ВЫЯВЛЕНИЕ ГОЛОСОВЫХ ПОДДЕЛОК ТИПА ПОВТОРНОЕ ВОСПРОИЗВЕДЕНИЕ С ИСПОЛЬЗОВАНИЕМ КОЭФФИЦИЕНТОВ МОДИФИЦИРОВАННОЙ ФУНКЦИИ ГРУППОВОЙ ЗАДЕРЖКИ

Р.С. Насретдинов, И.Д. Ильяшенко
Алтайский государственный университет, г. Барнаул

Технология автоматической верификации дикторов всё чаще применяется как инструмент аутентификации в информационных системах. От неё требуется не только высокая точность работы для образцов, записанных в различных акустических средах, но также устойчивость к попыткам взлома [1,2]. В данный момент к наиболее опасным видам атак относятся голосовые подделки типа повторное воспроизведение ввиду доступности устройств звукозаписи и простоты методов, необходимых для создания подделок. Атаки подобного типа заключаются в записи злоумышленником речи целевого диктора и ее повторном воспроизведении при аутентификации в информационной системе.

Для привлечения внимания специалистов к данной проблеме было проведено соревнование ASVspoof 2017 [3], задача которого состояла в создании системы выявления голосовых подделок типа повторное воспроизведение. Использованный в соревновании набор данных состоял из образцов, записанных во множестве различных сочетаний акустических сред, устройств записи и воспроизведения.

В данной работе предложен новый подход к выявлению голосовых подделок типа повторное воспроизведение. Он основан на использовании модифицированной функции групповой задержки для извлечения информативных признаков из аудиосигналов. Апробация метода производилась на наборе данных ASVspoof 2017, а результаты сравнивались с базовым методом, предложенным авторами конкурса [3].

В качестве метода извлечения признаков были выбраны диаграммы групповой задержки [4]. Они вычисляются на основе оконного преобразования Фурье [5]. Последнее может быть вычисленно по следующей формуле:

$$X(\omega,t) = |X(\omega,t)|e^{j\theta(\omega,t)}, \quad (1)$$

где $|X(\omega,t)|$ - амплитуда сигнала на частоте $\omega \ \ t$, $X(\theta,t)$ - фаза сигнала.

Функция групповой задержки [6] является отрицательной производной фазовой составляющей оконного преобразования Фурье:

$$\tau(\omega,t) = -\frac{d(\theta(\omega,t))}{d\omega} \quad (2)$$

Эта же функция может быть вычисленна с использованием только амплитудных значений:

$$\tau(\omega,t) = \frac{X_R(\omega,t)Y_R(\omega,t)+Y_I(w,t)X_I(\omega,t)}{|X(\omega,t)|^2} \quad (3)$$

где n – номер отсчета сигнала, X_R, X_R – действительная и мнимая части оконного преобразования Фурье сигнала $x(n)$ соотвественно, Y_R, Y_I – действительная и мнимая части оконного преобразования Фурье сигнала $nx(n)$ соотвественно.

Для устранения негативного эффекта пиков квадрата амплитуды спектра исходного сигнала в знаменателе (3) используется модифицицированная версия функции групповой задержки:

$$\tau_m(\omega,t) = \frac{\tau(\omega,t)}{|\tau(\omega,t)|}(|\tau(\omega,t)|)^{\alpha},\ 0 < \alpha \leq 1, \quad (4)$$

где:

$$\tau(\omega,t) = \frac{X_R(\omega,t)Y_R(\omega,t)+Y_I(w,t)X_I(\omega,t)}{|\xi(\omega,t)|^2},\ 0 < \gamma \leq 1, \quad (5)$$

где $\xi(\omega,t)$ – кепстрально сглаженный спектр сигнала.

Для получения диаграмм групповой задержки вектора признаков, вычисленные на основе модифицированной функции групповой задержки всех фреймов сигнала, конкатенировались, образуя двумерную матрицу. Полученные данные были обрезаны, либо дозаполнены нулями до 256 отсчетов по временной шкале для того, чтобы их размер был одинаковым при подаче на вход классифицирующей нейронной сети.

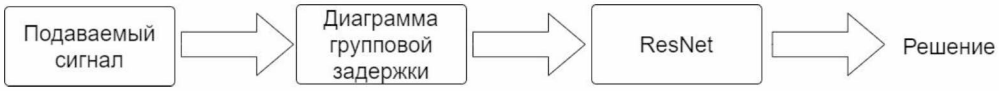
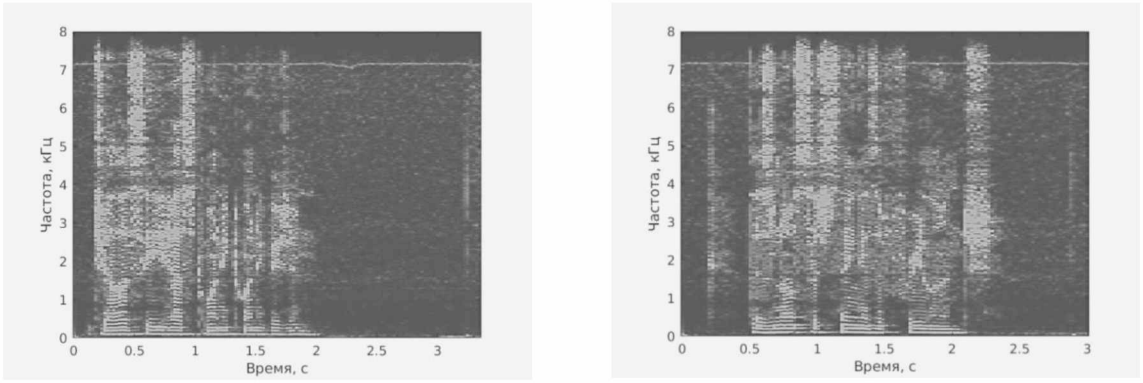


Рисунок 1 – Схема работы метода.

При извлечении диаграмм групповой задержки из аудиофайлов использовалось окно Хэмминга длительностью 50мс и сдвигом 25мс, в следствие чего количество коэффициентов Фурье для одного фрейма составляло 513. Значения параметров α и γ (4, 5) были подобраны эмпирически, и составляли 0,4 и 0,9 соответственно.

Во время предварительного анализа спектральных характеристик для образцов из набора данных ASVspoof 2017 были выявлены особенности для некоторых поддельных аудиозаписей. У перезаписанных аудиофайлов (рисунок 3) наблюдаются характерные полосы на частотах 2 и 5 кГц, чего не было замечено для настощих образцов (рисунок 2).



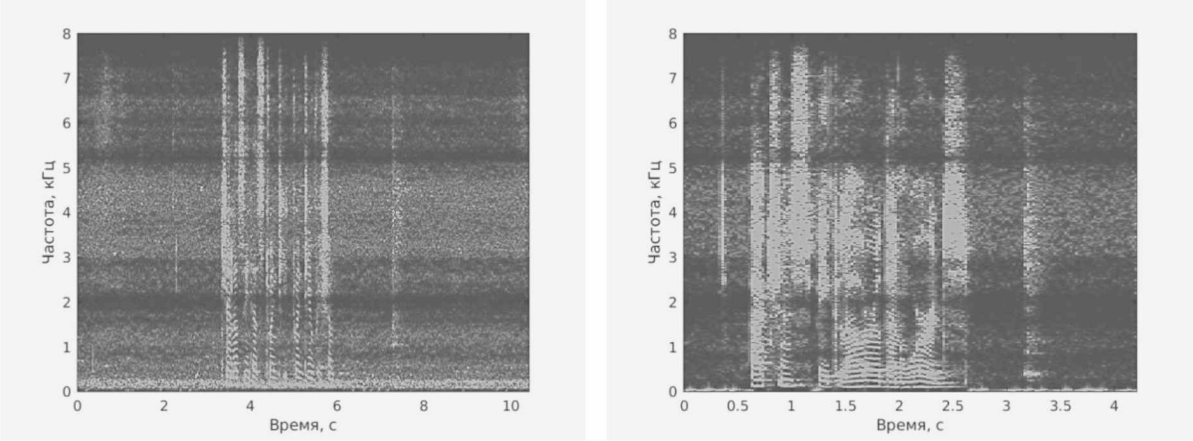
а) б)

Рисунок 2 – Диаграммы групповой задержки настоящих образцов датасета ASVspoof 2017: а – для аудиофайла D_1000759.wav, б – для аудиофайла D_1000760.wav.

В качестве классификатора использовалась глубокая нейронная сверточная сеть ResNet. Она была предложена [7] для решения проблемы «затухания» градиента при использовании метода обратного распространения ошибки для обучения моделей. Использованные в ней остаточные связи между слоями предотвращают обнуление градиента, что позволяет существенно увеличить «глубину» сети. Авторы [7] представили различные вариации архитектуры ResNet с разным количеством слоев, количечество обучаемых параметров которых приведено в таблице 1.

Таблица 1 – Количество параметров для различных архитектур ResNet.

Нейронная сеть	Количество параметров
ResNet18	11x10^6
ResNet34	23x10^6
ResNet50	25x10^6
ResNet101	44x10^6
ResNet152	60x10^6



а) б)

Рисунок 3 – Диаграммы групповой задержки поддельных образцов датасета ASVspoof 2017: а – для аудиофайла D_1000764.wav, б – для аудиофайла D_1000769.wav.

Архитектура нейронных сетей ResNet18 и ResNet34 состояла из начального сверточного слоя с фильтром 7x7 и блоков сверточных слоев с фильтрами 3x3. Такие блоки содержали сверточные слои с одинаковой глубиной, которая увеличивалась в 2 раза при переходе на следующий блок. Размер выходных данных сокращался вдвое при помощи операции максимального пулинга 2x2 после каждого блока. Для соблюдения размерностей перед выходным слоем данные сжимались при помощи глобального усредняющего пулинга.

Обучение и тестирование предложенного метода производилось на наборе данных ASVspoof 2017 [3]. Он состоял из настоящих и поддельных образцов речевых сигналов. Настоящие образцы взяты из набора данных RedDots, который был ранее собран волонтёрами для решения задачи текстозависимой автоматической верификации дикторов. Поддельные образцы были получены путём воспроизведения и записи настоящих образцов, используя 26 различных устройств воспроизведения и 25 устройств записи звука в 26 различных акустических средах.

Набор данных ASVspoof 2017 разделён на 3 подвыборки: обучающая, валидационная и оценочная. В таблице 2 представлено их описание. Набор данных содержит метки образцов настоящих/поддельных классов, описания среды записи/перезаписи, использованных микрофон и звуковоспроизводящих устройств.

Таблица 2 – Структура набора данных ASVspoof 2017 v.2

Множество	Дикторы	Сессии перезаписи	Конфигурации перезаписи	Образцы	
				Настоящие	Поддельные
Обучение	10	6	3	1507	1507
Валидация	8	10	10	760	950
Оценка	24	161	57	1298	12008

Для тестирования предложенного метода использовались различные сочетания выборок из набора данных ASVspoof 2017 версии 2. Подробности формирования подмножеств для обучения и оценки приведены в таблице 3. В качестве классификатора были использованы нейронные сети ResNet18 и ResNet34, так как использование более глубоких архитектур приводило к переобучению моделей. При обучении моделей в качестве метода оптимизации использовался Adam [8]. Оптимизируемой ошибкой была функция перекрёстной энтропии с дополнительным регуляризационным членом L2 [9], предотвращающим переобучение модели.

Значения скорости обучения и регуляризации составляли 0.0001 и 0.01 соответственно. Работа с нейронными сетями осуществлялась при помощи библиотеки PyTorch для языка программирования python.

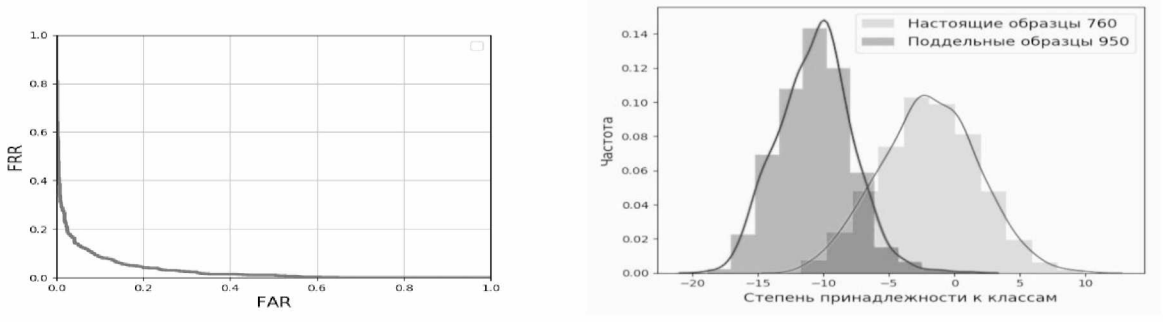
Качество работы моделей оценивалось распространенной для биометрических систем метрикой EER. Она является порогом, при котором ошибки первого рода равны ошибкам второго. Также были построены графики DET-кривых [10].

В таблице 3 представлены результаты апробации предложенного метода для выявления атак типа повторное воспроизведение. Применение архитектур ResNet50 и глубже приводило к переобучению сети, поэтому сравнивалось качество при использовании нейронных сетей ResNet18 и ResNet34 в качестве классификатора. Из таблицы видно, что для 1, 2 и 4 вариантов формирования подмножеств для обучения и оценки лучшее качество показывает нейронная сеть ResNet34, а для 3 варианта - ResNet18. Возможно, это связано с тем, что оценочная выборка содержит большее количество обучаемых образцов в сравнении с обучающей и валидационной, и в связи с этим нейронная сеть ResNet34 обладает меньшей обобщающей способностью при обучении на ней. Были построены DET-кривые и распределения значений степени принадлежности к классам для 1 и 2 вариантов сочетания выборок для обучения и тестирования.

Таблица 3 – Результаты тестирования предложенного подхода

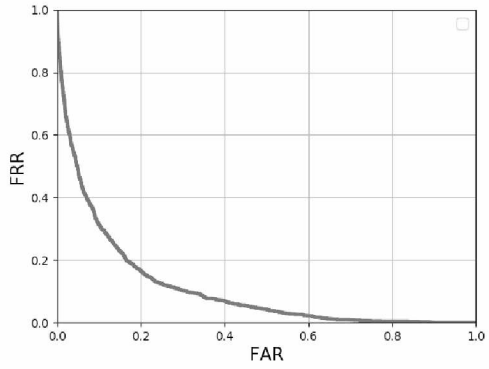
Вариант	Подмножество для обучения	Подмножество для оценки	Метод	EER	EER для базовой модели
1	Обучение	Валидация	ResNet 18	0.18	0.10
			ResNet34	**0.09**	
2	Обучение	Оценка	ResNet18	0.22	0.30
			ResNet34	**0.18**	
3	Оценка	Валидация	**ResNet18**	**0.13**	-
			ResNet34	0.18	
4	Обучение+валидация	Оценка	ResNet18	0.09	-
			ResNet34	**0.07**	

Разработанный метод сравнивался с базовой системой, предложенной авторами конкурса ASVspoof 2017. Она основана на применении кепстральных коэффициентов константного Q-преобразования в качестве метода извлечения признаков. Классификатор GMM использовался для выявления поддельных аудиообразцов.

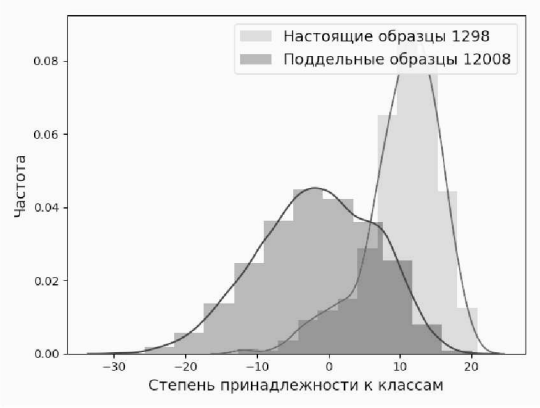


а) б)

Рисунок 4 – Характеристики предложенного метода для 1 варианта сочетания выборок обучения и оценки: а) DET-кривая, б) распределения значений степени принадлежности к классам.



а)



б)

Рисунок 5 – Характеристики предложенного метода для 2 варианта сочетания выборок обучения и оценки: а) DET-кривая, б) распределения значений степени принадлежности к классам.

Результат сравнения методов представлен в таблице 3. Качество базового метода было взято из научной статьи [11]. Из таблицы 3 видно, что предложенный подход показывает лучшее качество в сравнении с базовым методом.

В данной работе был представлен новый метод к выявлению голосовых подделок типа повторное воспроизведение. Предложенный подход показал лучшее качество в сравнении с базовым методом, предложенным авторами конкурса ASVspoof 2017, на основе метрики EER.

Библиографический список

1. Ильяшенко И., Насретдинов Р. Анализ признаков из wavenet автоэнкодера в задаче обнаружения искусственных искажений в аудиофайлах // Проблемы правовой и технической защиты информации – 2018. - № 6.– С.39-45.
2. Ильяшенко И., Насретдинов Р., Лепендин А. Применение WaveNet-автоэнкодера в задаче обнаружения искусственных искажений аудиофреймов // Высокопроизводительные вычислительные системы и технологии - № 1. – 2018. – С.40-45.
3. Delgado H., Todisco M., Sahidullah M. ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements [Электронный ресурс] / 2018. URL: http://www.asvspoof.org/data2017/asvspoof-2017-version-2-cameraReady.pdf
4. Francis T., Mohit J., Prasenjit D. End-To-End Audio Replay Attack Detection Using Deep Convolutional Networks with Attention [Электронный ресурс] // 2018. – URL: https://www.researchgate.net/publication/327388690_End-To-End_Audio_Replay_Attack_Detection_Using_Deep_Convolutional_Networks_with_Attention
5. Xuedong H., Acero A., Hsiao-Wuen H. Spoken Language Processing // M.: Prentice Hall – 2001. - с. 980
6. Hegde R.M. Murthy H. A., Gadde V.R.R. Significance of the modified group delay feature in speech recognition // IEEE Transactions on Audio, Speech, and Language Processing - №1. – 2007. - С. 190-201.
7. Kaiming H., Xiangyu Z., Shaoqing R. Deep Residual Learning for Image Recognition [Электронный ресурс] // 2015. - URL: https://arxiv.org/pdf/1512.03385.pdf
8. Yoshua B., Nicolas B., Razvan P. Advances in optimizing recurrent networks [Электронный ресурс] // 2012. – URL: https://arxiv.org/pdf/1212.0901v2.pdf
9. Ian G. Yoshua B., Aaron C. Deep Learning // M.:The MIT Press, 2016 - С.781

10. ISO/IEC 1975-1:2006. 6(Information technology — Biometric performance testing and reporting — Part 1: Principles and framework
11. Лаврентьева Г., Новосёлова С., Козлов А. Методы детектирования спуфинг-атак повторного воспроизведения на голосовые биометрические системы // Научно-технический вестник информационных технологий, механики и оптики - 2018 – Т.18. - № 3. - С.428-436.