Федеральный исследовательский центр «Информатика и управление»
Российской академии наук

На правах рукописи

УДК 004.89

Макаров Виктор Витальевич

Методы и алгоритмы автоматической классификации психофизиологических характеристик человека

05.13.17 – Теоретические основы информатики

ДИССЕРТАЦИЯ
на соискание ученой степени
кандидата технических наук

Научный руководитель
д. ф.-м. н., проф.
Цурков Владимир Иванович

Москва – 2022

Оглавление

Введение

Актуальность темы исследования.

Несмотря на постоянное развитие систем безопасности, человеческий фактор все еще остается одним из самых незащищенных элементов. Использование полиграфа позволяет уменьшить такие риски. Психофизиологические исследования решают следующие прикладные задачи: выявление негативных факторов в прошлом опыте кандидатов на должность, проверка на лояльность и соблюдение внутренних регламентов организации, осуществление оперативно-розыскной деятельности, проведение корпоративных и антикоррупционных расследований и т.д.

Полиграфная проверка является эффективным, но достаточно трудозатратным и требовательным к квалификации специалиста способом выявления скрываемой информации. Именно поэтому одной из основных целей данной работы является создание системы для автоматических рекомендаций полиграфологу. «Второе мнение» поможет оперативно принять решение или скорректировать саму процедуру полиграфной проверки.

Распознавание эмоций человека является важной научно-исследовательской проблемой, которая затрагивает такие дисциплины, как медицина и психология.

Распознавание эмоций решает прикладные задачи в следующих сферах деятельности: онлайн-обучение - построение учебного плана с учетом динамики вовлеченности учащихся на каждом этапе; банковское дело - дополнение скоринговых моделей информацией о возможных искажениях для выявления мошенников; колл-центры - управление удовлетворенностью во время звонка, составление независимого индекса потребительской лояльности; организация транспортной безопасности - контроль за состоянием водителя, сигнализирование о возможном переутомлении; производство эмпатичных роботов - дополнительный инструмент для выбора оттенков диалога.

Цели и задачи диссертационной работы:

В работе были поставлены следующие **цели**:

- Повысить точность методов и алгоритмов классификации эмоционального состояния человека на голосовой фонограмме.
- Разработать методы и алгоритмы для классификации эмоционального состояния человека на видеозаписи, устойчивые к изменениям условий съемки.
- Создать методы и алгоритмы классификации силы реакции организма на стимулы при помощи регистрируемых полиграфом параметров (дыхание, сердечно-сосудистая и электродермальная активность).

Для достижения поставленных целей были решены следующие **задачи**:

- Создание, исследование и подбор алгоритмов обработки голосовых фонограмм для классификации эмоционального состояния говорящего
- Исследование и разработка методов классификации эмоций человека на видеозаписях
- Нормализация психофизиологических характеристик, учитывающих индивидуальные особенности испытуемого.
- Разработка методов и алгоритмов автоматической классификации параметров, регистрируемых при помощи полиграфа – КГР, плетизмограмма, дыхательные циклы.
- Создание тестового приложения и проведение вычислительных экспериментов по определению работоспособности перечисленных методов.

Научная новизна. диссертационной работы состоит в следующем:

1. Создана архитектура нейронной сети для автоматической классификации голосовых фонограмм с высокой точностью;

2. Созданы новые методы классификации эмоций на видео, отличающиеся высокой устойчивостью при работе с материалами, записанными в нестудийных условиях;

3. Предложен метод нормализации психофизиологических характеристик, полученных при помощи полиграфа, учитывающий индивидуальные особенности испытуемого.

4. Созданы 3 новых метода автоматической классификации силы реакции человека (балльная оценка) на предъявляемый стимул при помощи оценки регистрируемых независимых параметров: дыхательных циклов, электрической активности кожи (КГР), сердечных ритмов (плетизмограммы).

Теоретическая и практическая значимость.

Результаты, изложенные в диссертации, применены для создания интеллектуальной системы оценки факторов риска при трудоустройстве и проведения служебных опросов.

Получено 2 свидетельства о государственной регистрации программы для ЭВМ:

1. №2021615620, «Программное обеспечение по оценке эмоционального состояния человека по видеопотоку в режиме реального времени с использованием искусственного интеллекта»;

2. №2022661019 «Программный комплекс на основе инновационной стандартизированной и валидизированной методики для проведения полиграфных проверок»;

Результаты работы реализованы и используются в следующих системах:

- Система автоматической балльной оценки проведенного тестирования в Профессиональном компьютерном полиграфе «Финист»;

- Модуль по оценке эмоционального состояния собеседника в аппаратно-программном комплексе «Детектрон».

ФГБУ «Фонд содействия развитию малых форм предприятий в научно-технической сфере» подтвердил практическую значимость указанных выше систем и обеспечил поддержку развития указанных выше модулей в форме грантов: №3493ГС1/57463 «Разработка прототипа программного обеспечения по оценке эмоционального состояния человека по видеопотоку в режиме реального времени с использованием программного обеспечения искусственного интеллекта» - 2020 г.; №240ГС1ЦТС10-D5/65720 «Разработка прототипа программного комплекса на основе инновационной стандартизированной и валидизированной методики для проведения полиграфных проверок» - 2021 г.

Положения, выносимые на защиту:

- Предложена и программно реализована архитектура нейронной сети для решения задачи определения эмоции на голосовой фонограмме с высокой точностью.
- Предложена и программно реализована архитектура многослойной нейронной сети для решения задачи определения эмоции человека на видеозаписи, подготовленной в нестудийных условиях.
- Предложен метод нормализации психофизиологических характеристик, полученных при помощи полиграфа, учитывающий индивидуальные особенности испытуемого.
- Создан модуль для автоматической классификации силы реакции человека на предъявляемые стимулы при помощи оценки регистрируемых полиграфом параметров (дыхание, сердечно-сосудистая и электродермальная активность).

Степень достоверности и апробация результатов.

Достоверность результатов подтверждена экспериментальной проверкой результатов предлагаемых методов на реальных данных, в том числе сторонними организациями; публикациями результатов исследования в рецензируе-

мых научных изданиях и конференциях по машинному обучению; воспроизводимостью результатов исследования при использовании различных тестовых наборов данных из публичных репозиториев; корректным использованием математического аппарата известных алгоритмов машинного обучения, стандартных метрик качества классификации, современных средств программирования и библиотек машинного обучения; публикациями результатов в рецензируемых научных изданиях, в том числе рекомендованных ВАК;

Основные результаты диссертации докладывались на следующих конференциях: II Всероссийская научная конференция с международным участием «От идеи – к практике: социогуманитарное знание в цифровой среде» - Новосибирск, 2022; 14-я международная научная конференция студентов и магистрантов «Современный специалист-профессионал: теория и практика» - Барнаул, 2022; Всероссийская научная конференция молодых ученых, посвященная Году науки и технологии в России «Наука. Технологии. Инновации» - Новосибирск, 2021.

Соответствие паспорту научной специальности. Область исследования и содержание диссертации соответствуют паспорту специальности 05.13.17 «Теоретические основы информатики (технические науки)», в частности по следующим пунктам:

пункт 5: Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений.

пункт 7: Разработка методов распознавания образов, фильтрации, распознавания и синтеза изображений, решающих правил. Моделирование формирования эмпирического знания.

Публикации. Материалы диссертации опубликованы в 3 печатных работах, из них 3 статьи в рецензируемых журналах из списка ВАК.

Личный вклад автора. Содержание диссертации и основные положе-

ния, выносимые на защиту, отражают персональный вклад автора в опубликованные работы. Подготовка к публикации полученных результатов проводилась совместно с соавторами, причем вклад диссертанта был определяющим. Все представленные в диссертации результаты получены лично автором.

Структура и объем диссертации. Диссертация состоит из введения, 3 глав, заключения и библиографии. Общий объем диссертации 104 страницы, из них 86 страниц текста, включая 14 рисунков. Библиография включает 159 наименований на 16 страницах.

Краткая характеристика содержания работы. Диссертация включает в себя 3 главы и одно приложение.

Глава 1. Методы распознавания эмоций на голосовой фонограмме. Сделан краткий обзор методов распознавания эмоций на голосовой фонограмме. Описаны методы классификации с использованием разложения аудиосигнала на эмпирические моды и вейвлет-анализа. Предложена архитектура сверточной нейронной сети для распознавания с высокой точностью.

Глава 2. Методы распознавания эмоций на видеозаписи. Сделан краткий обзор методов распознавания эмоций на изображениях и видеозаписях. Описан метод покадровой классификации с применением локальных бинарных паттернов. Предложен метод синтеза нового набора данных для распознавания эмоций на видео и подход к переобучению сверточной нейронной сети. В результате этого была создана архитектура сети для автоматической классификации эмоции на видеозаписях, устойчивая к различным изменениям условий видеосъемки.

Глава 3. Методы классификации в исследованиях на полиграфе. Описан принцип работы полиграфа. Исследованы возможности применения часто используемых архитектур нейронных сетей для оценки психофизиологических характеристик, зарегистрированных при помощи полиграфа. Предложена архитектура трансформера для решения данной задачи. Визуализированы полученные результаты классификации.

История развития метода выявления сокрытия информации.

Потребность выявлять ложь существовала всегда. Исторические хроники повествуют о том, что методы выявления были абсолютно различными: от сложных ритуалов и божественного суда до физических пыток.

Н.М. Карамзин в своих комментариях к первому своду российских законов пишет: «древние россияне, подобно другим народам, употребляли железо и воду для изобличения преступников — обыкновение безрассудное и жестокое. Обвиняемый брал в голую руку железо раскаленное или вынимая ею кольцо из кипятка, после чего судьям надлежало обвязать и запечатать оную. Ежели через три дня не оставалось язвы или знака на его коже, то невиновность была доказана. Народ думал, что Богу легко сделать чудо для спасения невиновного». Аналогичные нормы и законы были закреплены и в заподноевропейских варварских правдах.

Но очевидно, что установление истины и изобличение лживых показаний такими методами носит случайный характер и представляется сомнительным.

В то же время развивался и метод наблюдений. Люди обратили внимание на то, что при сокрытии информации лжец ведет себя особенным образом. Прежде всего в случаях, когда человек осознавал возможное наказания и испытывал сильный страх перед разоблачением. Наблюдатель мог зафиксировать физиологические изменения у лжеца по внешним признакам.

Например, китайцы определяли пересыхание слюнных желез при помощи измерения влажности горсти рисовой муки во рту подозреваемого во время зачитывания конкретных обвинений. В Индии предлагали называть ассоциативный ряд на нейтральные и явно связанные с совершенным преступлением фразы, в то же время ударяя в гонг. Они обратили внимание на то, что в таких случаях причастные к нарушению закона сопровождали ответы на критические для них стимулы более громким звоном [20].

Древние римляне оценивали возможность сокрытия своих эмоций, чтобы понять предрасположенность к заговорам. При отборе телохранителей их могли

бить по лицу и задавать провокационные вопросы. Те, кто краснел и показывал свои явные эмоции, были в приоритете.

Первое упоминание про анализ физиологических процессов в мировой литературе можно встретить у Д. Дефо, автора знаменитой книги про приключения Робинзона Крузо. В 1730 году писатель опубликовал трактат «Эффективный проект непосредственного предупреждения уличных ограблений и пресечения всяких иных беспорядков по ночам». Дефо обратил внимание, что «у вора существует дрожь (тремор) в крови, которая, если ею заняться, разоблачит его». Такой подход был применим даже для психологически подготовленных преступников: «Некоторые из них настолько закостенели в преступлении, что даже смело встречают преследователя; но схватите его за запястье и вы обнаружите его виновность».

Так или иначе издревле оценивалось поведение человека при предъявлении какого-то стимула, а все измерения проводились методом наблюдения по причине отсутствия специального инструментария. Тем самым начал формироваться психофизиологический способ выявления скрываемой информации, в основе которого лежит неслучайность зависимости динамики физиологических процессов и внутреннего состояния человека. В результате метод стал безопасным и беспристрастным.

Приоритет в формулировании генерального принципа психофизиологического метода выявления скрываемой информации по праву принадлежит психологу А.Р. Лурия [13], который в 1920-х годах писал, что «единственная возможность изучить механику внутренних «скрытых» процессов сводится к тому, чтобы соединить эти скрытые процессы с каким-либо одновременно протекающим рядом доступных для объективного наблюдения процессов, в которых внутренние закономерности и соотношения находили бы свое отражение».

Развитие приборов для измерения артериального давления и кровонаполнения.

Прародителем полиграфа можно считать гидроплетизмограф – устройство

для графического определения колебаний объема различных частей тела в зависимости главным образом от кровенаполнения.

Данный прибор был выполнен из суживающегося с одной стороны стеклянного цилиндра, соединенного трубкой с резервуаром для воды. Также в колбе имелось три отверстия: одно для слива воды после завершения процедуры, второе для соединения с записывающим устройством в виде механического осциллометра, третье для ввода и фиксации объекта для исследования. После установки руки в устройство, ее обтягивали резиновым рукавом, а в резервуар наливали воду до полного вытеснения воздуха из цилиндра. Таким образом создавалась герметически замкнутое пространство, соединенное с осциллометром, перо которого фиксировало изменения объема руки.

В 1877 году при помощи гидроплетизмографа итальянский физиолог А. Моссо зафиксировали, что во время приема в клинике внезапно и без явных причин увеличились пульсации у одной пацентки. Моссо описал этот случай: «это поразило меня, и я спросил женщину, как она себя чувствует. Она сказала, что хорошо. Я тщательно проверил прибор, чтобы убедиться, что все в порядке. Тогда я попросил пациентку рассказать мне, о чем та думала две минуты назад. Она ответила, что, рассматривая отсутствующим взором книжную полку, висевшую напротив, остановила свой взгляд на черепе, стоявшем среди книг, и была напугана им, так как он напомнил ей о болезни». В результате Моссо опубликовал материалы и результаты своих экспериментов в монографии под названием «Страх».

В 1895 году методики выявления стресса изложил выдающийся итальянский криминалист Чезаре Ломброзо в своей широко известной книге «Преступный человек» практический опыт применения гидроплетизмографа в ходе проверки фигуранта по уголовному делу об ограблении. При проведении психофизиологического исследования автор не смог зафиксировать видимые изменения динамики артериального давления в ответ на предъявление стимулов, связанных с ограблением, но обнаружил резкое уменьшение базовой линии осцилло-

метра по другому делу, связанному с хищением паспортов. Позднее оперативными методами удалось подтвердить правильность выводов Ломброзо.

В 1902 году он привлекался к расследованию уголовного дела об изнасиловании и убийстве девочки. Ломброзо вновь применил гидроплетизмограф и обнаружил незначительные изменения в пульса у подозреваемого, когда он делал математические вычисления в уме. В то же время у него не было внезапных изменений динамики артериального давления при предъявлении фотографии убитой девочки. В дальнейшем результаты расследования подтвердили невиновность подозреваемого.

В 1854 году немецким врачом Карлом фон Фирордтом был изобретен сфигмограф. Основной задачей прибора является графическое отображение свойств артериального пульса. По полученным результатам можно оценить динамику изменения кровяного давления в артериях и ритм сердечных сокращений.

В 1860 году французский физиолог и изобретатель Э. Маре сконструировал усовершенствованную версию сфигмографа. Данный прибор регистрировал колебания пульса лучевой артерии на движущейся пластинке при помощи рычага. Колебания передаются на рычаг через пелоту. Пелота накладывается на пульсирующую артерию и закрепляется винтом, который помогает преодолеть толстый слой кожи, обеспечивая необходимое надавливание на пластинку, и связывает рычаг с артерией. Спереди устройства располагается штифт с винтовой поверхностью. При его движении вверх-вниз происходит зацепление с зубчатым колесом, посаженным на ось. Далее данное колесо двигает полоску закопченной бумаги, на которой и рисуется сфигмограмма. Все это фиксируется на руке при помощи шин.

В начале 20-го века итальянский криминалист Э.Ферри опубликовал труд «Уголовная социология», в котором предложил в качестве одного из методов проверки истинности показаний подозреваемых использовать сфигмограф.

Развитие приборов для измерения кожно-гальванической реакции.

Другим каналом регистрации физиологических изменений в полиграфе

является электрокожное сопротивление. Французский ученый Дюбуа-Реймон первым заметил электробиотоки на изолированной коже лягушки, которые по своей величине превосходили нервные и мышечные. Подобные эффекты в дальнейшем были названы кожно-гальванической реакцией (КГР).

В 1888 г. Ч. Фере, обследуя больную с жалобами на электрические покалывания в кистях и ступнях, обнаружил, что при пропускании слабого тока через предплечье происходили отклонения стрелки включенного в цепь гальванометра в моменты сенсорных или эмоциональных воздействий. Независимо от Фере в 1890 г. И. Тарханов показал, что электрические сдвиги наблюдаются и без приложения внешнего тока. Он установил, что любое раздражение, нанесенное человеку, через 1-10 секунд латентного периода вызывает сначала легкое и медленное, а затем ускоряющееся отклонение стрелки гальванометра, иногда даже выходящее за пределы шкалы. Оба метода, как показатели состояния организма, дают идентичные результаты, только латентный период изменения сопротивления кожи несколько выше, чем при изменении потенциалов кожи.

Карл Юнг рассматривал данный сигнал, как объективное физиологическое «окно» в сферу бессознательного, подлежащего изучению через психоанализ. Он же первым выявил прямую зависимость между величиной КГР и силой эмоционального переживания.

Развитие приборов для измерения дыхания.

Пневмограф - аппарат для измерения и графической регистрации дыхательных движений грудной клетки или животаго изображения дыхательных движений. Он был изобретен русским ученым Г.Н. Пио-Ульским в 1900 году.

Прибор состоит из манжетки от сфигмоманометра или полой резиновой трубки, соединенной с капсулой Марея. В регистрирующую систему вводят немного воздуха и закрывают краном от внешнего мира. После этого изменения объема легких визуально фиксируется на чернильной ленте. Аппарат дает возможность проанализировать дыхательные ритмы, длительность дыхательных фаз (вдох, выдох, пауза).

На сегодняшний день для фиксации дыхательных движений используют не только пневматические, но и пьезоэлектрические датчики.

В 1914 году итальянец В. Бенусси использовал анализ динамики процесса дыхания (изменения частоты и глубины) при проведении допросов подозреваемых в совершении преступлений. На тот момент не существовало методики по оценке дыхательных ритмов, но она начала формироваться.

Аналоги полиграфных устройств.

В 1951 году Дж. Даусон разработал технику фотографической суперпозиции (наложения) кривых ЭЭГ, которые могут возникать при многократном предъявления одного и того же конкретного стимула. Такие исследования называют техникой «вызванных потенциалов» (ВП).

Активность головного мозга можно измерить при помощи ЭЭГ, причем возможен анализ как позитивных (положительных) и негативных (отрицательных) волн, вызванных конкретным стимулом, с периодом развития реакции около 300 мс.

Особенность и его новизна для применения в области детекции лжи заключалась в изучении ответов головного мозга на конкретных стимул, вместо анализа разрозненных и малодифференцированных мозговых процессов.

К сожалению, применение ЭЭГ требует особых условий для проведения исследований, такие как:

• Постоянная фиксация глаз для избежания артефактов, вызванных морганием, особенно при записи активности от передних отделов мозга.

• Расслабленное состояние мышц головы и шеи на всем протяжении исследования для уменьшения электрической активности.

• Повторение одного и того же стимула от десяти до нескольких тысяч раз в зависимости от вопросов, которые подлежат проверке.

Очевидно, что соблюсти такие серьезные ограничения на протяжении длительного времении, а тем более в «полевых» условиях попросту невозможно. В то же время такие каналы информации, как КГР и частота сердечных сокра-

щений являются очень устойчивыми к тоническому мышечному напряжению.

В 1968 году была проведена первая запись электромагнитного поля мозга человека Д. Коеном. Несмотря на невысокую первоначальную чувствительность методы, изобретение сверхпроводникового квантомеханического интерферационного датчика, работающего на жидком гелии, позволило на порядок повысить точность и пространственную разрешающую способность метода.

Данный метод мог бы претендовать на более точные и устойчивые измерения, но применение его на практике сильно усложняется стоимостью оборудования, основанного на криотехнологиях, и серьезными требованиями к магнитной защищенности помещения.

Именно по этим причинам полиграф остается самым подходящим для массового применения в целях выявления скрываемой информации.

Распознавание эмоции в оценке поведения человека.

Одно из определений эмоций описывает их как «психическое отражение в форме непосредственного пристрастного переживания отношения конкретных явлений и ситуаций к потребностям» [3]. В таком толковании делается акцент на возможность удовлетворения (или фрустрации) потребности, причем в неразрывной связи с такими событиями, явлениями и предметами окружающей действительности. Возникновению эмоций неизбежно предшествует появление мотива деятельности (шире – поведения).

В деятельностной концепции эмоции в узком смысле этого слова определяют как отношение результата деятельности к ее мотиву. Так, говоря абстрактно, радость возникает у человека, когда мотив его деятельности реализован, страх возникает, когда под угрозой находится мотив самосохранения, раздражение возникает в том случае, если на пути к реализации мотива человек сталкивается с каким-либо непредвиденным препятствием и т.д.

Таким образом, эмоции являются очень важным индикатором, говорящим об истинных мотивах поведения конкретного человека.

Глава 1

Методы распознавания эмоций на голосовой фонограмме

1.1. Краткий обзор существующих подходов и постановка задачи

Рассматривая задачу распознавания эмоций человека на голосовой фонограмме, стоит обратить внимание на существующие на настоящий момент решения в этой области [132, 126]. Несмотря на определенную субъективность при оценке такой характеристики, как проявление эмоционального состояния на аудизаписи, некоторые наборы данных отвечают всем необходимым требованиям [33, 125].

Чаще всего для оценки эмоций выделяют просодические (характеризующие речевую мелодию, темпоральные и тембральные особенности голоса) и спектральные характеристики аудиофайла с последующей классификацией полученных данных.

Различные исследования показали, что для распознавания эмоций требуется анализировать такие характеристики голоса, как высота тона, энергия, длительность фонем, перцептивные линейные прогностические коэффициенты (PLP), линейные прогностические кепстральные коэффициенты (LPCC), мел-частотные кепстральные коэффициенты и их комбинации.

Одним из часто используемых методов в исследовании звуковых сигналов является вейвлет-анализ [146]. Он применяется для шумоподавления, обнаружения, сжатия, классификации и других операций с аудио данными [57, 124, 44, 75, 74, 66, 24, 30].

Для решения задачи классификации эмоций с применением вейвлет-анализа применяются следующие методы:

- Вычисление вейвлет-характеристик вместе с мел-кепстральными коэффициентами MFCC и результатами применения дифференциального энергетического оператора (TEO) [152];
- Расчет энтропии биортогонального вейвлета [155].
- Вычисление следующих вейвлет-преобразований: непрерывное (CWT), бионическое (BWT) и синхронно-сжатое (SSWT) [141].
- Расчет характеристик стационарного вейвлет-преобразования, энергии пакета вейвлета и энтропийных характеристики [111].
- Настроенное вейвлет-преобразования Q-фактора (TQWT) и пакетное вейвлет-преобразование (WPT) [157].
- Пакетное вейвлет-преобразование на основе энергии и нелинейной энтропии [66].
- Повышения точности классификации на основе пакетного вейвлет-преобразования с помощью последовательного плавающего прямого поиска (SFFS) [148].

Классифицировать семь эмоций сложнее, чем классифицировать шесть эмоций для базы данных EMODB [32, 135, 145].

Согласно исследованию набора данных EESDB (The elderly emotional speech database) точность распознавания эмоций пожилых людей ниже, чем у молодых [50]. Возможно это связано с изменением голосовых связок с возрастом.

Выбор признаков, который важен для распознавания эмоций [26, 51, 99], можно разделить на две категории: контролируемые и неконтролируемые методы. Недавно был предложен ряд новых стратегий выбора признаков, таких как полууправляемый метод [39] и разработка паранепротиворечивых признаков [63].

В этом исследовании используется последовательный плавающий поиск вперед (SFFS), который является одним из видов методов выбора признаков, впервые предложенных в [123]. Он широко использовался в области классификации речевых эмоций [141, 147]. Чтобы уменьшить количество векторов признаков и улучшить производительность классификации, был выбран итеративный подход SFFS для повышения производительности распознавания [148].

В статье [148] были описаны особенности пакета вейвлета для изучения их влияния на производительность распознавания эмоций на голосовых фонограммах. Производительность классификации с использованием вейвлет-пакетов сравнима с мел-кепстральными коэффициентами MFCC. Чтобы уменьшить пространство признаков, был применен метод последовательного плавающего прямого поиска (SFFS). При этом была достигнуто увеличение точности распознавания.

В [94] для распознавания эмоций была предложена модель Маркова со скрытой глубокой нейронной сетью (DNN-HMM), которая широко использовалась для работы с речью.

Коэффициенты вейвлет-пакетов оцениваются для каждого кадра. Как показано в (1.42), $C_{j,k}$ есть k-й коэффициент в j-уровневой декомпозиции речи. После 5-уровневого разложения имеется 32 коэффициента: $C_{5,0}, C_{5,1}, C_{5,2}, \ldots, C_{5,31}$ соответственно.

Исследования [54, 29, 41, 34, 62, 42] показали, что признаки, содержащие глобальную информацию, лучше локальных с точки зрения вычислительной эффективности и производительности распознавания. Итак, максимум коэффициента рассчитывается как в [29, 62, 42] из EESDB.

На сегодняшний день несколько исследователей разрабатывают методы распознавания эмоций с акцентом на независимость от диктора. Авторы выделили признаки, связанные со статистикой высоты тона, формант и энергетических контуров, а также спектр, кепстр, перцептивные и временные признаки, автокорреляцию и другие - всего 2327 признаков [92]. В другой работе были вы-

делены признаки с использованием энергии, скорости пересечения нуля, MFCC и параметров Фурье [145]. Дикторонезависимые признаки дают более устойчивые результаты при распознавании эмоций [92].

Мел-частотные кепстральные коэффициенты (MFCC) были впервые введены в работе [45]. Исследования в [29, 92, 145] использовали данные показатели для распознавания речи. Мел-частотные кепстральные характеристики описывают восприятие человеческого слуха, благодаря своей шкале.

Также были рассчитаны глобальные признаки, поскольку они могут уменьшить количество признаков. Пакет вейвлетов поддерживает множество алгоритмов, таких как вейвлет-фильтр Добеши и фильтр Габора [44].

Вейвлеты Добеши широко применяются для решения задач обработки сигналов, особенно речевых сигналов. В текущем исследовании было выбраны семейства вейвлетов Добеши [44]. DB2 изначально применялась для распознавания эмоций.

Семейство вейвлетов Добеши широко используется в распознавании речи [95], сжатии речи [93] и во многих других областях обработки сигналов [124, 56, 53]. Однако данный метод редко применяется для распознавания речевых эмоций.

Производительность системы распознавания эмоций на аудиозаписях зависит от качества: функций, используемых для различения эмоций, классификаторов и набора данных, используемых для обучения.

Для распознавания эмоций на голосовых фонограммах используются разные методы классификации. Например, опорные вектора (SVM) [109, 154, 37, 117], метод смеси Гауссовых распределений (GMM) [133], скрытая марковская модель (HMM), нейронные сети (NN) [122, 133, 37], рекуррентные нейронные сети (RNN) [98, 110, 80] и линейная регрессия (LR) [79].

Также имеет смысл рассмотреть методы выделения информативных признаков из голосового сигнала.

Наиболее популярным способом получения дополнительных признаков из

сигнала является быстрое преобразование Фурье (FFT). При помощи него функция представляется в виде суммы гармонических колебаний с разными частотами. Для упрощения вычислений можно использовать метод быстрого преобразования Фурье. Данные подходы основаны на гипотезах линейности и стационарности сигнала. Тем не менее, быстрое преобразование Фурье метод теряет часть информации, которая может быть полезна для задачи классификации во временной области [144].

Для минимизации проблемы стационарности возможно использование кратковременного преобразования Фурье (STFT) [67, 144, 112]. Даный метод заключается в повторении умножения сигнала на короткие временные окна со сдвигом и выполнении преобразования Фурье для полученного сигнала. В то же время, краткосрочное преобразование Фурье (STFT) также имеет ограничения в соответствии с фундаментальным принципом неопределенности. Согласно которому время и частота не могут быть одновременно определены с одинаковой точностью. Данный метод не решает проблему нелинейности сигнала [52].

Для преодоления сложностей повышения точности распознавания эмоций по речи, связанных с вариативностью и незначительностью изменений нелинейных характеристик в проявлениях разных человеческих чувств речевой эмоции, используется метод эмпирических мод (EMD) [70]. Сигнал раскладывается на составляющие и анализируется без потери первоначальных свойств, несмотря на линейность и/или нестационарность. Метод эмпирических мод (EMD) успешно применялся в области распознавания эмоций в голосовых фонограммах [96, 97, 128, 158].

Достижения в области нетрадиционного анализа речевых сигналов хорошо описан в [130]. Метод эмпирических мод (EMD) раскладывает временной ряд на внутренние колебания (IMF), имеющие самую высокую локальную частоту. Однако такой подход имеет недостаток в виде смешивания мод [149, 58]. С одной стороны это связано с дефектами сигнала, с другой - с изъянами самого метода. Для решения описанной проблемы используется метода ансамбля декомпо-

зиции эмпирических мод (EEMD). Данный подход заключается в добавлении к исходному сигналу гауссовского белого шума, среднее значение которого равно нулю.

Метод эмпирических мод EMD в сочетании с оператором энергии также используется в качестве альтернативного метода повышения эффективности частотно-временного анализа сигналов [27]. Для корректировки элементов поддиапазона могут использовать различные функции. Достижение наиболее продуктивных результатов возможно при совместном использовании метода эмпирических мод (EMD) и оператора энергии Тигера-Кайзера (TKEO) [96, 97]. Такое отображение может отслеживать мгновенные амплитуду и частоту компонентов модулированного сигнала (AM-FM) в любой момент времени.

Также комбинация метода эмпирических мод (EMD) и оператора энергии Тигера-Кайзера (TKEO) показывает хорошие результаты частотно-временного анализа при демодуляции сигнала [83].

Для повышения точности классификации также могут быть применены алгоритмы удаления избыточной и нерелевантной информации из результатов обработки сигнала [137]. Например, можно извлечь признаки при помощи линейного дискриминантного анализа (LDA), метода главных компонент (PCA), последовательного прямого выбора (SFS), рекурсивного удаления признаков (RFE).

Близкими к данному исследованию выступают архитектуры, разработанные в университетах Пассау в Германии [46], Калифорнии [36] и Техасса [22]. Сравнительный анализ вышеуказанных алгоритмов, основанных на CREPE, привел к выявлению следующих недостатков:

1. Использование частичного обучением учителя в [46] приводит к нестабильным промежуточным результатам и потере точности.

2. Применение генеративно-состязательных сетей в [36] подразумевает повышение качественных требований к набору данных, а также усложненному процессу обучения и генерации результатов.

3. Сложный алгоритм, описанный в [22], имеет в основе адаптацию алгоритмов обучения с помощью метода опорных векторов, примененного к синтетическим данным (автоматически сгенерированных алгоритмом), для дальнейшего использования с реальными данными (доменная адаптация). Помимо сложности имплементации такая система имееть увеличенную вычислительную стоимость и базируется на условных правилах (так называемая “rule-based система”). При наличии образцов данных на момент прогнозирования, выходящих за установленный набор, точность такой системы окажется ниже расчетной.

У всех представленных алгоритмов имеется следующий недостаток: отсутствие непосредственного анализа информации сигнала, так как обучение производится при помощи данных, либо полученных от внутренних преобразований сети, либо от препроцессирующих алгоритмов.

Учитывая описанные выше моменты, особое внимание было уделено алгоритму CREPE, представленному в [84] и являющегося продолжением работ над алгоритмами YIN [40] и pYIN [108]. Упомянутые публикации являются инновационными для задач определения частоты основного тона (также называемой F_0 или Fundamental Frequency) в монофоническом аудиоматериале.

В центральном месте алгоритма CREPE находится сверточная нейронная сеть, производящая обучение на непосредственно аудиосигнале во временной области.

К сожалению, на сегодняшний день отсутствуют русскоязычные наборы данных для распознавания эмоций на голосовых фонограммах. Исследование будет продолжено по факту подготовки таких материалов.

1.2. Набор данных для классификации

Рассматривается следующая задача классификации:

X - множество голосовых фонограмм x;

$x_i = \{x_{i1}, \ldots, x_{il}\}$ - аудиозапись, длительностью t с.

$Y = \{$Спокойствие, Радость, Грусть, Злость, Страх, Удивление, Отвращение$\}$ - множество классов из семи эмоций.

$X^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, где m - размер набора данных

Требуется построить алгоритм $a : X \to Y$, способный классифицировать произвольный объект $x \in X$

Для обучения был выбран набор данных Ravdess [100], представляющий из себя 1440 аудиозапией, на которых актеры произносят два предложения на английском языке по 2 раза каждое (обычное произношение и нараспев). Предложения произносятся 2 раза для записи сильного и слабого проявления. Каждая запись длится в среднем 4 с, в первой и последней секундах присутствует лидирующий и заключительный отрезок без звука.

Аудиоматериал записан в стереоформате, частота семплирования равна 48 кГц. Каждый аудиофайл имеет метку с эмоцией (радость, спокойствие, грусть, злость, страх, отвращение, удивление), которую испытывал актер при записи.

Выбранный набор данных RAVDESS является общедоступным и наиболее полно аннотированным по сравнению с остальными.

1.3. Метод сверточных нейронных сетей

В соответствии с основной идеей работы алгоритма CREPE (непосредственная работа над характеризующей сигнал графической информацией) были рассмотрены несколько вариантов сверточных нейронных сетей с некоторыми внутренними различиями, которые включают в себя: серьезные отличия архитектур, разное количество слоев свертки, групп слоев, применение дополнительных техник предотвращения эффекта переобучения (over-fitting, dropout, regularization).

Нейронная сеть, как известно, носит такое название в силу того, что состоит из некоторого количества вычислительных единиц – нейронов. Эти единицы способны получать, обрабатывать и отправлять любую информацию дальше.

Делятся нейроны на три основных вида (входной, выходной, скрытый) и два вспомогательных (нейроны смещения, контекстный). Чтобы улучшить обработку информации при наличии большого количества нейронов, их совмещают в слой. Они также разделяются на входной, выходной и скрытый слой. Общий принцип работы основан на том, что каждый нейрон имеет два параметра: входные и выходные данные. Дальнейшие действия сводятся к простому циклу: входной нейрон или слой получает введенную информацию, после чего обрабатывает ее и отдает на скрытый нейрон или слой. Во всех последующих скрытых нейронах или слоях информация обрабатывается, и каждая последующая передача сопровождается собранной информацией каждого нейрона или слоя. В конце функция активации нормализует все полученное и отдает на выходной нейрон или слой, который выводит результат.

Чтобы решить более значительную задачу, например задачу классификации, нейроны собирают в общую систему - искусственную нейронную сеть [14]. Как известно, при наличии в нейронной сети более одного скрытого слоя такую сеть принято называть глубокой [134]. В большом многообразии различных архитектур были выделены и рассмотрены лишь подходящие для цели исследования. В качестве оптимальных для данной задачи изначально рассматривались: полносвязная нейронная сеть, когнитрон, перцептрон и сверточная нейронная сеть.

В полносвязной нейронной сети присутствует множество простых процессоров, которые сами по себе могут только совершать тривиальные операции. Каждому такому процессору (т.е. нейрону) назначается одна из задач: входные принимают набор данных, обрабатывающие совершают простые математические операции над набором, выходные используются для дальнейшей передачи. В итоговом счете каждому пикселю изображения ставится в отношение один нейрон. Это имеет место в большинстве вариантов таких архитектур. Такой подход в машинном обучении прост в использования. Однако расчеты занимают большое количество времени и задействованных нейронов, а качественная

оценка результатов может разниться из-за плохого качества изображения или наличия шума, не видного человеческому глазу. Упомянутые причины снимают приоритет с данного выбора.

Как известно, когнитрон и перцептрон являются двумя сходными архитектурами. Оба варианта в основе имеют принцип обработки изображения человеческим мозгом зрительной корой, но есть различия во внутренней архитектуре. В перцептроне клетки одного слоя не связаны между собой, но соседние слои полностью связаны. При обработке объекта нейроны реагируют на него и дают сигнал (по аналогии с реакцией зрительной коры мозга на попадание света на сетчатку глаза). В когнитроне имеется иерархическая многослойная организация, в которой нейроны между слоями связаны только локально. Несомненно, достоинством, общим для двух архитектур является то, что когнитрон и перцептрон дают более точные результаты, по сравнению с полносвязными нейронными сетями. Но стоит отметить, что даже малейшие изменения изображения могут восприниматься ими как совершенно новый объект изучения (что требует постоянного дополнения набора данных для более полного охвата предметной области задачи).

Как известно, в сверточной нейронной сети имеются слои, выполняющие операцию свертки. Каждый фрагмент изображения умножается на матрицу (ядро) свертки поэлементно, а результат суммируется и записывается в аналогичную позицию выходного изображения. Информация проходит распределение определенных свойств изображения, в которых выделяются более абстрактные детали. Структуру и распределение этих абстрактных признаков и ядро свертки нейронная сеть определяет самостоятельно в процессе обучения, обретая способность фильтрации деталей и выделения существенных признаков.

По причине того, что сверточная нейронная сеть нацелена на высокую точность распознавания образов и лучшую из предложенных работу по классификации изображения, данный вариант рассматривался как наиболее приоритетный, что выделяет эту архитектуру как самую эффективную для дальнейшей

работы.

Для работы с аудиофайлами и их последующей обработки предварительно необходимо рассмотреть схему создания звуковых волн в речевом тракте. Несмотря на то, что при исследовании нет возможности создать трехмерную схему траектории движения звуковой волны, достаточно описать общие характеристики данных акустических процессов с учетом аэродинамических свойств. Теория речеобразования достаточно полно описывает приведенную схему.

Как известно, речевым сигналом называется функция возбуждения с откликами линейных фильтров. В этом случае в качестве функции возбуждения выступает шум. В пределах 90-300 Гц колеблется основной тон человеческой речи, который является уникальным для каждого отдельно взятого индивида. В пределах 90-180 Гц располагается частота мужских голосов и в пределах 185-300 Гц – частота женских и детских голосов. Набор гармоник, кратных основному тону, представляет щелчок голосовой щели. Падение уровня энергии гармоник напрямую зависим от увеличения частоты, 18 кГц — это максимальная граничная частота речевого сигнала, но для тракта достаточно частоты до 3500 Гц. При таком частотном ряде часть фонем не воспринимается человеческим ухом.

Резонансные полости речевого тракта напрямую используются щелчком голосовой щели. В этот момент часть гармоник, кратных основному тону, резонируют и созданные в спектре локальные максимумы образуют области концентрации энергии, которые называются формантами. Четыре форманты служат для создания гласных фонем, а любые другие изменения образуют согласные звуки. Все вышеперечисленное называют фонемами. Однако форманта также может служить для составления метрик на аудиоматериале речи человека, так как принадлежит к статическим характеристикам речи.

Если рассматривать образование речи как создание легкими, бронхами и трахеей акустической волны, которая образует речь посредством изменения траектории в голосовом тракте, то голосовой тракт (совокупность вышеназванных

органов) можно представить как резонатор с несколькими пиками амплитудной частотной характеристики, частоты которых определяют вид произносимой фонемы и соответственно состояние человека.

Реализованный на начальном этапе исследований простой алгоритм производит перевод итогового аудиосигнала в соответствующий набор параметров в рамках описанного теоретического материала и в последствии –, в графический вид. Благодаря своей информативности в сравнении с остальными вариантами была выбрана спектрограмма – двумерная диаграмма с прямой зависимостью, где по вертикальной оси показана частота, по горизонтальной оси – время, а амплитуда на определенной частоте в каждый конкретный момент времени представлена цветом.

Однако, несмотря на большую меру информативности спектрограмм, на этапе первичного обучения сверточной нейронной сети не было получено должной ориентировочной точности классификации эмоций, что привело к выдвижению гипотезы о применении психофизической шкалы.

Известно, что человеческое ухо более чувствительно к изменениям звука на низких частотах, чем на высоких. Это значит, что если частота звука изменится со 100 Гц на 120 Гц, то человек с очень высокой вероятностью распознает это изменение. Однако изменение частоты с 10000 Гц на 10020 Гц сложнее для восприятия человеческим ухом.

Такая особенность слуха учтена в одной из единиц измерения высоты звука — мел. Она основана на психофизиологическом восприятии звука человеком и логарифмически зависит от частоты, что непосредственно приводит к использованию мел-спектрограмм:

$$m = 1127 \ln(1 + \frac{f}{700}) \tag{1.1}$$

где m – высота звука в мелах, f – частота звука в Гц.

Мел-спектрограмма - это вариант спектрограммы, где частота выражена

не в Гц (рис.1.1), а в мелах (рис. 1.2). Переход к мелам происходит с помощью применения шкалирования исходной спектрограммы.

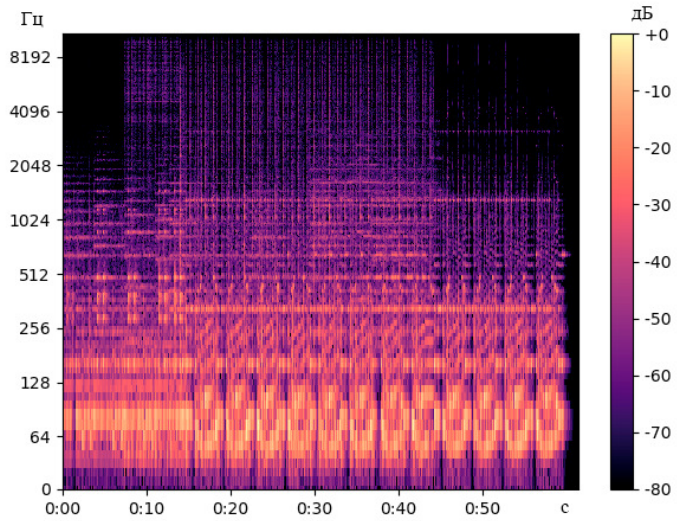

Рис. 1.1. Пример изображения со спектрограммой аудиофайла

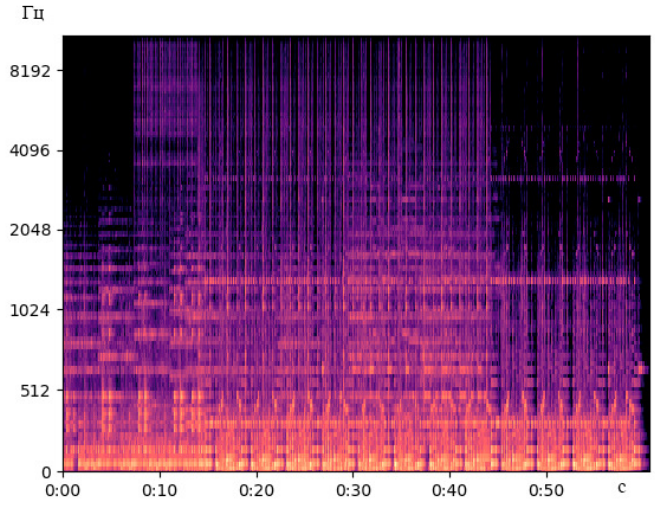

Рис. 1.2. Пример изображения с мел-спектрограммой аудиофайла

Фильтрация шумов

Чтобы в дальнейшем система работала достаточно устойчиво на практике, требуется уменьшить помехи и шумы в таких аудиозаписях. Можно сказать, что требуется приблизить их качество к студийному. Такое изменение качества фонограммы можно сделать следующим образом:

• Рассчитывается результат быстрого преобразования Фурье по всему аудиофайлу;

• При помощи быстрого преобразования Фурье рассчитываются частотные статистические параметры;

• На основе этих параметров устанавливается порог шума и желаемая чувствительность алгоритма;

• Рассчитывается быстрое преобразование Фурье на зашумленном сигнале аудиофайла;

• Создается маска, сравнением быстрого преобразования Фурье шумного сигнала и порога;

• Маска выравнивается при помощи фильтра по частоте и временному домену сигнала;

• Маска применяется к быстрому преобразованию Фурье аудио и результат инвертируется.

Выбор архитектуры сверточной нейронной сети.

Данная система имеет следующие входные и выходные данные. В качестве входных данных взяты 1024 выдержки из аудиосигнала в моноформате во временной области с частотой дискретизации 22 кГц. Они обрабатываются при помощи шести сверточных слоев.

Выходными данными является тензор размерностью 2048, который затем передается на полносвязный выходной слой классификации с активирующей функцией сигмоидой размерностью в 360 нейронов. Каждый из 360 элементов выходного вектора соответствует конкретному значению высоты звука, выражаемой в центах.

Цент — единица частотного интервала, равная $1/1200$ части октавы. Таким образом данная шкала покрывает диапазон звуков в диапазоне частот от 32.70 Гц до 1975.5 Гц.

Ключевой характеристикой голосового сигнала является частота основного тона. С музыкальной точки зрения — это образующая для всех остальных

звуков натурального звукоряда, а для человеческой речи — частота колебаний голосовых связок. Она присуща непосредственно их обладателю, а ее повышение воспринимается слушателем как повышение высоты звука. Таким образом, возможно следующее предположение: решение задачи определения эмоций по монофоническому аудиоматериалу можно осуществить, использовав набор инструментов, схожий с задачей определения частоты основного тона алгоритма CREPE.

При помощи вспомогательных библиотеки функций librosa и matplotlib производится первичная обработка файлов: приведение материалов к монофор-мату, децимация аудиофайлов до частоты 22 кГц, подготовка изображения с мел-спектрограммой аудизаписи с разрешением 640 на 480 пикселей (рис. 1.2). Именно на этом этапе все аудизаписи были преобразованы в изображения, в которых по горизонтальной оси приведено время, по вертикальной оси — частота. Третье измерение с указанием амплитуды на определенной частоте в конкретный момент времени представлено интенсивностью желтого цвета каждой точки изображения.

Посредством библиотеки Keras производится загрузка, нормализация, разделение на обучающую и тестовую выборки, обучение и выбор наилучших архитектур, составление матрицы ошибок и классификационного отчета по распознаванию эмоций.

Первоначальный вариант архитектуры, который показал неудовлетворительный для данного исследования результат, выглядел следующим образом

блок №1: 1 слой свертки и 1 слой пулинга.

блок №2: 1 слой свертки и 1 слой пулинга.

слой выравнивания

слой выброса

полносвязный слой

Для повышения точности классификации были использованы следующие методы: использование различных комбинаций увеличения блоков и сверточ-

ных слоев, нормализация пакета данных для одной итерации (батчей), тюнинг гиперпараметров с помощью алгоритма RandomSearch.

По результатам проведенных исследований были получены следующих результаты:

1conv_2blocks – 80.2,

2conv_2blocks – 78.29,

1conv_3blocks – 72.56,

2conv_3blocks – 78.81.

По результатам обучения отмечено следующее: сверточные сети, имеющие в своей архитектуре один сверточный слой, показали более выскоую точность классификации по сравнению с результатами сетей, имеющих два сверточных слоя;

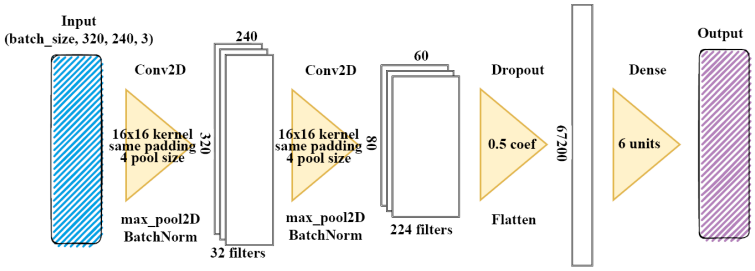

Рис. 1.3. Архитектура нейронной сети

Наивысший показатель валидационной точности имеет сеть с архитектурой, представленной на рис. 1.3:

- входной слой (Input),
- Блок №1: сверточный слой, слоя нормализации пакетов (Batch Normalization), функция активации ReLU,
- Блок №2: сверточный слой, слоя нормализации пакетов (Batch Normalization), функция активации ReLU,
- слой выброса (Dropout),
- выравнивающий слой (Flatten),

- полносвязный слой (Dense) с функцией активации softmax.

Итоговая точность распознавания эмоций на голосовых фонограммах с применением сверточных нейронных сетей при помощи анализа мел-спектрограмм составила 80,2%.

1.4. Метод эмпирических мод (EMD)

В этом разделе рассмотрен алгоритм распознавания эмоций на аудиозаписи с применением метода эмпирических мод (EMD) [151]. Сочетание данного подхода с энергетическим оператором Тигера-Кайзера (TKEO) дает информативную частотно-временную характеристику нестационарных сигналов [82]. Такая совокупность методов дает возможность анализировать локальные особенности голосовых фонограмм.

Звуковой сигнал раскладывается на колебательные компоненты, называемые эмпирическими модами (IMF). Энергетический оператор Тигера-Кайзера (TKEO) используется для оценки изменяющейся во времени огибающей амплитуды и мгновенной частоты сигнала, который может быть получен при помощи амплитудно-частотной модуляции. Было выбрано подмножество эмпирических мод (IMF), которое использовалось для извлечения признаков из речевого сигнала для распознавания различных эмоций.

Далее извлекаются спектральные (MS) и частотные (MFF) характеристики модуляции, которые получаются при помощи кепстральных параметров. Сочетание всех функций позволяет достичь высокой эффективности в распознавании эмоций. Для задачи классификации был использован метод опорных векторов (SVM).

В работе описаны функции расчета признаков спектральной (MS) и частотной (MFF) модуляции, основанные на демодуляции (AM-FM) и отслеживании формант. Они объединяются с кепстральными характеристиками: энергетические кепстральные коэффициенты (ECC), частотно-взвешенные энергетические

кепстральные коэффициенты (EFCC) и мел-кепстральные коэффициенты реконструированного сигнала (SMFCC).

Эффективность функций улучшается при помощи нормализации голосовых фонограмм. Машина опорных векторов (SVM) используется для классификации эмоций. Такой выбор был сделан по результатам работы [81].

В итоге сигнал был обработан следующим образом:

1) Декомпозиция сигнала (препроцессинг): разложение на эмпирические моды (EMD) и применение к ним нелинейного энергетического оператора Тигера-Кайзера (TKEO).

2) Вычисление признаков частотной и спектральной модуляции, энергетических кепстральных коэффициентов (ECC), частотно-взвешенных энергетических кепстральных коэффициентов (EFCC) и мел-кепстральных коэффициентов реконструированного сигнала (SMFCC).

3) Классификация эмоций при помощи метода опорных векторов (SVM)

1.4.1. Модуляция сигнала (AM-FM)

Для анализа голоса используется метод амплитудно-частотных модуляции, которая представляет речевой сигнал как сумму формантных резонансных сигналов. [105] описывает этот речевой резонанс $r_i(t)$ следующим образом:

$$r_i(t) = Re\left(a_i(t)e^{j\phi_i(t)}\right) \tag{1.2}$$

А речевой сигнал может быть представлен так:

$$x(t) = \sum_{i=1}^{N} r_i(t) + res(t) \tag{1.3}$$

где $res(t)$ — последние несколько компонентов, которые представляют собой низкочастотные колебания, исключаемые из речевого сигнала [130], Re - действительная часть, $\phi_i(t)$ - фаза, $a_i(t)$ – мгновенная амплитуда, $f_i(t)$ – мгно-

венная частота i-й эмпирической моды:

$$\phi_i(t) = 2\pi \int f_i(t)dt \tag{1.4}$$

В [121] речевой резонансный сигнал $r(t)$ извлекается из речевого сигнала $x(t)$ с применением фильтра Габора. Звуковой сигнал при помощи эмпирического разложения представляет из себя сумму мод $r_i(t)$. Затем применяется энергетический оператор Тайгера Кайзера (TKEO) для демодуляции резонансных сигналов $r_i(t)$ в амплитудную $a_i(t)$ и частотную $f_i(t)$ огибающие. Далее каждый элемент будет рассмотрен по отдельности.

1.4.2. Разложение на эмпирические моды

Рассматривается метод эмпирических мод, который может разложить любой нестационарный сигнал на набор внутренних колебаний, которые представляют собой однокомпонентные сигналы амплитудно-частотной модуляции. Их извлечение нелинейно, но реконструкция сигнала линейна. Фактически добавление всех мод позволяет линейно восстанавливать исходный сигнал без потери и без искажения исходной информации.

Функция называется функцией эмпирической моды, если она удовлетворяет следующим свойствам:

• количество экстремумов (максимумов+минимумов) в сигнале должно быть равно или отличаться не более чем на единицу от количества пересечений нуля;

• Среднее значение огибающих, определяемых локальными максимумами и минимумами, всегда должно быть равно нулю.

Процедура декомпозиции для извлечения функций эмпирических мод называется процессом просеивания и описывается следующим образом:

Вход: речевой сигнал $x(t)$

Выход: набор эмпирических мод (IMF).

Шаг 1. Вычисление всех локальных экстремумов в сигнале $x(t)$: локальные максимумы и минимумы;

Шаг 2. Построение верхней огибающей $E_u(t)$ и нижнюю огибающую $E_l(t)$, соединив локальные максимумы и минимумы кубическим сплайном в заданном сигнале $x(t)$;

Шаг 3. Вычисление среднего значения огибающей: $m(t) = \frac{E_u(t)+E_l(t)}{2}$;

Шаг 4. Получение новой последовательности $r(t)$, из которой удалена низкая частота: $r(t) = x(t) - m(t)$;

Шаг 5. Повторение шагов 1–4 до тех пор, пока $r(t)$ не станет удовлетворять условиям эмпирической моды (IMF);

Шаг 6. Вычитание эмпирической моды(IMF) $r(t)$ из исходного сигнала $res(t) = x(t) - r(t)$;

Шаг 7. Повторение шагов 1–6 до тех пор, пока в остаточном сигнале $res(t)$ не останется ни одоной эмпирической моды (IMF).

Процесс завершается, когда остаточный элемент $res(t)$ является константой либо монотонной функцией.

Таким образом можно получить N эмпирических мод (EMD) $r_1(t), r_2(t)$, ..., $r_N(t)$ и остаточный сигнал $res_N(t)$. Следовательно, исходная последовательность данных $x(t)$ может быть представлена в следующем виде:

$$x(t) = res_N(t) + \sum_{i=1}^{N} r_i(t) \tag{1.5}$$

В этом методе каждый входной сигнал разлагается на конечное число эмпирических мод, каждую из которых можно проанализировать отдельно, чтобы получить признаки для классификации эмоций. В [127] авторы ограничиваются первыми пятью коэффициентами при условии, что они дают достаточную информацию об энергии и высоте тона. Как правило, после определенного порядкового номера эмпирические моды не являются информативными.

1.4.3. Энергетический оператор Тигера-Кайзера (TKEO)

Эмпирические моды сами по себе несут не очень много информации, но применение к ним оператора энергии Тигера-Кайзера (TKEO) позволяет оценить изменяющуюся во времени огибающую амплитуды и мгновенную частоту, что имеет определенный физический смысл.

Оператор энергии Тигера-Кайзера — это нелинейный оператор, вычисляющий энергию однокомпонентных сигналов как произведение квадрата амплитуды и частоты сигнала. Мгновенные характеристики этих сигналов затем могут быть получены при помощи алгоритма дискретного разделения энергии (DESA-2) [105].

Такая операция улучшает оценку мгновенных характеристик данных вибрации по сравнению с другими широко используемыми методами, например, с преобразованием Гильберта. Метод, основанный на алгоритме декомпозиции эмпирических мод с применением оператора Тигера-Кайзера (TKEO) называется преобразованием Тигера-Хуанга (THT) [71].

$$\Psi[r_i(t)] = [\dot{r}_i(t)]^2 - r_i(t)\ddot{r}_i(t) \tag{1.6}$$

где $\dot{r}_i(t)$ и $\ddot{r}_i(t)$ — первая и вторая производные по времени от $r_i(t)$ соответственно. Для дискретного сигнала времени $r_i(n)$ производная [107] будет выглядеть так:

$$\Psi[r_i(n)] = r_i^2(n) - r_i(n+1)r_i(n-1) \tag{1.7}$$

где n — дискретный индекс.

Следующие уравнения точно описывают мгновенную частоту $f(n)$ и мгновенную амплитуду $a(n)$ в любой момент времени эмпирической моды $r_i(n)$ [105]:

$$f(n) = \frac{1}{2}\arccos\left(1 - \frac{\Psi[x(n+1) - x(n-1)]}{2\Psi[x(n)]}\right) \tag{1.8}$$

$$|a(n)| = \frac{2\Psi[x(n)]}{\sqrt{\Psi[x(n+1) - x(n-1)]}} \tag{1.9}$$

В [120] обратили внимание, что приближенное значение оператора энергии содержит в себе высокочастотную составляющую ошибки. Поэтому было предложено устранить ее при помощи биномиального фильтра нижних частот.

1.4.4. Извлечение признаков

Далее будет описан процесс извлечения различных признаков из голосовых фонограмм.

Мел-кепстральные коэффициенты реконструированного сигнала (SMFCC)

Кепстральные признаки учитывают физические особенности слуховой системы человека и вычисляются при помощи спектра сигнала.

Мел-кепстральные коэффициенты (MFCC) широко используется для распознавания речевых эмоций [154]. Можно предположить, что речевой сигнал является кратковременным стационарным процессом. В таком случае мел-кепстральные коэффеценты содержали бы в себе основные информативные признаки. Но на самом деле речевой сигнал имеет сложные и случайные изменения, и наличие тренда сигнала создает большую погрешность в спектральном анализе мощности в частотной области или поиске корреляций во временной области.

Поэтому удаление сигнального тренда повысит информативность признаков. Метод поиска мел-кепстральных коэффициентов после удаления сигнального тренда $T(n)$ подробно описан в [96]. Такие численные параметры дают более точное описание распределения энергии в частотной области. Вычисление мел-кепстральных коэффициентов реконструированного сигнала (SMFCC) схематически показано на Рис. 1.4.

Тренд сигнала может быть вычислен как сумма эмпирических мод, удо-

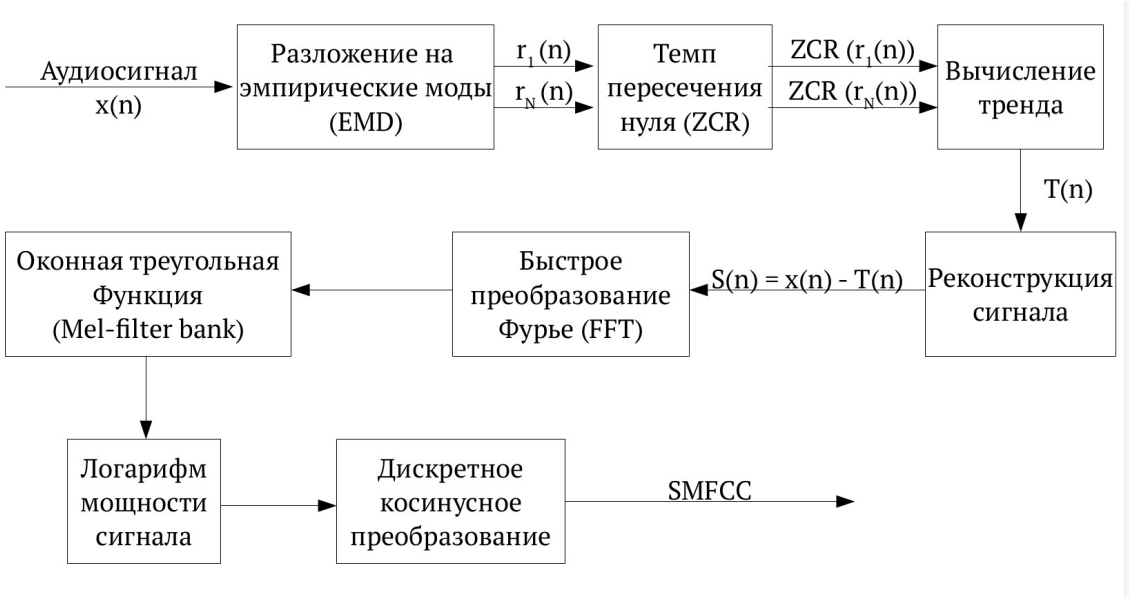

Рис. 1.4. Схема вычисления мел-кепстральных коэффициентов реконструированного сигнала (SMFCC).

влетворяющих следующему условию:

$$\frac{ZCR_{r_i}}{ZCR_{r_1}} < 0.01 (i = 2, \ldots, n) \tag{1.10}$$

где ZCR - темп пересечения сигналом нуля (показатель, характеризующий частоту изменения сигнала с отрицательного на положительный и наоборот).

$$ZCR = \frac{1}{T-1} \sum_{t=1}^{T-1} 1_{\mathbb{R}<0}(r_t r_{t-1}) \tag{1.11}$$

где $r(t)$ - сигнал длины T, $1_{\mathbb{R}<0}$ - индикаторная функция.

Впоследствии окончательный сигнал $S(n)$ получается путем вычитания тренда $T(n)$ из исходных данных $x(n)$ [97].

$$T(n) = \sum_i r_i(n) \tag{1.12}$$

$$S(n) = x(n) - T(n) \tag{1.13}$$

Мел-кепстральные коэффициенты реконструированного сигнала (SMFCC) извлекаются из полученных восстановленных данных $S(n)$ с помощью алгорит-

ма быстрого преобразования Фурье (FFT) и дискретного косинусного преобразования (DCT).

Энергетические кепстральные коэффициенты (ECC) и частотно-взвешенные энергетические кепстральные коэффициенты (EFCC).

В [130] авторы подтверждают, что распределение спектральной энергии меняется при разных эмоциях. Это означает, что энергетические частотные диапазоны разных эмоциональных состояний могут пересекаться. Поэтому гораздо важнее проанализировать распределение энергии в гильбертовом спектре (во временной области).

Считается, что мгновенная энергия каждой эмпирической моды пропорциональна амплитуде и не зависит от мгновенной частоты. Но согласно физическому смыслу в энергетическую оболочку $a(t)$ входит не только мгновенная энергия, но и мгновенная частота $f(t)$. На основе приведенных выше соображений исследователи разработали мгновенную взвешенную по частоте энергию (EFCC) для улучшения характеристик ECC, представленных выше.

Стандартная реализация расчета кепстральных коэффициентов энергии (ECC) и частотно-взвешенных кепстральных коэффициентов энергии (EFCC) [97] показана на рисунке 1.5.

Первым этапом обработки является разложение речевого сигнала на эмпирические моды. Их мгновенная амплитуда ($a(i,n)$) и мгновенная частота ($f(i,n)$) оцениваются с помощью энергетического оператора Тигера Кайзера на втором этапе. На третьем шаге обработки регистрируется мгновенная амплитуда и частота коротких кадров с перекрытием (длительностью 250 мс, наложением 64 мс), что дает $a_k(i,n)$ и $f_k(i,n)$. Четвертым этапом обработки является вычисление маргинального гильбертового спектра. Он предлагает меру общей амплитуды каждой частоты. Поэтому спектр разбивается на 12 разных полос частот и вычисляется мощность каждой. Затем рассчитывается натуральный логарифм энергии поддиапазона и дополняется дискретным косинусным преобразованием.

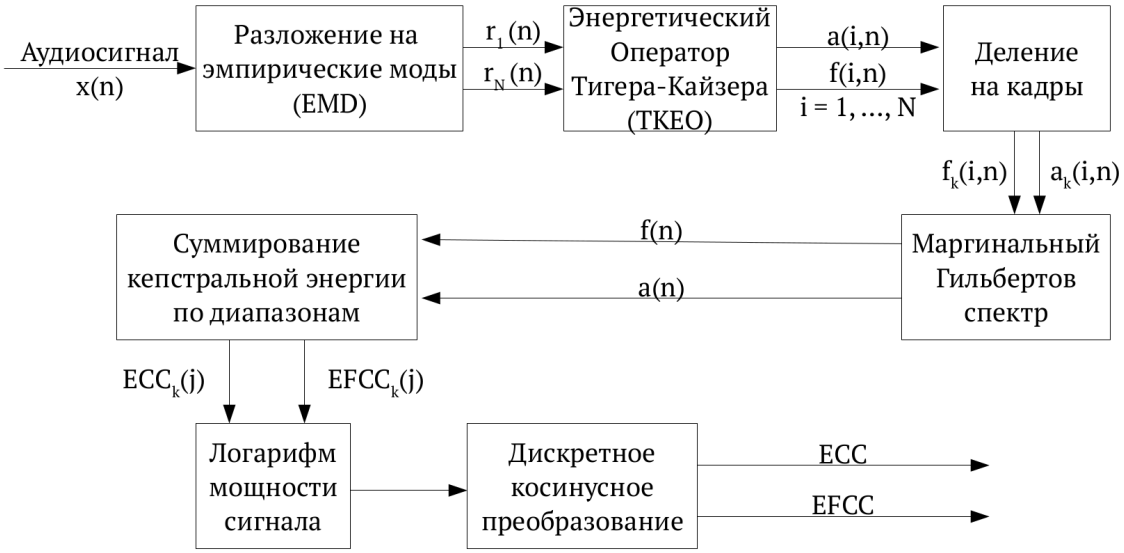

Рис. 1.5. Схема вычисления кепстральных коэффициентов энергии (ECC) и частотно-взвешенных кепстральных коэффициентов энергии (EFCC).

Первые 12 коэффициентов DCT дают значения ECC и EFCC, используемые в процессе классификации.

Рассчитать маргинальный гильбертовый спектр можно следующим образом:

$$h_j(f) = \sum_{n=1}^{L_f} H(f,n)\mathbb{1}_{B_j}(f) \tag{1.14}$$

где L_f - длина кадра в отсчетах, а $H(f,n)$ — Гильбертов спектр. Он определяется как мгновенная огибающая энергии в частотно-временном пространстве, которая представляет собой квадрат величины огибающей амплитуды.

$$H(f,n) = \sum_{i=1}^{N} a(i,n)^2 \mathbb{1}_{\{f(i,n)\}}(f) \tag{1.15}$$

где i - номер эмпирической моды, а B_j - соответствующий поддиапазон.

$$B_j = [f_c(j-1), f_c(j+1)] \tag{1.16}$$

где f_c центр частот.

Индикаторная функция подмножества множества Ω определяется как:

$$\mathbb{1}_{\Omega}(x) = \begin{cases} 1, & \text{если } x \in \Omega; \\ 0, & \text{если } x \notin \Omega. \end{cases} \tag{1.17}$$

Расчет энергетических кепстральных коэффициентов (ECC) для непрерывной функции подробно описана в [130]:

$$ECC(B_j, S_i) = \int\limits_{f \in B_j} h_j(f)df, t \in S_i, j = 1, \ldots, 12 \tag{1.18}$$

где B_j обозначает поддиапазон, S_i - речевой кадр.

В данной работе эти характеристики вычисляются с использованием дискретного речевого сигнала следующим образом:

$$ECC_k(j) = \sum_{i=1}^{N} \frac{1}{L_F} \sum_{n=1}^{L_F} a_k^2(i,n) \mathbb{1}_{B_j} \left(f_k(i,n) \right), j = 1, \ldots, 12 \tag{1.19}$$

Частотно-взвешенные энергетические кепстральные коэффициенты (EFCC)

Непрерывная функция вычисления данных компонентов описана в [130]:

$$EFCC(B_j, S_i) = \int\limits_{f \in B_j} f(t)h_j(f)df, t \in S_i, j = 1, \ldots, 12 \tag{1.20}$$

где B_j обозначает поддиапазон, S_i - речевой кадр.

Для дискретного случая EFCC может быть вычислен следующим образом:

$$EFCC_k(j) = \sum_{i=1}^{N} \frac{1}{L_F} \sum_{n=1}^{L_F} f_k(i,n) a_k^2(i,n) \mathbb{1}_{B_j}(f_k(i,n)), j = 1, \ldots, 12 \tag{1.21}$$

где B_j - поддиапазон, i - эмпирическая мода, k - речевой кадр.

Были извлечены первые 12 коэффициентов ECC и EFCC для голосовых фонограмм. Для каждого из них были вычислены среднее значение, дисперсия, а также коэффициенты вариации, эксцесса и асимметрии. Каждый вектор признаков ECC и EFCC состоит из 60 точек.

Спектральные характеристики модуляции (MSF)

Спектральные характеристики модуляции (MSF), представленные в [150] извлекаются для решения проблемы краткосрочных спектральных характеристик (MFCC) и для моделирования природы слухового восприятия человека. Метод основан на имитации спектрально-временной обработки, выполняемой в слуховой системе человека, и учитывает как обычную акустическую частоту, так и частоту модуляции. Эти функции основаны на разложении речевого сигнала слуховыми фильтрами и вычислении гильбертовой огибающей каждой частотной полосы.

В данной работе эти признаки извлекаются с помощью метода модуляции (AM-FM). Этапы выделения спектральных характеристик модуляции (MS) изображены на рисунке 1.6 .

После использования энергетического оператора Тигера-Кайзера (TKEO) к эмпирическим модам (IMF), фильтры модуляции дополнительно применяются к мгновенной амплитуде для выполнения частотного анализа.

Спектральный состав сигналов модуляции называется спектром модуляции, а предлагаемые признаки называются спектральными признаками модуляции (MS).

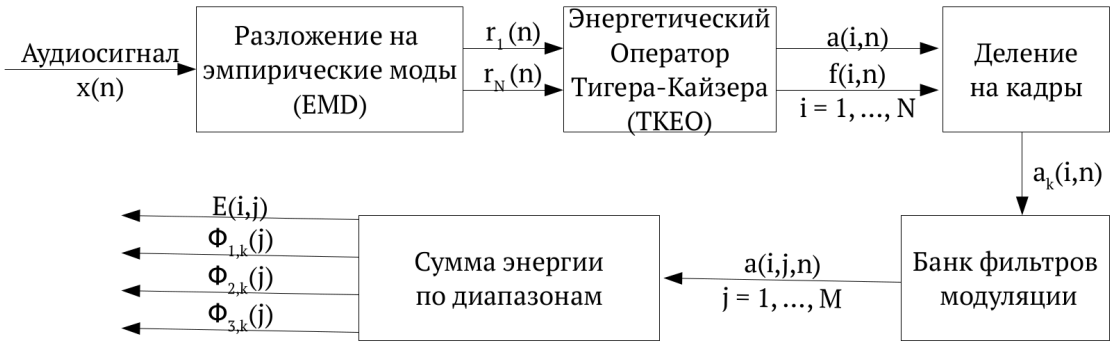

Рис. 1.6. Схема вычисления спектральных признаков модуляции (MS).

Энергия всех кадров в каждой спектральной полосе, дает характеристику $E(i, j)$. Энергия в каждой спектральной полосе определяется следующим

образом:

$$E(i,j) = \sum_{k=1}^{N_f} E_k(i,j) \tag{1.22}$$

Где $E_k(i,j)$ — энергия по каналам, N_f — количество кадров для $1 \leqslant j \leqslant 8$. Для каждого кадра k значение $E(i,j)$ нормализуется к единице энергии перед дальнейшими вычислениями:

$$\sum_{i,j} E_k(i,j) = 1 \tag{1.23}$$

Затем для каждого кадра рассчитываются три спектральные меры Φ_1, Φ_2, Φ_3 [150]. Первая рассчитывается как среднее значение отсчетов энергии, принадлежащих j-й полосе модуляции ($1 \leqslant j \leqslant 8$):

$$\Phi_{1,k}(j) = \frac{\sum_{i=1}^{N} E_k(i,j)}{N} \tag{1.24}$$

Для кадра k $\Phi_{2,k}(j)$ — это спектральная плоскостность, которая определяется как отношение среднего геометрического Φ_1 к среднему арифметическому. Таким образом, Φ_2 определяется следующим образом:

$$\Phi_{2,k}(j) = \frac{\sqrt[N]{E_k(1,j)E_k(2,j)\dots E_k(N,j)}}{\Phi_{1,k}(j)} \tag{1.25}$$

В текущем исследовании Ф2 вычисляется в логарифмическом масштабе следующим образом:

$$\log \Phi_{2,k}(j) = \frac{1}{N}\sum_{i=1}^{N} \log E_k(i,j) - \log \Phi_{1,k}(j) \tag{1.26}$$

Последняя используемая мера Φ_3 - это спектральный центроид, который определяет центр масс в каждой полосе модуляции. Для j-й полосы модуляции Φ_3 определяется следующим образом:

$$\Phi_{3,k}(j) = \frac{\sum_{i=1}^{N} f(i)E_k(i,j)}{\sum_{i=1}^{N} E_k(i,j)} \tag{1.27}$$

где $f(i)$ - индекс i-го фильтра критической полосы, для упрощения можно считать $f(i) = i$.

Различные статистические данные, среднее значение, дисперсия, коэффициент вариации, эксцесс и асимметрия извлекаются из энергии в каждой спектральной полосе. Наряду с этими статистическими данными оцениваются среднее значение и дисперсия спектральной энергии, спектральной плоскостности и спектрального центроида, которые используются в качестве признаков.

Вычисления говорят о том, что эмоции грусти и спокойствия обладают значительно большей низкочастотной акустической энергией, чем эмоция гнева, [150]. Для эмоции гнева дисперсия еще больше. Однако менее выразительные эмоции, такие как грусть, демонстрируют более выраженные спектральные формы низкочастотной модуляции, что предполагает более низкую скорость речи.

Характеристики частотной модуляции (MFF)

Исследования восприятия речи показали, что наиболее важная перцептивная информация находится на низких частотах модуляции [28]. Их анализ может быть полезен в распознавании эмоционального состояния на голосовых фонограммах.

Метод модуляции (AM-FM) используется для анализа сигнала, далее будет исследовано частотное распределение энергии речевого сигнала в полосе частот. Производится разложение на эмпирические моды, а мгновенная огибающая амплитуды и частотная функция каждой из них оцениваются с использованием TKEO. Затем они используются для получения краткосрочных оценок средней мгновенной частоты и полосы пропускания по каждому кадру. Средняя амплитуда, взвешенная по мгновенной частоте F_w, и средняя взвешенная по амплитуде мгновенная ширина полосы частот B_w описаны для непрерывного

временного сигнала в [130]:

$$F_w = \frac{\int_{t_o}^{t_0+T} f(t)a^2(t)dt}{\int_{t_0}^{t_0+T} a^2(t)dt} \tag{1.28}$$

$$B_w^2 = \frac{\int_{t_0}^{t_0+T} [\{\dot{a}/2\pi\}^2 + \{f(t) - F_w\}^2 a^2(t)]dt}{\int_{t_0}^{t_0+T} a^2(t)dt} \tag{1.29}$$

где t_0 и T представляют собой начало и продолжительность анализируемого кадра, $f(t)$ и $a(t)$ - мгновенная огибающая частоты и амплитуды каждого сигнала (AM-FM), а $\dot{a}(n)$ вычисляется для дискретного случая следующим образом:

$$\dot{a}(t) = a(t+1) - a(t) \tag{1.30}$$

При помощи огибающей амплитуды и мгновенной частоты вычисляются следующие характеристики сигнала: средняя мгновенная частота ($\bar{F}$), средняя мгновенная ширина полосы (B), взвешенная по средней амплитуде мгновенная частота ($\bar{F}_w$) и средняя амплитуда взвешенной мгновенной ширины полосы пропускания (B_w). Их можно посчитать для дискретного случая следующим образом:

$$\bar{F}_k(i) = \frac{1}{L_F} \sum_{n=1}^{L_F} f_i(k,n) \tag{1.31}$$

$$B_k(i) = \sqrt{\frac{1}{L_F} \sum_{n=1}^{L_F} (f_k(i,n) - \bar{F}_k(i))} \tag{1.32}$$

$$\bar{F}_k^w(i) = \frac{\sum_{n=1}^{L_F} f_k(i,n) a_k^2(i,n)}{\sum_{n=1}^{L_F} a_k^2(i,n)} \tag{1.33}$$

$$B_k^w(i) = \sqrt{\frac{\sum_{n=1}^{L_F} \{\dot{a}_k(i,n)/2\pi\}^2 + \{f_k(i,n) - \bar{F}_k^w(i)\}^2 a_k^2(i,n)}{\sum_{n=1}^{L_F} a_k^2(i,n)}} \tag{1.34}$$

где i представляет номер эмпирической моды, L_F представляет количество выборок на кадр, f_k и a_k представляют соответственно мгновенную частоту и амплитудную частоту в конкретном кадре речи k. В этой работе среднее, максимальное и минимальное из этих двух значений (B, B_w) и среднее значение $(\bar{F}, \bar{F}_w)$ были вычислены для каждого признака и использовались для классификации.

Для каждой эмпирической моды краткосрочная оценка частоты $\bar{F}_k(i)$ рассчитывается для каждого кадра.

Продолжительность анализируемого кадра голосовой фонограммы составляет 25 мс со сдвигом кадра 10 мс.

Эмоция гнева обычно смещается в сторону высоких частот (между 600 и 1300 Гц). а грусть - к низким частотам.

Итоговая точность распознавания эмоций на голосовых фонограммах с применением метода эмпирических мод составила 74.6%.

1.5. Метод вейвлет-анализа

Коэффициенты вейвлет-пакета (WPC)

Вейвлеты можно определить как в [103]. Вейвлет-преобразование сигнала $f \in L^2(R)$ во время u и в масштабе s вычисляется путем корреляции f с атомом вейвлета по формуле :

$$Wf(u,s) =< f, \psi_{u,s} >= \int_{-\alpha}^{+\alpha} f(t) \frac{1}{\sqrt{s}} \psi^* \left(\frac{t-u}{s} \right) dt \tag{1.35}$$

Вейвлет-преобразование также может быть описано при помощи свертки:

$$Wf(u,s) = \int_{-\alpha}^{+\alpha} f(t) \frac{1}{\sqrt{s}} \psi^* \left(\frac{t-u}{s} \right) dt = f * \overline{\psi_s}(u) \tag{1.36}$$

Пакетное вейвлет-преобразование (WPT) [103] затрагивает и низкие и высокие частоты. Это может быть реализовано с помощью многомерного банка

фильтров [131]. Он рассчитывается при помощи наложения фильтров во временной области на субсигнал, полученный из частотных составляющих в каждом субдиапазоне [104]. Многомерный банк фильтров вычисляется при помощи разложения вейвлет-пакета.

$$\psi^i_{j,k}(t) = 2^{\frac{j}{2}}\psi^j(2^j t - k), i = 1, 2, 3, \ldots \tag{1.37}$$

где i — параметры модуляции, j — параметры масштаба, k — параметры трансляции, ψ^j — вейвлет-функция, которую можно определить рекурсивно:

$$\psi^{2j}(t) = \sqrt{2}\sum_{-\alpha}^{+\alpha} h(k)\psi^j(2t - k) \tag{1.38}$$

$$\psi^{2j+1}(t) = \sqrt{2}\sum_{-\alpha}^{+\alpha} g(k)\psi^j(2t - k) \tag{1.39}$$

j-уровневое разложение $f(t)$ можно определить следующим образом:

$$f(t) = \sum_{i=1}^{2j} f^i_j(t) \tag{1.40}$$

Коэффициенты вейвлет-пакета задаются в таком виде:

$$f^i_j(t) = \sum_{-\alpha}^{+\alpha} c^i_{j,k}(t)\psi^i_{j,k}(t) \tag{1.41}$$

$$c^i_{j,k}(t) = \sum_{-\alpha}^{+\alpha} f(t)\psi^i_{j,k}(t)dt \tag{1.42}$$

Таким образом, сигнал может быть описан с помощью комбинации коэффициентов пакета вейвлетов, а также их разности первого и второго порядка. Исследована четырехуровневая и пятиуровневая вейвлет-пакетная декомпозиция.

В текущей работе был выбран метод опорных векторов (SVM) в качестве классификатора.

Таблица 1.1. Таблица итоговой точности распознавания эмоций на голосовых фонограммах с применением различных методов

Метод	Точность распознавания
Метод эмпирических мод	74,6%
Вейвлет-анализ	72,7 %
Предложенная архитектура	80,2 %

По результатам вычислений можно сделать вывод о том, что коэффициенты вейвлета-пакета эффективны для распознавания эмоций радости, грусти, злости и отвращения, но неэффективны для распознавания нейтральных эмоций.

Итоговая точность распознавания эмоций на голосовых фонограммах при помощи вейвлет-анализа составила 72,7%.

1.6. Выводы к первой главе

Результаты первой главы опубликованы в работах [9, 47].

Предложена и программно реализована архитектура нейронной сети для решения задачи определения эмоции на голосовой фонограмме с высокой точностью 1.1.

Благодаря поиску эффективной архитектуры для распознавания эмоций на голосовых фонограммах при помощи анализа мел-спектрограмм сверточными нейронными сетями удалось достичь высоких показателей точности.

В дальнейших исследованиях планируется оценить возможность повышения точности классификации при обогащении аудиосигналов информацией об уровне стресса исследуемых лиц.

Итоговая точность распознавания эмоций на голосовых фонограммах представлена в таблице 1.1:

Глава 2

Методы распознавания эмоций на видеозаписи

2.1. Краткий обзор существующих подходов

В данной главе рассматриваются подходы к распознаванию эмоций человека по визуальным признакам лица. Применяется глубокое обучение многослойных нейронных сетей.

Существует огромное разнообразие алгоритмов, способных распознавать эмоции человека по мимике лица [1, 4, 11]. Однако качество этих систем уменьшается из-за следующих обстоятельств:

маленькая выборка для обучения, расхождение в пропорциях лица, наигранность эмоций, освещенность во время съемки, окклюзия, различный угол поворота головы, внутриклассовое различие и межклассовое сходство, этническая принадлежность, пол, возраст.

Использование многослойных нейронных сетей направлено на повышение точности определения эмоций на изображениях.

Наиболее точно на сегодняшний день эмоции человека были описаны Полом Экманом в работе [49], где каждая эмоция была представлена при помощи кодирования лицевых движений, но данный подход сложно автоматизировать [91, 90]. Это связано с тем, что он содержит 46 основных категорий и более 50 дополнительных, а эмоции являются комбинациями таких групп. Поэтому для подготовки такие датасеты (пронумерованный набор изображений, фонограмм или видеозаписей с указанием исследуемых признаков каждого элемента) обрабатываются психологами вручную с учетом указанных обстоятельств [8].

Особый интерес представляет датасет Aff-Wild [87, 88, 89], который состоит из фрагментов видеороликов платформы YouTube, они не являются предзаписанными в видеостудии. Это значит, что условия записи материалов максималь-

но приближены к действительности.

Существует два принципиально разных подхода в распознавании эмоций:

с предварительным алгоритмическим извлечением визуальных признаков и последующей машинной классификацией [23], с использованием глубоких нейронных сетей без предварительного извлечения признаков [153].

Визуальные признаки могут быть извлечены при помощи выявления:

геометрических объектов лица (брови, нос, рот, глаза и др.) [138, 68] методами дескриптора line edge map, сравнения направленности градиентов [76, 60], метода активной формы ASM [73], курвлет-преобразования [35], использования структурных моделей [19] и др.;

текстурных особенностей методами фильтра Габора, дискретным вейвлет-преобразованием [113] и др.;

глобальных и локальных объектов методами главных компонент [140], оптического потока [159], морфологическими преобразованиями [6] и др.

Но при условии наличия достаточного датасета наиболее высокой точности классификации удается достичь именно при помощи автоматического выявления признаков и классификацией глубокими нейронными сетями [6, 18, 136, 7].

Можно выделить три основных направления исследований в области распознавания эмоций по мимике:

1. Объединение групп мимических мышц в единицу действия (ЕД), в таком случае эмоции можно выразить при помощи движений нескольких ЕД. Так из входного изображения можно определить соответствующее эмоциональное состоянии при помощи декодирования ЕД. Но мышечные движения в основном небольшие, поэтом сложно достичь высокой точности в их обнаружении [156]. Это сильно ограничивает развитие такого подхода.

2. Выделение признаков соответствующих эмоций на лице. Данный подход включает в себя три шага. На первом этапе необходимо найти лицо на изображении с помощью выявления ориентиров в области лица. На втором этапе происходит извлечение признаков из области лица с помощью гистограммы

ориентированного градиента (HOG), локальных бинарных паттернов (LBP) и вейвлет-методов Габора. На третьем этапе производится классификация при помощи полученных параметров.

3. Использование методов глубокого обучения, в которых в отличие от традиционных подходов, признаки анализируемого изображения создаются непосредственно в сверточных слоях [142, 77, 143].

Хотя методы с применением глубокого обучения работают зачастую лучше обычных, но у них есть серьезные ограничения, связанные со временем обработки и потреблением памяти. Иногда в прикладных задачах приходится работать на слабом оборудовании, поэтому необходимо изучить альтернативные методы для распознавания по мимике семи основных эмоций (спокнойствие, радость, удивление, печаль, страх, гнев и отвращение).

Метод локальных бинарных паттернов (LBP) достаточно эффективно справляется с задачей для описания текстур. Изменение выражения лица создает различные текстуры отдельными группами мышц, что делает возможным применение данного метода для выявления эмоциональных состояний. Объединение данного подхода с ORB (Oriented FAST and Rotated BRIEF) позволяет обеспечить еще более высокую скорость вычислений [55].

Совместное применение аудио и видео систем может увеличить точность распознавания эмоций [38], но требования к пользователю сильно возрастают для корректного размещения записывающего оборудования.

2.2. Набор данных для классификации

В текущей главе в основном использовались наборы данных RAVDESS (описан ранее) и Aff-Wild [88].

Набор данных Aff-Wild состоит из 298 видеороликов, общей продолжительностью более 30 часов. Основная задача, которую ставили перед собой ее авторы - собрать спонтанное поведение лица в произвольных условиях записи.

Все видеоролики были собраны на веб-сайте обмена видео Youtube по ключевому слову «реакция». Такие фрагменты показывают, как люди реагируют на различные стимулы (раздражители). Это может быть последствием неожидинного поворота в сюжете фильма, обратная связь на необычную еду, впечатления после экстримальных событий. В данных видеороликах присутствуют, как положительные, так и отрицательные эмоции.

Каждый ролик аннотирован по шкале валентности (позитив-негатив) и возбуждения (расслабленность - тонус) по методике, описанной в [43]. Значения каждого показателя изменялись от -1 до 1. Для удобства работы метки проставлялись с помощью джойстика. В этом наборе данных представлно 200 субъектов, из них 130 мужчин и 70 женщин. Корректность аннотирования достигалась при помощи инструктажа в устной форме и выдачи инструкции в виде многостраничного документа с объяснением процедуры. Описание включало в себя краткий список хорошо идентифицированных эмоциональных сигналов для возбуждения и валентности. Перед началом комментирования данных, каждый аннотатор просматривал видео целиком, чтобы знать порядок реакций и эмоций в текущем фрагменте.

Выбор набора данных играет важную роль. Благодаря этой работе для исследования предоставлен строго аннотированный и разнообразный материал для обучения нейронных сетей.

Рассмотрим следующую задачу классификации:

X - множество видеозаписей x;

$x_i = \{x_{i1}, \ldots, x_{il}\}$ - видеозапись продолжительностью t, состоящая из l кадров x_{ij}

, где l - длительность видеозаписи в кадрах,

x_{ij} - матрица целых чисел размера $w \times h \times c$, где w - ширина кадра, h - высота кадра, c - цветность кадра ($c = 3$ для цветного, $c = 1$ для черно-белого);

$Y = \{$Спокойствие, Радость, Грусть, Злость, Страх, Удивление, Отвращение$\}$ - множество классов из семи эмоций.

Существует неизвестная целевая зависимость - отображение $y^* : X \to Y$, значения которой известны на объектах набора данных

$X^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, где m - размер набора данных

Требуется построить алгоритм $a : X \to Y$, способный классифицировать произвольный объект $x \in X$

2.3. Метод локальных бинарных паттернов

В данной главе представлена структура алгоритма распознавания выражения лица. Лица обнаруживаются и извлекаются при помощи библиотеки Dlib из-за ее высокой скорости обработки, далее применяется метод локальных бинарных паттернов LBP [55].

Метод локальных бинарных паттернов (LBP) может зафиксировать мелкие текстурные детали при помощи мелкомасштабного дескриптора. Такой подход устойчив к изменениям освещения на видеозаписи и справляется с задачей кодирования мелких визуальных деталей на лице.

Для применения данного метода требуется извлечь лицо из кадра. Для обнаружения лиц была выбрана библиотека Dlib лиц из-за ее высокой скорости обработки. Выбранные методы содержат детектор ориентиров лица с предварительно обученной архитектурой [85, 78], которая выделяет 68 точек, описанные в таблице 2.1.

В этом методе создается дерево регрессии для нахождения этих ориентиров лица непосредственно по интенсивности пикселей без извлечения признаков; таким образом, процесс обнаружения может быть достаточно быстрым, чтобы преодолеть проблемы с точностью и качеством, связанные с анализом в реальном времени.

Обычно эмоции в основном передаются через глаза, нос, брови и некоторые области лица, поэтому другие части лица, такие как уши и лоб, могут быть исключены из дальнейшего анализа. Этот алгоритм обнаружения лица

Таблица 2.1. Таблица соответствия выделенных библиотекой dlib точек и областей лица

Область лица	Номера точек
челюсть	1-17
левая бровь	18-22
правая бровь	23-27
левый глаз	37-42
правый глаз	43-48
нос	28-36
внешняя граница губы	49-60
внутренняя граница губы	61-68

подходит для получения точных областей лица, а точки 1–27 используются для извлечения области лица из исходных изображений.

Локальные бинарные шаблоны (LBP). LBP могут эффективно описывать текстурную информацию изображения [114, 115, 116]. Оператор LBP определяется для окрестностей 3x3. Он принимает каждый пиксель за центральный и оценивает 8 значений вокруг выбранного на основе заданного порога. Результирующий фрагмент изображения с двоичным значением формирует локальный дескриптор изображения [129].

Оператор LBP принимает следующую форму:

$$LBP(x_c, y_c) = \sum_{n=0}^{7} 2^n s(i_n - i_c) \tag{2.1}$$

где c - центральный пиксель, i_c и i_n - значения серого цвета у c и у 8-ми его соседей с порядковым номером n,

$$s(u) = \begin{cases} 1, & \text{если } u \geqslant 0; \\ 0, & \text{если } u < 0. \end{cases} \tag{2.2}$$

Если оператор LBP содержит не более одного перехода 0-1 и один переход 1-0 в двоичном коде, то существует универсальный шаблон. Равномерный шаблон содержит примитивную структурную информацию для краев и углов.

Длина вектора признаков для одной ячейки составляет 256. Размер области лица составляет 130 × 130, а лицо LBP имеет размер 128 × 128.

Итоговая точность распознавания эмоций на видеозаписях с применением метода локальных бинарных паттернов (LBP) составила 58,7%.

2.4. Метод сверточных нейронных сетей

Сначала осуществляется предварительная обработка. Этот этап предподготовки информации позволяет минимизировать перечисленные ранее причины ошибок. В первую очередь выполняется обнаружение области лица, обрезка изображения и масштабирование для подведения под нужный размер. Затем производится изменение контрастности для уменьшения ошибок, возникающих из-за большой разницы освещения, и выделение общих визуальных особенностей лица. Под этими особенностями понимаются компоненты лица, такие, как губы, нос, рот, брови и т. д., а также немаловажный признак – текстура кожи. Этот этап нужен для более точной постановки задачи перед алгоритмом классификации.

Среди известных подходов обнаружения лиц на изображении существуют две категории.

1. Методы, построенные на конкретном наборе составленных правил, основанных на выделении независимых свойств изображений лиц. Здесь имеет место два этапа построения:

установка явных признаков, характерных для изображений лица, обработка найденных признаков.

2. Методы, в которых задействован вычислительный вектор признаков, разделяющий изображение на два типа: лицо и не лицо. Выбор метода зави-

сит от установленных ограничений и условий в процессе выполнения задачи. Выделяются следующие возможные ограничения:

пространственные характеристики положения лиц;

наличие или отсутствие ограничений на возможные искусственные помехи на лице;

количество лиц на изображении;

условия освещенности объектов;

цветность изображения;

приоритет в минимизации ложных обнаружений или в количестве обнаруженных лиц;

масштаб лиц и разрешение изображения.

Для анализа видеопотока была выбрана библиотека для проектирования нейросетей - Theano. Это библиотека, которая используется для разработки систем машинного обучения как сама по себе, так и в качестве вычислительного бекэнда для более высокоуровневых библиотек, в данном случае - Lasagne. Также применяется Nolearn как дополнительная и вспомогательная библиотека машинного обучения.

Вычисления проводились для различных наборов данных. Например, в RAVDESS [100] каждый видеоматериал разбивался на части по 0.5 с, и при обработке они перемешивались. Для датасета CK+ [101] из каждой директории брали 3-5 последних кадров, они отбирались многократно и случайным образом, а затем была проведена аугментация в виде поворота кадра в случайный угол на 10°, для того чтобы достичь баланса в классификации.

Для датасета Aff-Wild из видео выбирались 4-16 кадров случайного отрезка, а сами видеоматериалы определялись также многократно и случайным образом, после проводилась аугментация для достижения баланса. Дальнейшая работа заключалась в фиксации лица и его переформировании в конкретное изображение лица в формат 256x256, чтобы пропустить его через сверточную нейронную сеть.

После прохождения нейронной сетью обучения требуется тестирование ее полученной архитектуры. Происходит это посредством обработки нейронной сетью новых данных. Для этого была сделана выборка из датасета RAVDESS. Для работы над видеоматериалами были так же взяты выборки из CK+ и Aff-Wild. На основе этой выборки была проведена оптимизация гиперпараметров для получения более точного результата.

Целевая функция использовала кортеж гиперпараметров и возвращала связанные с ними потери. Использовался случайный перебор всех комбинаций на выборку их случайным образом при помощи библиотеки Keras Tuner, а именно с применением алгоритмов случайного поиска и HyperBand.

2.5. Синтез нового набора данных

В рамках проведения экспериментов были получены следующие результаты тестирования нейронной сети для видеофиксации.

Датасет RAVDESS, точность – 69.35%. Классификация по эмоциям:

- грусть,
- злость,
- нейтральное состояние,
- отвращение,
- радость,
- спокойствие,
- страх,
- удивление.

Датасет CK+: точность – 82.3%. Классификация по эмоциям:

- грусть,
- злость,
- отвращение,
- презрение,
- радость,
- страх,
- удивление.

Далее был синтезирован набор данных Aff-Wild-Em (Рис. 2.1), включающий в себя следующее сопоставление

$c_i \in$ Happy для $(0.3 < x \leqslant 1)$ и $(0.3 < y \leqslant 1)$

$c_i \in$ Calm для $(0.3 < x \leqslant 1)$ и $(-1 \leqslant y < -0.3)$

$c_i \in$ Sad для $(-1 \leqslant x < -0.3)$ и $(-1 \leqslant y < -0.3)$

$c_i \in$ Scare для $(-1 \leqslant x < -0.3)$ и $(0.3 < y \leqslant 1)$

$c_i \in$ Neutral для $(-0.3 \leqslant x \leqslant 0.3)$ и $(-0.3 \leqslant y \leqslant 0.3)$

Набор данных Aff-Wild-Em: точность – 60.7%. Классификация по состояниям: Классификация по состояниям (эмоциям): нейтральное состояние (Нейтральность), возбужденное позитивное (Радость), возбужденное негативное (Страх), расслабленное позитивное (Спокойствие), расслабленное негативное (Грусть).

Здесь в каждой из трех групп вводятся неотрицательные коэффициенты (вероятности), сумма которых равна единице. Цель распознавания - нахождение максимального коэффициента.

Остановимся на результатах датасета Aff-Wild. Он является единственным из представленных, в котором материал для анализа содержит изображения в различных ракурсах. Из этого делается следующий вывод: в датасетах CK+ и RAVDESS наличие схожих кадров лиц неизбежно привело бы к ухудшению показателей результата.

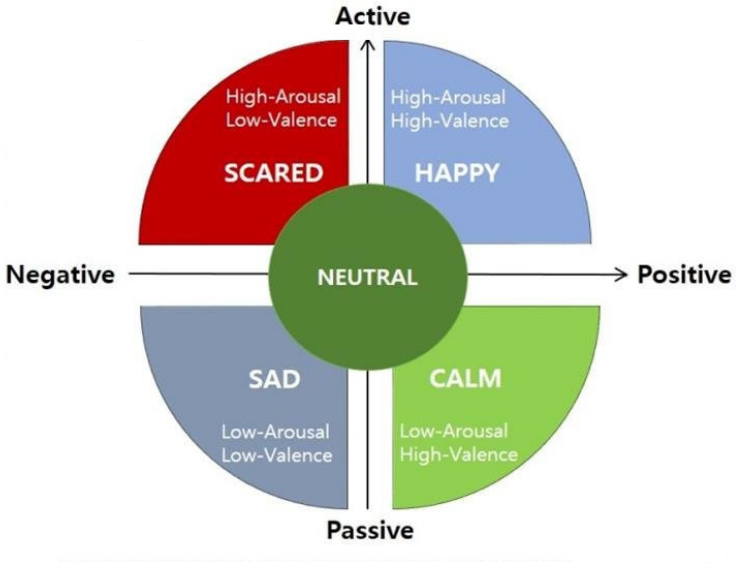

Рис. 2.1. Схема подготовки набора данных Aff-Wild для задачи классификации

На предыдущем этапе разработки была создана и обучена нейросеть, которая показывает хорошую точность в условиях, близким к идеальным: при правильном освещении, фоне, расстоянии от камеры до лица. Но при ухудшении условий точность результатов падает. Поэтому следующей задачей является создание нейросети, которая обучена уже на данных, приближенных к реальным.

Для обучения был выбран датасет Aff-Wild, который участвовал в предыдущем этапе разработки, но в качестве датасета результативной выборки. Сама организация процесса обучения выглядит следующим образом.

Обучаем первичную сеть на простой задаче, удаляя первый класс. При этом датасет сбалансируется, так как в исходной версии было много нейтральных кадров с нулевой результативностью. На этом этапе кадры обрабатываются по отдельности без последовательности. Точность на этом этапе составляет 72

Из обученной в предыдущем этапе нейросети убираем последние слои, дойдя до слоя укрупнения (maxpool) с предыдущего блока слоев. Полученную нейросеть делаем первичной. После чего пропускаем через эту нейросеть датасет RAVDESS и сохраняем найденный промежуточный результат.

Таким образом получаем промежуточный датасет для обучения нейросетей по определению эмоций и их силы. В результате такой комбинации архитек-

тур нейросети будут достаточно точны при пропускании через них материалов, приближенных к реальным условиям, потому что первичная нейросеть уже умеет с ними работать, а студийные условия выступают лишь как частный случай.

Нейросети должны объединять в себе точность, скорость работы и простоту реализации. Работа с изображением все так же предполагает работу со сверточными нейронными сетями в силу их спецификации. Точность сети можно предварительно оценить, исходя из результатов теста cifar-10. В тесте убирается первый слой, а после через нейросеть пропускаются картинки размером 32x32, разбитые на 10 классов.

Подходящие под все три критерии сети обладают большим количеством простых слоев и относятся к одному из двух видов:

полносвязные (dense), в которых результат свертки объединяется с исходными данными;

остаточные (residual), в которых результат свертки (или нескольких) суммируется с исходными данными.

Особенность обеих архитектур состоит в том, что градиент ошибки, являющийся фактором обучения, не угасает от слоя к слою, а равномерно обучает все слои сети. Кроме того, обе архитектуры используют после каждого слоя (либо перед каждым слоем) нормализацию внутри партии. Это значит, что из исходных данных вычитается среднее, а отклонение делается равным единице. Этот процесс заметно стабилизирует и ускоряет обучение и заодно повышает точность, однако замедляет работу примерно на 30

Выбран вариант нейросети с нарастанием. Используются тонкие свертки с нелинейностью elu и нормализацией результата, потому что исходные данные уже нормализованы, при добавлении новых данных остаются нормализованными, и нет смысла повторять вычисления.

Также для удобства реализации создан слой плотной свертки, в котором последовательно объединены:

тонкая свертка с ядром 1x1,

основная свертка с ядром 3x3 и нелинейностью,

нормализация по партии,

объединения с исходными данными.

Две первые операции эквивалентны обычной свертке, но требуют на порядок меньше вычислений, а значит и времени. Для нормализации обучаются параметры β и γ, и итоговое значение будет

$$result = \frac{(conv - conv_{mean})}{conv_{std}}\gamma + \beta \tag{2.3}$$

где $conv_{mean}$ - среднее по осям партии, длине и ширине свертки, $conv_{std}$ – стандартное отклонение по осям партии, длине и ширине свертки.

В отличие от стандартной архитектуры, здесь удалены полносвязные слои, образующие итоговый набор классов. Вместо них при достижении слоем размеров менее 5x5 происходит усреднение внутри слоев.

Видео читается покадрово при помощи библиотеки Scikit-video. Для удобства работы написан итератор, который перебирает элементы контейнерного класса без необходимости пользователю знать реализацию определенного контейнерного класса. Он определяет характеристики видео из заголовка и затем читает видео посекундно. Лишние кадры удаляются, а на оставшихся итератор находит лицо при помощи библиотеки Dlib, и изменяет его размер до 256x256 при помощи библиотек Scikit-image или CV2. Полученные изображения объединяем в партию и обрабатываем первичной нейросетью. Полноценно архитектура сети выглядит следующим образом:

первые слои сети - нормализация внутри партии и обнуление (dropout) с вероятностью 0.1.

далее следует свертка с ядром 5x5 и шагом 2.

4-8 плотных сверток с ядром 3x3 и 15 каналами.

укрупнение размера промежуточного изображения.

слой нормализации.

слой обнуления.

свертки 1x1 с числом каналов, равным половине входящих.

последний блок, который впоследствии будет отбрасываться.

обнуление с вероятностью 0.3.

свертка размером 90x1x1.

6 плотных сверток с ядром 1x1.

усреднение слоев.

нормализация по партии.

усреднение каналов до 5 созданных классов.

Контрольная выборка была взята из расширенного датасета RAVDESS, в которой три актера произносили одну фразу с разными эмоциональными оттенками. Итоговая нейросеть показала точность в 91%, тем самым можно сделать вывод об улучшении качества классификации на 21% по сравнению с предварительными итогами.

Результаты исследования в этой главе были взяты за основу SaaS платформы оценки поведения человека по видеопотоку в режиме реального времени, которая была включена в реестр участников проекта создания и обеспечения функционирования инновационного центра «Сколково» под номером 1123236 на основании решения Некоммерческой организации Фонд развития Центра разработки и коммерциализации новых технологий о присвоении статуса участника проекта создания и обеспечения функционирования инновационного центра «Сколково».

2.6. Выводы ко второй главе

Результаты второй главы опубликованы в работах [2, 25].

Предложена и программно реализована архитектура многослойной нейронной сети для решения задачи определения эмоции человека на видеозаписи, подготовленной в нестудийных условиях, точность классификации которой ука-

Таблица 2.2. Таблица итоговой точности распознавания эмоций на видеозаписях с применением различных методов

	Aff-Wild-Em	RAVDESS
Local Binary Patterns	-	58,7%
Предложенная архитектура	-	69,4%
После смены набора данных	60,7%	91,0 %

зана в таблице 2.2.

Рассматривается возможность повышения точности распознавания с добавлением к текущим данным информации об уровне стресса говорящего, которые можно получить при помощи полиграфа.

Глава 3

Методы классификации в исследованиях на полиграфе

3.1. Краткий обзор существующих подходов

Полиграф является медико-биологическим прибором для одновременного измерения динамики физиологических показателей определенных психофизиологических реакций.

Согласно психофизиологическому феномену, сознание индивида неразрывно связано с физиологическими процессами, протекающими в его организме. Следовательно, изменение психологического состояния, а именно динамику уровня стресса, можно отследить при помощи полиграфа.

Центры ВНС контролируются подкорковыми ядрами мозга, поэтому сознательно человек не может контролировать их деятельность без специальной подготовки. В данном случае исключение составляют, например, йоги. Однако даже им для изменения своего физиологического состояния необходимо использовать медитативные методики, которые легко диагностируется визуально (внешний вид и поведение), либо по характерным речевым тенденциям.

Полиграф считывает следующие психофизиологические показатели:

- Динамика дыхательных циклов;
- Электрокожное сопротивление (кожно-гальваническая реакция);
- Артериальное давление;
- Перефирическое кровонаправление (фотоплетизмограмма);
- Двигательная активность (тремор).

Как известно, полиграф работает так: имеется круг вопросов из некоторой предметной области. К испытуемому индивиду подключаются датчики для вышеуказанных измерений. Принимается ответ: либо «да», либо «нет». Все во-

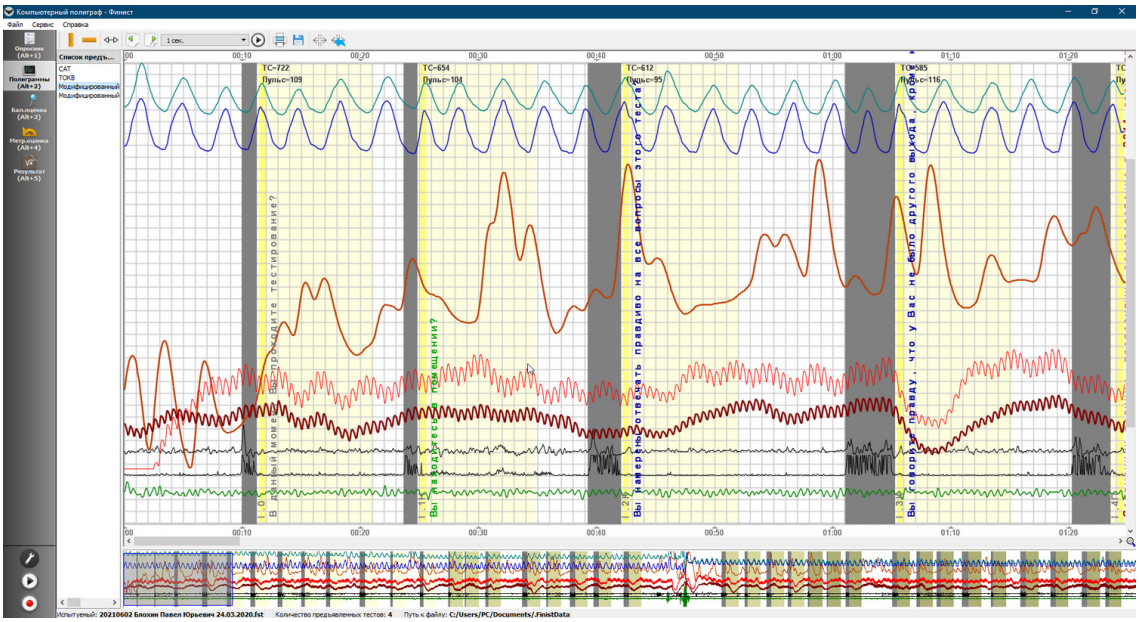

Рис. 3.1. Скриншот полиграммы в интерфейсе профессионального компьютерного полиграфа «Финист». Каналы съема психофизиологических характеристик отмечены следующими цветами: грудно дыхание - синий, дифрагмально дыхание - бирюзовый, кожно-гальваническая реакция - коричневый, пьезоплетизмограмма - бордовый, фотоплетизмограмма - красный, тремор - зеленый цвет

просы повторяются не менее 3-х раз. Полиграфолог анализирует полученные функции с помощью системы трехбалльной оценки и окончательно констатирует о справедливости ответа [16].

Суть такой системы заключается в том, что реакциям на каждый вопрос теста выставляются баллы в следующем порядке:

- (сильная) максимальная по выраженности и близкие к ней реакции получают по 2 балла,
- (средняя) вторая по выраженности реакция и близкие к ней - 1 балл,
- (слабая) все остальные реакции - 0 баллов.

Критерии визуальной оценки силы реакции на вопросы по каждому показателю отличаются. Они подробно описаны в профильной литературе. Реакция на каждый вопрос обсчитывается отдельно по респираторному каналу, кардиоканалу и электро-дермальному каналу. В итоге получается 3 числа – по одному на показатель. Для принятия решения суммируются баллы по 3 каналам по

всем повторениям.

Например, если в результате трех повторений вопроса получилось 11 баллов, то можно говорить о высокой ситуативной значимости реакции и возможном сокрытии информации, если 2 балла – то имеет место слабая реакция (это указывает на то, что индивид не испытывает стресс и ему нечего скрывать).

Для однозначности понимания терминологии далее в этой главе балльная оценка реакции заменяется на равнозначный процесс классификации силы реакции на слабую, среднюю и сильную.

По каждому каналу регистрации психофизиологических характеристик можно выделить следующие особенности их балльной оценки полиграфологами [5, 16].

Канал кожно-гальванической реакции (КГР)

Электрокожное сопротивление - разность потенциалов между двумя участками на поверхности кожи. Канал кожно-гальванической реакции в полиграфе регистрирует именно такую биоэлектрическую активность на поверхности ладони или на двух пальцах руки.

Изменения в данном канале при реакциях обусловлены снижением электрического сопротивления на поверхности кожи вследствие выделения желез, реагирующих на изменение эмоционального состояния и уровня стресса.

Анализ информации о структуре и механизмах возникновения и регуляции кожно-гальванической реакции, а также ее информативных признаков позволяет сделать следующие выводы:

- динамика проявления кожно-гальванической реакции может являться критерием уровня стресса и эмоционального напряжения;
- интенсивность кожно-гальванической реакции зависит от новизны стимула, свойств нервной системы, уровня мотивации исследуемого и его функционального состояния.

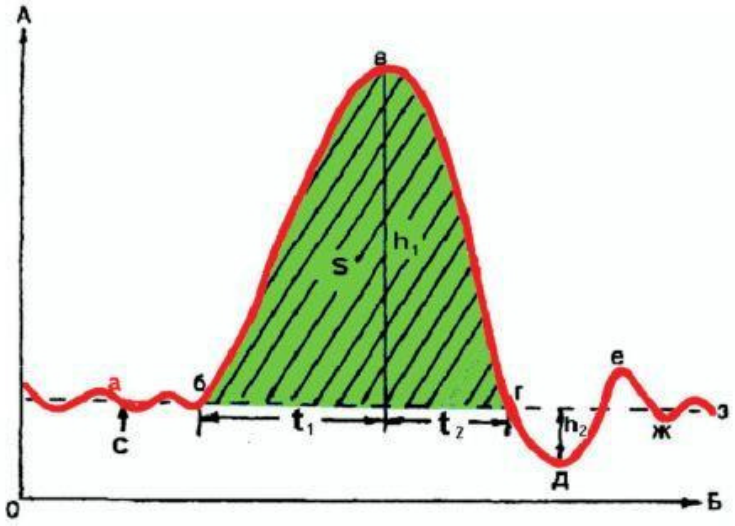

Рис. 3.2. Схематическое изображение реакции в канале КГР

На Рис. 3.2 изображено схематическое изображение реакции и характеристики, используемые для последующего анализа в канале кожно-гальванической реакции, а именно:

- отрезок А-Б: время запаздывания реакции;
- отрезок Б-В: длина восходящей части сигнала (сила активирующих процессов возбуждения);
- отрезок В-Г: длина нисходящей части сигнала (интенсивность активации тормозных процессов);
- отрезок t_1: время реакции, за которое она достигает локального максимума (характеризует подвижность процессов возбуждения);
- отрезок h_1: максимальная амплитуда сигнала относительно средней линии (характеризует силу ответной реакции центральной нервной системы исследуемого лица на стимул);
- площадь участка S: интегральный показатель, определяющийся амплитудой (h_1) и общей длительностью реакции;

- отрезок h_2 амплитуда отрицательной части колебания ()определяет степень тормозных процессов в центральной нервной системе).

Канал «Дыхание»

Дыхание обеспечивает одну из основных функций организма человека: функцию газообмена между организмом и окружающей средой, то есть поступление кислорода в организм и удаление углекислого газа.

В состоянии стресса потребление кислорода в организме человека происходит интенсивнее, чем в состоянии покоя. Это обусловлено тем, что стресс мобилизует функциональные резервы организма человека. И для адаптации к стрессогенным условиям в организме активизируются дополнительные физиологические процессы. Соответственно, в норме, увеличение уровня стресса ведет за собой изменение характеристик дыхания.

Данное положение позволяет сделать вывод, что дыхание является показателем степени эмоционального напряжения, эмоционального стресса исследуемого лица. Это обусловливает необходимость использования показателей дыхания при оценке степени стресса и эмоционального напряжения исследуемого лица при проведении проверок с использованием полиграфа.

При проведении полиграфных проверок производится регистрация показателей дыхания, определяющегося изменением объема легких на вдохе и на выдохе.

Показатели данного физиологического параметра регистрируются при помощи двух датчиков, опоясывающих грудную клетку и реагирующих на изменение силы их натяжения. Таким образом фиксируется грудное и диафрагмальное дыхание.

Для оценки силы реакции в канале дыхания учитывают:

- динамику амплитуды колебаний сигнала на 3 дыхательных циклах;
- отклонение средней линии сигнала;

Автор	Год	Использовавшиеся критерии
Benussi	1914	Изменение соотношения вдох – выдох на кривой дыхания
Larson	1923	Уменьшение амплитуды дыхания, появление реакции задержки (на выдохе)
Keeler	1930	Уменьшение амплитуды дыхания
Winter	1936	Уменьшение амплитуды дыхания, появление реакции задержки
Lee	1937	Уменьшение амплитуды дыхания
	1943	Уменьшение и увеличение амплитуды дыхания
Marston	1938	Изменение соотношения вдох – выдох, уменьшение и увеличение амплитуды дыхания
Trovillo	1942	Изменения амплитуды, появление реакции задержки (на выдохе) – по дыхательной кривой
Inbau	1942 1948	Изменение амплитуды дыхания Изменения амплитуды, изменение базовой линии, появление реакции задержки (на выдохе)
Haney	1944	Изменения амплитуды, базовой линии и появление реакции задержки (на выдохе)

Рис. 3.3. Информативные признаки, используемые при ручном анализе канала дыхания.

- задержки дыхания;

Важно учитывать, что ряд изменений дыхательных циклов буду иметь ситуационный характер, такие как: ослабление дыхания после одного цикла глубокого вдоха/выдоха, глотание, приготовление к даче ответа, дыхательный цикл при ответе, попытка противодействия.

На Рис. 3.3 представлена хронология изменений определения критериев информативных признаков дыхания.

Канал плетизмограммы.

Канал плетизмограммы отвечает за фиксацию динамики показателей сердечной деятельности в зависимости от фазы кардиоцикла, а также оптической плотности ткани, с помощью регистрации амплитуды колебаний объема крови в сосудах путем просвечивания участка ткани. Съем сигнала осуществляется с пальца руки.

Дополнительная запись плетизмограммы производится при помощи пьезодатчика. Он регистрирует микроизменения давления кожи от пульсации крови в подушечке пальца.

Оценка плетизмограммы позволяет сделать оценку функционального состояния организма. Сердце неизбежно реагирует на изменение эмоционального состояния и уровня стресса, что проявляется в динамике кровенаполняемости сосудов и их эластичности.

Для оценки силы реакции в канале плетизмограммы учитывают:

- динамику изменения амплитуды колебаний сигнала;
- повышение/понижение огибающих сигнала;
- частоту колебаний;
- время возвращения избыточного давления крови к исходному значению;

К данному моменту применяют только прямые методы обсчета полиграмм [59, 12, 15]. В результате вычисляются и сравниваются следующие характеристики:

- длина линии сигнала верхнего (грудного) и нижнего (диафрагмального) дыхания,
- количество дыхательных циклов,
- максимальная амплитуда КГР,
- площадь под графиком сигнала КГР,
- смещение средней линии плетизмограммы,
- и другие характеристики.

К сожалению, индивидуальные физиологические особенности каждого человека привносят неравномерные искажения по одному или нескольким сигналам. Прямые методы обсчетов не позволяют сделать автоматическую подстройку к таким изменениям, поэтому точность таких подходов является недостаточной для самостоятельного принятия решения по результатам проведенного психофизиологического исследования. Далее ставится задача исправить этот недостаток при помощи нейронных сетей и машинного обучения.

Применение полиграфа требует высокой квалификации специалиста, требующей учет различных факторов при подготовке и проведении проверки, но в то же время нет аналогов по точности и удобству применения в задачах выявления скрываемой информации.

Понимание механизма эмоциональной обратной связи и расчет балльной оценки по проведенному тестированию может помочь создать в дальнейшем инструментарий для подготовки автоматического «второго мнения» для полиграфолога.

Интересным направлением развития является интеграция полиграфа с дополнительными биометрическими модальностями, показывающими хорошую чувствительность в близких задачах, а именно с трехмерной картой лица [86] и зрачковой реакцией [139].

3.2. Подготовка набора данных

Каждая реакция представляет 12 с записанной полиграммы с частотой сигнала 20 Гц, т.е. 240 точек по каждому каналу:

- КГР (электрическая активность кожи),
- фотоплетизмограмма,
- пьезоплетизмограмма,
- грудное дыхание,
- диафрагмальное дыхание.

Соответственно, рассматриваются 3 следующих независимых задачи классификации:

1) Для канала дыхания:

X_{breath} - множество зарегистрированных по каналам дыхания (грудного и диафрагмального) полиграфом реакций на вопросы, состоящее из векторов:

$x_i = (x_{i1}, \ldots, x_{i480}), x_{ij} \in \mathbf{Z}$;

$Y = \{$«Слабая реакция», «Средняя реакция», «Сильная реакция»$\}$;

$\{(x_1, y_1), \ldots, (x_m, y_m)\}$ - обучающая выборка

Требуется построить алгоритм $a : X_{breath} \to Y$, способный классифицировать произвольный объект $x \in X_{breath}$

2) Для канала кожно-гальванической реакции (КГР):

X_{kgr} - множество зарегистрированных по каналу КГР полиграфом реакций на вопросы, состоящее из векторов:

$x_i = (x_{i1}, \ldots, x_{i240}), x_{ij} \in \mathbf{Z}$;

$Y = \{$«Слабая реакция», «Средняя реакция», «Сильная реакция»$\}$;

$\{(x_1, y_1), \ldots, (x_m, y_m)\}$ - обучающая выборка

Требуется построить алгоритм $a : X_{kgr} \to Y$, способный классифицировать произвольный объект $x \in X_{kgr}$

3) Для канала плетизмограммы:

X_{pg} - множество зарегистрированных по каналам плетизмограммы (фото и пьезо датчики) полиграфом реакций на вопросы, состоящее из векторов:

$x_i = (x_{i1}, \ldots, x_{i480}), x_{ij} \in \mathbf{Z}$;

$Y = \{$«Слабая реакция», «Средняя реакция», «Сильная реакция»$\}$;

$\{(x_1, y_1), \ldots, (x_m, y_m)\}$ - обучающая выборка

Требуется построить алгоритм $a : X_{pg} \to Y$, способный классифицировать произвольный объект $x \in X_{pg}$

Разметка данных. По каждой реакции группа профессиональных полиграфологов проставляет силу реакции на вопрос (слабая, средняя, сильная) в соответствии с описанными выше правилами.

Так было обработано 90 психофизиологических исследований с разными испытуемыми и получено 8 000 классифицированных по каждому каналу отрезков.

3.3. Сравнительное тестирование архитектур

Рассматриваются встроенные архитектуры библиотеки scikit-learn [119, 118]. Сначала выбираем подготовленные “сырые” необработанные данные поочередно как в равном количестве, так и в несбалансированном отношении 7:3. Применяем следующие алгоритмы классификации данных, реализованные в указанной выше библиотеке. Они создают для каждого из трех показателей соответствующую архитектуру для расчета результатов полиграфного тестирования.

SVM (support vector machine) — набор схожих алгоритмов обучения с учителем, использующихся для задач классификации и регрессивного анализа [64, 65].

GPC (gaussian process) - случайный процесс Гаусса (набор случайных величин, индексированных по времени или пространству), такой, что каждый конечный набор этих случайных величин имеет многомерное нормальное распределение, т. е. каждая их конечная линейная комбинация обычно распределена.

GNB (gaussian naive bayes) — простой вероятностный классификатор, основанный на применении теоремы Байеса со строгими предположениями о независимости и нормальном распределении признаков в наборе данных.

AdaBoost (с параметрами: количество оценщиков – 2500, коэффициент обучения - 0.001) – адаптивный классификатор в том смысле, что последующие слабые ученики настраиваются в пользу тех экземпляров, которые были неправильно классифицированы предыдущими классификаторами.

MLP (multi-layer perceptron) — класс упреждения искусственной нейронной сети.

Gradient Boosting [61] - метод машинного обучения для регрессии, классификации и других задач, который создает модель прогнозирования в виде ансамбля слабых моделей прогнозирования, обычно деревьев решений.

DecisionTree — средство поддержки принятия решений, использующееся в

машинном обучении, анализе данных и статистике.

RandomForest — алгоритм машинного обучения, заключающийся в применении комитета решающих деревьев.

ExtraTrees [21] - ансамблевый алгоритм машинного обучения, который объединяет прогнозы из многих деревьев решений.

В итоге получено 27 моделей, по одной на каждый показатель и на каждый алгоритм классификаций. Они получают на вход прямые данные с полиграфа, а на выходе выдают класс, соответствующий силе реакции: слабая, средняя или сильная.

Выбираем три наиболее точных классификатора по каждому каналу. Предположим, что при оценке определенной реакции на вопрос с вероятностью 100% первые два алгоритма получили класс «Слабая реакция», а третий - класс «Сильная реакция». Для адекватного принятия решения, особенно в случаях несогласованности между классификаторами, применяем надстройку в виде VotingClassificator[72], которая присваивает весовые коэффициенты каждому из них. Итоговый результат будет рассчитан по формуле

$$\max(w_1 p_{11} + w_2 p_{21} + w_3 p_{31}, w_1 p_{12} + w_2 p_{22} + w_3 p_{32}, w_1 p_{13} + w_2 p_{23} + w_3 p_{33}) \quad (3.1)$$

где w_i - весовой коэффициент i-го классификатора, p_{ij} – вероятность принадлежности реакции к классу j, полученная при помощи классификатора i.

Пусть в указанном выше примере коэффициенты равны 0.2, 0.35 и 0.45 соответственно. По этим результатам принимается следующее решение: класс «Слабая реакция» - вероятность 55%, класс «Сильная реакция» - вероятность 45%. Поэтому в такому случает выбирается класс «Слабая реакция». Аналогично рассматриваются другие ситуации для выбора оптимального классификатора.

Точность классификации силы реакции по трем классам на архитектуре VotingClassificator, включающей в себя три наилучших алгоритмов, составила

52%, а из пяти – 59%.

При повторном анализе файлов полиграмм было отмечено, что классы "слабая"и "сильная"реакции, представляющие собой минимальные и максимальные проявления дестабилизации организма соответственно, являются наиболее различимыми и визуально и численно. В то же время среди экземпляров целевого класса "средней"реакции мера различимости мала.

При последующем анализе первоначальных вариантов классификаторов был выявлен следующий ряд их недостатков:

• Метод опорных векторов естественным образом дает меньшую результативность в силу требований к высокой мере различимости поставляемых данных;

• Гауссовский процесс, чаще применяемый для решения задач регрессии, в рамках задачи классификации требует сопроводительный материал в виде, например, статистических распределений или набора корреляционных данных, а не только, непосредственно, образцов сигнала;

• Гауссовский наивный байесовский классификатор подходит для проверки гипотез, однако, при потенциальном добавлении новых категорий, не представленных в оригинальном наборе данных – таким данным будет выставлена нулевая вероятность в момент прогнозирования. Кроме того, данный классификатор также имеет требования к высокой различимости категорий между собой;

• Алгоритм AdaBoost крайне требователен к качеству данных;

• Схожие требования (в силу наличия высокого числа параметров и полносвязности) к качеству набора данных имеет и алгоритм MLP. Помимо этого, данный классификатор является устаревшим, по сравнению с новыми вариациями методов;

• При внесении изменений в набор данных (например, при добавлении новых образцов) внутренняя структура дерева решений, полученная от алгоритма Decision Tree, может полностью измениться, что влечет за собой неста-

бильность, как при процессе обучения, так и прогнозирования;

• Классификатор на основе алгоритма Gradient Boost склонен к переобучению, кроме того, данный вариант классификатора крайне чувствителен к наличию краевых случаев в наборе данных;

• Алгоритм Random Forest является подобием “черного ящика”, что дает меньше контроля над поведением системы, построенной на его основе. Помимо этого, данный алгоритм – один из самых требовательных к вычислительной мощности аппаратуры, на которой выполняются эксперименты или полноценное развертывание;

• Extra Tree – это ансамблирующий метод для деревьев решений, следовательно, также не является лишенным недостатков, указанных при рассмотрении алгоритма Decision Tree;

3.4. Нормализация данных

Необходимо упомянуть, что данные при обучении с использованием указанных выше архитектур подавались в различных вариантах: без нормализации, нормализированные на всем наборе данных, масштабированные на всем наборе данных, нормализированные в рамках одного повторения одного блока вопросов одной полиграммы.

По результатам проведенных численных экспериментов наиболее высокую точность классификации удалось получить с применением нормализации в рамках одного повторения одного блока вопросов одной полиграммы.

Структура каждой полиграфной проверки выглядит следующим образом:

Тест №1: предъявление №1, вопросы №1, №2, ..., №12

Тест №1: предъявление №2, вопросы №1, №2, ..., №12

Тест №1: предъявление №3, вопросы №1, №2, ..., №12

(возможный перерыв)

Тест №2: предъявление №1, вопросы №1, №2, ..., №12

Тест №2: предъявление №2, вопросы №1, №2, ..., №12

Тест №2: предъявление №3, вопросы №1, №2, ..., №12

Согласно методологии полиграфных проверок [16], признаки силы реакции меняются от предъявления к предъявлению даже в рамках одного блока вопросов (теста). Индивидуальные особенности испытуемых сохраняются на протяжении всей проверки.

На основании этого был сделан вывод о том, что нормализация должна проходить по наиболее стабильному этапу проверки. Поэтому был применен метод z-масштабирования по каждому предъявлению:

$$z_{ijkl} = \frac{x_{ijkl} - \mu_{ijk}}{\sigma_{ijk}} \tag{3.2}$$

где i - идентификатор испытуемого (исследования), j - номер теста, k - номер предъявления, l - номер вопроса, μ_{ijk} и σ_{ijk} - среднее значение и стандартное отклонение значений психофизиологических показателей на соответствующем предъявлении.

3.5. Применение архитектуры трансформера

Для повышения точности остановимся подробнее на реализации классификатора для двух классов при помощи архитектуры трансформера, измененной в соответствии с задачей. По аналогии с рекуррентными нейронными сетями (РНС) трансформеры предназначены для обработки последовательностей, таких, как текст на естественном языке, и решения таких задач, как машинный перевод и автоматическое реферирование. В отличие от РНС, трансформеры не требуют обработки последовательностей по порядку. Например, если входные данные — это текст, то трансформеру не требуется обрабатывать конец текста после обработки его начала.

В классическом варианте трансформер представляет собой две части – кодировщик и декодировщик. В данном случае была использована структура,

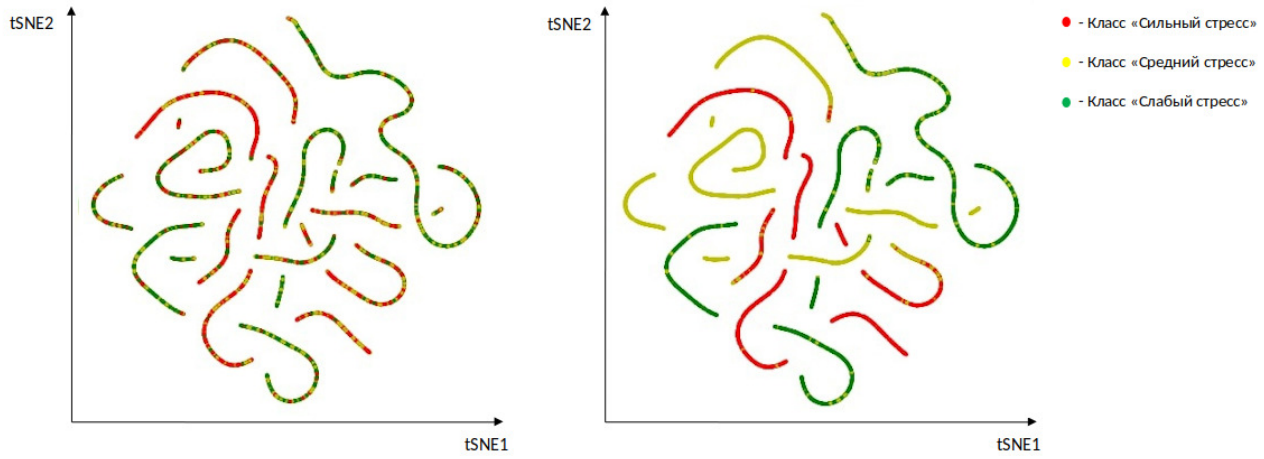

Рис. 3.4. Результаты кластеризации набора данных по каналу дыхания методом стохастического вложения соседей с t-распределением (t-SNE): экспертная разметка (слева) и результаты классификации предложенным трансформером (справа).

состоящая в части кодировщика из чередующихся слоев внимания (внимание на основе скалярного произведения) и многослойного перцептрона, а также декодировщика (простого классификатора). Обе части трансформера реализованы стандартными средствами библиотеки Keras.

Данные на вход подаются в размерности (26; 240), т.е. единовременно подается 26 образцов данных размером 240 значений. Затем данные кодируются при помощи одноразмерного (1D) PatchEncoder с шириной (8; 2), т.е. линейно трансформируются проецированием на вектор указанной выше размерности.

Кодированные данные поступают на группу чередующихся слоев внимания и перцептронов, а затем поступают на слой SeqPool для субдискретизации. Этап подвыборки последовательности нормируется (LayerNormalization) средствами библиотеки Keras и данные с него затем поступают на полносвязный слой в 600 нейронов с активацией SeLu. Результат работы кодировщика на этом этапе укрупняется слоем FeaturePooling до двух групп.

Главные характеристики:

размер набора данных (batch size): 60,

количество блоков (чередование внимания и перцептрона): 6,

размерность слоя внимания: 312,

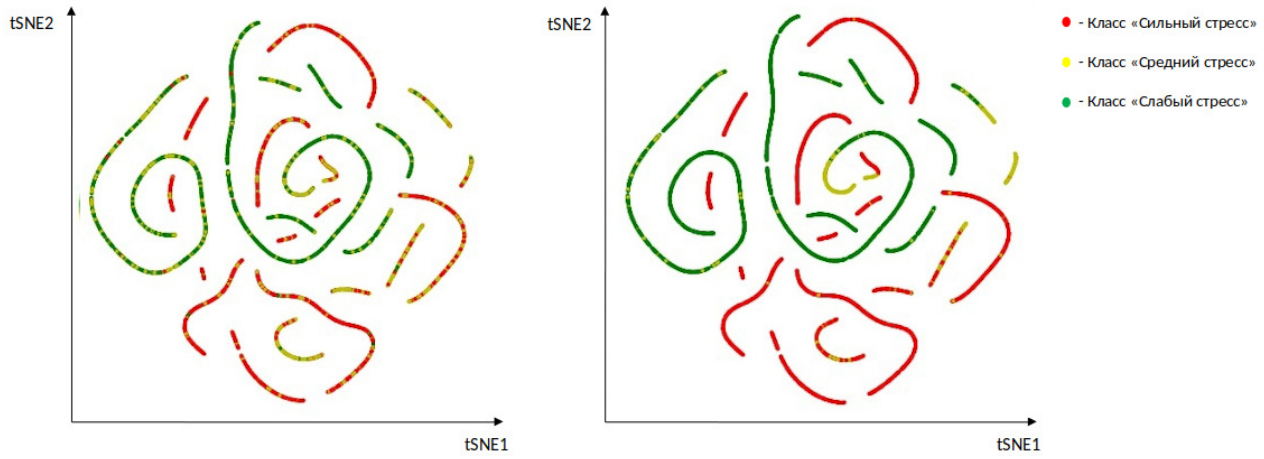

Рис. 3.5. Результаты кластеризации набора данных по каналу кожно-гальванической реакции методом стохастического вложения соседей с t-распределением (t-SNE): экспертная разметка (слева) и результаты классификации предложенным трансформером (справа).

выброс слоя внимания: 0.1 (10

функция трансформации ядра: softmax,

количество наборов матриц весов (запросов, ключей, значений): 6.

Текущая архитектура скомпилирована с использованием оптимизатора Adam (коэф. обучения 0.01) и функции потерь SquaredHinge (квадратичная кусочно-линейная функция потерь). Подгонка осуществлялась на протяжении 50 эпох с применением ранней остановки, если валидационная точность не увеличивается на протяжении 12 эпох и уменьшением коэффициент обучения на пяти эпохах.

Точность обучения на два класса («сильная» и «слабая» реакции на заданный вопрос) составляет для показателей:

плетизмограммы – 86.8%,

кожно-гальванического сопротивления – 95.3%,

дыхания – 72.7%.

Также были предприняты попытки обучить классификатор трансформера сразу на три класса, однако они показали заметно худшую точность и большой процент ошибок в сравнении классов «Средняя реакция»-«Слабая реакция» и «Средняя реакция»-«Сильная реакция». Используемый метод заметно уменьшает долю неясных средних реакций в пользу выраженных слабых и сильных.

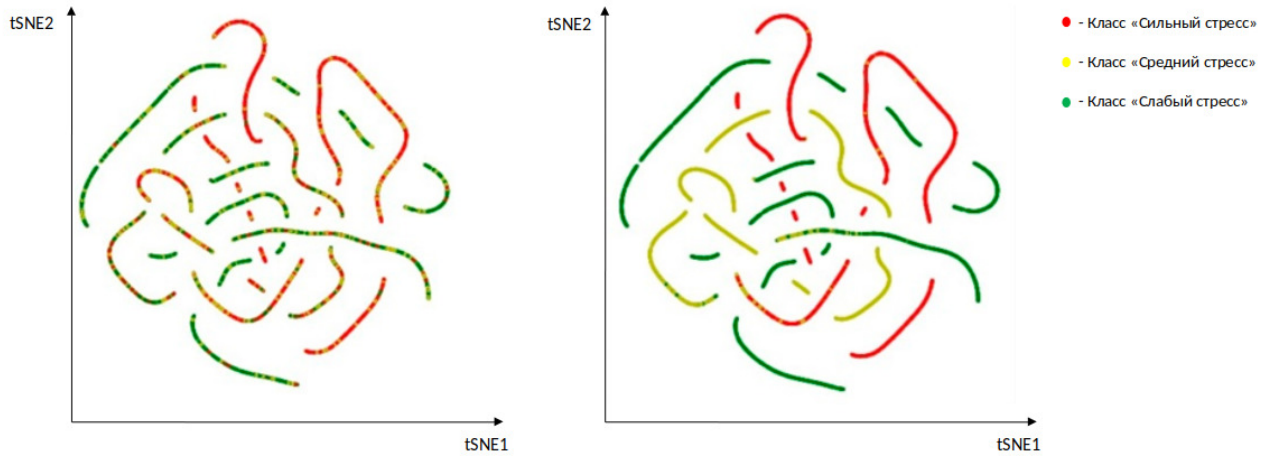

Рис. 3.6. Результаты кластеризации набора данных по каналу фото и пьезо плетизмограммы методом стохастического вложения соседей с t-распределением (t-SNE): экспертная разметка (слева) и результаты классификации предложенным трансформером (справа).

Метод стохастического вложения соседей с t-распределением.

Стохастическое вложение соседей (SNE) впервые было введено в [69]. Данный метод размещает объекты в пространстве более низкой размерности с условным сохранением расстояния между соседями.

Благодаря описанным выше свойствам SNE может создавать достаточно хорошую визуализацию, но алгоритм имеет большой недостаток в виде сложной функции потерь.

В [102] был представлен метод стохастического вложения соседей с t-распределением. Его целью является преобразование многомерного набора данных $X = (x_1, \ldots, x_n)$ в $Y = (y_1, \ldots, y_n)$ с уменьшением размерности объектров. Данный метод является более оптимизированным и обеспечивает более высокое качество визуализации за счет снижения тенденции к скоплению точек вместе в центре.

Функция потерь в TSNE отличается от его предшественника. В данном случае применяется симметричная версия SNE для уменьшения влияния проблемы выбросов.

Метод SNE преобразует многомерную евклидовую дистанцию в вероятность, описывающую сходство точек.

Ассимметричный вариант расчета SNE задается следующим образом;

$$q_{ij} = \frac{e^{-||y_i - y_j||^2}}{\sum_{k \neq i} e^{-||y_k - y_j||^2}} \tag{3.3}$$

где q_{ij} — попарное сходство в пространстве более низкой размерности. Такие же вероятности по изначальным объектам p_{ij} задаются следующим выражением

$$p_{ij} = \frac{e^{-||x_i - x_j||^2/2\sigma^2}}{\sum_{k \neq i} e^{-||x_k - x_j||^2/2\sigma^2}} \tag{3.4}$$

Эти уравнения называются симметричными, так как:

$$p_{ij} = p_{ji}, q_{ij} = q_{ji} \text{ для } \forall i, j \tag{3.5}$$

Другая уникальность TSNE заключается в применении t-распределения Стьюдента со степенью свободы $v = 1$. Оно имеет тяжелый хвост в пространстве малой размерности. Совместные вероятности для низкоразмерного отображения q_{ij} можно посчитать следующим образом:

$$q_{ij} = \frac{\left(1 + ||y_i - y_j||^2\right)^{-1}}{\sum_{k \neq i} \left(1 + ||y_i - y_j||^2\right)^{-1}} \tag{3.6}$$

Конечной целью метода TSNE является представление p_{ij} через q_{ij} как можно точнее, поэтому функция потерь задается выражением:

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{3.7}$$

Метод градиентов применяется для минимизации функции потерь:

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j) \left(1 + ||y_i - y_j||^2\right)^{-1} \tag{3.8}$$

Данное уравнение можно интерпретировать как сумму равнодействующей силы, тянущей y_i в направлении y_j или отталкивающей его в зависимости от того, является ли объект j соседом i.

Градиентный спуск инициализируется выборкой точки $Y^{(0)} = (y_1, \ldots, y_n)$ случайным образом из $(N)(0, 10^{-4}I)$. Импульс добавляется для ускорения оптимизации и избежания застревания в локальном оптимуме.

$$Y^{(t)} = Y^{t-1} + \zeta\frac{\partial C}{\partial Y} + \alpha(t)(Y^{t-1} - Y^{t-2}) \tag{3.9}$$

где $Y^{(t)}$ — решение на итерации t, ζ — скорость обучения, $\alpha(t)$ — импульс на итерации t.

Для визуализации и оценки результатов работы FeaturePooling данные были обработаны нейронной сетью без последнего слоя. Данные с ручной разметкой для каждой из характеристик нормируются и кластеризуются в t-SNE и сравниваются с результатами работы трансформера. На Рис. 3.4, 3.5, 3.6 можно заметить, что классифицированные машиной точки «Средняя реакция» не разбросаны хаотично, а образуют сплошные непрерывные линии после обработки трансформером.

Кроме того, как видно по точки категории «Средняя реакция» намного реже попадают в кластеры «Слабая реакция» и «Сильная реакция» по сравнению с ручной разметкой. Это означает наличие возможности повышения точности обучения на три класса.

По результатам проведенного исследования можно сделать следующие выводы:

- Увеличение числа блоков в архитектуре дает умеренную точность (до 4го блока зависимость сильная, после 8го – низкая). В отличии от зависимости точности и количества в оценке изображений.

- Нестабильные признаки образуют большое количество нейронов со значением, близким к нулю. Поэтому метод стохастического градиентного

спуска (SGD) показывает высокую эффективность на 1-2 эпохах.

- Для обучения архитектур потребовалось 5 эпох для сверточных сетей, 20 для трансформера. «Укрупнение» входных данных ускоряет обучение трансформера.

- Для предотвращения переобучения в обязательном порядке следует использовать методы автоблокировки (понижение скорости через 5 эпох, остановка через 20).

3.6. Выводы к третьей главе

Результаты третьей главы опубликованы в работах [10, 48].

Предложен метод нормализации психофизиологических характеристик, полученных при помощи полиграфа, учитывающий индивидуальные особенности испытуемого.

	Вопрос	ИИ1	Чел1	ИИ2	Чел2	ИИ3	Чел3	ИИ всего	Чел всего
		ФПГ	ФПГ	ФПГ	ФПГ	ФПГ	ФПГ	5	4
		КГР	КГР	КГР	КГР	КГР	КГР	2	3
1.	Вы находитесь в помещении?	Дых	Дых	Дых	Дых	Дых	Дых	2	3
		5	5	2	1	2	4	**9**	**10**
		ФПГ	ФПГ	ФПГ	ФПГ	ФПГ	ФПГ	4	3
		КГР	КГР	КГР	КГР	КГР	КГР	1	3
2.	Вы намерены солгать хотя бы на один вопрос этого теста?	Дых	Дых	Дых	Дых	Дых	Дых	2	5
		4	5	0	2	3	4	**7**	**11**
		ФПГ	ФПГ	ФПГ	ФПГ	ФПГ	ФПГ	1	4
		КГР	КГР	КГР	КГР	КГР	КГР	0	4
3.	[illegible]	Дых	Дых	Дых	Дых	Дых	Дых	4	4
		2	5	2	4	1	3	**5**	**12**
		ФПГ	ФПГ	ФПГ	ФПГ	ФПГ	ФПГ	4	5
		КГР	КГР	КГР	КГР	КГР	КГР	2	6
4.	[illegible]	Дых	Дых	Дых	Дых	Дых	Дых	3	4
		1	4	4	6	4	5	**9**	**15**
		ФПГ	ФПГ	ФПГ	ФПГ	ФПГ	ФПГ	3	3
		КГР	КГР	КГР	КГР	КГР	КГР	0	1
5.	[illegible]	Дых	Дых	Дых	Дых	Дых	Дых	3	4
		2	4	1	3	3	1	**6**	**8**
		ФПГ	ФПГ	ФПГ	ФПГ	ФПГ	ФПГ	2	2
		КГР	КГР	КГР	КГР	КГР	КГР	2	5
6.	[illegible]	Дых	Дых	Дых	Дых	Дых	Дых	4	5
		2	4	4	4	2	4	**8**	**12**
		ФПГ	ФПГ	ФПГ	ФПГ	ФПГ	ФПГ	1	3
		КГР	КГР	КГР	КГР	КГР	КГР	2	5
7.	[illegible]	Дых	Дых	Дых	Дых	Дых	Дых	2	4
		0	4	4	4	1	4	**5**	**12**

Рис. 3.7. Пример расчета балльной оценки при помощи предложенного метода в профессиональном компьютерном полиграфе «Финист».

Создан модуль для автоматической классификации силы реакции челове-

Таблица 3.1. Таблица итоговой точности автоматической классификации силы реакции на полиграммах

Канал регистрации	Voting3, 3 кл.	Voting5, 3 кл.	Трансформер, 2 кл.
Дыхание	45,6%	49,3%	72,7%
КГР	52,1%	59,3%	95,3%
ФПГ	46,9%	53,7%	86,8 %

ка на предъявляемые стимулы при помощи оценки регистрируемых полиграфом параметров (дыхание, сердечно-сосудистая и электродермальная активность).

В настоящий момент данные нейронные сети интегрированы в программное обеспечение профессионального компьютерного полиграфа «Финист». Получены результаты по эффективности для оценки факторов риска кандидатов при трудоустройстве.

Заключение

1. Предложена и программно реализована архитектура нейронной сети для решения задачи определения эмоции на голосовой фонограмме с высокой точностью.

2. Предложена и программно реализована архитектура многослойной нейронной сети для решения задачи определения эмоции человека на видеозаписи, подготовленной в нестудийных условиях.

3. Предложен метод нормализации психофизиологических характеристик, полученных при помощи полиграфа, учитывающий индивидуальные особенности испытуемого.

4. Создан модуль для автоматической классификации силы реакции человека на предъявляемые стимулы при помощи оценки регистрируемых полиграфом параметров (дыхание, сердечно-сосудистая и электродермальная активность).

Список литературы

1. Александров А.А., Кирпичников А.П., Ляшева С.А., Шлеймович М.П. Анализ эмоционального состояния человека на изображении // Вестн. технологического ун-та. 2019. Т. 22. № 8. С. 120–123.
2. Ахияров Ф.Р., Деревягин Л.А., Макаров В.В., Цурков В.И., Яковлев А.Н. Покадровое определение эмоций на видеозаписи с применением многослойных нейронных сетей // Изв. РАН ТИСУ. 2022. №2. С. 80-85.
3. Березанская Н.Б., Нуркова В.В. Психология. - Красногорск: Высшее образование, 2009. – 575 с.
4. Бобе А.С., Конышев Д.В., Воротников С.А. Система распознавания базовых эмоций на основе анализа двигательных единиц лица // Инженерный журнал: наука и инновации. 2016. № 9. С. 7.
5. Варламов В.А. Детектор лжи. - М.: Когито-центр, 2004, 540 с.
6. Визильтер Ю.В., Выголов О.В., Желтов С.Ю., Князь В.В. Метрический подход к семантико-морфологическому сравнению изображений // Вестн. компьютерных и информационных технологий. 2020. Т.17. № 5(191). С. 3-12.
7. Визильтер Ю.В., Горбацевич В.С., Желтов С.Ю. Структурно-функциональный анализ и синтез глубоких конволюционных нейронных сетей // Компьютерная оптика. 2019. Т.43. №5. С.886-900.
8. Гранская Ю.В. Распознавание эмоций по выражению лица: Автореф. дис. . . . канд. психологических наук по специальности 09.00.01. СПб., 1998.
9. Деревягин Л.А., Макаров В.В., Цурков В.И., Яковлев А.Н. Интеллектуальная система для определения эмоций на аудиозаписи с помощью мел-спектрограмм // Изв. РАН ТИСУ. 2022. №3. С. 116-121.
10. Деревягин Л.А., Макаров В.В., Молчанов А.Ю., Цурков В.И., Яковлев А.Н. Применение нейронных сетей в исследованиях на полиграфе // Изв. РАН ТИСУ. 2022. №4. С. 80-85.

11. Заболеева-Зотова А. В. Развитие системы автоматизированного определения эмоций и возможные сферы применения // Открытое образование. 2011. № 2. С. 59–62.
12. Леонтьев К.А., Панин С.Д., Холодный Ю.И.Оценка результатов тестирования на полиграфе методами регрессионного анализа // Наука и Образование: электронный журнал МГТУ им. Н.Э. Баумана. 2014. №10. С. 230-243.
13. Лурия А.Р. Диагностика следов аффекта. Психология эмоций. Тексты. – М.: Из-во Моск. ун-та, 1984, 288 с.
14. Люсин Д.В. Современные представления об эмоциональном интеллекте // Социальный интеллект: теория, измерение, исследования / Под ред. Д.В. Люсина, Д.В. Ушакова. М.: Изд-во Ин-та психологии РАН, 2004. С. 29–36.
15. Минакова Н.Н., Божич Е.В. Применение методов многомерного анализа данных при обработке полиграмм для изучения биофизических характеристик // Изв. АлтГУ. 2018. №1 (99). С. 34-38.
16. Оглоблин С.И., Молчанов А.Ю. Инструментальная «детекция лжи»: академический курс. Ярославль: Нюанс, 2004. 464 с.
17. Попова И.А., Попова А.А., Соболева Е.Д. Визуализация многомерных наборов данных при помощи алгоритмов снижения пространства признаков PCA и T-SNE// Научно-образовательный журнал для студентов и преподавателей «StudNet». 2020. №11
18. Рюмина Е.В., Карпов А.А. Аналитический обзор методов распознавания эмоций по выражениям лица человека // Научно-технический вестник информационных технологий, механики и оптики. 2020. № 2. С. 163-176.
19. Себряков Г.Г., Визильтер Ю.В. Разработка методики построения специализированных экспертных систем для анализа цифровых изображений в задачах обнаружения и идентификации сложных структурных объектов // Вестн. компьютерных и информационных технологий. 1997. № 3. С. 31.
20. Симонов П.В. Высшая нервная деятельность человека (мотивационно-эмоциональные аспекты). – М.: Наука, 1975. – 175 с.

21. Abdelkader B., Jaziri R., Bernard G. Deep Cascade of Extra Trees // Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). Macau, China, 2019 P. 117-129.
22. Abdelwahab M., Busso C. Incremental Adaptation Using Active Learning for Acoustic Emotion Recognition // IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Toronto. 2017. P. 5160-5164. Doi: 10.1109/ICASSP.2017.7953140.
23. Abdulrahman M., Eleyan A. Facial Expression Recognition Using Support Vector Machines // Proc. 23nd Signal Processing and Communications Applications Conf. (SIU 2015). Malatya, Turkey, 2015. P. 276–279.
24. Addison P.S. The Illustrated Wavelet Transform Handbook // Introductory Theory and Applications in Science, Engineering, Medicine And finance, CRC Press, 2017.
25. Akhiyarov F.R., Derevyagin L.A., Makarov V.V., Tsurkov V.I., Yakovlev A.N. Frame-by-Frame Determination of Emotions in a Video Recording Using Multilayer Neural Networks // Journal of Computer and Systems Sciences International, 2022, Vol. 60, No. 2
26. Anagnostopoulos C. N., Iliou T., Giannoukos I. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011 // Artif. Intell. Rev., 2012, p. 1–23.
27. Antoniadou I., Manson G., Dervilis N., Barszcz T., Staszewski W., Worden K. Use of the teager-kaiser energy operator for condition monitoring of a wind turbine gearbox // International Conference on Noise and Vibration Engineering 2012, ISMA 2012, including USD 2012: International Conference on Uncertainty in Structure Dynamics, 6, pp. 4255–4268.
28. Atlas L., Shamma S.A. Joint acoustic and modulation frequency // EURASIP J. Appl. Sign. Process., 2003, p. 668–675.
29. Ayadi M.E., Kamel M.S., Karray F. Survey on speech emotion recognition: features, classification schemes, and databases // Pattern Recogn, 2011, 44

(3), p. 572–587.

30. Balasubramanian G., Kanagasabai A., Mohan J. Music induced emotion using wavelet packet decomposition an EEG study // Biomed. Signal Process. Control, 2018, 42, 115–128.
31. Bhatnagar S., Ghosal D., Kolekar M. Classification of Fashion Article Images Using Convolutional Neural Networks // Fourth Intern. Conf. on Image Information Processing (ICIIP). Waknaghat, 2017, P. 1-6, Doi: 10.1109/ICIIP.2017.8313740.
32. Bitouk D., Verma R., Nenkova A. Class-level spectral features for emotion recognition // Speech Commun., 2010, 52 (7–8), p. 613–625.
33. Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W., Weiss B. A database of german emotional speech // Proceeding of the INTERSPEECH2005, 2005
34. Busso C., Lee S., Narayanan S. Analysis of emotionally salient aspects of fundamental frequency for emotion detection, audio, speech, and language processing // IEEE Trans., 2009, 17 (4), p. 582–596.
35. Candes E., Demanet L., Donoho D., Ying L. Fast Discrete Curvelet Transforms // Multiscale Modeling And Simulation. 2006. V. 5. №3. P. 861–899.
36. Chang J., Scherer S. Learning Representations of Emotional Speech with Deep Convolutional Generative Adversarial Networks // IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Toronto. 2017. P. 2746-2750. Doi: 10.1109/ICASSP.2017.7952656.
37. Chen R., Zhou Y., Qian Y. Emotion recognition using support vector machine and deep neural network // National Conference on Man-Machine Speech Communication. Springer. 2017. pp. 122–131.
38. Chen, S., Jin, Q. Multi-modal dimensional emotion recognition using recurrent neural networks // Brisbane, Australia. 2015.
39. Chen X., Yuan G., Nie F. Semi–supervised feature // Selection via Rescaled Linear Regression IJCAI. 2017. pp. 1525–1531.
40. Cheveigne A., Kawahara H. A Fundamental Frequency Estimator for

Speech and Music // Ircam-CNRS. Wakayama University, 2002. URL: http://recherche.ircam.fr/equipes/pcm/cheveign/ps/2002_JASA_YIN_proof.pdf

41. Choueiter G.F., Glass J.R. An implementation of rational wavelets and filter design for phonetic classification // Audio, speech, and language processing, IEEE Trans. 2007. 15 (3). p. 939–948.
42. Clavel C., Vasilescu I., Devillers L., Richard G., Ehrette T., Feartype emotion recognition for future audio-based surveillance systems // Speech Commun. 2008. 50. p. 487–503.
43. Cowie R., Cornelius R.R. Describing the emotional states that are expressed in speech // Speech communication. 2003. V. 40(1). P. 5–32
44. Daubechies I. The wavelet transform, time-frequency localization and signal analysis // IEEE Trans. Inf. Theory. 2009. v. 36 (5). p. 961–1005.
45. Davis S., Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences // IEEE Trans. Audio Speech Lang. Process. 1980. v. 28. p. 357–366.
46. Deng J., Xu X., Zhang Z., Frühholz S., Schuller B. Semisupervised Autoencoders for Speech Emotion Recognition // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2018. V.26 (№1). P. 31-43. Doi: 10.1109/TASLP.2017.2759338.
47. Derevyagin L. A., Makarov V. V., Tsurkov V. I., Yakovlev A. N. An Intelligent System for Identifying Emotions on Audio Recordings Using Chalk Spectrograms // Journal of Computer and Systems Sciences International, 2022, Vol. 61, No. 3
48. Derevyagin L. A., Makarov V. V., Tsurkov V. I., Yakovlev A. N. Using neural networks in polygraph research // Journal of Computer and Systems Sciences International, 2022, Vol. 62, No. 4
49. Ekman P. Facial Action Coding System. Palo Alto. USA: Consulting Psychologist Press, 1978.
50. Engberg I.S., Hansen A.V. Documentation of the Danish Emotional Speech

Database (DES) // Department of Communication Technology, Institute of Electronic System, Aalborg University, Denmark, 1996.

51. Eyben F., Batliner A., Schuller B. Towards a standard set of acoustic features for the processing of emotion in speech // Proceedings of Meetings on Acoustics. 2010. volume 9. pp. 1–12.
52. Fonseca-Pinto R. A new tool for nonstationary and nonlinear signals: The hilbert-huang transform in biomedical applications // Biomedical Engineering, Trends in Electronics, Communications and Software. InTech. 2011.
53. Fateri S., Boulgouris N.V., Wilkinson A., Balachandran W., Gan T.H. Frequency-sweep examination for wavelet mode identification in multimodal ultra sonic guided wavelet signal // ultrasonics, ferroelectrics, and frequency control, IEEE Trans. 2014. V. 61 (9). P. 1515–1524.
54. Fragopanagos N., Taylor J.G. Emotion recognition in human computer interaction // Neural Netw. 2005. V. 18 (4). P. 389–405.
55. Gao Z., Guo B., Niu B. Facial Expression Recognition with LBP and ORB Features // Hindawi, Computational Intelligence and Neuroscience, Volume 2021, Article ID 8828245
56. Germain F. The Wavelet Transform Applications // Music Information Retrieval, McGill University, Canada, 2009.
57. Ghanbari Y., Karami-Mollaei M.R. A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets // Speech Commun. 2006. V. 48 (8). P. 927–940.
58. Goel S.G. Speech emotion recognition using EEMD, SVM and ANN. Ph.D. thesis. 2014
59. Gordon N. Essentials of Polygraph and Polygraph Testing. Boca Raton: CRC Press, 2016. 320 p.
60. Greche L., Es-Sbai N., Lavendelis E. Histogram of Oriented Gradient and Multi Layer Feed Forward Neural Network for Facial Expression Identification // Proc. Intern. Conf. on Control, Automation and Diagnosis (ICCAD 2017).

Hammamet, Tunisia, 2017. P. 333–337.

61. Greenwell B., Boehmke B. Hands-On Machine Learning with R. Boca Raton: CRC Press,2019 – 488 p.
62. Grimm M., Kroschel K., Mower E., Narayanan S. Primitives based evaluation and estimation of emotions in speech // Speech Commun. 2007. V. 49. P. 787–800.
63. Guido R.C. Paraconsistent feature engineering // IEEE Signal Process. 2018. V. 36 (1). P. 154–158.
64. Gunn S.R. Support vector machines for classification and regression // ISIS Tech. 1998. V. 14 (1). P. 5–16.
65. Guyon I., Weston J., Barnhill S., Vapnik, V. Gene selection for cancer classification using support vector machines // Mach. Learn. 2002. V. 46 (1–3). P. 389–422.
66. Hariharan M., Sindhu R., Vijean V., Improved binary dragonfly optimization algorithm and wavelet packet based non-linear features for infant cry classification // Comput. Methods Progr. Biomed. 2018. V. 155. P. 39–51.
67. Haque A.F. Frequency analysis and feature extraction of impressive tools // Int. J. Adv. Innovat.Thoughts and Ideas. 2013. V. 2 (2)
68. Hernandez-Matamoros A., Bonarini A., Escamilla-Hernandez E., Nakano-Miyatake M., Perez-Meana H. A Facial Expression Recognition with Automatic Segmentation of Face Regions // Communications in Computer and Information Science. 2015. V. 532. P. 529–540.
69. Hinton G.E., Roweis S.T. Stochastic neighbor embedding // Advances in Neural Information Processing Systems. 2002. V. 15 (2/3). P. 833–840.
70. Huang N.E., Shen Z., Long S.R., Wu M.C., Shih H.H., Zheng Q., Yen N.C., Tung C.C., Liu H.H. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis // Proc. R. Soc. Lond. A Math. Phys. Eng. Sci. 1998. V. 454. P. 903–995.
71. Huang N.E., Shen Z., Long S.R., Wu M.C., Shih H.H., Zheng Q., Yen N.C.,

Tung C.C., Liu H.H. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis // Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences. 1998. V. 454. P. 903–995.

72. Hung C., Lee S. Adaptive Distance-Based Voting Classification // Intern. Conf. on Machine Learning and Cybernetics (ICMLC). Tianjin, China, 2013 P. 1671-1677.
73. Iqtait M., Mohamad F.S., Mamat M. Feature Extraction for Face Recognition Via Active Shape Model (ASM) and Active Appearance Model (AAM) // IOP Conf. Series: Materials Science and Engineering. Tangerang Selatan, Indonesia, 2018. V. 332. P. 1-8.
74. Islam M.T., Shahnaz C., Zhu W.P. Rayleigh modeling of teager energy operated perceptual wavelet packet coefficients for enhancing noisy speech // Speech Commun. 2017. V. 86. P. 64–74.
75. Islam M.T., Shahnaz C., Zhu W.P. Speech enhancement based on student t modeling of teager energy operated perceptual wavelet packet coefficients and a custom thresholding function // IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP). 2015. V. 23 (11). P. 1800–1811.
76. Jumani S.Z., Ali F., Guriro S., Kandhro I.A., Khan A., Zaidi A. Facial Expression Recognition with Histogram of Oriented Gradients Using CNN // Indian J. Science and Technology. 2019. V. 12. N 24. P. 1–8.
77. Kahou S.E., Michalski V., Konda K. Recurrent neural networks for emotion recognition in video // Proceedings of the ACM on International Conference Multimodal Interaction, Seattle, WA, USA. 2015. P. 467–474.
78. Kazemi V., Sullivan J. One millisecond face alignment with an ensemble of regression trees // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA. 2014. P. 1867–1874
79. Kerkeni L., Serrestou Y., Mbarki M., Mahjoub M., Raoof K., Cleder C. Speech emotion recognition: Recurrent neural networks compared to svm and

linear regression // Artificial Neural Networks and Machine Learning. 2017. P. 451–453.

80. Kerkeni L., Serrestou Y., Mbarki M., Raoof K., Mahjoub M.A. A review on speech emotion recognition: Case of pedagogical interaction in classroom // Advanced Technologies for Signal and Image Processing (ATSIP), 2017 International Conference on. IEEE. 2017. P. 1–7.
81. Kerkeni L., Serrestou Y., Mbarki M., Mahjoub M., Raoof K.Speech emotion recognition: Methods and cases study // International Conference on Agents and Artificial Intelligence (ICAART). 2018.
82. Kerkeni L., Serrestou Y., Raoof K., Mbarki M., Mahjoub M.A., Cleder C. Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO // Speech Communication, Volume 114, 2019, P 22-35, ISSN 0167-6393
83. Khaldi K., Boudraa A.O., Komaty A. Speech enhancement using empirical mode decomposition and the teager–kaiser energy operator // J. Acoust. Soc. Am. 2014. V. 135 (1). P. 451–459.
84. Kim J.W., Salamon J., Li P., Bello J.P.CREPE: A Convolutional Representation for Pitch Estimation // Music and Audio Research Laboratory. N. Y.: Center for Urban Science and Progress, New York University. 2018.
85. King D.E. Dlib-ml: a machine learning toolkit // Journal of Machine Learning Research. 2009. V. 10. P. 1755–1758.
86. Knyaz V.A., Matveev I.A., Murynin A.B. Applying Computer Stereovision Algorithms to Study of Correlation Between Face Asymmetry and Human Vision Pathology // Pattern Recognition and Image Analysis. 2009. V. 19 (4). P. 679-686.
87. Kollias D. Analysing Affective Behavior in the First ABAW 2020 Competition // arXiv preprint arXiv:2001.11409. 2020.
88. Kollias D. Deep Affect Prediction in-the-wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond // International J. Computer

Vision (IJCV). Berlin, Germany. 2019. №127. P. 907-929

89. Kollias D. Distribution Matching for Heterogeneous Multi-Task Learning: a Large-scale Face Study // arXiv preprint arXiv:2105.03790, 2020.
90. Kollias D. Face Behavior a la carte: Expressions, Affect and Action Units in a Single Network // arXiv preprint arXiv:1910.11111, 2019.
91. Kollias D., Zafeiriou S. Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace // arXiv preprint arXiv:1910.04855, 2019.
92. Kotti M., Patern F. Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema // Int. J. Speech Technol. 2012. P. 131–150.
93. Kumar A. The optimized wavelet filters for speech compression // Int. J. Speech Technol. 2013. V. 16 (2). P. 171–179.
94. Li L., Zhao Y., Jiang D., Zhang Y., Wang F., Gonzalez I., Valentin E., Sahli H. Hybrid deep neural network-hidden Markov model (DNN-HMM) based speech emotion recognition // Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction. 2013. P. 312–317.
95. Li Q., Zheng J., Tsai A. Robust endpoint detection and energy normalization for real-time speech and speaker recognition // Speech Audio Process. IEEE Trans. 2002. V. 10 (3). P. 146–157.
96. Li X., Zheng X., Zhang D. Emd-teo based speech emotion recognition // Life Syst. Model. Intell. Comput. 2010. P. 180–189.
97. Li X. Speech emotion recognition using novel hht-teo based features //JCP. 2011. V. 6 (5). P. 989–998.
98. Lim W., Jang D., Lee T. Speech emotion recognition using convolutional and recurrent neural networks // Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific. IEEE. 2016. P. 1–4.
99. Liu Z.T., Wu M., Cao W.H., Mao,J.W., Xu,J.P., Tan G.Z. Speech emotion recognition based on feature selection and extreme learning machine decision

tree // Neurocomputing. 2018. V. 273. P. 271–280.

100. Livingstone S.R., Russo F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English // PLoS ONE. 2018. V.13 № 5. С. 1-35.
101. Lucey P. The Extended Cohn-Kanade Dataset (CK+): A Complete Bataset for Action Unit and Emotion-specified Expression // Proc. IEEE CVPR Workshop on Biometrics. - San Francisco: IEEE Computer Society, 2010. P. 94–101.
102. Maaten L.V., Hinton G.E. Visualizing data using t-stochastic neighbor embedding // Journal of Machine Learning Research. 2008. V. 9. P. 2579–2605.
103. Mallat S., Zhong S. Characterization of signals from multiscale edges // IEEE Trans. Pattern Anal. Mach. Intell. 1992. V. 16. P. 710-732.
104. Mallat S.G. A theory for multiresolution signal decomposition: the wavelet representation // IEEE Trans. Pattern Anal. Mach. Intell. 1989. V. 11 (7). P. 674–693.
105. Maragos P., Kaiser J.F., Quatieri T.F. Energy separation in signal modulations with application to speech analysis // IEEE Trans. Sign. Process. 1993. V. 41 (10). P. 3024–3051.
106. Maragos P., Kaiser J.F., Quatieri T.F. On amplitude and frequency demodulation using energy operators // IEEE Trans. Sign. Process. 1993. V. 41 (4). P. 1532–1550.
107. Maragos P., Quatieri T.F., Kaiser J.F. Speech nonlinearities, modulations, and energy operators // Acoustics, Speech, and Signal Processing, ICASSP-91. 1991. P. 421–424.
108. Mauch M., Dixon S. PYIN: A Fundamental Frequency Estimator Using Probabilistic Threshold Distributions. London: Queen Mary University of London, Centre for Digital Music, 2014.
109. Milton A., Roy S.S., Selvi S.T. Svm scheme for speech emotion recognition using mfcc feature // Int. J. Comp. Appl. 2013. V. 69 (9).

110. Mirsamadi S., Barsoum E., Zhang C., Automatic speech emotion recognition using recurrent neural networks with local attention // Acoustics, Speech and Signal Processing (ICASSP). 2017. P. 2227–2231.
111. Muthusamy H., Polat K., Yaacob S. Particle swarm optimization based feature enhancement and feature selection for improved emotion recognition in speech and glottal signals // PloS One. 2015. V. 10 (3).
112. Nayak M., Panigrahi B.S. Advanced signal processing techniques for feature extraction in data mining // Int. J. Comp. Appl. 2011. V. 19 (9). P. 30–37.
113. Nigam S., Singh R., Misra A.K. Efficient Facial Expression Recognition Using Histogram of Oriented Gradients in Wavelet Domain // Multimedia Tools and Applications. 2018. V. 77 (21). P. 28725–28747.
114. Ojala T., Pietikainen M., Harwood D. A comparative study of texture measures with classification based on featured distributions // Pattern Recognition. 1996. V. 29 (1). P. 51–59.
115. Ojala T., Pietikainen M., Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002. V. 24 (7). P. 971–987.
116. Topi M., Timo O., Matti P., Maricor S. Robust texture classification by subsets of local binary pattern // Pattern Recognition. 2000. V. 7 (3). P. 335–338.
117. Pan Y., Shen P., Shen L. Speech emotion recognition using support vector machine // Int. J. Smart Home. 2012. V. 6 (2). P. 101–108.
118. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E. Scikit-learn: machine learning in python. J. Mach. Learn. 2012. Res. 12. P. 2825–2830.
119. Porcu V. Python for Data Mining Quick Syntax Reference. N.Y.: Apress Media LLC, 2018. 260 p.
120. Potamianos A., Maragos P. A comparison of the energy operator and the hilbert transform approach to signal and speech demodulation // Sign. Process. 1994.

V. 37 (1). P. 95–120.

121. Potamianos A., Maragos P. Speech analysis and synthesis using an am–fm modulation model // Speech Commun. 1999. V. 28 (3). P. 195–209.
122. Prasomphan S. Improvement of speech emotion recognition with neural network classifier by using speech spectrogram // Systems, Signals and Image Processing (IWSSIP). 2015 P. 73–76.
123. Pudil P., Novovicova J., Kittler J. Floating search methods in feature selection // Pattern Recogn. Lett. 1994. V. 15 (11). P. 1119–1125.
124. Rao K.D., Swamy M.N. Discrete wavelet transforms // Digital Signal Processing, Springer, Singapore. 2018. P. 619–691.
125. Saratxaga I., Navas E., Hernaez I., Luengo I. Designing and recording an emotional speech database for corpus based synthesis in basque // Proc. of fifth international conference on Language Resources and Evaluation (LREC). 2006. P. 2126–2129.
126. Schuller B., Batliner A., Steidl S., Seppi D. Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge // Speech Commun. 2011. V. 53 (9/10). P. 1062–1087.
127. Sethu V., Ambikairajah E., Epps J. Empirical mode decomposition based weighted frequency feature for speech-based emotion classification // Acoustics, Speech and Signal Processing. 2008. P. 5017–5020.
128. Shahnaz C., Sultana S., Fattah S.A., Rafi R.M., Ahmmed I., Zhu W.P., Ahmad M.O. Emotion recognition based on emd-wavelet analysis of speech signals // Digital Signal Processing (DSP). 2015. P. 307–310.
129. Shan C., Gong S., McOwan P.W. Facial expression recognition based on local binary patterns: a comprehensive study // Image and Vision Computing.2009. V. 27 (6) P. 803–816.
130. Sharma R., Vignolo L., Schlotthauer G., Colominas M.A., Rufiner H.L., Prasanna S. Empirical mode decomposition for adaptive am-fm analysis of speech: a review // Speech Commun. 2017.

131. Silva J., Narayanan S.S. Discriminative wavelet packet filter bank selection for pattern recognition // IEEE Trans. Signal Process. 2009. V. 57 (5). P. 1796–1810.

132. Sonmez Y.U., Varol A. New Trends in Speech Emotion Recognition // 7th Intern. Sympos. on Digital Forensics and Security (ISDFS).Barcelos 2019. P. 1-7. Doi: 10.1109/ISDFS.2019.8757528.

133. Sree G.D., Chandrasekhar P., Venkateshulu B. Svm based speech emotion recognition compared with gmm-ubm and nn // Int. J. Eng. Sci. 2016. P. 3293.

134. Stuhlsatz A., Meyer C., Eyben F., Zielke T., Meier G., Schuller B. Deep neural networks for acoustic emotion recognition: raising the benchmarks // Pro ceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). 2011. P. 5688–5691.

135. Swain M., Routray A., Kabisatpathy P. Databases, features and classifiers for speech emotion recognition: a review // Int. J. Speech Technol. 2018. V. 21 (1). P. 93–120.

136. Talegaonkar I., Joshi K., Valunj S., Kohok R., Kulkarni A. Real Time Facial Expression Recognition Using Deep Learning // Proc. of Intern. Conf. on Communication and Information Processing (ICCIP). 2019. URL: https://ssrn.com/abstract=3421486.

137. Tang J., Alelyani S., Liu H. Feature selection for classification: a review // Data Classificat. Algorith. Appl. 2014. P. 37.

138. Tripathi A., Pandey S. Efficient Facial Expression Recognition System Based on Geometric Features Using Neural Network // Lecture Notes in Networks and Systems. 2018. V. 10. P. 181–190.

139. Varchenko N.N., Gankin K.A., Matveev I.A. Using Binocular Pupillometry Method for Evaluating Functional State of Person // Sports Technology. 2015. V.8. P.67-75.

140. Varma S., Shinde M., Chavan S.S. Analysis of PCA and LDA Features for

Facial Expression Recognition Using SVM and HMM Classifiers // Techno-Societal 2018: Proc. 2nd Intern. Conf. on Advanced Technologies for Societal Applications. Berlin, Germany. 2019. V. 1. P. 109– 119.

141. Vasquez-Correa J.C. , T. Arias-Vergara, J.R. Orozco-Arroyave, J.F. Vargas-Bonilla, E. Noeth, Wavelet-based time-frequency representations for automatic recognition of emotions from speech, in: Proceedings of the ITG Symposium, VDE Speech Communication. 2016. V. 12. P. 1–5.
142. Viola P. Jones M.J. Robust real-time face detection // International Journal of Computer Vision. 2004. V. 57 (2). P. 137–154.
143. Walecki R., Rudovic O., Pavlovic V., Schuller B. Deep structured learning for facial expression intensity estimation // Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA. 2017. V. 259. P. 143–154.
144. Wang C.C., Kang Y. Feature extraction techniques of non-stationary signals for fault diagnosis in machinery systems // J. Sign. Inform. Process. 2012. V. 3 (01). P. 16.
145. Wang K., An N., Li B., Zhang Y. Speech emotion recognition using fourier parameters affective computing // IEEE Trans. 2015. V. 6 (1). P. 69–75.
146. Wang K., An N., Li L. Speech emotion recognition based on wavelet packet coefficient model // Proceeding of the 2014 Ninth International Symposium on Chinese Spoken Language Processing (ISCSLP) IEEE. 2014.
147. Wang K., Zhang Q.L., Liao S.Y. A database of elderly emotional speech // Proceeding of the International Symposium on Signal Processing Biomedical Engineering, and Informatics (SPBEI). 2014. P. 549–553.
148. Wang K., Su G., Liu L., Wang S. Wavelet packet analysis for speaker-independent emotion recognition // Neurocomputing. 2020. V. 398. P. 257-264.
149. Wu S. Recognition of human emotion in speech using modulation spectral features and support vector machines. Ph.D. thesis. 2009.
150. Wu S., Falk T.H., Chan W.Y. Automatic speech emotion recognition using

modulation spectral features // Speech Commun. 2011. V. 53 (5). P. 768–785.

151. Wu Z., Huang N.E. Ensemble empirical mode decomposition: a noise-assisted data analysis method // Adv. Adapt. Data Anal. 2009. V. 1. P. 1–41.
152. Zao L., Cavalcante D., Coelho R. Time-frequency feature and AMS-GMM mask for acoustic emotion classification // IEEE Signal Process. Lett. 2014. V. 21 (5). P. 620–624.
153. Zhang L., Wang S., Liu B. Deep learning for sentiment analysis: a survey // arXiv preprint arXiv:1801.07883. 2018.
154. Zhang W., Zhao D., Chai Z., Yang L.T., Liu X., Gong F., Yang S Deep learning and svm-based emotion recognition from chinese speech for smart affective services // Softw. Pract. Exp. 2017. V. 47 (8). P. 1127–1138.
155. Zhang Y.D., Yang Z.J., Lu H.M. Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation // IEEE Access. 20216. V. 4. P. 8375–8385.
156. Zhao K., Chu W.S., Zhang H. Deep region and multi-label learning for facial action unit detection // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA. 2016. P. 3391–3399.
157. Zheng B.S., Khairunizam W., Murugappan S.A. Effectiveness of tuned q-factor wavelet transform in emotion recognition among left-brain damaged stroke patients // Int. J. Simul.-Syst. Sci. Technol. 2018. V. 19 (3).
158. Zhuang N., Zeng Y., Tong L., Zhang C., Zhang H., Yan B. Emotion recognition from eeg signals using multidimensional information in emd domain // BioMed Res. Int. 2017.
159. Zhao J., Mao X., Zhang J. Learning Deep Facial Expression Features from Image and Optical Flow Sequences Using 3D CNN // Visual Computer. 2018. V. 34. №10. P. 1461–1475.

Список иллюстративного материала