УДК 004.934

АНАЛИЗ ВЛИЯНИЯ ЯЗЫКА НА СЛОЖНОСТЬ ОБНАРУЖЕНИЯ ГОЛОСОВЫХ ДИПФЕЙКОВ

Абрамова В. М. (ИТМО)
Научный руководитель – Чирковский А. Д.
(ООО "ЦРТ-Инновации").

Введение. С развитием речевых технологий голосовые дипфейки становятся всё более и более частым инструментом в руках мошенников для атак на системы голосовой биометрии, обмана людей и манипуляции общественным мнением. Несмотря на то, что наиболее совершенные методы генерации дипфейков и клонирования голоса в первую очередь разрабатываются на английском языке, проблема остается актуальной и на других языках. Некоторые системы способны генерировать речь в мультиязычном или даже кросс-язычном режимах. Системы дипфейков опираются на внутреннюю языковую модель, что сильно влияет на результаты детектирования дипфейков, при использовании неизвестной модели синтеза, в связи с чем подобные системы нельзя назвать строго языко-независимыми. Данные обстоятельства поднимают проблему языковой универсальности методов голосового антиспуфинга и детектирования голосовых дипфейков.

Основная часть. Поскольку виды дипфейков, существующих для разных языков отличаются, и во многом зависят от доступности данных на этих языках, а также модели синтеза речи, при оценке качества детекции мультиязычных голосовых дипфейков достаточно сложно разделить влияние лингвистических свойств языка (таких, как особенности фонации, уникальные фонетические и лексические проявления) и свойств методов создания дипфейков на этом языке. На первом этапе исследования был сформирован мультиязычный корпус голосовых данных, объединяющий синтетические записи из датасета MLAAD [1] (38 языков) и образцы натуральной речи из FLEURS [2]. Принципом разделения на обучающую, валидационную и тестовую выборки для синтезированной речи стал подход MLAAD Source Tracing [1]: обучающая выборка включала 8 языков (арабский, немецкий, английский и др.), тогда как валидационная и тестирующая охватили 20 языков. Вместо датасета M-AILABS, рекомендуемого авторами датасета MLAAD_V5 [1] в качестве живой речи для обучения модели и дальнейшего анализа был использован датасет FLEURS благодаря большему количеству представленных языков и разнообразию дикторов. Данные были разделены на обучающую и тестовые выборки в соответствии с рекомендациями автором датасета, а затем в каждой выборке были оставлены данные только тех языков, которые присутствуют в соответствующей выборке датасета MLAAD [1]. На втором этапе дообучена модель антиспуфинга [3] на базе архитектуры WAVLM [4] на мультиязычных данных: синтетические образцы из MLAAD [1] сочетались с натуральной речью из FLEURS [2]. Применялись техники аугментации, включая добавление шумов и искажение временных характеристик.

Экспериментальная оценка эффективности включала анализ метрики Equal Error Rate (EER - критерий равновероятного распределения ошибок) на тестовой выборке с акцентом на влияние знакомства модели с методом генерации дипфейков и языком. Были выявлены 4 тестовые группы в зависимости от присутствия языка и модели синтеза речи в обучающей выборке.

Выводы. Результаты выявили наихудшие показатели ошибки для записей в которых ни язык, ни метод генерации дипфейков речи не участвовали в дообучении модели в среднем 3,2%. Наилучшие результаты же были выявлены у данных, для которых и язык и модель использовались в обучении (0,05%). Для данных, где либо модель, либо язык применялся в дообучении разница EER составляет доли процента, 0,39% и 0,38%, соответственно, что дает нам возможность выдвинуть гипотезу о двойственной природе детекции дипфейков: языковой и зависимой от модели синтеза речи. Наибольшее влияние на результат оказывают модели

синтеза речи, которые могут синтезировать сразу несколько языков. Хорошие результаты из языков показал английский, за счёт использования наибольшего количества моделей.

Список использованных источников:

1. Müller N. M. et al. Mlaad: The multi-language audio anti-spoofing dataset //arXiv preprint arXiv:2401.09512. – 2024.
2. 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.
3. Aliyev A., Kondratev A. Intema system description for the ASVspoof5 Challenge: power weighted score fusion //Proc. ASVspoof 2024. – 2024. – С. 152-157.
4. Zhang F. et al. Audio is all in one: speech-driven gesture synthetics using WavLM pre-trained model //arXiv preprint arXiv:2308.05995. – 2023.