УДК 004.8

ОБНАРУЖЕНИЕ АКУСТИЧЕСКИХ СЛЕДОВ ПРИМЕНЕНИЯ АЛГОРИТМА ГРИФФИНА-ЛИМА

Мельник Д.А. (ИТМО)
Научный руководитель – Чирковский А.Д. (ООО «ЦРТ»)

Введение. В современном мире развитие речевых технологий привело к появлению систем, способных качественно обрабатывать и синтезировать человеческую речь. Это в свою очередь позволило использовать эти системы злоумышленникам для имитации речи целевого человека с целью мошенничества или получения некоторой информации. Помимо использования в качестве инструмента социальной инженерии, сгенерированную или обработанную речь также возможно использовать с целью взлома систем голосовой биометрии для получения доступа к защищаемой ими информации. Имитация речи в описанных целях называется спуфингом, или спуфинг-атакой.

Неотъемлемой частью систем генерации речи является вокодер – система для генерации фонограммы речи на основе заданных в некотором закодированном формате характеристик. Следовательно, для систем распознавания спуфинга, называемых системами голосового антиспуфинга, является важным умение детектировать следы применения вокодеров. Простейшим примером вокодера является алгоритм Гриффина-Лима [1], являющийся алгоритмом для порождения аудио на основе амплитудной спектрограммы. Несмотря на ряд посвящённых ему исследований, в литературе слабо освещена тема обнаружения следов его применения системами голосового антиспуфинга.

Основная часть. В работе исследована возможность детектирования спуфинг-атак, при получении которых использовался алгоритм Гриффина-Лима, современными системами голосового антиспуфинга. Для оценки качества работы таких систем на основе базы спуфинг-атак и живой речи на английском языке ASV spoof 2021 [2] была сгенерирована база данных. Для этого была проведена аугментация записей живой речи с помощью алгоритма Гриффина-Лима с различными параметрами, а именно: количеством итераций, характеристиками окна преобразования Фурье, а также реализациями алгоритма. Для каждой рассматриваемой комбинации параметров выбиралось по 200 записей живой речи из рассматриваемой базы, амплитудная спектрограмма которых передавалась на вход алгоритма. В результате была получена база данных, состоящая из 435 тысяч записей речи в формате «.flac» с частотой дискретизации 16 кГц. Полученная база содержит те же поля, что и ASV spoof 2021, а также поля, описывающие применяемые для генерации параметры алгоритма. Это позволяет исследовать зависимость качества работы систем антиспуфинга от этих параметров.

Оценка способности детектирования следов применения алгоритма Гриффина-Лима проводилась на основе современной системы голосового антиспуфинга, построенной на основе архитектуры, описанной в статье [3]. В данной архитектуре WavLM-Large используется в качестве энкодера, а нейронная сеть прямого распространения в качестве классификатора. Для увеличения способности сети к распознаванию записей, синтезированных алгоритмом Гриффина-Лима, была разработана аугментация данных, включающая шаг использования данного алгоритма. Аугментация применялась непосредственно во время обучения модели, то есть перед поступлением каждой записи на вход системы задействовались шаги аугментации, каждый с некоторой заданной вероятностью. На основе базы данных ASVspoof2019-LA были обучены следующие модели: базовая – без применения шага аугментации с рассматриваемым алгоритмом, модель с применением этого шага, а также модель, обученная только на записях живой речи из оригинальной базы с применением аугментации. Качество обученных моделей было измерено с помощью разработанной базы данных, а также записей из базы данных MLAADv3 [4], объединённой с базой данных FLEURS [5].

Качество обнаружения системами антиспуфинга записей, сгенерированных с помощью

алгоритма Гриффина-Лима, было оценено на основе метрики Equal Error Rate (EER). На основе сгенерированной базы данных было выяснено, что увеличение числа итераций алгоритма приводит к усложнению распознавания следов его применения. Также было замечено, что размер шага оконного преобразования Фурье, использующийся в алгоритме, сильно влияет на сложность распознавания. Шаг преобразования размером в половину окна делает синтезированные фонограммы заметно более простыми для детектирования. Также показательным является сравнение качества моделей на выборке из объединения баз данных MLAADv3 и FLEURS, включающей записи живой речи и речи, полученной с помощью алгоритма Гриффина-Лима. На этой выборке модели, обученные с использованием шага аугментации с данным алгоритмом, показывают значительно лучшие результаты метрики EER, чем модель, обученная без использования этого шага (0,0376 и 0,0319 против 0,1946).

Выводы. В работе была исследована зависимость качества обнаружения системами антиспуфинга фонограмм, синтезированных с помощью алгоритма Гриффина-Лима, от его параметров. Проведённое исследование также показало, что разработанные методы аугментации помогают увеличить качество распознавания подобных фонограмм. Разработанную аугментацию можно рекомендовать к использованию при обучении системы голосового антиспуфинга, которая должна иметь хорошее качество работы на записях, синтезированных с помощью алгоритма Гриффина-Лима.

Список использованных источников:

1. Griffin D., Lim J. Signal estimation from modified short-time Fourier transform //IEEE Transactions on acoustics, speech, and signal processing. – 1984. – Т. 32. – №. 2. – С. 236-243.
2. Yamagishi J. et al. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection //ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge. – 2021.
3. Aliyev A., Kondratev A. Intema system description for the ASVspoof5 Challenge: power weighted score fusion //Proc. ASVspoof 2024. – 2024. – С. 152-157.
4. Müller N. M. et al. Mlaad: The multi-language audio anti-spoofing dataset //arXiv preprint arXiv:2401.09512. – 2024.
5. Conneau A. et al. Fleurs: Few-shot learning evaluation of universal representations of speech //2022 IEEE Spoken Language Technology Workshop (SLT). – IEEE, 2023. – С. 798-805.