На правах рукописи

СТЕФАНИДИ АНТОН ФЕДОРОВИЧ

ИССЛЕДОВАНИЕ МУЛЬТИМОДАЛЬНЫХ АЛГОРИТМОВ БИОМЕТРИЧЕСКОЙ ИДЕНТИФИКАЦИИ НА ОСНОВЕ МЕТОДОВ ЦИФРОВОЙ ОБРАБОТКИ РЕЧЕВЫХ СИГНАЛОВ И ИЗОБРАЖЕНИЙ

2.2.13. Радиотехника, в том числе системы
и устройства телевидения

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата технических наук

Ярославль – 2022

Работа выполнена в Центре искусственного интеллекта и цифровой экономики ФГБОУ ВО «Ярославский государственный университет им. П.Г. Демидова»

Научный руководитель: **Хрящев Владимир Вячеславович**
кандидат технических наук, доцент,
доцент кафедры цифровых технологий
и машинного обучения ФГБОУ ВО «Ярославский государственный университет им. П.Г. Демидова»,
г. Ярославль.

Официальные оппоненты: **Медведева Елена Викторовна**
доктор технических наук, доцент,
профессор кафедры радиоэлектронных средств
ФГБОУ ВО «Вятский государственный университет» (ВятГУ), г. Киров.

Волохов Владимир Андреевич
кандидат технических наук, доцент,
научный сотрудник ООО «ЦРТ-инновации»,
г. Санкт-Петербург.

Ведущая организация: ФГБОУ ВО «Рязанский государственный радиотехнический университет имени В.Ф. Уткина» (РГРТУ), г. Рязань.

Защита диссертации состоится 21.09.2022 в 14.00 на заседании диссертационного совета Д 24.2.281.01 в ФГБОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых» по адресу: 600000, г. Владимир, ул. Горького, д. 87, ВлГУ, корп. 3, ауд. 301.

С диссертацией можно ознакомиться в библиотеке Владимирского государственного университета имени Александра Григорьевича и Николая Григорьевича Столетовых и на сайте диссертационного совета http://diss.vlsu.ru.

Автореферат разослан «24» июня 2022 г.

Отзывы на автореферат, заверенные печатью, просим направлять по адресу: 600000, г. Владимир, ул. Горького, д. 87, ВлГУ, РТ и РС Самойлову А.Г.

Ученый секретарь диссертационного совета Д 24.2.281.01, д.т.н., проф. А.Г. Самойлов

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Методы и алгоритмы цифровой обработки сигналов широко используются при построении радиотехнического и телевизионного оборудования, проектировании систем управления, создании систем хранения, поиска и сжатия информации. Особый интерес для различных практических приложений представляют цифровые алгоритмы анализа изображений и речевых сигналов. Например, если человек является объектом наблюдения, то его можно идентифицировать с помощью анализа таких сигналов, как оцифрованный отпечаток пальца, фотография лица или сетчатки глаза, запись речи. Системы распознавания личности на основе анализа уникальных физиологических и поведенческих черт индивида носят название биометрических. Такие методы получили массовое распространение, поскольку, в отличие от паролей и аппаратных ключей, физиологические особенности невозможно потерять или забыть.

Системы биометрической идентификации личности стали неотъемлемой частью повседневной жизни. Например, большинство современных мобильных телефонов оборудованы сканерами отпечатков пальцев или используют встроенную видеокамеру для аутентификации пользователя по лицу. Одной из актуальных задач в области биометрии является задача идентификации человека в сеансе видеоконференцсвязи (ВКС). На основе таких алгоритмов, например, строится система прокторинга для наблюдения и контроля за дистанционным испытанием. Такие системы уже используются в ведущих мировых ВУЗах. Также стоит отметить использование биометрических методов для построения систем контроля и управления доступом (СКУД) высокой надежности, что актуально для закрытых предприятий и объектов с высокими требованиями к безопасности.

В основе работы методов идентификации личности лежит анализ биометрических параметров человека. Запись речевого сигнала и изображение лица человека являются цифровыми «слепками» личности, однозначно ее определяющими. Однако следует отметить, что в практических приложениях использования данной технологии качество речевых сигналов и изображений лиц может быть существенно неидеальным, ввиду наличия ряда искажающих факторов. Системы идентификации человека (диктора) по голосу чувствительны к эффектам, возникающим при передаче и обработке данных, физиологическим особенностям говорящего, акустическим свойствам окружающей среды. Алгоритмы

распознавания пользователя по лицу имеют сильную зависимость от уровня освещенности, ракурса, качества фоторегистратора, а также чувствительны к возрастным изменениям и мимике. Системы идентификации личности на основе анализа одного биометрического параметра (унимодальные) уязвимы и могут быть взломаны в результате создания цифровой копии лица или голоса человека. Таким образом, с развитием технологий постоянно возникает потребность в разработке более совершенных алгоритмов идентификации.

Одним из перспективных направлений развития биометрических систем является разработка и исследование алгоритмов идентификации личности на основе двух и более биометрических параметров, так называемых мультимодальных решений. Подход на основе использования комбинации различных сигналов позволяет не только повысить устойчивость и точность работы биометрических систем, но и улучшить их надежность при попытках несанкционированного доступа.

Анализ современной научно-технической литературы показывает, что наиболее эффективным подходом для решения задачи распознавания образов на сегодняшний день является использование алгоритмов глубокого обучения. Так, сверточные нейронные сети (СНС) стали основным инструментом анализа видеоизображений в системах прикладного телевидения. Особенность данного подхода заключается в том, что признаки исследуемых объектов формируются автоматически в процессе обучения. Сгенерированные таким образом признаки позволяют, как правило, добиться наилучших результатов в задачах обнаружения, сегментации и распознавания объектов на цифровых изображениях, а также при идентификации диктора на основе анализа речевых сигналов.

Важнейший вклад в развитие области обработки визуальной информации и построения систем биометрической идентификации по изображению лица внесли отечественные и зарубежные ученые Ю.Б. Зубарев, М.И. Кривошеев, В.П. Дворкович, А.В. Дворкович, Ю.И. Журавлев, В.А. Сойфер, А.С. Конушин, Д.С. Ватолин, М.К. Чобану, Ю.С. Бехтин, Н.Н. Красильников, Ю.В. Визильтер, Э.М. Браверман, М.Н. Фаворская, Е.В. Медведева, П.Д. Филлипс, Э. Янг, А. Мартинес, А. Зиссерман, А. Ведальди, Р. Челлаппа, О.М. Пархи и др.

В области обработки речевых сигналов и систем идентификации диктора широкую известность получили работы таких ученых, как Л. Рабинер, Р. Шафер,

А. Оппенгейм, М. Сапажков, Д.А. Рейнольдс, Д. Хансен, Х. Ли, Т. Киннунен, Д. Повье, Х. Бейджи, Д. Гарсия-Ромеро и др.

Важнейшие результаты в теории построения нейросетевых моделей получены А.И. Галушкиным, К.В. Воронцовым, Я. Лекуном, Т. Кохоненом, Э. Энджи, И. Бенджио, Д. Хинтоном, Ф. Ли, Я. Гудфеллоу и др. В области построения мультимодальных биометрических систем следует выделить работы А.К. Джейна, А. Росса, Д. Фиерреса, Х. Ортега-Гарсии и др.

Таким образом, разработка новых нейросетевых алгоритмов идентификации личности для мультимодальных биометрических систем является актуальной научно-технической задачей и имеет практический интерес для развития цифровых систем в областях радиотехники и прикладного телевидения.

Целью работы является повышение точности систем идентификации личности путем разработки нейросетевых алгоритмов анализа речевых сигналов и изображений лиц.

Для достижения поставленной цели в диссертационной работе определены и решены следующие задачи:

- разработка комбинированного детектора голосовой активности;
- разработка нейросетевых алгоритмов идентификации личности на основе анализа речевых сигналов и изображений лиц;
- усовершенствование работы алгоритмов идентификации личности в условиях действия шумов и помех в речевых сигналах и наличия медицинской маски на изображениях лиц;
- разработка мультимодальных алгоритмов идентификации личности на основе комбинированного анализа речевых сигналов и изображений лиц.

Методы исследования. При решении поставленных задач применялись методы цифровой обработки сигналов и изображений, спектрального анализа, распознавания образов, теории нейронных сетей, машинного и глубокого обучения. Для практической реализации исследуемых алгоритмов применялись современные методы и инструменты программирования на языке Python, а также фреймворки глубокого обучения TensorFlow и Keras.

Объектом исследования являются алгоритмы биометрической идентификации, применяемые в системах прикладного телевидения и радиотехнических системах обработки и анализа цифровых сигналов.

Предметом исследования является разработка нейросетевых алгоритмов биометрической идентификации на основе анализа речевых сигналов и цифровых изображений лиц с целью повышения точности систем распознавания личности.

Научная новизна. В рамках диссертационной работы получены следующие результаты, обладающие научной новизной:

- комбинированный детектор голосовой активности для выделения речевых фрагментов на основе алгоритма решающих деревьев;
- робастный алгоритм голосовой биометрии на основе x-подобной нейросетевой структуры, обеспечивающий низкую деградацию качества в условиях действия шумов и помех;
- робастный алгоритм лицевой биометрии на основе сверточной нейронной сети, обеспечивающий низкую деградацию качества в условиях наличия медицинской маски;
- мультимодальные алгоритмы идентификации личности, выполняющие объединение модулей голосовой и лицевой биометрии на уровне принятия решения и слияния признаков.

Практическая значимость

- собран аудиовизуальный набор FaceSpeechDB, содержащий более 60 часов записи русскоязычной речи, а также набор аудиосигналов VADSpeakersDB, включающий 138000 фрагментов речи, шумов и пауз;
- разработаны робастные нейросетевые алгоритмы, для которых деградация точности в условиях зашумления речевых сигналов или наличия медицинской маски составляет в среднем 7-8%, что превосходит аналоги на 3-5% и более;
- установлено, что разработанные нейросетевые алгоритмы содержат в среднем в 15-25 раз меньше весовых параметров, что дает им существенное преимущество в скорости работы относительно аналогов;
- определено, что предложенные мультимодальные алгоритмы имеют преимущество в точности относительно унимодальных аналогов на 7% и более при зашумлении речевых сигналов, на 2% и более в условиях использования медицинской маски.

Разработанные алгоритмы биометрической идентификации требуют для своей практической реализации сравнительно небольших вычислительных ресурсов, что позволяет использовать их в системах обработки видеоизображений и речевых

сигналов, работающих в режиме реального времени, в том числе в задачах прокторинга при ВКС и при построении защищенных СКУД.

Результаты работы внедрены в соответствующие разработки ООО «Цифровые решения» г. Ярославль, ООО «ТЕКМЭН» г. Ярославль, ООО «СОФТ ВИЖН» г. Ярославль. Отдельные результаты диссертационной работы внедрены в учебный процесс ЯрГУ им. П.Г. Демидова в рамках дисциплин «Цифровая обработка речевых сигналов», «Цифровая обработка изображений», а также в научно-исследовательские работы при выполнении исследований в рамках гранта РФФИ № 19-37-90158 и грантов «Участник молодежного научно-инновационного конкурса» («УМНИК») и «СТАРТ» по договорам с Фондом содействия инновациям № 11758ГУ/2016 от 03.07.2017, № 3867ГС1/63173 от 24.12.2020.

Получены три свидетельства о государственной регистрации программ для ЭВМ (№ 2019613092, № 2021663249, № 2021681283).

Достоверность материалов диссертационной работы подтверждена корректным использованием инструментов математического моделирования и полученными экспериментальными результатами, согласующимися с теоретическими и практическими сведениями из научно-технических источников, апробацией трудов исследования на научно-практических конференциях различного уровня.

Апробация работы. Результаты работы докладывались и обсуждались на следующих научно-технических конференциях:

- 12-я международная научно-техническая конференция «Перспективные технологии в средствах передачи информации» (ПТСПИ), Суздаль, 2017.
- 7-я всероссийская конференция «Радиоэлектронные средства получения, обработки и визуализации информации» (РСПОВИ), Москва, 2017.
- 20-я, 22-я и 23-я международные конференции «Цифровая обработка сигналов и ее применение» (DSPA), Москва, 2018, 2020, 2021.
- 11-я международная конференция «International Conference on Machine Vision» (ICMV-2018), Мюнхен, Германия, 2018.
- 17-я и 18-я международные конференции «Новые информационные технологии и системы» (НИТиС), Пенза, 2020, 2021.
- 26-я международная конференция «Open Innovation Association» (FRUCT), Ярославль, 2020.

- 18-я международная конференция «IEEE East-West Design & Test Symposium (EWDTS-2020)», Варна, Болгария, 2020.
- 21-я международная конференция «Проблемы информатики в образовании, управлении, экономике и технике», Пенза, 2021.
- 77-я всероссийская конференция «Радиоэлектронные устройства и системы для инфокоммуникационных технологий» (REDS-2022), Москва, 2022.

Публикации. По теме диссертации опубликовано 16 научных работ, из них 3 статьи в журналах, рекомендованных ВАК, 3 работы, индексированных в SCOPUS, и 10 докладов на научных конференциях.

Структура и объем работы. Диссертация состоит из введения, четырех глав, заключения, списка литературы, содержащего 118 наименований, и 4 приложений. Она изложена на 130 страницах машинописного текста, содержит 48 рисунков и 18 таблиц.

Основные научные положения и результаты, выносимые на защиту:

- комбинированный детектор голосовой активности, для которого точность выделения речевых фрагментов составляет до 94%, что превосходит соответствующее аналоги на 2-3%;
- алгоритм идентификации диктора на основе x-подобной нейросетевой структуры, который может быть использован в зашумленной среде, где он превосходит аналоги в среднем на 5% и более;
- нейросетевой алгоритм идентификации личности по изображению лица, работающий в условиях присутствия медицинской маски, для которого деградация в точности составляет менее 7%, что превосходит аналогичные показатели на 3% и более;
- мультимодальные алгоритмы идентификации личности, превосходящие по точности унимодальные аналоги на 7% и более при зашумлении речевых сигналов, на 2% и более при использовании медицинской маски.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность выбранной темы, сформулированы цель и задачи исследования, изложены основные положения, выносимые на защиту, показаны научная новизна и практическая значимость работы.

В первой главе проведен анализ современного состояния разработок в области построения биометрических систем на основе анализа речевых сигналов и

цифровых изображений лиц. Описаны технологические барьеры, возникающие в процессе разработки алгоритмов на основе голосовой и лицевой биометрии. Рассмотрена классификация задач распознавания личности с использованием аудиосигналов и изображений.

Анализ научно-технической литературы показал, что на сегодняшнем этапе развития модели СНС являются главным инструментом для систем биометрической идентификации личности. Рассмотрены базовые структурные блоки для построения таких моделей, а также современные нейросетевые архитектуры. Проведен обзор методов комбинирования биометрических параметров. Особое внимание уделено анализу мультимодальных алгоритмов идентификации личности.

Описан этап подготовки двух оригинальных баз биометрических данных, содержащих записи русскоязычной речи. Подготовлена аудиовизуальная база FaceSpeechDB, содержащая 60 часов живой записи 104 человек. Еще одна собранная база аудиосигналов VADSpeakersDB включает 138000 фрагментов речи, шумов и пауз. Этот набор содержит записи живой речи 23 дикторов, которые разделены на фрагменты длительностью 10 мс. Каждый фрагмент имеет метку, определяющую его принадлежность к одному из классов – «речь» или «шум/пауза». Запись речи и видеоизображений производилась с использованием общедоступных приложений для ВКС в повседневных акустических и визуальных условиях, а также с применением записывающих устройств различного качества. Эти базы используются в дальнейшем для разработки и тестирования комбинированного детектора голосовой активности и нейросетевых алгоритмов идентификации личности.

Во второй главе исследуются нейросетевые алгоритмы идентификации личности на основе анализа речевых сигналов.

Такой тип сигналов состоит из речи, внешних шумов, шумов записывающего устройства и пауз. Наличие остановок в речевом сигнале с точки зрения систем идентификации диктора является негативным фактором. Для устранения данной проблемы разработан комбинированный детектор голосовой активности (КДГА). Он строится на основе объединения более простых детекторов: детектора анализа энергии фонограммы во временной области (далее – ДГА1), детектора на основе метода Тигера-Кайзера (далее – ДГА2) и детектора на основе частотного анализа фонограммы (далее – ДГА3). На Рисунке 1 представлена структурная схема разработанного алгоритма КДГА.

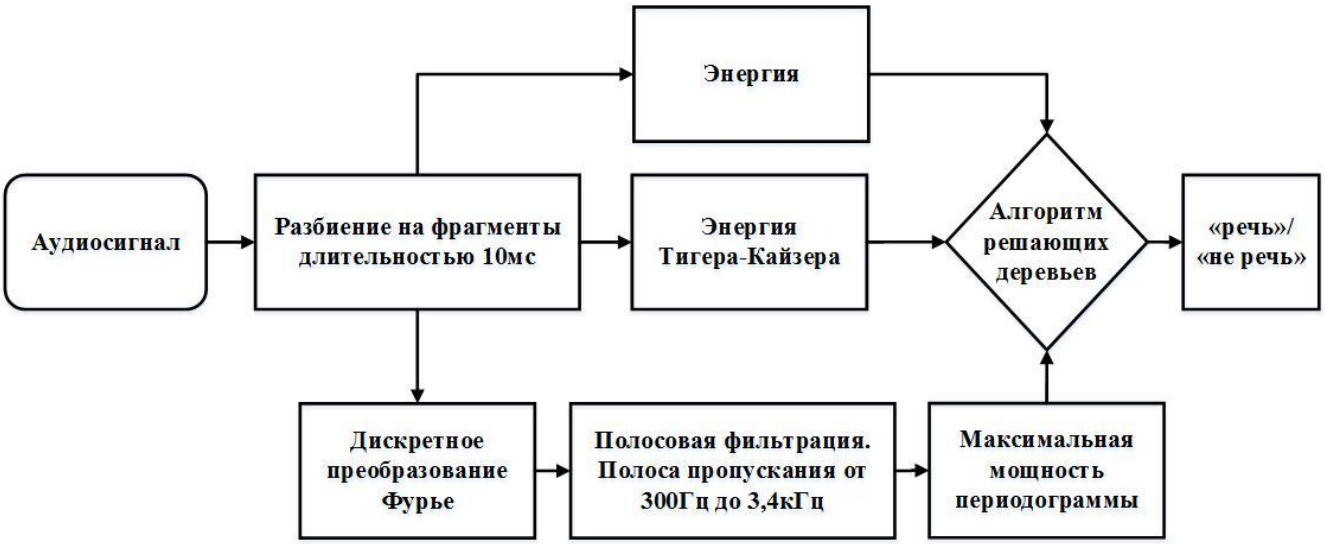

Рисунок 1 – Структурная схема комбинированного детектора голосовой активности

Тестирование выполняется с использованием собранного набора аудиосигналов VADSpeakersDB. Для анализа качества работы упомянутых детекторов используется набор метрик. Первая из них основана на оценке доли правильных ответов (*acc*). Подготовленный набор аудиосигналов является несбалансированным, поскольку фрагментов речи существенно больше фрагментов, содержащих паузы или помехи. Для учета этого свойства используется модификация метрики *acc* для несбалансированных данных (*accb*). Определим индикатор корректности распознавания фрагмента аудиосигнала:

$$c(x_i) = \begin{cases} 1, y_i = y_i' \\ 0, y_i \neq y_i' \end{cases},$$

где x_i – i-й фрагмент фонограммы, длительностью 10 мс; $c(x_i)$ – индикатор корректности распознавания i-го фрагмента; y_i – целевая метка фрагмента; y_i' – метка фрагмента, определяющая результат работы детектора. Тогда метрика *acc* определяется следующим образом:

$$acc = \frac{\sum_{i=1}^{n} c(x_i)}{n},$$

где n – количество всех фрагментов длительностью 10 мс в используемом наборе.

Доля правильных ответов на несбалансированных данных в задаче бинарной классификации рассчитывается по формуле:

$$accb = \frac{acc_{y_i=1} + acc_{y_i=0}}{2},$$

где $acc_{y_i=1}$ – доля правильно детектированных фрагментов класса «речь»; $acc_{y_i=0}$ – доля правильно детектированных фрагментов класса «шум/пауза».

Также для оценки качества работы детекторов используется гармоническое среднее между точностью и полнотой (*F*-мера, *F*):

$$F = 2 \cdot \frac{P \cdot R}{P + R},$$

где P – точность (precision), метрика, определяющая ошибки I рода, R – полнота (recall), метрика, определяющая ошибки II рода.

Дополнительно вычисляется *F*-мера на основе макро-усредняющего подхода, т.е. расчет метрики производится для каждого класса с нормировкой на общее количество классов:

$$F_{\text{макро}} = \frac{F_{y_i=1} + F_{y_i=0}}{2}.$$

Сравнительный анализ работы детекторов, приведенный в Таблице 1, показывает преимущество разработанного алгоритма КДГА, который позволяет повысить точность определения речевых фрагментов на 2-3%. Тестовая выборка насчитывала 20700 фрагментов длительностью 10 мс.

Таблица 1 – Сравнительный анализ детекторов голосовой активности

Метрики	**acc**	**accb**	**F**	**$F_{макро}$**
$ДГА_1$	0,90	0,88	0,93	0,88
$ДГА_2$	0,89	0,88	0,92	0,88
$ДГА_3$	0,89	0,87	0,92	0,87
КДГА	**0,91**	**0,90**	**0,94**	**0,90**

Комбинированный детектор голосовой активности используется для обработки базы FaceSpeechDB. В результате формируется набор для обучения и тестирования алгоритмов биометрической идентификации личности на основе голосовой биометрии. Набор включает более 32 тыс. аудиозаписей.

Для решения задачи идентификации личности на основе речевых сигналов необходимо определить уникальные акустические свойства и особенности голоса

диктора. Для этого используются частотные преставления речевых сигналов в виде спектрограмм и мел-частотных кепстральных коэффициентов (МЧКК). Основными этапами формирования МЧКК являются: выделение фрагмента речевого сигнала фиксированной длины, вычисление спектрограммы с использованием дискретного преобразования Фурье, применение банка мел-фильтров к периодограмме и вычисление логарифма, использование дискретного косинусного преобразования. Для идентификации личности с использованием речевых сигналов разработан нейросетевой алгоритм на основе x-векторной системы, показанный на Рисунке 2.

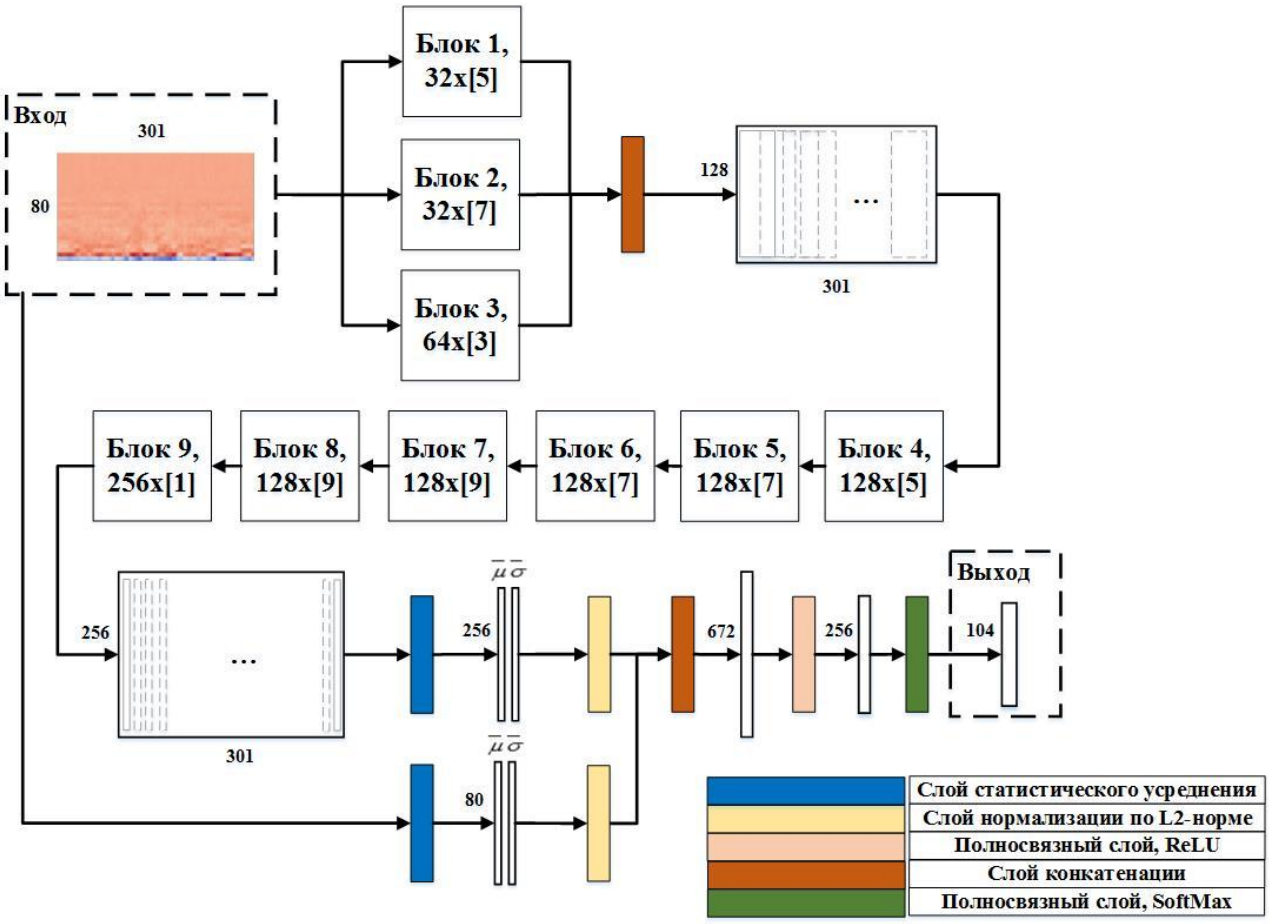

Рисунок 2 – Архитектура нейронной сети X-Speech

Базовыми структурными элементами такой архитектуры являются 9 блоков временной задержки с разным числом используемых фильтров и размерностью анализируемых фреймов. Каждый такой блок выполняет операцию свертки сигнала на входе. Все слои нейронной сети в качестве функции активации используют линейную ректификацию (ReLU) за исключением выходного, где применяется функция SoftMax. Отличительной особенностью разработанной архитектуры X-Speech является использование параллельного подключения нескольких блоков на входе архитектуры, а также дополнительный анализ карты МЧКК без прохода

через блоки временной задержки. Важно отметить, что архитектура X-Speech содержит менее 1 млн. весовых параметров, что на порядок меньше наиболее часто используемых современных аналогов.

В процессе обучения использовалась аугментация данных (генерирование новых данных на основе имеющихся). При этом сигналы подвергались следующим искажениям и преобразованиям: добавлению аддитивного белого гауссовского шума; смещению по времени; использованию эффекта реверберации; применению медианной фильтрации для разделения гармонических и ударных компонент сигнала. Дополнительно, для моделирования шумов, вызванных окружающей средой, использовался открытый набор аудиосигналов Urban Sound Dataset. Применение аугментации данных позволило увеличить размер обучающей выборки до 1,4 млн. речевых фрагментов.

Для проведения анализа робастности алгоритмов идентификации диктора подготовлен набор «Тест-Ш», состоящий из 1560 зашумленных речевых сигналов. Уровень зашумления определялся величиной отношения сигнал/шум с допустимым интервалом значений от 6 до 40 дБ. В Таблице 2 представлены результаты проведенного исследования с использованием оригинального и зашумленного набора. Сравнение предложенного алгоритма на базе сети X-Speech (МЧКК) выполняется с известными из литературы СНС на базе архитектур VGG-M, ResNet18, x-vectors. Обучение алгоритмов выполнено как на основе метода анализа спектрограмм (СП), так и с использованием МЧКК.

Таблица 2 – Анализ робастности алгоритмов идентификации диктора

-	**VGG-M (СП)**	**VGG-M (МЧКК)**	**ResNet18 (МЧКК)**	**x-vectors (СП)**	**x-vectors (МЧКК)**	**X-Speech (МЧКК)**
Тест	98,17%	98,24%	99,74%	91,34%	95,70%	98,37%
Тест-Ш	56,71%	77,08%	56,18%	86,52%	36,26%	91,54%

Результаты показывают, что алгоритм на базе предложенной архитектуры X-Speech (МЧКК) имеет высокую точность идентификации на исходном тестовом наборе данных «Тест» – 98%, уступая лишь решению на базе более сложной нейронной сети ResNet18. Результат на зашумленном тестовом наборе «Тест-Ш» показывает, что разработанный алгоритм лучше работает в условиях искажений и помех, чем исследуемые аналоги. Точность работы алгоритма в зашумленных условиях превышает 91%, что превосходит аналоги на 5% и более. Дополнительно

следует отметить, что алгоритм на базе предложенной сети X-Speech содержит в 10-20 раз меньше весовых параметров в сравнении с рассмотренными аналогами, что снижает его вычислительную сложность.

В третьей главе исследуются нейросетевые алгоритмы идентификации личности на основе анализа изображений лиц.

Для предварительного обнаружения (детектирования) лиц на изображениях использовался стандартный алгоритм на основе нейронной сети MTCNN (Multi-Task Cascaded Convolutional Networks). Видеоизображения из подготовленного набора FaceSpeechDB обрабатывались данным алгоритмом, что позволило подготовить набор из более чем 35 тыс. изображений лиц, разрешением 320x320 пикселей.

Для биометрической идентификации личности на основе анализа цифровых изображений лиц спроектирована архитектура нейронной сети CNN-Face. Ее структура показана на Рисунке 3.

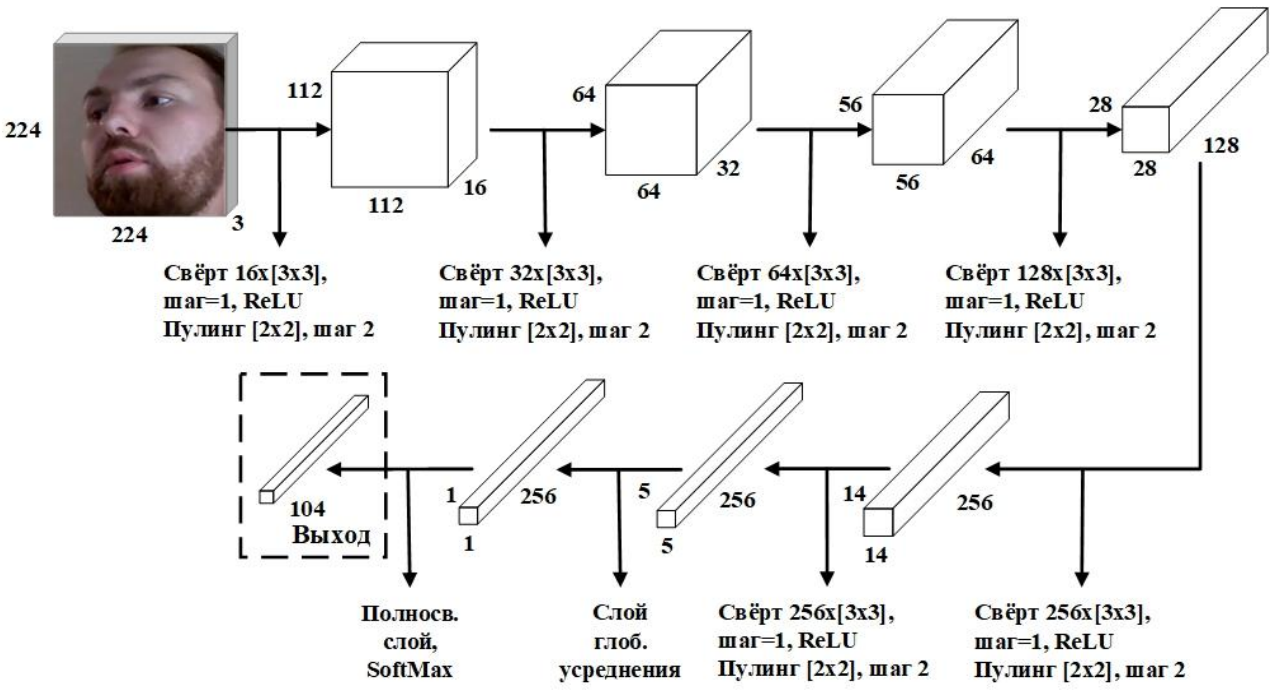

Рисунок 3 – Архитектура разработанной сети CNN-Face

Разработанная сеть состоит из последовательно идущих слоев свертки и пулинга. Общее количество проходов составляет шесть итераций. Далее выполняется операция глобального усреднения и используется полносвязный слой. На выходе сети формируется вектор вероятностей из 104 значений. Для предложенной нейронной сети количество обучаемых параметров (далее – КП) составляет 1 млн., что в 25-30 раз меньше наиболее часто используемых на текущем этапе развития нейросетевых аналогов.

В Таблице 3 представлен сравнительный анализ точности работы алгоритма на базе предложенной архитектуры CNN-Face со стандартными нейросетевыми решениями на базе архитектур VGG16, ResNet50 и SeNet50. Анализ результатов позволяет сделать вывод, что алгоритм на базе архитектуры CNN-Face, несмотря на существенно меньшее КП, также демонстрирует высокую точность работы.

Таблица 3 – Сравнительный анализ точности работы нейросетевых алгоритмов идентификации лиц

-	**КП**	**Обучение**	**Валидация**	**Тест**
VGG16	28 млн.	99,99%	99,86%	99,93%
ResNet50	25 млн.	99,78%	99,73%	99,87%
SeNet50	27 млн.	99,74%	99,41%	99,35%
CNN-Face	1 млн.	99,99%	99,93%	99,87%

В условиях появления пандемии Covid-19 появились новые технологические вызовы для систем распознавания лиц. В частности, обычная медицинская маска способна перекрывать до 70% площади лица. Возникает потребность в разработке робастных алгоритмов, способных работать в таких условиях. Для анализа работы алгоритмов идентификации лиц в ситуации использования медицинской маски, сформирована база изображений – «Тест-ММ», которая насчитывала 1560 примеров лиц с медицинской маской.

Анализ работы алгоритма на основе CNN-Face показал, что в условиях наличия медицинской маски точность распознавания снижается и составляет не более 74%. Для повышения робастности работы алгоритма предложена его модификация. Ее идея заключается в дополнительном анализе видимой области лица (лоб и линия глаз) в условиях наличия маски. При этом также сохраняется полноценный анализ всего детектируемого лица, поскольку алгоритм должен определять личность не только в ситуации наличия медицинской маски, но и в условиях ее отсутствия. В результате разработана архитектура сверточной нейронной сети – CNN-FaceMask, имеющая два входа и один выход, структура которой показана на Рисунке 4. Она состоит из двух эквивалентных модулей, каждый из которых представляет собой сеть архитектуры CNN-Face. Обучение осуществляется с использованием пар изображений – полноценного лица и его усеченной части.

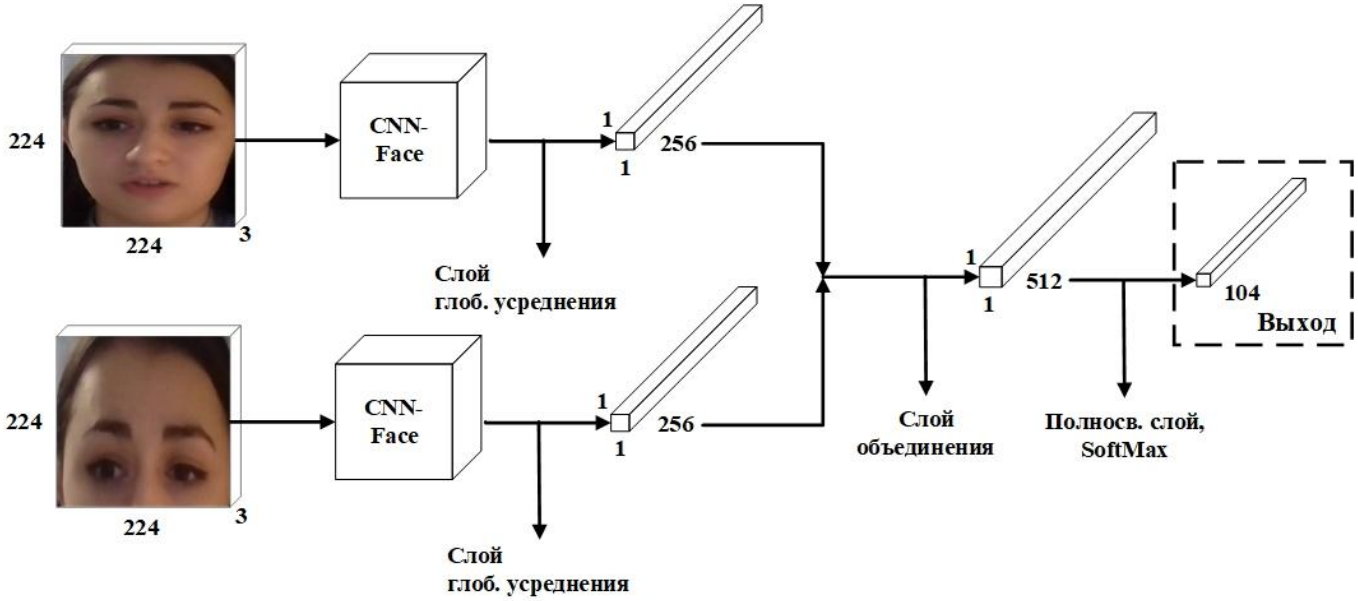

Рисунок 4 – Архитектура разработанной сети CNN-FaceMask

Результаты работы нейросетевого алгоритма на основе архитектуры CNN-FaceMask представлены в Таблице 4. Модификация нейросетевого алгоритма повысила точность работы разрабатываемого решения в условиях наличия маски на 20%. Уровень идентификации на тестовом наборе «Тест-ММ» составляет более 93%, что превосходит имеющиеся аналоги на 3% и более.

Таблица 4 – Анализ робастности работы алгоритмов
идентификации лиц в условиях использования медицинских масок

-	**VGG16**	**ResNet50**	**SeNet50**	**CNN-Face**	**CNN-FaceMask**
Тест	99,93%	99,87%	99,35%	99,87%	99,94%
Тест-ММ	90,23%	90,07%	77,49%	73,74%	93,10%

Также важно отметить, что дополнительный анализ области лба и линии глаз несущественно влияет на вычислительную сложность алгоритма. Для нейронной сети на базе CNN-FaceMask КП составляет 2 млн., что по-прежнему на порядок ниже, чем у алгоритмов, построенных на стандартных нейросетевых архитектурах VGG16, ResNet50 и SeNet50.

В четвертой главе выполняется разработка мультимодальных алгоритмов идентификации личности на основе анализа речевых сигналов и изображений лиц.

Для обучения алгоритмов используется набор FaceSpeechDB, обработанный с помощью детектора голосовой активности КДГА и детектора лиц MTCNN. С целью анализа работы алгоритмов подготовлены следующие наборы данных: тестовые изображения лиц «Тест-Л»; тестовые изображения лиц с медицинскими масками

«Тест-ММ»; тестовые речевые сигналы «Тест-Г»; тестовые речевые сигналы с добавлением искажений и шума «Тест-Ш». В результате сформированы четыре типа тестов: «Тест-Л, Тест-Г»; «Тест-Л, Тест-Ш»; «Тест-ММ, Тест-Г»; «Тест-ММ, Тест-Ш», каждый из которых насчитывает 1560 пар речевых сигналов и изображений лиц.

Разработка мультимодальных алгоритмов выполняется с применением двух типов объединения голосовой и лицевой биометрии: на уровне модуля принятия решений и на уровне слияния признаков. Первый тип объединения используется для комбинирования унимодальных алгоритмов на уровне принятия решений – МА-1. Модуль принятия решения строится на основе логических операций «ИЛИ» и «И» (далее – МА-1Д и МА-1К). В случае использования оператора «ИЛИ» (дизъюнкция) – если одна из модальностей определяет личность верно, то попытка классификации считается успешной. В ситуации использования оператора «И» (конъюнкция) – обе модальности должны идентифицировать пользователя верно, иначе попытка считается неудачной. Алгоритмы семейства МА-1 строятся с использованием решений на базе обученных моделей X-Speech и CNN-FaceMask.

Результаты тестирования мультимодальных алгоритмов представлены в Таблице 5. Установлено, что мультимодальный алгоритм МА-1Д лучше всего подходит для работы в ситуациях, когда определяющее значение имеют ошибки второго рода. Мультимодальный алгоритм МА-1К имеет высокую чувствительность к условиям эксплуатации. Тем не менее, выбор конъюнкции в качестве решающего правила является более надежным подходом, поскольку допуск инициализируется только в том случае, если пользователь верно классифицирован на основе анализа двух модальностей. В итоге достигается уменьшение ошибок первого рода.

Таблица 5 – Точность работы мультимодального алгоритма МА-1

-	**«Тест-Л, Тест-Г»**	**«Тест-Л, Тест-Ш»**	**«Тест-ММ, Тест-Г»**	**«Тест-ММ, Тест-Ш»**
МА-1Д	99,93%	99,87%	99,81%	99,10%
МА-1К	98,33%	91,53%	88,13%	82,04%

На следующем этапе выполняется разработка мультимодальных алгоритмов с применением метода объединения модальностей на уровне признаков (далее – МА-2). Первое решение представляет собой нейронную сеть, состоящую из 3-х модулей. На Рисунке 5 изображена структура мультимодального алгоритма МА-2.

На выходе модулей создаются векторы признаков, формируемые сверточными слоями. При создании признаков, описывающих лицо, предпочтение отдается подструктуре, анализирующей видимую часть лица в случае наличия медицинской маски. В результате анализа изображения лица и речевого сигнала формируются два вектора признаков одинаковой размерности. Далее они объединяются в общий вектор размерности 512. При данном способе конкатенации признаков влияние каждой из модальностей на результат классификации является равнозначным.

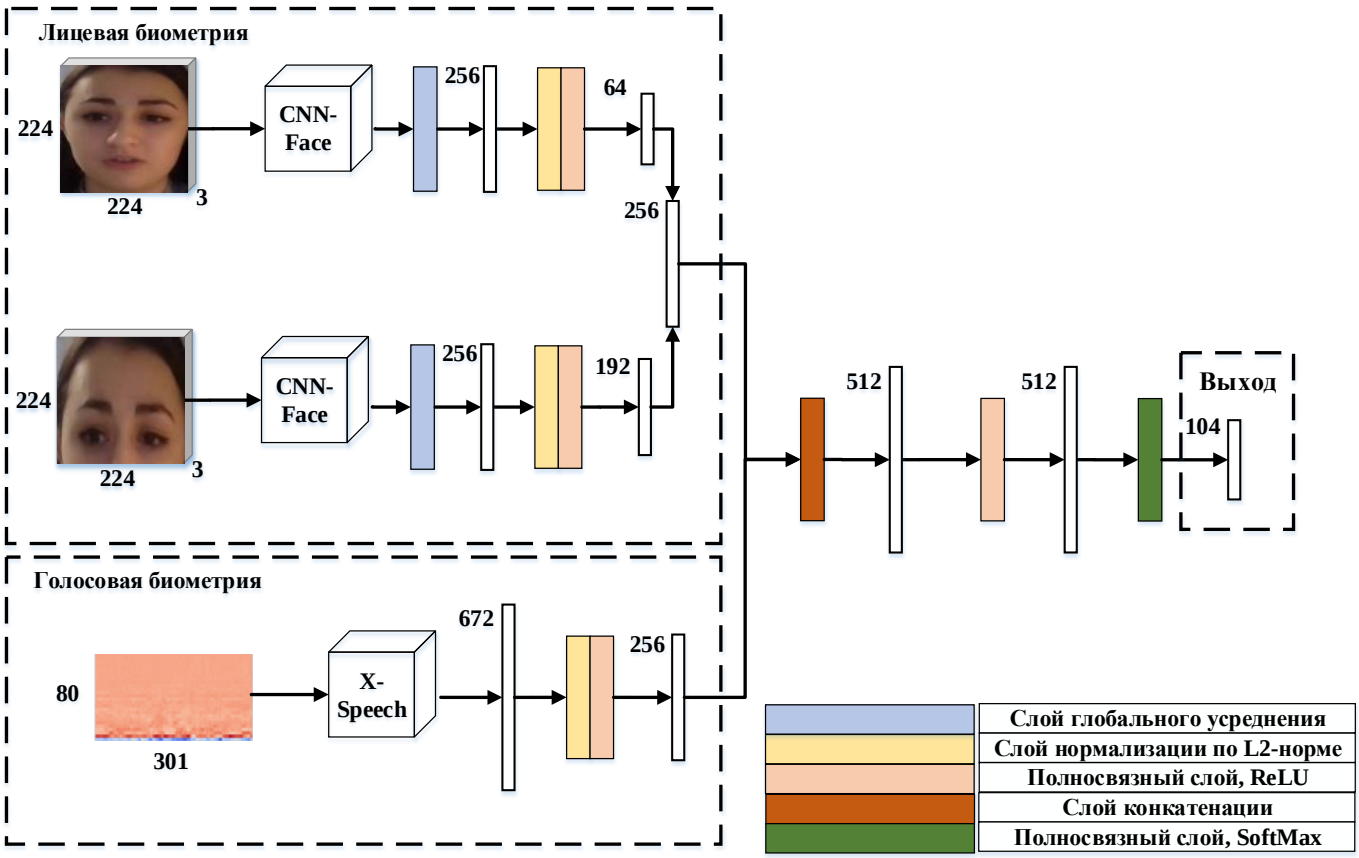

Рисунок 5 – Мультимодальный алгоритм МА-2

В Таблице 6 приводятся результаты тестирования мультимодального алгоритма МА-2 для четырех типов тестовых наборов мультимодальных данных.

Таблица 6 – Точность работы мультимодального алгоритма МА-2

-	**«Тест-Л, Тест-Г»**	**«Тест-Л, Тест-Ш»**	**«Тест-ММ, Тест-Г»**	**«Тест-ММ, Тест-Ш»**
МА-2	99,90%	98,43%	88,93%	88,28%

Из Таблицы 6 следует, что мультимодальный алгоритм МА-2 показывает лучшую робастность к появлению на лице медицинских масок и наличию помех в речевых сигналах по сравнению с алгоритмом МА-1К. В ситуации использования

медицинской маски выигрыш составляет в среднем до 1%, тогда как в условиях зашумления речевого сигнала преимущество увеличивается до 7%. На тестовом наборе «Тест-ММ, Тест-Ш», моделирующем более сложные условия эксплуатации мультимодальных биометрических систем, выигрыш алгоритма МА-2 составляет более 6%.

Важно отметить, что модуль сверточных слоев, анализирующий всю область лица, вносит меньший вклад в общий вектор признаковых параметров. Однако появление медицинской маски снижает точность классификации личности. Для того чтобы обойти данное ограничение разработана модификация алгоритма МА-2 (далее – МА-2М). Мультимодальный алгоритм МА-2М выполняет обработку исключительно видимой области лица. Анализ изображения осуществляется вне зависимости от наличия или отсутствия медицинской маски, то есть выполняется обработка области лба, глаз и части носа. Подструктура анализа речевых сигналов остается без изменений.

В Таблице 7 представлены результаты сравнения работы мультимодальных алгоритмов на основе слияния признаков, а также алгоритма МА-1К. Алгоритм МА-1Д не рассматривается, поскольку с точки зрения надежности он более уязвим, т.к. подвержен ошибкам ложной классификации.

Таблица 7 – Результаты работы мультимодальных алгоритмов

-	**«Тест-Л, Тест-Г»**	**«Тест-Л, Тест-Ш»**	**«Тест-ММ, Тест-Г»**	**«Тест-ММ, Тест-Ш»**
МА-1К	98,33%	91,53%	88,13%	82,04%
МА-2	**99,90%**	**98,43%**	88,93%	88,28%
МА-2М	99,80%	96,81%	**95,12%**	**94,68%**

Анализ результатов показал, что подходы на основе объединения модальностей на уровне слияния признаков лучше справляются с задачей идентификации. Так, алгоритм МА-2 лучше всего подходит для работы в условиях низкой зашумленности аудио- и видеоканала. Такое решение может также применяться в условиях наличия шумов в речевом сигнале, поскольку в этом случае деградация в точности работы составляет в среднем 1-2%. Алгоритм МА-2М лучше всего подходит для работы в условиях присутствия на лице медицинской маски. Он демонстрирует точность на уровне 95%, что превосходит результаты работы других мультимодальных аналогов на 6% и более. Особое внимание заслуживает результат

работы алгоритма МА-2М в условиях теста «Тест-ММ, Тест-Ш». Несмотря на высокую сложность проводимого эксперимента, деградация в точности результатов составляет менее 5%.

Также проведен сравнительный анализ рассмотренных уни- и мультимодальных алгоритмов. В ситуации зашумления речевых сигналов преимущество в точности работы мультимодального алгоритма МА-2 составляет в среднем 7% и более, а в условиях использования медицинской маски выигрыш решения на базе МА-2М составляет не менее 2% относительно унимодальных аналогов.

Таким образом, показано, что алгоритм объединения модальностей на уровне слияния признаков позволяет повысить точность и робастность идентификации по сравнению не только с алгоритмом объединения модальностей на уровне принятия решения, но и относительно рассмотренных унимодальных аналогов.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. Созданы русскоязычные базы аудиовизуальных данных, нейросетевые алгоритмы идентификации и зарегистрированное программное обеспечение для построения мультимодальных биометрических систем.
2. Собраны аудиовизуальный набор данных FaceSpeechDB, содержащий 60 часов русскоязычной записи 104 человек, и набор аудиосигналов VADSpeakersDB из 138 000 фрагментов.
3. Предложенный детектор голосовой активности улучшает качество речевых сигналов за счет фильтрации пауз, эффектов глотации, вдохов и шумов. Точность определения фрагментов голосовой активности при его использовании повышается на 2-3% в сравнении с имеющимися аналогами.
4. Разработанный нейросетевой алгоритм на основе x-векторной системы может использоваться для автоматической идентификации диктора. При сохранении точности идентификации на уровне 98-99%, он имеет в 10-20 раз меньше обучаемых параметров, что снижает его вычислительную сложность. В условиях воздействия шумов и помех он превосходит аналоги на 5% и более.
5. Разработанный алгоритм на базе предложенной нейросетевой архитектуры CNN-Face может использоваться в задачах лицевой биометрии. При сохранении точности идентификации на уровне до 99%, он содержит в 25-30 раз меньше обучаемых параметров, что снижает его вычислительную сложность. Его

модификация CNN-FaceMask показывает наилучшую в рассматриваемом классе робастность к присутствию медицинской маски на лице человека. В условиях наличия маски он превосходит стандартные алгоритмы на 3% и более.

6. Предложенный мультимодальный алгоритм МА-2 подходит для работы в условиях низкого зашумления канала связи. Также он показывает хорошую робастность к искажениям в аудиосигналах, поскольку деградация в точности составляет в среднем 1-2%.
7. Предложенный мультимодальный алгоритм МА-2М подходит для работы в условиях перекрытия лица медицинской маской. В данных условиях точность работы алгоритма на тестовом наборе данных определяется на уровне 95%, при этом показатель деградации составляет менее 5%.
8. Разработанные мультимодальные алгоритмы имеют преимущество в точности относительно унимодальных аналогов на 7% и более при зашумлении речевых сигналов, на 2% и более в условиях перекрытия части лица медицинской маской.
9. Цель и задачи диссертационной работы успешно выполнены. Результаты работы могут использоваться при создании мультимодальных биометрических систем, работающих в режиме реального времени, в том числе в задачах прокторинга при использовании ВКС и при построении СКУД повышенной надежности.

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в изданиях, рекомендуемых ВАК

1. Хрящев В.В., Приоров А.Л., Стефаниди А.Ф., Топников А.И. Разработка и исследование алгоритмов обработки и распознавания речевых сигналов и изображений для систем мультимодальной биометрии // Цифровая обработка сигналов. – 2017. – №3. – С. 45-49.
2. Стефаниди А.Ф., Приоров А.Л., Топников А.И., Хрящев В.В. Применение сверточных нейронных сетей в задаче мультимодальной идентификации // Цифровая обработка сигналов. – 2020. – №2. – С. 52-58.
3. Стефаниди А.Ф., Приоров А.Л., Топников А.И., Хрящев В.В. Модификация VGG-архитектуры в задачах унимодальной и мультимодальной биометрии // Цифровая обработка сигналов. – 2020. – №3. – С. 35-40.

Публикации, индексируемые в Scopus

4. Khryashchev V., Topnikov A., Stefanidi A., Priorov A. Bimodal person identification using voice data and face images // International Conference on Machine Vision (ICMV 2018), – SPIE, 2019. – vol. 11041. – pp. 296-303.

5. Stefanidi A., Topnikov A., Tupitsin G., Priorov A. Application of convolutional neural networks for multimodal identification task // Conference of Open Innovation Association (FRUCT). – IEEE, 2020. – pp. 423-428.
6. Stefanidi A., Topnikov A., Priorov A, Kosterin I. Modification of VGG Neural Network Architecture for Unimodal and Multimodal Biometrics // East-West Design & Test Symposium (EWDTS). – IEEE, 2020. – pp. 1-4.

Доклады на российских и международных конференциях

7. Стефаниди А.Ф., Лебедев А.А., Хрящев В.В., А.М. Шемяков. Разработка и исследование алгоритмов обработки и распознавания речевых сигналов и видеоизображений для систем мультимодальной биометрии // Перспективные технологии в средствах передачи информации (ПТСПИ-2017): докл. 12-й междунар. науч-техн. конф. – Суздаль, 2017. – Т. 1. – С. 174-177.
8. Хрящев В.В., Приоров А.Л., Стефаниди А.Ф., Степанова О.А. Разработка алгоритмов обработки цифровых сигналов и изображений для систем мультимодальной биометрии // Радиоэлектронные средства получения, обработки и визуализации информации (РСПОВИ-2017): докл. 7-й всерос. конф. – Москва, 2017. – С. 155-160.
9. Стефаниди А.Ф., Лебедев А.А., Матвеев Д.В. Исследование робастности алгоритмов распознавания лиц на изображениях // Цифровая обработка сигналов и ее применение (DSPA-2018): докл. 20-й междунар. конф. – Москва, 2018. – Т. 2. – С. 821-826.
10. Стефаниди А.Ф., Топников А.И., Приоров А.Л. Использование сверточных нейронных сетей в задаче распознавания диктора // Цифровая обработка сигналов и ее применение (DSPA-2020): докл. 22-й междунар. конф. – Москва, 2020. – С. 642-646.
11. Стефаниди А.Ф., Топников А.И., Приоров А.Л. Бимодальная идентификация личности на основе лицевой и речевой биометрии // Новые информационные технологии и системы: докл. 17-й междунар. конф. – Пенза, 2020. – С. 125-129.
12. Стефаниди А.Ф., Топников А.И., Приоров А.Л. Модификация нейросетевой VGG-архитектуры в задаче мультимодальной идентификации личности // Цифровая обработка сигналов и ее применение (DSPA-2021): докл. 23-й междунар. конф. – Москва, 2021. – С. 243-247.
13. Стефаниди А.Ф. Разработка алгоритма обнаружения голосовой активности в задаче мультимодальной идентификации личности // Новые информационные технологии и системы: докл. 18-й междунар. конф. – Пенза, 2021. – С. 145-150.
14. Сенников А.В., Стефаниди А.Ф. Разработка алгоритма детектирования средств индивидуальной защиты на видеоданных // Новые информационные технологии и системы: докл. 18-й междунар. конф. – Пенза, 2021. – С. 150-155.

15. Сенников А.В., Стефаниди А.Ф., Назаровский А.Е. Разработка алгоритма детектирования средств индивидуальной защиты на видеоданных // Проблемы информатики в образовании, управлении, экономике и технике: докл. 21-й междунар. науч.-техн. конф. – Пенза, 2021. – С. 56-63.
16. Стефаниди А.Ф. Применение методов цифровой обработки речевых сигналов и изображений для построения мультимодальных алгоритмов биометрической идентификации // Радиоэлектронные устройства и системы для инфокоммуникационных технологий (REDS-2022): докл. 77-й всерос. конф. (с междунар. участием) – Москва, 2022. 5 с.

Свидетельства о государственной регистрации

17. Стефаниди А.Ф., Топников А.И. Bimodal Human Identification 1.0 – программа для мультимодальной идентификации человека по голосу и лицу с помощью цифровых изображений и аудиосигналов // Свидетельство о государственной регистрации программы для ЭВМ № 2019613092 от 7 марта 2019.
18. Стефаниди А.Ф. Multimodal Identification ToolKit – программа для мультимодальной идентификации личности на основе голосовой и лицевой биометрии // Свидетельство о государственной регистрации программы для ЭВМ № 2021663249 от 13 августа 2021.
19. Стефаниди А.Ф. VoiceActivityDetector 1.0 – программа для анализа голосовой активности в задаче мультимодальной идентификации личности // Свидетельство о государственной регистрации программы для ЭВМ № 2021681283 от 20 декабря 2021.

Стефаниди Антон Федорович

ИССЛЕДОВАНИЕ МУЛЬТИМОДАЛЬНЫХ АЛГОРИТМОВ БИОМЕТРИЧЕСКОЙ ИДЕНТИФИКАЦИИ НА ОСНОВЕ МЕТОДОВ ЦИФРОВОЙ ОБРАБОТКИ РЕЧЕВЫХ СИГНАЛОВ И ИЗОБРАЖЕНИЙ

Автореферат диссертации на соискание ученой степени кандидата технических наук

Подписано в печать 23.06.2022.
Формат 60x84 1/16. Усл. печ. л. 1. Тираж 100 экз.
ИП Грязнухин Р.А. 150000, Ярославль, ул. Б. Октябрьская, 37/1.