УДК 004.056.53 **DOI** 10.15622/ia.24.1.8

М.В. Евсюков

СУБЪЕКТОЗАВИСИМЫЙ МЕТОД ОБНАРУЖЕНИЯ АТАК НА БИОМЕТРИЧЕСКОЕ ПРЕДЪЯВЛЕНИЕ В СИСТЕМАХ РАСПОЗНАВАНИЯ ДИКТОРА НА ОСНОВЕ ОБНАРУЖЕНИЯ АНОМАЛИЙ

Евсюков М.В. **Субъектозависимый метод обнаружения атак на биометрическое предъявление в системах распознавания диктора на основе обнаружения аномалий.**

Аннотация. Основная тенденция, присущая современным исследованиям в области обнаружения атак на биометрическое предъявление, заключается в том, что в большинстве работ применяется субъектонезависимый подход. Тем не менее, существует ряд исследований, свидетельствующих о перспективности применения субъектозависимого подхода, который подразумевает использование информации о предполагаемой личности субъекта для увеличения точности обнаружения спуфинга. В связи с этим, цель данной работы – реализация субъектозависимого метода обнаружения атак на биометрическое предъявление в системах распознавания диктора на основе обнаружения аномалий, а также его экспериментальная оценка применительно к задаче обнаружения синтезированной речи и преобразованного голоса. Для извлечения признаков используются искусственные нейронные сети, предобученные для задач обнаружения атак на биометрическое предъявление, распознавания диктора и распознавания звуковых паттернов. В качестве классификаторов применяется ряд моделей обнаружения аномалий, каждая из которых обучается на подлинных данных целевого диктора. Экспериментальная оценка предложенного метода с использованием набора данных ASVspoof 2019 LA показывает, что лучшая субъектозависимая система обнаружения атак на биометрическое предъявление, использующая нейронную сеть, предобученную для распознавания дикторов, обеспечивает EER (Equal Error Rate, равный процент ошибок) равный 4.74%. Данный результат свидетельствует о том, что признаки, извлечённые сетями, предобученными для распознавания диктора, содержат полезную информацию для обнаружения атак на биометрическое предъявление. Кроме того, предложенный метод позволил увеличить точность трёх базовых систем ОАБП, предназначенных для обнаружения синтезированного голоса. При проведении экспериментов с двумя базовыми системами на наборе данных ASVspoof 2019 LA улучшение EER составило 7.1% и 9.2%, а min t-DCF – 4.6%, относительно исходного результата. При проведении экспериментов с третьей базовой системой на наборе данных ASVspoof 2021 LA улучшение EER составило 3.9% относительно исходного результата с незначительным улучшением min t-DCF.

Ключевые слова: субъектозависимый подход, обнаружение спуфинга, обнаружение атак на биометрическое предъявление, биометрические системы, голосовая биометрия, трансфер обучения, обнаружение аномалий.

1. Введение. Современные методы распознавания диктора демонстрируют высокую точность при обработке подлинного человеческого голоса [1], однако их главным недостатком является уязвимость атакам на биометрическое предъявление [2]. Под термином «атака на биометрическое предъявление» (АБП) понимается предъявление биометрической системе скопированного,

сгенерированного, преобразованного или искажённого сигнала биометрической характеристики с целью вмешательства в процесс её функционирования [3]. Термин «спуфинг-атака» является синонимом термина «атака на биометрическое предъявление». В связи с высокой актуальностью угрозы АБП обнаружение атак на биометрическое предъявление (ОАБП) является важнейшим направлением исследований, а подсистема ОАБП является необходимой составной частью современных голосовых биометрических систем [3].

В то время как первые исследования в области ОАБП опирались на использование статистических моделей [4] и конструирование признаков [5], развитие методов машинного обучения повлекло за собой распространение глубоких нейронных сетей [6], что в свою очередь позволило создать сквозные системы (также известные как интегральные [7]), которые принимают на вход необработанное аудио без предварительного извлечения признаков [8]. В настоящий момент наиболее актуальными задачами в рассматриваемой области являются противодействие дипфейк-атакам, состязательным атакам на системы распознавания диктора и системы ОАБП, а также разработка систем распознавания диктора, обладающих встроенной защитой от данных видов атак. Именно на решение перечисленных задач была направлена недавняя конференция ASVspoof 5 [9].

Видное место на международных конференциях и конкурсах, посвящённых обнаружению АБП, направленных против голосовых биометрических систем, занимают работы, выполненные российскими учёными. Система, предложенная в работе [10], заняла второе место на конкурсе ASVspoof 2015, посвящённому обнаружению синтеза речи и преобразования голоса. Система, предложенная в работе [11], заняла первое место на конкурсе ASVspoof 2017, посвящённому обнаружению АБП, направленных против голосовых биометрических систем, использующих повторное воспроизведение. Системы, предложенные в работе [12] заняли первое место в конкурсе ASVspoof 2019 в секции Logical Access (обнаружение синтеза речи и преобразования голоса) среди одиночных систем и второе место среди систем-ансамблей, а также третье место в секции Physical Access (обнаружение повторного воспроизведения) среди одиночных систем и второе место среди систем-ансамблей. Системы, предложенные в работе [13], заняли первое место на конкурсе ASVspoof 2021 в секциях Logical Access и Deepfake, а также третье место в секции Physical Access.

Основная тенденция, присущая современным исследованиям в области ОАБП, заключается в том, что в большинстве работ применяется субъектонезависимый подход. Это означает, что

создатели систем ОАБП обучают модель машинного обучения на большом наборе данных, который содержит примеры голосов разных людей. Обученная таким образом модель ОАБП способна отличать подлинный голос от АБП, независимо от личности диктора, даже для дикторов, голоса которых не включены в обучающий набор данных.

Несмотря на то, что модель ОАБП, как правило, обучается с использованием субъектонезависимого подхода, системы ОАБП обычно функционируют во взаимодействии с системами верификации диктора, которые обладают информацией о предполагаемой личности субъекта. Существуют исследования, которые демонстрируют, что применение этой информации в рамках системы ОАБП позволяет повысить её точность.

Например, в работе [14], посвящённой обнаружению атак повтором, проанализировано влияние разнообразия дикторов в наборе данных на распределения голосовых признаков. Данное исследование приводит экспериментальное обоснование того, что подлинный голос и примеры АБП, использующие повторное воспроизведение, проще отличить друг от друга в случае распределений голосовых признаков одного диктора, чем в случае распределений множества дикторов. Кроме того, авторы работы [14] создают субъектозависимые системы ОАБП путём адаптации моделей смеси гауссовых распределений и нейронных сетей для конкретных дикторов, используя подлинные и сфабрикованные данные. В результате разработанные субъектозависимые системы при прочих равных демонстрируют бóльшую точность, чем их субъектонезависимые аналоги. Похожее исследование для обнаружения синтезированного голоса представлено в работах [15, 16]. Основное отличие данного исследования от работ [14 – 16] заключается в том, что в качестве классификаторов используется набор методов обнаружения аномалий.

В другой работе [17] обучается свёрточная нейронная сеть xResNet при помощи функции потерь OC-Softmax [18] для задачи субъектонезависимого обнаружения синтезированного голоса. Далее данная сеть используется для извлечения признаков, на которых обучается PLDA-модель (Probabilistic Linear Discriminant Analysis, вероятностный линейный дискриминантный анализ) [19], которая также выполняет субъектонезависимое ОАБП. Затем глобальная PLDA-модель адаптируется для целевых дикторов при помощи их подлинных данных.

Структура системы ОАБП, рассматриваемая данной работе, похожа на структуру, описанную в [17]. В обеих работах рассматриваются субъектозависимые модели ОАБП, обученные на

признаках, для извлечения которых используются глубокие нейронные сети. Однако отличие данной работы от [17] заключается в том, что в качестве классификатора применяется не PLDA, а набор моделей обнаружения аномалий, для обучения которых используются только подлинные данные. Другое отличие от работы [17] заключается в том, что в данной работе для обучения и оценки моделей используется набор данных ASVspoof 2019 LA, в то время как в работе [17] применяется набор данных, полученный путём объединения нескольких баз данных. Тем не менее, использование субъектозависимого подхода не позволило авторам работы [17] дополнительно улучшить значение EER, по сравнению с предложенной ими субъектонезависимой системой.

В работе [20] предлагается субъектозависимый вариант модели ОАБП AASIST (Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks) [21], а также рассматриваются различные способы инкорпорирования специфичной информации о голосе целевого диктора в систему ОАБП. В результате субъектозависимый вариант системы демонстрирует бóльшую точность, чем субъектонезависимый. Отличие данной работы от [20] заключается в том, что, в то время как система, предложенная в [20] является информированной о личности субъекта (speaker-aware) и способна обрабатывать голоса разных дикторов, системы, предложенные в данной работе и в работах [14, 16, 17], предназначены для обработки голоса конкретного целевого диктора.

Работы [14, 16, 17, 20] являются примерами успешного увеличения точности ОАБП за счёт использования информации о личности диктора, который проходит верификацию. Однако в статье [22], в которой исследуется обнаружение атак повтором при распознавании по геометрии лица, был реализован другой способ добавления субъектозависимой информации в модель ОАБП.

В условиях практического применения разработчику системы ОАБП доступна информация о личности предполагаемого диктора, а также образцы его подлинного голоса, которые использовались для регистрации в системе биометрического распознавания или были собраны в процессе её функционирования. Целесообразно исходить из предположения, что примеры сфабрикованных данных для произвольного диктора отсутствуют, поскольку самостоятельная генерация примеров АБП силами разработчика системы не только существенно повышает трудоёмкость создания системы ОАБП, но и не позволяет обеспечить достаточное разнообразие видов спуфинга, в связи с непрерывным появлением новых угроз. Перечисленные

ограничения привели авторов работы [22] к использованию моделей обнаружения аномалий для субъектозависимого ОАБП.

Подход к бинарной классификации, известный как обучение с одним классом или обнаружение аномалий, целесообразно использовать в том случае, когда один из классов, называемый положительным или целевым, хорошо характеризуется экземплярами в обучающих данных, а для другого класса, именуемого нецелевым или отрицательным, данные полностью отсутствуют, немногочисленны или не образуют статистически репрезентативной выборки генерального распределения отрицательного класса [23]. Механизм функционирования методов обнаружения аномалий позволяет учесть тот факт, что положительный класс более полно представлен в обучающем наборе данных, чем отрицательный [24]. В связи с этим в ходе обучения положительные примеры данных используются как основные, а отрицательные – как вспомогательные, позволяющие уточнить решающую границу. В случае использования методов обнаружения аномалий применительно к задаче ОАБП, в качестве положительного (нормального) класса выступают примеры подлинных данных, а в качестве отрицательного (аномального) – примеры данных, используемых для проведения АБП. Различные методы обучения с одним классом ранее были реализованы в рамках исследований, посвящённых обнаружению АБП, направленных против голосовых биометрических систем, и продемонстрировали высокую точность, выйдя на лидирующие позиции конкурсов ASVspoof 2015 [25] и ASVspoof 2019 [18].

Несмотря на это, в большинстве современных исследований задача ОАБП рассматривается как задача классификации с несколькими классами, что подразумевает равноправное использование подлинных и сфабрикованных обучающих данных. Основное преимущество данного подхода заключается в том, что он позволяет достигнуть высокой точности противодействия известным видам АБП. С другой стороны, его основной недостаток заключается в том, что модель, обученная таким образом, не обладает достаточной обобщающей способностью против неизвестных атак. Вследствие этого эффективность противодействия системы алгоритмам спуфинга, которые не представлены в обучающем наборе данных, оказывается недостаточно высокой.

Таким образом, существует ряд научных работ, свидетельствующих об эффективности использования субъектозависимого подхода [14 – 17, 20] и методов обнаружения аномалий [18, 25] применительно к обнаружению АБП, направленных

против голосовых биометрических систем. Кроме того, в работе [22] продемонстрированы преимущества совместного использования данных техник применительно к задаче обнаружения АБП, направленных против систем биометрического распознавания по геометрии лица. Тем не менее, совместное использование субъектозависимого подхода и методов обнаружения аномалий ранее не было исследовано применительно к обнаружению АБП, направленных против систем распознавания диктора.

2. Постановка задачи. Основная цель данной работы – реализация субъектозависимого метода ОАБП в системах распознавания диктора на основе обнаружения аномалий и его экспериментальная оценка применительно к задаче обнаружения синтезированного голоса.

Для достижения поставленной цели необходимо решить следующие задачи:

– описание системы ОАБП, построенной в соответствии с субъектозависимым методом ОАБП в системах распознавания диктора на основе обнаружения аномалий;

– оценка эффективности предлагаемого метода при извлечении признаков с использованием искусственных нейронных сетей, предобученных для решения различных задач в области обработки речи и звука;

– оценка эффективности применения различных моделей обнаружения аномалий в рамках систем, построенных в соответствии с предлагаемым методом.

3. Предлагаемый метод обнаружения атак на биометрическое предъявление. Структура системы ОАБП, построенной в соответствии с субъектозависимым методом ОАБП в системах распознавания диктора на основе обнаружения аномалий, представлена на рисунке 1.

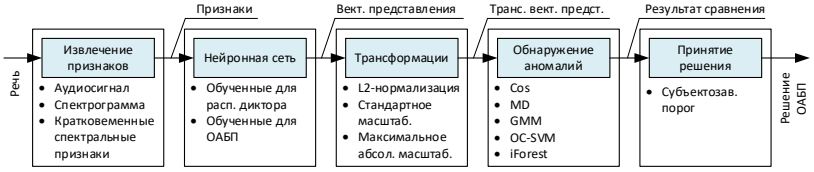

Рис. 1. Структура системы ОАБП, построенной в соответствии с субъектозависимым методом ОАБП в системах распознавания диктора на основе обнаружения аномалий

Для извлечения голосовых признаков используются предобученные искусственные нейронные сети, в связи с высокой эффективностью, которую они демонстрируют применительно к задаче обнаружения АБП, направленных против голосовых биометрических систем [6]. Поскольку разные нейронные сети обрабатывают разные голосовые признаки в качестве входных данных, выбор алгоритма извлечения признаков определяется требованиями конкретной модели машинного обучения. Далее в данном исследовании признаки, извлечённые при помощи нейронных сетей, называются векторными представлениями, чтобы подчеркнуть их отличие от признаков, для извлечения которых используются другие вычислительные методы.

Извлечённые при помощи искусственной нейронной сети векторные представления обрабатываются с использованием методов обнаружения аномалий. Известно, что точность некоторых методов обнаружения аномалий может быть повышена путём предварительного применения методов трансформации к обрабатываемым векторным данным [26]. В связи с этим, в рамках исследования предлагаемого метода оценивается влияние таких преобразований, как l2-нормализация [27], стандартное масштабирование [26], максимальное абсолютное масштабирование [26] и метод главных компонент [28], на точность ОАБП.

На этапе регистрации диктора субъектозависимая модель обнаружения аномалий обучается на трансформированных векторных представлениях с использованием примеров подлинного голоса целевого диктора. Кроме того, в соответствии с процедурой, продемонстрированной в экспериментальной части работы, вычисляется субъектозависимое пороговое значение. На этапе применения системы модель обнаружения аномалий применяется к трансформированным векторным представлениям для оценки степени подлинности предъявленных данных. Затем степень подлинности сравнивается с пороговым значением для получения решения классификации фрагмента речи: подлинный или АБП.

Процесс обучения системы ОАБП с использованием субъектонезависимого подхода представлен на рисунке 2. Процесс обучения системы ОАБП с применением субъектозависимого метода ОАБП в системах распознавания диктора на основе обнаружения аномалий представлен на рисунке 3. Из рисунка 2 видно, что субъектонезависимый подход предполагает использование как подлинных, так и сфабрикованных данных различных дикторов для обучения глобальной модели ОАБП. В то же время, как показано на

рисунке 3, субъектозависимый метод ОАБП в системах распознавания диктора на основе обнаружения аномалий предполагает использование только подлинных обучающих данных и создание собственной модели ОАБП для каждого целевого диктора.

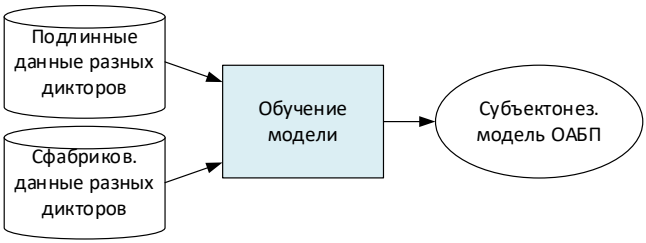

Рис. 2. Процесс обучения системы ОАБП с применением субъектонезависимого подхода

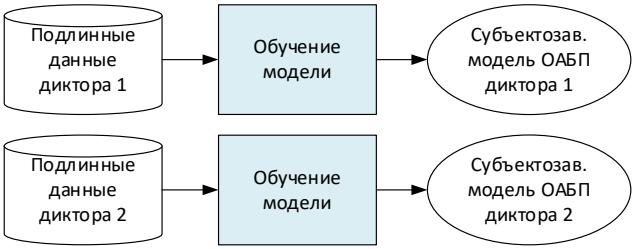

Рис. 3. Процесс обучения системы ОАБП с применением субъектозависимого метода ОАБП в системах распознавания диктора на основе обнаружения аномалий

3.1. Искусственные нейронные сети. В рамках данной работы для извлечения голосовых признаков используются три группы предобученных искусственных нейронных сетей:

– нейронные сети, предобученные для задачи ОАБП;

– нейронные сети, предобученные для задач идентификации и верификации диктора;

– нейронные сети, предобученные для задачи распознавания звуковых паттернов.

3.1.1. Нейронные сети, предобученные для задачи ОАБП. Эксперименты, в которых для извлечения признаков используются нейронные сети, предобученные для задачи обнаружения синтезированного голоса, проводятся для того, чтобы проверить возможность применения предлагаемого метода для увеличения точности существующих субъектонезависимых систем ОАБП.

Информация об используемых глубоких нейронных сетях, предобученных для задачи обнаружения синтезированного голоса, представлена в таблице 1.

Таблица 1. Используемые глубокие нейронные сети, предобученные для обнаружения синтезированного голоса

Исследование	Модель	Длительность фрагмента речи, сек.	Размер векторных представлений	Тестовый набор данных	EER, %
[8]	Res-TSSDNet	6.4	128, 64, 32	ASVspoof 2019 LA [30]	1.64
	Inc-TSSDNet		128, 64, 32		4.04
[29]	wav2vec 2.0 + AASIST	4	512	ASVspoof 2021 LA [31]	0.82 (7.65)

Все рассматриваемые сети, предобученные для задачи обнаружения синтезированного голоса, принимают необработанные фрагменты речи в качестве входных данных. Для модели, предложенной в работе [27], авторы заявляют EER равный 0.82%. Однако в открытом доступе имеется только версия модели, EER которой составляет 7.65%. Именно она используется в данной работе.

3.1.2. Нейронные сети, предобученные для задачи распознавания диктора.

Выводы, полученные в работе [32] свидетельствуют о существовании потенциала к переносу знаний от задачи распознавания диктора к задаче ОАБП при использовании методов многоцелевого обучения. В связи с этим в рамках данного исследования проводится серия экспериментов с искусственными нейронными сетями, предобученными для распознавания диктора, чтобы проверить, позволит ли применение методов обнаружения аномалий к векторным представлениям таких сетей найти решающую границу, обеспечивающую надёжное ОАБП. В случае получения успешных результатов экспериментов будет продемонстрирована возможность применения одной нейронной сети для распознавания диктора и ОАБП, что является предпосылкой для существенного снижения вычислительной нагрузки на биометрическую систему.

Информация об используемых глубоких нейронных сетях, предобученных для задачи распознавания диктора, представлена в таблице 2. Данные сети были реализованы и обучены в рамках исследований, направленных на автоматический поиск оптимальной сетевой архитектуры [33], распознавание диктора в условиях, отличных от лабораторных [34], и сквозное распознавание с использованием необработанного аудио [35, 36].

Таблица 2. Используемые глубокие нейронные сети, предобученные для задачи распознавания диктора

Исследование	Модель	Задача	Тип входных данных	Длительность фрагмента речи, сек.	Размер векторных представлений	Набор данных	EER, %
[33]	AutoSpeech	Идент.	Спектрог.	3	2048	VoxCeleb1 [37]	8.95
		Вериф.					
	ResNet18	Идент.			512		12.30
		Вериф.					
	ResNet34	Идент.					11.99
		Вериф.					
[34]	Thin ResNet VLAD	Идент.	Спектрог.	3	512	VoxCeleb2 [38]	3.22
[35]	SincNet	Идент.	Аудио	Произвольная	2048	LibriSpeech [39]	0.96
[36]	RawNet3	Идент.	Аудио	3	256	VoxCeleb1 &2 [37, 38]	0.89

3.1.3. Нейронные сети, предобученные для задачи распознавания звуковых паттернов. В работе [22] была продемонстрирована эффективность использования искусственных нейронных сетей, предобученных для задачи распознавания образов, применительно к ОАБП при защите биометрических систем распознавания по геометрии лица. В связи с этим, в рамках данной работы исследуется возможность создания субъектозависимых систем ОАБП, которые используют нейронные сети, предобученные для распознавания звуковых паттернов, с целью извлечения признаков.

В таблице 3 представлена информация об используемых в данном исследовании глубоких нейронных сетях, предобученных для задачи распознавания звуковых паттернов.

Таблица 3. Используемых глубокие нейронные сети, предобученные для задачи распознавания звуковых паттернов

Модель	Входные данные	Размер векторных представлений	mAP (показатель точности)	AUC (показатель точности)
Cnn14_16k	Логмел-спектрограмма 16 кГц	2048	0.427	0.973
Cnn14	Логмел-спектрограмма 32 кГц	2048	0.412	0.969
Cnn14_emb32	Логмел-спектрограмма 32 кГц	32	0.364	0.958
ResNet22	Логмел-спектрограмма 32 кГц	2048	0.430	0.973
Wavegram_Logmel _Cnn14	Логмел-спектрограмма 32 кГц	2048	0.439	0.973

Работа [40] является наиболее объёмным исследованием, посвящённым обучению искусственных нейронных сетей для решения задачи распознавания звуковых паттернов. В рамках неё применяется набор данных AudioSet [41], включающий в себя более 5000 часов аудиозаписей, которые являются примерами 527 различных типов звуков. Важной частью работы [40] является анализ возможности использования сетей, предобученных для задачи распознавания звуковых паттернов, применительно к другим задачам в области обработки звука посредством применения такой техники как перенос обучения [42].

В качестве входных данных рассматриваемые модели используют логмел-спектрограммы, извлечённые из аудиозаписей продолжительностью 3 секунды.

3.2. Методы обнаружения аномалий. Методы обнаружения аномалий предназначены для того, чтобы на этапе использования системы ОАБП оценить степень подлинности векторных представлений, извлечённых из предъявленного фрагмента речи. При этом применяемые в данном исследовании методы подразумевают необходимость регистрации диктора, которая, в зависимости от конкретного метода, может принимать форму обучения модели, оценки параметров распределения вероятностей или вычисления эталона для сравнения.

Исходя из того, что на практике для регистрации диктора используется ограниченное количество данных, в этом исследовании применяются простые методы обнаружения аномалий, которые включают в себя меры расстояния в пространстве признаков, вероятностные модели и поверхностные модели машинного обучения.

3.2.1. Косинусное сходство. Косинусное сходство (cos) – мера, отражающая степень подобия двух ненулевых векторов, численно равная косинусу угла между ними [43].

Косинусное сходство векторов $\vec{a}$ и $\vec{b}$ может быть рассчитано по формуле 1 [43]:

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\left|\vec{a}\right|\left|\vec{b}\right|}, \quad (1)$$

где θ – угол между векторами $\vec{a}$ и $\vec{b}$.

На этапе регистрации вычисляется среднее значение векторных представлений, рассчитанных для обучающего набора подлинных данных диктора, которое затем используется в качестве эталона для

сравнения. На этапе применения системы вычисляется косинусное подобие между полученным ранее эталоном и векторным представлением предъявляемого фрагмента речи, которое используется в качестве степени подлинности примера тестовых данных.

3.2.2. Расстояние Махаланобиса. Расстояние Махаланобиса (Mahalanobis Distance, MD) – мера расстояния между точкой и многомерным нормальным распределением. Расстояние Махаланобиса отличается от евклидового расстояния между точкой и средним значением некоторого распределения вероятностей инвариантностью к масштабированию, а также тем, что при вычислении расстояния Махаланобиса учитываются существующие корреляции между параметрами случайной величины [44].

Расстояние Махаланобиса между точкой x и распределением вероятностей F может быть рассчитано по формуле 2 [44]:

$$MD(x,F) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)}\,, \qquad (2)$$

где μ – математическое ожидание распределения вероятностей F, S – матрица ковариации распределения вероятностей F.

На этапе регистрации, исходя из предположения, что векторные представления фрагментов речи диктора распределены по нормальному закону, с использованием тренировочного набора данных оцениваются математическое ожидание и матрица ковариации распределения вероятностей. На этапе применения системы аппроксимированные параметры распределения используются для вычисления расстояния Махаланобиса между распределением и векторным представлением тестовых данных, которое используется в качестве степени фальсифицированности примера тестовых данных.

3.2.3. Машина опорных векторов с одним классом. Машина опорных векторов (Support Vector Machine, SVM) – модель бинарной классификации, которая предусматривает обучение с учителем. Её обучение заключается в аппроксимации параметров гиперплоскости, которая разделяет различные классы данных, так, чтобы расстояние от каждого класса до неё было максимальным [45].

В рамках данного исследования применяется разновидность машины опорных векторов, предназначенная для обнаружения аномалий, которая называется «машина опорных векторов с одним классом» (One-Class SVM, OC-SVM) [45]. Её основная особенность заключается в том, что при её обучении используются данные только положительного класса.

На этапе регистрации происходит обучение машины опорных векторов с одним классом с использованием подлинных данных диктора. На этапе применения системы в качестве степени подлинности высказывания используется расстояние со знаком в пространстве векторных представлений от точки, соответствующей фрагменту речи, до разделяющей многомерной поверхности.

В ходе экспериментов выявлено, что оптимальные результаты достигаются при использовании значений гиперпараметров по умолчанию ($\nu = 0.5$, тип ядра – радиальная базисная функция).

3.2.4. Модель смеси гауссовых распределений. Модель смеси гауссовых распределений (Gaussian Mixture Model, GMM) – вероятностная модель, которая аппроксимирует многомерное распределение вероятностей при помощи взвешенной суммы конечного набора нормальных распределений [46].

Плотность вероятности модели смеси гауссовых распределений может быть представлена формулой 3 [47]:

$$p(x) = \sum_{i=1}^{M} w_i p_i(x), \tag{3}$$

где $p_i(x)$ – плотность вероятности i-го компонента смеси, M – количество компонентов смеси, w_i – вес i-го компонента смеси. При этом веса удовлетворяют ограничению $\sum_{i=1}^{M} w_i = 1$.

В свою очередь, плотность вероятности i-го компонента смеси может быть представлена формулой 4 [47]:

$$p_i(x) = \frac{1}{(2\pi)^{D/2} \left|S_i\right|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)^T (S_i)^{-1}(x-\mu_i)\right\}, \tag{4}$$

где D – количество измерений, μ_i – математическое ожидание i-го компонента смеси, S_i – матрица ковариации i-го компонента смеси.

На этапе регистрации набор подлинных данных диктора используется для обучения модели смеси гауссовых распределений. На этапе применения системы в качестве степени подлинности экземпляра данных используется логарифм правдоподобия того, что соответствующая тестовому фрагменту речи точка принадлежит аппроксимируемому распределению.

В связи с ограниченным объёмом обучающего набора данных в большинстве экспериментов использовалась модель с одним компонентом и полной матрицей ковариации. Были испытаны такие способы инициализации модели как *k*-средних и случайная пятикратная инициализация, однако точность ОАБП при их использовании не отличалась. Кроме того, была протестирована возможность применения диагональной матрицы ковариации, однако её использование привело к ухудшению результатов.

3.2.5. Модель изолирующего леса. Модель изолирующего леса (Isolation Forest, iForest) – модель машинного обучения, которая использует бинарные деревья для обнаружения аномалий. Принцип её функционирования основан на предположении о том, что аномальные точки проще отделить от остальных данных, чем нормальные. Чтобы изолировать экземпляр данных, алгоритм рекурсивно генерирует разделяющие гиперплоскости, случайным образом выбирая атрибут, а также его значение для разделения точек на две части [48].

На этапе регистрации выполняется обучение модели на подлинных данных диктора. На этапе применения системы в качестве степени подлинности экземпляра данных используется глубина дерева изоляции, т.е. количество гиперплоскостей, которые необходимо провести, чтобы отделить выбранную точку от всех остальных.

В связи с ограниченным количеством обучающих данных, наибольшая эффективность модели в ходе экспериментов была достигнута при использовании значений гиперпараметров по умолчанию (100 деревьев, подвыборка не выполняется).

4. Методы проведения экспериментов

4.1. Данные. В рамках экспериментальной части работы используется набор данных ASVspoof 2019 LA, который состоит из трёх подмножеств, предназначенных для обучения, разработки и тестирования моделей машинного обучения. Информация о количестве данных в подмножествах набора данных ASVspoof 2019 LA представлена в таблице 4 (учитываются только те дикторы, для которых имеются примеры подлинных и сфабрикованных данных) [30].

Таблица 4. Количество данных в подмножествах набора данных ASVspoof 2019 LA

Подмножество набора данных	Дикторы-мужчины			Дикторы-женщины		
	Число дикторов	Число записей для каждого диктора		Число дикторов	Число записей для каждого диктора	
		Подлин.	Сфабр.		Подлин.	Сфабр.
Обучение	8	132	1176	12	127	1116
Разработка	4	140	1848	6	154	2484
Тестирование	21	68	936	27	146	1638

Одна из особенностей экспериментального исследования субъектозависимых систем ОАБП заключается в том, что необходимо наличие обучающих и тестовых данных для каждого диктора. Поскольку подмножества набора данных ASVspoof 2019 LA не имеют общих дикторов [30], была реализована специальная процедура разделения данных каждого диктора на обучающие и тестовые, представленная на рисунке 4.

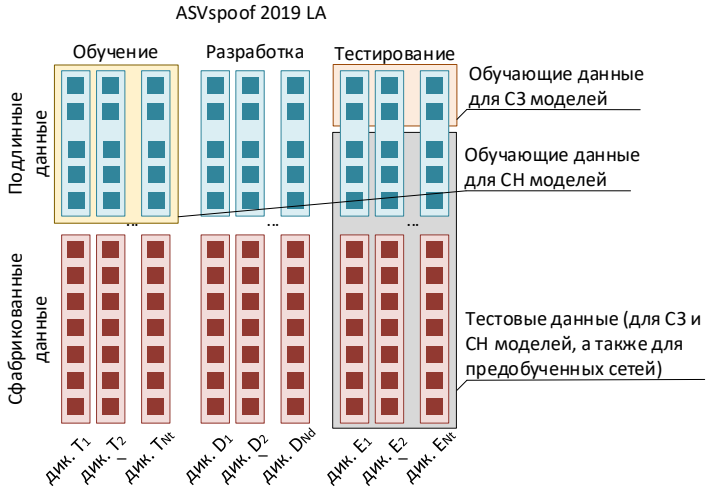

Рис. 4. Используемая процедура разделения данных на обучающие и тестовые («СЗ» – субъектозависимый, «СН» – субъектонезависимый, «дик.» – диктор)

Для обучения и тестирования субъектозависимых моделей используется тестовое подмножество набора данных ASVspoof 2019 LA, поскольку оно содержит наибольшее разнообразие примеров АБП. При этом в рамках исследования используются данные только тех дикторов, для которых имеются как подлинные, так и сфабрикованные данные. При обучении субъектозависимой модели для каждого диктора выделяются 30 подлинных фрагментов речи в качестве обучающего набора данных. Остальные подлинные данные, а также все сфабрикованные данные, используются для тестирования моделей. Для обучения субъектонезависимых моделей используется тренировочное подмножество набора данных ASVspoof 2019 LA. Экспериментальная оценка субъектонезависимых и субъектозависимых моделей проводится на одном и том же множестве фрагментов речи.

Кроме того, при исследовании предлагаемого метода применительно к системе ОАБП, описанной в работе [29], для оценки точности применяется тестовое (evaluation) подмножество набора данных ASVspoof 2021 LA, которое содержит 14816 примеров подлинных и 133360 примеров сфабрикованных аудио, распределённых неравномерно по 68 различным дикторам [31]. Как и в случае с набором данных ASVspoof 2019 LA, в рамках исследования используются данные только тех дикторов, для которых имеются подлинные и сфабрикованные данные (таких дикторов 48). Схема использования тестового подмножества набора данных ASVspoof 2021 LA аналогична схеме использования тестового подмножества набора данных ASVspoof 2019 LA, представленной на рисунке 4.

4.2. Показатели точности. Для исследования точности некоторого класса субъектозависимых систем ОАБП обучается набор идентичных систем, принадлежащих одному классу, т.е., использующих одинаковые метод обнаружения аномалий, нейронную сеть и набор алгоритмов трансформации. В качестве основного показателя точности используется среднее значение EER (Equal Error Rate, равный процент ошибок) [49] для систем исследуемого класса.

В качестве дополнительного показателя точности систем, использующих искусственные нейронные сети, предобученные для задачи обнаружения синтезированного голоса, используется среднее значение min t-DCF (минимальное значение тандемной функции стоимости обнаружения) для систем исследуемого класса, процедура вычисления которой представлена в статье [50]. При этом в текущем исследовании при вычислении min t-DCF для каждого конкретного диктора используются результаты попыток верификации, в которых он является целевым субъектом, предоставленные организаторами конкурсов ASVspoof 2019 [30] и ASVspoof 2021 [31].

Чтобы исключить влияние способа разделения подлинных данных диктора на обучающие и тестовые, эксперименты проводятся 21 раз для каждого класса систем с случайным разделением данных. Для оценки доверительных интервалов полученных значений показателей точности используется формула 5 [51]:

$$a_{CI} = \bar{a} \pm t_{0.05,20} \frac{s}{\sqrt{n}}, \quad (5)$$

где a_{CI} – 95%-ный доверительный интервал значения показателя точности, $\bar{a}$ – точечная оценка значения показателя точности (среднее

значение, полученное в ходе испытаний), $t_{0.05,20}$ – критическое значение t-распределения Стьюдента для двустороннего доверительного интервала с $\alpha = 0.05$ и 20 степенями свободы (равно 2.086), s – исправленное выборочное среднеквадратичное отклонение, n – количество экспериментов (равно 21).

4.3. Детали реализации. Экспериментальная часть исследования реализована на языке программирования Python, с применением ряда библиотек машинного обучения.

Нейронная сеть Thin ResNet VLAD [34] реализована с использованием библиотеки Keras. Остальные предобученные нейронные сети реализованы с использованием фреймворка PyTorch. При извлечении векторных представлений применялся графический процессор NVIDIA GeForce RTX 3060 GPU. Для реализации методов обнаружения аномалий использовалась библиотека Scikit-learn. При обучении и применении моделей обнаружения аномалий использовался процессор AMD Ryzen 5 5600X 6-Core 3.70 GHz.

Для оценки эффективности применения различных комбинаций методов обнаружения аномалий, искусственных нейронных сетей и алгоритмов трансформации исследовано более 570 субъектозависимых классов систем ОАБП. Для каждого класса обучено и протестировано 48 субъектозависимых систем ОАБП (в соответствии с количеством дикторов в тестовом подмножестве набора данных ASVspoof 2019 LA). Каждая из этих систем была обучена и протестирована 21 раз с различным случайным разделением данных на подлинные и тестовые. Среднее время обучения и оценки точности одной субъектозависимой системы ОАБП (для одного диктора, с единственным разбиением данных) составило 5.3 секунд.

Кроме того, были обучены 92 базовых субъектонезависимых системы ОАБП, использующих методы обнаружения аномалий.

5. Результаты экспериментов

5.1. Субъектозависимые системы, использующие нейронные сети, предобученные для распознавания диктора. В таблице 5 представлены средние EER субъектозависимых систем ОАБП, использующих методы обнаружения аномалий и искусственные нейронные сети, предобученные для задачи распознавания диктора. Для каждой комбинации нейронной сети и метода обнаружения аномалий был протестирован набор классов систем ОАБП, использующих разные трансформации, с целью определения их оптимальной конфигурации. В каждой ячейке таблицы отражён лучший результат, продемонстрированный системами ОАБП, использующими нейронную сеть, соответствующую строке таблицы

и метод обнаружения аномалий, соответствующий столбцу. Кроме того, для каждой субъектозависимой системы также приведён EER, который продемонстрировала субъектонезависимая система, использующая идентичные нейронную сеть, метод обнаружения аномалий и набор трансформаций.

Таблица 5. Средние EER (%) систем ОАБП, использующих нейронные сети, предобученные для задачи распознавания диктора, при исследовании на наборе данных ASVspoof 2019 LA («СЗ» – субъектозависимый, «СН» – субъектонезависимый)

Исследование	Модель	cos		GMM		iForest		MD		OC-SVM	
		СЗ	СН	СЗ	СН	СЗ	СН	СЗ	СН	СЗ	СН
[33]	AutoSpeech (иден.)	17.17	26.74	13.34	26.32	26.01	41.83	22.74	27.30	15.02	34.50
	AutoSpeech (вериф.)	16.20	25.04	13.17	23.16	23.88	40.72	22.49	25.81	14.90	24.59
	ResNet18 (иден.)	**4.74**	18.10	5.17	16.30	11.55	18.89	9.56	16.77	**5.06**	17.61
	ResNet18 (вериф.)	**4.79**	18.49	**4.93**	15.89	13.87	20.73	9.81	16.29	5.26	17.99
	ResNet34 (иден.)	8.21	21.42	8.64	22.43	12.83	24.16	13.47	23.47	8.61	21.73
	ResNet34 (вериф.)	6.29	19.17	7.31	18.41	10.86	19.70	12.09	19.33	6.92	19.12
[35]	SincNet	34.68	55.65	32.03	51.30	35.09	52.86	32.79	56.19	34.40	55.48
[36]	RawNet3	15.71	36.81	13.94	28.67	18.75	40.64	21.56	31.12	14.11	36.55
[34]	Thin ResNet VLAD	24.95	46.79	24.54	42.10	25.72	48.20	26.55	41.79	24.36	49.09

Анализ таблицы 5 показывает, что точность субъектозависимых систем ОАБП во всех случаях превосходит точность аналогичных субъектонезависимых систем. Кроме того, можно сделать вывод, что точность ОАБП в большей степени зависит от используемой предобученной нейронной сети, чем от метода обнаружения аномалий. Наилучший результат продемонстрировали системы ОАБП, использующие сети ResNet18, предобученные в рамках исследования [33].

Среди классов систем ОАБП, использующих одинаковую нейронную сеть, системы, в рамках которых применяются такие методы обнаружения аномалий как косинусное подобие, модель смеси гауссовых распределений и машина опорных векторов с одним классом, продемонстрировали наилучшие результаты. Применение леса изоляции и расстояния Махаланобиса приводит к ухудшению точности ОАБП, по сравнению с другими методами. При этом,

предположительного ввиду ограниченного количества данных, лес изоляции продемонстрировал наименьшую точность. По всей видимости, обработка векторных представлений размерностью 512 и более представляет существенную сложность для леса изоляции, поскольку, согласно результатам исследования [48], данный метод подвержен проблеме проклятия размерности (curse of dimensionality).

Что касается влияния трансформаций на точность ОАБП, согласно результатам экспериментов, косинусное подобие показывает наилучшие результаты при отсутствии трансформаций. Остальные методы обеспечивают наилучшие результаты при использовании l2-нормализации. Вопреки тому, что для машин опорных векторов часто рекомендуется применение масштабирования [26], его использование совместно с машинами опорных векторов с одним классом не позволило улучшить точность ОАБП. Использование метода главных компонент в качестве алгоритма трансформации в большинстве случаев приводило к заметному снижению точности. Вероятно, это связано с тем, что уменьшение размерности приводит к потере существенной части информации, необходимой для ОАБП.

В связи ограниченным объёмом обучающих данных, в ходе проведения экспериментов с субъектозависимыми системами, использующими модель смеси гауссовых распределений, наибольшая точность была достигнута при количестве компонентов смеси равным одному. Однако, поскольку субъектонезависимые системы обучаются на большем количестве данных, при их реализации использовались модели смеси гауссовых распределений с полной матрицей ковариации и количеством компонентов равным 1, 4, 16, 32 и 64. Для каждой из таких систем в таблице отражён наилучший результат. В большинстве случаев использование модели смеси гауссовых распределений с одним компонентом позволило обеспечить наименьший EER. Однако для сети ResNet34, предобученной для идентификации, и сети SincNet лучший результат был достигнут при использовании смеси с 64 компонентами, а для сети ResNet34, предобученной для верификации – при использовании смеси с 4 компонентами.

Четыре субъектозависимые системы ОАБП продемонстрировали EER равный 5.06% и менее с наибольшей предельной ошибкой 95%-ного доверительного интервала равной 0.09%. Для сравнения, в работе [8] представлены две передовые сквозные системы обнаружения синтезированного голоса, использующие нейронные сети Inc-TSSDNet и Res-TSSDNet, обученные на наборе данных ASVspoof 2019 LA, которые

демонстрируют EER 3.75% и 1.64%, соответственно. В то время как они обеспечивают более высокую точность, в ходе разработки систем, рассматриваемых в данном разделе, не использовались примеры АБП ни при обучении искусственных нейронных сетей, ни при регистрации дикторов (обучении субъектозависимых моделей обнаружения аномалий), что фактически делает все виды АБП неизвестными.

Наиболее точный класс субъектозависимых систем ОАБП, в рамках которого применяется сеть ResNet18, предобученная для задачи идентификации, и косинусное подобие продемонстрировал EER равный 4.74% (±0.07% на уровне значимости $\alpha = 0.05$). На рисунке 5 показаны значения EER, полученные для разных дикторов, в результате одной из итераций обучения и оценки систем данного класса.

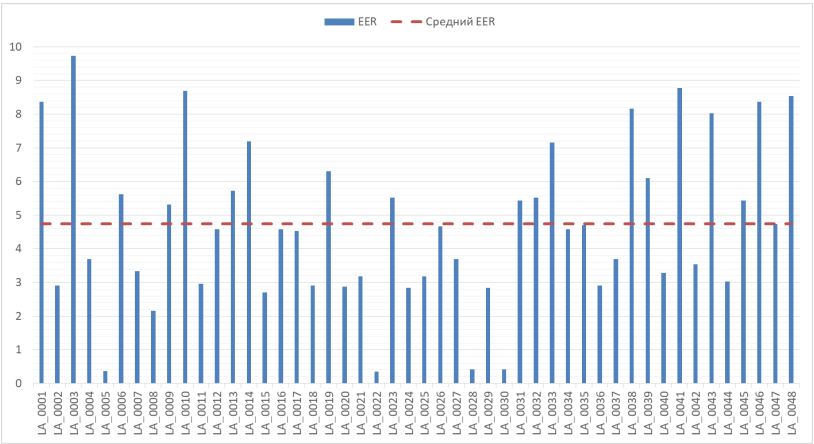

Рис. 5. Значения EER (%), полученные для разных дикторов, в результате одной из итераций обучения и оценки класса субъектозависимых систем ОАБП на наборе данных ASVspoof 2019 LA, использующих сеть ResNet18, предобученную для идентификации, и косинусное подобие

Приведённые наблюдения свидетельствуют о том, что некоторые искусственные нейронные сети, предобученные для задачи распознавания диктора, могут быть использованы для обнаружения синтезированного голоса. С практической точки зрения, полученные результаты свидетельствуют о возможности создания биометрических систем, в которых одна и та же искусственная нейронная сеть применяется для распознавания диктора и ОАБП.

5.2. Субъектозависимые системы, использующие нейронные сети, предобученные для задачи ОАБП. Таблица 6 аналогична таблице 5, представленной в предыдущем разделе. Основное отличие заключается том, что в данном разделе рассматриваются субъектозависимые системы ОАБП, которые используют нейронные сети, обученные для обнаружения синтезированного голоса в рамках исследований [8, 29]. В связи с этим, данные сети, без применения методов обнаружения аномалий, используются в качестве базовых систем ОАБП при сравнении значений показателей точности на наборе тестовых данных, проиллюстрированном на рисунке 4 (значения показателей точности для субъектонезависимых систем ОАБП, использующих методы обнаружения аномалий, не приводятся). Кроме того, для каждой системы ОАБП приведён не только EER, но также и min t-DCF, полученный в ходе испытаний.

Таблица 6. Средние значения показателей точности систем ОАБП, использующих нейронные сети, предобученные для обнаружения синтезированного голоса

Набор данных	Модель	Базовая система		cos		GMM		iForest		MD		OC-SVM	
		EER	min t-DCF	EER	min t-DCF	EER	min t-DCF	EER	min t-DCF	EER	min t-DCF	EER	min t-DCF
ASVspoof 2019 LA	Inc-TSSDNet	3.31	0.090	3.11	0.088	3.16	0.089	3.42	0.096	4.11	0.114	**3.09**	0.086
	Res-TSSDNet	1.42	0.068	**1.30**	0.065	1.84	0.079	1.87	0.087	2.89	0.089	1.36	0.067
ASVspoof 2021 LA	wav2vec 2.0 + AASIST	7.44	0.363	**7.16**	0.362	13.9	0.541	10.6	0.439	17.9	0.641	7.31	0.363

Поскольку исследуемые сети, предложенные в работе [8], имеют 4 полносвязных слоя, для каждой комбинации нейронной сети и метода обнаружения аномалий были исследованы не только различные комбинации трансформаций, но и возможность извлечения векторных представлений размеров 32, 64 и 128 значений. В ячейках таблицы отражены наилучшие результаты, полученные субъектозависимой системой, использующей указанные нейронную сеть и метод обнаружения аномалий.

Наблюдения касательно эффекта применения трансформаций на точность ОАБП, приведённые в предыдущем разделе, справедливы для экспериментов, результаты которых представлены в данном разделе.

Три класса субъектозависимых систем ОАБП, использующие сеть Inc-TSSDNet [8] для извлечения признаков, при тестировании на наборе данных ASVspoof 2019 LA [30] превзошли результат базовой системы, не использующей методы обнаружения аномалий. В рамках

данных систем применяется косинусное подобие, модель смеси гауссовых распределений и машина опорных векторов с одним классом в качестве классификаторов. Система, использующая машину опорных векторов с одним классом совместно с l2-нормализацией и обрабатывающая векторные представления размерности 64, продемонстрировала наилучший результат, относительно улучшив EER базовой системы на 7.1%, а min t-DCF – на 4.6%.

В то время как 95%-ный доверительный интервал для EER рассматриваемой субъектозависимой системы составил $3.09\% \pm 0.027\%$, 95%-ный доверительный интервал для EER базовой системы, использующей сеть Inc-TSSDNet, составил $3.31\% \pm 0.025\%$. Поскольку данные доверительные интервалы не пересекаются, приведённые результаты статистически значимы при $\alpha = 0.05$.

В то же время, 95%-ный доверительный интервал для min t-DCF рассматриваемой субъектозависимой системы составил 0.086 ± 0.0007. 95%-ный доверительный интервал для базовой системы, использующей сеть Inc-TSSDNet, составил 0.090 ± 0.0005. Поскольку данные доверительные интервалы не пересекаются, приведённые результаты статистически значимы при $\alpha = 0.05$.

Два класса субъектозависимых систем ОАБП, использующих сеть Res-TSSDNet [8], при тестировании на наборе данных ASVspoof 2019 LA [30] превзошли результат базовой системы, не использующей методы обнаружения аномалий. В качестве методов обнаружения аномалий в рамках данных систем используются косинусное подобие и машина опорных векторов с одним классом. В рамках субъектозависимых систем ОАБП обрабатываются векторные представления размерностью 64 и не используются алгоритмы трансформации. Система, использующая косинусное подобие, показывает наилучший результат, относительно превосходя EER базовой системы на 9.2%, а min t-DCF – на 4.6%.

В то время как 95%-ный доверительный интервал для EER рассматриваемой субъектозависимой системы ОАБП составил $1.30\% \pm 0.027\%$, 95%-ный доверительный интервал для EER соответствующей базовой системы составил $1.42\% \pm 0.046\%$. Поскольку данные доверительные интервалы не перекрываются, приведённые результаты статистически значимы при $\alpha = 0.05$.

В то же время, 95%-ный доверительный интервал для min t-DCF рассматриваемой субъектозависимой системы составил 0.065 ± 0.0004. 95%-ный доверительный интервал для базовой системы, использующей сеть Res-TSSDNet, составил 0.068 ± 0.0006. Поскольку

данные доверительные интервалы не пересекаются, приведённые результаты статистически значимы при $\alpha = 0.05$.

Два класса субъектозависимых систем ОАБП, использующих комбинацию сетей wav2vec 2.0 и AASIST [29] для извлечения признаков, при тестировании на наборе данных ASVspoof 2021 LA [31] превзошли результат базовой системы, не использующей методы обнаружения аномалий. В качестве методов обнаружения аномалий в рамках данных систем используются косинусное подобие (без применения трансформаций) и машина опорных векторов с одним классом (с применением l2-нормализации). Система, использующая косинусное подобие, показывает наилучший результат, относительно превосходя EER базовой системы на 3.9%.

В то время как 95%-ный доверительный интервал для EER рассматриваемой субъектозависимой системы ОАБП составил $7.16\% \pm 0.039\%$, 95%-ный доверительный интервал для EER соответствующей базовой системы составил $7.44\% \pm 0.059\%$. Поскольку данные доверительные интервалы не перекрываются, приведённые результаты статистически значимы при $\alpha = 0.05$. При этом уменьшение min t-DCF, по сравнению с базовой системой, незначительно.

Меньший относительный прирост точности в случае тестирования на наборе данных ASVspoof 2021 LA объясняется наличием искажений в тестовых данных, обусловленных их передачей по сетям связи, и применением различных кодеков [31].

Чтобы рассмотреть, как улучшения EER распределены по разным дикторам, на рисунке 6 продемонстрированы разности (улучшения) между EER базовой субъектонезависимой системы, и EER наилучшего класса субъектозависимых систем ОАБП, использующего сеть Res-TSSDNet, полученные для разных дикторов в результате одной итерации обучения и оценки систем данного класса. Как видно из рисунка 6, точность ОАБП удалось улучшить для 42 из 48 дикторов.

Таким образом, применение моделей обнаружения аномалий совместно с субъектозависимым подходом позволило увеличить точность нейронных сетей, используемых для обнаружения синтеза речи и преобразования голоса.

5.3. Субъектозависимые системы, использующие нейронные сети, предобученные для задачи распознавания звуковых паттернов. В таблице 7 представлены средние EER субъектозависимых систем ОАБП, использующих методы обнаружения аномалий и нейронные сети, предобученные для задачи

распознавания звуковых паттернов. В отличие от таблиц 5 и 6, в таблице 7 приведён EER только для субъектозависимых систем. Представленные результаты свидетельствуют о том, что сети, предобученные для задачи обнаружения звуковых паттернов в рамках исследования [40] не позволяют обеспечить эффективное извлечение признаков для дальнейшей обработки в рамках субъектозависимой системы обнаружения синтезированного голоса.

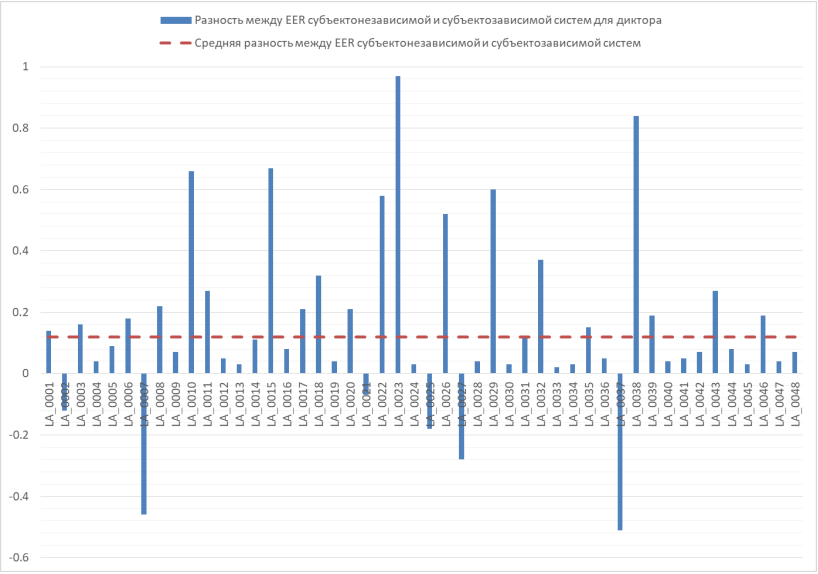

Рис. 6. Разности (улучшения) в п.п. между EER базовой субъектонезависимой системы, и EER наилучшего класса субъектозависимых систем ОАБП, использующего сеть Res-TSSDNet, полученные для разных дикторов в результате одной итерации обучения и оценки на наборе данных ASVspoof 2019 LA

Таблица 7. Средние EER (%) субъектозависимых систем ОАБП, использующих нейронные сети, предобученные для задачи распознавания звуковых паттернов, при использовании набора данных ASVspoof 2019 LA

Модель	cos	GMM	iForest	MD	OC-SVM
Cnn14	19.33	18.34	18.53	18.63	18.34
Cnn14_16k	23.13	20.21	19.79	23.45	21.22
Cnn14_emb32	24.99	29.49	29.47	30.28	24.38
ResNet22	22.84	24.75	22.67	26.37	21.78
Wavegram_Logmel_Cnn14	21.36	21.04	20.92	22.39	19.53

5.4. Пороговое значение. Методы обнаружения аномалий позволяют вычислить степень подлинности экземпляра тестовых данных. Однако для качественной работы системы ОАБП необходимо найти порог, который разделяет значения степени подлинности на подлинные и сфабрикованные таким образом, чтобы обеспечить приемлемое соотношение между количеством ложноположительных и ложноотрицательных ошибок [49].

В случае субъектонезависимой системы ОАБП существует прямолинейная процедура определения порогового значения. Вычисляются степени подлинности экземпляров тестовых данных и находится пороговое значение, которое обеспечивает требуемое соотношение процента ложных принятий (False Acceptance Rate, FAR) и процента ложных отказов (False Rejection Rate, FRR) [49].

Описанная выше процедура вычисления порогового значения неприменима для систем, построенных в соответствии с субъектозависимым методом ОАБП в системах распознавания диктора на основе обнаружения аномалий. Это связано с тем, что, во-первых, в работе [22] было продемонстрировано, что использование субъектозависимых пороговых значений совместно с субъектозависимыми моделями ОАБП обеспечивает более качественные результаты, по сравнению с использованием глобального порога, и, во-вторых, целесообразно полагать, что в сценарии практического применения субъектозависимой системы ОАБП примеры спуфинг-атак для конкретного диктора отсутствуют. В связи с этим, данном разделе статьи приводится пример выбора субъектозависимого порогового значения для систем ОАБП, построенных в соответствии с предлагаемым методом, в практическом сценарии применения.

С целью реализации данного примера для трёх дикторов из тестового подмножества набора данных ASVspoof 2019 LA были обучены идентичные субъектозависимые системы ОАБП. Данные системы используют сеть ResNet18, предобученную для идентификации диктора в рамках исследования [33], косинусное подобие и не используют алгоритмы трансформации.

Из-за отсутствия примеров АБП для конкретного диктора при выборе порогового значения возможно ориентироваться только на подлинные данные, которые могут быть использованы для вычисления FRR.

Предлагаемая процедура определения порогового значения для конкретной системы ОАБП в практическом сценарии применения состоит из двух шагов:

1. Определение FRR, который соответствует пороговому значению, задающему приемлемый средний FAR для класса субъектозависимых систем ОАБП (используются тестовые подлинные и сфабрикованные данные различных дикторов);

2. Выбор порогового значения для конкретной системы ОАБП, которое соответствует определённому ранее FRR (используются подлинные данные целевого диктора, которые не применялись для обучения модели ОАБП).

На рисунке 7 представлены кривые компромисса обнаружения и ошибок [52] для исследуемых систем.

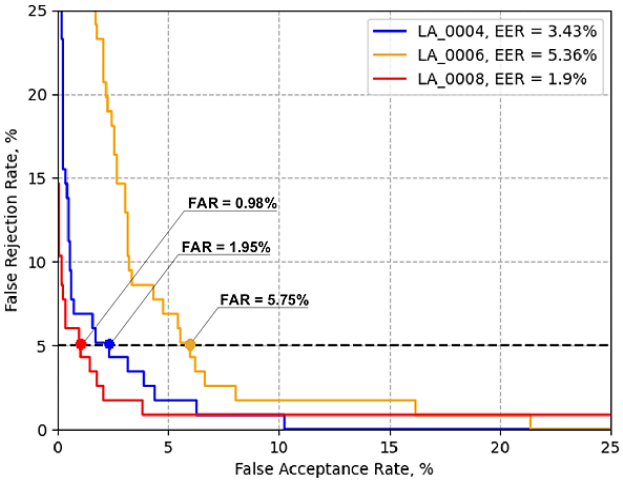

Рис. 7. Кривые компромисса обнаружения и ошибок для исследуемых систем

В качестве целевого значения FRR в данном примере выбрано 5%. Соответствующие значения FAR для выбранных дикторов, представлены на рисунке 7. В случае использования порогового значения, которое соответствует FRR равному 5%, для всех дикторов тестового подмножества набора данных ASVspoof 2019 LA, среднее значение FAR составит 4.29%.

Таким образом, знание вида кривой компромисса обнаружения и ошибок, характерной для некоторого класса субъектозависимых систем ОАБП, использующих обнаружение аномалий, позволяет определять субъектозависимое пороговое значение, регулируя FRR.

6. Заключение. В данной работе представлена реализация субъектозависимого метода ОАБП в системах распознавания диктора на основе обнаружения аномалий и проведена его экспериментальная

оценка применительно к задаче обнаружения синтезированного голоса. При этом в качестве классификатора использовались методы обнаружения аномалий, а в качестве инструмента извлечения голосовых признаков – предобученные искусственные нейронные сети.

Исследована возможность применения нейронных сетей, обученных для задачи распознавания диктора, с целью построения субъектозависимых систем ОАБП. Несмотря на то, что системы, использующие такие сети, уступают в точности передовым системам ОАБП, их производительность заслуживает внимания, поскольку примеры АБП не использовались ни при обучении данных искусственных нейронных сетей, ни при обучении моделей обнаружения аномалий. EER лучшей системы такого рода составил 4.74% при применении специального протокола испытаний с использованием набора данных ASVspoof 2019 LA. Представленные результаты подтверждают, что векторные представления сетей, предобученных для распознавания диктора, содержат ценную для задачи обнаружения синтезированного голоса информацию. Кроме того, они указывают на возможность разработки биометрической системы, которая использует одну искусственную нейронную сеть для распознавания диктора и ОАБП.

Предложенный метод позволил улучшить точность искусственных нейронных сетей, предобученных для задачи обнаружения синтезированного голоса, без их повторного обучения и без внесения каких-либо изменений в их архитектуру и параметры. При проведении экспериментов с двумя базовыми системами [8] на наборе данных ASVspoof 2019 LA [30] удалось улучшить EER на 7.1% и 9.2%, а min t-DCF – на 4.6% относительно исходных результатов. При проведении экспериментов с третьей базовой системой [29] на наборе данных ASVspoof 2021 LA [31] удалось улучшить EER на 3.9% относительно исходного результата с незначительным улучшением min t-DCF.

В то же время, применение искусственных нейронных сетей, предобученных для задачи распознавания звуковых паттернов, проявило себя как неэффективный способ извлечения признаков для задачи обнаружения синтезированного голоса.

Полученные результаты свидетельствуют о перспективности применения субъектозависимого подхода для увеличения точности современных систем ОАБП, однако требуются дальнейшие исследования для того, чтобы обеспечить более значительный прирост точности (в особенности на сложных данных, содержащих различные

искажения, примеры которых представлены в наборе данных ASVspoof 2021 LA).

В связи с этим, в ходе дальнейшей работы планируется исследовать возможные преимущества от применения субъектонезависимых примеров сфабрикованных обучающих данных, влияние количества субъектозависимых обучающих данных на прирост точности ОАБП, целесообразность использования методов аугментации для повышения качества субъектозависимых обучающих данных, а также перспективность обучения искусственных нейронных сетей с учётом субъектозависимого подхода.

Литература

1. Bai Z., Zhang X.-L. Speaker Recognition Based on Deep Learning: An overview // Neural Networks. 2021. vol. 140. pp. 65–99. DOI: 10.1016/j.neunet.2021.03.004.
2. Wang X., Yamagishi J. A Practical Guide to Logical Access Voice Presentation Attack Detection // Frontiers in Fake Media Generation and Detection. Singapore: Springer. 2022. pp. 169–214. DOI: 10.1007/978-981-19-1524-6_8.
3. ГОСТ Р 58624.1-2019. Информационные технологии. Биометрия. Обнаружение атаки на биометрическое предъявление. Часть 1. Структура. М.: Стандартинформ, 2019. 16 с.
4. Chettri B., Sturm B.L. A Deeper Look at Gaussian Mixture Model Based Anti-Spoofing Systems // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. pp. 5159–5163. DOI: 10.1109/ICASSP.2018.8461467.
5. Wei L., Long Y., Wei H., Li Y. New Acoustic Features for Synthetic and Replay Spoofing Attack Detection // Symmetry. 2022. vol. 14. no. 2. DOI: 10.3390/sym14020274.
6. Balamurali B.T., Lin K.E., Lui S., Chen J.-M., Herremans D. Toward Robust Audio Spoofing Detection: A Detailed Comparison of Traditional and Learned Features // IEEE Access. 2019. vol. 7. pp. 84229–84241. DOI: 10.1109/ACCESS.2019.2923806.
7. Марковников Н.М., Кипяткова И.С. Аналитический обзор интегральных систем распознавания речи // Труды СПИИРАН. 2018. № 3(58). С. 77–110. DOI: 10.15622/sp.58.4.
8. Hua G., Teoh A.B.J., Zhang H. Towards End-To-End Synthetic Speech Detection // IEEE Signal Processing Letters. 2021. vol. 28. pp. 1265–1269. DOI: 10.1109/LSP.2021.3089437.
9. Wang X., Delgado H., Tak H., Jung J., Shim H., Todisco M., Kukanov I., Liu X., Sahidullah M., Kinnunen T., Evans N., Lee K.A., Yamagishi J. ASVspoof 5: Crowdsourced Speech Data, Deepfakes, and Adversarial Attacks at Scale // arxiv preprint: arXiv:2408.08739v1. 2024.
10. Novoselov S., Kozlov A., Lavrentyeva G., Simonchik K., Shchemelinin V. STC Anti-spoofing Systems for the ASVspoof 2015 Challenge // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016. pp. 5475–5479. DOI: 10.1109/ICASSP.2016.7472724.
11. Lavrentyeva G., Novoselov S., Malykh E., Kozlov A., Kudashev O., Shchemelinin V. Audio Replay Attack Detection with Deep Learning Frameworks // Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2017. pp. 82–86. DOI: 10.21437/Interspeech.2017-360.
12. Lavrentyeva G., Novoselov S., Tseren A., Volkova M., Gorlanov A., Kozlov A. STC Antispoofing Systems for the ASVspoof2019 Challenge // Proceedings of the Annual

Conference of the International Speech Communication Association, Interspeech. 2019. pp. 1033–1037. DOI: 10.21437/Interspeech.2019-1768.
13. Tomilov A., Svishchev A., Volkova M., Chirkovskiy A., Kondratev A., Lavrentyeva G. STC Antispoofing Systems for the ASVspoof2021 Challenge // Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2021. pp. 61–67. DOI: 10.21437/ASVSPOOF.2021-10.
14. Suthokumar G., Sriskandaraja K., Sethu V., Ambikairajah E., Li H. An Analysis of Speaker Dependent Models in Replay Detection // APSIPA Transactions on Signal and Information Processing. 2020. vol. 9. no. 1. DOI: 10.1017/ATSIP.2020.9.
15. Евсюков М.В., Путято М.М., Макарян А.С. Исследование различимости подлинного и синтезированного голоса дикторов // Вопросы кибербезопасности. 2024. № 2(60). С. 44–52. DOI: 10.21681/2311-3456-2024-2-44-52.
16. Евсюков М.В., Путято М.М., Макарян А.С., Черкасов А.Н. Оценка точности субъектозависимого подхода к обнаружению синтезированного голоса // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2024. № 1. С. 77–93. DOI: 10.17308/sait/1995-5499/2024/1/77-93.
17. Castan D., Rahman M.H., Bakst S., Cobo-Kroenke C., McLaren M., Graciarena M., Lawson A. Speaker-Targeted Synthetic Speech Detection // Proc. of The Speaker and Language Recognition Workshop (Odyssey 2022). 2022. pp. 62–69. DOI: 10.21437/Odyssey.2022-9.
18. Zhang Y., Jiang F., Duan Z. One-Class Learning Towards Synthetic Voice Spoofing Detection // IEEE Signal Processing Letters. 2021. vol. 28. pp. 937–941. DOI: 10.1109/LSP.2021.3076358.
19. Brummer N., Swart A., Mosner L., Silnova A., Plchot O., Stafylakis T., Burget L. Probabilistic Spherical Discriminant Analysis: An Alternative to PLDA for length-normalized embeddings // Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2022. pp. 1446–1450. DOI: 10.21437/Interspeech.2022-731.
20. Liu X., Sahidullah M., Lee K.A., Kinnunen T. Speaker-Aware Anti-spoofing // Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2023. pp. 2498–2502. DOI: 10.21437/Interspeech.2023-1323.
21. Jung J.W. Heo H.S., Tak H., Shim H.J., Chung J.S., Lee B.J., Yu H.J., Evans N. AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022. pp. 6367–6371. DOI: 10.1109/ICASSP43922.2022.9747766.
22. Fatemifar S., Arashloo S.R., Awais M., Kittler J. Client-Specific Anomaly Detection for Face Presentation Attack Detection // Pattern Recognition. 2020. vol. 112. no. 8. DOI: 10.1016/j.patcog.2020.107696.
23. Seliya N., Zadeh A.A., Khoshgoftaar T.M. A Literature Review on One-Class Classification and its Potential Applications in Big Data // Journal of Big Data. 2021. vol. 8. no. 1. DOI: 10.1186/s40537-021-00514-x.
24. Khan S., Madden M. A Survey of Recent Trends in One Class Classification // Artificial Intelligence and Cognitive Science, Lecture Notes in Computer Science. 2009. vol. 6206. pp. 188–197. DOI: 10.1007/978-3-642-17080-5_21.
25. Villalba J., Miguel A., Ortega A., Lleida E. Spoofing Detection with DNN and One-Class SVM for the ASVspoof 2015 Challenge // Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2015. pp. 2067–2071. DOI: 10.21437/interspeech.2015-468.

26. Amorim L.B.V., Cavalcanti G.D.C., Cruz R.M.O. The Choice of Scaling Technique Matters for Classification Performance // Applied Soft Computing. 2023. vol. 133. DOI: 10.1016/j.asoc.2022.109924.
27. Wang C., Xu R., Xu S., Meng W., Zhang X. CNDesc: Cross Normalization for Local Descriptors Learning // IEEE Transactions on Multimedia. 2022. vol. 99. DOI: 10.1109/TMM.2022.3169331.
28. Dorabiala O., Aravkin A.Y., Kutz J.N. Ensemble Principal Component Analysis // IEEE Access. 2024. vol. 12. pp. 6663–6671. DOI: 10.1109/ACCESS.2024.3350984.
29. Tak H., Todisco M., Wang X., Jung J., Yamagishi J., Evans N. Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation // Proc. of The Speaker and Language Recognition Workshop (Odyssey 2022). 2022. pp. 112–119. DOI: 10.21437/Odyssey.2022-16.
30. Wang X. et al. ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech // Computer Speech & Language. 2020. vol. 64. DOI: 10.1016/j.csl.2020.101114.
31. Yamagishi J., Wang X., Todisco M., Sahidullah M., Patino J., Nautsch A., Liu X., Lee K.A., Kinnunen T., Evans N., Delgado H. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection // Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2021. pp. 47–54. DOI: 10.21437/asvspoof.2021-8.
32. Ge W., Tak H., Todisco M., Evans N. On the Potential of Jointly-Optimised Solutions to Spoofing Attack Detection and Automatic Speaker Verification // Proceedings of the 6th International Conference, IberSPEECH. 2022. pp. 51–55. DOI: 10.21437/iberspeech.2022-11.
33. Ding S., Chen T., Gong X., Zha W., Wang Z. AutoSpeech: Neural Architecture Search for Speaker Recognition // Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2020. pp. 916–920. DOI: 10.21437/Interspeech.2020-1258.
34. Xie W., Nagrani A., Chung J.S., Zisserman A. Utterance-Level Aggregation for Speaker Recognition in the Wild // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019. pp. 5791–5795. DOI: 10.1109/ICASSP.2019.8683120.
35. Ravanelli M., Bengio Y. Speaker Recognition from Raw Waveform with SincNet // IEEE Spoken Language Technology Workshop (SLT). 2018. pp. 1021–1028. DOI: 10.1109/SLT.2018.8639585.
36. Jung J.W., Kim Y., Heo H.S., Lee B.-J., Kwon Y., Son Chung J.S. Pushing the Limits of Raw Waveform Speaker Recognition // Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2022. pp. 2228–2232. DOI: 10.21437/Interspeech.2022-126.
37. Nagraniy A., Chungy J.S., Zisserman A. VoxCeleb: A large-scale speaker identification dataset // Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2017. pp. 2616–2620. DOI: 10.21437/Interspeech.2017-950.
38. Chung J.S., Nagrani A., Zisserman A. VoxCeleb2: Deep Speaker Recognition // Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2018. pp. 1086–1090. DOI: 10.21437/Interspeech.2018-1929.
39. Panayotov V., Chen G., Povey D., Khudanpur S. LibriSpeech: An ASR Corpus Based on Public Domain Audio Books // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015. pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

40. Kong Q., Cao Y., Iqbal T., Wang Y., Wang W., Plumbley M.D. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition // IEEE/ACM Transactions on Audio Speech and Language Processing. 2020. vol. 28. pp. 2880–2894. DOI: 10.1109/TASLP.2020.3030497.
41. Gemmeke G.F., Ellis D.P.W., Freedman D., Jansen A., Lawrence W., Moore R.C. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2017. pp. 776–780. DOI: 10.1109/ICASSP.2017.7952261.
42. Hosna A., Merry E., Gyalmo J., Alom Z., Aung Z., Azim M.A. Transfer Learning: A Friendly Introduction // Journal of Big Data. 2022. vol. 9. no. 1. DOI: 10.1186/s40537-022-00652-w.
43. Januzaj Y., Luma A. Cosine Similarity – A Computing Approach to Match Similarity Between Higher Education Programs and Job Market Demands Based on Maximum Number of Common Words // International Journal of Emerging Technologies in Learning. 2022. vol. 17. no. 12. pp. 258–268. DOI: 10.3991/ijet.v17i12.30375.
44. Ghorbani H. Mahalanobis Distance and its Application for Detecting Multivariate Outliers // Facta Universitatis, Series: Mathematics and Informatics. 2019. vol. 34. no. 3. pp. 583–595. DOI: 10.22190/fumi1903583g.
45. Alegre F., Amehraye A., Evans N. A One-Class Classification Approach to Generalised Speaker Verification Spoofing Countermeasures Using Local Binary Patterns // IEEE 6th International Conference on Biometrics: Theory, Applications and Systems (BTAS). 2013. pp. 1–8. DOI: 10.1109/BTAS.2013.6712706.
46. Scrucca L. Entropy-Based Anomaly Detection for Gaussian Mixture Modeling // Algorithms. 2023. vol. 16. no. 4. DOI: 10.3390/a16040195.
47. Reynolds D.A., Quatieri T.F., Dunn R.B. Speaker Verification Using Adapted Gaussian Mixture Models // Digital Signal Processing: A Review Journal. 2000. vol. 10. no. 1-3. pp. 19–41. DOI: 10.1006/dspr.1999.0361.
48. Liu F.T., Ting K.M., Zhou Z.H. Isolation forest // Proceedings of the Eighth IEEE International Conference on Data Mining (ICDM). 2008. pp. 413–422. DOI: 10.1109/ICDM.2008.17.
49. Hao B., Hei X. Voice Liveness Detection for Medical Devices // Design and Implementation of Healthcare Biometric Systems. 2019. pp. 109–136. DOI: 10.4018/978-1-5225-7525-2.ch005.
50. Kinnunen T., Lee K.A., Delgado H., Evans N., Todisco M., Sahidullah M., Yamagishi J., Reyonolds D.A. t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification // Proc. The Speaker and Language Recognition Workshop (Odyssey 2018), 2018. pp. 312–319.
51. Hazra A. Using the Confidence Interval Confidently // Journal of Thoracic Disease. 2017. vol. 9. no. 10. DOI: 10.21037/jtd.2017.09.14.
52. Martin A., Doggington G., Kamm T., Ordowski M. Przybocki M. The DET curve in assessment of detection task performance // Proceedings of the 5th European Conference on Speech Communication and Technology, Eurospeech (ISCA). 1997. pp. 1895–1898. DOI:10.21437/Eurospeech.1997-504.

Евсюков Михаил Витальевич — аспирант, кафедра кибербезопасности и защиты информации, Кубанский государственный технологический университет. Область научных интересов: обнаружение атак на биометрическое предъявление, голосовая биометрия, машинное обучение, постквантовая криптография. Число научных публикаций — 23. michael.evsyukov@gmail.com; улица Московская, 2, 350072, Краснодар, Россия; р.т.: +7(861)255-0346.

DOI 10.15622/ia.24.1.8

M. EVSYUKOV

SPEAKER-SPECIFIC METHOD OF SPOOFING ATTACK DETECTION BASED ON ANOMALY DETECTION

Evsyukov M. **Speaker-Specific Method of Spoofing Attack Detection Based on Anomaly Detection.**

Abstract. Most research in the field of voice presentation attack detection relies on the speaker-independent approach. Nevertheless, several scientific works indicate that using the speaker-specific approach, which involves utilizing prior knowledge about the identity of the claimed speaker to enhance the accuracy of spoofing detection, is likely to be beneficial. Therefore, the goal of this work is to propose a speaker-specific method of spoofing attack detection based on anomaly detection and to evaluate its applicability to the detection of synthesized speech and converted voice. Artificial neural networks pre-trained for the tasks of spoofing detection, speaker recognition, and audio pattern recognition are used for feature extraction. A set of anomaly detection models are used as backend classifiers. Each of them is trained on bonafide data of a target speaker. The experimental evaluation of the proposed method on the ASVspoof 2019 LA dataset shows that the best speaker-specific spoofing detection system, which uses an anomaly detection model and a neural network pre-trained for the task of speaker recognition, achieves an EER of 4.74%. This result suggests that embeddings extracted by networks pre-trained for speaker recognition contain information that can be utilized for spoofing detection. In addition, the proposed method allowed to increase the accuracy of three baseline systems pre-trained for the task of spoofing detection. Experiments with two baseline systems on the ASVspoof 2019 LA dataset showed relative improvement in terms of EER by 7.1% and 9.2%, and in terms of min t-DCF by 4.6%. Experiments with the third baseline system on the ASVspoof 2021 LA dataset showed relative improvement in terms of EER by 3.9% without significant improvement of min t-DCF.

Keywords: speaker-specific approach, spoofing detection, presentation attack detection, biometric systems, voice biometrics, transfer learning, anomaly detection.

References

1. Bai Z., Zhang X.-L. Speaker Recognition Based on Deep Learning: An overview. Neural Networks. 2021. vol. 140. pp. 65–99. DOI: 10.1016/j.neunet.2021.03.004.
2. Wang X., Yamagishi J. A Practical Guide to Logical Access Voice Presentation Attack Detection. Frontiers in Fake Media Generation and Detection. Singapore: Springer. 2022. pp. 169–214. DOI: 10.1007/978-981-19-1524-6_8.
3. GOST R 58624.1-2019. Informacionnye tehnologii. Biometrija. Obnaruzhenie ataki na biometricheskoe predjavlenie. Chast' 1. Struktura [Information technology. Biometrics. Biometric presentation attack detection. Part 1. Framework]. M.: Gosstandart Rossii, 2019. (In Russ.).
4. Chettri B., Sturm B.L. A Deeper Look at Gaussian Mixture Model Based Anti-Spoofing Systems. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. pp. 5159–5163. DOI: 10.1109/ICASSP.2018.8461467.
5. Wei L., Long Y., Wei H., Li Y. New Acoustic Features for Synthetic and Replay Spoofing Attack Detection. Symmetry. 2022. vol. 14. no. 2. DOI: 10.3390/sym14020274.
6. Balamurali B.T., Lin K.E., Lui S., Chen J.-M., Herremans D. Toward Robust Audio Spoofing Detection: A Detailed Comparison of Traditional and Learned Features. IEEE Access. 2019. vol. 7. pp. 84229–84241. DOI: 10.1109/ACCESS.2019.2923806.

7. Markovnikov N., Kipyatkova I. An Analytic Survey of End-to-End Speech Recognition Systems. Trudy SPIIRAN – SPIIRAS Proceedings. 2018. vol. 3. no. 58. pp. 77-110. DOI: 10.15622/sp.58.4. (In Russ.).
8. Hua G., Teoh A.B.J., Zhang H. Towards End-To-End Synthetic Speech Detection. IEEE Signal Processing Letters. 2021. vol. 28. pp. 1265–1269. DOI: 10.1109/LSP.2021.3089437.
9. Wang X., Delgado H., Tak H., Jung J., Shim H., Todisco M., Kukanov I., Liu X., Sahidullah M., Kinnunen T., Evans N., Lee K.A., Yamagishi J. ASVspoof 5: Crowdsourced Speech Data, Deepfakes, and Adversarial Attacks at Scale. arxiv preprint: arXiv:2408.08739v1. 2024.
10. Novoselov S., Kozlov A., Lavrentyeva G., Simonchik K., Shchemelinin V. STC Anti-spoofing Systems for the ASVspoof 2015 Challenge. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016. pp. 5475–5479. DOI: 10.1109/ICASSP.2016.7472724.
11. Lavrentyeva G., Novoselov S., Malykh E., Kozlov A., Kudashev O., Shchemelinin V. Audio Replay Attack Detection with Deep Learning Frameworks. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2017. pp. 82–86. DOI: 10.21437/Interspeech.2017-360.
12. Lavrentyeva G., Novoselov S., Tseren A., Volkova M., Gorlanov A., Kozlov A. STC Antispoofing Systems for the ASVspoof2019 Challenge. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2019. pp. 1033–1037. DOI: 10.21437/Interspeech.2019-1768.
13. Tomilov A., Svishchev A., Volkova M., Chirkovskiy A., Kondratev A., Lavrentyeva G. STC Antispoofing Systems for the ASVspoof2021 Challenge. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2021. pp. 61–67. DOI: 10.21437/ASVSPOOF.2021-10.
14. Suthokumar G., Sriskandaraja K., Sethu V., Ambikairajah E., Li H. An Analysis of Speaker Dependent Models in Replay Detection. APSIPA Transactions on Signal and Information Processing. 2020. vol. 9. no. 1. DOI: 10.1017/ATSIP.2020.9.
15. Evsyukov M.V., Putyato M.M., Makaryan A.S. [The Effect of Speaker Variability on Distinguishability of Bonafide and Synthetized Speech]. Voprosy kiberbezopasnosti – Cybersecurity issues. 2024. vol. 60. no. 2. pp. 44–52. DOI: 10.21681/2311-3456-2024-2-44-52. (In Russ.).
16. Evsjukov M.V., Putjato M.M., Makarjan A.S., Cherkasov A.N. [Assessing Accuracy of Speaker-Specific Approach to Logical Access Spoofing Detection]. Vestnik Voronezhskogo gosudarstvennogo universiteta. Serija: Sistemnyj analiz i informacionnye tehnologii – Proceedings of Voronezh State University. Series: Systems Analysis and Information Technologies. 2024. no. 1. pp. 77–93. DOI: 10.17308/sait/1995-5499/2024/l/77-93. (In Russ.).
17. Castan D., Rahman M.H., Bakst S., Cobo-Kroenke C., McLaren M., Graciarena M., Lawson A. Speaker-Targeted Synthetic Speech Detection. Proc. of The Speaker and Language Recognition Workshop (Odyssey 2022). 2022. pp. 62–69. DOI: 10.21437/Odyssey.2022-9.
18. Zhang Y., Jiang F., Duan Z. One-Class Learning Towards Synthetic Voice Spoofing Detection. IEEE Signal Processing Letters. 2021. vol. 28. pp. 937–941. DOI: 10.1109/LSP.2021.3076358.
19. Brummer N., Swart A., Mosner L., Silnova A., Plchot O., Stafylakis T., Burget L. Probabilistic Spherical Discriminant Analysis: An Alternative to PLDA for length-normalized embeddings. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2022. pp. 1446–1450. DOI: 10.21437/Interspeech.2022-731.

20. Liu X., Sahidullah M., Lee K.A., Kinnunen T. Speaker-Aware Anti-spoofing. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2023. pp. 2498–2502. DOI: 10.21437/Interspeech.2023-1323.
21. Jung J.W. Heo H.S., Tak H., Shim H.J., Chung J.S., Lee B.J., Yu H.J., Evans N. AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022. pp. 6367–6371. DOI: 10.1109/ICASSP43922.2022.9747766.
22. Fatemifar S., Arashloo S.R., Awais M., Kittler J. Client-Specific Anomaly Detection for Face Presentation Attack Detection. Pattern Recognition. 2020. vol. 112. no. 8. DOI: 10.1016/j.patcog.2020.107696.
23. Seliya N., Zadeh A.A., Khoshgoftaar T.M. A Literature Review on One-Class Classification and its Potential Applications in Big Data. Journal of Big Data. 2021. vol. 8. no. 1. DOI: 10.1186/s40537-021-00514-x.
24. Khan S., Madden M. A Survey of Recent Trends in One Class Classification. Artificial Intelligence and Cognitive Science, Lecture Notes in Computer Science. 2009. vol. 6206. pp. 188–197. DOI: 10.1007/978-3-642-17080-5_21.
25. Villalba J., Miguel A., Ortega A., Lleida E. Spoofing Detection with DNN and One-Class SVM for the ASVspoof 2015 Challenge. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2015. pp. 2067–2071. DOI: 10.21437/interspeech.2015-468.
26. Amorim L.B.V., Cavalcanti G.D.C., Cruz R.M.O. The Choice of Scaling Technique Matters for Classification Performance. Applied Soft Computing. 2023. vol. 133. DOI: 10.1016/j.asoc.2022.109924.
27. Wang C., Xu R., Xu S., Meng W., Zhang X. CNDesc: Cross Normalization for Local Descriptors Learning. IEEE Transactions on Multimedia. 2022. vol. 99. DOI: 10.1109/TMM.2022.3169331.
28. Dorabiala O., Aravkin A.Y., Kutz J.N. Ensemble Principal Component Analysis. IEEE Access. 2024. vol. 12. pp. 6663–6671. DOI: 10.1109/ACCESS.2024.3350984.
29. Tak H., Todisco M., Wang X., Jung J., Yamagishi J., Evans N. Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation. Proc. of The Speaker and Language Recognition Workshop (Odyssey 2022). 2022. pp. 112–119. DOI: 10.21437/Odyssey.2022-16.
30. Wang X. et al. ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech. Computer Speech & Language. 2020. vol. 64. DOI: 10.1016/j.csl.2020.101114.
31. Yamagishi J., Wang X., Todisco M., Sahidullah M., Patino J., Nautsch A., Liu X., Lee K.A., Kinnunen T., Evans N., Delgado H. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2021. pp. 47–54. DOI: 10.21437/asvspoof.2021-8.
32. Ge W., Tak H., Todisco M., Evans N. On the Potential of Jointly-Optimised Solutions to Spoofing Attack Detection and Automatic Speaker Verification. Proceedings of the 6th International Conference, IberSPEECH. 2022. pp. 51–55. DOI: 10.21437/iberspeech.2022-11.
33. Ding S., Chen T., Gong X., Zha W., Wang Z. AutoSpeech: Neural Architecture Search for Speaker Recognition. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2020. pp. 916–920. DOI: 10.21437/Interspeech.2020-1258.
34. Xie W., Nagrani A., Chung J.S., Zisserman A. Utterance-Level Aggregation for Speaker Recognition in the Wild. IEEE International Conference on Acoustics,

Speech and Signal Processing (ICASSP). 2019. pp. 5791–5795. DOI: 10.1109/ICASSP.2019.8683120.
35. Ravanelli M., Bengio Y. Speaker Recognition from Raw Waveform with SincNet. IEEE Spoken Language Technology Workshop (SLT). 2018. pp. 1021–1028. DOI: 10.1109/SLT.2018.8639585.
36. Jung J.W., Kim Y., Heo H.S., Lee B.-J., Kwon Y., Son Chung J.S. Pushing the Limits of Raw Waveform Speaker Recognition. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2022. pp. 2228–2232. DOI: 10.21437/Interspeech.2022-126.
37. Nagraniy A., Chungy J.S., Zisserman A. VoxCeleb: A large-scale speaker identification dataset. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2017. pp. 2616–2620. DOI: 10.21437/Interspeech.2017-950.
38. Chung J.S., Nagrani A., Zisserman A. VoxCeleb2: Deep Speaker Recognition. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech. 2018. pp. 1086–1090. DOI: 10.21437/Interspeech.2018-1929.
39. Panayotov V., Chen G., Povey D., Khudanpur S. LibriSpeech: An ASR Corpus Based on Public Domain Audio Books. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015. pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.
40. Kong Q., Cao Y., Iqbal T., Wang Y., Wang W., Plumbley M.D. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. IEEE/ACM Transactions on Audio Speech and Language Processing. 2020. vol. 28. pp. 2880–2894. DOI: 10.1109/TASLP.2020.3030497.
41. Gemmeke G.F., Ellis D.P.W., Freedman D., Jansen A., Lawrence W., Moore R.C. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2017. pp. 776–780. DOI: 10.1109/ICASSP.2017.7952261.
42. Hosna A., Merry E., Gyalmo J., Alom Z., Aung Z., Azim M.A. Transfer Learning: A Friendly Introduction. Journal of Big Data. 2022. vol. 9. no. 1. DOI: 10.1186/s40537-022-00652-w.
43. Januzaj Y., Luma A. Cosine Similarity – A Computing Approach to Match Similarity Between Higher Education Programs and Job Market Demands Based on Maximum Number of Common Words. International Journal of Emerging Technologies in Learning. 2022. vol. 17. no. 12. pp. 258–268. DOI: 10.3991/ijet.v17i12.30375.
44. Ghorbani H. Mahalanobis Distance and its Application for Detecting Multivariate Outliers. Facta Universitatis, Series: Mathematics and Informatics. 2019. vol. 34. no. 3. pp. 583–595. DOI: 10.22190/fumi1903583g.
45. Alegre F., Amehraye A., Evans N. A One-Class Classification Approach to Generalised Speaker Verification Spoofing Countermeasures Using Local Binary Patterns. IEEE 6th International Conference on Biometrics: Theory, Applications and Systems (BTAS). 2013. pp. 1–8. DOI: 10.1109/BTAS.2013.6712706.
46. Scrucca L. Entropy-Based Anomaly Detection for Gaussian Mixture Modeling. Algorithms. 2023. vol. 16. no. 4. DOI: 10.3390/a16040195.
47. Reynolds D.A., Quatieri T.F., Dunn R.B. Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing: A Review Journal. 2000. vol. 10. no. 1-3. pp. 19–41. DOI: 10.1006/dspr.1999.0361.
48. Liu F.T., Ting K.M., Zhou Z.H. Isolation forest. Proceedings of the Eighth IEEE International Conference on Data Mining (ICDM). 2008. pp. 413–422. DOI: 10.1109/ICDM.2008.17.

49. Hao B., Hei X. Voice Liveness Detection for Medical Devices. Design and Implementation of Healthcare Biometric Systems. 2019. pp. 109–136. DOI: 10.4018/978-1-5225-7525-2.ch005.
50. Kinnunen T., Lee K.A., Delgado H., Evans N., Todisco M., Sahidullah M., Yamagishi J., Reyonolds D.A. t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification. Proc. The Speaker and Language Recognition Workshop (Odyssey 2018), 2018. pp. 312–319.
51. Hazra A. Using the Confidence Interval Confidently. Journal of Thoracic Disease. 2017. vol. 9. no. 10. DOI: 10.21037/jtd.2017.09.14.
52. Martin A., Doggington G., Kamm T., Ordowski M. Przybocki M. The DET curve in assessment of detection task performance. Proceedings of the 5th European Conference on Speech Communication and Technology, Eurospeech (ISCA). 1997. pp. 1895–1898. DOI:10.21437/Eurospeech.1997-504.

Evsyukov Mikhail — Postgraduate student, Department of cybersecurity and information protection, Kuban State Technological University. Research interests: presentation attack detection, voice biometrics, machine learning, postquantum cryptography. The number of publications — 23. michael.evsyukov@gmail.com; 2, Moskovskaya St., 350072, Krasnodar, Russia; office phone: +7(861)255-0346.