Инженерный вестник Дона, №6 (2024)
ivdon.ru/ru/magazine/archive/n6y2024/9312

Математический аппарат и технологическая инфраструктура системы прогнозирования голосовых дипфейков

К.Г. Пономарёв, Е.А. Верещагина

Институт математики и компьютерных технологий Дальневосточного федерального университета, Владивосток

Аннотация: В статье рассмотрены математические модели по сбору и обработке голосового контента, на основании которых разработана принципиально-логическая схема системы прогнозирования синтетических голосовых дипфейков. Проведены эксперименты выбранных математических формул и наборов библиотек языка программирования «python», позволяющих проводить в режиме реального времени анализ звукового контента в организации. Рассмотрены программные возможности нейронных сетей по выявлению голосовых дипфейков и сгенерированной синтетической (искусственной) речи и определены основные критерии исследования голосовых сообщений. По результатам проведенных экспериментов сформирован математический аппарат, необходимый для положительных решений задач по выявлению голосовых дипфейков. Сформирован перечень технических стандартов, рекомендованных для сбора голосовой информации и повышению качества информационной безопасности в организации.

Ключевые слова: нейронные сети, выявление голосовых дипфейков, информационная безопасность, синтетическая голосовая речь, голосовые дипфейки, технические стандарты сбора голосовой информации, алгоритмы выявления аудио дипфейков, клонирование голоса, сверточные и рекуррентные нейронные сети, мел-спектрограммы

Введение. Генерация голосовых дипфейков путем применения социальной инженерии, методов спуфинг и фишинг-атак является современным инструментом злоумышленника. По своей сути, такие голосовые дипфейки базируются на реальных сообщениях людей, собранных различными способами сбора данных: телефонный разговор, мессенджеры быстрых сообщений, социальные сети или цифровые развлекательные сервисы. Речь человека становится ценностью, требующей новых подходов в области защиты информационной безопасности [1]. В сфере услуг клонирование голоса широко применяется для создания программных роботов колл-центров, адаптированной озвучки актеров в кинотеатрах, а также обучению иностранным языкам с учетом локального акцента произносящего. Клонирование голоса может также производить кибератаки на инфраструктуру и на сотрудников целевой организации, и, следовательно, несет угрозу информационной безопасности. Ранние теоретические заметки

исследований проблематики применения нейронных искусственных сетей для выявления инцидентов безопасности и аномальных действий инфраструктуры организации опубликованы в отдельной статье [2].

Векторами атаки потенциально являются мессенджеры и телеграмм аккаунты сотрудников компании, в том числе информационные системы, применяющие механизмы биометрической аутентификации. Ключевым принципом создания аудио дипфейков является применение и накопление базы данных реальных голосов жертв. Способы накопления голосовых наборов проводятся в развлекательной форме или целенаправленно по телефонному звонку, записывающего голос жертвы. Отметим наиболее известные зарубежные платные и условно-бесплатные цифровые сервисы «PlayHT», «VoiceAI», «Listnr», «Murf.AI», «Lovo», «Resemble.AI», позволяющие клонировать собственный голос пользователя с применением функционала состязательной или рекуррентной нейронной сети. Данные сервисы предлагают доработать родной голос пользователя под акцент, лингвистику выбранного языка или под манеру разговора публичного человека, и включить в звуковой контент эмоции собеседника [3].

По технологии искусственного интеллекта термин «deepfake» определен из двух слов: «deep learning» глубинное обучение и «fake» подделка. Методология данного термина включает математические методы машинного обучения для создания правдоподобных голосовых подделок. Следовательно, доказана потребность в создании системы прогнозирования голосовых дипфейков, обеспечивающая защиту от новых видов кибератак.

Постановка задачи исследования. Для создания информационной системы прогнозирования аудио дипфейков требуется разработать принципиально-логическую схему обработки голосового контента в режиме реального времени и включить механизмы обработки звуковых данных из внешних информационных ресурсов. В связи со спецификой исследуемой

области и задачи результаты анализа должны быть доступны после обработки аналитику безопасности для принятия быстрых решений к предотвращению кибератаки или попытки реализации злоумышленником социальной инженерии.

Математические модели в системе прогнозирования с наибольшей точностью должны определять реалистичность голосового произношения в анализируемой аудиозаписи, и определять возможную генерацию искусственным путем сравнения с базой знаний фейковых сообщений (обучающие выборки), а также особых меток (артефактов) в аудиозаписях.

Обзор и сравнение нейронных сетей позволит ускорить процесс изучения голосовых сообщений, а также организовать процедуру самообучения системы прогнозирования. Прототип системы должен иметь функционал искусственного интеллекта и консолидировать наиболее точные математические модели анализа и обработки аудиозаписей. Особое внимание должно быть уделено клонированной речи, созданной на профессиональном уровне специализированными нейронными сетями в связи со сложностью их выявления. Подобные типы задач в научных источниках рассматриваются узконаправленно по каждому анализируемому алгоритму и не объединяются в единую информационную систему прогнозирования.

Решение задачи в условиях неопределенности анализируемого голосового контента. Первичный блок системы должен быть представлен функционалом сбора и обработки голосовых данных, поступающих в организацию. Для этого используем метод переноса реального голоса в цифровой вид.

Все входящие голосовые сообщения обрабатываются с помощью автокодировщика (Encoder) и представляются в цифровом формате. Полученные данные в сжатом виде извлекает декодировщик (Decoder) по

следующим характеристикам: тембр, частота, что сказано человеком, и соответственно формируют образ «целевого человека» [4].

Декодировщик синтезирует по двум аудиозаписям новый результат, используя программный вокодер или другими словами синтезатор речи, позволяющего оцифровать голосовой поток новой звуковой записи путем оконного преобразования Фурье по следующей формуле (1):

$$F(w,t)=\int_{-\infty}^{\infty} f(\tau)W(\tau-t)e^{-iw\tau}d\tau \qquad (1)$$

где $W(\tau-t)$ оконная весовая функция, позволяющая вести работу по спектру звука.

Дискретное преобразование Фурье, используемое в алгоритмах цифровой обработки аудио сигналов, в математических операциях свертки применяется следующая формула (2):

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

где $k=0...N-1$

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

Для подробного изучения входящих голосовых сообщений используем различные виды спектрограмм, позволяющих визуализировать звуковой сигнал цифрового формата в графическом виде по частоте и временной шкале [5]. Спектрограмма позволяет работать по визуальным изображениям аудиозаписи, обозначая цветом, мощность звукового сигнала по затраченному промежутку времени. Дополнительно добавим формулу (3) в математический аппарат системы прогнозирования аудиофейков; единицу

измерения высоты звука – мел, которая учитывает психофизическое восприятие звука человеком с логарифмической зависимостью от частоты:

$$mel = 1127.01048 * \ln(1 + \frac{fred}{700}) \quad (3)$$

Приступим к эксперименту и на тестовой звуковой записи в формате .wav применим созданный математический аппарат по обработке голосовых сообщений на высокоуровневом языке программирования «Python».

Применим библиотеку «librosa» для графической визуализации голосовой записи и построим графическое изображение аудиозаписи. Библиотека «os» обеспечит работу со звуковыми файлами. Построим график зависимости в отношении амплитуды голосового сообщения, и затраченного на воспроизведение звука на определенной частоте. Результат построения, представляющий визуальную структуру аудиозаписи, указан на рис. 1:

```
import librosa
import os
dir='c:/audiotest'
file=dir+'/sample.wav'
signal, sr=librosa.load(file, sr = 22050)
print(signal)
```

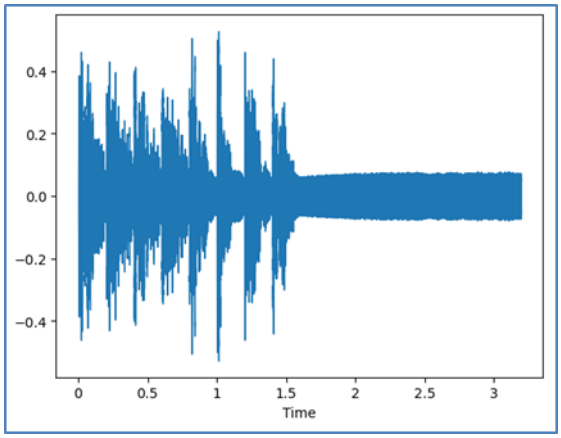

Рис. 1. – Простой метод визуализации тестовой аудиозаписи

Наиболее информативная спектрограмма тестовой выборки формируется через функции stft () и specshow (), которые позволяют построить график зависимости от уровня громкости звукового сигнала во времени. Результат указан на рис. 2. Каждая входящая звуковая выборка обрабатывается, и формируется уникальное изображение сигнала. Таким образом, организован сбор и накопление набора данных реальных и искусственных (синтетических) аудиозаписей. Спектрограмма является базовым инструментом работы со звуком, и имеет весовые показатели, характеризующие голосовые сообщения.

```
X = librosa.stft(signal)
s = librosa.amplitude_to_db(abs(X))
librosa.display.specshow(s, sr=sr, x_axis = 'time', y_axis='linear')
plt.pyplot.colorbar()
```

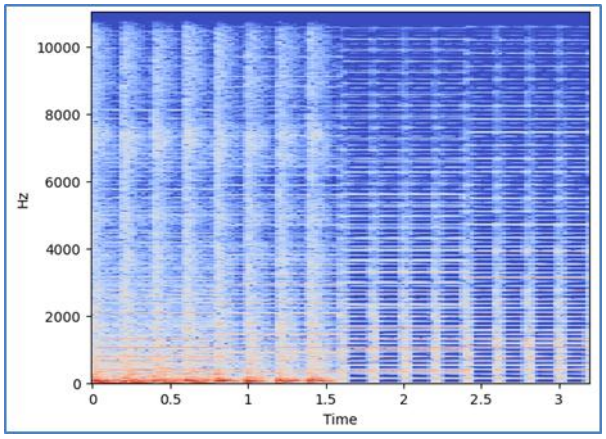

Рис. 2. - Спектрограмма тестовой аудиозаписи

Усложним задачу тестового эксперимента и построим мел-спектрограмму, где вместо частоты в герцах построим график в отношении измеримой величины мел (единица высоты звука) во времени [6]. Используем функцию melspectrogram () и создадим графическое изображение тестового аудиосигнала на рис. 3.

```
mel_spectrogram = librosa.feature.melspectrogram(y=signal, sr=sr)
plt.figure(figsize=(10, 4))
```

librosa.display.specshow (librosa.power_to_db(mel_spectrogram, ref=np.max), sr=sr, hop_length=512, y_axis="mel", x_axis="time")

plt.colorbar (format="%+2.0f dB")

plt.title ("Мел-спектрограмма")

plt.tight_layout ()

plt.show ()

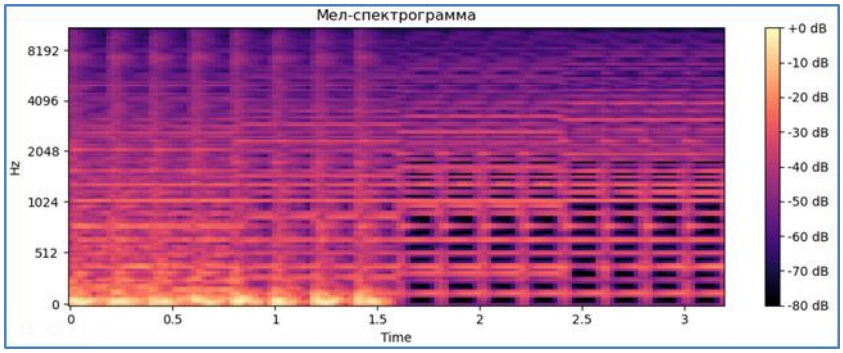

Рис. 3. – Мел-спектрограмма тестовой аудиозаписи

Для анализа тембральных аспектов звука применим мел-кэпстральные коэффициенты, что позволяет обеспечить идентификацию лингвистического содержания и спектральной энергии по каждому временному окну. Применим функцию librosa.feature.mfcc() для формирования массива коэффициентов.

mfccs = librosa.feature.mfcc(y=signal, sr=sr, n_mfcc = 40, hp_length=512)

По результатам анализа тестовой аудиозаписи построена матрица мел-кэпстральных коэффициентов Mel-frequency cepstral coefficients (далее MFCC), представленные следующим образом:

array([[-1.27967926e+02, -1.03908722e+02, -1.22817139e+02, ..., -2.49799377e+02, -2.34608398e+02, -1.92846222e+02], [7.85245514e+01, 9.29010849e+01, 1.04652451e+02, ..., 1.01514683e+01, 2.79959869e+01, 7.69037933e+01], [1.73341293e+01, 1.67533474e+01, 1.21460094e+01, ..., -5.07151871e+01, -4.21031418e+01, -2.86569290e+01], ..., [5.13937235e+00, 2.83761549e+00, -3.22143745e+00, ..., 6.50576115e+00, 9.42087364e+00,

1.15799751e+01], [2.93866277e-01, -3.16372663e-02, 4.88240033e-01, ..., 2.00537729e+00, 3.55618072e+00, 6.57195520e+00], [-4.05332041e+00, -4.58375549e+00, 4.04561329e+00, ..., -1.54050913e+01, -1.48462286e+01, -1.08517132e+01]], dtype=float32).

Проанализировать голосовое сообщение возможно по яркости звука и определить центр масс звукового сообщения. С помощью функции librosa.feature.spectral_centroid () построим спектральный центройд на рис. 4 и выявим характерные особенности голосового сообщения, определяя расположение центр масс звука в тестовой аудиозаписи [7].

```
cent = librosa.feature.spectral_centroid(y=signal, sr=sr)
plt.figure (figsize= (15, 5))
plt.semilogy (cent.T, label='Спектральный центройд')
plt.ylabel ('Гц')
plt.legend ()
```

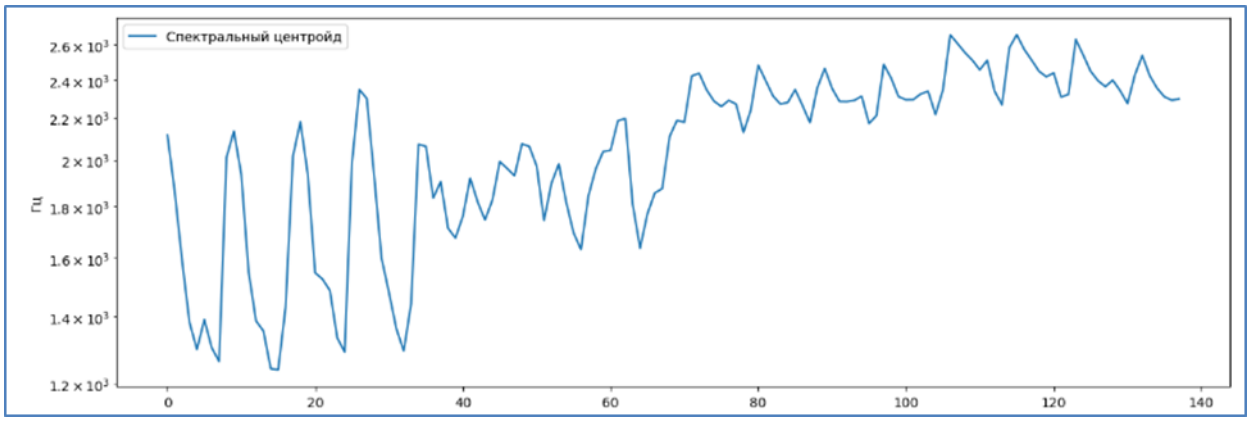

Рис. 4. – Спектральный центройд тестовой аудиозаписи

Полученные результаты эксперимента позволяют их использовать для второго модуля системы прогнозирования аудио дипфейков – оценка оцифрованной аудиозаписи и подготовка результата. Данный блок должен выполнять экспертную оценку аудиозаписей и предоставлять результат аналитику безопасности. Для этого используем рекуррентные и генеративно-состязательные нейронные сети.

Системы управления негативных событий и инцидентами информационной безопасности Security Information and Event Management (далее SIEM) требуют расширения функционала выявления аномалий сетевой активности и применения функционала оценки аудио контента. Функционал прогнозирования в SIEM-системах может быть представлен дополнительным модулем оценки голосовых сообщений, что позволит своевременно обеспечить защиту аудиоконтента в организации и обеспечить сбор и накопление базы знаний [8].

Входные данные представлены изображениями спектрограмм и мел-спектрограмм, позволяющими использовать функционал искусственных нейронных сетей для обучения и накопления баз знаний голосовых сообщений по изображениям.

При создании системы прогнозирования аудио-дипфейков в SIEM используем критерии надежности программного обеспечения Чидамберома и Кемерерома [9] и применим все ранее использованные методики для формирования самостоятельной базы знаний. Наиболее важным критерием надежности отметим Lack of Chesionin Methods (далее LCOM), подтверждающий необходимость расширения функционала создаваемой системы с максимально возможным математическим аппаратом, что впоследствии позволит расширить программный инструмент для исследований голосовых сообщений.

Обратимся к научным источникам, ведущим исследования нейронных сетей для выявления аудио-дипфейков. Сверточная нейронная сеть позволяет достичь точности выявления аудио-дипфейка до 99 % с ошибочными срабатываниями на аудио дипфейк не более 1 % [10]. Существует возможность обучить создаваемую во втором модуле системы нейронную сеть по публичным наборам данных: The M-AILABS Speech, Baidu Silicon Valley AI Lab cloned audio, Fake oR Real (FoR), AR-DAD: Arabic Diversified

Audio, H-Voice, ASV spoof 2021 Challenge, FakeAVCeleb и ADD. Отметим наличие множества экземпляров голосовых сообщений на различных языках мира по указанным наборам данных, позволяющих предварительно обучить создаваемую нейронную сеть.

Усилим систему выявления через создание простой модели генеративно-состязательной нейронной сети Generative Adversarial Networks (далее GAN), где генератором случайных значений определим сверточную нейронную сеть G с учетом поступающих реальных значений X_1 и будем формировать случайные пары чисел z_1, z_2. Дискриминатор D позволяет оценить входные данные на реальность поступающих данных: 0 – аудио дипфейк или 1 – реальные данные. Следовательно, генератор G настроен на максимальную ошибку в решениях дискриминатора D, а, соответственно, D максимально настроен на выявление случайно сгенерированных данных генератором G.

Приступим к проведению эксперимента и создадим простую модель сети GAN путем применения библиотеки «PyTorch». Для этого загрузим библиотеку «torchvision» и «torchvision.transforms» для обработки изображений. Нейронная сеть будет использовать вычислительные мощности центрального процессора Central Processing Unit или графического процессора Graphics Processing Unit. Укажем ключевые классы нейронной сети GAN для проведения успешного эксперимента. Для независимости вычислений от мощности компьютера создадим объект «gadget».

```
install -c pytorch torchvision=0.5.0
import torchvision
import torchvision.transforms as transforms
gadget = ""
if torch.cuda.is_available():
gadget = torch.device("cuda")
```

```
else:
gadget = torch.device("cpu")
```

Для загрузки изображений используем функцию transform с возможностью преобразования в тензор библиотеки «PyTorch». Диапазон значений функции transform.ToTensor представлен от 0 до 1.

```
transform = transforms.Compose(
[transforms.ToTensor(), transforms.Normalize((0.5,), (0.5,))])
```

Обучающие данные выборки загружаем набор данных MNIST с набором уже полученных мел-спектрограмм. Для работы с изображениями можно применить набор данных CIFAR-10 для классификации небольших изображений.

```
tr_set = torchvision.datasets.MNIST( root=".", train=True, download=True, transform=transform)
```

Отобразим полученные результаты через методы cpu() и detach() для отделения тензора и получим результат.

```
gen_samples = gen_samples.cpu().detach()
for i in range(16):
    ax = plt.subplot(4, 4, i + 1)
    plt.imshow(gen_samples[i].reshape(28, 28), cmap="colors")
    plt.xticks([])
    plt.yticks([])
```

По результатам проведения 50 эпох на рис. 5, получаем наиболее правдоподобное изображение мел-спектрограммы, сгенерированной состязательно-генеративной сетью GAN. Изображения мел-спектрограмм формируют базу знаний фейковых аудио-данных и реальных голосовых сообщений. В этом заключен принцип глубокого обучения нейронной сети, что позволяет самостоятельно генерировать правдоподобные изображения.

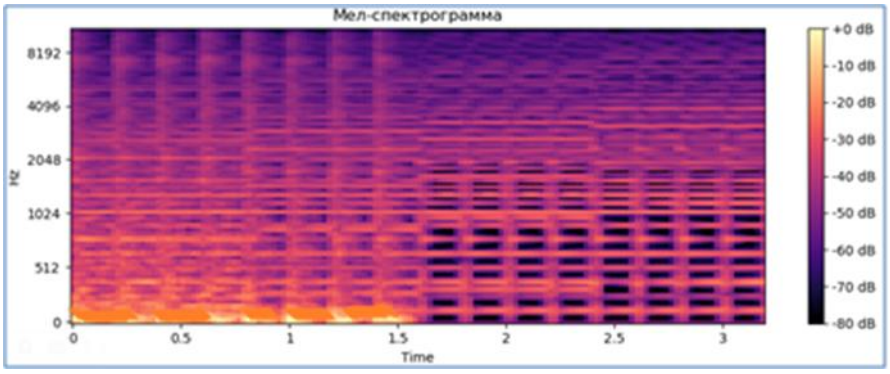

Рис. 5. – Мел-спектрограмма, сгенерированная генеративно-состязательной нейронной сетью GAN

Выводы. Сделаем вывод, что выбранная для эксперимента нейронная сеть данного вида позволяет выявлять голосовые дипфейки, а также их правдоподобно генерировать в случае накопления базы знаний. По результатам проведенных экспериментов на рис. 6 построим принципиально-логическую схему создаваемой системы прогнозирования аудио - дипфейков и определить 3 главных модуля программного обеспечения (сбор и обработка, оценочный модуль и итоговый модуль) с критерием оценки LCOM.

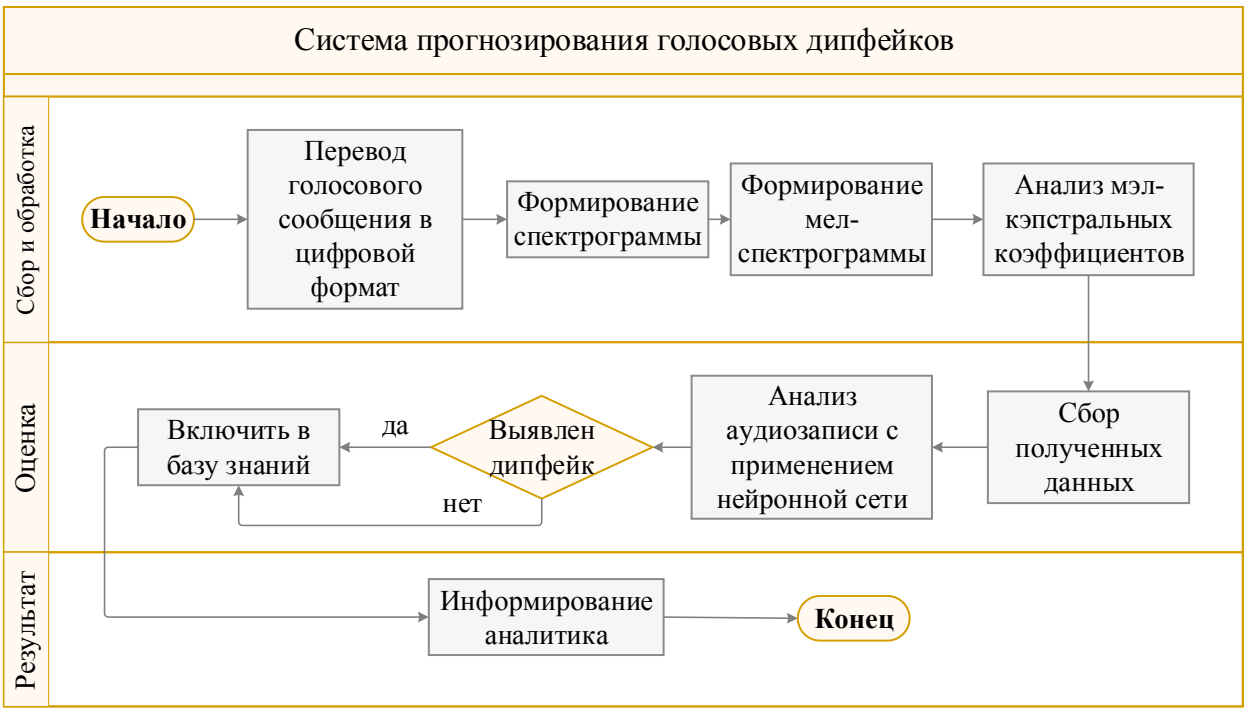

Рис. 6 – Принципиально-логическая схема системы прогнозирования аудио-дипфейков

Для модуля сбора и обработки данных целесообразно автоматизировать первичный сбор данных через технологию чат-ботов мессенджеров Телеграмм, WhatsApp, и технологию веб-хук по протоколу Application Programming Interface. При использовании в организации IP-телефонии целесообразно организовать сбор данных через протокол Voice over IP. Для управления речевым потоком и сжатия данных в целях их дальнейшей передачи и наиболее быстрой обработки в организации целесообразно использовать аудиокодек «SoundStream» [11].

Литература

1. Добробаба М.Б. Дипфейки как угроза правам человека // Lex Russica. 2022. №11. URL: cyberleninka.ru/article/n/dipfeyki-kak-ugroza-pravam-cheloveka

2. Пономарёв К.Г., Верещагина Е.А. Методы развития систем управления информацией и событиями безопасности с применением искусственных нейронных сетей // Сборник избранных статей научной сессии ТУСУР, 2023. С. 17-20. URL: elibrary.ru/item.asp? id=54623902

3. Пантюхин Д.В. Нейронные сети синтеза речи голосовых помощников и поющих автоматов // Речевые технологии Speech Technologies. 2021. №3-4. URL: cyberleninka.ru/article/n/neyronnye-seti-sinteza-rechi-golosovyh-pomoschnikov-i-poyuschih-avtomatov

4. Мосин Е.Д., Белов Ю.С. Модель вариационного автокодировщика для жанровой генерации музыки // E-Scio. 2023. №6. URL: cyberleninka.ru/article/n/model-variatsionnogo-avtokodirovschika-dlya-zhanrovoy-generatsii-muzyki

5. Белоножко П.Е., Белов Ю.С. Реализация вокодера для системы преобразования текста в речь на основе RNN модификации модели WaveNet

// E-Scio. 2023. №6. URL: cyberleninka.ru/article/n/realizatsiya-vokodera-dlya-sistemy-preobrazovaniya-teksta-v-rech-na-osnove-rnn-modifikatsii-modeli-wavenet

6. Волохов В.А., Махныткина О.В., Мещеряков И.Д., Шуранов Е.В., Методические указания к выполнению лабораторных работ по курсу «Цифровая обработка сигналов», Университет ИТМО, 2021. – 60 с.

7. Болдышев А.В., Медведева А.А., Прохоренко Е.И., Гайворонская Д.И. Построение спектрограмм звуковых сигналов на основе субполосных представлений // Экономика. Информатика. 2024. №1. URL: cyberleninka.ru/article/n/postroenie-spektrogramm-zvukovyh-signalov-na-osnove-subpolosnyh-predstavleniy

8. Пономарёв К.Г., Верещагина Е.А. Основные аспекты развития систем управления событиями и инцидентами информационной безопасности с применением методологии других направлений наук // Инженерное дело на Дальнем Востоке России, 2023. С 227-231. URL: elibrary.ru/item.asp?id=50278181

9. С.И. Носков, И.В. Овсянников, А.П. Медведев Агрегированный нелинейный критерий оценки надежности программного обеспечения // Инженерный вестник Дона, 2024, № 5. URL: ivdon.ru/ru/magazine/archive/n5y2024/9183

10. Zaynab Almutairi, Hebah Elgibreen A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions // Algorithms. - 04.2022. №15. URL: doi.org/10.3390/a15050155

11. Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, Marco Tagliasacchi SoundStream: An End-to-End Neural Audio Codec // ResearcheGate. 2021. URL: researchgate.net/publication/353066582_SoundStream_An_End-to-End_Neural_Audio_Codec

References

1. Dobrobaba M.B. Dipfejki kak ugroza pravam cheloveka. Lex Russica, 2022. №11. URL: cyberleninka.ru/article/n/dipfeyki-kak-ugroza-pravam-cheloveka

2. Ponomarjov K.G., Vereshhagina E.A. Metody razvitija sistem upravlenija informaciej i sobytijami bezopasnosti s primeneniem iskusstvennyh nejronnyh setej. Sbornik izbrannyh statej nauchnoj sessii TUSUR. 2023. pp. 17-20. URL: elibrary.ru/item.asp?id=54623902

3. Pantjuhin D.V. Nejronnye seti sinteza rechi golosovyh pomoshhnikov i pojushhih avtomatov. Rechevye tehnologii. Speech Technologies. 2021, № 3-4. URL: cyberleninka.ru/article/n/neyronnye-seti-sinteza-rechi-golosovyh-pomoschnikov-i-poyuschih-avtomatov

4. Mosin E.D., Belov Ju.S. Model' variacionnogo avtokodirovshhika dlja zhanrovoj generacii muzyki. E-Scio. 2023, №6. URL: cyberleninka.ru/article/n/model-variatsionnogo-avtokodirovschika-dlya-zhanrovoy-generatsii-muzyki

5. Belonozhko P.E., Belov Ju.S. Realizacija vokodera dlja sistemy preobrazovanija teksta v rech' na osnove RNN modifikacii modeli WaveNet. E-Scio. 2023, №6. URL: cyberleninka.ru/article/n/realizatsiya-vokodera-dlya-sistemy-preobrazovaniya-teksta-v-rech-na-osnove-rnn-modifikatsii-modeli-wavenet

6. Volohov V.A., Mahnytkina O.V., Meshherjakov I.D., Shuranov E.V., Metodicheskie ukazanija k vypolneniju laboratornyh rabot po kursu «Cifrovaja obrabotka signalov» [Guidance for performing laboratory work on the course "Digital signal processing"], Universitet ITMO, 2021, p. 60.

7. Boldyshev A.V., Medvedeva A.A., Prohorenko E.I., Gajvoronskaja D.I. Postroenie spektrogramm zvukovyh signalov na osnove subpolosnyh predstavlenij, Jekonomika. Informatika. 2024, №1. URL: cyberleninka.ru/article/n/postroenie-spektrogramm-zvukovyh-signalov-na-osnove-subpolosnyh-predstavleniy

8. Ponomarjov K.G., Vereshhagina E.A. Inzhenernoe delo na Dal'nem Vostoke Rossii, 2023, pp. 227-231. URL: elibrary.ru/item.asp?id=50278181

9. S.I. Noskov, I.V. Ovsjannikov, A.P. Medvedev. Inzhenernyj vestnik Dona, 2024, № 5. URL: ivdon.ru/ru/magazine/archive/n5y2024/9183

10. Zaynab Almutairi, Hebah Elgibreen A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions. Algorithms. 2022, №15. URL: doi.org/10.3390/a15050155

11. Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, Marco Tagliasacchi SoundStream: An End-to-End Neural Audio Codec. ResearcheGate. 2021. URL: researchgate.net/publication/353066582_SoundStream_An_End-to-End_Neural_Audio_Codec.

Дата поступления: 16.04.2024

Дата публикации: 5.06.2024