НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2024 Том 24 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2024 Vol. 24 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

ОБЗОРНЫЕ СТАТЬИ

REVIEW PAPERS

doi: 10.17586/2226-1494-2024-24-5-669-686
УДК 004.5, 004.93

Автоматический сурдоперевод: обзор нейросетевых методов распознавания и синтеза звучащей и жестовой речи

Денис Викторович Иванько[1]✉, Дмитрий Александрович Рюмин[2]

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

[1] ivanko.d@iias.spb.su✉, https://orcid.org/0000-0003-0412-7765
[2] ryumin.d@iias.spb.su, https://orcid.org/0000-0002-7935-0569

Аннотация

Введение. Представлен обзор современных методов и технологий автоматического машинного сурдоперевода, включающих распознавание и синтез как звучащей, так и жестовой речи. Рассмотренные методы предназначены для обеспечения эффективной коммуникации между глухими, слабослышащими и слышащими людьми. Предложенные решения могут найти применение в современных интерфейсах человеко-машинного взаимодействия. **Методы.** Рассмотрены ключевые аспекты новых технологий, включая методы распознавания и синтеза жестовой речи и аудиовизуальной речи, существующие наборы данных для обучения нейросетевых моделей, а также современные системы автоматического машинного сурдоперевода. Представлены актуальные нейросетевые подходы, включающие использование методов глубокого обучения, таких как сверточные и рекуррентные нейросети, а также трансформеры. Приведен анализ существующих наборов данных для обучения систем распознавания и синтеза речи, проблем и ограничений существующих систем машинного сурдоперевода. **Основные результаты.** Выявлены основные недостатки и конкретные проблемы текущих технологий автоматического машинного сурдоперевода. Определены перспективные пути их решения. Особое внимание уделено возможности применения автоматических систем машинного сурдоперевода в реальных условиях. **Обсуждение.** Показана необходимость дальнейших исследований в области сбора и разметки данных. Доказана целесообразность разработки новых методов и нейросетевых моделей, а также создания инновационных технологий для обработки аудио- и видеоданных с целью улучшения качества и эффективности существующих систем автоматического машинного сурдоперевода.

Ключевые слова

автоматическое распознавание речи, синтез речи, распознавание жестов, синтез жестов, автоматический сурдоперевод, машинное обучение

Благодарности

Раздел «Предмет исследования» выполнен при поддержке бюджетной темы (№ FFZF-2022-0005), остальные исследования выполнены при финансовой поддержке Российского научного фонда (проект № 23-71-01056).

Ссылка для цитирования: Иванько Д.В., Рюмин Д.А. Автоматический сурдоперевод: обзор нейросетевых методов распознавания и синтеза звучащей и жестовой речи // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 5. С. 669–686. doi: 10.17586/2226-1494-2024-24-5-669-686

Automatic sign language translation: a review of neural network methods for recognition and synthesis of spoken and signed language

Denis V. Ivanko[1]✉, Dmitry A. Ryumin[2]

[1,2] St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg (SPC RAS), 199178, Russian Federation

[1] ivanko.d@iias.spb.su✉, https://orcid.org/0000-0003-0412-7765
[2] ryumin.d@iias.spb.su, https://orcid.org/0000-0002-7935-0569

Abstract
A review of modern methods and technologies for automatic machine translation for the deaf and hard of hearing is presented, including recognition and synthesis of both spoken and sign languages. These methods aim to facilitate effective communication between deaf/hard-of-hearing and hearing individuals. The proposed solutions have potential applications in contemporary human-machine interaction interfaces. Key aspects of new technologies are examined, including methods for sign language recognition and synthesis, audiovisual speech recognition and synthesis, existing corpora for training neural network models, and current systems for automatic machine translation. Current neural network approaches are presented, including the use of deep learning methods such as convolutional and recurrent neural networks as well as transformers. An analysis of existing corpora for training recognition and synthesis systems is provided, along with an evaluation of the challenges and limitations of existing machine translation systems. The main shortcomings and specific problems of current automatic machine translation technologies are identified, and promising solutions are proposed. Special attention is given to the applicability of automatic machine translation systems in real-world scenarios. The need for further research in data collection and annotation, development of new methods and neural network models, and creation of innovative technologies for processing audio and video data to enhance the quality and efficiency of the existing automatic machine translation systems is highlighted.

Keywords
automatic speech recognition, speech synthesis, gesture recognition, gesture synthesis, automatic sign language translation, machine learning

Acknowledgements
The section “Research Subject” was carried out with the support of the budget topic (No. FFZF-2022-0005), the remaining studies were carried out with the financial support of the Russian Science Foundation (project No. 23-71-01056).

For citation: Ivanko D.V., Ryumin D.A. Automatic sign language translation: a review of neural network methods for recognition and synthesis of spoken and signed language. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 5, pp. 669–686 (in Russian). doi: 10.17586/2226-1494-2024-24-5-669-686

Введение

Автоматический машинный перевод речи с одного языка на другой играет ключевую роль в устранении языковых барьеров и обеспечении коммуникации между людьми, говорящими на разных языках. Это имеет огромное значение в международных деловых отношениях, туризме, образовании, научных исследованиях, и многих других сферах. За последние годы нейросети стали популярным инструментом для автоматического машинного перевода звучащей речи благодаря своей способности обучаться на больших объемах данных и выявлять сложные шаблоны и зависимости между языками [1].

В то же время автоматический машинный сурдоперевод — технология, которая способна значительно облегчить жизнь людям с нарушением слуха, обеспечивая им доступ к устной коммуникации. Однако, в отличие от звучащей речи, основным каналом передачи жестовой речи является визуальный. Потому для реализации надежной системы автоматического машинного сурдоперевода требуется многомодальный подход, учитывающий визуальную и акустическую составляющие. Несмотря на большой научный прогресс, достигнутый в области автоматического машинного перевода звучащей речи, а также в области компьютерного зрения, надежная технология автоматического машинного сурдоперевода до сих пор не реализована по нескольким причинам:

— сложность восприятия жестов: существует большое разнообразие жестов, которые используются в различных языках жестовой речи, и их интерпретация может быть сложной для автоматических систем;
— контекстуальная зависимость: понимание различных жестикуляций требует учета контекста и ситуации, что может быть сложно для методов обработки жестов на основе машинного обучения, особенно в разнообразных сценариях коммуникации;
— технические ограничения: точное распознавание и интерпретация жестов требует высокоточных сенсорных устройств и сложных методов машинного обучения для обработки данных, что может быть технически сложно и затратно;
— жестовые наборы данных (корпуса): на сегодняшний день не существует полностью репрезентативных корпусов жестовой речи пригодных для обучения современных моделей на основе глубоких нейросетей.

В целом автоматический машинный сурдоперевод представляет собой перспективную технологию, но требует дальнейших исследований и разработок, чтобы достичь полной реализации и обеспечить эффективное использование в повседневной жизни.

Учитывая, что область автоматического сурдоперевода находится на пересечении нескольких областей знаний, таких как распознавание речи, компьютерное зрение, машинное обучение и т. д., в настоящей работе рассмотрены достижения современной науки по основным аспектам, комплексирование которых приведет к созданию надежной системы автоматического машинного сурдоперевода. Например, приведен анализ методов распознавания и синтеза жестовой речи, методов распознавания и синтеза аудиовизуальной речи, анализ существующих жестовых корпусов для обучения нейросетевых моделей и анализ существующих систем автоматического машинного сурдоперевода.

Предмет исследования

Существующие методы автоматического распознавания и синтеза звучащей и жестовой речи уже применяются в некоторых практических приложениях.

Однако их качество и надежность при работе в реальных условиях остаются недостаточно высокими, что представляет серьезную научно-техническую проблему. В то же время систем, позволяющих осуществлять автоматический двухсторонний машинный сурдоперевод на основе распознавания и синтеза аудиовизуальной и жестовой речи практически не существует [2]. Для решения этих проблем необходимо продолжить разработку новых методов улучшения качества распознавания, включая увеличение объемов и качества тренировочных данных, а также использование более совершенных моделей машинного обучения и сенсорных технологий.

В настоящее время отсутствует единый подход к разработке интеллектуальных систем автоматического двустороннего машинного сурдоперевода как в России, так и за рубежом. Существует множество нерешенных, но важных проблем и задач, которые требуют внимания ученых со всего мира. Среди ключевых проблем в этой области можно выделить следующие: сбор, анализ и аннотирование представительных аудиовизуальных речевых и жестовых корпусов, записанных в естественных условиях. Отсутствие общепринятых нейросетевых архитектур и моделей распознавания и синтеза жестовой и звучащей речи. В целом эти проблемы ставят перед исследователями серьезные вызовы, и их решение требует совместных усилий и междисциплинарного подхода со стороны ученых и специалистов по обработке цифровых сигналов, машинному обучению и компьютерному зрению.

Автоматический машинный перевод с жестовых языков сопряжен со значительными трудностями по сравнению с обработкой звучащих языков. Это обусловлено рядом факторов, связанных с задачами компьютерного зрения и машинного обучения:

— окклюзии: жестовая речь часто сопровождается движениями различных частей тела, таких как руки, голова и туловище. Это может привести к перекрытию (окклюзиям) между различными частями тела, что усложняет точное распознавание и интерпретацию жестов и жестикуляций;
— различия в фоновом освещении: изменения в освещении могут привести к изменению теней и контраста, что делает сложным обнаружение и выделение жестов в различных условиях освещения;
— необходимость больших вычислительных ресурсов: обработка и анализ видеоданных, особенно при высоком разрешении, требует значительных вычислительных ресурсов для обеспечения высокой точности и скорости обработки;
— недостаточный объем корпусов: для эффективного обучения моделей машинного обучения необходимы большие и разнообразные корпуса, содержащие различные жесты в разных условиях. Недостаточный объем таких корпусов может привести к малой обученности нейросетевых моделей;
— нелинейная структура высказывания: в отличие от звучащих языков, жестовые языки обладают визуальной природой и обычно имеют нелинейную структуру высказывания (несколько жестов могут выполняться одновременно, имея различные пространственные координаты и контексты);
— динамическая природа жестов: жесты могут изменяться в зависимости от скорости и плавности движений. Некоторые жесты могут быть выполнены быстрее или медленнее в зависимости от ситуации, что требует от автоматической системы гибкости в интерпретации;
— разнообразие жестовых языков и диалектов: существует множество жестовых языков и диалектов, каждый из которых имеет свои уникальные жесты и правила, что усложняет создание универсальной модели перевода;
— пользовательский интерфейс и эргономика: разработка интуитивно понятных и удобных интерфейсов для пользователей с нарушением слуха, которые способны обеспечивать эффективное взаимодействие с автоматической системой перевода;
— лингвистические и культурные различия: жестовые языки часто содержат культурные и контекстуальные особенности, которые могут быть сложны для автоматического машинного перевода без учета соответствующих культурных контекстов;
— точность и достоверность перевода: обеспечение высокой точности машинного перевода критически важно, так как ошибки могут приводить к серьезным ошибкам. Например, в медицинских контекстах неправильный перевод может привести к неверной интерпретации симптомов или предписаний, что может серьезно повлиять на здоровье пациента.

Многочисленные междисциплинарные исследования подчеркивают важность визуальной информации в понимании звучащей речи [3]. Например, наблюдение за лицом собеседника значительно облегчает восприятие речи. Сигналы из визуальных и слуховых каналов взаимодополняются, помогают правильно воспринимать речь в сложных условиях, таких как динамические акустические шумы. Особенно это важно для людей со слабым слухом, которые часто опираются на визуальные данные, такие как движения губ и мимика лица. В результате во многих странах мира проводятся исследования и разработки автоматических систем аудиовизуального распознавания речи для основных мировых языков [4–6].

Помимо вышеупомянутых факторов, решение указанной научной проблемы осложняется ограниченностью русскоязычных корпусов по размеру и доступным данным. В отличие от аналогичных корпусов на других языках, таких как английский, немецкий или китайский, русскоязычных корпусов жестовой и аудиовизуальной речи значительно меньше. Это подчеркивает необходимость использования широкого спектра дополнительных подходов, методов и алгоритмов для решения поставленных задач в условиях ограниченного объема доступных данных. В частности, необходимо исследование новых подходов к аугментации данных и адаптации иноязычных ресурсов для русскоязычного контекста.

В последние десятилетия значительное внимание уделяется разработке и совершенствованию технологий обработки речи, включая распознавание аудиовизуальной и жестовой речи, а также синтезу акустической и жестовой речи. Именно поэтому в настоящей работе

рассматриваются основные методы и достижения в области распознавания и синтеза аудиовизуальной и жестовой речи.

Распознавание аудиовизуальной речи — процесс автоматического определения речевых звуков по видеозаписям говорящего. Он объединяет информацию из акустического и визуального потоков данных для более точного распознавания речи.

Синтез акустической речи — процесс генерации звучащей речи из текстовых данных. Современные методы синтеза акустической речи, используют глубокие нейросети для моделирования акустических признаков речи и создания естественно звучащей речи с высоким качеством, интонацией и эмоциями.

Распознавание жестовой речи — процесс автоматического определения жестов и движений рук и тела, используемых для передачи визуальной информации. Методы распознавания жестовой речи включают в себя как классические подходы [7], основанные на анализе признаков и классификации жестов, так и современные методы, такие как глубокие нейросети, использующиеся для извлечения и интерпретации динамических жестов.

Синтез жестовой речи — процесс генерации анимированных движений рук, тела и мимики лица, соответствующих передаваемой информации. Методы синтеза жестовой речи включают в себя использование моделей движения и анимации, основанных на данных о реальных жестах и выражениях лица, а также генеративных нейросетей, создающих реалистичные и естественные движения аватаров.

Отметим конкретные проблемы и недостатки текущих технологий автоматического машинного сурдоперевода:

— точность распознавания: одной из основных проблем современных систем автоматического машинного сурдоперевода является недостаточная точность распознавания жестов. Это вызвано множеством факторов, таких как разнообразие жестов, различные диалекты жестовой речи и ограниченные объемы обучающих данных;
— синхронизация жестов и аудиовизуальной речи: еще одной важной проблемой является синхронизация. Текущие технологии часто не обеспечивают необходимую синхронизацию распознавания звучащей и жестовой речи, что приводит к искажению смысла передаваемой информации;
— интеграция различных моделей в единую систему сурдоперевода: современные системы автоматического машинного сурдоперевода часто сталкиваются с трудностями интеграции различных компонентов, таких как распознавание жестов, синтез речи и обработка аудиовизуальных данных.

Распознавание аудиовизуальной и жестовой речи, а также синтез акустической и жестовой речи, представляют собой важные направления исследований в области обработки речи. Современные методы и технологии в этой области обладают потенциалом для создания более эффективных и интуитивно понятных систем коммуникации, обучения и развлечений, содействуя развитию более доступной среды человеко-машинного взаимодействия для всех пользователей.

Автоматическое распознавание жестовой речи

В последнее десятилетие ученые по всему миру активно проводят научно-технические исследования, особенно в областях компьютерного зрения, машинного обучения и обработки сигналов, и разрабатывают новые технологии автоматического распознавания жестовой речи для глухих людей. Основные методы распознавания жестовой речи приведены в табл. 1.

В работе [8] представлен метод оценки положения и классификации формы руки с применением многоуровневого метода композиции рандомизированных лесов решений. В [9] предложен метод отслеживания положения руки в реальном времени с использованием данных от датчиков глубины и 3D-модели руки, состоящей из 21 сегмента, а также применен метод леса случайных решений для классификации пикселей и совместной оценки местоположения. В [10] представлена генеративная нейросетевая модель и метод, основанный на данных о глубине, для отслеживания движений руки с использованием функции расстояния для моделирования ее геометрии и быстрой оптимизации при высокой частоте кадров. Предложенный подход позволял отслеживать взаимодействие между двумя руками или другими объектами.

Работа [11] посвящена разработке системы компьютерного зрения в реальном времени, предназначенной для помощи пациентам с нарушениями слуха в больничных условиях. Система задает пациентам ряд вопросов для определения цели их визита, принимая ответы через язык жестов. В работе предложено использовать временные накопительные признаки для распознавания изолированных жестов. Этот метод включает элементы, специфичные для жестового языка, для захвата его лингвистических характеристик, что позволило создать эффективную и быструю систему распознавания. В [12] представлен метод перевода жестового языка в письменный текст с использованием глубоких нейросетей. Сделана попытка улучшить системы перевода, включив в процесс токенизацию для более точного отображения лингвистической структуры жестового языка.

В [13] описан метод перевода жестовой речи в письменный текст с использованием глубоких нейросетей. Выполнен анализ лингвистической структуры жестовых языков с помощью нейросетевого метода, с целью оценить потенциал нейросетей для улучшения систем перевода. В работе [14] исследовано обучение для систем распознавания жестовой речи с использованием многопоточной модели CNN-LSTM-HMM с целью выявления последовательного параллелизма в видео. Осуществлено обучение модели на слабо размеченных данных и продемонстрирован потенциал машинного обучения для улучшения систем перевода.

В [15] изучена проблема мультиартикуляции и предложена многоканальная архитектура трансформера. Эта архитектура позволяет моделировать меж- и внутриконтекстные отношения между различными каналами, сохраняя при этом информацию, относящуюся к конкретному каналу. Представленная в работе архитектура объединила задачи распознавания и перевода в единую

Таблица 1. Основные методы распознавания жестовой речи
Table 1. Sign Language Recognition Methods

Ссылка	Метод	Описание
[8]	Многоуровневый метод рандомизированных лесов решений	Метод оценки положения и классификации формы руки с использованием композиции рандомизированных лесов решений
[9]	Датчики глубины, трехмерная (3D) модель руки	Метод отслеживания положения руки в реальном времени с использованием данных от датчиков глубины и 3D-модели руки на основе леса решений
[10]	Данные о глубине, функция расстояния, быстрая оптимизация, геометрия руки	Генеративная нейросетевая модель и метод отслеживания движений руки с использованием данных о глубине и функции расстояния
[11]	Временные накопительные признаки	Использование временных накопительных признаков для распознавания изолированных жестов рук
[12]	Глубокие нейросети, токенизация	Метод перевода жестового языка в письменный текст с использованием глубоких нейросетей
[13]	Нейросетевой метод анализа лингвистической структуры, перевод	
[14]	Многопоточная нейросетевая модель Convolutional Neural Network — Long Short-Term Memory — Hidden Markov Model (CNN-LSTM-HMM)	Обучение систем распознавания жестовой речи с использованием многопоточной модели CNN-LSTM-HMM
[15]	Трансформер, объединение задач распознавания и перевода	Нейросетевая архитектура на основе трансформера для объединения задач распознавания и перевода в единую модель
[16]	3DCNN, LSTM	Метод комбинирования 3DCNN и LSTM для мультимодального распознавания жестов
[17]	3DCNN	Метод улучшения динамического распознавания жестов рук с использованием 3DCNN
[18]	Глубокие CNN, RGB-D	Метод обучения многомерным функциям для распознавания жестов RGB-D с использованием глубоких CNN
[19]	Скоростные и мультимодальные нейросети	Метод распознавания жестов с использованием скоростных и мультимодальных нейросетей с временным расширением
[20]	Ансамбль CNN	Метод распознавания жестов с использованием ансамбля глубоких CNN
[21]	Ансамбль CNN, Oriented FAST and Rotated BRIEF (ORB) дескриптор, фильтр Габора	Метод извлечения жестовых признаков на основе CNN, ORB-дескриптора и фильтров Габора

модель. Это значительно повысило производительность по сравнению с обычными методами, где распознавание и перевод выполняются как отдельные процессы. Также в работе [15] рассмотрена совместная задача по переводу, в которой использована модель чистого текста. Цель исследования заключалась в улучшении перевода жестового языка в письменный текст с помощью данной методологии.

В [16] представлен метод, сочетающий 3DCNN и сверточный LSTM для мультимодального распознавания жестов, демонстрирующий эффективность такой комбинации. В работе [17] предложен метод, который улучшает динамическое распознавание жестов рук с использованием 3DCNN путем внедрения знаний из нескольких модальностей в отдельные сети. В [18] использован MultiD-CNN, метод обучения многомерным функциям для распознавания жестов RGB-D с использованием глубоких CNN. В [19] рассмотрен метод распознавания жестов с использованием многоскоростных и мультимодальных нейросетей с временным расширением, в которых применен алгоритм поиска для определения оптимальной комбинации архитектуры нейросети, временной информации о разрешении и модальности.

В работе [20] исследованы ансамблевые методы для изолированных жестов, а также метод с использованием ансамбля нескольких глубоких CNN. В свою очередь, в [21] предложено выполнять объединение сверточных нейросетей с ORB-дескриптором и фильтром Габора для более эффективного распознавания языка жестов по видео.

Наряду с вышеописанными методами, в [22] представлен всесторонний обзор методов распознавания и синтеза жестов рук, включая методы на основе компьютерного зрения, машинного обучения и носимых устройств. Следует также отметить ученых из Университета Карнеги-Меллона (США), которые одними из первых разработали решение с открытым исходным кодом для определения множества ориентиров скелета и лица (модель человеческого скелета) на отдельных изображениях в режиме реального времени. Подробное описание библиотеки с открытым исходным кодом OpenPose представлено в [23]. В то же время Google активно разрабатывает кроссплатформенную

среду с открытым исходным кодом MediaPipe [24], которая включает новые методы, основанные на глубоком обучении, для определения трехмерных ориентиров лица, рук и тела человека.

Автоматический синтез жестовой речи

Первоначально автоматический синтез жестовой речи в виде 3D-аватаров был предложен в качестве инструмента для облегчения общения и взаимодействия людей с ограниченными возможностями — глухими или слабослышащими людьми [25]. Для таких людей жестовая речь является естественным средством общения, и автоматический синтез жестовых аватаров позволяет им взаимодействовать с другими людьми и компьютерными системами наравне со слышащими людьми.

В современном мире 3D-аватары могут использоваться в различных сферах деятельности. Например, в развлекательной индустрии аватары применяются для создания интерактивных виртуальных персонажей в видеоиграх, фильмах или виртуальной реальности. Это позволяет пользователям взаимодействовать с виртуальными мирами и персонажами более естественным образом, что улучшает их игровой опыт и вовлеченность.

Кроме того, методы автоматического синтеза жестовой речи в виде 3D-аватаров могут быть полезны в медицинских приложениях. Например, в реабилитации пациентов с нарушениями речи или движения такие методы помогут восстановлению навыков общения и моторики. Также данные методы могут быть применены для обучения медицинского персонала в области коммуникации с людьми с ограниченными возможностями.

Одним из ключевых преимуществ автоматического синтеза жестовой речи в виде 3D-аватаров является его возможность улучшить виртуальное общение и телекоммуникации. В наше время, когда все больше людей переходят на удаленную работу и обучение, эта технология может стать важным средством для создания более естественного и привлекательного виртуального взаимодействия.

Нейросетевые методы синтеза жестовых аватаров используют глубокие нейросети для создания анимированных персонажей или аватаров, которые могут эмитировать жестовую речь и другие движения. Основные нейросетевые методы синтеза жестовой речи и 3D-аватаров, а также их характеристики приведены в табл. 2.

YOLO [26]. Методы глубокого машинного обучения версии YOLO, применяются для распознавания и моделирования последовательностей жестов рук в реальном времени на изображениях или видео. Основное преимущество YOLO заключается в его способности к выполнению локализации и классификации жестов за один проход, что обеспечивает высокую скорость работы и эффективность. В отличие от других методов, которые разделяют процессы локализации и классификации на несколько этапов, YOLO анализирует изображение или видеопоток в целом, одновременно определяя расположение рук и класс жеста.

cGANs [27]. Методы используют генеративные состязательные нейросети для синтеза жестовых аватаров на основе входных данных разной модальности. Они позволяют генерировать реалистичные изображения жестовых аватаров, учитывая входные параметры.

VAEs [28]. Методы применяются для сжатия и генерации жестовых аватаров на основе обучающих данных. Эти нейросетевые модели обучаются на больших жестовых наборах данных и могут генерировать 3D-аватары с учетом обученной нейросетевой модели. Они позволяют создавать реалистичные 3D-аватары, сохраняя при этом их выразительность и уникальные характеристики, что полезно в различных приложениях, таких как обучение жестовой речи, виртуальная реальность и анимация персонажей.

Таблица 2. Основные нейросетевые методы синтеза жестовой речи и 3D-аватаров
Table 2. Neural Sign Language Synthesis and 3D Avatar Methods

Ссылка	Метод	Описание
[26]	You Only Look Once (YOLO)	Распознавание жестов рук в реальном времени на основе глубокого обучения с использованием версий YOLO обеспечивает высокую точность и скорость классификации, что подходит для интерактивных приложений
[27]	conditional Generative Adversarial Networks (cGANs)	Методы генеративных состязательных нейросетей для синтеза жестовых аватаров
[28]	Variational Autoencoders (VAEs)	Методы на основе вариационных автокодировщиков для генерации жестовых аватаров
[29]	LSTMs	Рекуррентные нейросети с длинной кратковременной памятью для анализа последовательностей жестов
[30]	Convolutional Neural Networks (CNNs)	Методы на основе сверточных нейросетей для эффективного извлечения пространственных признаков из изображений жестов
[31]	Graph Convolutional Networks (GCNs)	Графовые нейросети для анализа структуры и взаимосвязи между жестами в пространстве
[32]	Механизмы внимания	Механизмы внимания позволяют модели сосредоточиться на определенных частях жеста
[33]	Трансформеры	Модели на основе трансформера могут эффективно моделировать долгосрочные зависимости

LSTMs [29]. Методы реализуются для моделирования долгосрочных зависимостей в жестовых последовательностях, позволяя эффективно улавливать контекст и выражения различных движений. Благодаря этому LSTMs способны генерировать аватары, которые не только соответствуют входным жестам, но и сохраняют связанные с ними эмоциональные и динамические особенности.

CNNs [30]. Сверточные нейросети эффективно извлекают пространственные признаки из изображений жестов, что делает их полезными для синтеза жестовых аватаров. Они способны адаптироваться к различным аспектам жестов, включая формы рук и их позы, а также выражения лица и жесты других частей тела. Эти особенности делают их подходящими для создания аватаров, которые не только точно отражают входные жесты, но и захватывают их выразительность и эмоциональную окраску.

GCNs [31]. Графовые нейросети способны анализировать сложные взаимосвязи между жестами в 3D-пространстве. Они оперируют на основе графовых структур, где узлы представляют собой жесты, а ребра — их взаимодействия и связи. Это позволяет методам GCNs эффективно учитывать контекст и зависимости между жестами, что особенно важно для синтеза жестовых аватаров с реалистичным поведением и выражением. В процессе машинного обучения GCNs учитывают геометрическую структуру пространства жестов, а также динамику изменения этих жестов с течением времени. Благодаря этому они способны обнаруживать сложные шаблоны и взаимосвязи, которые могут быть упущены другими методами, и использовать эту информацию для создания более реалистичных и выразительных жестовых аватаров.

Механизмы внимание [32]. В контексте синтеза жестовых аватаров данные методы играют важную роль, позволяя нейросетевым моделям сосредоточиться на наиболее значимых или информативных частях жеста. Этот принцип достигается за счет выделения ключевых аспектов жеста и уделения им особого внимания в процессе генерации аватара. Механизмы внимания могут включать в себя идентификацию ключевых точек в пространстве, определение эмоционального содержания жеста или выделение динамических аспектов движения. Путем акцентирования на этих важных деталях модели способны создавать более реалистичные и выразительные жестовые аватары, которые точнее передают эмоциональное и содержательное содержание жестов.

Трансформеры [33]. Нейросетевые модели на основе трансформеров способны эффективно моделировать долгосрочные зависимости в последовательностях жестов, что полезно для синтеза жестовых аватаров, учитывая контекст и динамику движения.

Таким образом, вышеперечисленные исследования направлены на решение задач эффективного комплексного интеллектуального анализа движений тела человека для автоматического синтеза жестового языка и аватаров. Отметим, что полностью абстрагироваться от цифровой сцены (видеоинформации) и анализировать только динамически меняющееся состояние (поведение) человека (в том числе и жесты) пока достаточно сложно. В настоящее время не существует полностью автоматических нейросетевых моделей и методов для машинных систем распознавания жестовой речи и 3D-аватаров для синтеза элементов жестового языка.

Автоматическое распознавание речи по аудиовизуальным данным

Традиционно системы аудиовизуального распознавания речи состоят из двух этапов обработки: извлечение признаков из аудио- и визуальной информаций с последующим распознаванием речи [34, 35]. При традиционных методах информативные признаки обычно извлекаются из интересующей области рта и из аудиосигнала, а затем объединяются [36, 37].

В последние годы, с развитием технологий глубокого машинного обучения и компьютерного зрения, было представлено множество нейросетевых методов, которые заменили этап извлечения признаков. Первый нейросетевой классификатор изображений сверточной нейросети для распознавания визем обучен в работе [38]. В [39] нейросетевые признаки использованы для распознавания слов, чтобы в полной мере применить глубокие сверточные слои. В работе [40] предложено использовать трехмерные сверточные фильтры для обработки пространственно-временной информации о губах, а в [41] применен механизм внимания к интересующей области рта.

В работах [42–44] рассмотрены интегральные (сквозное тестирование) нейросетевые архитектуры для систем автоматического распознавания речи, которые привлекли большое внимание исследователей по распознаванию аудиовизуальной речи. Основным преимуществом современного интегрального подхода является возможность как выделения признаков, так и этапов классификации в границах одной нейросети. Эти методы можно разделить на две группы. В первой группе одни и те же слои используются для извлечения признаков и моделирования временной динамики. Во второй — сверточные слои применены для извлечения признаков, за которыми следуют LSTM или Gated Recurrent Unit (GRU) для моделирования результатов распознавания.

В последнее время интегральные методы успешно используются для многих задач распознавания, синтеза речи и задач компьютерного зрения. Можно отметить работы [45, 46], в которых механизм внимания применялся как к интересующим областям рта, так и к мел-частотным кепстральным коэффициентам, а модель обучалась интегрально. Затем полносвязные слои, за которыми следует LSTM, используются для извлечения признаков из изображений и спектрограмм и выполнения классификации.

Первая интегральная модель, которая выполняла аудиовизуальное распознавание слов на большом наборе данных, описана в работе [47], где предложена двухпоточная модель для извлечения признаков. Каждый поток состоял из нейросети ResNet [48], который извлекал признаки из необработанных входных данных, за которыми следует двухуровневый двунаправленный Bidirectional GRU (BiGRU), который моделирует вре-

менную динамику в каждом потоке. Этот метод позволяет нейросетевой модели эффективно обрабатывать как аудио-, так и видеоданные и извлекать значимые признаки для распознавания слов. Чтобы построить интегральную нейросеть в [49] использована рекуррентная нейросеть с длинной кратковременной памятью для извлечения признаков из необработанных данных. Обычно существующие методы обрабатывают интересующую область рта целиком, однако в работе [50] предложено использовать отдельные части (области) губ. Исследователи провели сравнительный анализ нейросетей для аудиовизуального распознавания речи, начиная с использования кросс-энтропийной функции потерь и затем переходя к коннекционистской временной классификации.

Первоначально архитектура трансформера была предложена в машинном переводе [51], после чего было проведено множество исследований по ее применению не только в акустическом, но и в аудиовизуальном распознавании речи. Она способна вычислять глобальный контекст для всех входных данных, что приводит к повышению производительности и более стабильному машинному обучению [52]. В работе [53] трансформер был объединен с рекуррентной нейросетью с длинной кратковременной памятью. Таким образом, сочетание современных методов глубокого машинного обучения и крупномасштабных аудиовизуальных корпусов позволяет достигать значительных результатов в точности распознавания.

Существующие методы распознавания аудиовизуальной речи кратко систематизированы в табл. 3.

В последние годы развитие технологий глубокого машинного обучения существенно изменило подходы к аудиовизуальному распознаванию речи. Отказ от традиционных методов извлечения признаков в пользу нейросетевых подходов привел к созданию эффективных интегральных архитектур, объединяющих этапы извлечения признаков и классификации в одной модели. Эти архитектуры позволяют моделировать как пространственную, так и временную динамику речевых данных, что значительно улучшает точность распознавания речи. В основе глубоких сверточных и рекуррентных нейросетей, а также моделей на основе трансформера, лежат эффективные методы, успешно применяемые к аудио- и визуальным данным. Это позволяет достигать высокой точности в распознавании аудиовизуальной речи даже в сложных условиях. Такой подход обладает значительным потенциалом и является активным направлением исследований в области распознавания аудиовизуальной речи.

Автоматический синтез акустической речи

Синтез речи играет ключевую роль в современном мире, применяясь в различных областях, от помощи людям с ограниченными возможностями до улучшения пользовательских интерфейсов в мобильных устройствах и автомобилях [54]. Например, для людей со слабым зрением синтез речи становится важным инструментом, преобразующим текстовые данные в аудиоформат, что делает информацию более доступной для восприятия. Это особенно актуально для чтения электронных документов, интернет-страниц и других текстовых материалов.

Кроме того, синтез речи улучшает пользовательские интерфейсы в различных устройствах и интеллектуальных приложениях. Например, в мобильных устройствах и умных домашних системах синтез речи используется для озвучивания текстовых уведомлений, команд голосового управления и других элементов интерфейса, упрощая взаимодействие пользователя с устройством. Это особенно важно для людей с ограниченными возможностями или пожилых, которым сложно использовать клавиатуру или сенсорный экран.

Актуальным направлением исследований в области синтеза речи является разработка голосовых ассистен-

Таблица 3. Систематизация методов распознавания аудиовизуальной речи
Table 3. Systematization of Audio-Visual Speech Recognition Methods

Ссылка	Метод	Описание
[34, 37]	Извлечение признаков из области рта и аудиосигнала	Традиционные методы аудиовизуального распознавания речи включают извлечение признаков из интересующей области рта, таких как движения губ и языка, а также из аудиосигнала, таких как мел-частотные кепстральные коэффициенты или спектрограммы. Затем эти признаки объединяются для создания комплексного представления речи, которое используется в дальнейшем для распознавания речевых слов или команд
[38–41]	2D-3D сверточные нейросетевые признаки	Нейросетевые методы используются для извлечения признаков из изображений и аудиосигналов. Примеры включают сверточные нейросети для изображений и использование 3D сверточных фильтров для обработки информации о губах
[42–50]	Интегральные нейросетевые архитектуры для распознавания речи	Методы на основе интегральных нейросетевых архитектур объединяют этапы извлечения признаков и классификации в рамках одной нейросети. Это позволяет более эффективно моделировать временную динамику речи
[51–53]	Модели с использованием трансформера	Модель трансформера, исходно разработанная для машинного перевода, успешно применяется в распознавании аудиовизуальной речи. Это позволяет моделировать глобальный контекст для всех входных данных, повышая производительность и стабильность машинного обучения

тов [55], которые позволяют пользователям взаимодействовать с компьютерами и другими устройствами через голосовые команды, что делает процесс взаимодействия более естественным и удобным. Например, голосовые ассистенты используют синтез речи для предоставления ответов на вопросы пользователей и выполнения команд.

Еще одним важным аспектом синтеза речи является его роль в развитии искусственного интеллекта. Технологии синтеза речи используются в различных системах искусственного интеллекта для создания более интеллектуальных и адаптивных интерфейсов. Например, синтез речи может применяться для создания персонализированных рекомендаций и подбора контента, учитывая предпочтения и интересы пользователей.

Нейросетевые методы синтеза акустической речи представляют собой подходы, использующие глубокие нейросети для генерации речи из текста или других модальностей. Несколько ключевых методов и их особенностей приведены в табл. 4.

Нейросетевые методы синтеза акустической речи продемонстрировали значительный прогресс в последние годы, обеспечивая высокое качество синтеза речи с естественным звучанием. Они широко используются в приложениях распознавания речи, голосовых помощниках, аудиокнигах и других областях, где требуется генерация человекоподобной речи.

Таблица 4. Нейросетевые методы синтеза звучащей речи
Table 4. Neural Network Methods for Speech Synthesis

Ссылка	Метод синтеза речи	Описание
[56]	WaveNet	Нейросеть, разработанная компанией Google DeepMind, которая использует генеративную модель для синтеза речи. Она создает аудиофайлы, имитирующие человеческую речь, с высоким качеством и естественностью
[57]	Tacotron	Нейросеть, разработанная Google, которая преобразует текст в аудиофайлы с речью. Этот метод использует механизм внимания для преобразования текста в голос
[58]	Deep Voice	Серия нейросетей, разработанных компанией Baidu, которые используются для синтеза речи из текстовых данных. Эти сети обучаются на больших объемах речевых данных для достижения высокого качества генерации речи
[59]	Transformer-TTS	Архитектура нейросетевой модели синтеза речи, основанная на трансформере, которая показывает высокую производительность в задачах генерации речи. Метод использует механизм внимания для преобразования текста в речь
[60]	FastSpeech	Метод синтеза речи, который позволяет синтезировать речь из текста. Метод отличается высокой скоростью работы и качеством сгенерированной речи
[61]	WaveGlow	Нейросеть, разработанная компанией Nvidia, основана на комбинации вариационных автокодировщиков и нормализационных потоков для генерации аудиофайлов. Она работает, пропуская случайный шум через многослойную нейросеть, чтобы создать аудио, имитирующее человеческую речь, с высокой степенью реалистичности и плавности
[62]	MelGAN	Архитектура нейросети, разработанная для синтеза речи, которая использует генеративную модель для генерации аудиосигналов на основе мел-спектрограмм
[63]	Parallel WaveGAN	Архитектура нейросети, основанная на генеративных состязательных сетях GAN и использующая конволюционные нейросети для синтеза аудио. Метод работает, пропуская случайный шум через генератор, который обучается создавать высококачественные аудиофайлы, имитирующие человеческую речь. Дискриминатор оценивает качество синтезированной речи, помогая генератору улучшать свои результаты
[64]	LPCNet	Гибридная архитектура, комбинирующая линейное предсказание коэффициентов (Linear Predictive Coding, LPC) с нейросетями для синтеза речи, объединяет классический метод LPC для моделирования основной формы речи с возможностями нейросетей в выявлении сложных закономерностей данных
[65]	Mel-Spectrogram GAN	Генеративная модель, которая использует мел-спектрограммы для синтеза речи
[66]	HiFi-GAN	Нейросеть, разработанная для синтеза высококачественной речи с помощью генеративно-состязательного метода
[67]	Tacotron 2	Улучшенная версия Tacotron [57], которая использует механизмы внимания для улучшения качества синтеза речи
[68]	Flowtron	Нейросеть, разработанная для синтеза речи с использованием модели потока для генерации аудиосигналов
[69]	WaveGrad	Метод синтеза речи, который использует градиентные методы для генерации аудиосигналов
[70]	ClariNet	Архитектура нейросети, разработанная для синтеза речи на основе глубокого машинного обучения, которая использует спектрограммы

Корпуса жестовой и аудиовизуальной речи

На сегодняшний день научным сообществом и крупными техническими корпорациями собрано и аннотировано множество аудиовизуальных речевых и жестовых корпусов для решения задач распознавания и синтеза аудиовизуальной и жестовой речи. Такие корпуса играют важную роль в разработке и обучении нейросетевых моделей компьютерного зрения и искусственного интеллекта для распознавания и синтеза звучащей и жестовой речи.

В работе [71] рассмотрены проблемы сбора корпусов жестовой речи для обучения нейросетевых моделей машинного обучения. Данное исследование предоставляет информацию о сложностях сбора высококачественных данных и подчеркивает важность учета конфиденциальности и этических соображений. Междисциплинарное исследование в [72] предоставило всесторонний обзор наборов данных жестовых корпусов, классифицируя их по различным факторам, таким как модальность, язык и применение, а также проведя анализ их пригодности для различных задач. Также в работе [72] выполнен анализ ограничений текущих корпусов и предложены будущие направления для улучшения, что делает их работу важным ресурсом для исследователей и практиков в области распознавания жестовых языков.

Приведем несколько примеров, наиболее часто используемых для задач машинного обучения жестовых корпусов.

ASL-LEX [73]. Корпус американского жестового языка (American Sign Language, ASL), содержащий информацию о лексических и семантических свойствах американских жестов.

RWTH-PHOENIX-Weather 2014T [74]. Корпус жестовой речи, содержащий видеозаписи жестов, связанных с обсуждением погоды. Используется для исследований по распознаванию жестов и мультимодальной обработке данных.

MSR Gesture 3D Dataset [75]. Корпус содержит трехмерные видеозаписи жестов, собранные с помощью камер глубины. Предназначен для разработки методов распознавания жестов с использованием трехмерной информации.

PHOENIX 2014T [76]. Крупномасштабный корпус жестовой речи, содержащий видеозаписи жестов и соответствующие текстовые транскрипции на немецком языке.

ChaLearn Looking at People Dataset [77]. Корпус содержит различные виды жестов, включая жесты рук, лица и тела, собранные в различных контекстах и для различных задач, таких как распознавание эмоций и действий.

TheRuSLan [78]. Корпус мультимедийных материалов по русскому жестовому языку, содержащий лексические единицы, связанные с продуктами питания в супермаркете. Все видеоматериалы представлены в высоком разрешении, а также включают карты глубины, собранные с помощью устройства MS Kinect v2.

AUTSL [79]. Крупномасштабный корпус турецкого жестового языка, содержащий данные, записанные с использованием MS Kinect v2. Корпус включает видеоданные, карту глубины и координаты скелета для каждого жеста.

HaGRID [80]. Крупномасштабный корпус от компании Сбер предназначенный для разработки систем распознавания жестов рук с учетом взаимодействия с устройствами. Корпус содержит больше 554 тысяч изображений и аннотации ограничивающих рамок с метками жестов, предназначенные для решения задач обнаружения руки и классификации жестов. Он создан с учетом возможности распознавания не только статических, но и динамических жестов. Для обеспечения разнообразия корпус собран с использованием краудсорсинговых платформ, при участии более 37 тыс. людей в различных сценах с разными условиями естественного освещения.

Эти корпуса представляют собой ценные ресурсы для исследований в области жестовой речи, обеспечивая данные для обучения и тестирования нейросетевых моделей распознавания и интерпретации жестов. Они помогают улучшить точность и эффективность систем распознавания жестов и мультимодальных интерфейсов.

Аудиовизуальные речевые корпуса представляют собой коллекции данных, включающие аудио- и видеозаписи речи, которые используются для разработки и оценки алгоритмов распознавания речи, синтеза речи, распознавания говорящего и других задач обработки речи. Перечислим примеры нескольких наиболее часто используемых в научных работах аудиовизуальных корпусов.

AVLetters Dataset [81]. Корпус содержит видеозаписи, на которых показано произношение букв английского алфавита, с соответствующими звуками.

GRID Corpus [82]. Корпус включает аудио- и видеозаписи говорящих, произносящих фразы на английском языке, а также текстовые транскрипции.

Lip Reading in the Wild [83]. Корпус содержит видеозаписи говорящих, произносящих короткие слова на английском языке.

MOBIO Dataset [84]. Данный корпус включает аудио- и видеозаписи говорящих на разных языках, а также биометрические данные лица и голоса. Используется для исследований по мультимодальному биометрическому распознаванию и аутентификации.

MIRACL-Voice Dataset [85]. Корпус содержит аудио- и видеозаписи различных речевых команд на нескольких языках, включая английский, французский и немецкий. Используется для разработки и оценки систем распознавания речи и управления голосом.

Перечисленные корпуса позволяют проводить исследования в области аудиовизуальной обработки речи, включая обучение и тестирование нейросетевых моделей, а также их оценку и сравнение с существующими методами, в том числе сравнение с лучшими на данный момент моделями (State-of-the-Art, SOTA).

Однако существующие аудиовизуальные и жестовые корпуса часто ограничены как по количеству данных, так и по разнообразию сценариев и условий. Это ограничение означает, что нейросетевые модели, обученные на этих наборах данных, могут оказаться недостаточно обобщающими для реальных сценариев. Более того,

такие корпуса могут не учитывать разнообразие различных культур, диалектов и аспектов поведения, что осложняет создание универсальных и точных систем распознавания и синтеза звучащей и жестовой речи. Для обучения реальных систем необходимы более разнообразные и объемные корпуса, которые охватывают широкий спектр сценариев и условий, чтобы обеспечить их эффективную работу в различных контекстах и для различных пользователей.

Кроме того, существующие корпуса часто могут быть недостаточно размечены или содержать ограниченное количество аннотированных данных. Также существует проблема доступности данных, поскольку некоторые корпуса могут быть закрытыми или недоступными для широкой общественности, что затрудняет исследования. В целом, для развития более точных, эффективных и универсальных систем распознавания и синтеза звучащей и жестовой речи необходимо больше усилий по созданию и расширению разнообразных и крупномасштабных корпусов, а также по улучшению их разметки и доступности для исследователей.

Существующие системы автоматического машинного сурдоперевода

В последние годы прогресс в области технологий обработки речи привел к разработке и улучшению систем перевода жестового языка для людей с нарушениями слуха или речи. Эти системы играют важную роль в обеспечении коммуникации и доступа к информации для данных групп пользователей. Однако, несмотря на успешные тесты в контролируемых условиях, многие системы сурдоперевода и их отдельные компоненты (распознавание/синтез) сталкиваются с проблемами в реальных условиях применения. Основные причины включают: изменяющиеся условия освещения (в реальных условиях освещение может сильно варьироваться, что негативно влияет на качество видео и, соответственно, на точность распознавания жестов); шум и отвлекающие факторы (в реальных ситуациях окружающий визуальный шум и другие отвлекающие факторы могут мешать точному распознаванию жестов); вариативность жестов (разные люди могут выполнять одни и те же жесты по-разному, что требует от систем высокой гибкости и способности адаптироваться к индивидуальным особенностям); ограниченные обучающие данные (для эффективного обучения систем требуются большие и разнообразные жестовые корпуса, которые часто недоступны или ограничены).

Таким образом, автоматические системы машинного перевода жестовой речи можно разделить на два основных типа: основанные на компьютерном зрении и на сенсорах. Первые используют камеры для захвата жестов и их распознавания с помощью методов машинного обучения, в то время как вторые применяют сенсоры для обнаружения движений рук и пальцев и их преобразования в звучащую речь.

Одним из ключевых аспектов достижения высоких результатов является интеграция различных методов и подходов. Например, использование сверточных нейросетей для обработки визуальных данных в сочетании с разновидностями рекуррентных нейросетей или трансформерами для обработки последовательностей может значительно повысить точность распознавания жестовой речи. Сверточные нейросети эффективно выделяют пространственные признаки жестов, тогда как рекуррентные и трансформеры успешно обрабатывают временные зависимости о различных жестикуляциях и движениях. Кроме того, комбинирование распознавания аудиовизуальных данных с синтезом жестовой речи может существенно улучшить производительность системы в реальных условиях. Например, система может использовать информацию о движениях губ для повышения точности распознавания речи в шумной обстановке. В свою очередь, синтез речи может быть улучшен за счет использования информации о жестах, что обеспечивает более естественное и синхронизированное воспроизведение.

Комбинирование различных методов и подходов в системах автоматического сурдоперевода открывает новые возможности для повышения точности, производительности и естественности перевода. Существующие системы автоматического машинного перевода/сурдоперевода приведены в табл. 5.

Представленные в табл. 5 системы — только небольшая часть существующих технологий перевода жестового языка, и каждая из них имеет свои уникальные особенности и преимущества. Они играют важную роль в обеспечении коммуникации и доступа к информации для людей со слуховыми или речевыми нарушениями. В то же время стоит отметить, что на сегодняшний день не существует надежной системы, комплексирующей распознавание и синтез акустической и жестовой речи.

На основе проведенного анализа можно обозначить следующие основные проблемы в создании систем автоматического машинного сурдоперевода и выделить основные тренды и пути решения для их преодоления.

1) Шум в данных. Одной из основных проблем является наличие шума в данных, который может существенно снизить точность распознавания жестов и синтеза речи. Это вызвано плохим освещением, фоновым шумом или нечеткими изображениями.

Решение: для уменьшения влияния шума необходимо применять методы предобработки данных, такие как фильтрация изображений, улучшение контраста, шумоподавление и другие методы цифровой обработки изображений. Использование данных из различных источников и в разных условиях также способствует обучению более устойчивых к шуму нейросетевых моделей.

2) Вариативность жестов. Жесты могут значительно варьироваться в зависимости от индивидуальных особенностей людей, их стиля жестовой речи и скорости выполнения жестов.

Решение: для повышения точности распознавания необходимо использовать более разнообразные и крупномасштабные корпуса данных, которые учитывают различные стили жестовой речи. Аугментация данных и использование методов регуляризации также способны помочь нейросетевым моделям лучше обобщать новые примеры.

Таблица 5. Системы автоматического машинного сурдоперевода
Table 5. Automatic Machine Sign Language Translation Systems

Система	Основные характеристики	Предназначения/особенности
SignAll [86]	Использует технологии распознавания и синтеза жестовой речи	— Предоставляет возможность автоматического сурдоперевода с жестовой речи в текст и обратно; — имеет гибкую архитектуру, позволяющую интегрировать систему в различные среды обучения и коммуникации
Google Live Transcribe [87]	Распознает аудиосигналы и выводит текстовую транскрипцию	— Предназначена в первую очередь для людей с нарушениями слуха, чтобы они могли читать речь в реальном времени; — имеет возможность перевода на несколько языков
DeepHand [88]	Использует методы машинного обучения для распознавания различных жестикуляций и жестов	— Предназначена для автоматического перевода жестовой речи на текст; — имеет возможность адаптации к различным жестовым языкам и стилевым особенностям
MotionSavvy [89]	Использует камеру для распознавания жестов и перевода их в текст	— Ориентирована на общение с людьми, не владеющими жестовым языком, путем автоматического перевода жестов в текст и наоборот; — имеет возможность обучения системы новым жестам и адаптации к индивидуальным особенностям пользователей
Microsoft Translator [90]	Использует интеллектуальные технологии для автоматического распознавания и синтеза речи	— Предназначена для автоматического перевода аудио- и текстовых сообщений; — имеет поддержку множества языков и диалектов
Motion Gesture Recognition [91]	Использует камеру для распознавания движений и жестов	— Позволяет пользователю контролировать устройства и взаимодействовать с компьютером через жестовую речь; — может быть интегрирована в различные платформы, такие как мобильные приложения и игровые консоли

3) Сложность жестов. Некоторые жесты очень похожи друг на друга, что затрудняет их различие.

Решение: применение более сложных нейросетевых архитектур, таких как трансформеры, способных учитывать контекст и последовательность жестов позволит повысить точность распознавания. Также дополнительное использование многомодальных данных, таких как видео, аудио, карт глубины и т. д., способно улучшить различение подобных жестов.

4) Проблемы с производительностью. Методы, основанные на глубоких нейросетях, часто требуют значительных вычислительных ресурсов для обучения.

Решение: для снижения вычислительных затрат необходимо применять методы компрессии и оптимизации нейросетей, такие как квантизация, дистилляция знаний и использование легковесных архитектур.

Заключение

Область распознавания и синтеза звучащей и жестовой речи является одной из наиболее активных и перспективных в современной компьютерной науке. Эти технологии находят широкое применение в различных областях, таких как системы управления, обработка естественного языка, компьютерное зрение, медицина, образование и т. д. В последние годы с развитием методов глубокого машинного обучения наблюдается значительный прогресс в области распознавания и синтеза, что позволяет создавать более точные, быстрые и эффективные системы двухстороннего машинного сурдоперевода.

В настоящей работе рассмотрены несколько ключевых аспектов разработки надежной системы автоматического двухстороннего машинного сурдоперевода. Например, методы распознавания и синтеза жестовой и аудивизуальной речи, а также существующие корпуса и системы сурдоперевода. Проведен тщательный анализ и сделаны выводы о современном состоянии области исследования и наиболее перспективных направлениях по каждому из аспектов.

В последние годы методы глубокого машинного обучения (сверточные, рекуррентные, генеративно-состязательные нейросети и трансформеры), стали основными инструментами для решения задач распознавания и синтеза, как звучащей, так и жестовой речи. Эти методы позволяют создавать более точные и эффективные нейросетевые модели, способные работать с различными типами данных и обеспечивать хорошее качество необходимого результата.

Отметим, что существующих аудиовизуальных и жестовых корпусов недостаточно для обучения реальных систем распознавания и синтеза звучащей и жестовой речи. Это связано с тем, что создание качественных корпусов требует больших временных и финансовых затрат, а также экспертных знаний в области сбора и разметки данных. Большинство существующих корпусов содержат ограниченное количество примеров, что затрудняет обучение нейросетевых моделей на реальных данных и приводит к недостаточному качеству их работы в реальных условиях.

Кроме того, существующие системы машинного сурдоперевода имеют определенные ограничения (низ-

кая точность распознавания и медленная скорость синтеза). В результате существующие методы и модели не всегда способны точно обрабатывать разнообразные сценарии и условия использования, такие как различные диалекты, акценты, скорость и интонация речи, а также жесты и мимику лица.

Для преодоления выявленных ограничений и повышения эффективности систем автоматического машинного сурдоперевода в реальных условиях можно выделить следующие перспективные направления для их усовершенствования: улучшение алгоритмов предобработки данных (разработка более эффективных методов для фильтрации шума и улучшения качества видео в условиях плохого освещения; интеграция многомодальных данных (аудио, видео, текст, карты глубины, данные сенсоров и т. д.); обучение на расширенных корпусах (создание и использование более крупномасштабных и разнообразных жестовых и аудиовизуальных корпусов); оптимизация моделей (применение современных методов машинного обучения и оптимизации моделей для повышения их производительности и устойчивости к изменяющимся условиям).

Для улучшения качества систем распознавания и синтеза речи необходимо продолжить исследования в области сбора и разметки данных, разработки новых методов и моделей глубокого машинного обучения, а также создания инновационных технологий для обработки аудио- и видеоданных. Также важно учитывать особенности различных культур и языков, чтобы обеспечить широкую доступность и использование этих технологий для всех пользователей.

Литература

1. Mehrish A., Majumder N., Bharadwaj R., Mihalcea R., Poria S. A review of deep learning techniques for speech processing // Information Fusion. 2023. V. 99. P. 101869. https://doi.org/10.1016/j.inffus.2023.101869
2. 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
3. Axyonov A., Ryumin D., Ivanko D., Kashevnik A., Karpov A. Audio-visual speech recognition in-the-wild: multi-angle vehicle cabin corpus and attention-based method // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2024. P. 8195–8199. https://doi.org/10.1109/ICASSP48485.2024.10448048
4. Ma P., Haliassos A., Fernandez-Lopez A., Chen H., Petridis S., Pantic M. Auto-AVSR: audio-visual speech recognition with automatic labels // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023. P. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096889
5. Wang X., Mi J., Li B., Zhao Y., Meng J. CATNet: Cross-modal fusion for audio-visual speech recognition // Pattern Recognition Letters. 2024. V. 178. P. 216–222. https://doi.org/10.1016/j.patrec.2024.01.002
6. Ryumin D., Axyonov A., Ryumina E., Ivanko D., Kashevnik A., Karpov A. Audio-visual speech recognition based on regulated transformer and spatio-temporal fusion strategy for driver assistive systems // Expert Systems with Applications. 2024. V. 252. Part. A. P. 124159. https://doi.org/10.1016/j.eswa.2024.124159
7. Ryumin D., Karpov A. Towards automatic recognition of sign language gestures using kinect 2.0 // Lecture Notes in Computer Science. 2017. V. 10278. P. 89–101. https://doi.org/10.1007/978-3-319-58703-5_7
8. Keskin C., Kıraç F., Kara Y.E., Akarun L. Hand pose estimation and hand shape classification using multi-layered randomized decision forests // Lecture Notes in Computer Science. 2012. V. 7577. P. 852–863. https://doi.org/10.1007/978-3-642-33783-3_61
9. Keskin C., Kıraç F., Kara Y.E., Akarun L. Real time hand pose estimation using depth sensors // Consumer Depth Cameras for Computer Vision: Research Topics and Applications. 2013. P. 119–137. https://doi.org/10.1007/978-1-4471-4640-7_7
10. Taylor J., Tankovich V., Tang D., Keskin C., Kim D., Davidson P., Kowdle A., Izadi S. Articulated distance fields for ultra-fast tracking of hands interacting // ACM Transactions on Graphics. 2017. V. 36. N 6. P. 1–12. https://doi.org/10.1145/3130800.3130853
11. Kındıroğlu A.A., Özdemir O., Akarun L. Temporal accumulative features for sign language recognition // Proc. of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). 2019. P. 1288–1297. https://doi.org/10.1109/ICCVW.2019.00164
12. Orbay A., Akarun L. Neural sign language translation by learning tokenization // Proc. of the 15th International Conference on Automatic Face and Gesture Recognition (FG). 2020. P. 222–228. https://doi.org/10.1109/FG47880.2020.00002

References

1. Mehrish A., Majumder N., Bharadwaj R., Mihalcea R., Poria S. A review of deep learning techniques for speech processing. *Information Fusion*, 2023, vol. 99, pp. 101869. https://doi.org/10.1016/j.inffus.2023.101869
2. 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
3. Axyonov A., Ryumin D., Ivanko D., Kashevnik A., Karpov A. Audio-visual speech recognition in-the-wild: multi-angle vehicle cabin corpus and attention-based method. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 8195–8199. https://doi.org/10.1109/ICASSP48485.2024.10448048
4. Ma P., Haliassos A., Fernandez-Lopez A., Chen H., Petridis S., Pantic M. Auto-AVSR: audio-visual speech recognition with automatic labels. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096889
5. Wang X., Mi J., Li B., Zhao Y., Meng J. CATNet: Cross-modal fusion for audio-visual speech recognition. *Pattern Recognition Letters*, 2024, vol. 178, pp. 216–222. https://doi.org/10.1016/j.patrec.2024.01.002
6. Ryumin D., Axyonov A., Ryumina E., Ivanko D., Kashevnik A., Karpov A. Audio-visual speech recognition based on regulated transformer and spatio-temporal fusion strategy for driver assistive systems. *Expert Systems with Applications*, 2024, vol. 252. part. A, pp. 124159. https://doi.org/10.1016/j.eswa.2024.124159
7. Ryumin D., Karpov A. Towards automatic recognition of sign language gestures using kinect 2.0. *Lecture Notes in Computer Science*, 2017, vol. 10278, pp. 89–101. https://doi.org/10.1007/978-3-319-58703-5_7
8. Keskin C., Kıraç F., Kara Y.E., Akarun L. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. *Lecture Notes in Computer Science*, 2012, vol. 7577, pp. 852–863. https://doi.org/10.1007/978-3-642-33783-3_61
9. Keskin C., Kıraç F., Kara Y.E., Akarun L. Real time hand pose estimation using depth sensors. *Consumer Depth Cameras for Computer Vision: Research Topics and Applications*, 2013, pp. 119–137. https://doi.org/10.1007/978-1-4471-4640-7_7
10. Taylor J., Tankovich V., Tang D., Keskin C., Kim D., Davidson P., Kowdle A., Izadi S. Articulated distance fields for ultra-fast tracking of hands interacting. *ACM Transactions on Graphics*, 2017, vol. 36, no. 6, pp. 1–12. https://doi.org/10.1145/3130800.3130853
11. Kındıroğlu A.A., Özdemir O., Akarun L. Temporal accumulative features for sign language recognition. *Proc. of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 1288–1297. https://doi.org/10.1109/ICCVW.2019.00164
12. Orbay A., Akarun L. Neural sign language translation by learning tokenization. *Proc. of the 15th International Conference on Automatic Face and Gesture Recognition (FG)*, 2020, pp. 222–228. https://doi.org/10.1109/FG47880.2020.00002

13. Camgoz N.C., Hadfield S., Koller O., Ney H., Bowden R. Neural sign language translation // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2018. P. 7784–7793. https://doi.org/10.1109/CVPR.2018.00812
14. Koller O., Camgoz N.C., Ney H., Bowden R. Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020. V. 42. N 9. P. 2306–2320. https://doi.org/10.1109/TPAMI.2019.2911077
15. Camgoz N.C., Koller O., Hadfield S., Bowden R. Multi-channel transformers for multi-articulatory sign language translation // Lecture Notes in Computer Science. 2020. V. 12538. P. 301–319. https://doi.org/10.1007/978-3-030-66823-5_18
16. Narayana P., Beveridge J.R., Draper B.A. Gesture recognition: focus on the hands // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2018. P. 5235–5244. https://doi.org/10.1109/CVPR.2018.00549
17. Zhu G., Zhang L., Shen P., Song J. Multimodal gesture recognition using 3-D convolution and convolutional LSTM // IEEE Access. 2017. V. 5. P. 4517–4524. https://doi.org/10.1109/ACCESS.2017.2684186
18. Abavisani M., Joze H.R.V., Patel V.M. Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019. P. 1165–1174. https://doi.org/10.1109/CVPR.2019.00126
19. Elboushaki A., Hannane R., Afdel K., Koutti L. MultiD-CNN: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences // Expert Systems with Applications. 2020. V. 139. P. 112829. https://doi.org/10.1016/j.eswa.2019.112829
20. Amangeldy N., Kudubayeva S., Kassymova A., Karipzhanova A., Razakhova B., Kuralov S. Sign language recognition method based on palm definition model and multiple classification // Sensors. 2022. V. 22. N 17. P. 6621. https://doi.org/10.3390/s22176621
21. Damaneh M.M., Mohanna F., Jafari P. Static hand gesture recognition in sign language based on convolutional neural network with feature extraction method using ORB descriptor and Gabor filter // Expert Systems with Applications. 2023. V. 211. P. 118559. https://doi.org/10.1016/j.eswa.2022.118559
22. Núñez-Marcos A., Perez-de-Viñaspre O., Labaka G. A survey on sign language machine translation // Expert Systems with Applications. 2023. V. 213. Part. B. P. 118993. https://doi.org/10.1016/j.eswa.2022.118993
23. Bohacek M., Hrúz M. Learning from what is already out there: few-shot sign language recognition with online dictionaries // Proc. of the 17th International Conference on Automatic Face and Gesture Recognition (FG). 2023. P. 1–6. https://doi.org/10.1109/FG57933.2023.10042544
24. Wei S.E., Ramakrishna V., Kanade T., Sheikh Y. Convolutional pose machines // Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. P. 4724–4732. https://doi.org/10.1109/CVPR.2016.511
25. Naert L., Larboulette C., Gibet S. A survey on the animation of signing avatars: from sign representation to utterance synthesis // Computers and Graphics. 2020. V. 92. P. 76–98. https://doi.org/10.1016/j.cag.2020.09.003
26. Mujahid A., Awan M.J., Yasin A., Mohammed M.A., Damaševičius R., Maskeliūnas R., Abdulkareem K.H. Real-time hand gesture recognition based on deep learning YOLOv3 model // Applied Sciences. 2021. V. 11. N 9. P. 4164. https://doi.org/10.3390/app11094164
27. Wang Y., Yu B., Wang L., Zu C., Lalush D.S., Lin W., Wu X., Zhou J., Shen D., Zhou L. 3D conditional generative adversarial networks for high-quality PET image estimation at low dose // NeuroImage. 2018. V. 174. P. 550–562. https://doi.org/10.1016/j.neuroimage.2018.03.045
28. Vahdat A., Kautz J. NVAE: A deep hierarchical variational autoencoder // Proc. of the Neural Information Processing Systems (NeurIPS). 2020. P. 19667–19679.
29. Ma C., Guo Y., Yang J., An W. Learning multi-view representation with LSTM for 3-D shape recognition and retrieval // IEEE Transactions on Multimedia. 2019. V. 21. N 5. P. 1169–1182. https://doi.org/10.1109/TMM.2018.2875512
30. Vasileiadis M., Bouganis C.-S., Tzovaras D. Multi-person 3D pose estimation from 3D cloud data using 3D convolutional neural

13. Camgoz N.C., Hadfield S., Koller O., Ney H., Bowden R. Neural sign language translation. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7784–7793. https://doi.org/10.1109/CVPR.2018.00812
14. Koller O., Camgoz N.C., Ney H., Bowden R. Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, vol. 42, no. 9, pp. 2306–2320. https://doi.org/10.1109/TPAMI.2019.2911077
15. Camgoz N.C., Koller O., Hadfield S., Bowden R. Multi-channel transformers for multi-articulatory sign language translation. *Lecture Notes in Computer Science*, 2020, vol. 12538, pp. 301–319. https://doi.org/10.1007/978-3-030-66823-5_18
16. Narayana P., Beveridge J.R., Draper B.A. Gesture recognition: focus on the hands. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5235–5244. https://doi.org/10.1109/CVPR.2018.00549
17. Zhu G., Zhang L., Shen P., Song J. Multimodal gesture recognition using 3-D convolution and convolutional LSTM. *IEEE Access*, 2017, vol. 5, pp. 4517–4524. https://doi.org/10.1109/ACCESS.2017.2684186
18. Abavisani M., Joze H.R.V., Patel V.M. Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1165–1174. https://doi.org/10.1109/CVPR.2019.00126
19. Elboushaki A., Hannane R., Afdel K., Koutti L. MultiD-CNN: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences. *Expert Systems with Applications*, 2020, vol. 139, pp. 112829. https://doi.org/10.1016/j.eswa.2019.112829
20. Amangeldy N., Kudubayeva S., Kassymova A., Karipzhanova A., Razakhova B., Kuralov S. Sign language recognition method based on palm definition model and multiple classification. *Sensors*, 2022, vol. 22, no. 17, pp. 6621. https://doi.org/10.3390/s22176621
21. Damaneh M.M., Mohanna F., Jafari P. Static hand gesture recognition in sign language based on convolutional neural network with feature extraction method using ORB descriptor and Gabor filter. *Expert Systems with Applications*, 2023, vol. 211, pp. 118559. https://doi.org/10.1016/j.eswa.2022.118559
22. Núñez-Marcos A., Perez-de-Viñaspre O., Labaka G. A survey on sign language machine translation. *Expert Systems with Applications*, 2023, vol. 213. part. B, pp. 118993. https://doi.org/10.1016/j.eswa.2022.118993
23. Bohacek M., Hrúz M. Learning from what is already out there: few-shot sign language recognition with online dictionaries. *Proc. of the 17th International Conference on Automatic Face and Gesture Recognition (FG)*, 2023, pp. 1–6. https://doi.org/10.1109/FG57933.2023.10042544
24. Wei S.E., Ramakrishna V., Kanade T., Sheikh Y. Convolutional pose machines. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4724–4732. https://doi.org/10.1109/CVPR.2016.511
25. Naert L., Larboulette C., Gibet S. A survey on the animation of signing avatars: from sign representation to utterance synthesis. *Computers and Graphics*, 2020, vol. 92, pp. 76–98. https://doi.org/10.1016/j.cag.2020.09.003
26. Mujahid A., Awan M.J., Yasin A., Mohammed M.A., Damaševičius R., Maskeliūnas R., Abdulkareem K.H. Real-time hand gesture recognition based on deep learning YOLOv3 model. *Applied Sciences*, 2021, vol. 11, no. 9, pp. 4164. https://doi.org/10.3390/app11094164
27. Wang Y., Yu B., Wang L., Zu C., Lalush D.S., Lin W., Wu X., Zhou J., Shen D., Zhou L. 3D conditional generative adversarial networks for high-quality PET image estimation at low dose. *NeuroImage*, 2018, vol. 174, pp. 550–562. https://doi.org/10.1016/j.neuroimage.2018.03.045
28. Vahdat A., Kautz J. NVAE: A deep hierarchical variational autoencoder. *Proc. of the Neural Information Processing Systems (NeurIPS)*, 2020, pp. 19667–19679.
29. Ma C., Guo Y., Yang J., An W. Learning multi-view representation with LSTM for 3-D shape recognition and retrieval. *IEEE Transactions on Multimedia*, 2019, vol. 21, no. 5, pp. 1169–1182. https://doi.org/10.1109/TMM.2018.2875512
30. Vasileiadis M., Bouganis C.-S., Tzovaras D. Multi-person 3D pose estimation from 3D cloud data using 3D convolutional neural

networks // Computer Vision and Image Understanding. 2019. V. 185. P. 12–23. https://doi.org/10.1016/j.cviu.2019.04.011
31. Lin J., Yuan Y., Shao T., Zhou K. Towards high-fidelity 3D face reconstruction from in-the-wild images using graph convolutional networks // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. P. 5890–5900. https://doi.org/10.1109/cvpr42600.2020.00593
32. Liu R., Shen J., Wang H., Chen C., Cheung S.-C., Asari V. Attention mechanism exploits temporal contexts: real-time 3D human pose reconstruction // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. P. 5063–5072. https://doi.org/10.1109/cvpr42600.2020.00511
33. Zhang Z., Sun L., Yang Z., Chen L., Yang Y. Global-correlated 3D-decoupling transformer for clothed avatar reconstruction // Proc. of the Neural Information Processing Systems (NeurIPS). 2023. P. 7818–7830.
34. Dupont S., Luettin J. Audio-visual speech modeling for continuous speech recognition // IEEE Transactions on Multimedia. 2000. V. 2. N 3. P. 141–151. https://doi.org/10.1109/6046.865479
35. Ivanko D., Ryumin D., Axyonov A., Kashevnik A. Speaker-dependent visual command recognition in vehicle cabin: methodology and evaluation // Lecture Notes in Computer Science. 2021. V. 12997. P. 291–302. https://doi.org/10.1007/978-3-030-87802-3_27
36. Ivanko D., Ryumin D., Kipyatkova I., Axyonov A., Karpov A. Lip-reading using pixel-based and geometry-based features for multimodal human-robot interfaces // Smart Innovation, Systems and Technologies. 2020. V. 154. P. 477–486. https://doi.org/10.1007/978-981-13-9267-2_39
37. 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. https://doi.org/10.17586/2226-1494-2023-23-4-767-775
38. Petridis S., Pantic M. Deep complementary bottleneck features for visual speech recognition // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016. P. 2304–2308. https://doi.org/10.1109/ICASSP.2016.7472088
39. Takashima Y., Aihara R., Takiguchi T., Ariki Y., Mitani N., Omori K., Nakazono K. Audio-visual speech recognition using bimodal-trained bottleneck features for a person with severe hearing loss // Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2016. P. 277–281. https://doi.org/10.21437/Interspeech.2016-721
40. Ninomiya H., Kitaoka N., Tamura S., Iribe Y., Takeda K. Integration of deep bottleneck features for audio-visual speech recognition // Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2015. P. 563–567. https://doi.org/10.21437/Interspeech.2015-204
41. Ivanko D., Ryumin D., Karpov A. A review of recent advances on deep learning methods for audio-visual speech recognition // Mathematics. 2023. V. 11. N 12. P. 2665. https://doi.org/10.3390/math11122665
42. 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
43. Hong J., Kim M., Choi J., Ro Y.M. Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023. P. 18783–18794. https://doi.org/10.1109/CVPR52729.2023.01801
44. Li G., Deng J., Geng M., Jin Z., Wang T., Hu S., Cui M., Meng H., Liu X. Audio-visual end-to-end multi-channel speech separation, dereverberation and recognition // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2023. V. 31. P. 2707–2723. https://doi.org/10.1109/TASLP.2023.3294705
45. Burchi M., Timofte R. Audio-visual efficient conformer for robust speech recognition // Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2023. P. 2257–2266. https://doi.org/10.1109/WACV56688.2023.00229
46. Chang O., Liao H., Serdyuk D., Shahy A., Siohan O. Conformer is all you need for visual speech recognition // Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP).

networks. *Computer Vision and Image Understanding*, 2019, vol. 185, pp. 12–23. https://doi.org/10.1016/j.cviu.2019.04.011
31. Lin J., Yuan Y., Shao T., Zhou K. Towards high-fidelity 3D face reconstruction from in-the-wild images using graph convolutional networks. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5890–5900. https://doi.org/10.1109/cvpr42600.2020.00593
32. Liu R., Shen J., Wang H., Chen C., Cheung S.-C., Asari V. Attention mechanism exploits temporal contexts: real-time 3D human pose reconstruction. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5063–5072. https://doi.org/10.1109/cvpr42600.2020.00511
33. Zhang Z., Sun L., Yang Z., Chen L., Yang Y. Global-correlated 3D-decoupling transformer for clothed avatar reconstruction. *Proc. of the Neural Information Processing Systems (NeurIPS)*, 2023, pp. 7818–7830.
34. Dupont S., Luettin J. Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2000, vol. 2, no. 3, pp. 141–151. https://doi.org/10.1109/6046.865479
35. Ivanko D., Ryumin D., Axyonov A., Kashevnik A. Speaker-dependent visual command recognition in vehicle cabin: methodology and evaluation. *Lecture Notes in Computer Science*, 2021, vol. 12997, pp. 291–302. https://doi.org/10.1007/978-3-030-87802-3_27
36. Ivanko D., Ryumin D., Kipyatkova I., Axyonov A., Karpov A. Lip-reading using pixel-based and geometry-based features for multimodal human-robot interfaces. *Smart Innovation, Systems and Technologies*, 2020, vol. 154, pp. 477–486. https://doi.org/10.1007/978-981-13-9267-2_39
37. 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). https://doi.org/10.17586/2226-1494-2023-23-4-767-775
38. Petridis S., Pantic M. Deep complementary bottleneck features for visual speech recognition. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2304–2308. https://doi.org/10.1109/ICASSP.2016.7472088
39. Takashima Y., Aihara R., Takiguchi T., Ariki Y., Mitani N., Omori K., Nakazono K. Audio-visual speech recognition using bimodal-trained bottleneck features for a person with severe hearing loss. *Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2016, pp. 277–281. https://doi.org/10.21437/Interspeech.2016-721
40. Ninomiya H., Kitaoka N., Tamura S., Iribe Y., Takeda K. Integration of deep bottleneck features for audio-visual speech recognition. *Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015, pp. 563–567. https://doi.org/10.21437/Interspeech.2015-204
41. Ivanko D., Ryumin D., Karpov A. A review of recent advances on deep learning methods for audio-visual speech recognition. *Mathematics*, 2023, vol. 11, no. 12, pp. 2665. https://doi.org/10.3390/math11122665
42. 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
43. Hong J., Kim M., Choi J., Ro Y.M. Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 18783–18794. https://doi.org/10.1109/CVPR52729.2023.01801
44. Li G., Deng J., Geng M., Jin Z., Wang T., Hu S., Cui M., Meng H., Liu X. Audio-visual end-to-end multi-channel speech separation, dereverberation and recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023, vol. 31, pp. 2707–2723. https://doi.org/10.1109/TASLP.2023.3294705
45. Burchi M., Timofte R. Audio-visual efficient conformer for robust speech recognition. *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 2257–2266. https://doi.org/10.1109/WACV56688.2023.00229
46. Chang O., Liao H., Serdyuk D., Shahy A., Siohan O. Conformer is all you need for visual speech recognition. *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,

2024. P. 10136–10140. https://doi.org/10.1109/icassp48485.2024.10446532
47. Wand M., Koutník J., Schmidhuber J. Lipreading with long short-term memory // Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016. P. 6115–6119. https://doi.org/10.1109/ICASSP.2016.7472852
48. Assael Y.M., Shillingford B., Whiteson S., De Freitas N. LipNet: end-to-end sentence-level lipreading // arXiv. 2016. arXiv:1611.01599. https://doi.org/10.48550/arXiv.1611.01599
49. Shi B., Hsu W.N., Mohamed A. Robust self-supervised audio-visual speech recognition // Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2022. P. 2118–2122. https://doi.org/10.21437/interspeech.2022-99
50. Ivanko D., Ryumin D., Kashevnik A.M., Axyonov A., Kitenko A., Lashkov I., Karpov A. DAVIS: driver's audio-visual speech recognition // Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2022. P. 1141–1142.
51. Zhou P., Yang W., Chen W., Wang Y., Jia J. Modality attention for end-to-end audio-visual speech recognition // Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019. P. 6565–6569. https://doi.org/10.1109/ICASSP.2019.8683733
52. Makino T., Liao H., Assael Y., Shillingford B., Garcia B., Braga O., Siohan O. Recurrent neural network transducer for audio-visual speech recognition // Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). 2019. P. 905–912. https://doi.org/10.1109/ASRU46091.2019.9004036
53. Li J., Li C., Wu Y., Qian Y. Unified cross-modal attention: robust audio-visual speech recognition and beyond // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2024. V. 32. P. 1941–1953. https://doi.org/10.1109/TASLP.2024.3375641
54. Tan X., Qin T., Soong F., Liu T.Y. A survey on neural speech synthesis // arXiv. 2021. arXiv:2106.15561. https://doi.org/10.48550/arXiv.2106.15561
55. de Barcelos Silva A., Gomes M.M., da Costa C.A., da Rosa Righi R., Barbosa J.L.V., Pessin G., de Doncker G., Federizzi G. Intelligent personal assistants: a systematic literature review // Expert Systems with Applications. 2020. V. 147. P. 113193. https://doi.org/10.1016/j.eswa.2020.113193
56. Oord A., Li Y., Babuschkin I., Simonyan K., Vinyals O., Kavukcuoglu K., Driessche G., Lockhart E., Cobo L., Stimberg F., Casagrande N., Grewe D., Noury S., Dieleman S., Elsen E., Kalchbrenner N., Zen H., Graves A., King H., Walters T., Belov D., Hassabis D. Parallel wavenet: fast high-fidelity speech synthesis // Proc. of the 35th International Conference on Machine Learning (ICML). 2018. P. 3918–3926.
57. Wang Y., Skerry-Ryan R.J., Stanton D., Wu Y., Weiss R.J., Jaitly N., Yang Z., Xiao Y., Chen Z., Bengio S., Le Q., Agiomyrgiannakis Y., Clark R., Saurous R.A. Tacotron: towards end-to-end speech synthesis // Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2017. P. 4006–4010. https://doi.org/10.21437/Interspeech.2017-1452
58. Arık S.Ö., Chrzanowski M., Coates A., Diamos G., Gibiansky A., Kang Y., Li X., Miller J., Ng A., Raiman J., Sengupta S., Shoeybi M. Deep voice: real-time neural text-to-speech // Proc. of the 34th International Conference on Machine Learning (ICML). 2017. P. 195–204.
59. Li N., Liu S., Liu Y., Zhao S., Liu M. Neural speech synthesis with transformer network // Proc. of the AAAI Conference on Artificial Intelligence. 2019. V. 33. N 1. P. 6706–6713. https://doi.org/10.1609/AAAI.V33I01.33016706
60. Ren Y., Ruan Y., Tan X., Qin T., Zhao S., Zhao Z., Liu T.Y. Fastspeech: fast, robust and controllable text to speech // Proc. of the Neural Information Processing Systems (NeurIPS). 2019. P. 1–10.
61. Prenger R., Valle R., Catanzaro B. Waveglow: a flow-based generative network for speech synthesis // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019. P. 3617–3621. https://doi.org/10.1109/ICASSP.2019.8683143
62. Kumar K., Kumar R., De Boissiere T., Gestin L., Teoh W.Z., Sotelo J., de Brébisson A., Bengio Y., Courville A.C. Melgan: generative adversarial networks for conditional waveform synthesis // Proc. of the Neural Information Processing Systems (NeurIPS). 2019. P. 320–335.
63. Yamamoto R., Song E., Kim J.M. Parallel WaveGAN: a fast waveform generation model based on generative adversarial networks

2024, pp. 10136–10140. https://doi.org/10.1109/icassp48485.2024.10446532
47. Wand M., Koutník J., Schmidhuber J. Lipreading with long short-term memory. *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6115–6119. https://doi.org/10.1109/ICASSP.2016.7472852
48. Assael Y.M., Shillingford B., Whiteson S., De Freitas N. LipNet: end-to-end sentence-level lipreading. *arXiv*, 2016. arXiv:1611.01599. https://doi.org/10.48550/arXiv.1611.01599
49. Shi B., Hsu W.N., Mohamed A. Robust self-supervised audio-visual speech recognition. *Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2022, pp. 2118–2122. https://doi.org/10.21437/interspeech.2022-99
50. Ivanko D., Ryumin D., Kashevnik A.M., Axyonov A., Kitenko A., Lashkov I., Karpov A. DAVIS: driver's audio-visual speech recognition. *Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2022, pp. 1141–1142.
51. Zhou P., Yang W., Chen W., Wang Y., Jia J. Modality attention for end-to-end audio-visual speech recognition. *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6565–6569. https://doi.org/10.1109/ICASSP.2019.8683733
52. Makino T., Liao H., Assael Y., Shillingford B., Garcia B., Braga O., Siohan O. Recurrent neural network transducer for audio-visual speech recognition. *Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 905–912. https://doi.org/10.1109/ASRU46091.2019.9004036
53. Li J., Li C., Wu Y., Qian Y. Unified cross-modal attention: robust audio-visual speech recognition and beyond. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024, vol. 32, pp. 1941–1953. https://doi.org/10.1109/TASLP.2024.3375641
54. Tan X., Qin T., Soong F., Liu T.Y. A survey on neural speech synthesis. *arXiv*, 2021. arXiv:2106.15561. https://doi.org/10.48550/arXiv.2106.15561
55. de Barcelos Silva A., Gomes M.M., da Costa C.A., da Rosa Righi R., Barbosa J.L.V., Pessin G., de Doncker G., Federizzi G. Intelligent personal assistants: a systematic literature review. *Expert Systems with Applications*, 2020, vol. 147, pp. 113193. https://doi.org/10.1016/j.eswa.2020.113193
56. Oord A., Li Y., Babuschkin I., Simonyan K., Vinyals O., Kavukcuoglu K., Driessche G., Lockhart E., Cobo L., Stimberg F., Casagrande N., Grewe D., Noury S., Dieleman S., Elsen E., Kalchbrenner N., Zen H., Graves A., King H., Walters T., Belov D., Hassabis D. Parallel wavenet: fast high-fidelity speech synthesis. *Proc. of the 35th International Conference on Machine Learning (ICML)*, 2018, pp. 3918–3926.
57. Wang Y., Skerry-Ryan R.J., Stanton D., Wu Y., Weiss R.J., Jaitly N., Yang Z., Xiao Y., Chen Z., Bengio S., Le Q., Agiomyrgiannakis Y., Clark R., Saurous R.A. Tacotron: towards end-to-end speech synthesis. *Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017, pp. 4006–4010. https://doi.org/10.21437/Interspeech.2017-1452
58. Arık S.Ö., Chrzanowski M., Coates A., Diamos G., Gibiansky A., Kang Y., Li X., Miller J., Ng A., Raiman J., Sengupta S., Shoeybi M. Deep voice: real-time neural text-to-speech. *Proc. of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 195–204.
59. Li N., Liu S., Liu Y., Zhao S., Liu M. Neural speech synthesis with transformer network. *Proc. of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, no. 1, pp. 6706–6713. https://doi.org/10.1609/AAAI.V33I01.33016706
60. Ren Y., Ruan Y., Tan X., Qin T., Zhao S., Zhao Z., Liu T.Y. Fastspeech: fast, robust and controllable text to speech. *Proc. of the Neural Information Processing Systems (NeurIPS)*, 2019, pp. 1–10.
61. Prenger R., Valle R., Catanzaro B. Waveglow: a flow-based generative network for speech synthesis. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3617–3621. https://doi.org/10.1109/ICASSP.2019.8683143
62. Kumar K., Kumar R., De Boissiere T., Gestin L., Teoh W.Z., Sotelo J., de Brébisson A., Bengio Y., Courville A.C. Melgan: generative adversarial networks for conditional waveform synthesis. *Proc. of the Neural Information Processing Systems (NeurIPS)*, 2019, pp. 320–335.
63. Yamamoto R., Song E., Kim J.M. Parallel WaveGAN: a fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. *Proc. of the IEEE International*

with multi-resolution spectrogram // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. P. 6199–6203. https://doi.org/10.1109/ICASSP40776.2020.9053795
64. Valin J.M., Skoglund J. LPCNet: Improving neural speech synthesis through linear prediction // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019. P. 5891–5895. https://doi.org/10.1109/icassp.2019.8682804
65. Asimopoulos D.C., Nitsiou M., Lazaridis L., Fragulis G.F. Generative adversarial networks: a systematic review and applications // SHS Web of Conferences. 2022. V. 139. P. 03012. https://doi.org/10.1051/shsconf/202213903012
66. Kong J., Kim J., Bae J. Hifi-gan: generative adversarial networks for efficient and high fidelity speech synthesis // Proc. of the Neural Information Processing Systems (NeurIPS). 2020. P. 17022–17033.
67. Fang W., Chung Y.A., Glass J. Towards transfer learning for end-to-end speech synthesis from deep pre-trained language models // arXiv. 2019. arXiv:1906.07307. https://doi.org/10.48550/arXiv.1906.07307
68. Valle R., Shih K., Prenger R., Catanzaro B. Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis // Proc. of the 9th International Conference on Learning Representations (ICLR). 2021. P. 1–6.
69. Chen N., Zhang Y., Zen H., Weiss R.J., Norouzi M., Chan W. Wavegrad: estimating gradients for waveform generation // Proc. of the 9th International Conference on Learning Representations (ICLR). 2021. P. 1–8.
70. Ping W., Peng K., Chen J. Clarinet: Parallel wave generation in end-to-end text-to-speech // Proc. of the 7th International Conference on Learning Representations (ICLR). 2019. P. 1–7.
71. Camgöz N.C., Koller O., Hadfield S., Bowden R. Sign language transformers: joint end-to-end sign language recognition and translation // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. P. 10020–10030. https://doi.org/10.1109/CVPR42600.2020.01004
72. Bragg D., Koller O., Caselli N., Thies W. Exploring collection of sign language datasets: privacy, participation, and model performance // Proc. of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility. 2020. P. 1–14. https://doi.org/10.1145/3373625.3417024
73. Caselli N.K., Sehyr Z.S., Cohen-Goldberg A.M., Emmorey K. ASL-LEX: a lexical database of american sign language // Behavior Research Methods. 2017. V. 49. N 2. P. 784–801. https://doi.org/10.3758/s13428-016-0742-0
74. Forster J., Schmidt C., Koller O., Bellgardt M., Ney H. Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-Weather // Proc. of the 9th International Conference on Language Resources and Evaluation (LREC). 2014. P. 1911–1916.
75. Azad R., Asadi-Aghbolaghi M., Kasaei S., Escalera S. Dynamic 3D hand gesture recognition by learning weighted depth motion maps // IEEE Transactions on Circuits and Systems for Video Technology. 2019. V. 29. N 6. P. 1729–1740. https://doi.org/10.1109/TCSVT.2018.2855416
76. Chen Y., Wei F., Sun X., Wu Z., Lin S. A simple multi-modality transfer learning baseline for sign language translation // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022. P. 5110–5120. https://doi.org/10.1109/CVPR52688.2022.00506
77. Escalera S., Baró X., Gonzàlez J., Bautista M.A., Madadi M., Reyes M., Ponce-López V., Escalante H.J., Shotton J., Guyon I. ChaLearn looking at people challenge 2014: dataset and results // Lecture Notes in Computer Science. 2015. V. 8925. P. 459–473. https://doi.org/10.1007/978-3-319-16178-5_32
78. Kagirov I., Ivanko D., Ryumin D., Axyonov A., Karpov A. TheRuSLan: database of russian sign language // Proc. of the 12th International Conference on Language Resources and Evaluation (LREC). 2020. P. 6079–6085.
79. Sincan O.M., Keles H.Y. AUTSL: a large scale multi-modal turkish sign language dataset and baseline methods // IEEE Access. 2020. V. 8. P. 181340–181355. https://doi.org/10.1109/ACCESS.2020.3028072
80. Kapitanov A., Kvanchiani K., Nagaev A., Kraynov R., Makhliarchuk A. HaGRID — HAnd gesture recognition image dataset // Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2024. P. 4560–4569. https://doi.org/10.1109/WACV57701.2024.00451

Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6199–6203. https://doi.org/10.1109/ICASSP40776.2020.9053795
64. Valin J.M., Skoglund J. LPCNet: Improving neural speech synthesis through linear prediction. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5891–5895. https://doi.org/10.1109/icassp.2019.8682804
65. Asimopoulos D.C., Nitsiou M., Lazaridis L., Fragulis G.F. Generative adversarial networks: a systematic review and applications. *SHS Web of Conferences*, 2022, vol. 139, pp. 03012. https://doi.org/10.1051/shsconf/202213903012
66. Kong J., Kim J., Bae J. Hifi-gan: generative adversarial networks for efficient and high fidelity speech synthesis. *Proc. of the Neural Information Processing Systems (NeurIPS)*, 2020, pp. 17022–17033.
67. Fang W., Chung Y.A., Glass J. Towards transfer learning for end-to-end speech synthesis from deep pre-trained language models. *arXiv*, 2019, arXiv:1906.07307. https://doi.org/10.48550/arXiv.1906.07307
68. Valle R., Shih K., Prenger R., Catanzaro B. Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis. *Proc. of the 9th International Conference on Learning Representations (ICLR)*, 2021, pp. 1–6.
69. Chen N., Zhang Y., Zen H., Weiss R.J., Norouzi M., Chan W. Wavegrad: estimating gradients for waveform generation. *Proc. of the 9th International Conference on Learning Representations (ICLR)*, 2021, pp. 1–8.
70. Ping W., Peng K., Chen J. Clarinet: Parallel wave generation in end-to-end text-to-speech. *Proc. of the 7th International Conference on Learning Representations (ICLR)*, 2019, pp. 1–7.
71. Camgöz N.C., Koller O., Hadfield S., Bowden R. Sign language transformers: joint end-to-end sign language recognition and translation. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10020–10030. https://doi.org/10.1109/CVPR42600.2020.01004
72. Bragg D., Koller O., Caselli N., Thies W. Exploring collection of sign language datasets: privacy, participation, and model performance. *Proc. of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, 2020, pp. 1–14. https://doi.org/10.1145/3373625.3417024
73. Caselli N.K., Sehyr Z.S., Cohen-Goldberg A.M., Emmorey K. ASL-LEX: a lexical database of american sign language. *Behavior Research Methods*, 2017, vol. 49, no. 2, pp. 784–801. https://doi.org/10.3758/s13428-016-0742-0
74. Forster J., Schmidt C., Koller O., Bellgardt M., Ney H. Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-Weather. *Proc. of the 9th International Conference on Language Resources and Evaluation (LREC)*, 2014, pp. 1911–1916.
75. Azad R., Asadi-Aghbolaghi M., Kasaei S., Escalera S. Dynamic 3D hand gesture recognition by learning weighted depth motion maps. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019, vol. 29, no. 6, pp. 1729–1740. https://doi.org/10.1109/TCSVT.2018.2855416
76. Chen Y., Wei F., Sun X., Wu Z., Lin S. A simple multi-modality transfer learning baseline for sign language translation. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5110–5120. https://doi.org/10.1109/CVPR52688.2022.00506
77. Escalera S., Baró X., Gonzàlez J., Bautista M.A., Madadi M., Reyes M., Ponce-López V., Escalante H.J., Shotton J., Guyon I. ChaLearn looking at people challenge 2014: dataset and results. *Lecture Notes in Computer Science*, 2015, vol. 8925, pp. 459–473. https://doi.org/10.1007/978-3-319-16178-5_32
78. Kagirov I., Ivanko D., Ryumin D., Axyonov A., Karpov A. TheRuSLan: database of russian sign language. *Proc. of the 12th International Conference on Language Resources and Evaluation (LREC)*, 2020, pp. 6079–6085.
79. Sincan O.M., Keles H.Y. AUTSL: a large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 2020, vol. 8, pp. 181340–181355. https://doi.org/10.1109/ACCESS.2020.3028072
80. Kapitanov A., Kvanchiani K., Nagaev A., Kraynov R., Makhliarchuk A. HaGRID – HAnd gesture recognition image dataset. *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 4560–4569. https://doi.org/10.1109/WACV57701.2024.00451
81. Petridis S., Wang Y., Ma P., Li Z., Pantic M. End-to-end visual speech recognition for small-scale datasets. *Pattern Recognition Letters*,

81. Petridis S., Wang Y., Ma P., Li Z., Pantic M. End-to-end visual speech recognition for small-scale datasets // Pattern Recognition Letters. 2020. V. 131. P. 421–427. https://doi.org/10.1016/j.patrec.2020.01.022
82. Cooke M., Barker J., Cunningham S., Shao X. An audio-visual corpus for speech perception and automatic speech recognition // The Journal of the Acoustical Society of America. 2006. V. 120. N 5. P. 2421–2424. https://doi.org/10.1121/1.2229005
83. Chung J.S., Zisserman A. Lip reading in the wild // Lecture Notes in Computer Science. 2017. V. 10112. P. 87–103. https://doi.org/10.1007/978-3-319-54184-6_6
84. Sequeira A.F., Monteiro J.C., Rebelo A., Oliveira H.P. MobBIO: a multimodal database captured with a portable handheld device // Proc. of the 9th International Conference on Computer Vision Theory and Applications (VISAPP). 2014. P. 133–139. https://doi.org/10.5220/0004679601330139
85. Parekh D., Gupta A., Chhatpar S., Yash A., Kulkarni M. Lip reading using convolutional auto encoders as feature extractor // Proc. of the IEEE 5th International Conference for Convergence in Technology (I2CT). 2019. P. 1–6. https://doi.org/10.1109/I2CT45611.2019.9033664
86. Leeson L., Sheikh H. SIGNALL: a european partnership approach to deaf studies via new technologies // Proc. of the INTED. 2009. P. 1270–1279.
87. Loizides F., Basson S., Kanevsky D., Prilepova O., Savla S., Zaraysky S. Breaking boundaries with live transcribe: expanding use cases beyond standard captioning scenarios // Proc. of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility. 2020. P. 1–6. https://doi.org/10.1145/3373625.3417300
88. Sinha A., Choi C., Ramani K. DeepHand: robust hand pose estimation by completing a matrix imputed with deep features // Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. P. 4150–4158. https://doi.org/10.1109/CVPR.2016.450
89. Ee L.W.S., Ramachandiran C.R., Logeswaran R. Real-time sign language learning system // Journal of Physics: Conference Series. 2020. V. 1712. P. 12011. https://doi.org/10.1088/1742-6596/1712/1/012011
90. Junczys-Dowmunt M. Microsoft translator at wmt 2019: towards large-scale document-level neural machine translation // Proc. of the Conference on Machine Translation. 2019. P. 225–233. https://doi.org/10.18653/v1/W19-5321
91. Hong F., You S., Wei M., Zhang Y., Guo Z. MGRA: motion gesture recognition via accelerometer // Sensors. 2016. V. 16. N 4. P. 530. https://doi.org/10.3390/s16040530

81. Petridis S., Wang Y., Ma P., Li Z., Pantic M. End-to-end visual speech recognition for small-scale datasets // Pattern Recognition Letters. 2020, vol. 131, pp. 421–427. https://doi.org/10.1016/j.patrec.2020.01.022
82. Cooke M., Barker J., Cunningham S., Shao X. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 2006, vol. 120, no. 5, pp. 2421–2424. https://doi.org/10.1121/1.2229005
83. Chung J.S., Zisserman A. Lip reading in the wild. *Lecture Notes in Computer Science*, 2017, vol. 10112, pp. 87–103. https://doi.org/10.1007/978-3-319-54184-6_6
84. Sequeira A.F., Monteiro J.C., Rebelo A., Oliveira H.P. MobBIO: a multimodal database captured with a portable handheld device. *Proc. of the 9th International Conference on Computer Vision Theory and Applications (VISAPP)*, 2014, pp. 133–139. https://doi.org/10.5220/0004679601330139
85. Parekh D., Gupta A., Chhatpar S., Yash A., Kulkarni M. Lip reading using convolutional auto encoders as feature extractor. *Proc. of the IEEE 5th International Conference for Convergence in Technology (I2CT)*, 2019, pp. 1–6. https://doi.org/10.1109/I2CT45611.2019.9033664
86. Leeson L., Sheikh H. SIGNALL: a european partnership approach to deaf studies via new technologies. *Proc. of the INTED*, 2009, pp. 1270–1279.
87. Loizides F., Basson S., Kanevsky D., Prilepova O., Savla S., Zaraysky S. Breaking boundaries with live transcribe: expanding use cases beyond standard captioning scenarios. *Proc. of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, 2020, pp. 1–6. https://doi.org/10.1145/3373625.3417300
88. Sinha A., Choi C., Ramani K. DeepHand: robust hand pose estimation by completing a matrix imputed with deep features. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4150–4158. https://doi.org/10.1109/CVPR.2016.450
89. Ee L.W.S., Ramachandiran C.R., Logeswaran R. Real-time sign language learning system. *Journal of Physics: Conference Series*, 2020, vol. 1712, pp. 12011. https://doi.org/10.1088/1742-6596/1712/1/012011
90. Junczys-Dowmunt M. Microsoft translator at wmt 2019: towards large-scale document-level neural machine translation. *Proc. of the Conference on Machine Translation*, 2019, pp. 225–233. https://doi.org/10.18653/v1/W19-5321
91. Hong F., You S., Wei M., Zhang Y., Guo Z. MGRA: motion gesture recognition via accelerometer. *Sensors*, 2016, vol. 16, no. 4, pp. 530. https://doi.org/10.3390/s16040530

Авторы

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

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, ryumin.d@iias.spb.su

Authors

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

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

Статья поступила в редакцию 07.05.2024
Одобрена после рецензирования 28.05.2024
Принята к печати 16.09.2024

Received 07.05.2024
Approved after reviewing 28.05.2024
Accepted 16.09.2024