УДК 004.8

ДЕТЕКТИРОВАНИЕ СЛЕДОВ НЕЙРОВОКОДЕРОВ И АКУСТИЧЕСКИХ МОДЕЛЕЙ В ЗАДАЧЕ ОБНАРУЖЕНИЯ ДИПФЕЙКОВ

Мельник Д. А.[1]
Научный руководитель – Чирковский А.Д.[2]
[1]Университет ИТМО, [2]ООО «ЦРТ-Инновации»
den26012002@mail.ru

Введение

Современные речевые технологии позволяют качественно обрабатывать и синтезировать человеческую речь. Помимо полезного применения, например, в голосовых ассистентах, эти технологии также могут использоваться для создания дипфейков, то есть имитации речи целевого человека в целях мошенничества, например, в виде социальной инженерии. Обнаружение голосовых дипфейков является важной областью исследований.

В общем случае системы генерации или преобразования речи можно рассматривать как системы, состоящие из акустической модели и вокодера. Акустическая модель позволяет преобразовать текст или фонемы в некоторый вектор акустических признаков. Вокодер является алгоритмом для синтеза фонограммы на основе заданных в некотором закодированном формате характеристик, роль которых в системе генерации речи и выполняют акустические признаки. В статье [1] показано, что детектирование следов вокодеров позволяет улучшить качество распознавания синтезированной речи. Однако применение вокодеров не всегда является признаком дипфейка, они могут применяться, например, для удаления фонового шума или устранения дефектов речи диктора. Следовательно, важной задачей становится разделение задач обнаружения следов вокодеров и акустических моделей.

Основная часть

Для исследования возможности детектирования следов вокодеров и акустических моделей был проведён ряд экспериментов. В качестве примера современной системы обнаружения дипфейков была рассмотрена система на основе архитектуры, описанной в [2]. В данной архитектуре WavLM-Large используется в качестве энкодера, а нейронная сеть прямого распространения в качестве классификатора. Для увеличения способности модели к распознаванию записей, обработанных нейровокодерами, при обучении применялась аугментация, включающая в себя в том числе операции применения вокодеров Diffwave [3] и MelGAN [4] к мел-спектрограммам записей. Аугментация применялась в онлайн-режиме, то есть перед поступлением каждой записи на вход системы последовательно использовались операции аугментации, каждая с некоторой заданной вероятностью. С использованием описанных архитектуры и аугментаций на основе баз данных ASVspoof2019-LA и ASVspoof5 был обучен ряд моделей по сценариям «вокодер – признак дипфейка» и «вокодер – не признак дипфейка». В первом сценарии любая речь, обработанная нейровокодерами, считалась дипфейком. Во втором сценарии живая речь, обработанная вокодерами, продолжала считаться живой речью, дипфейком считалась только изначально сгенерированная речь, то есть речь, при создании которой применялась акустическая модель.

Для оценки качества обученных моделей на данных, обработанных нейровокодерами, была составлена тестовая база данных на основе открытых баз LibriTTS, LibriSeVoc и закрытой базы OpenLibriTTS. Тестовая база была составлена из трёх частей, а именно: тестовой выборки LibriSeVoc (18487 записей); подвыборки из OpenLibriTTS, содержащей речь, синтезированную FastSpeech и TransformerTTS и различными вокодерами (20000

записей); и выборки, сгенерированной синтезом и клонированием речи на основе LibriTTS с помощью Tacotron и нейровокодеров WaveRNN, HifiGAN, DiffWave (8991 запись). Для базовой оценки способности детектирования дипфейков использовались базы данных ASVspoof2021-DF и ASVspoof5. Оценка производилась на основе метрики Equal Error Rate. Модели, обученные на ASVspoof2019-LA, не показали улучшения метрики при включении применения нейровокодеров в операции аугментации, однако продемонстрировали улучшение способности по распознаванию живой речи. Модели, обученные на ASVspoof5 для сценария «вокодер – не признак спуфинга», показали более высокое качество. Применение нейровокодеров как операций аугментации позволило улучшить результат этих моделей не только на составленной тестовой базе данных (2,17% против 7,88%), но и на ASVspoof5 (0,57% против 0,60%). Такие результаты позволяют сделать вывод о способности этих моделей обнаруживать следы акустических моделей вне зависимости от наличия следов применения вокодеров.

Выводы

В работе исследована способность систем обнаружения голосовых дипфейков к детектированию дипфейков на основе нейросетевых вокодеров, а также возможность разделению задач обнаружения следов акустических моделей и нейровокодеров. Было продемонстрировано улучшение качества моделей обнаружения дипфейков при обучении на ASVspoof5 и применении нейровокодеров как операции аугментации. Полученные результаты подтверждают возможность разделения рассмотренных задач.

Литература

1. Wang X., Yamagishi J. Spoofed training data for speech spoofing countermeasure can be efficiently created using neural vocoders //ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2023. – С. 1-5.
2. Aliyev A., Kondratev A. Intema system description for the ASVspoof5 Challenge: power weighted score fusion //Proc. ASVspoof 2024. – 2024. – С. 152-157.
3. Kong Z., Ping W., Huang J., Zhao K., Catanzaro B. Diffwave: A versatile diffusion model for audio synthesis //International Conference on Learning Representations. – 2021. – С. 862-878.
4. Kumar K., Kumar R., de Boissiere T., Gestin L., Teoh W.Z., Sotelo J., de Brebisson A., Bengio Y., Courville A.C. Melgan: Generative adversarial networks for conditional waveform synthesis //Advances in neural information processing systems. – 2019. – Т. 32. – С. 14843-14854.