Представлена новая ИИ-модель для борьбы с голосовым мошенничеством

Ученые Института AIRI и МТУСИ предложили новую модель детекции поддельных сгенерированных голосов под названием AASIST3. Представленная архитектура вошла в топ-10 лучших решений международного соревнования ASVspoof 2024 Challenge. Модель применима для противодействия голосовому мошенничеству и повышению безопасности систем, использующих голосовую аутентификацию.

Системы голосовой биометрии (ASV) помогают идентифицировать людей на основе их голосовых характеристик. Их используют для аутентификации пользователей при проведении финансовых транзакций и эксклюзивном контроле доступа в смарт-устройствах, а также в противодействии телефонному мошенничеству нового поколения.

Модели распознавания голоса могут быть уязвимы к состязательным атакам, когда определенным образом настроенное небольшое изменение входного аудио приводит к значительному изменению результатов работы модели, для человека же оно незаметно или незначительно. В поиске способов обойти преграды систем безопасности, злоумышленники научились генерировать синтетический голос с помощью преобразования текста в речь (text-to-speech, TTS) и преобразования голоса (voice conversion, VC). Для эффективного противодействия таким атакам необходимо внедрение систем защиты от подмены голоса.

ИИ-модель AASIST для анализа аудиоряда была продемонстрирована коллективом ученых из Южной Кореи и Франции в 2021 году и показала высокую надежность, подтвержденную многочисленными исследованиями. В то же время, с бурным развитием генеративного ИИ после 2022 года ей перестало хватать качественного функционала для обнаружения синтетических голосов. Использовав AASIST в качестве базы, команда «Доверенные и безопасные интеллектуальные системы» AIRI и команда НИО «Интеллектуальные решения» МТУСИ при участии аспиранта Сколтеха сформировала новую архитектуру для выявления поддельных синтезированных голосов.

Применение сети Колмогорова-Арнольда (KAN), дополнительных слоев и предварительного обучения, лучшего feature extractor, а также специальных обучающих функций, позволило улучшить производительность модели более чем в два раза по сравнению с базовым решением. Кроме того, созданная модель демонстрирует лучшую обобщающую способность к новым видам атак.

"Важно использовать современные нейросети для противодействия голосовому спуфингу, потому что злоумышленники постоянно совершенствуют свои инструменты. Технологии TTS и VC позволяют создавать синтетические голоса, которые уже сейчас очень трудно отличить от настоящих. Преимущество KAN-сетей заключается в их способности учитывать контекст и знания о голосовых данных, позволяя более эффективно различать подлинный голос и его подделку. Такие сети не только распознают подделки с высокой точностью, но и способны адаптироваться к новым типам угроз. Внедрение подобных передовых методов существенно повышает уровень безопасности и защищенности от атак, основанных на подмене голоса"

Олег Рогов, руководитель научной группы “Доверенные и безопасные интеллектуальные системы” AIRI

Задачу голосового антиспуфинга можно решать с помощью двух подходов. Первый — бинарная классификация того, является ли речь в аудио подлинной человеческой или искусственно сгенерированной. Второй — в связке с системой голосовой биометрией, когда необходимо разрешить авторизацию при предъявлении подлинного голоса спикера А, но не при подаче речи спикера Б или искусственной речи спикера А.

Процесс создания модели и выбора подхода к обучению носил итеративный характер: исследователи проверяли разные гипотезы, выбирали лучшие и старались объединить подходы так, чтобы усилить метрики качества, например, EER (уровень, при котором частота ошибки первого рода равна частоте ошибки второго рода) и t-DCF, которая взвешенно учитывает вклады ошибок при разных сценария авторизации (для обоих метрик — чем меньше, тем лучше).

На валидационных данных нам удалось достичь t-DCF 0.2657 в сравнение с 0.5671 у обычного AASIST. На тестовых данных (спикеры и типы атак не были представлены в обучающей и валидирующих выборках), наши модели показали t-DCF 0.5357 и EER 22.67% для закрытого сценария (нельзя использовать дополнительные данные и предобученные модели) и t-DCF 0.1414 и EER 4.89% для открытого сценария соревнования.

“AASIST3 демонстрирует потенциал для практического применения в различных сферах, включая финансовый сектор и телекоммуникации. Основная цель разработки — противодействие голосовому мошенничеству и повышение безопасности систем, использующих голосовую аутентификацию.

Интеграция в бизнес может осуществляться различными способами, от внедрения отдельного программного решения до встраивания в существующие системы безопасности через API. Потребность в подобных технологиях высока, учитывая растущую угрозу атак с использованием синтетических голосов”

Грач Мкртчан, Руководитель НИО «Интеллектуальные решения» МТУСИ

Полный текст статьи доступен по ссылке https://arxiv.org/pdf/2408.17352. Работа была представлена на одной из наиболее известных научных конференций в сфере автоматической обработки речи “Interspeech 2024”, прошедшей на о. Кос, Греция.

Вопросы: pr@airi.net, pr@mtuci.ru

Научно-исследовательский институт AIRI *— автономная некоммерческая организация, занимающаяся фундаментальными и прикладными исследованиями в области искусственного интеллекта. На сегодняшний день более 180 научных сотрудников AIRI задействовано в исследовательских проектах Института для работы совместно с глобальным сообществом разработчиков, академическими и индустриальными партнерами.*

МТУСИ – *Московский технический университет связи и информатики (МТУСИ) – ведущее высшее учебное заведение Центральной России по подготовке специалистов в сфере ИТ, информационной безопасности, телекоммуникаций, радиотехники, телевидения и цифровой экономики, подведомственное Министерству цифрового развития, связи и массовых коммуникаций РФ. С момента основания, в 1921 году, история университета насчитывает уже более 100 лет.*