ИССЛЕДОВАНИЕ РАЗЛИЧИМОСТИ ПОДЛИННОГО И СИНТЕЗИРОВАННОГО ГОЛОСА ДИКТОРОВ

Евсюков М. В.[1], Путято М. М.[2], Макарян А. С.[3]

DOI: 10.21681/2311-3456-2024-2-44-52

Цель работы: *исследование статистических различий обучающих данных, используемых для реализации субъектонезависимого и субъектозависимого подходов к обнаружению синтезированного голоса при противодействии спуфинг-атакам на системы распознавания личности по голосу.*

Методы исследования: *в качестве голосовых признаков используются линейно-частотные кепстральные коэффициенты (LFCC). Для аппроксимации вероятностных распределений голосовых признаков используется модель смеси гауссовых распределений. Для визуализации голосовых признаков используется алгоритм уменьшения размерности t-SNE. Для оценки степени различия вероятностных распределений используется расстояние Кульбака-Лейблера, рассчитываемое при помощи метода Монте-Карло.*

Результаты исследования: *мы обнаружили, что данные, принадлежащие различным дикторам, разделены на кластеры в пространстве голосовых признаков, используемых для обнаружения синтезированного голоса. Полученные результаты свидетельствуют о том, что использование субъектозависимых распределений признаков, вместо субъектонезависимых, увеличивает различимость подлинного и синтезированного голоса. Это подтверждает наше предположение о том, что разнообразие дикторов в обучающем наборе данных является запутывающим фактором при обнаружении спуфинга. Следовательно, использование субъектозависимых моделей обнаружения спуфинга, с высокой вероятностью, позволит повысить точность обнаружения синтезированного голоса.*

Научная новизна: *при помощи статистических методов, мы подтверждаем, что разнообразие дикторов в обучающем наборе данных – существенный запутывающий фактор при обучении моделей обнаружения синтезированного голоса.*

Ключевые слова: *спуфинг, атака на биометрическое предъявление, биометрия, расстояние Кульбака-Лейблера, синтезированный голос, голосовая аутентификация, распознавание по голосу, распознавание личности, модель смеси гауссовых распределений, LFCC.*

THE EFFECT OF SPEAKER VARIABILITY ON DISTINGUISHABILITY OF BONAFIDE AND SYNTHETIZED SPEECH

Evsyukov M. V.[4], Putyato M. M.[5], Makaryan A. S.[6]

The purpose of the research: *studying statistical differences between training data used for implementing speaker-independent and speaker-specific logical access voice spoofing countermeasures.*

Methods: *Linear Frequency Cepstral Coefficients (LFCC) are used as voice features. Gaussian mixture models are used for approximating probability distributions of the features. The t-SNE method is used for visualizing voice features. The Kullback-Leibler divergence is used for estimating distinguishability of the probability distributions. The value of the Kullback-Leibler divergence is computed with the help of the Monte Carlo method.*

1 Евсюков Михаил Витальевич, аспирант, Кубанский государственный технологический университет, Краснодар, Россия. ORCID: 0000-0001-7101-6251. Scopus Author ID: 57274464300. E mail: michael.evsyukov@gmail.com

2 Путято Михаил Михайлович, доцент, Кубанский государственный технологический университет, Краснодар, Россия. ORCID: 0000-0003-0414-6034. Scopus Author ID: 57226388985. E mail: putyato.m@gmail.com

3 Макарян Александр Самвелович, кандидат технических наук, доцент, заведующий кафедрой кибербезопасности и защиты информации, Кубанский государственный технологический университет, Краснодар, Россия. ORCID: 0000-0002-1801-6137. Scopus Author ID: 57226384905. E mail: msanya@yandex.ru.

4 Mikhail. V. Evsyukov, postgraduate, Kuban State Technological University, Krasnodar, Russia. ORCID: 0000-0001-7101-6251. Scopus Author ID: 57274464300. E-mail: michael.evsyukov@gmail.com

5 Mikhail M. Putyato, Associate Professor, Kuban State Technological University, Krasnodar, Russia. ORCID: 0000-0003-0414-6034. Scopus Author ID: 57226388985. E-mail: putyato.m@gmail.com

6 Alexander S. Makaryan, Ph. D., Associate Professor, Head of Department of Cybersecurity and Information Protection, Kuban State Technological University, Krasnodar, Russia. ORCID: 0000-0002-1801-6137. Scopus Author ID: 57226384905. E-mail: msanya@yandex.ru

Results: *we discovered that data belonging to different speakers is separated into clusters in the space of features used for detection of synthetized speech. Our findings suggest that using speaker-specific feature distributions, rather than speaker independent ones, enables distinguishing between bonafide and spoofed speech more easily. This supports our assumption that speaker variability in the training dataset is a confusing factor for spoofing detection. Therefore, eliminating it by using speaker-specific machine learning models is likely to increase accuracy of synthetized voice detection.*

Scientific novelty: *by using statistical methods, we confirm that speaker variability in a training dataset is a significant confusing factor while training logical access spoofing detection models.*

Keywords: *spoofing, antispoofing countermeasures, presentation attack detection, biometrics, synthetized voice, speaker recognition, biometric authentication, Gaussian mixture model, LFCC.*

Введение

Существующие методы распознавания личности по голосу демонстрируют высокую точность при обработке подлинного человеческого голоса, однако их главным недостатком является уязвимость к спуфингу. Под спуфингом биометрических систем понимаются действия злоумышленника, направленные на подделку предъявляемых биометрических характеристик, таким образом, чтобы биометрическая система распознала злоумышленника в качестве другого субъекта. В связи с высокой актуальностью угрозы спуфинга, противодействие ему является важнейшим направлением исследований в области распознавания личности по голосу, а подсистема обнаружения спуфинга является необходимой частью современных голосовых биометрических систем [1,2].

Основным регулярным событием в области исследования обнаружения спуфинга являются конференции ASVspoof [3,4], в ходе которых выходят в свет большое количество научных работ, посвящённых применению различных методов машинного обучения и обработки сигналов для обнаружения спуфинга. В то время как одни исследования направлены на разработку [5] и обучение [6] голосовых признаков, обеспечивающих наибольшую различимость подлинного голоса и спуфинга, другие предлагают использование новых архитектур нейронных сетей [7], функций потерь [8] и методов аугментации данных [9]. Сравнительно недавним событием является появление конференции SASV [10], организаторы которой стимулируют участников разрабатывать системы распознавания личности по голосу с «встроенной» защитой от спуфинга.

Основной тенденцией, присущей современным исследованиям методов обнаружения спуфинг-атак на голосовые биометрические системы, является доминирование субъектонезависимого подхода. Это означает, что создатели систем обнаружения спуфинга обучают систему на большом наборе данных, который содержит примеры голосов разных людей [11,12]. Несмотря на это, существуют исследования, свидетельствующие о перспективности применения субъектозависимого подхода к обнаружению спуфинга [13,14]. В то время как субъектонезависимый подход подразумевает обучение универсальных моделей подлинных и сфабрикованных данных для последующего обнаружения спуфинга, без привязки к голосу конкретного диктора, субъектозависимый подход подразумевает обучение отдельной модели подлинных, а также, возможно, сфабрикованных данных для каждого диктора.

Субъектозависимый подход показал высокую точность применительно к задаче обнаружения спуфинга биометрической системы распознавания по геометрии лица [13], а также при защите систем распознавания диктора от спуфинг-атак, использующих повторное воспроизведение записи голоса [14]. Тем не менее, эффективность его использования на данный момент не была изучена применительно к обнаружению широкого класса голосовых спуфинг-атак, использующих методы синтеза голоса.

Задачи исследования

Основное отличие между субъектозависимым и субъектонезависимым подходами к обнаружению спуфинга заключается в использовании разных наборов данных при обучении моделей обнаружения спуфинга. В связи с этим, мы полагаем целесообразным начать исследование использования субъектозависимого подхода, применительно к обнаружению синтезированного голоса, с статистического анализа различий используемых обучающих данных.

Мы предполагаем, что разнообразие дикторов в обучающем наборе данных является запутывающим фактором при обнаружении синтезированного голоса и, следовательно, использование субъектозависимых моделей обнаружения спуфинга может быть более выгодно, по сравнению с использованием субъектонезависимых моделей.

Чтобы проверить наше предположение, мы планируем ответить на следующие вопросы:

- влияет ли присутствие разных дикторов в датасете на характер распределений голосовых признаков подлинных и сфабрикованных данных?
- проще ли различить распределения подлинных и сфабрикованных голосовых признаков одного диктора, по сравнению с соответствующими распределениями множества дикторов?

Заметим, что аналогичное исследование различимости распределений подлинных и сфабрикованных голосовых признаков было проведено для спуфинг-атак повторным воспроизведением звукозаписи [14]. Однако на данный момент отсутствуют научные работы, исследующие каким образом наличие разных дикторов в наборе данных влияет на различимость подлинного и синтезированного голоса.

Используемый набор данных и метод извлечения голосовых признаков

В данном исследовании используется датасет ASVspoof 2019 LA [3], который содержит примеры подлинного и синтезированного голоса. Датасет ASVspoof 2019 LA состоит 3 подмножеств, предназначенных для обучения моделей, настройки гиперпараметров и тестирования. В рамках данного исследования мы используем обучающее подмножество поскольку, оно содержит достаточное количество как подлинных, так и сфабрикованных данных, и хорошо сбалансировано по признаку пола диктора. Обучающее подмножество обладает следующим распределением записей голоса по дикторам (табл.1).

В качестве алгоритма извлечения голосовых признаков используются линейно-частотные кепстральные коэффициенты (LFCC), которые широко применяются для обнаружения синтезированного голоса [15]. При этом, мы использовали параметры извлечения признаков, аналогичные параметрам базовой системы обнаружения спуфинга, представленной организаторами конкурса ASVspoof 2021 [4]:

- количество коэффициентов первого порядка – 20;
- длина окна – 30 мс;
- сдвиг окна – 15 мс;
- вместе с коэффициентами первого порядка извлекаются Δ и $\Delta\Delta$ коэффициенты.

Перед извлечением LFCC выполняется предварительная обработка записи фрагмента речи. Во-первых, запись нормализуется. Во-вторых, тихие кадры записи удаляются при помощи энергетического детектора активности голоса (VAD) с порогом 21 дБ. В-третьих, применяется преэмфасис. Схему извлечения голосовых признаков можно представить следующим образом (рис.1).

Таблица 1

Распределение записей голоса по дикторам в обучающем подмножестве датасета ASVspoof 2019 LA [3]

Пол дикторов	Идентификаторы дикторов	Количество дикторов	Количество подлинных записей для каждого диктора	Количество сфабрикованных записей для каждого диктора
Мужской	LA_0082 LA_0083 LA_0089 LA_0092 LA_0093 LA_0094 LA_0095 LA_0096	8	132	1176
Женский	LA_0079 LA_0080 LA_0081 LA_0084 LA_0085 LA_0086 LA_0087 LA_0088 LA_0090 LA_0091 LA_0097 LA_0098	12	127	1116

Визуализация разнообразия дикторов в пространстве голосовых признаков

Мы используем алгоритм уменьшения размерности t-SNE [16], чтобы спроецировать значения в пространстве голосовых признаков на двухмерную плоскость. Каждая точка соответствует среднему значению LFCC одной обработанной записи голоса диктора. Для удобства восприятия, в данной визуализации используются голоса 10 дикторов: LA_0089-LA_0098.

Проекции голосовых признаков подлинных записей голоса дикторов представлены на (рис.2), а проекции голосовых признаков записей синтезированного голоса – на (рис.3). Точки, принадлежащие одному диктору, обозначены одинаковым цветом.

Представленные рисунки показывают, что голосовые признаки, соответствующие разным дикторам, в значительной степени кластеризованы в пространстве, как в случае подлинных, так и в случае сфабрикованных данных. Кроме того, можно заметить, что кластеры данных, изображённые на (рис.2), выглядят несколько более отчётливо, чем на (рис.3). Данное наблюдение может свидетельствовать о том, что разнообразие дикторов в большей степени влияет на подлинные данные, чем на сфабрикованные.

Рис. 1. Схема извлечения голосовых признаков

Количественная оценка степени кластеризованности данных по признаку диктора

Пусть X – множество голосовых признаков, $P_i(x)$, $P_j(x)$, $P_k(x)$ – распределения голосовых признаков i-го, j-го, k-го дикторов, соответственно. В случае, если данные кластеризованы по признаку диктора, можно ожидать, что среднее значение разницы между распределениями, соответствующими любым двум различным дикторам $P_i(x)$ и $P_j(x)$, будет больше, чем среднее значение разницы между распределением произвольного диктора $P_k(x)$ и универсальным распределением голосовых признаков всех дикторов в наборе данных $P(x)$ [14].

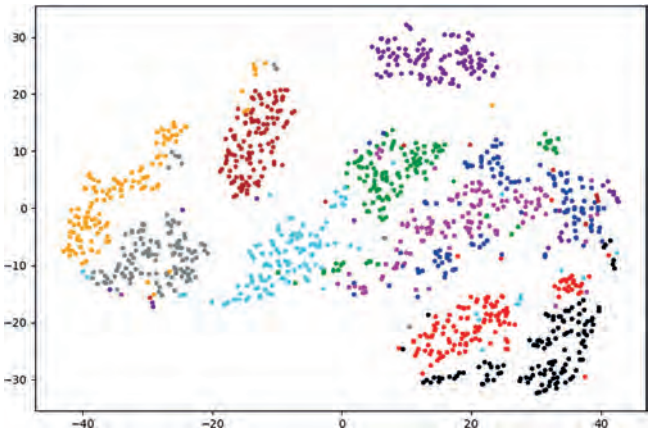

Рис. 2. Проекция средних значений LFCC каждой подлинной аудиозаписи на двухмерную плоскость для дикторов из обучающего подмножества набора данных ASVspoof 2019 LA

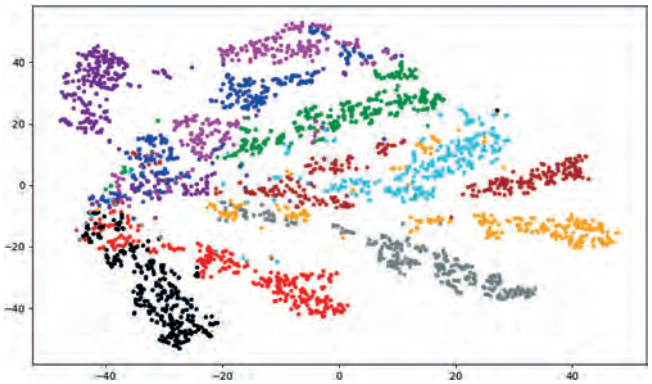

Рис. 3. Проекция средних значений LFCC каждой сфабрикованной аудиозаписи на двухмерную плоскость для дикторов из обучающего подмножества набора данных ASVspoof 2019 LA

Для моделирования распределений вероятности голосовых признаков в данной работе используется модель смеси гауссовых распределений [17] с количеством компонентов равным 512, по аналогии с конфигурацией базовой системы обнаружения спуфинга в конкурсе ASVspoof 2021 [4]. Мы выбрали данную вероятностную модель в связи с тем, что она наиболее широко используется для аппроксимации вероятностных распределений различных голосовых признаков [17].

В качестве множества X выступает пространство возможных значений коэффициентов LFCC. Универсальное распределение голосовых признаков всех дикторов $P(x)$ моделируется путём обучения модели смеси гауссовых распределений при помощи алгоритма максимизации ожидания [17]. Модели смеси гауссовых распределений конкретных дикторов формируются путём MAP-адаптации [17] модели смеси гауссовых распределений, соответствующей универсальному распределению всех дикторов. Обучение моделей на подлинных и сфабрикованных данных происходит раздельно. Схема процесса обучения моделей смеси гауссовых распределений, аппроксимирующих изучаемые распределения голосовых признаков, имеет следующий вид (рис.4).

Для количественной оценки степени различия вероятностных распределений используется расстояние Кульбака-Лейблера [18] между соответствующими моделями смеси гауссовых распределений. Выбор расстояния Кульбака-Лейблера обоснован тем, что это – наиболее часто применяемая числовая характеристика различия между двумя вероятностными распределениями, которая многократно использовалась в различных исследованиях, посвящённых распределениям голосовых признаков [19].

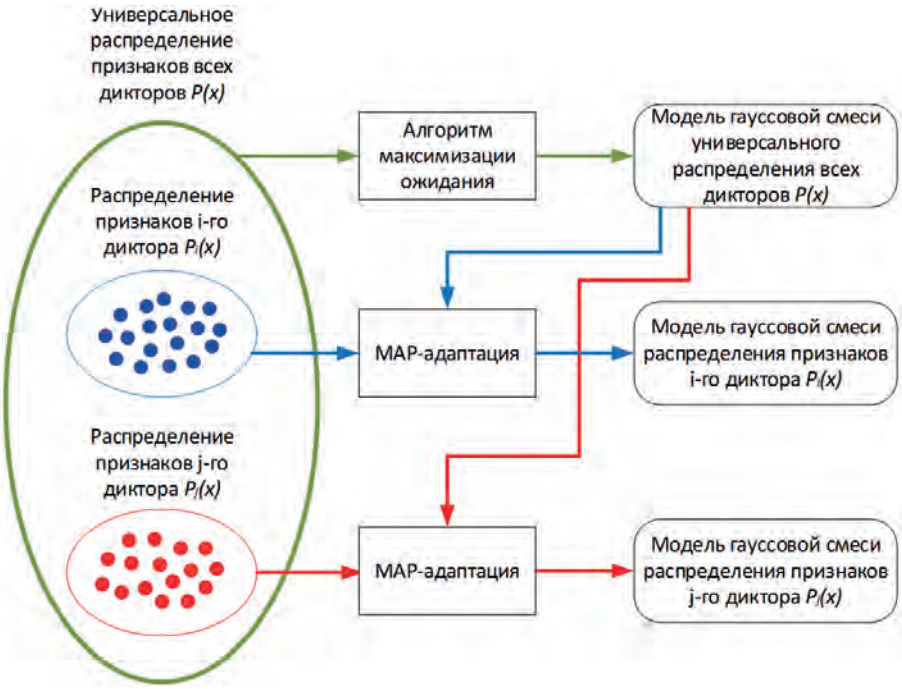

Рис. 4. Схема процесса обучения моделей смеси гауссовых распределений, аппроксимирующих изучаемые распределения голосовых признаков

Кроме того, существуют исследования, в которых расстояние Кульбака-Лейблера применялось в качестве основной метрики, используемой для обнаружения спуфинга [14], что свидетельствует о целесообразности применения данного математического метода в нашем исследовании.

Пусть P_1 и P_2 – абсолютно непрерывные D-мерные вероятностные распределения с функциям плотности вероятности $p_1(x)$ и $p_2(x)$, соответственно, определённые на множестве $X \subseteq \mathbb{R}^D$. Тогда расстояние Кульбака-Лейблера для P_2 относительно P_1 определяется как [18]:

$$KL(P_1,P_2) = \int_X p_1(x) \ln\left(\frac{p_1(x)}{p_2(x)}\right) dx \quad (1)$$

Значение формулы 1 не может быть рассчитано аналитически для моделей смеси гауссовых распределений. В связи с этим, для приблизительного вычисления расстояния Кульбака-Лейблера мы используем метод Монте-Карло [19]. Для расчёта каждого значения по формуле 1 в данной работе мы проводили два миллиона итераций метода Монте-Карло. При этом, среднеквадратическое отклонение значения формулы 1 для оценки расстояния Кульбака-Лейблера между субъектонезависимыми распределениями подлинных и сфабрикованных данных составило 0.0035.

Поскольку расстояние Кульбака-Лейблера – асимметричная мера, то есть $KL(P_1,P_2) \neq KL(P_2,P_1)$, симметричное расстояние Кульбака-Лейблера определяется следующим образом [14]:

$$S_{KL}(P_1,P_2) = \frac{1}{2}(KL(P_1,P_2) + KL(P_2,P_1)) \quad (2)$$

Для оценки различия между распределениями разных дикторов вычисляется среднее симметричное расстояние Кульбака-Лейблера для i–го диктора относительно всех остальных дикторов – $I_{KL}(i)$, которое рассчитывается следующим образом [14]:

$$I_{KL}(i) = \frac{1}{N_{spk} - 1} \sum_{j=1}^{N_{spk}} S_{KL}(P_i,P_j);\ i \neq j \quad (3)$$

где N_{spk} – количество дикторов в наборе данных, а P_i и P_j – распределения голосовых признаков i-го и j-го дикторов, соответственно.

Для оценки различия между распределением признаков i–го диктора и общим распределением голосовых признаков вычисляется симметричное расстояние Кульбака-Лейблера для i–го диктора относительно универсального распределения голосовых признаков – $U_{KL}(i)$, которое рассчитывается следующим образом [14]:

$$U_{KL}(i) = S_{KL}(P_i,P) \quad (4)$$

где P_i – распределение голосовых признаков i-го диктора, а P – универсальное распределение голосовых признаков.

Значения симметричного расстояния Кульбака-Лейблера относительно остальных дикторов (I_{KL}), а также симметричного расстояния Кульбака-Лейблера относительно универсального распределения голосовых признаков (U_{KL}) изображены на (рис.5)

для распределений признаков подлинного голоса и на (рис.6) – для распределений признаков синтезированного голоса. Также на обоих рисунках изображены средние значения данных величин.

В (табл.2) представлены усреднённые по всем дикторам значения среднего симметричного расстояния Кульбака-Лейблера относительно остальных дикторов ($I_{KL_{cp}}$), а также симметричного расстояния Кульбака-Лейблера относительно универсального распределения голосовых признаков ($U_{KL_{cp}}$).

Из (рис.5) и (рис.6) видно, что среднее симметричное расстояние Кульбака-Лейблера относительно остальных дикторов (I_{KL}) превышает симметричное расстояние Кульбака-Лейблера относительно универсального распределения голосовых признаков (U_{KL}) для всех дикторов, как для распределений подлинных голосовых признаков, так и для распределений сфабрикованных голосовых признаков. Данное наблюдение свидетельствует о заметном пространственном разделении голосовых признаков,

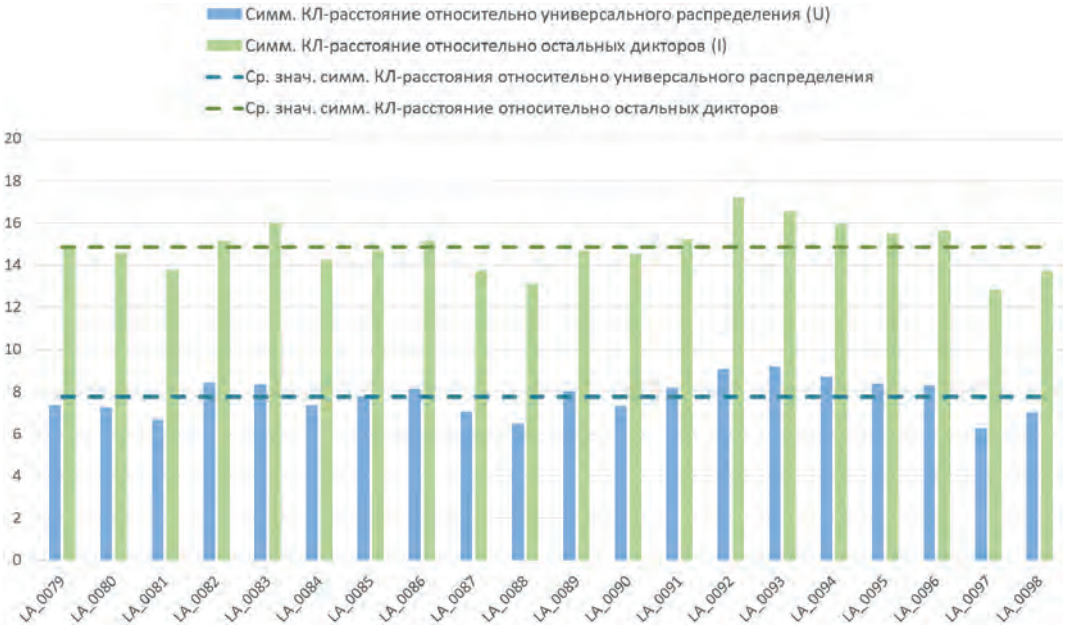

Рис. 5. Сравнение симметричных расстояний Кульбака-Лейблера между распределением подлинных данных диктора и (а) универсальным распределением подлинных данных; (б) распределениями подлинных данных других дикторов

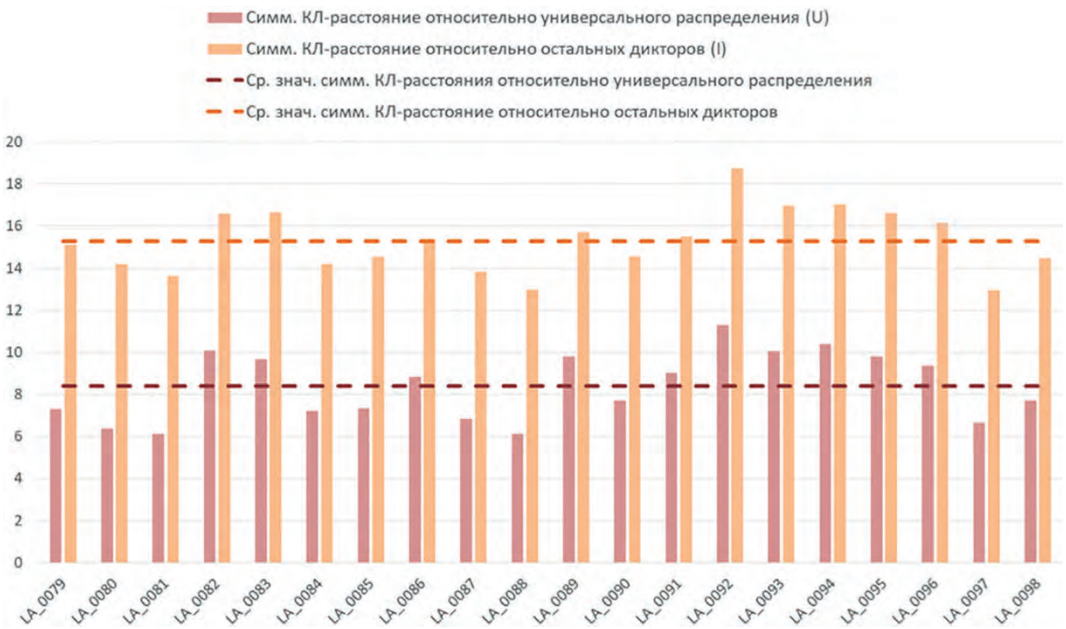

Рис. 6. Сравнение симметричных расстояний Кульбака-Лейблера между распределением сфабрикованных данных диктора и (а) универсальным распределением сфабрикованных данных; (б) распределениями сфабрикованных данных других дикторов

Таблица 2.

Усреднённые по всем дикторам значения I_{KL} и U_{KL}

Значение	Подлинные данные	Спуфинг
I_{KLcp} – усреднённое по всем дикторам среднее симметричное расстояние Кульбака-Лейблера относительно распределений голосовых признаков остальных дикторов	14.87	15.23
U_{KLcp} – усреднённое по всем дикторам симметричное расстояние Кульбака-Лейблера относительно универсального распределения голосовых признаков	7.77	8.39
I_{KLcp} / U_{KLcp} – отношение исследуемых усреднённых значений	1.91	1.81

принадлежащих разным дикторам. Таким образом, кластеризованный характер распределения голосовых признаков в наборе данных, содержащем голоса различных дикторов, визуализированный на (рис. 2) и (рис. 3), подтверждается статистическими методами.

Анализ (табл.2) показывает, что значения отношения характеристик I_{KLcp} и U_{KLcp} различаются незначительно между подлинными и сфабрикованными данными. Таким образом, вопреки наблюдаемой разнице между визуализациями распределений, представленными на (рис. 2) и (рис. 3), в ходе исследования статистическими методами не выявлено различий в степени кластеризации распределений подлинных и сфабрикованных голосовых признаков.

Анализ различимости распределений голосовых признаков подлинного и синтезированного голоса

Результаты, представленные в предыдущих разделах статьи, указывают на то, что распределения как подлинных, так и синтезированных голосовых признаков имеют кластеризованный характер ввиду присутствия разнообразия дикторов в наборе данных. Проверим, позволит ли использование субъектозависимых моделей распределений голосовых признаков упростить обнаружение спуфинга. Для этого проведём количественную оценку различимости подлинных и сфабрикованных распределений голосовых признаков при помощи статистических методов, применяя различные виды моделей.

Для обнаружения спуфинга при помощи моделей смеси гауссовых распределений используется пара моделей, одна из которых соответствует распределению подлинных данных, а вторая – распределению сфабрикованных данных [14]. Рассмотрим пары моделей, которые могут быть использованы для обнаружения спуфинга:

- субъектонезависимая модель распределения подлинных данных и субъектонезависимая модель распределения сфабрикованных данных;
- субъектозависимая модель распределения подлинных данных и субъектонезависимая модель распределения сфабрикованных данных;
- субъектозависимая модель распределения подлинных данных и субъектозависимая модель распределения сфабрикованных данных.

Заметим, что чем лучше пара моделей смеси гауссовых распределений способна функционировать в качестве классификатора, тем существеннее различие между распределениями данных, соответствующими им, и, следовательно, тем большее значение принимает симметричное расстояние Кульбака-Лейблера между ними [14].

В связи с этим, для того, чтобы оценить какая из перечисленных выше пара моделей обладает наибольшей способностью к обнаружению спуфинга, вычислим 3 вида симметричных расстояний Кульбака-Лейблера.

Во-первых, рассчитаем симметричное расстояние Кульбака-Лейблера между универсальным распределением подлинных голосовых признаков и универсальным распределением сфабрикованных голосовых признаков по следующей формуле [14]:

$$DU_{KL} = S_{KL}(P_g, P_s) \quad (5)$$

где P_g – универсальное распределение подлинных признаков, а P_s – универсальное распределение сфабрикованных признаков.

Во-вторых, рассчитаем симметричные расстояния Кульбака-Лейблера между распределением подлинных голосовых признаков i-го диктора и универсальным распределением сфабрикованных голосовых признаков для каждого диктора по следующей формуле [14]:

$$D1_{KL}(i) = S_{KL}(P_i^g, P_s) \quad (6)$$

где P_i^g – распределение подлинных признаков i-го диктора, а P_s – универсальное распределение сфабрикованных признаков.

В-третьих, рассчитаем симметричные расстояния Кульбака-Лейблера между распределениями подлинных и сфабрикованных голосовых признаков i-го диктора для каждого диктора по следующей формуле [14].

$$D2_{KL}(i) = S_{KL}(P_i^g, P_i^s) \quad (7)$$

где P_i^g – распределение подлинных признаков i-го диктора, а P_i^s –распределение сфабрикованных признаков i-го диктора.

Рассчитанные по формулам 5–7 значения представлены на (рис.7).

Из (рис.7) видно, что $D1_{KL}$ и $D2_{KL}$ превышают значение DU_{KL} для всех дикторов. Следовательно, субъектозависимое распределение подлинных признаков проще отличить от распределений сфабрикованных признаков, чем субъектонезависимое. Это подтверждает наше предположение о том, что разнообразие дикторов в обучающем наборе данных является запутывающим фактором при обнаружении спуфинга и, следовательно, использование субъектозависимых моделей обнаружения спуфинга может быть более выгодно, по сравнению с использованием субъектонезависимых моделей.

Мы предполагаем, что $D2_{KL}$ превышает $D1_{KL}$ для всех дикторов с связи с тем, что такие голосовые признаки как LFCC содержат информацию не только об акустических артефактах, которые позволяют обнаружить спуфинг, но и уникальные голосовые признаки, позволяющие распознать личность диктора. Субъектозависимое распределение подлинных LFCC «ближе» к субъектозависимому распределению сфабрикованных LFCC, чем к универсальному, поскольку оба субъектозависимых распределения более «похожи» из-за того, что содержат уникальные голосовые признаки одного диктора.

Выводы

В рамках данного исследования при помощи статистических методов выявлено наличие значительного пространственного разделения голосовых признаков, принадлежащим разным дикторам, что свидетельствует о кластеризованном характере распределения как подлинных, так и сфабрикованных голосовых признаков разных дикторов в наборе данных.

Выявлено, что субъектозависимое распределение голосовых признаков подлинного голоса проще отличить от распределений голосовых признаков синтезированного голоса, чем субъектонезависимое. Это подтверждает наше предположение о том, что разнообразие дикторов в обучающем наборе данных является запутывающим фактором при обнаружении спуфинга и, следовательно, использование субъектозависимых моделей обнаружения синтезированного голоса может быть более выгодно, по сравнению с использованием субъектонезависимых моделей.

Достоверность полученных выводов подтверждается применением релевантных, качественных и широко используемых методов извлечения голосовых признаков, методов моделирования вероятностных распределений, а также методов оценки различия между вероятностными распределениями.

В ходе дальнейших исследований мы планируем оценить в какой степени применение субъектозависимых моделей позволяет увеличить точность обнаружения синтезированного голоса.

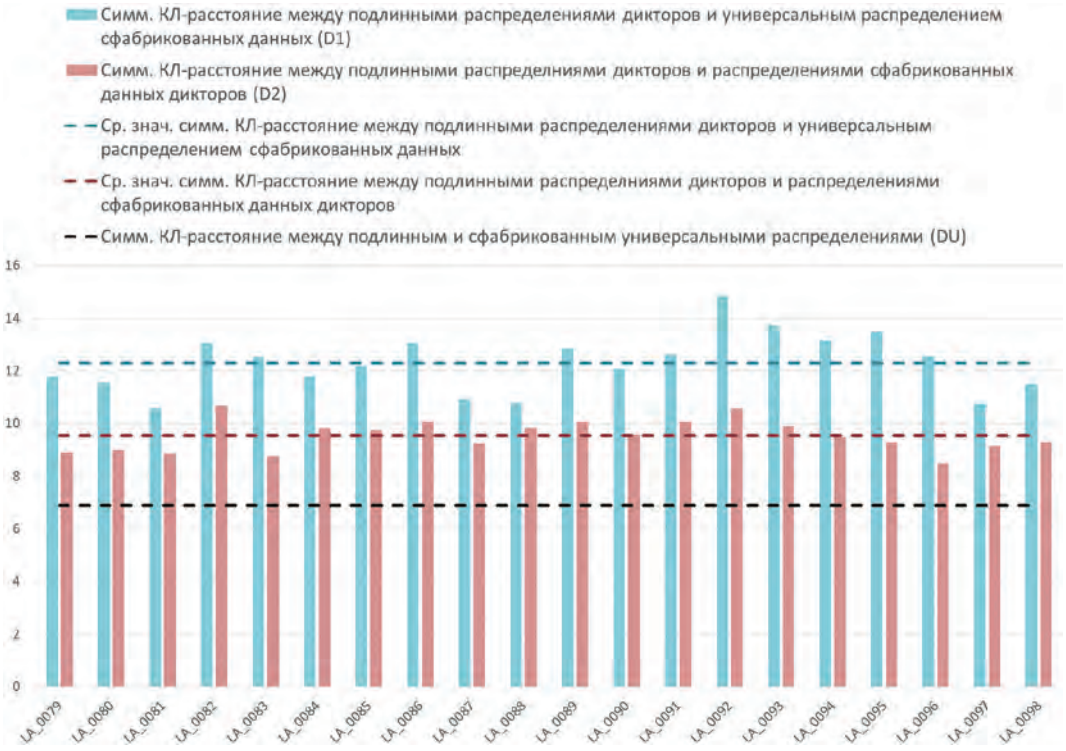

Рис. 7. Сравнение симметричных расстояний Кульбака-Лейблера между распределениями подлинных и сфабрикованных данных

Литература

1. *Evsyukov M., Putyato M., Makaryan A. Methods of protection in speaker verification systems // AIP Conference Proceedings. – 9 March 2023. – Vol. 2700. DOI: 10.1063/5.0137244.*
2. *Evsyukov M. V., Putyato M. M., Makaryan A. S. Antispoofing Countermeasures in Modern Voice Authentication Systems // CEUR Workshop Proceedings. – Yalta, Crimea, 20–22 September 2021. – Vol. 3057. – P. 197–202.*
3. *Nautsch A. et al. ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech // IEEE Transactions on Biometrics, Behavior, and Identity Science. – 2021. – Vol. 3, No. 2. – P. 252–265. DOI: 10.1109/tbiom.2021.3059479.*
4. *Yamagishi J. et al. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection // ASVspoof 2021 Workshop – Automatic Speaker Verification and Spoofing Coutermeasures Challenge. – Virtual, France, September 2021. DOI: 10.21437/asvspoof.2021-8.*
5. *Gunendradasan T., Irtza S., Ambikairajah E., Epps J. Transmission Line Cochlear Model Based AM-FM Features for Replay Attack Detection // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – Brighton, UK, 12–17 May 2019. – P. 6136–6140. DOI: 10.1109/ICASSP.2019.8682771.*
6. *Balamurali B. T., Lin K. W. E., Lui S., Chen J-R., Herremans D. Toward Robust Audio Spoofing Detection: A Detailed Comparison of Traditional and Learned Features // IEEE Access. – 2019. – Vol. 7. – P. 84229–84241. DOI: 10.1109/ACCESS.2019.2923806.*
7. *Lavrentyeva G. et al. STC antispoofing systems for the ASVspoof 2019 challenge // Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech 2019). – Graz, Austria, 15–19 September 2019. – P. 1033–1037. DOI: 10.21437/Interspeech.2019-1768.*
8. *Zhang Y., Jiang F., Duan Z. One-Class Learning Towards Synthetic Voice Spoofing Detection // IEEE Signal Processing Letters. – 2021. – Vol. 28. – P. 937–941. DOI: 10.1109/LSP.2021.3076358.*
9. *Cohen A., Rimon I., Aflalo E., Permuter H. H. A study on data augmentation in voice anti-spoofing // Speech Communication. – 2022. – Vol. 141. – P. 56–67. DOI: 10.1016/j.specom.2022.04.005.*
10. *Teng Z. et al. SA-SASV: An End-to-End Spoof-Aggregated Spoofing-Aware Speaker Verification System // Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech 2022). – Incheon, Korea, 2022. – P. 4391–4395. DOI: 10.21437/interspeech.2022-11029.*
11. *Khan A., Malik K., Ryan J., Saravanan M. Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures // Artificial Intelligence Review. – 2023. – Vol. 56. – P. 1–54. DOI: 10.1007/s10462-023-10539-8.*
12. *Wang X., Yamagishi J. A Practical Guide to Logical Access Voice Presentation Attack Detection // Frontiers in Fake Media Generation and Detection / ed. M. Khosravy, I. Echizen, N. Babaguchi. Singapore: Springer, 2022. – P. 169-214. DOI: 10.1007/978-981-19-1524-6_8.*
13. *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. – P. 107696. DOI: 10.1016/j.patcog.2020.107696.*
14. *Suthokumar G. et al. 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. *Hao B., Hei X. Voice Liveness Detection for Medical Devices // Design and Implementation of Healthcare Biometric Systems / ed. D. R. Kisku, P. Gupta, J. K. Sing. Hershey, USA: IGI Global, 2019. – P.109-136. DOI: 10.4018/978-1-5225-7525-2.ch005.*
16. *Cai T. T., Ma R. Theoretical foundations of t-SNE for visualizing high-dimensional clustered data // The Journal of Machine Learning Research. – 2022. – Vol. 23, No. 1. – P. 13581-13634.*
17. *Kamiński K. A., Dobrowolski A. P. Automatic Speaker Recognition System Based on Gaussian Mixture Models, Cepstral Analysis, and Genetic Selection of Distinctive Features // Sensors. – 2022. – Vol. 22, No. 23. – P. 9370. DOI: 10.3390/s22239370.*
18. *Bulinski A., Dimitrov D. Statistical estimation of the Kullback–Leibler divergence // Mathematics. – 2021. – Vol. 9, No. 5. – P. 1–36. DOI: 10.3390/math9050544.*
19. *Hansen J. H., Bokshi M., Khorram S. Speech variability: A cross-language study on acoustic variations of speaking versus untrained singing // The Journal of the Acoustical Society of America. – 2020. – Vol. 148, No. 2. – P. 829–844. DOI:10.1121/10.0001526.*