ЯРОСЛАВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
им. П.Г. ДЕМИДОВА

На правах рукописи

Стефаниди Антон Федорович

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

Специальность 2.2.13
Радиотехника, в том числе системы и устройства телевидения

ДИССЕРТАЦИЯ
на соискание ученой степени
кандидата технических наук

Научный руководитель:
кандидат технических наук, доцент
Хрящев Владимир Вячеславович

Ярославль – 2022

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность темы. Методы и алгоритмы цифровой обработки сигналов и изображений широко используются при построении радиотехнического и телевизионного оборудования, проектировании систем управления, создании систем хранения, поиска и сжатия информации [1-9]. Особый интерес для различных практических приложений представляют цифровые алгоритмы анализа изображений и речевых сигналов. Если человек является объектом наблюдения, то его можно идентифицировать с помощью анализа таких цифровых сигналов, как оцифрованный отпечаток пальца, фотография лица или запись речи. Системы распознавания личности на основе анализа уникальных физиологических и поведенческих черт индивида носят название биометрических [10, 18, 25, 40, 90].

Системы биометрической идентификации личности стали неотъемлемой частью нашей повседневной жизни. Можно отметить, что сейчас большинство современных мобильных телефонов оборудованы сканерами отпечатков пальцев или используют встроенную камеру для аутентификации пользователя по лицу [68, 77, 80, 87]. Одним из актуальных приложений биометрии является идентификация человека в сеансе видеоконференцсвязи (ВКС). В частности, к ним относится задача прокторинга – процедура наблюдения и контроля за дистанционным испытанием (от англ. «proctor» – человек, который следит за ходом экзамена в университете). Такая технология уже давно используется 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 повышенной секретности [1, 12, 13].

В основе работы методов идентификации личности лежит анализ биометрических параметров человека. В частности, это может быть цифровое изображение отпечатка пальца, лица или сетчатки глаза [65, 80, 82]. Запись речевого сигнала также является цифровым «слепком» личности,

однозначно ее определяющим. Биометрические методы аутентификации получили массовое распространение, поскольку, в отличие от паролей и аппаратных ключей, физиологические особенности невозможно потерять или забыть. Однако следует отметить, что в перечисленных выше актуальных примерах использования данной технологии (прокторинг при ВКС и биометрия при построении СКУД) качество речевого сигнала и изображения лица может быть существенно неидеальным, ввиду наличия ряда искажающих факторов [91].

Системы идентификации диктора чувствительны к эффектам, возникающим в процессе передачи и обработки данных, физиологическим особенностям говорящего, акустическим свойствам окружающей среды [38, 96]. Методы распознавания пользователя по лицу имеют сильную зависимость от уровня освещенности, ракурса, качества фоторегистратора, а также чувствительны к возрастным изменениям и мимике [68, 91, 95]. Системы идентификации личности на основе анализа одного биометрического параметра (унимодальные) можно обойти в случае создания цифровой копии лица или голоса человека. В результате возникает потребность в разработке более совершенных алгоритмов идентификации [90-93].

Одним из перспективных направлений развития биометрических систем является разработка и исследование алгоритмов идентификации личности на основе двух и более биометрических параметров, так называемые мультимодальные решения. Подход на основе комбинирования модальностей позволяет не только повысить устойчивость и точность работы биометрических систем, но и улучшить надежность работы при попытках несанкционированного доступа [70, 84, 88-94, 97, 98].

Анализ научно-технической литературы показывает, что наиболее эффективным подходом для автоматического распознавания образов является использование алгоритмов глубокого обучения [23, 26, 30, 31]. Так, сверточные нейронные сети (СНС) стали одним из главных инструментов

анализа изображений в области построения систем прикладного телевидения (СПТ). Особое место они занимают в задачах биометрической идентификации на основе анализа голоса и лица человека [54, 55, 68, 83, 89, 102]. Особенность данного подхода заключается в том, что признаки (дескрипторы) исследуемых объектов формируются автоматически в процессе обучения. Операция свертки является основным структурным блоком для сетей данного типа. Сгенерированные таким образом дескрипторы позволяют, как правило, добиться лучших результатов в задачах обнаружения, сегментации и распознавания объектов на цифровых изображениях, а также при идентификации диктора на основе анализа речевых сигналов [101-103].

Важнейший вклад в развитие области обработки визуальной информации и построения систем биометрической идентификации по лицу 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, 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 общую 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 области 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, Х. Галбалли и др.

Таким образом, можно сделать вывод о том, что разработка новых нейросетевых алгоритмов идентификации личности для мультимодальных

биометрических систем является актуальной научно-технической задачей и несет 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 изображений и речевых сигналов, работающих в режиме реального времени, в том числе в задачах прокторинга при ВКС и при построении 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 ICMV-2018 (International Conference on Machine Vision), 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-26», 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.

Благодарности. Автор выражает искреннюю признательность кафедре цифровых технологий и машинного обучения Ярославского государственного университета им. П.Г. Демидова, в особенности научному руководителю, кандидату технических наук Хрящеву Владимиру Вячеславовичу, и доктору технических наук Приорову Андрею Леонидовичу. Особая благодарность кандидату технических наук Топникову Артему Игоревичу за поддержку на всех этапах исследования и активное участие в формирование научного направления диссертационной работы.

Отдельно хотелось бы выразить слова благодарности родным и близким за возможность заниматься научной деятельностью и поддержку во время написания данной работы.

ГЛАВА 1
ТЕКУЩЕЕ СОСТОЯНИЕ ОБЛАСТИ ИССЛЕДОВАНИЙ

1.1 Вводные замечания

Системы биометрической идентификации личности все чаще используются в повседневной жизни людей. Среди возможных биометрических признаков наибольший интерес для анализа и усовершенствования соответствующих алгоритмов представляют лицо и голос человека [1, 11-13].

Алгоритмы распознавания личности на основе анализа изображений лиц обладают рядом преимуществ относительно других биометрических подходов [1, 33]. При использовании таких систем не нужен физический контакт с регистрирующими устройствами, достаточно просто пройти мимо или остановиться на небольшой промежуток времени вблизи камеры или фоторегистратора. Это отличает подобные системы от, например, систем идентификации по радужной оболочке глаза или отпечаткам пальцев, которые предъявляют жесткие требования к процедуре взятия биометрических признаков [11].

К недостаткам процедуры распознавания человека по изображению лица можно отнести сильную зависимость от степени освещенности и угла поворота головы. Качество оптического устройства также влияет на точность работы такого рода биометрических систем [10, 11, 33]. Это особенно важно при мониторинге в местах массового скопления людей, таких как стадионы, метро и аэропорты, где расстояние от видеокамеры до людей может измеряться десятками метров. Также алгоритмы распознавания лиц чувствительны к возрастным изменениям. Со временем человек может изменить прическу, могут появиться борода или усы, а также очки, что в итоге усложняет задачу определения личности [104, 107]. В условиях пандемии Covid-19 возникают новые вызовы для систем распознавания лиц.

В частности, обычная медицинская маска способна перекрывать до 70% лица. Существенная часть информации, описывающая исключительные свойства лица, такие как губы, нос и подбородок, остается под маской [113, 114]. Вследствие этого возникает потребность в разработке робастных алгоритмов идентификации личности на основе анализа лиц, способных работать в реальных практических условиях. Потребность в такого рода алгоритмах остается на высочайшем уровне [33, 99, 100, 118].

Системы анализа голоса и распознавания дикторов приобретают все большую массовость и популярность по мере развития речевых технологий [15, 41]. Существует постоянно растущая потребность в приложениях для поиска и распознавания аудиоматериалов, голосовых помощниках – Siri, Google Assistant, Яндекс Алиса и др. Потребность в таких потребительских решениях дает качественный скачок для всей индустрии речевых технологий [13, 14].

Методы распознавания личности по голосу используются не только в задачах контроля и управления доступом, но также являются важным инструментом для борьбы с телефонным терроризмом и в криминалистике в целом [13]. Из-за массовой доступности мобильные устройства стали не только средством связи, но и способом мошенничества для преступников. Голос, записанный как часть доказательства виновности, может являться важной уликой для правоохранительных органов. Однако преступники целенаправленно стараются изменить свой голос, который может быть замаскирован или искажен. Это существенно усложняет идентификацию личности в задачах криминалистического анализа. Согласно принятой в этой области классификации процедура определения личности по голосу может осуществляться [13]:

- неподготовленным человеком;
- экспертом в области криминалистического анализа;
- системами автоматической идентификации диктора.

В настоящее время активно ведется разработка новых методов и алгоритмов автоматического распознавания диктора [40-47]. Качество и точность работы таких систем непрерывно растет и уже сопоставимо со способностью человека воспринимать и различать звуки [13, 14].

Однако, несмотря на широкое распространение методов голосовой биометрии, системы распознавания диктора обладают рядом недостатков, в частности, зависимостью от эффектов канала передачи информации и микрофона, физиологических особенностей говорящего, акустических свойств окружающей среды [43, 78]. Алгоритм идентификации может столкнуться с проблемой, когда регистрация пользователей производится в близких к идеальным условиям, а тестирование и эксплуатация устройства происходит в зашумленной среде. Отсутствие возможности контроля внешних факторов и несоблюдение правил сбора биометрических данных может существенно снизить точность работы такой системы [14, 72].

1.2 Классификация задач распознавания личности

Различают два типа задач распознавания личности – идентификация и верификация (аутентификация). На этапе регистрации пользователей, например в системе контроля и управления доступом, различия между этими задачами отсутствуют. Данные пользователей вносят в базу, где для каждого из них формируется унифицированный объект, полученный в результате анализа его индивидуальных биометрических параметров – изображения лица, образца речевого сигнала (фонограмма) или отпечатков пальцев. В качестве объекта выступает набор параметров, например, вектор признаков или цифровая модель. Так, модели гауссовых смесей популярны для использования в задачах распознавания диктора [68, 69].

После того, как база пользователей сформирована, переходят к этапу эксплуатации системы. На этапе проверки потенциального пользователя система осуществляет считывание биометрических данных, и на их основе формирует тестовый объект. Он анализируется и сравнивается с объектами

из базы. Степень схожести определяется с помощью метрик правдоподобия [13].

При решении задачи идентификации требуется определить личность из ограниченного набора зарегистрированных в системе людей. Сравнение проводится по принципу «один ко многим». В общем случае результатом такого процесса является вывод кандидата, эталонный объект которого по своим параметрам и свойствам более всего схож с тестовым объектом.

Представленная на Рисунке 1.1 схема иллюстрирует принцип построения биометрической системы идентификации личности на основе анализа лиц. Стоит отметить, что после процедуры обнаружения лица на изображении, выполняются этапы предобработки и выделения признаков. Особенности данных этапов зависят от типа используемого классификатора. Как правило, предварительная обработка обнаруженного в видеопотоке лица позволяет улучшить точность процедуры идентификации.

Рисунок 1.1 – Схема биометрической системы для задачи идентификации по лицу

При решении задачи верификации (проверки подлинности) система обладает информацией о том, в качестве какой личности потенциальный

пользователь планирует пройти аутентификацию. Сопоставление проводится по принципу «один к одному». По сути, тестовый объект сравнивается с эталонным объектом заявленного пользователя, хранимым в базе. В результате принимается положительное либо отрицательное решение об их соответствии (Рисунок 1.2) [13].

Рисунок 1.2 – Схема биометрической системы для задачи верификации

Любая система распознавания личности определяется одним из режимов работы: работой на закрытом или открытом множестве. В первом случае все потенциальные пользователи известны системе (закрытый сценарий). В случае, если условия эксплуатации подразумевают проверку пользователей, которые не зарегистрированы в системе, то говорят о распознавании на открытом множестве. При такой постановке задачи система должна инициализировать отказ неизвестным людям [13].

Задачи распознавания человека (диктора) по голосу аналогичны задачам анализа по лицу, то есть определяются по типу распознавания и режиму работы. На Рисунке 1.3 изображена общая схема системы голосовой биометрии или системы распознавания диктора. В случае необходимости в схему работы системы голосовой биометрии может быть включен блок диаризации (разделения говорящих). Под ней понимается процесс разделения входного аудиосигнала на однородные сегменты в соответствии с принадлежностью к конкретному диктору. Данный блок необходим при

использовании в ситуациях, когда входной аудиосигнал содержит речь двух и более дикторов. В частности, это может быть в условиях записи интервью, телефонного разговора или сеанса видеоконференцсвязи [16, 17].

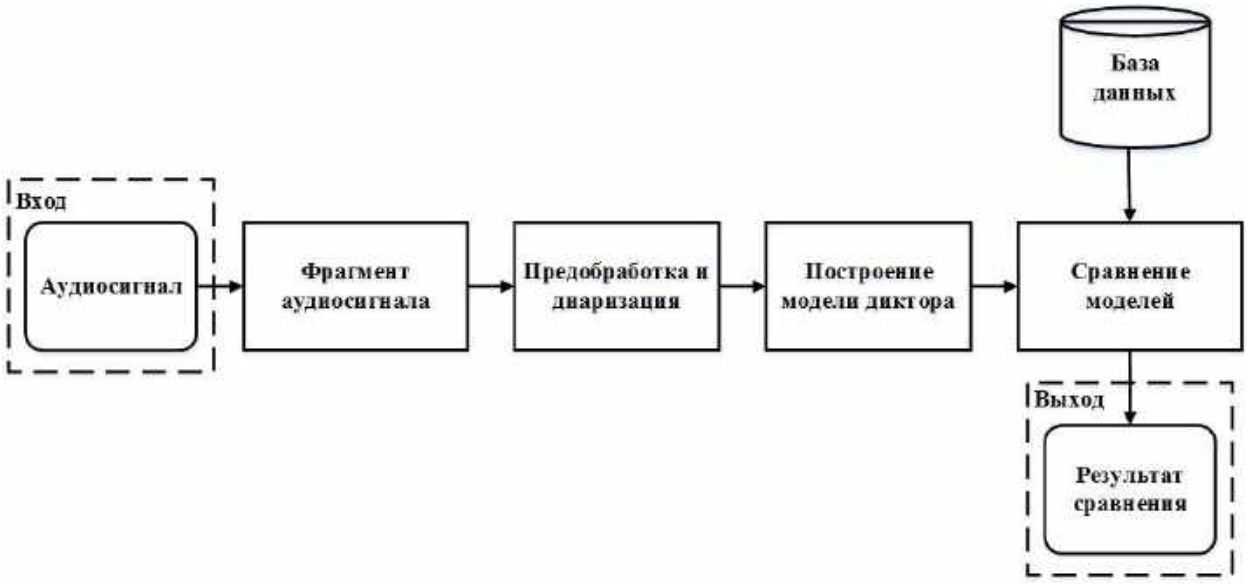


Рисунок 1.3 – Схема системы распознавания диктора

Анализ научно-технической литературы показывает, что наиболее эффективным подходом для решения задач идентификации и верификации является использование алгоритмов глубокого обучения [21, 22, 23, 24, 26]. Сверточные нейронные сети стали одним из главных инструментов анализа голоса и лица человека [33, 40, 43, 58, 64, 116]. Рассмотрим их построение более подробно.

1.3 Сверточные нейронные сети

В 1989 г. СНС предложены французским ученым Яном Лекуном [19]. С помощью таких моделей сегодня решается огромное количество практических задач. В частности, методы и алгоритмы на основе СНС показывают высокие результаты идентификации личности с использованием изображений лица и речевых сигналов. Также они широко используются в задачах обработки и анализа текста, в медицине, биохимических

исследованиях, робототехнике и др. Алгоритмы такого рода относятся к классу «глубокого обучения», поскольку архитектура таких нейросетевых систем является многосвязной и иерархичной, а сам процесс обучения занимает длительное время [18-20].

Присутствие в названии определения термина «сверточная» говорит о том, что при построении таких сетей используется соответствующая математическая операция. Свертка является операцией на двух функциях вещественного аргумента. В терминологии нейронных сетей сигнал x, например, представляет входной сигнал, а вторая функция w является ядром. При их взаимодействии формируется выходной сигнал s или карта свойств. Можно записать выражение для операции дискретной свертки следующим образом:

$$s(t) = (x * w)(t) = \sum_{a=-\infty}^{\infty} x(a) w(t-a).$$

В глубоком обучении вход обычно представляет собой многомерный массив данных, а ядро – многомерный массив параметров, которые адаптируются в процессе обучения. Выход представляет собой карту признаков. Так как цифровое изображение является двумерным массивом, то и ядро также должно быть двумерным. В итоге двумерная свертка имеет вид [18, 19]:

$$S(i,j) = (I * K)(i,j) = \sum_{m} \sum_{n} I(m,n) K(i-m, j-n),$$

где K – двумерное ядро, а I – входное изображение.

На Рисунке 1.4 представлен пример операции двумерной свертки, где вход представлен некоторой областью изображения.

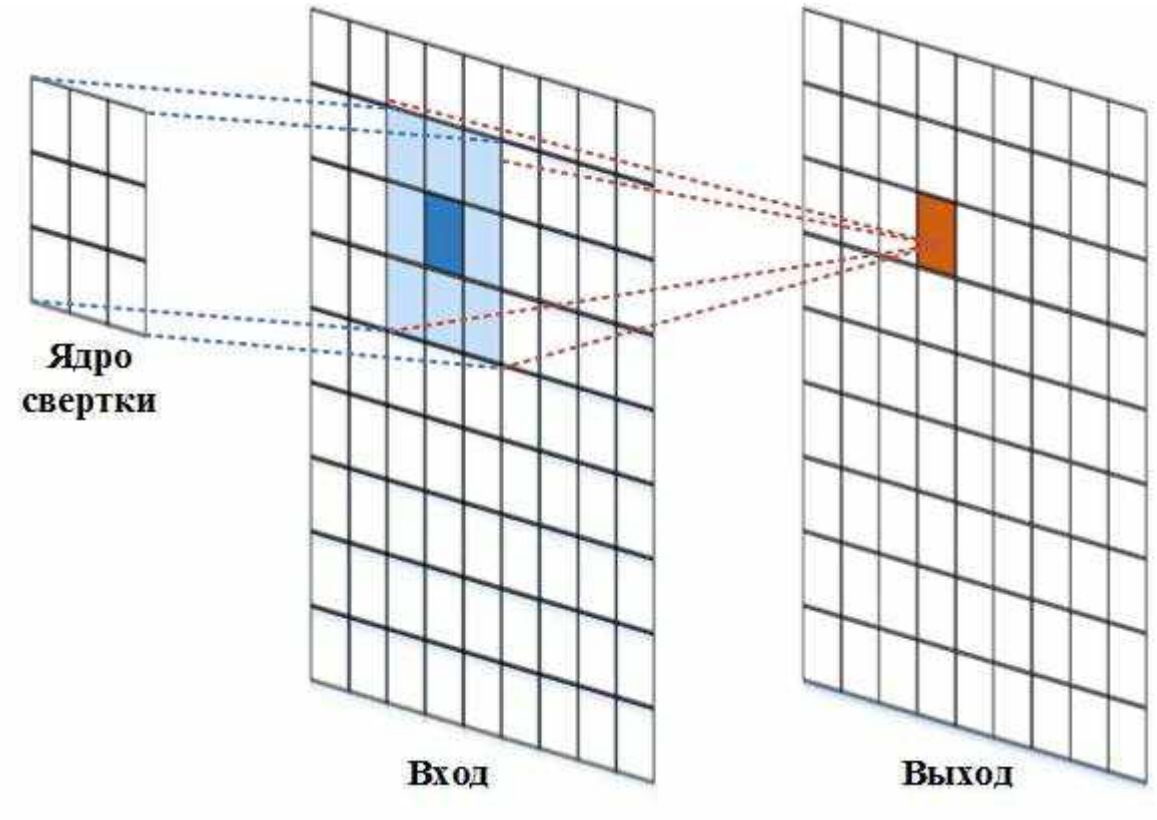


Рисунок 1.4 – Иллюстрация операции двумерной свертки

Одним из важнейших элементов СНС является функция активации. Она определяет, будет ли нейрон активирован вследствие входного воздействия. В случае активации нейрона сигнал продолжает свое движение в направлении более глубоких слоев. В сверточных слоях весовые параметры определяются ядром. Адаптация весовых параметров осуществляется в процессе обучения нейронной сети. На Рисунке 1.5 представлен нейрон сверточного слоя, где ядро свертки имеет размер 2×2 [21, 22]. Математически активацию нейрона можно описать следующим образом:

$$z = \sum_{i=0}^{n} \omega_i x_i,$$

$$\hat{y} = f(z),$$

где z – взвешенная сумма входов, $f(z)$ – функция активации, $\hat{y}$ – результат активации нейрона.

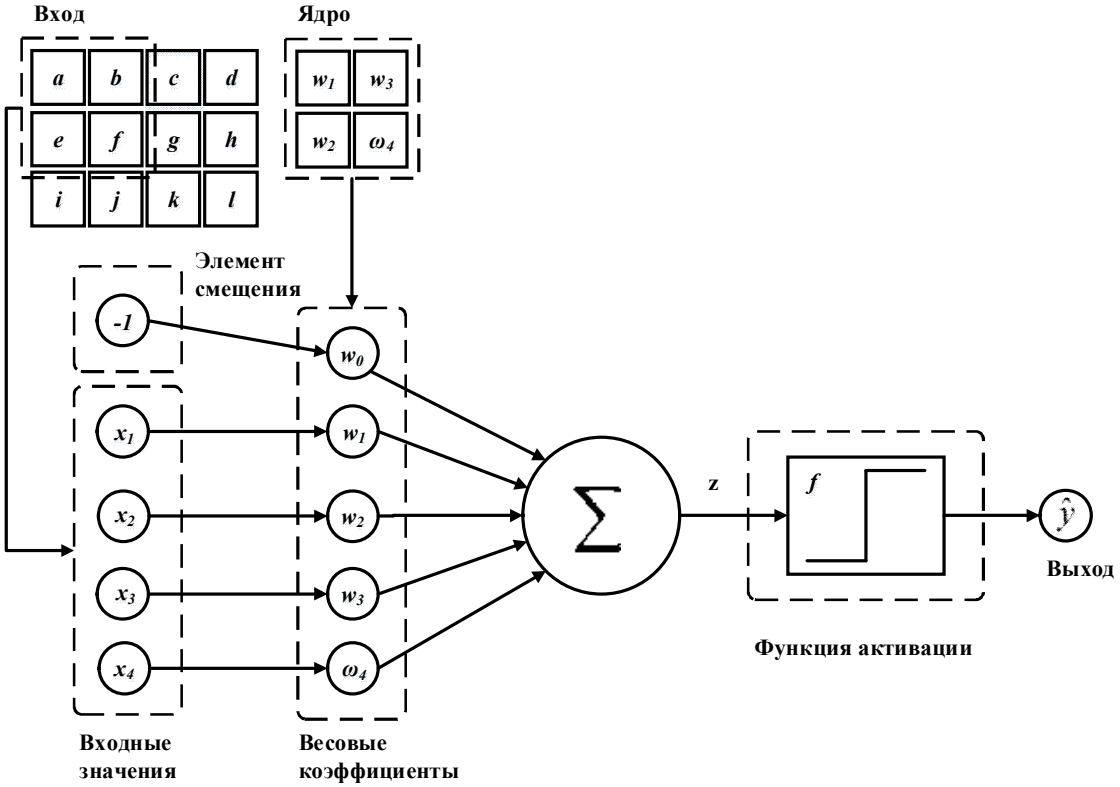


Рисунок 1.5 – Нейрон сверточного слоя с функцией активации

В настоящее время наиболее часто используемой функцией активации является блок линейной ректификации (Rectified Linear Unit, ReLU), а также ряд её модификаций. Такая функция описывается следующим образом [22]:

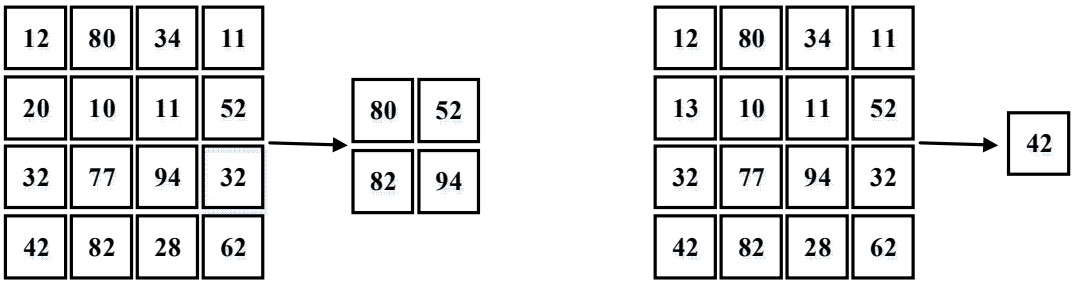
$$f(z) = \max(0, z).$$

Функция ReLU обладает рядом преимуществ относительно своих аналогов. В процессе обучения сети и адаптации весовых параметров быстро считается производная от данной функции, которая равна 0 для отрицательных значений и 1 для положительных. Поэтому производная блока линейной ректификации остается большой всюду, где блок активен. Градиенты не только велики, но еще и согласованы. Также, в отличие от сигмоидной и гиперболической функций, ReLU позволяет активировать часть нейронов. В результате слои становятся разряженными, что снижает вычислительную нагрузку в процессе обучения. Кроме того, в результате применения блока линейной ректификации увеличивается скорость сходимости градиентных методов [21, 22, 27].

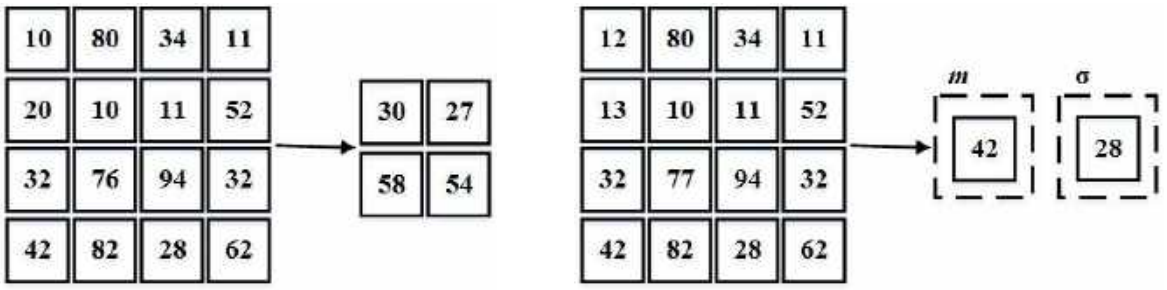
Ещё одним важным элементом глубоких нейронных сетей является операция субдискретизации или пулинга (down sampling, pooling layer). Пулинг с помощью обобщения выделяемых признаков и их уплотнения позволяет сформировать инвариантное представление входных данных относительно малых переносов входа. Некоторая часть информации теряется, но при этом сокращается размерность. Операция пулинга позволяет уменьшить вычислительную сложность СНС за счет прореживания карт признаков [21, 22, 27-29].

Частным случаем субдискретизации является операция уменьшения размерности с выбором максимального значения (max-pooling). Данная операция представляет собой нелинейное уплотнение признаков. Из области карты свойств, например, размером 2×2, выбирается максимальное значение. Далее операция повторяется для всей карты признаков с фиксированным шагом [28, 30].

Существуют и другие виды операции пулинга: глобальное усреднение (global average pooling); локальное усреднение (average pooling); статистический пулинг (statistics pooling). Последний подход вычисляет математическое ожидание и среднеквадратическое отклонение для всей карты признаков [27-30]. Иллюстрации таких операций приведены на Рисунке 1.6.



а) б)



в) г)

Рисунок 1.6 – Примеры субдискретизации: а) с выбором максимального значения с шагом 2; б) глобальное усреднение; в) локальное усреднение с шагом 2; г) статистический пулинг

Количество сверточных слоев теоретически быть неограниченным, однако чем их больше, тем требовательнее СНС становится к вычислительным ресурсам. На выходе каждого слоя формируется тензор, внутри которого содержатся параметры изучаемого объекта. Эти параметры являются абстрактным представлением входных данных [20, 32]. На Рисунке 1.7 изображена структурная схема сверточного слоя классической СНС.

Рисунок 1.7 – Структурная схема сверточного слоя СНС

На выходе классической СНС используются полносвязные слои. Их может быть несколько. Последний из них образует N-мерный вектор, где N – число исследуемых объектов. В случае биометрической идентификации это количество уникальных личностей. Выходной слой имеет нейроны softmax-группы в качестве функции активации. Она имеет следующий вид для *i*-го нейрона:

$$y_i = \frac{e^{z_i}}{\sum_{i=1}^{N} e^{Z_i}} .$$

На выходе СНС каждое число в выходном наборе параметров представляет собой «вероятность» конкретного класса. Сумма всех выходов равняется единице. На практике в качестве «победителя» выбирается тот класс, который имеет больший вес. На Рисунке 1.8 изображена классическая архитектура многослойной СНС [31, 32].

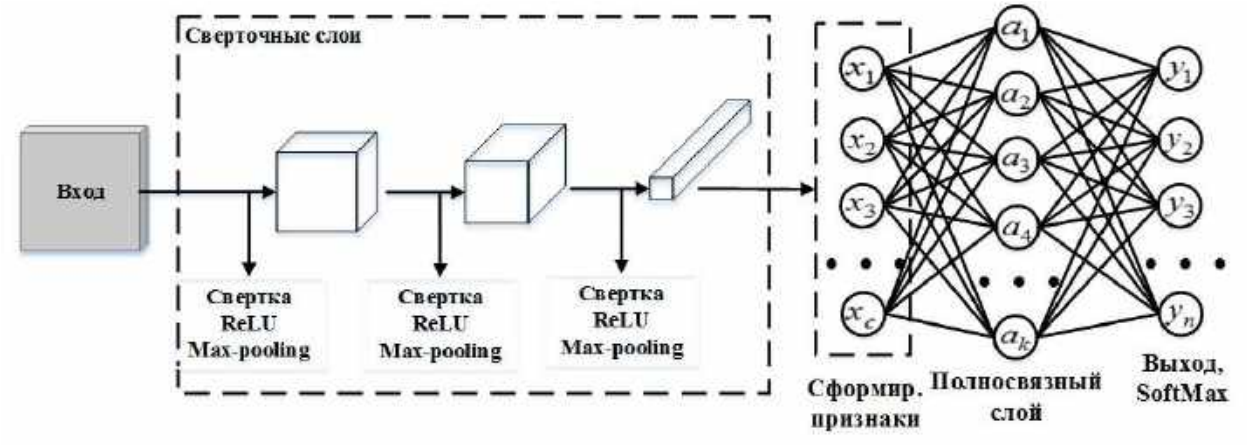


Рисунок 1.8 – Классическая архитектура СНС

В процессе обучения СНС слои линейной ректификации и субдискретизации представляют собой постоянные функции. Адаптация параметров проходит в сверточных и полносвязных слоях путем применения метода обратного распространения ошибки. Такой подход является модификацией классического метода градиентного спуска [18, 24].

1.4. Применение сверточных нейронных сетей в задачах распознавания лиц

Одной из самых известных архитектур СНС в данном классе задач является VGG16 [32], показанная на Рисунке 1.9. На вход сети поступают изображения лиц размером 224x224 пикселя. Сеть состоит из тринадцати

сверточных слоев с размером ядра 3x3 и трех полносвязных слоев. Выход сети представляет собой softmax-слой, размер которого регулируется параметром N. Значение данного параметра определяется количеством исследуемых классов.

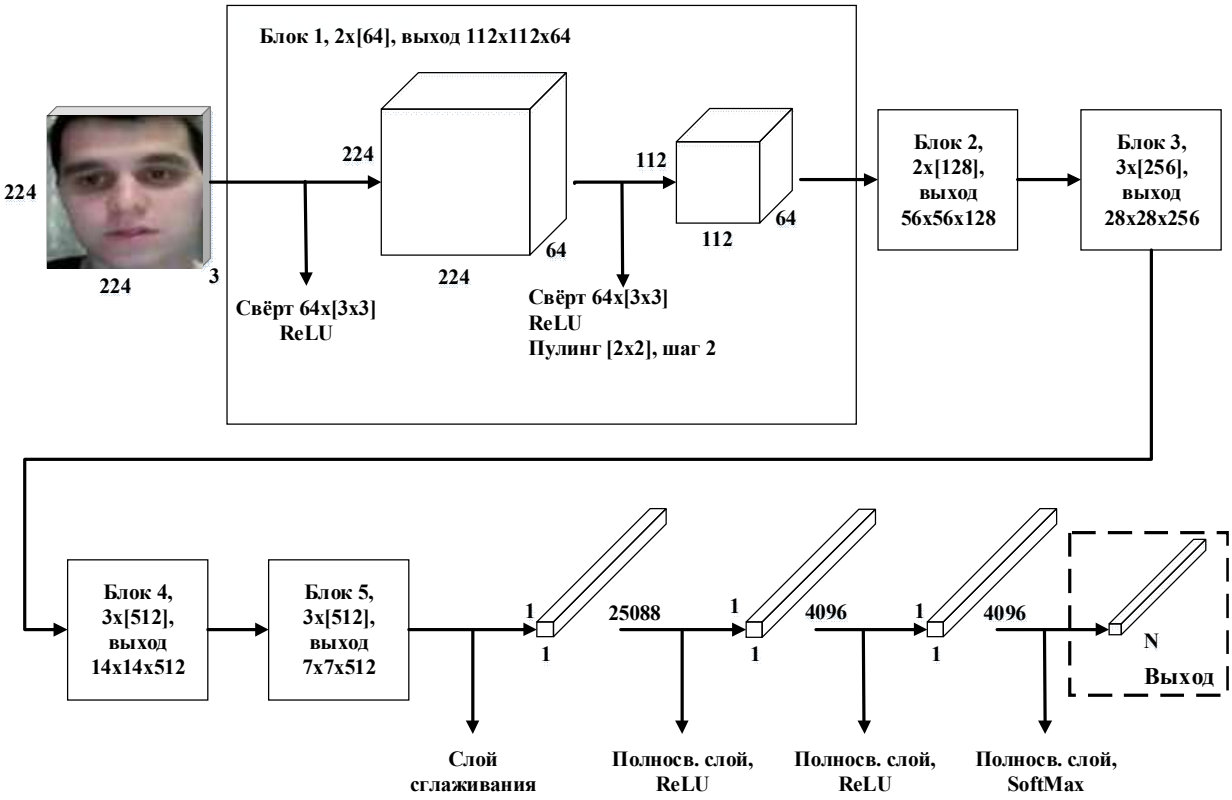


Рисунок 1.9 – Нейросетевая архитектура VGG16

Архитектура VGG16 представлена в одной из ключевых работ [33] по распознаванию лиц, где для обучения сети сформирован набор VGG Face Dataset, состоящий из более чем 2,6 млн. изображений лиц. Общее количество анализируемых пользователей в наборе составляло 2622 класса. Авторами исследования доказана эффективность использования данной архитектуры в задаче распознавания лиц. Однако стоит отметить, что сеть VGG16 обладает рядом недостатков. В частности, такая архитектура содержит большое количество обучаемых параметров. В зависимости от конфигурации их число изменяется от 133 до 144 млн. Ввиду этого, обучение

такой сети выполняется в течение длительного времени и требует высоких вычислительных затрат.

Другой популярной архитектурой СНС является ResNet (Residual Network, остаточная сеть) [34, 35]. Сверточные нейронные сети на основе данной архитектуры демонстрируют высокие результаты при распознавании лиц [36, 37]. Особенность этой архитектуры заключается в использовании связей быстрого доступа (shortcuts, skip-connections), что позволяет существенно увеличить эффективную глубину обучаемых сетей [34]. Данные связи применяются в пределах последовательно идущих свёрток, которые формируют разностный блок (residual unit) – основной структурный элемент ResNet [36, 38]. Сама связь осуществляется при помощи простого поэлементного сложения входа и выхода разностного блока, как показано на Рисунке 1.10. Ее можно описать следующим образом:

$$F(x_i, W_i, W_{i+1}) = W_2 f(W_1 x_i),$$
$$x_{i+1} = f(F(x_i, W_i, W_{i+1}) + x_i),$$

где x_i и x_{i+1} – вход и выход i-го блока, F – разностная функция, W_i и W_{i+1} – весовые параметры разностного блока, f – функция активации ReLU.

Использование разностных блоков позволяет решить две задачи. Первая относится к решению проблемы затухающего градиента. При использовании алгоритма обратного распространения ошибки, чем глубже слой сети, тем меньше становятся градиенты для обновления его весов. В результате глубокие слои практически не обучаются, что влияет на сходимость обучаемой сети. Дополнительно, использование разностных блоков позволяет существенным образом уменьшить общее количество обучаемых параметров. Так, одна из разновидностей ResNet, включающая 152 сверточных слоя, содержит до 60 млн. параметров, тогда как VGG16 включает в себя до 144 млн. параметров. Это достигается за счет использования ядер свертки размерности 1x1 внутри разностных блоков [34].

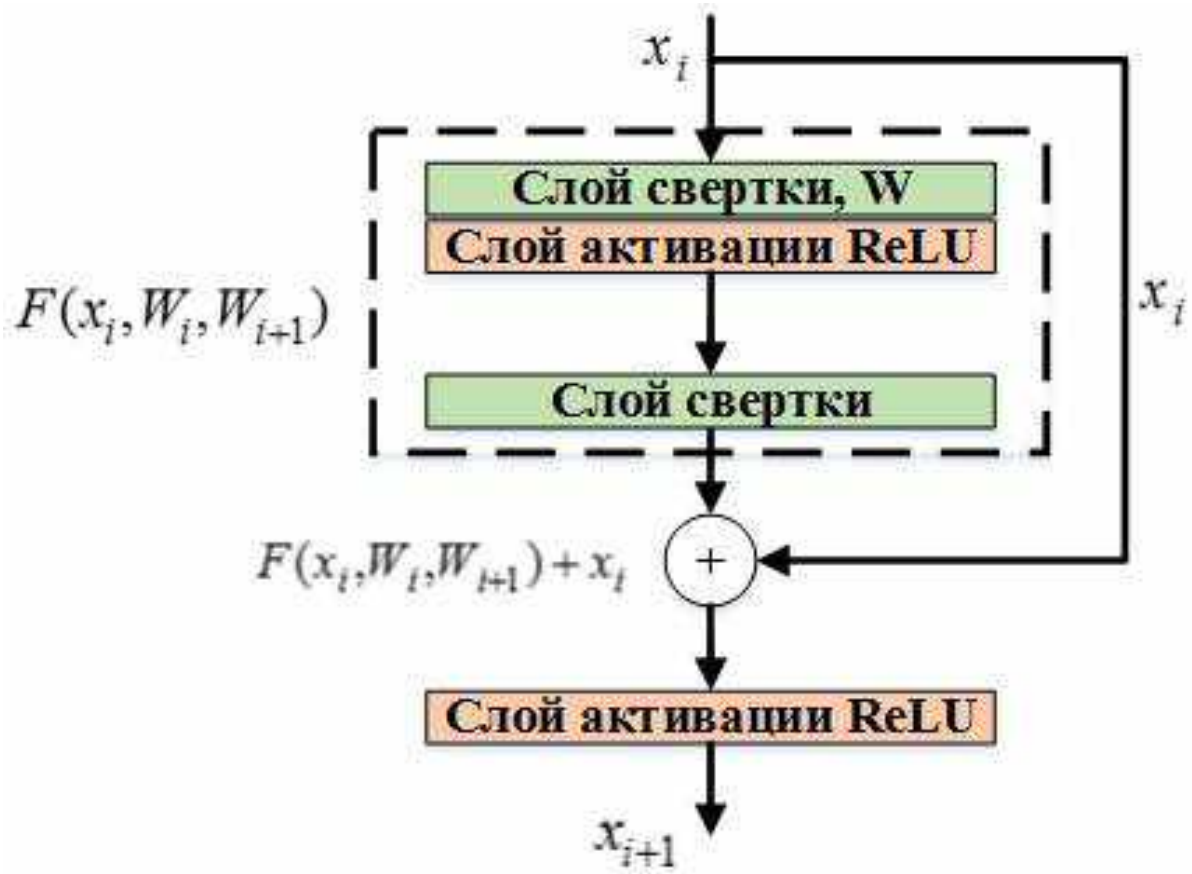


Рисунок 1.10 – Разностный блок архитектуры ResNet

В литературе предлагается ряд вариаций архитектуры ResNet, различие между которыми определяется глубиной сети. Увеличение количества разностных блоков теоретически позволяет повысить точность работы сети, однако на практике это чаще всего не выполняется [34, 36]. Разница в точности работы между сетями из 18 и 152 слоев незначительна, а ресурсы, необходимые для обучения, существенно возрастают, поэтому в данной работе используется относительно простая архитектура ResNet18 (Рисунок 1.11).

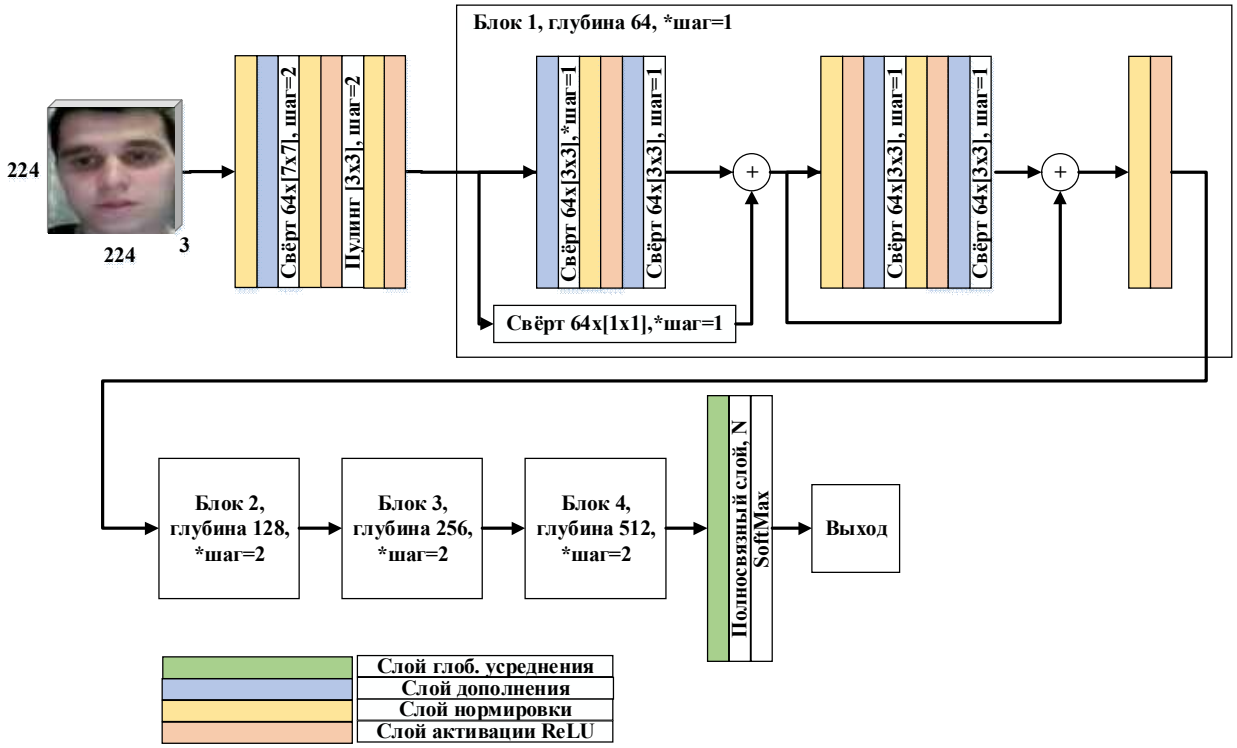


Рисунок 1.11 – Нейросетевая архитектура ResNet18

Еще одним примером архитектуры СНС, которая может использоваться в задачах распознавания лиц, является сеть SENet50 (Squeeze and Excitation Networks, сеть сжатия и возбуждения) [39]. Ее структурная схема приведена на Рисунке 1.12. В рассмотренных выше сетях VGG и ResNet пространственная и канальная информация смешиваются с использованием операции свертки, результатом которой является процесс формирования карт признаков. Особенность архитектуры SENet заключается в применении оригинального SE-блока, который адаптивно калибрует отклики функций по каналам за счет определения весовых параметров. В результате взвешивания выходных карт признаков происходит выделение информативных функций и подавление менее полезных. Важно отметить, что SE-блок не требователен к вычислительным ресурсам и незначительно усложняет архитектуру СНС. Доказано, что использование такого блока «сжатия и возбуждения» позволяет улучшить не только точность работы алгоритма, но и ускорить процесс сходимости в процессе обучения глубоких нейронных сетей [39].

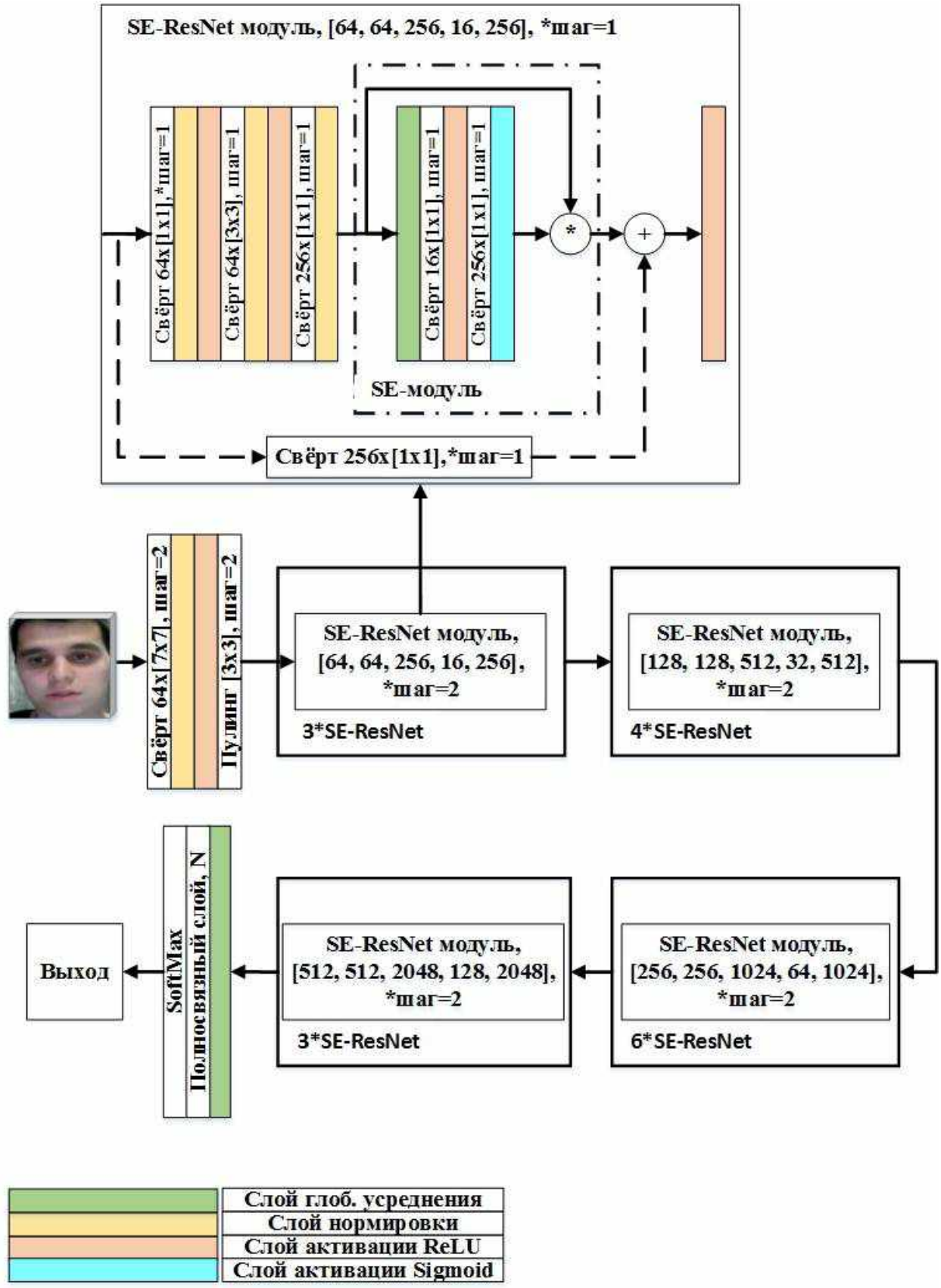


Рисунок 1.12 – Нейросетевая архитектура SeNet50

1.5. Применение сверточных нейронных сетей в задаче распознавания диктора

Одним из самых популярных подходов в задаче распознавания диктора является использование нейронных сетей на основе блоков временной задержки (Time Delay Neural Network, TDNN). Основная идея данного подхода заключается в анализе частотного представления сигнала с позиции временного ряда. Скрытые слои в сети анализируют разные по ширине временные фрагменты контекста, размер которых увеличивается по мере перехода к более глубоким слоям [40, 41, 46].

Ещё одним распространенным подходом является анализ двумерного представления речевого сигнала, которое можно интерпретировать как одну из разновидностей цифрового изображения [42, 43, 72, 111]. В этом случае задача распознавания диктора сводится к задаче идентификации визуальных образов, для которой используются СНС на базе ядер двумерной свертки. Так, в [72] авторами предлагается алгоритм на основе модификации СНС VGG-M, архитектура которой представлена на Рисунке 1.13. Данный подход анализирует спектральную карту признаков по принципу анализа изображений, используя ядра двумерной свертки.

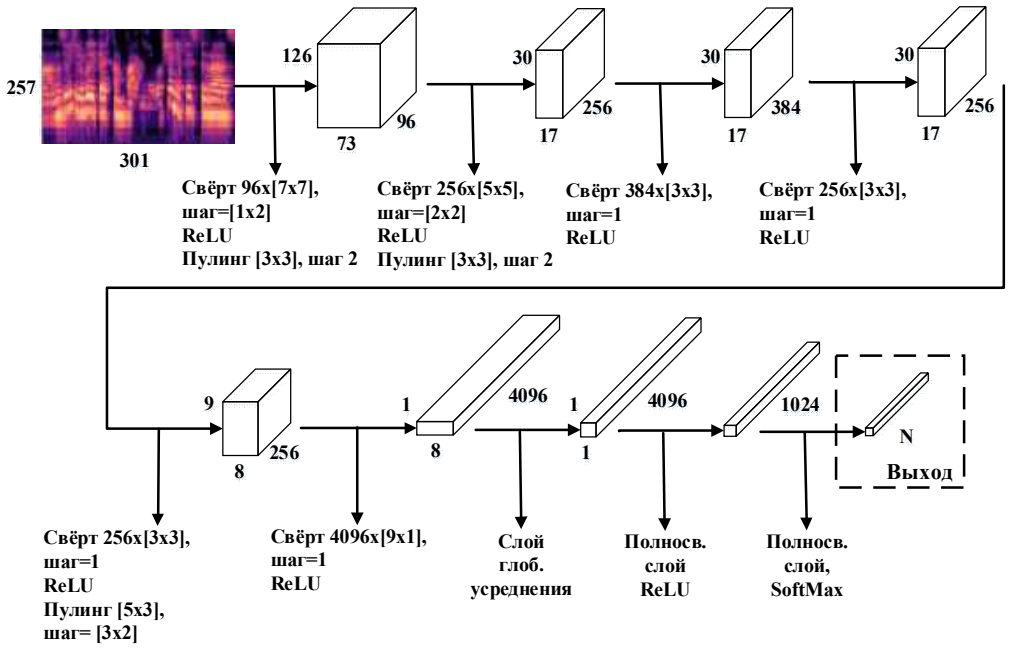


Рисунок 1.13 – Нейросетевая архитектура VGG-M

Архитектура нейронной сети ResNet, описанная выше в п. 1.4, получила распространение также в области распознавания речи и диктора. В частности, такие её разновидности, как ResNet18, ResNet34 и ResNet50, позволяют добиться высоких результатов в задаче биометрической идентификации личности по голосу [42, 43].

В последние несколько лет широкое распространение в задаче распознавания диктора получили x-векторные системы (x-vectors) [40-44]. Базовыми структурными блоками x-векторных систем являются сети типа TDNN. Архитектура таких систем позволяет добиваться высоких результатов в задаче распознавания диктора в условиях действия телефонного и микрофонного каналов. Подобная модель – метод представления голосового сегмента аудиозаписи в сжатой и в то же время богатой индивидуальными для говорящего признаками форме [41]. Такой подход показывает высокие результаты в задаче распознавания речи [45, 46]. В качестве x-подобного алгоритма в дальнейших исследованиях выбрано решение, представленное в работе [47]. Его архитектура изображена на Рисунке 1.14.

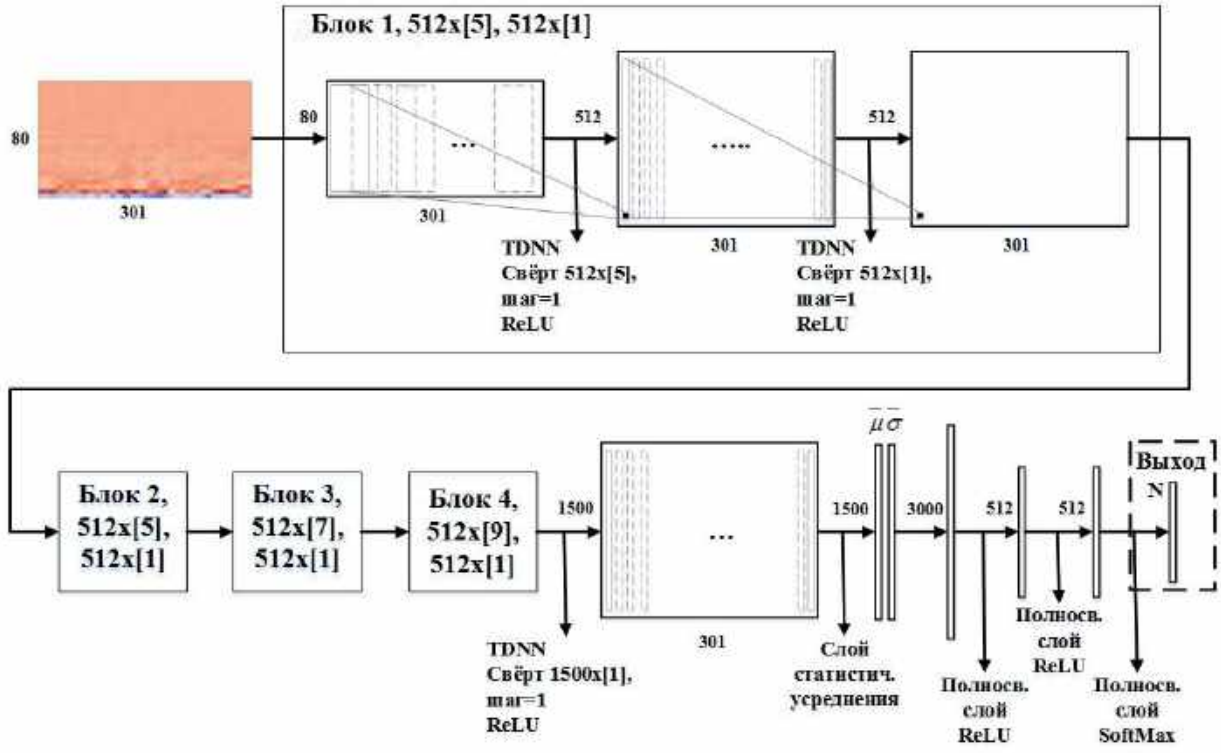


Рисунок 1.14 – x-векторная архитектура на базе TDNN-блоков

1.6 Мультимодальные биометрические системы и алгоритмы

1.6.1 Классификация методов комбинирования биометрических параметров

Классические системы идентификации личности строятся на основе использования одной биометрической модальности (унимодальные системы). Модальность отвечает за тип биометрических данных, используемый для проверки и подтверждения личности. Это могут быть, например, отпечатки пальцев, изображение лица, фонограмма, радужная оболочка глаза и т.п. Любая унимодальная система распознавания личности обладает характерным рядом ограничений. Например, подходы на основе анализа отпечатков пальцев или анализа радужной оболочки глаз требуют контакта с человеком, что уменьшает область их практической применимости. Системы распознавания пользователя по лицу имеют сильную зависимость от уровня освещенности, ракурса, качества фоторегистратора, кроме того, они чувствительны к возрастным изменениям, мимике и углу поворота лица. Система идентификации диктора зависит от качества микрофона и канала передачи информации. Таким образом, формируется потребность в создании решений, способных повысить надежность и устойчивость процесса распознавания личности [48, 49, 106].

Одним из направлений развития биометрических систем является создание мультимодальных алгоритмов. Основная идея данного подхода заключается в комплексном анализе двух и более биометрических параметров [59, 105]. Мультимодальный подход обладает свойством универсальности – использовать один из нескольких биометрических параметров при проверке личности [48, 49]. Это особенно актуально в ситуациях, когда человек имеет врожденные или приобретенные физиологические ограничения, такие как отсутствие пальцев или глухонемота.

Мультимодальные системы биометрической идентификации состоят из четырех составных блоков – модуль захвата данных, модуль формирования признаков, модуль сравнения и модуль принятия решения, как показано на Рисунке 1.15. В задаче идентификации личности на закрытом множестве модуль сравнения отсутствует, поскольку он используется только в задаче верификации, и заменяется модулем классификации [104, 115].

Рисунок 1.15 – Структурная схема мультимодальной системы биометрической идентификации

Модуль захвата данных представляет собой устройство, которое выполняет функцию считывания биометрических данных, например, камера, микрофон, сканер отпечатков пальцев. Дополнительно в данный модуль входит предобработка исходных данных: повышение качества, удаление шума, фильтрация.

Модуль формирования признаков необходим для создания компактного и информативного представления данных. Например, в задачах анализа изображений это могут быть признаки из классических алгоритмов

машинного обучения: HOG-признаки, локальные бинарные шаблоны [65, 66]. В задаче анализа речевых сигналов часто используют i-вектора [67].

Модуль классификации является алгоритмом, выполняющим функцию поиска закономерностей в признаковых данных и определяющим личность. При использовании алгоритмов на основе СНС модуль формирования признаков и модуль классификации представляют собой единый блок. На выходе мультимодальной системы используется блок принятия решения (блок постобработки), который анализирует результат работы классификатора и выносит финальное решение о результате распознавания личности [48, 49, 108].

Мультимодальные биометрические системы проектируются по двум схемам – последовательно и параллельно. В первом случае выходные результаты анализа одной биометрической модальности используются для сужения области поиска потенциальных пользователей и передачи данной информации на вход следующему блоку. При параллельном построении системы биометрическая информация разной природы анализируется единовременно. При этом существует три возможных уровня объединений: на уровне признаков; в процессе принятия итогового решения; на этапе сравнения. Последний вариант далее не рассматривается, поскольку относится только к задачам верификации пользователя [58, 63, 111].

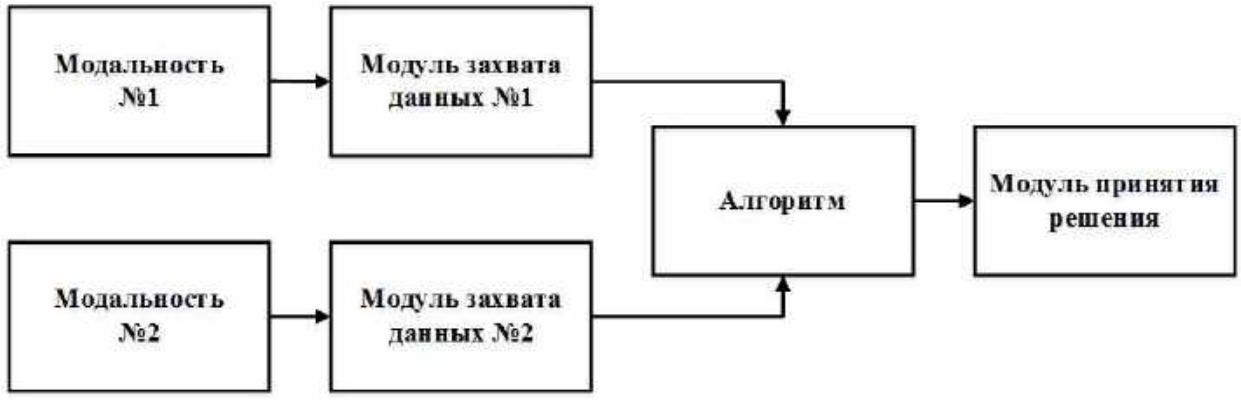
На Рисунке 1.16 представлен мультимодальный подход объединения модальностей на уровне признаков. Анализ биометрических данных разной природы выполняется независимо и параллельно. В процессе анализа биометрических данных на входе формируются признаки, характеризующую конкретную модальность. Далее они объединяются, образуя общий набор признаков, который используется для классификации и принятия итогового решения [60, 64, 112, 117, 118].

Рисунок 1.16 – Схема объединения модальностей
на уровне сформированных признаков

На Рисунке 1.17 представлен подход объединения модальностей на уровне принятия решения. При таком подходе каждая модальность обрабатывается и классифицируется независимо. Результаты классификации поступают на вход модуля принятия решения, где выполняется постобработка на основе решающего правила [48, 55, 104].

Рисунок 1.17 – Схема объединения модальностей
на уровне принятия решения

Существуют и другие подходы в области комбинированного анализа биометрических данных. В частности, ведутся исследования в области создания мультиалгоритмических систем. Их идея заключается в том, чтобы проводить анализ с использованием одной модальности, но с помощью разных алгоритмов, как показано на Рисунке 1.18. Конкретный режим работы алгоритмов определяется условиями эксплуатации системы. В частности, при анализе изображения лица один из алгоритмов системы может работать в условиях дневного освещения, тогда как другой запускается в режиме вечерней и ночной съемки. Возможно построение системы, где алгоритмы функционируют одновременно, а итоговый результат принимается модулем принятия решения [48, 49].



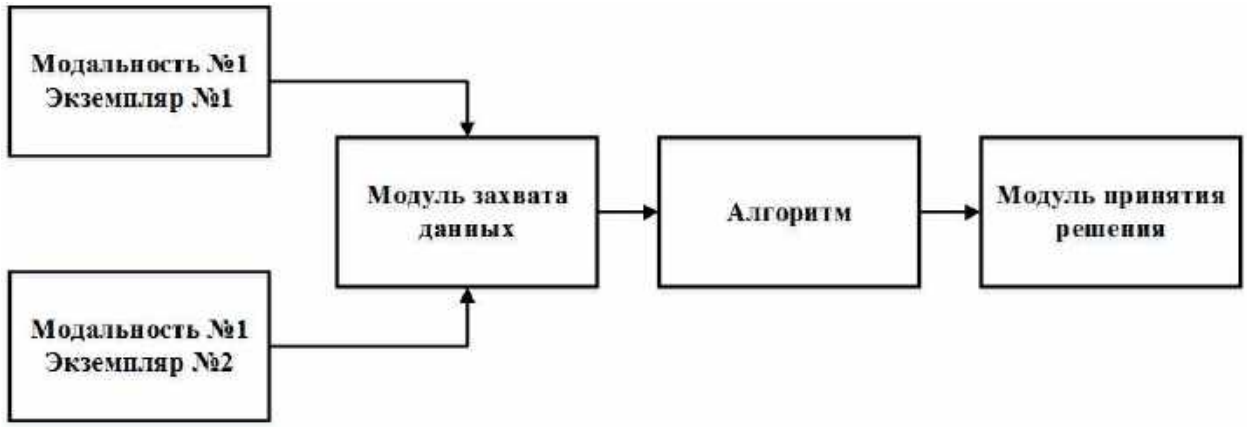
а)

б)

Рисунок 1.18 – Схемы построения комплексных биометрических систем:
а) мультимодальная; б) мультиалгоритмическая

Помимо анализа нескольких модальностей и применения набора алгоритмов к одному биометрическому параметру существует ещё один подход, посвященный разработке мультисенсорных биометрических систем. В данном случае анализ одного биометрического параметра выполняется с помощью двух и более физических датчиков разного действия. Так, в задаче для распознавания лиц может применяться одновременно камера видимого диапазона и тепловизор (камера инфракрасного диапазона). Еще одним способом в классификации комбинационных биометрических систем является мультиэкземплярный подход (Рисунок 1.19). Он основан на анализе нескольких биометрических экземпляров одной биометрической модальности. Например, производится анализ лица во фронтальном положении и при некотором изменении угла [48, 49].

а)



б)

Рисунок 1.19 – Схемы построения комплексных биометрических систем: а) мультисенсорная; б) мультиэкземплярная

1.6.2 Развитие мультимодальных биометрических алгоритмов

Разработкой систем мультимодальной биометрии активно начали заниматься в начале 2000-х. Одними из первых мультимодальных решений являлись подходы на основе различных вариаций комбинированного анализа цифрового изображения лица с отпечатками пальцев, геометрией кисти руки или отпечатком внутренней стороны ладони. Решения основывались на таких алгоритмах, как метод k-ближайших соседей, метод опорных векторов (SVM), решающие деревья (Random Forest), Байесовские методы. Понижение размерности входных данных и формирование признаков осуществлялись с применением метода главных компонент (PCA), линейно дискриминантного анализа (LDA), банка корреляционных фильтров на основе преобразования Фурье [50, 51]. Более поздняя работа посвящена анализу исключительно биометрических параметров рук [52], где предложен подход на основе обработки отпечатка пальцев, геометрии кисти руки и отпечатка внутренней стороны ладони. Для выделения информативных признаков с использованием изображений отпечатков пальцев и ладони используется дискретное вейвлет-преобразование (ДВП).

Чуть позже стали появляться мультимодальные решения с применением голосовой биометрии. В частности, в [53] рассматривается мультимодальный алгоритм на основе голосовой и лицевой биометрии, где используется Байесовская сеть для оценки надежности каждой модальности. Следует отметить, что оценка строится не только на основе точности работы классификаторов, но и анализа условий окружающей среды. В работе [54] предлагается комбинированная система на основе анализа лица, голоса и движения губ. Особое место в развитии мультимодальных биометрических систем занимают подходы на основе анализа отпечатка пальца и радужной оболочки глаза [55, 56]. В работе [57] предлагается подход на основе анализа изображений лиц и радужной оболочки глаза.

С развитием вычислительных мощностей появилась возможность достаточно быстро и качественно обучать алгоритмы на основе СНС. Их универсальность заключается в том, что они способны самостоятельно формировать признаковое представление входной информации и выполнять на его основе классификацию. На основе анализа современной научно-технической литературы можно с уверенностью утверждать, что наилучшие результаты в задаче распознавания личности достигаются при использовании именно СНС [33, 40, 72, 112].

Как правило, разработка архитектур СНС сводится к созданию конструкции, где сверточные слои используются для формирования признаков биометрических данных, а полносвязные применяются в качестве модуля классификации. Однако нередко сверточные слои предобученных моделей используются в качестве экстракторов признаков, а классификатор строится на базе классических алгоритмов машинного обучения.

Примером такого подхода является работа [58], где рассматриваются мультимодальные алгоритмы на основе анализа электрокардиограммы (ЭКГ) и отпечатков пальцев. Для формирования признаков в работе используются слои сети VGG16, которая ранее обучалась на стандартном наборе изображений ImageNet [33]. Выбор ЭКГ в данном случае обусловлен тем, что её можно использовать не только в качестве источника биометрической информации, но и как дополнительную проверку факта присутствия живого человека, что повышает надежность системы. В работе предлагаются два мультимодальных алгоритма. Первый представляет собой последовательную мультимодальную систему, где объединение осуществляется на уровне принятия решения, как показано на Рисунке 1.20. Из схемы видно, что в случае провала проверки на присутствие живого человека, система моментально инициализирует отказ в доступе.

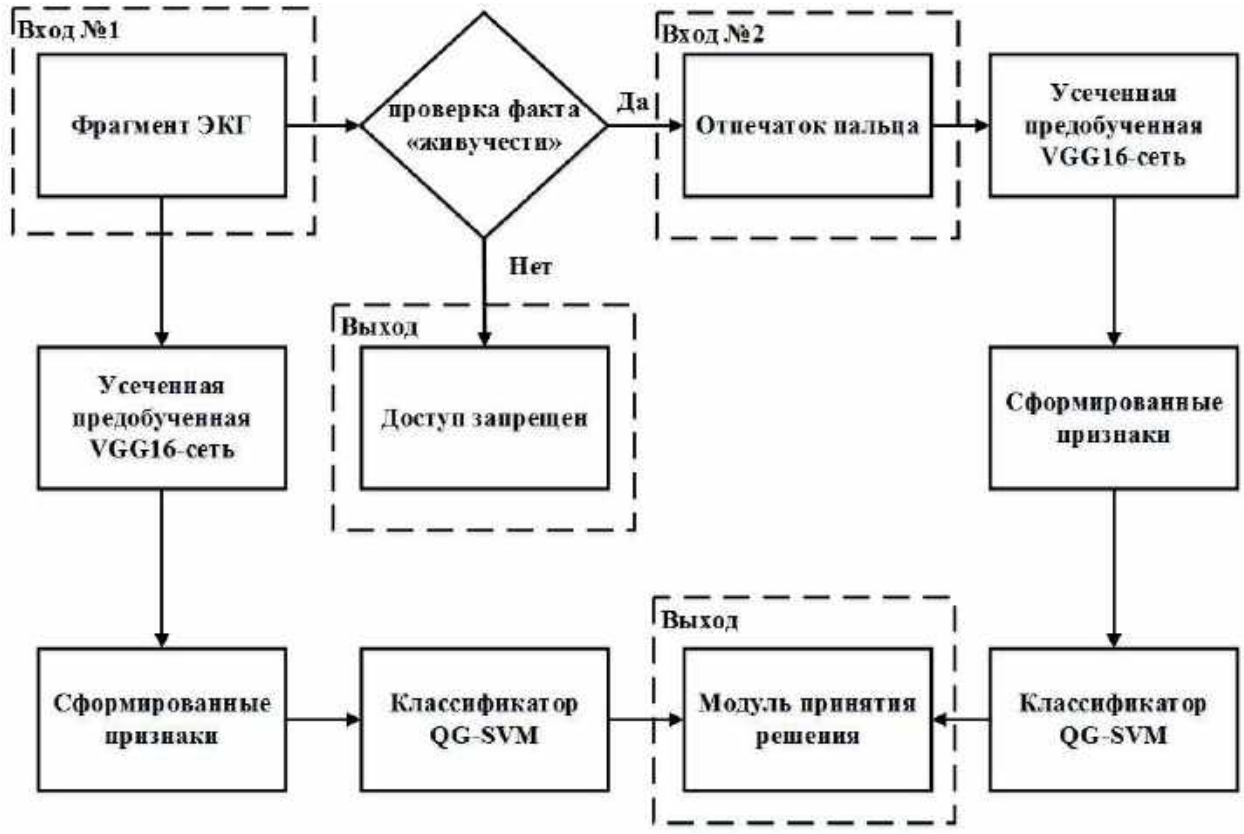


Рисунок 1.20 – Схема последовательной мультимодальной системы на основе анализа ЭКГ и отпечатка пальца

Второй алгоритм представляет собой параллельную мультимодальную систему. В данном случае объединение осуществляется на уровне признаков. Признаки объединяются в один общий вектор, который анализируется классификатором на базе модификации метода опорных векторов, как показано на Рисунке 1.21.

Использование сверточных слоев в качестве экстракторов признаков в комбинации с классическими алгоритмами машинного обучения является распространенной практикой в задаче мультимодальной биометрии, однако в большинстве своем наилучшее результаты достигаются при использовании решений на базе исключительно нейросетевых архитектур [59-64].

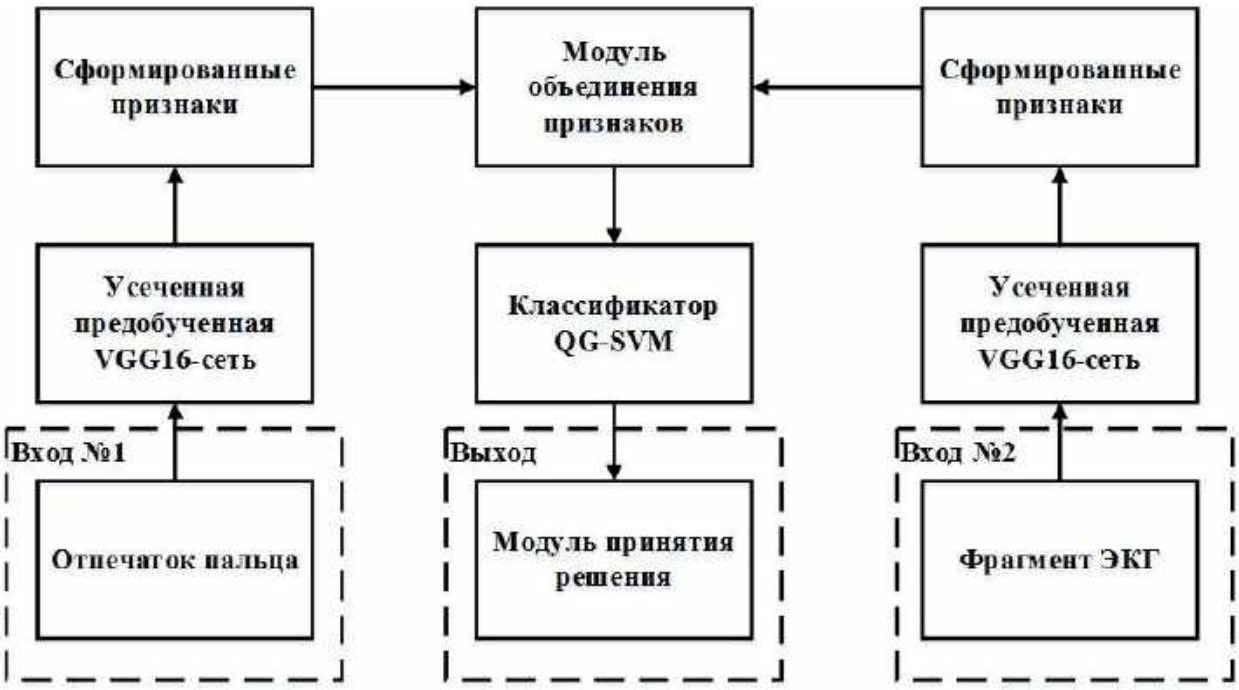


Рисунок 1.21 – Схема параллельной мультимодальной системы на основе анализа ЭКГ и отпечатка пальца

Для разработки алгоритмов биометрической идентификации личности на основе нейросетевых алгоритмов требуется большой объем данных. От размера и качества используемых данных зависят основные характеристики биометрических систем: точность, устойчивость работы в условиях изменения свойств окружающей среды, отсутствие жестких технических требований к устройствам сканирования биометрических данных. Далее будет рассмотрен этап подготовки и сбора речевых сигналов и изображений лиц для разработки и исследования алгоритмов распознавания личности.

1.7 Создание наборов биометрических данных

1.7.1 Текстозависимое и текстонезависимое распознавание диктора

В задаче распознавания диктора выделяют два типа систем: текстозависимые и текстонезависимые [13, 15]. К первым из них относятся системы, обладающие знанием о том, какая фраза должна быть озвучена пользователем во время идентификации – это может быть, например,

комбинация чисел или короткий фрагмент текста. В этом случае на этапе регистрации пользователю предлагается произнести уникальную ключевую фразу. Голосовой образец преобразуется из аналогового в цифровой формат. Далее извлекаются признаки, характеризующие голос пользователя. На последнем этапе формируется новый вектор признаков или модель диктора.

Большинство текстозависимых систем используют концепцию скрытых марковских моделей (СММ, Hidden Markov Models, HMMs) – случайных моделей, обеспечивающих статистическое представление звуков, создаваемых человеком. Они описывают изменения в речевом сигнале на основе анализа интенсивности, длительности и высоты тона. Другие методы основываются на модели гауссовых смесей (МГС, Gaussian Mixture Model, GMM), которые также часто используются и в текстонезависимых приложениях. Эти подходы используют информацию о голосе для создания вектора состояний, характеризующего физиологические особенности конкретного человека [68-71]. В процессе контроля происходит двухфакторная аутентификация – анализ сказанного и биометрическая идентификация.

Текстонезависимые системы не используют априорную информацию о фразе, произнесенной пользователем. Такая система является более гибкой, поскольку появляется возможность распознавания диктора в ситуациях, когда пользователь не желает быть опознанным. Это особенно актуально для случаев телефонного терроризма или мобильного мошенничества [12, 14].

В дальнейших исследованиях ставится задача на разработку текстонезависимых алгоритмов. Данный выбор определяется рядом причин. Во-первых, для систем прокторинга характерна спонтанная речь участников на этапе сдачи контрольных мероприятий. Во-вторых, использование текстонезависимых алгоритмов в системах контроля и управления доступом позволяет повысить надежность идентификации, а также расширить область их практического использования. Вследствие этого формируется требование

на подготовку текстонезависимого набора биометрических данных на русском языке.

1.7.2 Существующие текстонезависимые аудиовизуальные наборы данных

На сегодняшнем этапе развития в мире существует большое количество свободно распространяемых аудиовизуальных наборов голосовых данных, которые являются текстонезависимыми. Они, как правило, ориентированы на англоязычную речь. Наиболее популярными из них являются базы Voxceleb1, Voxceleb2, VGG-Sound, AVSpeech. В качестве источника биометрических данных они обычно используют сервис YouTube [72-75].

Аудиовизуальный набор данных Voxceleb1 состоит из коротких аудио- и видеофрагментов. Он включает в себя речь спикеров, охватывающих широкий спектр национальностей, стилей произношения и возрастных категорий. Набор содержит более 150 тыс. аудио- и 22 тыс. видеозаписей. Всего для записей использовался 1251 человек. Набор VoxCeleb2 является обновленной версией рассматриваемой базы. Он содержит более 1,1 млн. аудио- и 150 тыс. видеозаписей для 6112 различных записанных людей [72, 73].

Другой крупномасштабный набор VGG-Sound содержит аудио- и видеоданные, записанные в сложных акустических и визуальных условиях. Он насчитывает более 200 тыс. видеоклипов и аудиозаписей, которые принадлежат 309 различным людям. Аудио- и видеозаписи разделены на фрагменты одинаковой длины с длительностью 10 сек. Однако данный набор содержит не только фрагменты записей людей. Данные разделяются на следующие категории: люди, животные, музыка, спорт, природа, транспорт, дом, инструменты и другое [74].

Набор AVSpeech более идеализирован, поскольку включает в себя аудио- и видеоданные высокого качества. Записи сделаны в условиях

отсутствия фоновых шумов и помех, а также с использованием профессионального оборудования. Этот набор содержит более 290 тыс. записей высококачественных лекций и обучающих видеороликов. Общая длительность набора составляет 4700 часов, охватывающих более 150 тыс. людей. Уникальность данного набора заключается в его мультиязычности. Тем не менее, стоит отметить, что большая часть набора AVSpeech содержит фрагменты исключительно англоязычной речи [75].

Поскольку русскоязычные текстонезависимые наборы оказались недоступны, в ходе проведения исследования была сформулирована задача на сбор и подготовку собственного набора речевых сигналов и цифровых изображений лиц.

1.7.3 Подготовка требований к базе видеоданных и речевых сигналов

Перед стартом работ по подготовке базы биометрических данных был составлен сценарий для сбора цифровых изображений и речевых сигналов. Данная инструкция содержит описание технических требований и условий, необходимых для корректной записи биометрических параметров. Также она включала в себя 49 вопросов, на которые отвечал записываемый человек (далее – респондент).

В процессе записи и накопления аудио- и видеоданных использовалось популярное в настоящий момент приложение Zoom [76], предназначенное для видеоконференций. Данный выбор обусловлен тем, что программа Zoom позволяет подготовить аудиовизуальный набор, который по своим параметрам копирует процедуру проведения дистанционных испытаний. Требования на технические характеристики записывающих устройств не выставлялись с целью создания аппаратно-независимого набора данных.

Для того чтобы собрать наиболее репрезентативную выборку, перед респондентами ставился ряд требований к условиям записи, а также к речи и лицу, как к источникам биометрической информации. Одним из дополнительных требований являлось отсутствие длительных пауз,

поскольку на этапе предобработки данных использовался детектор голосовой активности (ДГА, Voice Activity Detector, VAD) [77]. Основным назначением ДГА является определение участков, содержащих речь с последующей фильтрацией пауз. Если диктор делает большое количество остановок в речи, то общее время записи живого разговора может быть недостаточным для обучения и тестирования нейросетевых алгоритмов идентификации диктора.

1.7.4 Создание набора аудио- и видеоданных FaceSpeechDB

В соответствии с требованиями, описанными в предыдущем пункте, подготовлен аудиовизуальный набор FaceSpeechDB, предназначенный для обучения и тестирования нейросетевых алгоритмов распознавания личности. На Рисунке 1.22 представлена схема процесса записи аудио- и видеоданных.

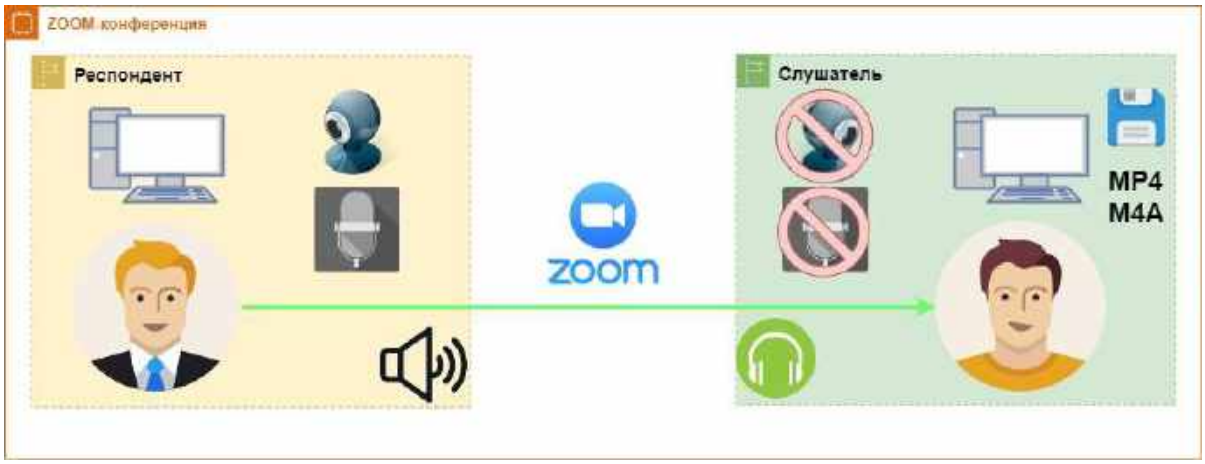


Рисунок 1.22 – Схема процесса записи аудио- и видеоданных

На Рисунке 1.23 представлены примеры кадров из собранного набора видеоданных. В процессе видеозаписи респонденты изменяли угол поворота головы. Жесткие требования к условиям освещения не предъявлялись с целью приближения экспериментальных данных к реальным условиям видеоконференцсвязи. Также стоит отметить, что видео записывалось с разным разрешением и качеством: 640x360, 640x480 и 1280x720 пикселей.

Рисунок 1.23 – Кадры изображений из собранной базы видеоданных

В результате сбора и предварительной проверки биометрических данных подготовлен набор, включающий в себя 60 часов записи разговоров людей в реальных акустических и визуальных условиях:

- Zoom-видеозаписи продолжительностью 30 часов;
- Zoom-аудиозаписи продолжительностью 30 часов;
- общее количество респондентов составило 104 человека.

Набор FaceSpeechDB является универсальным, поскольку может быть использован как для разработки унимодальных, так и мультимодальных биометрических алгоритмов. Другими словами, возможно комбинирование голосовой и лицевой биометрии, а также применение биометрических параметров в качестве самостоятельных и независимых модальностей.

Подготовленный набор данных FaceSpeechDB, в отличие от наборов из открытых источников, обладает рядом особенностей. Во-первых, он содержит фрагменты, ориентированные на русскоязычную речь. Во-вторых, набор содержит исключительно фрагменты аудио- и видеозаписей людей. В-третьих, в наборе содержатся шумы, помехи и дополнительные искажения, возникающие в процессе использования приложения Zoom. Возникновение нежелательных эффектов обусловлено алгоритмами сжатия, которые

применяются при передаче мультимедийных сигналов по каналу связи, а также различным качеством пользовательских устройств.

Таким образом, подготовлен набор данных FaceSpeechDB, который имеет высокую степень сходства с реальными условиями эксплуатации систем прикладного телевидения и видеоконференцсвязи и может быть использован для разработки биометрических алгоритмов идентификации на основе анализа русскоязычной речи и изображений лиц.

1.7.5 Создание набора аудиоданных VADSpeakersDB

Человек в спонтанной речи не может обойтись без остановок, поэтому паузы всегда присутствуют в записанных голосовых наборах. Наличие таких остановок в речи с точки зрения систем распознавания диктора является негативным фактором [78]. Это связано с тем, что работа большинства нейросетевых методов стоится на анализе коротких речевых фрагментов длительностью 2-4 секунды [42, 72, 105]. Этого достаточно, чтобы с высокой точностью распознать говорящего.

Дополнительно стоит отметить, что фонограмма является крайне сложным представлением информации, которое напрямую зависит от качественных характеристик записывающих устройств, а также степени зашумленности канала передачи и акустических свойств окружающей среды. В итоге любой из анализируемых кратковременных фрагментов может полностью или частично содержать паузы без речи или только шумы, что может существенно ухудшать качество работы алгоритмов распознавания диктора. Для борьбы с этим на этапе предобработки из фонограммы выделяются фрагменты, содержащие исключительно речь. Для этого используют ДГА алгоритмы. В данном исследовании рассматривается работа классических алгоритмов выделения голосовых фрагментов, а также предлагается оригинальный комбинированный детектор на основе объединения независимых ДГА.

Для разработки детектора голосовой активности подготовлен оригинальный набор речевых сигналов VADSpeakersDB. Набор представляет собой запись живой русскоязычной речи 23 дикторов. Речевые данные являются сбалансированными по гендерному признаку: 55% составляет речь мужчин и 45% речь женщин. Каждого диктора записывали в течение 60 секунд. В качестве программного обеспечения для записи использовалось приложение Zoom. Данные в ручном режиме размечались специалистами. Каждый фрагмент прослушивался специалистом, который в итоге выставлял метку «речь» или «шум/пауза». В Таблице 2.1 представлены основные характеристики подготовленного набора фонограмм.

Таблица 1.1 – Характеристики фонограмм
из набора VADSpeakersDB

Язык	Русский
Частота дискретизации	16 кГц
Количество дикторов	23
Суммарная длительность	23 мин.
Длительность одного фрагмента	10 мс
Общее количество фрагментов	138 000
Количество фрагментов класса «речь»	68,28%
Количество фрагментов класса «шум/пауза»	31,72%

Таким образом, на этапе подготовки и сбора биометрических данных размечены и структурированы два оригинальных набора: FaceSpeechDB и VADSpeakersDB. Далее они будут использованы для разработки и тестирования унимодальных и мультимодальных алгоритмов идентификации личности, а также для разработки и тестирования детектора голосовой активности.

1.8 Краткие выводы

Результаты проведенного анализа существующих задач, современных методов и алгоритмов в области систем распознавания личности на основе обработки цифровых изображений лиц и речевых сигналов, позволяют сделать следующие основные выводы:

- Сверточные нейронные сети на сегодняшнем этапе развития являются главным инструментом в задачах биометрической идентификации личности как на основе анализа речевых сигналов, так и в случае использования цифровых изображений лиц.
- В настоящее время одной из практических проблем, которая возникает на пути использования алгоритмов распознавания личности на основе лицевой и голосовой биометрии, является зависимость точности работы биометрических алгоритмов от искажающих факторов. В частности, системы идентификации диктора зависят от эффектов, возникающих в канале передачи и микрофона, физиологических особенностей говорящего, акустических свойств окружающей среды. Точность распознавания пользователя по лицу зависит от уровня освещения, ракурса, качества фоторегистратора, кроме того, она чувствительна к возрастным изменениям и мимике. Возникает задача разработки комбинированных методов идентификации личности на основе двух и более биометрических параметров, что позволит не только повысить устойчивость и точность работы биометрических систем, но и улучшить надежность работы при попытках несанкционированного доступа.
- Относительно новым вызовом для систем лицевой биометрии является распространение ситуации наличия на лице человека медицинской маски, которая существенно усложняет процесс идентификации личности.

– В процессе разработки алгоритмов распознавания личности на основе глубоких СНС возникает необходимость использования больших наборов речевых сигналов и цифровых изображений лиц. Аудивизуальные наборы, которые учитывали бы такие факторы, как русскоязычность речи, запись с использованием большого разнообразия аудио- и видеоустройств, в условиях сильной изменчивости акустических свойств и освещения, в настоящий момент недоступны. Поэтому перед исследованием возникла и решена задача сбора и разметки собственных наборов аудиовизуальных данных.
– Собран аудиовизуальный набор данных FaceSpeechDB для обучения и тестирования алгоритмов биометрической идентификации личности. Он содержит 60 часов русскоязычной записи 104 человек. Акустические свойства набора имеют высокую степень сходства с реальными условиями эксплуатации систем прокторинга.
– Собран набор аудиосигналов VADSpeakersDB для разработки детектора голосовой активности. Набор содержит записи русскоязычной речи. Общее количество подготовленных фрагментов составляет 138 000 штук.

Таким образом, проведенный анализ позволил сформулировать следующие основные задачи диссертации:

- 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.

ГЛАВА 2
ИССЛЕДОВАНИЕ НЕЙРОСЕТЕВЫХ АЛГОРИТМОВ ИДЕНТИФИКАЦИИ ЛИЧНОСТИ НА ОСНОВЕ АНАЛИЗА РЕЧЕВЫХ СИГНАЛОВ

2.1 Вводные замечания

Рассмотрим задачу предобработки аудиосигналов. Проведем тестирование классических ДГА и разработку на их основе комбинированного детектора голосовой активности (КДГА). Исследование выполняется с использованием подготовленного набора аудиосигналов VADSpeakersDB. Применение ДГА для предобработки аудиосигналов обусловлено повышением точности работы алгоритмов идентификации диктора вследствие очистки фонограмм от пауз, эффектов глотации, вдохов и шумов. Далее будет описан этап обработки фонограмм, который включает применение разработанного КДГА, формирование частотного представления речевых сигналов, создание выборок данных из набора FaceSpeechDB для обучения и тестирования нейросетевых алгоритмов идентификации личности. В заключительной части выполняется исследование стандартных алгоритмов голосовой биометрии и разработка робастного алгоритма на основе x-векторной системы. Исследование точности работы алгоритмов выполняется в условиях, близких к условиям эксплуатации биометрических систем, а также в условиях сильного зашумления речевых сигналов. Дополнительно проводится анализ вычислительной сложности рассматриваемых нейросетевых алгоритмов.

2.2 Метрики оценки качества работы детектора голосовой активности

Поскольку алгоритм ДГА решает задачу бинарной классификации, то для оценки качества работы может быть использована такая метрика, как доля правильных ответов (*acc*). В процессе оценки *acc* важно обратить

внимание на несбалансированность данных, поскольку фрагментов речи существенно больше фрагментов, содержащих паузы или помехи. Для учета данного свойства вводится аналогичная оценка доли правильных ответов для несбалансированных данных (*accb*). Необходимо также определить индикатор корректности распознавания фрагмента речевого сигнала:

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

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 ДГА. Тогда метрику *acc* можно определить, как:

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

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

Для подсчета доли правильных ответов на несбалансированных данных в задаче бинарной классификации необходимо воспользоваться выражением:

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

где $acc_{y_i=1}$ – доля верно детектированных фрагментов, представляющих класс «речь»; $acc_{y_i=0}$ – доля верно детектированных фрагментов, определяющих класс «шум/пауза».

Также для оценки качества работы ДГА воспользуемся гармоническим средним 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 [79].

Ранее отмечалось, что набор VADSpeakersDB обладает дисбалансом в данных. В соответствии с данным свойством определим F на основе макро-усредняющего подхода, то есть расчет метрики внутри каждого класса («речь», «шум/пауза») с последующей нормировкой на общее количество классов:

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

Рассмотрим для начала работу классических ДГА.

2.3 Классические алгоритмы анализа голосовой активности

Аудиосигнал неразрывно связан с понятием энергии. В теории сигналов энергия является количественной характеристикой, отражающей определенные свойства сигнала и динамику изменения его значений во времени, в пространстве или по любым другим аргументам. Она определяется как квадрат функции амплитуды сигнала. Дополнительно энергия может быть рассчитана как интеграл от мощности по всему интервалу существования рассматриваемого сигнала. В радиотехнических системах энергия сигнала является одной из определяющих характеристик.

Одним из самых простых видов ДГА является алгоритм на основе анализа энергии [77, 78]. Сигнал во временной области делится на окна длиной 10-30 мс. Далее в каждом окне считается квадрат амплитуды всех отсчетов внутри окна. После этого проводится пороговая фильтрация. Если значение энергии больше заданного порога θ, то фрагмент оставляют. Логика этого процесса определяется тем, что речевые фрагменты имеют высокий уровень энергии, тогда как фрагменты, содержащие шум/паузы обладают слабой энергетической составляющей. Исключением из такой логики является импульсный шум. Алгоритм на основе анализа энергии математически можно описать следующим образом:

$$S = \{\vec{S}_1, \vec{S}_2, \vec{S}_3 .... \vec{S}_j\}, \text{где } \vec{S}_j = (s_1, s_2, s_3 ..... s_N),$$

$$E_j = \sum_{i=1}^{N} E(i) = \sum_{i=1}^{N} s^2(i),$$

$$\vec{V} = \begin{cases} \vec{S}_j, \textit{если}\ \theta <= E_j \\ 0, \textit{если}\ \theta > E_j, \end{cases}$$

$$S' = \left\{ \vec{V}_1, \vec{V}_2, \vec{V}_3, \ldots, \vec{V}_w \right\},$$

где S – исходный речевой сигнал, $\vec{S_j}$ – j-ый фрагмент исходного сигнала, $s(i)$ – амплитуда i-го отсчета, $E(i)$ – энергия i-го отсчета, E_j – энергия j-го фрагмента исходного сигнала, N – длина окна, $\vec{V}$ – речевой фрагмент в соответствии с заданным порогом θ, w – количество окон, содержащих речь, S' – обработанный сигнал [77, 80].

Значение порога θ определяется следующим образом. Для каждого фрагмента фонограммы $\vec{S_j}$ высчитывается энергия, после чего определяется минимальное и максимальное значение энергии для всей фонограммы. Затем эмпирически подбирается коэффициент k, изменяющийся в диапазоне $[2 * 10^{-5},\ 2 * 10^{-3}]$ Регулируя его, можно контролировать порог θ, рассчитываемый следующим образом:

$$\theta = k \cdot (E_{\max} - E_{\min}),$$

где E_{max}, E_{min} – максимальное и минимальное значение энергии фонограммы.

В Таблице 2.1 представлены результаты тестирования алгоритма ДГА на основе анализа энергии фонограмм (далее – $ДГА_1$). Данный подход демонстрирует высокую точность выделения речевых фрагментов с учетом того, что анализируемый набор VADSpeakersDB подготовлен с использованием широкого спектра устройств записи, а также в условиях изменчивости акустических свойств среды. В рамках эксперимента определено оптимальное значение настраиваемого параметра $k = 2*10^{-4}$.

Таблица 2.1 – Результаты тестирования $ДГА_1$ на базе VADSpeakersDB

Метрики	k							
	$2*10^{-8}$	$2*10^{-7}$	$2*10^{-6}$	$2*10^{-5}$	$2*10^{-4}$	$2*10^{-3}$	$2*10^{-2}$	$2*10^{-1}$
acc	0,77	0,77	0,86	0,89	**0,90**	0,88	0,78	0,53
accb	0,63	0,64	0,79	0,85	**0,88**	0,89	0,83	0,66
F	0,85	0,86	0,91	0,92	**0,93**	0,90	0,81	0,48
$F_{макро}$	0,64	0,65	0,82	0,87	**0,88**	0,87	0,77	0,53

Другим возможным методом построения ДГА является подход на основе анализа энергии Тигера-Кайзера (далее – $ДГА_2$) [80, 81]. Соответствующий алгоритм основывается на том, что сигнал не разбивается на окна, а для каждого временного отсчета вычисляется энергия следующим образом:

$$E(i) = s^2(i) - s(i-1)s(i+1),$$

где *s(i) –i*-ый отсчет фонограммы. В Таблице 2.2 представлены результаты тестирования алгоритма $ДГА_2$. Они показывают, что при значении порога θ, равном $3*10^{-6}$, достигаются наиболее высокие результаты, которые в целом сопоставимы с алгоритмом $ДГА_1$.

Таблица 2.2 – Результаты тестирования $ДГА_2$ на базе VADSpeakersDB

Метрики	θ							
	$2*10^{-9}$	$2*10^{-8}$	$2*10^{-7}$	$2*10^{-6}$	$3*10^{-6}$	$4*10^{-6}$	$2*10^{-5}$	$2*10^{-4}$
acc	0,77	0,86	0,89	0,89	**0,89**	0,89	0,85	0,69
accb	0,64	0,78	0,84	0,87	**0,88**	0,88	0,87	0,77
F	0,86	0,91	0,92	0,92	**0,92**	0,92	0,88	0,71
$F_{макро}$	0,65	0,81	0,86	0,88	**0,88**	0,87	0,84	0,69

Также методы анализа фонограмм реализуются с использованием частотного представления сигнала на основе дискретного преобразования Фурье (ДПФ). Для проведения исследований в работе реализован алгоритм ДГА на основе частотного анализа фонограмм (далее – $ДГА_3$). Опишем подробно основные этапы работы такого детектора. В начале исходная фонограмма обрабатывается с помощью ДПФ, что в результате формирует

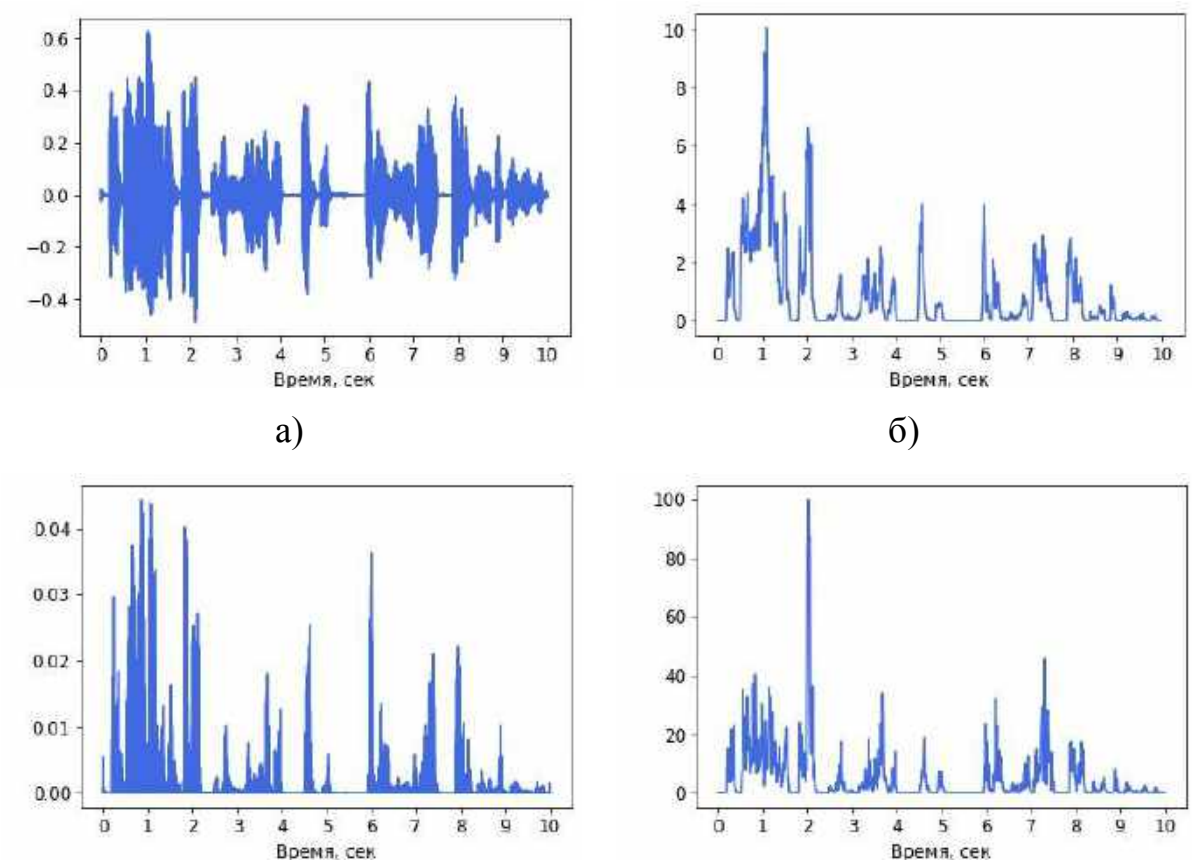
спектрограмму. Следующим этапом вычисляется периодограмма, которая определяется как модуль квадрата спектрограммы. Затем периодограмма обрабатывается полосовым фильтром, поскольку речь человека составляет ограниченный диапазон частот от 300 Гц до 3,4 кГц. На завершающем этапе обработки для каждого фрагмента определяется максимальное значение величины мощности спектра. Если значение устанавливается выше определяемого порога θ, то фрагмент помечается как участок фонограммы, содержащий исключительно голос диктора [82].

В Таблице 2.3 представлены результаты тестирования алгоритма $ДГА_3$. Рассмотренное решение при значении порога θ, равном $2*10^{-3}$, демонстрирует точность определения речевых фрагментов более 87%.

Таблица 2.3 – Результаты тестирования $ДГА_3$ на базе VADSpeakersDB

Метрики	θ							
	$2*10^{-8}$	$2*10^{-6}$	$2*10^{-5}$	$2*10^{-4}$	$2*10^{-3}$	$2*10^{-2}$	$2*10^{-1}$	$2*10^{0}$
acc	0,77	0,84	0,87	0,89	**0,89**	0,87	0,79	0,62
accb	0,63	0,75	0,80	0,85	**0,87**	0,87	0,83	0,72
F	0,85	0,89	0,91	0,92	**0,92**	0,90	0,82	0,62
$F_{макро}$	0,64	0,78	0,83	0,86	**0,87**	0,85	0,78	0,62

На Рисунке 2.1 изображены примеры представления речевого сигнала длительностью 10 сек. Анализируя данные зависимости, можно отметить относительную эффективность работы рассмотренных детекторов. Они явным образом выделяют области с высокой интенсивностью и участки со слабой энергетической составляющей. Результаты данной части работы показали, что классические ДГА обладают точностью обнаружения речевых фрагментов на уровне 0,87-0,9. Стоит отметить, что поскольку рассмотренные ДГА не требовательны к вычислительным ресурсам, то появляется практическая возможность для создания более сложного комбинированного алгоритма.



а) б)

в) г)

Рисунок. 2.1 – Примеры представления речевого сигнала в детекторах:
а) осциллограмма; б) $ДГА_1$; в) $ДГА_2$; г) $ДГА_3$

Для улучшения качества выделения участков, содержащих речь, в работе разработан детектор голосовой активности с использованием одного из типов ансамблевого объединения независимых алгоритмов – стекинга (Stacked Generalization, Stacking) [83-85].

2.4 Разработка комбинированного детектора голосовой активности

Подход на основе стекинга состоит в том, чтобы использовать каждый ДГА в качестве независимого детектора, а их выходы объединить для последующего анализа обобщающим классификатором [77]. На Рисунке 2.2 представлена структурная схема предложенного решения – комбинированный детектор голосовой активности.

Рисунок 2.2 – Структурная схема предложенного алгоритма КДГА

На вход алгоритма поступает фрагмент фонограммы длительностью в 10 мс. Далее каждый из детекторов анализирует входной образец. После этого формируются три независимых предсказания, которые подаются на вход обобщающего классификатора. Он принимает итоговое решение, к какому из классов отнести входной фрагмент фонограммы – «речь» или «не речь». В качестве такой модели используется классификатор на основе ансамбля решающих деревьев (случайный лес) [86, 87].

Для того чтобы обучить предложенный классификатор, необходимо подготовленный набор данных VADSpeakersDB разложить на обучающую и тестовую выборки. Для проведения эксперимента в процессе обучения использовался подход на основе *k*-блочной перекрестной проверки (K-Fold Cross Validation) [88, 89]. Параметр *k* определяет, на сколько равных частей будет разбит обучающий набор. Затем на *k-1* частей обучается модель, а оставшаяся часть используется в качестве проверочного множества. Обучение повторяется *k* раз. В итоге каждая из *k*-частей участвует в проверке. После чего оценочная характеристика усредняется. В исследовании параметр *k* задавался равным 10.

В Таблице 2.4 показана статистика разделения набора VADSpeakersDB. На тестирование выделялось 15% примеров от общей суммы фрагментов

речевых данных. В процессе обучения и применения перекрестной проверки 10% данных использовались для оценки работы классификатора.

Таблица 2.4 – Разделение набора данных VADSpeakersDB

-	Обучающая выборка	Тестовая выборка	Общее количество
Количество фрагментов	117300	20700	138000

На Рисунке 2.3 представлена схема процесса обучения и тестирования разработанного алгоритма КДГА. Классификатор на базе ансамбля решающих деревьев обладает широким набором настраиваемых параметров. Для настройки используются такие показатели, как общее количество деревьев, выбор критерия разделения, максимальная глубина деревьев, количество анализируемых признаков в каждом из узлов для принятия решения. Подбор оптимальных настроек для классификатора выполняется на основе построения сетки параметров [90]. Для этого выполняется перебор возможных значений для каждого из параметров. Такой подход является достаточно требовательным к временным и вычислительным ресурсам, поскольку количество обучаемых классификаторов может достигать десятков тысяч. После этого определяется оптимальный классификатор в соответствии с целевой метрикой. В работе в качестве такой метрики использовалась F-мера на основе макро-усредняющего подхода.

В процессе обучения и поиска оптимального классификатора проведен анализ более 2100 моделей. Выполнен выбор модели, показавшей наилучшие результаты работы на обучающей и проверочной выборках. На финальном этапе точность классификации проверялась с использованием тестовой подвыборки VADSpeakersDB.

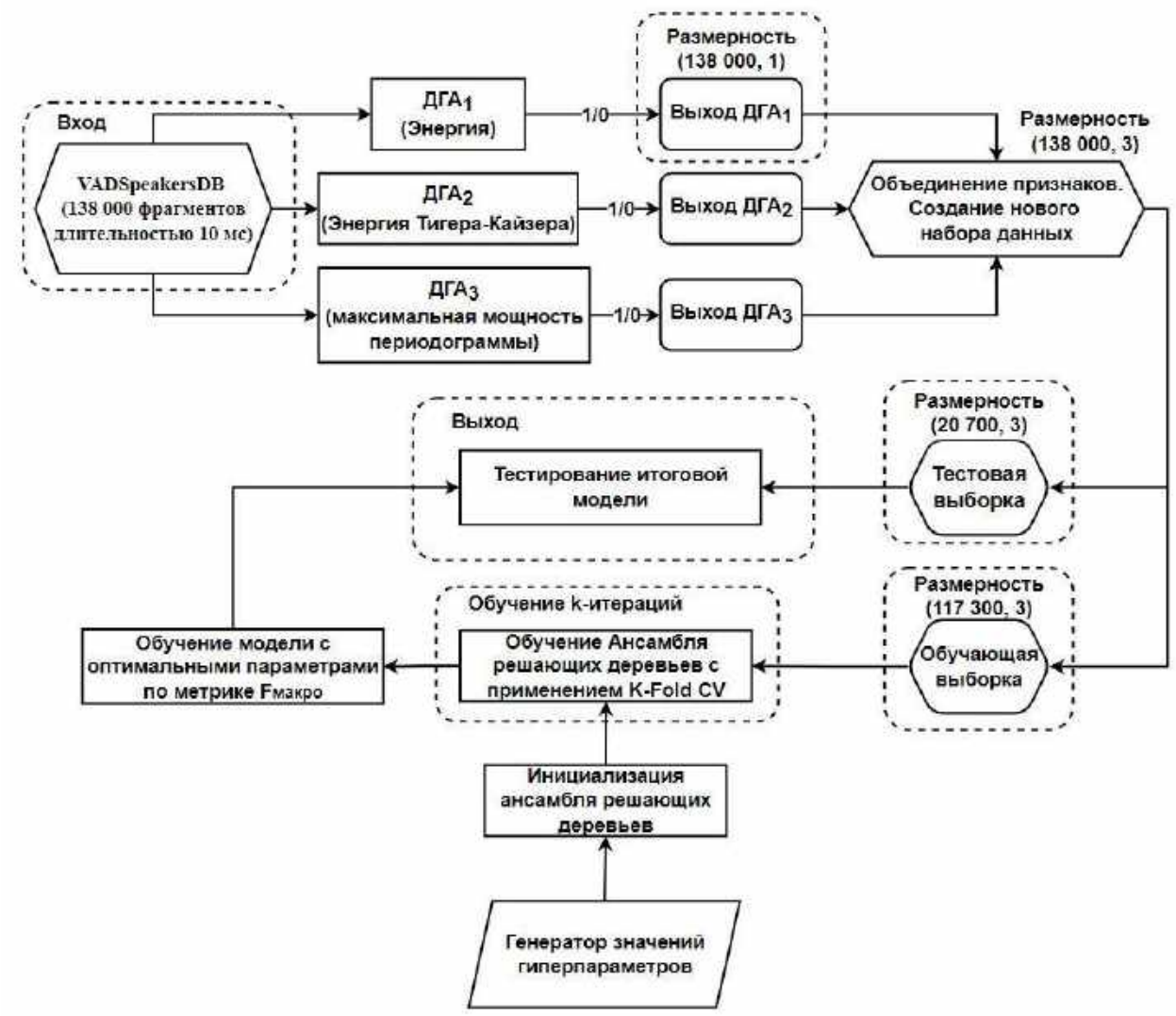


Рисунок 2.3 – Схема процесса обучения и тестирования КДГА

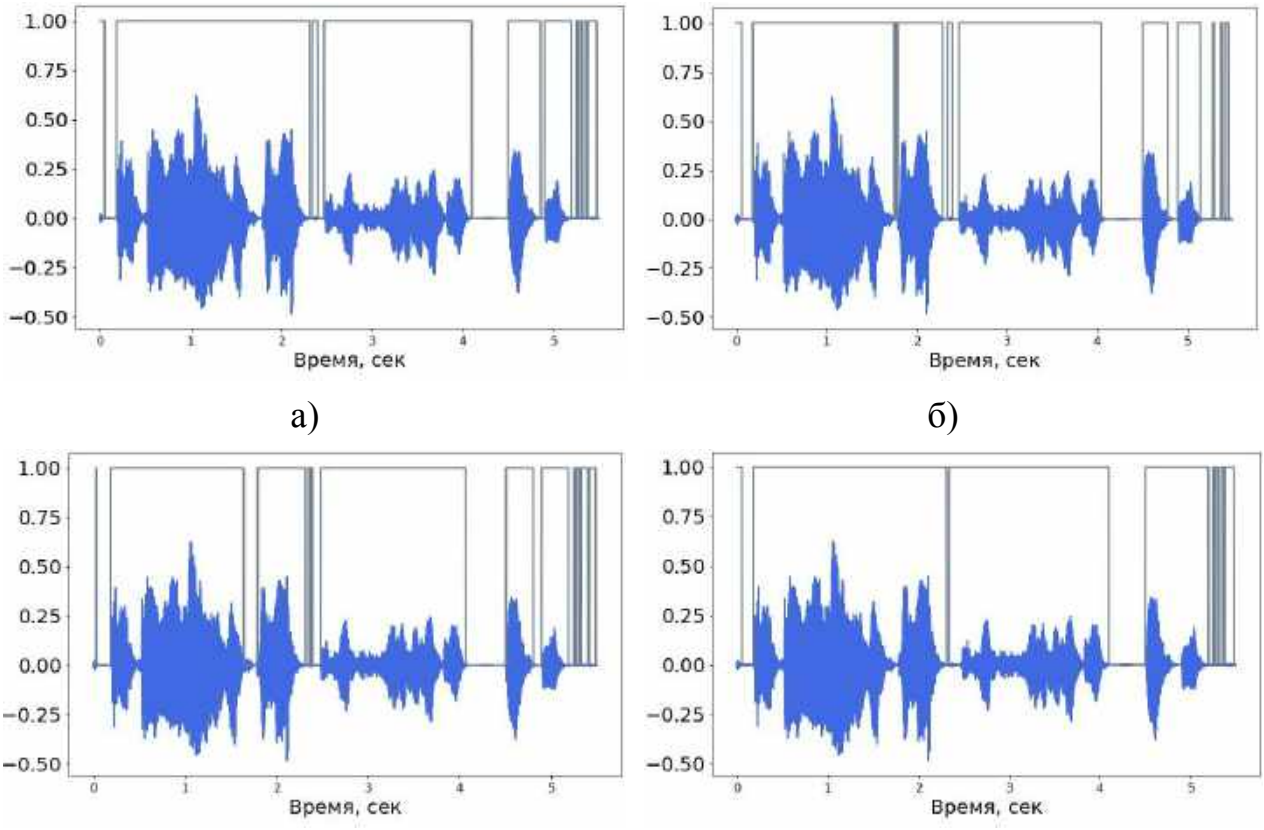
В Таблица 2.5 представлен сравнительный анализ разработанного алгоритма с классическими подходами. Из полученных результатов видно, что применение КДГА с последующим обучением обобщающего классификатора позволяет повысить точность детектирования голосовых фрагментов на 2-3%. Улучшение в точности работы обусловлено объединением трёх более «слабых учеников». Так, прогнозы, полученные на выходе каждого из детекторов, объединяются для анализа «сильным учеником». В процессе обучения модели сформированы весовые параметры для каждого детектора. В итоге каждый ДГА имеет индивидуальный вес при

принятии итогового решения. Комбинирование детекторов позволяет повысить точность определения речевых фрагментов.

Таблица 2.5 – Сравнительный анализ детекторов голосовой активности

Метрики	ДГА$_1$ ($k=2*10^{-4}$)	ДГА$_2$ ($\theta=3*10^{-6}$)	ДГА$_3$ ($\theta=2*10^{-3}$)	КДГА
acc	0,90	0,89	0,89	**0,91**
accb	0,88	0,88	0,87	**0,90**
F	0,93	0,92	0,92	**0,94**
$F_{макро}$	0,88	0,88	0,87	**0,90**

На Рисунке 2.4 представлены результаты работы рассмотренных детекторов. Визуально можно оценить улучшение в выделении зон голосовой активности при использовании алгоритма КДГА.



а) б)

в) г)

Рисунок 2.4 – Результаты работы детекторов голосовой активности: а) на основе анализа энергии; б) на основе анализа энергии Тигера-Кайзера; в) на основе частотного анализа сигнала; г) предложенный алгоритм КДГА

Разработанный детектор голосовой активности КДГА далее применялся на этапе предобработки речевых сигналов. Это позволило повысить точность работы алгоритмов идентификации личности путем очистки фонограмм от пауз, эффектов глотации, вдохов и шумов.

2.5 Обработка речевых сигналов набора FaceSpeechDB

2.5.1 Частотное представление речевых сигналов

Речь является акустическим сигналом и может быть описана функцией времени. Для успешного решения задачи распознавания диктора необходимо выполнить анализ частотных свойств речевого сигнала, характеризующих индивидуальные особенности голоса [91].

Мел-частотные кепстральные коэффициенты (МЧКК) являются одним из самых эффективных и популярных видов представления речевых данных. Данный подход используется как при решении задачи идентификации диктора, так и в системах распознавания речи [92-94]. Идея подхода заключается в том, что в процессе анализа частотного представления аудиосигнала учитываются особенности слуховой системы человека. В частности, человеческому уху легче определять небольшие изменения высоты звука на низких частотах, чем на высоких. Мел-шкала связывает воспринимаемую человеком высоту звука с его фактически измеренной частотой, как показано на Рисунке 2.5 [91].

Прямая и обратная связь между мел-значениями и частотой в герцах определяется следующими выражениями:

$$m = M(f) = 1127 \times \ln\left(1 + \frac{f}{700}\right), \quad (1)$$

$$f = M^{-1}(m) = 700 \times \left(exp\left(\frac{m}{1127}\right) - 1\right), \quad (2)$$

где m – частота в мелах, f – частота в герцах [91].

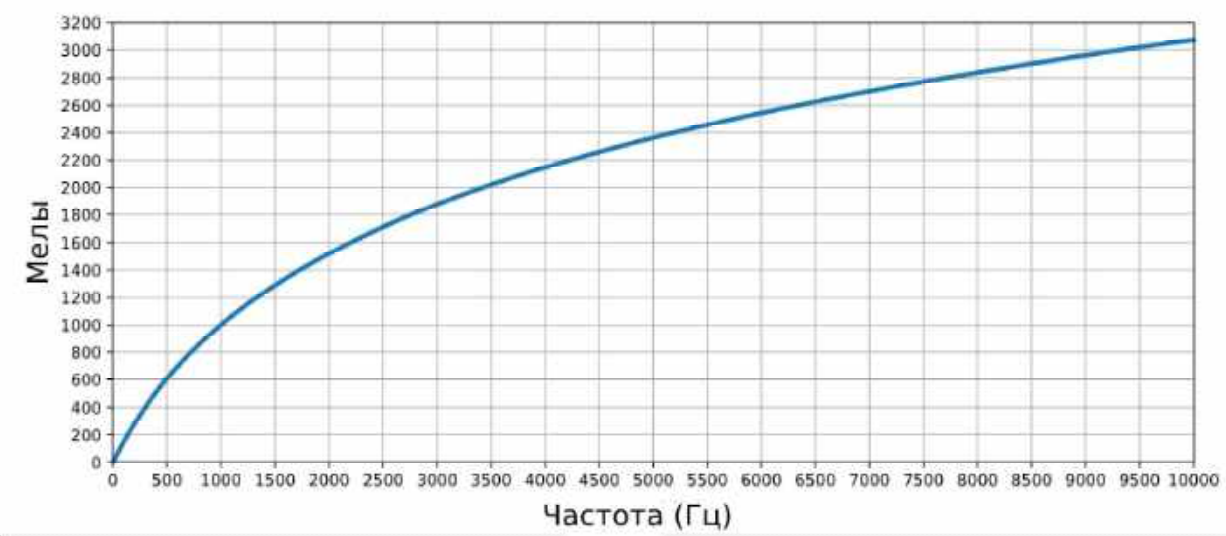


Рисунок 2.5 –Зависимость частоты в мелах от частоты в герцах

С помощью мел-шкалы формируется набор (банк) треугольных фильтров. Банк фильтров применяется к периодограмме, сворачивая ее вдоль оси частот (Рисунок 2.6).

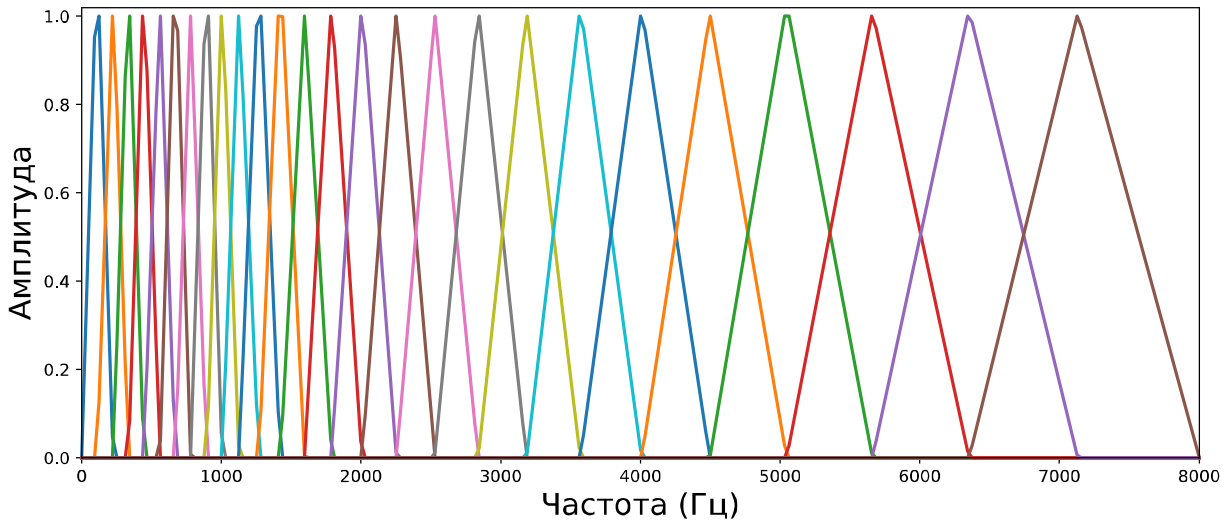


Рисунок 2.6 – Банк треугольных мел-фильтров

Фильтры позволяют аккумулировать спектральную энергию в области центральных частот. Количество частот определяется размером гребенки треугольных фильтров. Ширина пропускания каждого фильтра зависит от частотного диапазона. В районе нулевой частоты фильтры имеют более

узкую полосу, нежели в области высоких частот, что также определяется особенностями слуховой системы человека. На Рисунке 2.7 представлена структурная схема вычисления коэффициентов МЧКК.

Рисунок 2.7 – Схема вычисления коэффициентов МЧКК
(ДКП – дискретное косинусное преобразование)

Подробное описание алгоритма вычисления коэффициентов МЧКК представлено в Приложении А к настоящей работе.

2.5.2 Предобработка речевых сигналов

Для предобработки аудиоданных из набора FaceSpeechDB применялся предложенный алгоритм КДГА. Его использование позволило повысить качество анализируемых данных за счет очистки сигналов от участков, которые не содержат речь. На следующем этапе речевые сигналы разделялись на неперекрывающиеся фрагменты длительностью в 3 секунды. Далее выполнялся анализ подготовленного набора на смещение в данных. В частности, посчитано общее количество речевых фрагментов, описывающих каждого диктора. Анализ показал, что в среднем на класс приходится от 180 до 400 фрагментов. Для повышения чистоты проводимого эксперимента решено сделать сбалансированную проверочную и тестовую выборки. Для этого в каждый из наборов отобрано по 15 речевых фрагментов с каждого класса. В Таблице 2.6 представлено распределение речевых фрагментов по соответствующим выборкам.

Таблица 2.6 – Распределение речевых фрагментов по выборкам

-	Обучение	Валидация	Тест	Суммарно
Речевые фрагменты	28956	1560	1560	32076
Общее количество классов	104			

После распределения данных по выборкам вычислялись частотные представления речевых сигналов – спектрограмм и МЧКК. Для перехода в частотную область использовалось ДПФ со следующими параметрами: частота дискретизации 16 кГц, размер оконной функции 32 мс, что эквивалентно 512 временным отсчетам, шаг оконной функции 10 мс. Для формирования МЧКК использовался банк из 80 треугольных мел-фильтров. Таким образом, каждый речевой фрагмент длительностью в 3 секунды преобразовывался в спектрограмму и матрицу МЧКК, имеющие размеры 257x301 и 80x301 соответственно. После того как данные прошли этап предобработки, включающий выделение речевых фрагментов фонограммы и переход в частотную область, их можно использовать для обучения и тестирования нейросетевых алгоритмов.

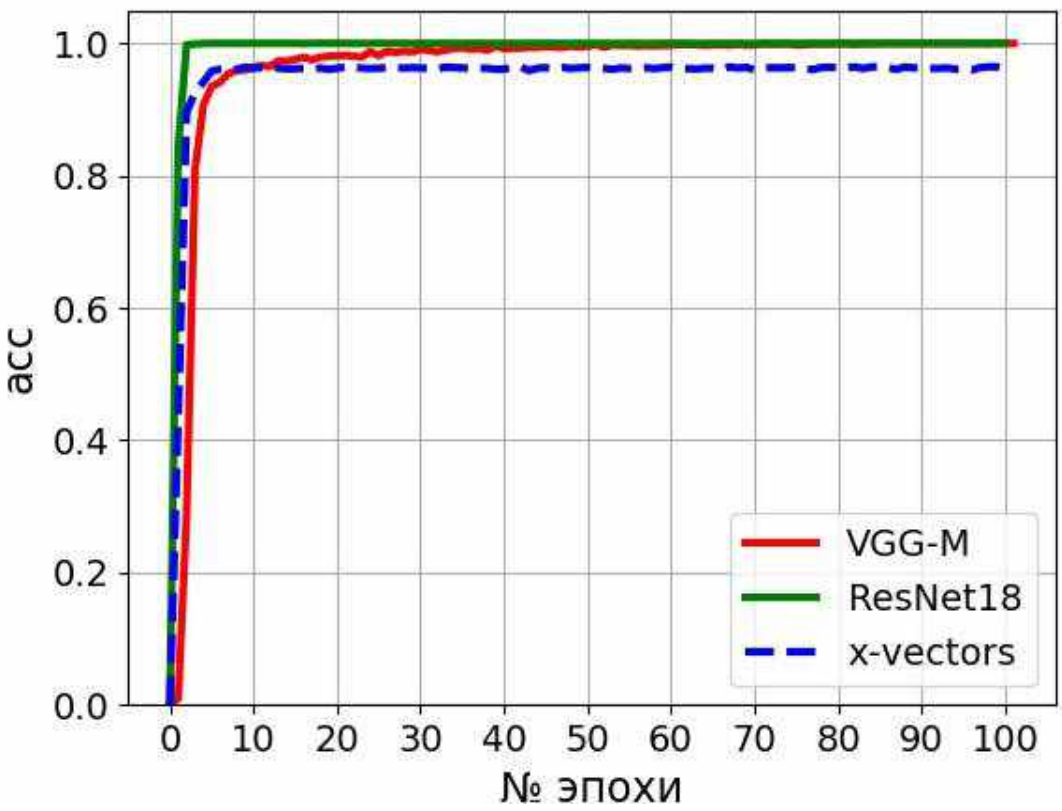
2.6 Тестирование стандартных нейросетевых алгоритмов идентификации диктора на наборе FaceSpeechDB

В качестве алгоритмов идентификации диктора использовались СНС на основе современных архитектур VGG-M, ResNet18 и x-vectors, описанные в п. 1.5. Обучение и тестирование нейросетевых алгоритмов выполнялось на основе аудиосигналов из подготовленного набора FaceSpeechDB. Данные проходили предобработку, которая включала три этапа: применение алгоритма КДГА, нарезка речевых сигналов на фрагменты равной длины, формирование частотных представлений для речевых фрагментов в виде спектрограмм (далее – СП) и МЧКК. В результате предобработки создано 32076 речевых фрагментов. В соответствие каждому фрагменту, длительностью в 3 секунды, сформирована спектрограмма и матрица МЧКК.

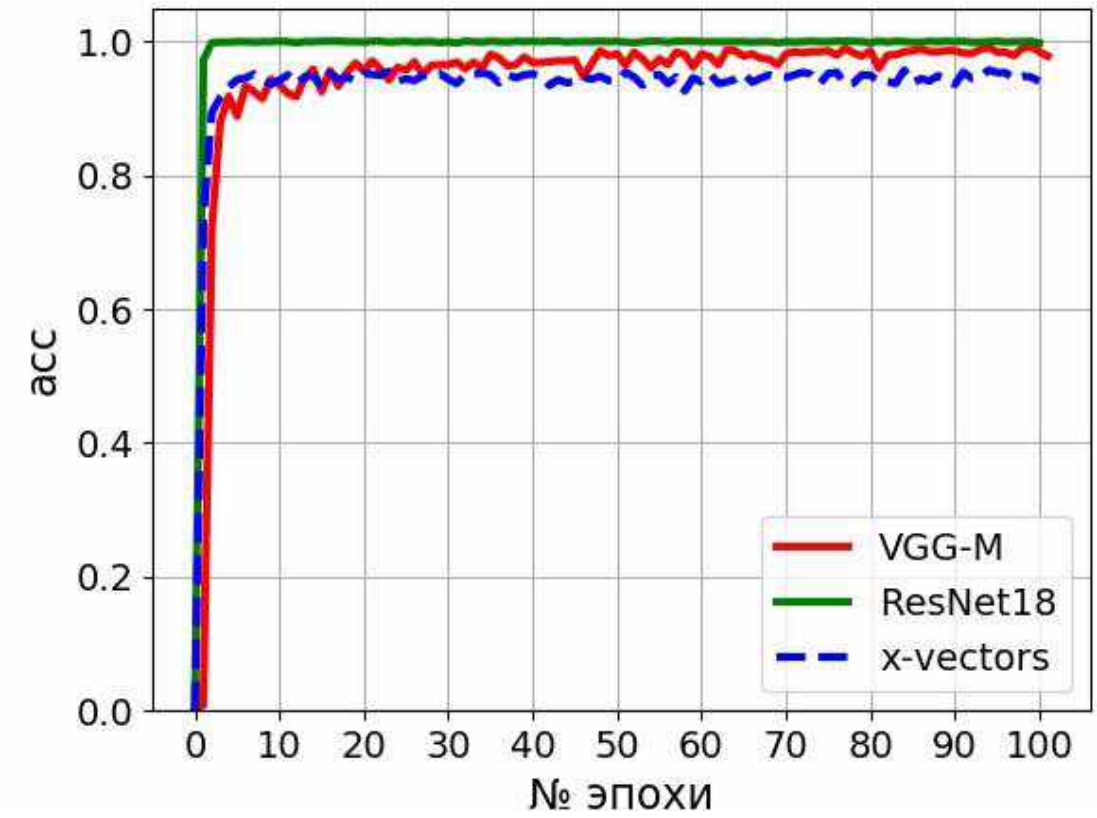
Для оценки качества работы алгоритмов идентификации личности в данном и последующих пунктах будет применяться метрика *acc,* описанная в п. 2.2.

Обучение и тестирование алгоритмов выполнялось с использованием GPU NVIDIA GeForce 2080 RTX, размер пакета данных (батча), который формируется и подается на вход СНС, составлял 32 примера. Этап, при котором анализируется один батч и корректируются весовые параметры нейронной сети, называется итерацией. Эпохой называется контрольная точка обучения, при которой весь набор исходных данных полностью проходит через нейронную сеть. В итоге обучение длилось в процессе выполнения более 90 000 итераций, что соответствует 100 эпохам.

На Рисунке 2.8 представлены кривые обучения алгоритмов идентификации диктора на основе анализа МЧКК. Анализируя их, можно сделать вывод, что в процессе обучения алгоритмы достаточно быстро сходятся. Дополнительно стоит отметить, что зависимость, описывающая результаты работы алгоритмов на валидационном наборе, указывает на отсутствие признаков переобучения.



а)



б)

Рисунок 2.8 – Зависимости изменения доли правильных ответов в процессе обучения нейросетевых алгоритмов идентификации диктора:
а) на обучающем наборе; б) на валидационном наборе

В Таблице 2.7 представлены результаты работы нейросетевых алгоритмов идентификации диктора с использованием набора FaceSpeechDB.

Таблица 2.7 – Результаты работы стандартных нейросетевых алгоритмов идентификации диктора для набора FaceSpeechDB

-	Кол-во параметров	Обучение	Валидация	Тест
VGG-M (СП)	16,7 млн.	99,92%	98,76%	98,17%
VGG-M (МЧКК)	14,6 млн.	98,47%	98,63%	98,24%
ResNet18 (СП)	11,8 млн.	94,46%	72,31%	71,83%
ResNet18 (МЧКК)	11,8 млн.	99,99%	99,90%	99,74%
x-vectors (СП)	9,8 млн.	92,60%	90,49%	91,34%
x-vectors (МЧКК)	9,4 млн.	96,53%	94,34%	95,70%

Исследуемые алгоритмы показали высокую точность – более 90%. Исключением является алгоритм анализа спектрограмм на базе ResNet18. Следует отметить, что использование МЧКК в качестве представления речевого сигнала позволяет улучшить точность идентификации по сравнению с подходом, основанным на спектрограммах.

Несмотря на высокую точность работы, алгоритмы на основе рассмотренных архитектур являются требовательными к вычислительным ресурсам. Количество весовых параметров для каждой модели насчитывает десятки миллионов. Применение данных алгоритмов накладывает ряд ограничений на аппаратные ресурсы устройств и, как следствие, на скорость работы конечной системы. Это также увеличивает стоимостные характеристики программно-аппаратных решений, поскольку для их запуска в режиме реального времени требуются высокопроизводительные GPU. С учетом этого задача разработки быстрого и робастного нейросетевого алгоритма идентификации диктора является актуальной.

2.7 Разработка и тестирование алгоритма идентификации диктора на основе x-векторной системы

Разработка нового нейросетевого алгоритма идентификации диктора осуществлялась на основе x-векторной системы. Данное решение обусловлено рядом причин. Во-первых, такой алгоритм показал высокую точность работы в задаче идентификации диктора. Во-вторых, топология сети основывается на операции одномерной свертки, которая является менее вычислительно затратной, в отличие от двумерной версии, которая используется в VGG-M или ResNet18. Кроме того, данный подход анализирует временные фрагменты речевого сигнала по всей полосе частот, тогда как решения на базе двумерной свертки выполняют исследование локальных областей и их особенностей. В качестве частотного представления речевого сигнала использовались МЧКК. Структурная схема разработанной сверточной нейронной сети X-Speech представлена на Рисунке 2.9.

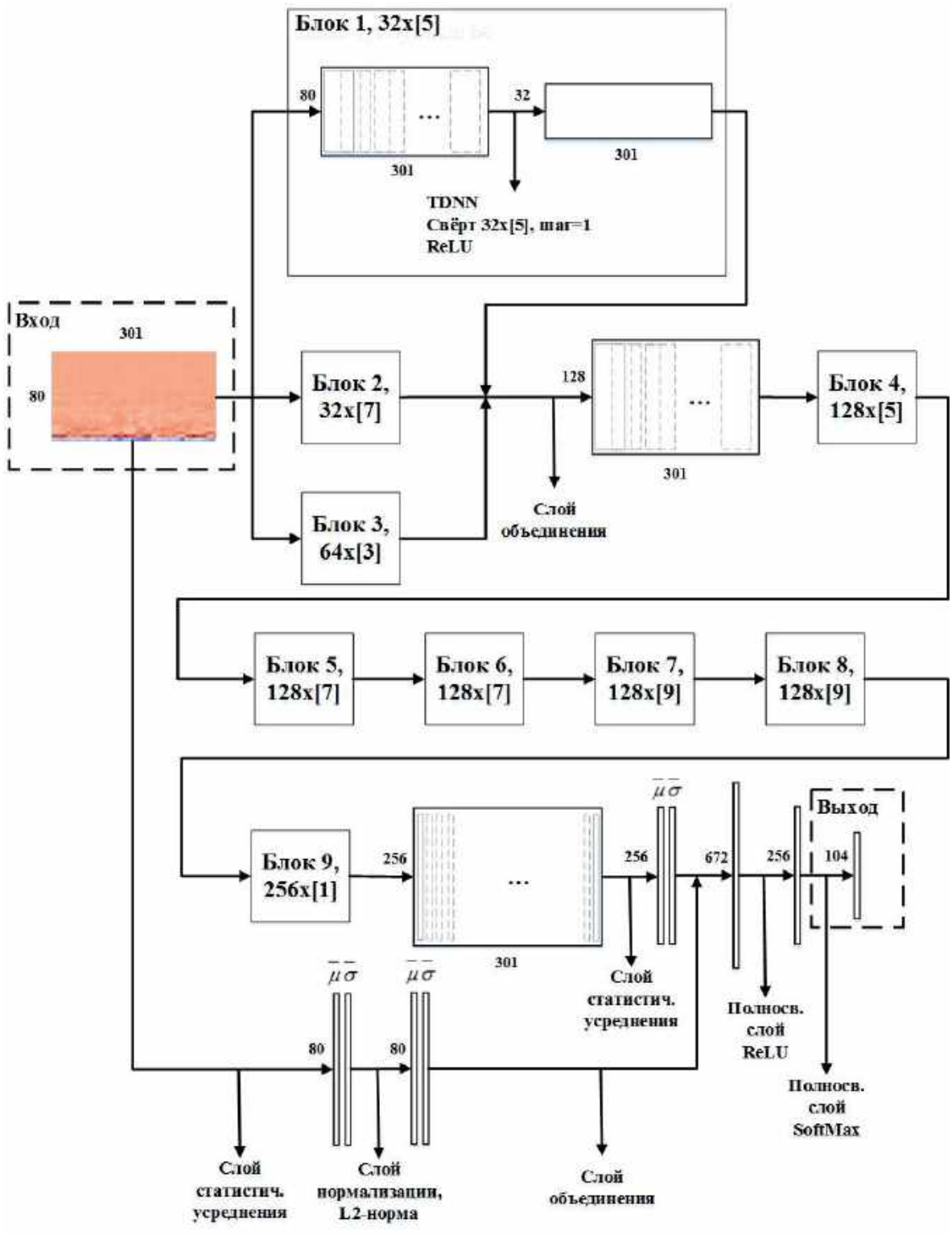


Рисунок 2.9 – Архитектура нейронной сети X-Speech

На входе сеть имеет 3 блока нейронных сетей с временной задержкой. Первый и второй блоки имеют по 32 фильтра с размерами фреймов 5 и 7, соответственно. Один фрейм спектрограммы представляет собой частотное разложение речевого сигнала в пределах ширины окна, которое используется

при анализе сигнала во временной области. В исследовании ширина окна составляет 512 отсчетов, что соответствует временному участку исходного сигнала длительностью в 32 мс. Третий блок состоит из 64 фильтров. Выходы с данных блоков объединяются, формируя карту размером 128x301. Далее последовательно идут 5 блоков, где количество фильтров остается неизменным и равняется 128. Далее идет «Блок 9», который увеличивает ширину карты признаков за счет использования 256 фильтров.

На следующем уровне используется слой статистического усреднения, который высчитывает среднее значение и среднеквадратическое отклонение для каждого признака. Вычисление осуществляется вдоль временной оси. В итоге формируются два 256-мерных вектора – $\overline{\mu_1}$ и $\bar{\sigma}_1$. Далее данные векторы объединяются с векторами $\overline{\mu_2}$ и $\bar{\sigma}_2$. Последние формируются в результате использования ещё одного слоя статистического усреднения, на вход которого напрямую поступает карта МЧКК. Вследствие объединения $\overline{\mu_1}$, $\bar{\sigma}_1$, $\overline{\mu_2}$ и $\bar{\sigma}_2$ формируется общий вектор признаков длиною 672. Далее общий вектор поступает на полносвязный слой, после чего используется выходной слой с функцией активации SoftMax, где количество выходных нейронов соответствует количеству исследуемых классов – 104. Все сверточные слои в качестве функции активации используют ReLU. Общее количество слоев составляет 18, из которых 9 являются сверточными.

В Таблице 2.8 представлено сравнение по количеству параметров для нейросетевых алгоритмов идентификации личности на основе голосовой биометрии. Результаты исследования показали, что разработанный алгоритм менее требователен к вычислительным ресурсам. Предлагаемая нейросетевая архитектура содержит примерно на порядок меньше весовых параметров относительно своих аналогов.

Таблица 2.8 – Сравнительный анализ количества параметров нейросетевых алгоритмов идентификации диктора

-	VGG-M (СП)	VGG-M (МЧКК)	ResNet18 (МЧКК)	X-vectors (СП)	X-vectors (МЧКК)	X-Speech (МЧКК)
Кол-во параметров	16,7 млн.	14,6 млн.	11,8 млн.	9,8 млн.	9,4 млн.	0,89 млн.
Кол-во слоев	11	11	90	16	16	18

Скорость распознавания – одна из важнейших характеристик систем идентификации диктора. Однако не менее значимым свойством является устойчивость к шумам. Телефонный и микрофонный каналы связи подвержены негативному воздействию шумов и помех. Их источником может быть окружающая среда или эффекты, возникающие при передаче информации в канале связи. Также чистота речевого сигнала определяется качеством записывающих, ретранслирующих и принимающих устройств. Вследствие этого для повышения обобщающей способности обучаемой модели на базе разработанной сверточной нейронной сети X-Speech, применялся метод размножения (аугментации) данных. В процессе обучения речевые сигналы подвергались искажениям и преобразованиям: добавление аддитивного белого гауссовского шума; смещение сигнала по времени; использование эффекта реверберации, позволяющего изменять свойства аудиосигнала, меняя представления о масштабе и глубине акустической сцены; использование медианной фильтрации для разделения гармонических и ударных компонент сигнала.

Для моделирования шумов, вызванных окружающей средой, использовался открытый набор аудиосигналов Urban Sound Dataset (UrbanSound8K). Набор включает 8 732 записи, каждая длиной не менее 4 секунд, представляющие собой характерные для города звуки. Речевые сигналы набора FaceSpeechDB случайным образом смешивались с сигналами из набора UrbanSound8K. Стоит отметить, что уровень зашумления речевых

сигналов контролировался величиной отношения сигнал/шум с допустимым интервалом значений 6-40 дБ [95-98].

Обучение разрабатываемого алгоритма осуществлялось на основе анализа МЧКК размером 80x301. Изначально обучающий набор речевых сигналов содержал 28 956 примеров, но, поскольку для повышения обобщающей способности использовалась аугментация данных, то итоговый размер обучающей выборки составил более 1,4 млн. речевых сигналов. В процессе обучения размер батча составлял 32 экземпляра МЧКК. Обучение останавливалось после совершения порядка 45 000 итераций, что соответствует 50 эпохам.

Для проведения анализа робастности алгоритмов идентификации диктора обрабатывался исходный тестовый набор («Тест»). Для этого 1 560 речевых сигналов из него зашумлялись с использованием подхода на основе аугментации данных. В результате сформирован дополнительный тестовый набор «Тест-Ш».

В Таблице 2.9 представлен анализ робастности работы алгоритмов идентификации диктора с использованием оригинального и зашумленного набора речевых сигналов. Алгоритм на базе предложенной СНС X-Speech имеет высокую точность идентификации на исходном тестовом наборе данных «Тест» – 98%, уступая лишь решению на базе более сложной сети ResNet18. Анализ на зашумленном тестовом наборе «Тест-Ш» показал, что разработанный алгоритм более робастен к искажениям и помехам, чем исследуемые аналоги. Точность работы алгоритма в этих условиях превышает 91%, что превосходит аналоги на 5% и более.

Таблица 2.9 – 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%

В результате проведенного исследования можно сделать вывод, что нейросетевой алгоритм на базе сети X-Speech демонстрирует высокую точность идентификации, а также устойчивость к шумам и при этом содержит в 10-20 раз меньше весовых параметров в сравнении с рассмотренными нейросетевыми аналогами.

2.8 Краткие выводы

Результаты проведенных исследований по разработке нейросетевых алгоритмов идентификации диктора с использованием речевых сигналов позволяют сделать следующие основные выводы:

- Разработанный алгоритм КДГА улучшает качество речевых сигналов за счет фильтрации фонограмм от пауз, эффектов глотации, вдохов и шумов. Предложенный алгоритм повышает точность определения фрагментов голосовой активности на 2-3% в сравнении с имеющимися аналогами.
- Разработанный нейросетевой алгоритм на основе x-векторной системы может быть использован для автоматической идентификации диктора с применением анализа речевых сигналов. При сохранении точности идентификации на уровне до 98-99%, разработанный алгоритм имеет в 10-20 раз меньше весовых параметров, что дает ему преимущество в скорости работы относительно существующих аналогов.
- Разработанный нейросетевой алгоритм может быть использован в условиях действия шумов и помех, где деградация в точности его работы составляет в среднем 7%. В этих условиях он превосходит аналоги на 5% и более.

ГЛАВА 3
ИССЛЕДОВАНИЕ НЕЙРОСЕТЕВЫХ АЛГОРИТМОВ ИДЕНТИФИКАЦИИ ЛИЧНОСТИ НА ОСНОВЕ АНАЛИЗА ИЗОБРАЖЕНИЙ ЛИЦ

3.1 Вводные замечания

В данной главе рассматривается задача идентификации личности с использованием цифровых изображений лиц. В качестве исходных данных для обучения и тестирования нейросетевых алгоритмов используются изображения из подготовленного аудиовизуального набора FaceSpeechDB. Вначале описывается этап предобработки изображений с применением алгоритма обнаружения (детектирования) лиц. Предобработка обусловлена необходимостью определения области нахождения лица на изображении с последующим его отделением от остального фона. Далее выполняется сравнение работы алгоритмов идентификации личности на основе архитектур СНС VGG16, ResNet50 и SENet50, которые подробно описаны в п. 1.4. Проводится разработка и исследование нового алгоритма идентификации лиц на базе СНС. Также рассматривается робастная модификация алгоритма, способного работать в условиях использования медицинской маски и перекрытия до 70% лица.

3.2 Алгоритмы обнаружения лиц на изображениях

Работа алгоритмов распознавания лиц неразрывно связана с задачей обнаружения лиц на изображении. Прежде чем распознавать личность, необходимо определить область нахождения лица на изображении. В работе для обнаружения лиц использовался алгоритм поиска на основе сверточной нейронной сети MTCNN (Multi-task Cascaded Convolutional Networks) схематически показанный на Рисунке 3.1 [99, 100].

Рисунок 3.1 – Структурная схема алгоритма обнаружения лиц на базе нейронной сети MTCNN

Алгоритм обнаружения лиц представляет собой каскад из нескольких СНС. На первом этапе исходное изображение подвергается операции изменения размера для того, чтобы построить пирамиду изображений. Далее пирамида анализируется первой, самой компактной сетью из каскада P-Net (Proposal Network). Каждое из изображений анализируется скользящим окном, имеющим разрешение 12x12 пикселей, что соответствует размеру входа для P-Net. На выходе сети формируются значения, характеризующие области-кандидаты наличия лиц. Области, имеющие сильное пересечение между собой, удаляются с использованием NMS-алгоритма (non-maximum suppression, NMS) [101]. На втором этапе оставшиеся фрагменты-кандидаты приводятся к единому разрешению 24x24 пикселей и передаются на вход второй, более сложной сети R-Net (Refinement Network). Она анализирует данные фрагменты и отклоняет большое количество ложных срабатываний, выделенных P-Net. Оставшиеся кандидаты также анализируются NMS-алгоритмом. На третьем этапе фрагменты приводятся к формату 48x48 пикселей, где анализируются самой производительной сетью O-Net (Output Network). Она используется для принятия финального решения и вывода

координат объектов, которые определены как лица. Дополнительно сеть O-Net вычисляет координаты пяти ключевых лицевых точек, соответствующих расположению глаз, носа, уголков губ. Несмотря на многоступенчатую каскадную архитектуру, рассматриваемый нейросетевой алгоритм не требователен к вычислительным ресурсам и способен работать в режиме реального времени.

Видеоданные из набора FaceSpeechDB обработаны с помощью алгоритма на базе сети MTCNN. Все видеопоследовательности из подготовленного набора обрабатывались независимо. На вход алгоритма подавался каждый 90-й кадр, то есть поиск лиц осуществлялся с временным промежутком в 3 секунды. Это связано с тем, что в процессе записи человек за короткий промежуток времени слабо изменяет свое положение и мимику лица. Если уменьшить интервал, то набор лиц будет содержать большое количество одинаковых или сильно коррелированных изображений, что негативно повлияет на результаты обучения алгоритмов распознавания лиц. В частности, это приводит к эффекту переобучения. Дополнительно, чтобы исключить ситуации ложного срабатывания, доверительный порог на каждом из этапов устанавливался равным 0,7. Если алгоритм обнаружения на каком-то из этапов определял наличие лица на изображении с уверенностью меньше заданного порога, то данный образец пропускался и не учитывался. На Рисунке 3.2 представлены примеры выделенных изображений лиц.

В результате подготовлен набор для разработки алгоритмов идентификации лиц, включающий 35 660 изображений. В ходе визуальной проверки установлено, что все изображения действительно содержат исключительно лица людей. Для стандартизации изображений все образцы приведены к единому разрешению – 320x320 пикселей.

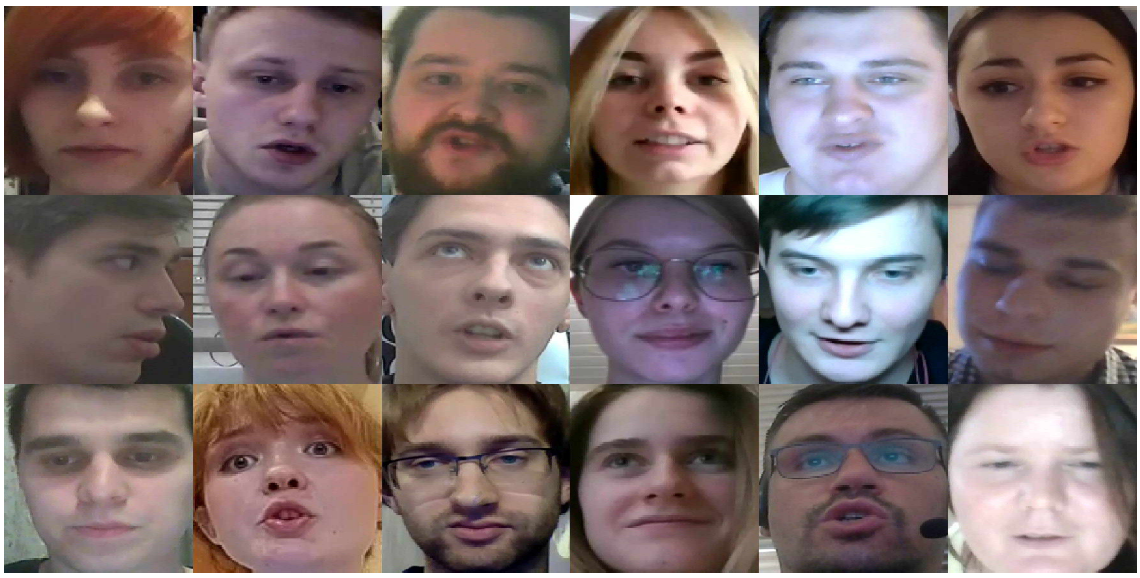


Рисунок 3.2 – Результаты работы алгоритма обнаружения лиц

Дополнительно, как и в случае с речевыми сигналами, выполнено исследование подготовленного набора изображений на наличие эффекта дисбаланса в данных. В частности, посчитано количество изображений лиц, описывающих каждую из личностей набора FaceSpeechDB. Анализ показал, что в среднем на класс приходится от 280 до 450 изображений. В данных прослеживается эффект смещения, поэтому фомирование проверочной и тестовой выборки выполнялось по аналогичному принципу подготовки речевых сигналов. Для этого в каждый из этих наборов отобрано по 15 изображений от каждого класса. В Таблице 3.1 представлено статистика распределения изображений лиц по соответствующим выборкам.

Таблица 3.1 – Распределение изображений лиц по выборкам

-	Обучение	Валидация	Тест	Суммарно
Изображения лиц	32540	1560	1560	35660
Общее количество классов	104			

В следующем пункте исследования выполняется тестирование стандартных нейросетевых алгоритмов идентификации лиц на подготовленном наборе изображений лиц.

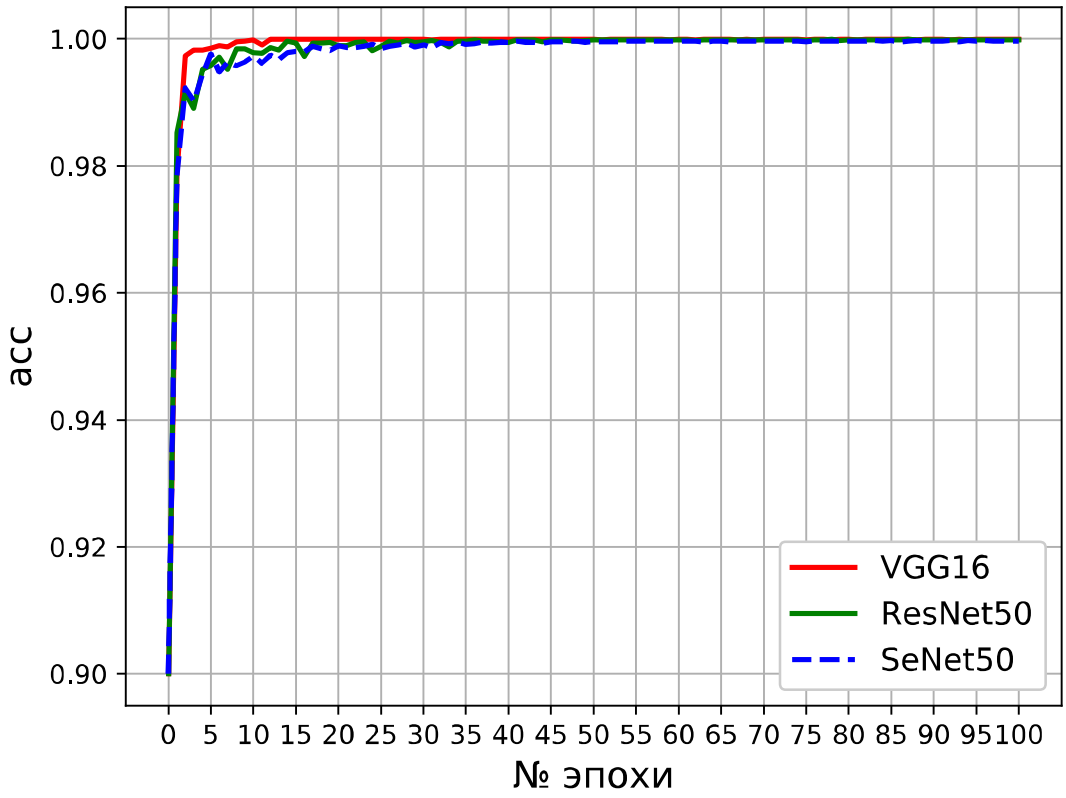
3.3 Тестирование стандартных нейросетевых алгоритмов идентификации лиц на наборе FaceSpeechDB

Обучение глубоких моделей нейронных сетей является вычислительно затратным процессом, поскольку количество настраиваемых весов составляет десятки и сотни миллионов параметров. Для решения данной проблемы применяют подход на основе трансферного обучения [102, 103]. Так, в исследовании использовались предобученные модели на основе ранее рассмотренных в п. 1.4 архитектур VGG16, ResNet50 и SENet50. Их сверточные слои отлично подходят для выделения признаков, описывающих особенности лиц. Выбранные реализации заранее обучены на большом стандартном наборе лиц VGG-Face [33].

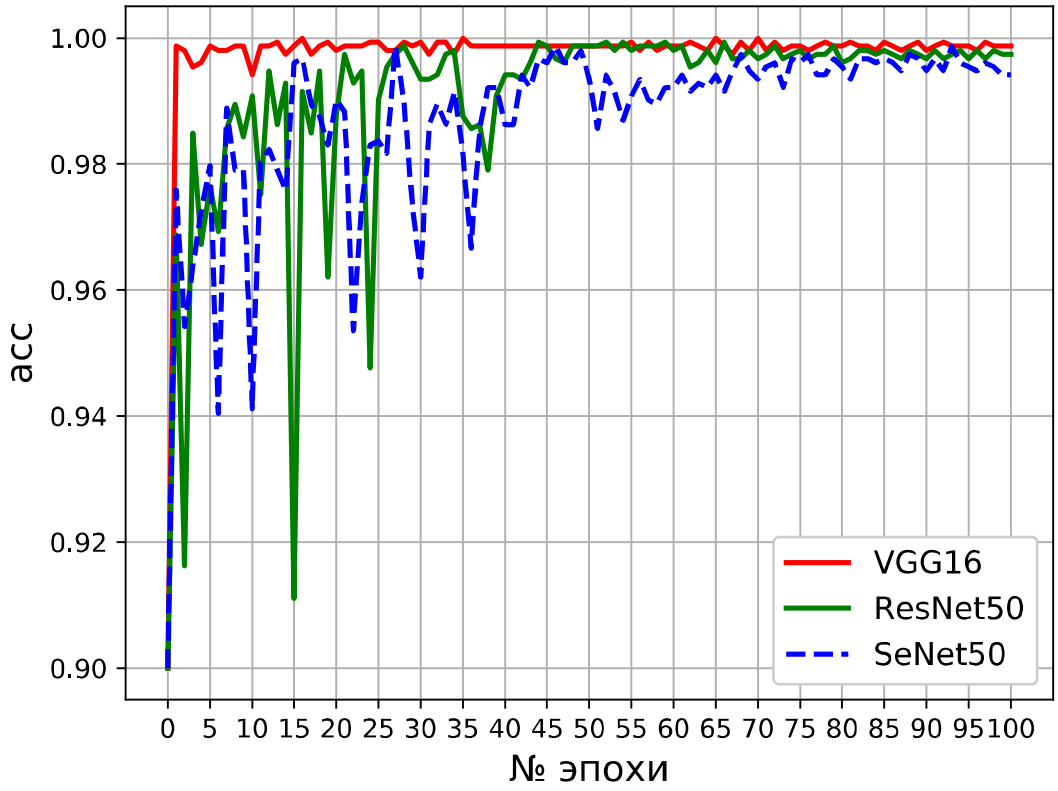
Трансферное обучение выполнялось с помощью набора FaceSpeechDB. Для этого сверточные слои выделялись из предобученных моделей и далее объединялись с классификатором, состоящим из последовательно идущих полносвязных слоев. На Рисунке 3.3 изображена схема классификатора, построенного таким образом. Выходом является 104-мерный softmax-слой.

Рисунок 3.3 – Схема классификатора, используемого в процессе трансферного обучения

На Рисунке 3.4 показаны кривые изменения доли правильных ответов в процессе трансферного обучения сетей. Из приведенных зависимостей видно, что модели показывают высокую точность классификации уже на первых эпохах обучения. В частности, на первой эпохе обучения точность классификации на валидационном множестве составляет более 96%. Высокая скорость сходимости обусловлена тем, что весовые параметры сверточных слоев заранее адаптированы под задачу анализа изображений лиц, поэтому данные слои способны качественно выделять лицевые признаки. Адаптировать сверточные слои в процессе обучения не имеет смысла, поэтому они не изменяются. Обучение и коррекция весовых параметров выполняется исключительно в полносвязных слоях, которые являются классификатором, анализирующим признаковое представление лиц. В итоге трансферное обучение позволяет ускорить процесс сходимости нейросетевых алгоритмов и снизить вычислительную сложность за счет уменьшения количества обучаемых параметров сети.



а)



б)

Рисунок 3.4 –Изменение доли правильных ответов в процессе трансферного обучения: а) на обучающем наборе; б) на валидационном наборе

Рассмотренные решения демонстрируют высокие показатели точности и могут применяться в задаче биометрической идентификации. Однако алгоритмы на основе данных архитектур имеют высокие требования к вычислительным ресурсам и не всегда могут быть использованы при работе в режиме реального времени.

3.4 Разработка и исследование нейросетевого алгоритма идентификации лиц на основе сети CNN-Face

На Рисунке 3.5 представлена архитектура предложенной сверточной нейронной 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. В качестве отклика на входное изображение отбирается наибольшее значение, которое соответствует конкретному классу.

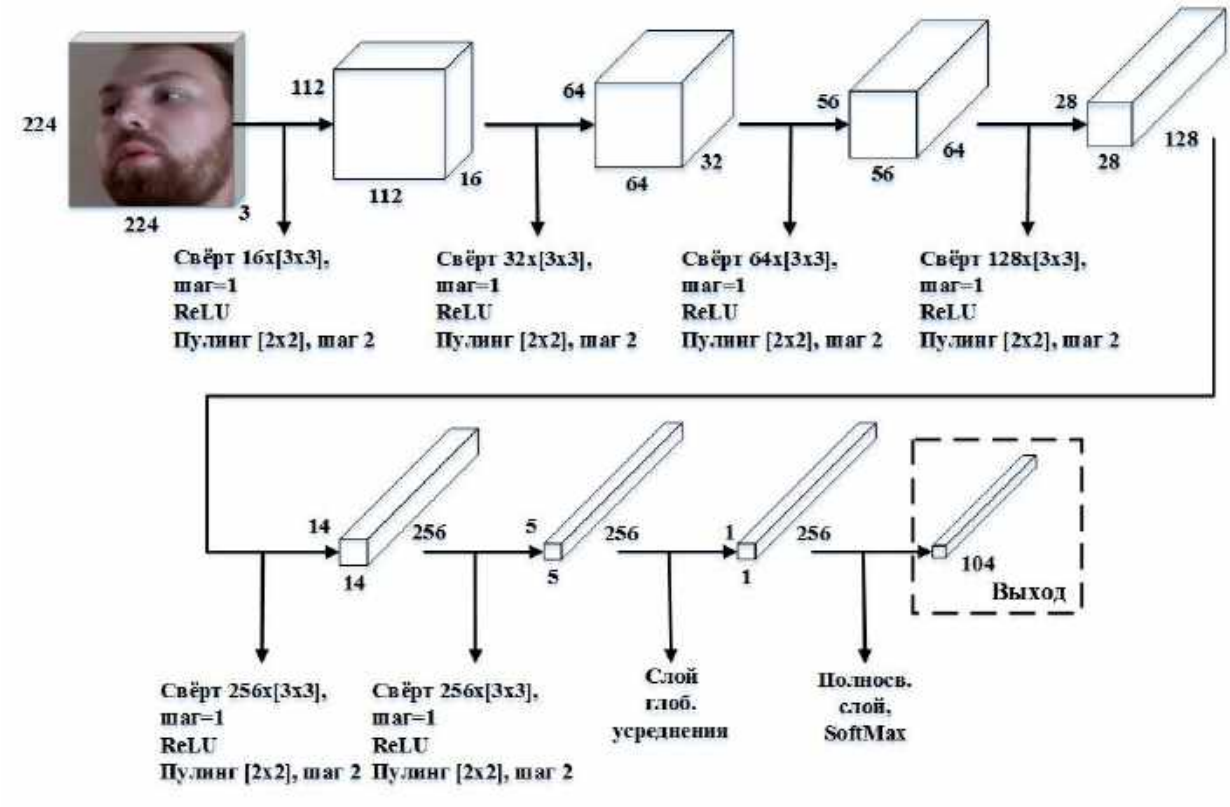


Рисунок 3.5 – Архитектура разработанной сети CNN-Face

Важно отметить, что данное решение более чем в 25-28 раз проще по вычислительной сложности относительно рассмотренных ранее нейросетевых алгоритмов (Таблица 3.2).

Таблица 3.2 – Сравнительный анализ параметров нейронных сетей в задаче идентификации лиц

	VGG16	ResNet50	SeNet50	CNN-Face
Количество весовых параметров	28 млн.	25 млн.	27 млн.	1 млн.
Количество слоев в сети	23	178	290	18

В Таблице 3.3 представлен анализ работы стандартных нейросетевых алгоритмов идентификации лиц и алгоритма на основе сети CNN-Face.

Точность на тестовом наборе данных для всех алгоритмов составляет более 99%, что указывает на высокую обобщающую способность алгоритмов. Алгоритм на базе предложенной архитектуры CNN-Face, несмотря на свою компактность, также демонстрирует высокую точность идентификации лиц.

Таблица 3.3 – Сравнительный анализ точности работы нейросетевых алгоритмов идентификации лиц

	Обучение	Валидация	Тест
VGG16	99,99%	99,86%	99,93%
ResNet50	99,78%	99,73%	99,87%
SeNet50	99,74%	99,41%	99,35%
CNN-Face	99,99%	99,93%	99,87%

Результаты исследования показывают, что алгоритм на базе разработанной архитектуры CNN-Face может быть использован при проектировании эффективных систем биометрической идентификации личности на основе анализа лиц. Важно отметить, что применение данного решения позволяет существенно снизить требования к аппаратным ресурсам в сравнении с существующими нейросетевыми аналогами.

3.5 Исследование и модификация алгоритма идентификации лиц в ситуации наличия медицинской маски

Во время пандемии COVID-19 возникли новые вызовы для систем биометрической идентификации личности [113, 114]. Часто используемая медицинская маска способна перекрывать до 70% лица. В результате существенная часть информации, описывающая исключительные свойства лица, такие как губы, нос и подбородок, остается перекрытой областью маски.

Для проведения анализа робастности алгоритмов идентификации лиц выполнена модификация тестового набора из 1560 изображений

(Рисунок 3.6). К каждому изображению лица в графическом редакторе ручным способом добавлены медицинские маски.

Рисунок 3.6 – Изображения лиц с добавленными медицинскими масками

В итоге сформирован новый тестовый набор изображений лиц («Тест-ММ»). Анализ робастности работы алгоритмов идентификации в условиях наличия медицинской маски показал, что их точность снизилась для всех рассматриваемых решений. Так, алгоритмы на базе глубоких архитектур VGG16, ResNet50 и SeNet50 показывают деградацию в точности работы на 10-22%. Низкую устойчивость в условиях наличия медицинской маски показал алгоритм на базе архитектуры CNN-Face. Его уровень точности снизился более чем на 26%.

С целью повышения робастности работы разрабатываемого алгоритма в условиях использования медицинских масок проводилась модификация архитектуры СНС CNN-Face (Рисунок 3.7). Разработан алгоритм на основе СНС, состоящей из двух эквивалентных модулей (далее – CNN-FaceMask).

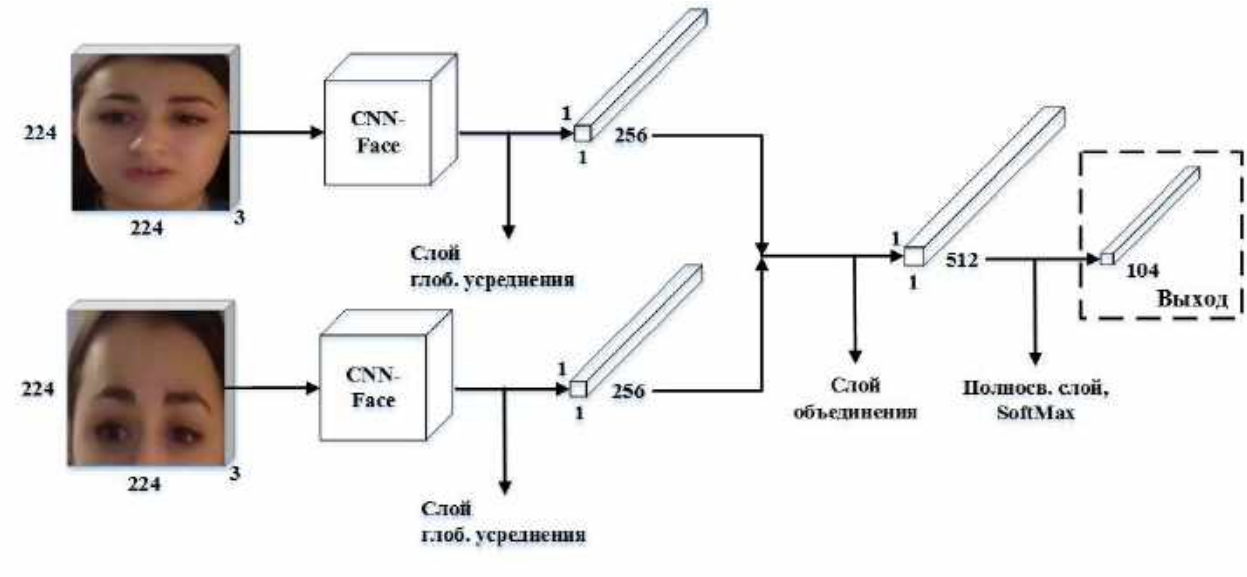


Рисунок 3.7 – Архитектура разработанной сети CNN-FaceMask

Основная идея модификации заключается в дополнительном анализе видимой области лица (лба и линии глаз) в условиях наличия маски. При этом важно также сохранить полноценный анализ всего обнаруженного лица, поскольку алгоритм должен определять личность не только в ситуации наличия медицинской маски, но и в условиях полной видимости. Каждый из модулей представляет собой сеть архитектуры CNN-Face, описанной ранее. Сеть имеет два входа и единый выход.

Обучение осуществлялось с использованием набора пар изображений (Рисунок 3.8). Одно из них представляло собой полноценную область лица, как это описывалось ранее. Дополнительно к каждому такому изображению сформировано и определено усеченное изображение, описывающее отмасштабированную верхнюю часть лица. В итоге обучающий набор увеличился в 2 раза и составил 65080 изображений.

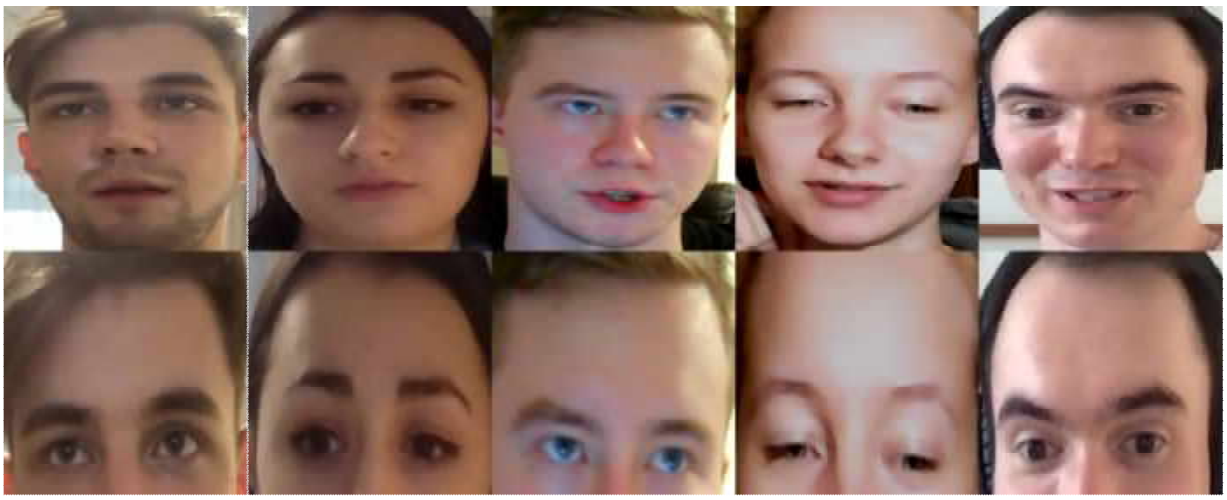


Рисунок 3.8 – Примеры изображений для обучения сети CNN-FaceMask

В Таблице 3.4 представлен анализ работы алгоритмов идентификации лиц в условиях отсутствия и использования медицинских масок.

Таблица 3.4 – Анализ устойчивости 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%

Модификация нейросетевого алгоритма CNN-Face позволила повысить устойчивость работы предлагаемого решения. Точность идентификации в условиях наличия маски на тестовом наборе данных составляет более 93%, что превосходит лучшие аналоги почти на 3%. Также важно отметить, что дополнительный анализ области лба и линии глаз несущественно повлиял на вычислительную сложность алгоритма. Количество весовых параметров для нейронной сети на базе CNN-FaceMask составляет 2 млн., что по-прежнему значительно ниже, чем у алгоритмов, построенных на стандартных архитектурах VGG16, ResNet50 и SeNet50.

3.6 Краткие выводы

Результаты проведенных исследований по разработке нейросетевых алгоритмов идентификации личности с использованием изображений лиц позволяют сделать следующие основные выводы:

- Разработанный алгоритм на базе предложенной архитектуры CNN-Face может эффективно использоваться в задачах лицевой биометрии. При высокой точности идентификации на тестовом наборе данных на уровне 99%, разработанный алгоритм содержит в 25-30 раз меньше весовых параметров, что дает ему существенное преимущество в скорости работы относительно имеющихся аналогов.
- Предложенный нейросетевой алгоритм на базе модифицированной архитектуры CNN-FaceMask демонстрирует наилучшую в рассматриваемом классе робастность к присутствию медицинской маски на лице человека. Деградация в точности идентификации составляет менее 7%, что превосходит аналогичные показатели для стандартных нейросетевых алгоритмов на 3% и более.

ГЛАВА 4
ИССЛЕДОВАНИЕ МУЛЬТИМОДАЛЬНЫХ АЛГОРИТМОВ ИДЕНТИФИКАЦИИ ЛИЧНОСТИ

4.1 Построение мультимодальных алгоритмов на основе сверточных нейронных сетей

При разработке мультимодальных алгоритмов идентификации на основе СНС наиболее популярным типом объединения модальностей является объединение на уровне признаков [59-64]. Существует несколько вариантов построения систем такого типа. Первый из них основывается на операции конкатенации. Признаки, сформированные в результате анализа независимых модальностей, объединяются в один вектор признаков с использованием слоя конкатенации, как показано на Рисунке 4.1.

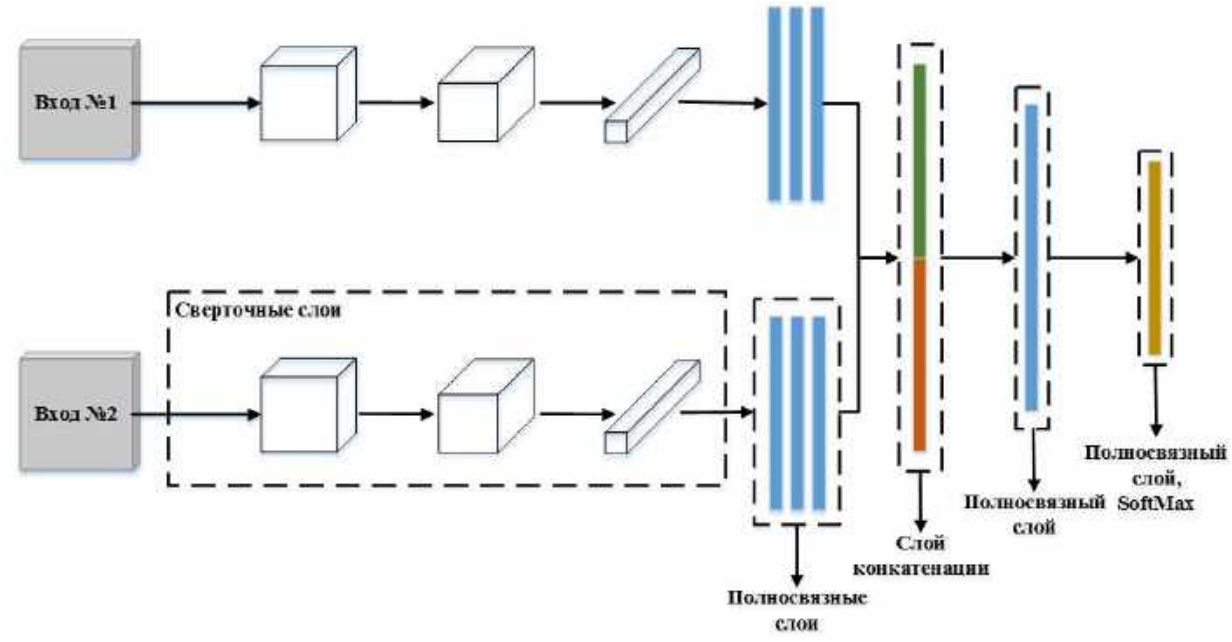


Рисунок 4.1 – Схема объединения признаков на основе конкатенации

Математически описать процедуру конкатенации можно следующим образом. Пусть s_{m1} – набор из k признаков, сформированный при анализе

модальности m_1, а f_{m2} – набор из n признаков, сформированный при анализе модальности m_2:

$$s_{m1} = \{s_1, s_2, s_3 \dots s_k\ \},$$
$$f_{m2} = \{f_1, f_2, f_3 \dots f_n\ \}.$$

Далее с использованием конкатенации формируется комбинированный набор признаков v, который можно описать следующим образом:

$$v = \{s_{m1}, f_{m1}\},$$

$$v = \{s_1, s_2, s_3 \dots s_k, f_1, f_2, f_3 \dots f_n\},$$

$$v \in R^{n+k}.$$

После процедуры конкатенации используется полносвязный слой меньшего размера с целью понижения размерности пространства признаков. Данный тип объединения используется, например, в работе [59], где в качестве базового алгоритма применяется СНС на базе архитектуры VGG16. В качестве биометрических параметров анализируются изображения лиц, радужной оболочки глаз и вен на пальцах. Такой же подход используется в работе [60], где в качестве базовой нейронной сети выбрана архитектура VGG19. Авторы исследуют мультимодальный биометрический алгоритм на основе анализа лица и радужной оболочки глаза.

Второй подход, который применяется для объединения модальностей на уровне признаков основан на билинейном слиянии [60-62]. Билинейное слияние наборов признаков s_{m1} и f_{m1} можно описать следующим образом [61]:

$$s_{m1} = \{s_1, s_2, s_3 \dots s_n\ \}, f_{m2} = \{f_1, f_2, f_3 \dots f_n\ \},$$

$$v = s_{m1} * f_{m2},$$

$$v = \{v_1, v_2, v_3 \dots v_n\}, v \in R^n.$$

Далее выполняется нормировка результирующего вектора v:

$$k_n = \begin{cases} 1, v_n \geq 0 \\ -1, \text{иначе}' \end{cases}$$

$$k = \{k_1, k_2, k_3 \dots k_n\}, k \in R^n,$$

$$y = k\sqrt{|v|}.$$

$$z = \frac{y}{\|y\|_2}.$$

Важно отметить, что при таком подходе количество признаков в наборах s_{m1} и f_{m1} должно быть равным. На Рисунке 4.2 изображена схема объединения признаков на основе билинейного слияния.

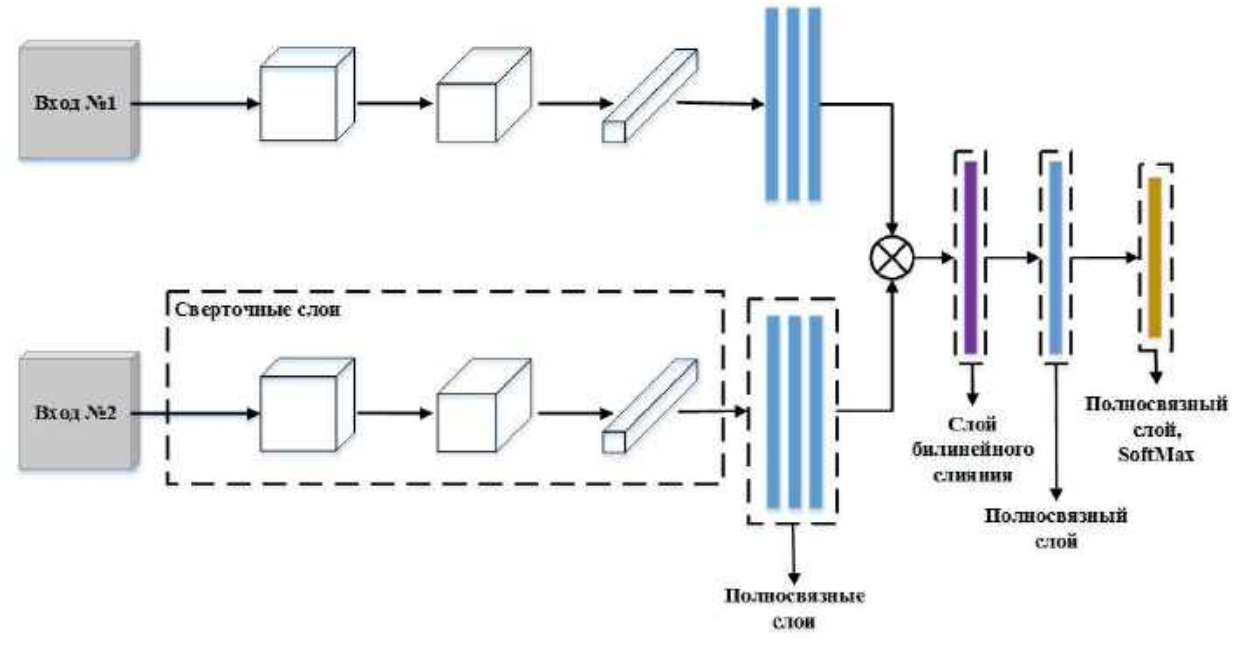


Рисунок 4.2 – Схема объединения признаков на основе билинейного слияния

Третий подход основывается на объединении признаков с разных уровней СНС. Идея заключается в слиянии нескольких блоков информации, характеризующих отличные друг друга представления входных данных. Данный способ слияния признаков получил название мультиабстрактного объединения [63, 64]. Его структурная схема показана на Рисунке 4.3.

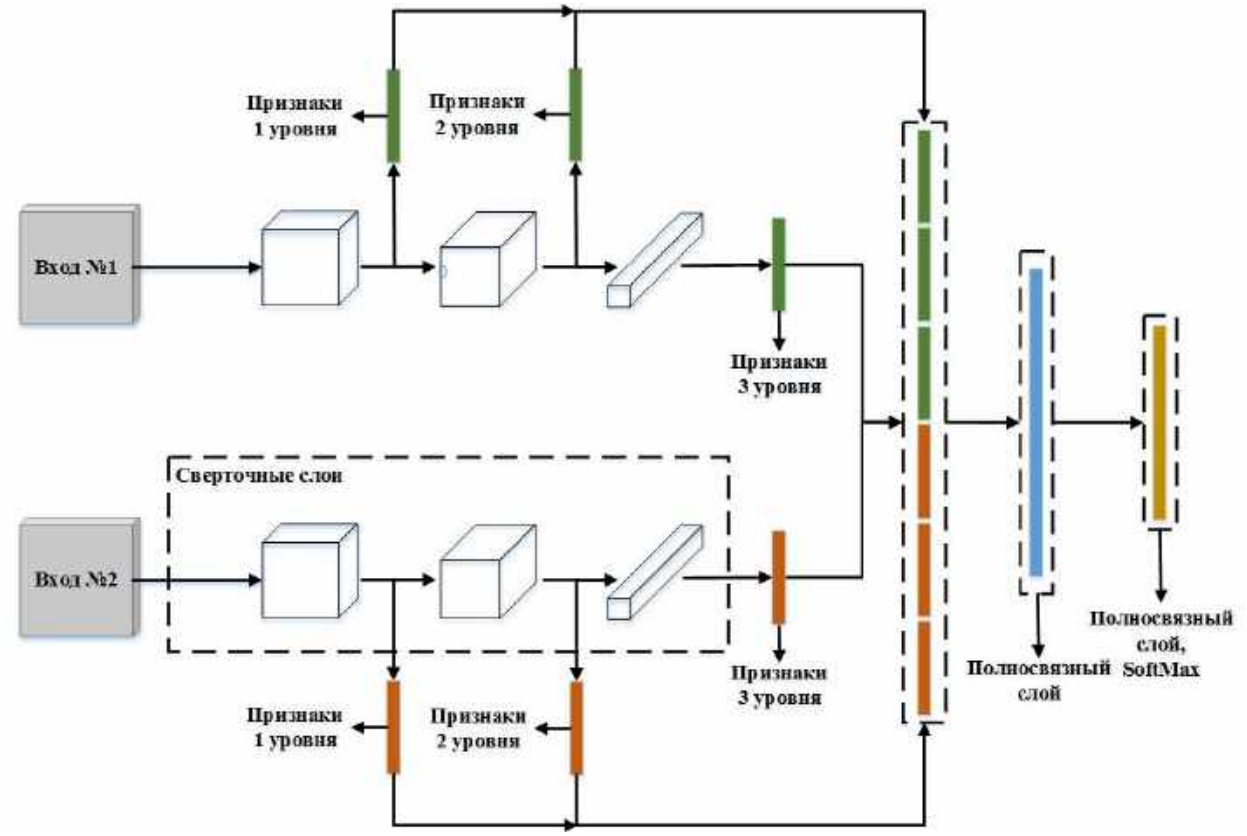


Рисунок 4.3 – Схема объединения признаков на основе мультиабстрактного слияния

Важно отметить, что карты признаков от уровня к уровню имеют разную размерность, поэтому для их объединения необходимо выполнить дополнительные операции. В частности, может быть использован слой глобального усреднения или комбинация из слоев пулинга и свертки, как показано на Рисунке 4.4. После этого признаки могут быть объединены с использованием слоя конкатенации.

В работах [63, 64] данный подход используется для слияния признаков при разработке мультимодальных алгоритмов. В частности, рассматриваются системы на основе различных комбинаций лица, радужной оболочки глаза и отпечатков пальцев. Дополнительно рассматриваются мультиэкземплярные подходы на основе комплексного анализа кадров лица с разного ракурса, а также с использованием указательных и больших пальцев обеих рук. В качестве базовой архитектуры используется сеть VGG19.

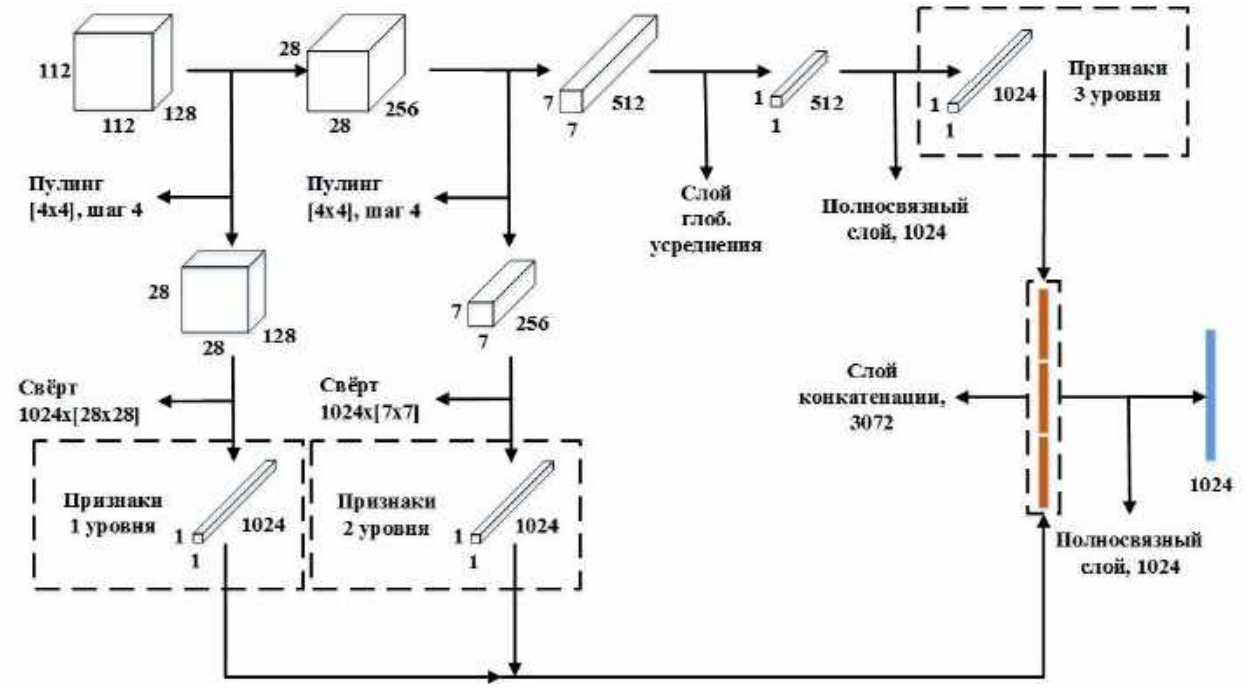


Рисунок 4.4 – Формирование признаков на разных уровнях СНС с использованием комбинации слоев пулинга и свертки

Ранее в п. 1.6.1 и п. 1.6.2 выполнен анализ существующих подходов в области разработки мультимодальных алгоритмов идентификации личности на основе классических алгоритмов и СНС. Большинство из существующих решений основываются на комбинированной идентификации по лицу с отпечатками пальцев или радужной оболочкой глаза. Важно отметить, что сканеры отпечатков пальцев чувствительны к отрицательным температурам, а сам процесс сканирования биометрического параметра является контактным методом. В результате ограничивается область использования такого рода систем. Анализ по радужной оболочке глаза остается одним из самых дорогостоящих решений. При этом системы такого типа накладывают жесткие требования на процесс сканирования. Хорошей альтернативой является подход на основе комбинированного анализа лица и голоса. Далее описан процесс разработки мультимодальных алгоритмов на основе анализа лицевой и голосовой биометрии.

4.2 Разработка и тестирование мультимодальных алгоритмов, выполняющих объединение модальностей на уровне принятия решения

Ранее в главах 2 и 3 настоящей работы выполнена разработка унимодальных алгоритмов на основе СНС. Так, например, разработан нейросетевой алгоритм на базе архитектуры CNN-FaceMask. Этот алгоритм с высокой точностью идентифицирует личность на основе анализа цифровых изображений лиц и способен работать в условиях наличия медицинской маски. Также разработан алгоритм распознавания диктора на основе x-векторной системы, обладающий повышенной устойчивостью к шумам и искажениям. Оба указанных алгоритма имеют высокую практическую значимость для реальных биометрических систем. Они взяты за основу для разработки мультимодальных подходов. Все алгоритмы, рассмотренные в данном и следующем разделах, структурно являются параллельными, то есть анализ модальностей в них выполняется одновременно.

Тестирование мультимодальных алгоритмов осуществляется на основе комбинирования тестовых данных:

- тестовые изображения лиц (далее – Тест-Л);
- тестовые изображения лиц с медицинскими масками (Тест-ММ);
- тестовые голосовые сигналы (далее – Тест-Г);
- тестовые голосовые сигналы с искажениями и шумами (Тест-Ш).

В результате составлены 4 комбинации тестов: «Тест-Л, Тест-Г»; «Тест-Л, Тест-Ш»; «Тест-ММ, Тест-Г»; «Тест-ММ, Тест-Ш».

Разработанные унимодальные алгоритмы могут быть использованы и в качестве независимых систем, поэтому самый простой вариант их комбинирования выполняется на уровне принятия решения (далее – МА-1). Такой модуль строится на основе логических операций «ИЛИ» и «И» (далее – МА-1Д и МА-1К). В случае использования оператора «ИЛИ» (дизъюнкция) – если одна из модальностей определила личность корректно, то попытка классификации считается успешной. В ситуации использования оператора

«И» (конъюнкция) – обе модальности должны распознать пользователя верно, иначе попытка считается неудачной. На Рисунке 4.5 представлена мультимодальная система на основе МА-1 алгоритма.

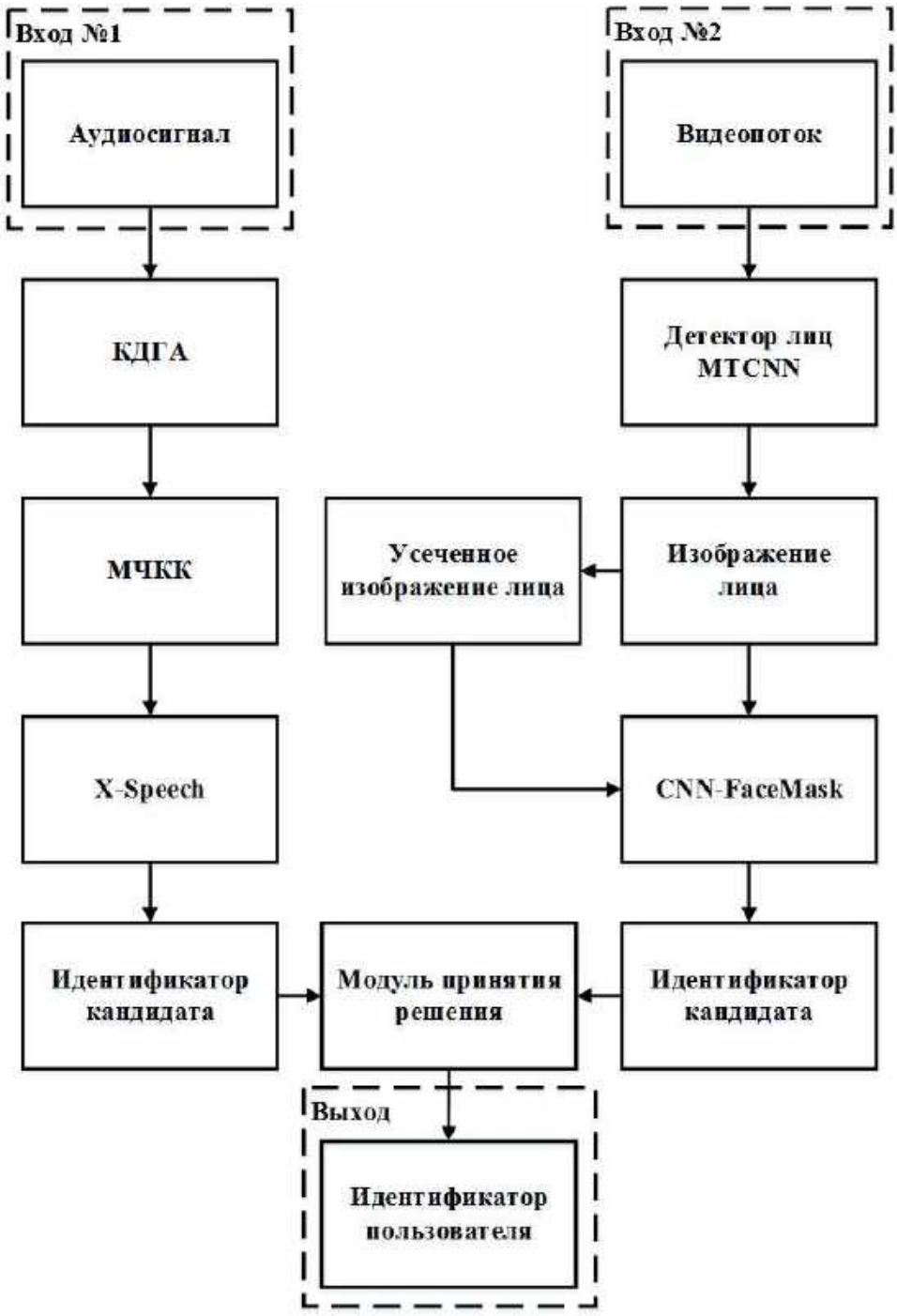


Рисунок 4.5 – Мультимодальная система на основе алгоритма МА-1

Исследование системы выполнялось в результате комбинирования унимодальных алгоритмов на базе обученных моделей X-Speech и

CNN-FaceMask. В Таблице 4.1 приведены результаты тестирования мультимодального алгоритма МА-1.

Таблица 4.1 – 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%

В начале рассмотрим подход, где правило объединения выполняется на основе операции дизъюнкции. В данном случае достигается высокий уровень идентификации на всех тестах, в том числе и в условиях использования медицинской маски и зашумления речевого сигнала. В частности, точность идентификации при проведении самого сложного теста «Тест-ММ, Тест-Ш» составила выше 99%. Другой подход, на основе конъюнкции, оказался чувствителен к изменениям в данных. Так, точность идентификации при проведении аналогичного теста «Тест-ММ, Тест-Ш» снизилась более чем на 17%.

Исходя из полученных результатов, можно сделать вывод, что модуль принятия решения на основе логической операции дизъюнкции в рамках мультимодальной системы, лучше всего подходит для работы в ситуациях, когда определяющее значение имеют ошибки второго рода. Дополнительно данный подход может быть использован в условиях наличия медицинской маски или зашумления речевых данных.

Модуль принятия решения на основе конъюнкции имеет высокую чувствительность к шумам и помехам. 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. Таким образом достигается уменьшение ошибок первого рода. Алгоритм МА-1К подходит для работы в условиях низкой зашумленности аудио- и видеоканала, а также

в условиях полной видимости лица. В такого рода условиях эксплуатации («Тест-Л, Тест-Г») точность идентификации составляет до 98%.

4.3 Разработка и тестирование мультимодальных алгоритмов, выполняющих объединение модальностей на уровне слияния признаков

В этом разделе выполняется разработка мультимодальных алгоритмов с использованием метода объединения модальностей на уровне признаков. При разработке архитектур СНС в качестве базиса выбраны рассмотренные ранее топологии CNN-FaceMask и X-Speech.

Первый предложенный алгоритм представляет собой СНС, состоящую из 3-х трактов (далее – МА-2), как показано на Рисунке 4.6.

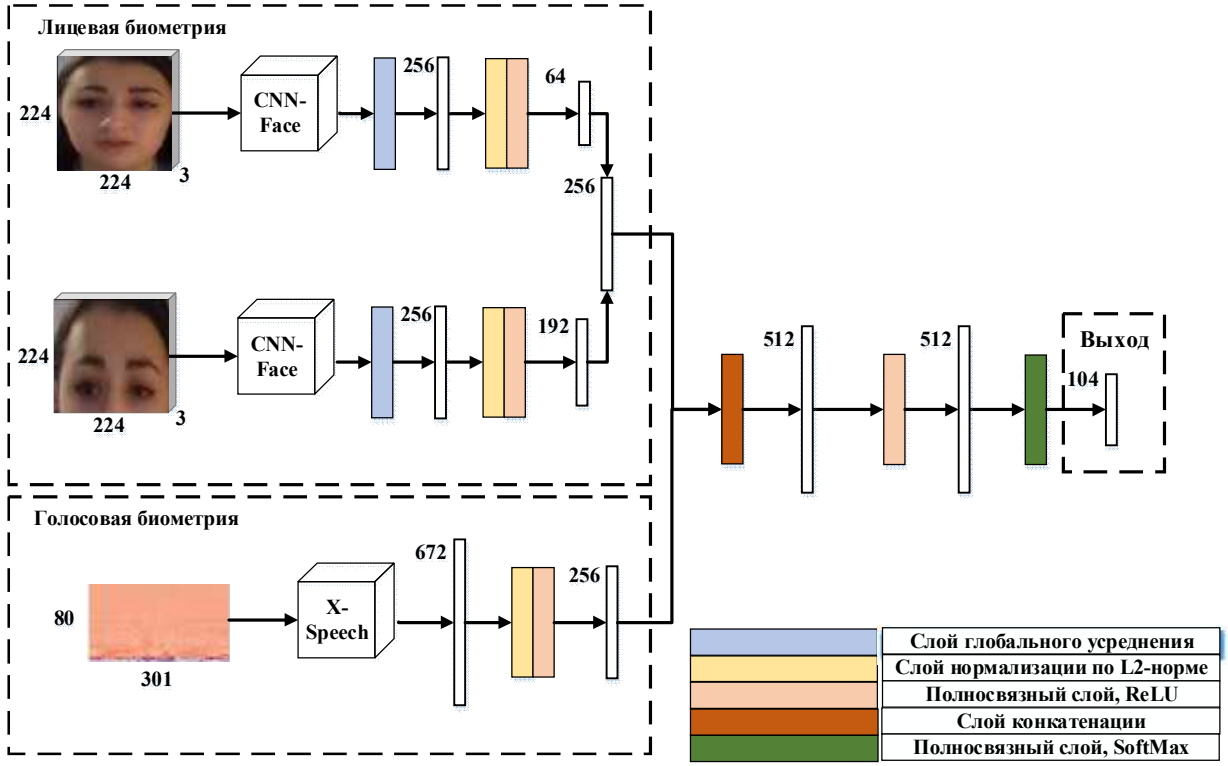


Рисунок 4.6 – Мультимодальный алгоритм МА-2

На выходе трактов МА-2 формируются векторы признаков размером 64, 192, 256. В процессе формирования признаков, описывающих лицо, предпочтение отдается модулю, анализирующему видимую область лица в случае использования медицинской маски. В результате анализа

изображений лица и речевого сигнала формируются два вектора признаков одинаковой размерности. Далее векторы объединяются в результирующий вектор размера 512. При данном способе конкатенации признаков влияние каждой модальности на результат классификации является равнозначным.

В Таблице 4.2 представлены результаты тестирования работы мультимодального алгоритма МА-2. Алгоритм показывает более высокие результаты относительно мультимодального решения МА-1К. В данной части исследования алгоритм МА-1Д не рассматривается, поскольку с точки зрения взлома и попыток несанкционированного доступа он более уязвим, то есть подвержен ошибкам ложной классификации.

В частности, в ситуации использования медицинской маски выигрыш составляет не более 1%, тогда как в условиях зашумления речевого сигнала преимущество увеличивается до 7%. На тестовом наборе данных «Тест-ММ, Тест-Ш», который моделирует более сложные условия эксплуатации биометрических систем, выигрыш мультимодального алгоритма МА-2 составляет более 6%.

Таблица 4.2 – 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%

При анализе архитектуры алгоритма МА-2 важно отметить одну особенность. Тракт 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, предложена еще одна модификация алгоритма, обозначаемая далее, как МА-2М.

На Рисунке 4.7 изображена архитектура СНС для мультимодального подхода МА-2М. При проектировании блока анализа лицевой биометрии алгоритм МА-2М выполняет обработку исключительно видимой области

лица. Анализ изображения осуществляется вне зависимости от наличия или отсутствия медицинской маски. В данном случае выполняется обработка области лба, глаз и части носа. Модуль анализа речевых сигналов остается без изменений.

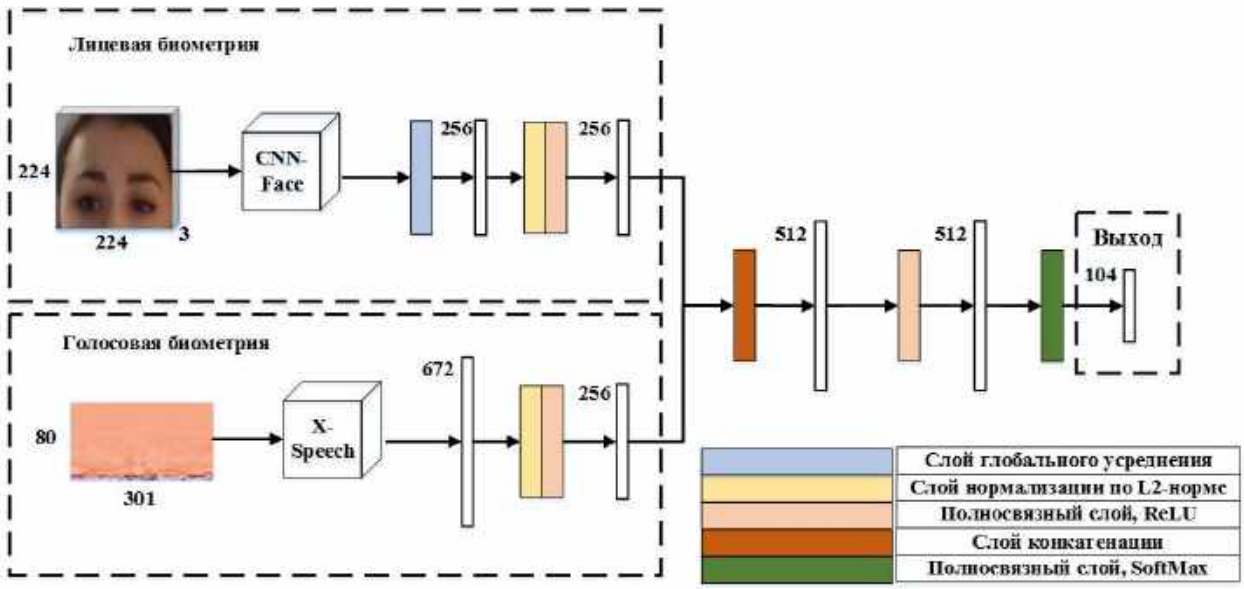


Рисунок 4.7 – Мультимодальный алгоритм МА-2М

В Таблице 4.3 представлены результаты сравнения мультимодальных алгоритмов на основе слияния признаков с алгоритмом МА-1К.

Таблица 4.3 – Сравнение точности 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%**

Анализируя полученные результаты, можно сделать вывод, что подходы на основе объединения модальностей на уровне слияния признаков лучше справляются с задачей идентификации по сравнению с алгоритмом комбинирования нейросетевых алгоритмов на основе конъюнкции. Так, МА-2 лучше всего подходит для работы в условиях низкой зашумленности

аудио- и видеоканала. Также данное решение может применяться в условиях наличия шумов в канале связи или при использовании микрофона низкого качества, поскольку в данных условиях деградация в точности работы составляет менее 2%. Алгоритм MA-2M лучше всего подходит для работы в условиях наличия медицинской маски. В частности, решение демонстрирует точность на уровне 95%, что превосходит результаты работы других мультимодальных аналогов на 6% и более. На Рисунке 4.8 представлена структурная схема мультимодальной системы на основе алгоритма MA-2M.

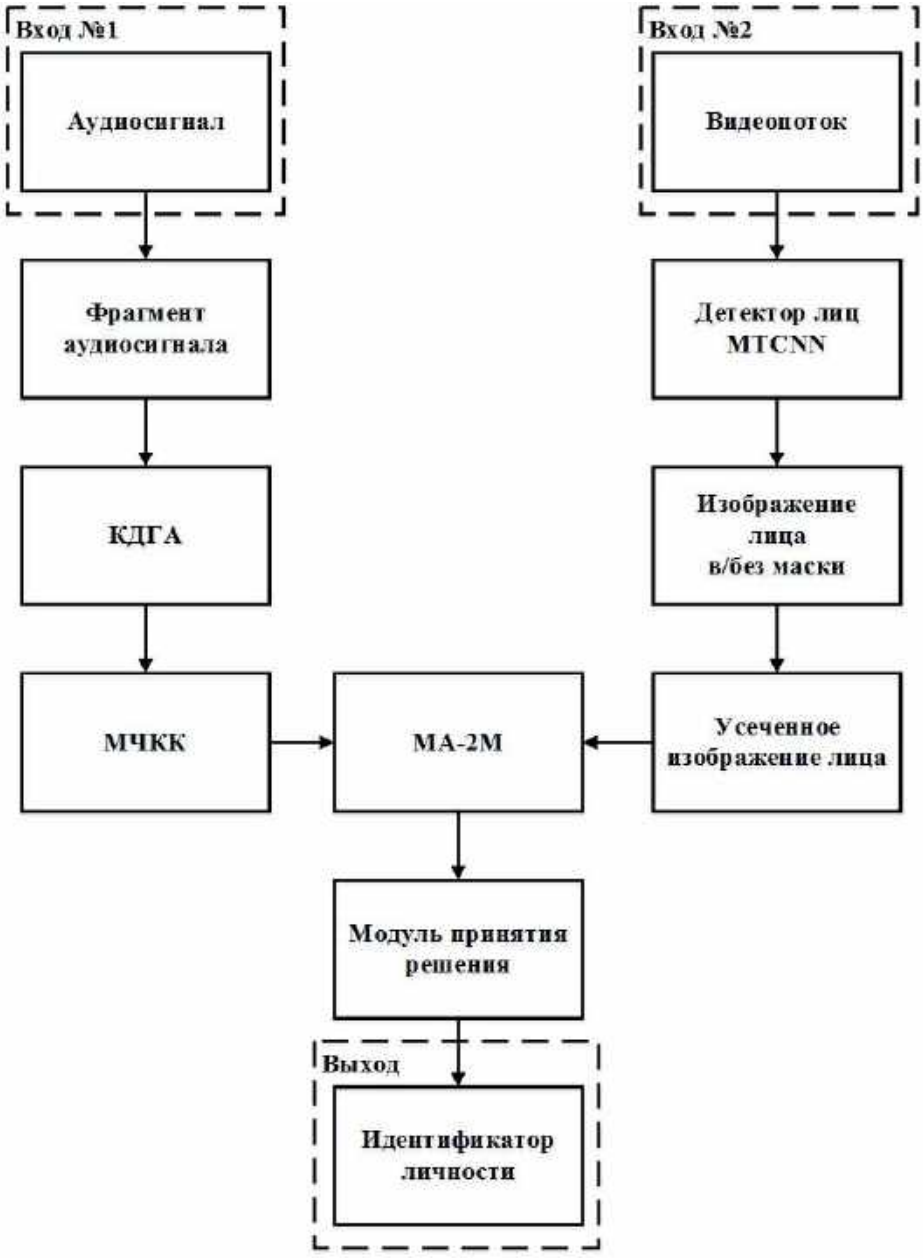


Рисунок 4.8 – Схема мультимодальной системы на основе алгоритма MA-2M

4.4 Сравнительный анализ унимодальных и мультимодальных алгоритмов

Проведенное ранее исследование работы мультимодальных алгоритмов показало, что объединение модальностей на уровне слияния признаков позволяет улучшить точность и робастность идентификации личности в условиях зашумления речевых сигналов, а также в условиях использования медицинской маски. Помимо этого, важно провести сравнительный анализ работы мультимодальных алгоритмов и рассмотренных ранее унимодальных аналогов.

В Таблице 4.4 представлены результаты тестирования разработанных в работе мультимодальных и унимодальных алгоритмов. Важно отметить, что стандартные нейросетевые алгоритмы в данной части исследования не рассматриваются, поскольку уступают в робастности решениям на базе СНС CNN-FaceMask и X-Speech.

Таблица 4.4 – Сравнительный анализ работы мультимодальных и унимодальных алгоритмов

Унимодальный алгоритм голосовой биометрии				
-	«Тест-Г»		«Тест-Ш»	
X-Speech	98,37%		91,54%	
Унимодальный алгоритм лицевой биометрии				
-	«Тест-Л»		«Тест-ММ»	
CNN-FaceMask	99,94%		93,10%	
Мультимодальные алгоритмы				
-	«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%**

В условиях зашумления речевых сигналов лучший показатель точности демонстрирует мультимодальный алгоритм МА-2. Преимущество подхода относительно унимодального алгоритма на базе СНС X-Speech составляет в

среднем 7%. С задачей идентификации личности на основе анализа лиц в условиях использования медицинской маски лучше всего справляется алгоритм МА-2М. Выигрыш относительно унимодального решения на базе СНС CNN-FaceMask составляет не менее 2%. Особое внимание заслуживает результат работы алгоритма МА-2М в условиях теста «Тест-ММ, Тест-Ш», где одновременно моделируется ситуация использования медицинских масок и зашумление речевых сигналов. Несмотря на высокий уровень сложности проводимого эксперимента, деградация в точности мультимодального алгоритма МА-2М составляет менее 5%.

По результатам проведенного исследования можно сделать вывод, что объединение модальностей на уровне слияния признаков позволяет повысить точность и робастность идентификации по сравнению не только с алгоритмами объединения модальностей на уровне принятия решения, но и также относительно рассмотренных унимодальных подходов.

Мультимодальная система на основе разработанного алгоритма МА-2 имеет практическую значимость для систем прокторинга, осуществляемого в сеансе ВКС. В этом случае возможна деградация качества аудиосигнала ввиду: ухудшения соединения или потери части информации в канале передачи; использования микрофона низкого качества; влияния шумов и помех внешней среды. Для систем контроля и управления доступом высокой надежности практическую важность имеет система на основе мультимодального алгоритма МА-2М. Данное решение может быть использовано на предприятиях с обязательным использованием средств индивидуальной защиты, а также на закрытых объектах с высокой степенью секретности.

4.5 Краткие выводы

Результаты проведенных исследований в части разработки мультимодальных алгоритмов идентификации личности с использованием

речевых сигналов и цифровых изображений лиц позволяют сделать следующие основные выводы:

- Разработанный мультимодальный алгоритм МА-2 основывается на принципе объединения модальностей на уровне слияния признаков. Он подходит для работы в условиях низкой зашумленности аудио- и видеоканала, когда его точность составляет до 99% на тестовом наборе данных. Алгоритм МА-2 устойчив к искажениям в аудиосигналах, поскольку деградация в точности работы составляет в среднем 1-2%.
- Разработанный 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%.
- Установлено, 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.
- Мультимодальный алгоритм МА-2 имеет практическую значимость для биометрических систем в задачах прокторинга, а его модификация МА-2М может быть использована при построении прикладных систем контроля и управления доступом высокой надежности.

ЗАКЛЮЧЕНИЕ

Основные выводы и результаты диссертационной работы можно сформулировать в следующем виде.

1. Собран аудиовизуальный набор данных FaceSpeechDB для обучения и тестирования алгоритмов биометрической идентификации личности. Он содержит 60 часов русскоязычной записи 104 человек. Акустические свойства набора имеют высокую степень сходства с реальными условиями эксплуатации систем прокторинга.
2. Собран набор аудиосигналов VADSpeakersDB для разработки детектора голосовой активности. Набор содержит записи русскоязычной речи. Общее количество подготовленных фрагментов составляет 138 000 штук.
3. Разработанный алгоритм КДГА улучшает качество речевых сигналов за счет фильтрации фонограмм от пауз, эффектов глотации, вдохов и шумов. Предложенный алгоритм повышает точность определения фрагментов голосовой активности на 2-3% в сравнении с имеющимися аналогами.
4. Разработанный нейросетевой алгоритм на основе x-векторной системы может быть использован для автоматической идентификации диктора с применением анализа речевых сигналов. При сохранении точности идентификации на уровне до 98-99%, он имеет в 10-20 раз меньше весовых параметров, что дает ему преимущество в скорости работы относительно существующих аналогов.
5. Разработанный нейросетевой алгоритм может быть использован в условиях действия шумов и помех, где деградация в точности его работы составляет в среднем 7%. В этих условиях он превосходит аналоги на 5% и более.
6. Разработанный алгоритм на базе предложенной архитектуры CNN-Face может эффективно использоваться в задачах лицевой биометрии. При высокой точности идентификации на тестовом наборе данных на уровне 99%, он содержит в 25-30 раз меньше весовых параметров, что дает ему

существенное преимущество в скорости работы относительно имеющихся аналогов.

7. Предложенный нейросетевой алгоритм на базе модифицированной архитектуры CNN-FaceMask демонстрирует наилучшую в рассматриваемом классе робастность к присутствию медицинской маски на лице человека. Деградация в точности идентификации составляет менее 7%, что превосходит аналогичные показатели для стандартных нейросетевых алгоритмов на 3% и более.
8. Разработанный мультимодальный алгоритм МА-2 основывается на принципе объединения модальностей на уровне слияния признаков. Он подходит для работы в условиях низкой зашумленности аудио- и видеоканала. Алгоритм МА-2 устойчив к искажениям в аудиосигналах, поскольку деградация в точности работы составляет в среднем 1-2%.
9. Разработанный 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%.
10. Установлено, 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.
11. Мультимодальный алгоритм МА-2 имеет практическую значимость для биометрических систем в задачах прокторинга, а его модификация МА-2М может быть использована при построении прикладных систем контроля и управления доступом высокой надежности.
12. Разработанные алгоритмы требуют для своей практической реализации сравнительно небольших вычислительных ресурсов, что позволяет использовать их при создании биометрических систем, работающих в режиме реального времени, в том числе в задачах прокторинга при использовании ВКС и при построении СКУД.

13. Цель работы успешно достигнута, а задачи выполнены. Разработанные алгоритмы интегрированы в следующие программы для ЭВМ:

- VoiceActivityDetector 1.0 – программа для анализа голосовой активности в задаче мультимодальной идентификации личности;
- Multimodal Identification ToolKit – программа для мультимодальной идентификации личности на основе голосовой и лицевой биометрии;
- Bimodal Human Identification 1.0 – программа для мультимодальной идентификации человека по голосу и лицу с помощью цифровых изображений и аудиосигналов.

Для данных программных продуктов получены свидетельства о государственной регистрации программы для ЭВМ (Приложение В).

ЛИТЕРАТУРА

1. Кухарев Г.А. Методы обработки и распознавания изображений лиц в задачах биометрии / Г.А. Кухарев, Е.И. Каменская, Ю.Н. Матвеев, Н.Л. Щеголева; под ред. М.В. Хитрова. – СПб.: Политехника, 2013. - 388 с.
2. Дворкович В.П. Цифровые видеоинформационные системы (теория и практика) / В.П. Дворкович, А.В. Дворкович. – М.: Техносфера, 2012. – 1009 с.
3. Безруков В.Н. Системы цифрового вещательного и прикладного телевидения: учебное пособие для вузов / В.Н. Безруков, В.Г. Балобанов; под ред. В.Н. Безрукова. – М.: Гор. линия-Телеком, 2015. – 608 с.
4. Приоров А.Л. Цифровая обработка изображений: учеб. пособие / А.Л. Приоров, И.В. Апальков, В.В. Хрящев; Яросл. гос. ун-т. – Ярославль: ЯрГУ, 2007. – 235 с.
5. Лукьяница А.А. Цифровая обработка видеоизображений / А.А. Лукьяница, А.Г. Шишкин. – М.: Ай-Эс-Эс Пресс, 2009. – 518 с.
6. Рабинер Л.Р. Цифровая обработка речевых сигналов / Л.Р. Рабинер, Р.В. Шафер; под ред. М.В. Назарова и Ю.Н. Прохорова. – М.: Радио и связь, 1981. – 496 с.
7. Форсайт Д.А. Компьютерное зрение. Современный подход / Д.А. Форсайт, Д. Понс. – М.: Вильямс/Диалектика, 2018. – 960 с.
8. Гашников М.В. Методы компьютерной обработки изображений / М.В. Гашников, Н.И. Глумов, Н.Ю. Ильясова и [др.]; под ред. В.А. Сойфера. – М.: Физматлит, 2001. – 784 с.
9. Басараб М.А. Цифровая обработка сигналов и изображений в радиофизических приложениях / М.А. Басараб, О.В. Горячкин, В.Ф. Кравченко; под ред. В.Ф. Кравченко. – М.: Физматлит, 2007. – 544 с.
10. Скопченко А.А., Дорофеев В.А. Анализ методов распознавания лиц // Технологии Microsoft в теории и практике программирования: докл. 13-й всеросс. конф. – Томск: Изд-во ТПУ, 2016. – С. 176-178.

11. Лебеденко Ю.И. Биометрические системы безопасности: учеб. пособие; Тульский гос. ун-т. – Тула: ТулГУ, 2012. – 160 с.
12. Aron J. How innovative is Apple's new voice assistant, Siri? // The New Scientist. – 2011. – vol. 212. – №. 2836. – p. 24.
13. Hansen J.H.L., Hasan T. Speaker Recognition by Machines and Humans: A tutorial review // In IEEE Signal Processing Magazine. – 2015. – vol. 32. – №. 6. – pp. 74-99.
14. Woodward J.D., Orlans N.M., Higgins P.T. Biometrics // New York: McGraw Hill Osborne. – 2003. – p. 464.
15. Первушин Е.А. Обзор основных методов распознавания дикторов // Математические структуры и моделирование. – 2011. – №. 3 (24). – С. 41-54.
16. Anguera X., Bozonnet S., Evans N., Fredouille C., Friedland G. Speaker diarization: A review of recent research // IEEE Transactions on audio, speech, and language processing. – 2012. – vol. 20. – №. 2. – pp. 356-370.
17. Wang Q., Downey C., Wan L., Mansfield P.A., Moreno I.L. Speaker diarization with LSTM // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2018. – pp. 5239-5243.
18. LeCun Y., Bengio Y. Convolutional networks for images, speech, and time series // The handbook of brain theory and neural networks, MIT Press Cambridge. – 1998. – pp. 255-258.
19. LeCun Y., Boser B., Denker J.S., Henderson D., Howard R., Hubbard W., Jackel L.D. Backpropagation applied to handwritten zip code recognition // Neural Computation. – 1989. – Vol. 1. – № 4. –pp. 541–551.
20. Lee H., Grosse R. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations // In Proceedings of the 26th International Conference on Machine Learning. – 2009. – pp. 609-616.
21. Николенко С. Глубокое обучение / С. Николенко, А. Кадурин, Е. Архангельская. – СПб.: Питер. – 2018. – 480 с.

22. Гудфеллоу Я. Глубокое обучение / Я. Гудфеллоу, Б. Иошуа, А. Курвилль; пер. с англ. А.А. Слинкина. – 2-е изд. – М: ДМК-Пресс. – 2022. – 652 с.
23. Хайкин С. Нейронные сети: полный курс / С. Хайкин. – 2-е издание. – М.: Вильямс, 2008. – 1104 с.
24. Лекун Я. Как учится машина: Революция в области нейронных сетей и глубокого обучения / Я. Лекун; пер.с фр. – М.: Альпина, 2021. – 335 с.
25. Jordan M.I., Mitchell T.M. Machine learning: Trends, perspectives, and prospects // Science. – 2015. – vol. 349. – №. 6245. – pp. 255-260.
26. Рашка С. Python и машинное обучение / С. Рашка; пер. с англ. А.В. Логунова. – М: ДМК Пресс, 2017. – 418 с.
27. Boureau Y.L., Ponce J., LeCun Y. A theoretical analysis of feature pooling in visual recognition // Proceedings of the 27th international conference on machine learning (ICML-10). – 2010. – pp. 111-118.
28. Graham B. Fractional Max-Pooling // In Cornell University Library, Computer Vision and Pattern Recognition. – 2014.
29. Boureau Y.L., Bach F., LeCun Y., Ponce J. Learning mid-level features for recognition // IEEE computer society conference on computer vision and pattern recognition. – 2010. – pp. 2559-2566.
30. Scherer D., Müller A., Behnke S. Evaluation of pooling operations in convolutional architectures for object recognition // International conference on artificial neural networks. – Springer, Berlin, Heidelberg. – 2010. – pp. 92-101.
31. Jain V., Murray J.F., Roth F., Turaga S., Zhigulin V. Supervised learning of image restoration with convolutional networks // In Proceedings 11th International Conference on Computer Vision. – 2007. – pp. 1-8.
32. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition // In International Conference on Learning Representations. – 2015.
33. Parkhi O.M., Vedaldi A., Zisserman A. Deep face recognition. – 2015.

34. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2016. – pp. 770-778.
35. Zagoruyko S., Komodakis N. Wide residual networks //arXiv preprint arXiv:1605.07146. – 2016.
36. Masi I., Wu Y., Hassner T., Natarajan P. Deep face recognition: A survey // Conference on graphics, patterns and images. – IEEE, 2018. – pp. 471-478.
37. Ling H., Wu J., Wu L., Huang J., Chen J., Li P. Self residual attention network for deep face recognition // IEEE Access. – 2019. – vol. 7. – pp. 55159-55168.
38. He K., Zhang X., Ren S., Sun J. Identity mappings in deep residual networks // European conference on computer vision. – 2016. – pp. 630-645.
39. Hu J., Shen L., Sun G. Squeeze-and-excitation networks // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2018. – pp. 7132-7141.
40. Snyder D., Garcia-Romero D., Sell G., Povey D., Khudanpur S. X-Vectors: Robust DNN Embeddings for Speaker Recognition // International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2018. – pp. 5329–5333.
41. Виноградова А.Р. Реализация текстонезависимой верификации диктора по голосу на основе X-векторной системы во фреймворке общего назначения // Сборник трудов IX Конгресса молодых ученых. – 2021. – С. 185-190.
42. Gusev A., Volokhov V. Vinogradova A., Andzhukaev T., Shulipa A. et al. STC-Innovation Speaker Recognition Systems for Far-Field Speaker Verification Challenge 2020 // Interspeech. – 2020. – pp. 3466-3470.
43. Hajavi A., Etemad A. A deep neural network for short-segment speaker recognition // arXiv preprint arXiv:1907.10420. – 2019.
44. Snyder D., Garcia-Romero D., Povey D., Khudanpur S. Deep Neural Network Embeddings for Text-Independent Speaker Verification // Interspeech. – 2017. – pp. 999-1003.

45. Font R., Grau T. The Biometric Vox System for the Albayzin-RTVE 2020 Speech-to-Text Challenge // Proceedings of the Iber-SPEECH, Valladolid, Spain. – 2021. – pp. 24-25.
46. Peddinti V., Povey D., Khudanpur S. A time delay neural network architecture for efficient modeling of long temporal contexts // Sixteenth annual conference of the international speech communication association. – 2015.
47. Garcia-Romero D., Snyder D., Sell G., McCree A., Povey D., Khudanpur S. X-Vector DNN Refinement with Full-Length Recordings for Speaker Recognition // Interspeech. – 2019. – pp. 1493-1496.
48. Sanjekar P.S., Patil J.B. An overview of multimodal biometrics // Signal & Image Processing. – 2013. – vol. 4. – №. 1. – C. 57.
49. Ashish M. Multimodal Biometrics it is: Need for Future Systems // International Journal of Computer Applications. – 2010. – vol. 3. – № 4. – pp. 28-33.
50. Ross A., Jain A. Information Fusion in Biometrics // Journal of Pattern Recognition Letters. – 2003. – vol. 24. – pp. 2115-2125.
51. Yan Y., Zhang Y.J. Multimodal biometrics fusion using correlation filter bank // 19th International Conference on Pattern Recognition. – 2008. – pp. 1-4.
52. Yang F., M. Baofeng. Two Models Multimodal Biometric Fusion Based on Fingerprint, Palm-print and Hand-Geometry // IEEE. – 2007.
53. Kryszczuk K., Richiardi, J., Prodanov P., Drygajlo A. Reliability-based decision fusion in multimodal biometric verification systems // EURASIP Journal on advances in signal processing. – 2007. – vol. 2007. – pp. 1-9.
54. Frischholz R.W., Dieckmann U. BiolD: a multimodal biometric identification system // Computer. – 2000. – vol. 33. – №. 2. – pp. 64-68.
55. Jagadeesan A., Duraiswamy K. Secured cryptographic key generation from multimodal biometrics: feature level fusion of fingerprint and iris // arXiv preprint arXiv:1003.1458. – 2010.
56. Conti V., Militello C., Sorbello F., Vitabile S. A frequency-based approach for features fusion in fingerprint and iris multimodal biometric identification

systems // IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). – 2010. – vol. 40. – №. 4. – pp. 384-395.

57. Ammour B., Boubchir L., Bouden T., Ramdani M. Face–iris multimodal biometric identification system // Electronics. – 2020. – vol. 9. – №. 1. – p. 85.
58. Hammad M., Liu Y., Wang K. Multimodal biometric authentication systems using convolution neural network based on different level fusion of ECG and fingerprint // IEEE Access. – 2018. – vol. 7. – pp. 26527-26542.
59. Alay N., Al-Baity H. Deep learning approach for multimodal biometric recognition system based on fusion of iris, face, and finger vein traits // Sensors. – 2020. – vol. 20. – №. 19. – p. 5523.
60. Talreja V., Valenti M.C., Nasrabadi N. M. Deep hashing for secure multimodal biometrics // IEEE Transactions on Information Forensics and Security. – 2020. – vol. 16. – pp. 1306-1321.
61. Lin T.Y., RoyChowdhury A., Maji S. Bilinear CNN models for fine-grained visual recognition // Proceedings of the IEEE international conference on computer vision. – 2015. – pp. 1449-1457.
62. Chowdhury A.R., Lin T.Y., Maji S. Learned-Miller E. One-to-many face recognition with bilinear cnns // IEEE Winter Conference on Applications of Computer Vision (WACV). – 2016. – pp. 1-9.
63. Soleymani S., Dabouei A., Kazemi H., Dawson J., Nasrabadi N.M. Multi-level feature abstraction from convolutional neural networks for multimodal biometric identification // 24th International Conference on Pattern Recognition (ICPR). – 2018. – pp. 3469-3476.
64. Soleymani S., Torfi A., Dawson J., Nasrabadi N.M. Generalized bilinear deep convolutional neural networks for multimodal biometric identification // 25th IEEE International Conference on Image Processing. – 2018. – pp. 763-767.
65. Dalal N., Triggs B. Histograms of oriented gradients for human detection // IEEE computer society conference on computer vision and pattern recognition (CVPR'05). – 2005. – vol. 1. – pp. 886-893.

66. Guo Z., Zhang L., Zhang D. A completed modeling of local binary pattern operator for texture classification // IEEE transactions on image processing. – 2010. – vol. 19. – №. 6. – pp. 1657-1663.
67. Saon G., Soltau H., Nahamoo D., Picheny M. Speaker adaptation of neural network acoustic models using i-vectors // IEEE Workshop on Automatic Speech Recognition and Understanding. – 2013. – pp. 55-59.
68. Reynolds D.A. Gaussian mixture models // Encyclopedia of biometrics. – 2009. – vol. 741. – C. 659-663.
69. Povey D., Burget L., Agarwal M., Akyazi P. et al. The subspace Gaussian mixture model - A structured model for speech recognition // Computer Speech & Language. – 2011. – vol. 25. №. 2. – pp. 404-439.
70. Varga A.P., Moore R.K. Hidden Markov model decomposition of speech and noise // International Conference on Acoustics, Speech, and Signal Processing. IEEE. – 1990. – pp. 845-848.
71. Schuller B., Rigoll G., Lang M. Hidden Markov model-based speech emotion recognition // IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03). – 2003. – Vol. 2. – pp. 1-4.
72. Nagrani A., Chung J.S., Zisserman A. VoxCeleb: a large-scale speaker identification dataset // arXiv preprint arXiv:1706.08612. – 2017.
73. Chung J.S., Nagrani A., Zisserman A. VoxCeleb2: Deep Speaker Recognition // In Proceedings Interspeech. – 2018. – pp. 1086-1090.
74. Chen H., Xie W., Vedaldi W., Zisserman A. Vggsound: A large-scale audio-visual dataset // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2020. – pp. 721–725.
75. Ephrat A., Mosseri I., Lang O., Dekel T., Wilson K., Hassidim A. et al. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation // arXiv preprint arXiv:1804.03619. – 2018.
76. Tsarapkina J.M., Anisimova A.V., Grigoriev S.G., Alekhina A.A., Mironov A.G. Application of Zoom and Mirapolis Virtual Room in the context

of distance learning for students // Journal of Physics: Conference Series. – IOP Publishing, 2020. – Т. 1691. – №. 1. – С. 012094.

77. 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.
78. Матвеев Ю.Н. Технология биометрической идентификации личности по голосу и другим модальностям // Вестник МГТУ им. Н.Э. Баумана. Сер. «Приборостроение». – 2012. – № 3. – С. 46–61.
79. Powers D. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation // arXiv preprint arXiv:2010.16061. – 2020.
80. Sohn J., Kim N.S., Sung W. A statistical model-based voice activity detection // IEEE signal processing letters. – 1999. – vol. 6. – №. 1. – pp. 1-3.
81. Ramırez J., Segura J. C., Benıtez C., De La Torre A., Rubio A. Efficient voice activity detection algorithms using long-term speech information // Speech communication. – 2004. – vol. 42. – №. 3-4. – pp. 271-287.
82. Moattar M.H., Homayounpour M.M. A simple but efficient real-time Voice Activity Detection algorithm // 17th European Signal Processing Conference. – 2009. – pp. 2549-2553.
83. Sagi O., Rokach L. Ensemble learning: A survey // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. – 2018. – vol. 8. – №. 4. – pp. 1249.
84. Dzeroski S., Zenko B. Is combining classifiers with stacking better than selecting the best one? // Machine learning. – 2004. – vol. 54. – №. 3. – pp. 255-273.
85. Sikora R. A modified stacking ensemble machine learning algorithm using genetic algorithms // Handbook of research on organizational transformations through big data analytics. – 2015. – pp. 43-53.

86. Чистяков С.П. Случайные леса: обзор // Труды Карельского научного центра Российской академии наук. – 2013. – №. 1. – pp. 117–136.
87. Breiman L. Random forests // Machine learning. – 2001. – vol. 45. – №. 1. – pp. 5-32.
88. Refaeilzadeh P., Tang L., Liu H. Cross-validation // Encyclopedia of database systems. – 2009. – vol. 5. – pp. 532-538.
89. Arlot S., Celisse A. A survey of cross-validation procedures for model selection // Statistics surveys. – 2010. – vol. 4. pp. 40-79.
90. Bergstra J., Bengio Y. Random search for hyper-parameter optimization // Journal of machine learning research. – 2012. – vol. 13. – №. 2. – pp. 281-305.
91. Kopparapu S.K., Laxminarayana M. Choice of Mel filter bank in computing MFCC of a resampled speech // 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010). – 2010. – pp. 121-124.
92. Tiwari V. MFCC and its applications in speaker recognition // International journal on emerging technologies. –2010. – vol. 1. – №. 1. pp. 19-22.
93. Ittichaichareon C., Suksri S., Yingthawornsuk T. Speech recognition using MFCC // International conference on computer graphics, simulation and modeling. – 2012. – pp. 135-138.
94. Dave N. Feature extraction methods LPC, PLP and MFCC in speech recognition // International journal for advance research in engineering and technology. – 2013. – vol. 1. – №. 6. pp. 1-4.
95. Salamon J., Bello J.P. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification // In IEEE Signal Processing Letters. – 2017. – vol. 24. – № 3. – pp. 279-283.
96. Park D.S., Chan W., Zhang Y., Chiu C.C., Zoph B., Cubuk E.D. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition // arXiv preprint arXiv:1904.08779. – 2019.
97. Ko T., Peddinti V., Povey D., Khudanpur S. Audio augmentation for speech recognition // In INTERSPEECH-2015. – 2015. – pp. 3586-3589.

98. Salamon J., Jacoby C., Bello J.P. A Dataset and Taxonomy for Urban Sound Research // 22nd ACM International Conference on Multimedia. – 2014.
99. Zhang K., Zhang Z., Li Z., Qiao Y. Joint face detection and alignment using multitask cascaded convolutional networks // IEEE Signal Processing Letters. – 2016. – vol. 23. – №. 10. – pp. 1499-1503.
100. Li X., Yang Z., Wu H. Face detection based on receptive field enhanced multi-task cascaded convolutional neural networks // IEEE Access. – 2020. – vol. 8. – pp. 174922-174930.
101. Neubeck A., Van Gool L. Efficient non-maximum suppression // 18th International Conference on Pattern Recognition (ICPR'06). – 2006. – vol. 3. – pp. 850-855.
102. Tan C., Sun F., Kong T., Zhang W., Yang C., Liu C. A survey on deep transfer learning // International conference on artificial neural networks. – Springer, Cham, 2018. – pp. 270-279.
103. Lu J., Behbood V., Hao P., Zuo H., Xue S., Zhang G. Transfer learning using computational intelligence: A survey // Knowledge-Based Systems. – 2015. – vol. 80. – pp. 14-23.
104. 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.
105. 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.
106. 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.
107. 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): Материалы 12-й междунар. науч-техн. конф. – Суздаль. – 2017. – Т. 1. – С. 174-177.

108. 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): Сб. докладов 7-ой всеросс.конф. – Москва. – 2017. – С. 155-160.

109. 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.

110. 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.

111. Стефаниди А.Ф., Топников А.И., Приоров А.Л. Бимодальная идентификация личности на основе лицевой и речевой биометрии // Новые информационные технологии и системы (НИТиС-2020): докл. 17-й междунар. конф. – Пенза. – 2020. – С. 125-129.

112. 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.

113. Сенников А.В., Стефаниди А.Ф. Разработка алгоритма детектирования средств индивидуальной защиты на видеоданных // Новые информационные технологии и системы (НИТиС-2021): докл. 18-й междунар. конф. – Пенза. – 2021. – С. 150-155.

114. 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. конф. – Пенза. – 2021. – С. 56-63.
115. Khryashchev V.V., Topnikov A.I., Stefanidi A.F., Priorov A.L. Bimodal person identification using voice data and face images // Proc. SPIE 11041, Eleventh International Conference on Machine Vision (ICMV 2018). – 2019. – Vol. 11041. – pp. 296-303.
116. Stefanidi A., Topnikov A., Tupitsin G., Priorov A. Application of convolutional neural networks for multimodal identification task // Proceedings of 26th Conference of Open Innovations Association FRUCT. – 2020. – pp. 423-428.
117. Stefanidi A., Topnikov A., Priorov A, Kosterin I. Modification of VGG Neural Network Architecture for Unimodal and Multimodal Biometrics // Proceedings of 18th IEEE East-West Design & Test Symposium (EWDTS-2020), Varna, Bulgaria. – 2020. – pp. 1-4.
118. 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 всероссийской конференции (с международным участием). – Москва, 2022.

ПРИЛОЖЕНИЕ А. Алгоритм вычисления мел-частотных кепстральных коэффициентов

Рассмотрим работу алгоритма выделения мел-частотных кепстральных коэффициентов. Исходный речевой сигнал делится на небольшие фрагменты, длительностью 20-40 мс. Пусть $x_j(n)$ – j-й фрагмент исходного сигнала, а $0 \leq n < N$, где N – длина окна. Далее будет описываться работа алгоритма в рамках одного окна речевого сигнала. К фрагменту применяется дискретное преобразование Фурье:

$$X_j(k) = \sum_{n=0}^{N-1} x_j(n) w(n) e^{\frac{-2\pi}{N} kn}, 0 \leq k < N,$$

где $w(n)$ – оконная (весовая) функция [15, 91]. В качестве такой функции, как правило, используют окно Хэннинга, Хэмминга, Блэкмана или Кайзера [91]. Далее в работе используется окно Хэннинга:

$$w(n) = 0.5 - 0.5\cos\left(\frac{2\pi n}{N-1}\right).$$

Длина оконной функции и фрагмента речевого сигнала эквивалентны. После применения ДПФ формируется спектрограмма. Следующим этапом определяется периодограмма, путем вычисления квадрата модуля:

$$P_j(k) = \frac{\left|X_j(k)\right|^2}{N}.$$

Коэффициенты разложения ДПФ, номера которых расположены симметрично относительно $\frac{N}{2}$, образуют комплексно-сопряженные пары. При вычислении модуля эти значения становятся эквивалентными, поэтому во время анализа спектра они не несут в себе дополнительной информации и отбрасываются. Поэтому далее спектральные отсчеты будут рассматриваться только при $0 \leq k \leq \frac{N}{2} - 1$ [15, 92].

Далее необходимо рассчитать банк треугольных фильтров. Метод расчета представляет собой ряд последовательных действий [91-93]:

- Выбирается интересующий диапазон частот – (f_{low}, f_{max}), который переводится в мелы – (m_{low}, m_{max}) по формуле (1).
- Определяется количество фильтров в банке – C. Как правило, используется от 12 до 80 фильтров в зависимости от задачи.
- Вычисляются значения центральных частот общим количеством $C + 2$ в мелах. Частоты должны располагаться линейно между определенным ранее диапазоном (m_{low}, m_{max}):

$$m_i = m_{low} + i\frac{m_{\max} - m_{low}}{C+1}, \; при \; i = 0....C+1.$$

- Значения центральных частот переводятся из мелов в герцы по формуле (2):

$$f_i = M^{-1}(m_i), \; при \; i = 0....C+1.$$

- Полученные частоты переводятся в номера спектральных отсчетов:

$$k(f_i) = f_i \times \left(\frac{N}{F_s}\right),$$

где F_s – частота дискретизации речевого сигнала, $k(f_i)$ – номер спектрального отсчета для частоты f_i. Номер отсчета должен быть целым числом, поэтому полученные значения округляются.

- Для каждого треугольного фильтра рассчитывается амплитудно-частотная характеристика (АЧХ):

$$H_i = \begin{cases} 0, \; при \; k < k(f_{i-1}) \\ \dfrac{k - k(f_{i-1})}{k(f_i) - k(f_{i-1})}, \; при \; k(f_{i-1}) \le k < k(f_i) \\ \dfrac{k(f_{i+1}) - k}{k(f_{i+1}) - k(f_i)}, \; при \; k(f_i) \le k < k(f_{i+1}) \\ 0, \; при \; k > k(f_{i+1}) \end{cases},$$

где $i = 1 \ldots C, k = 0 \ldots \frac{N}{2} - 1$, H_i – АЧХ i-го фильтра.

- Полученный банк фильтров применяется к периодограмме:

$$S_j(i) = \sum_{k=0}^{\frac{N}{2}-1} H_i(k) P_j(k), \text{ где } i = 1....C.$$

Результатом вычислений является вектор коэффициентов $S_j(i)$ для j-го фрагмента речевого сигнала.

- Рассчитывается натуральный логарифм от $S_j(i)$:

$$L_j(i) = \ln(S_j(i)).$$

- На последнем этапе проводится расчет ДКП 2-го типа от $L_j(i)$:

$$U_j(n) = \sum_{i=1}^{C} L_j(i) \cos\left(\frac{\pi}{2C}(2i-1)(n-1)\right) \cdot \begin{cases} \sqrt{1/C}, n = 1 \\ \sqrt{2/C}, 2 \le n \le C. \end{cases}$$

В данном случае $U_j(n)$ – n-й коэффициент ДКП, где $n = 1 \dots C$. В итоге получаем мел-частотные кепстральные коэффициенты.

Использование ДКП обусловлено тем, что фильтры внутри гребенки имеют области пересечения, и в итоге коэффициенты $L_j(i)$обладают высокой корреляцией, а применение ДКП позволяет их декоррелировать. Дополнительно, использование ДКП позволяет более компактно представить входные данные, уменьшая их размерность [92, 93].

ПРИЛОЖЕНИЕ Б. Акты внедрения

(а)вижн

Общество с ограниченной ответственностью
«А-ВИЖН»
Юридический адрес: 150054, г. Ярославль, ул. Угличская, 31-43 Почтовый адрес: 150000, г. Ярославль, ул. Большая Октябрьская, д. 45
тел: +7 (4852) 26 50 10, e-mail: connect@a-vsn.ru
Р/сч 40702810177030005062
в Калужском отделении №8608 ПАО Сбербанк
к/с 30101810100000000612 БИК 042908612
ИНН 7604082087 КПП 760401001

«УТВЕРЖДАЮ»
Директор ООО «А-ВИЖН», к.т.н.
_____ И.В. Апальков
«24» марта 2022 г.

АКТ

О внедрении результатов диссертационной работы Стефаниди А.Ф. на тему «Исследование мультимодальных алгоритмов биометрической идентификации на основе методов цифровой обработки речевых сигналов и изображений»

Комиссия в составе: председатель – руководитель проектов Костин В.В., члены комиссии – заместитель директора по поддержке бизнеса Игнатов И.С., начальник технического отдела Нестеров М.С, рассмотрев диссертацию Стефаниди А.Ф. составила настоящий акт о том, что ее результаты нашли применение в работе ООО «А-Вижн». Особый практический интерес представляет следующий результат диссертации:

- разработан нейросетевой алгоритм идентификации личности на базе предложенной архитектуры CNN-FaceMask, который способен работать при наличии медицинской маски на лице человека.

Предложенный алгоритм применен при разработке системы контроля и управления доступом в составе модуля анализа биометрических данных. Применение нейросетевого алгоритма, разработанного Стефаниди А.Ф., позволило повысить точность идентификации в условиях наличия медицинских масок.

Председатель комиссии _____ Костин В.В.

Члены комиссии _____ Игнатов И.С.

_____ Нестеров М.С.

ЦИФРОВЫЕ РЕШЕНИЯ

Общество с ограниченной ответственностью «ЦИФРОВЫЕ РЕШЕНИЯ»

150047, Ярославская обл., г. Ярославль,
ул. Лермонтова, д.44а, кв.5

ИНН 7606119310
КПП 760601001
ОГРН 1197627002198

«УТВЕРЖДАЮ»

Генеральный директор
ООО «Цифровые решения», к.т.н.

Матвеев Д.В.
«05» апреля 2022г

АКТ

внедрения результатов диссертационной работы Стефаниди Антона Федоровича на тему «Исследование мультимодальных алгоритмов биометрической идентификации на основе методов цифровой обработки речевых сигналов и изображений»

Комиссия в составе: председатель комиссии – к.т.н. Голубев М.Н., члены комиссии – ведущий инженер Гомулин С.А., программист Федькина А.А., рассмотрев диссертационную работу Стефаниди А.Ф. составила настоящий акт о том, что ее результаты нашли применение в разработке коммерческих программных продуктов ООО «Цифровые решения». Особый практический интерес представляет следующий результат диссертации:

– разработан мультимодальный алгоритм идентификации личности с повышенной устойчивостью к шумам и помехам в речевых сигналах.

Процесс тестирования алгоритма в реальных условиях видеоконференцсвязи подтвердил возможность достижения точности идентификации на уровне выше 98%. Алгоритм интегрирован в состав системы прокторинга (проведение проверочных мероприятий и экзаменов в онлайн-режиме) для идентификации экзаменуемых. Применение мультимодального алгоритма идентификации позволило повысить точность и устойчивость работы модуля проверки личности в реальных условиях видеоконференцсвязи.

Председатель комиссии _____ / Голубев М.Н.
(Подпись)

Члены комиссии _____ / Гомулин С.А.
(Подпись)

_____ / Федькина А.А.
(Подпись)

«УТВЕРЖДАЮ»
Ректор Ярославского
государственного университета
им. П.Г. Демидова
_____А.И. Русаков
«20» мая 2022 г.

АКТ

внедрения результатов диссертационной работы Стефаниди Антона Федоровича на тему «Исследование мультимодальных алгоритмов биометрической идентификации на основе методов цифровой обработки речевых сигналов и изображений» в учебный процесс

Мы, нижеподписавшиеся, заведующий кафедрой цифровых технологий и машинного обучения, доцент, к.ф.-м.н. М.В. Чистяков и профессор кафедры цифровых технологий и машинного обучения, доцент, д.т.н. А.Л. Приоров составили настоящий акт о том, что результаты диссертационной работы А.Ф. Стефаниди внедрены в учебный процесс на кафедре цифровых технологий и машинного обучения физического факультета ЯрГУ (направление «Радиотехника»):

- в курсе «Цифровая обработка речевых сигналов» – усовершенствованный комбинированный детектор голосовой активности;
- в курсе «Цифровая обработка изображений» – мультимодальные алгоритмы идентификации личности на основе сверточных нейронных сетей.

Заведующий кафедрой цифровых
технологий и машинного обучения,
к.ф.-м.н., доцент _____ М.В. Чистяков

Профессор кафедры
цифровых технологий
и машинного обучения, д.т.н., доцент _____ А.Л. Приоров

«УТВЕРЖДАЮ»
Ректор Ярославского
государственного университета
им. Т.Г. Демидова
_____ А.И. Русаков
«20» мая 2022 г.

АКТ

внедрения результатов диссертационной работы Стефаниди Антона Федоровича на тему «Исследование мультимодальных алгоритмов биометрической идентификации на основе методов цифровой обработки речевых сигналов и изображений» в научно-исследовательские работы

Результаты диссертационной работы А.Ф. Стефаниди, представленной на соискание ученой степени кандидата технических наук по специальности 2.2.13 Радиотехника, в том числе системы и устройства телевидения, использованы в научно-исследовательской работе:

– при выполнении НИР «Разработка алгоритмов идентификации и верификации личности по речевой и видеоинформации для систем мультимодальной биометрии» (грант РФФИ 19-37-90158) внедрены результаты исследования мультимодальных алгоритмов идентификации личности с использованием анализа речевых сигналов и изображений лиц.

Заведующий кафедрой цифровых
технологий и машинного обучения,
доцент, к.ф.-м.н. _____ М.В. Чистяков

ПРИЛОЖЕНИЕ В. Свидетельства о государственной регистрации интеллектуальной собственности

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2021681283

VoiceActivityDetector 1.0 – программа для анализа голосовой активности в задаче мультимодальной идентификации личности

Правообладатель: ***Общество с ограниченной ответственностью «СОФТ ВИЖН» (RU)***

Автор(ы): ***Стефаниди Антон Федорович (RU)***

Заявка № **2021680843**

Дата поступления **13 декабря 2021 г.**

Дата государственной регистрации в Реестре программ для ЭВМ ***20 декабря 2021 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2021663249

Multimodal Identification ToolKit - программа для мультимодальной идентификации личности на основе голосовой и лицевой биометрии

Правообладатель: ***Общество с ограниченной ответственностью «СОФТ ВИЖН» (RU)***

Автор(ы): ***Стефаниди Антон Федорович (RU)***

Заявка № **2021662262**
Дата поступления **02 августа 2021 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***13 августа 2021 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2019613092

Bimodal Human Identification 1.0 - программа для мультимодальной идентификации человека по голосу и лицу с помощью цифровых изображений и аудио-сигналов

Правообладатели: ***Стефаниди Антон Федорович (RU), Топников Артем Игоревич (RU)***

Авторы: ***Стефаниди Антон Федорович (RU), Топников Артем Игоревич (RU)***

Заявка № **2019611974**
Дата поступления **27 февраля 2019 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***07 марта 2019 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

ПРИЛОЖЕНИЕ Г. Сертификаты, дипломы и грамоты


ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
ПЕНЗЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
PENZA STATE UNIVERSITY
PNZGU.RU
ПОЧЕТНАЯ ГРАМОТА
НАГРАЖДАЕТСЯ
аспирант ФГБОУ ВО «Ярославский государственный университет им. П. Г. Демидова»
Стефаниди Антон Федорович
за лучший секционный доклад
на XVIII Международной научно-технической конференции
«Новые информационные технологии и системы»
Ректор университета
А.Д. Гуляков
ПЕНЗА • 2021

2018 The 11th International Conference on Machine Vision

CERTIFICATE OF APPRECIATION

THIS CERTIFICATE IS AWARDED TO

Mr. Anton Stefanidi

P. G. Demidov Yaroslavl State University, Russia

In Honor of your significant contribution to the success of 2018 The 11th International Conference on Machine Vision (ICMV 2018) as a Listener
Munich, Germany, November 1-3, 2018

Dmitry Nikolaev
Session Chair

Prof. Antanas Verikas
Prof. Antanas Verikas
Halmstad University, Sweden
Conference Chair


FRUCT
CERTIFICATE
We hereby confirm that
Anton Stefanidi
(P.G. Demidov Yaroslavl State University, Russia)
took part in the
26th FRUCT Conference
on 23-24 April 2020, Yaroslavl, Russia
(distant participation)
FRUCT26
Sergey Balandin
FRUCT President
IEEE
IEEE ComSoc