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 19 с., 2 рис., 1 табл., 2 ист, 1 прил.

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

ВВЕДЕНИЕ

Аудио дипфейки относятся к тем группам аудио-файлов, которые изготовленные методами глубокими нейронными сетями, которые изучают движения звуко-записей до такой степени, что они могут воспроизводить реалистично звучащий поддельный звук. Обычно их используют для имитации голосов людей, хотя иногда они могут быть забавными, такими методами можно злоупотреблять для распространения дезинформации, которая может привести к пагубным последствиям, некоторые из способов с помощью которых могут использоваться методы дипфейки, включают онлайн-травлю, влияние на политические движения и выдачу себя за людей.

Синтетический или фейковый контент существует уже много лет, но контент, созданный с помощью нейронных сетей, то есть дипфейк, существует всего несколько лет. В то время как фотографии и видео, синтезированные с помощью искусственного интеллекта, получили большое внимание, синтетические человеческие голоса также претерпели значительные изменения, достигнув беспрецедентного качества и эффективности. Однако растущий реализм и доступность синтетических человеческих голосов также таят в себе значительные риски. В то время как методы обнаружения изображений и видео, синтезированных с помощью нейронных сетей, были тщательно изучены, методы обнаружения синтетических человеческих голосов получили меньше внимания и недостаточно развиты.

В данной научной-исследовательской работе рассмотрим следующие задачи:

1. Понятие цифровое аудио;
2. Процесс преобразование аналоговый голоса в цифровой и обратно;
3. Понятие Синтезированное аудио и известны подходы генерации синтетического звука;
4. Классификация и обзор известных методов обнаружения синтезирование звука.

1 Аудио Дипфейк

Под терменом дипфейковым звуком обычно понимается любой звук, важные атрибуты которого были изменены с помощью технологий нейронной сети, но при этом сохраняется его воспринимаемая естественность. Предыдущие исследования в основном включали пять видов дипфейкового звука:

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

также в таблице (1.1), проведено классификация аудио дипфейков по способу генерации.

1.1. Преобразование текста в речь

Преобразование текста в речь (TTS) [8], широко известное, направлено на синтез понятной и естественной речи, заданной любым произвольным текстом, с использованием моделей, основанных на машинном обучении. Модели TTS могут генерировать реалистичную и похожую на человеческую речь с развитием глубоких нейронных сетей [1]. Системы TTS в основном включают в себя модули анализа текста и генерации речевых сигналов. Существует два основных метода генерации речевых сигналов [13]:

– конкатенативный;
– статистический параметрический TTS.

1.2. Преобразование голоса

Преобразование голоса (VC) [8] относится к клонированию голоса человека в цифровом виде. Она направлена на то, чтобы изменить тембр и просодию речи данного говорящего на тембр речи другого говорящего, в то время как

содержание речи остается прежним. Вводом в систему VC является естественное высказывание данного говорящего. Существует примерно три основных подхода к технологиям VC [18]: статистический параметрический, частотное искажение и выбор единиц измерения. Статистическая параметрическая модель также имеет вокодер, который аналогичен таковому в статистических параметрических TTS [22].

1.3. Подделка эмоций

Подделка эмоций [25] направлена на изменение звука таким образом, что меняется эмоция речи, в то время как другая информация остается прежней, например, личность говорящего и содержание речи. Изменение эмоций голоса часто приводит к изменению семантики.

1.4. Подделка сцен

Подделка сцены [27] включает в себя сопоставление акустической сцены исходного высказывания с другой сценой с помощью технологий улучшения речи, при этом личность говорящего и содержание речи остаются неизменными.

1.5. Частично подделка

частично подделка [28] фокусируется на изменении лишь нескольких слов в высказывании. Фальшивое высказывание создается путем манипулирования исходными высказываниями с помощью подлинных или синтезированных аудиоклипов. Спикер оригинального высказывания и фейковых клипов — один и тот же человек. Синтезированные аудиоклипы, сохраняя при этом личность говорящего неизменной.

Таблица 1.1 – Класификации аудио дипфейков по способу генерации

Поддельный тип	Поддельная черта	Поддельная продолжительность	С помощью нейронной сети
Преобразование текста в речь	Личность спикера, Речевое содержание	полностью	да
Преобразование голоса	Личность спикера	полностью	да
Подделка эмоций	эмоция спикера	полностью	да
Подделка сцен	Акустическая сцена	полностью	да
Частично подделка	Речевое содержание	частично	да

2 Отличительные признаки аудио для изучения

Извлечение признаков является ключевым модулем классификатора аудио дипфейков. Целью извлечения признаков является изучение отличительных признаков путем выделения звуковых поддельных артефактов из речевых сигналов. Большое количество усилий показало важность полезных функций для обнаружения поддельных атак. Признаки, использованные в исследованиях, можно условно разделить на четыре категории [48]:

– спектральные особенности;

– просодические особенности;

– глубокие особенности.

2.1. спектральные особенности

Спектральные характеристики в анализе звука относятся к характеристикам, которые показывают распределение энергии по различным частотам в сигнале, эти спектральные характеристики вычисляются с помощью математических преобразований, таких как быстрое преобразование Фурье (FFT), и имеют решающее значение для извлечения значимой информации из аудиосигналов для различных применений, спектральные характеристики могут быть классифицированы на краткосрочные и долгосрочные в зависимости от масштаба времени, в течение которого они вычисляются.

2.1.1. Кратковременные спектральные особенности

Кратковременные спектральные характеристики, извлеченные из коротких кадров, обычно длительностью 20-30 мс, описывают кратковременную спектральную огибающую, включающую акустический коррелят тембра голоса. Кратковременные спектральные характеристики вычисляются главным образом путем применения кратковременного преобразования Фурье (STFT) к речевому сигналу [52]. Учитывая речевой сигнал x(t), предполагается, что он квазистационарен в течение короткого периода. STFT речевого сигнала x(t) формулируется следующим образом:

$$X(t,\omega) = |X(t,\omega)|e^{j\phi(\omega)} \tag{2.1}$$

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 состоят из кратковременных характеристик, основанных на магнитуде и фазе. Обычно несколько характеристик, основанных на магнитуде, непосредственно выводятся из спектра магнитуд, но большинство из них выводятся из спектра мощности. Характеристики, основанные на фазе, выводятся из фазового спектра.

2.1.2. Долгосрочные спектральные особенности

Кратковременные спектральные признаки не очень хорошо улавливают временные характеристики траекторий речевых признаков из-за того, что они вычисляются покадрово [60]. Поэтому были предложены долгосрочные спектральные характеристики для получения информации на большом расстоянии из речевых сигналов, и исследования показали, что они имеют решающее значение для обнаружения поддельной речи [49].

2.2. Просодические особенности

Просодия относится к несегментарной информации речевых сигналов, включая ударение на слоге, интонационные паттерны, темп речи и ритм [78]. В отличие от кратковременных спектральных характеристик с короткой продолжительностью, обычно составляющей 20-30 мс, они охватывают более длинные сегменты, такие как телефоны, слоги, слова, высказывания и т.д. Важные просодические параметры включают основную частоту (F0), длительность, распределение энергии, скорость разговора и т.д. Предыдущие исследования [78] по обнаружению поддельного звука в основном рассматривали три основные просодические характеристики:

– F0;

– продолжительность;

– энергию.

Эти функции менее чувствительны к канальным эффектам по сравнению со спектральными функциями [8]. Они могут предоставлять дополнительную

информацию к спектральным функциям для повышения эффективности обнаружения поддельного звука.

2.3. Глубокие возможности

Вышеупомянутые спектральные особенности и просодические особенности - это почти все функции, созданные вручную, обладающие сильными и желательными репрезентативными способностями. Однако их дизайн испорчен предубеждениями из-за ограничений представлений ручной работы [51]. Таким образом, глубокие функции мотивированы заполнить пробел. Глубокие функции изучаются с помощью глубоких нейронных сетей, которые можно условно разделить на: обучаемые спектральные функции, контролируемые функции встраивания и самоконтролируемые функции встраивания.

На рисунке (2.1) представлено, класификации аудио по признаками для обучение:

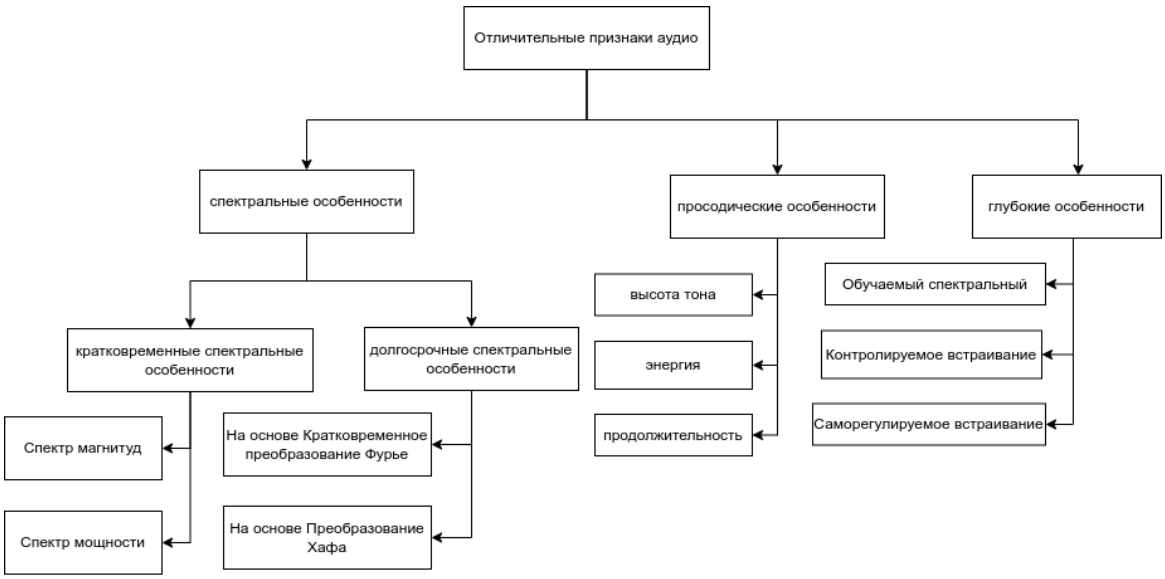

Рис. 2.1 – Классификации аудио по признаками

3 Методы обнаружения поддельного звука (Аудио Дипфейк)

В системах, которые используются для обнаружения поддельного звука, важнешей фактором является особенности аудио для изучения и внутренний классификатор очень важен для глубокого распознавания аудио, целью которого является изучение высокоуровневого представления функций входного интерфейса и моделирование превосходных возможностей обнаружения. Внутренний классификатор которые часто используются для обнаружения аудио дипфейков в основном делятся на две категории:

- Статистические методы;
- С применением глубоких нейронных сетей.

3.1. Статистические методы обнаружения Аудио Дипфейк

Под терменом статистические методы в статей пишут алгоритмы машиное обучение, алгоритмы машиное обучение изучают характеристики аудио и для решение задачой обнаружения аудио дипфейк, используется алгоритмы бинарное классификации. В связм с этим, для обнаружения фальшивой речи было использовано множество классических подходов к классификации паттернов. Самые популярные методы обнаружения аудио дипфейк с применением машиное обучение являются:

- Машина опорных векторов (SVM);
- Гауссовые моделей смеси (GMM).

3.1.1. Машина опорных векторов (SVM)

SVM - это контролируемый метод обучения, который основывается в основном на двух предположениях [1]:

1. Преобразование данных в многомерное пространство может свести сложные проблемы классификации со сложными поверхностями принятия решений к более мелким проблемам, которые могут быть решены путем их линейного разделения;

2. Только обучающие шаблоны вблизи поверхности принятия решений обеспечивают наиболее чувствительную детали для классификации.

Так как, проблема обнаружения аудио дипфейков представляет собой бинарную классификацию с линейно разделяемыми векторами $x_i \in R^n$, в качестве поверхности принятия решения, используемой для классификации паттерна как принадлежащего к одному из двух классов, используется гиперплоскость H_0. Если x это случайный вектор n * R, тогда мы определяем:

$$f_{(x)} = w.x + b \tag{3.1}$$

В формуле (.) это скалярное произведение, набор всех x-векторов, удовлетворяющих уравнению $f_{(x)} = 0$, обозначается как H_0. Предполагая две гиперплоскости, H1 и H2, расстояние между ними называется их границей, которую можно представить следующим образом:

$$\begin{cases} H_1 = \{x \in R^n | f_{(x)} > 0\} \\ H_2 = \{x \in R^n | f_{(x)} < 0\} \end{cases} \tag{3.2}$$

Гиперплоскость решения H0 зависит от векторов, ближайших к двум параллельным гиперплоскостям, называемым опорными векторами. Запас должен быть максимальным, чтобы получить классификатор, который не очень адаптирован к обучающим данным.

На рисунке (3.1) представлено, машина опорных векторов:

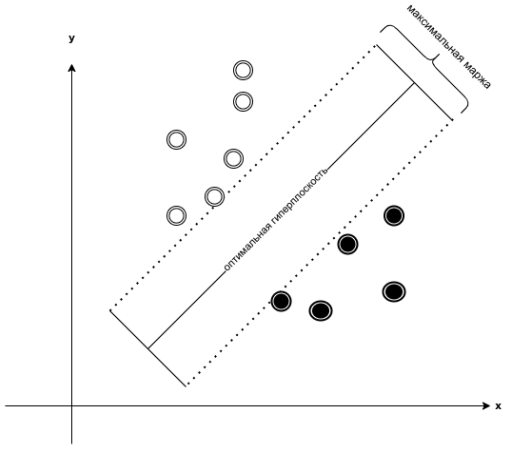

Рис. 3.1 – Машина опорных векторов (SVM)

В многих статей по класификации аудио дипфейков используется метод SVM, который помогает классифицировать аудиосигналы дипфейк. Он хорошо

работает с четким разделением выборок и эффективен в средах с высокой размерностью. Он использует подмножество точек обучения в функции принятия решения, что делает его эффективным с точки зрения памяти.

Методы с применением SVM

В статей [110] и [68], используют машина опорных векторов в качестве классификатора, в статей [110] предполагают, что классификаторы SVM по своей сути устойчивы к атакам с искусственным подменой сигналов. Однако очень трудно определить точную природу атак с подменой в практических сценариях. В статей [68] предложили одноклассовый классификатор SVM, обученный только использованию подлинных высказываний для классификации реальных и поддельных голосов, который хорошо подходит для неизвестных атак с использованием подделки.

3.1.2. Гауссовые моделей смеси (GMM)

Модель гауссовой смеси (GMM), как следует из названия, представляет собой смесь нескольких гауссовых распределений. Речевые признаки представлены в виде векторов в n-мерном пространстве. Распределение этих векторов признаков представлено смесью гауссовых плотностей.
Для n-мерного вектора признаков x функция плотности смеси для класса s с параметром модели λ^s определяется как [2]:

$$p(x|\lambda^s) = \sum_{i=1}^{\mathcal{M}} \alpha_i^s f_i^s(x) \tag{3.3}$$

Функция плотности смеси представляет собой взвешенную линейную комбинацию M унимодальных гауссовых плотностей компонентов $f_i^s(.)$. Каждая функция гауссовой плотности $f_i^s(.)$ параметризуется вектором среднего μ_i^s и ковариационной матрицей Σ_i^s с использованием:

$$f_i^s(x) = \frac{1}{\sqrt{(2\pi)^n|\Sigma_i^s|}} \exp\left(-\frac{1}{2}(x-\mu_i^s)^T\Sigma_i^{s-1}(x-\mu_i^s)\right) \tag{3.4}$$

где Σ_i^s является ковариационной матрией и $(\Sigma_i^s)^{-1}$ называется обратной ковариационной матрией.

Методы с применением GMM

В статей [32] и [114], в качестве классификатора используется GMM. в статей [32] используется в качестве классификатора человеческих и синтезированных голосов. в статей [114], предложят метод, который определяет разницу между реальной и преобразованной речью, используя логарифмическое отношение правдоподобия, основанное на модели GMM для реальной и преобразованной речи.

3.2. Методы с применением глубоких нейронных сетей

Классификаторы новейших систем обнаружения аудио дипфейков в основном основаны на методах глубокого обучения, которые значительно превосходят классификаторы на основе SVM и GMM благодаря их мощным возможностям моделирования [145]. На сегодняшний день в большинство методах обнаружения аудио дипфейков, которые в качестве классификатора применяются глубокие нейронные сети, используют сверточные нейронные сети и глубокой остаточной сети.

3.2.1. Сверточные нейронные сети

Часто когда речь идет о нейронной сети, то в большинство случях это сверточные нейронные сети, мелспектрогр

ЗАКЛЮЧЕНИЕ

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Deepfake audio detection via MFCC features using machine learning / A. Hamza [и др.] // IEEE Access. — 2022. — Т. 10. — С. 134018—134028.
2. *Jothilakshmi S.*, *Gudivada V.* Chapter 10 - Large Scale Data Enabled Evolution of Spoken Language Research and Applications // Cognitive Computing: Theory and Applications. Т. 35 / под ред. V. N. Gudivada [и др.]. — Elsevier, 2016. — С. 301—340. — (Handbook of Statistics). — DOI: https://doi.org/10.1016/bs.host.2016.07.005. — Режим доступа, URL: https://www.sciencedirect.com/science/article/pii/S0169716116300463.

ПРИЛОЖЕНИЕ А