УНИВЕРСИТЕТ ИТМО

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
июль–август 2021 Том 21 № 4 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
July–August 2021 Vol. 21 No 4 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК УНИВЕРСИТЕТ ИТМО
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2021-21-4-545-552
УДК 004.934; 004.056.53

Алгоритм выявления синтезированного голоса на основе кепстральных коэффициентов и сверточной нейронной сети

Роман Андреевич Муртазин[1], Александр Юрьевич Кузнецов[2], Евгений Андреевич Фёдоров[3], Ильнур Мидхатович Гарипов[4], Анна Викторовна Холоденина[5], Юлия Батоевна Балданова[6], Алиса Андреевна Воробьева[7]✉

[1,2,3,4,5,6,7] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[3] ООО «Лаборатория ППШ», Санкт-Петербург, 199178, Российская Федерация

[1] murtazinroman3161@gmail.com, https://orcid.org/0000-0003-3669-7586
[2] al.ur.kouznetsov@gmail.com, https://orcid.org/0000-0002-5702-3786
[3] fyodorov.1997@gmail.com, https://orcid.org/0000-0003-2911-5509
[4] i_garipov@mail.ru, https://orcid.org/0000-0003-3108-5484
[5] annaholodenina@gmail.com, https://orcid.org/0000-0003-1911-3710
[6] yuxx1004@gmail.com, https://orcid.org/0000-0002-6751-8993
[7] Alice_w@mail.ru✉, https://orcid.org/0000-0001-6691-6167

Аннотация

Предмет исследования. Рассмотрены существующие подходы для выявления синтетической речи, базирующиеся на проблемах синтезирования голосовой последовательности. Представлено описание этапов и итоговая схема алгоритма выявления спуфинг-атак на голосовые биометрические системы. Основное внимание уделено обнаружению синтезированного голоса как наиболее опасного вида атак. Создан программный комплекс для проведения экспериментальных исследований, представлена его структура. **Метод.** Предложен алгоритм выявления синтезированного голосового образа. Алгоритм основан на использовании мел-частотных и Q-константных кепстральных коэффициентов для извлечения речевых признаков. Для построения модели пользователя использована модель гауссовых смесей. В качестве классификатора для принятия решения о подлинности голоса выбрана сверточная нейронная сеть. **Основные результаты.** Для сопоставления выбраны два базовых решения противодействия спуфинг-атакам, предложенные авторами конкурса ASVspoof2019. В одном из решений в качестве извлекаемых речевых признаков использованы линейно-частотные кепстральные коэффициенты, в другом — Q-константные. В обоих решениях в качестве классификатора применена модель гауссовых смесей. Для оценки эффективности предложенного решения и сравнения его с другими выбраны метрики EER и minDCF и сформирована голосовая база. Экспериментальные результаты продемонстрировали преимущество разработанного алгоритма перед другими рассмотренными вариантами. Достоинство представленного решения — применение извлекаемых речевых признаков, имеющих высокие результаты и для идентификации пользователя. Это позволяет оптимизировать голосовую биометрическую систему с внедренной защитой от спуфинг-атак посредством синтеза голоса. Сам алгоритм при внесении незначительных модификаций может быть использован для голосовой идентификации. **Практическая значимость.** Голосовые биометрические системы имеют высокий потенциал применения в банковской сфере. Такие системы позволят финансовым организациям ускорить и упростить осуществление денежных операций, и предоставить пользователям расширенный функционал в удаленном режиме. Внедрение систем голосовой биометрической идентификации осложняется их уязвимостью для спуфинг-атак, в частности посредством синтеза голоса. Предложенное решение может быть интегрировано в системы голосовой биометрии с целью повышения их надежности.

Ключевые слова

биометрия, голосовые биометрические системы в банковской сфере, синтезированная речь, выявление фальсификации голоса, кепстральный анализ, сверточная нейронная сеть

Благодарности

Работа выполнена в Университете ИТМО в рамках темы НИР № 50449 «Разработка алгоритмов защиты киберпространства для решения прикладных задач обеспечения кибербезопасности организаций банковской сферы».

Ссылка для цитирования: Муртазин Р.А., Кузнецов А.Ю., Фёдоров Е.А., Гарипов И.М., Холоденина А.В., Балданова Ю.Б., Воробьева А.А. Алгоритм выявления синтезированного голоса на основе кепстральных коэффициентов и сверточной нейронной сети // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21, № 4. С. 545–552. doi: 10.17586/2226-1494-2021-21-4-545-552

The speech synthesis detection algorithm based on cepstral coefficients and convolutional neural network

Roman A. Murtazin[1], Aleksandr Yu. Kuznetsov[2], Evgeny A. Fedorov[3], Ilnur M. Garipov[4], Anna V. Kholodenina[5], Yulia B. Baldanova[6], Alisa A. Vorobeva[7]✉

[1,2,3,4,5,6,7] ITMO University, Saint Petersburg, 197101, Russian Federation
[3] Laboratory PPS Ltd, Saint Petersburg, 199178, Russian Federation

[1] murtazinroman3161@gmail.com, https://orcid.org/0000-0003-3669-7586
[2] al.ur.kouznetsov@gmail.com, https://orcid.org/0000-0002-5702-3786
[3] fyodorov.1997@gmail.com, https://orcid.org/0000-0003-2911-5509
[4] i_garipov@mail.ru, https://orcid.org/0000-0003-3108-5484
[5] annaholodenina@gmail.com, https://orcid.org/0000-0003-1911-3710
[6] yuxx1004@gmail.com, https://orcid.org/0000-0002-6751-8993
[7] Alice_w@mail.ru✉, https://orcid.org/0000-0001-6691-6167

Abstract

The existing approaches to detecting synthesized speech, based on the current issues of synthesizing voice sequences, are considered. The stages of the algorithm for detecting spoofing attacks on voice biometric systems are described, and its final workflow is presented. The research focuses mainly on detecting synthesized speech, as it is the most dangerous type of attacks. The authors designed a software application for an experimental study, present its structure and propose the detection synthesized speech algorithm. This algorithm uses mel-frequency and constant Q cepstral coefficients to extract speech features. A Gaussian mixture model is used to construct a user model. Convolutional neural network was chosen as a classifier to determine the voice's authenticity. Two basic methods for combating spoofing attacks, proposed by the authors of the ASVspoof2019 competition, were selected for making comparisons. One of these methods involved using linear frequency cepstral coefficients as speech features, while the other method used constant Q. Both solutions used Gaussian mixture models for classification. To evaluate the effectiveness of the proposed solution and compare it with other methods, a voice database was created. The selected EER and minDCF metrics were applied. The experimental results demonstrated the advantages of the proposed algorithm in comparison with the other algorithms. An advantage of the proposed solution is that it uses extracted speech features that perform efficiently when it comes to user identification. This makes it possible to use the algorithm to optimize a voice biometric system that has embedded protection against spoofing attacks that is built on speech synthesis. In addition, it is possible to use the proposed method for voice identification with minimal modifications required. Voice biometric identification systems have excellent opportunities in the banking sector. Such systems allow banks to simplify and accelerate the process of financial transactions and provide their users with advanced banking functions remotely. The implementation of voice biometric systems is difficult by their vulnerability to spoofing attacks, particularly to those conducted by means of speech synthesis. The proposed solution can be integrated into voice biometric systems to improve their security.

Keywords

biometric, automatic speaker verification in banking, synthetic speech, spoofing detection, cepstral analysis, convolutional neural network

Acknowledgements

The paper was prepared at ITMO University within the framework of the scientific project No. 50449 "Development of cyberspace protection algorithms for solving applied problems of ensuring cybersecurity of banking organizations".

For citation: Murtazin R.A., Kuznetsov A.Yu., Fedorov E.A., Garipov I.M., Kholodenina A.V., Baldanova Yu.B., Vorobeva A.A. The speech synthesis detection algorithm based on cepstral coefficients and convolutional neural network. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2021, vol. 21, no. 4, pp. 545–552 (in Russian). doi: 10.17586/2226-1494-2021-21-4-545-552

Введение

Использование голосовой биометрии является одним из наиболее перспективных направлений в области биометрической идентификации по ряду причин. Во-первых, человеческий голос — динамический, а не статический биометрический признак, что усложняет процесс его подделки. Также для осуществления голосовой идентификации требуется гораздо менее дорогостоящее оборудование[1], и данный биометрический признак один из немногих позволяет осуществить подтверждение личности удаленно [1].

Из последних двух факторов следует, что потенциально использование голосовой биометрии станет одним из самых доступных способов идентификации и сможет использоваться в большом количестве самых разных областей [2].

По прогнозу экспертов[2] в ближайшие годы темпы роста рынка систем голосовой биометрии будут ежегодно составлять более 20 %, что является ведущим показателем среди систем, использующих биометрические признаки.

Развитие технологий идентификации личности по голосу открывает новые возможности и в банковском

[1] Мультимодальная система доступа с использованием голосовой биометрии [Электронный ресурс]. Режим доступа: https://indeed-id.ru/blog/multimodalnaya-sistema-dostupa-s-is/ (дата обращения: 04.04.2021).

[2] Обзор международного рынка биометрических технологий и их применение в финансовом секторе [Электронный ресурс]. Режим доступа: http://www.cbr.ru/ (дата обращения: 05.03.2021).

секторе — появляются возможности расширения списка предоставляемых функций и услуг, повышается клиентоориентированность в обслуживании пользователей [3]. Тем не менее, повсеместное использование данной биометрической характеристики для идентификации личности осложняется недостаточно высокой точностью идентификации пользователей [4] и уязвимостью голосовых биометрических систем (ГБС) к различным спуфинг-атакам. Атаки можно различить на виды: имперсонализацию, воспроизведение, морфинг, синтез голоса.

По совокупности факторов область технологий синтеза голоса — наиболее активно развивающаяся и потенциально представляет для ГБС опасность. В связи с этим возникает необходимость построения подсистемы защиты ГБС от спуфинг-атак посредством синтеза голоса, основывающейся на алгоритме его выявления. Для исключения влияния акустической среды между системой преобразования текста в речь (“text-to-speech”, TTS) и ГБС, первую чаще всего подключают к устройству связи и не воспроизводят выходной сигнал системы синтеза перед данным устройством. На выявление таких атак через прямое подключение TTS-системы к устройству связи и нацелено предлагаемое решение.

Основные этапы алгоритма выявления синтезированного голоса

Сформулированные в работе [5] принципы построения защищенных ГБС и обозначенные в работах [6] проблемы синтеза[1] голоса позволяют выделить три основных подхода к построению защиты от атак посредством синтезирования голосового образа.

Первый основывается на определении и отслеживании параметров речи, зависящих от психофизического состояния пользователя. Второй — выявление свойственной каждому пользователю семантики (логическое ударение, членение речи, стиль построения речевых конструкций, смысловая нагрузка). Третий — анализ самого входного сигнала: контроль за уровнем дискретизации и сглаженностью сигнала, поиск спекл-шумов, следов склеек и одинаковых звуковых элементов, перепадов в уровне сигнала. Использование первых двух подходов почти невозможно из-за недостаточной проработки данных областей. Таким образом, третий подход в настоящее время — наиболее эффективен и служит основой для разработанного алгоритма выявления атак на ГБС посредством синтеза голоса.

Реализация ГБС в банковской сфере предусматривает возможность идентификации пользователя при его звонке с обычного смартфона. В таком случае входной сигнал может содержать шумы или помехи. Их наличие негативно сказывается на качестве построенных на основе речевых признаков моделей пользователей, используемых для распознавания личности и оценки подлинности голоса. Таким образом, первый этап алгоритма выявления синтезированного голосового образа (ВСГО) — предобработка входного сигнала.

Для повышения точности извлечения речевых признаков из цифрового сигнала, принятого на вход системой, выделяются те участки, на которых присутствует речь, т. е. из звуковой дорожки удаляется неинформативный сигнал (рис. 1).

Затем производится нормализация громкости, так как диапазоны громкости для разных звуковых дорожек могут не совпадать. На данной стадии громкость звука в исходных файлах нормализуется относительно выбранного эталона. В разработанном алгоритме предлагается применение способа RMS (root mean square) так как он больше остальных подходит для человеческого уха:

$$RMS = \sqrt{\frac{a_1^2 + a_2^2 + ... + a_n^2}{n}},$$

где a — отдельный образец голоса; n — количество образцов голоса.

Для приведения голосовых образцов, поступающих на вход в систему, к единому формату используется нормализация частоты дискретизации – децимация. Ее следствием является уменьшение объема данных, что сокращает время обучения системы:

[1] Устойчивость обучения GAN [Электронный ресурс]. Режим доступа: https://habr.com/ru/post/416531/ (дата обращения: 01.04.2021).

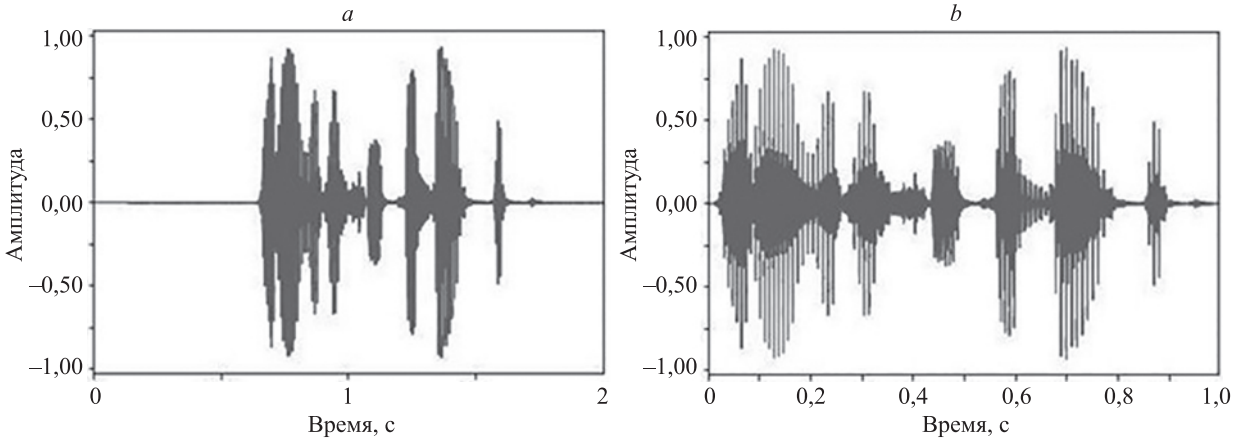

Рис. 1. Пример голосового образца без удаления (*a*) и с удалением (*b*)

Fig. 1. An example of speech without (*a*) and with (*b*) removal of silence

$$y[n] = \sum_{k=0}^{M-1} x[nM + k],$$

где $y[n]$ и $x[n]$ — отсчеты нормализованного и входного сигналов соответственно; M — длительность строба; k — коэффициент децимации.

Так как большинство систем синтеза речи генерируют звук в один канал (моно), а большая часть файлов с исходной речью состоит из двух каналов (стерео), все двухканальные файлы преобразуются в один канал посредством микширования каналов.

После обработки входного сигнала из него извлекаются речевые признаки. В разработанном алгоритме используются мел-частотные (MFCC) и Q-константные кепстральные коэффициенты (CQCC). Кепстр является ортогональным разложением спектра. В его основе лежит отображение N коэффициентов Фурье на значительно меньшее количество независимых кепстральных коэффициентов q, содержащих наиболее значимую информацию из спектра. MFCC считаются стандартными методами извлечения признаков при обработке речи, а CQCC показывают лучшую производительность обнаружения, особенно для неизвестных атак [7].

Установлено, что показатели ВСГО при применении CQCC выше, чем MFCC. Но обучение на MFCC позволяет получить более низкий коэффициент равной ошибки (EER) для распознавания синтезированного голоса при недостаточной тренировке алгоритма на других видах атак [8]. Также преимуществом данных признаков является их применимость для идентификации пользователей, независимость от текста, диктора и языка. В результате было принято решение использовать оба метода извлечения признаков.

MFCC используются для описания характеристик фонемы, для их извлечения исходный речевой сигнал записывается в дискретном виде:

$$x[n],\ 0 \le n < N,$$

где $x[n]$ — отсчет сигнала в определенный момент времени; n — номер отсчета; N — длина сигнала.

Затем к нему применяется преобразование Фурье для получения спектра исходного сигнала, и составляется набор треугольных фильтров (оконной функции). Вычисляется энергия для каждого фильтра в логарифмической шкале:

$$S[m] = \ln\left(\sum_{k=0}^{N-1} |X_a[k]|^2 \times H_m[k]\right),\ 0 \le m < M,$$

где $S[m]$ — энергия фильтра; k — индекс частоты; $X_a[k]$ — k-ая комплексная амплитуда, составляющая часть исходного сигнала; H_m — весовые коэффициенты полученных фильтров; M — число фильтров; m — треугольный фильтр.

В результате вычисляется кепстральный коэффициент c. Повторные вычисления[1] позволяют получить набор MFCC.

[1] Мел-кепстральные коэффициенты (MFCC) и распознавание речи [Электронный ресурс]. Режим доступа: https://habr.com/ru/post/140828/ (дата обращения: 25.03.2021).

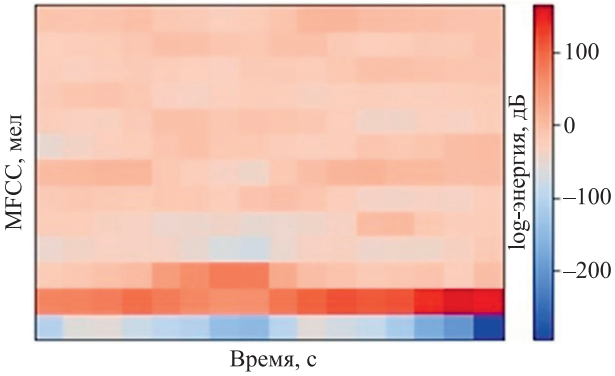

Рис. 2. Представление MFCC в графическом виде
Fig. 2. Graphical representation of MFCC

$$c[n] = \sum_{m=0}^{M-1} S[m]\cos\left(\frac{\pi n\left(m + \frac{1}{2}\right)}{M}\right),\ 0 \le n < M.$$

Пример графического представления MFCC представлен на рис. 2.

Постоянное Q-преобразование (CQT) дает более высокое разрешение в области низких частот и большее временное разрешение в области высоких частот. CQCC извлекаются в соответствии с формулой:

$$CQCC[p] = \sum_{l=1}^{L} \log|X^{CQ}[l]|^2 \cos\left[\frac{p\pi\left(l - \frac{1}{2}\right)}{L}\right],$$
$$0 \le p < L,$$

где $\log|X^{CQ}[l]|^2$ — линейный логарифмический спектр мощности; l — пересчитанные частотные значения; L — номер линейного элемента разрешения по частоте.

На рис. 3 представлена обобщенная схема алгоритма получения CQCC.

Пересчет частотных значений необходим из-за преобразования геометрического пространства в линейное, в результате которого: изменяется расстояние между октавами в зависимости от исходных частотных значений; для каждой следующей октавы по сравнению с предыдущей удваивается частота дискретизации [7].

После извлечения речевых признаков MFCC и CQCC подаются на вход в модель гауссовых смесей (Gaussian Mixture Model, GMM) [9–11] для построения модели пользователя. Количество компонент определено впоследствии эмпирически и равняется 512.

$$P(\mathbf{x}|G_s) = \sum_{i=1}^{M} w_i G_i(\mathbf{x}|\boldsymbol{\mu}_\mathbf{i}, \sum\mathbf{i}),$$

где $\mathbf{x}$ — вектор признаков; параметры модели, характеризующие голос диктора в виде вероятностной функций плотности: w_i — вес смеси, $\boldsymbol{\mu}_\mathbf{i}$ — вектор математического ожидания и $\sum\mathbf{i}$ — ковариационная матрица.

Для определения подлинности голоса используется классификатор на основе методов машинного обучения. С учетом всех особенностей и недостатков различных способов машинного обучения был выбран «с частичным привлечением учителя» ввиду его

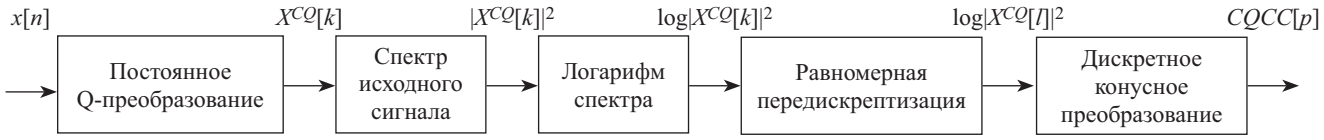

Рис. 3. Алгоритм извлечения CQCC

Fig. 3. An algorithm for extracting CQCC

гибкости, практичности и успешного опыта в выявлении спуфинг-атак [12]. В качестве классификатора в разработанном алгоритме ВСГО используется сверточная нейронная сеть (Convolutional Neural Network, CNN). К достоинствам CNN можно отнести обобщение подаваемой на вход информации при обучении, а не запоминание каждого пиксела, из-за небольшого количества настраиваемых весов. Дополнительное повышение уровня абстрактного представления данных может способствовать выявлению ключевых взаимосвязей и факторов, позволяющих с большей точностью проводить бинарную классификацию голосовых образцов.

В состав CNN входят: три слоя свертки с ядром (3,3) и функцией активации ReLU; два слоя пулинга — (2,2) и maxpooling; сглаживающий слой; три полносвязных слоя — ReLU; один полносвязный слой — sigmoid.

Общее количество параметров — 566 337. Обучение модели осуществлялось на графическом процессоре, состояло из пяти эпох, благодаря чему удалось снизить значение функции loss с 0,324 до 0,047, а параметра EER с 10,19 % до 4,79 %.

Таким образом, модель пользователя, сформированная на основе MFCC и CQCC с помощью GMM, в виде n-мерной матрицы подается на вход CNN. Результат обработки данных — ответ нейронной сети о факте выявления синтезированного голоса. Итоговая блок-схема алгоритма ВСГО представлена на рис. 4.

Данный подход к формированию структуры алгоритма основывается на предположении, что унификация форматов данных и представлений признаков MFCC и CQCC с помощью GMM перед подачей на вход в CNN, одновременно с высокими показателями самих MFCC, CQCC и CNN в выявлении синтеза голоса, позволит повысить точность бинарной классификации голосовых образцов.

Преимущество предложенного решения — оптимизация количества извлекаемых речевых признаков при использовании данного алгоритма в ГБС за счет высоких показателей MFCC и CQCC в идентификации пользователей. Более того, данный алгоритм при внесении некоторых модификаций может отдельно использоваться для подтверждения личности.

Структура и описание основных функций и модулей программного комплекса

Программный комплекс (ПК ВСГО) предназначен для выявления синтезированного голосового образа при реализации спуфинг-атак на ГБС банка. Комплекс осуществляет автоматический сбор речи пользователя в режиме реального времени при звонке в колл-центр банка, полный цикл обработки речи и предоставляет оператору информацию о подлинности голоса пользователя.

Структура ПК ВСГО представлена на рис. 5 и включает в себя подсистемы ввода голосовых данных пользователя (предназначена для оцифровки речи пользователя и ее записи) и вывода результата. Также в его состав входят модули предобработки и параметризации

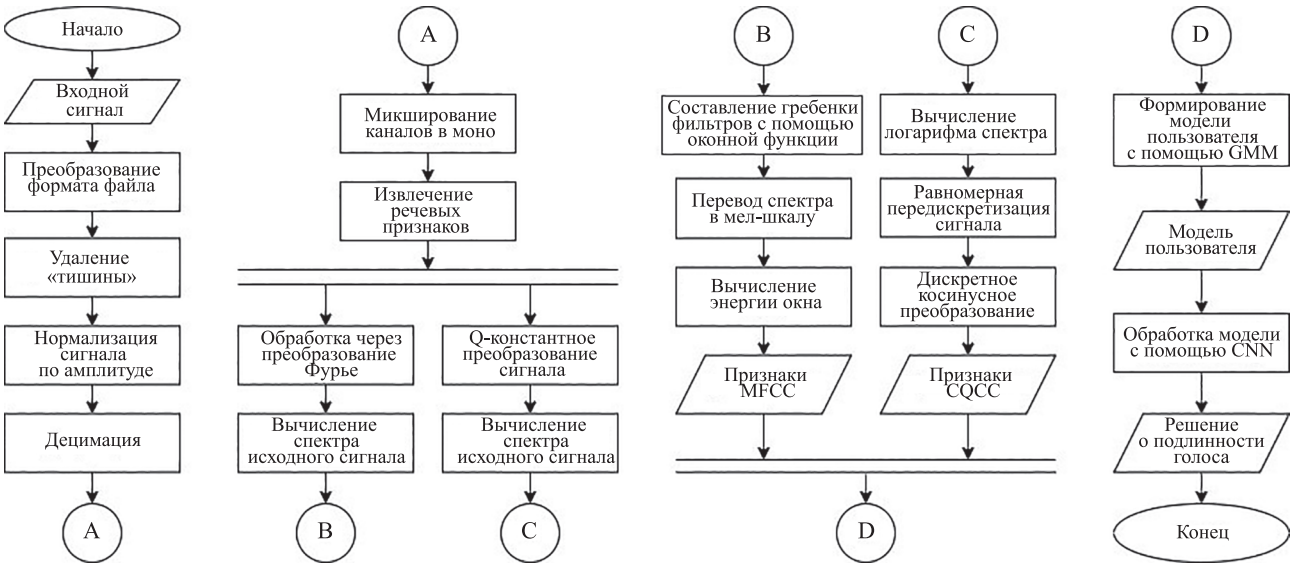

Рис. 4. Блок-схема алгоритма выявления синтезированного голоса

Fig. 4. A block diagram of the algorithm for detecting synthesized speech

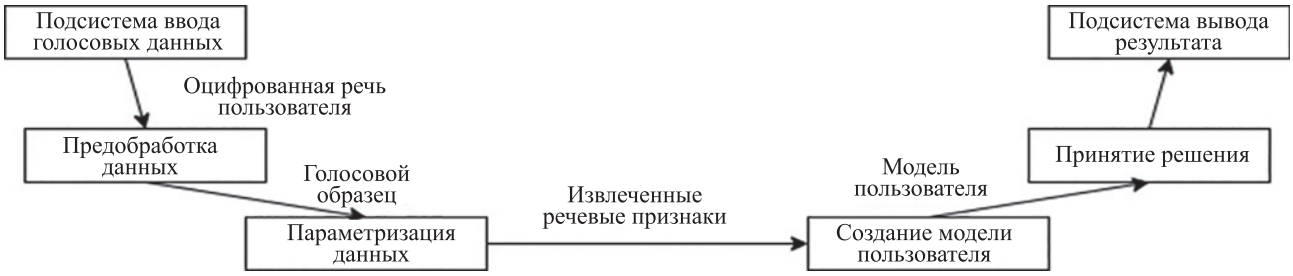

Рис. 5. Структура программного комплекса выявления синтезированного голосового образа

Fig. 5. Structure of the software application for detecting synthesized speech

данных, создания модели диктора и принятия решения. Функции модулей заключаются в выполнении одноименных этапов алгоритма ВСГО.

Экспериментальные исследования

Для проверки разработанного алгоритма выполнены лабораторные исследования с использованием специально сформированной голосовой базы (табл. 1) на основе подлинных и сгенерированных с помощью 14 различных алгоритмов образцов. Для генерации выборок обучения и отладки использованы 6 алгоритмов, для оценки – 8 (S7–S14), таким образом точность решений оценивалась только на неизвестных атаках. Голосовые образцы содержат разный текст, продолжительностью от 3 до 10 с, с частотой дискретизации 16 кГц и глубиной квантования 16 бит. Данные были отобраны из датасетов конкурсов ASVspoof15[1], ASVspoof19[2] и содержат равное количество мужских и женских голосов.

Для оценки точности выявления синтезированного голоса использовалось два параметра: EER и минимальная функция стоимости обнаружения (minDCF), которая определяется значениями ошибок первого (FAR) и (FRR). Под параметром FAR в данном случае подразумевается вероятность признания системой синтезированного голоса подлинным, а под FRR — вероятность признания подлинного голоса синтезированным.

$$\text{minDCF} = (0{,}01P_{fa} + 0{,}1P_{fr}) \times 100\ \%,$$

где P_{fa} — вероятность FAR; P_{fr} — вероятность FRR.

Для сравнения с предложенным решением использованы алгоритмы, предложенные авторами ASVspoof19, как наиболее эффективные решения с доступной программной реализацией: линейно-частотные кепстральные коэффициенты (LFCC-) в сочетании с GMM и CQCC-GMM. Алгоритмы обучены и протестированы на сформированной голосовой базе. Полученные результаты отражены в табл. 2.

Из табл. 2 следует, что представленное решение показывает лучшие результаты для большинства неизвестных атак на фоне используемых для сравнения алгоритмов, а также демонстрирует наименьшее значение метрики minDCF.

Важность снижения EER заключается в его влиянии на существенное количество параметров функционирования ГБС вместе с подсистемой защиты: средняя наработка на отказ, вероятность безотказной работы, вероятность реализации атаки, вероятность принятия системой корректного решения. Эти параметры определяются посредством анализа данных об эксплуатации, но использование решения с наименьшим EER еще на стадии разработки системы позволит снизить возможные риски.

Тем не менее, полученный результат не может считаться лучшим на данный момент в связи с рядом трудностей сравнения решений в данной области: недоступность программных реализаций большинства алгоритмов, использование собственных закрытых голосовых баз, не использование иных методик оценки эффективности алгоритмов противодействия спуфингу в сочетании с ГБС.

[1] Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2015) Database [Электронный ресурс]. Режим доступа: https://datashare.ed.ac.uk/handle/10283/853 (дата обращения: 05.05.2021).

[2] ASVspoof 2019: The 3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge database [Электронный ресурс]. Режим доступа: https://datashare.ed.ac.uk/handle/10283/3336 (дата обращения: 05.05.2021).

Таблица 1. Информация об используемой голосовой базе

Table 1. Voice database information

Параметр	Обучение	Отладка	Оценка	Всего
Дикторы	40	50	110	200
Подлинные, шт.	6200	5900	16 500	28 600
Синтезированные, шт.	20 100	34 700	89 300	144 100
Суммарно, шт.	26 300	40 600	105 800	172 700
Объем данных, ГБ	2,32	3,75	10,41	16,48

Таблица 2. Результаты экспериментальных исследований, %
Table 2. Experimental research results, %

Номер алгоритма синтеза голоса	Исследуемое решение		
	LFCC + GMM	CQCC + GMM	GMM (MFCC, CQCC) + CNN
S7	0,69	0,31	0,03
S8	10,05	9,12	0,76
S9	6,33	7,56	4,19
S10	0,85	0,64	0,37
S11	6,73	0,89	0,01
S12	10,98	7,33	8,12
S13	6,01	20,41	9,32
S14	17,09	3,91	16,51
Avg. EER	7,34	6,27	4,79
minDCF	1,312	0,989	0,623

Заключение

Разработан и реализован в виде программного комплекса алгоритм выявления синтезированного голоса на основе MFCC, CQCC и CNN с использованием GMM для построения моделей пользователя. В рамках проведенной экспериментальной оценки на специально подготовленной базе голосовых образцов данное сочетание показало наилучший результат с показателями EER равным 4,79 %, и minDCF равным 0,623.

Основными преимуществами предложенного решения являются оптимизация защищенной голосовой биометрической системы за счет использования MFCC, CQCC и низкие требования к внесению модификаций для выполнения задачи идентификации личности.

Литература

1. Мартынова А.Б., Пашковский М.Ю. Электронный банкинг и мобильный банкинг // Научно-техническое творчество аспирантов и студентов: материалы 45-й научно-технической конференции студентов и аспирантов ФГБОУ ВПО «КнАГТУ». Комсомольск-на-Амуре, 2015. С. 333–335.
2. Шилов Н.М. Области применения идентификации личности по голосу // Инновации. Наука. Образование. 2021. № 27. С. 1292–1297.
3. Маслова Е.В. Развитие рынка биометрических технологий в банковской сфере // Современные проблемы и перспективы развития банковского сектора России: Материалы III Всероссийской научно-практической конференции с международным участием. Тамбов: Тамбовский государственный университет им. Г.Р. Державина, 2018. С. 109–118.
4. Васильев Р.А., Николаев Д.Б. Анализ возможностей применения голосовой идентификации в системах разграничения доступа к информации // Научный результат. Информационные технологии. 2016. Т. 1. № 1. С. 48–57. https://doi.org/10.18413/2518-1092-2016-1-1-48-57
5. Kuznetsov A.Yu., Murtazin R.A., Garipov I.M., Fedorov E.A., Kholodenina A.V., Vorobeva A.A. Methods of countering speech synthesis attacks on voice biometric systems in banking // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21. № 1. С. 109–117. https://doi.org/10.17586/2226-1494-2021-21-1-109-117
6. Кузнецов Д.А., Кузнецов А.В., Тезин А.В., Басов О.О. Сравнительный анализ синтезаторов речи для подсистемы оповещения интеллектуального зала совещаний // Научный результат. Информационные технологии. 2018. Т. 3. № 3. С. 9–14. https://doi.org/10.18413/2518-1092-2018-3-3-0-2
7. Todisco M., Delgado H., Evans N. A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients // Odyssey 2016: Speaker and Language Recognition Workshop. 2016. P. 283–290. https://doi.org/10.21437/Odyssey.2016-41
8. Paul D., Sahidullah M., Saha G. Generalization of spoofing countermeasures: A case study with ASVspoof 2015 and BTAS 2016 corpora // Proc. of the IEEE International Conference on Acoustics,

References

1. Martynova A.B., Pashkovskii M.Iu. *Electronic and mobile banking. Scientific and Technological Works of Students: The Proceedings of the 45th Scientific and Technological Student Conference*. Komsomolsk-on-Amur, KnASTU, 2015, pp. 333–335. (in Russian)
2. Shilov N.M. Applications of voice recognition. *Innovation. Science. Education*, 2021, no. 27, pp. 1292–1297. (in Russian)
3. Maslova E.V. Biometrics Banking Market. *Modern Problems and Prospect of Banking Development in Russia: Proceedings of the 3th all-Russian scientific and practical conference with international participation*. Tambov, Tambov State University, 2018, pp. 109–118. (in Russian)
4. Vasiliev R.A., Nikolaev D.B. Analyzing the possible use of voice identification in the systems of access to information. *Research Result. Information Technologies*, 2016, vol. 1, no. 1, pp. 48–57. (in Russian). https://doi.org/10.18413/2518-1092-2016-1-1-48-57
5. Kuznetsov A.Yu., Murtazin R.A., Garipov I.M., Fedorov E.A., Kholodenina A.V., Vorobeva A.A. Methods of countering speech synthesis attacks on voice biometric systems in banking. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2021, vol. 21, no. 1, pp. 109–117. https://doi.org/10.17586/2226-1494-2021-21-1-109-117
6. Kuznetsov D.A., Kuznetsov A.V., Tezin A.V., Basov O.O. The comparative analysis of the speech synthesizers for the notification subsystem of smart hall. *Research Result. Information Technologies*, 2018, vol. 3, no. 3, pp. 9–14. (in Russian). https://doi.org/10.18413/2518-1092-2018-3-3-0-2
7. Todisco M., Delgado H., Evans N. A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients. *Odyssey 2016: Speaker and Language Recognition Workshop*, 2016, pp. 283–290. https://doi.org/10.21437/Odyssey.2016-41
8. Paul D., Sahidullah M., Saha G. Generalization of spoofing countermeasures: A case study with ASVspoof 2015 and BTAS 2016 corpora. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2047–2051. https://doi.org/10.1109/ICASSP.2017.7952516
9. Bilmes J.A. A *Gentle Tutorial of the EM Algorithm and Its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov*

Speech and Signal Processing (ICASSP). 2017. P. 2047–2051. https://doi.org/10.1109/ICASSP.2017.7952516
9. Bilmes J.A. A Gentle Tutorial of the EM Algorithm and Its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models: technical report ICSI-TR-97-021. Berkeley: University of Berkeley, 1998. 13 p.
10. Чернецова Е.А., Шишкин А.Д. Алгоритм идентификации личности по голосу для санкционирования доступа к информации // Международный научно-исследовательский журнал. 2019. № 2(80). С. 59–64. https://doi.org/10.23670/IRJ.2019.80.2.010
11. Chow D., Abdulla W.H. Robust speaker identification based on perceptual log area ratio and Gaussian mixture models // Proc. 8th International Conference on Spoken Language Processing, (ICSLP 2004). 2004. P. 1761–1764.
12. Sholokhov A., Sahidullah M., Kinnunen T. Semi-supervised speech activity detection with an application to automatic speaker verification // Computer Speech & Language. 2018. V. 47. P. 132–156. https://doi.org/10.1016/j.csl.2017.07.005

Models. Technical report ICSI-TR-97-021. Berkeley, University of Berkeley, 1998, 13 p.
10. Chernetsova E.A., Shishkin A.D. Algorithm for personal identification based on voice for information access authorization. *International Research Journal*, 2019, no. 2(80), pp. 59–64. (in Russian). https://doi.org/10.23670/IRJ.2019.80.2.010
11. Chow D., Abdulla W.H. Robust speaker identification based on perceptual log area ratio and Gaussian mixture models. *Proc. 8th International Conference on Spoken Language Processing*, (ICSLP 2004), 2004, pp. 1761–1764.
12. Sholokhov A., Sahidullah M., Kinnunen T. Semi-supervised speech activity detection with an application to automatic speaker verification. *Computer Speech & Language*, 2018, vol. 47, pp. 132–156. https://doi.org/10.1016/j.csl.2017.07.005

Авторы

Муртазин Роман Андреевич — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0003-3669-7586, murtazinroman3161@gmail.com
Кузнецов Александр Юрьевич — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57195326635, https://orcid.org/0000-0002-5702-3786, al.ur.kouznetsov@gmail.com
Фёдоров Евгений Андреевич — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация; технический специалист, ООО «Лаборатория ППШ», Санкт-Петербург, 199178, Российская Федерация, https://orcid.org/0000-0003-2911-5509, fyodorov.1997@gmail.com
Гарипов Ильнур Мидхатович — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0003-3108-5484, i_garipov@mail.ru
Холоденина Анна Викторовна — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0003-1911-3710, annaholodenina@gmail.com
Балданова Юлия Батоевна — инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0002-6751-8993, yuxx1004@gmail.com
Воробьева Алиса Андреевна — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57191359167, https://orcid.org/0000-0001-6691-6167, Alice_w@mail.ru

Authors

Roman A. Murtazin — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0003-3669-7586, murtazinroman3161@gmail.com
Aleksandr Yu. Kuznetsov — PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57195326635, https://orcid.org/0000-0002-5702-3786, al.ur.kouznetsov@gmail.com

Evgeny A. Fedorov — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation; Technical Specialist, Laboratory PPS Ltd, Saint Petersburg, 199178, Russian Federation, https://orcid.org/0000-0003-2911-5509, fyodorov.1997@gmail.com

Ilnur M. Garipov — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0003-3108-5484, i_garipov@mail.ru
Anna V. Kholodenina — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0003-1911-3710, annaholodenina@gmail.com
Yulia B. Baldanova — Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0002-6751-8993, yuxx1004@gmail.com
Alisa A. Vorobeva - PhD, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57191359167, https://orcid.org/0000-0001-6691-6167, Alice_w@mail.ru

Статья поступила в редакцию 16.05.2021
Одобрена после рецензирования 25.06.2021
Принята к печати 25.07.2021

Received 16.05.2021
Approved after reviewing 25.06.2021
Accepted 25.07.2021