Формальная модель преобразования речевого сигнала при реализации телефонной связи по протоколу VoLTE

У. В. Токарева

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

tokareva.ulv@gmail.com

А. Д. Шульженко

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

adshulzhenko@etu.ru

Аннотация. **В последние пять лет отмечается тенденция роста числа целевых атак с использованием AI-преобразования голоса [1, 2]. Соответственно, для обеспечения личной безопасности граждан и минимизации вероятности успешного воздействия на граждан – сотрудников организаций, в том числе относящихся к объектам критической информационной инфраструктуры, разрабатываются методы детекции таких преобразований [3,4]. На точность работы таких методов влияет сохранность детектируемых артефактов синтеза на этапе транспортировки сигнала. В статье представлена формальная модель преобразования речевого сигнала на этапе подготовке к передаче по протоколу VoLTE [5]. Модель описывает математические операции, применяемые к уже сформированному, в том числе с использованием технологий искусственного интеллекта, сигналу в рамках стандартизированного транспортного конвейера. Рассмотрены три последовательных этапа: дискретизация и квантование в соответствии с теоремой Котельникова; компрессия кодеком AMR-WB с математическим описанием линейного предсказания (LPC) и алгебраического возбуждения (ACELP) [6] и, наконец, инкапсуляция кадров в пакеты RTP согласно RFC 4867 [7] с учетом временной маркировки и нумерации. Формализовано влияние параметров качество обслуживания (QCI=1, гарантированная полоса пропускания) архитектуры IMS на целостность пакетов. Показано, что сжатие с потерями и сетевая транспортировка вносят вторичные искажения, способные маскировать первичные артефакты AI-преобразования голосовых сигналов. Разработанная модель создает теоретическую основу для проектирования детекторов, устойчивых к воздействию транспортного протокола, и может быть использована операторами связи для обеспечения безопасности пользователей.**

Ключевые слова: VoLTE; AMR-WB; RTP-пакетизация; формальная модель; сжатие речи; вишинг; детекция дипфейков; IMS; качество обслуживания (QoS); цифровая обработка сигналов

I. Введение

Актуальность исследования обусловлена стремительным развитием технологий искусственного интеллекта и двойственностью применения этих технологий. С одной стороны, системы преобразования голоса на базе глубокого обучения открывают перспективы для создания персонализированных голосовых ассистентов и иммерсивных интерфейсов. С другой стороны, доступность данных технологий приводит к росту числа преступлений, связанных с голосовым фишингом (вишингом). Согласно отчёту компании BI.ZONE Brand Protection, количество фишинговых атак в Российской Федерации увеличилось с 111 тыс. в 2022 г. до 350 тыс. в 2024 г., причём значительная часть инцидентов связана с использованием синтезированных голосовых сообщений [1].

Первый крупный случай мошенничества с применением нейросетевого клонирования голоса был зафиксирован в 2019 году, когда сотрудник энергетической компании перевёл 243 тыс. долларов США на счёт злоумышленника, представившегося генеральным директором [1]. В 2024 году в России были выявлены случаи, когда мошенники выдавали себя за высокопоставленных государственных служащих, используя технологии генерации голоса на базе искусственного интеллекта [4]. Подобные атаки сочетают в себе эффективную комбинацию социальной инженерии и технических инноваций, угрожая конфиденциальности данных и подрывая доверие пользователей к цифровым системам.

В ответ на растущую угрозу разрабатываются методы детекции искусственно преобразованных речевых сигналов (дипфейков). Современные исследования подтверждают эффективность подходов, основанных на анализе акустических и статистических признаков: мел-частотных кепстральных коэффициентов (MFCC), кепстральных коэффициентов на основе постоянного Q (CQCC), статистических свойств остатка линейного предсказания (LPC-residual), параметров контура основной частоты (джиттер, шиммер) [1, 4, 5]. Однако точность работы данных методов существенно зависит от формата представления сигнала. Сжатие с потерями, применяемое в сетях мобильной связи, и особенности формирования пакетов для передачи могут вносить вторичные искажения, способные «размывать» артефакты синтеза [2, 12].

При передаче голосового сигнала по протоколу VoLTE аудиосигнал подвергается многоступенчатой обработке. Отсутствие формальной математической модели, описывающей преобразование речевого сигнала на каждом этапе обработки, затрудняет оценку устойчивости методов детекции артефактов применения средств искусственного интеллекта в реальных условиях эксплуатации сетей мобильной связи. Большинство существующих решений тестируются на «чистых»

аудиофайлах без учёта влияния алгоритмов сжатия и применения транспортного стека операторов связи [2].

Задачей настоящей работы является разработка формальной модели преобразования речевого сигнала при подготовке к передаче по протоколу VoLTE.

II. Обзор существующих решений

A. *Существующие решения*

Современные исследования в области детекции голосовых дипфейков можно классифицировать по типу используемых признаков и применяемым алгоритмам классификации.

Наиболее распространённым подходом является анализ спектральных характеристик сигнала. В работе [1] показано, что кепстральные коэффициенты на основе постоянного преобразования (CQCC) обладают высокой чувствительностью к артефактам, вносимым нейро-вокодерами систем преобразования голоса. Авторами показана достижимая точность классификации ~93,7% (AUC=0,937, EER=0,139) на наборе данных ASVspoof 2019 при использовании гауссовых смесей (GMM) в качестве классификатора. Мел-частотные кепстральные коэффициенты (MFCC), учитывающие нелинейное восприятие звука человеческим ухом, также широко применяются в качестве базового набора признаков [1].

Относительно новым направлением является использование статистических свойств остатка линейного предсказания (LPC-residual). Также в работе [1] экспериментально установлено, что для сигналов, преобразованных с использованием нейросетевого вокодера, коэффициент остроты распределения амплитуд остатка составляет $k_{AI} \approx 5{,}72 \pm 0{,}41$, тогда как для естественной речи $k_{nat} \approx 3{,}24 \pm 0{,}33$. Данное различие обусловлено фазовыми искажениями, вносимыми на этапе восстановления формы волны. Авторами показано достижимое значение точности классификации ~93,9% (AUC=0,989, EER=0,135) на наборе данных AVSPOOF2019 также при использовании гауссовых смесей (GMM) в качестве классификатора.

B. *Недостатки существующих решений*

Несмотря на высокую эффективность в лабораторных условиях, существующие методы детекции имеют ряд ограничений при применении в реальных сетях мобильной связи.

Во-первых, большинство алгоритмов детекции предполагают анализ сигнала, оцифрованного с частотой дискретизации 16 кГц и выше и представленного в несжатом формате. В сетях мобильной связи речевой сигнал подвергается сжатию с потерями (кодеки AMR, AMR-WB, EVS), что может нивелировать или исказить детектируемые артефакты [2].

Во-вторых, существующие исследования анализируют статические аудиофайлы, не учитывая влияние передачи по протоколам реального времени (RTP/RTCP).

Таким образом, поставленная задача разработки формальной модели, описывающей преобразование речевого сигнала в стандартизированном транспортном конвейере сетей мобильной связи, является актуальной для оценки устойчивости методов детекции к артефактам алгоритмов регистрации и передачи сигнала по протоколу VoLTE.

III. Предлагаемое решение

Формальная модель преобразования речевого сигнала состоит из трех последовательных этапов подготовке к передаче по протоколу VoLTE: дискретизация и квантование, компрессия кодеком и инкапсуляция в пакеты.

A. *Дискретизация и квантование*

Пусть $x(t)$ – непрерывный аналоговый речевой сигнал. В соответствии с теоремой Котельникова, для точного восстановления сигнала необходимо, чтобы частота дискретизации f_s удовлетворяла условию:

$$f_s \geq 2 \cdot f_{max}$$

где f_{max} – максимальная частота в спектре сигнала. Для широкополосной речи в VoLTE применяется частота дискретизации 16 кГц, что обеспечивает полосу пропускания до 8 кГц [6].

Дискретизированный сигнал представляется последовательностью отсчётов

$$x[n] = x(n \cdot T_s), T_s = \frac{1}{f_s}$$

Квантование отображает непрерывные значения отсчётов в дискретный набор уровней. При равномерном квантовании с шагом Δ:

$$x_q[n] = Q(x[n]) = \Delta \cdot round\left(\frac{x[n]}{\Delta}\right) \quad (1)$$

Ошибка квантования определяется как:

$$q[n] = x_q[n] - x[n], |q[n]| \leq \frac{\Delta}{2} \quad (2)$$

Мощность шума квантования для равномерного распределения ошибки традиционно оценивается как [5]:

$$\sigma_q^2 = \frac{\Delta^2}{12} \quad (3)$$

Данное выражение справедливо при условии равномерной плотности вероятности сигнала или при условии, что сигнал значительно превышает шаг квантования. Для сигналов, преобразованных с использованием искусственного интеллекта, данное условие нарушается вследствие специфических статистических артефактов [1].

B. *Компрессия кодеком*

Кодек AMR-WB использует модель линейного предсказания для представления спектральной огибающей сигнала. Согласно модели, каждый отсчёт сигнала аппроксимируется линейной комбинацией p предыдущих отсчётов:

$$\hat{x}[n] = \sum_{k=1}^{p} a_k \cdot x[n-k] \quad (4)$$

где p – порядок предсказателя, a_k – коэффициенты линейного предсказания.

Вектор оптимальных коэффициентов $a = [a_1, a_2, \dots, a_p]^T$, находится из решения уравнения Юла-Уокера [7]:

$$R \cdot a = r$$

где R – автокорреляционная матрица сигнала размером $p \times p$ с элементами $R_{ij} = E[x[n-i] \cdot x[n-j]]$, а r – вектор автокорреляции $r_k = E[x[n] \cdot x[n-k]]$.

Разностный сигнал, подлежащий дальнейшему кодированию, определяется с учетом (4) как ошибка предсказания:

$$d[n] = x[n] - \hat{x}[n] = x[n] - \sum_{k=1}^{p} a_k \cdot x[n-k]$$

В теории линейного предсказания $d[n]$ интерпретируется как возбуждающий сигнал (LPC – остаток) $\varepsilon[n]$. Для естественной речи распределение амплитуд остатка близко к гауссовскому с коэффициентом остроты $k \approx 3$. Для сигналов, преобразованных с использованием искусственного интеллекта, наблюдается значительный рост остроты распределения [1]:

$$k_{AI} = \frac{E[\varepsilon^4[n]]}{\sigma_\varepsilon^4} \approx 5.72 \pm 0.41, \quad k_{nat} \approx 3.24 \pm 0.33 \qquad (5)$$

На этапе алгебраического кодирования возбуждения остаток $d[n]$ квантуется с использованием алгебраической кодовой книги. Оператор квантования $Q(\cdot)$ отображает непрерывный сигнал разности в дискретный набор уровней $d_q[n]$. При равномерном квантовании с шагом Δ, учитывая правило (1):

$$d_q[n] = Q(d[n]) = \Delta \cdot round\left(\frac{d[n]}{\Delta}\right)$$

Ошибка квантования, с учетом правила (2), равняется $q[n] = d_q[n] - d[n]$ и имеет равномерное распределение в интервале $[-\frac{\Delta}{2}, +\frac{\Delta}{2}]$ только для сигналов с равномерной плотностью вероятности.

Для сигналов с эксцессом $k \neq 3$ при использовании компандера традиционная оценка мощности шума квантования (3) требует коррекции. Наличие «тяжёлых хвостов» распределения у AI-сигнала $(k > 3)$ приводит к увеличению вероятности попадания амплитуды сигнала на границы уровней квантования, что увеличивает дисперсию ошибки квантования.

Пусть мощность шума квантования $\sigma_{q,AI}^2$ для сигнала с эксцессом $k/3$ пропорциональна отклонению остроты от нормального распределения. Плотность вероятности ошибки предсказания $p(d)$ для AI-сигнала аппроксимируется обобщенным нормальным распределением, параметризованным остротой k.

Дисперсия ошибки квантования для нелинейного компандера зависит от производной функции компандирования $c'(d)$ и плотности вероятности $p(d)$ [5]:

$$\sigma_q^2 \approx \frac{\Delta^2}{12} \int_{-\infty}^{\infty} \frac{p(d)}{(c'(d))^2} dd$$

Для AI-сигнала с повышенной остротой k_{AI} вероятность больших амплитуд (выбросов) выше. Введём коэффициент чувствительности квантователя β, зависящий от характеристики компандера. Разложив интеграл в ряд Тейлора относительно отклонения остроты $(k - 3)$ получаем новую зависимость, с учетом (3):

$$\sigma_{q,AI}^2 = \frac{\Delta^2}{12} \cdot (1 + \beta \cdot (k_{AI} - 3)) \qquad (6)$$

где $\sigma_{q,AI}^2$ – мощность шума квантования AI-сигнала; Δ – шаг квантования; k_{AI} – острота распределения LPC-остатка AI-сигнала (2); β – коэффициент нелинейности квантователя.

Следовательно, коэффициент дисторсии сжатия DPCM для AI-сигнала определяется как отношение мощностей шума:

$$K_{DPCM-AI} = \frac{\sigma_{q,AI}^2}{\sigma_{q,nat}^2} \approx 1 + \beta \cdot (k_{AI} - k_{nat}) \qquad (7)$$

Подставляя экспериментальные значение $(k_{AI} = 5.72, k_{nat} = 3.24, \beta = 0.18)$, получаем:

$$K_{DPCM-AI} \approx 1 + 0.18 \cdot (5.72 - 3.24) \approx 1.45 \qquad (8)$$

Это говорит о том, что сжатие методом DPCM голосового сигнала, преобразованного с использованием искусственного интеллекта, приводит к увеличению мощности шума квантования на 45% по сравнению с естественной речью при идентичных параметрах кодека. Это создаёт математически детектируемый артефакт на этапе сжатия.

Получение аналитической зависимости мощности шума квантования от эксцесса распределения LPC-остатка для сигналов, преобразованных с использованием искусственного интеллекта, составляет научную новизну данного исследования.

C. Инкапсуляция в пакеты

На заключительном этапе подготовки к передаче в сети VoLTE сжатые кадры речи инкапсулируются в пакеты протокола реального времени (real-time protocol, RTP). Данный процесс регламентирован спецификацией RFC 4867 [7] в соответствии с профилем GSMA VoLTE [6]. Транспортный уровень обеспечивает доставку битового потока, сформированного кодеком AMR-WB, до абонентского устройства получателя.

Заголовок пакета RTP имеет фиксированную длину 12 октетов и содержит служебную информацию, необходимую для синхронизации и сборки потока: порядковый номер, метка времени, идентификатор источника и тип полезной нагрузки.

Важно отметить, что протокол RTP является прозрачным для содержимого полезной нагрузки. Он не модифицирует битовый поток сжатого сигнала, сформированный на этапе сжатия, следовательно, статистические характеристики шума квантования и эксцесс LPC-остатка, являющиеся яркими маркерами AI-преобразования, сохраняются внутри пакета неизменными [2]. Однако целостность детектируемых признаков зависит от вероятности доставки пакета.

Для обеспечения корректного воспроизведения речи метка времени для пакета вычисляется рекуррентно:

$$T_n = T_0 + n \cdot \Delta T \quad (9)$$

где T_0 – начальная метка времени сессии, ΔT – инкремент метки времени на один кадр. Для кодека AMR-WB при частоте дискретизации 16 кГц и длительности кадра 20 мс инкремент составляет:

$$\Delta T = f_s \cdot t_{frame} = 16000 \cdot 0.02 = 320 \text{ единиц} \quad (10)$$

Тело пакета содержит один или несколько кадров AMR-WB. Каждый кадр предваряется заголовком таблицы содержимого (ToC), указывающим режим кодирования и статус кадра. Размер полезной нагрузки варьируется от 35 до 61 байта в зависимости от выбранного режима кодека [6]. Таким образом, прямое влияние пакетизации на статистические свойства сигнала отсутствует, однако транспортные искажения (потеря пакетов, джиттер) могут привести к выпадению участков сигнала, содержащих артефакты, что эквивалентно потере детектируемых признаков.

D. Влияние VoLTE на артефакты AI-преобразования

Архитектура политики и управления тарификацией (PCC) в сетях мобильной связи обеспечивает выделение гарантированных ресурсов для голосового трафика. Для услуг голосовой связи назначается идентификатор класса качества обслуживания QCI=1 [6]. Это говорит о том, что вероятность потери настолько мала, что даже в случае потери данных их доля остается незначительной, что обеспечивает устойчивость оценок мощности шума квантования к сетевым искажениям. Более того, гарантированная полоса пропускания минимизирует потери из-за перегрузки канала и исключает массовое выпадение кадров, способное нарушить работу алгоритмов линейного предсказания [7]. Таким образом, параметры QoS VoLTE обеспечивают сохранность сжатого битового потока и позволяют детектору анализировать артефакты, внесённые именно кодеком, а не сетевой транспортировкой.

Несмотря на то, что прохождение сигнала через конвейер VoLTE не устраняет артефакты AI-преобразования, оно трансформирует их проявление как было показано в разделе эксперимента.

Во-первых, в несжатом сигнале основным признаком служит высокий эксцесс. После VoLTE прямой эксцесс сглаживается, но возникает вторичный артефакт – аномально высокая мощность шума квантования.

Во-вторых, коэффициент $K_{DPCM-AI}$ является наиболее устойчивым признаком, так как менее чувствителен к потерям.

В-третьих, так как артефакт формируется на этапе кодирования [6], он сохраняется неизменным при дальнейшей инкапсуляции и транспортировке, что позволяет размещать детектор как на стороне оператора, так и на стороне абонента.

IV. Заключение

В рамках исследования была разработана формальная модель преобразования речевого сигнала при регистрации и подготовке к передаче по протоколу VoLTE. Модель описывает три последовательных этапа обработки: дискретизацию и квантование, компрессию кодеком AMR-WB с использованием линейного предсказания и алгебраического кодирования возбуждения, инкапсуляцию кадров в пакеты протокола реального времени.

Научная новизна работы заключается в получении аналитической коррекции классической формулы мощности шума квантования для сигналов с негауссовским распределением остатка линейного предсказания, характерным для речи, преобразованной с использованием искусственного интеллекта.

Проведённый анализ влияния VoLTE на проявление артефактов синтеза позволил установить, что инкапсуляция в пакеты RTP сама по себе не вносит искажений в содержимое полезной нагрузки.

Полученные результаты создают теоретическую основу для улучшения детектора артефактов применения моделей искусственного интеллекта для синтезирования голосовых дипфейков [1], устойчивых к воздействию транспортного протокола, и могут быть использованы операторами связи для повышения уровня информационной безопасности систем голосовой связи.

Перспективы дальнейших исследований связаны с экспериментальной верификацией предложенной формальной модели на реальных наборах данных, включающих как естественную, так и искусственно преобразованную речь, прошедшую через транспортный конвейер оператора мобильной связи.

Список литературы

[1] Мирошников Н.Ю. Способ определения ИИ-подделок голосовых аудиозаписей: выпускная квалификационная работа. СПб.: СПбГЭТУ «ЛЭТИ», 2026. 117 с.

[2] Ишутин И.А. Способ идентификации типов сжатых аудиофайлов по результатам спектрального анализа: выпускная квалификационная работа. СПб.: СПбГЭТУ «ЛЭТИ», 2026. 117 с.

[3] Токарева У.В., Шульженко А.Д. Принцип формирования тестовых наборов сигналов в речевом диапазоне частот для исследования алгоритмов обработки аудиозаписей // Труды конференции «Информационная безопасность». СПб. 2025. С. 45–52.

[4] Мирошников Н.Ю., Шульженко А.Д. Модуль выявления применения AI-VC в аудиосообщениях // Сборник научных трудов ЛЭТИ. 2025. Вып. 4. С. 12–18.

[5] Jayant N.S., Noll P. Digital Coding of Waveforms. Englewood Cliffs, NJ: Prentice-Hall, 1984. 624 p.

[6] GSMA. VoLTE Service Description and Implementation Guidelines. Version 1.1. 2014. 121 p.

[7] Makhoul J. Linear Prediction: A Tutorial Review // Proceedings of the IEEE. 1975. Vol. 63. No. 4. P. 561–580. DOI: 10.1109/PROC.1975.9792.

[8] 3GPP TS 26.190. Speech codec speech processing functions; Adaptive Multi-Rate – Wideband (AMR-WB) speech codec; Transcoding functions. Release 16. 2021.

[9] IETF RFC 4867. RTP Payload Format and File Storage Format for the Adaptive Multi-Rate (AMR) and Adaptive Multi-Rate Wideband (AMR-WB) Audio Codecs. 2007.

[10] 3GPP TS 23.203. Policy and charging control architecture. Release 17. 2022

[11] Todisco M., Wang X., Vestman V., Sahidullah M., Delgado H., Nautsch A., Evans N., Lee K.A., Kinnunen T., Yamagishi J. The ASVspoof 2019

[12] Hicsonmez S., Uzun E., Sencar H. T. Methods for Identifying Traces of Compression in Audio // 2013 International Conference on Computer Systems and Applications. IEEE, 2013.

[13] Frank J., Schönherr L. WaveFake: A Data Set to Facilitate Audio Deepfake Detection // 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. 2021.