УДК 004.8

ОБОБЩАЮЩАЯ СПОСОБНОСТЬ EGEMAPS В ЗАДАЧЕ ДЕТЕКТИРОВАНИЯ ДИПФЕЙКОВ

Ходжаметова М.И[1]
Научный руководитель – Чирковский А.Д[1]
Научный консультант – Аусев Е.В.[2]
[1]Университет ИТМО, [2]ООО «ЦРТ-Инновации»
milana.hodzhametova12@gmail.com

Введение

Современные нейросетевые методы синтеза речи позволяют создавать реалистичные дипфейк-аудио, угрожающие биометрическим системам. Для обнаружения таких аудио часто используется набор акустических признаков eGeMAPS [1]. Хотя классификаторы на его основе показывают эффективность, их способность к обобщению на новые голоса и методы синтеза изучена недостаточно [2, 3]. Исследования просодии [4, 5] показывают, что внутридикторская вариативность признаков может быть сильнее различий между реальной и синтезированной речью.

Основная часть

В данном исследовании выдвигается и проверяется ключевая гипотеза, согласно которой естественная вариативность акустических характеристик, наблюдаемая в пределах записей одного и того же реального диктора, может оказаться более значительной, чем усредненные различия этих же характеристик между подлинной речью и синтезированными дипфейк-аудио. Иными словами, внутриклассовый разброс признаков живого голоса способен перекрывать межклассовые различия, что потенциально затрудняет детекцию.

Для эмпирической проверки этого предположения был организован и проведен комплекс экспериментов на базе специализированного датасета «In-the-Wild» [4], включающего реальные и синтезированные образцы речи 58 известных публичных личностей, собранные из открытых источников, что гарантирует высокое акустическое разнообразие и приближает условия эксперимента к реальным сценариям использования систем детекции. Методологически работа была выстроена следующим образом: для каждого аудиофайла с помощью инструментария OpenSMILE были рассчитаны 88 акустических дескрипторов, входящих в стандартизированный набор eGeMAPS. Далее был проведен анализ внутридикторской вариативности путём сравнения межклассовых различий со стандартными отклонениями признаков внутри одного диктора. В качестве модели машинного обучения для разграничения подлинных и синтезированных образцов был применен ансамблевый алгоритм Random Forest. Ключевым аспектом работы стала схема валидации: для объективной оценки способности модели к обобщению эксперименты проводились с использованием трёх различных подходов, а именно стандартной случайной кросс-валидации, кросс-валидации по дикторам, при которой обучение происходит на одних людях, а тестирование на других, и кросс-валидации по источникам синтеза, подразумевающей обучение на дипфейках, созданных одними методами, и тестирование на синтезе от других генераторов. Применение этой дифференцированной системы валидации позволило не просто измерить точность классификации, но и количественно оценить фундаментальную способность модели адаптироваться к новым условиям, что дало возможность понять, действительно ли алгоритм выявляет общие акустические закономерности мошеннических записей или же его успех ограничен запоминанием специфических артефактов, присутствовавших в обучающей выборке.

Выводы

Результаты показали, что для большинства признаков (частота основного тона, энергия, дрожание и мерцание) внутридикторская вариативность превышает межклассовые различия (отношение разницы к вариативности составило 0.29–0.82), исключение тембральный признак MFCC1 (отношение выше 1 в 40% случаев). Точность классификации при случайной кросс-валидации достигла 91.8%, но на новых дикторах и типах синтеза упала до 76.8% и 75.0% (в одной группе — до 58.3%). Это подтверждает ограниченную обобщающую способность eGeMAPS и необходимость перехода к анализу динамических паттернов.

Литература

1. Eyben F., Scherer K.R., Schuller B.W. et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing // IEEE Transactions on Affective Computing. 2016. Vol. 7, No. 2. P. 190–202.
2. Pascu O., Oneata D., Cucu H., Müller N.M. Easy, Interpretable, Effective: openSMILE for voice deepfake detection // arXiv preprint arXiv:2408.15775. 2024.
3. Müller N., Czempin P., Dieckmann F. et al. Speech is Silver, Silence is Golden: What do ASVspoof-trained Models Really Learn? // Proceedings of the ASVspoof 2021 Workshop. 2021. P. 55–62.
4. Müller N., Czempin P., Dieckmann F. et al. Does Audio Deepfake Detection Generalize? // arXiv preprint arXiv:2203.16263. 2022.
5. Todisco M., Wang X., Vestveit R. et al. ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection // Proc. Interspeech 2019. 2019. P. 1008–1012.