DOI 10.37882/2223–2966.2022.04–2.30

О ПЕРСПЕКТИВАХ ВНЕДРЕНИЯ БИОМЕТРИЧЕСКОЙ ИНФОРМАЦИИ В ГРАФИЧЕСКИЙ ФОРМАТ РЕЧЕВОГО СИГНАЛА

ON THE PROSPECTS FOR THE INTRODUCTION OF BIOMETRIC INFORMATION INTO A GRAPHICAL SPEECH SIGNAL FORMAT

I. Savelyev
A. Antipenko

Summary. Currently, due to the improvement of machine learning algorithms, as well as a significant reduction in the cost of computing power of server equipment, deepfake audio technologies are becoming more and more popular, which allow artificially synthesizing (faking) the speaker's voice. The article discusses the main areas of application of deepfake audio, the threats that these technologies carry, as well as ways to stop them, one of which is the use of a speech signature. A software and hardware complex are presented that allows, based on the method of calculating the phase characteristics of voice vocalisms, including embedding biometric information in the graphic format of a speech signal.

Keywords: voice vocalisms, speech information protection, intelligibility, speech signal, phase characteristics, sinusoidal model, biometrics.

Савельев Иван Андреевич
К.т.н., доцент, Финансовый университет при Правительстве РФ, г. Москва
IASavelyev@fa.ru
Антипенко Антон Олегович
Аспирант, Финансовый университет при Правительстве РФ, г. Москва
An-go-55@yandex.ru

Аннотация. В настоящее время из-за совершенствования алгоритмов машинного обучения, а также существенного удешевления вычислительной мощности серверного оборудования становятся всё более доступными, и, следовательно, популярными технологии дипфейк-аудио, позволяющие искусственно синтезировать (подделывать) голос диктора. В статье рассматриваются основные сферы применения дипфейк-аудио, угрозы, которые несут в себе данные технологии, а также способы их купирования, одним из которых является использование речевой подписи. Представлен программно-аппаратный комплекс, позволяющий на основе метода вычисления фазовых характеристик голосовых вокализмов в том числе встраивать биометрическую информацию в графической формат речевого сигнала.

Ключевые слова: голосовые вокализмы, защита речевой информации, разборчивость, речевой сигнал, фазовые характеристики, синусоидальная модель, биометрия.

Введение

Одним из самых перспективных и быстроразвивающихся направлений в области информационных технологий в 2010-х годах стали разработки, основанные на распознавании и синтезе голоса человека. Это напрямую происходит из их естественности и удобства для пользователя [1]. При этом математический аппарат, позволяющий проводить соответствующие преобразования, был в основном разработан ещё в 80-е годы прошлого века, однако тогда такие вычисления оказались слишком трудоёмкими, мощностей серверов не хватало даже для низкокачественного синтеза речи. Сейчас ситуация кардинально изменилась — стоимость вычислений падает из года в год, а развитие нейросетевых технологий позволяет существенно упростить решение задач, которые ранее требовали колоссальных ресурсов. Вместе с этим появилось и большое количество как библиотек для разработчиков, так и готовых сервисов, предоставляющих функции высококачественного распознавания, анализа, синтеза голоса. Однако, как это часто и бывает, новые технологии быстро осваивают и злоумышленники, применяя их с целью нарушения целостности систем голосовой связи путём создания так называемых дипфейк-аудио [2]. Одним из методов эффективного купирования новых угроз может быть, в том числе, применение речевой подписи на основе фазовых характеристик голосовых вокализмов.

Речь и основные угрозы речевой информации

Многие учёные говорят о том, что именно устная речь стала одним из главных факторов развития человечества в прошлом, позволив обмениваться знаниями между индивидуумами, передавать достижения и жизненный опыт из поколения в поколение. Под речью учёные понимают исторически сложившуюся форму общения людей посредством языковых конструкций, создаваемых на основе определённых правил [3]. Базовым компонентом речи являются звуки, которые формируются в голосовом аппарате человека посредством колебаний, далее они переносятся по среде передачи, например воздухе, и только после этого колебания попадают на барабанную перепонку человека, где преобразуются

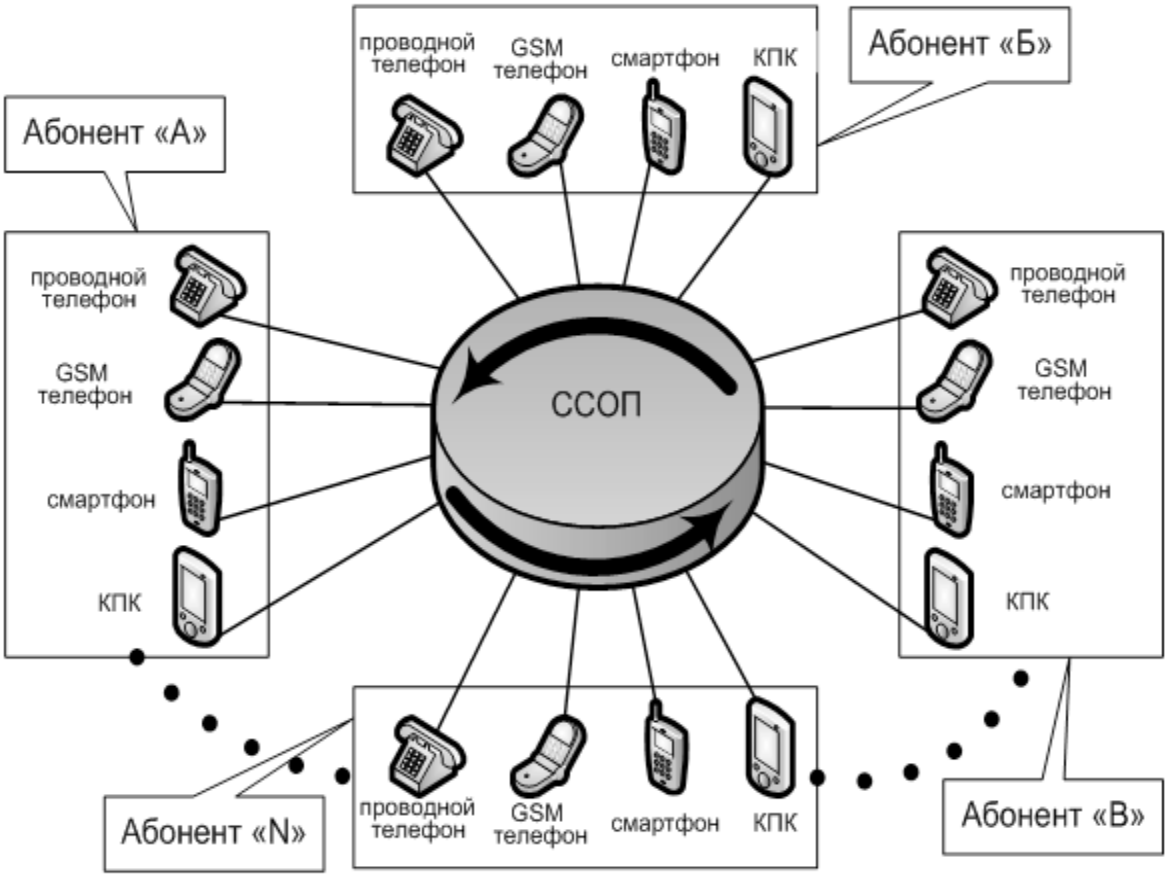

Рис. 1. Современная модель голосовой связи

в электрические сигналы, которые интерпретируются в головном мозге.

К основным характеристикам речевого сигнала относят энергетические характеристики (плотность энергии, интенсивность), временные (темп), амплитудные (громкость), и частотные (диапазон). Вместе с тем существуют ещё и фазовые характеристики речевых сигналов, которым уделяется мало внимания, но, однако они несут достаточно много полезной информации о личности диктора [4, 5].

В современном мире человечество перешло от моделей личного голосового взаимодействия к виртуальным пространствам путём организации различных распределённых информационных систем. Такие информационные системы намного сложнее традиционных, и могут включать в себя миллионы подвижных абонентов, единиц вышек связи, серверного и сетевого оборудования. Визуальная модель такой современной системы связи представлена на рисунке 1.

На системы речевой связи действуют все те же самые угрозы информационной безопасности, как и на любые другие информационные системы, а именно угрозы конфиденциальности, доступности и целостности.

При этом вместе с нарастающей сложностью систем связи соответственно увеличивается и поверхность возможных на них атак, а также количество уязвимостей. Например, если у нас есть простейшая система связи для проведения переговоров по конфиденциальным вопросам, включающая в себя двух людей и специальное помещение, то по сути единственным классом угроз для неё будут угрозы конфиденциальности — злоумышленник может перехватить переговоры с помощью специальных средств (например подслушивающих устройств), либо, если наше помещение совсем плохо защищено от технических каналов утечки информации, просто физически подслушать разговор из смежного помещения в необходимый момент времени (сделать это возможно и без каких-либо технических средств).

Вместе с тем, если эксплуатируется сложная разветвлённая система связи, то мы сталкиваемся с возможностью реализации всех классов угроз информационной безопасности — и угроз конфиденциальности, и угроз целостности, и угроз доступности. К угрозам доступ-

Рис. 2. Методы защиты речевой информации

ности речевой информации можно отнести, например, поломку сетевого оборудования. В таком случае связь будет недоступна, абонент не сможет дозвониться и провести необходимые переговоры. Другой причиной недоступности может служить заражение серверов вредоносным программным обеспечением, в результате чего их нормальная работа может быть нарушена. Угроза конфиденциальности может быть реализована путём перехвата речевой информации. Угрозы нарушения целостности подробно будут рассмотрены ниже, однако отметим, что к таким угрозам относят несанкционированное изменение аудиоинформации или нарушение её аутентичности. На рисунке 2 представлены основные методы защиты речевой информации от угроз целостности, доступности и конфиденциальности.

В последнее время внимание злоумышленников сконцентрировано на возможности нарушения целостности речевой информации. Новые современные алгоритмы, построенные на машинном обучении и нейросетевых технологиях, а также стремительное увеличение вычислительной мощности как домашних компьютеров, так и серверов позволило значительно облегчить задачу анализа и синтеза искусственной речи. Таким образом технологии, которые ранее были недоступны широкой публике, стали обыденными. Вместе со снижением порога вхождения данную область смогли использовать и среднестатистические злоумышленники, хотя ранее атаки на целостность речевой информации были под силу лишь государственным структурам.

Примером нарушения целостности голоса с помощью нейросетевых технологий может служить случай, произошедший в Великобритании в 2019 году. Тогда злоумышленникам удалось построить модель речи одного из состоятельных клиентов банка, которого обслуживал личный менеджер. Построив модель речи и узнав телефон управляющего, злоумышленники позвонили ему от имени клиента, а тот, узнав голос, без лишних вопросов и проверок перевёл по личному указанию на сторонний нелегитимный счёт около $240'000 [6]. Аналогичный случай произошёл с одним из банков ОАЭ в 2020 году, тогда злоумышленникам удалось вывести порядка $400'000 [7].

Для купирования многих угроз целостности речевой информации, в том числе представленных выше, можно использовать технологию речевой подписи на основе фазовых характеристик голосовых вокализмов, о которой мы расскажем далее. Эта технология, с одной стороны, позволит увеличить точность распознавания голоса диктора, а с другой стороны встроить в графическое изображение сигнала некоторые биометрические признаки человека, такие как собственноручную подпись или изображение отпечатка пальца.

Построение фазограмм голосового сигнала

Для анализа и синтеза речи используют различные варианты математических моделей, каждая из которых имеет свои достоинства и недостатки, при этом идеальной модели, которая одинаково хорошо подходит для решения всех задач речеобработки и речепреобразования, не существует. Вместе с тем подавляющее большинство моделей не учитывают (отбрасывают) фазовые характеристики голосовых вокализмов из-за их неочевидной пользы, а также сложности вычислений. Авторы статьи считают, что именно внедрение фазовых характеристик в существующие модели позволит значительно масштабировать их сферу применения.

В качестве базовой модели в исследовании авторы предлагают рассмотреть описание голосового сигнала на основе синусоидальной модели Куатьери и МакАуэля [8]:

$$S_{\widetilde{SR}}(n) = \sum_{K=1}^{L} A_k \cos(\varphi_k + n\Omega_k),$$

в которой L — число синусоид (изменяется во времени), а $\widetilde{SR}$ обозначает синусоидальное представление.

Модель можно сделать более эффективной, если знать, что основное количество информации в голосовом сигнале сконцентрировано в низких частотах, тогда:

$$S_{\widetilde{HR}}(n) = \sum_{k=1}^{L(\Omega_0)} A_k \cos(n\Omega_0 k + \varphi_k),$$

в формуле $\widetilde{HR}$ обозначает гармоническое представление аудиосигнала, Ω_0 — частота основного тона, $L(\Omega_0)$ — количество гармоник заданной частоты.

Предложен следующий способ учёта вариативности акустического описания речевого сигнала. На вокализованных участках протяжённостью Δt, в точках анализа с шагом r, речевой сигнал может быть представлен суммой M его составляющих гармоник:

$$S_r(t) = \sum_{k=0}^{M-1} A_k \cos(\omega_k t + ц_k)$$

Для данного временного участка речи Δt и точки анализа r вводятся понятия вектора начальных фаз $ц_r$ и вектора приведённых начальных фаз $\overline{ц_r}$:

$$ц_r = \begin{bmatrix} ц_{M-1} \\ ц_{M-2} \\ \vdots \\ ц_1 \\ ц_0 \end{bmatrix}$$

$$\overline{ц_r} = \begin{bmatrix} ц_{M-1} - ц_0 \\ ц_{M-2} - ц_0 \\ \vdots \\ ц_1 - ц_0 \\ 0 \end{bmatrix}$$

Отмечено, что в качестве опорной начальной фазы в $\overline{ц_r}$ можно взять фазу любой гармонической составляющей речевого сигнала, а не только $ц_0$. Такое приведение фаз всех имеющихся на анализируемом участке гармоник к одной опорной необходимо для снятия неопределённости, связанной с выбором точки начала отсчёта при выполнении процедур анализа.

Предлагается метод нахождения вектора приведённых начальных фаз. Устранив из изначальной формулы $S_r(t)$ амплитуду гармоник, получаем следующее описание данного вокализованного участка речи протяжённостью Δt:

$$\overline{S_r(t)} = \sum_{k=0}^{M-1} \cos(щ_k t + ц_k)$$

Один из моментов времени t_0на отрезке анализируемого вокализованного участка речи принимается за начало отсчёта. Далее строится система уравнений

$$\begin{cases} \overline{S_r(t_0)} = \sum_{k=0}^{M-1} \cos(\text{щ}_k t_0 + \text{ц}_k) \\ \overline{S_r(t_0 + r)} = \sum_{k=0}^{M-1} \cos(\text{щ}_k (t_0 + r) + \text{ц}_k) \\ \vdots \\ \overline{S_r(t_0 + (M-1)\cdot r)} = \sum_{k=0}^{M-1} \cos(\text{щ}_k (t_0 + (M-1)\cdot r) + \text{ц}_k) \\ \overline{S_r(t_0 + M\cdot r)} = \sum_{k=0}^{M-1} \cos(\text{щ}_k (t_0 + M\cdot r) + \text{ц}_k) \end{cases}$$

Формула 1

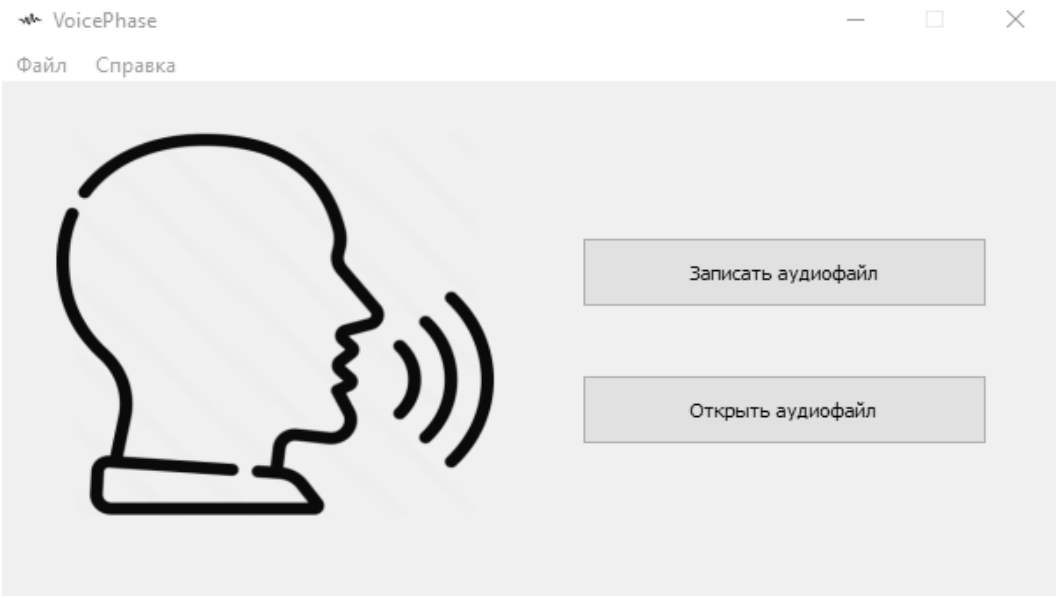

Рис. 3. Начальное окно программы в среде ОС Microsoft Windows

размерностью равной числу гармоник M, находящихся в интервале анализа (формула 1).

Результатом решения этой системы являются векторы $\{\text{cosц}_r\}$ и $\{cos\overline{\text{ц}_r}\}$. Причём последний вектор получается посредством несложного пересчёта из первого вектора. Процедура решения системы повторяется на всём протяжении вокализованного участка Δt для более точного определения вектора косинусов приведённых начальных фаз $\{cos\overline{\text{ц}_r}\}$.

Выражение $\{cos\overline{\text{ц}_r}\}$ может приниматься за эталонное описание в системах верификации. Аналитическая часть процесса верификации будет состоять из сравнения эталонного вектора косинусов приведённых начальных фаз с аналогичным вектором, вычисляемым из анализируемого звука по заданной мере со степенью точности, отмеченной ранее.

Внедрение биометрических характеристик в графический формат речевого сигнала

Для автоматизированного вычисления фазовых характеристик и анализа голосовых вокализмов авторами был разработан специализированный мультифункциональный программный комплекс, который позволяет не только проводить анализ, но и совершать различные

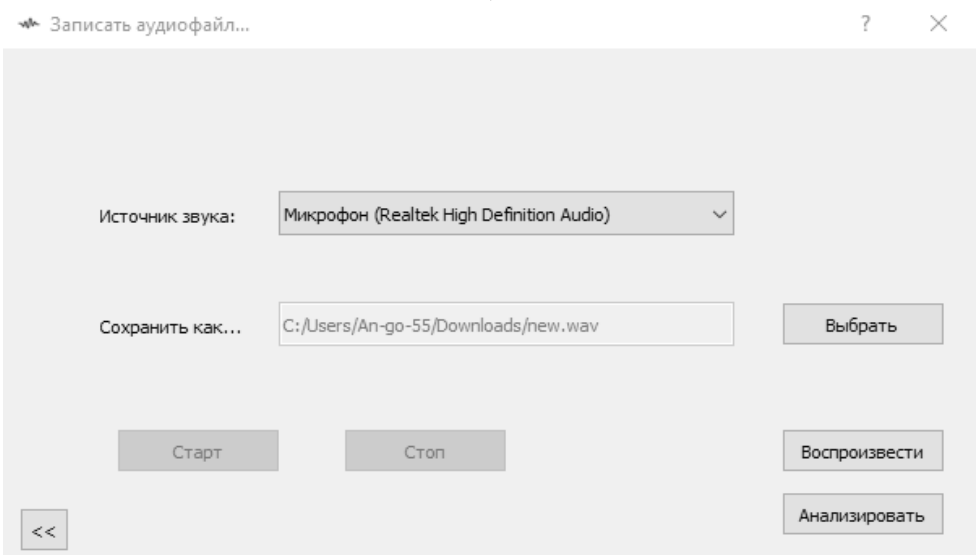

Рис. 4. Окно записи аудиофайла в среде ОС Microsoft Windows

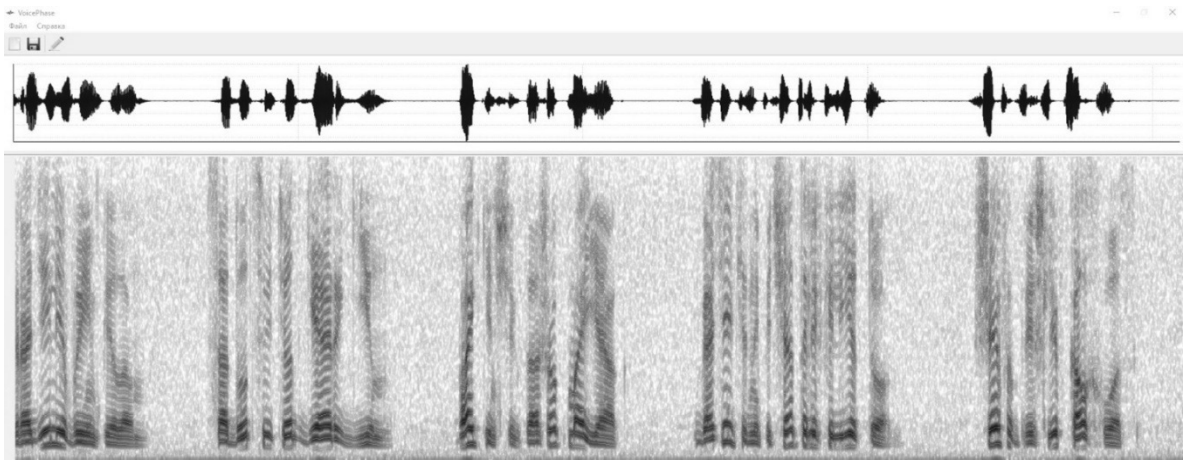

Рис. 5. Основное рабочее окно программы с анализом аудиофайла

звуковые преобразования (например, шумоочистку), строить фазограммы, а также включать в них избыточную информацию, в том числе биометрическую.

Программа разработана на языке программирования C++ с использованием кроссплатформенного фреймворка Qt 5.15.2, который позволяет запускать программное обеспечение в различных операционных системах — как в настольных (Windows, Linux, MacOS), так и в мобильных (Android, iOS), с соответствующей адаптацией интерфейса к размеру и разрешению дисплея [9, 10]. Скриншот стартового экрана программы представлен на рисунке 3. В основе вычисления фазовых характеристик голосовых вокализмов, и, соответственно, построения на их основе фазограмм, находятся, в том числе, методы и алгоритмы, представленные ранее в данной статье.

Алгоритм работы специалиста с программой выглядит следующим образом:

- Эксперт в главном окне (рисунок 3) может либо создать новую аудиозапись определённого зафиксированного формата для последующего анализа, либо же открыть имеющуюся запись голоса;
- В случае, если эксперт на предыдущем шаге выбрал «Записать аудиофайл», то открывается новое окно программного обеспечения «Диктофон» (рисунок 4). В данном окне эксперт может выбрать источник аудиосигнала, а также название нового файла. Отметим, что файл кодируется заранее определённым алгоритмом на основе Waveform Audio File Format (.WAV), который оптимизирован для дальнейшего анализа про-

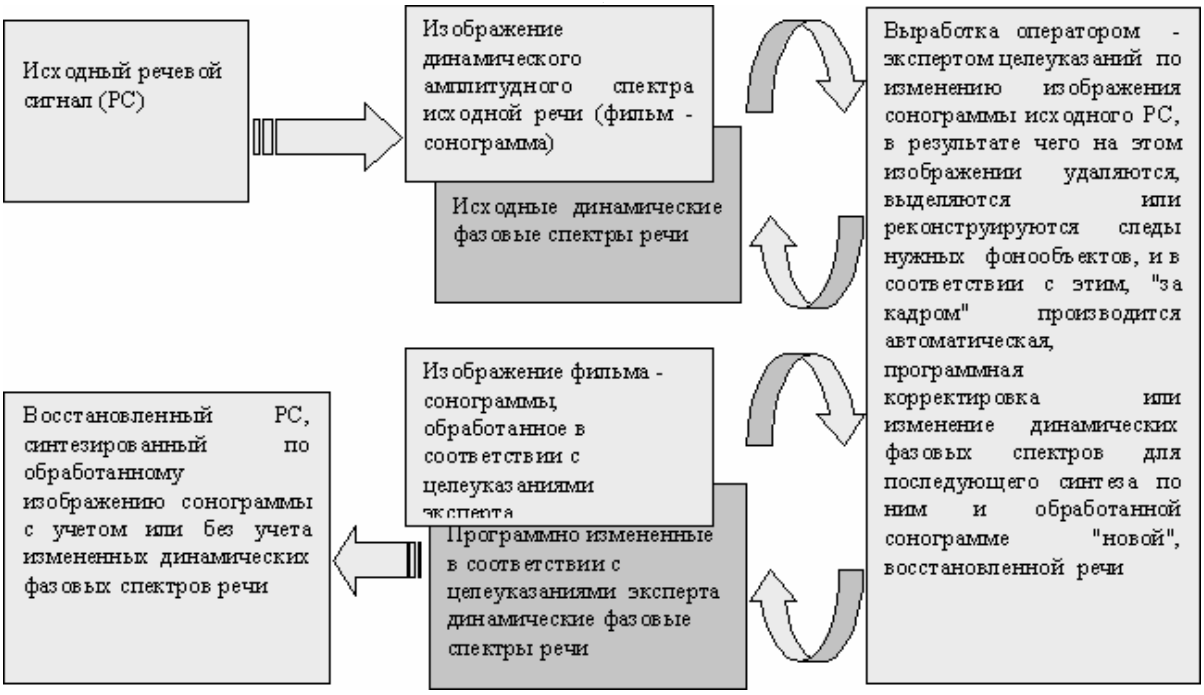

Рис. 6. Диаграмма перехода «Речевой сигнал — графический файл — речевой сигнал»

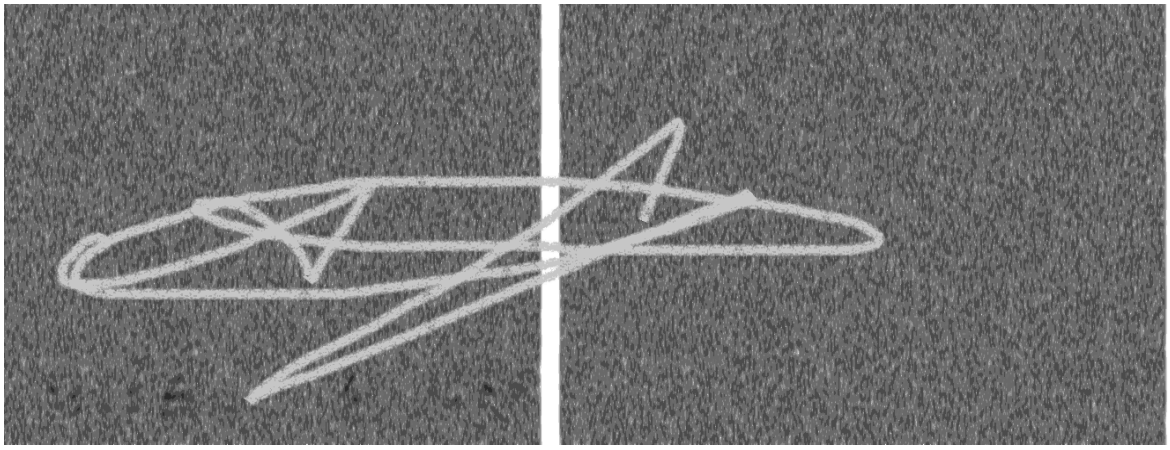

Рис. 7. Пример фазограммы с наложением рукописной подписи

граммой, вместе с тем воспроизвести аудиофайл можно любым современным аудиопроигрывателем. После записи аудиофайла эксперт может его прослушать, приступить к его анализу, либо же записать новый аудиофайл;

- В случае, если в главном окне эксперт выбрал «Открыть аудиофайл», то открывается новое окно программы. В окне эксперт выбирает, какой из заранее записанный файлов он хочет анализировать, после загрузки его можно воспроизвести, перейти к анализу, либо же загрузить другой файл;
- Основное окно программы, изображённое на рисунке 5, представляет собой рабочую область, поделённую на три зоны. В верхней части окна представлена панель инструмента, которые можно применить к аудиосигналу, в средней части окна можно наблюдать построенную осциллограмму загруженного звукового файла, в нижней части окна располагается его фазограмма.

Фактически представленное программное средство может обеспечить переход «Речевой сигнал — графиче-

ский файл — речевой сигнал» (образный анализ), а диаграмма такого перехода представлена на рисунке 6 [11].

Фазограмма является многомерной визуализацией звука (в данном случае речи), поскольку кроме частоты, а также времени в ней в градациях серого выражены ещё мощность и фаза. Подобный формат представления кажется удобным, поскольку в него можно встроить биометрическую информацию о говорящем, например рукописную подпись или изображение отпечатка пальца. Для этого в программу встроены возможности по редактированию фазограммы — можно дорисовывать и удалять соответствующие линии звукового сигнала, тем самым модифицируя его, а также синтезировать абсолютно новый звуковой файл «с нуля».

В качестве примера включения биометрии на рисунке 7 представлена фазограмма с наложением рукописной подписи говорящего. Таким же образом в фазограмму можно встроить, например, изображение отпечатка пальца диктора.

Таким образом внедрение речевой подписи на основе фазовых характеристик голосовых вокализмов позволит купировать многие современные угрозы целостности речевой информации. Кроме того, по мнению авторов, фазовые характеристики могут найти широкое применение и во многих других областях защиты речевой информации — например в целях анализа защищённости специального выделенного помещения от технических каналов утечки информации, или для уточнения алгоритмов искусственного синтеза речи.

Заключение

Таким образом в условиях увеличивающихся рисков нарушения целостности речевых записей становится необходимым разрабатывать и внедрять в практику использования и новых средств защиты. Одним из таких средств может выступать внедрение речевой подписи в виде биометрических данных в графические форматы изображения голоса, полученные благодаря вычислению фазовых характеристик голосовых вокализмов и построения соответствующих фазограмм.

Вместе с тем фазовые характеристики могут найти широкое применение и во многих иных областях защиты речевой информации, начиная от увеличения точности моделей распознавания диктора и заканчивая разработкой новых методик по оценке защищённости выделенного помещения от технических каналов утечки информации.

ЛИТЕРАТУРА

1. Федоринин М. Они нас слышат: куда развиваются речевые технологии? // Электронный ресурс https://www.forbes.ru/ tekhnologii/331035-oni-nas-slyshat-kuda-razvivayutsya-rechevye-tekhnologii (дата обращения: 24.04.2022 года).
2. Немцева М. Невооруженным ухом: как аудиодипфейки делают из мошенников крупных боссов // Электронный ресурс https://iz.ru/1209251/mariia-nemtceva/nevooruzhennym-ukhom-kak-audiodipfeiki-delict-iz-moshennikov-krupnykh-bossov (дата обращения: 24.04.2022 года).
3. Ахманова О.С. Словарь лингвистических терминов. М.: КомКнига, 2007. 607 с.
4. Дворянкин С.В., Уленгов С.В., Устинов Р.А., Дворянкин Н.С., Антипенко А.О. Системное моделирование речеподобных сигналов и его применение в сфере безопасности, связи и управления // Безопасность информационных технологий. 2019. Т. 26, № 4. С. 101–119.
5. Оппенгейм А.В. Применение цифровой обработки сигналов. М.: Мир.— 1980.— 552 с.
6. Stupp C. Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case // Электронный ресурс https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 (дата обращения: 24.04.2022).
7. Quach K. Bank manager tricked into handing $35m to scammers using fake 'deep voice' tech // Электронный ресурс https://www.theregister.com/2021/10/16/ai_in_brief/ (дата обращения: 24.04.2022 года).
8. McAulay R.J., Quatieri T.F. Speech analysis/Synthesis based on a sinusoidal representation // Article in IEEE Transactions on Acoustics Speech and Signal Processing. 1986. ASSP-34(4). pp. 744–754.
9. Прата С. Язык программирования C++: лекции и упражнения. М.: Вильямс, 2017.— 1248 с.
10. Шлее М. Qt 5.10: профессиональное программирование на C++. Санкт-Петербург: БХВ-Петербург.— 2018.— 1072 с.
11. Дворянкин С.В., Дворянкин Н.С., Устинов Р.А. Развитие технологий образного анализа-синтеза акустической (речевой) информации в системах управления, безопасности и связи // Безопасность информационных технологий.— 2019.— Т. 26.— № 1.— С. 64–76.

Журнал «Современная наука: актуальные проблемы теории и практики»