УДК 621.391

Д. А. Лютова

ОСНОВНЫЕ ЗАДАЧИ И МЕТОДЫ ТЕХНОЛОГИЙ РАСПОЗНАВАНИЯ ГОВОРЯЩЕГО ПО ГОЛОСУ

Статья посвящена биометрическому способу установления личности – распознаванию говорящего по голосу. Развитие информационных технологий, постоянное увеличение потоков передачи данных по различным каналам связи, которые требуют защиты от несанкционированного доступа, а также широта круга практических приложений, в которых используются технологии идентификации и верификации по голосу, делают данную область столь привлекательной для теоретического и практического исследования. Цель данной статьи – рассмотреть основные задачи и методы распознавания дикторов. Задача автора – исследовать сущность систем распознавания по голосу, описать индивидуальные речевые характеристики говорящих, охарактеризовать основные методы распознавания говорящих по голосу, выявить основные проблемы технологий идентификации / верификации говорящих.

Ключевые слова: распознавание говорящего; идентификация; верификация; речевой сигнал; индивидуальные характеристики; внутридикторская изменчивость; «отпечаток» голоса; имитация голоса; маскировка голоса.

Распознавание людей по голосу в наше время является одним из важных приложений речевых технологий. Это обусловлено развитием информатизации общества, постоянным наращиванием потоков передачи данных по различным каналам связи, которые требуют защиты от несанкционированного доступа, ростом значения автоматической обработки данных, а также нарастающим использованием автоматических речевых технологий. Системы распознавания говорящего по индивидуальным характеристикам вызывают большой интерес – как в научных, так и коммерческих кругах.

Распознавание устной речи относится к биометрическим способам установления личности наравне с установлением личности по отпечаткам пальцев, по сетчатке глаза или структуре генов. Одним из главных стимулов к исследованию процессов распознавания устной речи служит стремление к более достоверному опознаванию личности, так как уникальность биометрических характеристик обеспечивает более высокую надежность идентификации, кроме того, биометрические свойства организма не могут быть утрачены или забыты.

Еще одним фактором интереса к опознаванию личности по голосу является широкий спектр возможных приложений, в которых

используются данные технологии. Технологии и средства идентификации по голосу применяются в ряде областей, связанных с задачами обеспечения безопасности, таким как контроль за физическим доступом к вычислительным системам, банковским счетам, различным устройствам, служебным помещениям, каналам связи; автоматическим контролем за выполнением деловых операций, совершаемых по телефонным каналам (банковские сделки, запрос о состоянии банковских счетов, оплата различных счетов или бронирование авиабилетов). Существуют и другие области, где установление личности по голосу очень важно. В качестве примера можно назвать криминалистическую экспертизу: доказательства в суде, анализ записей переговоров при различных аварийных ситуациях, анализ записей переговоров при их санкционированном прослушивании [4].

Системы распознавания говорящего по голосу строятся на принципе отличия голосов разных людей друг от друга. Распознавание говорящего – это процесс определения на основе характеристик речевого сигнала, принадлежит ли данное высказывание конкретному говорящему. Системы распознавания человека по голосу подразделяются на два вида: верификацию и идентификацию говорящего. При верификации полученный образец голоса диктора сравнивается с эталоном и устанавливается его идентичность данному эталону. Необходимо принять одно из двух возможных решений: является ли диктор тем, за кого он себя выдает, или не является. Для вынесения такого решения используется совокупность параметров, содержащих необходимую информацию об индивидуальности диктора и измеряемых по одной или нескольким фразам. Измеренные значения сравниваются с аналогичными параметрами эталонных образцов подлежащего опознанию диктора.

Задача идентификации диктора существенно отличается от задачи верификации. «Задача заключается в выделении из общей совокупности *M* дикторов той личности, которая по своим голосовым характеристикам, заэталонированным заранее будет совпадать с характеристиками голоса принадлежащего опознаванию <…> Для идентификации неизвестной личности запись его речи должна поочередно сравниваться с эталонными записями всех *M* возможных дикторов, в результате чего процесс идентификации оказывается зависимым от значения *M*. От числа *M* оказывается при этом зависимым <…> величина вероятности ошибки идентификации» [11, с. 110].

Существует еще один «важный параметр устной речи, используемый в системах распознавание – это контролируемости речевых оборотов. В тех прикладных системах, где говорящий субъект желает быть опознанным и, следовательно, способствует этому, используются лексемы фиксированного текста, или тексто-зависимые речевые обороты. В тех же случаях, когда подобный контроль невозможен (по причине противодействия говорящего) или должен проводиться незаметно для него, требуется применение лексем произвольного текста, т. е. тексто-независимых оборотов речи. Качество распознавания в тексто-зависимых приложениях будет всегда выше, чем в тексто-независимых, поскольку в первом случае возможна более тщательная калибровка входных лексем с помощью идентичного речевого материала, а возможность контроля за входными сообщениями нередко распространяется на самого говорящего и его окружение. В тексто-независимых приложениях отсутствие таких возможностей в какой-то мере компенсируется тем, что распознавание часто производится по более длинным речевым фрагментам, а это позволяет улучшать качество идентификации» [4, с. 131].

Несмотря на внешнее сходство задач идентификации и верификации говорящего, эти системы отличаются друг от друга по областям применения, речевому материалу, который используется в этих системах, а также параметрам, по которым сравниваются голосовые образцы с эталонами.

Очевидно, что в системах верификации возможен индивидуальный подход к подбору парольной фразы для каждого диктора с учетом его артикуляционных особенностей. В данных системах также возможна стабилизация манеры произношения заданных слов пользователем, в том случае, когда контрольная фраза подсказывается системой устно. Также большу́ю роль в точности распознавания играет то, что пользователь сам способствует правильному распознаванию.

Для систем идентификации диктора одну из главных трудностей представляет невозможность контроля за всеми аспектами задачи, а именно, использование тексто-независимых оборотов речи, а также то, что подозреваемый всеми способами стремится помешать правильному опознанию. В этих системах, помимо использования сегментной информации (отдельных фонем и лексем), необходимо подключать к распознаванию и информацию супрасегментную (ритм, тембр, мелодика, временны́е характеристики речи, систему ударений

и т. д.), что в большой степени усложняет задачу правильного распознавания говорящего.

Подходы к задачам опознавания человека по голосу делятся на субъективные, в которых решение принимают эксперты, и объективные, где целью является исключение человеческого фактора из процесса принятия решения. К субъективным методам относятся распознавание дикторов на слух и визуальное сравнение спектрограмм, объективным способом оценки является полная автоматизация систем распознавания [9].

Суть метода распознавания дикторов на слух заключается в том, что группе экспертов предоставляют для прослушивания речевой материал, который необходимо сравнить с эталонными речевыми образцами и принять решение об идентификации личности. Распознавание дикторов на слух основано на том, что человеческое ухо – очень тонкий инструмент, генетически предрасположенный к распознаванию различных звуков. «Слух человека, не обладая способностью дать точную количественную оценку тем или иным характеристикам воспринимаемого речевого сигнала тем не менее может произвести глобальное решение относительно определенных свойств речи и голоса и указать на ряд трудноуловимых его особенностей» [11, с. 41].

Эксперименты показали, что правильность распознавания говорящего на слух зависит от многих факторов, таких как представленность и качество речевого материала, числа дикторов контрольной группы, промежутком времени между моментом получения эталона и моментом сравнения с голосом диктора и многих других. На качестве распознавания в сильной мере сказываются как акустические условия, так и состояние говорящего и слушающего. В этой области проводился целый рад исследований, направленный на установление переменных, влияющих на качество распознавания голосов на слух и перцептивных основ распознавания дикторов. К примеру, эксперименты показали, что идентификация голосов оказывается более точной, если образец голоса и сам идентифицируемый голос записаны одинаковым способом (например, прямая запись на пленку и запись через телефонные линии). Точность понижается, когда идентифицируемый голос и голосовые образцы записаны разными способами [9].

Достоинство этого метода заключается в том, что слуховое восприятие допускает более широкий выбор условий и стратегий распознавания. Например, слушающие могут узнавать говорящих по голосам,

не прибегая к сопоставлению в явном виде каких-либо двух фрагментов устной речи, а просто будучи знакомыми с этими голосами.

К недостаткам можно отнести то, что в зависимости от условий и задач распознавания способность эксперта изменяется в широких пределах. Эксперименты также показали, что два образца речи, воспринимаемые на слух как одинаковые, сильно отличаются по акустическим параметрам. Еще одним минусом являются предельные возможностями человеческого восприятия, способного одновременно контролировать ограниченное число параметров. Тем не менее данный метод можно применять в комбинации с другими, что способствует улучшению результата идентификации.

В качестве другого экспертного метода распознавания дикторов может быть назван метод визуального выявление сходства между говорящими путем сравнения видимых картин произносимых слов – «отпечатков» голоса. Этот метод привлек к себе внимание в 60-х гг. XX в., «после того как в лаборатории Белла был предложен новый подход к представлению спектрографических картин речи. Оригинальный способ получения так называемых контурных спектрограмм, напоминающих рельефные карты, позволяет представить произносимые слова и фразы в виде трехмерных изображений в координатах время – интенсивность – частота» [11, с. 55].

«В основе данного подхода лежат предположения, что: а) спектрографические структуры различных произнесений одних и тех же слов или звуков одним говорящим обладают релевантным сходством и б) речь разных говорящих на спектрограмме значительно отличается» [9, с. 13].

Метод установления личности по отпечаткам голоса имеет ряд существенных недостатков, которые не позволяют считать его результаты достаточным доказательством в суде. Этот метод не имеет четко разработанной объективной процедуры и остается лишь сопоставлением изображений, но самое главное, не существует определенных критериев, которые могут служить подтверждением сходства или различия спектрограмм, достаточных для принятия решения.

Однако идея о том, что голос имеет индивидуальный отпечаток, оказалась очень привлекательной для дальнейшей разработки. Экспертные методы идентификации дали большой толчок и теоретическую базу для развития автоматических методов распознавания личности по голосу.

К объективным методам распознавания следует отнести компьютерные системы идентификации / верификации, так как процесс может быть полностью автоматизирован и полностью исключать принятие решения экспертом.

Компьютерное распознавание диктора начинается с цифровой обработки речи. Речевой сигнал представляется с использованием таких методов цифровой обработки, которые сохраняют индивидуальные особенности диктора. Следующий этап – сравнение с имеющимися эталонными описаниями зарегистрированного числа дикторов в базе данных компьютера. Решение о распознавании принимается согласно сходству или различию сравниваемого образца с эталоном. Если в системах верификации входящий речевой сигнал имеет степень отличия меньше заданной величины, значит, отождествление прошло успешно, при отличии большем заданного, система признает результат опознания отрицательным. При автоматической идентификации система выбирает того зарегистрированного диктора, чей контрольный образец наиболее близок к входному сигналу. Результат распознавания в большой степени зависит от выбранных параметров речевых характеристик, по которым производится сравнение [17].

При всей кажущейся схожести задачи идентификации / верификации говорящего отличаются стратегиями принятия решения. «Наиболее типичная стратегия тексто-зависимой верификации диктора состоит в том, чтобы создать эталонный файл речевых сигналов (в функции времени) для каждого пользователя, а затем в процессе верификации сравнивать параметры речи неизвестного говорящего субъекта с параметрами эталона в эквивалентных точках временной оси. Это означает, что входной речевой сигнал, соответствующий голосу неизвестного диктора, совмещается во времени с предлагаемой системой эталонным сигналом, а затем вычисляется расстояние между соответствующими точками на оси времени, которое усредняется по всему речевому фрагменту. Такой общий метод стал основой почти всех подходов к решению задачи тексто-зависимой верификации дикторов. Исключением из этого правила является подход, при котором вычисляются усредненные по времени статистические оценки параметров речевых кадров, и решения по отождествлению голоса основываются на вычислении оценки максимального правдоподобия для эталонного голоса» [4, с. 141].

Стратегия тексто-независимой идентификации основывается на двух других подходах. «Суть первого подхода состоит в использовании

характеристик, усредненных по большому интервалу времени. Это значит, что определенные отличительные признаки речевого сигнала вычисляются для каждого кадра речевых данных и затем усредняются по всему фрагменту речи. Решение по распознаванию принимается на основе вычисления статистического правдоподобия усредненного идентифицирующего вектора в соответствии с гипотезой принадлежности голоса конкретному диктору <…> Второй, интуитивно весьма привлекательный подход, состоит в том, чтобы отыскивать в принимаемом речевом сигнале характерные фонетические явления, а затем сопоставлять характеристики отобранных и продетектированных фонетических явлений с характеристиками соответствующих фонетических явлений в эталонных голосах» [4, с. 139].

Основополагающую роль при анализе речевого материала помимо выбора стратегии распознавания играет выбор параметров речевого сигнала, по которым происходит сравнение с эталонными образцами. С развитием компьютерных технологий появилось большое количество программ, позволяющих работать с оцифрованным звуковым сигналом. Эти программы позволяют измерять целые комплексы практически всех известных акустических параметров речевого сигнала. В экспертных и автоматизированных системах, даже при использовании одинаковых параметров для анализа, алгоритмы принятия решения отличаются. Именно комплексный объективный подход, учитывающий большое количество параметров, позволяет автоматическим системам выгодно отличаться от экспертных систем распознавания дикторов.

Самой главной задачей в системах распознавания говорящего по голосу является выделение основных параметров, по которым можно отличить одного диктора от другого.

В каждом языке существуют инвариантные структуры, которые помогают людям понимать друг друга, но все же голос и речь каждого из нас индивидуальна, и именно этот тезис лежит в основе распознавания говорящего по голосу.

Индивидуальные голосовые качества разных людей определяются целым рядом параметров, включающих анатомические, артикуляционные, возрастные, социальные характеристики, дефекты речи, а также различные сочетания этих параметров.

«Речевое общение начинается с того, что в мозгу диктора возникает в абстрактной форме некоторое сообщение. В процессе речеобразования это сообщение преобразуется в акустическое речевое

колебание. Информация, содержащаяся в сообщении, представлена в акустическом колебании весьма сложным образом. Сообщение сначала преобразуется в последовательность нервных импульсов, управляющих артикуляторным аппаратом (т. е. перемещением языка, губ, голосовых связок и т. д.). В результате воздействия нервных импульсов артикуляторный аппарат приходит в движение, результатом которого является акустическое речевое колебание, несущее информацию об исходном сообщении» [10, с. 9].

Звучащая речь является результатом последовательного взаимодействия четырех артикуляционных процессов, на каждом из которых формируются индивидуальные признаки.

1. Из легких выталкивается струя воздуха, которая является основой формирования звучания.

«В зависимости от емкости легких и характера дыхания (ключичный, грудной или брюшной тип) человек при этом делает неодинаковое количество вдохов и выдохов и затрачивает на них различное время» [11, с. 15]. Это определяет цикл речевого дыхания, ритмическую структуру речи и силу звучания.

2. Воздушный поток начинает вибрировать, проходя через голосовые связки, расположенные в гортани.

«Размеры гортани и голосовых связок существенно меняются от индивида к индивиду. У мужчин они крупнее, чем у женщин и тем более у детей. Голос уже на данном этапе характеризуется определенной высотой, силой и тембром. Последние две характеристики, однако, с прохождением звуковых колебаний через глотку, затем ротовую и носовую полости существенно видоизменяются в зависимости от параметров этих полостей – резонаторов. Что же касается высоты звука, то она сохраняется до конца, представляя одну из основных особенностей индивидуального голоса.

Эта особенность, т. е. высота голоса, находится в прямой зависимости от колебания голосовых связок, которые, в свою очередь, зависят от длины, толщины и натяжения последних. Длинные, толстые и слабо натянутые связки обеспечивают низкие по высоте звуки. Увеличение натяжения связок, осуществляемое с помощью мышечного аппарата гортани, влечет за собой повышение высоты звука» [11, с. 17–18].

3. Вибрация в струе воздуха обретает особую форму благодаря резонаторам, сформированным в ротовой и носовой полостях органами артикуляции.

Здесь голосу придается индивидуальная тембровая окраска, благодаря целенаправленной регуляции человеком взаимоположений подвижных (язык, губы, увула) и неподвижных (мягкое и твердое нёбо, зубы) органов. В зависимости от того, в каком объеме воздушный поток направляется в носовую полость или полость рта, появляются назализованные или неназализованные звуки.

4. Распространение воздушной волны особой формы в окружающую среду.

Форма речевой волны однозначно определяется источником и фильтром. «Речевой тракт представляет собой сложный акустический фильтр с рядом резонансов, создаваемых полостями рта, носа и носоглотки, т. е. с помощью артикуляционных органов речи» [1, с. 46]. «Положения и движения артикуляторных органов определяют размеры резонирующих полостей. Деятельность дыхательных органов и мышц гортани определяют свойства голосового источника при сонорных звуках (к ним отнесены те звуки речи, которые образуются при участии голоса, независимо от того, гласные это или согласные), в основном не зависимо от артикуляции. В отличие от этого, для глухих согласных свойства источника определяются как дыхательными органами, так и положением органов, осуществляющих артикуляцию» [14, с. 17].

«При описании речеобразования резонансные частоты трубы голосового тракта называют формантными частотами, или просто формантами. Формантные частоты зависят от конфигурации и размеров голосового тракта: произвольная форма тракта может быть описана набором формантных частот. Спектральные свойства речевого сигнала изменяются во времени в соответствии с изменением формы голосового тракта» [10, с. 44–45]. «Положения частотных пиков в спектрограммах относительно устойчивы к спектральным искажениям. Формантные контуры сохраняют стабильность при телефонной передаче и, следовательно, могут служить базой при идентификации говорящего. Существуют различные методы параметризации и классификации формантных контуров» [9, с. 256].

В системах распознавания используются «низкоуровневые» акустические параметры, к числу которых относятся частота основного тона (ЧОТ), частоты формант (частоты максимумов спектра), ширина полосы каждой из них и характерные аномалии голоса. Все эти характеристики могут быть измерены как функции времени. В качестве

переменных распознавания могут также выступать соответствующие статистические характеристики, усредненные по большому интервалу времени [4].

«Исследования показывают, что одни звуковые сегменты играют более важную роль при идентификации личности говорящего, чем другие. Например, глухой фрикативный согласный **/с/** является потенциальным индикатором вариабельности речи разных говорящих. Оценки распределения энергии могут быть получены на основе анализа традиционных спектральных изображений, сделанных через определенные интервалы на всей длительности звука. Такие оценки, а в некоторых случаях также и вычисление усредненного энергетического спектра, вместе со звуковыми спектрограммами являются важными для описания акустических характеристик говорящих, в которых отражаются их индивидуальные особенности, связанные с артикуляцией гласных и согласных сегментов. Спектрограммы несут также информацию о длительности сегментов речевой цепочки. В ряде случаев эта информация может быть полезна, особенно тогда, когда в анализируемых голосовых образцах наблюдаются необычные или патологические отклонения от нормы» [9, с. 19].

«Речевой сигнал подвержен влиянию различных факторов. Параметры речевого сигнала передают как экстралингвистическую и паралингвистическую информацию, так и собственно лингвистическую информацию, участвуют в сегментной и супрасегментной организации высказываний» [8, с. 135]. «Под сегментами понимаются звуки речи или их сочетания, располагающиеся последовательно в речевом потоке. К супрасегментным единицам относятся ударение во всех его видах, мелодика, темп, громкость. Они отличаются от сегментных единиц тем, что не могут существовать сами по себе, без звуков (сегментов), а как бы накладываются на них» [8, с. 167].

Исследования основного тона выявили, что «особенностью основного тона является изменение в значительных пределах длительности интервалов при произношении отдельных фраз, а также наличие у многих людей разного основного тона для произношения тех или иных фраз. Первое явление носит название мелодии основного тона. Она характерна, например, для вопросительных и восклицательных предложений. Но у ряда людей и в ряде иностранных языков она часто встречается и в обычных фразах. По этой особенности можно опознать голос говорящего. Так как в этом случае происходит плавное

изменение тона, то оно может быть выделено с помощью приборов» [12, с. 180]. «В речи, как и во всякой информационной среде, все структурные компоненты взаимосвязаны и взаимозависимы. Звуковая фактура определяет произносительные особенности фигур слов и словосочетаний и обусловливает их просодическую (супрасегментную) специфику. В свою очередь, просодия действует на сцепления слов, словосочетаний и фактуру звуков так, что в каждый момент времени происходят фонетические изменения речевого потока» [9, с. 263].

Эти наблюдения показывают, что при распознавании дикторов по голосу можно опираться не только на пофонемный анализ отдельных звуковых сегментов, но и на просодическую информацию, которая несет дополнительные индивидуальные признаки. Просодическая информация может оказаться более полезной при идентификации говорящего, чем при верификации. Это связано с тем, что идентификация опирается на тексто-независимый анализ, где часто можно использовать длительные речевые фрагменты.

«К просодическим характеристикам речи относят частоту основного тона, длительность и интенсивность. Анализ результатов исследований, проведенных на супрасегментном уровне, показывает, что частота основного тона — акустический коррелят высоты тона — является одной из самых универсальных супрасегментных характеристик. Почти все виды интонационной информации могут быть переданы с помощью модификаций ЧОТ.

Фиксируемые значения ЧОТ несут информацию о характере работы голосового источника, который, в свою очередь, обусловливается как минимум тремя факторами: квазипериодичностью колебаний голосовых связок, индивидуальными особенностями голосового источника и эмоциональным состоянием говорящего. К числу параметров ЧОТ обычно относят средний уровень ЧОТ, частотный диапазон, частотный интервал, скорость изменения (подъема или падения) ЧОТ» [5, с. 70–71].

«Интерес представляет исследование такого параметра, как темп, его вариативность в речи одного и того же говорящего и в речи разных говорящих. Проведено исследование, как изменение темпа речи одного и того же индивида отражается на определенных акустических параметрах, в частности на значениях ЧОТ, формант гласных, долговременного спектра. Результаты показали, что для большинства говорящих (а) изменение темпа речи существенно не влияет на

долговременный спектр; (б) при быстром темпе речи наблюдается незначительное уменьшение площади гласных, вычисляемой по формуле F1xF2; (в) при быстром темпе речи заметно меняется распределение ЧОТ; (г) повышается средняя ЧОТ. Анализ вариативности ЧОТ в речи разных говорящих показал, что при быстром темпе речи вариативность ЧОТ понижается. Темповые показатели не являются величиной постоянной, а варьируются в зависимости от индивидуальных особенностей говорящих. Наблюдаемые изменения ЧОТ говорят о том, что в судебной фонетике при сравнении, основанном на параметрах ЧОТ, необходимо учитывать темп речи» [9, с. 256].

В ходе изучения систем распознавания голоса был выявлен феномен «внутридикторской» изменчивости, который заключается в том, что часто голос одного и того же диктора не похож сам на себя. «Являясь формой человеческого поведения, речь подвержена влиянию широкого ряда еще не полностью изученных факторов. Такие внутренние факторы, как усталость, болезнь, наличие алкоголя в крови и психическое состояние могут воздействовать на речевой сигнал на сегментном и просодическом уровнях. На речевой сигнал также влияют реакции говорящего на изменения во внешней коммуникативной среде» [9, с. 20].

«Среди всех прочих факторов, подлежащих учету, этот аспект представляет серьезную проблему для разработчиков систем распознавания речи. Как, например, можно гарантировать, что пользователь в каждом сеансе верификации будет произносить слова с одинаковой частотой и с одним и тем же напряжением голоса? Или, что более важно, как оградить систему от влияния на параметры глосса разного рода респираторных заболеваний, например обычной простуды? Ключом к решению указанных проблем является контролирование специфических статистических особенностей речевого сигнала, а не его усредненных параметров. «Типичный входной речевой сигнал может всегда обеспечивать высокое качество распознавания, и потому уровень ошибок в системе может определяться именно аномалиями речевых данных. Углубленный контроль речевых сигналов пользователей с учетом устойчивых к различным факторам характерных особенностей речи – вот ключ к созданию высокоэффективных систем верификации дикторов» [4, с. 141].

В случае внутридикторской вариативности эмоциональные состояния в большой степени затрагивают основные просодические характеристики (частоту основного тона, интенсивность, длительность).

«Параметр длительности на эмотивном уровне используется в большинстве языков для повышения степени речевой эмфазы. Темп в значительной степени определяется эмоциональным содержанием речевой ситуации. Возрастание эмоциональной напряженности связано с замедлением или ускорением темпа речи» [9, с. 21]. «Наличие трудностей, связанных с использованием параметров ЧОТ при идентификации говорящего, подтверждены результатами анализа средних значений ЧОТ, полученных из записи телефонных разговоров, имеющих отношения к преступлениям, и образцов голосов подозреваемых (в данном случае анализировались мужские голоса). Оказалось, что средние значения ЧОТ в реальных криминальных ситуациях намного выше, чем средние значения ЧОТ, полученные в лабораторных условиях. Также обнаружено, что данный параметр менялся в пределах 30 Гц у говорящего в зависимости от того, был ли голос записан во время телефонного разговора или специально для голосового образца. Установлено, что средние значения ЧОТ и значение стандартного отклонения, полученные из записей криминальных разговоров, не могут автоматически считаться действительными характеристиками голоса преступника в нормальных условиях. Чтобы получить достоверные данные о нормальном распределении значений ЧОТ анонимного говорящего по телефону, необходимо с помощью слухового анализа (опираясь как на лингвистическую, так и на акустическую информацию) выделить речевые сегменты довольно большой длительности, которые содержат речь, не подверженную влиянию анормальных психологических и / или ситуационных факторов» [9, с. 255–256].

Речевой сигнал представляет собой сложное, многоплановое явление, параметры которого зависят от целого ряда особенностей каждого человека. Выбор анализируемых параметров, по которым может проводиться распознавание диктора, должен опираться на цели, методы и задачи распознавания. Современные системы распознавания говорящих по голосу стремятся использовать как можно больше информации более высокого уровня, которая тесно связана с информацией низкого уровня. Вероятность правильного распознавания говорящего возрастает пропорционально тому, какие признаки анализируются, количество признаков и их корреляция.

Сущность систем распознавания говорящего по голосу предполагает, что рано или поздно любая система верификации сталкивается с попыткой несанкционированного доступа, а система

идентификации – с попыткой маскировки голоса с целью сокрытия личности. Эти проблемы являются весьма существенными в данных системах, так как в случае легкости процесса имитации и маскировки голосов данные системы теряют свою практическую значимость. Возникают закономерные вопросы: какие механизмы лежат в основе изменения голоса, какие индивидуальные параметры голоса возможно скопировать либо сильно деформировать, какими методами можно выявить данные изменения.

Имитация и маскировка голоса при видимом сходстве цели, достигаются разными способами и имеют разную степень легкости выполнения. «Как показали аудитивные опыты, слуховое восприятие человека, обладая широкими возможностями анализа голоса тем не менее вырабатывает свое мнение о специфике голоса человека путем оценки некоторых его особо характерных качеств, не ставя при этом акцента на количественной стороне разнообразных мелких компонент речевого сигнала.

Безупречное воспроизведение голосов и произношения, осуществляемое профессиональными имитаторами, возможно лишь в том случае, когда подражаемый субъект характеризуется ярко выраженными особенностями произношения (интонационной картиной, акцентом, темпом речи) или тембра (гнусавостью, шепелявостью, картавостью) и имитатор четко чувствует и воспринимает их. Именно этим следует объяснить тот факт, что даже профессиональным имитаторам очень сложно подделать ничем не примечательный, ординарный голос.

Автомат же, наоборот, не имея способности уловить общий характер голоса, в то же время может быть привязан к любым параметрам речевого сигнала, производя при этом их точный количественный анализ. Это обстоятельство позволяет предположить, что слух человека и распознающий автомат по-разному и на различных уровнях оценивают специфику голоса.

Формантный анализ речевых сигналов, соответствующих нормальным произношениям тестовых фраз, а также произношениям, измененным (с целью сокрыть свой голос) показал, что этим изменениям сопутствуют значительные деформации в формантной структуре сигнала. Однако отмечено, что сделать такую деформацию целенаправленной, намеренно стараясь достичь заранее известных значений формантных параметров, имитаторам удается крайне редко» [11, с. 165–167].

В отличие от имитации, маскировка голоса не требует никаких специальных навыков. К способам маскировки можно отнести маскировку возраста в речи, маскировку пола, имитацию иностранного акцента, имитацию дефектов речи, использование резкого / скрипучего голоса.

«Наблюдения показывают, что слушающие с трудом сопоставляют образцы замаскированного голоса с образцами нормального голоса одного и того же человека. Более того, слушающие пытаются искать совпадения поисковых признаков там, где их нет» [9, с. 25–26].

«Эксперименты показали, что как визуальное сравнение спектрографических “отпечатков” голоса при естественном и измененном произнесении, так и количественное сопоставление измеряемых параметров сигнала не дают основание надеяться на обнаружение каких-либо инвариантных признаков, характерных для данного диктора в условиях преднамеренного изменения голоса. Существенно изменчивыми и не подчиняющимися никаким закономерностям оказались и сами отклонения измеряемых параметров, что в свою очередь, не оставляет надежды обнаружить какие-либо корректирующие процедуры для компенсации изменчивости акустических параметров и для нормализации характерной для диктора индивидуальной природы голоса» [11, с.194–195].

«В качестве своего рода маскировки может выступать перекодирование речевого сигнала <…> Коварность перекодировки речевого сообщения для судебного фонетиста заключается в том, что такое сообщение не допускает какой-либо точной алгоритмической перестановки (между голосовыми образцами) фонетических, социолингвистических или других диалектальных признаков, а также в том, что в сообщении не может быть ошибок, выдающих замаскированный голос. Слуховой и акустический виды анализа могут не дать желаемого результата. Вместе с тем при наличии достаточной информации о самом говорящем и о содержании его поступка можно в определенной степени судить о вероятной причастности данного говорящего к исследуемому факту» [9, с. 26–27].

Как видно из вышесказанного, имитация голоса оказывает влияние на субъективное восприятие говорящего и легко выявляется автоматическими системами. Явление имитации голоса используется, в основном, в системах верификации, которые в настоящее время автоматизированы, таким образом, вероятность обмана системы злоумышленником достаточно мала. Чем больше различных параметров

голоса будет оцениваться системой, тем меньше вероятность, что имитатор сможет «обмануть» систему.

Если при верификации решением может являться использование автоматических систем, то при идентификации пока нет четких характеристик, по которым можно выявить сходства между замаскированным голосом и эталоном, так как пока не существует алгоритмов приведения замаскированного голоса к «нормальному» образцу. Сочетание субъективных и объективных методов могут повысить точность опознания говорящего, но вероятность правильного определения говорящего в случае маскировки голоса до сих пор остается невысокой.

Системы распознавания дикторов еще имеют большой потенциал для изучения. С развитием информатизации общества появляется все больше сфер, где могут быть использованы данные технологии. Российские и зарубежные организации продолжают проводить исследования в данной области, и если в системах верификации дикторов уже достигнута почти стопроцентная точность распознавания, то системы идентификации еще требуют серьезных доработок. Развитие компьютерных технологий и создание программных комплексов для анализа речевых сигналов позволяют вести работу над поиском новых параметров речевых сигналов, выработкой успешных стратегий сравнения входного сигнала с эталоном, созданием речевых баз для проведения экспериментальных исследований в данной области.

Большой научный и практический интерес представляет исследование маскировки / имитации голоса в системах идентификации дикторов. Основной областью, где используются лингвистические доказательства принадлежности голоса конкретному человеку, является судебная экспертиза, которая может повлиять на справедливость приговора. Поэтому необходимо иметь строгую доказательную базу и исключить ошибки идентификации. Для этого необходима разработка строгой системы выявления факта маскировки, а также четких алгоритмов и процедур поиска преступника из круга подозреваемых. Это требует всестороннего исследования всех возможных способов и нюансов маскировки колоса, выявления степени изменения основных акустических характеристик голоса при маскировке, поиска инвариантных признаков, характерных для каждого конкретного диктора в условиях изменения голоса. Большое количество еще не решенных задач данной области и их практическая значимость делают исследования маскировки / имитации голоса столь интересными и важными для изучения.

СПИСОК ЛИТЕРАТУРЫ

1. *Акустика*: справочник / Ефимов А. П. [и др.] ; под ред. М. А. Сапожкова. – 2-е изд. перераб. и доп. – М. : Радио и связь, 1989. – 336 с.
2. *Бондарко Л. В.* Основы общей фонетики : учеб. пособие / Л. В. Бондарко, Л. А. Вербицкая, М. В. Гордина. – СПб. : Изд-во С.-Петербургского ун-та, 1991. – 152 с.
3. *Галунов В. И.* О возможности определения эмоционального состояния говорящего по речи // Речевые технологии. – 2008. – № 1. – С. 63–66.
4. *Доддингтон Дж.* Распознавание дикторов: Идентификация людей по голосу // ТИИЭР. – 1985. – № 11 (73). – С. 129–146.
5. Общая и прикладная фонетика: учеб. пособие / Л. В. Златоустова, Р. К. Потапова, В. В. Потапов, В. Н. Трунин-Донской. – М. : Изд-во МГУ, 1997. – 416 с.
6. Информационно-издательский центр «CONNECT!» [Электронный ресурс] – М., 1996–2008. – Режим доступа: http://www.connect.ru/article.asp?id=7122, свободный. – Загл. с экрана.
7. *Михайлов В. Г.* Измерение параметров речи / В. Г. Михайлов, Л. В. Златоустова ; под ред. М. А. Сапожкова. – М. : Радио и связь, 1987. – 168 с.
8. *Потапова Р. К.* Речь: коммуникация, информация, кибернетика : учеб. пособие. – 3-е изд., стереотип. – М. : Едиториал УРСС, 2003. – 568 с.
9. *Потапова Р. К.* Язык, речь, личность / Р. К. Потапова, В. В. Потапов. – М. : Языки славянской культуры, 2006. – 496 с.
10. *Рабинер Л. Р.* Цифровая обработка речевых сигналов / Л. Р. Рабинер, Р. В. Шафер ; пер. с англ. / под ред. М. В. Назарова, Ю. Н. Прохорова. – М. : Радио и связь, 1981. – 496 с.
11. *Рамишвили Г. С.* Автоматическое опознавание говорящего по голосу. – М. : Радио и связь, 1981. – 224 с.
12. *Сапожков М. А.* Речевой сигнал в кибернетике и связи. – М. : Связьиздат, 1963. – 450 с.
13. *Сорокин В. Н.* Фундаментальные исследования речи и прикладные задачи речевых технологий // Речевые технологии. – 2008. – № 1. – С. 18–48.
14. *Фант Г.* Акустическая теория речеобразования : монография / пер. с англ. Л. А. Варшавского, В. И. Медведева ; под ред. В. С. Григорьева. – М. : Наука, 1964. – 284 с.
15. *Фланаган Дж. Л.* Анализ, синтез и восприятие речи : монография : пер. с англ. / под ред. А. А. Пирогова. – М. : Связь, 1968. – 360 с.
16. *Чистович Л. А.* Физиология речи. Восприятие речи человеком / Л. А. Чистович, А. В. Венцов. – Л. : Наука, 1976. – 388 с.
17. *Furui S.* Digital speech processing, synthesis, and recognition. – 2nd ed. – USA : Marcell Dekker Ink., 2001. – 453 p.
18. *Lindh J.* Handling the “Voiceprint” Issue // Proceedings, Phonetic 2004 Dept. of Linguistics, Stockholm : Stockholm University, 2004. – P. 72–75.