БЕЗОПАСНОСТЬ ЦИФРОВЫХ ТЕХНОЛОГИЙ. – 2022. – № 3 (106). – 62–80

МЕТОДЫ И СИСТЕМЫ ЗАЩИТЫ ИНФОРМАЦИИ, ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ

УДК 004.056 DOI: 10.17212/2782-2230-2022-3-62-80

ИССЛЕДОВАНИЕ ПОДХОДОВ К СИНТЕЗУ И ДЕТЕКТИРОВАНИЮ КЛОНИРОВАННЫХ ГОЛОСОВ (DEEPFAKE)*

А.В. ИВАНОВ[1], С.А. ПРИМАК[2], В.А. МАЗУРЕНКО[3]

[1] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, кандидат технических наук, заведующий кафедрой защиты информации. E-mail: andrej.ivanov@corp.nstu.ru*
[2] *630087, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, аспирант кафедры защиты информации. E-mail: sprimak@rit-it.com*
[3] *630087, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, магистрант кафедры защиты информации. E-mail: mazurenko.2017@stud.nstu.ru*

Современные методы защиты персональных данных часто предусматривают использование биометрических данных о голосе владельца данных для идентификации пользователя. Озвучивая кодовую фразу, владелец подтверждает свою личность. Однако злоумышленники пользуются несовершенством подобных систем и разрабатывают способы клонирования и подмены голоса, целью которых является создание двойника голоса для кибератаки на системы защиты персональных данных.

В рамках настоящей статьи предпринимаются попытки исследовать существующие методы детекции клонированных голосов в целях защиты информации и противодействия кибератакам. Также для достижения результатов системы детекции будут испытаны на выборке из русскоязычных голосовых записей, взятых в открытых источниках. Проводится сравнительная оценка существующих подходов с точки зрения их практической применимости. Учитывались требования к занимаемой памяти вычислительного устройства, вычислительной сложности, сложности в реализации и сборе данных для обучения.

Помимо этого, проведен анализ существующих предпосылок и тенденций к использованию систем синтеза и подмены голосов, описаны потенциальные риски и приведены примеры возможного ущерба при краже биометрических данных.

Также выполнена попытка описать процедуру эксперимента для оценки эффективности работы рассмотренных методов с заданием конкретизирующих и уточняющих условий. Заданы критерии верификации и валидации результатов, которые позволяют делать выводы об эффективности работы систем.

* Статья получена 08 августа 2022 г.

Ключевые слова: защита персональных данных, идентификация пользователя, способы клонирования и подмены голоса, детекция клонированных голосов

ВВЕДЕНИЕ

За последние десять лет произошел огромный прогресс в области машинного обучения с внедрением сложных алгоритмов, которые могут манипулировать мультимедийным контентом и создавать на его основе материал, не существующий в реальном мире. Это может привести к целому ряду проблем, связанных с использованием этих механизмов в преступных целях. Учитывая легкость создания и распространения ложной информации, можно использовать механизмы машинного обучения для дискредитации публичных личностей, манипулирования общественным мнением и т. п. Становится всё труднее отличать правду, что может привести к кризису доверия современным сетям распространения информации.

Доступность экономичных цифровых интеллектуальных устройств, таких как мобильные телефоны, планшеты, ноутбуки и цифровые камеры, привела к экспоненциальному росту мультимедийного контента в киберпространстве. Благодаря этому каждый человек может получить доступ к огромному объему мультимедийных данных. При этом для большого количества людей можно найти образцы видео, фото, звука, содержащие их изображение или речь.

Не стоит также забывать о системах биометрической идентификации и аутентификации. Системы клонирования голоса позволяют осуществлять атаки на системы, которые используют голос в своих алгоритмах. Идентификацию на основе голоса используют многие банки РФ [1].

1. ПОСТАНОВКА ЗАДАЧИ

Одной из исследовательских проблем является отсутствие единой системы верифицирования результатов. Широко используемым подходом является тестирование на базе AVSpoof [2]. Эта база записей содержит аудиофайлы, произведенные различными методами спуфинга. Однако, во-первых, с выходом новых работ по спуфингу речи база данных становится менее релевантна и, во-вторых, записи речи в базе AVSpoof произведены только на английском языке, из-за чего результаты тестирования детекторов могут отличаться от результатов при тестировании на данных других языков. Особенно важно это для языков с сильно отличающимся фонемным набором (например, для китайского).

В этой работе стоит задача сравнительного анализа рассмотренных методов в рамках их применимости для синтеза и детекции клонированных голосов русскоязычных людей. С этой целью из открытых источников был собран набор записей с русской речью. В качестве источников использовались радиопередачи, аудиокниги, телепрограммы, видеоролики, а также обезличенные записи разговоров из открытых данных, находящихся в открытом доступе. Чтобы на каждой аудиозаписи преобладал голос одного человека, аудиозаписи были разбиты на куски меньшей длины. На рис. 1 и 2 визуализированы характеристики собранных записей. Соотношение выборки по полу говорящего важно для валидации результатов, поскольку половая принадлежность оказывает сильное влияние на голосовые характеристики.

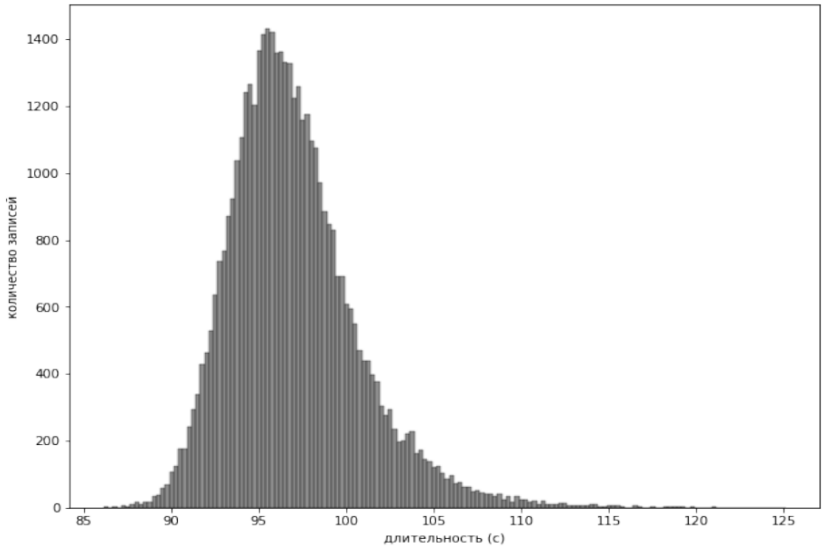


Рис. 1. Отношение длительности записи разговора к их количеству в выборке

Fig. 1. The ratio of the duration of a call recording to their number in the sample

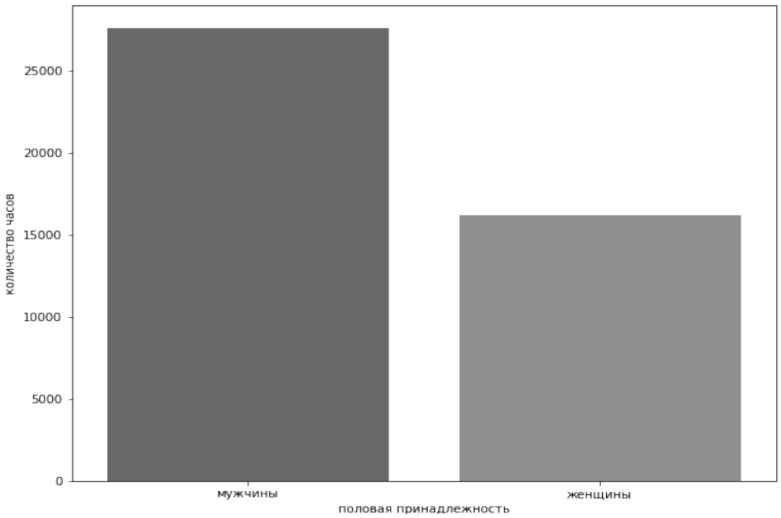


Рис. 2. Половая принадлежность голосов в записях разговоров, представленных в выборке

Fig. 2. Gender of voices in recordings of conversations presented in the sample

2. СРАВНЕНИЕ МЕТОДОВ

В рассмотренных статьях авторы использовали различный размер обучающей выборки, поскольку разные алгоритмы обладают разной скоростью сходимости. Однако на практике сбор достаточного количества данных представляет собой отдельную задачу. Поскольку обучающая выборка должна отражать характеристики данных, с которыми алгоритм должен работать, сбор специфичных данных может оказаться сложной задачей. О требуемом количестве данных для разных алгоритмов будет сказано позднее. Опираясь на вышесказанное, для обучения и тестирования всех алгоритмов были использованы одинаковые разбиения. Для обучения было использовано 80 % данных, для тестирования – 20 % данных. Из обучающей выборки было выделено 15 % валидационной выборки. Валидационная выборка не участвует напрямую в процессе обновления параметров модели, однако она используется для процедуры ранней остановки, что косвенно оказывает влияние на параметры модели.

Для каждого детектора была поставлена серия экспериментов, в которых в качестве записей с клонированным голосом были использованы записи, по-

лученные с одного из синтезаторов. Результатом тестирования является матрица, сопоставляющая результаты работы систем клонирования голоса и систем детектирования (рис. 3).

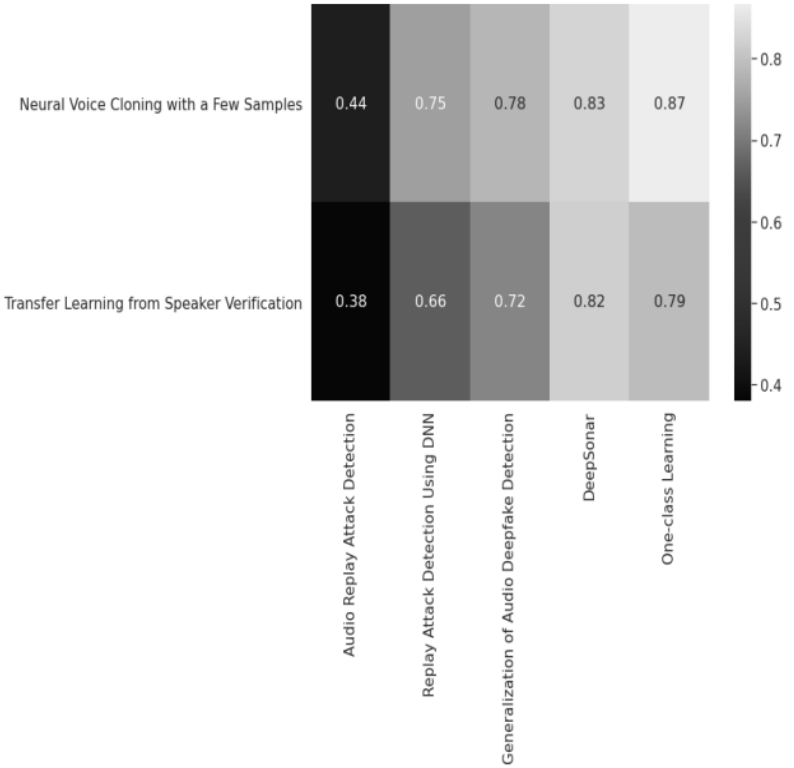


Рис. 3. Сравнение качества работы детекторов относительно современных средств клонирования голоса

Fig. 3. Comparison of the quality of the detectors versus modern voice cloning tools

Как видно из рис. 3, при анализе методов детекции лучшие результаты среди детекторов демонстрируют модели, направленные на анализ высокоуровневых признаков. Лучший результат показывают модели нейронных сетей с высокой обобщающей способностью и механизмами для борьбы с переобучением. Сравним качество работы на данных из обучающей выборки и из тестовой (результаты работы систем клонирования голоса были перемешаны) (табл. 1).

Таблица 1

Table 1

Сравнение работы нейросетей

Comparison of the work of neural networks

Модель	F1 – мера на обучающих данных	F2 – мера на тестовых данных
GMM	0,92	0,41
DNN	0,96	0,71
Полносверточная сеть	0,89	0,75
Система DeepSonar	0,9	0,82
Сеть, обученная одноклассовым методом	0,91	0,83

Как видно из табл. 1, модели склонны переобучаться на обучающей выборке, показывая худший результат на тестовых данных. Наиболее сильно этот эффект выражен у модели на основе гауссовых смесей. Нейронные сети лучше обобщаются, однако если в обучении не использовались механизмы борьбы с переобучением (например, ранняя остановка, аугментация данных, одноклассовый метод обучения), то этот эффект также будет проявляться и в сетях с мощной обобщающей способностью.

3. ИССЛЕДОВАНИЕ СОВРЕМЕННЫХ МЕТОДОВ ДЕТЕКЦИИ КЛОНИРОВАННЫХ ГОЛОСОВ

Использование высокочастотных полос для обучения модели детекции

Метод рассматривается в статье «Audio Replay Attack Detection Using High-Frequency Features» [3].

Сильная фильтрация нижних частот на частоте среза около 7,25 кГц и рассеяние временного спектра, вызванное сглаживающим эффектом работы сверточных фильтров, видны на нижней спектрограмме записи.

Был обучен классический классификатор без применения технологии нейронных сетей, представляющий модель гауссовых смесей (GMM), который обучается на кепстральных и CQCC-характеристиках, извлеченных для нескольких частотных диапазонов. Авторы рассмотрели полосы частот с нижним диапазоном частот от 1 до 7,5 кГц, в то время как самая высокая частота оставалась постоянной на частоте Найквиста, т. е. на уровне 8 кГц. Результаты EER, полученные с использованием набора данных разработки

ASVspoof [2], изображенного на рисунке, показывают, что установка нижней границы частоты в диапазоне от 4 до 6 кГц приводит к наименьшему уровню ошибки.

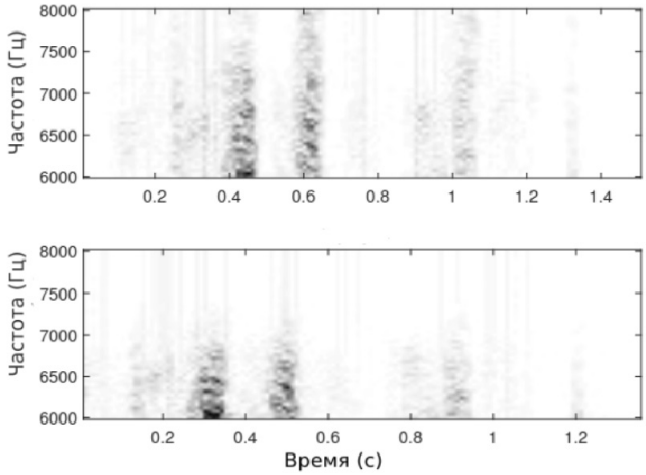


Рис. 4. Спектр записанной речи (вверху) и спектр искусственно синтезированной речи (внизу)

Fig. 4. The spectrum of recorded speech (above) and the spectrum of artificially synthesized speech (below)

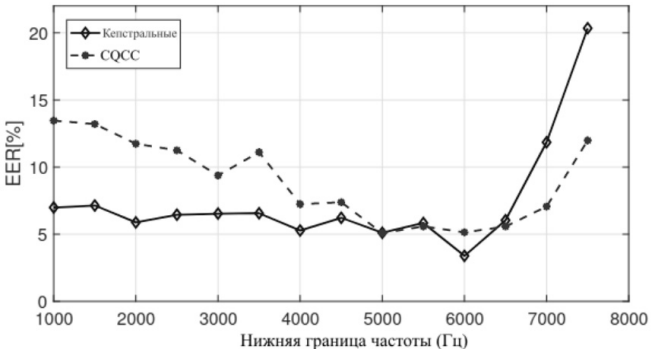


Рис. 5. Уровень ошибок в зависимости от выбранной нижней границы частоты

Fig. 5. Error level depending on the selected lower frequency limit

Такой детектор представляет собой классификатор на основе гауссовых примесей. Модель обладает значительно меньшим количеством параметров, поэтому требования к объему памяти значительно ниже, чем у моделей на основе нейронных сетей.

Однако для достижения высокой точности такие модели требуют моделирования большого количества гауссовых распределений, что является задачей, в меньшей степени поддающейся распараллеливанию, чем применение сверточных фильтров или вычисление прямого прохода полносвязной или рекуррентной сети. В связи с этим вычислительная сложность модели на основе гауссовых примесей выше, чем у нейронных сетей. GMM плохо обобщается на новые данные и не содержит механизмов для интерпретации работы метода.

Многоклассовый классификатор на основе DNN

Метод представлен в статье «Replay Attack Detection Using DNN for Channel Discrimination» [4]. Многие меры противодействия спуфингу голоса оказались успешными при использовании двух моделей GMM (по одной для подлинных и поддельных классов), каждая из них используется для получения оценок вероятности класса. Модель гауссовых смесей GMM предполагает, что распределение данных можно представить как конечную сумму гауссовых распределений. Количество параметров модели растет экспоненциально относительно качества моделирования.

Работу нейронной сети можно сформулировать как задачу бинарной классификации, т. е. различения подлинной речи от записанной или синтезированной.

Таблица 2

Table 2

Сравнение ERR для моделей бинарной и мультиклассовой классификации

ERR comparison for binary and multiclass classification models

Модель	ERR на обучающих данных	ERR на тестовых данных
Модель бинарной классификации	3,2 %	18,1 %
Модель мультиклассовой классификации	7,6 %	11,5 %

Как видно из сравнения ERR, модель, обученная на задаче мультиклассовой классификации, обладает лучшей обобщающей способностью. Однако это требует сбора дополнительных данных о методах атаки.

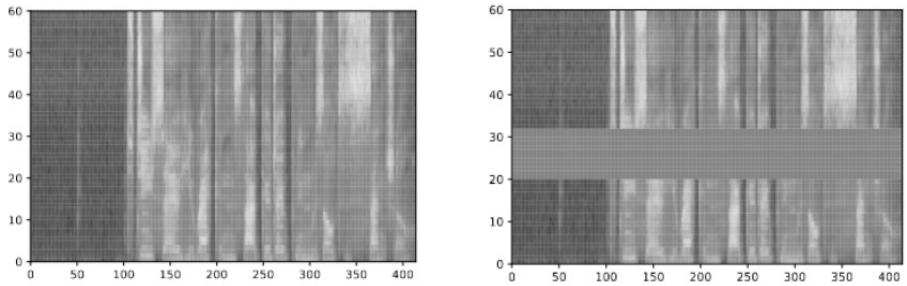
Архитектура сети основана на сверточных ядрах, поэтому количество параметров сети составляет около 5 млн. Требования к используемой памяти составляют 20…40 Мб.

Для обучения модель требует большого количества данных. В процессе обучения отсутствуют механизмы аугментации, которые могли бы искусственно увеличить набор данных. Кроме того, представленный метод обучения требует дополнительных меток класса. Данный классификатор не содержит механизма для интерпретации результатов своей работы.

Полносверточные нейронные сети для детекции подмены

Метод представлен в статье «Generalization of Audio Deepfake Detection» [5]. Наибольшей проблемой системы обнаружения спуфинга на данный момент является ее способность к обобщению. Традиционно исследователи обработки сигналов пытались решить эту проблему, создав различные низкоуровневые спектрально-временные особенности (например, кепстральные коэффициенты с постоянной добротностью (CQCC) [6], косинусная нормализованная фаза и модифицированная групповая задержка (MGD) [7]). Хотя эти работы подтвердили эффективность различных методов обработки звука при обнаружении синтетической речи, они примечательны тем, что сузили пробел в обобщении данных с помощью последних улучшенных технологий преобразования текста в речь. Однако последние результаты показывают, что ни одна из этих акустических характеристик не может быть однозначно обобщена на данные, полученные с помощью неизвестных моделей техник синтезирования голоса. В настоящей работе авторы решают эту проблему с другой точки зрения. Вместо того чтобы исследовать различные особенности звука низкого уровня, исследователи пытаются повысить обобщающую способность самой модели. Для этого мы используем функцию потерь косинуса с большим запасом (LMCL), которая изначально использовалась для распознавания лиц. Цель LMCL состоит в том, чтобы максимизировать разницу между подлинным и синтезированным классом и в то же время минимизировать внутриклассовую дисперсию. Кроме того, авторы используют алгоритм SpecAugment [8] путем добавления в сеть слоя FreqAugment, который случайным образом маскирует соседние частотные каналы во время обучения нейронной сети, чтобы еще больше повысить способность модели к обобщению.

В качестве входных значений для работы нейронной сети авторы используют линейные блоки фильтров (LFB). Применение этих признаков снижает риск переобучения сети во время обучения. Подобные кепстральные особенности основаны на банке фильтров, таких как линейные частотные кепстральные коэффициенты (LFCC) [9].



а *б*

Рис. 6. LFB аудиозаписи (*а*), результат работы механизма FreqAugment (*б*)

Fig. 6. LFB audio recordings (*a*), the result of the FreqAugment mechanism (*b*)

В силу большого количества параметров используемой сети требования к объему памяти устройства составляют около 130 Мб. Это делает затруднительным использование такого детектора на конечных устройствах с малым объемом памяти запоминающего устройства.

Благодаря использованию механизма аугментации аудиозаписей FreqAugment нейросеть может быть обучена на меньшем количестве данных, компенсируя недостаток данных с помощью внесения изменений в уже существующие. Такой классификатор не содержит механизма для интерпретации результатов своей работы, кроме тех, которые свойственны для всех нейронных сетей.

Система DeepSonar

Работа представлена в статье «DeepSonar: Towards Effective and Robust detection of AI-synthesized fake voices» [10]. Предлагается [11] использовать сигнатуры акустической среды в качестве важной функции для обнаружения подделки звука путем проверки целостности цифрового сигнала. Одной из наиболее значимых работ является статья о применении биспектральных характеристик для анализа аудиозаписей [12].

В этой статье вместо исследования артефактов в необработанных голосах, представленных в синтезе, мы исследуем новый способ, отслеживая поведение нейронов SR-систем на основе DNN с помощью простого бинарного классификатора, чтобы различать настоящие и поддельные голоса. Послойное поведение нейронов может улавливать более тонкие особенности различения реальных и фальшивых голосов.

Мониторинг поведения нейронов – важный метод для поиска различий между набором входных данных для DNN и исследования их внутреннего

поведения, обеспечения качества [13–16], безопасности интерпретации DNN [17, 18] и т. д.

С вычислительной точки зрения система DeepSonar представляет одну из наименее требовательных моделей. Используемый для классификации детектор имеет малое количество параметров. Однако для извлечения активаций используется более сложная нейронная сеть. DeepSonar позволяет выбрать компромисс между вычислительной сложностью и точностью путем замены сети извлекателя признаков.

Требования к занимаемой памяти системы DeepSonar также зависят от выбранного извлекателя признаков. Поскольку извлекатель признаков должен обладать достаточной выразительностью, сеть должна содержать достаточное количество признаков: 1…30 млн параметров. Поэтому занимаемое системой место составляет около 50…160 Мб.

Количество требуемых данных для обучения зависит от того, насколько сильно отличаются предполагаемые условия работы от условий, в которых была обучена сеть извлекателя признаков.

С точки зрения возможностей для интерпретации решений детектора DeepSonar предлагает визуализацию векторов представлений активаций нейронов. Поскольку анализ большинства извлекателей признаков анализирует звуковую последовательность, используя скользящее окно, возможным становится извлечение вектора активаций нейронов для каждого окна. Имея представления активаций нейронов для каждого куска записи, можно определить, какой из отрывков находится ближе к кластеру записей с подменой голоса, т. е. определить, какой отрывок является наиболее подозрительным.

Одноклассовое обучение обнаружению подмены голоса

Метод рассматривается в статье «One-class Learning Towards Synthetic Voice Spoofing Detection» [16]. Авторы статьи предлагают свою архитектуру модели для обнаружения спуфинг атак на системы голосовой аутентификации.

Для борьбы с проблемой детекции атак, неизвестных на стадии обучения, авторы статьи предлагают использовать модификацию широко используемой бинарной функции потерь «софтмакс». Функция «софтмакс» определяется следующим образом:

$$L_s = -\mathbb{M}\log\frac{e^{w_{y_i}^T x_i}}{e^{w_{y_i}^T x_i} + e^{w_{1-y_i}^T x_i}} = -\mathbb{M}\log(1 + e^{((w_{1-y_i} - w_{y_i})x_i)}), \quad (1)$$

где $x \in RD$ и $y \in \{0, 1\}$ являются вектором представлений и меткой класса, w обозначает параметры линейного классификатора.

Модификация этой функции ошибки позволяет улучшить уровень обучения сети путем добавления в функцию углового отступа. Функция «софтмакс с угловым отступом» формулируется так:

$$L_s = -\mathbb{M}\log\frac{e^{\alpha(\widehat{w}_{y_i}^T - m)\widehat{x}_i}}{e^{\alpha(\widehat{w}_{y_i}^T - m)\widehat{x}_i} + e^{\alpha\widehat{w}_{1-y_i}^T\widehat{x}_i}} = -\mathbb{M}\log(1 + e^{\alpha(m-(\widehat{w}_{1-y_i} - \widehat{w}_{y_i}))\widehat{x}_i}), \qquad (2)$$

где α является гиперпараметром, определяющим степень влияния фактора углового отступа; m – значение границы, начиная с которого пример считается однозначно классифицирован.

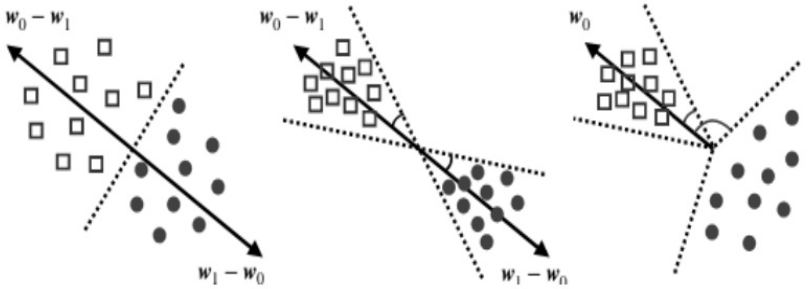


Рис. 7. Иллюстрация кластеров и границ принятия решений для моделей с разными функциями потерь. Слева направо: софтмакс, софтмакс с угловым отступом, софтмакс для сжатия одного класса

Fig. 7. Illustration of clusters and decision boundaries for models with different loss functions. Left to right: softmax, softmax with corner padding, softmax for single-class compression

Как видно на рис. 7, при обучении с помощью функции «софтмакс» и «софтмакс с угловым отступом» векторы представлений имеют тенденцию группироваться вдоль двух противоположных направлений.

Классификатор, обученный с применением софтмакса для сжатия одного класса, представляет собой классическую глубокую нейронную сеть.

Количество занимаемой памяти вычислительного устройства также сопоставимо с другими нейросетевыми подходами, такими как DeepSonar. Однако, в отличие от системы DeepSonar, данный детектор требует существенного объема тренировочных данных и не предлагает широкого выбора моделей для извлечения признаков.

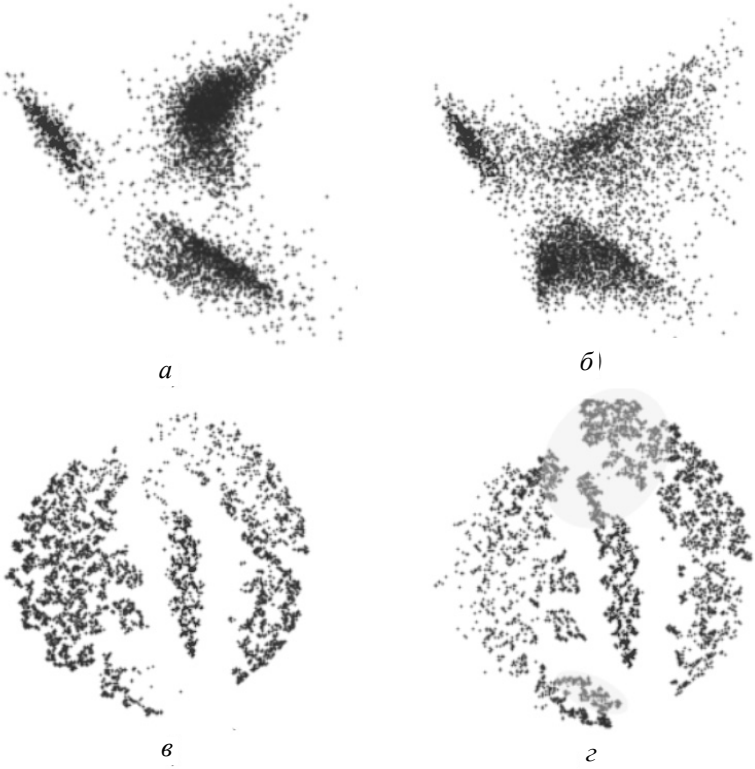


Рис. 8. Визуализация векторов представлений, полученных с помощью методов линейного понижения размерности анализом главных компонент (PCA) и tSNE:

а – PCA-представления, полученные на обучающих данных; *б* – PCA-представления, полученные на тестовом наборе данных; *в* – tSNE-представления, полученные на обучающем наборе данных; *г* – tSNE-представления, полученные на тестовом наборе данных

Fig. 8. Visualization of vectors of representations obtained using the methods of linear dimensionality reduction by principal component analysis (PCA) and tSNE:

a – PCA representations obtained on training data; *b* – PCA representations obtained on a test dataset; *c* – tSNE representations obtained on a training set data; *d* – tSNE representations obtained on the test dataset

Как показано на рис. 8, векторы представления, полученные с предпоследнего слоя модели, являются спутанными. Хотя такое распределение позволяет разделить большинство точек на классы, с практической точки зрения эта визуализация не позволяет достаточно подробно интерпретировать работу модели.

ЗАКЛЮЧЕНИЕ

В настоящей статье были рассмотрены последние разработки в области клонирования и детектирования клонированных голосов. Был проведен сравнительный анализ с учетом требований к реализации алгоритмов и эффективности работы методов.

Лучшие результаты показывают методы, основанные на глубоких нейронных сетях. Полносверточные сети показывают лучший результат в определении клонированных голосов. Тенденция развития архитектур нейронных сетей заключается в применении более эффективных комбинаций сверток.

Наиболее частым начальным преобразованием звука является вычисление мел-кепстральных характеристик (MFCC). Несмотря на то что подбор входных преобразований может улучшить показатели точности слабых детекторов, построение модели с более высокой обобщающей способностью дает бóльший прирост точности.

Разные системы синтеза голоса различно влияют на детекторы. Модель синтеза, основанная на переносе обучения, генерирует данные, на которых детекторы ошибаются чаще. При этом система DeepSonar показывает близкое значение метрики для всех используемых систем синтеза. Это показывает, что анализ признаков, основанный на отношениях высокого порядка, является одним из наиболее значимых методов.

Существующие детекторы синтезированных голосов в основном полагаются на фиксированные особенности существующих кибератак с использованием методов машинного обучения, включая неконтролируемую кластеризацию и контролируемые методы классификации и поэтому показывают скромный результат для неизвестных методов подмены. Методы, содержащие механизм для борьбы с этим, показывают лучший результат при переходе от обучающих данных к тестовым.

СПИСОК ЛИТЕРАТУРЫ

1. *Каледина А.* ВТБ24 первым запустит голосовую идентификацию // Известия. – 2016. – 28 октября. – URL: https://iz.ru/news/641241 (дата обращения: 29.08.2022).

2. ASVspoof 2019: a large-scale publicdatabase of synthetized, converted and replayed speech / X. Wang, et. al. // Computer Speech and Language. – 2020. – Vol. 64. – P. 101114. – DOI: 10.1016/j.csl.2020.101114.

3. Audio replay attack detection using high-frequency features / M. Witkowski, S. Kacprzak, P. Żelasko, K. Kowalczyk, J. Gałka // Proceedings Interspeech 2017. – Stockholm, Sweden, 2017. – P. 27–31. – DOI: 10.21437/Interspeech.2017-776.

4. Replay attack detection using DNN for channel discrimination / P. Nagarsheth, E. Khoury, K. Patil, M. Garland // Proceedings Interspeech 2017. – Stockholm, Sweden, 2017. – P. 97–101. – DOI: 10.21437/Interspeech.2017-1377.

5. Generalization of Audio Deepfake detection / T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, E. Khoury // Proceedings The Speaker and Language Recognition Workshop (Odyssey 2020). – Tokyo, Japan, 2020. – P. 132–137. – DOI: 10.21437/Odyssey.2020-19.

6. *Todisco M., Delgado H., Evans N.* A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients // The Speaker and Language Recognition Workshop (Odyssey 2016). – Bilbao, Spain, 2016. – P. 283–290.

7. *Wu Z., Chng E.S., Li H.* Detecting converted speech and natural speech for antispoofing attack in speaker recognition // Proceedings Interspeech 2012. – Portland, OR, USA, 2012. – P. 1700–1703. – DOI: 10.21437/Interspeech.2012-465.

8. SpecAugment: a simple data augmentation method for automatic speech recognition / D.S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E.D. Cubuk, Q.V. Le // Proceedings Interspeech 2019. – Graz, Austria, 2019. – P. 2613–2617. – DOI: 10.21437/interspeech.2019-2680.

9. *Sahidullah M., Kinnunen T., Hanilçi C.* A comparison of features for synthetic speech detection // Proceedings Interspeech 2015. – Dresden, Germany, 2015. – P. 2087–2091. – DOI: 10.21437/Interspeech.2015-472.

10. DeepSonar: towards effective and robust detection of ai-synthesized fake voices / R. Wang, F. Juefei-Xu, Y. Huang, Q. Guo, X. Xie, L. Ma, Y. Liu // MM '20: The 28th ACM International Conference on Multimedia. – ACM, 2020. – P. 1207–1216. – DOI: 10.1145/3394171.3413716.

11. *Zhao H., Malik H.* Audio recording location identification using acoustic environment signature // IEEE Transactions on Information Forensics and Security. – 2013. – Vol. 8 (11). – P. 1746–1759. – DOI: 10.1109/TIFS.2013.2278843.

12. *AlBadawy E.A., Lyu S., Farid H.* Detecting AI-synthesized speech using bispectral analysis // IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019. – IEEE, 2019. – P. 104–109.

13. DeepGauge: Multi-granularity testing criteria for deep learning systems / L. Ma, F. Juefei-Xu, F. Zhang, J. Sun, M. Xue, B. Li, C. Chen, T. Su, L. Li, Y. Liu, J. Zhao, Y. Wang // ASE 2018: Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering. – ACM, 2018. – P. 120–131. – DOI: 10.1145/3238147.3238202.

14. TensorFuzz: debugging neural networks with coverage-guided fuzzing / A. Odena, C. Olsson, D. Andersen, I. Goodfellow // Proceedings of Machine Learning Research. – 2019. – Vol. 97. – P. 4901–4911.

15. DeepXplore: automated whitebox testing of deep learning systems / K. Pei, Y. Cao, J. Yang, S. Jana // Proceedings of the 26th Symposium on Operating Systems Principles (SOSP '17). – ACM, 2017. – DOI: 10.1145/3132747.3132785.

16. DeepHunter: a coverage-guided fuzz testing framework for deep neural networks / X. Xie, L. Ma, F. Juefei-Xu, M. Xue, H. Chen, Y. Liu, J. Zhao, B. Li, J. Yin, S. See // Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2019). – ACM, 2019. – P. 146–157. – DOI: 10.1145/3293882.3330579.

17. NIC: detecting adversarial samples with neural network invariant checking / S. Ma, Y. Liu, G. Tao, W.-C. Lee, X. Zhang // Proceedings of the 26th network and distributed system security symposium (NDSS 2019). – The Internet Society, 2019. – DOI: 10.14722/ndss.2019.23415.

18. Attacks meet interpretability: Attribute-steered detection of adversarial samples / G. Tao, S. Ma, Y. Liu, X. Zhang // Advances in Neural Information Processing Systems. – 2018. – Vol. 31. – P. 7717–7228.

Иванов Андрей Валерьевич, кандидат технических наук, заведующий кафедрой защиты информации Новосибирского государственного технического университета. Область научных интересов – обработка звука от шума, техническая защита информации от утечки по каналам связи. E-mail: andrej.ivanov@corp.nstu.ru

Примак Степан Александрович, аспирант кафедры защиты информации Новосибирского государственного технического университета. Область научных интересов – обработка звука от шума, техническая защита информации от утечки по каналам связи. E-mail: sprimak@rit-it.com

Мазуренко Виктор Александрович, магистрант кафедры защиты информации Новосибирского государственного технического университета. Область научных интересов – обработка звука от шума, техническая защита информации от утечки по каналам связи. E-mail: mazurenko.2017@stud.nstu.ru

DOI: 10.17212/2782-2230-2022-3-62-80

Study of approaches to the synthesis and detection of cloned voices (DeepFake)*

A.V. Ivanov[1], S.A. Primak[2], V.A. Mazurenko[3]

[1] *Novosibirsk State Technical University, 20 Karl Marx Prospekt, 630073, Novosibirsk, Russian Federation, candidate of technical sciences, Head of the Information Security Department. E-mail: andrej.ivanov@corp.nstu.ru*
[2] *Novosibirsk State Technical University, 20 Karl Marx Prospekt, 630087, Novosibirsk, Russian Federation, postgraduate student of the Information Security Department. E-mail: sprimak@rit-it.com*
[3] *Novosibirsk State Technical University, 20 Karl Marx Prospekt, 630087, Novosibirsk, Russian Federation, master's student of the Information Security Department. E-mail: mazurenko.2017@stud.nstu.ru*

Modern methods of protecting personal information often uses the voice biometric data of the owner of the information to identify the user. When the owner of the information voices the passphrase, he confirms his identity. However, attackers take advantage of the imperfection of such systems and develop methods for voice cloning, to create a twinkly voice for a cyberattack on personal data protection systems.
Within the framework of this article, an attempt is made to explore existing methods for detecting cloned voices in order to protect information and counteract cyberattacks. Also, to achieve results, detection systems will be tested on a sample of Russian-language voice recordings taken from open sources. A comparative assessment of existing approaches is carried out in terms of their practical applicability. In particular, the requirements for the occupied memory of a computing device, computational complexity, complexity in implementation and data collection for training were taken into account.
In addition, an analysis of the existing prerequisites and trends for the use of voice synthesis and substitution systems was carried out, potential risks were described, and examples of possible damage from the theft of biometric data were given.
An attempt was also made to describe the experimental procedure for evaluating the performance of the considered methods with specifying and clarifying conditions. The criteria for verification and validation of the results are set, which allow drawing conclusions about the efficiency of the systems.

Keywords: deepfake, voice cloning, cloned voice detection, adversarial attacks, spoofing

REFERENCE

1. Kaledina A. VTB24 pervym zapustit golosovuyu identifikatsiyu [VTB24 will be the first to launch voice identification]. *Izvestiya*, 2016, 28 October. (In Russian). Available at: https://iz.ru/news/641241 (accessed 29.08.2022.).

* *Received 08 August 2022*.

2. Wang X., et al. ASVspoof 2019: a large-scale publicdatabase of synthetized, converted and replayed speech. *Computer Speech and Language*, 2020, vol. 64, p. 101114. DOI: 10.1016/j.csl.2020.101114.

3. Witkowski M., Kacprzak S., Żelasko P., Kowalczyk K., Gałka J. Audio replay attack detection using high-frequency features. *Proceedings Interspeech 2017*, Stockholm, Sweden, 2017, pp. 27–31. DOI: 10.21437/Interspeech.2017-776.

4. Nagarsheth P., Khoury E., Patil K., Garland M. Replay attack detection using DNN for channel discrimination. *Proceedings Interspeech 2017*, Stockholm, Sweden, 2017, pp. 97–101. DOI: 10.21437/Interspeech.2017-1377.

5. Chen T., Kumar A., Nagarsheth P., Sivaraman G., Khoury E. Generalization of Audio Deepfake detection. *Proceedings The Speaker and Language Recognition Workshop (Odyssey 2020)*, Tokyo, Japan, 2020, pp. 132–137. DOI: 10.21437/Odyssey.2020-19.

6. Todisco M., Delgado H., Evans N. A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients. *The Speaker and Language Recognition Workshop (Odyssey 2016)*, Bilbao, Spain, 2016, pp. 283–290.

7. Wu Z., Chng E.S., Li H. Detecting converted speech and natural speech for antispoofing attack in speaker recognition. *Proceedings Interspeech 2012*, Portland, OR, USA, 2012, pp. 1700–1703. DOI: 10.21437/Interspeech.2012-465.

8. Park D.S., Chan W., Zhang Y., Chiu C.-C., Zoph B., Cubuk E.D., Le Q.V. SpecAugment: a simple data augmentation method for automatic speech recognition. *Proceedings Interspeech 2019*, Graz, Austria, 2019, pp. 2613–2617. DOI: 10.21437/interspeech.2019-2680.

9. Sahidullah M., Kinnunen T., Hanilçi C. A comparison of features for synthetic speech detection. *Proceedings Interspeech 2015*, Dresden, Germany, 2015, pp. 2087–2091. DOI: 10.21437/Interspeech.2015-472.

10. Wang R., Juefei-Xu F., Huang Y., Guo Q., Xie X., Ma L., Liu Y. DeepSonar: towards effective and robust detection of ai-synthesized fake voices. *MM '20: The 28th ACM International Conference on Multimedia*. ACM, 2020, pp. 1207–1216. DOI: 10.1145/3394171.3413716.

11. Zhao H., Malik H. Audio recording location identification using acoustic environment signature. *IEEE Transactions on Information Forensics and Security*, 2013, Vol. 8 (11), PP. 1746–1759. DOI: 10.1109/TIFS.2013.2278843.

12. AlBadawy E.A., Lyu S., Farid H. Detecting AI-synthesized speech using bispectral analysis. *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019*. IEEE, 2019, pp. 104–109.

13. Ma L., Juefei-Xu F., Zhang F., Sun J., Xue M., Li B., Chen C., Su T., Li L., Liu Y., Zhao J., Wang Y. DeepGauge: Multi-granularity testing criteria for deep learning systems. *ASE 2018: Proceedings of the 33rd ACM/IEEE International*

Conference on Automated Software Engineering. ACM, 2018, pp. 120–131. DOI: 10.1145/3238147.3238202.

14. Odena A., Olsson C., Andersen D., Goodfellow I. TensorFuzz: debugging neural networks with coverage-guided fuzzing. *Proceedings of Machine Learning Research*, 2019, vol. 97, pp. 4901–4911.

15. Pei K., Cao Y., Yang J., Jana S. DeepXplore: automated whitebox testing of deep learning systems. *Proceedings of the 26th Symposium on Operating Systems Principles (SOSP '17)*. ACM, 2017. DOI: 10.1145/3132747.3132785.

16. Xie X., Ma L., Juefei-Xu F., Xue M., Chen H., Liu Y., Zhao J., Li B., Yin J., See S. DeepHunter: a coverage-guided fuzz testing framework for deep neural networks. *Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2019)*. ACM, 2019, pp. 146–157. DOI: 10.1145/3293882.3330579.

17. Ma S., Liu Y., Tao G., Lee W.-C., Zhang X. NIC: detecting adversarial samples with neural network invariant checking. *Proceedings of the 26th network and distributed system security symposium (NDSS 2019)*. The Internet Society, 2019. DOI: 10.14722/ndss.2019.23415.

18. Tao G., Ma S., Liu Y., Zhang X. Attacks meet interpretability: Attribute-steered detection of adversarial samples. *Advances in Neural Information Processing Systems*, 2018, vol. 31, pp. 7717–7228.

Для цитирования:

Иванов А.В., Примак С.А., Мазуренко В.А. Исследование подходов к синтезу и детектированию клонированных голосов (DeepFake) // Безопасность цифровых технологий. – 2022. – № 3 (106). – С. 62–80. – DOI: 10.17212/2782-2230-2022-3-62-80.

For citation:

Ivanov A.V., Primak S.A., Mazurenko V.A. Issledovanie podkhodov k sintezu i detektirovaniyu klonirovannykh golosov [Study of approaches to the synthesis and detection of cloned voices (DeepFake)]. *Bezopasnost' tsifrovykh tekhnologii = Digital Technology Security*, 2022, no. 3 (106), pp. 62–80. DOI: 10.17212/2782-2230-2022-3-62-80.