АКУСТИЧЕСКИЙ ЖУРНАЛ, 2023, том 69, № 4, с. 497–505

ОБРАБОТКА АКУСТИЧЕСКИХ СИГНАЛОВ. КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ

УДК 612.85

ДЕТЕКТИРОВАНИЕ УДАЛЕННОЙ РЕЧИ

© 2023 г. В. Н. Сорокин*

Институт проблем передачи информации, Российская академия наук, Большой Каретный пер. 19, стр. 1, Москва, 127051 Россия

**e-mail: vns@iitp.ru*

Поступила в редакцию 15.08.2022 г.
После доработки 09.11.2022 г.
Принята к публикации 22.12.2022 г.

Исследуются амплитудные и фазовые характеристики речевых сигналов, записанных на разном расстоянии от диктора микрофонами различных типов, в свободном пространстве и замкнутом помещении. Отношения средней энергии амплитудного спектра в различных диапазонах частот и средний наклон линейной компоненты фазы демонстрируют различия для слога, записанного вблизи микрофона, и такого же слога, записанного на удалении, и вновь воспроизведенного вблизи от микрофона. Наибольшее различие наблюдается в отношениях средней энергии в диапазонах частот 0–1 и 1–8 кГц, а также 3–4 и 4–6 кГц. Наклон линейной компоненты вычисляется в диапазоне 4–8 кГц. Степень различия зависит от гласного звука.

Ключевые слова: уязвимость верификации диктора, спектр удаленного сигнала, фаза удаленного сигнала, воспроизведение удаленного сигнала

DOI: 10.31857/S0320791923600282, **EDN:** QRYPVJ

1. ВВЕДЕНИЕ

Подтверждение личности по голосу (верификация, аутентификация) является важным элементом систем удаленного доступа к источникам информации, управления финансовыми или технологическими процессами. К алгоритмам верификации предъявляются противоречивые требования, такие как удобство для пользователя, устойчивость к изменениям голоса пользователя и характеристикам микрофонов, а также противодействие подмене голоса пользователя самозванцем.

Существует несколько способов взлома (spoofing) системы верификации диктора помимо вторжения в процесс принятия решения на конечном этапе:

1. имитация голоса целевого пользователя другим человеком,

2. запись речи целевого пользователя с последующим воспроизведением,

3. перехват речевого сигнала в канале связи,

4. преобразование голоса самозванца к голосу целевого пользователя с помощью вокодерной технологии или машинного обучения (Voice Convertion),

5. формирование модели голоса целевого диктора методами машинного обучения с помощью компиляционного синтезатора.

Методы борьбы (anti-spoofing) с попытками взлома систем верификации рассматриваются в [1–5]. Очевидно, что не может быть создано абсолютно надежной защиты от проникновения самозванца в систему верификации. Можно лишь рассчитывать на разработку такой защиты, преодоление которой обойдется злоумышленнику дороже, чем ожидаемая выгода. Необходимо также располагать системами защиты различной сложности с тем, чтобы учитывать риск, т.е. вероятность попытки взлома и возможные потери.

Имитация голоса пользователя не требует специальных технических средств, за исключением формирования речевой базы для тренировки имитатора. Такая имитация наиболее опасна для систем, в которых для верификации используются параметры голосового источника, в том числе и частота основного тона. Эта частота довольно легко воспроизводится, увеличивая риск пропуска самозванца. При этом успех имитации зависит не только от способностей имитатора, но и от особенностей голоса целевого пользователя [6, 7]. Преобразование голоса самозванца в голос целевого пользователя и синтез голоса целевого пользователя наиболее сложны с технической точки зрения, но и наиболее опасны, поскольку методы машинного обучения позволяют воспроизвести характерные особенности голоса [8, 9].

Воспроизведение подслушанной речи пользователя (replay) требует минимум технических

средств. Этот прием особенно опасен для систем верификации с фиксированным паролем, и даже для произвольных паролей ошибка может доходить до 100% [10–13]. Разработка методов противодействия атаке системы верификации диктора посредством воспроизведения подслушанного сигнала должна основываться на объективных различиях сигналов, записанных и воспроизведенных на разных расстояниях. Такие различия, как правило, ищут экспериментально, перебирая варианты формирования параметров речевого сигнала.

Обзор параметров, влияющих на восприятие расстояния до источника звука в помещении, был представлен в [14]. В [15–17] сообщается, что на восприятие расстояния влияет энергия в спектральной области ниже 500 Гц и отношение энергии прямого сигнала к реверберирующему.

Реверберация помещения тем больше искажает сигнал, чем на большем расстоянии от диктора производится запись. В [18] установлено, что в помещении с умеренной реверберацией (с задержкой около 6 мс) гистограмма энергии речевых сигналов длительностью 2 с смещается в сторону низких частот, а степень реверберации можно оценить по распределению энергии сигнал-остатка при линейном предсказании.

Вместе с тем, эксперименты по восприятию расстояния в ближнем (<1 м) и дальнем акустическом поле в свободном пространстве и помещениях с реверберацией демонстрируют противоречивые результаты в зависимости от спектра звука и отношения энергии в полосе частот ниже 3 кГц к энергии в высокочастотной области [19]. В других экспериментах обнаруживается, что важно не только относительное изменение объективных параметров речевого сигнала в зависимости от расстояния до источника звука. Так, в [20] было установлено, что оценка человеком расстояния до источника неизвестного звука выполняется с большой погрешностью, тогда как для известных звуков погрешность заметно меньше. Парное восприятие одного и того же слова, записанного на разных расстояниях от диктора, обычно позволяет лучше ощутить разницу в условиях записи.

Простейший детектор попытки воспроизведения состоит в сравнении принятого пароля с одним из сохраненных в базе обучения пользователя. Если мера сходства оказывается выше некоторого порога, то регистрируется попытка взлома. В системе [21] с этой целью пользователю предлагается дважды произнести одно и то же слово в парольной фразе. Этот прием основан на том факте, что многократно произнесенные слова всегда несколько отличаются друг от друга. Если записанные самозванцем сигналы не подвергаются преднамеренному небольшому искажению или не используется многократная запись одного и того же пароля, то вероятность успеха такой атаки существенно снижается. Другой детектор попытки воспроизведения записанного сигнала основан на искажении этого сигнала вследствие повторного наложения шумов и реверберации помещения (особенно в случае записи в одном помещении, а воспроизведения – в другом), в результате чего искажается амплитудный спектр сигнала и пауз [12].

Детектирование воспроизведения записанного речевого сигнала выполняется либо путем статистического анализа различных амплитудно-частотных характеристик, либо на основе различия в акустике исходного и воспроизведенного сигнала. В [22] обращают внимание на двойное преобразование аналог-цифра и цифра-аналог при записи и воспроизведении речевого сигнала, что приводит к искажению спектра в полосе частот 6–8 кГц при частоте отсчетов 16 кГц. В [23] рассматриваются характеристики амплитудной и частотной модуляции на выходе гребенки фильтров Габора. Такая же гребенка фильтров используется в [24], где исследуется влияние реверберации на характеристики так называемого нелинейного оператора Тигера [25]. В [26] сравниваются пики траекторий гармоник исходного и записанного сигналов. Наряду с параметрами кратковременного спектра мощности, для детектирования воспроизведения применяются и фазовые характеристики, такие как групповая задержка с нормировкой фазы [27, 28]. В [29] исследовалась роль линейной компоненты фазы в оценке параметров речевого сигнала.

Несмотря на значительные усилия по разработке методов противодействия взлома системы верификации диктора с помощью воспроизведения записанного сигнала, эта проблема требует дальнейшего теоретического и экспериментального исследования. Одно из перспективных направлений состоит в анализе фазовых характеристик.

2. МАТЕМАТИЧЕСИКЕ МОДЕЛИ

Анализ акустики ближнего и дальнего поля может указать на параметры речевого сигнала, которые наиболее перспективны для обнаружения записанного сигнала. Предположим, что при атаке на систему верификации записанный сигнал воспроизводится примерно на том же расстоянии от микрофона системы, что и целевой диктор при стандартной процедуре. Тогда отличие заключается в расстоянии до микрофона злоумышленника. Кроме того, тип микрофона злоумышленника может отличаться от типа микрофона системы верификации. Рассмотрим случай, когда при обучении и верификации расстояние от пользователя до микрофона не больше 20 см, а расстояние до микрофона злоумышленника существенно больше 20 см.

В качестве приемников звука используются различные виды микрофонов: угольные, пьезоэлектрические, электродинамические, электромагнитные, ленточные и конденсаторные; прямого и дифференциального действия. В зависимости от конструкции микрофон может быть приемником звукового давления, градиента давления, скорости или ускорения колебаний воздушных частиц. Наиболее качественные микрофоны представлены конденсаторным типом, в котором электрический сигнал создается изменением емкости конденсатора, представляющего собой две мембраны. Если одна из мембран неподвижна, а другая колеблется под воздействием акустической волны, то такой микрофон является приемником давления. Если и другая мембрана колеблется под влиянием акустической волны, приходящей с обратной стороны микрофона, то такой микрофон является приемником градиента давления.

В качестве модели механизма распространения речевого сигнала в пространстве примем модель колебаний поверхности сферы. Уравнение сферической распространяющейся волны акустического давления P есть

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial P}{\partial r}\right)-\frac{1}{c_0^2}\frac{\partial^2 P}{\partial t^2}=0,\quad r>r_0, \tag{1}$$

где r – расстояние от центра сферы, r_0 – радиус сферы, t – время, c_0 – скорость звука в воздухе. Акустическое давление P и радиальная скорость колебаний воздушных частиц V связаны соотношениями

$$\frac{\partial P}{\partial r}=-\rho_0\frac{\partial V}{\partial t}, \tag{2}$$

$$\frac{\partial P}{\partial t}=\frac{\rho_0 c_0^2}{r^2}\frac{\partial (r^2 V)}{\partial r}, \tag{3}$$

ρ_0 – плотность воздуха. Решение уравнения (1) для расходящихся волн давления и скорости имеет вид:

$$P=\frac{\rho_0}{r}\frac{\partial f(t-r/c_0)}{\partial t}, \tag{4}$$

$$V=\frac{1}{r^2}f(t-r/c_0)+\frac{1}{rc_0}\frac{\partial f(t-r/c_0)}{\partial t}. \tag{5}$$

Из (4) и (5) получается связь между скоростью и давлением в расходящейся волне

$$\rho_0 V=\frac{P}{c_0}+\frac{1}{r}\int_{-\infty}^{t}P\,dt. \tag{6}$$

Из (6) следует, что скорость частиц воздуха V на больших расстояниях от поверхности сферы при $r\to\infty$ почти пропорциональна акустическому давлению P, тогда как на малых расстояниях V зависит и от давления и от его интеграла. Интегрирование в (6) приводит также и к сдвигу фазы волны. Сравнение (4) и (5) позволяет уточнить эти выводы. Дифференцирование по времени эквивалентно умножению на частоту в спектральной области. Поэтому на высоких частотах в (5) доминирует второй член, и на всех расстояниях разница между давлением и скоростью мала, что препятствует обнаружению разницы в расстоянии микрофона от диктора. На низких частотах в (5) доминирует первый член, и скорость существенно отличается от давления. Более конкретная оценка диапазона частот, в котором эта разница может быть обнаружена, основана на понятиях ближнего и дальнего акустического поля, которые описывают эффекты дифракции акустических волн в зависимости от размера источника звука и длины волны.

Пусть сфера радиуса r_0 пульсирует по гармоническому закону как $v_r(t)=v_0e^{-jKc_0t}$, где $v_r(t)$ – радиальная скорость ее поверхности, а K – волновое число, $K=\omega/c_0$, ω – частота. Тогда для гармонических колебаний решение волнового уравнения (1) относительно давления и скорости при произвольном K есть [30]

$$P=\frac{r_0V_0\rho_0c_0}{r}\frac{jKr_0}{1+jKr_0}e^{-jK(r-r_0)}, \tag{7}$$

$$V=\frac{r_0V_0}{r}\frac{jKr_0}{1+jKr_0}\frac{1-jKr}{jKr}e^{-jK(r-r_0)}. \tag{8}$$

В ближнем акустическом поле, т.е. при $Kr_0\ll 1$,

$$\frac{jKr_0}{1+jKr_0}\xrightarrow[Kr_0\to 0]{}jKr_0.$$

Конкретизируя условие ближнего поля как $Kr_0=0.1$, оценим максимальную частоту $F_{\max}=0.1c_0/2\pi a$, для которой выполняется условие ближнего поля. При $r_0=10$ см (что примерно соответствует среднему радиусу головы), $F_{\max}\approx 54$ Гц, а при $r_0=1$ см (что примерно соответствует эквивалентному радиусу ротового отверстия), $F_{\max}\approx 540$ Гц. Последняя оценка близка к данным [15], согласно которым восприятие человеком расстояния до источника звука определяется компонентами сигнала в диапазоне частот ниже 500 Гц. На рис. 1 показаны амплитудно-частотные характеристики отношения $jKr_0/(1+jKr_0)$, из которых следует, что в свободном пространстве высокочастотные компоненты давления (7) и скорости (8) для удаленного сигнала значительно ниже, чем в ближнем поле. Отношение средней амплитуды в диапазоне частот от 4–8 кГц к средней амплитуде в диапазоне частот ниже 500 Гц для дальнего поля равно 23.1, а для ближнего поля – 11.5. На очень

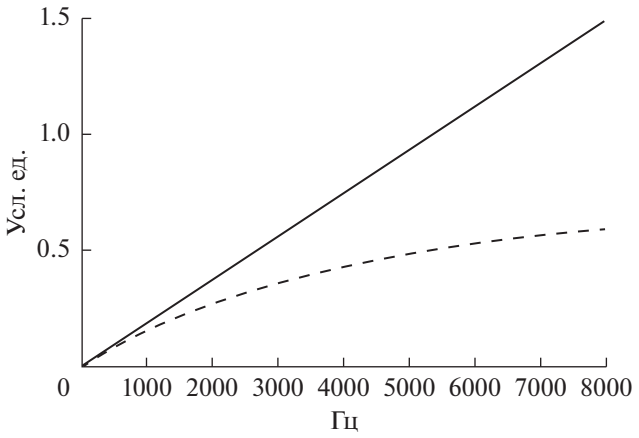

Рис. 1. Амплитудно-частотная характеристика компонент акустической волны в ближнем поле (–) и дальнем поле (---). По оси ординат – условные единицы.

малых расстояниях скорость (8) обратно пропорциональна расстоянию, поскольку

$$\frac{1 - jKr}{jKr} \xrightarrow[r\to 0]{} \infty.$$

Еще одно отличие оригинального сигнала от воспроизведенного состоит в искажении характеристик излучения. Поскольку отношение радиуса ротового отверстия к диаметру головы находится в диапазоне 0.08–0.14 [30], то свойства излучения речевого сигнала из ротового отверстия близки к свойствам излучения поршня, вставленного в бесконечный экран. Удельный безразмерный импеданс такого излучения, по [31], аппроксимируется как

$$Z = 1 - \frac{J_1(Kh)}{Kh} - j\frac{K_1(2Kh)}{2(Kh)^2}, \qquad (9)$$

где h – эквивалентный радиус ротового отверстия, J_1 – функция Бесселя первого рода, K_1 – функция Бесселя второго рода. С погрешностью около 10% (9) приводится к равенству, которое легко интерпретируется

$$Z = \frac{(\omega h)^2}{2c_0} - j\frac{8\omega h}{3\pi c_0}. \qquad (10)$$

Из (10) следует, что потери на излучение, а следовательно, и затухание компонент сигнала пропорциональны квадрату частоты и эквивалентному радиусу раскрытия ротового отверстия. В то же время, умножение на ω соответствует дифференцированию сигнала, что приводит к возрастанию амплитуд частотных компонент сигнала пропорционально частоте и сдвигу фаз.

Новый подход к разработке систем детектирования попытки проникновения в систему верификации диктора может быть основан на свойствах уравнения (6). Из этого уравнения следует, что разница в расстояниях до близкого и удаленного микрофона обнаруживается, если речевой сигнал принимается одновременно двумя микрофонами – приемниками звукового давления и скорости колебаний. Тогда, обозначив разность между нормированными сигналами, пропорциональными давлению P и V, как $\Delta_{PV} = V - P/\rho_0 c_0$, получим оценку расстояния до источника звука

$$r = \frac{1}{\rho_0 \Delta_{PV}} \int_{-\infty}^{t} P dt.$$

В воспроизводимом сигнале на характеристики излучения оригинального сигнала накладываются характеристики излучения динамика воспроизводящего устройства. Поэтому в спектрах оригинального и воспроизведенного речевого сигнала следует ожидать различия в высокочастотной области спектра.

3. ЭКСПЕРИМЕНТЫ

Эксперименты по обнаружению отличий между речевыми сигналами выполнялись на записях вблизи от микрофона и на некотором удалении от него. Эти записи производились в осеннем лесу и жилой комнате с размерами 12 × 3 ×2.5 м. Использовались три типа микрофонов: профессиональный направленный микрофон с шумоподавлением и встроенным преобразователем аналог-цифра, встроенный микрофон ноутбука и смартфон.

При этом расстояние от диктора до микрофона ноутбука составляло около 20 см. Для симуляции подслушивания использовались направленный микрофон и смартфон, расположенные на расстоянии около 5, 20 и 80 см от диктора. В экспериментах по воспроизведению записанных на смартфон речевых сигналов на расстоянии около 80 см, эти сигналы записывались либо на направленный микрофон, либо на микрофон ноутбука при положении ноутбука на расстоянии около 20 см.

Речевой материал состоял из последовательности слогов */ата этэ ото уту ити ыты/* с ударением на последнем гласном. Последовательность звуков произносилась как одна фраза с небольшими паузами между слогами. В комнате запись выполнялась параллельно на пары микрофонов “направленный микрофон/смартфон” и “ноутбук/смартфон” на разных расстояниях от диктора и друг от друга. В записях участвовал один диктор мужского пола. Обработка речевого сигнала заключалась в вычислении кратковременного спектра Фурье (КПФ) с гауссовым окном на интервале в 256 отсчетов при частоте отсчетов 16 кГц со сдвигом окна на 1 отсчет на всем сегменте последнего гласного каждого слога и последующем анализе параметров амплитудного и фазового спектра.

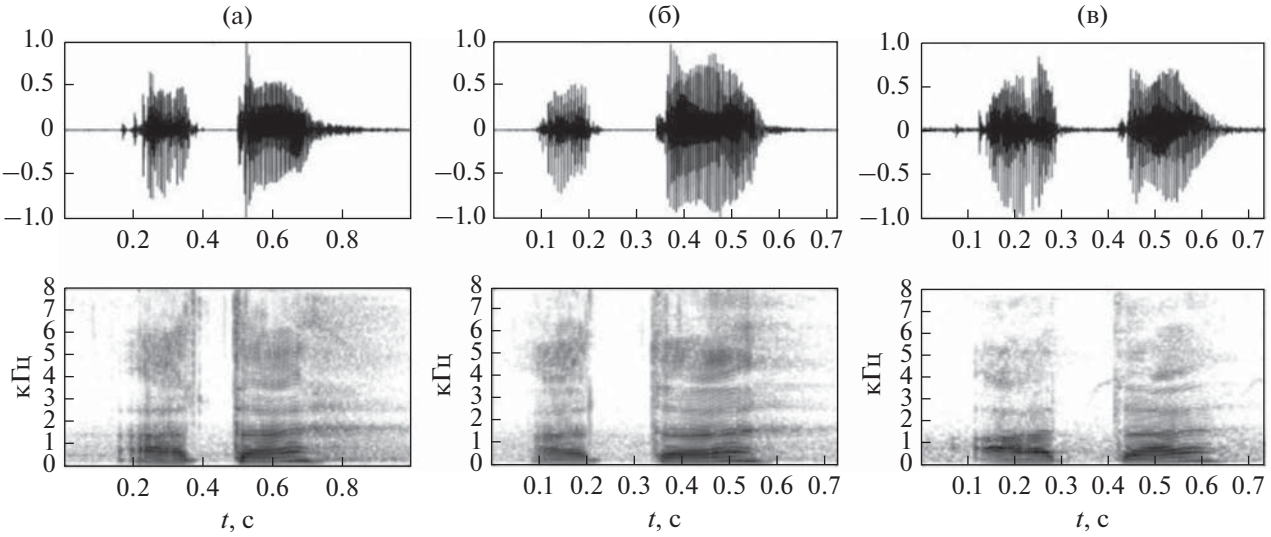

Рис. 2. Сонограммы трех произнесений слога */ama/* на разных расстояниях от микрофона: (а) – $r = 5$ см, (б) – $r = 20$ см, (в) – $r = 80$ см.

При записи речевых сигналов в лесу для каждого положения микрофона фраза с последовательностью слогов произносилась заново. Поэтому в характеристиках звуков могли присутствовать естественные различия в длительности записи и параметрах речевого сигнала (рис. 2, слог */ama/*). При записи в помещении одно и то же произнесение одновременно регистрировалось двумя микрофонами на разных расстояниях от диктора. Это исключает естественные вариации произнесения при поиске наиболее информативных различий, связанных с типом и расположением подслушивающего микрофона.

Первый цикл экспериментов был выполнен для записей в лесу. Цель этих экспериментов состояла в оценке различий в амплитудном и фазовом спектре при распространении речевого сигнала в свободном пространстве без влияния реверберации.

На рис. 3 показаны амплитудные и фазовые спектры последнего (ударного) гласного */a/* в слогах, записанных в лесу на расстоянии 5, 20 и 80 см. На интервале гласного амплитудный спектр усреднялся, нормировался к максимуму, и затем из него вычиталась постоянная составляющая, равная 0.3. Фазовые спектры вычислялись двумя способами. В первом способе фаза находилась также по комплексному спектру кратковременного преобразования Фурье (КПФ), и затем усреднялась на интервале гласного. Во втором способе комплексный спектр вычислялся через быстрое преобразование Фурье (БПФ) на всем интервале гласного.

В экспериментах с записью речевых сигналов в лесу было обнаружено, что наибольшее отличие амплитудных спектров гласных заключается в отношении среднего уровня амплитуд в диапазоне частот 4–6 и 3–4 кГц. Фазовые спектры различаются по среднему наклону в диапазоне частот 4–8 кГц. В соответствии с этим явлением были сформированы параметры $\delta A_r = 1 - (\overline{A}_{4,6}/\overline{A}_{3,4})$ и m_r, где $\overline{A}_{3,4}$ – среднее значение амплитудного спектра в диапазоне частот 3–4 кГц, $\overline{A}_{4,6}$ – среднее значение амплитудного спектра в диапазоне частот 4–6 кГц, m_r – средний наклон фазы, вычисленной по БПФ в диапазоне частот 4–8 кГц, $r = 5$, 20 и 80 см. Этот параметр лучше описывает нелинейность фазового спектра. Различие между фазовыми спектрами, вычисленными по КПФ, оказалось незначительным. Поэтому в табл. 1 эти параметры не показаны.

Представленные в табл. 1 данные свидетельствуют о заметном различии параметров δA_r и m_r для сигналов, записанных на разных расстояниях, но характер и степень различия зависит от типа гласного.

Таблица 1. Параметры амплитудных и фазовых спектров в свободном пространстве, δA, %

Лес	*а*	*э*	*о*	*у*	*и*	*ы*
δA_5	4.0	6.2	–2.6	–23.7	1.8	–3.6
δA_{20}	13.2	10.9	7.8	0.9	7.8	5.6
δA_{80}	20.0	23.1	–12.2	–7.3	17.5	9.8
m_5	17.4	18.1	11.7	15.5	11.6	4.8
m_{20}	25.2	29.6	19.5	29.9	22.1	15.0
m_{80}	28.1	27.6	24.3	18.5	13.4	17.7

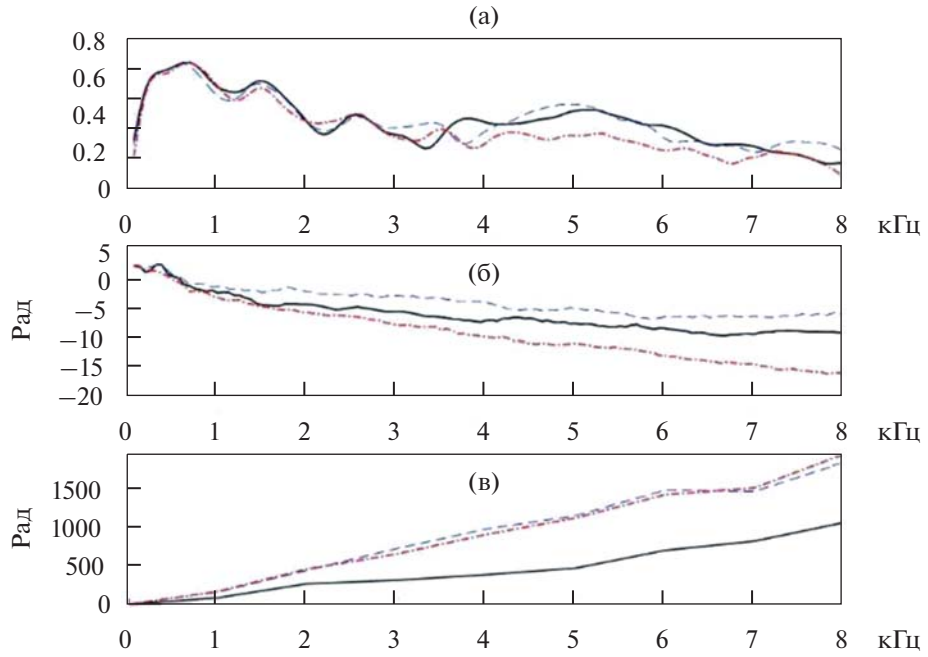

Рис. 3. Слог /*a*/: (а) – амплитудный спектр; (б) – фазовый спектр по КПФ; (в) – фазовый спектр по БПФ. Микрофон на расстоянии 5 см (—), 20 см (---), 80 см (- • -).

Таблица 2. Параметры амплитудных и фазовых спектров в помещении. Смартфон, δA, %

Комната	*а*	*э*	*о*	*у*	*и*	*ы*
δA_{20}	11, **24**	18, **36**	17, 29	−19, −3	−1, **21**	−19, −4
δA_{100}	11, **34**	18, **23**	18, 26	−19, 2	−1, **13**	−22, 6
m_{20}	11, **19**	20, 35	13, **26**	7, 14	**14**, 19	11, 14
m_{100}	15, **34**	15, 33	14, **18**	7, 14	9, 20	9, 13

Таблица 3. Одновременная запись через направленный микрофон и смартфон. Расстояние до диктора около 20 см. δA, %

Комната	*а*	*э*	*о*	*у*	*и*	*ы*
δA_{20}	**20.9**	**30.8**	**29.2**	**−3.0**	**21.1**	**−8.5**
δA_{80}	**11.4**	**18.2**	**20.6**	**−18.8**	**−0.8**	**−19.4**
m_{20}	18.6	34.7	18.2	**14.5**	**18.2**	14.1
m_{80}	17.3	31.1	18.1	**7.3**	**10.9**	11.9

Таблица 4. Одновременная запись через направленный микрофон на расстоянии около 20 см и смартфон на расстоянии около 80 см. δA, %

Комната	*а*	*э*	*о*	*у*	*и*	*ы*
δA_{20}	14.9	16.2	21.4	−20.0	**7.4**	**−5.3**
δA_{80}	16.6	18.8	20.4	−19.5	**−1.1**	**−22.4**
m_{20}	30.8	30.9	26.8	**16.5**	**29.8**	**13.2**
m_{80}	29.9	32.9	13.9	**9.5**	**9.5**	**9.5**

Если в помещении микрофон расположен достаточно далеко от диктора, то обычно возникает субъективное ощущение отклика помещения (реверберации). При этом характеристики одного и того же слога при многократном произнесении различаются не только в силу естественной вариации артикуляции, но также и из-за вариации расстояния и направления на микрофон. В табл. 2 приведены диапазоны оценок амплитудных и фазовых параметров речевых сигналов, записанных через смартфон, который находится на расстоянии около 20 или 80 см от диктора. Видно, что эти диапазоны в значительной степени перекрываются, но для некоторых гласных одна из границ заметно смещена в зависимости от расстояния до микрофона.

Различие между разными типами микрофонов при одновременной записи речевого сигнала иллюстрируется табл. 3–6, где жирными цифрами отмечены параметры, отличающиеся примерно в 1.5 раза. Здесь наибольшее различие в типах микрофона обнаруживается на малом расстоянии от диктора.

В дополнение к параметрам, описывающим отношение средних амплитуд в диапазонах 3–4 и 4–8 кГц между оригинальной и воспроизведенной записью (δA_{orig} и δA_{repl}), при анализе характеристик воспроизведенного сигнала обнаружилась разница средних амплитуд в диапазонах 0–1 и 1–8 кГц (rA_{orig} и rA_{repl}), где $rA_r = 1 - (\bar{A}_{0,1}/\bar{A}_{1,8})$ (рис. 4). Это связано с отмеченной в разделе 2 раз-

Таблица 5. Одновременная запись через микрофон ноутбука и смартфон. Расстояние до диктора около 20 см. δA, %

Комната	*а*	*э*	*о*	*у*	*и*	*ы*
δA_{20}	**11.6**	**22.9**	**24.1**	**4.5**	**8.6**	3.4
δA_{80}	**20.2**	**30.2**	**17.6**	**14.4**	**12.4**	–3.7
m_{20}	**21.3**	20.7	**6.1**	11.1	**23.5**	**19.7**
m_{80}	**11.1**	18.0	**16.6**	10.8	**13.0**	**11.1**

Таблица 6. Одновременная запись через микрофон ноутбука на расстоянии около 20 см и смартфон на расстоянии около 100 см. δA, %

Комната	*а*	*э*	*о*	*у*	*и*	*ы*
δA_{20}	**18.1**	18.9	**29.2**	**3.9**	**21.8**	**–3.7**
δA_{80}	**10.7**	22.4	**18.4**	**–14.7**	**12.6**	**–13.8**
m_{20}	**26.2**	20.2	17.3	**5.1**	21.9	8.9
m_{80}	**15.1**	15.7	15.2	**14.1**	20.0	10.8

ницей в уровне низкочастотных компонент в ближнем и дальнем поле.

Сравнительные характеристики записи через направленный микрофон и микрофон ноутбука при воспроизведении на расстоянии около 20 см сигналов от смартфона, записанных на расстоянии около 1 м, показаны в табл. 7 и 8. Здесь параметры δA_{orig}, rA_{orig} и m_{orig} представлены дважды для двух независимых сессий, так что отображаются естественные вариации параметров разных произнесений.

4. ОБСУЖДЕНИЕ

Обнаруженные особенности амплитудного спектра в зависимости от расстояния между микрофоном и диктором согласуются с высказанными ранее предположениями о роли низкочастотных и высокочастотных компонент в восприятии расстояния до источника звука [15]. Подтверждаются также результаты экспериментов [20], свидетельствующие о возможности оценки расстояния до источника звука при сопоставлении характеристик конкретного звукового образа. На это указывает зависимость параметров речевого сигнала от типа гласного, которая найдена в описанных экспериментах.

Помимо амплитудных характеристик, найдена и зависимость линейной составляющей фазы от расстояния до источника звука. Из сравнения табл. 1 и 2 следует, что, в отличие от распространения звука в свободном пространстве, реверберация замкнутого помещения приводит к ухудшению различия амплитудных характеристик в зависимости от расстояния до источника звука. При этом различимость фазовых параметров не ухудшается.

В результате формируется трехмерное пространство признаков $(rA, \delta A, m)$, в котором можно попытаться детектировать попытку взлома системы верификации с помощью воспроизведения

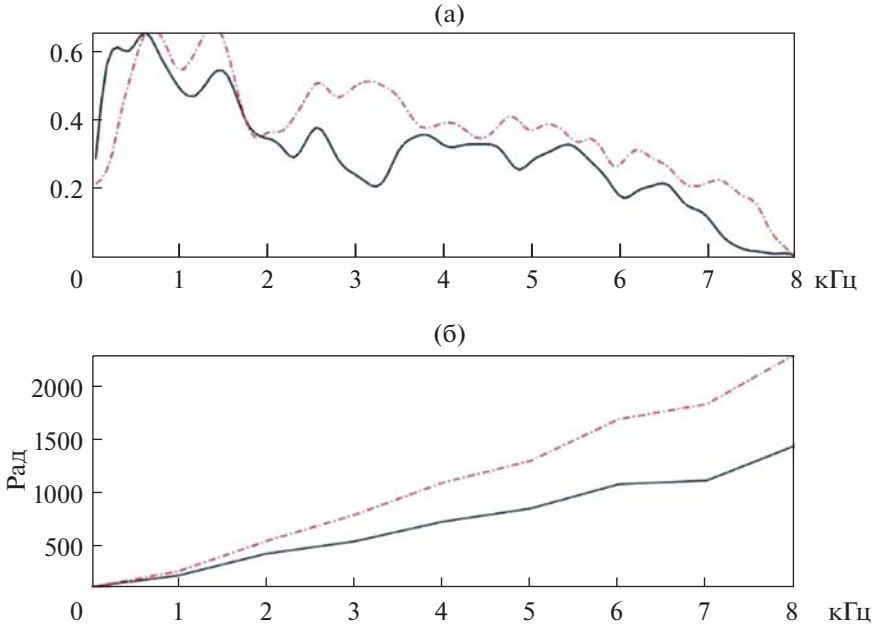

Рис. 4. Слог */ama/*: (а) – амплитудный спектр; (б) – фазовый спектр. Микрофон на расстоянии 20 см (—), 80 см (- • -).

Таблица 7. Параметры двух записей через направленный микрофон на расстоянии 20 см (rA_{orig}, δA_{orig}, m_{orig}) и воспроизведения удаленной записи на смартфон на расстоянии 20 см от микрофона (rA_{repl}, δA_{repl}, m_{repl})

Комната	*а*	*э*	*о*	*у*	*и*	*ы*
rA_{orig}	51.1, 44.5	46.0, 46.9	62.3, 58.1	65.3, 63.3	48.9, 31.7	42.3, 44.7
rA_{repl}	25.6	26.9	54.8	51.7	13.0	11.1
δA_{orig}	20.8, 14.9	30.8, 16.1	29.2, 21.4	−3.1, −20.0	21.1, 7.4	−8.6, −5.2
δA_{repl}	25.9	35.1	7.8	−10.1	4.3	0.6
m_{orig}	18.6, 30.8	34.7, 30.9	18.2, 26.8	14.5, 16.5	18.2, 29.8	14.1, 13.2
m_{repl}	31.5	24.3	18.3	5.2	13.0	15.6

Таблица 8. Параметры двух записей через микрофон ноутбука на расстоянии 20 см (rA_{orig}, δA_{orig}, m_{orig}) и воспроизведения удаленной записи на смартфон на расстоянии 20 см от ноутбука (rA_{repl}, δA_{repl}, m_{repl})

Комната	*а*	*э*	*о*	*у*	*и*	*ы*
rA_{orig}	46.9, 44.7	44.2, 47.1	66.8, 61.6	66.3, 63.5	34.7, 39.4	44.8, 48.2
rA_{repl}	20.3	20.7	38.5	45.5	5.1	8.9
δA_{orig}	11.6, 18.1	22.9, 18.9	24.1, 29.2	4.5, 3.9	8.6, 21.8	3.3, −3.7
δA_{repl}	19.8	6.1	19.8	−5.1	7.2	5.8
m_{orig}	21.3, 26.2	20.7, 20.2	6.1, 17.3	11.1, 5.1	23.5, 21.9	19.7, 8.9
m_{repl}	36.0	23.5	24.3	5.6	16.7	14.6

речевого сигнала, подслушанного на удалении от диктора. Две компоненты этого пространства зависят только от амплитудного спектра принятого сигнала. Это отношение среднего уровня амплитуд $\overline{A}_{0,1}/\overline{A}_{4,8}$ в диапазоне частот 0–1 и 4–8 кГц, и отношение $\overline{A}_{4,6}/\overline{A}_{3,4}$ в диапазоне 4–6 и 3–4 кГц. Третья компонента представляет собой линейную составляющую среднего фазового спектра на интервале частот 4–8 кГц. При этом необходимо учитывать неоднократно отмеченную в речевых исследованиях сильную зависимость фазового спектра от параметров алгоритма преобразования Фурье. Как показано на рис. 3, вид фазового спектра может отличаться и диапазоном значений, и даже знаком. Это пространство признаков должно быть сформировано отдельно для каждого гласного, который используется в голосовом пароле.

В дополнение к рассмотренным выше амплитудным и фазовым параметрам, эта оценка может послужить еще одним признаком при формировании детектора попытки взлома системы верификации диктора с помощью воспроизведения подслушанного речевого сигнала.

Следует отметить, что даже тот крайне ограниченный объем речевых данных, который использовался в описанных экспериментах, указывает на заметное перекрытие диапазонов параметров, перспективных для обнаружения речевого сигнала, записанного на большом расстоянии. Для того чтобы оценить возможность успешного применения найденных параметров для защиты систем верификации диктора, необходимо выполнить статистически достоверные исследования с различными типами микрофонов.

5. ЗАКЛЮЧЕНИЕ

Наблюдается различие в параметрах речевого сигнала, записанного в ближнем акустическом поле, и параметров этого же сигнала, записанного в дальнем акустическом поле и воспроизведенного в ближнем поле. В число этих параметров входят относительные значения энергии спектра в диапазонах низких, средних и высоких частот, а также средний наклон линейной компоненты фазового спектра в высокочастотной области. Степень различия зависит от типа гласного звука.

СПИСОК ЛИТЕРАТУРЫ

1. *Wu Z., Evans N., Kinnunen T., Yamagishi J., Alegre F., Li H.* Spoofing and countermeasures for speaker verification: A survey // Speech Communication. 2015. V. 66. P. 130–153.

2. *Kinnunen T., Sahidullah M., Delgado H., Todisco M., Evans N., Yamagishi J., Lee K.A.* The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection // InterSpeech 2017.

3. *Sahidullah M., Delgado H., Todisco M., Kinnunen T., Evans N., Yamagishi J., Lee K.A.* Introduction to voice presentation attack detection and recent advances // Handbook of Biometric Anti-Spoofing. 2019. P. 321–361. Springer, Cham.

4. *Lee K.A., Sadjadi O., Li H., Reynolds D.* Two decades into Speaker Recognition. Evaluation – are we there yet? // Computer Speech & Language. 2020. V. 61. 101058.

5. *Kamble M.R., Sailor H.B., Patil H.A., Li H.* Advances in anti-spoofing: from the perspective of ASVspoof challenges // APSIPA Transactions on Signal and Information Processing. 2020. V. 9. № 1. e2. https://doi.org/10.1017/ATSIP.2019.21

6. *Lau Y.W., Wagner M., Tran D.* Vulnerability of speaker verification to voice mimicking // IEEE Int. Symp. Intelligent Multimedia, Video and Speech Proc. 2004. P. 145–148. Hong Kong, 2004.

7. *Campbell J.P.* Speaker recognition: a tutorial // Proc. IEEE. 1997. V. 85. P. 1437–1462.

8. *Khodabakhsh A., Mohammadi A., Demiroglu C.* Spoofing voice verification systems with statistical speech synthesis using limited adaptation data // Computer Speech and Language. 2017. V. 42. P. 20–37.

9. *Sisman B., Yamagishi J., King S., Li H.* An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning // IEEE/ACM Trans. on Audio, Speech and Language Proc. 2021. V. 29. P. 132–157.

10. *Lindberg J., Blomberg M.* Vulnerability in speaker verification – A study of technical impostor techniques // Proc. European Conference on Speech Communication and Technology (Eurospeech). 1999. P. 1211–1244.

11. *Villalba J., Lleida E.* Preventing replay attacks on speaker verification systems // IEEE Int. Carnahan Conf. on Security Technology (ICCST). 2011. https://doi.org/10.1109/CCST.2011.6095943

12. *Wang Z.F., Wei G., He Q.H.* Channel pattern noise based playback attack detection algorithm for speaker recognition // Proc. IEEE Int. Conf. Machine Learning and Cybernetics (ICMLC). 2011. P. 1708–1713.

13. *Gałka J., Grzywacz M., Samborski R.* Playback attack detection for text-dependent speaker verification over telephone channels // Speech Comm. 2015. V. 67. P. 143–153.

14. *Kolarik A.J., Moore B.C.J., Zahori P., Cirstea S., Pardhan S.* Auditory distance perception in humans: A review of cues, development, neuronal bases, and effects of sensory loss // Atten., Percept. Psychophys. 2016. V. 2. № 78. P. 373–395.

15. *Скучик Е.* Основы акустики. М.: ИИЛ, 1959. Т. 2.

16. *Kopco N., Shinn-Cunningham B.G.* Effect of stimulus spectrum on distance perception for nearby sources // J. Acoust. Soc. Am. 2011. V. 130. № 3. P. 1530–1541.

17. *Prud'homme L., Lavandier M.* Do we need two ears to perceive the distance of a virtual frontal sound source? // J. Acoust. Soc. Am. 2020. V. 148. № 3. P. 614–1623.

18. *Georganti E., May T., Par S.V.D., Harma A., Mourjopoulos J.* Speaker distance detection using a single microphone // IEEE Trans. Audio Speech Lang. Process. 2011. V. 19. P. 1949–1961. https://doi.org/10.1109/TASL.2011.2104953

19. *Spiousas I., Etchemendy P.E., Eguia M.C., Calcagno E.R., Abregú E., Vergara R.O.* Sound spectrum influences auditory distance perception of sound sources located in a room environment // Frontiers in Psychology. 2017. V. 8. P. 969.

20. *Coleman P.D.* Failure to localize the source distance of an unfamiliar sound // J. Acoust. Soc. Am. 1962. V. 34. P. 345–346.

21. *Сорокин В.Н., Цыплихин А.И.* Верификация диктора по спектрально-временным параметрам речевого сигнала // Информационные процессы. 2010. Т. 10. № 2. С. 87–104.

22. *Witkowski M., Kacprzak S., Zelasko P., Kowalczyk K., Gałka J.* Audio replay attack detection using high-frequency features // InterSpeech. 2017. P. 27–31.

23. *Kamble M.R., Tak H., Patil H.A.* Amplitude and frequency modulation-based features for detection of replay spoof speech // Speech Communication. 2020. V. 125. P. 114–127.

24. *Kamble M.R., Patil H.A.* Detection of replay spoof speech using Teager energy feature cues // Computer Speech & Language. 2021. V. 65. 101140.

25. *Teager H.* Some observations on oral airflow during phonation // IEEE Trans. Acoust. Speech Signal Proc. 1980. V. 28. № 5. P. 599–601.

26. *Shang W., Stevenson M.* Detection of speech playback attacks using robust harmonic trajectories // Computer Speech & Language. 2021. V. 65. 101133.

27. *Oo Z., Wang L., Phapatanaburi K., Liu M., Nakagawa S., Iwahashi M., Dang J.* Replay attack detection with auditory filter-based relative phase features // EURASIP Journal on Audio, Speech, and Music. 2019. Art. number 8.

28. *Liu M., Wang L., Danga J., Lee K.A., Nakagawa S.* Replay attack detection using variable-frequency resolution phase and magnitude features // Computer Speech & Language Volume. 2021. V. 66. 101161.

29. *Сорокин В.Н., Леонов А.С.* Фазовые модуляции в речевом сигнале // Акуст. журн. 2022. Т. 68. № 2. С. 218–232.

30. *Фланаган Дж.* Анализ, синтез и восприятие речи. М.: Связь, 1968.

31. *Морз Ф.* Колебания и звук. М.–Л.: ГИТТЛ, 1949.