10.1109/TITS.2019.2918328
13. Jang S.W., Ahn B. Implementation of detection system for drowsy driving prevention using image recognition and IoT. *Sustainability*, 2020, vol. 12, no. 7, pp. 3037. https://doi.org/10.3390/su12073037
14. Mishra R.K., Urolagin S., Jothi J.A.A., Gaur P. Deep hybrid learning for facial expression binary classifications and predictions. *Image and Vision Computing*, 2022, vol. 128, pp. 104573. https://doi.org/10.1016/j.imavis.2022.104573
15. Sunitha G., Geetha K., Neelakandan S., Pundir A.K.S., Hemalatha S., Kumar V. Intelligent deep learning based ethnicity recognition and classification using facial images. *Image and Vision Computing*, 2022, vol. 121, pp. 104404. https://doi.org/10.1016/j.imavis.2022.104404

17. Hou J.C., Wang S.S., Lai Y.H., Tsao Y., Chang H.W., Wang H.M. Audio-visual speech enhancement using multimodal deep convolutional neural networks // IEEE Transactions on Emerging Topics in Computational Intelligence. 2018. V. 2. N 2. P. 117–128. https://doi.org/10.1109/TETCI.2017.2784878
18. Chan Z.M., Lau C.Y., Thang K.F. Visual speech recognition of lips images using convolutional neural network in VGG-M model // Journal of Information Hiding and Multimedia Signal Processing. 2020. V. 11. N 3. P. 116–125.
19. Zhu X., Cheng D., Zhang Z., Lin S., Dai J. An empirical study of spatial attention mechanisms in deep networks // Proc. of the IEEE/CVF International Conference on Computer Vision. 2019. P. 6688–6697. https://doi.org/10.1109/iccv.2019.00679
20. Bhaskar S., Thasleema T.M. LSTM model for visual speech recognition through facial expressions // Multimedia Tools and Applications. 2023. V. 82. N 4. P. 5455–5472. https://doi.org/10.1007/s11042-022-12796-1
21. Hori T., Cho J., Watanabe S. End-to-end Speech recognition with word-based RNN language models // Proc. of the 2018 IEEE Spoken Language Technology Workshop (SLT). 2018. P. 389–396. https://doi.org/10.1109/SLT.2018.8639693
22. Serdyuk D.D., Braga O.P.F., Siohan O. Transformer-based video front-ends for audio-visual speech recognition for single and multi-person video // Proc. of the INTERSPEECH. 2022. P. 2833–2837. https://doi.org/10.21437/interspeech.2022-10920
23. Chen C.F.R., Fan Q., Panda R. CrossViT: Cross-attention multi-scale vision transformer for image classification // Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021. P. 347–356. https://doi.org/10.1109/iccv48922.2021.00041
24. Pan S.J., Yang Q. A survey on transfer learning // IEEE Transactions on Knowledge and Data Engineering. 2010. V. 22. N 10. P. 1345–1359. https://doi.org/10.1109/tkde.2009.191
25. Романенко А.Н., Матвеев Ю.Н., Минкер В. Перенос знаний в задаче автоматического распознавания русской речи в телефонных переговорах // Научно-технический вестник информационных технологий, механики и оптики. 2018. Т. 18. № 2. С. 236–242. https://doi.org/10.17586/2226-1494-2018-18-2-236-242
26. Sui C., Bennamoun M., Togneri R. Listening with your eyes: towards a practical visual speech recognition system using deep boltzmann machines // Proc. of the IEEE International Conference on Computer Vision (ICCV). 2015. P. 154–162. https://doi.org/10.1109/iccv.2015.26
27. Ahmed N., Natarajan T., Rao K.R. Discrete cosine transform // IEEE Transactions on Computers. 1974. V. C-23. N 1. P. 90–93. https://doi.org/10.1109/T-C.1974.223784
28. Xanthopoulos P., Pardalos P.M., Trafalis T.B. Linear discriminant analysis // Robust Data Mining. Springer New York, 2013. P. 27–33. https://doi.org/10.1007/978-1-4419-9878-1_4
29. Томашенко Н.А., Хохлов Ю.Ю., Ларшер Э., Эстев Я., Матвеев Ю.Н. Использование в системах автоматического распознавания речи GMM-моделей для адаптации акустических моделей, построенных на основе искусственных нейронных сетей // Научно-технический вестник информационных технологий, механики и оптики. 2016. Т. 16. № 6. С. 1063–1072. https://doi.org/10.17586/2226-1494-2016-16-6-1063-1072
30. Ma P., Petridis S., Pantic M. End-to-end audio-visual speech recognition with conformers // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021. P. 7613–7617. https://doi.org/10.1109/ICASSP39728.2021.9414567
31. Ryumin D., Ivanko D., Ryumina E. Audio-visual speech and gesture recognition by sensors of mobile devices // Sensors. 2023. V. 23. N 4. P. 2284. https://doi.org/10.3390/s23042284
32. Huang J., Kingsbury B. Audio-visual deep learning for noise robust speech recognition // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing. 2013. P. 7596–7599. https://doi.org/10.1109/ICASSP.2013.6639140
33. Ivanko D., Ryumin D., Kashevnik A., Axyonov A., Kitenko A., Lashkov I., Karpov A. DAVIS: Driver's audio-visual speech recognition // Proc. of the International Conference INTERSPEECH. 2022. P. 1141–1142.
34. Zhou P., Yang W., Chen W., Wang Y., Jia J. Modality attention for end-to-end audio-visual speech recognition // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019. P. 6565–6569. https://doi.org/10.1109/ICASSP.2019.8683733

16. Yuan Y., Tian C., Lu X. Auxiliary loss multimodal GRU model in audio-visual speech recognition. *IEEE Access*, 2018, vol. 6, pp. 5573–5583. https://doi.org/10.1109/ACCESS.2018.2796118
17. Hou J.C., Wang S.S., Lai Y.H., Tsao Y., Chang H.W., Wang H.M. Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2018, vol. 2, no. 2, pp. 117–128. https://doi.org/10.1109/TETCI.2017.2784878
18. Chan Z.M., Lau C.Y., Thang K.F. Visual speech recognition of lips images using convolutional neural network in VGG-M model. *Journal of Information Hiding and Multimedia Signal Processing*, 2020, vol. 11, no. 3, pp. 116–125.
19. Zhu X., Cheng D., Zhang Z., Lin S., Dai J. An empirical study of spatial attention mechanisms in deep networks. *Proc. of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6688–6697. https://doi.org/10.1109/iccv.2019.00679
20. Bhaskar S., Thasleema T.M. LSTM model for visual speech recognition through facial expressions. *Multimedia Tools and Applications*, 2023, vol. 82, no. 4, pp. 5455–5472. https://doi.org/10.1007/s11042-022-12796-1
21. Hori T., Cho J., Watanabe S. End-to-end Speech recognition with word-based RNN language models. *Proc. of the 2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 389–396. https://doi.org/10.1109/SLT.2018.8639693
22. Serdyuk D.D., Braga O.P.F., Siohan O. Transformer-based video front-ends for audio-visual speech recognition for single and multi-person video. *Proc. of the INTERSPEECH*, 2022, pp. 2833–2837. https://doi.org/10.21437/interspeech.2022-10920
23. Chen C.F.R., Fan Q., Panda R. CrossViT: Cross-attention multi-scale vision transformer for image classification. *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 347–356. https://doi.org/10.1109/iccv48922.2021.00041
24. Pan S.J., Yang Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 2010, vol. 22, no. 10, pp. 1345–1359. https://doi.org/10.1109/tkde.2009.191
25. Romanenko A.N., Matveev Yu.N., Minker W. Knowledge transfer for Russian conversational telephone automatic speech recognition. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2018, vol. 18, no. 2, pp. 236–242 (in Russian). https://doi.org/10.17586/2226-1494-2018-18-2-236-242
26. Sui C., Bennamoun M., Togneri R. Listening with your eyes: towards a practical visual speech recognition system using deep boltzmann machines. *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 154–162. https://doi.org/10.1109/iccv.2015.26
27. Ahmed N., Natarajan T., Rao K.R. Discrete cosine transform. *IEEE Transactions on Computers*, 1974, vol. C-23, no. 1, pp. 90–93. https://doi.org/10.1109/T-C.1974.223784
28. Xanthopoulos P., Pardalos P.M., Trafalis T.B. Linear discriminant analysis. *Robust Data Mining*, Springer New York, 2013, pp. 27–33. https://doi.org/10.1007/978-1-4419-9878-1_4
29. Tomashenko N.A., Khokhlov Yu.Yu., Larcher A., Estève Ya., Matveev Yu.N. Gaussian mixture models for adaptation of deep neural network acoustic models in automatic speech recognition systems. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2016, vol. 16, no. 6, pp. 1063–1072. (in Russian). https://doi.org/10.17586/2226-1494-2016-16-6-1063-1072
30. Ma P., Petridis S., Pantic M. End-to-end audio-visual speech recognition with conformers. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7613–7617. https://doi.org/10.1109/ICASSP39728.2021.9414567
31. Ryumin D., Ivanko D., Ryumina E. Audio-visual speech and gesture recognition by sensors of mobile devices. *Sensors*, 2023, vol. 23, no. 4, pp. 2284. https://doi.org/10.3390/s23042284
32. Huang J., Kingsbury B. Audio-visual deep learning for noise robust speech recognition. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7596–7599. https://doi.org/10.1109/ICASSP.2013.6639140
33. Ivanko D., Ryumin D., Kashevnik A., Axyonov A., Kitenko A., Lashkov I., Karpov A. DAVIS: Driver's audio-visual speech recognition. *Proc. of the International Conference INTERSPEECH*, 2022, pp. 1141–1142.
34. Zhou P., Yang W., Chen W., Wang Y., Jia J. Modality attention for end-to-end audio-visual speech recognition. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*

35. Ivanko D., Axyonov A., Ryumin D., Kashevnik A., Karpov A. RUSAVIC Corpus: Russian audio-visual speech in cars // Proc. of the 13th Language Resources and Evaluation Conference (LREC). 2022. P. 1555–1559.
36. Kashevnik A., Lashkov I., Axyonov A., Ivanko D., Ryumin D., Kolchin A., Karpov A. Multimodal corpus design for audio-visual speech recognition in vehicle cabin // IEEE Access. 2021. V. 9. P. 34986–35003. https://doi.org/10.1109/ACCESS.2021.3062752
37. Lugaresi C., Tang J., Nash H., McClanahan C., Uboweja E., Hays M., Zhang F., Chang C.-L., Yong M., Lee J., Chang W.-T., Hua W., Georg M., Grundmann M. MediaPipe: A framework for perceiving and processing reality // Proc. of the 3rd Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR). 2019. V. 2019. P. 1–4.
38. Zhang H., Cisse M., Dauphin Y.N., Lopez-Paz D. MixUp: Beyond empirical risk minimization // Proc. of the ICLR Conference. 2018. P. 1–13.
39. Feng D., Yang S., Shan S. An efficient software for building LIP reading models without pains // Proc. of the IEEE International Conference on Multimedia & Expo Workshops (ICMEW). 2021. P. 1–2. https://doi.org/10.1109/ICMEW53276.2021.9456014
40. Kim M., Hong J., Park S.J., Ro Y.M. Multi-modality associative bridging through memory: speech sound recollected from face video // Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021. P. 296–306. https://doi.org/10.1109/iccv48922.2021.00036
41. Zhong Z., Lin Z.Q., Bidart R., Hu X., Daya I.B., Li Z., Zheng W., Li J., Wong A. Squeeze-and-attention networks for semantic segmentation // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. P. 13065–13074. https://doi.org/10.1109/cvpr42600.2020.01308
42. Аксёнов А.А., Рюмин Д.А., Кашевник А.М., Иванько Д.В., Карпов А.А. Метод визуального анализа лица водителя для автоматического чтения речи по губам при управлении транспортным средством // Компьютерная оптика. 2022. Т. 46. № 6. С. 955–962. https://doi.org/10.18287/2412-6179-CO-1092

(ICASSP), 2019, pp. 6565–6569. https://doi.org/10.1109/ICASSP.2019.8683733
35. Ivanko D., Axyonov A., Ryumin D., Kashevnik A., Karpov A. RUSAVIC Corpus: Russian audio-visual speech in cars. *Proc. of the 13th Language Resources and Evaluation Conference (LREC)*, 2022, pp. 1555–1559.
36. Kashevnik A., Lashkov I., Axyonov A., Ivanko D., Ryumin D., Kolchin A., Karpov A. Multimodal corpus design for audio-visual speech recognition in vehicle cabin. *IEEE Access*, 2021, vol. 9, pp. 34986–35003. https://doi.org/10.1109/ACCESS.2021.3062752
37. Lugaresi C., Tang J., Nash H., McClanahan C., Uboweja E., Hays M., Zhang F., Chang C.-L., Yong M., Lee J., Chang W.-T., Hua W., Georg M., Grundmann M. MediaPipe: A framework for perceiving and processing reality. *Proc. of the 3rd Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR)*, 2019, vol. 2019, pp. 1–4.
38. Zhang H., Cisse M., Dauphin Y.N., Lopez-Paz D. MixUp: Beyond empirical risk minimization. *Proc. of the ICLR Conference*, 2018, pp. 1–13.
39. Feng D., Yang S., Shan S. An efficient software for building LIP reading models without pains. *Proc. of the IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2021, pp. 1–2. https://doi.org/10.1109/ICMEW53276.2021.9456014
40. Kim M., Hong J., Park S.J., Ro Y.M. Multi-modality associative bridging through memory: speech sound recollected from face video. *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 296–306. https://doi.org/10.1109/iccv48922.2021.00036
41. Zhong Z., Lin Z.Q., Bidart R., Hu X., Daya I.B., Li Z., Zheng W., Li J., Wong A. Squeeze-and-attention networks for semantic segmentation. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13065–13074. https://doi.org/10.1109/cvpr42600.2020.01308
42. Axyonov A.A., Ryumin D.A., Kashevnik A.M., Ivanko D.V., Karpov A.A. Method for visual analysis of driver's face for automatic lip-reading in the wild. *Computer Optic*, 2022, vol. 46, no. 6, pp. 955–962. (in Russian). https://doi.org/10.18287/2412-6179-CO-1092

Авторы

Аксёнов Александр Александрович — младший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 57203963345, https://orcid.org/0000-0002-7479-2851, axyonov.a@iias.spb.su

Рюмина Елена Витальевна — младший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 57220572427, https://orcid.org/0000-0002-4135-6949, ryumina.e@iias.spb.su

Рюмин Дмитрий Александрович — кандидат технических наук, старший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 57191960214, https://orcid.org/0000-0002-7935-0569, dl_03.03.1991@mail.ru

Иванько Денис Викторович — кандидат технических наук, старший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 57190967993, https://orcid.org/0000-0003-0412-7765, ivanko.d@iias.spb.su

Карпов Алексей Анатольевич — доктор технических наук, профессор, заведующий лабораторией, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 57219469958, https://orcid.org/0000-0003-3424-652X, karpov@iias.spb.su

Authors

Alexandr A. Axyonov —Junior Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation, SC 57203963345, https://orcid.org/0000-0002-7479-2851, axyonov.a@iias.spb.su

Elena V. Ryumina — Junior Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation, SC 57220572427, https://orcid.org/0000-0002-4135-6949, ryumina.e@iias.spb.su

Dmitry A. Ryumin — PhD, Senior Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation, SC 57191960214, https://orcid.org/0000-0002-7935-0569, dl_03.03.1991@mail.ru

Denis V. Ivanko — PhD, Senior Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation, SC 57190967993, https://orcid.org/0000-0003-0412-7765, ivanko.d@iias.spb.su

Alexey A. Karpov — D.Sc., Professor, Head of Laboratory, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation, SC 57219469958, https://orcid.org/0000-0003-3424-652X, karpov@iias.spb.su

Статья поступила в редакцию 12.04.2023
Одобрена после рецензирования 17.05.2023
Принята к печати 24.07.2023

Received 12.04.2023
Approved after reviewing 17.05.2023
Accepted 24.07.2023