МРНТИ 28.23.15 **https://doi.org/10.58805/kazutb.v.2.31-1613**

# СОВРЕМЕННЫЕ МЕТОДЫ ДЕТЕКЦИИ СИНТЕТИЧЕСКОЙ РЕЧИ И ЗАЩИТЫ СИСТЕМ АВТОМАТИЧЕСКОЙ ВЕРИФИКАЦИИ ДИКТОРА ОТ АТАК ПОДМЕНЫ ГОЛОСА

**Ә.Е. Ержан, А.Т. Ахмедиярова, Ж.К. Алимсеитова**
*Некоммерческое акционерное общество «Казахский национальный исследовательский технический университет им. К.И.Сатбаева», Алматы, Казахстан*
*Корреспондент-автор: a.yerzhan@satbayev.university*

В данной статье представлен систематический аналитический обзор современных методов обнаружения синтезированной речи и защиты систем автоматической верификации диктора от атак с подменой голоса. Рассмотрены основные классы атак с подменой голоса, включая синтез речи из текста, преобразование голоса, атаки с воспроизведением, аудиодипфейки и их комбинированные сценарии. Проведен сравнительный анализ используемых наборов данных, акустических характеристик, архитектур моделей, парадигм обучения и метрик оценки.

Установлено, что современное состояние исследований характеризуется доминированием наборов данных ASVspoof и подходов к обучению с учителем, что обеспечивает сопоставимость результатов, но ограничивает обобщающую способность моделей в условиях переноса между наборами данных и сдвига предметной области. Показано, что остаются ключевые нерешенные проблемы: устойчивость к ранее не наблюдавшимся типам атак, чувствительность к шуму и искажениям канала, а также ограниченная интерпретируемость существующих моделей.

На основе проведенного анализа были выявлены основные методологические ограничения и сформулированы приоритетные направления дальнейших исследований, связанные с разработкой кросс-доменных, надежных и объяснимых систем обнаружения, предназначенных для практической интеграции в схемы защиты ASV. В работе также предложена классификация методов детекции синтетической речи, учитывающая тип используемых признаков и архитектурные особенности моделей.

**Ключевые слова:** синтетическая речь, спуфинг, ASV, аудиодипфейк, TTS, преобразование голоса, атака воспроизведения, антиспуфинг, самоконтролируемое обучение, кросс-доменная устойчивость.

# СИНТЕТИКАЛЫҚ СӨЙЛЕУДІ ДЕТЕКЦИЯЛАУДЫҢ ЖӘНЕ ДИКТОРДЫ ДАУЫСТЫ АУЫСТЫРУ ШАБУЫЛДАРЫНАН АВТОМАТТЫ ВЕРИФИКАЦИЯЛАУ ЖҮЙЕЛЕРІН ҚОРҒАУДЫҢ ҚАЗІРГІ ЗАМАНҒЫ ӘДІСТЕРІ

**Ә.Е. Ержан, А.Т. Ахмедиярова, Ж.К. Алимсеитова**
*«Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті» коммерциялық емес қоғамы, Алматы, Қазақстан,*
*e-mail: a.yerzhan@satbayev.university*

Бұл мақала синтезделген сөйлеуді анықтау және автоматты спикерді растау жүйелерін дауыстық алдау шабуылдарынан қорғаудың қазіргі әдістеріне жүйелі аналитикалық шолу ұсынады. Дауыс жалғандау шабуылдарының негізгі кластары қарастырылған, соның ішінде мәтіннен сөйлеу синтезі, дауыс трансформациясы, қайта ойнатқыш шабуылдары, аудиодипфейк және олардың аралас сценарийлері. Пайдаланылатын деректер жиынтықтары, акустикалық сипаттамалары, модельдердің архитектурасы, оқыту парадигмалары және бағалау метрикалары бойынша салыстырмалы талдау жүргізілді.

Зерттеудің қазіргі жағдайы ASVspoof деректер жиынтығы мен бақылаулы оқыту тәсілдерінің үстемдігімен сипатталады, бұл нәтижелерді салыстыруға мүмкіндік береді, бірақ деректер жиынтығы арасындағы ауысу және домендік ауысу жағдайында модельдердің жалпылау қабілетін шектейді. Негізгі шешілмеген мәселелер әлі де сақталып отыр: бұрын байқалмаған шабуыл түрлеріне

төзімділік, шу мен арна бұрмалануына сезімталдық, сондай-ақ қолданыстағы модельдердің интерпретациялау мүмкіндігінің шектеулілігі.

Талдау негізінде негізгі әдіснамалық шектеулер анықталып, ASV қорғау схемаларына практикалық енгізуге арналған кросс-домендік, берік және түсіндірілетін анықтау жүйелерін дамытуға қатысты әрі қарай зерттеудің басым бағыттары қалыптастырылды. Сонымен қатар мақалада қолданылатын ерекшеліктер түрі мен модельдердің архитектуралық сипаттамаларын ескере отырып, синтетикалық сөйлеуді анықтау әдістерінің жіктемесі ұсынылады.

**Түйін сөздер:** синтетикалық сөйлеу, алдау, ASV, аудио аудиодипфейк, TTS, дауыс трансформациясы, қайта ойнату шабуылы, алдауға қарсы шаралар, өзін-өзі бақылау негізіндегі оқыту, домендер арасындағы беріктік.

## MODERN METHODS FOR DETECTING SYNTHETIC SPEECH AND PROTECTING AUTOMATED SPEAKER VERIFICATION SYSTEMS AGAINST VOICE SPOOFING ATTACKS

**A.Y. Yerzhan✉, A.T. Akhmediyarova, Zh.K. Alimseitova**

*Non-Profit Joint-Stock Company «K.I. Satbayev Kazakh National Research Technical University», Almaty, Kazakhstan*
*e-mail: a.yerzhan@satbayev.university*

The article represents a systematical analytical review of the current methods in the detection of synthesized speech and protection against voice spoofing attacks of automatic speaker verification systems. The key types of voice spoofing attacks are discussed, such as text-to-speech synthesis, voice transformation, playback attacks, audio deepfakes, and their hybrid situations.

There is a comparison of the datasets used, acoustic properties, model design, training regimes, and performance measures. It is known that the current state of research is linked with the commonness of the ASVspoof datasets and supervised learning methods, which assure the comparability of the findings, but limit the applicability of models in the situation of cross-dataset transfer and domain shift. They show that the key unsolved problems are: resistance to previously unknown types of attacks, noise sensitivity and channel distortion sensitivity and poor interpretability of existing models.

Based on the analysis, the main methodological limitations were identified, and priority areas for further research were formulated, related to the development of cross-domain, reliable, and explainable detection systems intended for practical integration into ASV protection schemes. The paper also proposes a classification of synthetic speech detection methods that takes into account the type of features used and the architectural characteristics of the models.

**Keywords:** synthetic speech, spoofing, ASV, audio deepfake, TTS, voice transformation, replay attack, anti-spoofing, self-supervised learning, cross-domain robustness.

**Введение.** Стремительное развитие технологий синтеза и модификации речи существенно изменило ландшафт угроз в системах голосовой аутентификации. Современные нейронные методы генерации аудио позволяют создавать высокореалистичные аудиодипфейки, практически неотличимые от подлинной речи, что делает системы автоматической верификации диктора (ASV) уязвимыми к атакам подмены голоса, включая синтез речи, преобразование голоса и воспроизведение. В условиях широкого применения голосовой биометрии в удалённой аутентификации, включая банковские сервисы и KYC-процедуры, проблема достоверности речевого сигнала приобретает критическое значение, что требует разработки надёжных методов детекции синтетической речи и защиты ASV.

Технологии дипфейка, синтетической и поддельной речи в настоящее время рассматриваются как самостоятельная проблемная область на пересечении автоматической верификации диктора, цифровой аудиофорензики и более широких задач аудиодетекции [1- 3]. Значимость этой

области определяется не только самим развитием средств генерации и подмены речи, но и необходимостью оценивать устойчивость детектирующих систем в условиях спуфинг, дипфейк и состязательных сценариев, всё более приближённых к реальной эксплуатации [4,5]. Особую важность проблема приобретает потому, что атаки, основанные на синтезе и модификации голоса, непосредственно затрагивают защищённость систем голосовой биометрии и достоверность проверки аудиоподлинности в прикладных контекстах [1,2]. Это особенно заметно в таких задачах, как KYC-проверка живости голоса и детекция речевой социальной инженерии, где различение подлинной и поддельной речи имеет прямое практическое значение [6, 7]. В этих условиях разработка надёжных методов детекции синтетической речи становится необходимым условием защиты систем верификации диктора и смежных контуров голосовой аутентичности [1, 5, 7].

Практическая значимость данной проблемы особенно отчётливо проявляется в тех прикладных ситуациях, где голос используется как значимый сигнал для принятия решений и где поддельная речь способна непосредственно подорвать достоверность проверки [1, 7]. Для систем автоматической верификации диктора угроза не ограничивается одним типом воздействия, поскольку в рассматриваемых источниках прямо представлены атаки логического доступа на основе TTS/VC, атака воспроизведения, а также более широкие комбинации имитации, воспроизведения, преобразования голоса и синтеза речи, направленные против голосовой биометрии [1,8,9]. Тот же круг рисков распространяется и на KYC-сценарии с проверкой живости голоса, где различение подлинной и поддельной речи выступает как практический предварительный этап последующей аутентификации [6, 10]. Помимо этого, речевая социальная инженерия, включая вишинг, показывает, что угрозы, связанные с подменой или манипуляцией голосовым контентом, затрагивают и пользовательские коммуникационные контуры [7]. Именно поэтому защита ASV и смежных задач голосовой аутентичности всё в большей степени зависит от надёжных методов противодействия как синтетическим, так и ориентированным на воспроизведение атакам спуфинга [1,9,11].

Целью настоящей статьи является систематизация и сравнительный анализ современных методов детекции синтетической речи и антиспуфинга, а также выявление ключевых ограничений и направлений развития систем защиты автоматической верификации диктора (ASV) в условиях эволюции голосовых атак.

Научная новизна работы заключается в формировании единой аналитической рамки, объединяющей задачи детекции синтетической речи, антиспуфинга и защиты ASV, а также в разработке расширенной классификации методов детекции с учётом типа атак, используемых представлений сигнала, архитектурных решений и уровня обобщающей способности моделей. В отличие от существующих обзорных работ, предложенный подход ориентирован на анализ устойчивости методов в условиях междатасетного переноса, сдвига домена и появления ранее не встречавшихся атак.

В данной работе:

- систематизированы современные датасеты и экспериментальные протоколы оценки в задачах антиспуфинга;

- предложена классификация методов детекции синтетической речи по уровню представления признаков и архитектурной сложности;

- проведён сравнительный анализ подходов с точки зрения устойчивости, переносимости и практической применимости в системах ASV;

- выявлены ключевые методологические ограничения и сформулированы приоритетные направления дальнейших исследований.

Представленные результаты формируют целостное представление о современном состоянии области и её ключевых направлениях развития.

**Материалы и методы.** Данная работа выполнена в формате аналитического обзора и направлена на систематизацию и сравнительный ана-

лиз исследований в области детекции синтетической речи, аудиодипфейков и спуфинг-атак, а также методов защиты систем автоматической верификации диктора от атак подмены голоса [1-3]. В рамках принятого методологического подхода указанные направления рассматриваются как взаимосвязанные компоненты единой задачи обеспечения устойчивости голосовых систем к искусственно сгенерированным и модифицированным речевым воздействиям. Такой подход обеспечивает сопоставимость существующих решений по используемым признакам, архитектурам моделей, метрикам оценки и экспериментальным сценариям при сохранении целостной аналитической рамки.

В настоящем обзоре рассматривается совокупность научных публикаций, отобранных по тематической релевантности задачам детекции синтетической речи и антиспуфинга в системах автоматической верификации диктора. Основной массив работ относится к современному этапу развития области и включает преимущественно исследования последних лет, что позволяет сосредоточить анализ на актуальных архитектурных, датасетных и методических подходах. При этом целью данной работы не является воспроизведение формализованной процедуры систематического обзора; акцент сделан на последовательном описании логики отбора публикаций и их аналитической интерпретации в рамках единой исследовательской рамки.

В развитие обсуждения прикладной значимости голосовых атак следует отметить, что по данной тематике уже существуют обзорные и постановочные публикации, однако они, как правило, охватывают либо атаки синтеза речи на голосовую биометрию в конкретных прикладных контекстах, либо более широкий круг вопросов дипфейк-генерации, ИИ-детекции и цифровой форензики [1,2]. Наряду с этим представлены работы, рассматривающие применение архитектур трансформеров в задачах аудиодетекции, а также публикации, сосредоточенные на вишинге и состязательных или антифорензических угрозах, то есть на смежных или более широких аспектах проблемы [3,4,7]. В отличие от такого более общего или частично специализированного охвата, настоящий обзор сосредоточен именно на методах детекции синтетической речи и защите систем автоматической верификации диктора от атак подмены голоса, включая подходы, ориентированные на обобщённую детекцию, устойчивость, спуфинг-агностичные и одноклассовые постановки, а также на более экономные по разметке схемы обучения [12- 16]. Детальное сравнение с близкими обзорными и смежными работами приведено в Таблице 1.

**Таблица 1 - Сравнение профильных обзорных работ и настоящего обзора**

| Авторы | Год | Основной фокус / вклад | Ограничения | Релевантность для настоящего обзора |
|---|---|---|---|---|
| Кузнецов и др. [1] | 2021 | Обзор атак синтеза речи на системы голосовой биометрии в банковском контексте и практических контрмер для защиты ASV | Работа носит преимущественно концептуально-прикладной характер и не ориентирована на единый экспериментальный протокол сопоставления методов | Близка к теме защиты голосовой биометрии, но сосредоточена на банковском прикладном контексте |
| Bendiab и др. [2] | 2025 | Обзор генерации дипфейков, ИИ-детекции и вызовов цифровой форензики | Имеет широкий обзорный охват; количественная сопоставимость методов в единой экспериментальной рамке выражена ограниченно | Релевантна как контекст детекции дипфейков и цифровой форензики, но не фокусируется специально на антиспуфинге ASV |

| | | | | |
|---|---|---|---|---|
| Zaman и др. [3] | 2025 | Обзор применения архитектур трансформеров в задачах аудиодетекции, включая направление дипфейков | Сравнение подходов основано на неоднородных постановках исходных исследований; отдельный фокус на защите от спуфинга не является центральным | Полезна для архитектурного контекста, но шире темы детекции синтетической речи для ASV |
| Ain Q.U. и др. [4] | 2026 | Комплексный обзор состязательных и генеративных антифорензических атак и мер защиты в аудиовизуальной детекции дипфейков | Работа охватывает широкий мультимодальный контекст, а сравнение результатов зависит от неоднородности первичных исследований | Значима для вопросов надёжности и антифорензики, но шире и мультимодальнее, чем фокус на детекции синтетической речи и защите ASV |
| Andreas Triantafyllopoulos и др. [7] | 2025 | Обзор вишинга как задачи речевой социальной инженерии с акцентом на риски, данные и направления развития методов детекции | Основной акцент сделан на постановке проблемы и прикладном контексте; единый эталонный контур для сопоставления методов не является центральным элементом работы | Важна для прикладного контекста голосовых атак, но охватывает более узкий класс сценариев по сравнению с антиспуфингом для ASV |
| Настоящий обзор | - | Обзор методов детекции синтетической речи и защиты систем распознавания диктора / ASV от атак подмены голоса с систематизацией типов атак, датасетов, признаков, моделей, парадигм обучения, метрик и ограничений | Сфокусирован на задачах детекции синтетической речи и антиспуфинга в контуре распознавания диктора | Непосредственно соответствует заявленной теме и объединяет методологический, оценочный и прикладной аспекты в рамках защиты ASV |

Сопоставление работ, представленных в таблице 1, показывает, что существующие обзорные публикации охватывают главным образом либо отдельные прикладные контексты защиты голосовой биометрии, либо более широкий круг вопросов deepfake-генерации, аудиодетекции, цифровой форензики, vishing и anti-forensic угроз. При этом их фокус, как правило, смещён в сторону смежных или более общих аспектов проблемы, тогда как вопросы детекции синтетической речи и защиты систем автоматической верификации диктора от атак подмены голоса не выступают в них как единый центральный предмет анализа. В этом отношении рассматриваемый обзор занимает более целевую позицию, поскольку объединяет именно synthetic speech detection, anti-spoofing и защиту ASV в одной аналитической рамке, что делает его логически обоснованным и тематически специализированным по сравнению с близкими обзорными работами.

Критерием включения публикаций в обзор служила их прямая содержательная связь с задачами детекции синтетической речи, спуфинг-атак, антиспуфинга, детекции атак воспроизведения, преобразования голоса, синтеза речи, проверкой живости голоса и защитой систем автоматической верификации диктора от атак подмены голоса. Приоритет отдавался работам, в которых явно рассматривались методы обнаружения искусственно сгенерированной или подменённой речи, уязвимости голосовых систем либо подходы к противодействию таким воздействиям. Напротив, исследования, сосредоточенные исключительно на генерации речи, общем развитии ASR или смежных речевых технологиях без выраженной связи с безопасностью, детекцией или защитой от атак, не рассматривались как центральные для настоящего обзора и учитывались лишь постольку, поскольку они были необходимы для контекстуализации основной проблематики.

Отобранные публикации анализировались по единой системе параметров, обеспечивающей их содержательную сопоставимость в пределах общей исследовательской рамки. В центре внимания находились исследовательский фокус каждой работы, основная идея предлагаемого подхода, применяемые методы и алгоритмические решения, типы рассматриваемых атак, используемые признаки и представления аудиосигнала, характеристики задействованных датасетов, а также практический контекст применения. Дополнительно учитывались заявленные авторами преимущества, ограничения, элементы научной новизны и направления дальнейшего развития. Такая схема анализа позволила не только описать отдельные результаты, но и выявить повторяющиеся методологические закономерности, различия в экспериментальных постановках и устойчивые исследовательские акценты.

На завершающем этапе аналитической обработки публикации были сопоставлены и тематически сгруппированы по ключевым направлениям, включая детекцию синтетической речи, антиспуфинг для автоматической верификации диктора, детекцию атак воспроизведения, методы, связанные с проверкой живости, и смежные прикладные аспекты голосовой безопасности. Такая тематическая организация материала позволила перейти от описания отдельных работ к выявлению более устойчивых исследовательских конфигураций, проследить, какие подходы оказываются наиболее продуктивными в различных постановках, и определить те ограничения, которые сохраняются при переходе от лабораторных эталонных сценариев к более реалистичным условиям применения. Именно в этой аналитической рамке стало возможным выделить основные тенденции развития области, сильные стороны существующих решений и ключевые исследовательские пробелы, сохраняющие значение для дальнейшего развития методов детекции синтетической речи и защиты систем распознавания диктора от атак подмены голоса.

**Результаты и обсуждение.** Датасеты, используемые для детекции синтетической речи и антиспуфинга в системах распознавания диктора. В исследованиях по детекции синтетической речи и антиспуфингу выбор датасета выступает не вспомогательным, а определяющим методологическим фактором, поскольку именно совокупность используемых корпусов задаёт типы атак, структуру речевого материала, степень контролируемости условий записи и допустимые режимы сопоставления моделей. В этой области эталонные корпуса непосредственно влияют на корректность сравнения алгоритмов, воспроизводимость результатов, оценку устойчивости и выводы об обобщающей способности систем, поскольку различия между корпусами затрагивают как характер сценариев спуфинга, так и соотношение между лабораторными и более реалистичными условиями использования [3,5]. Для задач защиты систем автоматической верификации диктора это особенно важно, так как датасет определяет, идет ли речь об атаках синтеза речи по тексту, преобразования голоса, атаках воспроизведения, задачах проверки живости голоса или о более широком контексте аудиодипфейков, а следовательно, формирует и критерии экспериментальной валидности [8,9,10,17].

Используемые в данной области наборы данных целесообразно классифицировать прежде всего по условиям экспериментальной оценки, поскольку именно режим проверки определяет, какие свойства модели оцениваются и насколько корректно интерпретируются её результаты. В одной группе находятся контролируемые эталонные условия, ориентированные на строгое сопоставление подлинной и поддельной речи в рамках заранее заданных протоколов, что типично для постановок логического доступа и физического доступа, включая сценарии TTS/VC и воспроизведения [8,9,10,12,14,15,17,18,19].

Другая группа связана с задачами проверки живости голоса и KYC-ориентированной оценки, где на первый план выходит способность модели различать подлинную и поддельную речь в предварительном контуре голосовой аутентификации [6, 10]. Наконец, отдельный класс образуют кросс-доменные, мультимодальные и соревновательные условия оценки, в которых осо-

бенно важны переносимость, устойчивость к шуму и каналу, чувствительность к сдвигу домена, а также поведение модели за пределами строго контролируемой лабораторной среды [5,9,13,20,21,22,23]. Такая классификация позволяет рассматривать датасеты не только по типу атаки, но и по тому, какие аспекты обобщающей способности, устойчивости и внешней валидности проверяются в соответствующем исследовании.

Наиболее устойчивое положение в рассматриваемом поле занимает семейство ASVspoof, которое фактически выполняет функцию основной эталонной платформы для детекции синтетической речи и защиты ASV [5,8,9,12,13,14,24,25]. В его рамках ASVspoof 2019 LA и связанные сценарии логического доступа особенно часто используются для оценки методов, ориентированных на выявление TTS/VC-подделок, одноклассовую детекцию и спуфинг-агностичное моделирование, что объясняет их высокую популярность как в классических, так и в более современных архитектурных работах [8,12,14,15,18,19].

Для исследований, ориентированных на воспроизведение, сопоставимую роль играют ASVspoof 2017, ASVspoof 2019 PA и ASVspoof 2021 PA, поскольку они позволяют исследовать влияние переизлучения, канальных условий и постановки физического доступа на качество детекции [9,10,11,17]. Существенное значение приобрёл и ASVspoof 5, представленный как более реалистичный краудсорсинговый корпус с семью непересекающимися по дикторам разделами и 32 атакующими алгоритмами, что делает его важным ориентиром для современных исследований спуфинга, дипфейков и детекции состязательных воздействий [5]. Именно семейство ASVspoof в наибольшей степени задаёт общую экспериментальную рамку области, обеспечивая сопоставимость результатов между работами и формируя доминирующую логику современных исследований антиспуфинга.

Наряду с ASVspoof в рассматриваемых работах используются и другие публичные датасеты, расширяющие исследовательское поле и позволяющие проверять модели в дополнительных условиях. К числу речевых ресурсов относятся прежде всего LibriSeVoc, применяемый для анализа артефактов нейронных вокодеров [18], а также In-The-Wild, используемый в кросс-доменных проверках на менее контролируемом речевом материале [3]. В постановках, ориентированных на воспроизведение и проверку живости, важную роль играют POCO как ресурс для проверки живости голоса [6,10] и ReplayDF как более реалистичная база для валидации атак воспроизведения [9].

В более широком исследовательском контексте привлекаются также соревновательные наборы данных ADD 2023 для шумовых и мультиатаковых условий [21] и мультимодальные ресурсы, такие как FakeAVCeleb [22]. В совокупности эти датасеты играют дополняющую роль: они ослабляют зависимость анализа от одной эталонной среды и позволяют более содержательно оценивать переносимость и устойчивость методов за пределами стандартных протоколов ASVspoof.

С точки зрения характеристик датасеты различаются по объёму, разнообразию говорящих, типам подделки и степени приближения к реальным условиям. Наиболее явно эти параметры артикулированы в ASVspoof 5, где подчёркивается масштаб корпуса, широкое разнообразие дикторов и атакующих алгоритмов, что делает его одним из немногих ресурсов, специально спроектированных как более реалистичный эталонный набор для новой волны исследований спуфинга и дипфейков [5]. В противоположность этому ряд широко используемых наборов данных сохраняет более контролируемую экспериментальную структуру, в которой атаки и условия записи заранее типизированы и потому удобны для строгого сопоставления методов, как это характерно для ASVspoof LA/PA и ASVspoof 2017 [8,9,12,24]. Вместе с тем часть исследований целенаправленно обращается к менее контролируемым или более прикладным ресурсам, чтобы приблизить проверку к реальным условиям и оценить устойчивость к шуму, каналу и междатасетному переносу [3,6,9,21].

Основные ограничения существующих дата-

сетов связаны с тем, что значительная часть эталонных корпусов формируется в контролируемых условиях и потому не всегда отражает сложность реального речевого взаимодействия. Это проявляется в зависимости результатов от конкретного корпуса, в чувствительности к сдвигу домена и в ухудшении качества при переходе к ранее не встречавшимся генераторам, новым каналам, шумовым условиям и иным распределениям подлинной и поддельной речи [9,18,19,26].

В работах, ориентированных на воспроизведение, дополнительно подчёркивается, что оценка на одном эталонном сценарии ограничивает выводы о переносимости на более новые и разнообразные условия, из-за чего возникает необходимость расширения междатасетной оценки [9,17]. Даже там, где используются более реалистичные ресурсы, сохраняется потребность в дальнейшем расширении языков, каналов, типов атак и протоколов оценки, что особенно явно сформулировано для ASVspoof 5 и более сложных кросс-доменных дипфейк-сценариев [5]. В совокупности это означает, что модели нередко рискуют переобучаться на корпус-специфические признаки, а не на действительно устойчивые признаки подмены голоса, что снижает надёжность их обобщения за пределами исходного набора данных [9,19,26].

Поэтому развитие более репрезентативных, разнообразных и приближённых к реальным условиям датасетов остаётся одной из центральных задач области. Необходимость таких ресурсов обусловлена не только требованиями к большей внешней валидности эталонных корпусов, но и тем, что именно датасетная база определяет, какие признаки оказываются информативными, какие архитектуры демонстрируют устойчивость и какие стратегии детекции сохраняют эффективность при междатасетном переносе и сдвиге домена [5,9,19,26]. В этом смысле анализ методов детекции синтетической речи и защиты ASV не может рассматриваться изолированно от используемых наборов данных, поскольку различия между ними непосредственно формируют исследовательский ландшафт, в котором оцениваются преимущества и ограничения существующих решений. Именно поэтому дальнейшее рассмотрение методов детекции синтетической речи и защиты систем распознавания диктора целесообразно вести в тесной связи с той датасетной базой, на которой эти методы проектируются, обучаются и проверяются.

***Методы детекции синтетической речи и защиты ASV от атак подмены голоса.*** Методы детекции синтетической речи и защиты систем автоматической верификации диктора существенно различаются по используемым признаковым представлениям, архитектурным решениям и уровню обобщающей способности. Прежде всего, различия определяются типом признаков, на которых основаны модели: от явно заданных акустических характеристик до автоматически извлекаемых глубинных репрезентаций. Кроме того, подходы варьируются по уровню анализа сигнала от локальных спектральных и просодических признаков до моделирования временных, межканальных и контекстных зависимостей. Существенным фактором также является степень устойчивости методов к новым сценариям спуфинга и их практическая применимость в контуре защиты ASV, что особенно проявляется при переходе от классических схем к современным подходам обучения представлений и архитектурам, ориентированным на безопасность [1-3].

К числу базовых и сохраняющих значимость подходов относятся методы, основанные на использовании ручных и полуавтоматически извлекаемых признаков. В рамках данного подхода синтетическая или подменённая речь интерпретируется как источник аномалий в различных акустических представлениях сигнала, включая кепстральные, спектральные, фазовые и просодические характеристики. Именно поэтому в ряде работ используются мультикепстральные представления, дельта-признаки, графовые спектральные преобразования, признаки речевой акустики, а также специально сконструированные наборы признаков, ориентированные на выявление артефактов атак воспроизведения, TTS- и VC-атак [9,11,17,24,27,28]. Существенное достоинство таких методов состо-

ит в их относительной интерпретируемости, поскольку исследователь может соотносить решение модели с конкретными свойствами сигнала, такими как спектральные искажения, нестабильность фазовой структуры, атипичная просодия или особенности канала [1,29]. Однако зависимость от заранее выбранного признакового пространства ограничивает их адаптивность в условиях быстро развивающихся дипфейк- и спуфинг-атак, когда новые генеративные механизмы начинают производить подделки, менее уязвимые для традиционных акустических индикаторов [9,17,28].

В современных исследованиях доминирующее положение в задачах детекции синтетической речи и антиспуфинга занимают методы глубокого обучения, обеспечившие переход от фиксированных признаковых представлений к автоматическому извлечению дискриминативных репрезентаций непосредственно из аудиосигнала. Наибольшее распространение получили сверточные и остаточные архитектуры, модели на основе механизмов внимания, а также гибридные подходы, интегрирующие спектральную обработку с моделированием временной динамики сигнала.

В ряде работ используются решения, подобные ResNeXt, остаточные сверточные блоки с механизмами внимания, адаптивные схемы возмущений, модели с многомерным взаимодействием признаков и архитектуры, ориентированные на усиление слабых артефактов подделки [21,25,26,30,31,32]. Их преимущество по сравнению с классическими подходами, основанными на признаках, связано не только с более высокой точностью на эталонных корпусах, но и со способностью совместно учитывать несколько уровней представления сигнала, извлекая признаки, которые трудно задать вручную. Вместе с тем рост архитектурной сложности сам по себе не устраняет ограничений, поскольку переносимость таких моделей по-прежнему существенно зависит от датасета, протокола обучения и характера атак [21,25,30].

Современный этап развития методов детекции синтетической речи характеризуется переходом к архитектурам на основе трансформеров и подходам самоконтролируемого обучения, ориентированным на формирование универсальных и устойчивых представлений аудиосигнала. Эти методы направлены на снижение зависимости от корпусно-специфических шаблонов и повышение обобщающей способности моделей. Трансформерные архитектуры демонстрируют эффективность в моделировании долгосрочных зависимостей и интеграции разнородных признаков, что подтверждается как специализированными решениями, так и обзорными исследованиями в области аудиодетекции [3,8]. Параллельно усиливается интерес к самоконтролируемому обучению и связанным с ним схемам обучения представлений, включая методы с неконтрастивным обучением, метаобучением и раздельным обучением, которые направлены не только на повышение точности, но и на улучшение обобщающей способности, устойчивости к ранее не встречавшимся атакам и междатасетному переносу [14,16]. Однако эти подходы не следует трактовать как универсальное решение: в существующих работах они скорее обозначают важное направление эволюции поля, чем полностью преодолённый набор проблем, поскольку их эффективность по-прежнему зависит от характера данных, протокола обучения и сценария оценки [3,8,14,16].

Отдельное направление составляют методы, ориентированные на защиту систем автоматической верификации диктора. В данном контексте антиспуфинг рассматривается не как изолированная задача бинарной классификации, а как компонент комплексной архитектуры защиты, в которой существенное значение имеет не только точность различения подлинной и поддельной речи, но и влияние принимаемых решений на функционирование самой системы верификации.

В существующих исследованиях представлены как автономные схемы детекции, реализуемые в виде внешнего фильтра перед ASV, так и интегрированные подходы, учитывающие требования голосовой биометрии, задачи проверки живости и противодействие атакам воспроизве-

дения [1,9,10,24]. Такое различие принципиально важно, поскольку прикладная задача состоит не просто в обнаружении подделки как таковой, а в снижении уязвимости системы аутентификации в реальных эксплуатационных сценариях, включая банковскую биометрию и KYC-контуры [1,6,10]. Поэтому в исследованиях защиты ASV всё большее значение приобретают не только показатели детекции, но и то, насколько модуль антиспуфинга согласован с логикой функционирования всей системы и её требованиями к устойчивости и безопасности [9,14,20].

В современных исследованиях ключевое значение приобретают обобщённые, устойчивые и спуфинг-агностичные подходы, включая одноклассовые и экономные по разметке методы, ориентированные на условия быстро эволюционирующего ландшафта атак. Эти методы направлены на снижение зависимости моделей от фиксированного набора известных подделок и повышение их обобщающей способности. Одноклассовые и обобщённые схемы детекции позволяют формулировать задачу в более реалистичной постановке, предполагающей отсутствие априорной информации о типах атак [12,13,15].

Робастные и унифицированные подходы, включая методы с состязательным обучением, обратными возмущениями и иными механизмами повышения устойчивости, направлены на то, чтобы модель сохраняла качество при усложнении акустических условий и смене типа спуфинг-воздействия [9,14,21,31]. Наконец, экономные по разметке и основанные на представлениях решения свидетельствуют о стремлении преодолеть зависимость детекции от чрезмерно подробной разметки и узких обучающих сценариев, что особенно важно в условиях расширяющегося спектра дипфейк- и технологий синтетической речи [16]. В совокупности эти направления показывают, что современная область постепенно смещается от распознавания отдельных известных атак к построению более общих механизмов противодействия подделке голоса.

В целом развитие методов детекции синтетической речи и защиты ASV демонстрирует переход от более узких, корпусно-зависимых и признаково-ориентированных решений к архитектурно более сложным, устойчивым и практически значимым подходам. Этот переход сопровождается расширением исследовательского внимания к вопросам обобщающей способности, междатасетного переноса и интеграции механизмов антиспуфинга в реальные контуры голосовой аутентификации. Вместе с тем накопленные результаты не устраняют фундаментальных ограничений, связанных со смещением датасета, сдвигом домена, интерпретируемостью и реальным развёртыванием моделей за пределами лабораторных эталонных сценариев [9,4,16,20]. Именно эти сохраняющиеся противоречия и подводят к рассмотрению ключевых ограничений существующих решений и тех исследовательских пробелов, которые продолжают определять развитие данной области.

На основе проведённого анализа предлагается концептуальная архитектура устойчивой системы детекции синтетической речи и защиты систем автоматической верификации диктора от атак подмены голоса (рисунок 1).

В отличие от изолированного рассмотрения отдельных методов, в настоящей работе предлагается концептуальная архитектура устойчивой системы детекции синтетической речи, ориентированная на интеграцию различных уровней анализа аудиосигнала и обеспечение защиты ASV в условиях разнообразных атак.

Предлагаемая архитектура включает несколько функциональных уровней.

На первом уровне осуществляется предварительная обработка сигнала, включающая нормализацию и подавление шума. На втором уровне реализуется модуль извлечения признаков, обеспечивающий формирование спектральных, кепстральных и временных представлений сигнала. На следующем уровне используется ансамбль моделей детекции, включающий сверточные, трансформерные и модели обучения представлений. Далее реализуется модуль обобщённой детекции, ориентированный на выявление ранее не наблюдавшихся типов атак. На заключительном этапе осуществляется интеграция с систе-

мой автоматической верификации диктора, где результаты детекции используются для принятия решения.

Предложенная архитектура отражает переход от изолированных методов детекции к комплексным системам, ориентированным на устойчивость, обобщающую способность и практическую применимость.

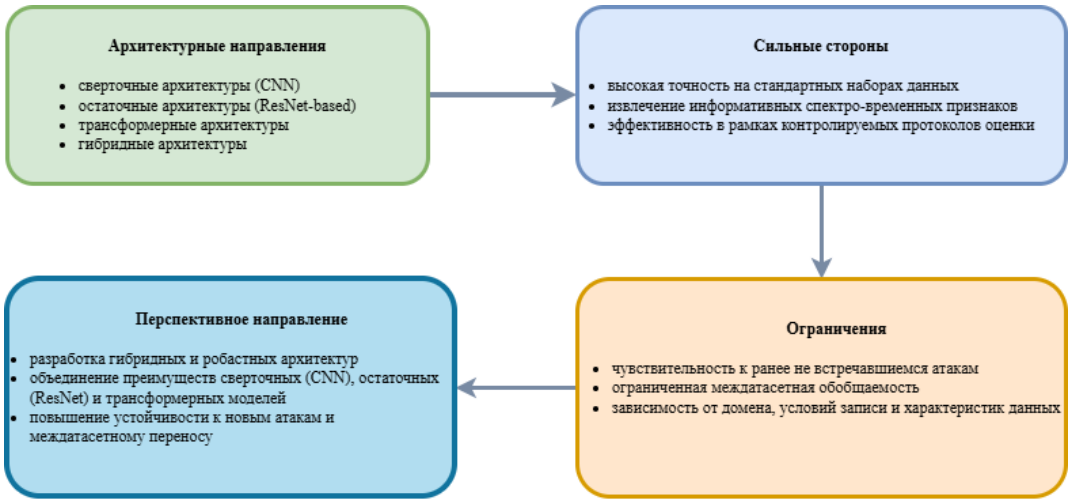



**Рис.1 - Концептуальная архитектура устойчивой системы детекции синтетической речи и защиты ASV от атак подмены голоса**

**Таблица 2 - Сравнительный анализ методов детекции синтетической речи**

| Класс методов | Представление признаков | Устойчивость к новым атакам | Зависимость от датасета | Интерпретируемость | Применимость в ASV |
|---|---|---|---|---|---|
| Ручные признаки | Явно заданные (MFCC, CQCC) | Низкая | Высокая | Высокая | Ограниченная |
| CNN/ResNet | Автоматически извлекаемые | Средняя | Высокая | Средняя | Высокая |
| Transformer | Контекстные представления | Средневысокая | Средняя | Низкая | Перспективная |
| Self-supervised | Универсальные эмбеддинги | Высокая | Низкая | Низкая | Очень высокая |
| One-class / spoof-agnostic | Аномалии | Высокая | Средняя | Средняя | Ключевая |

Обобщая рассмотренные подходы, следует отметить, что в задачах детекции синтетической речи и антиспуфинга наибольшее распространение получили архитектуры глубокого обучения, включая сверточные, остаточные, трансформерные и гибридные модели. Их широкое применение обусловлено высокой способностью извлекать информативные признаки и обеспечивать убедительные результаты в рамках стандартных экспериментальных протоколов. Вме-

сте с тем проведённый анализ показывает, что использование только одного класса архитектур не всегда обеспечивает достаточную устойчивость к новым типам атак, междатасетному переносу и условиям реальной эксплуатации. В связи с этим в качестве перспективного направления представляется комплексный подход, ориентированный на объединение сильных сторон различных архитектурных решений с целью повышения устойчивости, обобщающей способности и практической применимости системы (таблица 2).

***Классификация методов детекции синтетической речи.***

На основе проведённого анализа современные методы детекции синтетической речи и антиспуфинга целесообразно классифицировать по уровню представления признаков и архитектурной организации моделей.

В рамках данной классификации можно выделить следующие основные классы:

**1. Методы на основе ручных и инженерных признаков.** Данный класс включает подходы, использующие заранее заданные акустические характеристики, такие как MFCC, CQCC, спектральные и фазовые признаки, а также их производные. Эти методы характеризуются высокой интерпретируемостью, однако обладают ограниченной устойчивостью к новым типам атак.

**2. Глубокие сверточные и гибридные модели.** К этому классу относятся архитектуры CNN, ResNet, ResNeXt и их модификации, способные автоматически извлекать дискриминативные признаки из аудиосигнала. Они демонстрируют высокую точность на эталонных датасетах, но остаются чувствительными к сдвигу домена.

**3. Трансформерные и attention-ориентированные модели.** Методы данного класса используют механизмы самовнимания для моделирования долгосрочных зависимостей в сигнале. Их преимуществом является более гибкое представление аудиоданных, однако они требуют значительных объёмов данных и вычислительных ресурсов.

**4. Методы обучения представлений (self-supervised и label-efficient).** Данный класс включает подходы, ориентированные на извлечение универсальных представлений аудиосигнала с минимальной зависимостью от разметки. Они являются перспективными с точки зрения обобщающей способности и устойчивости к новым атакам.

**5. Обобщённые и спуфинг-агностичные методы.** Эти методы направлены на выявление аномалий в аудиосигнале без жёсткой привязки к конкретным типам атак. Они наиболее соответствуют реальным условиям эксплуатации, однако пока не обеспечивают стабильного превосходства во всех сценариях.

Предложенная классификация позволяет перейти от описания отдельных моделей к более системному пониманию структуры исследовательского поля и выявлению направлений, обладающих наибольшим потенциалом для развития устойчивых систем защиты ASV.

**Ограничения и исследовательские пробелы.** Несмотря на значительный прогресс в области детекции синтетической речи и антиспуфинга, современное состояние исследований показывает, что высокая точность, достигаемая на отдельных наборах данных, не обеспечивает эквивалентного уровня практической надёжности методов. Накопленные результаты свидетельствуют о существенном развитии как признаково-ориентированных подходов, так и методов глубокого обучения и обучения представлений. Однако их эффективность остаётся ограниченной вследствие неоднородности используемых датасетов, вариативности сценариев атак и зависимости от условий функционирования систем голосовой биометрии [1-3]. В результате одна и та же архитектурная концепция может демонстрировать высокие показатели в рамках конкретного экспериментального протокола, но существенно терять устойчивость при переносе на иные акустические, эксплуатационные и ориентированные на безопасность условия [14,20].

Одной из ключевых проблем остаются смещение датасета, сдвиг домена и ограниченная обоб-

щающая способность моделей, тесно связанные с неполной сопоставимостью результатов между работами. Значительная часть современных решений оценивается в пределах эталонных сценариев ASVspoof и близких к ним корпусов, что обеспечивает удобную экспериментальную воспроизводимость, но одновременно усиливает зависимость моделей от корпусно-специфических признаков и особенностей конкретного протокола [9,20]. В этих условиях высокая результативность может отражать не столько устойчивое распознавание признаков подделки, сколько успешную адаптацию к структуре конкретного корпуса, типу записи и заданному набору условий спуфинга [13,14]. При переносе на новые каналы записи, шумовые режимы, иные языковые и доменные условия, а также на менее контролируемые данные устойчивость таких моделей часто снижается, что непосредственно ограничивает их внешнюю валидность и практическую ценность [5,9,20,23].

Не менее существенное ограничение связано с быстрым развитием самих технологий подделки голоса. Прогресс в области синтеза речи по тексту, преобразования голоса, нейронных вокодеров и более широких схем генерации аудиодипфейков приводит к тому, что многие корпусно-специфические решения устаревают быстрее, чем успевают стать по-настоящему универсальными [2,4,18]. По мере усложнения генеративных моделей артефакты, на которых основывались ранние и промежуточные детекторы, становятся менее выраженными, а граница между подлинной и поддельной речью оказывается всё более трудноуловимой для методов, ориентированных на фиксированный набор известных атак [5, 8]. Именно поэтому в современной литературе усиливается интерес к обобщённым, спуфинг-агностичным и устойчивым подходам, стремящимся выявлять не частный шаблон конкретного спуфинга, а более общие признаки искусственной модификации речи и аномального распределения сигнала [12,13,14,15,16,21]. Однако и эти направления пока не устраняют проблему полностью, а лишь обозначают наиболее перспективный путь преодоления зависимости от ранее не встречавшихся атак.

Отдельный круг ограничений возникает при переходе от автономной детекции спуфинга к реальной защите систем автоматической верификации диктора. В прикладных системах качество антиспуфингового модуля определяется не только способностью различать подлинную и поддельную речь, но и тем, насколько такое решение согласуется с логикой контура ASV и требованиями реального контура аутентификации [1,9,10]. Для банковской биометрии, KYC-сценариев и голосовой аутентификации в целом это особенно важно, поскольку избыточно жёсткая детекция способна ухудшать практическую применимость системы, тогда как недостаточно строгая сохранять уязвимость к атакам воспроизведения, синтетическим и комбинированным атакам [1,6]. Тем самым проблема защиты ASV оказывается шире простой бинарной классификации: она требует оценки безопасности в связке с ASV, учитывающей совместную работу детекции, аутентификации и механизмов принятия решения в условиях реальной эксплуатации [1,5,9].

Сохраняются и существенные методологические проблемы, связанные с интерпретируемостью, объяснимостью, дефицитом разметки, ограниченным разнообразием данных и воспроизводимостью результатов. По мере усложнения моделей растёт разрыв между высокой эмпирической эффективностью и возможностью надёжно интерпретировать, какие именно свойства сигнала лежат в основе решения и насколько эти свойства устойчивы вне обучающего распределения [2,3,33]. Это особенно заметно в глубоком обучении и гибридных архитектурах, где увеличение ёмкости модели не всегда сопровождается ростом прозрачности. Одновременно ограниченность разметки стимулирует развитие экономных по разметке и связанных с ними слабо контролируемых направлений, однако такие решения пока лишь частично компенсируют нехватку репрезентативных обучающих данных [16]. В совокупности это означает, что многие современные методы становятся технически бо-

лее мощными, но не всегда более объяснимыми, воспроизводимыми и надёжными в прикладном смысле [4,33].

На этом фоне основные исследовательские пробелы просматриваются достаточно отчётливо. Область нуждается в более реалистичных, разнообразных и, по возможности, многоязычных эталонных ресурсах, способных поддерживать не только внутридатасетную, но и полноценную междатасетную оценку в условиях новых каналов, языков и сценариев атак [5]. Не менее важна разработка более содержательной оценки безопасности в связке с ASV, в которой антиспуфинговые методы оценивались бы не изолированно, а в контуре реальной голосовой верификации с учётом эксплуатационных ограничений и комбинированных угроз [1,5,9]. Существенным направлением остаются устойчивые одноклассовые, спуфинг-агностичные и экономные по разметке методы, ориентированные на снижение зависимости от узких эталонных протоколов и от закрытого перечня известных атак [12,15,16]. Наконец, сохраняется потребность в практико-ориентированных решениях, которые были бы устойчивы к новым типам подделки, к сдвигу домена и к реальным условиям развёртывания, а не только к лабораторно контролируемым сценариям [9,14,20].

Таким образом, дальнейшее развитие области определяется не только усложнением моделей и расширением архитектурного инструментария, но и необходимостью улучшать их обобщающую способность, устойчивость, воспроизводимость и прикладную ценность. Современный исследовательский ландшафт показывает, что прогресс в детекции синтетической речи и защите ASV должен оцениваться не по изолированным эталонным результатам, а по степени готовности методов к работе в изменчивых и эксплуатационно сложных условиях [9,14,16,20]. Именно эта двойственная ситуация - заметный технологический прогресс при сохранении фундаментальных ограничений - и определяет ключевые направления дальнейших исследований в данной области.

**Выводы.** Проведённый анализ показал, что детекция синтетической речи и защита систем автоматической верификации диктора сформировались как комплексная междисциплинарная задача, в которой эффективность решений определяется не только архитектурой модели, но и условиями обучения, характеристиками датасетов и сценариями атак. Установлено, что современные методы демонстрируют высокую результативность в рамках эталонных протоколов, однако их устойчивость существенно снижается при междатасетном переносе, изменении акустических условий и появлении новых типов спуфинг-атак.

Систематизация существующих подходов позволила выделить ключевую тенденцию развития области, заключающуюся в переходе от признаково-ориентированных и корпусно-зависимых методов к архитектурам, основанным на обучении представлений, самоконтролируемом обучении и спуфинг-агностичных стратегиях детекции. Показано, что наибольшим потенциалом обладают методы, ориентированные на обобщающую способность и устойчивость к ранее не наблюдавшимся атакам, однако их практическая эффективность пока ограничена зависимостью от данных и сложностью интеграции в реальные системы ASV.

В ходе исследования выявлено, что основной методологической проблемой остаётся разрыв между лабораторной оценкой и условиями реальной эксплуатации. Доминирование датасетов семейства ASVspoof обеспечивает воспроизводимость экспериментов, но одновременно усиливает риск переобучения моделей на корпусно-специфические признаки. Это приводит к снижению надёжности детекции в прикладных сценариях, включая голосовую биометрию, KYC-проверки и системы голосовой аутентификации.

На основе проведённого анализа обоснована необходимость перехода к более комплексным архитектурным решениям, интегрирующим различные уровни представления сигнала, включая спектральные, временные и контекстные характеристики, а также использующим гибридные подходы, сочетающие глубокое обучение и мето-

ды обучения представлений. Показано, что перспективным направлением является разработка устойчивых, спуфинг-агностичных систем, способных выявлять не конкретные типы атак, а общие признаки искусственной модификации речевого сигнала.

Таким образом, дальнейшее развитие области должно быть направлено на создание более репрезентативных и многоязычных датасетов, расширение междатасетной оценки, разработку методов с высокой обобщающей способностью, а также интеграцию механизмов антиспуфинга в единый контур защиты ASV. Полученные результаты формируют основу для разработки более надёжных и practically-oriented систем детекции синтетической речи, способных эффективно функционировать в условиях постоянно эволюционирующего ландшафта голосовых атак.

**Литература**

1. Кузнецов А.Ю., Муртазинт Р.А., Гарипов И.М., Фёдоров Е.А., Холоденина А.В., Воробьева А.А. Методы противодействия атакам посредством синтеза речи на голосовые биометрические системы в банковской сфере. //Научно-технический Вестник. -2021.-Т.21. № 1.- С.109-117. DOI 10.17586/2226-1494-2021-21-1-109-117.

2. Gueltoum Bendiab, Houda Haiouni, Isidoros Moulas, Stavros Shiaeles. Deepfakes in digital media forensics: Generation, AI-based detection and challenges. //Information Security and Applications. -2025.-Vol.88 103935. DOI 10.1016/j.jisa.2024.103935.

3. Khalid Zaman, Kai Li, Melike Sah, Cem Direkoglu, et.al. Transformers and audio detection tasks: An overview. //Digital Signal Processing. -2025.-Vol.158:104956. DOI 10.1016/j.dsp.2024.104956.

4. Qurat Ul Ain, Fatima Khalid, Hafsa Ilyas, Ali Javed, et.al. Adversarial and generative AI-based anti-forensics in audio-visual deepfake detection: A comprehensive review and analysis. //Information Fusion. -2026.-Vol.132: 104120. DOI 10.1016/j.inffus.2025.104120.

5. Xin Wang, Héctor Delgado, Hemlata Tak, Jee-weon Jung, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, Nicholas Evans, et.al. ASVspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech. //Computer Speech & Language. -2026.-Vol.95:101825. DOI 10.1016/j.csl.2025.101825.

6. Buğra Eyidoğan, Gökberk Özsoy, Pedram Khatamino,et.al. Voice Liveness Detection KYC Project: Distinguishing Genuine and Spoofed Voices Using Deep Learning. //Procedia Computer Science. -2025.-Vol.255.-P.63-72. DOI 10.1016/j.procs.2025.02.261.

7. Andreas Triantafyllopoulos, Anika A. Spiesberger, Iosif Tsangko, Xin Jing, et.al.Vishing: Detecting social engineering in spoken communication - A first survey & urgent roadmap to address an emerging societal challenge. //Computer Speech & Language. -2025.-Vol.94: 101802. DOI 10.1016/j.csl.2025.101802.

8. Awais Khan, Khalid Mahmood Malik. SpoTNet: A spoofing-aware Transformer Network for Effective Synthetic Speech Detection.2nd ACM International Workshop on Language. - AI against Disinformation. //History. -2023.-P.10-18. DOI 10.1145/3592572.3592841.

9. Şule Bekiryazıcı, Cemal Hanilçi, Neyir Ozcan. Toward robust replay attack detection in Automatic Speaker Verification: A study of spectrum estimation and channel magnitude response modeling. //Computer Speech & Language. -2026.-Vol.98:101906.DOI 10.1016/j.csl.2025.101906.

10. Arth J. Shah, Madhu R. Kamble. Leveraging Kolmogorov-Arnold networks for voice liveness detection in anti-spoofing systems. //Speech Communication.-2026.-Vol.178:103364. DOI 10.1016/j.specom.2026.103364.

11. Tingting Wang,Marvin Borsdorf, Qoquan Zhang, et.al. Going deeper with log-graph Fourier transform-based feature extraction for playback speech detection. //Applied Acoustics. -2026.-Vol.245:111212. DOI 10.1016/j.apacoust.2025.111212.

12. Jiahong Ye, Diqun Yan, Songyin Fu, Bin Ma, Zhihua Xia. One-class network leveraging spectro-temporal features for generalized synthetic speech detection. //Speech Communication. -2025.-Vol.169:103200. DOI 10.1016/j.specom.2025.103200.

13. Jingze Lu, Yuxiang Zhang, Zhuo Li,et.al. Leveraging distance information for generalized spoofing speech detection. //Computer Speech & Language. -2025.-Vol.94: 101804. DOI 10.1016/j.csl.2025.101804.

14. Zhenyu Wang, John H. L. Hansen. Toward Improving Synthetic Audio Spoofing Detection Robustness via Meta-Learning and Disentangled Training with Adversarial Examples. // IEEE Access. -2024.-Vol.12.-P. (99):1-1. DOI 10.1109/ACCESS.2024.3421281.

15. Jianqiang Zhang, Yushui Geng, Peng Zhang,et.al. One-class neural network with hybrid pooling on dual-band frequency for spoofing speech detection. //Computer Speech & Language. -2026.-Vol.99: 101937. DOI 10.1016/j.csl.2026.101937.

16. Falih Gozi Febrinanto, Kristen Moore, Chandra Thapa,et.al. SIGNL: A label-efficient audio deepfake detection system via spectral-temporal graph non-contrastive learning. //Expert Systems with Applications. -2026.-Vol.307:131056. DOI 10.1016/j.eswa.2025.131056.

17. Rodrigo Colnago Contreras, Monique Simplicio Viana, Everthon Silva Fonseca, et.al.Exploring multicepstral features in a new classical machine learning-based framework for replay attack detection. //Computers and Electrical Engineering. -2025.-Vol.127(PB). DOI 10.1016/j.compeleceng.2025.110570.

18. Jingxi Xue, Sanshuai Cui, Weinan Zhang. Leveraging Neural Vocoder Artifacts for Improved Synthetic Speech Detection.2024 7th International Conference on Sensors, Signal and Image. //Processing (SSIP). -2024.-P.162-166. DOI 10.1145/3725949.3725971.

19. Burak Celik, Ayhan Akbal. A multi-level and iterative feature engineering framework for deepfake audio detection. //Expert Systems with Applications. -2026.-Vol.303:130722. DOI 10.1016/j.eswa.2025.130722.

20. Soyul Han, Jaejin Seo, Sunmook Choi, Taein Kang,et.al. Enhancing voice spoofing detection in noisy environments using frequency feature masking augmentation. //Engineering Science and Technology, an International Journal. -2025.-Vol.63: 101972. DOI 10.1016/j.jestch.2025.101972.

21. Hao Meng, Wei Ou, Ju Huang, Haozhe Liang,et.al. A robust unified spoofing audio detection scheme. Computers and Electrical Engineerin g.-2025.-Vol.122: 109974. DOI 10.1016/j.compeleceng.2024.109974.

22. Yuanfan Jin, Yongfang Wang. CMALDD-PTAF: Cross-modal adversarial learning for deepfake detection by leveraging pre-trained models and cross-attention fusion. //Image and Vision Computing. -2026.-Vol.166: 105885. DOI 10.1016/j.imavis.2025.105885.

23. Tianle Yang, Chengzhe Sun, Siwei Lyu, Phil Rose. Forensic deep-fake audio detection using segmental speech features. //Forensic Science International. -2026.-Vol.379:112768. DOI 10.1016/j.forsciint.2025.112768.

24. Arsalan Rahman Mirza, Abdulbasit K. Al-Talabani. Spoofing countermeasures for fake speech detection using brute force features. //Computer Speech & Language. -2025.-Vol.90:101732. DOI 10.1016/j.csl.2024.101732.

25. Gul Tahaoglu, Daniele Baracchi, Dasara Shullani, et.al. Deepfake audio detection with spectral

features and ResNeXt-based architecture. //Knowledge-Based Systems.2025; 323: 113726. DOI 10.1016/j.knosys.2025.113726.

26. Dengtai Tan, Boao Tan, Deyi Yang, Yang Yang, Chengyu Niu. Mutual-guidance framework for audio DeepFake detection via multi-dimensional feature interaction. //Neurocomputing. -2026. – Vol.672:132732. DOI 10.1016/j.neucom.2026.132732.

27. Gökay Dişken. Delta feature maps with application to spoofed speech detection. // Computers and Electrical Engineering. -2025.-Vol.128(PartB):110748. DOI 10.1016/j.compeleceng.2025.110748.

28. Роговой В., Коржук В. М., Александров Д. С. Метод выявления синтетической речи в цифровых аудиозаписях с использованием признаков речевой акустики. //Вестник УрФО. Безопасность в информационной сфере. -2025.-№.2(56). DOI 10.14529/secur250201.

29. Ariel Cohen, Denis Shyrman, Aleksandr Solonskyi, Roman Frenkel, et.al. Robust prosody modeling for synthetic speech detection. //Speech Communication. -2025.-Vol.174: 103283. DOI 10.1016/j.specom.2025.103283.

30. Hamed Al-Tairi, Ali Javed, Tasawer Khan,et.al. DeepLASD countermeasure for logical access audio spoofing. //Scientific Reports. -2025.-Vol.15:20839. DOI 10.1038/s41598-025-04808-5.

31. Xue Ouyang, Chunhui Wang, Bin Zhao, Hao Li. Adaptive reverse perturbation network for audio deepfake detection. //Neurocomputing. -2025.-Vol.658: 131466. DOI 10.1016/j.neucom.2025.131466.

32. Kaijun Mai, Chen Chen, Yuhongxu Feng, Ao Li, Liang Xi. Dual-level principal component fusion networks for synthetic speech detection. //Information Fusion. -2026.-Vol.127(PartA): 103738. DOI 10.1016/j.inffus.2025.103738.

33. Jagabandhu Mishra, Manasi Chhibber, Hye-jin Shim, Tomi H. Kinnunen. Towards explainable spoofed speech attribution and detection: A probabilistic approach for characterizing speech synthesizer components. //Computer Speech & Language. -2026.-Vol.95:101840. DOI 10.1016/j.csl.2025.101840.

**References**

1. Kuznecov A.Ju., Murtazint R.A., Garipov I.M., Fjodorov E.A., Holodenina A.V., Vorob’ eva A.A. Metody protivodejstvija atakam posredstvom sinteza rechi na golosovye biometricheskie sistemy v bankovskoj sfere. //Nauchno-tehnicheskij Vestnik. -2021.-T.21. № 1.- S.109-117. DOI 10.17586/2226-1494-2021-21-1-109-117. [in Russian]

2. Gueltoum Bendiab, Houda Haiouni, Isidoros Moulas, Stavros Shiaeles. Deepfakes in digital media forensics: Generation, AI-based detection and challenges. //Information Security and Applications. -2025.-Vol.88 103935. DOI 10.1016/j.jisa.2024.103935.

3. Khalid Zaman, Kai Li, Melike Sah, Cem Direkoglu, et.al. Transformers and audio detection tasks: An overview. //Digital Signal Processing. -2025.-Vol.158:104956. DOI 10.1016/j.dsp.2024.104956.

4. Qurat Ul Ain, Fatima Khalid, Hafsa Ilyas, Ali Javed, et.al. Adversarial and generative AI-based anti-forensics in audio-visual deepfake detection: A comprehensive review and analysis. //Information Fusion. -2026.-Vol.132: 104120. DOI 10.1016/j.inffus.2025.104120.

5. Xin Wang, Héctor Delgado, Hemlata Tak, Jee-weon Jung, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, Nicholas Evans, et.al. ASVspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech. //Computer Speech & Language. -2026.-Vol.95:101825. DOI 10.1016/j.csl.2025.101825.

6. Buğra Eyidoğan, Gökberk Özsoy, Pedram Khatamino,et.al. Voice Liveness Detection KYC Project:

Distinguishing Genuine and Spoofed Voices Using Deep Learning. //Procedia Computer Science. -2025.-Vol.255.-P.63-72. DOI 10.1016/j.procs.2025.02.261.

7. Andreas Triantafyllopoulos, Anika A. Spiesberger, Iosif Tsangko, Xin Jing, et.al.Vishing: Detecting social engineering in spoken communication - A first survey & urgent roadmap to address an emerging societal challenge. //Computer Speech & Language. -2025.-Vol.94: 101802. DOI 10.1016/j.csl.2025.101802.

8. Awais Khan, Khalid Mahmood Malik. SpoTNet: A spoofing-aware Transformer Network for Effective Synthetic Speech Detection.2nd ACM International Workshop on Language. - AI against Disinformation. //History. -2023.-P.10-18. DOI 10.1145/3592572.3592841.

9. Şule Bekiryazıcı, Cemal Hanilçi, Neyir Ozcan. Toward robust replay attack detection in Automatic Speaker Verification: A study of spectrum estimation and channel magnitude response modeling. //Computer Speech & Language. -2026.-Vol.98:101906. DOI 10.1016/j.csl.2025.101906.

10. Arth J. Shah, Madhu R. Kamble. Leveraging Kolmogorov-Arnold networks for voice liveness detection in anti-spoofing systems. //Speech Communication. -2026.-Vol.178:103364. DOI 10.1016/j.specom.2026.103364.

11. Tingting Wang,Marvin Borsdorf, Qoquan Zhang, et.al. Going deeper with log-graph Fourier transform-based feature extraction for playback speech detection. //Applied Acoustics. -2026.-Vol.245:111212. DOI 10.1016/j.apacoust.2025.111212.

12. Jiahong Ye, Diqun Yan, Songyin Fu, Bin Ma, Zhihua Xia. One-class network leveraging spectro-temporal features for generalized synthetic speech detection. //Speech Communication. -2025.-Vol.169:103200. DOI 10.1016/j.specom.2025.103200.

13. Jingze Lu, Yuxiang Zhang, Zhuo Li,et.al. Leveraging distance information for generalized spoofing speech detection. //Computer Speech & Language. -2025.-Vol.94: 101804. DOI 10.1016/j.csl.2025.101804.

14. Zhenyu Wang, John H. L. Hansen. Toward Improving Synthetic Audio Spoofing Detection Robustness via Meta-Learning and Disentangled Training with Adversarial Examples. // IEEE Access. -2024.-Vol.12.-P. (99):1-1. DOI 10.1109/ACCESS.2024.3421281.

15. Jianqiang Zhang, Yushui Geng, Peng Zhang,et.al. One-class neural network with hybrid pooling on dual-band frequency for spoofing speech detection. //Computer Speech & Language. -2026.-Vol.99: 101937. DOI 10.1016/j.csl.2026.101937.

16. Falih Gozi Febrinanto, Kristen Moore, Chandra Thapa,et.al. SIGNL: A label-efficient audio deepfake detection system via spectral-temporal graph non-contrastive learning. //Expert Systems with Applications. -2026.-Vol.307:131056. DOI 10.1016/j.eswa.2025.131056.

17. Rodrigo Colnago Contreras, Monique Simplicio Viana, Everthon Silva Fonseca, et.al.Exploring multicepstral features in a new classical machine learning-based framework for replay attack detection. //Computers and Electrical Engineering. -2025.-Vol.127(PB). DOI 10.1016/j.compeleceng.2025.110570.

18. Jingxi Xue, Sanshuai Cui, Weinan Zhang. Leveraging Neural Vocoder Artifacts for Improved Synthetic Speech Detection.2024 7th International Conference on Sensors, Signal and Image. //Processing (SSIP). -2024.-P.162-166. DOI 10.1145/3725949.3725971.

19. Burak Celik, Ayhan Akbal. A multi-level and iterative feature engineering framework for deep-fake audio detection. //Expert Systems with Applications. -2026.-Vol.303:130722. DOI 10.1016/j.eswa.2025.130722.

20. Soyul Han, Jaejin Seo, Sunmook Choi, Taein Kang,et.al. Enhancing voice spoofing detection in noisy

environments using frequency feature masking augmentation. //Engineering Science and Technology, an International Journal. -2025.-Vol.63: 101972. DOI 10.1016/j.jestch.2025.101972.

21. Hao Meng, Wei Ou, Ju Huang, Haozhe Liang,et.al. A robust unified spoofing audio detection scheme. Computers and Electrical Engineerin g.-2025.-Vol.122: 109974. DOI 10.1016/j.compeleceng.2024.109974.

22. Yuanfan Jin, Yongfang Wang. CMALDD-PTAF: Cross-modal adversarial learning for deepfake detection by leveraging pre-trained models and cross-attention fusion. //Image and Vision Computing. -2026.-Vol.166: 105885. DOI 10.1016/j.imavis.2025.105885.

23. Tianle Yang, Chengzhe Sun, Siwei Lyu, Phil Rose. Forensic deep-fake audio detection using segmental speech features. //Forensic Science International. -2026.-Vol.379:112768. DOI 10.1016/j.forsciint.2025.112768.

24. Arsalan Rahman Mirza, Abdulbasit K. Al-Talabani. Spoofing countermeasures for fake speech detection using brute force features. //Computer Speech & Language. -2025.-Vol.90:101732. DOI 10.1016/j.csl.2024.101732.

25. Gul Tahaoglu, Daniele Baracchi, Dasara Shullani, et.al. Deepfake audio detection with spectral features and ResNeXt-based architecture. //Knowledge-Based Systems.2025; 323: 113726. DOI 10.1016/j.knosys.2025.113726.

26. Dengtai Tan, Boao Tan, Deyi Yang, Yang Yang, Chengyu Niu. Mutual-guidance framework for audio DeepFake detection via multi-dimensional feature interaction. //Neurocomputing. -2026. – Vol.672:132732. DOI 10.1016/j.neucom.2026.132732.

27. Gökay Dişken. Delta feature maps with application to spoofed speech detection. // Computers and Electrical Engineering. -2025.-Vol.128(PartB):110748. DOI 10.1016/j.compeleceng.2025.110748.

28. Rogovoj V., Korzhuk V. M., Aleksandrov D. S. Metod vyjavlenija sinteticheskoj rechi v cifrovyh audiozapisjah s ispol’ zovaniem priznakov rechevoj akustiki. //Vestnik UrFO. Bezopasnost’ v informacionnoj sfere. -2025.-№. 2(56). DOI 10.14529/secur250201. [in Russian]

29. Ariel Cohen, Denis Shyrman, Aleksandr Solonskyi, Roman Frenkel, et.al. Robust prosody modeling for synthetic speech detection. //Speech Communication. -2025.-Vol.174: 103283. DOI 10.1016/j.specom.2025.103283.

30. Hamed Al-Tairi, Ali Javed, Tasawer Khan,et.al. DeepLASD countermeasure for logical access audio spoofing. //Scientific Reports. -2025.-Vol.15:20839. DOI 10.1038/s41598-025-04808-5.

31. Xue Ouyang, Chunhui Wang, Bin Zhao, Hao Li. Adaptive reverse perturbation network for audio deepfake detection. //Neurocomputing. -2025.-Vol.658: 131466. DOI 10.1016/j.neucom.2025.131466.

32. Kaijun Mai, Chen Chen, Yuhongxu Feng, Ao Li, Liang Xi. Dual-level principal component fusion networks for synthetic speech detection. //Information Fusion. -2026.-Vol.127(PartA): 103738. DOI 10.1016/j.inffus.2025.103738.

33. Jagabandhu Mishra, Manasi Chhibber, Hye-jin Shim, Tomi H. Kinnunen. Towards explainable spoofed speech attribution and detection: A probabilistic approach for characterizing speech synthesizer components. //Computer Speech & Language. -2026.-Vol.95:101840. DOI 10.1016/j.csl.2025.101840.

***Сведения об авторах***

Ержан Ә.Е.- докторант, НАО «Казахский национальный исследовательский технический университет им. К.И.Сатбаева», Алматы, Казахстан, e-mail: a.yerzhan@satbayev.university;

Ахмедиярова А. Т.- PhD, ассоциированный профессор, НАО «Казахский национальный исследовательский технический университет им. К.И.Сатбаева», Алматы, Казахстан, e-mail: a.akhmediyarova@satbayev.university;

Алимсеитова Ж.К.- PhD, профессор, НАО «Казахский национальный исследовательский технический университет им. К.И.Сатбаева», г Алматы, Казахстан, e-mail: a.akhmediyarova@satbayev.university.

***Information about the authors***

Yerzhan A.-doctoral student, K.I. Satbayev Kazakh National Research Technical University, Almaty, Kazakhstan, e-mail: a.yerzhan@satbayev.university;

Akhmediyarova A.-Ph.D., associate professor, K.I. Satbayev Kazakh National Research Technical University, Almaty, Kazakhstan, e-mail: a.akhmediyarova@satbayev.university;

Alimseitova Zh.-Ph.D., professor, K.I. Satbayev Kazakh National Research Technical University, Almaty, Kazakhstan, e-mail: a.akhmediyarova@satbayev.university.