Инженерный вестник Дона, №8 (2025)
ivdon.ru/ru/magazine/archive/n8y2025/10326

Методы обнаружения ложных голосовых сигналов

В.Ц. Дашинимаева, А.Е. Боршевников

Дальневосточный Федеральный Университет, Владивосток

Аннотация: В статье анализируются различные подходы к генерации и обнаружению аудиодипфейков. Особое внимание уделяется предобработке акустических сигналов, выделению параметров голосовых сигналов и классификации данных. В работе рассмотрены три группы классификаторов: метод опорных векторов (Support Vector Machines – SVM), метод K-ближайших соседей (K-Nearest Neighbors – KNN) и нейронные сети. Для каждой группы были выделены эффективные методы, и на основе общего анализа определены наиболее результативные подходы. В ходе работы было выявлено два подхода, демонстрирующих высокую точность и надежность: детектор на основе временных сверточных сетей (Temporal Convolutional Network – TCN), анализирующий кепстрограмму мел-частотных кепстральных коэффициентов (Mel-Frequency Cepstral Coefficients – MFCC), показал равный уровнь ошибок (Equal Error Rate – EER) 0.07%, тогда как SVM с ядром в виде радиальной базисной функции (Radial Basis Function – RBF) достиг EER = 0.5%, а также показал следующие значения метрик на наборе данных ASVspoof 2021: точность (accuracy) = 99.6%, F1-score = 0.997, Precision = 0.998 и полнота (recall) = 0.994.

Ключевые слова: аудиодипфейк, предобработка акустических сигналов, метод опорных векторов, метод K-ближайших соседей, нейронные сети, временные сверточные сети, обнаружение дипфейков.

Введение

Современные технологии голосовой идентификации и аутентификации играют ключевую роль в повседневной жизни человека. Они используются в широком спектре приложений — от голосовых помощников до систем биометрической аутентификации в мобильных устройствах, банках и системах контроля доступа.

Голосовая аутентификация стала простым и удобным способом подтверждения личности. Однако, наряду с преимуществами, она привлекла внимание злоумышленников, эксплуатирующих её уязвимости. Одной из ключевых угроз стало использование ложных голосовых сигналов — искусственно созданных речевых данных с помощью технологий синтеза речи, имитации или модификации реальных записей.

Современные технологии позволяют генерировать естественную речь, практически неотличимую от реальной. Безусловно, развитие этих

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

технологий значительно повышает удобство нашей жизни во многих сферах. Однако они также представляют серьёзную угрозу для общественной безопасности и политико-экономической стабильности, если кто-то злоупотребит так называемыми дипфейк-технологиями в преступных целях [1]. Например, поддельные сигналы могут применяться для:

- несанкционированного доступа к конфиденциальным данным;
- мошеннических операций в банковской сфере;
- обхода систем контроля доступа;
- манипуляции голосовыми помощниками (например, выполнение вредоносных команд).

Проблематикой исследования является то, что стремительное развитие методов генерации дипфейков привело к появлению множества разнородных подходов к их детектированию. Это требует системного анализа и сравнения их эффективности. В частности, за последние годы предложены десятки методов, основанных на анализе спектральных и временных характеристик сигнала, а также выявлении возможных артефактов синтеза.

В данной статье проводится обзор и сравнительный анализ современных методов детектирования ложных голосовых сигналов, в результате которого определятся наиболее эффективные подходы для решения указанной задачи.

Метрики оценки эффективности методов обнаружения ложных голосовых сигналов

Несмотря на разнообразие методов детектирования ложных голосовых сигналов, ключевое значение для их практического использования имеет точная оценка эффективности. Для сравнения различных подходов исследователи используют стандартизированные метрики, позволяющие количественно измерить точность, устойчивость к ошибкам и общую

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

надежность системы. В рамках проведенного анализа были определены следующие основные метрики, используемые в исследованиях:

- Равный уровень ошибок (Equal Error Rate – EER) — это коэффициент, при котором уровень ошибки второго рода или ошибка ложного принятия (False Acceptance Rate – FAR) и ошибки первого рода или ошибка ложного отклонения (False Rejection Rate – FRR) эквивалентны:

$$\mathrm{EER} = \mathrm{FAR} = \mathrm{FRR}. \quad (1)$$

- Функция тандемной оценки стоимости обнаружения (tandem detection cost function – t-DCF) — это метрика для оценки эффективности систем обнаружения поддельного голоса, особенно в контексте систем проверки голоса. Низкие значения t-DCF свидетельствуют о высокой точности системы.
- Точность (Accuracy) — это доля правильно классифицированных образцов относительно общего числа образцов:

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}, \quad (2)$$

где TP (True Positive) — количество правильно классифицированных образцов, TN (True Negative) — количество правильно отклоненных образцов, FP (False Positives) — количество неправильно классифицированных образцов, FN (False Negatives) — количество неправильно отклоненных образцов.

- Precision — доля правильно классифицированных образцов среди всех объектов, которые алгоритм классифицировал:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \quad (3)$$

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

- Полнота (Recall) — доля правильно классифицированных образцов среди всех образцов, которые должны быть классифицированы алгоритмом:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (4)$$

- F1-score — это гармоническое среднее метрик Precision и Recall [2]:

$$\text{F1} - \text{score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (5)$$

- Площадь под кривой (Area Under the Curve – AUC) — это площадь под кривой ошибок (Receiver Operating Characteristic – ROC), которая используется для оценки качества работы классификатора в задачах бинарной классификации. ROC-кривая показывает соотношение истинно положительных результатов (совокупности значений TP и TN) и ложноположительных результатов (совокупности значений FP и FN) при различных пороговых значениях [3].

Рассмотренные метрики чаще всего применяются исследователями для оценки качества алгоритмов машинного обучения, поэтому в дальнейшем в работе анализируемые методы будут оценены согласно указанных метрик.

Методы обнаружения голосовых дипфейков

Развитие технологий генерации синтетической речи требует создания надежных методов детектирования голосовых дипфейков. В настоящее время для решения этой задачи применяются различные подходы машинного

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

обучения, включая классические алгоритмы (такие как метод опорных векторов (Support Vector Machines – SVM) и K-ближайших соседей (K-Nearest Neighbors – KNN)), а также современные нейросетевые архитектуры. Каждый из этих методов обладает своими преимуществами и ограничениями.

Метод опорных векторов– это линейный алгоритм, используемый в задачах классификации и регрессии. Данный алгоритм имеет широкое применение и может решать как линейные, так и нелинейные задачи. Суть работы метода опорных векторов проста: алгоритм создает линию или гиперплоскость, которая разделяет данные на классы [4]. В контексте задач обнаружения поддельных аудиозаписей SVM часто используется в сочетании с методами извлечения признаков, такими как мел-частотные кепстральные коэффициенты (Mel-Frequency Cepstral Coefficients – MFCC), и другими классификаторами, что делает его универсальным инструментом для анализа аудиоданных.

Авторы исследования [5] предложили использовать гибридную нейросетевую архитектуру лёгкой сверточной рекуррентной нейронной сети (Light Convolutional Gated Recurrent Neural Network – LC-GRNN) в качестве метода предварительной обработки данных, после чего в работе производилось исследование по использованию данных предобработки для анализа классификаторами: SVM, линейный дискриминантный анализ (Linear Discriminant Analysis – LDA) и вероятностный линейный дискриминантный анализ (Probability Linear Discriminant Analysis –PLDA).

На этапе предобработки речевые сигналы сегментируются с применением окна Блэкмана длиной 16 мс и шагом 4 мс, после чего из них извлекаются логарифмические спектрограммы с 256 частотными бинами, которые служат входными данными для нейронной сети. После извлечения

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

признаков нейронная сеть объединяет их в вектор идентификации, который затем подается на классификатор.

Система прошла тестирование на различных наборах данных серии ASVspoof, однако представленные результаты и дальнейший анализ сосредоточены на наиболее актуальной версии — ASVspoof 2019 LA. Этот набор данных предназначен для оценки уязвимостей систем автоматической верификации голоса (Automatic Speaker Verification – ASV) к спуфинг-атакам с использованием синтезированной (Text-to-Speech – TTS) и преобразованной (Voice Conversion – VC) речи в условиях логического доступа (прямая передача аудиосигнала в систему).

В рамках логического доступа (Logical Access – LA) система LC-GRNN с классификатором LDA достигла min-tDCF 0.1523, а EER составило 6.28%. Модель LC-GRNN с PLDA показала min-tDCF 0.1552 и EER 6.34%, тогда как LC-GRNN и SVM продемонстрировала min-tDCF 0.1873 и EER 7.12%.

Авторами статьи [6] были предложены способы распознавания речи с помощью кепстральной и биспектральной статистики. Для этого были использованы алгоритмы линейного метода опорных векторов (Linear SVM) и квадратичного метода опорных векторов (Quadratic SVM – Q-SVM).

Linear SVM строит линейную разделяющую гиперплоскость, максимизирующую зазор между классами, и эффективен для линейно разделимых данных. В отличие от него, Q-SVM применяет квадратичное ядро, которое преобразует данные в пространство более высокой размерности, учитывая не только линейные, но и квадратичные зависимости между признаками.

Основное преимущество квадратичного метода опорных векторов заключается именно в этой особенности: способность работать с нелинейно разделимыми данными, например, когда классы нельзя разделить простой

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

гиперплоскостью. Это делает метод особенно полезным для сложных речевых паттернов, где зависимости часто имеют нелинейный характер.

Методика работы включает сбор речевых образцов, предобработку, извлечение биспектральных и кепстральных характеристик (MFCC, Δ-Кепстр и Δ^2-Кепстр), а также классификацию. В процессе предобработки каждый речевой образец обрезается до средней длины 5 секунд. Аудиоданные также приводятся к монофоническому формату, если они изначально были стерео. Для выделения параметров авторы используют несколько ключевых методов. Звуковой сигнал y(k) разбивается на частоты с помощью преобразования Фурье.

После того как было получено частотное представление сигнала, следующий шаг — анализ спектра мощности. Спектр мощности показывает, насколько сильна каждая частота в сигнале. Данный параметр может быть определен по формуле:

$$P(\omega) = Y(\omega) \cdot Y^*(\omega), \quad (6)$$

где $Y^*(\omega)$ — это комплексно-сопряженное значение $Y(\omega)$.

Чтобы изучить более сложные взаимодействия между частотами, используется биспектр. Биспектральный анализ позволяет выявить высокоуровневые корреляции в частотной области. Биспектр сигнала может быть определен:

$$B(\omega_1, \omega_2) = Y(\omega_1) \cdot Y(\omega_2) \cdot Y^*(\omega_1 + \omega_2). \quad (7)$$

Чтобы упростить расчеты и сделать результаты более интерпретируемыми, используется нормализованный биспектр, также известный как бикогерентность:

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

$$B_c(\omega_1, \omega_2) = \frac{Y(\omega_1) \cdot Y(\omega_2) \cdot Y^*(\omega_1 + \omega_2)}{\sqrt{|Y(\omega_1) \cdot Y(\omega_2)|^2 \cdot |Y(\omega_1 + \omega_2)|^2}}. \quad (8)$$

Бикогерентность всегда находится в диапазоне от 0 до 1, что позволяет легко сравнивать разные звуковые сигналы. Значение, близкое к 1, указывает на сильное взаимодействие между частотами, а значение, близкое к 0, — на отсутствие такого взаимодействия.

Звуковой сигнал разбивается на маленькие отрезки, называемые окнами. Для каждого окна рассчитывается биспектр, а затем все результаты усредняются, чтобы получить общую картину. Это позволяет выявить скрытые закономерности в звуке, которые не видны при обычном анализе спектра мощности. Кроме того, для анализа звука используются мел-частотные кепстральные коэффициенты. MFCC представляют кратковременный спектр мощности аудио и используются для анализа речи на основе голосового тракта. Поскольку синтезированная речь искусственного интеллекта (ИИ) не создается из голосовых трактов, значения MFCC для человеческой и синтезированной речи различаются.

К спектру мощности применяются мел-частотные фильтры, которые имитируют восприятие звука человеком. После фильтрации выполняется дискретное косинусное преобразование (Discrete Cosine Transform – DCT) мел-логарифмического спектра мощности, чтобы выделить основные характеристики звука. MFCC представляют амплитуду полученного спектра после выполнения DCT.

Также дополнительно используются параметры для учета изменений звука во времени:

- Δ-Кепстр — разница между текущим и предыдущим значением MFCC.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

- Δ^2-Кепстр — разница между текущим и предыдущим значением Δ-Кепстр.

Модель была протестирована на данных, включающих человеческую речь и синтезированную речь от различных систем (Natural Reader, Spik.AI, Replica AI). На тестовых данных значение ROC AUC для квадратичного метода опорных векторов составило 0.84, используя только величину и фазу бикогерентности в качестве признака для классификации, и 0.99 с добавлением MFCC, Δ-Кепстр и Δ^2-Кепстр. Для линейного метода опорных векторов с величиной и фазой бикогерентности результат составил 0.80, а с добавлением MFCC, Δ-Кепстр и Δ^2-Кепстр 0.98.

Авторы статьи [7] предлагают подход, основанный на анализе аудиохарактеристик и применении различных классификаторов, включая SVM, для бинарной классификации речи на реальную и синтетическую. В работе представлен новый набор данных DEEP-VOICE, который включает реальную речь восьми известных личностей и их синтетические версии. Общая длительность записей составила 62 минуты 22 секунды, причём аудиодорожки были обрезаны до максимум 10 минут для обеспечения единообразия данных.

Для предобработки использовалась модель Spleeter, которая разделяет аудио на два основных компонента: вокалы и аккомпанемент (фоновые звуки). Это важно для того, чтобы сохранить фоновые звуки, такие как аплодисменты или шум, при преобразовании голоса. Разделение позволяет изолировать голос говорящего от фонового шума, что упрощает дальнейшее преобразование и анализ. Вокалы, извлечённые с помощью Spleeter, затем преобразовывались с использованием модели голосового преобразования на основе поиска (Retrieval-based Voice Conversion – RVC). Эта модель позволяет преобразовать голос одного человека в голос другого, сохраняя при этом интонацию и стиль речи. После преобразования вокалы

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

объединяются с оригинальным аккомпанементом, чтобы создать синтетическую аудиодорожку, которая звучит естественно, но при этом является искусственно созданной.

Для анализа аудио разбивается на короткие временные блоки, обычно длительностью 1 секунда. Это позволяет извлекать характеристики для каждого блока и проводить более детальный анализ. Сегментация также помогает уменьшить вычислительную сложность и упрощает обработку данных в реальном времени. Авторы извлекают параметры для каждого предобработанного 1-секундного отрезка аудиосигнала. Всего извлекается 26 признаков. Эти признаки включают хроматограмму (chromagram), которая может быть рассчитана на основе кратковременного преобразования Фурье X сигнала на кадре m и частотном интервале k:

$$SG(m,k) = |X(m,k)|. \tag{9}$$

После чего выполняется нормализация хроматических полос, представляющих собой частотные интервалы спектрограммы. На основе этого определяется спектральный центроид (Spectral Centroid – SC), который представляет собой положение центра масс в спектре:

$$SC(m) = \frac{\sum_k k \cdot SG(m,k)}{\sum_k SG(m,k)}. \tag{10}$$

Аналогично определяется спектральная ширина полосы (Spectral Bandwidth – SB), которая представляет собой разницу частот вокруг центроида:

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

$$SB(m) = \sqrt{\frac{\sum_k \left(k - SC(m)\right)^2 \cdot SG(m,k)}{\sum_k SG(m,k)}}. \quad (11)$$

Спектральный спад (Spectral Rolloff – SR) рассчитывается по частоте, составляющей менее 85% от общей спектральной энергии. Согласно спектральным характеристикам, скорость пересечения нуля (Zero-Crossing Rate – ZCR) измеряется путем наблюдения за скоростью, с которой сигнал (содержащий N отсчетов) меняет знак с помощью функции sgn. То есть, как часто сигнал пересекает x=0 и меняется с положительного на отрицательный или наоборот:

$$ZCR(m) = \frac{1}{2N} \sum_{n=1}^{N-1} |sgn(x[n]) - sgn(x[n-1])|. \quad (12)$$

Затем определяется среднеквадратичное значение (Root Mean Square – RMS) для аудиосигнала $x[n]$ на кадре m:

$$RMS(m) = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} x[n]^2}. \quad (13)$$

Наконец, вычисляются первые 20 MFCC. Мощности кратковременного преобразования Фурье (Short-time Fourier Transform – STFT) отображаются на мел-шкалу с помощью треугольного окна. Для каждого логарифма берется спектр мощности, и применяется DCT.

Учитывая, что каждый человек используется для создания семи поддельных треков, данные крайне несбалансированы. По этой причине

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

данные балансируются в соотношении 1:1 между реальной и поддельной речью.

В рамках исследования SVM достиг точности 72.3% и ROC AUC 0.723, а значение F1-score – 0.675. Время вывода для метода опорных векторов составило 0.605 миллисекунды для классификации 1 секунды аудио.

Для детектирования поддельного аудио авторы работы [8] используют извлечение признаков с применением MFCC для получения спектральных характеристик аудио и классификацию с использованием различных классификаторов, включая SVM, свёрточные нейронные сети (Convolutional Neural Networks – CNN), градиентный бустинг (Gradient Boosting) и другие. Чтобы извлечь ключевые звуковые характеристики также применяется стандартный масштабатор (Standard Scaler) для нормализации извлеченных признаков, что позволяет уменьшить влияние вариаций амплитуды и интенсивности в различных аудио образцах.

В целях повышения качества и устойчивости модели исследователи применяют комплексный подход к предобработке звука, включающий исследовательский анализ данных (Exploratory Data Analysis – EDA) для визуализации распределения данных и выявления закономерностей, а также аугментацию данных, такую как добавление шума, изменение высоты тона, растяжение и сдвиг аудиозаписей. Эти методы помогают модели лучше адаптироваться к реальным условиям и улучшают её обобщающую способность.

В качестве эталонного набора данных используется Fake or Real (FoR) dataset, который включает поднаборы данных с разной длительностью аудио и скоростью передачи данных:

- FOR-REREC DATASET: набор данных специально собран для обнаружения глубокой подделки звука. Вероятно, он содержит записи,

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

которые были перезаписаны или изменены для симуляции различных акустических сред или условий.

- FOR-2SEC: все аудио примеры в наборе данных обработаны для стандартизированной длительности в 2 секунды. Такая единообразная длительность обеспечивает согласованность в извлечении признаков и обучении модели.
- FOR-NORM: процесс извлечения признаков включает использование MFCC для обработки звуковых сигналов. Извлеченные признаки MFCC нормализуются, чтобы привести их к общей шкале, уменьшая влияние вариаций амплитуды и интенсивности в различных аудио образцах.
- FOR-ORIGINAL: эта категория относится к подмножеству набора данных, содержащему оригинальные, неизмененные аудиозаписи. Эти образцы служат базовыми или подлинными экземплярами, с которыми можно сравнивать обработанные или подделанные аудиозаписи. Включение оригинальных данных крайне важно для обучения модели машинного обучения, чтобы она могла отличать подлинные записи от поддельных.

Исследование показывает, что метод опорных векторов демонстрирует хорошие результаты: на наборе FOR-REREC точность составляет 0.866, Precision – 0.871, Recall – 0.866, F1 – 0.866; на наборе FOR-2SEC точность – 0.869, Precision – 0.872, Recall – 0.869, F1 – 0.869; на наборе FOR-NORM точность – 0.866, Precision – 0.871, Recall – 0.866, F1 – 0.866; на наборе FOR-ORIGINAL точность – 0.866, Precision – 0.871, Recall – 0.866, F1 – 0.866.

В статье [9] рассматривается новый подход к обнаружению аудио-дипфейков с использованием легковесной техники извлечения признаков. Авторы предлагают метод, который сочетает модифицированную архитектуру ResNet50 с линейным дискриминантным анализом для

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

повышения точности и эффективности обнаружения поддельных аудиозаписей.

В исследовании особое внимание уделяется использованию SVM в качестве классификатора. В данной работе метод опорных векторов применяется с ядром в виде радиальной базисной функции (Radial Basis Function – RBF), что позволяет эффективно работать с нелинейными данными.

Для предобработки звука используются мел-спектрограммы, которые представляют собой визуальное отображение спектра частот аудиосигнала. мел-спектрограмма строится с использованием нелинейной мел-шкалы, которая лучше соответствует восприятию звука человеком. Формула для преобразования частоты в мел-шкалу:

$$m = 2595 \cdot log_{10}\left(1 + \frac{f}{700}\right), \quad (14)$$

где m — значение в мел-шкале, а f — частота в герцах.

Аудиосигнал разбивается на кадры, к каждому из которых применяется дискретное преобразование Фурье (Discrete Fourier Transform – DFT) для перехода из временной области в частотную. Затем спектр мощности преобразуется в децибелы, что позволяет получить мел-спектрограмму.

Для извлечения признаков из аудио авторы исследования используют модифицированную ResNet50. Остаточная нейронная сеть (Residual Network – ResNet) — это предобученная модель CNN, которая принимает изображение в качестве входных данных для извлечения признаков. ResNet50 возвращает 2048 глубоких признаков для каждого аудио. Для уменьшения матрицы признаков применяется линейный дискриминантный анализ, позволяющий вернуть один вектор признаков для всех примеров аудиозаписей.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

Модель обучается на наборе данных ASVspoof 2019 LA, а тестируется на наборе данных ASVspoof 2021 Deepfake. Для оценки устойчивости модели к шуму используется набор данных DECRO, в который добавлен белый шум с уровнем 10 дБ (по метрике отношения сигнал-шум). Отношение сигнал-шум (Signal-to-noise ratios – SNR) — это метрика, характеризующая соотношение мощности полезного сигнала к мощности фонового шума.

При использовании SVM с ядром RBF модель показала лучшие результаты среди всех классификаторов, достигнув EER 0.5% и точности 99.6% на наборе данных ASVspoof 2021. На зашумленном наборе данных DECRO модель показала EER 1% при уровне шума 10 дБ.

Для детектирования аудио-дипфейков в работе [10] авторы предлагают два подхода: классификация на основе признаков и классификация на основе изображений. В первом подходе аудиофайлы анализируются для извлечения спектральных признаков, которые затем используются в моделях машинного обучения (например, метод опорных векторов). Во втором подходе аудио преобразуется в мел-спектрограммы, которые используются в качестве входных данных для моделей глубокого обучения.

Для предобработки данных использовали библиотеку librosa. С помощью нее отдельный аудиофайл загружается в качестве входных данных. Это дает временной ряд и частоту дискретизации, которые представляют цифровую форму аудиофайла. Функции библиотеки librosa также применялись для извлечения признаков аудиозаписей. Для каждого признака вычисляется среднее значение массива, что дает итоговое значение признака для данного аудиофайла. Эти признаки затем передаются в алгоритмы машинного обучения для классификации аудиозаписей как реальных или поддельных.

Каждый аудиообразец в наборе данных представлен в виде вектора из 37 признаков. Это было сделано с использованием метода скользящего окна,

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

поэтому признаки являются временными последовательностями. Извлеченные признаки включают: среднюю квадратичную энергия (Root Mean Square Energy – RMSE); хроматические признаки (Chroma Features), которые представляют собой частотный спектр в виде 12 классов высоты тона; спектральный центроид; спектральную ширину; спектральный спад; частота пересечения нуля и мел-частотные кепстральные коэффициенты (в рассматриваемой работе было рассмотрено 20 MFCC).

Аудиофайлы преобразуются в спектральные признаки или мел-спектрограммы. Для классификации на основе признаков используются методы: SVM, случайный лес (Random Forest – RF), KNN, экстремальный градиентный бустинг (eXtreme Gradient Boosting – XGBoost) и легкий градиентный бустинг (LightGBM – LGBM). Для классификации на основе изображений используются временная сверточная сеть (Temporal Convolutional Network – TCN) и пространственная трансформаторная сеть (Spatial Transformer Network – STN). Результаты показывают, что классификация на основе признаков с использованием метода опорных векторов достигает точности 67%, Precision – 0.69, Recall – 0.67, F-score – 0.65.

Для наглядного сопоставления рассмотренных подходов к детектированию аудиодипфейков на основе метода опорных векторов, ключевые характеристики их эффективности систематизированы в Таблице 1.

Таблица № 1.

Эффективность SVM в классификации аудиодипфейков

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

Исследование	Классификатор	Предобработка	Выделение параметров	EER (%)	min-tDCF	ROC AUC	Accuracy (%)	F1-score	Precision	Recall
1	2	3	4	5	6	7	8	9	10	11
5	SVM	Сегментация с окном Блэкмана, логарифмические спектрограммы	LC-GRNN	7.12	0.187	-	-	-	-	-
6	Q-SVM	Обрезка до 5 сек, приведение к моно	Величина и фаза бикогерентности	-	-	0.84	-	-	-	-
	Linear SVM			-	-	0.80	-	-	-	-
	Q-SVM		**Величина и фаза бикогерентности+MFCC & Δ и Δ^2Кепстр**	**-**	**-**	**0.99**	**-**	**-**	**-**	**-**
	Linear SVM			-	-	0.98	-	-	-	-
1	2	3	4	5	6	7	8	9	10	11
7	SVM	Spleeter, RVC, сегментация на 1-секундные блоки, балансировка данных.	Chromagram, SC, SB, SR, ZCR, RMS, 20 MFCC	-	-	0.72	72.3	0.675	0.815	0.576
8	SVM (for-rerec)	EDA, аугментация, (Standard Scaler)	MFCC	-	-	-	82.1	0.820	0.833	0.821
	SVM (for-2sec)			-	-	-	86.6	0.866	0.871	0.866
	SVM (for-norm)			-	-	-	86.9	0.869	0.872	0.869
	SVM (for-original)			-	-	-	86.6	0.866	0.871	0.866
9	**SVM (RBF)**	**мел-спектрограммы (преобразование в мел-шкалу, DFT, спектр мощности в dB)**	**Модифицированная ResNet50 + LDA**	**0.5**	**-**	**-**	**99.6**	**0.997**	**0.998**	**0.994**
10	SVM	Librosa (нормализация, загрузка аудио)	RMSE, Chroma, SC, SB, SR, ZCR, 20 MFCC	-	-	-	67	0.65	0.69	0.67

На основе анализа данных, проведенных в Таблице 1, можно сделать вывод, что наиболее эффективными методами классификации являются квадратичный метод опорных векторов с использованием величины и фазы бикогерентности, а также MFCC и их производные (Δ и Δ² Кепстр) [6], и

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

метод опорных векторов с ядром радиальной базисной функции из исследования [9].

Q-SVM демонстрирует наивысшее качество работы, о чем свидетельствует значение ROC AUC, равное 0.99, что близко к идеальному показателю и указывает на исключительную способность модели разделять классы. Использование комбинации признаков, таких как бикогерентность и MFCC, позволяет классификатору эффективно улавливать сложные закономерности в данных, что делает его подходящим для задач, требующих высокой точности и надежности.

В свою очередь, SVM (RBF) демонстрирует выдающиеся результаты по всем ключевым метрикам: точность достигает 99.6%, F1-score составляет 0.997, Precision – 0.998, а Recall – 0.994. Кроме того, показатель EER равен всего 0.5%, что свидетельствует о высокой надежности модели и минимальном уровне ошибок. Использование RBF в качестве ядра позволяет модели эффективно разделять данные даже в сложных нелинейных случаях, что делает её универсальным и мощным инструментом для задач классификации.

Таким образом, оба метода — Q-SVM с комбинированными признаками и SVM с ядром RBF — можно рекомендовать как наиболее эффективные для задач классификации среди предложенных методов опорных векторов, связанных с обработкой сигналов, аудиоаналитикой и другими областями, где требуется высокая точность и надежность предсказаний. Выбор между ними может зависеть от специфики задачи и предпочтений по используемым метрикам.

Несмотря на выдающиеся результаты, показанные методами на основе SVM, в задачах детектирования аудиодипфейков также активно применяются и другие алгоритмы машинного обучения, в том числе более простые и интерпретируемые. Одним из таких методов является K-ближайших соседей,

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

который, хотя и уступает в сложности методу опорных векторов с радиальным ядром или квадратичному методу опорных векторов, остается популярным благодаря своей прозрачности и устойчивости в условиях ограниченного объема данных.

Метод K-ближайших соседей — это один из базовых алгоритмов машинного обучения, основанный на принципе близости объектов в пространстве признаков. Его ключевая идея заключается в классификации объекта на основе классов его ближайших соседей в обучающей выборке. KNN часто используется в задачах бинарной классификации, таких как обнаружение синтезированной речи или аудиодипфейков, благодаря своей простоте и интерпретируемости.

Авторы работы [6] рассматривают задачу обнаружения синтезированной речи с использованием методов машинного обучения. В статье используются несколько классификаторов, включая метод K-ближайших соседей.

Взвешенный KNN (Weighted KNN) учитывает расстояния между объектами и присваивает больший вес ближайшим соседям, что улучшает точность классификации. Взвешенный KNN показывает хорошую эффективность при использовании только биспектральных признаков (AUC = 0.82), что свидетельствует о способности модели различать человеческую и синтезированную речь на основе взаимодействия частот. При добавлении кепстральных коэффициентов эффективность модели значительно улучшается, достигая AUC = 0.99. Это указывает на то, что комбинация биспектральных и кепстральных признаков позволяет более точно классифицировать аудиозаписи.

В исследовании [7] для бинарной классификации речи на реальную и синтетическую в качестве классификатора также был использован метод K-ближайших соседей. Он показал точность 81.5% и ROC AUC 0.815, а

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

значение F1-score – 0.817. Время вывода для KNN составило 0.143 миллисекунды для классификации 1 секунды аудио.

Авторы статьи [9] в качестве классификатора помимо метода опорных векторов используют метод K-ближайших соседей. Он показал хорошие результаты среди всех классификаторов, достигнув EER 0.7% и точности 99.5% на "чистых" данных и EER 1.7% и точности 98% на зашумленных данных. Точность и устойчивость метода K-ближайших соседей к шуму несколько уступают методу опорных векторов, однако он может быть полезен в случаях, где важна простота реализации и интерпретации.

В работе [10] для обнаружения аудиодипфейков в классификации на основе признаков результаты показывают, что классификация с использованием KNN достигает Accuracy – 62%, Precision – 0.61, Recall – 0.61, F1-score – 0.61.

В статье [11] рассматривается новый подход к обнаружению аудиодипфейков, основанный на комбинации MFCC и признаков спектрального контраста. Для классификации авторы используют несколько моделей машинного обучения, включая искусственную нейронную сеть (Artificial Neural Network – ANN), метод K-ближайших соседей и линейную регрессию. KNN, в частности, показал себя как эффективный классификатор благодаря своей способности работать с многомерными данными и высокой точности в задачах классификации.

При выделении параметров авторы используют несколько ключевых шагов. Аудиосигнал сначала разбивается на короткие перекрывающиеся кадры, после чего применяется оконная функция Хэмминга для уменьшения эффектов краев кадров. Для каждого кадра выполняется быстрое преобразования Фурье (Fast Fourier Transform – FFT) для получения спектра мощности.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

Спектр мощности преобразуется в мел-шкалу с использованием набора мел-фильтров, что позволяет лучше отразить восприятие звука человеком. Далее применяется логарифмическое сжатие для усиления важных частотных компонентов, и, наконец, выполняется DCT для получения MFCC-коэффициентов. Помимо MFCC, авторы также используют их производные – Δ и Δ^2, которые отражают скорость его изменения во времени и ускорение этих изменений соответственно, что позволяет захватить временные вариации в аудиосигнале.

Дополнительно применяется спектральный контраст, который вычисляет разницу в энергии между соседними частотными полосами. Этот метод помогает захватить динамику спектра и тональные изменения, что особенно полезно для различения реального и поддельного аудио. Формула для расчета спектрального контраста выглядит следующим образом:

$$C_i = E_{i+1} - E_i, \quad (15)$$

где C_i – контраст в i-й подполосе, а E_i – энергия в i-й подполосе.

Далее модели ANN, KNN и линейная регрессия обучаются на извлеченных признаках. Для оценки производительности используются метрики точности и F1-меры.

Наилучшие результаты были достигнуты с использованием искусственной нейронной сети, где точность составила 98.42%. Метод K-ближайших соседей также показал высокую точность, достигая 97.67% при использовании комбинированных признаков, а F1-мера составила 0.9809, что подтверждает его эффективность.

В статье [12] рассматривается подход с использованием KNN, а также других методов, таких как случайный лес, логистическая регрессия (Logistic Regression – LR) и гауссовский наивный байесовский классификатор (Gaussian Naive Bayes – GNB).

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

Авторы отбирали аудиофайлы с расширением Waveform Audio File Format (WAV) для дальнейшей обработки и отбрасывали остальные. Затем они закодировали целевой класс с помощью LabelEncoder, преобразовав метки так, что поддельное аудио кодируется как 1, а реальное – как 0. Кодирование было совершено с целью обеспечения единообразия в данных, что важно для корректного обучения моделей. Это позволяет избежать ошибок, связанных с неправильной интерпретацией меток алгоритмами машинного обучения.

Для выделения параметров звука также используется метод извлечения MFCC. Применяется библиотека librosa, и для каждого аудиосигнала извлекается 13 MFCC-признаков, которые представляют собой кратковременный спектр мощности сигнала. После извлечения MFCC-признаков они используются для обучения модели случайного леса, и на основе предсказанных вероятностей классов генерируются новые признаки, которые затем используются для обучения классификаторов, включая метод K-ближайших соседей.

Для детектирования дипфейка используется метод трансферного обучения (MFCC-Random Forest – MfC-RF). Предложенный метод сначала извлекает MFCC-признаки, затем генерирует признаки вероятности предсказания класса с помощью случайного леса. Они используются для обучения классификаторов, включая KNN, что позволяет улучшить точность обнаружения поддельного аудио. В исследовании применяется набор данных SceneFake, содержащий 12668 аудиофайлов (6334 реальных и 6334 поддельных).

Результаты экспериментов показывают, что метод K-ближайших соседей с MFCC-признаками достигает точности 90%, с Precision – 0.92, Recall – 0.87 и F1-мерой – 0.87. При использовании трансферных признаков KNN достигает точности 96%, с Precision – 0.97, Recall – 0.95 и F1-мерой –

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

0.96. Также в рамках работы стоит отметить, что вычислительная сложность метода K-ближайших соседей составляет 0.033 секунды, что делает его одним из самых быстрых методов.

Исследователи статьи [13] предлагают легковесный подход, основанный на вручную созданных аудио-признаках, который позволяет эффективно различать подлинные и поддельные аудиозаписи. В ходе предобработки данных исследователи воспользовались библиотекой librosa. С её помощью каждый звуковой файл загружался в систему, преобразуясь во временной ряд и частоту дискретизации – цифровое представление аудиозаписи. Частота дискретизации настраивается в соответствии с исходным значением аудиофайла. Для анализа аудиосигнала применяется FFT с длиной окна 2048 и шагом 512, что позволяет анализировать последовательные кадры аудио.

В данной работе из аудиосигналов извлекаются такие же ключевые признаки, как и в исследовании [10]. MFCC используются для анализа спектральных характеристик аудио, включая преобразование сигнала из временной области в частотную. Для выделения низкочастотных компонент применяется совокупность мел-фильтрового банка и дискретное косинусное преобразование для декорреляции коэффициентов.

В качестве классификаторов были использованы KNN, RF, XGBoost и многослойный перцептрон (Multi-layer Perceptron – MLP). Модель обучается на наборах данных ASVSpoof2019, FakeAVCelebV2 и In-The-Wild, а затем тестируется на неизвестных данных. Для повышения точности применяется протокол, независимый от субъекта, чтобы избежать переобучения на идентичности говорящих. Для повышения прозрачности модели использовался метод аддитивного объяснения Шепли (SHapley Additive exPlanations – SHAP), который помогает понять, какие признаки наиболее важны для классификации аудио.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

Также в статье предложенный метод был сравнен с подходами статьи [9], использующими те же классификаторы, но с разными наборами признаков. Авторы демонстрируют, что предложенный набор признаков (MFCC + спектральные признаки + RMSE + ZCR + хроматические признаки) значительно улучшает производительность классификаторов по сравнению с традиционными подходами, такими как использование только MFCC или MFCC в сочетании с линейным дискриминантным анализом. Например, для метода K-ближайших соседей EER снижается с 0.43 (только MFCC) до 0.2142 (предложенный метод).

Для систематизации результатов применения метода K-ближайших соседей в задачах детектирования аудиодипфейков, рассмотренных в представленных исследованиях, ключевые показатели эффективности сведены в Таблице 2.

Таблица № 2.

Эффективность методов с KNN в классификации аудиодипфейков

Исследование	Классификатор	Предобработка	Выделение параметров	EER (%)	ROC AUC	Accuracy (%)	F1-score	Precision	Recall
6	**Weighted KNN**	Обрезка до 5 сек, приведение к моно	Величина и фаза бикогерентности	-	0.82	-	-	-	-
			Величина и фаза бикогерентности+ MFCC + Δ и Δ^2Кепстр	**-**	**0.99**	**-**	**-**	**-**	**-**
7	KNN	Spleeter, RVC, сегментация на 1-секундные блоки, балансировка данных.	Chromagram, SC, SB, SR, ZCR, RMS, 20 MFCC	-	0.82	81.5	0.817	0.808	0.827
9	**KNN**	**мел-спектрограммы**	Модифицированная ResNet50 + LDA	0.7	-	99.5	0.996	0.996	0.992

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

		(преобразование в мел-шкалу, DFT, спектр мощности в dB)	MFCC	0.43	-	59	0.59	0.59	0.595
			MFCC+ LDA	**0.36**	**-**	**79**	**0.79**	**0.79**	**0.78**
10	KNN	Librosa (нормализация, загрузка аудио)	RMSE, Chroma, SC, SB, SR, ZCR, 20 MFCC	-	-	62	0.61	0.61	0.61
11	**KNN**	**Разбиение на кадры, оконная функция Хэмминга**	MFCC	-	-	93.42	-	-	-
			MFCC+Δ	-	-	93.45	-	-	-
			MFCC+Δ+Δ^2	-	-	93.45	-	-	-
			Spectral Contrast	-	-	95.40	-	-	-
			MFCC+SC	-	-	95.73	-	-	-
			MFCC+Δ+SC	-	-	95.67	-	-	-
			MFCC+Δ+Δ^2+SC	**-**	**-**	**97.67**	**-**	**-**	**-**
12	**KNN**	**Отбор wav файлов; кодирование меток**	13 MFCC	-	-	90	0.87	0.92	0.87
			MFCC + novel transfer features	**-**	**-**	**96**	**0.96**	**0.97**	**0.95**
13	KNN	Преобразование в цифровой формат (Librosa), FFT	RMSE, Chroma, SC, SB, SR, ZCR, 20 MFCC	0.214	-	81.08	0.8	0.84	0.8

На основе анализа данных, представленных в Таблице 2, можно сделать вывод, что среди рассмотренных классификаторов наиболее эффективными являются несколько методов, каждый из которых демонстрирует высокие результаты в зависимости от используемых подходов и признаков.

Наивысшую эффективность в задаче классификации демонстрирует метод взвешенного KNN, который использует данные о величине и фазе бикогерентности вместе с MFCC и их производными (Δ и Δ^2 Кепстр) [6]. Данный подход достигает значения ROC AUC, равного 0.99, что практически соответствует идеальному показателю.

Еще одним высокоэффективным методом является KNN, основанный на комбинации модифицированной архитектуры ResNet50 и линейного дискриминантного анализа [9]. Этот подход демонстрирует высокие результаты по всем основным метрикам. Модель достигает EER всего 0.7%, что свидетельствует о низкой вероятности ошибок и высокой надежности.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

Точность классификации составляет 99.5%, а значения F1-score (0.996), Precision (0.996) и Recall (0.992) приближаются к единице, что подтверждает качество работы модели. Это свидетельствует о сбалансированности модели и её способности эффективно работать как с положительными, так и с отрицательными классами.

Также высокие результаты демонстрирует метод K-ближайших соседей с использованием трансферных признаков [12], который достигает точности 96%, F1-score = 0.96, Precision = 0.97 и Recall = 0.95. Высокая эффективность этого метода, обусловлена использованием специализированных признаков, которые лучше учитывают особенности данных и позволяют модели более точно разделять классы.

Кроме того, метод K-ближайших соседей в работе [11] демонстрирует наивысшую точность (Accuracy = 97.67%) за счет комбинации признаков, включающей MFCC, их первые и вторые производные (Δ и Δ^2), а также спектральный центроид. Такое сочетание позволяет учитывать более полную информацию о звуковых сигналах, что значительно улучшает качество классификации.

Как показал анализ, методы на основе KNN демонстрируют высокую эффективность в детектировании аудиодипфейков, особенно при использовании комбинаций признаков (MFCC, бикогерентность, спектральный контраст) и сложных подходов к предобработке данных. Однако, несмотря на их результативность, метод K-ближайших соседей, как и метод опорных векторов, остаются методами с жестко заданной логикой классификации, требующими тщательного проектирования признаков и ручной настройки параметров.

В отличие от традиционных методов машинного обучения, нейронные сети обладают уникальной способностью автоматически извлекать сложные и иерархические признаки из аудиоданных, что делает их особенно

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

эффективными в задачах, где данные характеризуются высокой степенью нелинейности и сложными зависимостями. Благодаря своей архитектуре нейронные сети способны моделировать как временные, так и пространственные зависимости, что критически важно для анализа аудиосигналов.

Особое внимание уделяется гибридным методам, которые сочетают в себе преимущества различных подходов, таких как сверточные нейронные сети, рекуррентные нейронные сети, механизмы внимания и остаточные нейронные сети. Такие комбинации позволяют более эффективно выявлять сложные артефакты, характерные для поддельных аудиозаписей. Гибридные методы способны учитывать как временные, так и частотные характеристики аудиосигналов, что делает их особенно мощными в задачах обнаружения дипфейков.

В работе [14] исследуется возможность использования различных спектрограмм для обнаружения дипфейков в речи. Авторы сравнивают спектрограммы по их производительности, требованиям к ресурсам и скорости обработки. Основной целью является определение наиболее эффективных спектрограмм для обнаружения дипфейков, что важно для защиты систем биометрии речи от атак с использованием синтетических голосов.

В процессе предобработки была произведена нормализация после обрезки аудиозаписей, что позволило минимизировать влияние различий в громкости на результаты анализа. Удаление шумов и тишины из аудиозаписей также является важным этапом предобработки. В работе сегменты с тишиной (ниже 40 дБ) удаляются, чтобы уменьшить влияние шума и повысить качество данных. Это особенно важно для анализа спектральных характеристик, так как шум может искажать результаты. После выполнения

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

основных операций предобработки аудиозаписи преобразуются в формат WAV, который является стандартным для обработки аудиоданных.

Исследование включает создание и обучение нескольких детекторов дипфейков на основе временной сверточной сети с использованием различных спектрограмм. Для анализа спектральных характеристик используются кратковременное преобразование Фурье, постоянное Q-преобразование (Constant-Q Transform – CQT), вариативное Q-преобразование (Variable-Q Transform – VQT), временно-частотное представление с использованием многоскоростного фильтрового банка, состоящего из фильтров бесконечных импульсных характеристик (Infinite Impulse Response – IIR), мел-спектрограммы, MFCC и хромаграммы. Эти спектрограммы представляют собой визуальные отображения частотно-временных характеристик аудиосигналов и служат входными данными для классификации.

Аудиосигналы преобразуются в спектрограммы с использованием библиотеки librosa, а выделенные параметры подаются на вход TCN для классификации подлинных и поддельных аудиозаписей.

Результаты показывают, что детектор на основе MFCC-кепстрограммы работает лучше всего, достигая EER 0.07%. Детектор на основе STFT-спектрограммы также демонстрирует высокую производительность (EER 0.11%) (Таблица 3). Более того, общая производительность сравнима с результатами соревнования ASVspoof2021, где детекторы в задаче глубоких подделок достигли EER между 15.64 и 29.75% [15].

Таблица № 3.

EER для каждого обучающего набора данных и спектрограммы, (%)

Spectrogram (detector)	AS19	F2S	FREC	WF	Total

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

mel	0.26	0.06	0.02	0.01	0.13
stft	0.21	0.02	0.01	0.00	0.11
cqt	0.20	0.04	0.01	0.02	0.10
vqt	0.22	0.05	0.01	0.00	0.10
iirt	0.26	0.08	0.02	0.17	0.20
mfcc	0.16	0.02	0.01	0.02	0.07
chroma	0.27	0.26	0.10	0.34	0.27

Также в работе была проведена кросс-валидация набора данных. Результаты вычисления EER по всем валидационным наборам для каждого детектора (Таблица 4) показали, что мел-спектрограммы и VQT-спектрограммы являются наиболее точными при работе с неизвестными данными. Их производительность составила 45.85% и 45.88% соответственно.

Таблица № 4.

EER при проверке перекрестного набора данных, (%)

Спектрограммы	EER
Mel	45.85
STFT	46.25
CQT	46.43
VQT	45.88
IIRT	49.05
MFCC	48.23
Chroma	46.58

По итогам обоих экспериментов детектор на основе STFT-спектрограммы оказался одним из лучших. Его способность хорошо

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

работать как на знакомых, так и на незнакомых данных предполагает, что он может быть оптимальным выбором для обнаружения дипфейков.

В статье [16] авторы предлагают глубокие остаточные нейронные сети для задачи классификации. Было разработано три варианта моделей: MFCC-ResNet, CQCC-ResNet и Spec-ResNet, которые обрабатывают входные признаки MFCC, кепстральные коэффициенты на основе постоянного Q-преобразования (Constant Q Cepstral Coefficients – CQCC) и логарифмированную амплитуду STFT (которая, по сути, является спектрограммой) соответственно.

Эти три варианта имеют почти идентичную архитектуру, но различаются:

- размером входных данных (для соответствия разной размерности признаков);
- количеством нейронов в первом полносвязном слое (расположенном после последнего остаточного блока).

MFCC извлекается через STFT, затем применяется мел-фильтр и дискретное косинусное преобразование, при этом в расчет берутся первые 24 коэффициента, а также их первые и вторые производные (Δ-MFCC и Δ^2-MFCC), что улучшает захват динамики звука. CQCC задействует постоянное Q-преобразование, обеспечивающее лучшее частотное разрешение на низких частотах и временное разрешение на высоких частотах.

После CQT вычисляется спектр мощности, который затем логарифмируется и преобразуется с помощью DCT. Логарифмическая амплитуда STFT вычисляется с применением окна Хэмминга размером 2048 точек и 25% перекрытием. Полученная логарифмическая амплитуда спектра подается непосредственно на вход нейронной сети без дополнительной обработки.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

Каждый блок состоит из сверточных слоев, пакетной нормализации, активации и исключений (dropout). Функция потерь – взвешенная кросс-энтропия с соотношением весов 9:1 для балансировки классов, а оптимизация выполняется с помощью Adam с темпом обучения (learning rate) 5×10^{-5}.

Для детектирования дипфейков используется оценка противодействия (Countermeasure – CM), которая вычисляется как логарифмическое отношение правдоподобия по формуле (16). Высокая оценка CM указывает на настоящую речь, а низкая – на подделку.

$$CM(s) = log\big(p(\,bona\ fide \mid s;\theta\,)\big) - log\big(p(\,spoof \mid s;\theta\,)\big), \qquad (16)$$

где s – входной аудиофайл; θ – параметры модели, которые были обучены на тренировочных данных; $p(\,bona\ fide \mid s;\theta\,)$ – вероятность того, что аудиофайл является настоящим; $p(\,spoof \mid s;\theta\,)$– вероятность того, что аудиофайл является поддельным.

Методика работы включает обучение моделей на наборе данных ASVSpoof2019. Объединение моделей (fusion) выполняется путем взвешенного усреднения их оценок *CM*.

Результаты показывают, что в сценарии логического доступа объединение моделей достигает нулевых значений метрик t-DCF и EER на подготовительном наборе, содержащем известные типы атак. На оценочном наборе (включая неизвестные атаки) объединение моделей улучшает t-DCF на 25% (t-DCF = 0.1569) и EER на 25% (EER = 6.02). В сценарии физического доступа (атаки с повторным воспроизведением) объединение моделей улучшает t-DCF на 71% (t-DCF = 0.0693) и EER на 75% (EER = 2.78).

В работе [17] предложен метод Deep4SNet, основанный на сверточной нейронной сети, который использует в качестве входных данных гистограммы голосовых записей. Модель обладает свойством тексто-независимости, что делает её универсальной и применимой к любым

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

голосовым записям, независимо от их содержания. Для повышения обобщающей способности используется горизонтальное отражение гистограмм, а для предотвращения переобучения в архитектуру CNN интегрирован dropout – механизм, который случайным образом отключает часть нейронов во время обучения.

Deep4SNet успешно классифицирует как оригинальные, так и поддельные голосовые записи, независимо от метода их создания. Этот метод подтверждает эффективность использования глубокого обучения для решения задач обнаружения поддельных аудиозаписей, особенно в условиях, когда данные могут быть несбалансированными или содержать сложные нелинейные зависимости.

На рис. 1 представлена архитектура модели, где обозначены ключевые параметры: f – размер фильтра, s – размер шага, p – размер заполнения (padding). Модель состоит из трёх сверточных слоев (CONV) с последующими слоями пулинга (POOL). Размер выхода сверточного слоя может быть определен:

$$\frac{|n+2p-f|}{s}+1\times\frac{|m+2p-f|}{s}+1\times n_f, \qquad (17)$$

где n и m – количество строк и столбцов входного изображения, а n_f – количество фильтров.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

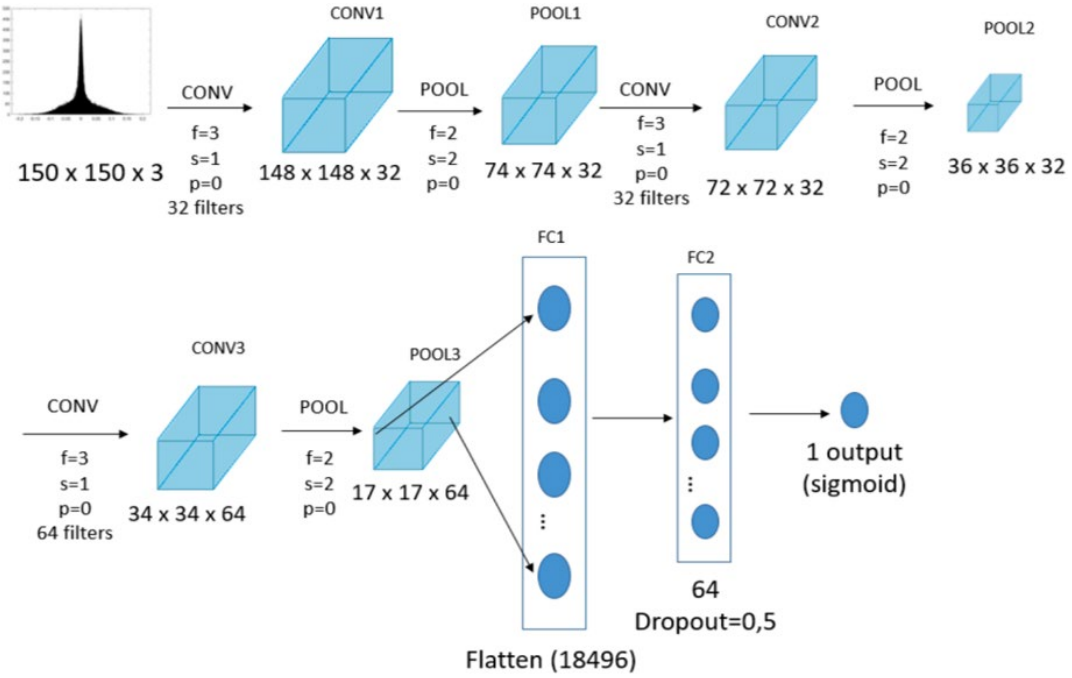

Рис. 1 – Архитектура модели CNN метода Deep4SNet

Например, для первого сверточного слоя (CONV1) с параметрами $n = m = 150$, $n_f = 32$, $f = 3$, $s = 1$ и $p = 0$ выходной размер составит $148 \times 148 \times 32$.

После сверточных слоев данные преобразуются в одномерный вектор с помощью слоя выравнивания (flatten). Затем следует полносвязный скрытый слой с dropout и выходной слой с сигмоидной активацией. В скрытых слоях применяется функция активации – выпрямленная линейная функция (Rectified Linear Unit – ReLU).

Для обучения модели используется функция потерь на основе бинарной кросс-энтропии, которая измеряет расхождение между предсказаниями модели и реальными данными:

$$L(y, \tilde{y}) = -\frac{1}{N}\sum_{i=0}^{N} y_i \, log(\tilde{y_i}) + (1 - y_i) \, log(1 - \tilde{y_i}), \qquad (18)$$

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

где y_i – истинные метки, $\tilde{y}_i$ – предсказанные вероятности, а N – количество примеров.

Модель демонстрирует высокую точность в обнаружении поддельных аудио: для записей, созданных методом Imitation, метрики Precision и Recall составили 0.997. Для записей, созданных методом Deep Voice, метрика Precision составила 0.985, а Recall – 0.944. Общая точность модели составила 0.985.

В работе [18] авторы предложили два новых подхода на основе сверточных нейронных сетей: EfficientCNN и RES-EfficientCNN. Эти модели предназначены для эффективного извлечения признаков из аудиоданных и обнаружения синтетических речевых сигналов с высокой точностью.

В исследовании процесс предобработки включает несколько ключевых шагов. Все аудиообразцы обрезаются до 4 секунд, затем каждый аудиофайл преобразуется в спектрограмму с частотой дискретизации 16 кГц, 1728 точек быстрого преобразования Фурье и окном Хэмминга длиной 108 мс с шагом 10 мс. После получения спектрограммы её значения логарифмируются и Z-нормализуются по формулам (19) и (20), что становится стандартным представлением, применяемым в экспериментах:

$$X_{log} = log\ (|X(m,k) + \varepsilon|), \tag{19}$$

$$X_{norm} = \frac{X_{log} - \mu}{\sigma}, \tag{20}$$

где $X(m,k)$ – это комплексное значение спектрограммы, ε – малое число, μ – среднее значение логарифмированной спектрограммы, σ – стандартное отклонение логарифмированной спектрограммы.

EfficientCNN – это легковесная сверточная сеть, которая принимает на вход нормализованную логарифмическую спектрограмму аудиозаписи.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

Архитектура модели включает в себя блок обработки входных данных, сверточные блоки и блок классификации:

- Входной блок обработки состоит из 2D сверточного слоя с ядром 5x5, за которым следуют слой активации ReLU, пакетная нормализация и слой макс-пулинга.
- Каждый сверточный блок состоит из свертки 1x1, активации ReLU, пакетной нормализации, свертки 3x3, ещё одной активации ReLU, пакетной нормализации и слоя макс-пулинга 2x2.
- Блок классификации включает линейный слой, активацию ReLU, пакетную нормализацию и ещё один линейный слой, с применением dropout (0.2) перед каждым линейным слоем. Выход последнего линейного слоя подается на слой софтмакс (softmax) для получения предсказаний.

Также авторы предлагают остаточную версию EfficientCNN, называемую RES-EfficientCNN, где каждый сверточный блок имеет соответствующий остаточный блок. Остаточный блок добавляет входные данные к выходу сверточного блока, что улучшает стабильность и точность обучения.

При тестировании на наборе данных ASVspoof 2019 модель RES-EfficientCNN продемонстрировала более высокий показатель F1-score, равный 97.61, что значительно превзошло результаты базовой модели EfficientCNN, которая показала F1-score 94.14. Остаточные соединения в архитектуре RES-EfficientCNN позволили модели лучше учитывать сложные признаки аудиоданных, что привело к повышению Accuracy и стабильности обучения. Эти результаты подтверждают эффективность предложенного подхода для обнаружения синтетических речевых сигналов.

В статье [19] авторы предлагают подход, основанный на использовании CNN и сиамской сверточной нейронной сети (Siamese CNN).

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

Для предобработки звука применяют гауссовскую смесь (Gaussian mixture models – GMM), которая моделирует характеристики речи. GMM обучается на наборе данных, содержащем как подлинную, так и поддельную речь. Для каждого кадра речи вычисляются логарифмические вероятности на компонентах гауссовской смеси, что позволяет учитывать вклад каждого гауссовского компонента в итоговую оценку. В модели гауссовской смеси расчет плотности вероятности смеси имеет следующий вид:

$$p(x) = \sum_{i=1}^{M} \omega_i p_i(x), \tag{21}$$

где ω_i – это вес i-го компонента смеси, а $p_i(x)$ – гауссовская плотность.

Авторы статьи используют два типа акустических признаков: CQCC и линейно-частотные кепстральные коэффициенты (Linear-frequency cepstral coefficients – LFCC). CQCC получаются с помощью CQT, которое лучше подходит для обработки речевых сигналов, а LFCC похожи на MFCC, но используют линейно-масштабированный фильтрбанк. Эти признаки извлекаются из речевых сигналов и используются для обучения моделей.

Для классификации авторы предлагают два типа классификаторов: одномерная сверточная нейронная (1-D CNN) сеть и сиамская сверточная нейронная сеть. 1-D CNN учитывает не только вероятности кадров на компонентах GMM, но и локальные взаимосвязи между кадрами. CNN использует сверточные слои, пулинг и полностью связанные слои для классификации. Siamese CNN состоит из двух идентичных CNN, которые обучаются одновременно на подлинной и поддельной речи. Выходные векторы двух сетей объединяются и передаются в полностью связанный слой для классификации.

Формула для логарифмического отношения правдоподобия в базовой системе выглядит так:

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

$$score_{baseline} = log\, p(X|\lambda_h) - log\, p(X|\lambda_s), \quad (22)$$

где λ_h и λ_s – GMM для подлинной и поддельной речи соответственно.

Для детектирования поддельной речи авторы используют гауссовские вероятностные признаки, которые включают распределение оценок для каждого компонента GMM. Это позволяет учитывать не только общую оценку, но и вклад каждого гауссовского компонента, что улучшает точность обнаружения. Формула для гауссовских вероятностных признаков:

$$f_{ij} = log\left(\omega_j \cdot p_j(x_i)\right), \quad (23)$$

где ω_j– вес j-го компонента, а $p_j(x_i)$ – вероятность кадра x_i на j-м компоненте.

Для начала GMM обучается на подлинной и поддельной речи без меток. Затем для каждого кадра речи вычисляются гауссовские вероятностные признаки. После этого нейронные сети (CNN и Siamese CNN) обучаются на извлеченных признаках для классификации подлинной и поддельной речи. Оценка производительности системы проводилась на наборе данных ASVspoof 2019, включающем два ключевых сценария: логический доступ, который был описан ранее, и физический доступ (Physical Access – PA).

Сценарий PA специально разработан для оценки устойчивости к атакам повторного воспроизведения (replay attacks), где используются реальные записи голоса, воспроизводимые через различные акустические системы в разных условиях. База данных PA содержит 9 различных конфигураций атак, варьирующихся по: расстоянию между источником звука и микрофоном, и качеству используемой аудиоаппаратуры для записи и воспроизведения.

Результаты экспериментов показывают, что предложенные методы значительно улучшают производительность по сравнению с базовыми

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

системами (GMM и CNN). Для сценария LA система LFCC + Siamese CNN улучшает показатели min-tDCF и EER и значения равны 0.093 и 3.79% соответственно. Для сценария PA система LFCC + Siamese CNN показывает улучшение min-tDCF (0.195) и EER (7.98%).

В работе [20] авторы предлагают метод перекрестной проверки, который позволяет сравнивать сомнительные записи с доверенными эталонными записями. Основное внимание уделяется выравниванию аудиозаписей и классификации кадров на совпадающие и несовпадающие, что позволяет эффективно обнаруживать подделки и подтверждать подлинность реальных записей.

В процессе предобработки звука применяются стандартные методы обработки речевых сигналов. Аудиозаписи преобразуются в последовательности мел-частотных кепстральных коэффициентов, которые широко применяются в задачах анализа речи. Параметры MFCC включают 13-мерные коэффициенты с Δ и Δ^2 признаками (всего 39 измерений), длину анализа кадра 25 мс и шаг между кадрами 10 мс.

Для классификации кадров на совпадающие и несовпадающие авторы экспериментируют с несколькими моделями, включая сети с долгой краткосрочной памятью (Long Short-Term Memory – LSTM) (однослойные и многослойные), двунаправленные LSTM (Bidirectional Long Short-Term Memory – BiLSTM) и трансформеры (Transformer).

Детектирование поддельных аудиозаписей основано на перекрестной проверке с доверенными эталонными записями. Процесс включает выравнивание запроса и эталона с использованием модифицированного алгоритма Нидлмана-Вунша (Subsequence Needleman Wunsch Time Warping – SNWTW) и классификацию кадров на совпадающие и несовпадающие с помощью моделей LSTM или Transformer. Формула для вычисления матрицы накопленных затрат D в алгоритме SNWTW выглядит следующим образом:

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

$$D[i,j] = \begin{cases} C[i,j] & i = 0, \\ \alpha + D[i-1,j] & i > 0, j = 0, \\ min\begin{pmatrix} \gamma + D[i,j-1], \alpha + D[i-1,j], \\ C[i,j] + D[i-1,j-1] \end{pmatrix} & i > 0, j > 0, \end{cases} \quad (24)$$

где $C[i,j]$ – попарная стоимость между кадрами, α и γ – штрафы за пропуск в запросе и эталоне соответственно.

Использование алгоритма SNWTW для выравнивания и моделей BiLSTM для классификации позволяет достичь высокой точности даже в сложных условиях. Этот подход может быть полезен для защиты мировых лидеров от распространения ложной информации.

В качестве набора данных используется база, состоящая из 50 аудиозаписей выступлений Дональда Трампа, собранных с официального YouTube-канала Белого дома за период с 2017 по 2018 год. Каждая запись длится от 1 до 3 минут и включает как чистую речь, так и фрагменты с фоновой музыкой.

Результаты работы системы демонстрируют высокую эффективность. Лучшая модель (3-слойная BiLSTM) достигает EER = 0.43%. Система практически не зависит от качества аудио (битрейта), что делает её применимой даже для низкокачественных записей.

В статье [21] представлен способ обнаружения подделки голоса с использованием комплекснозначных нейронных сетей. Предложенный метод объединяет преимущества обработки спектрограмм и «сырого» аудио, сохраняя фазовую информацию, что повышает как точность модели, так и её объяснимость. Это позволяет не только эффективно выявлять дипфейки, но и интерпретировать решения модели с помощью методов объяснимого ИИ.

Для предобработки звука задействуется постоянное Q-преобразование, которое преобразует временной аудиосигнал в частотное представление. Формула для CQT выглядит следующим образом:

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

$$Z_Q[t,k] = \frac{1}{N_k}\sum_{n=0}^{N_k-1} W_k[n]X[n+t]e^{-2\pi iQnN_k^{-1}}, \quad (25)$$

где $Z_Q[t,k]$ – комплексное частотное представление, Q – количество циклов для каждой частоты, N_k – размер окна для частоты k, W_k – окно (например, Гауссово или Ханна), а X – временной сигнал.

После постоянного Q-преобразования применяется логарифмическое преобразование для учета логарифмического восприятия звука человеком:

$$|z|e^{i\theta} \longmapsto max(\varepsilon, -log_e|z| + c)\cdot\alpha\cdot e^{i\theta}, \quad (26)$$

где ε – нижний порог для магнитуды, а α и c – обучаемые параметры.

Также из спектрограммы CQT извлекаются комплексные признаки, включающие как магнитуду, так и фазу. Это позволяет сохранить информацию, которая обычно теряется при использовании только магнитудных спектрограмм.

Для классификации используется комплекснозначная нейронная сеть (Complex-Valued Neural Networks – CVNN), которая обрабатывает комплексные входные данные. Основные компоненты сети включают комплексные свертки, комплексную выпрямленную линейную функцию (Complex Rectified Linear Unit – CReLU), пакетную нормализацию (Batch-Normalization) и Dropout. Модель обучается различать два класса: подлинное аудио (bona-fide) и поддельное аудио (spoofed). Она использует фазовую информацию для обнаружения несоответствий между магнитудой и фазой, что характерно для синтезированной записи.

Методика работы включает предобработку аудио, преобразование в комплексную спектрограмму CQT, обучение на наборах данных ASVspoof 2019 и "In-the-Wild", а также оценку модели на независимых данных. Для

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

объяснимости решений модели применяются методы объяснимого ИИ, такие как SmoothGrad, которые визуализируют важные признаки.

Модель продемонстрировала высокую эффективность на наборе данных "In-the-Wild", достигнув EER 26.95% ± 3.12. Исследования подтвердили важность фазовой информации: при её удалении EER увеличивается на 4%, а при рандомизации – на 8%.

В статье [22] представлен метод DeepSonar, основанный на анализе поведения нейронов в глубокой нейронной сети (Deep Neural Network – DNN), что позволяет эффективно различать реальные и синтезированные голоса. В отличие от традиционных методов, которые ищут артефакты в аудиоданных, DeepSonar фокусируется на внутренних процессах нейронной сети, что делает его более устойчивым к манипуляциям.

В исследовании применяются три набора данных: FoR (английские TTS-голоса, синтезированные с помощью коммерческих продуктов), Sprocket-VC (английские VC-голоса, созданные с использованием инструмента sprocket) и MC-TTS (китайские TTS-голоса, синтезированные системой Baidu). Эти наборы данных обеспечивают разнообразие в языках, техниках синтеза и источниках голосов.

Для выделения ключевых параметров применяются два критерия: среднее количество активированных нейронов (Average Count Neuron – ACN) и топ-k наиболее активных нейронов (Top-k Activated Neurons – TKAN). ACN подсчитывает количество активированных нейронов в каждом слое, где нейрон считается активированным, если его выходное значение превышает порог δ_l, который определяется как среднее значение выходов нейронов в слое l:

$$\delta_l = \frac{\sum_{x \in X, i \in I} \varphi(x, i; \theta)}{|I| \cdot |X|}, \tag{27}$$

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

где x – нейрон в -м слое, i – входные данные в обучающем наборе данных I, X – набор нейронов в слое l, φ – функция активации для расчета выходного значения нейрона для входных данных i с обученными параметрами θ, $|X|$ и $|I|$ представляют количество нейронов в слое l и количество входных данных в обучающем наборе данных I, соответственно.

TKAN, в свою очередь, выбирает нейроны с наибольшими выходными значениями в каждом слое, что позволяет выделить наиболее значимые нейроны, которые играют ключевую роль в представлении данных.

Для классификации используется бинарный классификатор на основе неглубокой нейронной сети. Входными данными для классификатора являются векторы, представляющие активацию нейронов в разных слоях DNN. Это позволяет классификатору эффективно различать реальные и поддельные голоса, даже если они были подвергнуты манипуляциям, таким как изменение высоты тона или добавление шумов.

Метод тестировался на разных наборах данных, включая голоса на английском и китайском языках, и показал высокую точность (более 98.1%) в обнаружении подделок. EER составила менее 2%, что подтверждает эффективность метода. Метод показывает стабильную производительность даже при высоких уровнях шума (SNR > 35). Например, при добавлении шумов, таких как дождь или ветер, точность DeepSonar снижается незначительно, что делает его пригодным для использования в реальных условиях, где такие шумы могут присутствовать.

В исследованиях [23, 24] предложен метод детектирования аудиодипфейков на основе сверточной нейронной сети, использующей мел-спектрограммы MFCC для классификации аудиосигналов. На первом этапе авторы сформировали графические изображения аудиосигнала мел-спектрограмм для обучения сверточной нейронной сети. Предобработка

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

исходного аудио проводилась через дискретное преобразование Фурье по формуле:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N}kn} = \sum_{n=0}^{N-1} x_n \left(cos\left(\frac{2\pi kn}{N}\right) - i \cdot sin\frac{2\pi kn}{N}\right), \qquad (28)$$

где N - количество значений сигнала, измеренных за период, а также количество компонент разложения, X_n, n = 0,….., N-1 – измеренные значения сигнала в дискретных временных точках, X_k, k = 0,…., N-1, -N комплексных амплитуд синусоидальных сигналов.

Также было выполнено масштабирование изображений с помощью функции ImageDataGenerator (rescale=1./255) с коэффициентом 1/255.

На рис. 2 приведена функциональная блок-схема модели. Показаны следующие блоки:

- входные параметры (мел-спектрограммы установленного размера, первичный входной слой нейронной сети);
- сверточная нейронная сеть (два сверточных слоя с нелинейной функцией активации RELU, слой выравнивания, скрытый слой с повторной активацией);
- выходной слой с двумя нейронами определяет истинность или фальсификацию аудиосигнала, разделив их на два класса: реальный и фейковый.

Для обучения модели была использована функция model.fit фреймворка «TensorFlow 2.0». Проведенный эксперимент установил, что созданная модель сверточной нейронной сети достигает точность вычислений в 78%.

Гибридные методы сочетают в себе преимущества различных подходов, таких как сверточные нейронные сети, рекуррентные нейронные сети, механизмы внимания и остаточные сети, что позволяет более эффективно выявлять сложные артефакты, характерные для поддельных

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

аудиозаписей. Гибридные методы способны учитывать как временные, так и частотные характеристики аудиосигналов, что делает их особенно мощными в задачах обнаружения дипфейков.

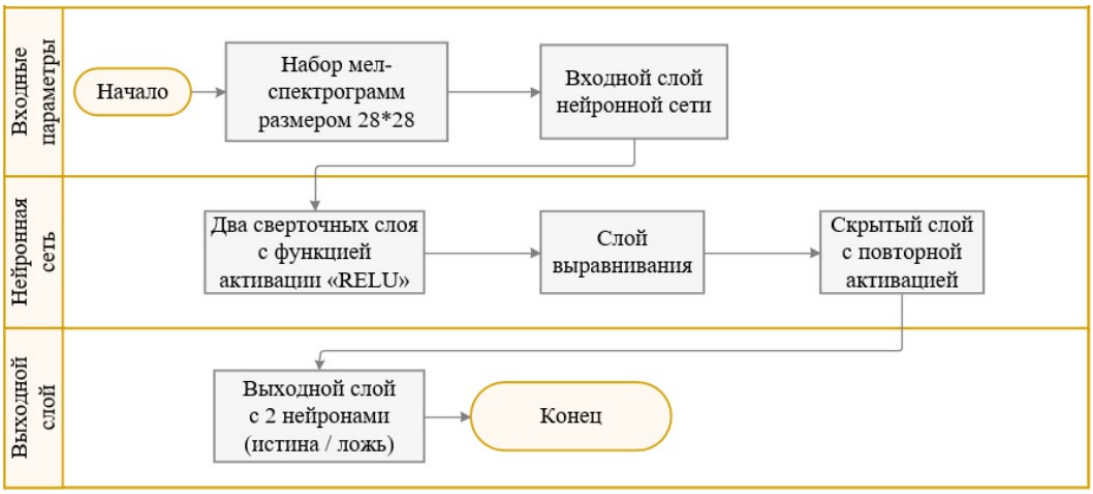

Рис. 2 – Функциональная блок-схема сверточной нейронной сети для выявления аудио-дипфейков

В публикации [25] предложен уникальный подход к обнаружению поддельного аудио с использованием самообучающихся методов. Авторы предлагают схему самоконтролируемого обнаружения поддельного аудио (Self-supervised Spoofing Audio Detection – SSAD), которая включает три модели: большая облегчённая свёрточная сеть (Lite Convolutional Neural Network-big – LCNN-big), малая LCNN (LCNN-small) и Squeeze-and-Excitation Network 12 (SENet12).

LCNN-big и LCNN-small – это легкие сверточные нейронные сети с разным количеством параметров, а SENet12 – сеть с архитектурой, основанной на Squeeze-and-Excitation блоках. Классификаторы обучаются с использованием функции потерь Angular Softmax (A-Softmax), что улучшает разделение классов.

Для предобработки звука задействуется восьмислойная сверточная сеть, которая обрабатывает исходный аудиосигнал. Каждый слой состоит из

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

одномерной свертки, пакетной нормализации и активации параметрического ReLU (Parametric ReLU – PReLU).

Также применяется временная сверточная сеть, которая позволяет эффективно захватывать долгосрочные зависимости в аудиосигнале. Временная сверточная сеть задействует расширенные свертки (dilated convolutions), что позволяет увеличить рецептивное поле без потери деталей. Формула расширенной свертки выглядит следующим образом:

$$F(s) = (x_d \cdot f)(s) = \sum_{i=0}^{k-1} f(i) \cdot x_{s-d \cdot i}\,, \tag{29}$$

где d – коэффициент расширения, k – размер фильтра, x – входной сигнал. а $(s - d \cdot i)$ учитывает направление в прошлое.

При выделении параметров аудио используются три регрессионных рабочих модуля, которые предсказывают следующие характеристики: логарифмический спектр мощности (Log power spectrum – LPS), LFCC и CQCC. LPS вычисляется из исходного сигнала без предварительной обработки, LFCC извлекаются из 20 логарифмических фильтров с добавлением первых и вторых производных, а CQCC извлекаются с использованием специальных методов [26], также с добавлением производных. Эти параметры специально подобраны для задачи обнаружения поддельного аудио.

Для детектирования поддельного аудио используется бинарный рабочий модуль, который решает задачу Congener Info Max (CIM), минимизируя расстояние между похожими аудиозаписями (например, двумя реальными записями) и максимизируя расстояние между разными (например, реальной и поддельной записью). Формула потерь для CIM выглядит следующим образом:

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

$$L1 = E_{S_r}\left[log\left(d(s_a, s_r)\right)\right], \quad (30)$$

$$L2 = E_{S_f}\left[log\left(1 - d(s_a, s_f)\right)\right], \quad (31)$$

$$L = L1 + L2, \quad (32)$$

где s_a – якорный пример (anchor), s_r – положительный пример (positive), s_f – отрицательный пример (negative), а d – функция дискриминатора.

Методика работы SSAD включает в себя обучение и тестирование. На этапе обучения кодировщик преобразует аудиосигнал в высокоуровневое представление, а рабочие модули (регрессионные и бинарный) обучаются на предсказании целевых признаков и разделении реальных и поддельных записей. Используется оптимизатор Adam с полиномиальным снижением скорости обучения.

В данном исследовании применяется набор данных ASVspoof 2019. Результаты экспериментов показывают, что SSAD с TCN демонстрирует улучшение на 2.49% по EER в наборе данных Eval по сравнению с PASE+. SSAD также превосходит базовые методы, такие как LFCC и CQCC, с EER 6.55% для SENet12 и 5.31% для LCNN-big.

В статье [27] предложен метод обнаружения поддельных аудиозаписей, основанный на анализе спектро-временных модуляционных признаков, полученных с помощью двумерного дискретного косинусного преобразования (2 dimensional DCT – 2D-DCT) над лог-мел-спектрограммой.

Суть метода заключается в том, что поддельные аудиозаписи, созданные с использованием современных технологий, таких как нейронные вокодеры и генеративно-состязательные сети, часто содержат артефакты, которые можно выявить с помощью анализа спектро-временных паттернов. Эти артефакты возникают из-за несовершенства алгоритмов синтеза речи,

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

например отсутствия временной согласованности между кадрами или искажений на высоких частотах.

При предобработке аудиоданные обрезается до 4 секунд. Для вычисления спектрограммы используется FFT с размером окна 1024, шагом 256 и 128 Мел-фильтрами. К лог-мел спектрограмме применяется 2D-DCT, что позволяет выделить долгосрочные спектро-временные модуляции, ключевые для обнаружения поддельных аудиозаписей. Полученные 2D-DCT признаки могут быть нормализованы (например, с использованием l1-нормализации) для улучшения стабильности и производительности модели. Преимущество метода заключается в анализе глобальных спектро-временных паттернов, что особенно полезно для выявления артефактов, связанных с синтетически сгенерированным аудио. Дополнительно для повышения точности используется взвешенное голосование между предсказаниями базовой модели и модели с предложенной 2D-DCT характеристикой.

Для реализации этого метода авторы используют классификатор на основе сверточная нейронной сети с остаточными блоками. Базовая модель CNN состоит из начального сверточного слоя, трех остаточных блоков, двунаправленных управляемых рекуррентных блоков (Gated Recurrent Units – GRU) и слоя самовнимательного пулинга. После обработки данных через эти слои применяется многослойный перцептрон с одним скрытым слоем для получения вероятности предсказания.

Математически метод описывается следующим образом. Входной сигнал $x(n)$ сначала преобразуется в частотную область с помощью быстрого преобразования Фурье. Затем вычисляется мощность спектра, которая пропускается через банк фильтров Мел-частот для получения Мел-спектра. Далее к Мел-спектру применяется логарифмическое преобразование, в результате чего формируется лог-мел-спектрограмма. Наконец, лог-мел-спектрограмма подвергается двумерному дискретному косинусному

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

преобразованию, что позволяет получить спектро-временные модуляционные признаки.

Результаты экспериментов на данных ASVspoof 2019 показали, что предложенный метод достигает t-DCF 0.0737, что на 14% ниже, чем у лучших одиночных систем [28]. Авторы также протестировали свой подход на других наборах данных, таких как FoR и RTVCspoof, и получили высокую точность обнаружения поддельных аудиозаписей – от 90% до 98%. Эти результаты свидетельствуют о том, что предложенный метод обладает хорошей способностью к обобщению и может быть успешно применен к различным типам данных, что делает его надежным инструментом для защиты систем верификации говорящего от мошеннических атак.

В работе [29] представлена система ASSERT, основанная на DNN, включая SENet, Mean-Std ResNet и Dilated ResNet. Данная система использует бинарную и многоклассовую классификацию для обнаружения поддельных аудиозаписей.

Для предобработки звука в системе ASSERT применяют два типа акустических признаков: CQCC – 30-мерные кепстральные коэффициенты, включающие статические, дельта и двойные дельта коэффициенты, и логарифмические спектры мощности (Log Power Magnitude Spectra – Logspec) с размерностью 257. Особенность предобработки заключается в том, что обнаружение активности голоса (Voice Activity Detection – VAD) и нормализация исключаются, так как это ухудшает результаты.

Также используются два следующих подхода: унифицированная карта признаков (Unified Feature Map) и целое высказывание (Whole Utterance). В первом подходе аудиозаписи разбиваются на сегменты длиной 400 кадров с перекрытием 200 кадров, что позволяет обрабатывать записи переменной длины. Во втором подходе записи обрабатываются целиком, с дополнением нулями до длины самой продолжительной записи в мини-батче. Эти методы

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

позволяют эффективно извлекать признаки из аудиоданных, что является важным этапом для последующей классификации.

Методика работы системы ASSERT включает несколько этапов. Сначала извлекаются акустические признаки, затем модели DNN обучаются с использованием оптимизатора Adam и планировщика learning rate. Модели обучаются с мини-батчами размером 64 для SENet и 32 для Mean-Std ResNet. После обучения лучшие модели комбинируются с помощью логистической регрессии для улучшения результатов. Этот подход позволяет достичь высокой точности в обнаружении аудио спуфинг-атак.

Результаты на корпусе ASVspoof 2019 демонстрируют высокую эффективность: для PA минимальная t-DCF составила 0.016, а EER – 0.59%. Для LA t-DCF составила 0.155, а EER – 6.70% (на этапе оценки).

В работе [30] предложен метод, основанный на комбинации расширенного локального тернарного паттерна (Extended Local Ternary Pattern – ELTP) и LFCC, которые используются для обучения глубокой двунаправленной сети LSTM (Deep Bidirectional Long Short-Term Memory – DBiLSTM). Эта модель обрабатывает временные последовательности данных, что делает её подходящей для анализа аудио-сигналов. Она состоит из 10 слоев BiLSTM, каждый с 64 скрытыми единицами.

Процесс начинается с предобработки звука, где аудио-сигнал разбивается на неперекрывающиеся кадры длиной l, и для каждого кадра применяется оконное преобразование, например Хэмминга, чтобы уменьшить эффекты краев кадра. Затем каждый кадр преобразуется в частотную область с помощью FFT для получения спектра сигнала. Для выделения параметров аудио звука используется расширенный локальный тернарный паттерн, который анализирует локальные изменения в аудио-сигнале. ELTP кодирует аудио-сигнал в тернарные значения (+1, 0, -1) в зависимости от разницы между центральным образцом c и его соседями s^i.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

Динамически вычисляемый порог θ зависит от стандартного отклонения σ каждого кадра. Этот подход делает ELTP устойчивым к шуму и изменениям в аудио-сигнале:

$$\theta = \alpha \times \sigma \quad (0 < \alpha \leq 1), \tag{33}$$

где α – масштабирующий коэффициент, а σ – стандартное отклонение, вычисленное для каждого кадра аудио.

Дополнительно используются LFCC, которые извлекаются из спектра сигнала с использованием линейного банка фильтров. LFCC помогают захватывать высокочастотные характеристики речи, которые важны для обнаружения поддельных голосов. Формула для вычисления LFCC:

$$LFCC(i) = \sum_{k=1}^{K} log(g_k)\, cos\left(\frac{(2k-1)i\pi}{2K}\right) \quad (1 \leq i \leq I), \tag{34}$$

где K и I представляют количество фильтров и количество коэффициентов LFCC соответственно.

Выходные данные последнего слоя передаются в полносвязный слой и слой softmax для классификации. Метод детектирования дипфейка основан на обнаружении алгоритмических артефактов. ELTP захватывает динамические изменения в речевом тракте подлинного голоса и артефакты, возникающие при синтезе или преобразовании голоса. LFCC дополняют ELTP, захватывая частотные характеристики, которые отличают подлинный голос от поддельного. Динамический порог, используемый в ELTP, делает его более устойчивым к шуму и изменениям в аудио-сигнале по сравнению с предыдущими методами.

Результаты на наборе данных ASVspoof 2019 показали, что метод превосходит базовые методы и современные подходы, такие комбинации

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

методов как Contrastive-1 и Primary [31]. Общая производительность метода составила EER 0.74% и t-DCF 0.008.

Для комплексного сравнения эффективности нейросетевых и гибридных подходов основные метрики производительности систематизированы в Таблице 5.

Таблица № 5.

Эффективность нейронных сетей и гибридных методов в классификации аудиодипфейков

Исследование	Классификатор	Предобработка	Выделение параметров	EER (%) LA/PA	min t-DCF LA/PA	Accuracy (%)	Precision	Recall	F1-score
1	2	3	4	5	6	7	8	9	10
14	**TCN**	**Нормализация, удаление шумов и тишины (<40дБ), конвертация в WAV**	**MFCC**	**0.07**	**-**	**-**	**-**	**-**	**-**
			STFT	0.11	-	-	-	-	-
16	ResNet	Окно Хэмминга, логарифмирование амплитуды спектра	STFT	9.68/ 3.81	0.274/ 0.099	-	-	-	-
		STFT, мел-фильтр, DCT	MFCC	9.33/-	0.204/-	-	-	-	-
		CQT, спектр мощности, логарифмирование, DCT	CQCC	7.69/ 4.43	0.217/ 0.107	-	-	-	-
		Объединение признаков Spec, MFCC, CQCC	CM	6.02/ 2.78	0.157/ 0.069	-	-	-	-
17	**CNN: Imitation**	**Гистограммы амплитуд сигнала, горизонтальное отражение (аугментация)**	**Выделение признаков с помощью CNN**	**-**	**-**	**98.5**	**0.997**	**0.997**	**-**
	CNN: Deep Voice			-	-		0.985	0.944	-
18	EfficientCNN	**Обрезка до 4с., преобразование в спектрограмму**	**Логарифмирование и Z-нормализация**	-	-	-	-	-	94.14
	RES-EfficientCNN			**-**	**-**	**-**	**-**	**-**	**97.61**
1	2	3	4	5	6	7	8	9	10
19	Siamese CNN	Обучение GMM на подлинной и поддельной речи без меток, вычисление гауссовских вероятностных признаков	CQCC, LFCC	3.79/ 7.98	0.093/ 0.195	-	-	-	-
	GMM			7.59/ 12.9	0.208/ 0.287	-	-	-	-
	CNN			4.28/ 9.48	0.122/ 0.241	-	-	-	-

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

20	BiLSTM	Стандартная обработка, SNWTW	13 MFCC $+\Delta + \Delta^2$	0.43	-	-	-	-	-
21	CVNN	CQT, логарифмическое преобразование магнитуды	Комплексные признаки (магнитуда + фаза)	26.95 ± 3.12	-	-	-	-	-
22	DeepSonar	Анализ активации нейронов в DNN	ACN, TKAN	2.0	-	98.1	-	-	-
25	SSAD (LCNN-big)	8-слойная сверточная сеть (1D-свертка+BatchNorm+PReLU), TCN+dilated convolutions	LPS, LFCC, CQCC	5.31	-	-	-	-	-
	SSAD (SENet12)			6.55	-	-	-	-	-
27	CNN: ResNet+GRU	Обрезка аудио до 4 секунд	2D-DCT над лог-мел спектр-мой	-	0.074	-	-	-	-
29	**ASSERT SENet34**	**CQCC, Logspec**	**Unified Feature Map, Whole Utterance**	**0.13**	**0.003**	**-**	**-**	**-**	**-**
	ASSERT Mean-Std ResNet			0.59	0.016	-	-	-	-
30	DBiLSTM	Разбитие на неперекрывающиеся кадры длиной l с окном Хэмминга, FFT	ELTP, LFCC	0.74	0.008	-	-	-	-
23, 24	CNN	DFT, масштабирование изображений (ImageDataGenerator)	Мел-спектрограммы MFCC	-	-	78	-	-	-

На основе анализа данных, представленных в Таблице 5, можно сделать вывод, что наиболее эффективными методами классификации являются TCN (MFCC) [14], ASSERT (SENet34) [29], CNN (записи Imitation) [17] и RES-EfficientCNN [18], каждый из которых демонстрирует высокие показатели в различных аспектах.

Временная сверточная сеть использует в качестве входных данных MFCC, что позволяет эффективно извлекать ключевые особенности из аудиосигналов. TCN (MFCC) выделяется наилучшим результатом по метрике EER, который составляет всего 0.07%, что свидетельствует о его высокой способности минимизировать ошибки классификации и делает его надёжным для задач, связанных с обработкой аудиосигналов.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

ASSERT (SENet34) также демонстрирует выдающиеся результаты, с EER 0.129% и минимальным значением min t-DCF (0.003), что указывает на его высокую точность и низкую вероятность ошибок, обусловленную использованием архитектуры SENet, которая улучшает извлечение и обработку признаков.

Сверточная нейронная сеть, обученная на записях Imitation, показывает высокие значения точности (Accuracy – 98.5%), Precision (0.997) и Recall (0.997), что делает его предпочтительным выбором для задач, где критически важны как точность, так и полнота классификации. RES-EfficientCNN, в свою очередь, демонстрирует наивысшую точность (Accuracy – 97.61%), что на 3.47% превышает показатель классификатора EfficientCNN, что свидетельствует о его высокой эффективности, вероятно, достигнутой за счёт усовершенствованной архитектуры с остаточными связями.

Таким образом, каждый из этих методов может быть рекомендован для использования в задачах, где критически важны высокая точность, минимальная вероятность ошибок и надежность классификации.

Сравнительный анализ

Анализ представленных в работе методов показывает, что современные подходы к обнаружению аудиодипфейков демонстрируют высокую эффективность. Наиболее успешные методы обнаружения ложных голосовых сигналов включают использование спектральных и кепстральных признаков, таких как MFCC и CQCC.

Эти методы позволяют выявлять артефакты, характерные для синтетических аудиозаписей, благодаря анализу частотных и временных характеристик сигналов. Кроме того, методы машинного обучения, такие как метод опорных векторов и метод K-ближайших соседей, показывают высокую точность классификации, особенно при использовании комбинации признаков, таких как биспектральные и кепстральные коэффициенты.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

Например, Q-SVM и Weighted KNN достигают ROC AUC 0.99, что свидетельствует о почти идеальной способности разделять классы подлинных и поддельных записей.

Нейронные сети, такие как CNN и TCN, также демонстрируют выдающиеся результаты благодаря своей способности автоматически извлекать сложные и иерархические признаки из аудиоданных. Методы на их основе, такие как TCN (MFCC) и RES-EfficientCNN, показывают минимальные значения EER (0.07% и 0.129% соответственно), что делает их одними из самых надежных подходов для обнаружения дипфейков.

Гибридные методы, сочетающие сверточные и рекуррентные нейронные сети, а также механизмы внимания, также показывают высокую эффективность в задачах обнаружения сложных артефактов. Сравнительные характеристики наиболее результативных методов детектирования аудиодипфейков представлены в Таблице 6.

Таблица № 6.

Сводная таблица наиболее эффективных методов в классификации

Исследование	Классификатор	Предобработка	Выделение параметров	EER (%)	min-tDCF	ROC AUC	Accuracy (%)	F1-score	Precision	Recall
1	2	3	4	5	6	7	8	9	10	11
6	Q-SVM	Обрезка до 5с., приведение к моно	Величина и фаза бикогерентности +MFCC & Δ и Δ^2Кепстр	-	-	0.99	-	-	-	-
	Weighted KNN			-	-	0.99	-	-	-	-
18	RES-Efficient CNN	Обрезка до 4с, преобразование в спект-му	Логарифмирование и Z-нормализация спектрограммы	-	-	-	-	97.61	-	-
1	2	3	4	5	6	7	8	9	10	11
14	**TCN**	**Нормализация, удаление шумов и тишины (<40дБ), WAV**	**MFCC**	**0.07**	**-**	**-**	**-**	**-**	**-**	**-**

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

29	ASSERT (SENet34)	CQCC, Logspec	Unified Feature Map, Whole Utterance	0.13	0.003	-	-	-	-	-
9	**SVM (RBF)**	**мел-спектрограммы (преобразование в мел-шкалу, DFT, спектр мощности в dB)**	**Модифицированная ResNet50+LDA**	**0.5**	**-**	**-**	**99.6**	**0.997**	**0.998**	**0.994**
	KNN		MFCC+ LDA	0.36	-	-	79	0.79	0.79	0.78
12	KNN	Отбор wav файлов; кодирование меток	MFCC + novel transfer features	-	-	-	96	0.96	0.97	0.95
17	CNN: Imitation	Гистограммы амплитуд сигнала, горизонтальное отражение (аугментация)	Выделение признаков с помощью CNN	-	-	-	98.5	-	0.997	0.997
11	KNN	Разбиение на кадры, оконная функция Хэмминга	MFCC+Δ+Δ^2+ SC	-	-	-	97.7	-	-	-

Наиболее эффективными из всех представленных в Таблице 6 методами обнаружения ложных голосовых сигналов являются TCN (MFCC) [14] и SVM (RBF) [9]. Эти подходы демонстрируют выдающиеся результаты благодаря особенностям обработки аудиоданных, извлечения признаков и классификации.

В работе [14] используется временная сверточная сеть в сочетании с мел-частотными кепстральными коэффициентами. Основной акцент делается на предобработке аудиоданных, которая включает нормализацию, удаление шумов и тишины, а также преобразование аудиозаписей в формат WAV. Аудиосигналы преобразуются в спектрограммы с помощью библиотеки librosa, после чего полученные параметры передаются на вход временной сверточной сети для классификации подлинных и поддельных записей.

В работе [9] основное внимание уделяется использованию метода опорных векторов с ядром радиальной базисной функции. Предобработка аудиоданных включает преобразование аудиосигналов в мел-спектрограммы, которые лучше соответствуют восприятию звука человеком. Для извлечения

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

признаков используется модифицированная архитектура ResNet50, которая возвращает 2048 глубоких признаков для каждого аудио. Для снижения размерности матрицы признаков применяется линейный дискриминантный анализ.

Оба метода демонстрируют высокую эффективность благодаря тщательной предобработке данных, использованию современных подходов к извлечению признаков и применению мощных классификаторов.

Заключение

В данной статье рассмотрено множество современных методов обнаружения аудиодипфейков, что подчеркивает актуальность и сложность задачи в условиях стремительного развития технологий синтеза речи и искусственного интеллекта. Эффективность классификации подлинных и поддельных голосовых сигналов во многом зависит от качества предобработки данных, выбора ключевых параметров и используемых классификаторов.

Современные подходы, такие как анализ спектральных и кепстральных признаков, а также применение методов машинного обучения, включая опорные вектора и нейронные сети, демонстрируют высокую эффективность. В частности, временные сверточные сети выделяются своей способностью обрабатывать долгосрочные временные зависимости в аудиосигналах, что особенно важно для обнаружения артефактов, характерных для синтетических записей. Благодаря использованию расширенных сверток TCN обеспечивают высокую точность анализа, особенно в сочетании с такими признаками, как MFCC. Это подтверждается минимальными значениями ошибок (EER), что делает TCN одним из наиболее перспективных инструментов для детектирования дипфейков.

Метод опорных векторов также заслуживает внимания, особенно в сочетании с радиальной базисной функцией. SVM демонстрирует высокую

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

точность классификации, эффективно работая с нелинейными данными и успешно разделяя классы подлинных и синтетических записей. Его устойчивость к шуму и способность адаптироваться к сложным условиям делают его важным инструментом в борьбе с аудиодипфейками.

Однако, несмотря на значительные успехи, ключевой проблемой остается постоянное совершенствование технологий синтеза речи и клонирования голоса. Это делает существующие методы уязвимыми к новым типам атак, что требует непрерывного обновления и адаптации алгоритмов. Кроме того, эффективность многих методов напрямую зависит от качества и разнообразия обучающих данных. Недостаточный объем данных или их низкое качество могут существенно снизить точность и надежность моделей, что особенно критично в условиях, где требуется высокая достоверность результатов.

В будущем стоит ожидать появления более совершенных методов, которые будут учитывать эти аспекты и обеспечивать еще более высокую точность и надежность в обнаружении ложных голосовых сигналов. Разработка устойчивых к шуму и адаптивных алгоритмов, а также улучшение интерпретируемости моделей станут ключевыми направлениями для дальнейших исследований.

Исследование проведено при финансовой поддержке Фонда поддержки молодых ученых имени Геннадия Комиссарова в рамках конкурса «Специальная стипендия имени Евгения Касперского».

Литература

1. Yi J., Wang C., Tao J., Zhang X., Zhang C. Y., Zhao Y. Audio Deepfake Detection: A Survey. arXiv. 2023. URL: arxiv.org/pdf/2308.14970.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

2. Voigt I., Boeckmann M., Bruder O., Wolf A., Schmitz T., Wieneke H. A deep neural network using audio files for detection of aortic stenosis. Clinical Cardiology. 2022. V. 45. №6. pp. 657-663.

3. Evsyukov D., Glinscaya A., Kukartsev A., Volneikina E., Kukartseva S. Analyzing the influence of parameters on water quality using logistic regression. BIO Web of Conferences. 2024. V. 130. №03001. 7 p.

4. Алёшин Н.А. Метод Опорных Векторов (SVM). Техника и технологии: теория и практика. Пенза. МЦНС «Наука и Просвещение». 2020. С. 9-11.

5. Gomez-Alanis A., Peinado A.M., Gonzalez J.A., Gomez A.M. A Light Convolutional GRU-RNN Deep Feature Extractor for ASV Spoofing Detection. Interspeech. 2019. pp. 1068-1072.

6. Singh A.K., Singh P. Detection of AI-Synthesized Speech Using Cepstral & Bispectral Statistics. 2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR). 2021. pp. 412-417.

7. Bird J.J., Lotfi A. Real-time detection of ai-generated speech for deepfake voice conversion. arXiv. 2023. URL: arxiv.org/abs/2308.12734.

8. Kumar C.S., Rao G.S.N., Vivek P., Sivaram G., Deepak M.V., Raj G.T. Leveraging Deep Learning Architectures for Deepfake Audio Analysis. Proceedings of 6th International Conference. 2024. V. 19. pp. 266-276.

9. Chakravarty N., Dua M. A lightweight feature extraction technique for deepfake audio detection. Multimedia Tools and Applications. 2024. V. 83. №26. pp. 67443-67467.

10. Khochare J., Joshi C., Yenarkar B., Suratkar S., Kazi F. A deep learning framework for audio deepfake detection. Arabian Journal for Science and Engineering. 2021. V. 47. №3. pp. 3447-3458.

11. Jellali A., Fredj I.B., Ouni K. Pushing the boundaries of deepfake audio detection with a hybrid MFCC and spectral contrast approach. Multimedia Tools and Applications. 2024. V. 84. pp. 20249–20268.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

12. Al-Shamayleh A.S., Riasat H., Alluhaidan A.S., Raza A., El-Rahman S.A., AbdElminaam D.S. Novel transfer learning based acoustic feature engineering for scene fake audio detection. Scientific Reports. 2025. V. 15. №1. 8066 p.

13. Bisogni C., Loia V., Nappi M., Pero C. Acoustic features analysis for explainable machine learning-based audio spoofing detection. Computer Vision and Image Understanding. 2024. V. 249. 104145 p.

14. Firc A., Malinka K., Hanáček P. Deepfake Speech Detection: A Spectrogram Analysis. Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing. 2024. pp. 1312-1320.

15. 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. arXiv. 2021. URL: arxiv.org/abs/2109.00537.

16. Alzantot M., Wang Z., Srivastava M. B. Deep residual neural networks for audio spoofing detection. arXiv. 2019. URL: arxiv.org/abs/1907.00501.

17. Ballesteros D.M., Rodriguez-Ortega Y., Renza D., Arce G. Deep4SNet: deep learning for fake speech classification. Expert Systems with Applications. 2021. V. 184. 115465 p.

18. Subramani N., Rao D. Learning efficient representations for fake speech detection. Proceedings of the AAAI Conference on Artificial Intelligence. 2020. V. 34. №04. pp. 5859-5866.

19. Lei Z., Yang Y., Liu C., Ye J. Siamese Convolutional Neural Network Using Gaussian Probability Feature for Spoofing Speech Detection. Interspeech. 2020. pp. 1116-1120.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

20. Shan M., Tsai T.J. A cross-verification approach for protecting world leaders from fake and tampered audio. arXiv. 2020. URL: arxiv.org/pdf/2010.12173.

21. Müller N.M., Sperl P., Böttinger K. Complex-valued neural networks for voice anti-spoofing. arXiv. 2023. URL: arxiv.org/abs/2308.11800.

22. Wang R., Juefei-Xu F., Huang Y., Guo Q., Xie X., Ma L., Liu Y. Deepsonar: Towards effective and robust detection of ai-synthesized fake voices. arXiv. 2020. URL: arxiv.org/pdf/2005.13770.

23. Пономарёв К.Г., Верещагина Е.А. Применение сверточных нейронных сетей и алгоритмов глубокого обучения для прогнозирования и идентификации голосовых дипфейков // Инженерный вестник Дона. 2025. №.2. URL: ivdon.ru/ru/magazine/archive/n2y2025/9859.

24. Пономарёв К.Г., Верещагина Е.А. Математический аппарат и технологическая инфраструктура системы прогнозирования голосовых дипфейков //Инженерный вестник Дона. 2024. №. 6. URL: ivdon.ru/ru/magazine/archive/n6y2024/9312.

25. Jiang Z., Zhu H., Peng L., Ding W., Ren Y. Self-Supervised Spoofing Audio Detection Scheme. Interspeech. 2020. pp. 4223-4227.

26. Todisco M., Delgado H., Evans N. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. Computer Speech & Language. 2017. V. 45. pp. 516-535.

27. Gao Y., Vuong T., Elyasi M., Bharaj G., Singh R. Generalized spoofing detection inspired from audio generation artifacts. arXiv. 2021. URL: arxiv.org/pdf/2104.04111.

28. Todisco M., Wang X., Vestman V., Sahidullah M., Delgado H., Nautsch A., Yamagishi J., Evans N., Kinnunen T., Lee K.A. ASVspoof 2019: Future horizons in spoofed and fake audio detection. arXiv. 2019. URL: arxiv.org/pdf/1904.05441.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

29. Lai C.I., Chen N., Villalba J., Dehak N. ASSERT: Anti-spoofing with squeeze-excitation and residual networks. arXiv. 2019. URL: arxiv.org/abs/1904.01120.

30. Arif T., Javed A., Alhameed M., Jeribi F., Tahir A. Voice spoofing countermeasure for logical access attacks detection. IEEE Access. 2021. V. 9. pp. 162857-162868.

31. Das R.K., Yang J., Li H. Long Range Acoustic Features for Spoofed Speech Detection. Interspeech. 2019. pp. 1058-1062.

References

1. Yi J., Wang C., Tao J., Zhang X., Zhang C. Y., Zhao Y. Audio Deepfake Detection: A Survey. arXiv. 2023. URL: arxiv.org/pdf/2308.14970.

2. Voigt I., Boeckmann M., Bruder O., Wolf A., Schmitz T., Wieneke H. Clinical Cardiology. 2022. V. 45. №6. pp. 657-663.

3. Evsyukov D., Glinscaya A., Kukartsev A., Volneikina E., Kukartseva S. BIO Web of Conferences. 2024. V. 130. №03001. 7 p.

4. Alyoshin N.A. Tekhnika i tekhnologii: teoriya i praktika [Engineering and technology: theory and practice]. 2020. pp. 9-11.

5. Gomez Alanis A., Peinado A.M., Gonzalez J.A., Gomez A.M. Interspeech. 2019. pp. 1068-1072.

6. Singh A.K., Singh P. 2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR). 2021. pp. 412-417.

7. Bird J.J., Lotfi A. arXiv. 2023. URL: arxiv.org/abs/2308.12734.

8. Kumar C.S., Rao G.S.N., Vivek P., Sivaram G., Deepak M.V., Raj G.T. Proceedings of 6th International Conference. 2024. V. 19. pp. 266-276.

9. Chakravarty N., Dua M. Multimedia Tools and Applications. 2024. V. 83. №26. pp. 67443-67467.

10. Khochare J., Joshi C., Yenarkar B., Suratkar S., Kazi F. Arabian Journal for Science and Engineering. 2021. V. 47. №3. pp. 3447-3458.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

11. Jellali A., Fredj I.B., Ouni K. Multimedia Tools and Applications. 2024. V. 84. pp. 20249–20268.

12. Al Shamayleh A.S., Riasat H., Alluhaidan A.S., Raza A., El Rahman S.A., AbdElminaam D.S. Scientific Reports. 2025. V. 15. №1. 8066 p.

13. Bisogni C., Loia V., Nappi M., Pero C. Computer Vision and Image Understanding. 2024. V. 249. 104145 p.

14. Firc A., Malinka K., Hanáček P. Proceedings of the 39th ACM SIGAPP Symposium on Applied Computing. 2024. pp. 1312-1320.

15. Yamagishi J., Wang X., Todisco M., Sahidullah M., Patino J., Nautsch A., Liu X., Lee K.A., Kinnunen T., Evans N., Delgado H. arXiv. 2021. URL: arxiv.org/abs/2109.00537.

16. Alzantot M., Wang Z., Srivastava M. B. arXiv. 2019. URL: arxiv.org/abs/1907.00501.

17. Ballesteros D.M., Rodriguez Ortega Y., Renza D., Arce G. Expert Systems with Applications. 2021. V. 184. 115465 p.

18. Subramani N., Rao D. Proceedings of the AAAI Conference on Artificial Intelligence. 2020. V. 34. №04. pp. 5859-5866.

19. Lei Z., Yang Y., Liu C., Ye J. Interspeech. 2020. pp. 1116-1120.

20. Shan M., Tsai T.J. arXiv. 2020. URL: arxiv.org/pdf/2010.12173.

21. Müller N.M., Sperl P., Böttinger K. arXiv. 2023. URL: arxiv.org/abs/2308.11800.

22. Wang R., Juefei Xu F., Huang Y., Guo Q., Xie X., Ma L., Liu Y. arXiv. 2020. URL: arxiv.org/pdf/2005.13770.

23. Ponomarev K.G., Vereshchagina E.A. Inzhenernyj vestnik Dona. 2025. №.2. URL: ivdon.ru/ru/magazine/archive/n2y2025/9859.

24. Ponomarev K.G., Vereshchagina E.A. Inzhenernyj vestnik Dona. 2024. №. 6. URL: ivdon.ru/ru/magazine/archive/n6y2024/9312.

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025

25. Jiang Z., Zhu H., Peng L., Ding W., Ren Y. Interspeech. 2020. pp. 4223-4227.

26. Todisco M., Delgado H., Evans N. Computer Speech & Language. 2017. V. 45. pp. 516-535.

27. Gao Y., Vuong T., Elyasi M., Bharaj G., Singh R. arXiv. 2021. URL: arxiv.org/pdf/2104.04111.

28. Todisco M., Wang X., Vestman V., Sahidullah M., Delgado H., Nautsch A., Yamagishi J., Evans N., Kinnunen T., Lee K.A. arXiv. 2019. URL: arxiv.org/pdf/1904.05441.

29. Lai C.I., Chen N., Villalba J., Dehak N. arXiv. 2019. URL: arxiv.org/abs/1904.01120.

30. Arif T., Javed A., Alhameed M., Jeribi F., Tahir A. IEEE Access. 2021. V. 9. pp. 162857-162868.

31. Das R.K., Yang J., Li H. Interspeech. 2019. pp. 1058-1062.

Дата поступления: 4.05.2025

Дата публикации: 12.08.2025

© Электронный научный журнал «Инженерный вестник Дона», 2007–2025