БИОМЕТРИЧЕСКИЕ ПРИЗНАКИ В ПАРАМЕТРАХ РЕЧЕПОДОБНЫХ СИГНАЛОВ ДЛЯ АУТЕНТИФИКАЦИИ ПОЛЬЗОВАТЕЛЯ В СИСТЕМАХ ГОЛОСОВОГО ВВОДА И УПРАВЛЕНИЯ

BIOMETRIC FEATURES IN THE SETTINGS OF THE SPEECH-LIKELY SIGNALS FOR USER AUTHENTICATION IN VOICE DRIVING SYSTEMS

S. Dvoryankin
N. Dvoryankin

Summary. Voice communications have been and remain one of the most common and convenient ways of human communication and remote human-machine interface. In recent years, there have been many reports of the successful creation and application of computer technology of speech cloning, with which you can successfully attack the system of speech control and automatic speech recognition, built-in information systems for various purposes. In this regard, the creation of methods for modeling speech-like signals with specified properties, used as specific audio markers, to confirm the authenticity of speech commands and messages is very important. Moreover, often in the process of remote interaction requires confirmation of the presence of a particular live subscriber with the verification of his authority, and not a robot acting in his place. The paper shows the possibility of using as identification audio markers, biometric characteristics of the legitimate user of protected resources, which in the form of graphic information (images) can be inserted into the spectral-time parameters of speech-like signals transmitted together with the passphrase (voice-key) via speeech communication channels during the negotiations of the user with the employee of the company, providing remote access to the protected resource.

Keywords: speech information protection, image analysis-synthesis, speech signal, binarization of spectrogram images, short-term Fourier transform.

Дворянкин Сергей Владимирович
Д.т.н., профессор, РГУ нефти и газа (НИУ) имени И. М. Губкина
s_dvrn@mail.ru
Дворянкин Никита Сергеевич
Аспирант, Национальный исследовательский ядерный университет «МИФИ»
nik.dvrn@gmail.com

Аннотация. Голосовые коммуникации были и остаются одним из распространенных и удобных способов человеческого общения и удаленного человеко-машинного интерфейса. В последнее время появилось много сообщений об успешном создании и применении компьютерных технологий речевого клонирования, с помощью которых можно успешно атаковать системы речевого управления и автоматического распознавания речи, встроенные в информационные системы различного назначения. В этой связи создание методов моделирования речеподобных сигналов с заданными свойствами, используемых как специфические аудиомаркеры, для подтверждения подлинности речевых команд и сообщений представляется весьма актуальным. Более того, часто в самом процессе удаленного взаимодействия требуется подтверждение присутствия конкретного живого абонента с верификацией его полномочий, а не робота, выступающего вместо него. В работе показана возможность использования в качестве идентификационных аудиомаркеров, биометрических признаков легитимного пользователя защищаемых ресурсов, которые в виде графической информации (образов) могут быть вставлены в спектрально-временные параметры речеподобных сигналов, передаваемым вместе с парольной фразой (словом) по голосовым каналам связи во время ведения переговоров пользователя с сотрудником компании, предоставляющим дистанционный доступ к защищенному ресурсу.

Ключевые слова: защита речевой информации, образный анализ-синтез, речевой сигнал, бинаризация изображений спектрограмм, кратковременное преобразование Фурье.

Введение

На сегодняшний день существует достаточно широкий набор методов и средств обработки и защиты значимой акустической (речевой) информации от НСД. В зависимости от требований безопасности и условий применимости используют тот или иной способ или их комбинации [1].

Отдельного внимания исследователей заслуживает технология образного анализа-синтеза (ОАС) акустических и речевых сигналов (РС) как оригинальная база моделирования существующих способов защиты и обработки акустической речевой информации (РИ) с возможностью оценки эффективности их работы, так и как платформа создания и использования ранее не применявшихся способов речепреобразования для их реализации в различных перспективных системах голосового управления, безопасности и связи, в которых циркулирует РИ [2–4].

Суть указанной технологии ОАС или по-другому технологии «звук-изображение-звук» состоит в преобразовании звукового или речевого сигнала в изображение узкополосной спектрограммы с применением к нему

развитых эффективных методов цифровой обработки изображений с последующим переходом от него к новой волновой форме звукового или речеподобного сигнала (РПС) с требуемыми характеристиками [2]. Понятно, что в таком случае на тех же базовых элементах и принципах можно реализовать «зеркальное» преобразование «изображение-звук-изображение», которое также может найти свое применение в указанных системах.

Исследованию возможностей ОАС изображений спектрограмм РС в различных областях применения: маскировании и скремблировании, аудио стеганографии, кодировании и компрессии речи, нейтрализации помех и искажений, идентификации говорящего и т.п. посвящено множество работ [2–4], в которых описаны различные варианты успешного применения технологии ОАС за счет реализации алгоритмов синтеза звуковых сигналов с требуемыми свойствами на основе заданного изображения спектрограммы. В качестве базовых в исследованиях использовались полутоновые (в уровнях серого цвета) изображения с достаточной информационной избыточностью.

В данной работе была сделана попытка оценить возможности прямого и обратного ОАС применительно к решению задач противодействия фальсификации голосовых команд и значимых речевых сообщений, их изменения и подмены клонами. Для достижения цели необходимо решение следующих задач: собственно обнаружения признаков подделки по анализу изображения спектрограмм РС и противодействия фальсификации посредством внедрения в РС специального PIN-кода, маскирование РС, маркирования защищаемого РС цифровыми водяными знаками (ЦВЗ) в виде буквенно-цифровых вставок, личных фото и подписи в спектрально-временных описаниях речевых команд и сообщений.

Особый интерес использования в качестве указанных ЦВЗ вызывает оценка возможности использования в их качестве биопризнакам легитимного пользователя путем преобразования их образов в спектрально-временные параметры речеподобного сигнала (РПС), имеющего основные характеристики подобные натуральной речи, что позволяет передавать РПС по всем существующим каналам речевой связи.

Усовершенствованная система образного анализа-синтеза речевых и речеподобных сигналов

Современная система анализа-синтеза речи с использованием свойств слухового восприятия (учета работы органов слуха, эффектов частотного и временного маскирования, психоакустики и др.) и контурного анализа узкополосных спектрограмм, соответствующая узкополосной модели РС, предложенной в разных вариациях в [1, 4], представлена на рис. 1.

Исходный РС можно рассматривать как выход линейной системы, представляющей характеристики речевого тракта при поступлении на нее сигнала возбуждения от голосовых связок. Согласно такому представлению процесса речеобразования формируемый РС при длительности анализируемого фрейма речи до 40 *мс* (оптимально $6–8$ *мс*) может быть представлен как:

$$s(n) = \sum_{i=1}^{L} A_i e^{\frac{-n^2}{2\sigma}} \cos(\omega_i n + \varphi_i) + e(n) \qquad (1)$$

где n — номер временного отсчета; L — количество значимых синусоид; A_i — амплитуда i-й синусоиды; ωi — частота i-й синусоиды; φ_i — фаза i-й синусоиды, σ — эффективная ширина окна функции Гаусса, $e(n)$ — остаточный сигнал.

В таком виде исходный речевой сигнал можно рассматривать как суперпозицию синусоидальных узкополосных сигналов или вейвлетов Морле. Такое представление *(1)* можно распространить и на другие акустические сигналы.

В блоке анализа входной дискретизированный речевой сигнал $S(n)$, подвергаясь кратковременному преобразованию Фурье (КПФ), периодически от фрейма к фрейму взвешивается временным окном $W(n)$ (например, усеченным окном Гаусса или окном Хэмминга), вычисляется его спектр $|S(f)|$. На каждом спектральном срезе $|S(f)|$ осуществляется отбор пиков локальных максимумов главных синусоидальных компонент РС, наиболее подходящих с точки зрения наилучшего перцептуального качества синтезированной речи, её слухового восприятия.

На получаемом по развертке спектральных срезов изображении спектрограммы отрисовываются треки или контура этих пиков для каждой узкополосной составляющей (УС).

В блоке синтеза по положениям контуров выбранных пиков определяются частоты наиболее мощных синусоид, определяющих основное звучание РС, и их амплитуды. Оригинальные фазы синусоид определяются по действительной и мнимой компонентам спектра $S(f)$ на соответствующих найденных частотах или вычисляются искусственным путем [1, 4] по функции развития фазы, определяемой также по изображению (разверткам) амплитудного спектра $|S(f)|$ или изображению иного содержания.

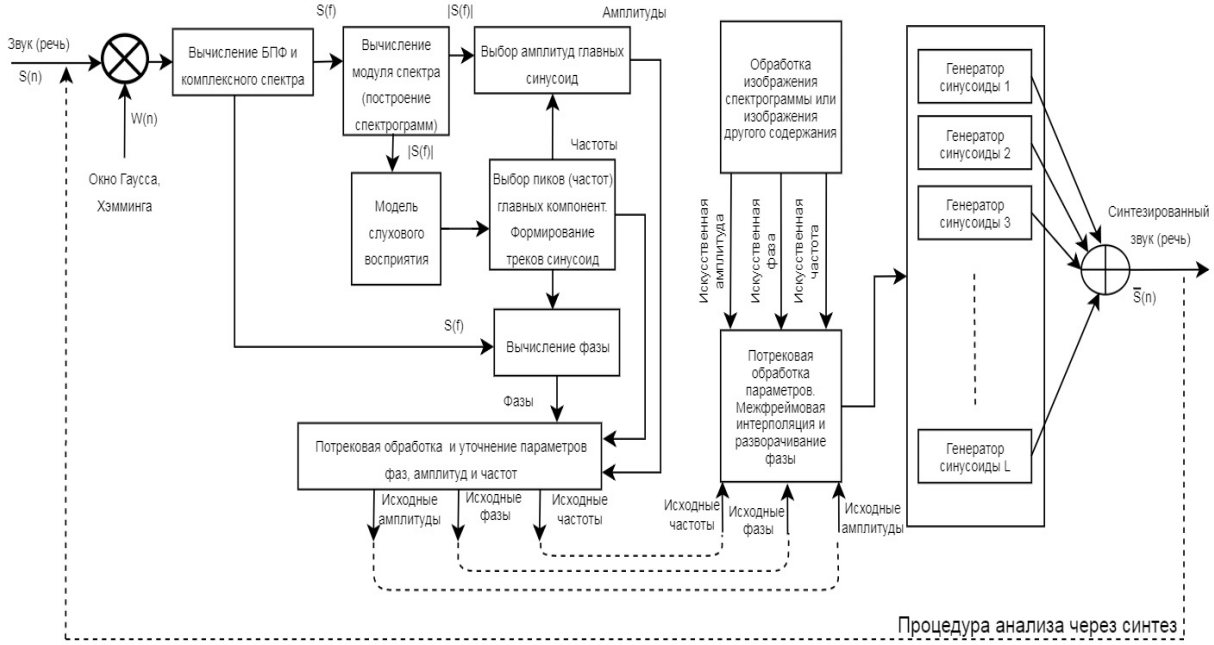


Рис. 1. Система образного анализа-синтеза речевого и речеподобного сигнала на основе узкополосной синусоидальной модели и свойств слуха.
(Fig.1. Image analysis and speech-likely signal synthesis system based on a sinusoidal narrowband model and hearing properties.)

Процесс синтеза новых речеподобных сигналов с заданными свойствами сводится либо к процедуре обратного КПФ для изменённого в зависимости от решаемой задачи спектрального среза с новым заданным пиковым подбором главных УС, либо к суммированию сгенерированных главных синусоидальных компонент с найденными для них в процессе анализа амплитудами, фазами и частотами (рис. 1).

При этом для получения в процессе синтеза приемлемого качества речи необходимо генерировать синусоиды, непрерывно изменяющиеся во времени. С этой целью применяется частотное упорядочивание синусоид и интерполяция их параметров от фрейма к фрейму [1, 4].

Система образного анализ-синтеза акустического (речевого) сигнала на основе синусоидальной узкополосной модели РС и учета свойств слуха, представленная на рис. 1, послужила основой для моделирования работы различных речепреобразующих устройств [2–4], когда на выходе анализирующей части системы (рис. 1) собирался ансамбль спектральных срезов модуля спектра (спектральные развертки), рассматриваемый в дальнейшем как изображение. На так получаемых графических образах РС определялись треки (линии) пиков локальных максимумов, по амплитудно-частотно-фазовым параметрам которых в блоке синтеза (рис. 1) генерировался новый речеподобный сигнал в соответствии с заданным порядком модифицированной спектрограммой.

Если изображение корректно рассчитанной спектрограммы не менялось, то синтезированный по ней РС и по волновой форме, и по звучанию практически совпадал с исходным. Для реализации различных процедур аудимаркирования значимого РС, подлежащего защите от подделки, изображение участков его спектрограммы подвергалось необходимым для этого трансформациям.

Построение и анализ полутоновых изображений спектрограмм РС на основе КПФ производится в анализирующей левой части системы ОАС (рис. 1). Заметим, что на всех далее приведенных изображениях спектрограмм по оси абсцисс задается время, по оси ординат — частота. В уровнях серого цвета — мощность на данной частоте в данное время. Максимальная — чёрный цвет, минимальная — белый. Максимальная частота на оси абцисс — $4\ КГц$, соответствовала половине частоты дискретизации РС в $8\ КГц$. Шаг анализа по времени, расстояние между столбцами спектральных срезов изображения спектрограммы — $8\ мс$.

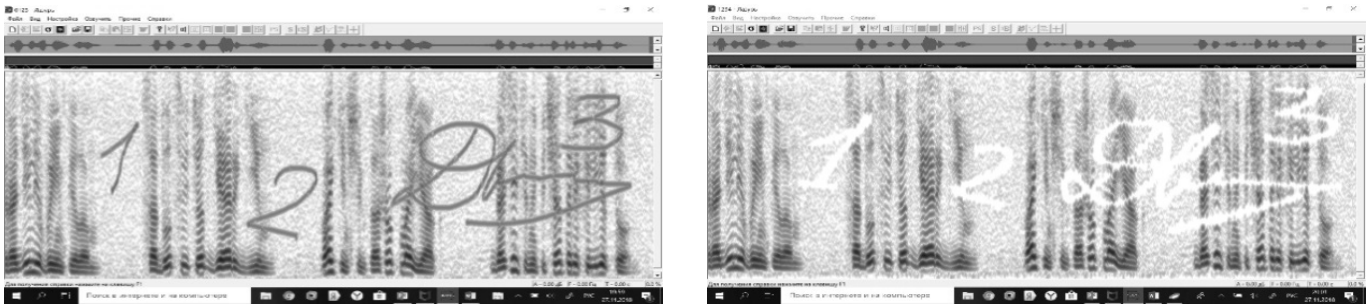


Рис. 2. Встраивание ЦВЗ в тело спектрограммы: а) «дорисовка» треков узкополосных составляющих; б) «продавливание» (стирание) треков с буквенно-цифровой и символьной информацией.
(Fig. 2. Embedding digital watermarks in the spectrogram body: a) "drawing" tracks of narrowband components; b) "pushing" (erasing) tracks with alphanumeric and symbolic information)

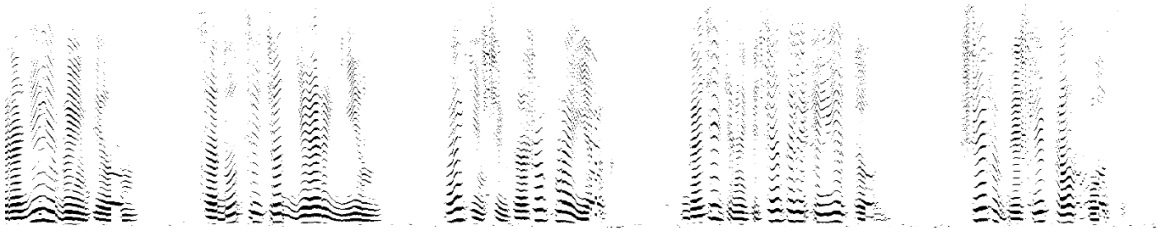


Рис. 3. Бинарное изображение спектрограммы речевого сигнала — бинарной речевой подписи.
(Fig. 3. Binary image of the speech signal spectrogram — speech signature)

Методы аудиомаркирования биопризнаками значимой речевой информации

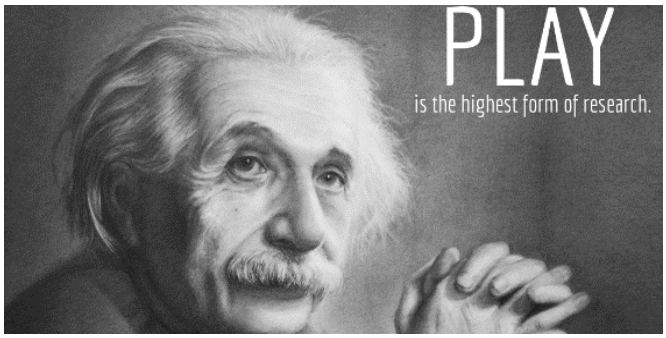
Как отмечалось, аудиомаркирование предполагает наличие в спектре значимого РС специального маркера или «цифрового водяного знака» (ЦВЗ), подтверждающего его связь именно с этим РС. Маркер проставляется на передающем конце канала голосовой связи во время отправки РС в паузах или в конце сообщения и выделяется для сравнения с записанным в кодовой книге уже при получении сообщения на приемном конце. Если сравнение происходит, значит РС не подвергся изменениям и подмене в процессе его передачи по каналам связи.

Возможно также встраивание ЦВЗ в тело спектрограммы значимого, подлежащего защите РС. Например, посредством «дорисовки» или стирания треков узкополосных составляющих, несущих буквенно-цифровую, символьную и иную информацию (см рис. 2).

Метод стирания оказался более эффективным поскольку его влияние на слуховое восприятие эксперта-аудитора практически не ощущалось. Дорисовка ЦВЗ, напротив, вносила мешающие восприятию свистки и помехи.

На основе использования системы ОАС (рис. 1) помимо метода внедрения в спектрограммы буквенно-цифровой информации (см рис. 2) были разработаны и исследованы следующие нижеперечисленные методы перевода образов биопризнаков в частотно-временные параметры речеподобных сигналов для «подписания» ими значимых РС в целях их защиты от модификаций и подмены. Это речь в виде бинарного образа в фоновых шумах в паузах или в конце значимого РС, бинарный образ отпечатка пальца, полутоновые или бинарные образы фото и рукописной подписи легитимного пользователя удаленного защищённого ресурса и др.

Действительно, помимо различных видов встраивания буквенно-цифровой и символьной информации в полутоновые спектрограммы значимых РС, можно встраивать в паузные и шумовые участки собственно бинарные спектрограммы РС, используемые в качестве ЦВЗ. Заметим, что бинарные и полутоновые корректно рассчитанные спектрограммы, получаемые по технологии ОАС, несут равнозначную информацию о разборчивости, узнаваемости, естественности и громкости исходного РС, являются взаимно обратимыми: из одной всегда можно получить другую без потери информативности.



а)

б)

Рис. 4. Фотопортрет — а) и его реализация в виде спектральной развертки осциллограммы звука-носителя — б)
(Fig. 4. Photo — a), and its implementation in the form of a spectral scan of the sound carrier oscillogram — b))

Бинарную спектрограмму РС (см рис. 3) также можно «продавливать» в фоновом паузном шуме или на вокализованных участках полутоновых спектрально-временных описаний значимого РС, по аналогии действий с буквенно-цифровой и символьной информацией

Заметим, что бинарные спектрограммы РС и отпечатка пальца крайне похожи и и используют практически одинаковые подходы к их обработке и защите методами ОАС.

Рассмотрим другие способы аудиомаркирования значимых РС биопризнаками легитимного пользователя.

Так на основе выше рассмотренных базовых элементов и принципах прямого ОАС можно реализовать его обратное «зеркальное» преобразование: «изображение-звук-изображение». Тогда изображение любого вида и содержания может быть преобразовано в звуковой речеподобный сигнал, спектрограмма которого будет похожа, например, на исходное фото, как это показано на рис. 4.

Звуковые сигналы со спектральными характеристиками в виде различных фотоснимков, образов могут найти своё применение в виде специфических аудиомаркеров, которыми можно, например, подтверждать, подлинность голосовых команд в системах речевого управления или доверенность среды передачи голосовых сообщений.

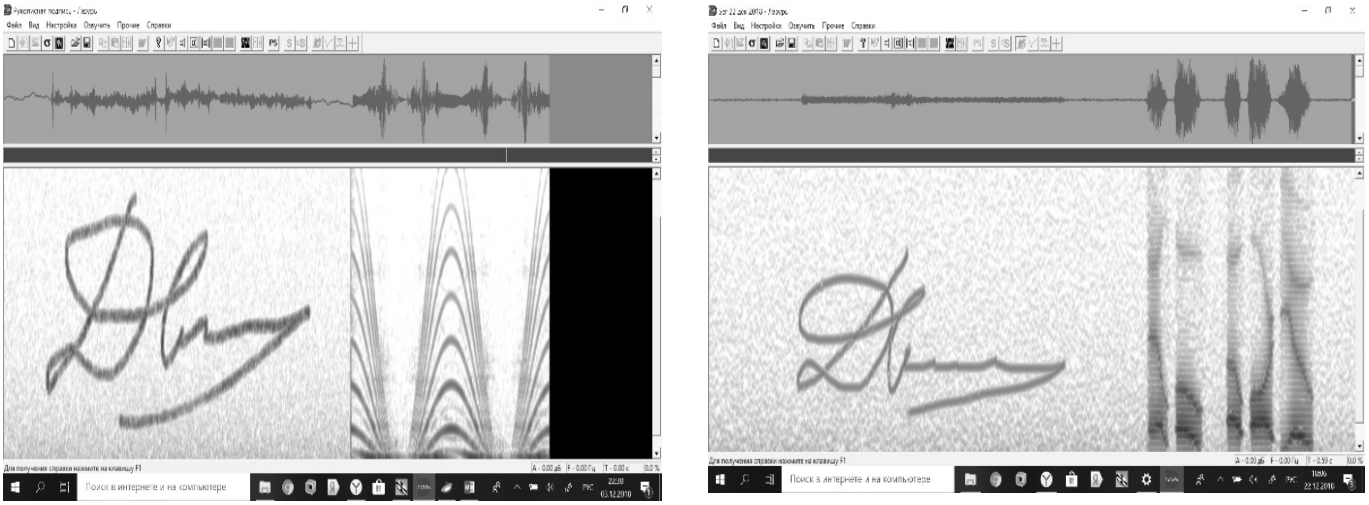
Аудиомаркеры можно использовать в системах голосовой аутентификации используя речеподобный сигнал с заданным селфи- фото, в качестве парольной фразы в процессе непрерываемого голосового общения с поставщиками необходимых мобильных услуг или call-центром.

Еще более интересные возможности открывает создание и применение в системах дистанционной аутентификации пользователя защищенного информационного ресурса речеподобного образа его рукописной подписи (рис. 5).

Причем в этом случае для сравнения с эталоном можно использовать не только статическое изображение самой рукописной подписи в спектре звука-носителя, но и динамические характеристики ее написания, сохраненные в виде изменений параметров частоты основного тона и количества формант в дополнительно создаваемом образе речеподобного сигнала, соответствующем движению пера (пальца) на экране смартфона.

Так, движение пера (пальца) на экране смартфона «вверх-вниз» изменяет частоту основного тона на спектрограмме синтезируемого речеподобного сигнала (РПС) от 100 до 250 *Гц*. Движение «слева-направо» соответствует двум формантам в спектре РПС с глобальными максимумами в области 200 и 500 *Гц*. Обратное движение — «справа-налево» характеризуется добавлением в спектральный образ РПС третьей форманты с глобальным максимумом в 1500 *Гц*.

Силе нажима на кончик пера в параметрах речеподобного сигнала будет соответствовать движение формант на спектрограмме РПС также сверху-вниз или снизу-вверх.



а) б)

Рис. 5. Представление рукописной подписи в качестве спектрограммы речеподобного сигнала, его статических и динамических характеристик: а) в частоте основного тона; б) в количестве и направленности формант
(Fig. 5. Sonification of the handwritten signature as a spectrogram, static and dynamic characteristics of the speech-like signal: a) in the frequency of the pitch tone; b) in the quality and direction of the formants)

Такая техника сонификации рукописной подписи как типовой частный случай применения ОАС проста в реализации и весьма перспективна.

Наличие для взаимного сравнения трёх образов рукописной подписи: один, как эталон-образец создается при первоначальном личном присутствии пользователя в службах сервиса и безопасности, а два других определяются из статических и динамических характеристик спектра РПС, передаваемого по каналу связи, существенно повышают надежность аутентификации удаленного пользователя защищенного информационного ресурса.

РПС с образцами подписей также может рассматриваться как некая парольная фраза и тогда можно ещё усилить надежность аутентификации, добавив к существующим хорошо отработанным процедурам сравнения рукописных подписей такую же хорошо отработанную процедуру и решения голосового паролирования и сравнения от известных российских и зарубежных производителей.

Возможны и другие не менее интересные приложения обратного преобразования «изображение-звук-изображение». Например, совмещение в системах удаленной аутентификации одновременно сразу двух вышеописанных приложений инверсного ОАС: речеподобных фото лица или отпечатка пальца и самой подписи.

Заметим, что помимо способов аудиомаркирования значимых РС возможно также их асинхронное маскирование от перехвата, распознавания и использования в своих целях злоумышленником. Здесь следует упомянуть два основных способа, использующих технологию ОАС.

Первый — «понижение уровня информативного сигнала», предполагает, например, скрытие бинарных спектрограмм значимых РС в шумах и помехах канала связи с общим уровнем мощности смеси, не превышающей уровень шума — носителя.

Второй — «повышение уровня сопутствующих помех», предусматривает закрытие значимого РС мощной речеподобной помехой, формируемой либо от внешнего источника, либо из самого передаваемого РС.

На приемном конце получателя значимые РС с использованием технологии ОАС демаскируются и затем используются в процессах голосового распознавания, информирования и управления.

Заключение

Работа посвящена развитию уже хорошо зарекомендовавшей себя в решении задач защиты и обработки речевой информации технологии образного анализа-синтеза РС.

В рамках указанной технологии исследован подход защиты и установления подлинности значимых сообщений и голосовых команд, заключающийся во внедрении в них на передающем конце канала связи и извлечении на приемном конце идентификационных данных и цифровых водяных знаков (ЦВЗ) или аудиомаркеров, находящихся в спектрально-временных описаниях РС (изображениях их спектрограмм). Внедрение аудиомаркеров практически не искажает качественные характеристики защищаемого ими РС. Его разборчивость и узнаваемость остаются на прежнем уровне.

В качестве таких аудиомаркеров в исследовании применялись преобразованные в спектрально-временные параметры речеподобного сигнала биопризнаки легитимного пользователя (бинарное изображение речевой подписи, бинарные и полутоновые изображения отпечатка пальца, фото лица, рукописной подписи с динамическими характеристиками ее написания), что не влияло на качество слухового восприятия речевого сигнала контейнера.

Такие ЦВЗ, подтверждающие подлинность речевого сообщения или голосовой команды переведенные в спектрально-временные характеристики речеподобного сигнала, рекомендуется добавлять в конец защищаемого РС, чтобы не нарушать качество его звучания и разборчивость.

Также использовались буквенно-цифровые вставки в изображения спектрограмм РС, подлежащего защите с последующим по ним синтезом нового защищенного РС. Причем использование при начертании на изображениях спектрограмм стирающих «следы» РС вставок (инструмент «ластик») оказалось предпочтительнее для сохранения речевой разборчивости чем использование «вставок», добавляющих (инструмент «карандаш»). Звучание последних в синтезируемой смеси РС плюс ЦВЗ воспринималось как фоновый свист.

Заметим, что в принципе, в качестве ЦВЗ можно использовать изображение любого содержания, о котором осведомлены участники речевого обмена «человек-человек» и «человек-машина» и которое затем на приемном конце связи будет сравниваться с изображением, заранее занесенным в кодовую книгу и соответствующим данному РС или голосовой команде. По результатам сравнения таких ЦВЗ принимается решение о подлинности или нет защищаемого ими РС.

Использование рукописной речевой подписи в качестве маркирующего РПС видится наиболее предпочтительным среди других биопризнаков, используемых в качестве ЦВЗ, поскольку в спектрально-временные описания РПС самой рукописной подписи могут быть вставлены ее статические (образ) и динамические (нажим, направление движения) характеристики. В результате РПС рукописной подписи может быть предъявлен к идентификационной проверке и как образец подписи и как уникальная специфическая парольная фраза, для которой уже применяются готовые решения для идентификации голосовых паролей. Таким образом, можно считать, что РПС рукописной подписи содержит в себе все необходимые данные для двухфакторной идентификации как защищаемого РС, так и самого пользователя а не его «аватара». Понятно, что такая идентификация обладает более высокой надежностью, чем просто парольная фраза (слово) и образец подписи по отдельности.

Также в качестве ЦВЗ рекомендуется использовать бинарные образы защищаемого РС, внедряемые в вокализованные, шумовые или в паузные участки его «тела» без нарушения РР.

Кроме использования в указанном приложении бинарные образы РС имеют большой потенциал для решения других задач защиты и обработки РС, в частности сжатия-восстановления, маскирования и др.

Так, помимо защиты собственно значимого РС от подмены, бинарные спектральные образы можно использовать для защиты бумажных и электронных документов от фальсификации и подделки, причем такие аудиомаркеры можно размещать на текстовом документе в явном и в скрытном виде, используя уже наработанные методы компьютерной стеганографии.

В традиционных приложениях компьютерной стеганографии бинарные изображения РС также могут оказаться предпочтительнее известных, поскольку их, например, гораздо проще встраивать в стегоконтейнеры по сравнению с полутоновыми и цветными изображениями разного содержания.

Бинарные образы РС также могут быть использованы в виде оригинальных цифровых водяных знаков для охраны авторских прав на аудиовизуальную и печатную продукцию.

На основе бинарного ОАС, реализуя сжатие речи через сжатие образов, возможно создание широкодиапазонного помехоустойчивого аудиокодека, работающего на скоростях от *1* до *64 Кбод* с плавной адаптацией к пропускной способности канала связи. С учетом использования голосовой базы данных конкретного диктора можно достичь нижней границы в *400–600 бит/с*, приближаясь вплотную к теоретическому пределу — *70 бит/с*.

ЛИТЕРАТУРА

1. Дворянкин С. В. Речевая подпись. М.: РИО МТУСИ. 2003. 184 с.
2. Дворянкин С.В., Нагорных И. М. К вопросу о технологии преобразования звук — изображение — звук. // Спецтехника и связь. 2013. № 1. С. 28–32.
3. Алюшин В.М., Дворянкин С. В. Технологии образного анализа в задачах цифровой обработки речевой информации. // Научная визуализация. 2013. Т. 5. № 3. С. 75–88.
4. McAulay R.J., Quatieri T.F, Speech Analysis/Synthesis Based on a Sinusoidal Representation, IEEE Trans. on Acoust., Speech and Signal Processing.— 1988. Vol. I. ASSP-34. P. 744–754.

Журнал «Современная наука: актуальные проблемы теории и практики»