Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н. Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н. Э. Баумана)

ФАКУЛЬТЕТ ФМОП «Факультет международных образовательных программ»

КАФЕДРА ИУ-7 «Программное обеспечение эвм и информационные технологии»

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

«Классификация методов обнаружения образцов голоса, синтезированных с помощью нейронных сетей»

Студент ИУ7И-74Б _____ Ахмад Халид Каримзай

(Подпись, дата)

Руководитель _____ А.С. Кострицкий

(Подпись, дата)

2023 г.

РЕФЕРАТ

Расчетно–пояснительная записка 23 с., 5 рис., 3 табл., 20 ист, 1 прил.

В работе рассматривается понятие синтезированного аудио, существующие типы и методы для генерации такого контента. Также проводится обзор существующих характеристик для изучения аудиозаписей и их классификации. Далее рассматривается понятие систем для обнаружения синтетического звука, включая внутренние компоненты и их взаимодействие.

Ключовые слови: синтезированный аудио, синтетический звук, система обнаружения синтетического звука, классификация синтетического звука.

СОДЕРЖАНИЕ

ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

В настоящей расчетно-пояснительной записке применяют следующие термины с соответствующими определениями.

1) GMM – Gaussian Mixture Model.

2) MFCC – Mel-frequency cepstral coefficients.

3) LFCC – Linear Frequency Cepstral Coefficients.

4) FFT – Fast Fourier Transform.

5) DCT – Discrete cosine transform.

6) CNN – Convolutional Neural Network.

7) STFT – Short Time Fourier Transform.

8) DeepFake – Фейковый контент, созданный с помощью нейронных сетей.

ВВЕДЕНИЕ

Аудио-дипфейки представляют собой категорию звуковых файлов, созданных при помощи глубоких нейронных сетей, способных анализировать и воспроизводить звуковые характеристики настолько реалистично, что созданный контент может звучать естественно и непринужденно. Чаще всего эти технологии применяются для имитации голосов людей, но, кроме того, могут вызывать веселье и забаву. Тем не менее, с увеличением популярности аудио-дипфейков возникают вопросы относительно их злоупотребления с целью распространения дезинформации [1].

Актуальность работы заключается в том, что синтетический или фейковый контент существует уже много лет, однако внимание к контенту, созданному с использованием нейронных сетей, также известного как дипфейк, стало значительным лишь в последние несколько лет. В то время как синтезированные фотографии и видео, порожденные нейронными сетями, привлекли большое внимание, синтетические человеческие голоса тоже достигли выдающегося качества и эффективности. Однако, несмотря на их улучшенную реалистичность и доступность, синтетические голоса также несут в себе существенные риски [2].

В рамках данной научной-исследовательской работы рассмотрим следующие цели и задачи:

1. Синтезирование аудио: Понятие и Типы;
2. Характеристики и особенности аудиоматериала для изучения;
3. Понятие и схема работы системы обнаружения синтетического звука;
4. Классификация и обзор известных методов обнаружения синтезированного звука.

1 Постановка задачи

Основной задачей системы обнаружения поддельного звука (аудио дипфейк) является процесс выявления поддельного звука в речевом потоке. В качестве входных данных используются звуковые сигналы, а на выходе представляются результаты классификации.

На рисунке (1.1) представлено, Формализация задачи обнаружения аудио дипфейков:

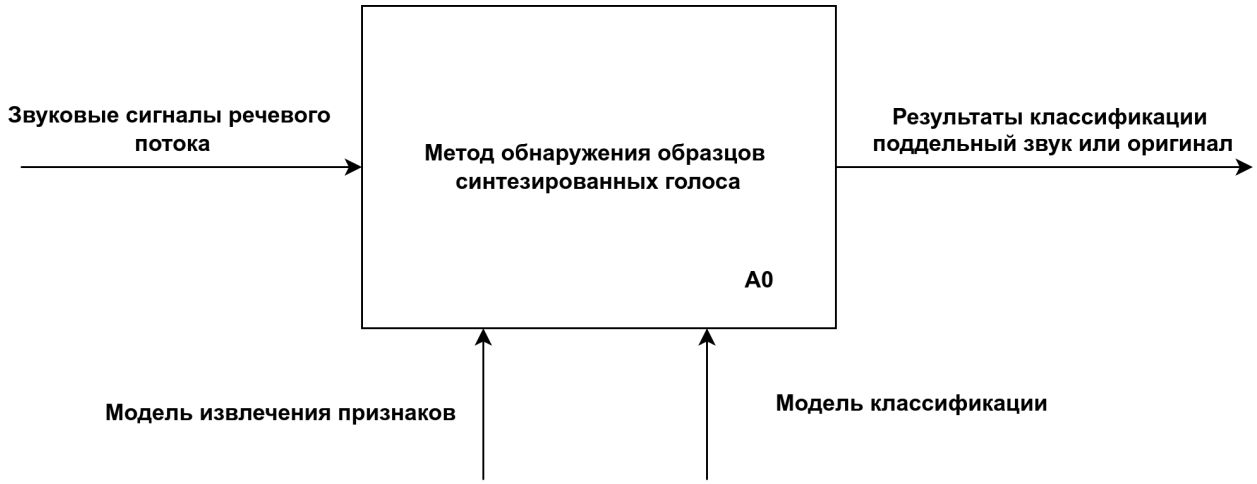

Рис. 1.1. Формализация задачи обнаружения аудио дипфейков

2 Анализ предметной области

Под термином ”Синтезирование голоса” обычно понимается любой аудио-сигнал, важные характеристики которого были изменены при помощи технологий нейронных сетей, сохраняя при этом воспринимаемую естественность. Ранее проведенные исследования в основном выделяли пять видов дипфейкового звука:

1. преобразование текста в речь;
2. преобразование голоса;
3. подделка эмоций;
4. подделка сцен;
5. частично подделка.

также в таблице (6.1) проведено классификация аудио-дипфейков по способу генерации. В первом столбце представлены типы поддельных

дипфейков, во втором столбце характеристики поддельных черт. Третий столбец отражает продолжительность дипфейков, указывая, являются ли они частично или полностью синтезированными. В четвертом столбце указано, что используется нейронная сеть при генерации.

2.1. Преобразование текста в речь

Преобразование текста в речь (TTS) [3] представляет собой широко применяемую технологию, ориентированную на синтез четкой и естественной речи из произвольного текста с использованием моделей, основанных на методах машинного обучения. Современные модели TTS в основном используют глубокие нейронные сети для генерации реалистичной речи, которая максимально приближена к человеческой.

Системы TTS обычно включают в себя два основных модуля: модуль анализа текста и модуль генерации речевых сигналов. Модуль анализа текста разбирает входной текст, определяя тон, интонацию и другие лингвистические аспекты, необходимые для правильной передачи смысла. Затем модуль генерации речевых сигналов создает звуковую последовательность, соответствующую заданному тексту.

2.2. Преобразование голоса

Преобразование голоса (VC) [3], или клонирование голоса в цифровой форме, фокусируется на изменении звучания речи одного говорящего, подражая тембру и просодии другого говорящего, при этом сохраняя содержание оригинального высказывания. Процесс введения в систему VC обычно включает в себя использование естественных высказываний данного говорящего как входных данных.

Существует несколько основных подходов к технологиям VC [4], включая статистический параметрический, частотное искажение и выбор единиц измерения. В частности, статистическая параметрическая модель включает в себя вокодер, аналогичный тому, который используется в статистических параметрических системах синтеза речи (TTS).

2.3. Подделка эмоций

Подделка эмоций [5], также известная как модификация эмоционального тонуса, представляет собой технологию, направленную на изменение акустических характеристик звука с целью создания впечатления изменения эмоционального состояния говорящего. Эта методика фокусируется на манипуляции параметрами, такими как тембр, интонация и темп речи, сохраняя при этом остальные аспекты звуковой информации, такие как личность говорящего и содержание высказывания.

2.4. Подделка сцен

Модификация сцены звучания, более известная как подделка сцены [6], представляет собой метод, направленный на сопоставление акустической обстановки оригинального высказывания с другой звуковой сценой, используя технологии улучшения речи. В этом процессе сохраняются как личность говорящего, так и содержание высказывания, при этом происходит изменение окружающей аудиообстановки.

2.5. Частично подделка

Частичная подделка [7], также известная как модификация части высказывания, представляет собой технику, прицельно изменяющую всего лишь несколько слов в оригинальном высказывании. Этот метод осуществляется путем манипулирования исходными аудиоклипами с использованием подлинных или созданных синтезом звуковых фрагментов. Однако при этом ключевым аспектом является сохранение неизменной личности говорящего.

Таблица 2.1. Класификации аудио дипфейков по способу генерации первая часть

Поддельный тип	Поддельная черта	Поддельная продолжительность	С помощью нейронной сети
Преобразование текста в речь	Речевое содержание	полностью	да

Таблица 2.2. Класификации аудио дипфейков по способу генерации вторая часть

Преобразование голоса	Личность спикера	полностью	да
Подделка эмоций	эмоция спикера	полностью	да
Подделка сцен	Акустическая сцена	полностью	да
Частично подделка	Речевое содержание	частично	да

3 Отличительные признаки аудио для изучения

Извлечение признаков представляет собой ключевой модуль классификатора аудио-дипфейков. Основной целью этого процесса является изучение характерных особенностей путем выделения акустических артефактов из речевых сигналов, которые могут свидетельствовать о наличии поддельных атак. Большое количество исследований подчеркнуло важность определения полезных признаков для эффективного обнаружения дипфейков.

В данной области уделено значительное внимание выявлению полезных функций, способных обнаруживать характерные аспекты поддельных атак. Признаки, использованные в проведенных исследованиях, условно могут быть разделены на три основные категории [8]:

– Спектральные характеристики;

– Просодические характеристики;

– Глубокие характеристики.

3.1. Спектральные характеристики

Спектральные характеристики в анализе звука относятся к характеристикам, которые отражают распределение энергии по различным частотам в сигнале. Эти характеристики вычисляются с использованием математических преобразований, таких как быстрое преобразование Фурье

(FFT), и имеют критическое значение для извлечения существенной информации из аудиосигналов для различных применений. Спектральные характеристики могут быть классифицированы на краткосрочные и долгосрочные в зависимости от временного масштаба, в течение которого они вычисляются.

3.1.1. Кратковременные спектральные характеристики

Кратковременные спектральные характеристики, извлеченные из коротких кадров обычно длительностью 20-30 мс, описывают кратковременную спектральную огибающую, которая включает в себя акустический коррелят тембра голоса. Кратковременные спектральные характеристики вычисляются, главным образом, путем применения кратковременного преобразования Фурье (STFT) к речевому сигналу [9]. При предположении, что речевой сигнал x(t) квазистационарен в течение короткого периода, STFT формулируется следующим образом:

$$X(t,\omega) = |X(t,\omega)|e^{j\phi(\omega)}, \tag{3.1}$$

где $|X(t,\omega)|$, это спектр магнитуд а $\phi(\omega)$ представляет собой фазовый спектр в кадре t и частотный диапазон ω. Спектр мощности определяется как $|X(t,\omega)|^2$.

Кратковременные спектральные характеристики в основном включают кратковременные характеристики, основанные на магнитуде и фазе. Обычно несколько характеристик, базирующихся на магнитуде, напрямую производятся из спектра магнитуды, но большинство из них вычисляются из спектра мощности. Характеристики, основанные на фазе, извлекаются из фазового спектра.

3.1.2. Долгосрочные спектральные характеристики

Кратковременные спектральные признаки не очень хорошо передают временные характеристики траекторий речевых признаков из-за того, что они вычисляются покадрово [10]. Поэтому были предложены долгосрочные спектральные характеристики для получения информации на более широких временных интервалах из речевых сигналов, и исследования показали, что они имеют решающее значение для обнаружения поддельной речи.

Долгосрочные спектральные характеристики охватывают более

продолжительные временные участки речевых сигналов и могут лучше отражать долгосрочные изменения в акустических свойствах речи. Это позволяет системе обнаружения более эффективно анализировать речевые траектории и выявлять особенности, связанные с поддельной речью. Такие характеристики могут включать в себя долгосрочные форманты, изменения в спектральной энергии на протяжении времени и другие параметры, охватывающие более широкие аспекты речевого сигнала [11].

3.2. Просодические характеристики

Просодия относится к несегментарной информации в речевых сигналах, включая ударение на слоге, интонационные паттерны, темп речи и ритм [12]. В отличие от кратковременных спектральных характеристик с короткой продолжительностью, обычно составляющей 20-30 мс, просодические особенности охватывают более длинные сегменты, такие как фоны, слоги, слова и высказывания. Важные просодические параметры включают основную частоту (F0), длительность, распределение энергии, скорость разговора и т.д. Предыдущие исследования [12] по обнаружению поддельного звука в основном рассматривали три основные просодические характеристики:

- Основная частота (F0): Показатель высоты голоса, который варьируется в зависимости от интонации и эмоционального состояния говорящего;
- Длительность: Временной параметр, отражающий продолжительность звуковых сегментов, таких как слоги и слова, в речевом потоке;
- Распределение энергии: Характеризует энергетические аспекты речи и может отражать эмфатическое выделение или особенности интонации.

Эти просодические характеристики могут быть использованы для выделения особенностей в произношении, что делает их важными для обнаружения поддельного звука, где подделка может влиять на натуральные просодические паттерны голоса. Однако они менее чувствительны к канальным эффектам по сравнению со спектральными функциями [3]. Они могут предоставлять дополнительную информацию к спектральным функциям для повышения эффективности обнаружения поддельного звука.

3.3. Глубокие характеристики

Упомянутые выше спектральные характеристики и просодические характеристики представляют собой большую часть ручных функций с сильными и желательными репрезентативными способностями. Однако их конструирование подвержено предвзятостям из-за ограничений представлений, внесенных человеческими разработчиками [13]. Таким образом, глубокие функции вдохновлены для заполнения этого пробела. Глубокие функции изучаются с использованием глубоких нейронных сетей, которые условно можно разделить на три категории: обучаемые спектральные функции, контролируемые функции встраивания и самоконтролируемые функции встраивания.

На рисунке (3.1) представлено, класификации аудио по признаками для обучение:

Рис. 3.1. Классификации аудио по признаками

4 Система обнаружения поддельного звука

В системах обнаружения поддельного звука в качестве классификатора до последних десяти лет применялись статистические методы классификации. Однако с появлением синтезированных аудио-потоков с использованием нейронных сетей, алгоритмы нейронных сетей завоевали большую популярность и показывают меньшую ошибку при классификации [14]. Существует два типа систем обнаружения поддельного звука:

1. Сквозная система - в этом варианте система обнаружения поддельного звука, получает на вход речевой поток;

2. Комбинированная система - в этом варианте система обнаружения поддельного звука, состоит из двух модулей:

   – Модуль извлечения признаков;

   – Модуль классификации.

4.1. Метод с гениализацией признаков

Метод стремится изучить преобразователь, который не изменяет характеристики подлинной речи, а лишь проецирует поддельную речь на другой выход, максимизируя разницу между подлинной и поддельной речью [15]. В данном методе применяется сверточная нейронная сеть (CNN) для обучения преобразователя генерации, а архитектура нейронной сети представлена на рисунке (4.2).

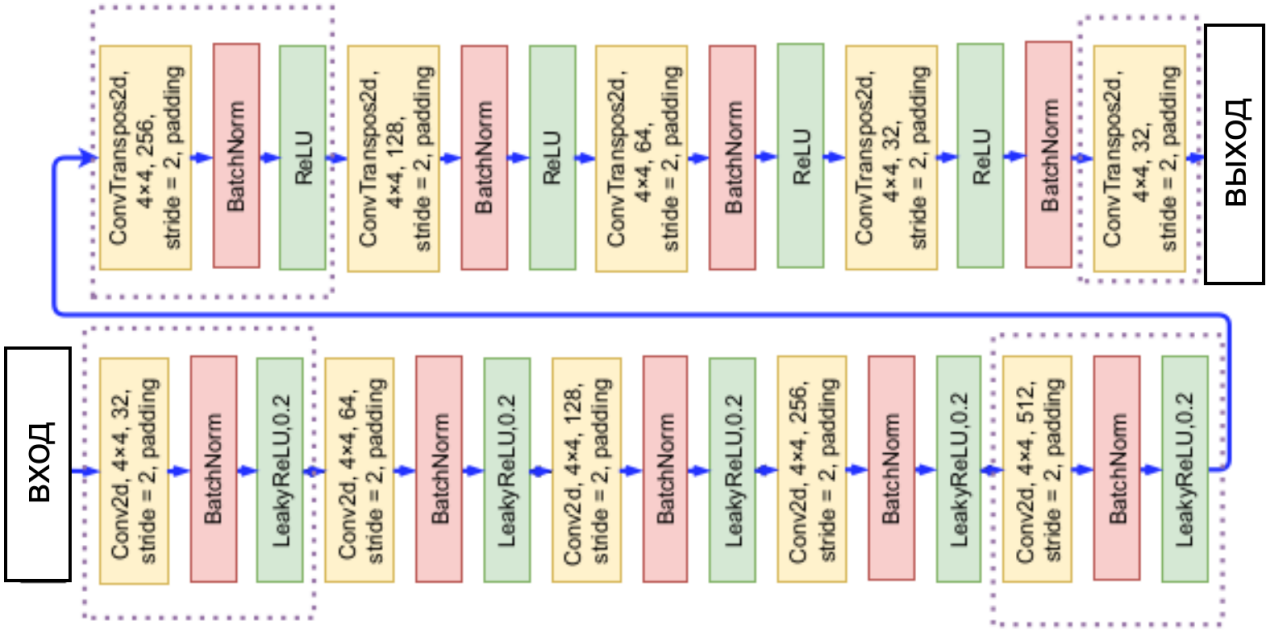

Рис. 4.1. Архитектура преобразователя генерации

Архитектура предлагаемого преобразователя генерации, изображенная на рисунке (4.2), состоит из двух функций:

1. Кодирования;

2. Декодирования.

На этапе кодирования входной сигнал сжимается через несколько последовательных сверточных слоев, а затем результат свертки проходит через дырчатую функцию ReLU. На этапе декодирования процесс кодирования обращается с использованием деконволюции, после чего применяется функция ReLU [15]. Таким образом, преобразователь действует как автокодировщик, изучающий характеристики подлинной речи [16]. Это способствует усилению различий между подлинной и поддельной речью в преобразуемой области. Данный метод является частью системы обнаружения поддельного звука. Он принимает на вход не речевой поток, а признаки для дальнейшей обработки.

4.2. Метод с использованием интегрированного спектрально временного подхода

Структура данного метода может быть описана следующим образом:

– Необработанный сигнал передается в кодер RawNet2 [17], который сначала извлекает признаки, используя предопределенные фильтры.

– Результат передается в остаточные блоки CNN. Затем строятся спектральные и временные графики с использованием внимания графов и их объединения.

– Графики объединяются путем выполнения операции сложения.

Далее применяется внимание к гетерогенному штабелируемому графу и операция максимального графа. Результат объединения используется для дальнейшей классификации. Основным преимуществом модели является относительно небольшое количество обучаемых параметров и высокая производительность. Однако, несмотря на хорошие результаты классификации с предложенным представлением скрытого пространства, эффективное обучение модели с самоконтролем представляет сложность [18].

Все эти операции графическим образом представлены на рисунке 4.2.

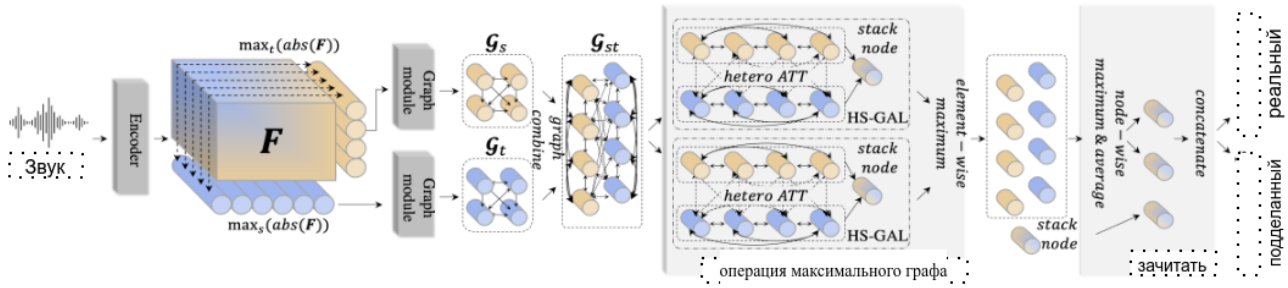

Рис. 4.2. Архитектура метода с использованием интегрированного спектрально-временного внимания

4.3. Метод с использованием трансферного обучения

Для обнаружения поддельной речи авторы предлагают использовать предварительно обученную модель ResNet на мел-спектрограммах в качестве представления звуковых характеристик. Существенным преимуществом этого подхода является более быстрое обучение. Тем не менее, необходимо отметить, что спектрограммы существенно отличаются от обычных изображений по своей структуре, обусловленной природой звука. Поэтому не очевидно, подойдет ли предварительно обученная модель визуального обнаружения объектов хорошо [19]. На рисунке (4.3) представлено, mel-спектрограмма аудио-сигнала:

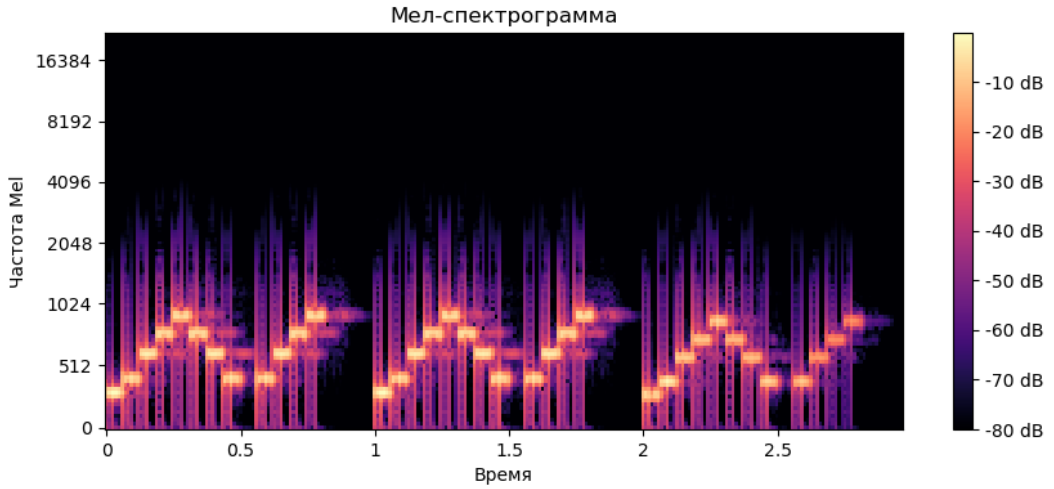

Рис. 4.3. Mel-спектрограмма аудио-сигнала

5 Критерии сравнения методов

Для правильной работы методов проводится обучение, оценка и тестирование, а затем результаты классификации проверяются по различным критериям.

В случае нейронных сетей сначала проводится обучение. Веса между слоями начально присваиваются случайным образом. Затем веса обновляются с использованием алгоритма обратного распространения ошибки, чтобы достичь желаемого результата. После этапа обучения проводится оценка работоспособности модели. Для этого из набора обучающих данных выбираются случайные примеры и проверяется их классификация. В тестировании примеры выбираются не из набора обучающих данных.

Для оценки производительности системы обнаружения в данном исследовании используются две оценочные метрики, обычно применяемые для обнаружения глубокой подделки звука:

1. Равная частота ошибок (EER);
2. Функция затрат на тандемное обнаружение $(mint - DCF)$ [20].

Функция затрат на тандемное обнаружение, предложенная конкурсом ASVspoof [20], описывается следующим образом:

$$P_{\text{ложный}}(\theta) = \frac{\text{количество фальшивых голосов с партитурой} > \theta}{\text{полное количество фальшивых голосов}}, \tag{5.1}$$

$$P_{\text{пропущенный}}(\theta) = \frac{\text{количество настоящих голосов со счетом} \leq \theta}{\text{полное количество истинных голосов}}, \tag{5.2}$$

$$ERR = P_{\text{ложный}}(\theta) = P_{\text{пропущенный}}(\theta), \tag{5.3}$$

$$mint - DCF = \min_{\theta}\{C_0 + C_1 P_{\text{пропущенный}}(\theta) + C_2 P_{\text{ложный}}(\theta)\}, \tag{5.4}$$

где EER обозначает частоту ошибок, при которой частота ложных срабатываний $P_{\text{ложный}}(\theta)$ и частота пропущенных срабатываний $P_{\text{пропущенный}}(\theta)$ равны, а θ_{ERR} обозначает пороговое значение, при котором $P_{\text{ложный}}(\theta)$ и $P_{\text{пропущенный}}(\theta)$ равны. Чем меньше EER, тем выше производительность системы обнаружения. Чем меньше $mint - DCF$, тем лучше обобщаемость системы обнаружения и тем меньше влияние на производительность системы автоматической проверки говорящего [18].

6 Классификация причисленных методов обнаружения синтетического звука

Для класификации причисленных трех методов обнаружения поддельной речи, можно использовать следующие Критерии:

1. К1 — Точность обнаружение поддельной речи, для этой цели рассматриваем оценка ошибки EER относителя корпус данных ASVSpoof [20], данное значение настолько меньше, настолько выше точность работы метода;

2. К2 — Устойчивость к различным типам поддельной речи, для этой цели рассматриваем оценку функция затрат на тандемное обнаружение $mint - DCF$ относителя корпус данных ASVSpoof [20], данное значение настолько меньше, настолько выше точность работы метода относительно различных видов синтезирования звука;

3. К3 — Принимает ли на вход аудиосигнал;

4. К4 — Требует ли обучение модель классификации.

Таблица 6.1. Класификации причисленных методов обнаружения синтетического звука

Метод	К1	К2	К3	К4
Метод с гениализацией признаков	4.07%	0.102	Нет	Да
Метод с использованием трансферного обучения	8.09%	0.2116	Нет	Да
Метод с использованием интегрированного спектрально-временного подхода	0.83%	0.0275	Да	Да

ЗАКЛЮЧЕНИЕ

В рамках данной работы было проведено изучение системы обнаружения аудиодипфейков, анализ предметной области, рассмотрение признаков аудио для изучения и обучения, а также проведена классификация и обзор методов обнаружения аудиодипфейков.

В итоге можно описать структуру работы системы обнаружения синтетического звука следующим образом:

1. На вход поступает аудиозапись.
2. Модель извлечения признаков предварительно обрабатывает запись.
3. В некоторых методах модели используют признаки для обучения, а в некоторых других просто принимают звуковую речь.
4. Модель классификации использует признаки или саму звуковую речь для обучения и распознавания.

Значительную роль в эффективности работы модели играют признаки, передаваемые в модель для классификации, а также сам процесс работы модели классификации.

Следует отметить, что огромное значение для работоспособности и результатов модели по классификации имеют корпусы данных, используемые для обучения и извлечения признаков, которое обнаруживается в процессе обучения модели.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Generalization of Audio Deepfake Detection. / T. Chen [и др.] // Odyssey. — 2020. — С. 132—137.
2. Add 2022: the first audio deep synthesis detection challenge / J. Yi [и др.] // ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). — IEEE. 2022. — С. 9216—9220.
3. Spoofing and countermeasures for speaker verification: A survey / Z. Wu [и др.] // speech communication. — 2015. — Т. 66. — С. 130—153.
4. An overview of voice conversion and its challenges: From statistical modeling to deep learning / B. Sisman [и др.] // IEEE/ACM Transactions on Audio, Speech, and Language Processing. — 2020. — Т. 29. — С. 132—157.
5. EmoFake: An initial dataset for emotion fake audio detection / Y. Zhao [и др.] // arXiv preprint arXiv:2211.05363. — 2022.
6. SceneFake: An Initial Dataset and Benchmarks for Scene Fake Audio Detection / J. Yi [и др.] // arXiv preprint arXiv:2211.06073. — 2022.
7. Half-truth: A partially fake audio detection dataset / J. Yi [и др.] // arXiv preprint arXiv:2104.03617. — 2021.
8. *Sahidullah M.*, *Kinnunen T.*, *Hanilçi C.* A comparison of features for synthetic speech detection. — 2015.
9. Spoofing speech detection using high dimensional magnitude and phase features: the NTU approach for ASVspoof 2015 challenge. / X. Xiao [и др.] // Interspeech. — 2015. — С. 2052—2056.
10. Synthetic speech detection using temporal modulation feature / Z. Wu [и др.] // 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. — IEEE. 2013. — С. 7234—7238.
11. *Das R. K.*, *Yang J.*, *Li H.* Long range acoustic and deep features perspective on ASVspoof 2019 // 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). — IEEE. 2019. — С. 1018—1025.
12. *Kinnunen T.*, *Li H.* An overview of text-independent speaker recognition: From features to supervectors // Speech communication. — 2010. — Т. 52, № 1. — С. 12—40.

13. LEAF: A learnable frontend for audio classification / N. Zeghidour [и др.] // arXiv preprint arXiv:2101.08596. — 2021.

14. *Almutairi Z.*, *Elgibreen H.* A review of modern audio deepfake detection methods: challenges and future directions // Algorithms. — 2022. — Т. 15, № 5. — С. 155.

15. Light convolutional neural network with feature genuinization for detection of synthetic speech attacks / Z. Wu [и др.] // arXiv preprint arXiv:2009.09637. — 2020.

16. Extracting deep bottleneck features using stacked auto-encoders / J. Gehring [и др.] // 2013 IEEE international conference on acoustics, speech and signal processing. — IEEE. 2013. — С. 3377—3381.

17. Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification / J.-w. Jung [и др.] // arXiv preprint arXiv:1904.08104. — 2019.

18. AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks / J.-w. Jung [и др.]. — 2021. — arXiv: 2110 . 01200 [eess.AS].

19. Audio Spoofing Verification using Deep Convolutional Neural Networks by Transfer Learning / R. T. P [и др.]. — 2020. — arXiv: 2008.03464 [eess.AS].

20. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection / J. Yamagishi [и др.]. — 2021. — arXiv: 2109.00537 [eess.AS].

ПРИЛОЖЕНИЕ А