УДК 004.934

ИССЛЕДОВАНИЕ УЯЗВИМОСТИ СИСТЕМ ГОЛОСОВОГО АНТИСПУФИНГА К ШУМОВОЙ МАСКИРОВКЕ ГОЛОСОВЫХ ДИПФЕЙКОВ

Галимов А. А. (ИТМО)
Научный руководитель – Чирковский А. Д.
(ООО «ЦРТ-Инновации»)

Введение. Развитие технологий синтеза речи и клонирования голосов на основе ИИ создаёт угрозы для систем голосовой аутентификации, требуя разработки надёжных методов антиспуфинга. Однако эффективность таких систем снижается в реальных условиях из-за шумовой маскировки [1]. В данной работе исследуется устойчивость моделей WavLM к шумовым искажениям, а также влияние аугментации шумами на качество обнаружения голосовых дипфейков. Результаты показывают компромисс между устойчивостью к шуму и точностью на чистых данных.

Основная часть. В данном исследовании были обучены две модели WavLM для оценки уязвимости систем голосового антиспуфинга к шумовой маскировке голосовых дипфейков. Первая модель была обучена на чистых данных, содержащих как живую речь, так и голосовые дипфейки, из набора ASVspoof2019_train. Вторая модель использовала тот же набор данных, но с добавлением аугментации шумом из набора MUSAN, который включает различные типы фонового шума, такие как музыка, шум и речь в равных пропорциях. Аугментация данных с использованием шума MUSAN является традиционным методом увеличения устойчивости моделей к шумам [2]. Для повышения устойчивости моделей к шумовым искажениям используется метод RawBoost. Этот метод добавляет шум и искажения к сырым аудиоданным, что улучшает обобщающую способность моделей, особенно в кросс-доменном обучении [3]. Для тестирования обоих моделей была создана база из записей ASVspoof2021_df_hidden_track, смешанная с шумами из MUSAN при различных уровнях соотношения сигнал/шум (SNR), что позволило оценить устойчивость к шумовой маскировке. Выборка Hidden Track была обусловлен двумя ключевыми факторами: во-первых, относительно небольшим объемом подвыборки, что обеспечивало возможность ее разносторонней аугментации; во-вторых, устранением зависимости от продолжительности пауз в записях, характерной для исходного набора данных ASVspoof_2021.

Для оценки производительности обеих моделей использовалась метрика Equal Error Rate (EER), которая позволяет измерить точность систем распознавания. EER для чистых данных равен 10% для обеих моделей; однако для модели без аугментации порог увеличивается на 5 процентных пунктов. Для шумов одного типа при уровне отношения сигнал/шум, равном 12 дБ, EER равен 14% для модели с аугментацией, а для модели без аугментации EER равен 16%. Стоит отметить, что для модели с аугментацией порог остаётся стабильным для всех уровней SNR, тогда как для модели без аугментации значение порога не стабильно. Из этих данных можно сделать вывод, что модель с аугментацией работает лучше. Однако, если взять замаскированные шумами дипфейки и сравнить их с чистой живой речью, оставив прежними отношение сигнал/шум и тип шумов, то EER для модели с аугментацией равен 10%, а для модели без аугментации EER равен 9%. В данном случае аугментация ухудшила качество модели на 1 процентный пункт. При этом порог остаётся стабильным для обеих моделей.

Выводы. Аугментация ухудшает качество модели на чистых данных, увеличивая EER, но улучшает её устойчивость на зашумленных данных. Это указывает на компромисс: аугментация помогает в сложных условиях, но снижает точность на чистых записях.

Список использованных источников:

1. 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.

2. Chen Y. et al. USTC-KXDIGIT System Description for ASVspoof5 Challenge //arXiv preprint arXiv:2409.01695. – 2024.

3. Tak H. et al. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation //arXiv preprint arXiv:2202.12233. – 2022.