УДК 004.056.2

ИССЛЕДОВАНИЕ МЕТОДОВ ДИАГНОСТИКИ ФАЛЬСИФИКАЦИИ ФОНОГРАММ ПУТЕМ СРАВНЕНИЯ ФОНОВЫХ ШУМОВ

© Лебедева Д.С., Максимов А.И.

Самарский национальный исследовательский университет имени академика С.П. Королева, г. Самара, Российская Федерация

e-mail: nirs@ssau.ru

Поскольку в настоящее время фонограммы активно используются как доказательство в судебных процессах, задача проверки аутентичности аудиозаписей в рамках криминалистической экспертизы стоит достаточно остро, а задача автоматизации такого процесса является крайне актуальной.

В данной работе исследуются возможности применения методов обработки сигналов и математической статистики к решению задачи диагностики фальсификации цифровой фонограммы, основанных на сравнении фрагментов фоновых шумов. Также были использованы методы машинного обучения и нейросетевые методы.

Фоновый шум является частью общего шума, поступающего от подвижных или стационарно расположенных источников, при отключении известных источников. В случае существенного различия метрик, характеристик, параметров взятых фрагментов шума, можно заключить, что в аудиозаписи присутствует вставка – ее части записаны в различных условиях.

Для данного исследования был самостоятельно создан тестовый датасет, представляющий собой набор фрагментов фонограмм, полностью состоящих из фоновых шумов. Фоновые шумы были записаны в различных условиях – с использованием различных записывающих устройств, в различных помещениях и в различное время суток. Датасет состоит из 58 фонограмм, из которых образовано 1653 пар аудиозаписей. 253 пары содержат записи, полученные в идентичных условиях, 1400 пар, соответственно, содержат записи, полученные в различных условиях. При сборе данного датасета было использовано 3 записывающих устройства, запись производилась в 2 различных помещениях в различное время суток.

В данной работе исследовалось 3 статистических критерия: хи-квадрат Пирсона, Стьюдента, Манна – Уитни – Вилкоксона.

Статистические критерии применялись для решения задачи диагностики фонограмм следующим образом: для диагностируемой пары аудиозаписей формировалась спектрограмма, в рамках одного частотного среза фонограмм вычислялся один из трех выбранных статистических критериев, полученные значения усреднялись по всем частотным срезам. Также исследовалось усредненное значение полученных критериев.

В данном исследовании полагалось, что две фонограммы идентичны, если значение усредненного критерия превысило порог 0,78, полученный эмпирическим образом.

С учетом установленного порога были получены следующие результаты точности работы статистических критериев:

– критерий Стьюдента – 0,30,
– критерия Манна – Уитни – Вилкоксона – 0,13,
– критерий хи-квадрат Пирсона – 0,15,
– среднее по трем критериям – 0,14.

В данной работе также исследовалась возможность использования методов машинного обучения для решения задачи. Был построен бинарный классификатор, в который в качестве вектора признаков подавался вектор из значений статистических критериев – $\left\{\bar{U}_S, \bar{U}_M, \bar{U}_{\chi^2}\right\}$. В результате работы классификатора были получены следующие значения:

– значение точности = 0,30,

– значение f-score = 0,08.

Поскольку применение классического машинного обучения не вызвало существенного повышения точности решения поставленной задачи, были исследованы возможности нейросетевых моделей.

Поскольку спектрограмма шумового сигнала является двумерным сигналом (изображением), было решено исследовать возможности синергии предназначенных для обработки изображений нейросетевых моделей и обработки звуковых сигналов.

Для решения выбранной задачи подходит архитектура сиамских нейронных сетей. Сиамская нейронная сеть состоит из двух идентичных нейронных подсетей, которые принимают на вход различные данные. Сиамская сеть высчитывает отображение входных данных в вектора, считает расстояние между ними и функцию потерь, после чего оценивает различие между входными данными.

В данной работе нейросеть строилась на основе модели VGG16. Данная модель содержит в себе 16 слоев. На выходе каждой подсети получался вектор-дескриптор входной спектрограммы аудиосигнала размерностью 4096. Для данной задачи обе подсети были перенастроены так, чтобы принимать на вход изображение 128x128. На выходе высчитывалось евклидово расстояние между полученными векторами. Полученные расстояния подавались на вход бинарному классификатору. В результате получились следующие значения:

– значение точности = 0,81,

– значение f1-score = 0,30.

Следующим шагом стало использование той же модели сиамской нейросети на мел-спектрограмме. Мел-спектрограмма строится путем перевода спектрограммы в шкалу мел. В результате работы классификатора в этом случае были получены следующие значения:

– значение точности = 0,83,

– значение f-score = 0,46.

Классификация на основе статистических критериев показала очень низкие результаты (точность 0,30 и f-мера 0,08), недостаточные для решения поставленной задачи.

Использование нейросетевой модели показало значительно более высокий результат, особенно при подаче на вход мел-спектрограммы – точность 0,83 и f-мера 0,46, что говорит о перспективности данного направления.

Планируется расширить список метрик расстояний, расширить датасет данными из открытых источников, а также исследовать использование MLP для векторов-дескрипторов.