УНИВЕРСИТЕТ ИТМО
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
май–июнь 2018 Том 18 № 3 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
May–June 2018 Vol. 18 No 3 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 004.93, 57.087.1

МЕТОДЫ ДЕТЕКТИРОВАНИЯ СПУФИНГ-АТАК ПОВТОРНОГО ВОСПРОИЗВЕДЕНИЯ НА ГОЛОСОВЫЕ БИОМЕТРИЧЕСКИЕ СИСТЕМЫ

Г.М. Лаврентьева[a,b], С.А. Новосёлов[a,b], А.В. Козлов[a], О.Ю. Кудашев[a,b], В.Л. Щемелинин [b, c], Ю.Н. Матвеев[a,b], М. Де Марсико[d]

[a] ООО «ЦРТ-инновации», Санкт-Петербург, 196084, Российская Федерация
[b] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[c] ООО «ЦРТ», Санкт-Петербург, 196084, Российская Федерация
[d] Университет Ла Сапиенца, Рим, 00198, Италия
Адрес для переписки: lavrentyeva@speechpro.com

Информация о статье
Поступила в редакцию 15.02.18, принята к печати 15.03.18
doi: 10.17586/2226-1494-2018-18-3-428-436
Язык статьи – русский

Ссылка для цитирования: Лаврентьева Г.М., Новосёлов С.А., Козлов А.В., Кудашев, О.Ю. Щемелинин В.Л., Матвеев Ю.Н., Де Марсико М. 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. doi: 10.17586/2226-1494-2018-18-3-428-436

Аннотация

Предмет исследования. Исследована задача детектирования атак повторного воспроизведения на голосовые биометрические системы. Подобные атаки, в силу простоты своей реализации, с большей вероятностью используются злоумышленниками и поэтому представляют собой особую опасность. В данной работе описана система детектирования атак повторного воспроизведения, которая была представлена на конкурсе ASVspoof 2017 по этой проблеме. **Метод.** Исследована эффективность подхода на основе глубоких нейронных сетей для решения описанной задачи, в частности, конволюционных нейронных сетей с Max-Feature-Map активационной функцией. **Основные результаты.** Результаты экспериментов на базе конкурса показали, что предложенный подход превосходит современные методы по качеству детектирования спуфинг-атак. Лучшая представленная система продемонстрировала ошибку EER, равную 6,73% на подмножестве неизвестных атак, что на 72% лучше базового метода, представленного на конкурсе. **Практическая значимость.** Результаты работы могут найти применение в области голосовой биометрии. Представленные методы могут быть использованы в системах автоматической верификации и идентификации дикторов по голосу для детектирования атак с целью взлома.

Ключевые слова
спуфинг, детектирование атак повторного воспроизведения, CNN, RNN, ASVspoof

Благодарности
Работа выполнена в рамках темы ПНИЭР «Разработка технологии автоматической бимодальной верификации по лицу и голосу с защитой от использования подложных биометрических образцов» при финансовой поддержке Министерства образования и науки Российской Федерации по соглашению о предоставлении субсидии №14.578.21.0189 от 03.10.2016 RFMEFI57816X0189.

AUDIO-REPLAY ATTACKS SPOOFING DETECTION FOR SPEAKER RECOGNITION SYSTEMS

G.M. Lavrentyeva[a,b], S.A. Novoselov[a,b], A.V. Kozlov[a], O.Yu. Kudashev[a,b], V.L. Shchemelinin[b, c], Yu.N. Matveev[a,b], M. De Marsico[d]

[a] STC-innovations Ltd., Saint Petersburg, 196084, Russian Federation
[b] ITMO University, Saint Petersburg, 197101, Russian Federation
[c] STC Ltd., Saint Petersburg, 196084, Russian Federation
[d] Sapienza University of Rome, Rome, 00198, Italy
Corresponding author: lavrentyeva@speechpro.com

Article info
Received 15.02.18, accepted 15.03.18
doi: 10.17586/2226-1494-2018-18-3-428-436
Article in Russian

For citation: Lavrentyeva G.M., Novoselov S.A., Kozlov A.V., Kudashev O.Yu., Shchemelinin V.L., Matveev Yu.N., De Marsico M. Audio-replay attacks spoofing detection for speaker recognition systems. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2018, vol. 18, no. 3, pp. 428–436 (in Russian). doi: 10.17586/2226-1494-2018-18-3-428-436

Abstract
Subject of Research. The present work considers the problem of detecting replay attacks on voice biometric systems. Due to their simplicity, these attacks are more likely to be used by the imposters, and that is why they are of special risk. This work describes the system for detecting replay attacks that was presented on the Automatic Speaker Verification Spoofing and Countermeasures (ASVspoof) Challenge 2017 focused on this problem. **Method.** We study the efficiency of deep learning approach in the described task, in particular, convolutional neural networks with Max-Feature-Map activation function. **Main Results.** Experimental results obtained on the Challenge corpora have demonstrated high performance of such approach in contrast to current state-of-the-art baseline systems. Our primary system achieved 6.73% EER on the evaluation part of the corpora which is 72% relative improvement over the ASVspoof 2017 baseline system. **Practical Relevance.** The results of the work can be applied in the field of voice biometrics. The presented methods can be used in systems of automatic speaker verification and identification for detecting spoofing attacks on them.
Keywords
spoofing, replay attack detection, CNN, RNN, ASVspoof

Acknowledgements
This work was financially supported by the Ministry of Education and Science of the Russian Federation, Contract 14.578.21.0189 from 3.10.2016 (ID RFMEFI57816X0189).

Введение

В последние годы из-за растущего интереса к обеспечению безопасности во всех аспектах нашей повседневной жизни возросла потребность в удобных и ненавязчивых методах аутентификации. Автоматическая аутентификация диктора по голосу предлагает недорогое и надежное решение проблемы идентификации при предоставлении голосовых услуг. Она уже используется в сфере социального обеспечения, иммиграционного контроля. Системы распознавания диктора по голосу широко используются в call-центрах, интернет-банкинге и других областях электронной коммерции. Однако, несмотря на то, что технология достигла точки массового распространения на рынке, она остается уязвимой к атакам с целью взлома (спуфингу) [1].

Согласно [2], спуфинг-атаки на голосовые биометрические системы классифицируются на направленные и ненаправленные атаки в соответствии с уровнем, на котором они совершаются. Ненаправленные атаки нацелены на внутренние модули системы и требуют определенного уровня доступа, например, атаки на модуль извлечения признаков, голосовые модели или итоговые оценки и решения. Ненаправленные атаки являются общими для всех биометрических систем. Защитой от них занимаются методы криптографии. Направленные атаки атакуют только устройство ввода и фокусируются на замене входных данных. В силу того, что аутентификация по голосу в основном используется в автоматических системах без непосредственного видимого контакта с пользователем, такие атаки с большей вероятностью будут использоваться злоумышленниками из-за простоты реализации. Наиболее известными типами атак являются имперсонализация, повторное воспроизведение, преобразование и синтез речи [3].

В отличие от других типов спуфинга, имперсонализация является лишь имитацией голоса зарегистрированного пользователя. Она не оставляет следов записи, устройств воспроизведения или обработки сигналов, поскольку это подлинная речь. Такую атаку можно обнаружить с помощью надежной системы аутентификации, так как мошенник в данном случае может подделать лишь манеру говорить, но не индивидуальные биометрические характеристики голоса [4]. Методы детектирования синтеза и преобразования речи были темой конкурса ASVspoof Challenge 2015 года [5]. Результаты конкурса продемонстрировали большой потенциал современных методов в обнаружении таких атак. По сравнению с ними атаки повторного воспроизведения значительно проще в реализации. Они не требуют специальных знаний в области обработки аудиосигнала. В случае повторного воспроизведения мошенник использует предварительно записанные образцы речи целевого диктора, которые можно легко подготовить с помощью недорогих записывающих устройств или смартфонов. По этой причине повторное воспроизведение является наиболее доступным и, следовательно, критическим методом спуфинга.

На сегодняшний день существует достаточно мало исследований, посвященных обнаружению атак повторного воспроизведения. Большая часть решений, представленных для тексто-зависимых систем аутентификации, основана на сравнении тестового высказывания с сохраненным высказыванием, записанным во время регистрации. Уязвимость тексто-независимых систем к спуфинг-атакам рассматривалась в [5]. Эта работа показывает значительное увеличение ошибки ложного пропуска в системе автоматической аутентификации в случае присутствия атак повторного воспроизведения. Методы обнаружения таких атак для тексто-независимого случая в основном основаны на обнаружении шума, характерного для тех или иных акустических условий.

Важным шагом на пути создания систем противодействия спуфинг-атакам является организация конкурса Automatic Speaker Verification Spoofing 2017 Challenge (ASVspoof 2017). Конкурс направлен на содействие в разработке контрмер против атак повторного воспроизведения, надежных в случаях как известных, так и неизвестных атак [6]. ASVspoof 2017 был сфокусирован на изолированной задаче обнаружения атак повторного воспроизведения, рассматриваемой без системы аутентификации и любых предварительно записанных данных регистрации.

Основной целью настоящей работы является исследование эффективности перспективного подхода конволюционных нейронных сетей (CNN) для решения задачи детектирования спуфинг-атак повторного воспроизведения. Успех CNN в таких задачах, как классификация видео [7], изображений [8, 9], распознавание лиц [10], а также детектирования атак на системы лицевой биометрии [11], был мощной мотивацией для применения таких подходов к задачам детектирования спуфинга. В [12] авторы исследовали применимость методов глубокого обучения для задачи детектирования синтеза и преобразования голоса на базе конкурса ASVspoof 2015 и подтвердили высокую эффективность применения глубоких нейронных сетей (DNN), CNN и рекуррентных нейронных сетей (RNN) для поставленной задачи. Авторы [13] предложили использовать архитектуру временной CNN (Temporal CNN) для обнаружения синтеза и преобразования речи и также достигли значительных результатов на базе конкурса ASVspoof 2015.

В данной работе описываются системы детектирования спуфинг-атак повторного воспроизведения, представленные компанией «Центр речевых технологий» на конкурсе Automatic Speaker Verification Spoofing and Countermeasures Challenge 2017.

Конкурс ASVspoof 2017

ASVspoof Challenge был организован, чтобы оценить возможности детектирования атак повторного воспроизведения «в дикой природе», в частности, в различных акустических условиях. Для этой цели на основе тексто-зависимой базы RedDots [14] была собрана база спуфинг-атак. Корпус RedDots был воспроизведен и повторно записан в различных акустических условиях (открытое офисное пространство, балкон и т.д.). Для записи попыток спуфинга использовались 15 различных устройств воспроизведения и 16 устройств записи, включая смартфоны, микрофоны и динамики ноутбуков, а также высококачественные динамики. Оригинальные записи RedDots использовались в качестве записей естественной речи. Этот набор данных был разделен на 3 части: для обучения системы, валидации и финального теста. Тестовое множество не содержит информации о спуфинге, устройствах или условиях записи. Более того спуфинг-атаки в нем были записаны с использованием устройств, которые не использовались в процессе записи баз для обучения и разработки. Таким образом, они представляли собой неизвестные атаки, т.е. атаки, детали реализации которых были неизвестны на момент обучения системы.

Базовая система

Авторы [15] считают, что методы детектирования спуфинга должны быть сосредоточены более на использовании информативных признаков, нежели на усложнении моделей. Они предложили систему, основанную на кепстральных коэффициентах константного Q-преобразования (CQCC). На рис. 1 продемонстрирована подробная схема извлечения CQCC-коэффициентов из акустического сигнала.

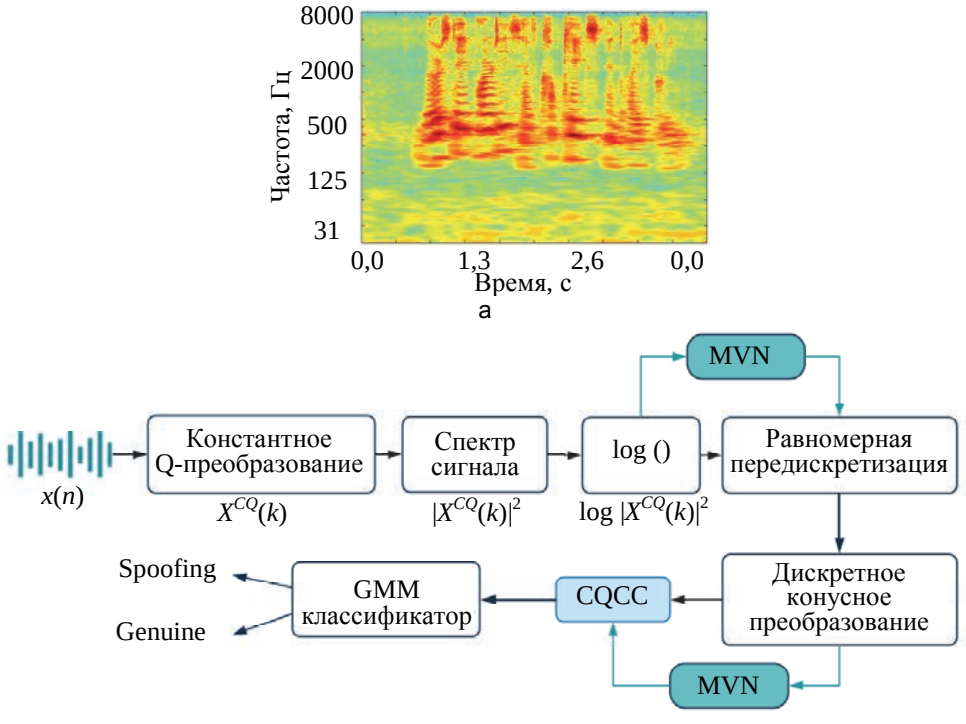

Рис. 1. CQT-спектр (а); система на основе CQCC-признаков и ее модификация (б)

Константное Q-преобразование (CQT) широко используется для обработки музыкальных сигналов. На рис. 1, а, представлен пример CQT-спектра. В качестве входных признаков для базовой системы использовались кепстральные коэффициенты CQT-преобразования, вычисленные в соответствии со схемой, представленной на рис. 1, б, где $x(n)$ – входной сигнал на временной шкале; n, k – индекс частотного бина; $X(k)$ – результат CQT-преобразования.

Для классификации на естественную и поддельную речь использовался стандартный двухклассовый классификатор на основе смеси гауссовых распределений (Gaussian Mixture Model, GMM). Для классов подлинной и поддельной речи были обучены 512-компонентные модели с помощью EM-алгоритма (Expectation Maximization) со случайной инициализацией. Для каждого высказывания из GMM-моделей были получены оценки правдоподобия, а конечный результат был рассчитан как отношение логарифмического правдоподобия: $\Lambda(x) = \log L\left(X\middle|\theta_g\right) - \log L(X|\theta_s)$, где X – последовательность признаков тестовой записи, L обозначает функцию правдоподобия, а θ_g и θ_s представляют GMM для подлинной и поддельной речи.

Этот подход показал впечатляющие результаты на базе конкурса ASVspoof 2015 и достиг улучшения в 72% по сравнению с лучшей системой, участвовавшей в том конкурсе. Исходя из этого, организаторами ASVspoof 2017 реализация этой системы была выбрана в качестве базовой.

В качестве альтернативного подхода также рассматривалась базовая система с дополнительными этапами нормализации спектра и кепстра сигнала по среднему и дисперсии (Mean and Variance Normalization, MVN-нормализация) (рис. 1, б).

Система на основе i-векторов

Наиболее эффективные для детектирования спуфинг-атак системы, представленные на конкурсе ASVspoof2015, были основаны на акустических признаках высокого уровня, моделируемых с помощью i-векторов. По этой причине в рамках настоящей работы были проведены эксперименты с множеством акустических признаков, используемых в рамках конкурса ASVspoof2015, таких как мел-частотные кепстральные коэффициенты (Mel Frequency Cepstarl Coefficients, MFCC), CosPhasePC и мел-кепстральные коэффициенты, полученные на основе использования вейвлет-преобразования (Mel Wavelet Principal Coefficients, MWPC). Как показали наблюдения, лучшей стала система, основанная на кепстральных коэффициентах линейного предсказания (Linear Predictive Cepstral Coefficients, LPCC).

В базовой системе, основанной на i-векторах, были использованы 78 LPCC-коэффициентов, полученных с использованием оконной функции Хэннинга для расчета спектра сигнала. I-вектора пространства полной изменчивости размером 200 были извлечены из всех высказываний диктора, они были отцентрованы и нормализованы по длине, после чего они подавались на вход классификатора на основе метода опорных векторов (Support Vector Machine, SVM) с линейным ядром.

Методы глубокого обучения

Задача детектирования атак повторного воспроизведения может быть сведена к детектированию локальных спектральных артефактов, присутствующих в воспроизведенной речи и отличающих ее от реальной речи. Для этой цели в данном исследовании в качестве экстрактора признаков использовались CNN. CNN часто используются для выделения высокоуровневых признаков из данных единой формы, например, изображений. Эта идея может быть легко применима в задачах, связанных с классификацией аудиосигналов, если в качестве входных данных использовать частотно-временное представление аудиосигналов, например, в виде спектрограмм. Но необходимо учитывать, что вход CNN должен быть унифицирован по форме, т.е. необходимо или требовать данные в единой частотно-временной форме для каждого произнесения, или использовать оконную обработку с окном фиксированного размера.

Извлечение признаков. В данном исследовании рассматривались два типа признаков, основанных на вычислении нормализованного спектра сигнала: одни были получены через CQT-преобразование, другие – с помощью быстрого преобразования Фурье (FFT). Также использовались два подхода для получения унифицированного частотно-временного представления признаков. Первый метод обрезает спектр вдоль временной оси до фиксированной длины. При этом спектры коротких файлов расширяются дублированием своего содержимого таким образом, чтобы достичь требуемой длины. Второй использует метод скользящего окна с фиксированным размером окна.

Конволюционные нейронные сети. Предлагаемый в этой работе метод детектирования атак повторного воспроизведения основан на использовании CNN-сетей с Max-Feature-Map (MFM) активационной функцией, которая является расширением Max-out активационной функции [16]. Функция MFM определяется как

$$y_{ij}^{k} = \max\left(x_{ij}^{k}, x_{ij}^{k+\frac{N}{2}}\right) \quad \forall i = \overline{1,H},\ j = \overline{1,W}, k = \overline{1,N/2},$$

где x – входной тензор размера $H \times W \times N$, y – результирующий тензор размера $H \times W \times N/2$, i, j указывает частотные и временные области, а k – индекс канала. Рис. 2 демонстрирует функцию MFM для конволюционного слоя.

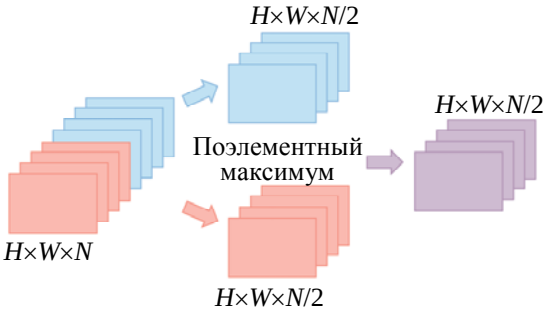

Рис. 2. MFM для конволюционного слоя [16]

Использование MFM позволило уменьшить архитектуру CNN, поэтому архитектура CNN называется легкой CNN (Light CNN, LCNN) [17]. В отличие от обычно используемой функции Rectified Linear Unit, которая «возбуждает» нейрон с помощью порога срабатывания, MFM делает это конкурентными отношениями. Таким образом, MFM играет роль селектора признаков. Мы использовали редуцированную версию CNN, предложенной в [17], с меньшим количеством фильтров в каждом слое (табл. 1).

Type	Filter / Stride	Output	#Params
Conv1	5 × 5 / 1 × 1	864 × 400 × 32	832
MFM1	–	864 × 400 × 16	–
MaxPool1	2 × 2 / 2 × 2	432 × 200 × 16	–
Conv2a	1 × 1 / 1 × 1	432 × 200 × 32	544
MFM2a	–	432 × 200 × 16	–
Conv2b	3 × 3 / 1 × 1	432 × 200 × 48	7.0K
MFM2b	–	432 × 200 × 24	–
MaxPool2	2 × 2 / 2 × 2	216 × 100 × 24	–
Conv3a	1 × 1 / 1 × 1	216 × 100 × 48	1.2K
MFM3a	–	216 × 100 × 32	–
Conv3b	3 × 3 / 1 × 1	216 × 100 × 64	13.9K
MFM3b	–	216 × 100 × 32	–
MaxPool3	2 × 2 / 2 × 2	108 × 50 × 32	–
Conv4a	1 × 1 / 1 × 1	108 × 50 × 64	2.1K
MFM4a	–	108 × 50 × 32	–
Conv4b	3 × 3 / 1 × 1	108 × 50 × 32	9.2K
MFM4b	–	108 × 50 × 16	–
MaxPool4	2 × 2 / 2 × 2	54 × 25 × 16	–
Conv5a	1 × 1 / 1 × 1	54 × 25 × 32	544
MFM5a	–	54 × 25 × 16	–
Conv5b	3 × 3 / 1 × 1	54 × 25 × 32	4.6K
MFM5b	–	54 × 25 × 16	–
MaxPool5	2 × 2 / 2 × 2	27 × 12 × 16	–
FC6	–	32 × 2	332K
MFM6	–	32	–
FC7	–	2	64
Total	–	–	371K

Таблица 1. Архитектура CNN-сети [16]

Каждый сверточный слой представляет собой комбинацию двух независимых сверточных частей, вычисленных на входе слоя. Затем используется функция активации MFM, чтобы вычислить максимальный размер этих частей. Слои пулинга с размером ядра 2×2 и шагом 2×2 используются для уменьшения размеров как по временной оси, так и по частотной. Полносвязный слой FC6 содержит низкоразмерное аудио-представление высокого уровня. Слой FC7 с функцией активации softmax используется, чтобы различать классы подлинной речи и спуфинга во время обучения сети. Данная CNN использовалась для получения высокоуровневых акустических признаков со слоя FC6. В полученном пространстве признаков для детектирования спуфинга достаточно использовать простой классификатор на основе GMM.

Рассматривалось несколько систем, основанных на LCNN с различными акустическими признаками. В системе $LCNN_{CQT}$ использовались усеченные признаки, полученные из нормализованных CQT-спектрограмм размером 864×400×1. Дополнительно исследовались признаки, основанные на FFT: $LCNN_{FFT}$ использовала усеченные признаки размером 864×400×1, а $LCNN_{FFT}^{SW}$ использовала скользящее окно для извлечения признаков размера 864×200×1 и 90% перекрытием вдоль временной оси.

Объединение конволюционной и рекуррентной нейронных сетей. Следуя работе [12], была рассмотрена объединенная архитектура CNN + RNN. Основная идея такой интеграции заключается в том, что CNN используется для выделения признаков, а RNN моделирует долгосрочные зависимости. CNN и RNN оптимизируются совместно посредством обратного распространения ошибки и образуют End-to-End решение.

Общая архитектура показана в табл. 2. Здесь CNN представляет собой редуцированную версию LCNN, но в отличие от предыдущих систем, пулинг по максимуму применяется с шагом 2 по оси частот для сжатия частотной информации и с шагом 1 вдоль временной оси для сохранения временной размерности.

Type	Filter / Stride	Output	#Params
Conv1	5 × 5 / 1 × 1	256 × 400 × 16	416
MFM1	–	256 × 400 × 8	–
MaxPool1	2 × 2 / 2 × 1	128 × 400 × 8	–
Conv2a	1 × 1 / 1 × 1	128 × 400 × 16	144
MFM2a	–	128 × 400 × 8	–
Conv2b	3 × 3 / 1 × 1	128 × 400 × 32	2.3K
MFM2b	–	128 × 400 × 16	–
MaxPool2	2 × 2 / 2 × 1	64 × 400 × 16	–
Conv3a	1 × 1 / 1 × 1	64 × 400 × 32	544
MFM3a	–	64 × 400 × 16	–
Conv3b	3 × 3 / 1 × 1	64 × 400 × 16	2.3K
MFM3b	–	64 × 400 × 8	–
MaxPool3	2 × 2 / 2 × 1	32 × 400 × 8	–
BGRU	–	16 × 2 × 8	40.3K
FC4	–	512 × 2	263K
MFM4	–	512	–
FC5	–	256 × 2	263K
MFM5	–	256	–
FC6	–	1	257
Total	–	–	572K

Таблица 2. Архитектура сети CNN+RNN [16]

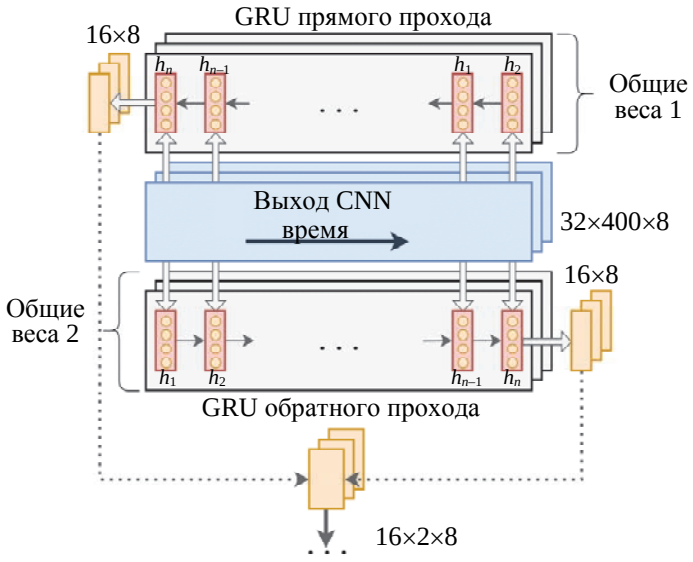

Рис. 3. Двунаправленный GRU [16]. h_i – выходные векторы прямого и обратного прохода

RNN состоит из двух управляемых рекуррентных блоков (Gated Recurrent Unit, GRU) [18], образующих двунаправленный управляемый рекуррентный блок (BGRU). Первый GRU отвечает за прямой проход и обрабатывает данные от первого входного вектора до последнего (рис. 3). Второй GRU обрабатывает данные от последнего входного вектора до первого, совершая обратный проход. Выходные векторы прямого и обратного прохода используются далее для формирования двух 16-мерных векторов. Такой блок BGRU применяется к каждому каналу выхода CNN, что приводит к тензору 16×2×8. Используются общие веса для блоков каждого канала для предотвращения переобучения. Сглаженный выход RNN используется в качестве входа на полносвязный слой с MFM-активацией, который выдает вероятность того, что высказывание является спуфингом.

Система, основанная на данной архитектуре $CNN_{FFT} + RNN$, использовала усеченные признаки, извлеченные из амплитудного FFT-спектра размера 256 ×400×1.

Экспериментальные результаты

Все эксперименты в данной работе были проведены на базах конкурса ASVspoof 2017. Подробное описание этих наборов данных можно найти в [6]. Для обучения всех систем, рассматриваемых в этой статье, использовалась только обучающая база. База валидации использовалась для подбора весовых коэффициентов для последующего объединения систем. Тестовая база включает записи новых дикторов, новые условия, новые комбинации звукозаписывающих устройств повторного воспроизведения и новые атаки, которые отличаются от атак обучающей и валидационной баз. Сравнение предлагаемых систем на тестовой базе является наиболее репрезентативным (табл. 3).

	Валидационное множество	Тестовое множество
Базовый метод	10,35	30,60
Модификация базового метода	9,85	17,31
SVM_{i-vect}	9,80	12,54
$\boldsymbol{LCNN_{FFT}}$	**4,53**	**7,37**
$LCNN_{FFT}^{SW}$	5,25	11,81
$LCNN_{CQT}$	4,80	16,54
$CNN_{FFT} + RNN$	7,51	10,69
$LCNN_{FFT}, SVM_{i-vect}, CNN_{FFT} + RNN$	4,5	6,78

Таблица 3. Результаты экспериментов на данных конкурса ASVspoof 2017

Наилучший результат на множествах валидации и оценки продемонстрировала система $LCNN_{FFT}$. Аналогичная система, использующая CQT-спектрограммы, показала слабую стабильность на тестовом множестве. Это объясняется низкой надежностью признаков CQT, которая также подтверждается результатами базовой системы. Метод скользящего окна показал худшие результаты по сравнению с усеченным подходом для извлечения признаков. Возможная причина этого – в том, что использование спектрограмм всего высказывания (в большинстве случаев) в качестве входа CNN приводит к более точной текстозависимой модели. Описанная комбинация CNN + RNN также показала худшее качество обнаружения атак повторного воспроизведения, чем индивидуальная LCNN. Такое ухудшение производительности объясняется уменьшением частотного разрешения спектра.

Основная система, представленная на конкурсе ASVspoof 2017, представляла собой объединение методов $LCNN_{FFT}$, SVM_{i-vect} и $CNN_{FFT} + RNN$ с помощью линейной комбинации финальных оценок каждой из подсистем. Данная система продемонстрировала уровень равновероятной ошибки (Equal Error Rate, EER), равный 3,95% и 6,73% на валидационном и тестовом множествах соответственно.

Заключение

В работе была исследована применимость подхода глубокого обучения для решения проблемы обнаружения спуфинг-атак повторного воспроизведения. В рамках этого исследования были рассмотрены индивидуальная CNN и в сочетании с подходами RNN. Наши эксперименты, проведенные на базах конкурса ASVspoof 2017, подтвердили высокую эффективность использования методов глубокого обучения для обнаружения спуфинга в реальных условиях. EER лучшей индивидуальной системы CNN составлял 7,34%. Лучшая система, основанная на объединении нескольких систем, обеспечила 6,73% EER на тестовой базе.

Литература

1. Sebastien M., Nixon M.S., Li S.Z. Handbook of Biometric Anti-Spoofing: Trusted Biometrics under Spoofing Attacks. Springer, 2014. 281 p. doi: 10.1007/978-1-4471-6524-8

References

1. Sebastien M., Nixon M.S., Li S.Z. *Handbook of Biometric Anti-Spoofing: Trusted Biometrics under Spoofing Attacks*. Springer, 2014, 281 p. doi: 10.1007/978-1-4471-6524-8

2. Faundez-Zanuy M., Hagmuller M., Kubin G. Speaker verification security improvement by means of speech watermarking // Speech Communication. 2006. V. 48. N 12. P. 1608–1619. doi: 10.1016/j.specom.2006.06.010
3. Wu Z., Evans N., Kinnunen T., Yamagishi J., Alegre F., Li H. Spoofing and countermeasures for speaker verification: A survey // Speech Communication. 2005. V. 66. P. 130–153. doi: 10.1016/j.specom.2014.10.005
4. Wu Z., Kinnunen T., Evans N., Yamagishi J., Hanilci C., Sahidullah M., Sizov A. ASVspoof: the automatic speaker verification spoofing and countermeasures challenge // IEEE Journal of Selected Topics in Signal Processing. 2017. V. 11. N 4. P. 588–604. doi: 10.1109/JSTSP.2017.2671435
5. Villalba J., Lleida E. Preventing replay attacks on speaker verification systems // Proc. IEEE Int. Carnahan Conf. on Security Technology. Barcelona, Spain, 2011. 8 p. doi: 10.1109/CCST.2011.6095943
6. Kinnunen T., Sahidullah M., Delgado H., Todisco M., Evans, N., Yamagishi J.,Lee K.A. The ACVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection // Proc. of Interspeech. Stockholm, Sweden, 2017. P. 2–6. doi: 10.21437/Interspeech.2017-1111
7. Karpathy A., Toderici G., Shetty S., Leung T., Sukthankar R., Li F.F. Large-scale video classification with convolutional neural networks // Proc. IEEE Conf. on Computer Vision and Pattern Recognition. Columbus, USA, 2014. P. 1725–1732. doi: 10.1109/CVPR.2014.223
8. Bengio Y., Courville A., Vincent P. Representation learning: a review and new perspectives // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013. V. 35. N 8. P. 1798–1828. doi: 10.1109/TPAMI.2013.50
9. Krizhevsky A., Sutskever I., Hinton G. E. ImageNet classification with deep convolutional neural networks // Advances in Neural Information Processing Systems. 2012. V. 2. P. 1097–1105.
10. Taigman Y., Yang M., Ranzato M., Wolf L. Deepface: Closing the gap to human-level performance in face verification // Proc. IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014. P. 1701–1708. doi: 10.1109/CVPR.2014.220
11. Волкова С.С., Матвеев Ю.Н. Применение сверточных нейронных сетей для решения задачи противодействия атаке спуфинга в системах лицевой биометрии // Научно-технический вестник информационных технологий, механики и оптики. 2017. Т. 17. № 4. С. 702–710. doi: 10.17586/2226-1494-2017-17-4-702-710
12. Zhang C., Yu C., Hansen J.H.L. An investigation of deep-learning frameworks for speaker verification anti-spoofing // IEEE Journal of Selected Topics in Signal Processing. 2017. V. 11. N 4. P. 684–694. doi: 10.1109/JSTSP.2016.2647199
13. Tian X., Xiao X., Siong C. E., Li H. Spoofing speech detection using temporal convolutional neural network // Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Jeju, South Korea, 2016. doi: 10.1109/APSIPA.2016.7820738
14. Lee K.A., Larcher A., Wang G. et al. The RedDots data collection for speaker recognition // Proc. of Interspeech. Dresden, Germany, 2015. P. 2996–3000.
15. Todisco M., Delgado H., Evans N. A new feature for automatic speaker verification antispoofing: Constant Q cepstral coefficients // Proc. Odyssey. Bilbao, Spain, 2016. doi: 10.21437/odyssey.2016-41
16. Lavrentyeva G., Novoselov S., Malykh E., Kozlov A., Kudashev O., Shchemelinin V. Audio replay attack detection with deep learning frameworks // Proc. of Interspeech. Stockholm, Sweden, 2017. P. 82–86. doi: 10.21437/Interspeech.2017-360
17. Wu X., He R., Sun Z., Tan T. A light CNN for deep face representation with noisy labels // arXiv: 1511.02683. 2015. 13 p.
18. Chung J., Gulcehre C., Cho K., Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling // arXiv:1412.3555. 2014.

2. Faundez-Zanuy M., Hagmuller M., Kubin G. Speaker verification security improvement by means of speech watermarking. *Speech Communication*, 2006, vol. 48, no. 12, pp. 1608–1619. doi: 10.1016/j.specom.2006.06.010
3. Wu Z., Evans N., Kinnunen T., Yamagishi J., Alegre F., Li H. Spoofing and countermeasures for speaker verification: a survey. *Speech Communication*, 2005, vol. 66, pp. 130–153. doi: 10.1016/j.specom.2014.10.005
4. Wu Z., Kinnunen T., Evans N., Yamagishi J., Hanilci C., Sahidullah M., Sizov A. ASVspoof: the automatic speaker verification spoofing and countermeasures challenge. *IEEE Journal of Selected Topics in Signal Processing*, 2017, vol. 11, no. 4, pp. 588–604. doi: 10.1109/JSTSP.2017.2671435
5. Villalba J., Lleida E. Preventing replay attacks on speaker verification systems. *Proc. IEEE Int. Carnahan Conf. on Security Technology*. Barcelona, Spain, 2011, 8 p. doi: 10.1109/CCST.2011.6095943
6. Kinnunen T., Sahidullah M., Delgado H., Todisco M., Evans, N., Yamagishi J.,Lee K.A. The ACVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. *Proc. of Interspeech*. Stockholm, Sweden, 2017, pp. 2–6. doi: 10.21437/Interspeech.2017-1111
7. Karpathy A., Toderici G., Shetty S., Leung T., Sukthankar R., Li F.F. Large-scale video classification with convolutional neural networks. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*. Columbus, USA, 2014, pp. 1725–1732. doi: 10.1109/CVPR.2014.223
8. Bengio Y., Courville A., Vincent P. Representation learning: a review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, vol. 35, no. 8, pp. 1798–1828. doi: 10.1109/TPAMI.2013.50
9. Krizhevsky A., Sutskever I., Hinton G. E. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 2012, vol. 2, pp. 1097–1105.
10. Taigman Y., Yang M., Ranzato M., Wolf L. Deepface: Closing the gap to human-level performance in face verification. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, USA, 2014, pp. 1701–1708. doi: 10.1109/CVPR.2014.220
11. Volkova S.S., Matveev Yu.N. Convolutional neural networks for face anti-spoofing. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2017, vol. 17, no. 4, pp. 702–710 (in Russian). doi: 10.17586/2226-1494-2017-17-4-702-710
12. Zhang C., Yu C., Hansen J.H.L. An investigation of deep-learning frameworks for speaker verification anti-spoofing. *IEEE Journal of Selected Topics in Signal Processing*, 2017, vol. 11, no. 4, pp. 684–694. doi: 10.1109/JSTSP.2016.2647199
13. Tian X., Xiao X., Siong C. E., Li H. Spoofing speech detection using temporal convolutional neural network. *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. Jeju, South Korea, 2016. doi: 10.1109/APSIPA.2016.7820738
14. Lee K.A., Larcher A., Wang G. et al. The RedDots data collection for speaker recognition. *Proc. of Interspeech*. Dresden, Germany, 2015, pp. 2996–3000.
15. Todisco M., Delgado H., Evans N. A new feature for automatic speaker verification antispoofing: Constant Q cepstral coefficients. *Proc. Odyssey*. Bilbao, Spain, 2016. doi: 10.21437/odyssey.2016-41
16. Lavrentyeva G., Novoselov S., Malykh E., Kozlov A., Kudashev O., Shchemelinin V. Audio replay attack detection with deep learning frameworks. *Proc. of Interspeech*. Stockholm, Sweden, 2017, pp. 82–86. doi: 10.21437/Interspeech.2017-360
17. Wu X., He R., Sun Z., Tan T. A light CNN for deep face representation with noisy labels. *arXiv*: 1511.02683, 2015, 13 p.
18. Chung J., Gulcehre C., Cho K., Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv*:1412.3555, 2014.

Авторы

Лаврентьева Галина Михайловна – научный сотрудник, ООО «ЦРТ-инновации», Санкт-Петербург, 196084, Российская Федерация; аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 56938815200, ORCID ID: 0000-0001-9474-098X, lavrentyeva@speechpro.com

Новосёлов Сергей Александрович – кандидат технических наук, старший научный сотрудник, ООО «ЦРТ-инновации», Санкт-Петербург, 196084, Российская Федерация; ведущий инженер, Университет ИТМО, Scopus ID: 56909843400, ORCID ID: 0000-0001-9474-098X, novoselov@speechpro.com

Козлов Александр Викторович – ведущий программист, ООО «ЦРТ-инновации», Санкт-Петербург, 196084, Российская Федерация, Scopus ID: 56352800900, ORCID ID: 0000-0002-6776-2996, kozlov-a@speechpro.com

Кудашев Олег Юрьевич – кандидат технических наук, руководитель проектов, ООО «ЦРТ-инновации», Санкт-Петербург, 196084, Российская Федерация; ведущий инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 55873653700, ORCID ID: 0000-0001-6520-2242, kudashev@speechpro.com

Щемелинин Вадим Леонидович – кандидат технических наук, руководитель отдела, ООО «ЦРТ», Санкт-Петербург, 196084, Российская федерация; инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 55873208400, ORCID ID: 0000-0002-5516-1544, shchemelinin@speechpro.com

Матвеев Юрий Николаевич – доктор технических наук, главный научный сотрудник, ООО «ЦРТ-инновации», Санкт-Петербург, 196084, Российская Федерация; заведующий кафедрой, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 7006613471, ORCID ID: 0000-0001-7010-1585, matveev@mail.ifmo.ru

Де Марсико Мария – магистр, доцент, Университет Ла Сапиенца, Рим, 00185, Италия, Scopus ID: 6508106114, ORCID ID: 0000-0002-1391-8502, demarsico@di.uniroma1.it

Authors

Galina M. Lavrentyeva – scientific researcher, STC-innovations Ltd., Saint Petersburg, 196084, Russian Federation; postgraduate, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 56938815200, ORCID ID: 0000-0001-9474-098X, lavrentyeva@speechpro.com

Sergey A. Novoselov – PhD, Senior scientific researcher, STC-innovations Ltd., Saint Petersburg, 196084, Russian Federation; Leading engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 56909843400, ORCID ID: 0000-0001-9474-098X, novoselov@speechpro.com

Alexander V. Kozlov – Leading programmer, STC-innovations Ltd., Saint Petersburg, 196084, Russian Federation, Scopus ID: 56352800900, ORCID ID: 0000-0002-6776-2996, kozlov-a@speechpro.com

Oleg Yu. Kudashev – PhD, Project manager, STC-innovations Ltd., Saint Petersburg, 196084, Russian Federation; Leading engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 55873653700, ORCID ID: 0000-0001-6520-2242, kudashev@speechpro.com

Vadim L. Shchemelinin – PhD, Department head, STC-innovations Ltd., Saint Petersburg, 196084, Russian Federation; engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 55873208400, ORCID ID: 0000-0002-5516-1544, shchemelinin@speechpro.com

Yu. N. Matveev – D.Sc., Senior researcher, STC-innovations Ltd., Saint Petersburg, 196084, Russian Federation; Head of Chair, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 7006613471, ORCID ID: 0000-0001-7010-1585, matveev@mail.ifmo.ru

Maria De Marsico – Master, Associate Professor, Sapienza University of Rome, Rome, 00185, Italy, Scopus ID: 6508106114, ORCID ID: 0000-0002-1391-8502, demarsico@di.uniroma1.it