УДК 004.934

ОБНАРУЖЕНИЕ СЛЕДОВ АУДИОМОНТАЖА ПРИ СПУФИНГ-АТАКЕ НА СИСТЕМУ ГОЛОСОВОЙ БИОМЕТРИИ

Самородова М.Э. (университет ИТМО)
Научный руководитель – Чирковский А.Д. (ООО «ЦРТ»)

Аннотация

Целью работы является изучение проблемы нестационарных спуфинг-атак в задаче голосового антиспуфинга. Для повышения качества работы системы антиспуфинга при предъявлении атак, созданных с помощью аудиомонтажа, в работе исследуется создание собственной базы данных, состоящей из нестационарных спуфинг-атак. Актуальность работы обусловлена тем, что в настоящее время слабо исследованы спуфинг-атаки, которые злоумышленники способны совершить с помощью аудиомонтажа.

Введение

С каждым годом увеличивается спрос на голосовые технологии, так как повышается их роль во многих аспектах жизни человека. Например, распознавание дикторов в онлайн-банкинге, определение ключевых слов в устройствах «умного дома» и т.д. Но параллельно с данными технологиями естественным образом совершенствуются и методы их взлома.

Под спуфинг-атакой (spoof) подразумевается попытка взлома голосовых биометрических систем, основанная на фальсификации голосовых характеристик другого человека [1]. Существует множество различных методов, которые используются для создания спуфинг-атак, среди них можно выделить атаки, совершённые с помощью синтеза речи (TTS), преобразования голоса (VC) или повторного воспроизведения записи целевого диктора. Для обнаружения подобного рода атак система верификации обычно работает в тандеме с системой антиспуфинга, детектирующей попытки предъявления злоумышленниками поддельных аудиозаписей. Однако, в то время как существующие системы антиспуфинга могут обнаруживать полностью поддельные высказывания, в которых целые аудиосигналы генерируются с помощью применения алгоритмов TTS и/или VC, существует необходимость в их адаптации или расширении к сценарию нестационарных спуфинг-атак. Под нестационарной спуфинг-атакой подразумевается аудиозапись, полученная с помощью аудиомонтажа истинных высказываний целевого диктора.

Существует множество возможных мотивов, по которым злоумышленники могут прибегнут к созданию спуфинг-атаки по средствам аудиомонтажа [2]. Например, злоумышленнику может быть достаточно замены или удаления конкретного слова или фразы в имеющемся высказывании для изменения его смысла. Так, добавление частицы отрицания «не» полностью меняет семантику утверждения в речи. В том числе до сих пор распространены текстозависимые голосовые биометрические системы. Примером таковой может являться верификация человека в банке, где необходимо назвать кодовое слово или шифр, состоящий из цифр. Таким образом, злоумышленник может произвести аудиомонтаж из имеющихся у него образцов речи целевого диктора для получения необходимой фразы. При этом данный метод фальсификации, возможно, один из самых простых в исполнении, поскольку его также могут выполнять непрофессионалы, используя свободно доступные инструменты редактирования звука, такие как Audacity или Ocean audio. Однако, несмотря на простоту создания, данная атака может представлять серьезную угрозу, поскольку обычно системы антиспуфинга используют операции агрегирования оценок с течением времени, и в этом случае короткие сегменты монтажа будут иметь мало влияния на итоговое решение при классификации образца на живую речь (genuine) и спуфинг-атаку (spoof).

Основная часть

В настоящее время в открытом доступе нет баз, в которых в качестве спуфинг-атак выступают частично поддельные высказывания, полученные с помощью аудиомонтажа

подлинных аудиозаписей целевого диктора, поэтому основной частью работы является создание базы данных.

На первом этапе создания спуфинг-атак имеющиеся высказывания целевых дикторов обрабатываются с помощью системы распознавания речи для получения разметки, включающей в себя начало и конец произнесения каждого слова. Целевые записи генерируются в трех возможных сценариях: одинарное, двойное и многократное сращивание аудиозаписей. Под одинарным подразумевается объединение двух фрагментов, то есть имеется одно место склейки аудиофрагментов. Под двойным – вставка фрагмента в исходное высказывание, при этом будет наблюдаться две точки на итоговой аудиозаписи, соответствующие местам склейки. Многократное же сращивание, соответственно, будет иметь n возможных мест склейки аудиофрагментов. Потенциально, чем больше мест монтажа на записи, тем легче будет происходить детектирование спуфинг-атак, то есть самыми сложными атаками являются атаки, полученные при одинарном сращивании.

Таким образом, генерация спуфинг-атак осуществляется в автоматическом режиме, длина каждого фрагмента берется случайным образом из нормального распределения. Начало фрагмента также определяется случайным образом по разметке, полученной с помощью распознавания речи, и соответствует началу произнесения какого-либо слова. Чтобы учесть возможные ошибки разметки, для определения начала и конца аудиофрагмента берутся слова продолжительностью не меньше 0,4 секунд.

Кроме того, при генерации данных необходимо учесть, что злоумышленник может использовать как одну запись целевого диктора, так и несколько. При этом сценарий с одной записью является более сложным для детектирования, поскольку в таком случае внешние условия не будут отличаться, что повысит качество атаки. Поэтому для каждого целевого диктора были получены спуфинг-атаки в каждом из трех рассматриваемых сценариях (одинарное, двойное и многократное сращивание) при объединении как одной аудиозаписи, так и нескольких, если таковые имелись для рассматриваемого диктора. При генерации базы данных также сохранялась разметка по местам аудиомонтажа, в которой место склейки характеризовалось небольшим фрагментом продолжительностью 200 мс.

После получения базы данных необходимо обучить модель с добавлением этих новых данных. При обучении в модель подаются случайные фрагменты аудиозаписей продолжительностью равной 3 секунды. Спуфинг-атаки из новой базы обрабатываются несколько иным образом с помощью разметки по местам аудиомонтажа: сначала случайным образом определяется одно из мест склейки, далее также случайным образом фиксируется его расположение в сегменте продолжительностью 3 секунды. Для повышения разнообразия обучающих данных место склейки было либо у краев сегмента, либо где-то в центре, что также определялось случайно. Таким образом гарантировалось, что спуфинг-атаки из нового набора данных всегда будут содержать хотя бы одно место аудиомонтажа.

Выводы

В настоящее время в научном сообществе голосового антиспуфинга слабо исследована проблема спуфинг-атак, полученных с помощью аудиомонтажа, одновременно с этим они представляют реальную угрозу для современных голосовых биометрических систем. Добавление данного вида атак в обучающий набор данных для модели антиспуфинга позволяет значительно повысить качество работы системы с нестационарными спуфинг-атаками, что повышает общую надежность системы верификации.

Список литературы

1. Wu Z. et al. Spoofing and countermeasures for speaker verification: A survey //speech communication. – 2015. – Т. 66. – С. 130-153.
2. Pan X., Zhang X., Lyu S. Detecting splicing in digital audios using local noise level estimation //2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2012. – С. 1841-1844.