Метод визуального анализа лица водителя для автоматического чтения речи по губам при управлении транспортным средством

А.А. Аксёнов [1], Д.А. Рюмин [1], А.М. Кашевник [1], Д.В. Иванько [1], А.А. Карпов [1]
[1] Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН), 199178, Российская Федерация, г. Санкт-Петербург, 14-я линия В.О., д. 39

Аннотация

В работе предложен метод визуального анализа и чтения речи по губам водителя при управлении транспортным средством. Автоматическое распознавание речи в акустически неблагоприятных динамических условиях является одной из актуальных задач искусственного интеллекта. Проблема эффективного автоматического чтения по губам во время дорожного движения на данный момент не решена из-за наличия различного рода помех (частые повороты головы, вибрация, динамическое освещение и т.п.). Кроме того, проблема усложняется отсутствием представительных баз данных визуальной речи. Для поиска и извлечения области интереса используется программная библиотека MediaPipe Face Mesh. Для анализа визуальной речи разработана интегральная нейросетевая архитектура (End-to-End). Визуальные признаки извлекаются из отдельного изображения с помощью свёрточной нейронной сети в связке с полносвязным слоем. Извлеченные нейросетевые признаки изображений являются входными данными для нейросети с длинной кратковременной памятью. В связи с небольшим объемом обучающих данных было предложено применять метод переноса обучения. Результаты по анализу и распознаванию визуальной речи водителя в процессе управления автомобилем представляют большие возможности для решения актуальной задачи автоматического чтения по губам. Экспериментальные исследования выполнены на собственном аудиовизуальном корпусе русской речи RUSAVIC, собранном в реальных условиях дорожного движения. Максимальная точность визуального распознавания 62 голосовых управляющих команд водителей составила 64,09%. Полученные результаты могут быть использованы в системах аудиовизуального распознавания речи, применяемых в акустически сложной обстановке дорожного движения (высокая скорость движения, открытые окна или люк в автомобиле, одновременное проигрывание музыки, слабая шумоизоляция и т.п.).

Ключевые слова: транспортное средство, водитель, визуальное распознавание речи, автоматическое чтение по губам, машинное обучение, End-to-End, CNN, LSTM.

Цитирование: **Аксёнов, А.А.** Метод визуального анализа лица водителя для автоматического чтения речи по губам при управлении транспортным средством / А.А. Аксёнов, Д.А. Рюмин, А.М. Кашевник, Д.В. Иванько, А.А. Карпов // Компьютерная оптика. – 2022. – Т. 46, № 6. – С. 955-962. – DOI: 10.18287/2412-6179-CO-1092.

Citation: Axyonov AA, Ryumin DA, Kashevnik AM, Ivanko DV, Karpov AA. Method for visual analysis of driver's face for automatic lip-reading in the wild. Computer Optics 2022; 46(6): 955-962. DOI: 10.18287/2412-6179-CO-1092.

Введение

Согласно отчету Всемирной организации здравоохранения, использование мобильных устройств (например, отправка текстовых сообщений, включение/выключение музыки или выполнение других действий) во время управления транспортным средством (ТС) ведет к четырехкратному возрастанию риска дорожно-транспортных происшествий (ДТП). Ежегодно общее количество травм различной степени тяжести, зачастую приводящих к инвалидности, достигает отметки в 20 – 50 млн человек [1]. Также, по статистике Главного управления по обеспечению безопасности дорожного движения Министерства внутренних дел Российской Федерации, за 2020 год в стране зафиксировано более 137 тыс. ДТП [2]. При этом часто водители отвлекаются от дороги и при взаимодействии с информационно-развлекательной системой автомобиля, помимо использования телефона. Возможно частично решить такую проблему отвлечения водителя за счет использования технологий автоматического распознавания речи для управления такими системами.

Несмотря на прогресс цифровых технологий, достигнутых в последние годы, системы автоматического распознавания речи не всегда способны функционировать с высокими показателями эффективности (точность, скорость распознавания). При наличии контролируемых офисных условий, ограниченного словаря и контролируемой грамматики распознаваемых команд точность современных систем распознавания речи по аудиомодальности (звучащей речи)

может приближаться к 100 % [3]. Тем не менее, в случае сложной динамической окружающей акустической среды (внешний шум, реверберация, помехи в микрофонном канале и т.д.) точность автоматического распознавания речи значительно снижается. В то же время важнейшим, если не решающим фактором использования систем распознавания речи для автоматизации определенных действий в кабине транспортного средства является их способность воспринимать речь водителя в условиях сложной акустической обстановки. Опираясь на вышеизложенное, можно считать, что акустический речевой сигнал служит основной модальностью в системах автоматического распознавания речи, но, помимо него, имеет смысл использовать и визуальную информацию о речи (движения губ диктора), благодаря чему возможно улучшить точность распознавания речи, особенно в условиях, когда акустический сигнал зашумлен или недоступен.

Распознавание визуальной речи (чтение речи по губам) является достаточно сложным навыком для человека, однако в акустически шумных условиях и при разговоре большого количества людей собеседники сами начинают обращать внимание на губы друг друга с целью лучшего понимания смысла высказываний [4]. Восприятие речи человеком – многомодальный процесс, и, основываясь на этом, в последние годы удалось улучшить показатели эффективности систем автоматического распознавания речи благодаря доступности представительных аудиовизуальных корпусов [5, 6] и усовершенствованию архитектур нейросетей [7, 8].

Задача настоящего исследования заключается в анализе визуальной речи водителя для распознавания речи и выполнения определенных действий (управление звонками, мультимедийными данными, навигационной системой и т.д.), что позволит взаимодействовать с мобильным устройством в шумных условиях, когда акустическая речь малоэффективна. Исследуется только дикторозависимый сценарий, так как обычно у персонального транспортного средства один владелец и он же водитель, и необходимо определять только его голосовые команды с высокой точностью.

1. Исследования в области визуального распознавания речи

В настоящее время ведущие научные институты и мировые промышленные корпорации, работающие в области искусственного интеллекта, активно проводят исследования, направленные на создание высокоэффективных систем распознавания визуальной речи [9 – 11]. В работах [12, 13] авторами выделяются следующие проблемы автоматического распознавания визуальной речи:

- устойчивое детектирование области интереса (область рта);
- извлечение наиболее информативных признаков из визуальной речи;
- эффективное моделирование и распознавание визуальной речи диктора (как изолированных слов, так и слитной речи).

Ряд современных работ [10, 14] посвящен методам извлечения визуальных признаков из заранее детектированной области интереса за счет комбинирования различных архитектур нейросетей с линейными классификаторами.

Стоит отметить, что этап детектирования области рта диктора оказывает существенное влияние на итоговую эффективность распознавания визуальной речи. Наиболее распространенные решения применительно к этой задаче (базовые подходы) включают в себя методы на основе примитивов Хаара [15] и методы на основе моделей активного внешнего вида (от англ. Active Appearance Model, AAM) [16].

На сегодняшний день классический подход к распознаванию аудиовизуальной речи постепенно заменяется End-to-End (E2E) интегральным подходом, т.е. каскадом нейронных сетей. В первом приближении E2E-подход близок к традиционным методам: последовательность изображений рта подается в сверточную нейросеть для извлечения признаков [17, 18], которые затем передаются во внутреннюю модель (RNN, LSTM, GRU или др.) для учета временной зависимости и классификации [19, 20]. Проведенные исследования демонстрируют, что извлеченные таким образом признаки больше подходят для автоматического чтения речи по губам, чем рассчитываемые традиционными методами.

Основным преимуществом современного подхода является то, что вся система состоит из единой нейросети. Таким образом, извлеченные признаки лучше связаны с данными, на которых обучается сеть. В работе [21] впервые было предложено использовать CNN для замены блока извлечения признаков. В свою очередь, в работе [22] впервые предложено использовать LSTM для задачи классификации. Позднее исследователи в [23] предложили нейросеть для извлечения акустических признаков и попытались объединить их с видеоинформацией.

Также следует выделить тот факт, что большая часть современных исследований используют аудиовизуальные корпусы, которые записаны в контролируемых офисных условиях без учета вариативности освещения, окклюзии, активных поворотов головы диктора и т.д. Очевидно, что для анализа визуальных данных водителя в момент управления транспортным средством такие корпусы не подходят, именно поэтому предыдущая работа авторов [24] посвящена разработке универсальной методики и программных средств для создания многомодальных речевых корпусов. Результатом реализации предложенной методики стал многодикторный аудиовизуальный корпус (бимодальная речевая база данных) слитной русской

речи с разноракурсными видеоданными, записанный в кабине транспортного средства (от англ. RUSsian Audio-Visual speech in Cars, RUSAVIC) [25, 26].

2. Многодикторный аудиовизуальный корпус RUSAVIC

Автоматическое распознавание речи, особенно в шумных условиях, представляет собой достаточно сложную задачу. Одним из важнейших этапов при распознавании речи является правильное определение границ речи во входящем аудиопотоке. Для изолированных слов данная проблема сводится к нахождению верной границы между словами, но если речь идет о слитной речи, то эта задача намного сложнее, из-за того что речь идет сплошным потоком, как правило, с минимальной паузой.

Запись многодикторного аудиовизуального корпуса RUSAVIC производилась в кабине автомобиля как в условиях дорожного движения, так и в режиме холостого хода транспортного средства, т.е. в натурных и полунатурных условиях, приближенных к реальным условиям вождения. Процесс автоматизированного аннотирования корпуса RUSAVIC осуществлялся при помощи разработанного метода, функциональная схема которого представлена на рис. 1.

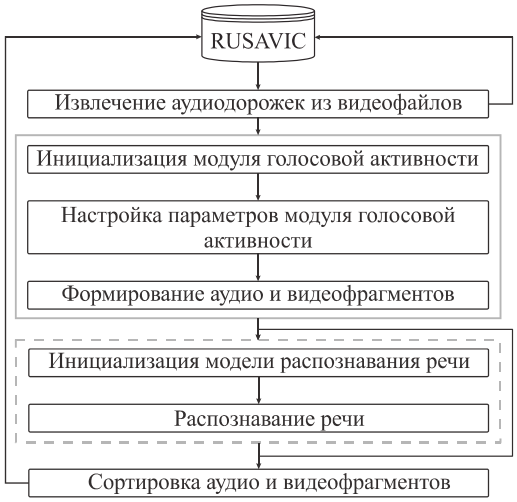

Рис. 1. Функциональная схема метода автоматического аннотирования аудиовизуальных речевых данных

Данный метод объединяет в себе современные алгоритмы машинного обучения и компьютерного зрения, что позволяет в автоматическом режиме разделять аудиовизуальные данные на фрагменты, которые отсортированы в соответствии с указанным фильтром (словарём распознавания). Логически фрагменты обеих модальностей представлены в виде модели иерархического типа для их физического хранения.

В рамках данного исследования для сбора корпуса использовалось мобильное приложение Drive Safely [27, 28], которое позволяет расставлять временные метки распознанных фраз водителя. В модуль голосовой активности поступают уже известные фразы на видеоданных, и нет необходимости использовать модель распознавания речи для определения речи на видеофрагменте, исходя из этого на функциональной схеме (рис. 1) блок с распознаванием речи выделен пунктиром, так как его использование является опциональным. Модуль голосовой активности в данном случае будет использоваться только для улучшения входных видеофрагментов за счет более точного определения границ речи в поданных на вход фрагментах речи (уменьшение ненужной информации).

Основные характеристики многодикторного аудиовизуального корпуса RUSAVIC представлены в табл. 1. Примеры видеоданных и словарь корпуса продемонстрированы на вебсайте [29]. Так как в статье нет примеров видеоданных.

Табл. 1. Основные характеристики корпуса RUSAVIC

Характеристика	Значение
Количество сеансов записи	10
Количество фраз в словаре	62
Количество дикторов	5
Частота кадров, к/с	60
Формат видеозаписи	mp4
Разрешение видео, пиксель	1920×1080
Частота дискретизации, Hz	48 000
Общий размер данных, ГБ	~50
Общее число фраз	3 100

3. Метод автоматического чтения речи водителя по губам

Результаты современных исследований [30 – 32] демонстрируют, что методы машинного обучения на основе интегральных нейросетевых моделей превосходят по точности базовые подходы автоматического распознавания визуальной речи, однако могут уступать им по скорости. В связи с этим для анализа визуальной речи диктора был разработан собственный метод чтения речи водителя по губам, который включает в себя создание нейросетевой модели путем ее обучения на записанных в реальных условиях видеоданных водителей, произносящих заданные фразы (в рамках создания корпуса RUSAVIC). Были проведены эксперименты (см. параграф 4) с несколькими архитектурами нейросетей, и выбрана наилучшая по точности распознавания на основе эмпирических экспериментов.

Для распознавания речи была предложена следующая последовательность операций (см. рис. 2): считывание видеоданных с камеры смартфона, поиск графических областей с лицами и пометка ближайшего лица как основного, слежение за основным лицом, обнаружение области губ, подача последовательности изображений, включающей области губ на вход нейросети, классификация этой последовательности путем отнесения к той или иной фразе, выполнение голосовой команды.

Обнаружение области губ базируется на современном подходе к определению 468 трехмерных (3D) лицевых ориентиров, который реализован в крос-

сплатформенной среде с открытым исходным кодом MediaPipe [33] (регрессионная модель Face Mesh). На рис. 3*а* представлены исходные изображения. Пример работы Face Mesh продемонстрирован на рис. 3*б*. На рис. 3*в* представлен ряд примеров извлеченных графических изображений рта.

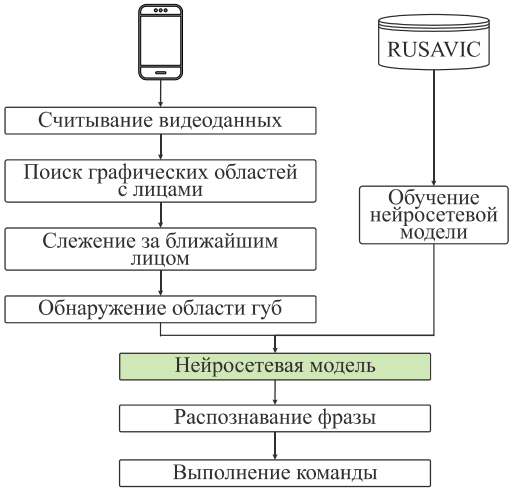

Рис. 2. Метод автоматического чтения речи водителя по губам

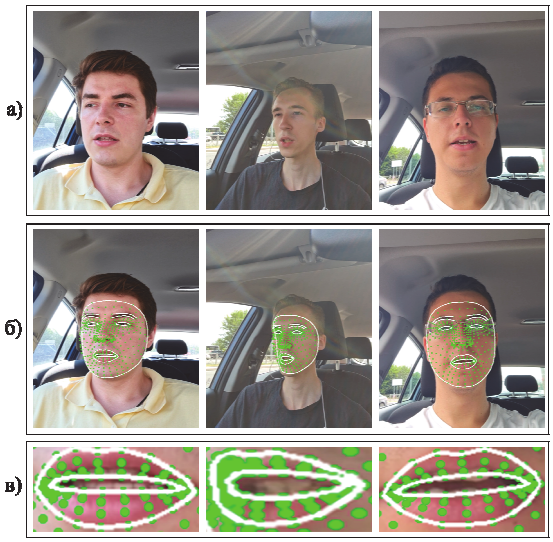

Рис. 3. Примеры обработки изображений: а) входные видеокадры информантов; б) результат работы MediaPipe Face Mesh; в) извлеченные области рта

Входными данными нейросетевой модели (выделено цветом на рис. 2) являются последовательности изображений области губ диктора. Сначала выполняется нормализация всех изображений до заданного размера (112×112 пикселей). Далее, в связи с небольшим обучающим набором данных (10 повторений для каждой фразы водителя) применяется метод переноса обучения (от англ. Transfer Learning [34]).

Архитектура предложенной нейросетевой модели представлена на рис. 4.

В рамках предложенной архитектуры нейросети входной информацией являются последовательности изображений длиной в 32 кадра (Sequence_Length) с разрешением 112×112 пикселей, которые проходят через 3D сверточный слой (3D Conv) и модифицированные остаточные блоки (Residual Blocks модели ResNet-18) с модулями внимания (Squeeze-and-Attention, SA). Данные блоки извлекают карты признаков размерностью 512×7×7 для каждого изображения из всей последовательности. Затем слой подвыборки (Global Average Polling) преобразует их в одномерные вектора, которые подаются на двунаправленные сети с длинной кратковременной памятью (BiLSTM) для последующего распознавания фраз водителя транспортного средства (см. рис. 4).

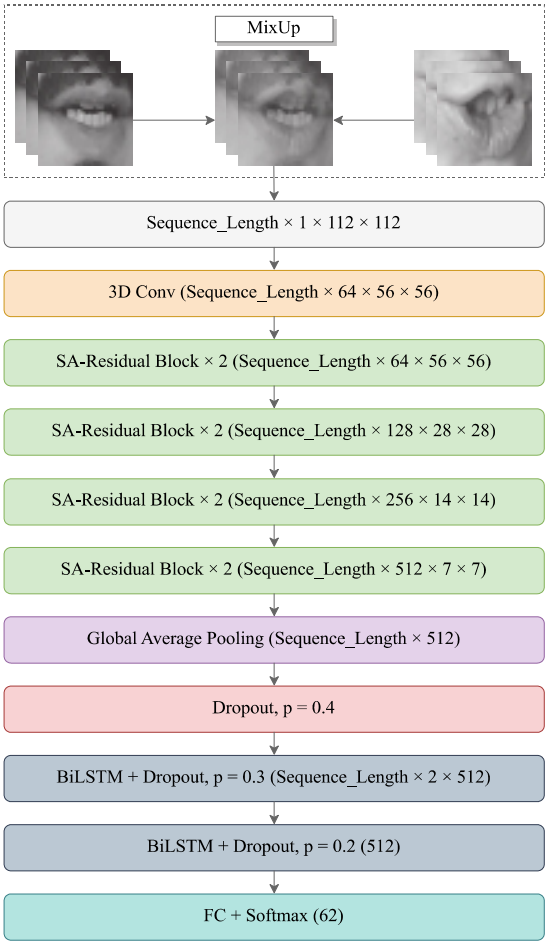

Рис. 4. Архитектура нейросетевой модели для распознавания визуальной речи

Поступающие на вход видеопоследовательности разделяются на сегменты одинаковой длины в 32 кадра с 50 % перекрытием (16 кадров), при этом если кадров не хватает (конец видео), то остаток заполняется последним кадром. Для уменьшения вычислительных затрат входные изображения преобразуются в градации серого и нормализуются до 112×112 пикселей. Помимо этого, выполняется выравнивание гистограммы яркости изображения [35].

Для предотвращения переобучения нейросетевой модели 40 % меток и изображений (выбранных случайным образом) применяется техника аугмен-

тации данных MixUp. Коэффициент объединения двух изображений и бинарных векторов варьировался от 20 до 80 % таким образом, чтобы сумма всегда была равна 100 % (нулевая прозрачность). При этом техника MixUp была применена как для изображений, так и для бинарных векторов по следующим формулам:

$$\tilde{x} = \lambda \times x_1 + (1-\lambda) \times x_2, \quad (1)$$

$$\tilde{y} = \lambda \times y_1 + (1-\lambda) \times y_2, \quad (2)$$

где $\tilde{x}$ и $\tilde{y}$ – сгенерированные новые изображение и вектор меток, λ – коэффициент объединения двух изображений и бинарных векторов, x_1 и x_2 – первое и второе случайные изображения, y_1 и y_2 – первый и второй бинарные вектора, соответствующие случайным изображениям. Процесс генерации новых изображений осуществлялся попарно для всех изображений из двух случайно выбранных последовательностей из корпуса RUSAVIC.

Для меток, к которым не применялся MixUp (оставшиеся 60 %), применено их сглаживание (Label Smoothing, LS [36]). Более детально процесс сглаживания можно описать следующим образом: бинарный вектор размерностью 62 (один элемент вектора имеет значение 1, а остальные – 0) преобразуется в вектор меток, в котором элемент со значением 1 принимает значение 0,8 (данное значение было определено эмпирически), оставшийся 61 элемент принимает значение, в сумме равное 0,2. Сглаживание меток выполняется по формуле:

$$\tilde{y} = (1-\alpha) \times y + \alpha / K, \quad (3)$$

где $\tilde{y}$ – сгенерированный новый вектор меток, α – коэффициент, отвечающий за степень сглаживания бинарного вектора, y – исходный бинарный вектор, K – количество классов-команд, равное 62.

Для извлечения признаков было предложено использовать модифицированную нейросетевую модель 3DResNet-18 [37] с добавлением модуля SA [38], которая была обучена с нуля, по причине того, что в архитектуру был добавлен модуль внимания SA с отключением последнего слоя.

Извлеченные признаки было предложено подавать на два слоя BiLSTM по 512 нейронов в каждом. Выходом первого слоя BiLSTM является «многие ко многим» (sequence-to-sequence), на входе и на выходе которого по 32 кадра (вектора признаков). При этом выходом второго слоя BiLSTM является «многие к одному» (sequence-to-one, для всей последовательности на выходе будет один вектор признаков размерностью 512). Последний полносвязный слой (Fully Connected, FC) с количеством нейронов, равным 62, выдает вектор вероятностных значений, сумма которых равна 1. Индекс правильно предсказанной фразы имеет наибольшее вероятностное значение.

4. Реализация метода и экспериментальные исследования

Разработанный метод был апробирован на данных из корпуса RUSAVIC, которые не использовались при обучении нейросетевой модели. Были проведены эксперименты как на предложенной архитектуре нейронной сети, так и на иных, наиболее эффективных и доступных на сегодняшний день для решения данной задачи. Реализация была выполнена на языке программирования Python v.3.8, для обучения использовался фреймворк PyTorch v.1.10.0 + TorchVision v.1.11.1.

Во время обучения нейросетевой модели было предложено использовать технику планировщика скорости обучения (Cosine Annealing Warm Restarts, Cosine WR), значение функции которой предложено варьировать от 0,0001 до 0,001 [39]. Скорость обучения вычисляется согласно следующей формуле [39]:

$$lr = \frac{lr_{start}}{2}\left(\cos\left(\frac{\pi \cdot mod\left(epoch_{curr} - 1, \left[\frac{epoch}{M}\right]\right)}{\left[\frac{epoch}{M}\right]}\right) + 1\right), \quad (4)$$

где lr_{start} – начальная скорость обучения, cos () – косинус числа, *mod* () – остаток от деления, $epoch_{curr}$ – текущая эпоха, *epoch* – количество эпох, *M* – количество циклов.

На рис. 5 представлен график изменения скорости обучения модели (*lr*) относительно количества эпох, которое было определено эмпирически.

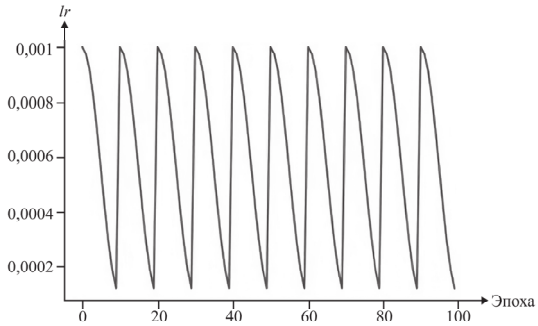

Рис. 5. Зависимость скорости обучения модели от количества эпох

Изначально максимальное количество эпох было предложено установить равным 100, а в случае, если на протяжении пяти эпох точность перестает расти, то обучение прерывается и фиксируется лучший результат, полученный за все время обучения. Рис. 5 показывает, что скорость обучения постепенно снижается с 0,001 до 0,0001 на протяжении 10 эпох, затем резко восстанавливается обратно до исходного значения. Таким образом, за 100 эпох выполняется 10 циклов снижения скорости обучения.

Значение скорости обучения позволяет управлять величиной корреляции весов в процессе всего обу-

чения, т.е. чем выше значение скорости обучения, тем выше будет и корреляция весов. Отсюда следует, что при высокой скорости обучения показатель эффективности может остановиться на локальном минимуме функции ошибки и не выявить глобальный. Для того чтобы не допустить этого, применяются различные планировщики скорости обучения. Используемый в данной работе планировщик скорости обучения имеет несколько значений скорости обучения в интервале от 0,001 до 0,0001. Это позволяет не останавливаться на локальном минимуме, а достичь оптимально возможного, в рамках экспериментальной установки, глобального минимума функции ошибки.

Сравнение различных известных архитектур нейросетевых моделей с предложенной в данной статье моделью представлено в табл. 2. Наилучший результат точности распознавания составляет 64,09 % (596 правильно распознанных фраз из 930).

Табл. 2. Результаты экспериментальных исследований по точности распознавания визуальной речи

№	Архитектура нейросетевой модели	Точность
1	3DResNet-18 + BiLSTM	46,45 % (432 правильных)
2	3DResNet-18 + BiLSTM + Cosine WR	48,28 % (449)
3	3DResNet-18 + MixUp + BiLSTM	49,14 % (457)
4	LS + 3DResNet-18 + BiLSTM	49,57 % (461)
5	SA + 3DResNet-18 + BiLSTM	55,59 % (517)
6	Предложенная архитектура: **LS + MixUp + SA + 3DResNet-18 + + BiLSTM + Cosine WR**	64,09 % (596)

Из табл. 2 следует, что при применении всех представленных техник (Cosine WR, MixUp, LS и SA) точность распознавания 62 голосовых управляющих команд водителей возросла с 46,45 % до 64,09 % (или на 17,64 %). Можно заметить, что значительный вклад в повышение точности достигается за счет модуля SA, а техники аугментации данных (MixUp и LS) дают приблизительно одинаковый прирост в точности. Меньшее влияние на точность оказывает Cosine WR. Однако несмотря на достигнутый результат точности (64,09 %), этого недостаточно для внедрения предложенного метода автоматического чтения речи водителя по губам в системы автоматизации определенных действий в кабине транспортного средства. Анализируя результаты распознавания, было также замечено, что на результат влияет также и предложенный словарь, т.к. в нем содержатся фразы с похожим звучанием и произношением. Например, система чаще всего ошибалась в конструкциях, где встречаются слова «Включить…» и «Отключить…», а также во фразах, в которых присутствует слово «Найти…». Следующим этапом для повышения точности планируется усовершенствование метода за счет добавления аудиомодальности.

Заключение

В работе предложен новый метод визуального анализа и чтения речи по губам водителя во время управления транспортным средством, который может использоваться в системах аудиовизуального распознавания речи, предназначенных для использования в акустически неблагоприятной обстановке, на которую влияют такие факторы, как различная скорость движения транспортного средства по дорогам с различным покрытием, степень открытия стекол, люка, наличие включенных радио /музыки, низкое качество шумоизоляции в автомобиле и т.д. Результаты экспериментов показали, что предложенный метод позволяет визуально распознавать произносимую водителем команду из словаря, содержащего 62 русскоязычных фразы, с точностью 64,09 %, что, несомненно, является хорошим показателем для такой задачи при использовании реальных зашумленных данных. В дальнейших исследованиях планируется разработать диктoронезависимую систему распознавания аудиовизуальной русской речи в кабине транспортного средства. Для этой цели будет увеличен объем данных и количество дикторов в многодикторном аудиовизуальном корпусе RUSAVIC, а также исследованы другие архитектуры на основе глубоких нейросетей для улучшения качества распознавания речи.

Благодарности

Работа выполнена при поддержке проекта фонда РФФИ № 19-29-09081-мк, ведущей научной школы НШ-17.2022.1.6, а также частично в рамках бюджетной темы № FFZF-2022-0005.

References

[1] Road traffic injuries. Source: ⟨https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries⟩.
[2] Indicators of road safety. Source: ⟨http://stat.gibdd.ru⟩.
[3] Ivanko D, Ryumin D. A novel task-oriented approach toward automated lip-reading system implementation. Int Arch Photogramm Remote Sens Spatial Inf Sci 2021; XLIV-2/W1-2021: 85-89. DOI: 10.5194/isprs-archives-XLIV-2-W1-2021-85-2021.
[4] McGurk H, MacDonald J. Hearing lips and seeing voices. Nature 1976; 264: 746-748.
[5] Chung JS, Zisserman A. Lip reading in the wild. Asian Conf on Computer Vision (ACCV) 2016: 87-103. DOI: 10.1007/978-3-319-54184-6_6.
[6] Yang S, Zhang Y, Feng D, Yang M, Wang C, Xiao J, Chen X. LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. Int Conf on Automatic Face and Gesture Recognition (FG) 2019: 1-8. DOI: 10.1109/FG.2019.8756582.
[7] Chen X, Du J, Zhang H. Lipreading with DenseNet and resBi-LSTM. Signal Image Video Process 2020; 14: 981-989. DOI: 10.1007/s11760-019-01630-1.
[8] Feng D, Yang S, Shan S. An efficient software for building LIP reading models without pains. Int Conf on Multimedia and Expo Workshops (ICMEW) 2021: 1-2. DOI: 10.1109/ICMEW53276.2021.9456014.

[9] Martinez B, Ma P, Petridis S, Pantic M. Lipreading using temporal convolutional networks. Int Conf on Acoustics, Speech and Signal Processing (ICASSP) 2020: 6319-6323. DOI: 10.1109/ICASSP40776.2020.9053841.
[10] Zhang Y, Yang S, Xiao J, Shan S, Chen X. Can we read speech beyond the lips? Rethinking RoI selection for deep visual speech recognition. Int Conf on Automatic Face and Gesture Recognition (FG) 2020: 356-363. DOI: 10.1109/FG47880.2020.00134.
[11] Ma P, Martinez B, Petridis S, Pantic M. Towards practical lipreading with distilled and efficient models. Int Conf on Acoustics, Speech and Signal Processing (ICASSP) 2021: 7608-7612. DOI: 10.1109/ICASSP39728.2021.9415063.
[12] Sui C, Bennamoun M, Togneri R. Listening with your eyes: Towards a practical visual speech recognition system using deep Boltzmann machines. Proc Int Conf on Computer Vision (ICCV) 2015: 154-162.
[13] Stafylakis T, Tzimiropoulos G. Combining residual networks with LSTMs for lipreading. Interspeech 2017: 3652-3656.
[14] Hlaváč M, Gruber I, Železný M, Karpov A. Lipreading with LipsID. Int Conf on Speech and Computer (SPECOM) 2020: 176-183. DOI: 10.1007/978-3-030-60276-5_18.
[15] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. Proc Computer Society Conf on Computer Vision and Pattern Recognition (CVPR) 2001; 1: 511-518. DOI: 10.1109/CVPR.2001.990517.
[16] Cootes TF, Edwards GJ, Taylor CJ. Active appearance models. IEEE Trans Pattern Anal Mach Intell 2001; 23(6): 681-685. DOI: 10.1109/34.927467.
[17] Xu B, Wang J, Lu C, Guo Y. Watch to listen clearly: Visual speech enhancement driven multi-modality speech recognition. Proc IEEE/CVF Winter Conf on Applications of Computer Vision 2020: 1637-1646.
[18] Ryumina E, Ryumin D, Ivanko D, Karpov A. A novel method for protective face mask detection using convolutional neural networks and image histograms. Int Arch Photogramm Remote Sens Spatial Inf Sci 2021; XLIV-2/W1-2021: 177-182. DOI: 10.5194/isprs-archives-XLIV-2-W1-2021-177-2021.
[19] Ryumina E, Karpov A. Facial expression recognition using distance importance scores between facial landmarks. Graphicon, CEUR Workshop Proceedings 2020; 2744: 1-10.
[20] Ivanko D, Ryumin D, Axyonov A, Kashevnik A. Speaker-dependent visual command recognition in vehicle cabin: Methodology and evaluation. In Book: Karpov A, Potapova R, eds. Speech and Computer (SPECOM). Lecture Notes in Computer Science 2021; 12997: 291-302. DOI: 10.1007/978-3-030-87802-3_27.
[21] Noda K, Yamaguchi Y, Nakadai K, Okuno HG, Ogata T. Lipreading using convolutional neural network. Proc Annual Conf of the Int Speech Communication Association (INTERSPEECH) 2014: 1149-1153.
[22] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997; 9(8): 1735-1780.
[23] Petridis S, Stafylakis T, Ma P, Cai F, Tzimiropoulos G, Pantic M. End-to-end audiovisual speech recognition. 2018 IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP) 2018; 6548-6552.
[24] 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; 9: 34986-35003. DOI: 10.1109/ACCESS.2021.3062752.
[25] Lashkov I, Axyonov A, Ivanko D, Ryumin D, Karpov A, Kashevnik A. Multimodal Russian Driver Multimodal database of Russian speech of drivers in the cab of vehicles (RUSAVIC – RUSsian Audio-Visual speech in Cars) [In Russian]. Database State Registration Certificate N2020622063 of October 27, 2020.
[26] Ivanko D, Axyonov A, Ryumin D, Kashevnik A, Karpov A. RUSAVIC Corpus: Russian audio-visual speech in cars. Proc Thirteenth Language Resources and Evaluation Conference (LREC'22) 2022: 1555-1559.
[27] Kashevnik A, Lashkov I, Gurtov A. Methodology and mobile application for driver behavior analysis and accident prevention. IEEE Trans Intell Transp Syst 2019; 21(6): 2427-2436.
[28] Kashevnik A, Lashkov I, Ponomarev A, Teslya N, Gurtov A. Cloud-based driver monitoring system using a smartphone. Sensors 2020; 20(12): 6701-6715.
[29] The multi-speaker audiovisual corpus RUSAVIC. Source: ⟨https://mobiledrivesafely.com/corpus-rusavic⟩.
[30] Fung I, Mak B. End-to-End Low-resource lip-reading with Maxout CNN and LSTM. Int Conf on Acoustics, Speech and Signal Processing (ICASSP) 2018: 2511-2515. DOI: 10.1109/ICASSP.2018.8462280.
[31] Xu K, Li D, Cassimatis N, Wang X. LCANet: End-to-end lipreading with cascaded attention-CTC. Int Conf on Automatic Face and Gesture Recognition (FG) 2018: 548-555. DOI: 10.1109/FG.2018.00088.
[32] Ma P, Petridis S, Pantic M. End-to-end audio-visual speech recognition with conformers. Int Conf on Acoustics, Speech and Signal Processing (ICASSP) 2021: 7613-7617. DOI: 10.1109/ICASSP39728.2021.9414567.
[33] Lugaresi C, Tang J, Nash H, McClanahan C, Uboweja E, Hays M, Zhang F, Chang CL, Yong M, Lee J, Chang WT, Hua W, Georg M, Grundmann M. Mediapipe: A framework for building perception pipelines. arXiv Preprint. 2019. Source: ⟨https://arxiv.org/abs/1906.08172⟩.
[34] Shin H, Roth H, Gao M, Lu L, Xu Z, Nogues I, Summers RM. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016; 35(5): 1285-31298. DOI: 10.1109/TMI.2016.2528162.
[35] Torchvision. Transforms. Source: ⟨https://pytorch.org/vision/stable/transforms.html?highlight=randomequalize#torchvision.transforms.RandomEqualize⟩
[36] Label smoothing. Source: ⟨https://paperswithcode.com/method/label-smoothing⟩.
[37] 3D ResNet. Source: ⟨https://pytorch.org/hub/facebookresearch_pytorchvideo_resnet/⟩.
[38] Zhong Z, Lin ZQ, Bidart R, Hu X, Ben Daya I, Li Z, Zheng W, Li J, Wong A. Squeeze-and-attention networks for semantic segmentation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition 2020; 13065-13074.
Cosine annealing warm restarts. Source: ⟨https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.CosineAnnealingWarmRestarts.html⟩.

Сведения об авторах

Аксёнов Александр Александрович , 1995 года рождения, в 2019 году окончил Университет ИТМО по специальности 09.04.04 «Программная инженерия», работает младшим научным сотрудником в лаборатории речевых и многомодальных интерфейсов СПб ФИЦ РАН. Область научных интересов: искусственный интеллект, машинное обучение, цифровая обработка изображений, нейронные сети, автоматическое распознавание визуальной речи. E-mail: *axyonov.a@iias.spb.su*.

.

Рюмин Дмитрий Александрович , 1991 года рождения, в 2016 году окончил факультет информационных технологий и программирования, а в 2020 году защитил кандидатскую диссертацию на тему «Модели и методы автоматического распознавания элементов русского жестового языка для человеко-машинного взаимодействия» в Университете ИТМО. Старший научный сотрудник лаборатории речевых и многомодальных интерфейсов СПб ФИЦ РАН. Область научных интересов: цифровая обработка изображений, распознавание образов, автоматическое распознавание визуальной речи, многомодальные интерфейсы, машинное обучение, нейронные сети, биометрия, человеко-машинные интерфейсы. E-mail: *ryumin.d@iias.spb.su* .

Кашевник Алексей Михайлович 1982 года рождения, кандидат технических наук (2008), доцент по специальности 05.13.11 (2020). Работает старшим научным сотрудником лаборатории интегрированных систем автоматизации СПб ФИЦ РАН с 2003 года и доцентом факультета информационных технологий и программирования Университета ИТМО с 2014 года. Область научных интересов: интеллектуальные транспортные системы, взаимодействие человека и компьютера, помощь водителю, системы мониторинга водителя, обнаружение отвлекающих факторов, обнаружение сонливости, искусственный интеллект, нейронные сети, компьютерное зрение и другие. Опубликовал более 200 научных работ в рецензируемых международных журналах, трудах международных конференций и книгах. E-mail: *alexey.kashevnik@iias.spb.su*

Иванько Денис Викторович, 1993 года рождения, в 2015 году окончил факультет информационных технологий и программирования, а в 2020 году защитил кандидатскую диссертацию на тему «Автоматическое распознавание аудиовизуальной русской речи» в Университете ИТМО. Старший научный сотрудник лаборатории речевых и многомодальных интерфейсов СПб ФИЦ РАН. Область научных интересов: автоматическое распознавание речи, обработка аудиовизуальной речи, чтение речи по губам диктора, цифровая обработка изображений, распознавание образов, машинное обучение, нейронные сети. E-mail: *ivanko.d@iias.spb.su* .

Карпов Алексей Анатольевич, 1978 года рождения, доктор технических наук (2013), доцент по специальности 05.13.11 (2012). Работает главным научным сотрудником (руководителем лаборатории) речевых и многомодальных интерфейсов СПб ФИЦ РАН. Область научных интересов: речевые технологии, автоматическое распознавание речи, обработка аудиовизуальной речи, многомодальные человеко-машинные интерфейсы, компьютерная паралингвистика и другие. E-mail: *karpov@iias.spb.su* .

ГРНТИ: 28.23.15
Поступила в редакцию 25 декабря 2021 г. Окончательный вариант – 30 апреля 2022 г.

Method for visual analysis of driver's face for automatic lip-reading in the wild

A.A. Axyonov [1], D.A. Ryumin [1], A.M. Kashevnik [1], D.V. Ivanko [1], A.A. Karpov [1]
[1] St. Petersburg Federal Research Center of the RAS (SPC RAS),
199178, St. Petersburg, Russia, 14th Line V.O. 39

Abstract

The paper proposes a method of visual analysis for automatic speech recognition of the vehicle driver. Speech recognition in acoustically noisy conditions is one of big challenges of artificial intelligence. The problem of effective automatic lip-reading in vehicle environment has not yet been resolved due to the presence of various kinds of interference (frequent turns of driver's head, vibration, varying lighting conditions, etc.). In addition, the problem is aggravated by the lack of available databases on this topic. A MediaPipe Face Mesh is used to find and extract the region-of-interest (ROI). We have developed End-to-End neural network architecture for the analysis of visual speech. Visual features are extracted from a single image using a convolutional neural network (CNN) in conjunction with a fully connected layer. The extracted features are input to a Long Short-Term Memory (LSTM) neural network. Due to a small amount of training data we proposed that a Transfer Learning method should be applied. Experiments on visual analysis and speech recognition present great opportunities for solving the problem of automatic lip-reading. The experiments were performed on an in-house multi-speaker audio-visual dataset RUSAVIC. The maximum recognition accuracy of 62 commands is 64.09 %. The results can be used in various automatic speech recognition systems, especially in acoustically noisy conditions (high speed, open windows or a sunroof in a vehicle, backgoround music, poor noise insulation, etc.) on the road.

Keywords: vehicle, driver, visual speech recognition, automated lip-reading, machine learning, End-to-End, CNN, LSTM.

Citation: Axyonov AA, Ryumin DA, Kashevnik AM, Ivanko DV, Karpov AA. Method for visual analysis of driver's face for automatic lip-reading in the wild. Computer Optics 2022; 46(6): 955-962. DOI: 10.18287/2412-6179-CO-1092.

Acknowledgements: This work was partly funded by the Russian Foundation for Basic Research under grant No. 19-29-09081 and the state research project No. 0073-2019-0005.

Authors' information

Alexandr Alexandrovich Axyonov (b. 1995), graduated from ITMO University in 2019 with a degree in 09.04.04 "Software Engineering", works as a junior researcher in the Laboratory of Speech and Multimodal Interfaces at the SPC RAS. Research interests: artificial intelligence, machine learning, digital image processing, neural networks, automatic visual speech recognition. E-mail: *axyonov.a@iias.spb.su*.

Dmitry Alexandrovich Ryumin (b. 1991) graduated from Information Technologies and Programming faculty in 2016. He defended Ph.D. thesis on Models and Methods for Automatic Recognition of Russian Sign Language Elements for Human-Machine Interaction in ITMO University in 2020. He is a senior researcher in Speech and Multimodal Interfaces Laboratory at the SPC RAS. Research interests are digital image processing, pattern recognition, automatic visual speech recognition, multimodal interfaces, machine learning, neural networks, biometrics, human-machine interfaces. E-mail: *ryumin.d@iias.spb.su*.

Alexey Mikhailovich Kashevnik (b. 1982) has Ph.D. degree in Computer Science (2008). He is associate professor (2020) and currently works as the senior researcher in Computer Aided Integrated Systems Laboratory at the SPC RAS as well as associate professor at Information Technologies and Programming faculty at ITMO University. Research interests are intelligent transport systems, human-computer interaction, driver decision support, driver monitoring systems, driver distraction, driver drowsiness, artificial intelligence, neural networks, computer vision and etc. He has published more than 200 research papers in reviewed international journals, proceedings of international conferences, and books. E-mail: *alexey.kashevnik@iias.spb.su*.

Denis Viktorovich Ivanko (b. 1993) graduated from Information Technologies and Programming faculty in 2015. He defended Ph.D. thesis on Audio-Visual Russian Speech Recognition in ITMO University in 2020. He is a senior researcher at the SPC RAS. Research interests are audio-visual speech recognition, automated lip-reading, digital image

processing, pattern recognition, automatic visual speech recognition, machine learning, neural networks. E-mail: *ivanko.d@iias.spb.su* .

Alexey Anatolievich Karpov (b. 1978) is Doctor of Technical Sciences (2013), Associate Professor (2012). Currently he works as the chief researcher and head of the Speech and Multimodal Interfaces Laboratory at the SPC RAS. Research interests are speech technology, automatic speech recognition, audio-visual speech processing, multimodal human-computer interfaces, and computational paralinguistics. E-mail: *karpov@iias.spb.su*.

Received December 25, 2021. The final version – April 30, 2022.