Федеральное государственное бюджетное учреждение науки
«Санкт-Петербургский Федеральный исследовательский центр
Российской академии наук»
(СПб ФИЦ РАН)

На правах рукописи

Величко Алёна Николаевна

Методы и программная система интегрального анализа деструктивных паралингвистических явлений в разговорной речи

Специальность 2.3.5 – Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель
д.т.н., профессор
Карпов Алексей Анатольевич

Санкт-Петербург – 2023

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность темы диссертации.

Деструктивное поведение пользователей при коммуникации в сети Интернет разрушительно влияет как на самого человека, так и на других. В связи с этим актуально выявление деструктивных (девиантных, агрессивных и враждебных) действий и обеспечение психологического комфорта пользователей социальных сетей (Станкевич М., Huang Z.).

Деструктивное поведение может проявляться как вербально (словами) или невербально (поведением). Объектом деструктивного поведения чаще всего являются эмоциональное и физическое состояния субъекта, предметы материального мира, социальные связи, коммуникация между людьми, их отношения и т.д. Под девиантным поведением чаще всего подразумевается поведение личности, которое отклоняется от общепринятого, устоявшихся и общественных норм (Майсак Н.).

В диссертационной работе рассматриваются различные деструктивные явления в поведении человека: передача ложных речевых сообщений (преднамеренная ложь/обман), депрессивные состояния, проявления агрессии к другим людям.

Существующие на данный момент автоматические программные решения по определению рассматриваемых деструктивных явлений в разговорной речи имеют следующие недостатки: 1) низкая эффективность распознавания явлений, 2) использование сложных нейросетевых архитектур, требовательных к вычислительным ресурсам, 3) большое время обучения моделей (обучение некоторых моделей может доходить до нескольких суток и даже недель), 4) отсутствие программных решений, анализирующих рассматриваемые деструктивные явления в совокупности. Таким образом, актуальна разработка программной системы, которая могла бы при низких требованиях к вычислительным ресурсам и малом количестве обучающих данных эффективно определять рассматриваемые деструктивные паралингвистические явления, в том числе с учетом взаимозависимостей между ними.

Степень разработанности темы.

Автоматическое определение деструктивных паралингвистических явлений в разговорной речи является относительно новой областью, но уже существуют многочисленные работы, представленные на конференциях, семинарах и соревнованиях по компьютерной паралингвистике. Такие российские ученые как Матвеев Ю.Н., Савченко А.В., Мельников С.Ю., Шуранов Е.В., Лякso Е.Е., Потапова Р.К., Комалова Л.Р., Мещеряков Р.В., Костюченко Е.Ю. и др., а также ряд зарубежных ученых, включая Schuller B., Batliner A., Rigoll G., Eyben F., Hirschberg J., Lefter I., Kaya H., Salah A.A., Minker W., Levitan S.I. и др. занимаются анализом паралингвистических аспектов в разговорной речи, в т.ч. задачей определения деструктивных явлений, негативных эмоций и аффективных состояний в речи. Однако известные подходы имеют ряд ограничений: недостаток обучающих данных в виду сложностей при записи речевых корпусов, содержащих рассматриваемые паралингвистические явления; дисбаланс данных для обучения и оценивания, который является естественным из-за того, что рассматриваемые паралингвистические явления не проявляются так же часто, как нейтральное состояние и т.д.

Цель диссертационной работы: повышение эффективности автоматического определения деструктивных паралингвистических явлений в разговорной речи.

Цель диссертационной работы предусматривает выполнение следующих задач:

1. Разработка новых методов автоматического определения различных деструктивных паралингвистических явлений в разговорной речи.
2. Разработка методики интегрального оценивания степени выраженности деструктивных паралингвистических явлений в разговорной речи.
3. Разработка программной системы интегрального анализа деструктивных паралингвистических явлений в разговорной речи и проведение экспериментальных исследований разработанных методов и программной системы

интегрального анализа деструктивных паралингвистических явлений в разговорной речи.

Важность и значимость решаемой задачи обусловлены возможностью применения фундаментальных результатов исследователями деструктивных паралингвистических явлений в разговорной речи, а также специалистами в области психологии для автоматизации первичного обследования пациентов путем бесконтактного определения деструктивных паралингвистических явлений в речи и предотвращения/уменьшения негативных последствий этих явлений.

Объектом исследования являются характеристики деструктивных паралингвистических явлений в разговорной речи.

Предметом исследования являются методы, модели и системы автоматического определения деструктивных паралингвистических явлений в разговорной речи.

Научная новизна исследования заключается в том, что:

1. Предложен комплекс методов анализа речевого сигнала для определения деструктивных паралингвистических явлений в разговорной речи, отличающийся использованием оригинальных наборов анализируемых акустических признаков и применением новых многоуровневых методов (для определения ложности/истинности и агрессии в разговорной речи), а также нейросетевого классификатора для табличных данных (для определения депрессии в разговорной речи).

2. Предложена методика интегрального оценивания степени выраженности деструктивных паралингвистических явлений в речевом сигнале диктора, отличающаяся использованием информации о взаимозависимостях между ложью, агрессией и депрессией для вычисления оценки степени выраженности рассматриваемых явлений в речи диктора.

3. Предложена архитектура программной системы интегрального анализа деструктивных паралингвистических явлений в разговорной речи, отличающаяся возможностью одновременного комплексного определения лжи, агрессии и

депрессии в разговорной речи с использованием предложенного комплекса методов и методики интегрального оценивания.

Теоретическая и практическая значимость работы. Теоретическая значимость заключается в разработке комплекса методов и новой методики для определения деструктивных паралингвистических явлений в разговорной речи. Разработанный комплекс методов предлагает новый подход к решению задачи эффективного определения деструктивного поведения человека по его речи. Он, прежде всего, ориентирован на универсальность, поэтому рассматривает несколько деструктивных паралингвистических явлений в разговорной речи, которые могут быть использованы как по отдельности (как самостоятельные средства для определения каждого рассматриваемого паралингвистического явления), так и в совокупности, в комплексном подходе, который учитывает взаимозависимости между рассматриваемыми паралингвистическими явлениями.

С практической точки зрения, разработанная программная система интегрального анализа деструктивных паралингвистических явлений в разговорной речи может быть использована как самостоятельно, так и в качестве системы комплексного анализа и распознавания многомодальной информации, полученной от человека. Такая система сможет учитывать не только аудио-, но и видеоинформацию, а также текстовые транскрипции речи, что может позволить улучшить результаты распознавания деструктивных паралингвистических явлений.

Методология и методы исследования. Для решения поставленных задач в работе используются и совершенствуются методы компьютерной паралингвистики, машинного обучения, глубокого обучения и искусственного интеллекта. В программной реализации системы использовались методы и алгоритмы, реализованные в открытых программных библиотеках Keras, TensorFlow, Scikit-learn, Catboost, XGBoost, LightGBM, TabNet, OpenSMILE и т.д.

Положениями, выносимыми на защиту, являются:

1. Комплекс методов анализа речевого сигнала на основе оригинальных наборов акустических признаков, новых многоуровневых методов и нейросетевого классификатора для табличных данных.

2. Методика интегрального оценивания степени выраженности деструктивных паралингвистических явлений в разговорной речи диктора.

3. Архитектура программной системы интегрального анализа деструктивных паралингвистических явлений в разговорной речи.

Соответствие диссертации научной специальности. Представленные результаты соответствуют специальности 2.3.5 – Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей.

Степень достоверности результатов диссертации обеспечивается посредством проведения аналитического обзора современных исследований и методов паралингвистического анализа речи для определения деструктивных явлений, машинного и глубокого обучения; подтверждается согласованностью полученных результатов, успешной апробацией программной системы интегрального анализа деструктивных паралингвистических явлений в речевом сигнале, а также выступлениями с докладами на международных и российских научных конференциях, публикациями результатов исследований в ведущих рецензируемых изданиях.

Апробация результатов работы. Основные результаты работы докладывались и обсуждались на следующих конференциях:

1. Информационные технологии в управлении (ИТУ-2018), г. Санкт-Петербург, 2018.

2. 20th International Conference on Speech and Computer SPECOM-2018, Leipzig, Germany, 2018.

3. 8-й междисциплинарный семинар «Анализ разговорной русской речи» (АРЗ-2019), г. Санкт-Петербург, 2019.

4. Intelligent Distributed Computing XIII (IDC 2019), г. Санкт-Петербург, 2019.

5. III международная конференция по инженерной и прикладной лингвистике «Пиотровские чтения 2019» (R. Piotrowski's Readings 2019), г. Санкт-Петербург, 2019.

6. International Conference on Computational Linguistics and Intellectual Technologies “Dialogue 2021”, г. Москва, 2021.

7. 23rd International Conference INTERSPEECH-2022, Incheon, Korea, 2022.

Результаты исследования были использованы в следующих проектах:

1. Разработка и исследование автоматической системы для выявления деструктивных паралингвистических явлений в разговорной речи, РФФИ № 20-37-90144-Аспиранты (Величко А.Н.), руководитель Карпов А.А., 2020-2022 гг.

2. Разработка методов и программных средств оценки ложности передаваемых речевых сообщений, РФФИ № 16-37-60085-мол_а_дк, руководитель Будков В.Ю., 2016-2019 гг.

3. Разработка и исследование интеллектуальной системы для комплексного паралингвистического анализа речи, РНФ № 18-11-00145, руководитель Карпов А.А., 2018-2020 гг.

4. Автоматическое бимодальное распознавание естественных эмоций в русской речи, РФФИ № 18-07-01407-а, руководитель Карпов А.А., 2018-2020 гг.

5. Интеллектуальная система многомодального распознавания аффективных состояний человека, РНФ № 22-11-00321, руководитель Карпов А.А., 2022-2024 гг.

Публикации. По результатам выполнения диссертационного исследования опубликовано 14 печатных работ (см. приложение А), включая 4 публикации в журналах из перечня рецензируемых научных изданий ВАК, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, 7 публикаций в изданиях, индексируемых в WoS/Scopus, 4 свидетельства о регистрации программ для ЭВМ в Роспатенте (см. приложение Б).

Личный вклад. Основные научные положения, теоретические выводы и практические решения, результаты тестирования сформулированы и изложены автором самостоятельно.

Структура и объем диссертационной работы. Диссертационная работа включает введение, три главы, заключение, список использованных источников (143 наименований) и три приложения. Основной текст изложен на 136 страницах машинописного текста, включая 16 рисунков и 15 таблиц.

В первой главе приводится аналитический обзор текущего состояния исследований в области компьютерной паралингвистики и рассматриваемых паралингвистических явлений, краткое описание существующих исследовательских данных, имеющиеся ограничения в разработке систем автоматического определения деструктивных паралингвистических явлений в разговорной речи, возможные пути решения и устранения этих ограничений, а также актуальные требования при разработке таких систем.

Во второй главе приводится описание и исследование методов вычисления акустических признаков, машинного и глубокого обучения, которые используются при разработке систем автоматического определения деструктивных паралингвистических явлений в речи, и приводится обоснование их выбора для программной реализации, дается подробное описание предложенных методов определения деструктивных паралингвистических явлений в разговорной речи. Приведено формальное описание методики интегрального оценивания степени выраженности деструктивных паралингвистических явлений в речи диктора.

В третьей главе описывается архитектура программной системы интегрального анализа деструктивных паралингвистических явлений в разговорной речи DesBDet, использованные открытые программные библиотеки и графический пользовательский интерфейс. Приводится подробная информация об исследовательских данных, показатели оценивания эффективности работы предложенных методов, результаты экспериментальных исследований и сравнение предложенных методов с аналогами, известными в литературе.

1 АНАЛИТИЧЕСКИЙ ОБЗОР МЕТОДОВ И СИСТЕМ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ ДЕСТРУКТИВНЫХ ПАРАЛИНГВИСТИЧЕСКИХ ЯВЛЕНИЙ В РАЗГОВОРНОЙ РЕЧИ

В данной главе приводится аналитический обзор текущего состояния исследований и разработок в области компьютерной паралингвистики, краткое описание корпусов, содержащих рассматриваемые деструктивные паралингвистические явления в разговорной речи, возможные пути решения и обхода ограничений при разработке автоматических систем паралингвистического анализа речи, а также актуальные требования к разработке таких систем. Кроме того, в главе приводятся определения рассматриваемых деструктивных паралингвистических явлений: ложь, депрессия, агрессия, их характеристики, существующие подходы определения этих паралингвистических явлений и описание речевых данных, которые используются для разработки таких подходов.

1.1 Систематизация деструктивных паралингвистических явлений

Областью паралингвистики является изучение различных невербальных аспектов в речи и коммуникации человека (например, интонации, эмоции, особенности произношения и параметров голоса диктора, его психофизиологические состояния, отражающиеся в устной речи). Компьютерная паралингвистика в свою очередь использует автоматизированные средства для усовершенствования систем анализа паралингвистических явлений в речи человека. Если рассматривать чистую речь без пауз, в среднем человек говорит 10-20 минут в день, при этом, на долю вербальной информации приходится всего около 7% от общего количества информации, передаваемой в процессе межличностной коммуникации [1]. Невербальная информация может быть передана по следующим каналам коммуникации: акустический; паралингвистический (тембр голоса, громкость голоса, ритмы речи, дикция, интонация); экстралингвистический (темп речи, паузы речи, неречевые звуки, особенности произношения); визуальный (мимика, кожная реакция, жесты, поза, взгляд, межличностное пространство); символический (внешний вид); тактильный

(прикосновения); ольфакторный (запахи). Физиологическое состояние человека очень тесно связано с его эмоциональным состоянием. Рассматриваемые паралингвистические исследования основываются на так называемом эффекте Липпольда [2], который заключается в том, что все мышцы человека, а также голосовые связки подвержены микроколебаниям с частотой 8-12 Гц. Чем выше частота этих колебаний (от 10 до 12 Гц), тем более уверенно можно сказать, что человек находится в неспокойном состоянии (спокойное состояние характеризуется частотами не более 10 Гц).

Аффект (т.е. проявленная эмоция) может как усилить утверждение, выраженное вербально, так и отрицать его. Выделяются 4 основных способа изменения невербального поведения: минимизация, нейтрализация, преувеличение, замещение эмоций и действий. Минимизация является попыткой подавления внешнего всплеска сильных переживаний. Преувеличение служит для попытки повлиять на окружающих. Нейтрализация – это попытка сокрытия эмоций при помощи спокойного состояния. Замещением называется способ сокрытия истинных переживаний подменой эмоций.

Для паралингвистики при выявлении невербальных характеристик в речи человека голосовые характеристики являются более важными, чем слова. В этом случае наиболее распространенными признаками являются просодические: частота основного тона (ЧОТ), форманты, темп речи, паузы, интонация и т.п. Также следует обращать внимание на слишком короткие и слишком длинные паузы в процессе речи в случае, если они встречаются часто.

Наиболее изученным признаком проявления эмоций посредством голоса является повышение частоты основного тона. Почти 70% экспериментов показали, что у людей, испытывающих подавленное настроение, высота голоса возрастает [3]. Также есть свидетельства того, что высота голоса падает при подавленном настроении. Скрыть эмоциональные изменения голоса нелегко.

На основе матрицы социальных девиаций поведение человека можно разделить на две большие категории: конструктивное и деструктивное. В свою очередь, деструктивное поведение можно разделить на аутодеструктивное

(саморазрушительное) и внешнедеструктивное (разрушение направлено вовне). Аутодеструктивное поведение включает в себя суицидальное (парасуицидальное и суицид) и аддиктивное (химическая и нехимическая зависимость) поведение, а внешнедеструктивное поведение – коммуникативные девиации и противоправные действия (делинквентное или предпреступное поведение, административные и криминальные или преступные правонарушения) [4]. Иерархическая систематизация представлена в виде дерева на рисунке 1, на котором также обозначено место деструктивных явлений, в том числе, рассматриваемых в данной работе. Конечные узлы дерева на рисунке 1 зависят от социальной одобряемости: социально одобряемое и просоциальное поведение, социально нейтральное поведение, социально неодобряемое поведение (асоциальное, антисоциальное). Ложь, депрессия и агрессия относятся к деструктивным паралингвистическим явлениям по следующим причинам:

1. Ложь может быть отнесена как к коммуникативным девиациям (проявление лживости, хитрости), так и к противоправному поведению (административные нарушения – мелкое хулиганство и воровство; делинквентное поведение – лживость, мелкое воровство; криминальное поведение – преступления корыстной направленности, например, махинации).

2. Депрессия может являться первопричиной аутодеструктивного поведения (суицидальное поведение – суицидальные тенденции и завершенный суицид, экстремальные хобби и профессиональная деятельность, различные модификации тела; химическая зависимость при аддиктивном поведении – употребление спиртных напитков, табакокурение, употребление наркотиков; нехимическая зависимость при аддиктивном поведении – трудоголизм, кибераддикции, нарушения пищевого поведения [4]). Также при депрессии могут проявляться признаки коммуникативных девиаций (вегетативность, неэстетичный имидж, нигилизм, активный отказ от жизни в «объективной реальности» и др.) и даже административных правонарушений (нежелание решать личные, семейные и производственные проблемы, уклонение от гражданского долга и обязанностей).

3. Агрессия может быть разделена на две подкатегории:

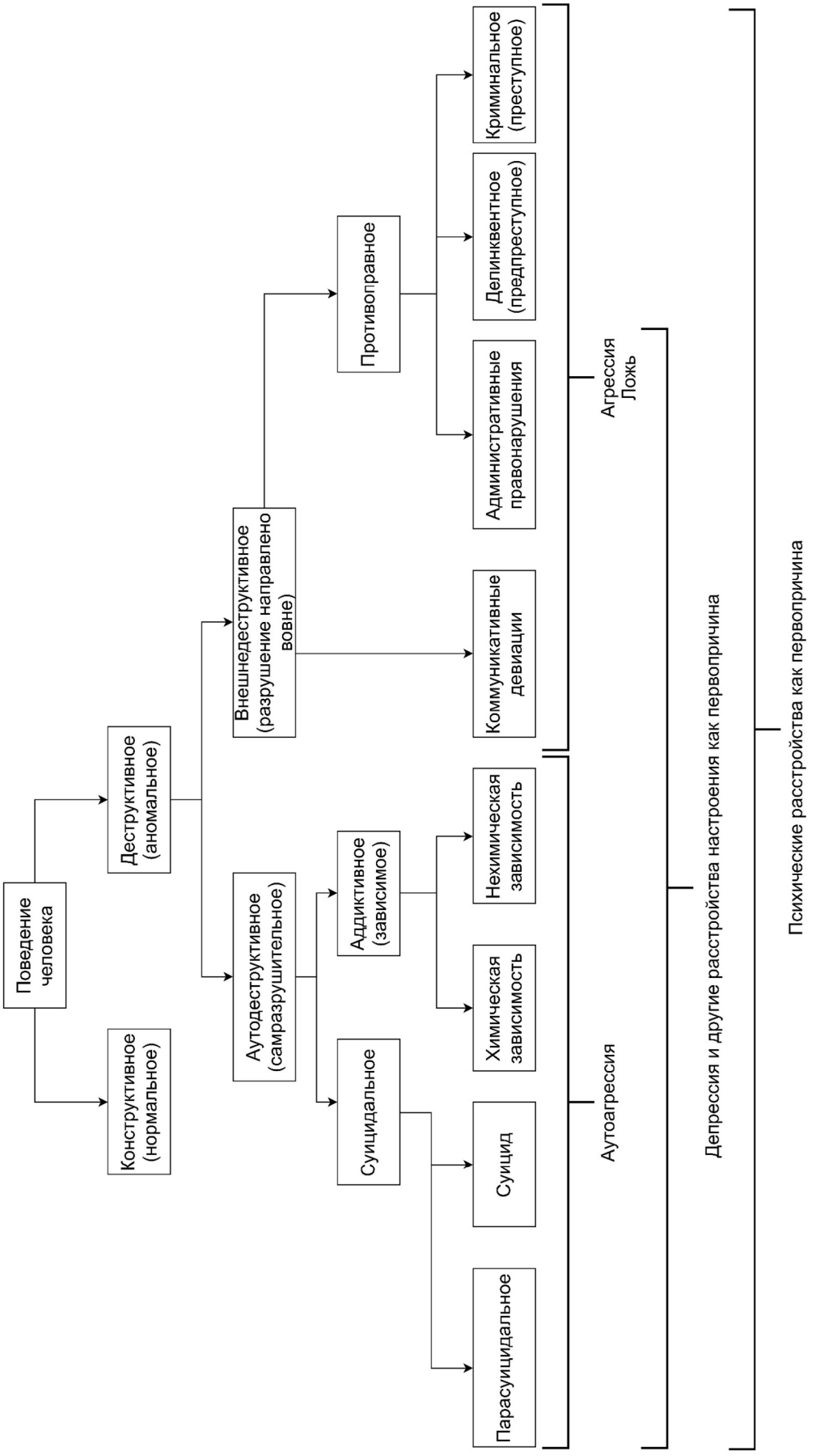

Рисунок 1 – Иерархическая систематизация деструктивных явлений на основе матрицы социальных девиаций (по матрице [4])

а. Аутоагрессия, направленная на себя, может быть одним из симптомов депрессии, расстройств настроения в целом или психических расстройств, что может проявляться в саморазрушительном поведении, как в аддиктивном, так и в суицидальном его аспектах.

б. Агрессия, направленная вовне может относиться как к коммуникативным девиациям (нарушение этикета, сквернословие, проявление жестокости, оппозиционность), так и к противоправному поведению (делинквентное поведение – агрессивные действия и убийства с целью самообороны и самозащиты, подростковые реакции оппозиции, вандализм, агрессивность, враждебность к окружающим, жестокость к младшим и животным; криминальное поведение – убийство во время войны, вендетта в некоторых современных государствах, инфантинацид, преступления против личности и группы агрессивной направленности [4]).

Чтобы избежать стигматизации психических расстройств, стоит отметить, что указанные на рисунке 1 область «Психические расстройства как первопричина» и «Депрессия как первопричина» могут как являться первопричиной отмеченных деструктивных явлений, так и не являться ею. Т.е. указанные деструктивные явления не обязательно могут проявляться под действием какого-либо психического заболевания или расстройства настроения (в т.ч. депрессии). Область «Психические расстройства как первопричина», равно как и область «Аутоагрессия» не рассматриваются далее в работе, т.к. имеющиеся корпуса содержат исключительно агрессию, направленную вовне, а также невозможно рассмотреть все психические заболевания в одной работе (даже в случае с расстройствами настроения в данной работе было выбрано только одно из них, депрессия). На рисунке видно, что некоторые конечные узлы дерева могут относиться к нескольким деструктивным явлениям, что указывает на комплексность природы этих явлений и может означать корреляцию между ними (которая подтверждается рядом теоретических и практических работ, рассмотренных далее в диссертации).

При разработке комплексной архитектуры программной системы для определения деструктивных паралингвистических явлений в речи стоит учитывать возможную корреляцию между рассматриваемыми явлениями. К примеру, в работе [5] на основе исследований выявлено, что, в соответствии с выдвинутыми авторами гипотезами, гнев и депрессия имеют тесную связь; склонность к размышлениям в значительной степени связана как с гневом, так и с депрессией. Таким образом, поведение, связанное со склонностью к размышлениям, может помочь объяснить, как депрессия связана с гневом. В работе [6] авторы указывают, что связи между гневом и депрессией образуют сложную сеть, а при лечении пациентов с депрессией часто оказывается полезным явное или скрытое обращение с гневом. В клинической реальности относительные объемы проявления гнева и обучения экспрессивному контролю, необходимые пациентам с депрессией, могут различаться для разных типов пациентов, разных фаз депрессивного расстройства и разных фаз терапии. Связь между гневом и депрессивным аффектом также изучалась как с помощью межиндивидуального, так и внутрииндивидуального анализа в работе [7], где было выявлено, что тенденция приписывать причину чувства гнева собственным действиям положительно связана с депрессивным аффектом, а тенденция подавлять выражение гнева положительно связана с уровнем депрессии.

Связь гнева, тревоги, депрессии и негативных эмоций рассматривалась в работе [8]. Результаты исследования показывают, что гнев, тревога, депрессия и негативные эмоции сильно коррелируют друг с другом; скорректированные корреляции этих переменных со стрессорами на рабочем месте и последствиями сходные (хотя также существовали некоторые различия); эти переменные в присутствии друг друга не объясняли уникальную дисперсию в некоторых коррелятах исследования. В другом похожем исследовании [9] авторы попытались определить взаимосвязь между ПТСР, депрессией, враждебностью, гневом, словесной и физической агрессией у вернувшихся ветеранов войн. В результате оценки 195 участников было выявлено, что симптомы депрессии частично объясняют связь между посттравматическим стрессовым расстройством (ПТСР),

вербальной и физической агрессией по отношению к другим объектам и самонаправленной физической агрессией. При этом, гнев и хроническая склонность к возбуждению гнева частично объясняют связь между посттравматическим стрессовым расстройством, вербальной и физической агрессией по отношению к объектам и другим людям. В работе [10] была построена многофакторная модель, которая включает в себя возраст, пол, симптомы депрессии, злой темперамент и внешний гнев было выявлено, что только злой темперамент значительно предсказывает суицидальные мысли независимо от симптомов депрессии, а связь между шкалами гнева и суицидальными мыслями не зависит от пола или симптомов депрессии и не опосредована межличностными трудностями.

Связь между агрессией и ложью рассматривалась в работе [11]. Обнаружено, что гнев приводит к более явному проявлению имплицитных установок в отличие от нейтрального или грустного настроения. Автор работы считает, что гнев имеет сходство со счастьем, которое вызывает аналогичный эффект, поскольку обе эти эмоции повышают уверенность человека в себе. Люди, которые уверены в своих эмоциональных состояниях, с большей вероятностью выскажут свое истинное мнение, свои внутренние чувства, в отличие от тех, кто менее уверен в себе. Когда человек сомневается в своих скрытых установках и менее уверен в себе, как это бывает в моменты грусти, он с меньшей вероятностью открыто проявит свои подлинные установки. Поскольку эмоции могут влиять на этическое поведение, в этом исследовании был проведен ряд экспериментов, в ходе которых авторы выявили, что случайный гнев способствует неэтичному поведению, потому что злые люди становятся менее чуткими, преследуя свои личные интересы [12].

1.2 Анализ исследований в области автоматического определения ложной и истинной информации в речевых сообщениях

Под ложью подразумевается преднамеренный акт введения собеседника в заблуждение путем передачи неверной или вводящей в заблуждение информации [2]. Ложная информация бывает как преднамеренная (дезинформация), так и

непреднамеренная (заблуждение). Ложь также может быть направлена как вовне, так и на самого человека (самоориентирована). Часто приводятся различия между явной ложью (ложь, диаметрально противоположна истине), преувеличением (сообщаемая информация или факты превосходят истинные данные) и тонкой ложью (сообщение практически является истинным, но составлено для заблуждения, уклонения от ответа или для умышленного опущения деталей) [13].

Несмотря на то, что использование полиграфа для определения ложной и истинной информации, является самой известной практикой, эта методика не является оптимальной, т.к. для проведения процедуры исследования с использованием полиграфа предъявляются жесткие требования к условиям работы с аппаратом, к месту проведения исследований (комфортный для участников температурный режим, влажность, шумоизоляция и иные ограничения), а также к самому испытуемому (например, наличие добровольного согласия на проведение исследований, отсутствие заболеваний соматических, психических расстройств и т.д.). Таким образом, появляется потребность в бесконтактных методах, не подразумевающих физический контакт с испытуемым, а исследующих речевую активность и невербальные сигналы, исходящие от испытуемого. У таких методов, однако, есть ограничения, которые делают их разработку комплексной задачей, в частности эти ограничения касаются разработки методов анализа звукового сигнала: индивидуальные особенности диктора (дефекты слуха, речи и пр.), наличие шумов при записи, лексическая неоднозначность языка и др.

Некоторые исследователи предполагают, что ложная/истинная информация может быть определена при помощи эмоциональных признаков, поскольку эмоциональная насыщенность речи проявляется в ее характеристиках: темп, тембр и громкость [14, 15, 16]. Вялая и лишенная эмоций речь свидетельствует о безразличии к излагаемой информации. Напротив, высокий темп речи может говорить об изменении эмоционального состояния (возбуждении) участника, о том, что тема разговора волнует его, поскольку он старается сказать больше из необходимости убедить собеседника в своей правоте. Отношение к теме разговора также выражается такой характеристикой голоса, как тембр. Чтобы подробнее

узнать о личности говорящего, можно обратить внимание на следующие признаки его речи как [17]:

– лексические (говорят об образовании, социальном статусе и возрасте);

– грамматические (свидетельствуют о том, насколько грамотен и образован человек);

– синтаксические (могут считаться признаком недостаточных навыков построения фраз или чрезмерного возбуждения);

– стилистические (являются отражением навыков использования речи при общении).

В работе [18] исследованы характеристики речи 19 дикторов, которые произносили ложные высказывания, с использованием таких характеристик как: скорость речи, время начала ответа, длительность и частота пауз. Выявлено, что при произнесении лжи заметно увеличивается темп речи, уменьшаются время начала ответа и длительность хезитаций.

Основная часть работ по автоматическому определению ложной и истинной информации в речевых сообщениях представлена на соревнованиях по компьютерной паралингвистике INTERSPEECH ComParE (http://www.compare.openaudio.eu). Эта тематика впервые была представлена в рамках соревнований в 2016 году. В качестве показателя качества работы систем используется показатель невзвешенной средней полноты UAR (Unweighted Average Recall). Детальное описание показателей оценки качества работы систем даны в главе 3.4. Участникам предоставлялись аудиоданные для обучения и отладки систем (речевой корпус Deceptive Speech Dataset, DSD), набор акустических признаков ComParE_2016, а также базовый результат системы, разработанной организаторами (UAR = 68,3%). В соревнованиях 2016 года участвовали более 20 команд из разных стран мира.

В работе [19] для определения лингвистических составляющих (просодических признаков и типов ответа) использовался программный инструментарий CMU-Sphinx. В ходе экспериментальных исследований достигнут

результат UAR = 74,9%. Авторы работы [20] предложили новый набор признаков (акустико-просодических, синтаксических, лексических и фонотактических), в котором каждый признак оценивался на его полезность для решения поставленной задачи. Помимо корпуса, представленного организаторами, в данной работе также использован корпус CSC (Columbia University, SRI International и University of Colorado Boulder). Авторам удалось добиться результата UAR = 67,7% и 62,2% при использовании предложенного ими набора признаков и набора признаков, предоставленного организаторами соревнований, соответственно. В работе [14] исследовано использование эмоциональных признаков (интенсивность эмоции (arousal) и ее валентность (valence)) в задаче определения ложной/истинной информации. С использованием такого подхода авторам удалось добиться результата UAR = 68,0%. Потенциал использования информативных признаков на основе автоматического распознавания фонов (фонем) в речи исследовался в работе [21]. Авторами был предложен высокоуровневый набор признаков (гласные, фоны, паузы, псевдослоги), из которых отобраны 29 статистических признака и добавлены к базовому набору признаков, предложенному организаторами соревнований. Результат классификации на наборе высокоуровневых признаков достиг показателя UAR = 58,6%, а при его объединении с базовым набором авторы смогли достичь результата распознавания 66,7% и 69,3% по показателю UAR для отладочного и тестового наборов, соответственно. В работе [22] использованы векторы Фишера и каскадная нормализация признаков для предобработки признаков, а метод экстремального обучения для классификации. Авторам удалось достигнуть результатов 75,2% и 66,6% по показателю UAR для отладочного и тестового наборов, соответственно.

Помимо работ, представленных на соревнованиях по компьютерной паралингвистике, имеется также ряд работ, представленных вне рамок соревнований. В работе [23] для вычисления акустических признаков использовано дробное преобразование Фурье. Если учитывать, что психоэмоциональное состояние человека влияет исключительно на частотные характеристики речевого сигнала, то возможно использование обычного оконного преобразования Фурье,

однако, его будет недостаточно в случае, если изменяется только фаза. По этой причине применены дробные мел-частотные кепстральные коэффициенты (MFCC). Результаты эксперимента показали, что при выборе оптимального порядка дробного преобразования Фурье для дробных кепстральных коэффициентов, точность распознавания ложности высказывания выше, чем при применении мел-частотных кепстральных коэффициентов. Средняя точность (Accuracy) распознавания ложной/истинной информации составила 56,2% и 59,9% для женщин и мужчин соответственно. Использование скрытых марковских моделей (СММ) улучшило эти показатели точности распознавания до 70,2% и 71,0%.

В работе [24] для определения ложной и истинной информации использовались акустические, просодические и лексические признаки речи диктора, а также информация о его половой и этнической принадлежности и личностных факторах. Авторы собрали корпус из записей диалогов 126 пар испытуемых, общей длительностью 93,8 часа речи. В ходе экспериментальных исследований с методом случайного леса авторам удалось достигнуть точности (Accuracy) распознавания ложной/истинной информации 61,2% на наборе акустико-просодических признаков и 63,03% при использовании нормализации признаков. При добавлении информации из результатов личностного теста NEO-FFI (Neuroticism-Extraversion-Openness Five-Factor Inventory), а также информации о половой принадлежности и родном языке дикторов точность распознавания выросла до 65,9%.

Авторы работы [25] выявили, что существуют индивидуальные различия при произнесении лжи: у некоторых дикторов повышается ЧОТ, у других наоборот понижается, а кто-то может рассмеяться при произнесении лжи. Был использован корпус, часть которого была размечена согласно этим наблюдениям, он состоит из записей диалогов людей длительностью 3-4 минуты, которые отвечают на простые открытые вопросы. Также был проведен тест NEO-FFI. Из аудиоданных выделены акустико-просодические и лексические признаки, после чего машинные классификаторы обучались определению пола, родного языка и личности

говорящего. Для обработки сигнала использован программный инструментарий Praat, а для выделения лексических признаков — LIWC [26]. При использовании метода адаптивного бустинга авторам удалось достичь значения точности (Accuracy) распознавания 61,0%.

В работе [27] авторы использовали многомодальный корпус Box of Lies для обучения и тестирования методов случайного леса. В экспериментальных исследованиях при многомодальной классификации получен лучший результат точности (Accuracy), равный 73,0%. Авторы работы [28] разработали макет автономного виртуального агента, который по их заявлениям, может соревноваться с людьми в игре, где необходимо лгать. С помощью игры авторы собрали корпус (содержит речь на английском языке и иврите), на котором обучили виртуального агента играть в игру, определяя, лжет ли оппонент. Лучшей моделью оказалась модель многослойного перцептрона с пятью признаками (MFCC, мел-частотные спектрограммы, признаки спектрального контраста, признаки короткого преобразования Фурье, представление тонального пространства Тоннец), которой удалось достичь 66,5% и 54,6% по показателям точности (Accuracy) и F1-меры, соответственно.

1.3 Анализ современного состояния исследований в области автоматического определения депрессии в разговорной речи

В последние годы в медицинской и научно-технической среде возрос интерес к задаче автоматического определения наличия депрессивного состояния у людей. Депрессия является одним из самых распространенных психических заболеваний, непосредственно влияющих на жизнь человека.

Согласно данным ВОЗ [29], депрессия является распространенным психическим расстройством и одной из основных болезней, которые приводят к ухудшению жизнедеятельности человека, и может стать причиной инвалидности. На 2018 год во всем мире около 264 млн. человек во всех возрастных группах страдали от депрессии [30].

В последние 10 лет возрос интерес к системам автоматического определения депрессии. На это повлияли многие причины: тяжесть заболевания и повсеместная распространенность, отсутствие лабораторных тестов и так далее. На данный момент наличие заболевания определяется путем беседы со специалистом-психотерапевтом и заполнения различного рода опросников: шкала депрессии (Patient Health Questionnaire, PHQ) [31], шкала депрессии Бека (Beck Depression Inventory, BDI) [32], самооценки депрессивных симптомов (Quick Inventory of Depressive Symptoms - Self-Report, QIDS-SR) [33], шкала Гамильтона для оценки депрессии (Hamilton Rating Scale for Depression, HRSD [34] и других. Однако профессиональная оценка может варьироваться в зависимости от компетентности специалиста и методов диагностики, которые он использует [35].

Многие работы представляют автоматические системы для определения состояния депрессии – существуют как одномодальные, так и многомодальные системы. Кроме того, часть систем решает задачу регрессии (определяя степень тяжести заболевания), а часть – задачу бинарной классификации (для определения наличия заболевания или его отсутствия). Задача определения депрессии была неоднократно представлена на соревнованиях AVEC (Audio-Visual Emotion Challenge) в 2013 [36], 2014 [37], 2016 [38], 2017 [39] и 2019 годах [40].

Большинство существующих работ при разработке автоматических систем опираются на гипотезу о том, что эмоциональное состояние диктора существенно влияет на акустические характеристики (спектральные и просодические) его речи. Лингвистические и нелингвистические факторы влияют на фонетические характеристики речи. Среди таких факторов можно отметить: физическое и психическое состояния говорящего, различные патологии мышления и психические болезни, ряд болезней, влияющих непосредственно на возможность речеобразования, и другие [17, 41, 42, 43].

Построение автоматической системы включает в себя, в том числе, и понимание исследователем рассматриваемой задачи. В случае с определением депрессии возможна векторизация и использование некоторых признаков депрессии для обучения моделей, а именно тех признаков, которые проявляются

вербально и невербально и на которые обращает внимание специалист при личной беседе с пациентом.

Аффективные состояния – психические состояния человека, отличающиеся заметной эмоциональной окрашенностью: различные эмоциональные состояния, состояние аффекта, настроение и подобные состояния. Изменения аффективного состояния являются естественной характеристикой поведения людей. Однако, когда эти изменения становятся интенсивными, длятся продолжительное время и при этом ухудшается жизнедеятельность человека, есть вероятность, что таким образом может проявляться аффективное расстройство. В отличие от кратковременных эмоций, настроение – длительное по времени аффективное состояние и, следовательно, клиническая депрессия – это расстройство настроения, которое может длиться неделями, месяцами и даже годами, изменяясь в тяжести заболевания, если не получено соответствующее лечение.

Расстройства настроения, несомненно, касаются естественных эмоциональных состояний. В частности, схема поведения людей, страдающих от таких расстройств настроения, как униполярная депрессия, показывает сильную временную корреляцию с аффективными величинами валентности, активации и доминации [36]. Данные аффективные величины используются для определения эмоционального состояния человека.

Специалисты выделяют два противоположных аффективных расстройства (расстройств настроения): депрессию (или большое депрессивное расстройство, БДР) и манию [29]. В психологии и психиатрии они обозначаются при помощи терминов униполярная депрессия (пациентов беспокоит депрессивное состояние) и биполярное аффективное расстройство (БАР, пациенты переживают как депрессию, так и манию). При этом депрессия и мания могут проявляться одновременно, что приводит к смешанному аффективному эпизоду. Кроме того, мания и депрессия могут проявляться в менее тяжелой форме (гипомания и дистимия соответственно) или могут быстро сменяться, что называют быстрой циркуляцией фаз. На рисунке 2 показаны фазы течения описанных заболеваний, приведенные в работе [44] (для удобства восприятия два графика течения

заболеваний объединены в один, переведены надписи на графиках, а также график представлен в монохромном виде). Верхняя граница рисунка обозначает состояние мании, нижняя указывает на состояние депрессии, а средняя – нормальное состояние. Пунктиром показаны менее тяжелые состояния гипомании и дистимии, которые также являются отклонением от нормального состояния, по горизонтальной оси указано течение времени, а по вертикальной – валентность заболевания.

Согласно диагностическому и статистическому руководству по психическим расстройствам в 5 издании (Diagnostic and Statistical Manual of mental disorders V, DSM-V) [45], для постановки диагноза депрессии необходимо, чтобы на протяжении как минимум двух недель присутствовали 5 или более симптомов (включая как минимум один из основных: подавленное настроение и/или потеря интереса и утрата способности получать удовольствие от приятной ранее деятельности): подавленное настроение; потеря интереса и утрата способности получать удовольствие от приятной ранее деятельности; расстройства сна и аппетита; психомоторное возбуждение/заторможенность; повышенная утомляемость и снижение энергичности; сниженная самооценка, чувство никчемности или неадекватное чувство вины; снижение способности к концентрации внимания или заторможенное мышление; суицидальные тенденции.

Крайней формой выражения депрессии является самоубийство, а риск самоубийства у пациентов с депрессией в течение жизни составляет 15%. В работе [46] проведен сравнительный метаанализ, который показал 3428 факторов риска суицида среди 365 лонгитюдных (длительных) исследований. Авторы сделали вывод, что все обозначенные факторы риска недостаточно точно могут предсказать суицид, что частично может быть вызвано методологическими ограничениями обследований. По мнению авторов работы [46], исследования показателей оценивания рисков суицида с использованием клинического инструментария, предсказывающего суицид, также недоработаны, а последние систематические

обзоры показывают, что на данный момент нет инструментария для определения риска суицида, который показывал бы высокую точность.

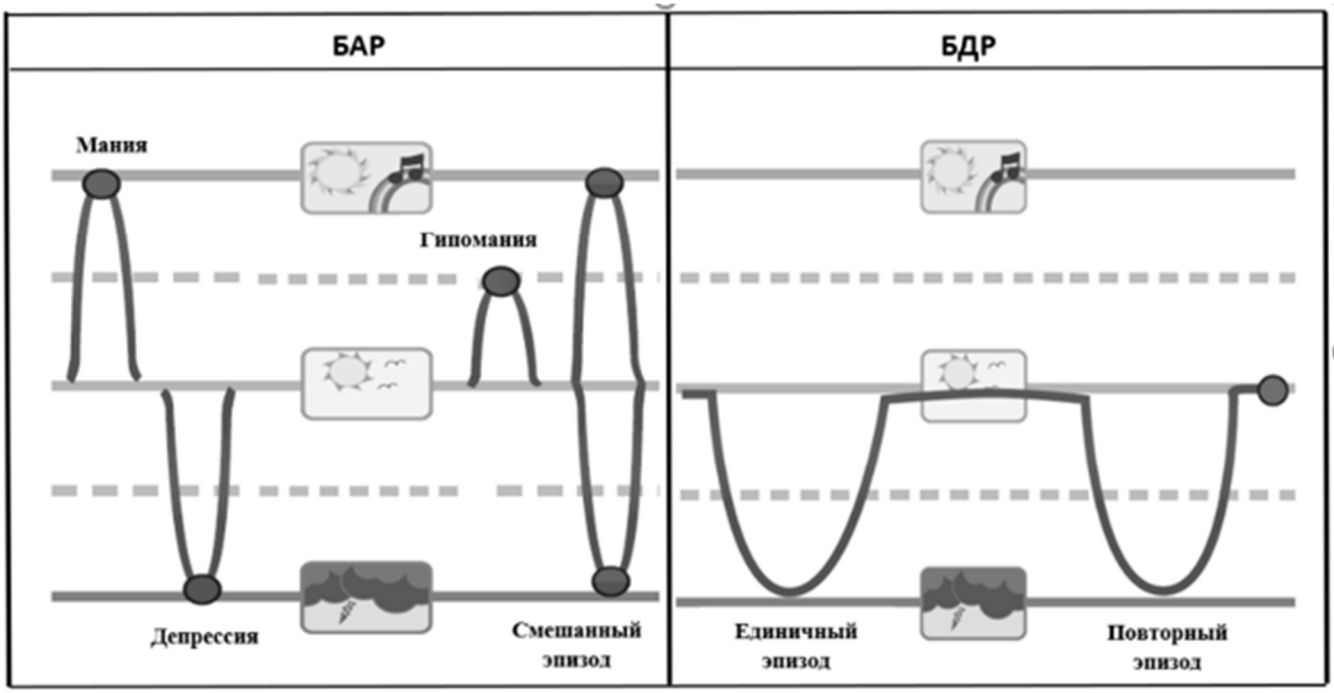

Рисунок 2 – Фазы течения заболевания при БАР и БДР (по рисунку [42])

В работе [47] проведен статистический анализ существующих предикативных моделей суицидального поведения на основе англоязычных рецензированных статей. В рассмотренных авторами исследованиях предикативные модели должны были предсказать смерть путем самоубийства или попытку самоубийства. Была проведена серия симуляций на гипотетической популяции индивидов с целью выявления преимуществ в статистическом моделировании, которые могли бы повысить способность предсказания попыток суицида и смертей. Было отмечено, что системы предсказания суицидального поведения должны проходить клинические тесты до внедрения в систему здравоохранения, чтобы показывать высокую точность именно в клинических случаях, а также предполагать более общие результаты, включая суицидальные тенденции, поскольку ошибочные гипотезы системы (как ложноположительные, так и ложноотрицательные) могут навредить пациентам.

Задача определения депрессии относится к задаче определения деструктивных паралингвистических явлений, которые также включают в себя

такие явления, как агрессия и ложь. Связь между агрессией и депрессией в известных теориях выражена неявно. Редкие случаи депрессии, предположительно, были обнаружены в обществах и группах, где агрессия могла быть сразу выражена, например, в военных сражениях и там, где процент случаев насильственной смерти велик [48]. Тем не менее, немногие межкультурные работы, существующие на момент написания статьи [48], не поддерживали гипотезу об обратной корреляции между агрессией и депрессией. Большинство рассмотренных автором статьи [48] работ подтверждали, что обязательным компонентом депрессии является чувство вины. Несмотря на то, что автор утверждает, что рассмотренные им работы имели методологические недоработки, сочетание признаков депрессии при постановке диагноза (депрессивное настроение, суточные циркадные изменения, усталость, бессонница, потеря интереса, потеря веса, периодичность и двухфазовая природа заболевания) все-таки обнаруживается во всех культурах.

Перед автоматическими системами определения депрессии, в том числе, стоит задача проверки данной гипотезы, а также использование выявленных корреляций с другими деструктивными паралингвистическими явлениями и аффективными величинами для получения более точного предсказания автоматической системы о наличии или отсутствии заболевания.

В рассмотренных далее работах, в основном, исследуется униполярная депрессия или БДР. Значительная часть работ представлена в рамках соревнований по аудиовизуальному определению эмоций AVEC. Лучшие результаты систем, представленных на соревнованиях AVEC-2019, отображены в таблице 1. В ней указаны авторы работы, модальности и признаки, которые использованы для обучения моделей, а также сами модели и результаты классификации, которых удалось добиться авторам по показателям точности CCC (Concordance Correlation Coefficient [49]) и RMSE (Root Mean Squared Error [50]). В качестве сокращений указаны Dev – набор данных для разработки (Development set), а Test – тестовый набор данных (Test set).

Таблица 1 – Результаты сравнения систем на соревнованиях AVEC

Работа	**Модальность**	**Признаки/ Классификатор**	**CCC**	**RMSE**
			Dev / Test	**Dev / Test**
Базовая работа Ringeval F. et al. [40]	Аудио	MFCCs	0,198 / –	7,28 / –
		eGeMAPS	0,076 / –	7,78 / –
		BoAW-M	0,102 / –	6,32 / –
		BoAW-e	0,272 / 0,045	6,43 / 8,19
		DS-DNet	0,165 / –	8,09/ –
		DS-VGG	0,305 / 0,108	8,00 / 9,33
	Видео	FAUs	0,115 / 0,019	7,02 / 10,0
		BoVW	0,107 / –	5,99 / –
		ResNet	0,269 / 0,120	7,72 / 8,01
		VGG	0,108 / –	7,69 / –
	Аудио Видео	Все признаки	0,336 / 0,111	5,03 / 6,37
Kaya H. et al. [53]	Аудио Видео	Объединение разработанных классификаторов	0,481 / 0,344	–
Ray A. et al. [55]	Аудио	Funct MFCC	– / –	5,11 / –
	Видео	BoVW	– / –	5,70 / –
	Текст	Текст	– / –	4,37 / –
Makiuchi M.R. et al. [56]	Аудио	CNN	0,338 / 0,199	5,97 / 7,02
		GCNN-LSTM	0,497 / –	5,70 / –
	Текст	LSTM	0,360 / 0,048	4,97 / 6,88
		8 CNN blocks-LSTM	0,685 / –	4,22 / –
	Видео	GCNN	0,372 / –	5,74 / –
	Аудио Текст	CNN and LSTM	0,452 / 0,213	5,08 / 6,42
	Аудио Текст	GCNN-LSTM и 7 CNN blocks-LSTM	0,696 / 0,403	3,86 / 6,11
	Аудио Текст Видео	GCNN-LSTM, 7 CNN blocks-LSTM и GCNN	0,624 / –	4,86 / –
Fan W. et al. [58]	Аудио Видео Текст	Ансамбль классификаторов	0,466 / 0,430	5,07 / 5,91
Yin S. et al. [59]	Аудио Видео Текст	Иерархическая двунаправленная LSTM	0,402 / 0,442	4,94 / 5,50

В соревнованиях AVEC-2019 [40] были представлены следующие темы: определение настроения (State-of-Mind Sub-challenge), использование искусственного интеллекта для определения депрессии (Detecting Depression with AI Sub-challenge) и определение эмоций в разных культурах (Cross-cultural Emotion Sub-challenge). Для задачи определения наличия депрессии был представлен корпус E-DAIC, расширенная версия WOZ-DAIC [51]. Аудиопризнаки состояли из

следующих наборов: MFCC (Mel-frequency Cepstral Coefficients), eGeMAPS, BoAW (Bag of Audio Words), Deep Spectrum. В качестве видеопризнаков использовались FAU (Face Action Units), BoVW (Bag of Video Words) [52], ResNet признаки.

Лучший результат на отладочном наборе данных для аудиомодальности получен при использовании признаков Deep Spectrum, вычисленных с использованием сети VGG-16, по показателю CCC = 0,305. При объединении всех представлений признаков результат, полученный на отладочном наборе, был улучшен до значений CCC = 0,336 и 0,111 на отладочном и на тестовом наборах соответственно. Значения RMSE также были улучшены до 5,03 и 6,37 на отладочном и тестовом наборах.

В работе [53] предложена аугментация (генерация) акустических признаков, предложенных организаторами, при помощи транскрипций, полученных с использованием технологии автоматического распознавания речи (Automatic Speech Recognition, ASR). Затем эти транскрипции были использованы для создания простой модели «мешка слов» (Bag of Words, BoW) [54444444444444444], после чего применялся анализ главных компонент для регрессии. Авторы использовали продолжительность реплик из транскрипций для получения общей длительности тишины и дыхания для каждого участника. Для моделирования автоматической сегментации на 7 классов (речь виртуального агента, дыхание, эксплетивные слова, звук губ, смех, тишина, речь субъекта) авторы экспериментировали с признаками eGeMAPS и Deep Spectrum (VGG-16). Супрасегментные признаки были смоделированы при использовании KELM (Kernel Extreme Learning Machine). Наилучший результат (CCC = 0,344) на тестовом наборе был получен при использовании простых признаков на основе транскрипции речи (подсчет слов, длительность и BoW). Помимо KELM использовалась модель ELM с взвешенным ядром (Weighted Kernel ELM). Наилучший результат сегментации на отладочном наборе в 4 из 17 аудиофайлов с использованием функционалов eGeMAPS LLDs и Weighted KELM, был UAR=65,75%. Результаты показали, что, хотя паттерны в части невербальных признаков сигнала важны, объединение их с лингвистической информацией

позволяет добиться лучших результатов без использования современных акустических и видеопризнаков.

Авторы работы [55] объединили признаки трех модальностей и использовали многоуровневую сеть с вниманием (attention network), которая обучалась взаимосвязям как между модальностями, так и внутри модальностей. Сеть использует несколько низкоуровневых и среднеуровневых признаков из аудио- и видеомодальностей, а также векторные представления предложений (sentence embeddings). В архитектуре предложенной сети контекстные признаки каждой модальности проходят через двуслойные сети прямого распространения, а выходные данные этих трех сетей объединяются в stacked BLSTM (Bidirectional LSTM). При помощи предложенной системы результат базовой системы был улучшен на 17,52% по показателю RMSE. Отдельные сети с вниманием для аудио- и видеомодальностей превзошли результат базовой системы по показателю UAR на 20,5%.

В работе [56] авторы использовали аудио- и текстовую модальность для определения депрессии. Для векторизации аудиомодальности использовались признаки, полученные из предобученной сети VGG-16 и применялась Gated Convolutional Neural Network (GCNN), а затем LSTM слой. Для получения лингвистических представлений были извлечены признаки BERT [57]. При использовании предложенного подхода удалось получить результаты CCC = 0,696 на отладочном наборе и CCC = 0,403 на тестовом наборе.

Авторы работы [59] для аудиомодальности использовали две категории признаков – 4096-размерный вектор, вычисленный из активации второго полносвязного слоя VGG-16, и 1920-размерный вектор, вычисленный из активации последнего слоя со сжатием в DenseNet-201. Для текстовой модальности использовалась RNN с кодировщиком-декодировщиком для создания вектора семантической репрезентации. Для создания вектора эмоциональных характеристик использовался словарь NRC Emotion Intensity Lexicon. Предложенный подход включает в себя две иерархические двунаправленные LSTM для объединения многомодальных признаков и предсказания тяжести

депрессии. На отладочном и тестовом наборах были достигнуты результаты 0,402 и 0,442 по показателю CCC; 4,94 и 5,50 по показателю RMSE, соответственно. Результаты систем, представленных вне соревнований AVEC, отображены в таблице 2. В ней указаны авторы системы, модальности, которые были использованы для обучения моделей, а также сами модели. Перечислены различные показатели качества, которые были использованы авторами, и результаты, полученные по этим показателям с использованием разработанных моделей.

В работе [60] для обучения моделей авторы использовали DAIC-WOZ корпус, взяв оттуда трехмерные изображения лица и речь информантов. Методика включает в себя использование векторных представлений на уровне предложений (sentence-level “summary" embedding), LSTM и casual-CNN. В качестве показателей качества были использованы показатель шкалы депрессии PHQ и бинарная классификация. Особенность предложенной системы заключается в том, что она не полагается на контекст интервью. Кроме того, она принимает на вход сырые данные (аудио, трехмерные модели лиц и транскрипция), которые суммируются в один вектор. Стоит отметить, что в подходе используются сделанные вручную и предобученные векторные представления на уровне слов на входе, то же самое используется и на уровне предложений. Модель показывает среднюю ошибку MAE = 3,67 по шкале депрессии, а также 83,3% чувствительности (Sensitivity), 82,6% специфичности (Specificity) [61] и F1-меру 76,9%.

В работе [62] также использовался DAIC-WOZ корпус. Авторы предложили архитектуру CombAtt: кодировщики модальностей, которые принимают на вход унимодальные признаки и на выходе выдают закодированные данные; сети с механизмом внимания для объединения сетей отдельных модальностей. Также они предложили сети для регрессии, которые на выходе предсказывают баллы по шкале PHQ-8. Использование предложенного подхода позволило получить результаты RMSE = 4,14 и MAE = 3,07, а также EVS (explained variance score) = 0,62.

Таблица 2 – Результаты сравнения систем вне соревнований AVEC

Работа	**Модальность**	**Классификатор**	**Показатель**	**Результат**
Haque A. et al. [60]	Аудио Видео	Casual CNN	F1-мера, Precision, Recall, Average Error	76,9%, 71,4%, 83,3%, 3,67
Qureshi S.A. et al. [62]	Аудио Видео	CombAtt network	RMSE, MAE, EVS	4,14, 3,34, 0,62
Niu M. et al. [63]	Аудио	Гибридная сеть (CNN, LSTM и DNN) и l_p-нормированное сжатие	RMSE, MAE	9,66, 8,02
Rohanian M. et al. [64]	Аудио Видео	LSTM с механизмом окна	F1-мера, Precision, MAE, RMSE	81,0%, 80,0%, 3,61, 4,99
Tao F. et al. [65]	Аудио	SVM	Accuracy	84,5%
Xezonaki D. et al. [66]	Текст	Двухуровневая иерархическая нейронная сеть с механизмом внимания	F1-мера (General Psychotherapy Corpus), F1-мера (DAIC-WOZ), UAR (DAIC-WOZ)	71,6%, 70,3%, 70,3%
Huang Zh. et al. [67]	Аудио	FVTC-CNN	UAR (SH2-FS), UAR (DAIC-WOZ)	68,0%, 88,0%
Zhao Z. et al. [68]	Аудио	Гибридная сеть (сеть с механизмом внутреннего внимания, глубокая сверточная сеть, SVR)	MAE (Corpus 2013), RMSE (Corpus 2013), MAE (Corpus 2014), RMSE (Corpus 2014)	9,65, 7,38, 9,57, 7,94
Seneviratne N. et al. [69]	Аудио	SVM	Accuracy	81,7%
Stankevich N. et al. [73]	Текст	SVM, модели TF-IDF со стилистическими и морфологическими признаками; SVM, векторные представления слов, стилистические признаки	F1-мера, Precision, Recall; Recall, F1-мера	63,0%, 65,0%, 61,0%; 84,0%, 61,0%
Enikolopov S.N. et al. [75]	Текст	Метод случайного леса, психолингвистические признаки и биграммы	F1-мера	73,0%

В работе [63] в качестве данных для обучения использовались два свободно доступных набора данных: AVEC2013 и AVEC2014. После вычисления и сегментирования MFCC признаков использовалось l_p-нормированное сжатие

пространства признаков, объединенное с LASSO (Least Absolute Shrinkage and Selection Operator), чтобы найти оптимальный параметр сжатия с целью последующей генерации признаков на уровне высказываний для определения депрессии. Эти данные использовались для обучения гибридной модели, которая содержит CNN, LSTM и DNN, а итоговое предсказание уровня депрессии проводилось с использованием SVR (Support Vector Regression). На тестовом наборе AVEC2013 были получены результаты RMSE = 9,79, MAE = 7,48, а на тестовом наборе AVEC2014 – RMSE = 9,66, MAE = 8,02.

Авторы работы [64] для обучения моделей использовали корпус DAIC-WOZ. Для данного исследования авторы решили прибегнуть к искусственному выравниванию между текстовыми, аудио- и видеопризнаками, чтобы получить точные временные метки каждого произнесенного слова. На каждой временной отметке выравнивали слова и соответствующие им отрезки аудиозаписей с использованием инструментария Penn Phonetics Lab Forced Aligner (P2FA), который может применяться для сравнения транскрипций с аудиофайлами, фонема за фонемой. Путем ручной проверки искусственное выравнивание было проделано с достаточно высокой точностью для изучения объединения модальностей. В модели использовались слои прямого распространения с оконным механизмом (реализация стохастического градиентного спуска), которые обучаются регулировать поток информации в сети, присваивая веса видео- и аудиоданным на каждой временной отметке. Результаты обучения LSTM с механизмом окна на признаках трех модальностей: F1-мера = 81,0%, Precision = 80,0%, MAE = 3,61, RMSE = 4,99.

В работе [65] для обучения системы использовались записи 110 человек, из которых 54 никогда не имели психических заболеваний (контрольная группа), а у 56 была диагностирована депрессия (группа людей с депрессией). В группе информантов с депрессией имелись следующие заболевания: большое депрессивное расстройство (19 случаев), биполярное расстройство в депрессивной фазе или с последним депрессивным эпизодом (13 случаев), реактивная депрессия (7 случаев), эндо-реактивная депрессия (6 случаев) и тревожно-депрессивное

расстройство (4 случая). Для остальных 7 информантов точный диагноз не был установлен. Все участники являлись носителями итальянского языка, их просили прочитать вслух басню Эзопа «Ветер и Солнце». В качестве акустических признаков использовался набор INTERSPEECH 2009 Emotion Challenge, который был расширен путем добавления признаков для определения скорости чтения и использования пауз. Обучение классификатора SVM, реализованного в библиотеке Scikit-learn, происходило при помощи метода leave-one-out. Авторы предположили, что люди с депрессией читают медленнее и чаще используют длинные паузы. Данное предположение подкрепляется исследованиями нейробиологов, которые показывают, что течение процессов в мозге, связанных с языком, занимают больше времени у людей с депрессией. В частности, было показано, что есть связь между депрессией и дисфункцией в некоторых зонах, участвующих в семантической обработке языка, включая фронтальную извилину и префронтальный кортекс. В проведенных экспериментах среднее время на чтение текста и стандартное отклонение 54,92±2,66 сек и 47,38±1,20 сек для людей с депрессией и без, соответственно, а скорость чтения 202,1 и 234,3 слова в минуту. Так, после добавления этих признаков, авторам удалось улучшить точность распознавания депрессии с 68,2% до 84,5%.

В работе [66] для обучения использовались корпусы General Psychotherapy Corpus и DAIC-WOZ. Авторы предложили подход с использованием иерархической архитектуры нейронной сети с вниманием для определения депрессии по транскрипциям клинических интервью. Было выдвинуто предположение, что эмоциональное содержание может быть отличительным фактором между характеристиками речи людей с депрессией и без. Основываясь на этом, они применили лингвистические знания об эмоциональном содержимом слов, рассмотрев эмоции, тональность, валентность и психолингвистическую аннотацию для слов. Для того, чтобы исследовать использование слов, которые отражают позитивную и негативную тональности, грусть и тревожность, авторы использовали инструментарий LIWC lexicon, в котором представлена психолингвистическая аннотация 18504 слов для 73 различных категорий слов.

Эксперименты показали, что дополнительная информация об эмоциях улучшает результат предложенной архитектуры. Авторам удалось добиться F1-меры 71,6% для корпуса General Psychotherapy Corpus, а также F1-меры и невзвешенной средней полноты 70,3% для корпуса DAIC-WOZ.

Работа [67] посвящена проблеме обобщения и предлагает несколько стратегий адаптации, которые модернизируют предобученные модели на основе расширяемых сверточных сетей с целью улучшить точность определения депрессии как в лабораторных, так и в естественных условиях. Для обучения сетей использовались два корпуса: SH2-FS (Free Speech) и DAIC-WOZ. Авторы использовали четыре набора акустических признаков: 3 форманты, 13 спектральных центроидных частот, 16 MFCC и 16 значений производных MFCC. В работе исследуется метод FVTC-CNN (full vocal tract coordination – convolutional neural networks). Авторам удалось добиться точности по невзвешенной средней полноте 68,0% для корпуса SH2-FS и 88,0% для корпуса DAIC-WOZ.

В работе [68] изучались преимущества гибридной сети, которая кодирует характеристики речи, относящиеся к депрессии. В работе использовался корпус AViD-Corpus. Для вычисления низкоуровневых признаков использовался набор eGeMAPS из программного инструментария openSMILE. Предложенный метод включает в себя сети внутреннего внимания, обученные на низкоуровневых акустических признаках, глубокую сверточную сеть, обученную на информации, полученной из трехмерных лог-мел-спектрограмм, и модуль предсказания степени депрессии по шкале Бека-2 [69] с использованием сверточной сети и регрессора опорных векторов. Для корпуса 2013 года были получены результаты 9,65 и 7,38 по показателям RMSE и MAE соответственно. Для корпуса 2014 года авторам удалось добиться результатов RMSE = 9,57 и MAE = 7,94.

Авторы работы [70] исследовали изменения в речи, которые происходят в результате психомоторной заторможенности, считающейся ключевым признаком БДР. Для этого применялись инверсированные переменные речевого тракта, которые получаются посредством системы инверсии речи, преобразующей акустический сигнал в шесть артикуляторных траекторий. Также были

использованы мел-частотные кепстральные коэффициенты и корреляционные признаки. В качестве данных использовался корпус Mundt, а классификатор SVM обучался с использованием подхода LOSO. Авторам удалось добиться точности (Accuracy) определения депрессии в 81,7%.

В цикле работ [71, 72, 73, 74, 75] авторы используют текстовую модальность для определения депрессии, а именно: частеречный анализ, TF-IDF [76], векторные преставления слов, n-граммы, классические текстовые и психолингвистические признаки, а также анализ тональности. Наилучший результат был получен авторами на основе набора данных CLEF/eRisk 2017, в который входит коллекция текстовых сообщений 887 пользователей социальной сети Reddit, из которых 135 текстов помечены как депрессивные. Использовались методы SVM и случайного леса, реализованные в библиотеке Scikit-learn. Лучшие результаты определения депрессии были получены при использовании SVM и модели TF-IDF со стилистическими и морфологическими признаками, 63,0% по показателю F1-меры, 61,0% полноты и 65,0% точности. При этом модель на основе SVM, векторного представления слов и стилистических признаков получила наилучший результат полноты, равный 84,0% и F1-меры равной 61,0%. В экспериментах со случайным лесом лучшей моделью оказалась TF-IDF с морфологическими признаками, с использованием которой был получен результат 79,0% точности и 62,0% F1-меры. Лучший результат на корпусах, собранных авторами, достиг F1-меры в 73,0% при использовании метода случайного леса и набора признаков, включающего психолингвистические признаки и биграммы. Кроме того, авторы собрали два корпуса: корпус эссе «Я, другие, мир» и корпус информации из профилей социальной сети «Вконтакте».

Среди рассмотренных работ можно выделить регрессионные и классификационные системы, в которых используются как нейросетевые, так и классические классификаторы. Так, можно отметить, что для задач классификации и регрессии при определении депрессии популярны в основном нейросетевые методы, а именно сложные архитектуры нейросетевых методов. Вероятно, данная тенденция прослеживается ввиду того, что такие методы обладают большей

устойчивостью к переобучению, большей способностью к обобщению, а также способностью к выявлению скрытых корреляций в признаковом пространстве.

Также стоит отметить, что в последние несколько лет актуальными являются системы, использующие различные архитектуры нейронных сетей, а наилучшие результаты были получены при использовании рекуррентных архитектур и нейросетей с механизмом внимания. Данная особенность присуща большинству систем, представленных как на соревнованиях AVEC, так и вне соревнований. Описанные работы показывают эффективность многомодального подхода при определении депрессии. Также, нельзя не заметить, что вербальная информация также играет важную роль и позволяет добиться высоких результатов, наряду с паттернами невербальной информации, что подтверждает опыт медицинских работников.

1.4 Анализ современного состояния исследований в области автоматического определения агрессии в разговорной речи

Под термином агрессия в европейской культуре подразумевается деструктивное поведение, которое является мотивированным, а также противоречит нормам сосуществования людей. Такое поведение может быть направлено как вовне (нанесение вреда или психологического дискомфорта окружающим людям, животным, предметам), так и на себя (самоповреждение, самобичевание) [77].

Согласно опроснику Басса-Дарки существует несколько видов агрессивных реакций [78]:

– Физическая агрессия – применение физической силы против собеседника.

– Косвенная агрессия – непрямым путем направленная на собеседника, или не направленная ни на кого.

– Раздражение – готовность к проявлению негативных чувств при малейшем возбуждении.

– Обида – ненависть или зависть к собеседнику по существующей или надуманной причине.

– Подозрительность – может находиться в интервале от недоверия и осторожности по отношению к окружающим до убежденности в том, что окружающие люди хотят нанести или наносят вред.

– Вербальная агрессия – вербальное проявление негативных чувств как через крик или визг, так и через словесные ответы.

– Чувство вины – выражение возможного убеждения субъекта в том, что он плохой человек и поступает плохо, субъект также ощущает угрызения совести.

Графически формы агрессивного поведения по классификации Басса представлены на рисунке 3.

Аутоагрессия является причинением субъектом вреда себе (как физического, так и психологического) и может быть отнесена к механизмам психологической защиты. Аутоагрессия может проявляться в самоунижении, самообвинении, нанесении себе телесных повреждений вплоть до самоубийства. К аутоагрессии также относится саморазрушительное поведение (алкоголизм, наркомания, выбор экстремальных видов спорта, опасных профессий, провоцирующее поведение). Аутоагрессия считается типичной для депрессивных личностей, также она может быть свойственна людям с мазохистическим характером [79].

В работе [80] было выявлено, что характеристиками агрессии являются такие акустические показатели как: высокая громкость речи и ее высокая вариативность, низкая высота основного тона и ее высокая вариативность, быстрая скорость речи, короткая длительность речи и короткая длительность пауз, а также малое количество пауз.

В 2021 году в рамках международной конференции INTERSPEECH на соревнованиях ComParE была представлена тема определения уровня агрессии. В качестве данных участникам был предложен набор данных, состоящий из двух речевых корпусов: Dataset of Aggression in Trains (TR) и the Stress at Service Desk Dataset (SD). Всего для обучения, отладки и оценивания моделей было предложено

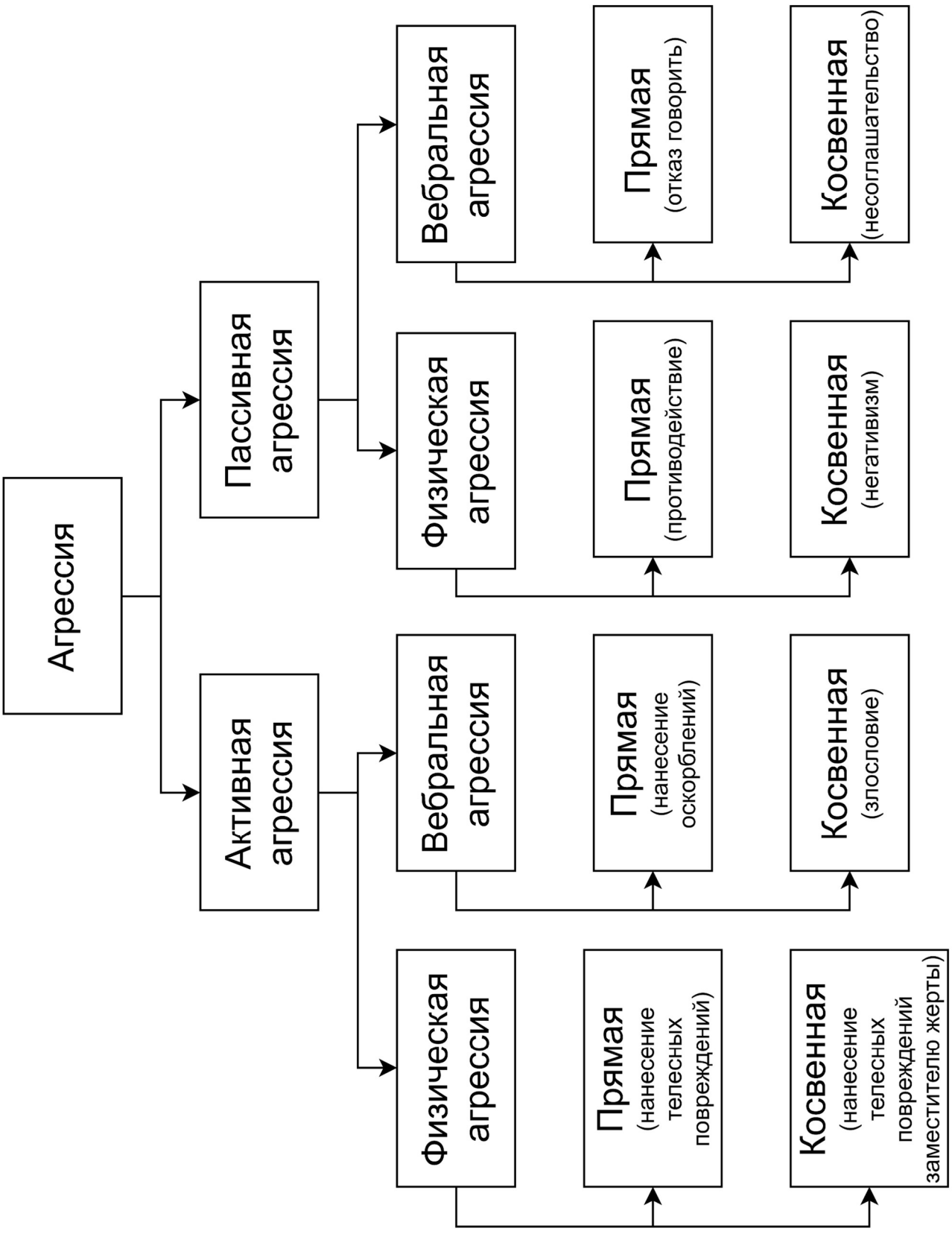

Рисунок 3 – Классификации видов агрессии по А. Бассу (по классификации [75])

911 аудиозаписей. Помимо набора данных были также предоставлены несколько наборов акустических признаков. С использованием набора признаков ComParE

2013 и метода опорных векторов организаторами конкурса был получен базовый результат 72,2% по показателю UAR. Победители данного соревнования [81] использовали X-вектора, вычисленные из спектрограмм, а также базовые признаки и векторы Фишера. Для классификации использовался метод опорных векторов, обучение каждого классификатора сначала проходило на отдельных наборах признаков, а затем предсказания объединялись. Авторам удалось добиться результатов на тестовом наборе данных в 61,5% по показателю невзвешенной средней полноты при использовании комбинации признаков с X-векторами и 63,2% при использовании комбинации признаков с векторами Фишера. На отладочном наборе авторы добились результата UAR = 77,8%.

В работе [82] авторы предлагают распознавать агрессию при помощи определения наложения речи (перебивания) субъектов. В большинстве случаев сегменты, в которых содержатся наложения речи, не учитываются в задаче определения агрессии, ввиду сложностей с определением акустических признаков высоты основного тона, но в данной работе такие сегменты учитываются. Авторами была разработана система терапии с использованием виртуальной реальности, целью которой является помощь пациентам судебно-медицинских клиник для борьбы со склонностью к агрессии. Помимо акустических признаков был использован вектор признаков, состоящий из информации о наложениях речи, представленных тремя категориями: короткий ответ, преждевременный коммуникативный ход и состязательное перебивание оппонента. В качестве метода классификации был выбран метод случайного леса. Авторы отметили, что использование наложений речи позволило улучшить результат определения агрессии в речи на 3% до 53,0% по показателю невзвешенной точности (Unweighted Accuracy).

Авторы работы [83] для определения агрессии в речи применяли признаки изменения давления воздуха в разных отделах голосового тракта. Они выявили, что одни и те же гласные, произнесенные с агрессией и без показывают различные значения давления воздуха в разных отделах голосового тракта. Для уменьшения размерности признакового пространства был применен метод главных компонент,

а для классификации использовался метод скрытых марковских моделей. Сначала выносилось решение для каждой гласной в аудиозаписи, после чего итоговое предсказание производилось за счет голосования по большинству. Авторам удалось добиться качества работы моделей по показателю точности (Accuracy) до 93,0%.

В работе [84] использовали акустическую и лексическую информацию для определения агрессии. Авторы использовали несколько наборов данных для экспериментов: TR и SD, RAVDESS, CREMA-D, SAVEE и TESS. Сначала авторы применили метод обнаружения голосовой активности во входном акустическом сигнале (Voice Activity Detector, VAD) для удаления сегментов, не содержащих речь. Затем из аудиоданных были извлечены MFCC признаки, которые впоследствии были поданы на вход предобученной нейронной сети ResNet-18. Для лексической составляющей был использован предобученный метод на основе энкодеров (Sentence-BERT, SBERT). На последнем этапе для классификации уровня агрессии был выбран метод опорных векторов. Лучшим результатом, полученным с использованием такого подхода, является результат 81,5% по показателю UAR для TR и SD наборов данных при использовании переноса обучения с наборов данных RAVDESS, CREMA-D, SAVEE и TESS.

1.5 Анализ баз данных для исследования задач автоматического определения деструктивных паралингвистических явлений в разговорной речи

Важным фактором, влияющим на результаты работы систем определения деструктивных паралингвистических явлений в разговорной речи, является набор данных, на которых были обучены их вероятностные модели. Многие исследователи сталкиваются с проблемой нехватки данных, так как речевых корпусов, содержащих рассматриваемые явления, не так много, что является естественным ввиду многих факторов: 1) задачи паралингвистики являются относительно новыми, хотя все больше ученых начинают проявлять интерес к исследованиям на эту тему; 2) процесс сбора специализированных аудиоданных является достаточно трудоемким и времязатратным, что, безусловно, сказывается

на объемах речевых корпусов и длительности аудиозаписей; 3) могут быть необходимы специальные сценарии, поскольку далеко не всегда возможно провести запись речевых сообщений, содержащих необходимые паралингвистические явления, в реальных естественных условиях. В данном разделе приведен анализ известных речевых и многомодальных корпусов, содержащих деструктивные паралингвистические явления в речи человека [85].

В мире существует всего несколько корпусов, содержащих ложную и истинную информацию. В таблице 3 приведено сравнение таких речевых корпусов по следующим параметрам: модальность (одна или несколько), язык дикторов, доступ (свободный, по запросу, закрытый), количество дикторов, длительность записей.

Речевой корпус DSD, был разработан в университете Аризоны (США). В записи аудиоданных принимали участие студенты университета, они были разделены на две группы, участники которых должны быть играть роли, определенные сценарием. Роль лжецов играли участники первой группы, они «украли» вопросы и ответы для экзамена на кафедре. Участники второй группы играли роль честных студентов, которые вернули документ в тот же кабинет. С каждым участником были проведены интервью, которые состояли из набора открытых вопросов, подразумевающих короткие ответы (в основном «да» или «нет»).

Корпус CSC был собран исследователями Университета Колумбии (США), некоммерческого института SRI и Колорадского Университета в Боулдере в 2013 году. Все участники являлись студентами или сотрудниками Университета Колумбии. Участников проинструктировали, что они участвуют в эксперименте, предназначенном для определения людей, чьи личностные качества являются схожими с характеристикой одного из ведущих бизнесменов Америки. Так, участники выполняли задания и отвечали на вопросы в шести областях знаний. После чего им сказали, что они набрали недостаточно баллов и могут попробовать еще раз. Интервьюер должен был определить то, как участники справились с заданиями, но ему было запрещено задавать им вопросы, не относящиеся к

заданиям. Участников попросили отмечать, был их ответ истинным или ложным, нажимая на одну из двух педалей, спрятанных под столом.

Корпус, собранный в Университете Ноттингема (Великобритания) состоит из записей речи мужчин, являющихся студентами или сотрудниками университета. Все они не имели каких-либо речевых или слуховых нарушений. Участникам были выданы жетоны, содержащие вербальную или графическую информацию, которую они должны были скрывать в течение всего интервью. Основное интервью состояло из нейтральных вопросов и его целью было собрать контрольные данные, содержащие истинную информацию. Другие два интервью должны были усилить напряжение участников с помощью наводящих вопросов.

Таблица 3 – Сравнение параметров корпусов, содержащих истинную/ложную информацию

Корпус	**Модальность**	**Язык дикторов**	**Кол-во дикторов**	**Длительность записей (мин)**	**Доступ**
Корпус DSD [86]	аудиоданные	английский	72	162	по запросу
Корпус CSC [87]	аудиоданные	американский английский	32	1920	по запросу
Корпус, разработанный в Университете Ноттингема [18]	аудиоданные	британский английский	19	нет данных	нет данных
Корпус, разработанный в Университете Сучжоу [23]	аудиоданные	китайский	50	нет данных	свободный доступ
Корпус CXD [24]	аудиоданные	американский английский, китайский	344	7350	нет данных
RLTDDD [88]	видеоданные	английский	56	49	свободный доступ
Box of Lies [89]	видеоданные	английский	26	144	свободный доступ

В Университете Сучжоу (Китай) речевой корпус собран с целью получить данные, максимально близкие к реальным условиям. Так, авторы разработали игру,

где участники были разделены на две группы (А и Б). Каждый участник группы А должен был рассказать историю, а участник группы Б мог задать произвольные вопросы о ней. Истории, вопросы и ответы были разными у всех дикторов. Участники группы Б не знали, являлась ли история правдой, а потому должны были принять решение самостоятельно, анализируя ответы на вопросы. В случае, если участники угадывали, они выигрывали игру, и наоборот, если не угадывали, выигрывал участник из группы А. Если история являлась специально придуманной ложью (блефом), рассказчик должен был приложить все усилия, чтобы не выдать себя и не проиграть. Участники группы Б могли задавать любые вопросы, касающиеся истории, чтобы заставить рассказчика нервничать и делать ошибки. Все истории, которые оказались ложью, использовались в качестве записей, содержащих ложную информацию. После чего была записана обычная речь участников, рассказывавших такие истории, при этом, темы рассказов могли быть любые (о себе, хобби, жизни и пр.).

Корпус CXD Corpus был собран в Университете Колумбии. В речи участников эксперимента исследованы акустические, просодические и лексические признаки, а также их этническая принадлежность, пол и личностные качества (участники проходили пятифакторный личностный тест NEO-FFI). Авторы разработали сценарий по типу игры и разделили участников на две группы. Было проведено биографическое интервью, которое состояло из 24 вопросов. Участники первой группы играли роль интервьюеров и им необходимо было определить, говорит ли оппонент правду, в то время как участники второй группы были интервьюируемыми (информантами) и должны были солгать во время ответа на случайную часть вопросов. После интервью группы менялись местами. На протяжении обоих интервью участники не имели визуального контакта. Интервьюеры записывали решение относительно правдивости ответов по шкале от 1 до 5, а интервьюируемые нажимали клавиши T (правда) или F (ложь), чтобы обозначить ответы, на которые они солгали.

Многомодальный корпус RLTDDD собран в Университете Мичигана (США) в 2016 году. Для сбора данных были использованы англоязычные ресурсы,

предоставляющие записи судебных слушаний, на которых можно определить поведение человека при произнесении ложной и правдивой информации. На аудио- и видеозаписях были точно определены подсудимый, и свидетель; достаточно четко видны их лица на протяжении большей части записи; качество видео является достаточно хорошим, чтобы определить мимику; также, качество аудиозаписей достаточное, чтобы слышать голоса и понимать, что говорят люди. Аудио- и видеозаписи были размечены экспертом как «лживые» и «правдивые» на основании вердикта суда: виновен, не виновен или оправдан. Так, если вердикт «виновен», аудио- и видеозаписи, содержащие ложную информацию, состоят из записей подсудимых, а видео, содержащие правдивую информацию – из записей свидетелей в тех же слушаниях.

Корпус Box of Lies содержит записи эпизодов из ТВ-шоу, где ведущий и гость рассказывают правду или ложь о предметах на столе. Участник, описывающий предмет, должен убедительно описать предмет (или скрыть его истинную суть) так, чтобы второй участник, который не видит этот предмет, ему поверил.

Значительная часть рассмотренных корпусов содержит речь носителей английского языка, однако они могут быть использованы в исследованиях по определению ложной или истинной информации в речи для языков из родственных языковых групп. Данная гипотеза подтверждается тем, что, хотя в различных языках и культурах существуют различия в поведении при произнесении лжи [90], языки и культуры из родственных языковых групп могут иметь схожие черты [91]. Некоторые из корпусов можно найти в свободном доступе (корпус, разработанный в Университете Сучжоу, RLTDDD и Box of Lies), однако, для доступа к большей части корпусов необходимо осуществить запрос у авторов.

Для создания автоматической системы определения депрессии необходимо иметь данные для обучения, которые содержат как речь информантов с установленной депрессией, так и информантов, у которых не было депрессии на момент записи. В таблице 4 представлены систематизированные данные об описанных ранее корпусах, содержащих речь людей с депрессией, а именно: язык,

количество дикторов / авторов текстов, методы оценивания заболевания, количество данных, доступность [35].

AViD-Corpus содержит аудио- и видеозаписи взаимодействия участников с компьютером, в то время как их действия записываются камерой и микрофоном. Поведение участников в ходе записей было определено заданием: произнесение букв, произнесение букв громким голосом, произнесение букв с улыбкой, повышение громкости голоса при выполнении задания. Участники читали отрывки из новелл и преданий, пели, рассказывали истории из своего прошлого, рассказывали вымышленные истории для тематического апперцепционного теста (апперцепция – восприятие, узнавание на основе прежних переживаний). Аннотация проводилась согласно опроснику депрессии, шкале Бека-2. Данный опросник является вторым пересмотром опросника Бека, принятым в 1996 году. Каждой записи было присвоено единственное значение метки наличия депрессии.

BlackDog – англоязычный многомодальный корпус, собранный в Австралии при помощи видеокамеры и микрофона. Метки бинарной классификации были проставлены вручную и содержат категории «тяжелая депрессия» и «здоровые субъекты». Интервью содержат вопросы с открытым ответом.

Pitt – англоязычный многомодальный корпус, собранный в Питтсбурге. Метки бинарной классификации содержат категории «тяжелая депрессия» и «легкая депрессия», проводилась ручная транскрипция. Опросник содержал вопросы из клинического интервью HRSD.

Distress Analysis Interview Corpus (DAIC) – многомодальная коллекция клинических интервью. Корпус разработан для симуляции стандартных процессов определения того, имеется ли у человека риск ПТСР (посттравматического стрессового расстройства) и большого депрессивного расстройства. Корпус содержит следующие типы интервью: очные интервью (лицом к лицу) между участниками и интервьюером; телеконференции, где интервью проводилось с

Таблица 4 – Сравнение корпусов, содержащих данные людей с депрессией

Корпус	**Язык данных**	**Количество информантов**	**Оценка заболеваний**	**Объем данных (мин или текстов)**	**Доступ**
AviD-Corpus [36]	Немецкий	292	Шкала Бека-2	14400	По запросу
BlackDog [92]	Английский	30	QIDS-SR	509	Не доступен
Pitt [93]	Английский	19	HRSD, QIDS-SR	355	Не доступен
Distress Analysis Interview Corpus (DAIC) [51]	Английский	219	5 опросников	4390	По запросу
Mundt [94]	Английский	35	Текущее лечение, шкала Гамильтона	Не известно	Не доступен
General Psychotherapy Corpus [95]	Английский	Нет данных	Оценка специалиста	Не известно	Платный доступ (30 дней бесплатного пользования)
SH2-FS (Free Speech) [96]	Английский	887	PHQ-9	960	Не доступен
RusNeuroPsych [97]	Русский	447 (246 детей до 18 лет и 209 взрослых)	3 опросника	643 (252 текста детей и 392 текста взрослых)	Открыт
Текстовый корпус эссе [71]	Русский	164	10 опросников	164 текста	Не доступен
Корпус информации из профилей социальной сети «Вконтакте» [72]	Русский	1330	Шкала депрессии Бека	Не известно	Не доступен

использованием телеконференций; автоматические интервью – интервью проводилось в автоматическом режиме с Элли; «Волшебник Оз» (WoZ) или «Гудвин» (Wizard of Oz, WoZ) – интервью проводилось анимированным

виртуальным интервьюером по имени Элли, которую контролировал интервьюер в другой комнате. Разработчиками корпуса Distress Analysis Interview Corpus (DAIC) были выбраны две группы жителей Лос-Анджелеса – ветераны военных сил США и гражданские лица. Все они были проверены опросниками на депрессию, ПТСР и тревожность. Интервью начинались с нейтральных вопросов, затем вопросы становились более специфичными (о симптомах, событиях), а заканчивались фазой спокойствия. Перед и после интервью участники заполняли ряд опросников, включающие базовые вопросы о биографии, измерение психологического стресса, а также измерение текущего настроения.

В записях в режиме «Волшебником Оз» в данных были найдены некоторые признаки стресса: участники, испытывающие стресс, медленнее начинали говорить, и использовали меньше заполненных пауз, чем участники, не испытывающие стресс. Кроме того, от типа стресса зависит, какие признаки наиболее предсказуемы. Если стандартное отклонение перед началом речи каждого участника диалога изменялось, это было лучшим признаком для предсказания депрессии. Однако для ПТСР более информативным было определение среднего количества заполненных пауз в сегменте. Время перед началом ответа при личных вопросах и длительность речи при ответах на вопросы для установления контакта указывают на наличие стресса. Участники использовали меньше заполненных пауз при диалоге с агентом, чем при диалоге с человеком. Участники выражали меньше страха или негативных проявлений, когда агент представлялся автоматическим, чем когда агент был представлен как управляемый человеком. Кроме того, участники показывали больше эмоций из категории «грусть», когда были уверены, что взаимодействуют с компьютером, а не с человеком.

RusNeuroPsych – текстовый русскоязычный корпус, разделен на две части: подкорпус «дети» (тексты написаны детьми школьного возраста от 12 до 17 лет) и подкорпус «взрослые» (тексты написаны взрослыми от 18 до 35 лет).

Mundt – речевой корпус, который содержит записи речи информантов за период в 6 недель. Одним из условий записи было начало фармакотерапевтического и/или психотерапевтического лечения депрессии.

Участники читали заранее подготовленный текст и описывали свои эмоциональные и физические ощущения.

General Psychotherapy Corpus – многомодальный корпус, состоит из транскрибированных терапевтических сессий, в которых представлены различные клинические подходы. Каждая сессия состоит из следующих друг за другом коммуникативных ходов, аннотированных как сторона терапевта и сторона клиента.

SH2-FS (Free Speech) – содержит записи речи в естественных условиях (дома, на работе, в машине) и оценки по самодиагностическому тесту PHQ-9. На всех записях присутствует фоновый шум.

Текстовый корпус эссе – корпус эссе длиной в одну страницу на тему «Я, другие, мир», записанный с целью определить лингвистические характеристики текстов людей с установленной депрессией и отсутствием депрессии. Также участников просили заполнить 10 опросников. Корпус информации из профилей социальной сети «Вконтакте» - были собраны данные профилей, также были получены баллы участников по шкале депрессии Бека. В корпусе также имеются изображения, собранные из альбомов, аватаров и постов в профилях социальной сети «Вконтакте» 398 волонтеров.

Существует несколько баз данных с агрессией в высказываниях и эмоциональных баз данных, содержащих категорию злость/агрессия (anger). Сравнение корпусов, содержащих высказывания с агрессией, детально представлено в таблице 5, где модальности указаны как A (аудио), B (видео), T (текст), C (сентимент).

Stress at Service Desk Dataset (SD) – содержит видеозаписи взаимодействий человек-человек в информационно-справочном центре. Для записи использовались 4 сценария, которые должны были вызвать стресс у испытуемых. В качестве испытуемых была выбрана группа актеров разных культурных групп (5 женщин и 4 мужчины), которые были разделены на две группы, а каждый сценарий воспроизводился дважды. Актерам не был предоставлен сценарий действий, а только краткое описание ситуации, им было необходимо импровизировать и

реагировать на оппонентов, таким образом были получены очень естественные сцены.

Таблица 5 – Сравнение корпусов, содержащих агрессивные высказывания

Корпус	**Язык данных**	**Количество модальностей**	**Количество информантов**	**Объем данных (мин)**	**Доступность**
SD [98]	Нидерландский, английский	3 (A, B, T)	8 актеров	32	По запросу
TR [99]	Английский	3 (A, B, T)	13 актеров	44	По запросу
CMU-MOSEI [100]	Английский	4 (A, B, T, C)	1000 людей	3953	Открытая
RAVDESS [101]	Английский	2 (A, B)	24 актера	7356 видеозаписей	По запросу
IEMOCAP [102]	Английский	3 (A, B, T)	10 актеров	720	По запросу
RAMAS [103]	Русский	3 (A, B, T)	10 актеров	420	По запросу

Aggression in Trains (TR) – корпус состоит из 21 сценария нежелательного поведения в поездах и на станциях (например, насилие, воровство, проезд без билета), которые были сыграны актерами. Сценарий действий не был предоставлен, только краткое описание ситуации.

CMU-MOSEI – многомодальная эмоциональная база данных, в которой содержатся видеозаписи, собранные с сервиса Youtube. Наиболее частые темы в базе данных среди 250: обзоры (16,2%), обсуждения/дискуссии (2,9%) и консультации/рекомендации (1,8%). Все видеозаписи были разбиты на предложения, которые были аннотированы по шкале сентимента. В аннотации также присутствуют следующие эмоции: счастье, грусть, злость, страх, отвращение и удивление. Все видеоданные были вручную транскрибированы.

RAVDESS является аудиовизуальной эмоциональной базой данных, в которой содержатся записи речи и пения профессиональных актеров. Среди размеченных эмоций в речи выделены спокойствие, радость, грусть, злость, страх, удивление и отвращение, а в пении – спокойствие, счастье, злость и страх. Каждая эмоция представлена двумя уровнями интенсивности.

База данных IEMOCAP содержит эмоционально окрашенные диалоги актеров, записанные с использованием специальных маркеров на лицах, голове и руках. Маркеры использовались для предоставления информации о выражении лиц актеров и их движений как в случае коммуникации по сценарию, так и в случае обычной разговорной речи. Разметка содержит следующие эмоции: счастье, злость, грусть, удивление, страх, отвращение, воодушевление, фрустрация/разочарование и нейтральное состояние. Также база данных аннотирована согласно категориям валентности, активации и доминантности.

RAMAS является русскоязычной многомодальной базой данных, в которой содержится речь актеров, воспроизводящих диалоги по короткому описанию различных ситуаций. Включает в себя такие эмоции как: злость, грусть, отвращение, счастье, страх и удивление. Помимо видео- и аудиозаписей база данных включает с датчиков движения и психофизиологические сигналы (например, электропроводимость кожи).

В ходе проведенного аналитического обзора был выявлен ряд основных сложностей решения задачи автоматического определения депрессии: 1) каждый поведенческий сигнал предоставляет только частичную информацию, которая может быть комбинированной формой более реалистичной модели определения поведенческих индикаторов депрессии; 2) некоторая полезная информация может быть недоступна или специально скрыта; 3) определение базового поведения может быть осложнено ввиду ограничений в поведенческих данных [35].

Наблюдаются изменения в речеобразовании у людей с депрессией после лечения, заключающиеся в изменениях тона голоса, громкости, частоты, артикуляции, беглости речи. В исследовании [75] показано, что психологические особенности человека влияют на особенности написанного им текста. Наиболее чувствительным к психологическим особенностям человека оказался показатель частоты лексики с аффективной семантикой. Авторы выявили, что при высоких показателях депрессивности и тревожности, чувстве собственной незначительности и сниженной стратегией самоконтроля чаще употребляется лексика протестного поведения [75].

1.6 Выводы по главе 1

В ходе проведения обзора обнаружен рост интереса к исследованиям, связанным с определением деструктивных паралингвистических явлений, проводятся различные соревнования и высокорейтинговые конференции, посвященные определению таких явлений. При этом многие из этих работ имеют ряд недостатков: 1) недостаточно высокое для применения систем на практике качество распознавания явлений, что особенно касается методов определения ложной и истинной информации в речи; 2) использование сложных нейросетевых архитектур, требовательных к вычислительным ресурсам, что ограничивает возможность экспериментальных исследований для значительного количества исследовательских групп ввиду невозможности регулярного обновления вычислительных мощностей; 3) большое время обучения моделей (обучение некоторых моделей может доходить до нескольких суток и даже недель), что может как вытекать из предыдущего пункта, так и идти в совокупности с ним; 4) отсутствие решений, анализирующих рассматриваемые деструктивные явления в совокупности, что, вероятно, является следствием того, что такая задача является комплексной. Кроме того, на данный момент не существует такого корпуса, который включал бы в себя все рассматриваемые в работе явления одновременно.

На основе проведенного анализа можно сформулировать потенциальные требования к программной системе определения деструктивных явлений в речи, а именно:

– Использование максимально возможного количества модальностей ввиду того, что специалистами учитываются все модальности при личной беседе. Кроме того, такой подход позволяет расширить возможности применения автоматических систем, так как будет возможность анализа дополнительных паралингвистических явлений.

– Результат верного распознавания деструктивных явлений должен быть как можно точнее, так как, например, как в случае с определением депрессии, сфера медицины относится к тем сферам применения, где ложные срабатывания

автоматической системы могут быть критическими. То же можно отнести и к определению лжи и агрессии, в этих случаях ложные срабатывания могут оказать негативное влияние на репутацию людей.

– Автоматические системы должны проходить тестирование в максимально приближенных к реальной жизни условиях.

– Апробация в реальной жизни на этапе тестирования должна проходить под контролем специалистов, которые могли бы подтвердить верные результаты классификации или скорректировать ложные.

Поскольку сбор данных для обучения в системах автоматического определения деструктивных явлений в речи является трудоемким и сложным процессом из-за специфики задач, то существующие на данный момент корпусы имеют относительно небольшое количество данных, а также зачастую имеют дисбаланс в количестве экземпляров в классах обучающих данных. Для решения такой проблемы при разработке таких систем возможно использование различных показателей точности работы, учитывающих дисбаланс в количестве экземпляров в классах, методов аугментации данных и выбора информативных признаков. В описанных выше работах для выбора информативных признаков использовались следующие методы: выбор признаков на основе корреляции, метод главных компонент, метод частных наименьших квадратов и др.

При разработке автоматических систем для определения деструктивных явлений в разговорной речи используются различные методы вычисления акустических признаков, машинного и глубокого обучения. Для всех рассматриваемых деструктивных паралингвистических явлений применяются как детерминированные методы машинного обучения, так и нейросетевые. А в качестве акустических признаков лидирующие места занимают экспертные признаки (например, openSMILE), однако в последние годы экспериментальные исследования также проводятся с нейросетевыми методами извлечения акустических признаков (например, ResNet, DenseNet и др.). Такие методы исследуются в следующей главе.

2 МАТЕМАТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДЛЯ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ ДЕСТРУКТИВНОГО ПОВЕДЕНИЯ В РАЗГОВОРНОЙ РЕЧИ

В данной главе приводится описание и исследование базовых и усовершенствованных методов вычисления акустических признаков, машинного и глубокого обучения, которые используются при разработке систем автоматического определения деструктивных паралингвистических явлений в разговорной речи. Кроме того, приводится математическая постановка решаемой задачи и формальное описание методики интегрального оценивания степени выраженности деструктивных паралингвистических явлений в речи диктора, используемой в предложенной программной системе.

2.1 Математическая постановка задачи

Формальная постановка задачи классификации объектов (аудиозаписей). Пусть имеется множество аудиозаписей $S = (s_1,\ldots,s_m)$ и множество меток классов $Y = (y_1,\ldots,y_m)$ этих аудиозаписей. Существует неизвестная целевая зависимость – отображение $A : X \to Y$, при этом ее метки классов известны только для векторов объектов-признаков аудиозаписей конечной обучающей выборки $X = \{(x_1,y_1),\ldots,(x_m,y_m)\}$, а X получен с использованием метода вычисления акустических признаков из аудиозаписей $F : S \to X$. Тогда требуется найти метод $A : X \to \hat{Y}$, который сможет классифицировать вектор объекта-признака x множества X. Здесь множество меток классов $Y = (y_1,\ldots,y_m)$ описывает истинные значения классов объектов обучения, а множество меток классов $\hat{Y} = (\hat{y}_1,\ldots, \hat{y}_m)$ описывает значения результатов классификации.

В нашем случае необходимо найти множество методов $A = \{A_{dec}, A_{agg}, A_{depr}\}$ для методов определения ложной/истинной информации, агрессии и депрессии в речи (A_{dec} – deception, A_{agg} – aggression, A_{depr} - depression)**:**

$$A_{dec} : X_{dec} \to \hat{Y}_{dec}, \quad A_{agg} : X_{agg} \to \hat{Y}_{agg}, \quad A_{depr} : X_{depr} \to \hat{Y}_{depr}, \tag{1}$$

где входные данные представлены вектором объектов-признаков $X = (x_1,x_2,\ldots,x_l)$ длины l, а целевые значения меток классов $\hat{y}$ множества $\hat{Y}$ представлены либо

бинарными значениями {0, 1}, где для множества $\hat{Y}_{dec}$ 0 обозначает истинное высказывание, а 1 – ложное; для множества $\hat{Y}_{depr}$ 0 обозначает отсутствие депрессии, а 1 – ее наличие, либо конечным множеством {0, 1, 2}, где для множества $\hat{Y}_{agg}$ 0 обозначает низкий уровень агрессии или ее отсутствие, 1 – средний уровень агрессии, а 2 – высокий уровень агрессии.

2.2 Комплекс методов анализа речевого сигнала для определения деструктивных паралингвистических явлений в разговорной речи

Комплекс методов анализа речевого сигнала для определения деструктивных паралингвистических явлений в разговорной речи представлен на рисунке 4. На вход программной системы подаются речевые данные *S*. Из этих данных с использованием метода вычисления интегральной совокупности акустических признаков вычисляются оригинальные наборы векторов-признаков $\{X_{dec}, X_{agg}, X_{depr}\}$. Полученные наборы признаков подаются на вход комплекса методов определения деструктивных паралингвистических явлений в речи *A*, состоящего из методов определения ложной и истинной информации A_{dec}, агрессии A_{agg} и депрессии A_{depr} в речи. Данные методы работают в иерархическом порядке: результаты классификации методов определения ложной и истинной информации и агрессии в речи $\hat{y}_{dec}$, $\hat{y}_{agg}$ в бинарном виде {0, 1} объединяются с вектором признаков X_{depr}, поступающим на вход метода определения депрессии в речи. Далее результаты классификации $\{\hat{y}_{dec}, \hat{y}_{agg}, \hat{y}_{depr}\}$ всех трех методов являются входными данными для предложенной методики, в результате использования которой интегральная оценка степени выраженности деструктивных паралингвистических явлений в речи I_{int}.

Данный комплекс построен на базе трех методов определения деструктивных паралингвистических явлений (лжи, агрессии, депрессии) и методики интегральной оценки деструктивных паралингвистических явлений в разговорной речи, которые описываются далее.

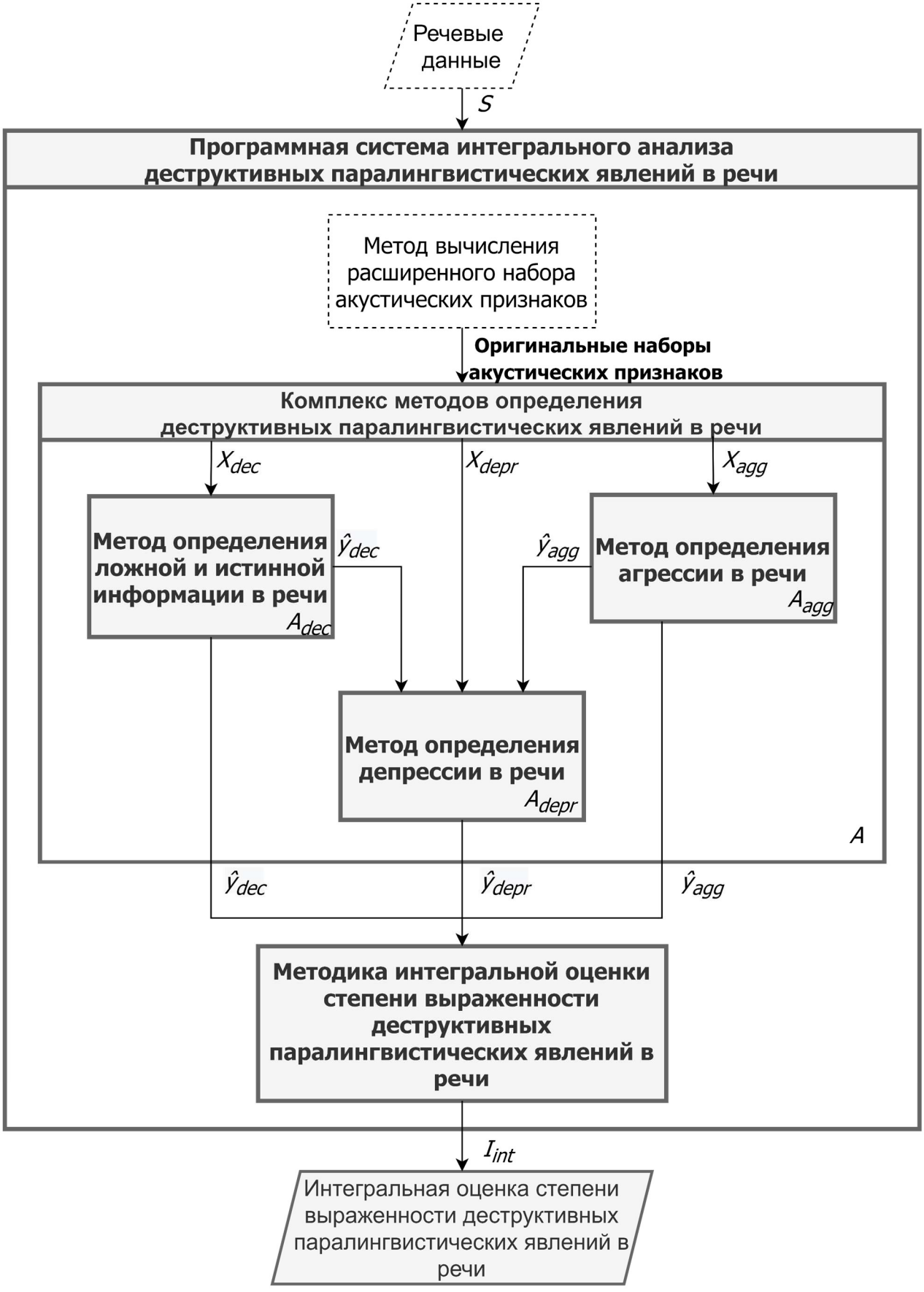

Рисунок 4 – Комплекс методов анализа речевого сигнала для определения деструктивных паралингвистических явлений в разговорной речи

2.3 Базовые методы вычисления акустических признаков для автоматического определения паралингвистических явлений в разговорной речи

Акустические признаки, имеющие корреляцию с психоэмоциональным состоянием диктора, проявляются на относительно длительных временных участках речевого сигнала, которые соответствуют как минимум одному произнесенному слову (супрасегментные признаки). Системы автоматического определения паралингвистических явлений в разговорной речи используют пространства акустических признаков большого размера, например, низкоуровневые энергетические, спектральные и просодические признаки, например: частота основного тона (ЧОТ, F0); мел-частотные кепстральные коэффициенты (MFCC); форманты (резонансные частоты голосового тракта); коэффициенты перцептивного линейного предсказания (PLP); модулированный спектр сигнала; энергетические признаки; характеристики вариативности и т.д.

Одним из наиболее популярных наборов таких признаков является набор openSMILE (для соревнований ComParE). Он содержит более 6 тысяч компонент в векторах признаков, которые можно вычислить при помощи открытого программного инструментария openSMILE (http://audeering.com/technology/opensmile/).

Эффективным также является использование искусственных нейронных сетей для вычисления информативных признаков из аудиосигнала. К таким нейронным сетям относятся: VGG-16 [104], AlexNet [105], DenseNet [106] в различных архитектурах и другие. Вектор признаков, полученный с их помощью, вычисляется следующим образом: спектральные изображения речи подаются на вход предобученной для распознавания образов сверточной нейронной сети CNN, а на выходе вычисляется вектор признаков.

Наборы акустических признаков openSMILE ComParE_2016, openXBOW BoAW, DeepSpectrum DenseNet и AuDeep популярны при определении эмоционального состояния диктора, например, они были использованы в работе [107].

На рисунке 5 представлены известные нейросетевые методы и программные инструментарии, использующиеся для вычисления акустических признаков при разработке систем определения паралингвистических явлений. Также на нем представлены основные информативные признаки депрессии, которые можно получить по аудиосигналу при использовании указанных методов вычисления признаков.

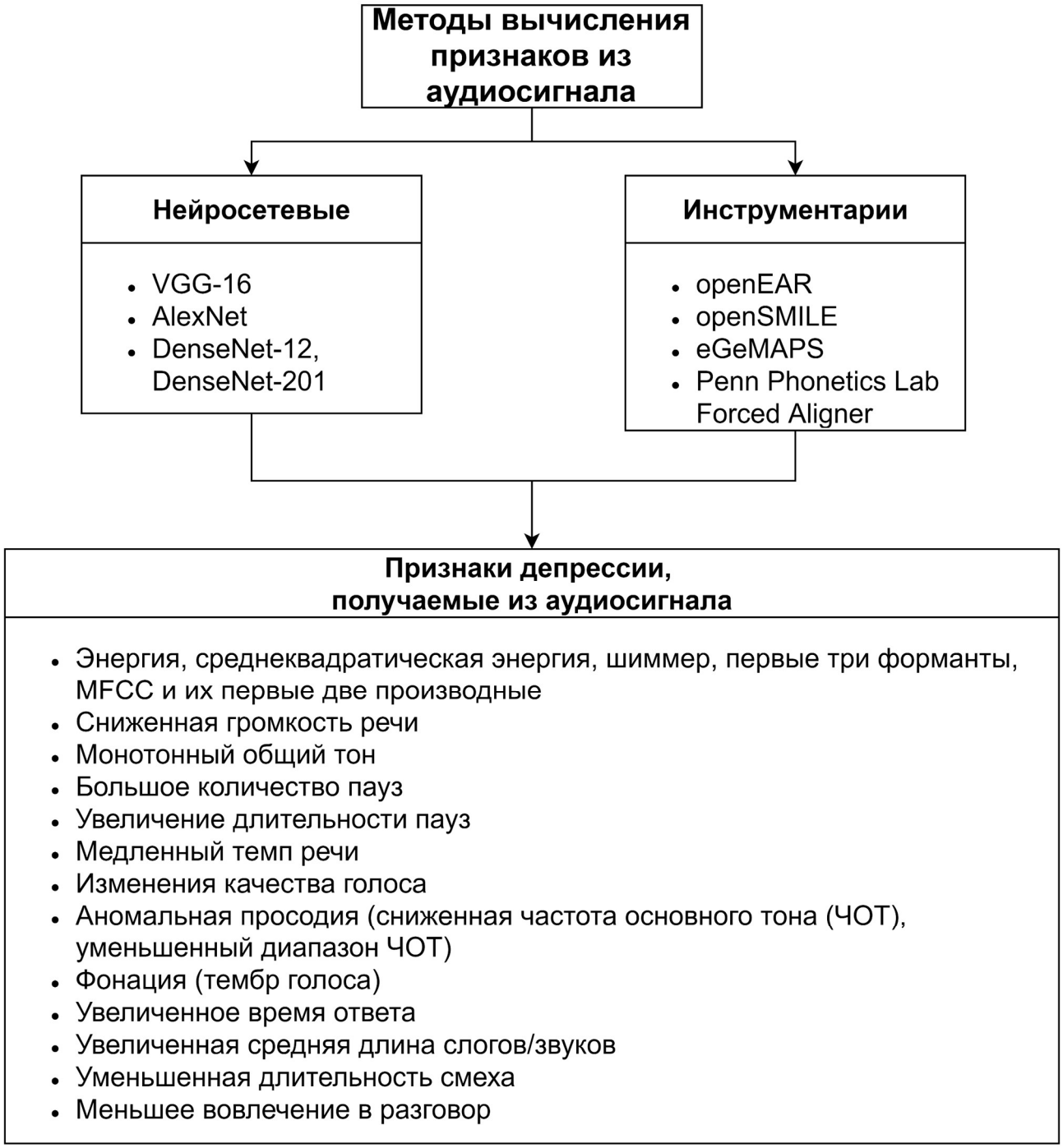

Рисунок 5 – Систематизация базовых методов вычисления информативных признаков из аудиосигнала и основные признаки депрессии

В данной работе применяются, исследуются и комплексируются как нейросетевые методы вычисления акустических признаков (VGG-16, DenseNet), так и экспертные наборы акустических признаков (openSMILE).

2.4 Базовые методы классификации для автоматического определения деструктивных явлений в разговорной речи

Для решения задач классификации существует множество методов, среди которых можно выделить две категории: классические (детерминированные) и нейросетевые методы. В этом и следующем разделах будут рассмотрены наиболее часто применяющиеся методы машинного обучения и классификации обеих категорий в задаче определения деструктивных явлений в разговорной речи.

Например, для решения задач определения деструктивных паралингвистических явлений чаще всего применяются методы классификации, представленные на рисунке 6. Подробное описание используемых методов машинного обучения представлено в разделах 2.4.1-2.4.2.

Реализации данных методов могут быть как запрограммированы самостоятельно, так и взяты из программных библиотек, например, WEKA [108], Scikit-learn, Keras, Tensorflow, и др.

2.4.1 Детерминированные методы классификации для автоматического определения деструктивных явлений в разговорной речи

Детерминированные методы классификации представляют собой классические статистические алгоритмы, в которых решение задачи происходит на основе данных. Классическое обучение применяется в случаях, когда имеются простые данные и понятные признаки, оно может быть как с учителем (классификация, регрессия), так и без учителя (кластеризация, поиск правил, общее уменьшение размерности). Отдельными ветвями машинного обучения являются ансамблевые методы (бустинг, стэкинг, бэггинг), обучение с подкреплением (генетические алгоритмы и др.) и нейросетевые методы (описаны далее в разделе 2.3.2), первые применяются в случаях, когда классические методы по каким-то причинам недостаточно хорошо справляются с поставленной задачей (чаще всего,

такими причинами являются качество данных или низкие показатели классификации).

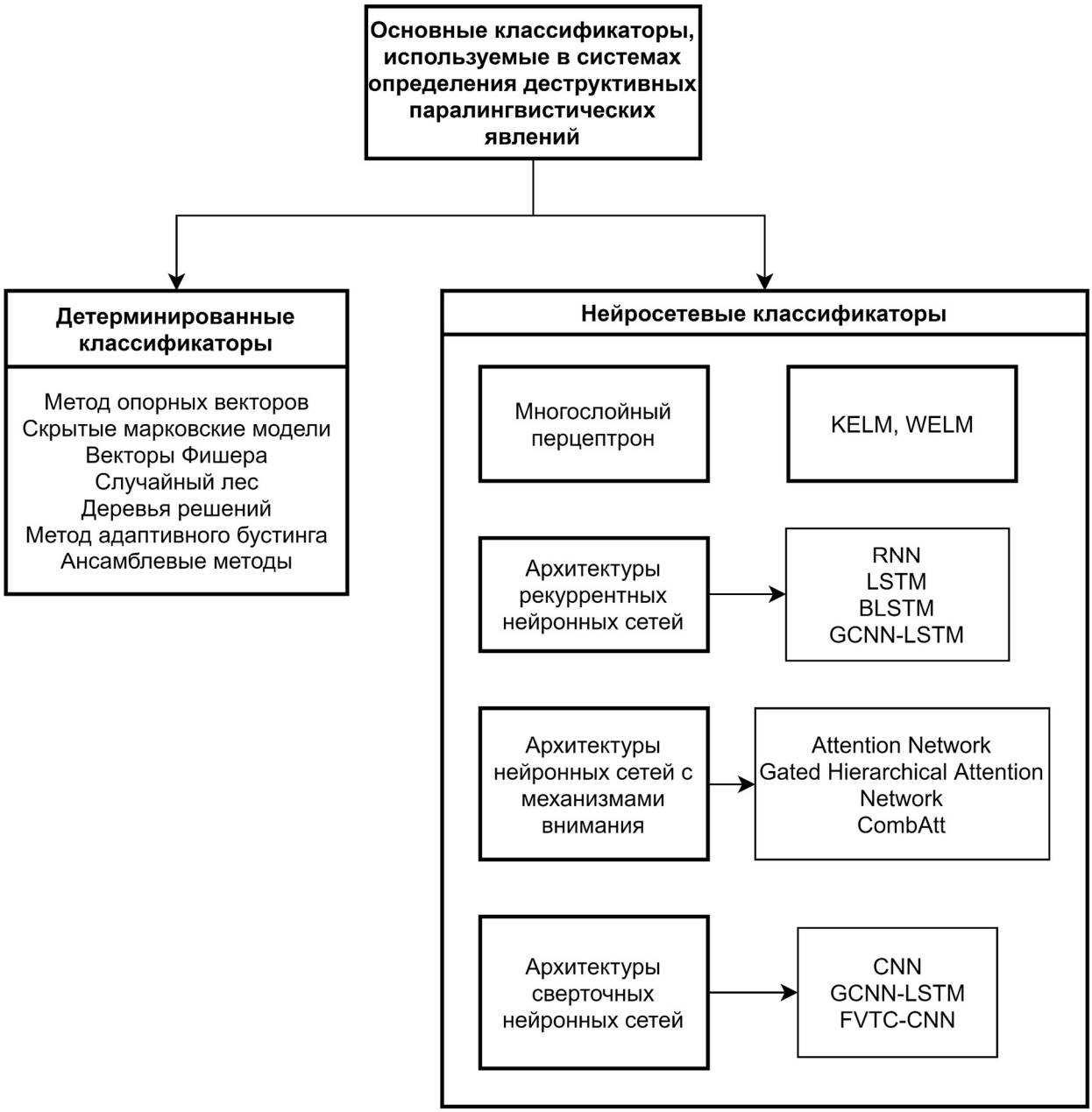

Рисунок 6 – Систематизация базовых методов классификации деструктивных паралингвистических явлений в разговорной речи

Бэггинг (Bagging) [109] – метод классификации, который использует композиции алгоритмов. Результат классификации определяется голосованием. В данном методе все элементарные классификаторы работают параллельно и проводят обучение независимо: при этом они не исправляют ошибки друг друга, а компенсируют их во время голосования.

Метод k ближайших соседей (k-Nearest Neighbours, k-NN) [110] –простейший метрический классификатор, который базируется на оценивании сходства объектов. Суть метода в том, что объект присваивается тому классу, которому принадлежат ближайшие к нему объекты (т.е. соседи) обучающей выборки. В основе метода лежит предположение о том, что в признаковом пространстве близким объектам соответствуют похожие метки. Алгоритм принимает во внимание не только количество определенных классов, но и удаленность от нового значения.

Классификация путем регрессии (Classification via Regression, CvR) [111]. Регрессия – метод, используемый для измерения отношения между зависимыми и независимыми переменными путем некоторой эмпирической функции. Используемый классификатор комбинирует обычное дерево решений с возможностью линейной регрессии на узлах.

Алгоритм адаптивного бустинга (Adaboost) [112] – алгоритм классификации, в процессе которого строится композиция из базовых классификаторов для улучшения их эффективности. В нем каждый классификатор обучается на поднаборах объектов, которые были плохо классифицированы предыдущими классификаторами. Сначала алгоритм вызывает слабый классификатор, затем после каждого вызова обновляется распределение весов, отвечающих за важность каждого из объектов обучающего множества для классификации. На каждой итерации веса каждого неверно классифицированного объекта возрастают, благодаря чему новый классификатор специализируется на распознавании таких объектов.

Последовательная минимальная оптимизация (Sequential Minimal Optimization, SMO) [113] – является одним из алгоритмов для решения задачи квадратичного программирования (через поиск двух множителей Лагранжа). Заключается в том, что выделяются пары переменных α для пары базовых векторов, наиболее близко примыкающих к границе классов и лежащих по разные стороны от границы. Для пары переменных с ограничениями решение задачи квадратичного программирования строится аналитически в явной форме. После

того, как пройден шаг оптимизации одна (или обе) переменная обращается в ноль, а соответствующие вектора исключаются из списка кандидатов. Алгоритм завершается, когда становится невозможным нахождение пары векторов, оптимизация параметров которых может улучшить решение.

Метод стохастического градиента (Stochastic Gradient Descent, SGD) [114] – использует алгоритм стохастического градиентного спуска для обучения линейных моделей. Стохастический градиентный спуск – это инкрементальный алгоритм с отсечением по времени, поэтому он может применяться к большим потокам или наборам данных.

Случайный лес (Random Forest, RF) [115] – множество деревьев решений, ответы которых усредняются (в задачах регрессии) или принимается решение голосованием по большинству из них (в задачах классификации). Деревья строятся независимо по определенной схеме: сначала из обучающей выборки берется подвыборка и по ней строится дерево (подвыборки разные для каждого дерева), затем вычисляется наилучший признак и по нему происходит расщепление. Дерево строится до тех пор, пока в листьях не останутся представители только одного класса, но существуют реализации, которые ограничивают высоту дерева, число объектов в листьях и число объектов в подвыборке, при котором происходит расщепление. Метод основан на бэггинге, но дополняет его, случайным образом выбирая подмножество признаков на каждом узле, чтобы сделать деревья более независимыми.

Метод деревьев решений (Decision Tree, DT) [116] – простой и наглядный метод классификации, который может использоваться в задачах регрессии, если предсказываемый результат можно обозначить как вещественное число. Структура дерева состоит из листьев, ветвей и узлов: в ветвях содержатся значения атрибутов, от которых зависит целевая функция, на листьях записываются значения этой функции, а узлы различаются на родительские и дочерние, по которым проходит разветвление.

Метод случайного дерева (Random Tree, RT) – метод на основе дерева, которое содержит k случайно выбранных атрибутов для каждого узла, при этом не используется обрезка ветвей.

PART [117] – метод обучения списков решений, основанный на повторяющемся построении частичных деревьев решений в стиле стратегии «отделения и захвата». Основным преимуществом PART перед другими похожими алгоритмами (C4.5, C5.0, RIPPER) является его простота, так как при объединении двух парадигм обучения правил, он создает хороший набор правил без необходимости глобальной оптимизации.

Метод одного правила (One Rule, OneR) [118] – простой способ классификации, который генерирует одно правило для каждого предсказания, а затем выбирает правило, которое имеет наименьшую ошибку, в качестве основного.

Классификация опорных векторов (Support Vector Classifier, SVC) [119] – задача классификации, решаемая методом опорных векторов. Метод опорных векторов решает задачи классификации и регрессии с помощью построения нелинейной плоскости, разделяющей решения. Метод основан на концепции гиперплоскостей, которые определяют границы гиперповерхностей. Гиперповерхность является обобщением трехмерной поверхности для случая евклидова пространства произвольной размерности. Гиперплоскость – это подпространство с размерностью, на единицу меньшей, чем объемлющее пространство. Гиперплоскость делит пространство соответствующей размерности на два полупространства, все точки каждого из которых определяются неравенствами. Разделяющая гиперплоскость – это гиперплоскость, которая отделяет группу объектов, имеющих различную классовую принадлежность.

Стэкинг (Stacking) [120] – популярный способ ансамблирования методов для различных задач машинного обучения. Суть метода состоит в обучении базовых методов с использованием перекрестной валидации, где на основе их прогнозов при помощи метамодели делается итоговое предсказание.

Градиентный бустинг (Gradient boosting) [121] – мощный метод машинного обучения, который позволяет достичь высоких результатов классификации/регрессии в различных практических задачах. Данный метод хорошо справляется в ситуациях, когда в данных присутствуют неоднородные признаки, шумы, сложные зависимости и пр. Градиентный бустинг, фактически, является процессом построения ансамбля предсказателей путем градиентного спуска в функциональное пространство. При градиентном спуске обновляются предсказания, основанные на скорости обучения, и ищутся значения, на которых функция потерь минимальна. Таким образом, предсказания обновляются так, чтобы сумма отклонения стремилась к нулю, а предсказанные значения были максимально близки к реальным. Алгоритмически весь процесс может быть описан следующими шагами:

1. Построение простых моделей и анализ ошибок.
2. Определение точек, не вписывающихся в простую модель.
3. Добавление моделей для обработки сложных случаев, выявленных на начальной модели.
4. Сбор построенных моделей и определение весов для каждого предсказателя.

В данной работе исследуются и сопоставляются все вышеописанные методы (реализации взяты из программной библиотеки Scikit-learn), в том числе три различных реализации градиентного бустинга на основе деревьев решений, которые являются лучшими на момент написания работы. Catboost (реализация градиентного бустинга, разработанная компанией Яндекс [122]), LightGBM (реализация градиентного бустинга, разработанная компаниями Microsoft and LightGBM Contributors [123]) и XGBoost (реализация градиентного бустинга, разработанная компанией The XGBoost Contributors [124]).

2.4.2 Нейросетевые методы для автоматического определения деструктивного поведения в разговорной речи

Нейросетевые методы являются одним из видов машинного обучения. В последнее время они стали крайне популярны в задачах классификации, в том числе и при определении деструктивного поведения в разговорной речи. Нейросетевые методы включают в себя несколько категорий: перцептроны, автоэнкодеры, генеративно-состязательные нейросети, рекуррентные нейросети и сверточные нейросети. Нейросетевые методы применяются тогда, когда имеются сложные данные и большое количество объектов обучения (при малом количестве объектов нейросетевым методам сложно качественно обучиться).

Искусственная нейронная сеть (ИНС, Artificial Neural Network, ANN [125]) – является математической моделью упрощенной биологической нейронной сети мозга живого организма. На данный момент существует большое количество типов искусственных нейронных сетей, однако в сфере анализа и распознавания речи чаще используются глубокие нейронные сети (Deep Neural Networks, DNN), свёрточные нейронные сети (Convolutional Neural Networks, CNN) и рекуррентные нейронные сети (Recurrent Neural Networks, RNN).

Рекуррентная нейронная сеть (Recurrent Neural Network, RNN [126]) – разновидность нейронной сети, в которой связи между элементами составляют направленную последовательность. Благодаря такой архитектуре они могут обрабатывать серии событий во времени или последовательные цепочки в пространстве. Используют внутреннюю память при обработке последовательностей произвольной длины.

Сеть длинной краткосрочной памяти (Long Short-Term Memory, LSTM [127]) является разновидностью рекуррентной нейронной сети, которая способна обучаться долговременным зависимостям.

TabNet [128] является глубокой нейронной сетью для работы с табличными данными. Основными качествами данной сети являются высокая производительность и, что важно, интерпретируемость. Сеть состоит из

полносвязных слоев с последовательным механизмом внимания. Алгоритмически процесс работы данной сети может быть описан следующими шагами:

1. Разреженный выбор объектов по экземплярам на основе обучающего набора данных.

2. Создание последовательной многоступенчатой архитектуры, основанную на выбранных функциях, где каждый шаг принятия решений может внести свой вклад в часть решения.

3. Улучшение способности к обучению за счет нелинейных преобразований выбранных функций.

4. Имитирование ансамблевого метода путем привлечения более точных измерений и увеличения шагов улучшения решения.

Интерпретируемость сети выполняется за счет маскирования каждого шага и каждого наблюдения выборки так, что можно получить агрегированную маску для всех шагов принятия решений, а маски выбора можно визуализировать.

В данной работе исследуются метод TabNet, реализация которого взята из библиотеки Pytorch.

2.5 Предложенный метод для автоматического определения ложных и истинных речевых сообщений

В данном разделе представлено описание предложенного метода для автоматического определения ложных и истинных речевых сообщений, разработанного на основе базовых методов, описанных в разделах 2.3 и 2.4.

Иерархический метод паралингвистического анализа речи Hierarchical Two-level Boosting-based Method for Deception detection – HTLBbM-Deception основывается на двухуровневом подходе на базе градиентных бустингов (Two-Level Boosting-based Method, TLBbM) и использовании паралингвистической информации о половой принадлежности и эмоциональном состоянии диктора. Метод HTLBbM-Deception включает в себя следующие шаги: 1) определение пола диктора; 2) определение эмоционального состояния диктора; 3) определение ложности/истинности речевого высказывания. Основанием для разработки

системы такой иерархией было значительное количество работ, в которых исследовалось влияние пола диктора на его эмоциональные проявления (например, [129, 130]) Детально схема иерархического метода паралингвистического анализа речи представлена на рисунке 7.

Данный метод отличается тем, что для вычисления акустических признаков используются несколько наборов акустических признаков. Кроме того, перед обучением расширен вектор признаков – использована информация о половой принадлежности диктора и информация о ее/его эмоциональном состоянии, которые предсказывались с использованием нейросетевого метода с временной задержкой и метода опорных векторов, соответственно.

Для обучения акустической модели на шаге 3 использованы три различных реализации метода градиентного бустинга: Catboost, XGBoost, LightGBM. Для объединения результатов классификации трех методов градиентного бустинга использовался метод стекинга (stacking). Стекинг в данном методе представляет собой двухуровневый подход, где на первом уровне находятся три метода градиентного бустинга, а на втором – логистическая регрессия. Логистическая регрессия здесь объединяет результаты классификации, полученные от классификторов первого уровня, и вычисляет итоговый результат классификации ложности или истинности речевого высказывания.

Для того, чтобы сбалансировать классы в данных для обучения, применяется метод искусственного увеличения количества объектов миноритарного класса SMOTE (подробнее об аугментации ниже в разделе). Ввиду того, что общая размерность полученного вектора признаков является большой, необходимо использовать методы уменьшения размерности признакового пространства.

Существует несколько подходов к работе с несбалансированными классами в данных:

1. Сбор большего количества данных.
2. Использование других методов классификации. Например, деревья решений хорошо справляются с классификацией при несбалансированных данных (методы C4.5, C5.0, CART, случайный лес и др.).

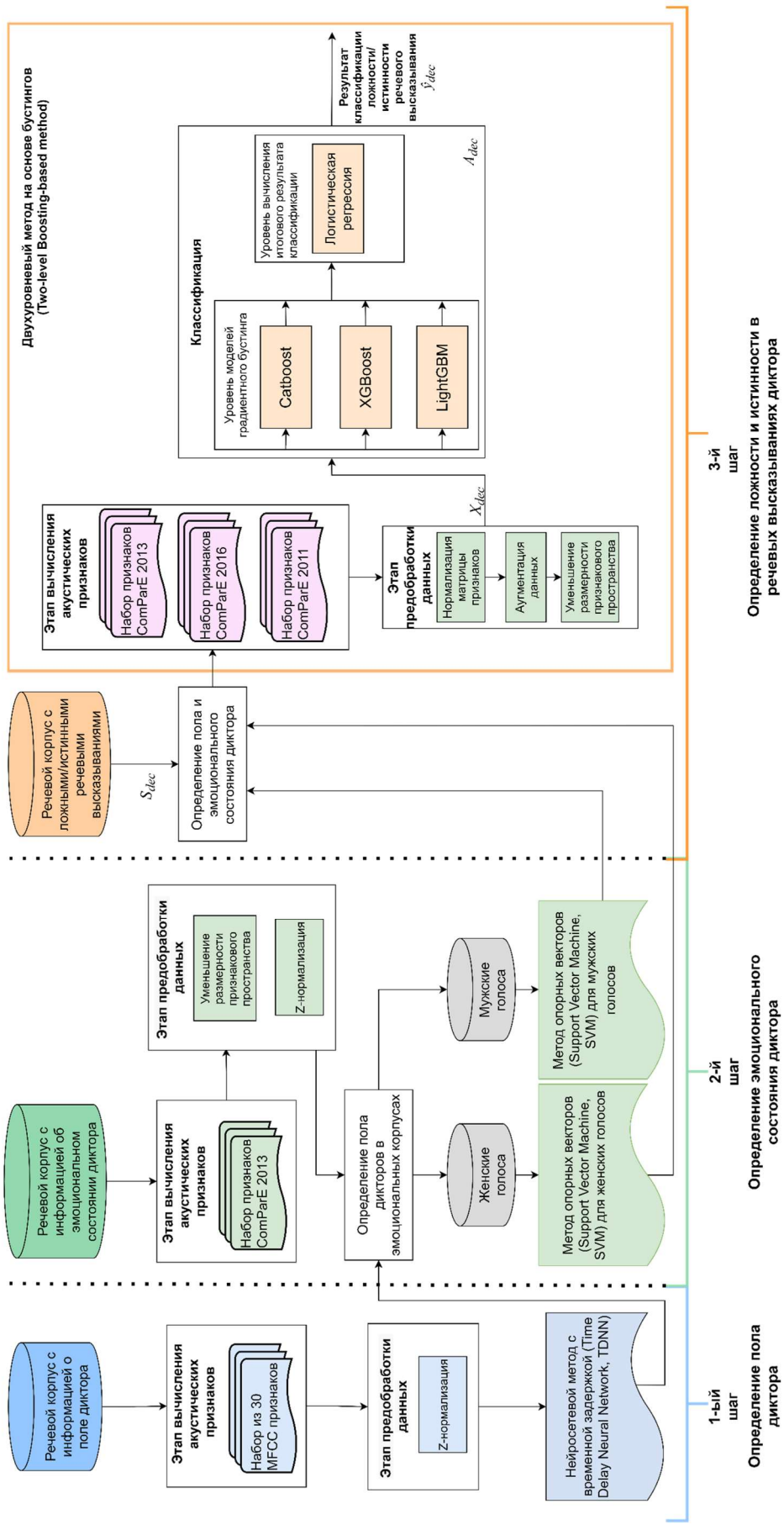

Рисунок 7 – Схема иерархического метода паралингвистического анализа речи для определения ложности в разговорной речи HTLBbM-Deception

3. Использование методов со штрафами. Такие методы предусматривают увеличение весов модели, которая сделала ошибку в предсказаниях миноритарного класса (метод линейного дискриминантного анализа со штрафами, метод опорных векторов со штрафами и др.). Такой подход может быть использован в случае, когда необходимо использовать конкретный метод классификации и при этом невозможно пересобрать набор данных для обучения.

4. Использование другого подхода к обучению моделей, например, поиска аномалий или определения изменений (в последнее время в некоторых практических задачах эффективен метод обучения по единичным объектам, one-shot learning).

Существует два основных направления обработки данных при работе с несбалансированными данными. Первое направление подразумевает удаление части объектов обучения мажоритарного класса (undersampling). При применении второго направления искусственно создаются новые объекты миноритарного класса (oversampling). Необходимо также отметить, что удаление части объектов используется, в основном, в тех случаях, когда количество данных большое (десятки и сотни тысяч объектов), а искусственное создание новых объектов миноритарного класса чаще используется, в ситуациях, когда данных меньше.

Удаление части объектов обучения мажоритарного класса может быть выполнено с использованием следующих методов:

– Случайное удаление части объектов обучения мажоритарного класса (Random undersampling). Данный подход заключается в том, что происходит подсчет количества объектов мажоритарного класса, которое необходимо удалить, чтобы достигнуть оптимального баланса классов в данных. После проведения этого подсчета случайным образом удаляются объекты мажоритарного класса с замещением или без замещения.

– Связи Томека (Tomek links) [131]. При применении данного подхода происходит удаление таких объектов мажоритарного класса, которые

накладываются или оказываются слишком близки к объектам миноритарного класса. Удаление происходит до тех пор, пока все ближайшие соседи не окажутся одного класса. Метод широко используется для удаления шумов из набора данных.

– Правило концентрации ближайших соседей (Condensed Nearest Neighbor Rule). Основной целью использования данного подхода является обучение классификатора обнаружению различий между похожими объектами обучения, принадлежащими к разным классам.

– Односторонняя выборка (One-side sampling, One-Sided Selection). Данный подход комбинирует подход связей Томека и подход правила концентрации ближайших соседей. Сначала при использовании подхода связей Томека определяются границы классов и идентифицируется мажоритарный класс, а затем используются правила концентрации ближайших соседей для удаления части таких объектов мажоритарного класса, которые находятся близко к границам классов.

– Правило «очищающего» соседа (Neighborhood cleaning rule). Основная цель данного подхода заключается в удалении всех объектов, которые негативно влияют на классификацию объектов миноритарного класса. На первом шаге происходит классификация данных с использованием метода k-ближайших соседей со значением k = 3. На втором шаге удаляются объекты мажоритарного класса, классифицированные верно и соседи неверно классифицированных объектов мажоритарного класса.

Увеличение количества объектов в миноритарном классе может быть выполнено с использованием следующих методов:

– Искусственное создание объектов миноритарного класса (oversampling). В зависимости от необходимого баланса между классами в данных, этот подход случайным образом выбирает объекты миноритарного класса для копирования.

– Искусственное создание новых объектов миноритарного класса (SMOTE, Synthetic Minority Oversampling Technique). В отличие от подхода выше

не копирует объекты миноритарного класса, а создает новые объекты, похожие на объекты миноритарного класса. С использованием метода k-ближайших соседей, создается вектор между объектами миноритарного класса, затем этот вектор умножается на случайное число в диапазоне от 0 до 1. Новые объекты создаются при сложении полученного значения и изначального значения. Степень похожести объектов может контролироваться изменением параметра k в методе k-ближайших соседей. Также возможно задать необходимое количество объектов для создания. Минусами данного подхода является увеличение размерности объектов миноритарного класса, что может создать искусственный шум в данных при равномерном распределении объектов мажоритарного класса.

– Адаптивное создание новых объектов миноритарного класса (ADASYN, Adaptive Synthetic Minority Oversampling). Данный подход похож на подход SMOTE, но отличается тем, что использует функцию размерности для автоматического определения количества объектов, которые необходимо создать для каждого объекта миноритарного класса. Таким образом, вес объектов миноритарного класса изменяется адаптивно в зависимости от того, насколько он является сложным для обучения классификатора. Так, новые объекты создаются, в основном, рядом с теми объектами, которые являются сложными для обучения классификатора.

2.6 Предложенный метод для определения депрессии в разговорной речи

Метод для определения депрессии в разговорной речи (Single Binary Classifier for Depression Detection – SBC-Depression) основан на индивидуальном классификаторе с использованием нескольких наборов акустических признаков. Схема предложенного метода определения депрессии в разговорной речи SBC-Depression представлена на рисунке 8. В этом методе также вычисляются и обрабатываются несколько наборов акустических признаков: eGeMAPS и признаков, вычисленных при помощи нейросети, DenseNet, из аудиоданных речевого корпуса.

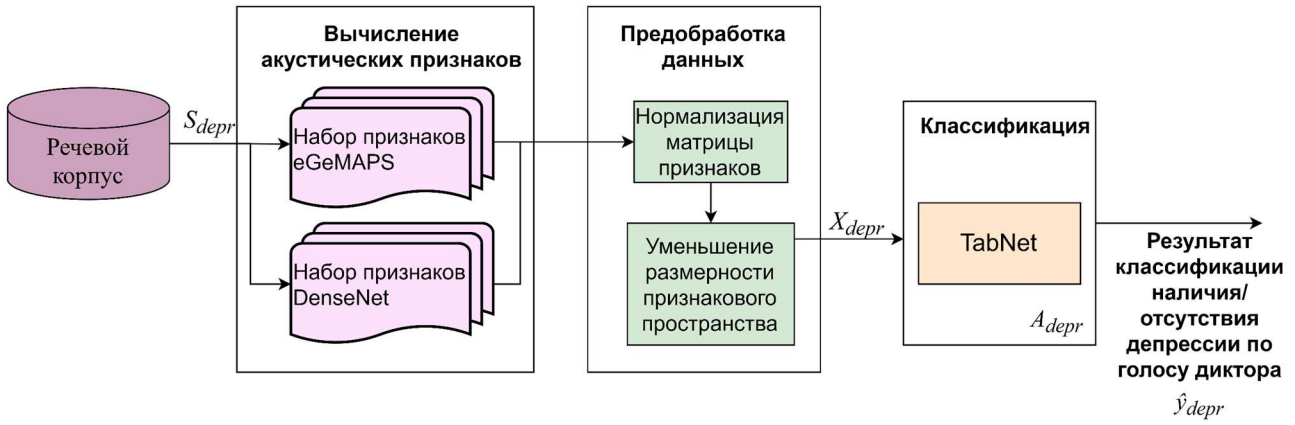

Рисунок 8 – Схема метода определения депрессии в разговорной речи SBC-Depression

Набор акустических признаков eGeMAPS, состоящий из 88 признаков, является расширенным набором признаков GeMAPS. Данный набор включает в себя множество категорий информации о речевом сегменте. Каждый признак выбран согласно его ассоциации с признаками качества голоса или эмоции. Для вычисления вектора признаков с использованием нейронной сети DenseNet используются спектрограммы аудиосигнала, которые поступают на вход предобученной нейронной сети. В результате на выходе нейронной сети получается вектор, содержащий 1024 признака.

Далее признаки подвергаются нормализации, а затем уменьшению размерности признакового пространства, в результате чего формируется оригинальный набор акустических признаков. Акустические признаки, хотя и вычислены оконным методом с наложением, в итоге усреднены так, что для каждого аудиофайла в наборах для обучения и тестирования записан один вектор признаков. При вычислении признаков DenseNet не использовано усреднение для каждого аудиофайла, признаки предоставлены в дискретном виде, и каждого окна имеется свой вектор признаков.

Для бинарной задачи определения депрессии (присутствует заболевание у говорящего или нет) используется архитектура глубокой нейронной сети TabNet. Данная архитектура разработана специально для табличных данных, которыми является вектор признаков, подающийся на вход классификатора. Нейросеть состоит из нескольких полносвязных слоев с последовательным механизмом

внимания. Каждый слой этой нейросети является шагом решения, содержащий в себе блок с полносвязными слоями для преобразования характеристик и механизм внимания для определения важности входных оригинальных характеристик. Кроме того, TabNet отличает то, что она требует небольших вычислительных ресурсов, причем обучение занимает малое количество времени, что несомненно повышает ее конкурентоспособность при ограниченных в ресурсах исследованиях.

2.7 Предложенный метод для определения агрессии в разговорной речи

Метод на основе нескольких наборов акустических признаков и ансамбля из методов случайного леса (Ensemble-based Model for Agression Detection) для решения задачи мультиклассовой классификации состояния агрессии диктора. Схема предложенного метода для определения агрессии по речи представлена на рисунке 9. На вход метода могут подаваться обучающие данные, которые размечены согласно трем уровням агрессии: низкий, средний и высокий. Из аудиоданных вычисляются наборы акустических признаков ComParE 2013, DenseNet и auDeep.

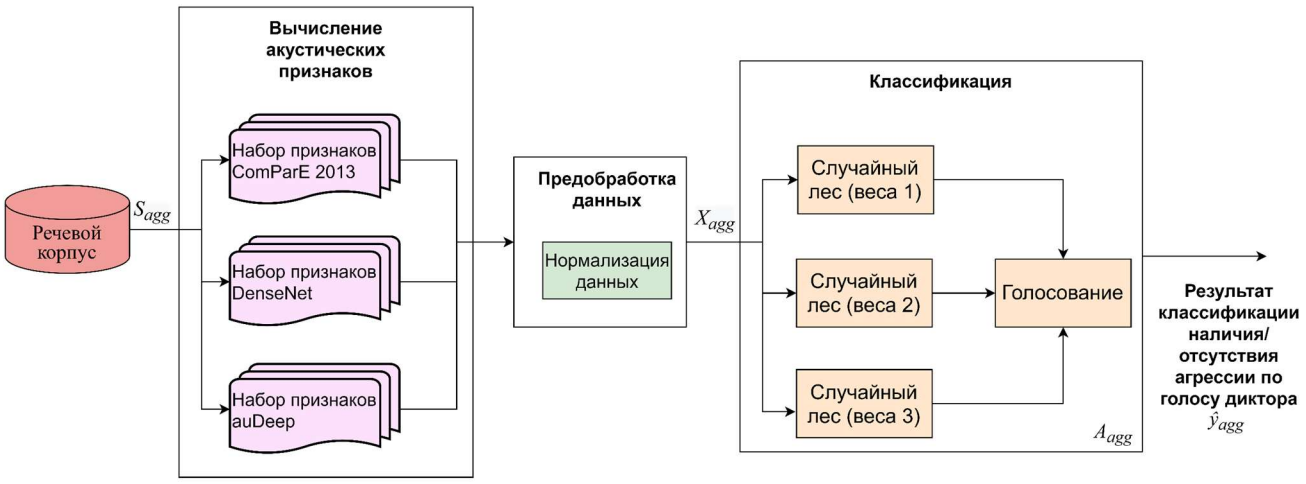

Рисунок 9 – Схема метода определения агрессии в разговорной речи EbM-Aggression

Набор признаков auDeep был вычислен с использованием нейронной сети AuDeep. Она моделирует последовательную природу аудиосигнала рекуррентными нейронными сетями внутри кодера и декодера. Сначала из сырых аудиофайлов вычисляются мел-спектрограммы и из них формируется набор данных. Чтобы избавиться от фоновых шумов, уровни мощности на этих

спектрограммах обрезаются ниже четырех заданных пороговых значений, что в результате дает четыре отдельных набора спектрограмм для каждого набора данных. После чего каждый автоэнкодер рекуррентной последовательности обучается (без учителя) на каждом из этих наборов.

Таким образом, обученные представления спектрограмм извлекаются в виде векторов признаков для каждого аудиофайла, после чего эти вектора признаков объединяются в финальный вектор из 4095 признаков. Затем выполняется нормализация матрицы признаков с использованием метода Normalizer из библиотеки Scikit-learn, в результате чего формируется оригинальный набор акустических признаков.

Далее применяется ансамбль из методов классификации случайного леса с различными значениями весов для классов и голосованием по большинству. В качестве весов для методов случайного леса эмпирическим путем были подобраны три весовых коэффициента. Каждый отдельный метод случайного леса обучается на всех трех наборах признаков, имеющих различные веса, т.е. внутри одного блока случайного леса на рисунке используются три метода для каждого набора признаков. Голосование по большинству происходит сначала между методами с одинаковыми весами (обученными на разных наборах признаков, имеющих одинаковые веса), а затем между методами с различными весами. Кроме того, для упрощения объединения в комплексе методов нескольких паралингвистических явлений выходные метки классов перекодируются из трех классов в два класса (отсутствие агрессии и наличие агрессии в аудиофайле). Еще одной особенностью данного ансамбля является и то, что, несмотря на ансамблирование методов случайного леса (комплексирование), он не требует большого количества вычислительных и временных ресурсов.

2.8 Методика интегрального оценивания степени выраженности деструктивных паралингвистических явлений в разговорной речи

Для интегрального оценивания наличия деструктивных паралингвистических явлений в разговорной речи предложена методика (рисунок 10), основанная на вычислении следующей формулы:

$$I_{int} = I_{agg} + I_{dec} + I_{depr} = w_{agg} \cdot \hat{y}_{agg} + w_{dec} \cdot \hat{y}_{dec} + w_{depr} \cdot \hat{y}_{depr}, \quad (2)$$

где: I_{int} – интегральная оценка $I_{int} = (I_1, \ldots, I_3)$, w_i – весовые коэффициенты (веса) значимости деструктивных явлений множества весовых коэффициентов значимости $W = (w_1, \ldots, w_3)$, $\hat{y}_i$ – результаты классификации методов определения деструктивных паралингвистических явлений множества $\hat{Y} = (\hat{y}_1, \ldots, \hat{y}_3)$.

В нашем случае имеется три частных результата классификации $\hat{y}$, на основе которых вычисляется интегральная оценка: ложность/истинность высказывания ($\hat{y}_{dec}$), наличие агрессии в высказывании ($\hat{y}_{agg}$) и наличие состояния депрессии у диктора ($\hat{y}_{depr}$).

Входными данными данной методики являются результаты классификации $\hat{Y}$, веса значимости явлений W, I_{int} – интегральная оценка степени выраженности деструктивных паралингвистических явлений в речи диктора.

Результаты классификации методов определения ложности/истинности и депрессии могут принимать бинарные значения {0, 1}, а результаты классификации метода определения агрессии могут принимать значения из множества {0, 1, 2}. Для удобства вычислений их необходимо привести к бинарному виду. Результат со значением 0 означает отсутствие агрессии в речевом высказывании, а значения 1 и 2, означающие средний и высокий уровни агрессии, соответственно, преобразуются в значение 1 – наличие агрессии в речи. Кроме того, результатами могут быть не только результаты классификации, усредненные по всему речевому сообщению, но и результаты классификации сегментов записи, а также вероятности принадлежности записи или сегментов к классам. Значения весов значимости удовлетворяют условию $w_{dec} + w_{agg} + w_{depr} = 1$.

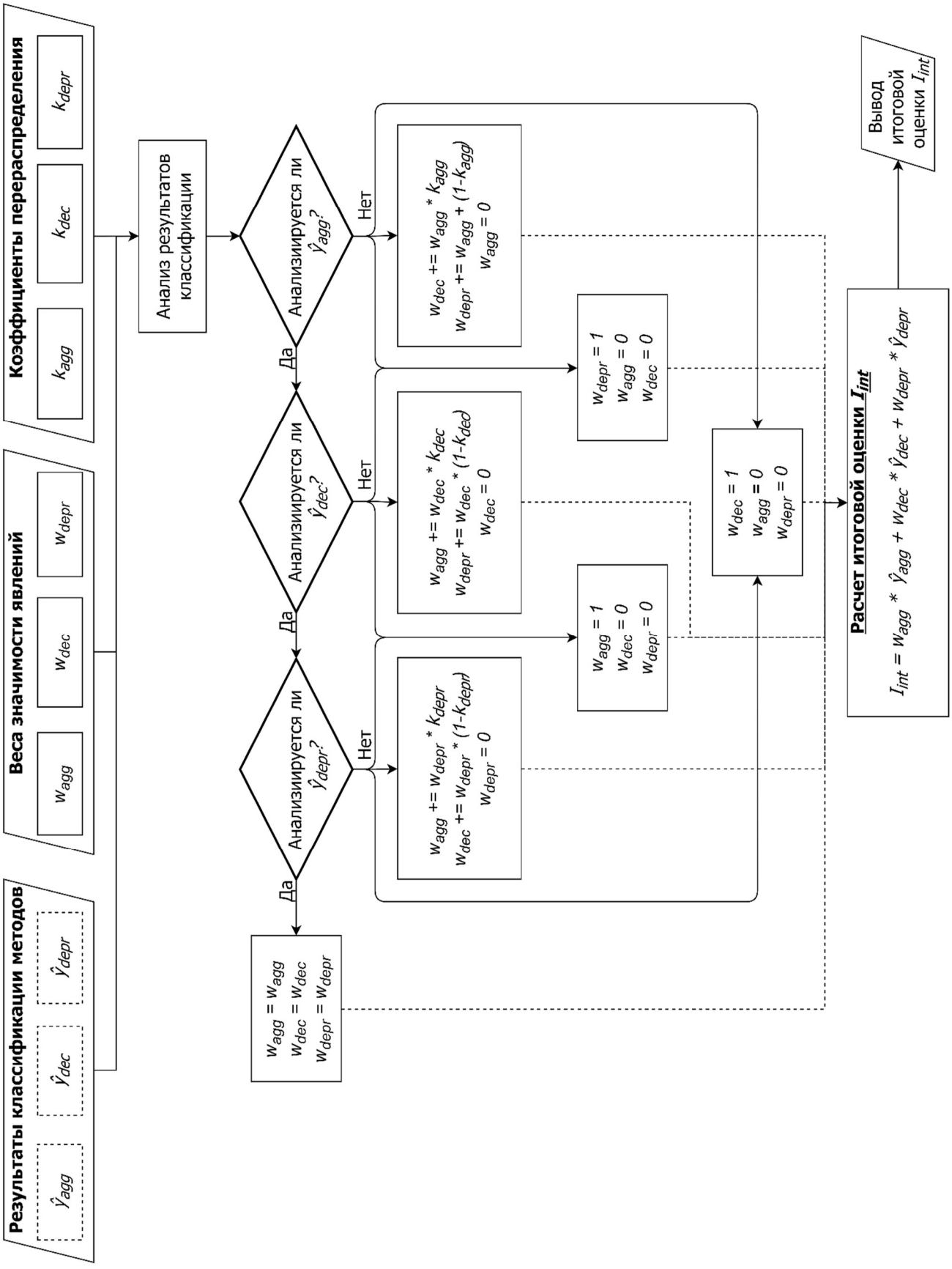

Рисунок 10 – Методика интегрального оценивания степени выраженности деструктивных паралингвистических явлений в речи диктора

Входные данные обрабатываются с использованием ряда формальных правил, которые основаны на экспертных оценках и теоретическом базисе корреляции между рассматриваемыми паралингвистическими явлениями (подробнее описаны в разделе 1.1 диссертации). Табличное представление предложенного набора формальных правил представлено в таблице 6.

Таблица 6 – Правила вычисления интегральной оценки степени выраженности деструктивных паралингвистических явлений в речи диктора при анализе различных деструктивных явлений

Результаты классификации методов			**Веса значимости результатов классификации**			**Интегральная оценка**
$\hat{y}_{dec}$	$\hat{y}_{agg}$	$\hat{y}_{depr}$	w_{dec}	w_{agg}	w_{depr}	I_{int}
Анализируются три деструктивных явления						
0	0	0	0,3	0,3	0,4	0
0	0	1	0,3	0,3	0,4	0,4
0	1	0	0,3	0,3	0,4	0,3
0	1	1	0,3	0,3	0,4	0,7
1	0	0	0,3	0,3	0,4	0,3
1	0	1	0,3	0,3	0,4	0,7
1	1	0	0,3	0,3	0,4	0,6
1	1	1	0,3	0,3	0,4	1
Анализируются два деструктивных явления						
–	0	0	–	0,3 + 0,1	0,4 + 0,2	0
–	0	1	–	0,3 + 0,1	0,4 + 0,2	0,6
–	1	0	–	0,3 + 0,1	0,4 + 0,2	0,4
–	1	1	–	0,3 + 0,1	0,4 + 0,2	1
0	–	0	0,3 + 0,1	–	0,4 + 0,2	0
0	–	1	0,3 + 0,1	–	0,4 + 0,2	0,6
1	–	0	0,3 + 0,1	–	0,4 + 0,2	0,4
1	–	1	0,3 + 0,1	–	0,4 + 0,2	1
0	0	–	0,3 + 0,2	0,3 + 0,2	–	0
0	1	–	0,3 + 0,2	0,3 + 0,2	–	0,5
1	0	–	0,3 + 0,2	0,3 + 0,2	–	0,5
1	1	–	0,3 + 0,2	0,3 + 0,2	–	1
Анализируется одно деструктивное явление						
–	–	0			0,4 + 0,6	0
–	–	1	–	–	0,4 + 0,6	1
–	0	–		0,3 + 0,7		0
–	1	–	–	0,3 + 0,7	–	1
0	–	–	0,3 + 0,7			0
1	–	–	0,3 + 0,7	–	–	1

Затем обработанные данные подаются на вход блока расчета интегральной оценки I_{int}, где происходит вычисление интегральной оценки I_{int} с использованием формулы 2.

В данной таблице используются веса значимости по умолчанию $w_{dec} = 0,3$, $w_{agg} = 0,3$, $w_{depr} = 0,4$. Вес значимости депрессии в данном случае выше, поскольку она является комплексным явлением.

Предложенная методика позволяет анализировать как все три результата классификации методов определения деструктивных паралингвистических явлений, так и отсутствие результатов классификации одного или двух методов с использованием коэффициента перераспределения весов.

В случае, когда не анализируются какие-либо результаты классификации, коэффициент перераспределения весов k равен 1/2 для депрессии и 1/3 для ложности и агрессии. Примеры представлены в таблице 6 в разделах «Анализируются два деструктивных явления» и «Анализируется одно деструктивное явление».

Выходными данными предложенной методики является интегральная оценка I_{int}, десятичные значения которой могут варьироваться в диапазоне [0, 1]. При этом значения интегральной оценки можно разделить по уровням выраженности деструктивных паралингвистических явлений: значения до 0,30 означают низкий уровень, от 0,31 до 0,6 – средний уровень, выше 0,61 – высокий уровень.

2.9 Выводы по главе 2

В данной главе приведено текстовое и графическое описание предложенных методов автоматического распознавания деструктивных паралингвистических явлений в разговорной речи, которые синтезированы на основе известных базовых методов вычисления акустических признаков и машинной классификации.

Все предложенные методы включают в себя следующие шаги обработки аудиоданных:

1. Предобработка одно- и многомодальных данных (в случае, если имеется многомодальный набор данных, эксперту необходимо извлечь из него

аудиодорожку и предобработать уже ее, то же можно сделать, например, с помощью методов обнаружения голосовой активности в аудиосигнале).

2. Вычисление акустических признаков из предобработанных аудиоданных.

3. Постобработка вычисленных векторов акустических признаков.

4. Обучение нейросетевой модели классификации с использованием вычисленных векторов признаков.

5. Машинная классификация деструктивных паралингвистических явлений.

Предложенные методы отличаются между собой как способами вычисления акустических признаков, так и подходами к обучению. Метод определения ложной/истинной информации в речевых сообщениях (HTLBbM-Deception) отличается тем, что в нем используется несколько наборов акустических признаков, а также сложная многоуровневая структура, что повышает устойчивость и точность распознавания. Метод определения депрессии отличается тем, что в нем используются уже не экспертные, а нейросетевые признаки, а также современная нейросетевая архитектура. Для метода определения агрессии были выбраны как экспертные, так и нейросетевые признаки, но уже в сочетании с ансамблевым методом и весовыми значениями для методов случайного леса.

Среди предложенных выше методов есть как методы на основе детерминированных методов машинного обучения, так и методы на основе нейронных сетей. В ходе исследований не найдено универсальных методов, которые бы хорошо подходили для решения задач всех даже в пределах одной области, что подтверждает необходимость разработки независимых методов для решения задач классификации отдельных деструктивных паралингвистических явлений.

Также приведена предложенная методика интегрального оценивания степени выраженности деструктивных паралингвистических явлений в речи диктора. На основе результатов классификации предложенных методов и ряда правил, выявленных при проведении аналитического обзора по обнаружению корреляций

между рассматриваемыми паралингвистическими явлениями, вычисляется интегральная оценка состояния диктора. Данная оценка используется далее в качестве одного из модулей архитектуры программной системы интегрального анализа деструктивных паралингвистических явлений в разговорной речи.

3 РАЗРАБОТКА И ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ ПРОГРАММНОЙ СИСТЕМЫ ИНТЕГРАЛЬНОГО АНАЛИЗА ДЕСТРУКТИВНЫХ ПАРАЛИНГВИСТИЧЕСКИХ ЯВЛЕНИЙ В РАЗГОВОРНОЙ РЕЧИ

В данной главе описывается архитектура разработанной программной системы интегрального анализа деструктивных паралингвистических явлений в разговорной речи, в том числе приводится информация об использованных открытых программных библиотеках. Приводится описание показателей оценивания качества работы предложенных программных реализаций методов и подробное описание данных, использованных в процессе обучения и оценивания методов. Приводятся результаты экспериментальных исследований, проведенных с программными реализациями методов, предложенных для автоматического определения деструктивных паралингвистических явлений в разговорной речи, а также сравнение результатов предложенных методов с аналогами, известными в литературе.

3.1 Архитектура программной системы интегрального анализа деструктивных паралингвистических явлений в разговорной речи

До этапа разработки программной системы в разделе 1.1 проведен анализ литературы в области психологии на предмет того, существует ли корреляция между рассматриваемыми деструктивными явлениями. Данный анализ показал, что существует связь между агрессией и депрессией: склонность к размышлениям в значительной степени связана как с гневом, так и с депрессией, а связи между гневом и депрессией образуют сложную связь. Также гнев, тревога, депрессия и негативные эмоции имеют корреляцию друг с другом. Выявлено, что симптомы депрессии частично объясняют связь между ПТСР, вербальной и физической агрессией по отношению к другим объектам и самонаправленной физической агрессией, а сам гнев и хроническая склонность к возбуждению гнева частично объясняют связь между посттравматическим стрессовым расстройством, вербальной и физической агрессией по отношению к объектам и другим людям.

Более того, злой темперамент значительно предсказывает суицидальные мысли независимо от симптомов депрессии. Существует также связь между агрессией и ложью: гнев приводит к более явному проявлению имплицитных установок в отличие от нейтрального или грустного настроения, имеет сходство со счастьем, которое вызывает аналогичный эффект, поскольку обе эти эмоции повышают уверенность (а также имеют высокий уровень активации). Таким образом, люди, уверенные в себе и своих эмоциональных состояниях, вероятно выскажут свое истинное мнение, свои внутренние чувства, в отличие от тех, кто менее уверен в себе. При этом, люди, менее уверенные в себе и своих эмоциональных состояниях и установках (в том числе, в моменты грусти), с меньшей вероятностью открыто проявят свои подлинные установки (что может свидетельствовать о положительной корреляции между депрессией и ложью).

В ходе диссертационного исследования разработана программная система интегрального анализа деструктивных паралингвистических явлений в разговорной речи, получившая название DesBDet (Destructive Behaviour Detection). Она построена по модульному принципу, ее архитектура включает в себя несколько независимых программных модулей (блоков): 1) предобработка исходных данных; 2) вычисление наборов акустических признаков из аудиоданных; 3) обработка полученного вектора акустических признаков с использованием нормализации (и аугментации) данных, а также уменьшения размерности признакового пространства; 4) получение итогового результата классификации от каждого модуля; 5) вычисление интегральной оценки анализа деструктивных явлений в речи диктора. Стоит отметить, что обучение происходит в иерархическом порядке: сначала параллельно работают модуль определения ложной/истинной информации и модуль определения агрессии, а затем их результаты классификации в бинарном виде {0, 1} добавляются в качестве дополнительных признаков в признаковое пространство, которое подается на вход модуля определения депрессии. При этом гипотезы о ложности и агрессии приобретают более высокий вес по сравнению с остальными признаками. После чего все три результата работы модулей выступают в качестве входных данных в

методике интегральной оценки степени выраженности деструктивных паралингвистических явлений в речи диктора. Предложенная архитектура программной системы представлена на рисунке 12.

С использованием предложенной программной системы можно получить результат классификации деструктивных паралингвистических явлений, как с применением отдельных методов, так и результат определения депрессии в речи с учетом корреляции акустических признаков в речевом сегменте с другими деструктивными явлениями (лжи/истинности и агрессии).

При разработке программной системы DesBDet использован объектно-ориентированный язык программирования высокого уровня Python версии 3.8 [128]. Данный язык был выбран для разработки, так как он отличается такими достоинствами, как расширяемость, кроссплатформенность и большое количество открытых программных библиотек, которые находят применение в решении конкретных задач в различных направлениях.

Python относится к высокоуровневым языкам программирования общего назначения, он отличается динамической типизацией и автоматическим управлением памятью. Данные его отличия имеют свои преимущества и недостатки, но во многих случаях являются гарантом повышения производительности разработчика, а также читаемости и качества кода.

Кроме того, Python является интерпретируемым и объектно-ориентированным языком программирования. В данных аспектах также присутствуют как преимущества, так и недостатки языка: иногда программы, написанные с использованием этого языка, могут потреблять значительное количество памяти и иметь низкую скорость работы, если сравнивать данные показатели с компилируемыми языками программирования [132].

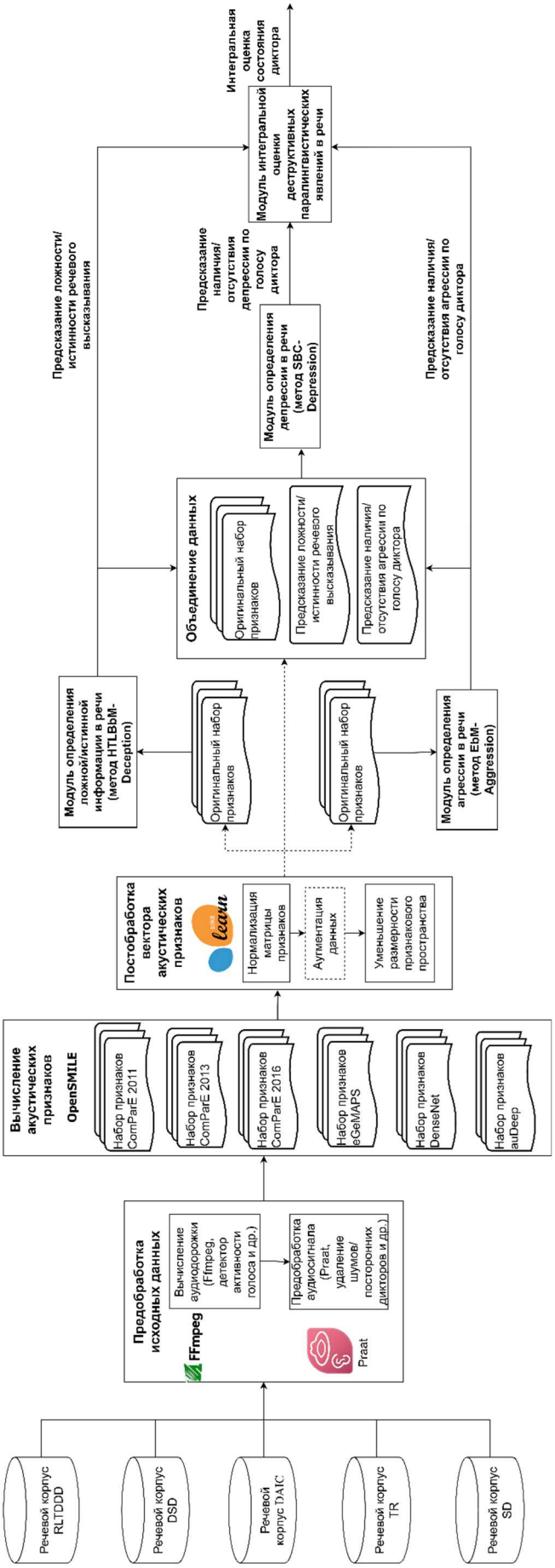

Рисунок 12 – Архитектура программной системы DesBDet

Для борьбы с подобными недостатками имеется возможность применения программных библиотек, написанных на компилируемых языках программирования, C и C++. Ввиду особенностей данных языков программирования появляется возможность реализации низкоуровневой вычислительной логики и работы с аппаратной платформой, что сказывается непосредственно на скорости и потреблении вычислительных ресурсов во многих задачах. Такие программные библиотеки существуют, в том числе и для области машинного обучения, которое необходимо при разработке программной системы, описанной в данной работе.

Для того, чтобы реализовать алгоритмы машинного обучения, а в частности, ИНС, были выбраны и использованы такие программные библиотеки как TensorFlow [133] и Keras [134]. Для обучения классических методов машинного обучения использованы реализации алгоритмов машинного обучения в программных библиотеках Scikit-learn [135], Catboost, XGBoost, LightGBM.

TensorFlow относится к открытыми программным библиотекам для машинного обучения. Эта библиотека разработана компанией Google и применяется для исследований и разработки как собственных продуктов компании, так и продуктов многих других известных компаний. Основным API при работе с библиотекой является реализация для Python. Основными особенностями библиотеки TensorFlow являются [133]:

– большое количество алгоритмов машинного обучения помимо ИНС;

– широкий инструментарий для гибкой настройки ИНС и автоматизация подбора гиперпараметров моделей;

– собственная система работы с матричными данными, которая упрощает работу с данными;

– интеграция с различными программно-аппаратными платформами (например, CUDA) позволяет повысить эффективность и скорость обучения;

– в библиотеке реализован базовый набор архитектур ИНС, который регулярно обновляется, при этом существует возможность разработки пользовательских архитектур при помощи программного интерфейса.

Keras является открытой библиотекой, которая написана на языке Python. Она является надстройкой над библиотеками Deeplearning4j, Theano, TensorFlow и обеспечивает взаимодействие с ИНС. Основными преимуществами данной библиотеки являются расширяемость, модульность и компактность. Keras является высокоуровневым инструментом для работы с ИНС и представляет более интуитивный набор абстракций, что делает работу с ИНС более удобной и простой вне зависимости от вычислительного ядра библиотеки. Такие преимущества этой библиотеки имеются в том числе и потому, что библиотека разработана в первую очередь для использования людьми и их удобства в работе с ней.

Scikit-learn является одним из самых используемых пакетов Python как для работы с данными, так и для машинного обучения. Этот пакет написан на языках Python, C, C++, Cython. С использованием Scikit-learn появляется возможность предобработки данных, уменьшения размерности признакового пространства, обнаружение аномалий в данных, использования популярных наборов данных, выбора моделей для машинного обучения, регрессионного анализа, классификации и кластеризации. Данный пакет не позволяет тонкую работу с ИНС, но в нем имеется множество реализаций классических алгоритмов машинного обучения, работа с которыми является простой и приятной за счет понятного синтаксиса, многофункциональности и кроссплатформенности. Большая часть методов, реализованных в данной работе, задействовала те или иные функции Scikit-learn, а именно: предобработка данных, уменьшение признакового пространства, некоторые модели машинного обучения [135].

Catboost, XGBoost и LightGBM являются различными реализациями градиентного бустинга, разработанными компаниями Яндекс, Microsoft and LightGBM Contributors, и The XGBoost Contributors соответственно.

Существенным отличием является то, что Catboost строит симметричные деревья решений, в отличие от двух других реализаций. Данная особенность

реализации помогает эффективнее использовать его CPU версию (предусмотрена также возможность вычислений на GPU), уменьшить время, необходимое для предсказания, и контролировать возможность переобучения. Еще одним отличием данной реализации является то, что она использует концепцию последовательного бустинга, подходе к обучению модели, основанного на перестановках, что позволяет избежать переобучения и утечек истинных значений для объектов. Кроме того, Catboost позволяет использовать все виды признаков: численные, категориальные, текстовые, тем самым сохраняя время и силы на этапе предобработки данных. При работе с текстом существует возможность использовать режим ранжирования, как и в двух других реализациях градиентного бустинга, однако Catboost предоставляет намного более мощные варианты. Благодаря поддержке работы с мультисерверными распределенными графическими процессорами (что позволяет использовать несколько хостов для ускоренного обучения), а также более ранними версиями GPU, обеспечивается масштабируемость. Наконец, Catboost предоставляет встроенные инструменты для анализа моделей, которые помогают понять, обнаружить ошибки и исправить модели машинного обучения с помощью статистики и визуализации (например, выбор наиболее релевантных признаков, графики анализа признаков, параметры модели, и пр.). Помимо описанных особенностей Catboost также имеет и другие полезные особенности: детектор переобучения, поддержка пропущенных значений в данных, инструменты для просмотра графиков и реализация перекрестной валидации [122].

XGBoost – алгоритм машинного обучения, основными преимуществами которого являются вычислительная скорость и производительность моделей. Модель поддерживает такие виды бустинга как:

– градиентный бустинг, контролируемый скоростью обучения;

– стохастический градиентный бустинг, который использует подвыборку в строке, столбце или столбце на каждом уровне разделения;

– регуляризированный градиентный бустинг использует L1 (Lasso) и L2 (Ridge) регуляризации.

Кроме того, у данной реализации также есть особенности, которые повышают производительность:

– использование кластера машин для обучения модели с использованием распределенных вычислений;

– использование всех доступных ядер CPU при построении дерева для обеспечения распараллеливания;

– выполнение вычислений во внешней памяти при работе с наборами данных, которые не помещаются в основную память;

– максимальное использование оборудования с оптимизацией кэша.

Также XGBoost может принимать в качестве входных данных различные типы данных, хорошо обрабатывать разреженные входные данные для деревьев и поддерживает использование пользовательских целевых функций и функций оценки качества.

LightGBM является распределенной программной библиотекой с высокой производительностью, которая использует деревья решений для задач ранжирования, классификации и регрессии. Она обладает следующими достоинствами:

– высокая скорость обучения моделей и точность благодаря тому, что LightGBM представляет собой алгоритм на основе гистограмм, который выполняет группировку значений (и требует меньше памяти для вычислений;

– поддерживает работу с большими и сложными наборами данных, но требует намного меньше времени для обучения;

– поддерживает как параллельное обучение, так и обучение с использованием GPU.

В отличие от горизонтального построения деревьев как в XGBoost, LightGBM использует вертикальное построение деревьев, что в результате дает уменьшение функции потерь и позволяет достичь более высокие показатели точности при

ускоренном обучении. Однако, это отличие может привести к переобучению на обучающих данных, что можно исправить использованием параметра максимальной глубины, который определяет, где будет проходить разделение.

Описанный стек технологий выбран, в первую очередь с учетом того, что для решения поставленной цели требовалась не разработка нестандартных решений при обучении, а максимальное использование уже проверенных методов и оптимизированных открытых библиотек. Данное требование является важным при решении поставленной задачи ввиду того, что конечную систему планируется использовать в режиме реального времени, а такие системы требуют применения оптимальных как по времени, так и по вычислительным ресурсам решений.

3.2 Графический пользовательский интерфейс программной системы DesBDet

Графический интерфейс разработанной программной системы является интуитивно понятным и разработан с использованием PyQt6, набора расширений графической библиотеки Qt для языка программирования Python. При запуске приложения предлагается записать аудиофайл, загрузить аудиофайл или загрузить базу данных для дальнейшей работы (рисунок 13).

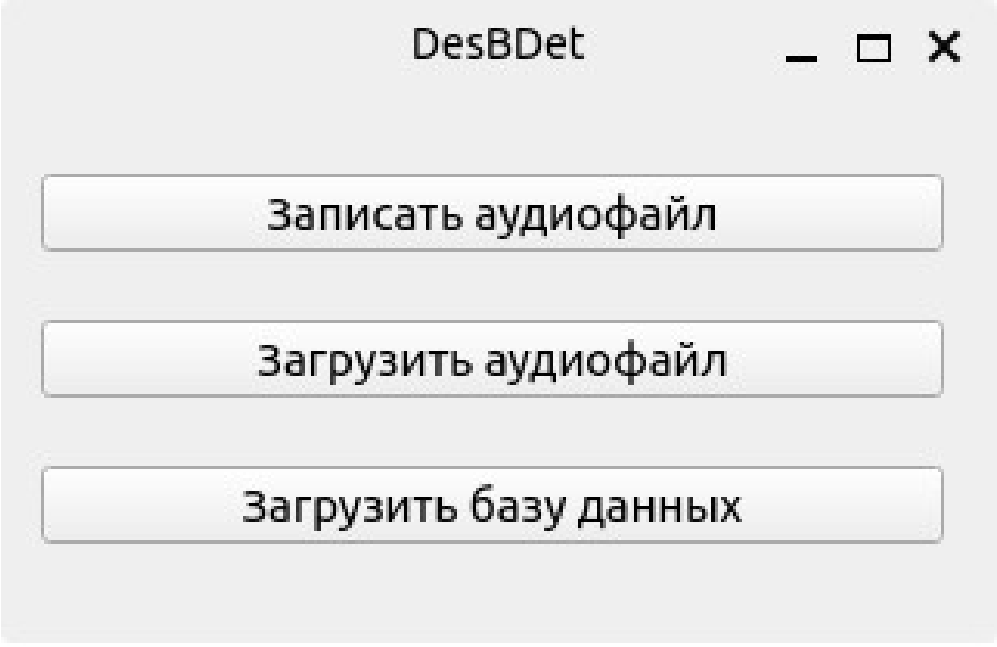

Рисунок 13 – Скриншот начального диалогового окна DesBDet

В случае, если на главном окне приложения пользователь выбрал «Записать аудиофайл», откроется окно для записи аудиофайла (рисунок 14). Данное окно имеет следующие графические элементы:

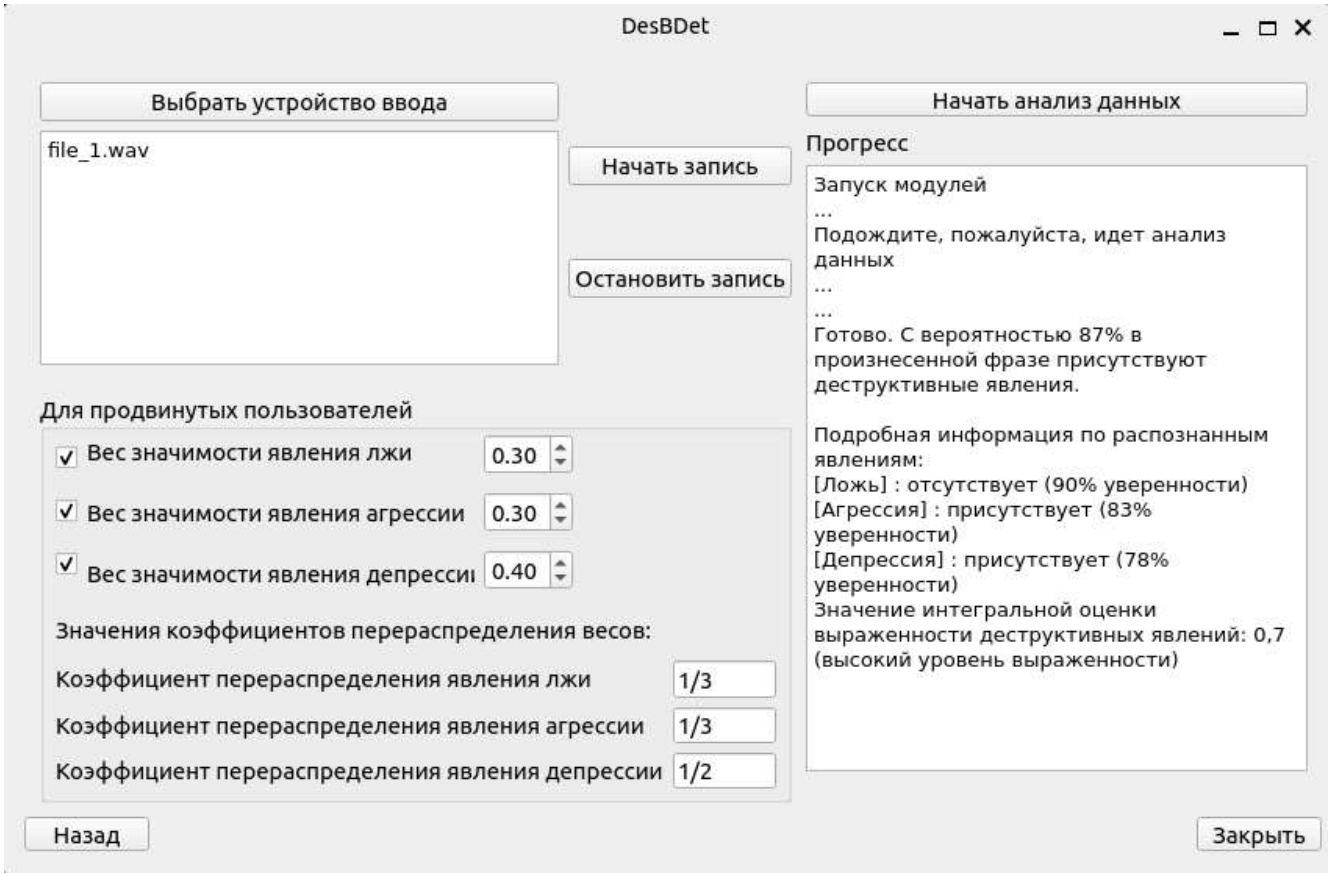

Рисунок 14 – Скриншот диалогового окна работы с модулем записи файлов

– Слева сверху расположены три кнопки: «Выберите устройство ввода», «Начать запись» и «Остановить запись». При нажатии первой пользователю предлагается из списка выбрать устройство, с помощью которого он хочет сделать запись, либо используется стандартное устройство записи, а в случае, если такового нет, пользователь увидит уведомление о том, что приложение не обнаружило подходящих устройств для записи. Когда устройство ввода выбрано, можно начать запись, нажав на кнопку с соответствующей надписью, кнопка «Остановить запись» останавливает запись и сохраняет ее в формате .wav в отдельную папку текущей сессии по месту расположения программной системы.

– Под списком файлов располагаются раздел для продвинутых пользователей, в котором можно задать веса значимости для отдельных явлений и подключить/отключить отдельные модули системы.

– Справа располагается окно прогресса работы системы, в котором система сообщает пользователю свое состояние.

– Над окном прогресса работы системы располагается кнопка, при нажатии которой запускается обработка записанного файла и работа модулей определения явлений.

– В нижней части окна находятся две кнопки: слева располагается кнопка «Назад», справа - «Закрыть». Первая предназначена для возвращения к предыдущему окну, вторая – для закрытия программы.

В случае, если пользователь выбирает «Загрузить аудиофайл», открывается окно работы с возможностью загрузки аудиофайлов (рисунок 15). Система может принимать на вход аудиофайлы в формате .wav. Диалоговое окно приложения имеет следующие графические элементы:

– Слева сверху располагается история выбранных аудиофайлов, которые пользователь может удалить в случае необходимости, кликнув по имени файла правой кнопкой мышки.

– Справа от списка файлов располагается кнопка, при нажатии которой запускается проводник, и пользователь может выбрать аудио- или видеофайлы (в процессе предобработки из видеофайлов берется только аудиосигнал).

– Под списком файлов располагаются раздел для продвинутых пользователей, в котором можно задать веса значимости для отдельных явлений и подключить/отключить отдельные модули системы.

– Справа располагается окно прогресса работы системы, в котором система сообщает пользователю свое состояние.

– Над окном прогресса работы системы располагается кнопка, при нажатии которой запускается обработка записанного файла и работа модулей определения явлений.

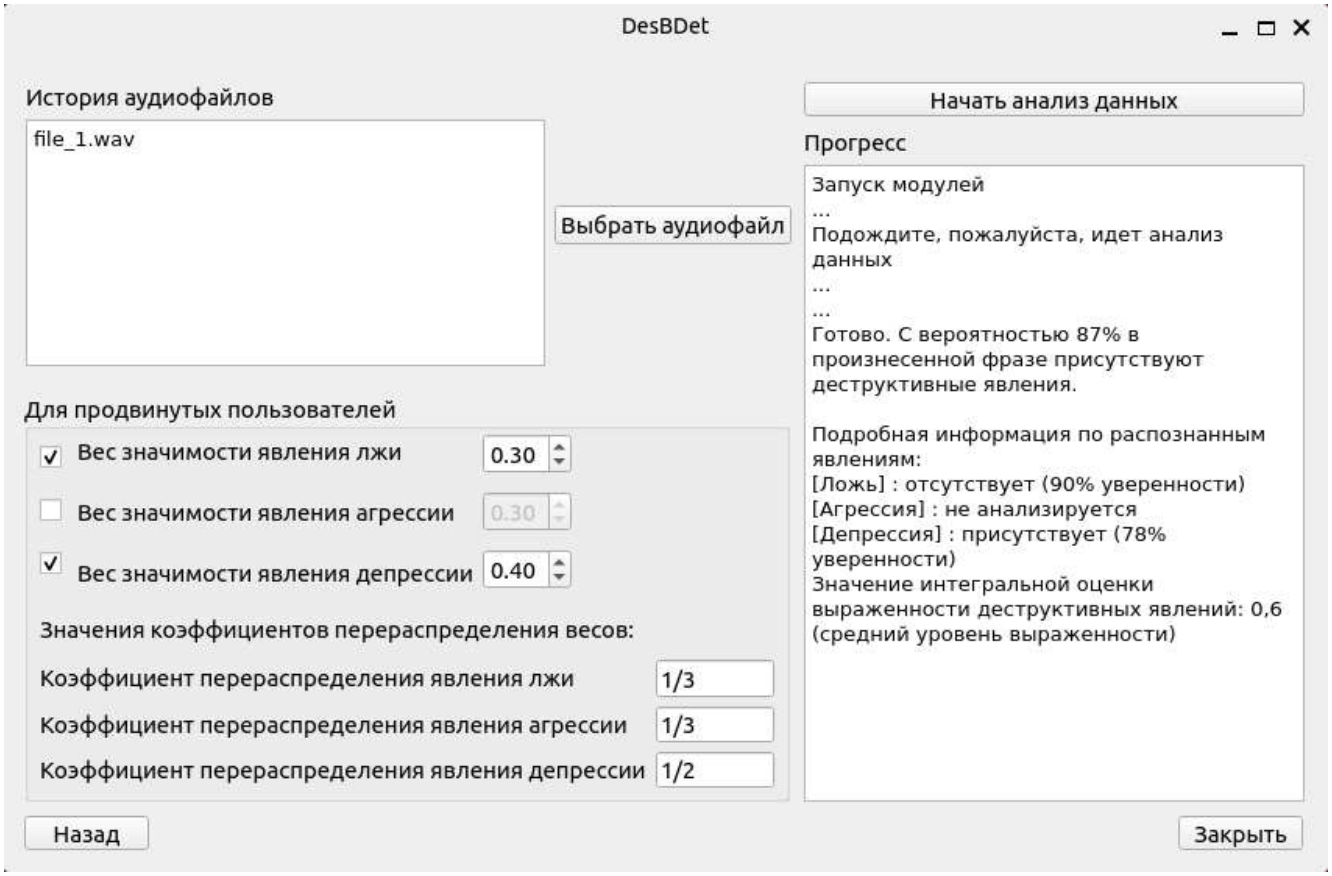

Рисунок 15 – Скриншот диалогового окна работы с модулем загрузки файлов

– В нижней части окна находятся две кнопки: слева располагается кнопка «Назад», справа - «Закрыть». Первая предназначена для возвращения к предыдущему окну, вторая – для закрытия программы.

В случае, если на главном окне приложения пользователь выбрал «Загрузить базу данных», откроется окно с возможностью загрузки аудиофайлов базы данных и файла разметки, содержащего метки классов загруженных файлов (рисунок 16). Это окно приложения имеет следующие графические элементы:

– Слева сверху располагается поле аудиофайлов базы данных, которые пользователь может удалить в случае необходимости, кликнув по имени файла правой кнопкой мышки.

– Под списком файлов располагаются раздел для продвинутых пользователей, в котором можно задать веса значимости для отдельных явлений и подключить/отключить отдельные модули системы.

– Справа от списка файлов располагается кнопка, при нажатии которой запускается проводник, и пользователь может выбрать папку с аудиофайлами.

– Под кнопкой загрузки аудиофайлов базы данных расположена кнопка для загрузки файла с разметкой базы данных.

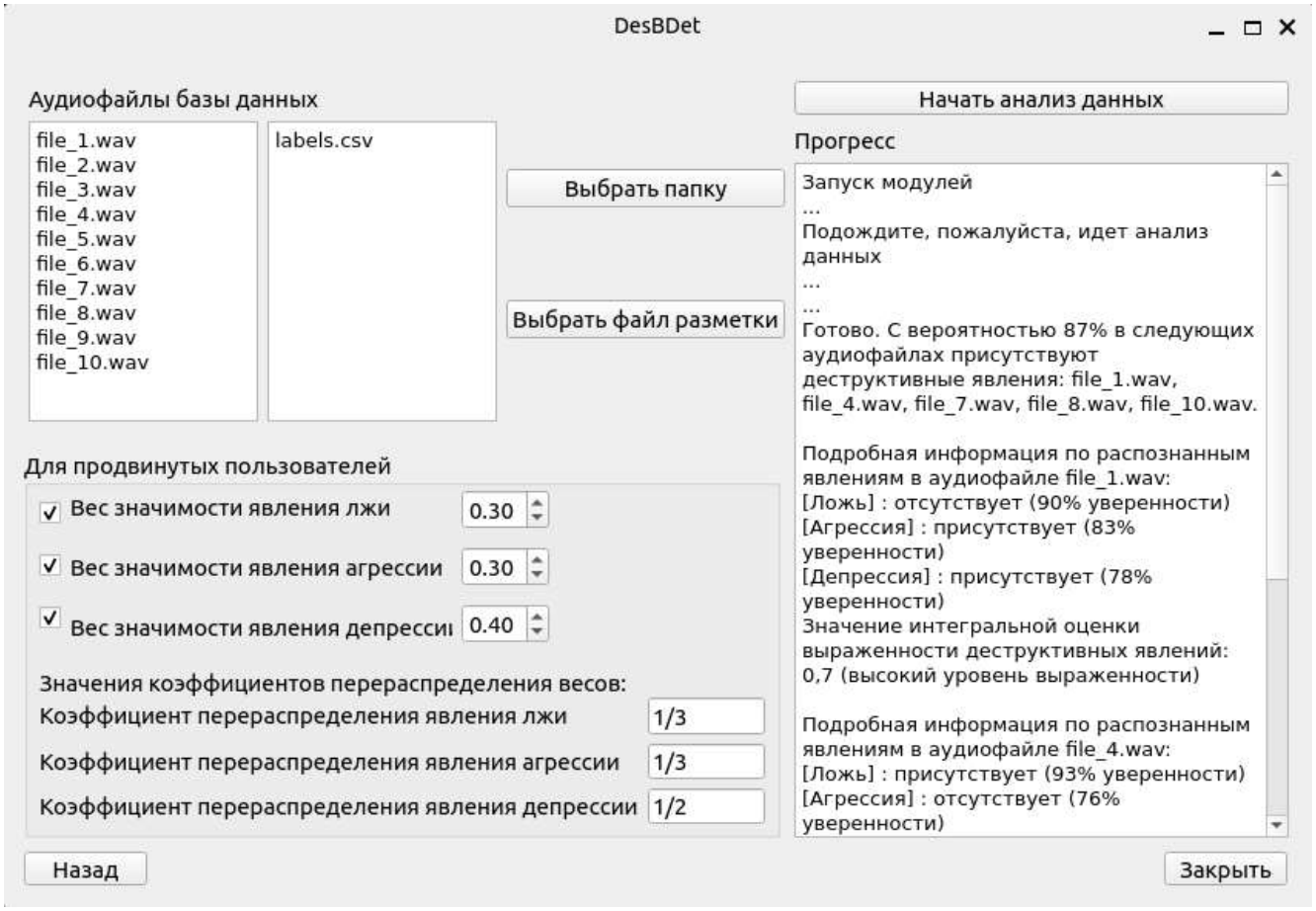

Рисунок 16 – Скриншот диалогового окна работы с модулем загрузки файлов базы данных

– Справа располагается окно прогресса работы системы, в котором система сообщает пользователю свое состояние.

– Над окном прогресса работы системы располагается кнопка, при нажатии которой запускается обработка базы данных и работа модулей.

– В нижней части окна находятся две кнопки: слева располагается кнопка «Назад», справа - «Закрыть». Первая предназначена для возвращения к предыдущему окну, вторая – для закрытия программы.

Можно отметить, что в окне прогресса работы системы находятся пояснения полученных результатов, а именно: в каких аудиофайлах обнаружены деструктивные паралингвистические явления и какие, степень уверенности системы в полученных результатах, а также приводится значение интегральной оценки и к какому уровню выраженности она относится.

3.3 Описание исследовательских речевых и многомодальных данных

В качестве экспериментальных данных для оценивания предложенных методов использованы несколько речевых и многомодальных корпусов, содержащих рассматриваемые деструктивные паралингвистические явления.

Корпус DSD содержит 1059 аудиозаписей в обучающей и отладочной частях, из которых 312 содержат ложную информацию, а 747 – истинную. Средняя продолжительность речевых сообщений 4-5 секунд. В данных содержатся аудиозаписи 22 женщин и 27 мужчин. Основные параметры речевого корпуса DSD представлены в таблице 7.

Таблица 7 – Параметры речевого корпуса DSD

Параметр	**Обучающая часть**	**Отладочная часть**	**Тестовая часть**
Кол-во дикторов (женщины, мужчины) (виновен, не виновен)	26 (11, 15) (14, 12)	23 (11, 12) (10, 13)	– – –
Кол-во аудиофайлов (виновен, не виновен) (ложные, истинные)	572 (308, 264) (182, 390)	487 (220, 267) (130, 357)	497 – –
Длительность диалога (сек) Среднее: мин.-макс.	7,6: 0,4-220	6,3: 1,1-236	6,5: 0,1-220
Длительность речевых сообщений (сек) Среднее: мин.-макс.	5,1: 0,2-214	4,0: 0,1-227	3,8: 0,1-211
Кол-во фонем Среднее: мин.-макс.	32,2: 1-1480	24,5: 1-1330	23,9: 1-1268

Корпус RLTDDD содержит 121 запись судебных заседаний, включая 61 - с ложными данными и 60 - с истинными. Средняя продолжительность видеозаписей по всей базе данных составляет 28 сек., в том числе 27,7 сек. и 28,3 сек. для ложных

и правдивых высказываний, соответственно. Данные содержат аудио- и видеозаписи 21 женщины и 25 мужчин, возраст информантов находится между 16 и 60 годами. Основные параметры корпуса RLTDDD представлены в таблице 8.

Таблица 8 – Параметры корпуса RLTDDD

Параметр	**Значение**
Общее количество видеозаписей (ложные, истинные)	121 (61, 60)
Количество участников (женщины, мужчины)	46 (21, 25)
Возраст участников (лет)	16-60
Средняя продолжительность видеозаписей (сек.) (ложные, истинные)	28 (27,7, 28,8)

Корпус DAIC включает в себя аудио-, видеоданные и записи сенсора глубины всех взаимодействий. В записи участвовали 219 человек (92 женщины и 127 мужчин). В коллекции многомодальных данных также присутствуют физиологические данные (гальваническая проводимость кожи, ЭКГ, дыхание). Использовались следующие опросники: The Positive and Negative Affect Scale (PANAS) для оценки настроения, PTSD Checklist – Civilian Version для оценки ПТСР, Patient Health Questionnaire для оценки психического здоровья, Depression module для оценки наличия и уровня депрессии, State-Trait Anxiety Inventory для оценки тревожности. Основные параметры корпуса представлены в таблице 9.

Таблица 9 – Параметры корпуса DAIC

Параметр	**Обучающая часть**	**Отладочная часть**	**Тестовая часть**
Количество информантов (женщины, мужчины)	163 (70, 93)	56 (22, 34)	56 (нет данных)
Длительность записей (ч:мин:сек)	43:40:20	14:47:31	14:52:42
Количество записей людей с/без депрессии	37 / 126	44 / 12	– / –

Корпус aGender [136] состоит из 49 часов записей телефонных разговоров на немецком языке, содержит 6 сессий. Общее количество речевых выражений –

65364. Длина записей варьируется от 1 до 6 секунд: слова-команды, названия месяцев, даты, время, телефонные номера, имена и фамилии. Данные разбиты на 7 классов: дети, молодые люди (женщины и мужчины), взрослые (женщины и мужчины) и пожилые (женщины и мужчины).

EmoDB [137] – корпус немецкой эмоциональной речи, записи которого содержат речь 10 актеров (5 женщин и 5 мужчин) в возрасте от 21 до 35 лет. Корпус состоит из 535 речевых выражений, длительность которых варьируется от 1,2 до 9 секунд с медианной длительностью 2,6 секунд. Данные размечены на 7 эмоциональных категорий: нейтральное состояние, гнев, страх, радость, грусть, отвращение и скука. Записи были сделаны в звукоизолированной комнате с использованием высококачественного оборудования. Были выбраны 5 коротких и 5 длинных фраз с нейтральным семантическим содержанием, которые могут быть использованы в ежедневных разговорах.

Таблица 10 – Параметры речевых корпусов SD и TR

Параметр	**Обучающая часть**	**Отладочная часть**	**Тестовая часть**
Количество записей по уровням агрессии (низкий, средний, высокий)	293 (156, 74, 63)	117 (69, 33, 15)	501 (–, –, –)
Средняя продолжительность записи (сек)	5,0		
Общая длительность записей (мин:сек)	75:23		

Корпус Ruslana [138] состоит из 3661 фонетически репрезентативных речевых выражения на русском языке, произнесенных 61 информантами (среди которых 49 женщин и 12 мужчин) в возрасте от 16 до 28 лет. По структуре корпус похож на корпус EmoDB, но в корпусе Ruslana большее число информантов. Записи были сделаны в звукоизолированной студии и размечены по следующим эмоциональным категориям: нейтральное состояние, удивление, счастье, гнев, грусть и страх. Длительность записей варьируется от 1,2 до 7,8 секунд с медианной длительностью 2,3 секунды.

Корпуса SD и TR включают в себя 893 записи речи актеров, которые представлены тремя классами: высокий уровень агрессии, средний уровень агрессии и низкий уровень агрессии. Основные параметры корпусов SD и TR представлены в таблице 10.

3.4 Показатели оценивания качества работы программных реализаций методов распознавания деструктивных паралингвистических явлений

При оценивании эффективности методов в данной работе используются различные показатели, большинство из которых основаны на матрице спутывания (Confusion Matrix, см. таблицу 11) [139], где:

TP – верно распознанные объекты класса 1,

FP – объекты класса 2, неверно распознанные как объекты класса 1,

FN – объекты класса 1, неверно распознанные как объекты класса 2,

TN – верно распознанные объекты класса 2.

Таблица 11 – Матрица спутывания для задачи бинарной классификации (2 класса)

	Предсказание для класса 1	**Предсказание для класса 2**
Истинное значение класса 1	Истинно-положительный (True Positive, TP)	Ложно-положительный (False Positive, FP)
Истинное значение класса 2	Ложно-отрицательный (False Negative, FN)	Истинно-отрицательный (True Negative, TN)

Полнота (Recall) распознавания класса, вычисляется как отношение верно классифицированных объектов к соответствующему количеству объектов в классе:

$$\text{Полнота}(Recall) = \frac{TP}{TP + TN} \quad (3)$$

Точность (Precision) – отношение верно классифицированных объектов класса 1 к количеству объектов, распознанных как класс 1:

$$\text{Точность}(Precision) = \frac{TP}{TP + FN} \quad (4)$$

F1-мера (F1) – гармоническое среднее между полнотой (Recall) и точностью (Precision).

$$F1 = 2\frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

Точность (Accuracy) вычисляется как отношение верно классифицированных объектов к общему количеству объектов:

$$\text{Точность}(Accuracy) = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

Невзвешенная средняя полнота (Unweighted Average Recall, UAR) – показатель на основе средней чувствительности и специфичности (mean of sensitivity and specificity), где $N_c^{(i)}$ описывает количество верно распознанных элементов *i*-го класса, $N_0^{(i)}$ описывает общее количество объектов в *i*-м классе, *N* описывает общее количество объектов, а *k* – количество классов:

$$UAR = \frac{1}{k}\sum_{i=1}^{k}\frac{N_c^{(i)}}{N_0^{(i)}} \quad (7)$$

Среднеквадратическое отклонение (Mean Squared Error, MSE) – среднее значение квадрата разницы между предсказанным значением и истинным значением.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2, \quad (8)$$

где *n* – количество объектов, Y_i – истинные значения класса, $\hat{Y}_i$ – предсказания для класса.

3.5 Экспериментальные исследования предложенного метода автоматического определения ложности/истинности в разговорной речи

В данном разделе представлены экспериментальные исследования предложенного метода определения ложной и истинной информации, описанного в разделе 2.5. В ходе экспериментальных исследований использованы два корпуса, содержащих ложные и истинные речевые высказывания: DSD и RLTDDD. В качестве акустических признаков для данных корпусов выбраны экспертные

признаки openSMILE, а для классификации – многоуровневый метод объединения классификаторов машинного обучения.

В качестве данных для исследования метода TLBbM-Deception (вне иерархического метода), описанного в разделе 2.5, использованы два корпуса: RLTDDD и DSD. Экспериментальные исследования проведены как с использованием одного корпуса (DSD) при обучении в иерархическом подходе, так и с использованием обоих корпусов при обучении в индивидуальном исследовании метода.

Из аудиоданных вычислены несколько наборов акустических признаков: INTERSPEECH ComParE 2013, ComParE 2016 (улучшенная версия набора 2013 года) [86] и ComParE 2011 (включает в себя акустические признаки, которые были использованы в соревнованиях по компьютерной паралингвистике в задаче автоматического определения состояния диктора) [140]. Общая размерность набора данных для обучения составила более 12000 признаков. Для уменьшения признакового пространства применен метод анализа главных компонент (PCA) в программной библиотеке Scikit-learn языка Python. Итоговый набор данных составил 1680 объектов для обучения и тестирования и 986 информативных признака для каждого объекта. Обучение проводилось с использованием 10-кратной перекрестной валидации с сохранением распределения между классами в каждом разбиении.

Для экспериментальных исследований метода выбраны следующие классификаторы: Catboost, XGBoost, LightGBM и эти же три классификатора, объединенные в двухуровневом методе стэкинга. Параметры всех методов обучения подобраны с использованием метода поиска по сетке.

Вне иерархического подхода лучшим результатом, полученным с использованием описанного подхода, является результат 85,6% по показателю F1-меры. Для сравнения, отдельно обученные реализации градиентного бустинга Cabtoost, XGBoost и LightGBM смогли достигнуть результатов по показателю F1-меры 84,1%, 84,6%, и 85,0% соответственно. Сравнение полученных результатов представлено в таблице 12.

Таблица 12 – Результаты сравнения метода TLBbM-Deception с аналогами

Метод	**Результаты классификации**
Метод на основе просодических признаков, типов ответа и метода опорных векторов [19]	UAR = 74,9%
Базовый метод на соревнованиях ComParE-2016 [86]	UAR = 68,3%
Метод на основе информации из четырех модальностей и объединения методов деревьев решений и случайного леса [88]	Точность (Accuracy, max) = 75,0%
Метод на основе акустических, лексических признаков и гибридной нейронной сети (LSTM + DNN) для показателя Precision и метода случайного леса для показателя F1-меры [141]	F1 = 63,9%, Precision = 76,1%
Catboost	F1 = 84,1%, UAR = 84,0%
XGBoost	F1 = 84,6%, UAR = 84,4%
LightGBM	F1 = 85,0%, UAR = 84,9%
Stacking (Catboost, XGBoost, LightGBM)	**F1 = 85,6%, UAR = 85,5%**

В иерархическом подходе HTLBbM-Deception (см. рисунок 7) при использовании базового метода, в которой не использовалась информация о половой принадлежности и эмоциональном состоянии диктора, удалось добиться результата 83,7±0,2% по показателю F1-меры. Результаты экспериментальных исследований базового метода, метода, использующего только эмоциональное состояние и комплексного метода представлены в таблице 13.

В этой таблице также сравниваются результаты классификации на основе истинных и предсказанных значений о поле диктора. Прочерками указывается, что информация о поле или эмоциональном состоянии не использовалась в течение процесса обучения.

Лучший результат получен с использованием эмоциональных признаков корпуса Ruslana и предсказанными метками пола как для корпуса Ruslana, так для корпуса DSD: 88,4%±1,5 по показателю F1-меры. Различие в результатах экспериментальных исследований между корпусами Ruslana и EmoDB может быть связано с тем, что в корпусе Ruslana содержится больше записей различных дикторов, а также в том, что условия записи близки к таковым в корпусе DSD.

Таблица 13 – Результаты экспериментальных исследований иерархического метода HTLBbM-Deception

Эмоциональный корпус	**Информация о половой принадлежности диктора**		**Результаты классификации, F1-мера (%)**
	В эмоциональном корпусе	**В корпусе DSD**	
Ruslana	–	–	85,5±0,2
	Предсказанные значения	Предсказанные значения	**88,4±1,5**
	Предсказанные значения	Истинные значения	88,0±1,2
	Истинные значения	Предсказанные значения	88,2±1,7
	Истинные значения	Истинные значения	85,4±0,3
EmoDB	–	–	85,4±0,6
	Предсказанные значения	Предсказанные значения	87,5±1,0
	Предсказанные значения	Истинные значения	87,7±1,3
	Истинные значения	Предсказанные значения	87,9±1,6
	Истинные значения	Истинные значения	85,4±0,1
Не используется	–	–	83,7±0,2

Проведенные экспериментальные исследования показывают эффективность использования нескольких наборов акустических признаков для задачи определения ложной и истинной информации в речевых сообщениях. Кроме того, объединение нескольких методов градиентного бустинга и использование информации о половой принадлежности диктора и его эмоциональном состоянии также позволило улучшить показатели определения рассматриваемого явления.

3.6 Экспериментальные исследования метода для автоматического определения депрессии в разговорной речи

При исследованиях метода определения депрессии, описанного в разделе 2.6, в качестве данных для обучения и тестирования использовались аудиозаписи корпуса DAIC, эти данные также использовались на соревнованиях AVEC-2019.

Общее количество аудиозаписей в обучающем и отладочном наборах составило 219, из которых 163 аудиозаписи в обучающем наборе и 56 – в отладочном.

Из аудиозаписей с использованием инструментария openSMILE вычислен набор акустических признаков eGeMAPS [142], после чего получен набор данных размерностью 163 объекта для обучения и 56 объектов для тестирования и 88 признаков для каждого объекта. В результате добавления набора акустических признаков DenseNet размерность признакового пространства оказалась на 1920 признаков больше. За счет особенностей метода вычисления набора признаков DenseNet из 163 объектов для обучения и 56 объектов для тестирования получено 310508 объектов для обучения и 109824 объектов для тестирования в оконном представлении.

Для классификации при разработке метода выбрана нейросетевая архитектура TabNet с преимущественно параметрами по умолчанию (оптимизированы параметры batch_size = 64 и virtual_batch_size = 32). Экспериментальные исследования проведены с использованием 10-кратной перекрестной валидации по показателям UAR и F1-мера. Результат распознавания депрессии с использованием предложенного метода достиг показателей F1-меры = 64,0% и UAR = 60,0%. В таблице 14 представлены результаты сравнения предложенного метода и известных в литературе аналогов.

Таблица 14 – Результаты сравнения метода SBC-Depression с аналогами

Методы определения депрессии по речи	**Данные**	**Результаты классификации**
Метод на основе обучения нейронной сети с механизмом внимания на двух корпусах (General Psychotherapy Corpus и DAIC) [65]	DAIC-WOZ	F1 = 70,3%, UAR = 70,3%
	General Psychotherapy Corpus	F1 = 71,6%
Метод на основе обучения расширяемой сверточной нейронной сети на двух корпусах (SH2-FS и DAIC) [66]	SH2-FS	UAR = 68,0%
	DAIC-WOZ	**UAR = 88,0%**
Предложенный метод SBC-Depression	DAIC-WOZ	F1 = 64,0%, UAR = 60,0%

3.7 Экспериментальные исследования метода для автоматического определения агрессии в разговорной речи

При исследованиях метода определения агрессии, описанного в разделе 2.7, для обучения и тестирования использовались корпуса SD и TR. В результате вычисления набора акустических признаков получено 6373 признака для 118 объектов для обучения и 296 объектов для тестирования. Для классификации при разработке метода EbM-Aggression выбран метод случайного леса с параметрами (n_estimators=10000, criterion = 'entropy', остальные параметры по умолчанию), подобранными при помощи поиска по сетке (см. рисунок 9).

Экспериментальные исследования проведены с использованием 5-кратной перекрестной валидации, при чем для каждого разбиения сохранено оригинальное распределение классов. В качестве показателя качества распознавания использован UAR. Результат распознавания агрессии с использованием предложенного метода достиг значения UAR = 76,5%. Результаты сравнения с известными аналогами представлены в таблице 15.

Таблица 15 – Результаты сравнения EbM-Aggression с альтернативными методами

Методы определения агрессии по речи	**Тестовые данные**	**Результат классификации (UAR, %)**
Базовый метод на соревнованиях ComParE-2021 на основе мешка аудио слов и метода опорных векторов [143]	SD + TR	72,2%
Метод на основе X-векторов и векторов Фишера и метода опорных векторов [80]	SD + TR	**77,8%**
Предложенный метод EbM-Aggression	SD + TR	76,5%

В результате проведенных экспериментальных исследований предложенного метода определения агрессии в разговорной речи было выявлено, что использование нескольких наборов акустических признаков и ансамблирование методов случайного леса с использованием весовых коэффициентов показывает результат классификации агрессии на уровне известных аналогов, а сам метод является конкурентоспособным.

На основе полученных в экспериментальных исследованиях результатов классификации для оценки предложенной программной системы можно вычислить интегральную среднюю F1-меру ($F1_{int}$) и интегральную невзвешенную среднюю полноту (UAR_{int}):

$$F1_{int} = \frac{1}{3}\sum_{i=1}^{k3} F1_i = 76{,}8\%, \quad (9)$$

где $F1_i$ – *F1-мера* i-го метода из 3 методов,

$$UAR_{int} = \frac{1}{3}\sum_{i=1}^{3} UAR_i = 75{,}0\%, \quad (10)$$

где UAR_i – *UAR* i-го метода из 3 методов.

3.8 Внедрение результатов диссертационного исследования

В ходе работы над диссертационным исследованием следующие основные научные результаты были применены при решении задач в грантах в СПб ФИЦ РАН. Архитектура программной системы интегрального анализа деструктивных паралингвистических явлений в речи (DesBDet) и методы автоматического определения деструктивных паралингвистических явлений в разговорной речи использовались при решении задачи разработки системы определения деструктивных паралингвистических явлений в речи в проекте Российского фонда фундаментальных исследований № 20-37-90144 Аспиранты «Разработка и исследование автоматической системы для выявления деструктивных паралингвистических явлений в разговорной речи». Метод анализа речевого сигнала на основе нескольких наборов акустических признаков и комплексирования моделей градиентного бустинга для определения ложной и истинной информации в речи использовался при решении задачи разработки определения ложной и истинной информации в речи в проекте Российского научного фонда № 18-11-00145 «Разработка и исследование интеллектуальной системы для комплексного паралингвистического анализа речи».

Следующие основные научные результаты диссертационного исследования были использованы при первичной диагностике пациентов путем компьютерного

тестирования в ООО «Первый психотерапевтический»: методика интегрального оценивания степени выраженности деструктивных паралингвистических явлений в речевом сигнале диктора; программная система для определения деструктивных паралингвистических явлений в речевом сигнале.

В обоих случаях принимающие комиссии отметили практическую значимость и новизну полученных в работе результатов (см. приложение В).

3.9 Выводы по главе 3

В данной главе описана архитектура и прототип разработанной программной системы интегрального анализа деструктивных паралингвистических явлений в разговорной речи DesBDet, с использованием которой можно получить как результат классификации отдельных методов, так и результат определения депрессии в разговорной речи с учетом корреляции акустических признаков в речевом сегменте с другими деструктивными явлениями (лжи/истинности и агрессии). С использованием разработанной программной системы можно оценить состояние диктора на присутствие деструктивных явлений в его речи. Представлено описание используемых открытых программных библиотек, а также обоснование выбора данных программных библиотек для разработки программной системы по критериям эффективности, производительности и пр.

Представлен интуитивно понятный пользовательский интерфейс разработанной программной системы, разработанный с использованием библиотеки PyQt6. Данный интерфейс предоставляет пользователю возможность загрузки аудиофайлов, записи аудиофайлов, а также оценки базы данных для оценки наличия деструктивных паралингвистических явлений, а также возможность самостоятельной корректировки весов значимости и подключения/отключения каких-либо программных модулей.

На основе проведенных экспериментальных исследований можно сделать вывод, что некоторые предложенные в работе методы для автоматического определения деструктивных паралингвистических явлений в разговорной речи в двух случаях превосходят известные аналоги по точности автоматического

определения деструктивных явлений. В частности, метод определения ложной и истинной информации в разговорной речи HTLBbM, с использованием которого удалось достичь результата 88,4% по показателю F1-меры. Также метод определения агрессии EbM-Aggression с результатом 76,5% по показателю F1-меры является конкурентоспособным с другими методами.

В ходе экспериментальных исследований выявлено, что увеличение количества данных для обучения и тестирования как естественным (использование нескольких исследовательских баз данных), так и искусственным путем (аугментация данных), оказывает положительное влияние на итоговый результат распознавания. Кроме того, можно отметить, что выбор ансамблевого подхода и многоуровневых методов классификации является успешным, что вероятно, следует из повышенной устойчивости таких подходов к переобучению и их способности к более качественному выявлению зависимостей в признаковом пространстве.

При проведении экспериментальных исследований предложенной системы интегрального анализа речи выявлено, что наиболее информативными эмоциональными состояниями при определении ложной и истинной информации оказались: агрессия, грусть и нейтральное состояние. Данный результат коррелирует с известными в литературе теоретическими психологическими и экспериментальными исследованиями (например, [3, 14]). Подтверждено, что проявление у диктора эмоции «счастье» также может быть маркером истинности/ложности в речи, что может быть объяснено тем, что некоторые люди могут использовать радость для сокрытия их настоящих эмоций, или чувство подъема от адреналина, который вырабатывается из-за обмана.

Разработанная программная система DesBDet является прототипом программного продукта, не имеющим известных аналогов в мире.

ЗАКЛЮЧЕНИЕ

В диссертационной работе сформулирована и решена новая научно-техническая задача повышения эффективности автоматического определения деструктивных паралингвистических явлений в разговорной речи. Решенная задача имеет важное значение для совершенствования методов и программных решений, используемых при паралингвистическом анализе разговорной речи в условиях недостаточного количества обучающих данных и их дисбалансе, а также при ограничениях в вычислительных ресурсах.

В процессе выполнения диссертационного исследования получены новые научные результаты, составляющие **итоги** исследования:

1. Предложен и исследован комплекс методов анализа речевого сигнала для определения деструктивных паралингвистических явлений в разговорной речи, включающий в себя: 1) метод определения ложности/истинности в разговорной речи HTLBbM-Deception; 2) метод определения депрессии в разговорной речи SBC-Depression; 3) метод определения агрессии в разговорной речи EbM-Aggression.

2. Разработана методика интегрального оценивания степени выраженности деструктивных паралингвистических явлений в разговорной речи диктора, которая учитывает результаты классификации каждого из трех методов и на их основе вычисляет интегральную оценку на с использованием ряда правил.

3. Разработана архитектура и прототип программной системы интегрального анализа деструктивных паралингвистических явлений в разговорной речи DesBDet, проведены экспериментальные исследования программных реализаций методов определения деструктивных паралингвистических явлений и получены следующие количественные результаты: 1) метод определения ложности/истинности в речевых высказываниях достигает F1-меры = 88,4%; 2) метод определения депрессивного состояния в речевых высказываниях достигает F1-меры = 64,0%; 3) метод агрессии в речевых

высказываниях достигает UAR = 76,5%; 4) интегральные F1-мера и невзвешенная средняя полнота для комплекса методов: $F1_{int}$ = 76,8%, UAR_{int} = 75,0%.

Рекомендации. Предложенные методы и разработанная программная система интегрального анализа деструктивных паралингвистических явлений в разговорной речи могут быть использованы как самостоятельно, так и в составе сложной системы комплексного анализа и распознавания речи человека. Такая система сможет учитывать не только аудио-, но также и видеоинформацию, и текстовые транскрипции, что, согласно исследованиям аналогичных многомодальных систем, может улучшить результаты распознавания деструктивных паралингвистических явлений. Ограничения предложенного комплекса методов связаны с относительно небольшим количеством исследовательских данных и дисбалансом классов в них, а также ограниченными доступными вычислительными ресурсами.

Основным эффектом от использования предлагаемого комплекса методов анализа речевого сигнала для определения деструктивных паралингвистических явлений в разговорной речи является психологический комфорт пользователей при взаимодействии в сети Интернет. Кроме того, разработанная программная система на основе предложенного комплекса методов может применяться для первичной оценки состояния пациентов при консультации с медицинскими специалистами в качестве одного из методов оценки психологического состояния пациента наряду с классическими подходами первичной оценки пациента (опросники, тесты и т.д.). Результаты диссертационного исследования были успешно апробированы в СПИИРАН, входящем в СПб ФИЦ РАН, и ООО «Первый психотерапевтический».

Перспективы дальнейшей разработки темы состоят в развитии предложенного комплекса методов и программной реализации в нескольких направлениях, в частности, увеличение количества модальностей, а именно анализ не только акустических характеристик речи, но и смысловой ее составляющей (текста высказываний), а также анализ визуальных образов (видеоданных лица человека). Развитие программной реализации возможно в направлении

использования новейших программных библиотек, которые смогут позволить повысить как скорость исполнения программного кода, так и улучшить дизайн и повысить удобство использования графического пользовательского интерфейса.

Полученные результаты соответствуют специальности 2.3.5 – Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей.

СПИСОК ТЕРМИНОВ И СОКРАЩЕНИЙ

UAR – Unweighted Average Recall, показатель невзвешенной средней полноты

PHQ – Patient Health Questionnaire, шкала депрессии

CCC – Concordance Correlation Coefficient, коэффициент корреляции согласованности

RMSE – Root Mean Squared Error, среднеквадратичная ошибка

MFCC – Mel-frequency Cepstral Coefficients, мел-частотные кепстральные коэффициенты

BoAW – Bag of Audio Words, мешок аудио слов

ResNet – Residual neural network, остаточная нейронная сеть

KELM – Kernel Extreme Learning Machine, метод экстремального обучения с ядерной функцией

LSTM – Long-Short Term Memory, сеть с долгой краткосрочной памятью

GCNN – Gated Convolutional Neural Network, управляемая свёрточная нейронная сеть

MAE – Mean Absolute Error, средняя абсолютная ошибка

LASSO – Least Absolute Shrinkage and Selection Operator, метод регрессионного анализа

TF-IDF – term frequency inverse document frequency, статистическая мера оценки текста

LIWC – Linguistic Inquiry and Word Count, программное обеспечение

k-NN – k-Nearest Neighbours, метод k ближайших соседей

CvR – Classification via Regression, классификация путем регрессии

SMO – Sequential Minimal Optimization, последовательная минимальная оптимизация

SGD – Stochastic Gradient Descent, метод стохастического градиента

CNN – Convolutional Neural Networks, свёрточная нейронная сеть

RNN – Recurrent Neural Networks, рекуррентная нейронная сеть

SMOTE – Synthetic Minority Oversampling Technique, синтетическое увеличение объектов миноритарного класса

ADASYN – Adaptive Synthetic Minority Oversampling, адаптивное синтетическое увеличение объектов миноритарного класса

PCA – Principal Component Analysis, метод анализа главных компонент

СПИСОК ЛИТЕРАТУРЫ

1. Карпов А.А., Кайа Х., Салах А.А. Актуальные задачи и достижения систем паралингвистического анализа речи // Научно-технический вестник информационных технологий, механики и оптики. – 2016. – Т. 16. – № 4. – С. 581–592. DOI: 10.17586/2226-1494-2016-16-4-581-592.

2. Горшков Ю.Г., Дорофеев А.В. Речевые детекторы лжи коммерческого применения // Информационный мост (ИНФОРМОСТ). Радиоэлектроника и Телекоммуникация. – 2003. – №6. – С. 13-15.

3. Ekman P. Telling Lies. Clues to Deceit in the Marketplace, Politics and Marriage // New York, USA: W. W. Norton & Company, Inc. P. 368.

4. Майсак Н.В. Матрица социальных девиаций: классификация типов и видов девиантного поведения // Современные проблемы науки и образования. – 2010. – № 4. – С. 78-86.

5. Balsamo M. Anger and Depression: Evidence of a Possible Mediating Role for Rumination // Psychological reports. 2010. Vol. 106. P. 3-12. DOI: 10.2466/PR0.106.1.3-12.

6. Luutonen S. Anger and depression - Theoretical and clinical considerations // Nordic journal of psychiatry. 2007. Vol. 61. P. 246-251. DOI: 10.1080/08039480701414890.

7. Robbins P., Tanck R. Anger and Depressed Affect: Interindividual and Intraindividual Perspectives // The Journal of psychology. 1997. Vol. 131. P. 489-500. DOI: 10.1080/00223989709603537.

8. Ng T., Sorensen K., Zhang Y., et al. Anger, anxiety, depression, and negative affect: Convergent or divergent? // Journal of Vocational Behavior. 2018. Vol. 110. P. 186–202 DOI: 10.1016/j.jvb.2018.11.014.

9. Bhardwaj V., Angkaw A., Franceschetti M., et. al. Direct and indirect relationships among posttraumatic stress disorder, depression, hostility, anger, and verbal and physical aggression in returning veterans // Aggressive Behavior. 2019. Vol. 45(4). P. 417-426.

10. Cui R., Owsiany M., Turiano N., et al. Association between anger and suicidal ideation // Current Psychology. 2022. DOI: 10.1007/s12144-021-02577-8.

11. Huntsinger J.R. Anger enhances correspondence between implicit and explicit attitudes // Emotion. 2013. Vol. 13(2). P. 350-357. DOI: 10.1037/a0029974.

12. Yip J.A., Schweitzer M.E. Mad and misleading: Incidental anger promotes deception // Organizational Behavior and Human Decision Processes. 2016. Vol. 137. P. 207–217. DOI: 10.1016/j.obhdp.2016.09.006.

13. Величко А.Н., Будков В.Ю., Карпов А.А. Аналитический обзор компьютерных паралингвистических систем для автоматического распознавания лжи в речи человека // Информационно-управляющие системы. – 2017. – №5 (90). – С. 30-41.

14. Amiriparian S., Pohjalainen J., Marchi E., Pugachevskiy S., Schuller B. Is Deception Emotional? An Emotion-Driven Predictive Approach // In Proc. Of INTERSPEECH-2016. 2016. P. 2011-2015.

15. Родькина О.Я., Никольская В.А. К проблеме распознавания психоэмоционального состояния человека по речи с использованием автоматизированных систем // Информационные технологии. – 2016. – Т.22. – №10. – С. 728-733.

16. Савченко В.В., Васильев Р.А. Анализ эмоционального состояния диктора по голосу на основе фонетического детектора лжи // Научные ведомости Белгородского государственного университета. – 2014. – Т.32/1. – № 21 (192). – С. 186-195.

17. Ляксо Е.Е., Фролова О.В., Гречаный С.В., Матвеев Ю.Н., Верхоляк О.В., Карпов А.А. Голосовой портрет ребенка с типичным и атипичным развитием // под ред. Ляксо Е.Е., Фроловой О.В. – СПб. – 2020. – 204 с.

18. Kirchhubel C., Stedmon A., Howard D.M. Analyzing Deceptive Speech. Engineering Psychology and Cognitive Ergonomics // Understanding Human Cognition. EPCE 2013. Lecture Notes in Computer Science. Berlin: Springer, Heidenberg, 2013. Vol 8019. P. 134-141.

19. Montacié C., Caraty M.-J. Prosodic Cues and Answer Type Detection for the Deception Sub-Challenge // In Proc. of INTERSPEECH-2016. 2016. P. 2016-2020. DOI: 10.21293/1818-0442-2016-19-2-56-60.

20. Levitan S.I., An G., Ma M., et al. Combining Acoustic-Prosodic, Lexical, and Phonotactic Features for Automatic Deception Detection // In Proc. of INTERSPEECH-2016. 2016. P. 2006-2010.

21. Herms R. Prediction of Deception and Sincerity from Speech using Automatic Phone Recognition-based Features // In Proc. of INTERSPEECH-2016. San Francisco, USA. 2016. P. 2036-2040.

22. Kaya H., Karpov A. Fusing Acoustic Feature Representations for Computational Paralinguistics Tasks // In Proc. of INTERSPEECH-2016. 2016. P. 2046-2050.

23. Pan X., Zhao H., Zhou Y. The Application of Fractional Mel Cepstral Coefficient in Deceptive Speech Detection // PeerJ. 2015. DOI: 10.7717/peerj.1194.

24. Levitan S.I., An G., Wang M., et al. Cross-Cultural Production and Detection of Deception from Speech // In Proc. of the ACM on Workshop on Multimodal Deception Detection. 2015. P. 1-8.

25. Levitan S.I., Levitan Y., An G., et al. Identifying Individual Differences in Gender, Ethnicity, and Personality from Dialogue for Deception Detection // In Proc. NAACL Workshop on Computational Approaches to Deception Detection. 2016. P. 40-44.

26. Pennebaker J.W., Booth R.J., Boyd R.L., Francis M.E. Linguistic Inquiry and Word Count: LIWC2015 // Austin, TX: Pennebaker Conglomerates (www.LIWC.net). 2015.

27. Zhang J., Levitan S.I., Hirschberg J. Multimodal Deception Detection Using Automatically Extracted Acoustic, Visual, and Lexical Features // In Proc. of INTERSPEECH-2020. P. 359-363. DOI: 10.21437/Interspeech.2020-2320.

28. Mansbach N., Neiterman E., Azaria A. An Agent for Competing with Humans in a Deceptive Game Based on Vocal Cues // In Proc. of INTERSPEECH-2021. 2021. P. 4134-4138. DOI: 10.21437/Interspeech.2021-83.

29. World Health Organization. Depression and Other Common Mental Disorders: Global Health Estimates // Technical Report. World Health Organization. 2017. Licence: CC BY-NC-SA 3.0 IGO.

30. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017 // The Lancet. 2018. DOI: 10.1016/S0140-6736(18)32279-7.

31. Spitzer R.L. Patient health questionnaire: PHQ // New York State Psychiatric Institute. 1999.

32. Beck A.T., Ward C.H., Mock J., et al. An inventory for measuring depression // Archives of General Psychiatry. 1961. Vol. 4. P. 561–571. DOI: 10.1001/archpsyc.1961.01710120031004.

33. Rush A.J., Trivedi M.H., Ibrahim H.M., et al. The 16-item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): A psychometric evaluation in patients with chronic major depression // Biological Psychiatry. 2003. Vol. 54(5). P. 573–583. DOI: 10.1016/S0006-3223(02)01866-8.

34. Gonzalez J.S., Shreck E., Batchelder A. Hamilton Rating Scale for Depression (HAM-D) // In: Gellman MD, Turner JR, editors. Encyclopedia of behavioral medicine. New York: Springer. 2013. P. 887–888. DOI: 10.1007/978-1-4419-1005-9_198.

35. Величко А.Н., Карпов А.А. Аналитический обзор систем автоматического определения депрессии по речи // Информатика и автоматизация. – 2021. – № 3 (20). – С. 497-529.

36. Valstar M., Schuller B., Smith K., et al. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge // In Proc. of the 3rd ACM International Workshop on Audio/visual Emotion Challenge (AVEC'13). 2013. P. 3–10. DOI: 10.1145/2512530.2512533.

37. Valstar M., Schuller B., Smith K., et al. AVEC 2014 – 3D dimensional affect and depression recognition challenge // In Proc. of the 4th ACM International Workshop on Audio/visual Emotion Challenge (AVEC'14). 2014. P. 3-10. DOI: 10.1145/2661806.2661807.

38. Valstar M., Gratch J., Schuller B., et al. Summary for AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge // In Proc. of the 24th ACM International Conference on Multimedia (MM '16). 2016. P. 1483–1484. DOI: 10.1145/2964284.2980532.

39. Ringeval F., Schuller B., Valstar M., et al. AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge // In Proc. of the 7th ACM International Workshop on Audio/visual Emotion Challenge (AVEC '17). 2017. P. 3–9. DOI: 10.1145/3133944.3133953.

40. Ringeval F., Schuller B., Valstar M., et al. AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition // In Proc. of the 9th ACM International Workshop on Audio/visual Emotion Challenge (AVEC '19). Association for Computing Machinery, New York, NY, USA. 2019. P. 3–12. DOI: 10.1145/3347320.3357688.

41. Kostyuchenko E., Meshcheryakov R., Ignatieva D., Pyatkov A., Choynzonov E., Balatskaya L.: Correlation criterion in assessment of speech quality in process of oncological patients rehabilitation after surgical treatment of the speech-producing tract // In: Bhatia, S.K., Tiwari, S., Mishra, K.K., Trivedi, M.C. (eds.) Advances in Computer Communication and Computational Sciences. 2019. Vol. 759. P. 209–216. DOI: 10.1007/978-981-13-0341-8_19.

42. Matveev Y., Matveev A., Frolova O., Lyakso E. Automatic Recognition of the Psychoneurological State of Children: Autism Spectrum Disorders, Down Syndrome, Typical Development // In Proc. of the 23th International Conference on Speech and Computer SPECOM 2021. Lecture Notes in Computer Science, Springer, Cham. 2021. Vol 12997. DOI: 10.1007/978-3-030-87802-3_38.

43. Потапова Р.К. Вариативность акустических параметров звучащей речи // Вестник Московского государственного лингвистического университета. Гуманитарные науки. – 2016. – Т. 740. – С. 137-147.

44. Stahl S.M. Stahl's essential psychopharmacology: Neuroscientific basis and practical applications // Cambridge: Cambridge University Press (4th ed.). 2013. P. 628.

45. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (5th ed.). 2013. P. 992. DOI: 10.1176/appi.books.9780890425596.

46. Franklin J.C., Ribeiro J.D., Fox K.R., et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research // Psychol Bull. 2017. Vol. 143(2). P. 187-232. DOI: 10.1037/bul0000084.

47. Belsher B.E., Smolenski D.J., Pruitt L.D., et al. Prediction Models for Suicide Attempts and Deaths: A Systematic Review and Simulation // JAMA Psychiatry. 2017. Vol. 76(6). P. 642–651.

48. Singer K. Depressive disorders from a transcultural perspective // Social Science & Medicine. 1975. Vol. 9. P. 289-301. DOI: 10.1016/0037-7856(75)90001-3.

49. Lin L.I. A concordance correlation coefficient to evaluate reproducibility // Biometrics. 1989. Vol 45(1). P. 255-268.

50. Willmott C.J., Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance // Climate Research. 2005. Vol. 30. P. 79–82. DOI: 10.3354/cr030079.

51. Gratch J., et al. The Distress Analysis Interview Corpus of Human and Computer Interviews // In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). 2014. P. 3123-3128.

52. Jun Y., Yu-Gang J., Alexander H., et al. Evaluating Bag-of-Visual-Words Representations in Scene Classification // In Proceedings of the international Workshop on Workshop on Multimedia Information Retrieval. 2007. Vol. 2. P. 197-206.

53. Kaya H., Fedotov D., Dresvyanskiy D., et al. Predicting depression and emotions in the crossroads of cultures, paralinguistics, and non-linguistics // In Proc. of the 9th ACM International Workshop on Audio/visual Emotion Challenge (AVEC '19). 2019. P. 27–35. DOI: 10.1145/3347320.3357691.

54. Harris Z. Distributional Structure // WORD. 1954. Vol. 10:2-3. P. 146-162. DOI: 10.1080/00437956.1954.11659520.

55. Ray A., Kumar S., Reddy R., et al. Multi-level Attention Network using Text, Audio and Video for Depression Prediction // In Proc. of the 9th ACM International Workshop on Audio/visual Emotion Challenge (AVEC '19). P. 81–88. DOI: 10.1145/3347320.3357697.

56. Makiuchi M.R., Warnita T., Uto K., et al. Multimodal Fusion of BERT-CNN and Gated CNN Representations for Depression Detection // In Proc. of the 9th ACM International Workshop on Audio/visual Emotion Challenge (AVEC '19). 2019. P. 55–63. DOI: 10.1145/3347320.3357694.

57. Devlin J., Chang M., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // In Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. Vol. 1. P. 4171–4186. DOI: 10.18653/v1/N19-1423.

58. Fan W., He Z., Xing X., et al. Multi-modality Depression Detection via Multi-scale Temporal Dilated CNNs // In Proc. of the 9th ACM International Workshop on Audio/visual Emotion Challenge (AVEC '19), Association for Computing Machinery. 2019. P. 73–80. DOI: https://doi.org/10.1145/3347320.3357695.

59. Yin S., Liang X., Ding H., et al. A Multi-Modal Hierarchical Recurrent Neural Network for Depression Detection // In Proc. of the 9th International Audio/Visual Emotion Challenge and Workshop (AVEC '19), Association for Computing Machinery, New York, NY, USA. 2019. P. 65-71. DOI: https://doi.org/10.1145/3347320.3357696.

60. Haque A., Guo M., Miner A.S., et al. Measuring Depression Symptom Severity from Spoken Language and 3D Facial Expressions // Machine Learning for Health (ML4H) Workshop at NeurIPS 2018. 2018. http://arxiv.org/abs/1811.0859.

61. Altman D., Bland J. Diagnostic tests. 1: Sensitivity and specificity // BMJ. 1994. Vol. 308(6943): 1552. DOI: 10.1136/bmj.308.6943.1552.

62. Qureshi S. A., Saha S., Hasanuzzaman M., et al. Multitask Representation Learning for Multimodal Estimation of Depression Level // In IEEE Intelligent Systems. 2019. Vol. 34(5). P. 45-52. DOI: 10.1109/MIS.2019.2925204.

63. Niu M., Tao J., Liu B., et al. Automatic Depression Level Detection via l_p-Norm Pooling // In Proc. of INTERSPEECH-2019. 2019. P. 4559-4563.

64. Rohanian M., Hough J., Purver M. Detecting depression with word-level multimodal fusion // In Proc. of INTERSPEECH-2019. 2019. P. 1443-1447.

65. Tao F., Esposito A., Vinciarelli A. Spotting the traces of depression in read speech: An Approach Based on Computational Paralinguistics and Social Signal Processing // In Proc. of INTERSPEECH-2020. 2020. P.1828-1832.

66. Xezonaki D., Paraskevopoulos G., Potamianos A., et al. Affective Conditioning on Hierarchical Networks applied to Depression Detection from Transcribed Clinical Interviews // In Proc. of INTERSPEECH-2020. 2020. P. 4556-4560.

67. Huang Zh., Epps J., Joachim D., et al Domain Adaptation for Enhancing Speech based Depression Detection in Natural Environmental Conditions Using Dilated CNNs // In Proc. of INTERSPEECH-2020. 2020. P. 4561-4565.

68. Zhao Z., Li Q., Cummins N., et al. Hybrid Network Feature Extraction for Depression Assessment from Speech // In Proc. of INTERSPEECH-2020. 2020. P. 4956-4960.

69. Beck A.T., Steer R.A., Brown G. Beck Depression Inventory–II // APA PsycTests. 1996. P.38. DOI: 10.1037/t00742-000.

70. Seneviratne N., Williamson J.R., Lammert A. C., et al. Extended Study on the Use of Vocal Tract Variables to Quantify Neuromotor Coordination in Depression // In Proc. of INTERSPEECH- 2020. 2020. P. 4551-4555.

71. Kuznetsova Y.M., Kiselnikova N.V., Enikolopov S.N. et al. Predicting Depression from Essays in Russian. Computational Linguistics and Intellectual Technologies // In Proc. of the International Conference “Dialogue 2019”. 2019. P. 647-657.

72. Stankevich M., Ignatiev N. Smirnov I. Predicting Depression with Social Media Images // In Proc. of the 9th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2020). 2020. P. 235-240.

73. Stankevich M., Isakov V., Devyatkin D., et al. Feature Engineering for Depression Detection in Social Media // In Proc. of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018). 2020. P. 426-431.

74. Stankevich M., Smirnov I., Kiselnikova N., et al. Depression Detection from Social Media Profiles // Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2019. Communications in Computer and Information Science. 2019. Vol. 1223. P. 181-194.

75. Ениклопов С.Н., Кузнецова Ю.М., Пенкина М.Ю., и др. Особенности текста и психологические особенности: опыт эмпирического компьютерного исследования // Труды Института системного анализа РАН. – 2019. – Т. 69. – №3. – С. 91-99.

76. Jones K.S. A statistical interpretation of term specificity and its application in retrieval // Document retrieval systems. Taylor Graham Publishing. GBR. 1988. P. 132–142.

77. Величко А.Н. Метод анализа речевого сигнала для автоматического определения агрессии в разговорной речи // Вестник ВГУ. Системный анализ и информационные технологии. – 2022. – № 4. – С. 180-188.

78. Buss A., Durkee A. An inventory for assessing different kinds of hostility // Journal of Consulting Psychology. 1957. Vol 21(4). P. 343–349. DOI: 10.1037/h0046900.

79. McWilliams N. Psychoanalytic diagnosis: Understanding personality structure in the clinical process. 2nd ed. Guilford Press. 2011. P. 426.

80. Sobin C., Alpert M. Emotion in Speech: The Acoustic Attributes of Fear, Anger, Sadness, and Joy // J Psycholinguist. 1999. Res 28. P. 347–365. DOI: 10.1023/A:1023237014909.

81. Egas-López J.V., Vetráb M., Tóth L., Gosztolya G. Identifying Conflict Escalation and Primates by Using Ensemble X-Vectors and Fisher Vector Features // In

Proc. of INTERSPEECH-2021. 2021. P. 476-480. DOI: 10.21437/Interspeech.2021-1173.

82. Lefter I., Jonker C.M. Aggression recognition using overlapping speech // Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). 2017. P. 299-304. DOI: 10.1109/ACII.2017.8273616.

83. Sahoo S., Routray A. Detecting Aggression in Voice Using Inverse Filtered Speech Features // IEEE Transactions on Affective Computing. 2018. Vol. 9. Issue 2. P. 217-226. DOI: 10.1109/TAFFC.2016.2615607.

84. Zhou Z., Xu Y., Li M. Detecting Escalation Level from Speech with Transfer Learning and Acoustic-Lexical Information Fusion. 2021. DOI: arXiv:2104.06004v2.

85. Величко А.Н., Карпов А. А., Будков В. Ю. Аналитический обзор речевых корпусов для систем определения ложных речевых сообщений // Материалы конференции «Информационные технологии в управлении» ИТУ-2018 в рамках МКПУ-2018, Санкт-Петербург, 2018, С. 638-642.

86. Schuller B., Steidl S., Batliner A., et al. The INTERSPEECH 2016 Computational Paralinguistic Challenge: Deception, Sincerity & Native Language // In Proc. of INTERSPEECH-2016. 2016. P. 2001-2005.

87. Hirschberg J., Benus S., Brenier J., et al. Distinguishing Deceptive from Non-Deceptive Speech // In Proc. of INTERSPEECH-2005. 2005. P. 1833-1836.

88. Pérez-Rosas V., Abouelenien M., Mihalcea R., Burzo, M. Deception detection using real-life trial data // In Proc. of the 2015 ACM International Conference on Multimodal Interaction. 2015. P. 59-66.

89. Soldner F., Pérez-Rosas V., Mihalcea R. Box of lies: Multimodal deception detection in dialogues // In Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. Vol. 1. P. 1768-1777.

90. Pérez-Rosas V., Mihalcea R. Cross-cultural Deception Detection // In Proc. of ACL 2014. 52nd Annual Meeting of the Association for Computational Linguistics. 2014. P. 440-445.

91. Kaya H., Karpov A. Efficient and effective strategies for cross-corpus acoustic emotion recognition // Neurocomputing. Vol 275(2). 2018. P. 1028-1034.

92. Alghowinem S., Goecke R., Wagner M., et al. From joyous to clinically depressed: Mood detection using spontaneous speech // In Proc. of FLAIRS Conference, G. M. Youngblood and P. M. McCarthy, Eds. AAAI Press. 2012. P. 141–146.

93. Yang Y., Fairbairn C., Cohn J. Detecting depression severity from vocal prosody // IEEE Transactions on Affective Computing. 2013. Vol. 4(2). P. 142–150.

94. Mundt J.C., Snyder P.J., Cannizzaro M.S., et al. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology // Journal of Neurolinguistics. 2007. Vol. 20(1). P. 50 – 64.

95. General Psychotherapy Corpus. URL: http://alexanderstreet.com. (дата обращения: 10.12.2020).

96. Huang Z., Epps J., Joachim D., et al. Depression detection from short utterances via diverse smartphones in natural environmental conditions // In Proc. of INTERSPEECH-2018. 2018. P. 3393–3397.

97. Litvinova T., Ryzhkova E., Litvinova O. Features of Written Texts of People with Different Profiles of the Lateral Brain Organization of Functions (on the Basis of RusNeuroPsych Corpus) // In Proc. of 7th Tutorial and Research Workshop on Experimental Linguistics, ExLing 2016. 2016. 2016. P. 107-110.

98. Lefter I., Burghouts G.J., Rothkrantz L.J. An audio-visual dataset of human–human interactions in stressful situations // Journal on Multimodal User Interfaces. 2014. Vol. 8(1). P. 29–41.

99. Lefter I., Rothkrantz L., Burghouts G. A comparative study on automatic audio–visual fusion for aggression detection using meta-information // Pattern Recognition Letters. 2013. Vol. 34(15). P. 1953-1964. DOI: 10.1016/j.patrec.2013.01.002.

100. Zadeh A., Liang P., Poria S., et al. Multi-attention recurrent network for human communication comprehension // In Proc. of the Thirty-Second AAAI Conference on Artificial Intelligence. 2018. P. 5642-5649.

101. 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. Vol. 13(5): e0196391. DOI: 10.1371/journal.pone.0196391.

102. Busso C., Bulut M., Lee C., et al. IEMOCAP: interactive emotional dyadic motion capture database // Language Resources Evaluation. 2008. Vol. 42(4). P. 335-359.

103. Perepelkina O., Kazimirova E., Konstantinova M. RAMAS: Russian Multimodal Corpus of Dyadic Interaction for studying emotion recognition // PeerJ Preprints 6:e26688v1. 2018. DOI: 10.7287/peerj.preprints.26688v1.

104. Simonyan K., Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition // International Conference on Learning Representations. 2015. arXiv:1409.1556v6.

105. Krizhevsky A., Sutskever I., Hinton G. ImageNet classification with deep convolutional neural networks // Commun. ACM 60. 2017. P. 84–90. DOI: 10.1145/3065386.

106. Huang G., Liu Z., Van Der Maaten L., Weinberger K. Densely Connected Convolutional Networks // In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. P. 2261-2269. DOI: 10.1109/CVPR.2017.243.

107. Amiriparian S., Sokolov A., Aslan I., et al. On the Impact of Word Error Rate on Acoustic-Linguistic Speech Emotion Recognition: An Update for the Deep Learning Era. 2021. arXiv: abs/2104.10121.

108. Frank E., Hall M.A., Witten I.H.: The WEKA Workbench // Online Appendix for “Data Mining: Practical Machine Learning Tools and Techniques”. 2016. 4th edn. Morgan Kaufmann.

109. Breiman L. Bagging predictors // Mach Learn. 1996. 24. P. 123–140. DOI: 10.1007/BF00058655.

110. Fix E., Hodges J.L. Discriminatory analysis, nonparametric discrimination: Consistency properties // Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas. 1951.

111. Frank E., Wang Y., Inglis S., et al. Using model trees for classification // Machine Learning. 1998. Vol. 32(1). P. 63-76.

112. Freund Y., Schapire R. Experiments with a new boosting algorithm // In: Thirteenth International Conference on Machine Learning. San Francisco. 1996. P. 148-156.

113. Platt J. Sequential minimal optimization: A fast algorithm for training support vector machines // Advances in kernel methods – support vector learning. 1998.

114. Robbins H., Monro S. A Stochastic Approximation Method // The Annals of Mathematical Statistics. 1951. Vol. 22(3). P. 400-407, DOI: 10.1214/aoms/1177729586.

115. Ho T.K. Random decision forests // In: Proceedings of 3rd international conference on document analysis and recognition. 1995. P. 278–82.

116. Breiman L. Random Forests // Machine Learning. 2001. Vol. 45(1). P.5-32.

117. Kiefer J., Wolfowitz J. Stochastic Estimation of the Maximum of a Regression Function // Annals of Mathematical Statistics. 1952. Vol. 23(3). P. 462-466.

118. Holte R.C. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets // Machine Learning 11. 1993. P. 63–90. DOI: 10.1023/A:1022631118932.

119. Cortes C. Vapnik V. Support-vector networks // Machine learning. 1995. Vol. 20(3). P. 273–297.

120. Wolpert D.H. Stacked generalization // Neural networks. 1992. Vol. 5(2). P. 241-259.

121. Friedman J.H. Greedy function approximation: a gradient boosting machine // Annals of statistics. 2001. P.1189–232.

122. Dorogush A.-V., Ershov V., Gulin A. CatBoost: gradient boosting with categorical features support // Workshop on ML Systems at NIPS 2017. 2017.

123. Ke G. Meng Q. et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree // Advances in Neural Information Processing Systems. 2017. P. 3146–3154.

124. Tianqi Ch. Guestrin C. XGBoost: A Scalable Tree Boosting System. // In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 785–794.

125. McCulloch W. S, Pitts W. A logical calculus of the ideas immanent in nervous activity // The bulletin of mathematical biophysics. 1943. Vol. 5(4). P.115–33.

126. Rumelhart D., Hinton G., Williams R. Learning representations by back-propagating errors // Nature. 1986. Vol. 323. P. 533–536. DOI: 10.1038/323533a0.

127. Hochreiter S., Schmidhuber J. Long Short-Term Memory // Neural Comput. 1997. Vol. 9(8). P. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

128. Arik S.O., Pfister T. TabNet: Attentive Interpretable Tabular Learning // In Proc. of the AAAI Conference on Artificial Intelligence. Vol. 35(8). P. 6679-6687. DOI: 10.1609/aaai.v35i8.16826

129. Sokolov A., Savchenko A.V.: Gender domain adaptation for automatic speech recognition // In Proc. of 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI). 2021. P. 413–418.

130. Sidorov M., Schmitt A., Semenkin E. et al. Could speaker, gender or age awareness be beneficial in speech-based emotion recognition? // In Proc. of Language Resources and Evaluation (LREC). 2016. P. 61–68.

131. Tomek I. An experiment with the edited nearest-neighbor rule // IEEE Transactions on Systems, Man, and Cybernetics. 1976. Vol. 6(6). P. 448-452.

132. About Python [Электронный ресурс]. Python.org. 2022. https:// www.python.org/about/.

133. Why TensorFlow [Электронный ресурс]. TensorFlow.org. 2022. https:// www.tensorflow.org/about.

134. Chollet F, et al. Keras [Электронный ресурс]. GitHub; 2015. https://github.com/fchollet/keras.

135. Pedregosa F., Varoquaux G., Gramfort A., et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. 2011. Vol. 12. P. 2825–2830.

136. Burkhardt F., Eckert M., Johannsen W., et al. A database of age and gender annotated telephone speech // In Proc. of Language Resources and Evaluation (LREC). Malta. 2010. P. 1562–1565.

137. Burkhardt F., Paeschke A., Rolfes M., et al. A database of german emotional speech // In Proc. of INTERSPEECH-2005. 2005. P. 1517–1520.

138. Makarova V., Petrushin V. Ruslana: a database of Russian emotional utterances // In Proc. of the 7th International Conference on Spoken Language Processing. 2002. Vol. 1. P. 2041–2044.

139. Stehman S. V. Selecting and interpreting measures of thematic classification accuracy // Remote Sensing of Environment. 1997. Vol. 62(1). P. 77–89. DOI: 10.1016/S0034-4257(97)00083-7.

140. Schuller B. Batliner A. et al. The INTERSPEECH 2011 speaker state challenge // In Proc. of INTERSPEECH-2011. 2011. P. 3201–3204.

141. Mendels G., Levitan S.I., Lee K., Hirschberg J. Hybrid acoustic-lexical deep learning approach for deception detection // In Proc. of INTERSPEECH-2017. 2017. P. 1472–1476.

142. Eyben F. et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing // In IEEE Transactions on Affective Computing. Vol. 7(2). P. 190-202. DOI: 10.1109/TAFFC.2015.2457417.

143. Schuller B., Batliner A., Bergler C., et al. The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates // In Proc. of INTERSPEECH-2021. 2021. P. 431-435. DOI: 10.21437/Interspeech.2021-19.

Приложение А. Список публикаций по теме диссертации

Публикации в ведущих рецензируемы научных журналах и изданиях из перечня рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук:

1. Величко А.Н. Метод анализа речевого сигнала для автоматического определения агрессии в разговорной речи // Вестник ВГУ. Системный анализ и информационные технологии. – 2022. – № 4. – С. 180-188.
2. Величко А.Н., Карпов А.А. Аналитический обзор систем автоматического определения депрессии по речи // Информатика и автоматизация. – 2021. – № 3 (20). – С. 497-529.
3. Величко А.Н., Карпов А.А., Будков В.Ю. Исследование методов классификации для автоматического определения истинной или ложной информации в речевых сообщениях // Научный вестник НГТУ. – 2018. – № 3. – С. 21-32.
4. Двойникова А.А., Маркитантов М.В., Рюмина Е.В., Уздяев М.Ю., Величко А.Н., Рюмин Д.А., Ляксо Е.Е., Карпов А.А. Анализ информационного и математического обеспечения для распознавания аффективных состояний человека // Информатика и автоматизация. – 2022. – № 6 (21). – С. 1097-1144.

Публикации в ведущих российских и иностранных научных изданиях, входящих в перечни WoS/Scopus:

1. Velichko A., Markitantov M., Kaya H., Karpov A. Complex Paralinguistic Analysis of Speech: Predicting Gender, Emotions and Deception in a Hierarchical Framework // In Proceedings of the International Conference INTERSPEECH-2022. 2022. P. 4735-4739.
2. Velichko A.N., Karpov A.A. Automatic Detection of Deceptive and Truthful Paralinguistic Information in Speech using Two-Level Machine Learning Model //

Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2021”. 20 (27). 2021. P. 698-704.

3. Verkholyak O., Dresvyanskiy D., Dvoynikova A., Kotov D., Ryumina E., Velichko A., Mamontov D., Minker W., Karpov A. Ensemble-Within-Ensemble Classification for Escalation Prediction from Speech // In Proceedings of the International Conference INTERSPEECH-2021. 2021. P. 481-485.

4. Velichko A., Karpov A. A Study of Data Scarcity Problem for Automatic Detection of Deceptive Speech Utterances // In Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019) Saint Petersburg, Russia, 2019. CEUR-WS. Vol. 2552. 2020. P. 38-46.

5. Velichko A., Budkov V., Kagirov I., Karpov A. Applying Ensemble Learning Techniques and Neural Networks to Deceptive and Truthful Information Detection Task in the Flow of Speech // Intelligent Distributed Computing XIII. IDC 2019. Studies in Computational Intelligence, Springer, Cham. Vol. 868. 2019. P. 477-482.

6. Levonevskii D., Shumskaya O., Velichko A., Uzdiaev M., Malov D. Methods for Determination of Psychophysiological Condition of User within Smart Environment Based on Complex Analysis of Heterogeneous Data // In Proceedings of 14th International Conference on Electromechanics and Robotics “Zavalishin's Readings”. Smart Innovation, Systems and Technologies. Springer, Cham. Vol. 154. 2019. P. 511-523.

7. Velichko A.N., Budkov V.Y., Kagirov I. A., Karpov A.A. Comparative Analysis of Classification Methods for Automatic Deception Detection in Speech // In Proceedings of the 20th International Conference on Speech and Computer SPECOM-2018. Springer, LNAI. Vol. 11096. 2018. P. 737-746.

Публикации в других изданиях:

1. Величко А.Н., Будков В.Ю. Разработка прототипа системы автоматического определения ложной и истинной информации в речи // Материалы семинара «Анализ разговорной русской речи» (АР3-2019), Санкт-Петербург, 2019, С. 17-20.

2. Величко А.Н., Карпов А. А., Будков В. Ю. Аналитический обзор речевых корпусов для систем определения ложных речевых сообщений // Материалы конференции «Информационные технологии в управлении» ИТУ-2018 в рамках МКПУ-2018, Санкт-Петербург, 2018, С. 638-642.

3. Dvoynikova A., Markitantov M., Ryumina E., Uzdiaev M., Velichko A., Kagirov I., Kipyatkova I., Lyakso E., Karpov A. An analysis of automatic techniques for recognizing human's affective states by speech and multimodal data. Proceedings of the 24th International Congress on Acoustics ICA-2022. 2022. P. 22-33.

Регистрация результатов интеллектуальной деятельности:

1. Величко А.Н. Программное обеспечение для определения депрессивного состояния по речи человека. Свидетельство № 2021680548. Зарегистрировано в Реестре программ для ЭВМ 13.12.2021.

2. Величко А.Н., Верхоляк О.В., Карпов А.А. Программная система для распознавания эмоций в речи (ProSpER – Program for Speech Emotion Recognition). Свидетельство № 2020664234. Зарегистрировано в Реестре программ для ЭВМ 10.11.2020.

3. Верхоляк О.В., Маркитантов М.В., Величко А.Н., Кипяткова И.С., Карпов А.А. Программная система комплексного анализа паралингвистических явлений в речи (ComPAS - Complex Paralinguistic Analysis of Speech). Свидетельство № 2020664233. Зарегистрировано в Реестре программ для ЭВМ 10.11.2020.

4. Величко А.Н., Будков В.Ю., Карпов А.А. Программная система для автоматического определения ложной и истинной информации в речи. Свидетельство № 2018662956. Зарегистрировано в Реестре программ для ЭВМ 17.10.2018.

Приложение Б. Копии зарегистрированных свидетельств и патентов на результаты интеллектуальной собственности

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2018662956

Программная система для автоматического определения ложной и истинной информации в речи

Правообладатель: ***Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН) (RU)***

Авторы: ***Величко Алёна Николаевна (RU), Будков Виктор Юрьевич (RU), Карпов Алексей Анатольевич (RU)***

Заявка № **2018619973**

Дата поступления **19 сентября 2018 г.**

Дата государственной регистрации в Реестре программ для ЭВМ ***17 октября 2018 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2021680548

Программное обеспечение для определения депрессивного состояния по речи человека

Правообладатель: ***Федеральное государственное бюджетное учреждение науки "Санкт-Петербургский Федеральный исследовательский центр Российской академии наук" (RU)***

Автор(ы): ***Величко Алёна Николаевна (RU)***

Заявка № **2021680259**
Дата поступления **13 декабря 2021 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***13 декабря 2021 г.***

Руководитель Федеральной службы по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 0x02A5CFBC00B1ACF59A40A2F08092E9A118
Владелец **Ивлиев Григорий Петрович**
Действителен с 15.01.2021 по 15.01.2035

Г.П. Ивлиев

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2020664234

Программная система для распознавания эмоций в речи (ProSpER – Program for Speech Emotion Recognition)

Правообладатель: ***Федеральное государственное бюджетное учреждение науки "Санкт-Петербургский Федеральный исследовательский центр Российской академии наук" (RU)***

Авторы: ***Величко Алёна Николаевна (RU), Верхоляк Оксана Владимировна (RU), Карпов Алексей Анатольевич (RU)***

Заявка № **2020663624**

Дата поступления **10 ноября 2020 г.**

Дата государственной регистрации в Реестре программ для ЭВМ ***10 ноября 2020 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2020664233

Программная система комплексного анализа паралингвистических явлений в речи (ComPAS - Complex Paralinguistic Analysis of Speech)

Правообладатель: ***Федеральное государственное бюджетное учреждение науки "Санкт-Петербургский Федеральный исследовательский центр Российской академии наук" (RU)***

Авторы: ***Верхоляк Оксана Владимировна (RU), Маркитантов Максим Викторович (RU), Величко Алена Николаевна (RU), Кипяткова Ирина Сергеевна (RU), Карпов Алексей Анатольевич (RU)***

Заявка № **2020663623**

Дата поступления **10 ноября 2020 г.**

Дата государственной регистрации в Реестре программ для ЭВМ ***10 ноября 2020 г.***

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

Приложение В. Акты о внедрении полученных научных результатов

ООО «Первый психотерапевтический»
ИНН 7839125797 / КПП 783901001

Тел. +7 911 780 84 73
https://1psycenter.ru/

АКТ

об использовании результатов диссертационной работы на соискание ученой степени кандидата технических наук «Методы, алгоритмы и программная система для автоматического определения деструктивных паралингвистических явлений в разговорной речи» Величко Алёны Николаевны
в ООО «Первый психотерапевтический» при первичной диагностике пациентов путем компьютерного тестирования.

Комиссия в составе: генерального директора Жихарева О.В. и врача-психотерапевта Жихаревой О.Б. составила настоящий акт о том, что основные научные результаты диссертационной работы Величко Алёны Николаевны, а именно:

- методика интегрального оценивания степени выраженности деструктивных паралингвистических явлений в речевом сигнале диктора;
- программная система для определения деструктивных паралингвистических явлений в речевом сигнале

были использованы в ООО «Первый психотерапевтический» при первичной диагностике пациентов путем компьютерного тестирования. Комиссия отмечает практическую значимость и новизну полученных в работе результатов.

Члены комиссии:

_____ Жихарев О.В.,

_____ Жихарева О.Б.

«2» 03 2023 г.

М.П. (печать организации)

МИНОБРНАУКИ РОССИИ

**Федеральное государственное бюджетное учреждение науки
«Санкт-Петербургский Федеральный исследовательский центр
Российской академии наук»
(СПб ФИЦ РАН)**

14-я линия, д. 39, г. Санкт-Петербург, 199178
Телефон: (812) 328-33-11, факс: (812) 328-44-50, e-mail: info@spcras.ru, web: http://www.spcras.ru
ОКПО 04683303, ОГРН 1027800514411, ИНН/КПП 7801003920/780101001

УТВЕРЖДАЮ
Директор СПб ФИЦ РАН
Профессор РАН
_____ Ронжин А.Л.
_____ 2023 года

Акт внедрения результатов диссертационного исследования Величко Алёны Николаевны «Методы и программная система автоматического определения деструктивных паралингвистических явлений в разговорной речи», представленного на соискание ученой степени кандидата наук по научной специальности 2.3.5 – Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей (технические науки)

Комиссия в составе: председателя – директора СПИИРАН, доктора технических наук, профессора Осипова Василия Юрьевича; старшего научного сотрудника лаборатории речевых и многомодальных интерфейсов, кандидата технических наук, доцента Кипятковой Ирины Сергеевны; старшего научного сотрудника лаборатории речевых и многомодальных интерфейсов, кандидата технических наук, Рюмина Дмитрия Александровича составила настоящий акт в том, что результаты диссертационного исследования Величко Алёны Николаевны «Методы и программная система автоматического определения деструктивных паралингвистических явлений в разговорной речи» были внедрены при выполнении научно-исследовательских работ в лаборатории речевых и многомодальных интерфейсов (Грант Российского фонда финансирования исследований № 20-37-90144 Аспиранты, 2020-2022; Грант Российского научного фонда № 18-11-00145, 2018-2020). С применением разработанных Величко А.Н. методов и архитектуры программной системы решались задачи выявления деструктивных паралингвистических явлений в разговорной речи, включая:

Архитектура программной системы (DesBDet) для комплексного определения различных деструктивных паралингвистических явлений в речи и методы автоматического определения деструктивных паралингвистических

явлений в разговорной речи использовались при решении задачи разработки определения ложной и истинной информации в речи в проекте: Грант Российского фонда финансирования исследований № 20-37-90144 Аспиранты «Разработка и исследование автоматической системы для выявления деструктивных паралингвистических явлений в разговорной речи».

Метод анализа речевого сигнала на основе нескольких наборов акустических признаков и комплексирования моделей градиентного бустинга для определения ложной и истинной информации в речи использовался при решении задачи разработки определения ложной и истинной информации в речи в проекте: Грант Российского научного фонда № 18-11-00145 «Разработка и исследование интеллектуальной системы для комплексного паралингвистического анализа речи».

Комиссия отмечает теоретическую, практическую значимость и новизну полученных в работе результатов.

Председатель комиссии:

Директор СПИИРАН,
доктор технических наук,
профессор — Осипов В. Ю.

Члены комиссии:

Старший научный сотрудник
лаборатории речевых и
многомодальных интерфейсов,
кандидат технических наук — Кипяткова И. С.

Старший научный сотрудник
лаборатории речевых и
многомодальных интерфейсов,
кандидат технических наук — Рюмин Д. А.