УДК 004.8

ДООБУЧЕНИЕ УНИВЕРСАЛЬНОЙ МОДЕЛИ ГОЛОСОВОГО АНТИСПУФИНГА ПРИ ДЕКОМПОЗИЦИИ ЗАДАЧИ ГОЛОСОВОГО АНТИСПУФИНГА

Чирковский А. Д.[1]
Научный руководитель – канд. физ.-мат. наук Рыбин С. В.[1]
[1]Университет ИТМО
artemioarbaletos@gmail.com

Введение

Системы голосового антиспуфинга крайне востребованы в сценариях защиты голосовой биометрии от мошенников, а также в криминалистической экспертизе фонограмм. Одним из ограничений современных систем голосового антиспуфинга является их низкая интерпретируемость. Частичной интерпретируемости можно достичь путём декомпозиции задачи голосового антиспуфинга, однако это усложняет модель, уменьшает производительность системы и в перспективе понижает качество работы системы. Данное исследование направлено на компенсацию указанных ограничений путем введения дополнительного этапа дообучения универсальной модели голосового антиспуфинга перед декомпозицией.

Основная часть

Частичного решения проблемы низкой интерпретируемости систем голосового антиспуфинга можно достичь путём декомпозиции задачи– то есть разделения общей задачи антиспуфинга на задачи обнаружения атак различного вида и обучение отдельных моделей под каждую из них. Данную концепцию возможно реализовать через использование мультиклассовой модели [1], либо через физическое разделение под каждую задачу. Первый подход активно применяется в задачах классификации типов дипфейк-систем и их компонент, но не исследован в ситуациях, когда системе антиспуфинга требуется разделять разнородные классы атак, такие как дипфейки и атаки повторного воспроизведения. Ограничениями второго являются пропорционально падающая производительность системы по количеству целевых атак, и скорость обучения финального решения. Частичное решение указанных проблем может быть достигнуто путем переиспользования весов разных моделей [2]. Данное исследование направлено на компенсацию указанных ограничений путем введения дополнительного этапа дообучения универсальной модели голосового антиспуфинга перед декомпозицией.

В работе предлагается сравнение подходов к обучению декомпозированных моделей голосового антиспуфинга на задачах обнаружения атак повторного воспроизведения и дипфейков. Подготовлена база данных на основе наборов данных из открытых источников, содержащая как дипфейки, так и атаки повторного воспроизведения, из таких наборов данных как ASVspoof5, InTheWild, ReMASC, SpoofCeleb. Сравнены следующие политики обучения декомпозированных моделей голосового антиспуфинга:

1. Обучение моделей под каждый из исследуемый тип атак отдельно;
2. Обучение универсальной модели голосового антиспуфинга, и последующее дообучение её под исследуемые типы атаки;
3. Обучение универсальной модели голосового антиспуфинга в мульти-класс режиме с дополнительными функциями потерь на все виды классов;
4. Дообучение универсальной модели, полученный в ходе экспериментов третьего пункта.

Исследовано влияние на качество работы системы во всех из указанных сценариях, а также проведено сравнение с точки зрения эксплуатационных характеристик, в том числе

скорости обучения, скорости работы, и сложности организации процесса с точки зрения MLOps процедур.

Выводы

Результаты показали, что подготовка специальной универсальной модели для последующего дообучения позволяет заметно ускорить скорость полученного решения и подготовки системы при сохранении общего уровня качества, что позволяет рекомендовать такой подход в сценариях, где скорость работы системы голосового антиспуфинга критична, а также требуется низкие задержки на обновление системы.

Литература

1. Klein, N., Chen, T., Tak, H., Casal, R., Khoury, E. Source Tracing of Audio Deepfake Systems // Proc. Interspeech 2024, 1100-1104, doi: 10.21437/Interspeech.2024-1283
2. Büker, A., Kurnaz, O., Bekiryazıcı, Ş., Demirtaş, S.C., Hanilçi, C. Evaluating Parameter Sharing for Spoofing-Aware Speaker Verification: A Case Study on the ASVspoof 5 Dataset // Proc. Interspeech 2025. P. 4573-4577. doi: 10.21437/Interspeech.2025-2618