Информационные процессы, Том 10, №2, стр. 87- 104

====== ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ТЕХНИЧЕСКИХ ======
====== И СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ СИСТЕМАХ =======

ВЕРИФИКАЦИЯ ДИКТОРА ПО СПЕКТРАЛЬНО-ВРЕМЕННЫМ ПАРАМЕТРАМ РЕЧЕВОГО СИГНАЛА

В.Н.Сорокин, А.И.Цыплихин

Институт проблем передачи информации, Российская академия наук, Москва, Россия
Поступила в редколлегию 19.02.2010

Аннотация. Верификация диктора выполняется на основе измерения формантных частот на стационарных участках и переходных процессах гласных звуков, спектральных признаков фрикативных звуков, а также длительности речевых сегментов. Для каждого слова из фиксированного словаря русских числительных от 0 до 9 были отобраны наилучшие признаки. Парольная фраза генерируется системой в случайном порядке при каждом новом акте верификации. Компенсация динамических помех и противодействие вторжению с помощью воспроизведения подслушанных и записанных слов диктора осуществляется с помощью требования повторного произнесения некоторых слов. В результате более чем 30 миллионов тестов на базе данных для 429 дикторов для максимальной длины парольной фразы в 10 слов получены суммарные вероятности ошибки 0.006% для мужских голосов и 0.025% - для женских, причем вероятности пропуска самозванца и ложного отказа в этом случае примерно равны.

Введение.

Существует ряд ситуаций, в которых человеку необходимо подтвердить свое право на распоряжение материальными или информационными ресурсами, доступ к информации или в помещение, сейф и т.д. Подтверждение такого права осуществляется с помощью документов (паспорта, удостоверения личности, пропуска), физических (ключи, кредитные карты) или электронных средств (коды авторизации, пароли). В ряде случаев такие средства верификации личности либо неудобны, либо не обеспечивают необходимой степени защиты. Согласно решению *Federal Financial Institution Examination Council, USA,* от 2005 года, использование однофакторной методологии аутентификации личности (т.е. подтверждения личности с помощью ПИН-кода или буквенно-цифрового пароля) является неадекватным средством защиты в системах удаленного доступа к финансам. Поэтому, в дополнение к таким традиционным средствам, целесообразно использовать биометрические параметры человека. Преимущество биометрии заключается в том, что эти параметры всегда находятся при человеке, их нельзя забыть, потерять, передать другому человеку, украсть и довольно трудно воспроизвести.

К биометрическим параметрам относятся анатомические характеристики (рост, вес), признаки лица, радужной оболочки или сетчатки глаза, отпечатки пальцев или ладони, индивидуальные признаки речевого сигнала. Каждый из этих методов имеет свои преимущества и недостатки, ограничивающие область применения. Так, распознавание кровеносных сосудов сетчатки глаза или радужной оболочки требует особых условий измерения и сопряжено с риском повреждения глаз. Распознавание отпечатков пальцев, ладони или рисунка радужной оболочки глаза или сетчатки требует непосредственного контакта с регистрирующей аппаратурой, что сильно ограничивает возможности удаленного доступа, в том числе и по причине возможного перехвата биометрических параметров в канале связи.

Для всех биометрических методов верификации характерно огромное различие между показателями эффективности в лабораторных условиях, которые сообщаются разработчиками, и результатами тестирования независимыми организациями. Например, имеются сообщения о чрезвычайно низкой вероятности ошибки верификации личности по радужной оболочке, порядка 10^{-6}. В реальных условиях эксплуатации такая ошибка никогда не достигается (см. Табл. 1 по [1]).

Таблица 1. Независимые оценки систем биометрии.

Тип биометрии	**Кто тестировал**	**Процент ложных отказов**	**Процент ложных пропусков**
Отпечатки пальцев (4 пальца)	Fingerprint Verification Competition (2004)	2 %	2 %
Характеристики лица	Face Recognition Vendor Test (2002)	10 %	1 %
Отпечатки ладони	Purdue University's Research (2005)	0.1 %	2 %
Рисунок радужной оболочки	International Biometrics Group, Independent Testing of Iris Recognition Technology (2005)	0.99 %	0.94 %
Параметры голоса	The National Institute of Standards and Technology (NIST), (2004)	5-10 %	2 – 5 %

Аналогично, компания Persay сообщает, что EER (*Equal Error Rate*) верификации диктора по голосу составляет 0.05%, не указывая об условиях тестирования, тогда как независимые испытания, выполненные в университете Канберры, Австралия, в 2005 г., показали, что ошибка составляет от 1.5% до 5.3% даже в идеализированных условиях. Показатель EER определяется как величина ошибки при условии равенства ошибок пропуска самозванца и ошибок ложного отказа. Суммарная ошибка, таким образом, есть удвоенная ошибка EER.

Результаты тестирования коммерческих систем голосовой биометрии, проведенного *NIST* (*The National Institute of Standards and Technology*) в 2008 году, оказались очень близки к результатам, полученным в 2004 году [2]. Отсюда можно сделать вывод, что характеристики анализированных систем голосовой биометрии если и улучшаются, то очень медленно.

Помимо ошибок типа ложных пропусков и отказов, существует и такой показатель, как отказ от регистрации пользователя. Так, в системах распознавания отпечатков пальцев вероятность отказа от регистрации составляет около 4%, а в системах распознавания рисунка радужной оболочки – 7%. Для коммерческих систем биометрии голоса этот показатель близок к 2%. Это связано как с алгоритмами распознавания, так и с особенностями измерения биометрических параметров. В частности, глубина бороздок папиллярных узоров пальцев может быть настолько малой, что сканер не в состоянии считать эти узоры.

Эффективность систем верификации зависит не только от ошибок первого и второго рода, но и от условий, в которых эти системы используются. Если вероятность злонамеренного вторжения очень мала, то и ошибки в несколько процентов могут не препятствовать применению системы. Однако даже при малой вероятности вторжения высокий риск, связанный с ценностью защищаемой информации, не допускает использования систем с ошибками в несколько процентов. Вероятность пропуска самозванца p_1 до некоторой степени может быть снижена за счет увеличения вероятности ложного отказа p_2 путем смещения порога принятия решения. Однако, во-первых, достигаемое таким образом уменьшение p_1 обычно ограничено пределом, который зависит от степени перекрытия вероятностных распределений, и этот предел не обязательно приемлем. Во-вторых, увеличение числа ложных отказов существенно снижает готовность пользователей взаимодействовать с такой системой.

Фактически, именно неприемлемо большая вероятность ложных отказов является причиной малой востребованности систем биометрии. Например, по оценкам *Gartrner Group*, системы верификации личности по голосу занимают лишь около 1% потенциального рынка таких систем.

Принципиальный недостаток всех методов биометрии, кроме речевого, состоит в постоянстве используемого биометрического кода, т.к. отпечатки пальцев или ладоней, рисунок радужной оболочки и черты лица неизменны для индивидуума. Этот недостаток препятствует применению этих методов в случаях, требующих особо высокой надежности идентификации личности, поскольку неизменный биометрический код может быть считан путем злонамеренного вторжения в программу распознавания.

В отличие от биометрии по фиксированным параметрам, верификации по голосу обладает практически неограниченным потенциалом для снижения ошибки за счет использования все более длинных речевых сообщений. Верификации по голосу может использоваться в темноте, на расстоянии, в частности, по стандартному телефонному каналу, в условиях, когда невозможно получить изображение лица.

Примеры конкретных применений верификации диктора охватывают широкий спектр приложений:

- распоряжение финансовыми процессами по электронным или телефонным каналам (управление банковским счетом, электронная коммерция, подтверждение права пользования кредитной картой);
- разрешение на смену пароля или PIN-кода;
- доступ к компьютеру или отдельным программам компьютера (вход в Интернет, доступ к конфиденциальным документам, базам данных и т.д.);
- разрешение на вход в помещение, открывание сейфа;
- управление механизмами и системами (например, запуск двигателя автомобиля);
- мониторинг того, кто, когда и к каким компьютерным ресурсам имел доступ.

С самого начала работ в области распознавания дикторов, исследования велись в двух направлениях: исследования специфических для диктора акустических параметров, и технического подхода, основанного на использовании формальных методов. Примером первого подхода является работа [3], а второго – весьма популярный в свое время метод "отпечатков голоса" [4]. В настоящее время доминирует второй, так называемый "*ignorance based*" подход, который минимально использует свойства речевого сигнала. Первичный анализ заключается в вычислении коэффициентов кепстрального разложения либо по логарифмическому спектру в шкале *мел* [5], либо по коэффициентам линейного предсказания [6]. Затем обычно применяются статистические процедуры принятия решения с использованием смеси нормальных распределений [1]или скрытых Марковских моделей [7, 8]. Детерминистский подход представлен искусственными нейронными сетями [9, 10].

Метод кепстрального представления, т.е. обратного Фурье-преобразования от логарифмического амплитудного спектра в масштабе частот *мел*, был разработан в результате длительных поисков такого описания речевого сигнала, которое было бы устойчиво к индивидуальным особенностям дикторов. Поэтому довольно странно было бы его использовать для решения противоположной задачи – поиска описания, подчеркивающего различия в голосах дикторов. Однако большинство работ либо игнорирует этот факт, либо занято поиском наиболее эффективных методов принятия решения в заданном пространстве кепстральных коэффициентов.

Иногда предпринимаются попытки связать характеристики речевого сигнала с артикуляторными параметрами. Однако, это направление сталкивается с большими трудностями в силу того, что задача определения параметров речевого тракта по акустическим параметрам является некорректной обратной задачей, т.е. не гарантирует устойчивого и однозначного решения.

Оценивая эффективность той или иной системы верификации, желательно знать, насколько ее характеристики близки к потенциально наименьшим ошибкам распознавания, и при каких условиях такие ошибки достижимы. В настоящей работе приводятся результаты исследования системы верификации диктора, в которой используются физически и физиологически обоснованные признаки, специфические для русского языка и словаря числительных от 0 до 9. Оригинальность и эффективность разработанных методов подтверждена патентом [11].

Пространство признаков.

Физической основой верификации по голосу служит анатомия речевого тракта, свойства системы управления артикуляцией и особенности голосового источника. Анатомия тракта определяет спектральные характеристики звуков речи, система управления артикуляцией влияет на темп речи, скорость переходных процессов и длительность речевых сегментов, а голосовой источник определяет частоту основного тона и тембральные характеристики речевого сигнала.

В данной работе исследовались только такие признаки, которые могут быть непосредственно измерены в речевом сигнале, и не предпринималось никаких попыток решения

обратных задач относительно формы речевого тракта, артикуляторных параметров или параметров голосового источника. В случае успешного решения этих задач можно ожидать дальнейшее повышение эффективности систем верификации диктора. Вместе с тем, как показывают результаты настоящего исследования, верификация дикторов в пространстве акустических параметров обеспечивает характеристики, удовлетворяющие большинству условий реального применения.

Узнаваемость голоса людей с удаленной гортанью для родственников и близких знакомых достаточно высока. Это указывает на то, что в анатомических параметрах речевого тракта и системе управления артикуляцией присутствует существенная информация об индивидуальности диктора. На то, что речевой тракт больше влияет на идентификацию голоса, чем голосовой источник, указывается в [12].

Длина речевого тракта и форма его функции поперечного сечения однозначно определяют его резонансные частоты, а податливость стенок и аэродинамические потери влияют на затухания временных мод. Аналогом резонансных частот в речевом сигнале являются формантные частоты, определяемые как частоты пиков огибающей спектра, или как полюса его передаточной функции, которые обычно вычисляются методом линейного предсказания. Формантные частоты определяют фонетическое качество ударных гласных, тогда как безударные гласные характеризуются формой речевого тракта, близкой к нейтральной. Известно, что диапазон значений первых трех формантных частот $\{F_1^{уг}, F_2^{уг}, F_3^{уг}\}$ ударных каждого языка значительно шире диапазона этих частот для каждого диктора в отдельности. Это создает основу для различения дикторов в пространстве этих частот.

Акустические параметры безударных гласных характеризуют общую форму речевого тракта и его длину. Поэтому формантные частоты $\{F_1^{бг}, F_2^{бг}, F_3^{бг}\}$ безударных гласных также несут информацию об анатомических особенностях ротовой полости диктора. Формантные частоты $\{F_1^{н}, F_2^{н}, F_3^{н}\}$ назальных сегментов /*м*,*н*/, а также назализованных гласных в окрестности назальной смычки, определяются как формой речевого тракта, так и формой носовой полости.

По мере движения артикуляторных органов к смычке при артикуляции согласных /*б, д, г, п, т, к, м, н, в, л, р*/, резонансные частоты тракта стремятся к значениям, соответствующим нулевой площади поперечного сечения тракта в так называемом тесте артикуляции. Эти частоты также зависят от формы тракта и положения смычки в тракте. Чем ближе к смычке измеряются формантные частоты, тем больше они характеризуют форму тракта при артикуляции согласных.

Фрикативные звуки /*с, ш, х, з, ж*/, а также аспиративные участки взрыва согласных, образуются с участием шумоподобного источника давления $F(x, t)$, создаваемого турбулентным потоком воздуха в области расширения тракта после сужения, являющегося местом артикуляции фрикативных. Здесь x – координата вдоль средней линии речевого тракта, t – время. Спектр этого источника зависит от формы расширения. Парциальные возбуждения $F_k(t)$ для каждой временной моды находятся как [13]:

$$F_k(t) = \frac{2}{\rho_0 l}\int_0^l F(x,t)\psi_k(x)dx$$

где ρ_0 – плотность воздуха, l – длина речевого тракта, ψ_k – k-я собственная функция акустического давления в тракте. В зависимости от распределения возбуждения по длине речевого тракта амплитуда k-*й* временной моды может быть больше или меньше, а в некоторых случаях эта мода вообще не возбуждается. Действительно, если возбуждение сосредоточено в одной точке x_0, то можно записать $F(x,t) = f(t)\delta(x - x_0)$, где δ — функция Дирака, f – изменение давления во времени в турбулентном потоке. Тогда амплитуда возбуждения k-й временной моды

$$F_k(t) = \frac{2}{\rho_0 l}\psi_k(x_0) f(t).$$

Если x_0 совпадает с пучностью собственной функции ψ_k, то эффект воздействия F_k на *k*-ю моду будет максимальным. Если же x_0 приходится на нуль ψ_k, то и F_k =0 и, следовательно, эта мода не возбуждается. Это явление не изменяется по существу и в том случае, если источник не сконцентрирован в одной точке, а распределен на некотором интервале.

Поскольку собственные функции тракта определяются его формой, то можно ожидать, что в спектре фрикативных звуков и аспиративных взрывов отражаются индивидуальные особенности диктора. Исследование признаков фрикативных участков речевого сигнала было выполнено в [14, 15]. Оценка чувствительности спектральных характеристик фрикативных к индивидуальным особенностям дикторов позволила определить ряд статических и динамических признаков. К числу статических признаков относятся частота центра тяжести $F_{\text{цт}}$ среднего спектра стационарного участка фрикативного, частота пересечения огибающей среднего спектра его среднего значения со стороны низких частот $F_{нч}^{ф}$ и частота пересечения огибающей среднего спектра его среднего значения со стороны высоких частот $F_{вч}^{ф}$. Средний наклон спектра $s_ф$ также характеризует особенности формы тракта.

Турбулентные процессы развиваются и затухают во времени. Признаки динамики этих процессов находятся как частоты $F_д^{ф+}$, $F_д^{ф-}$ и амплитуды $A_д^{ф+}$, $A_д^{ф-}$ локального максимума энергии динамического детектора в начале и конце фрикативного сегмента. Концепция динамических детекторов, описывающих спектрально-временные характеристики переходных процессов в речевом сигнале, была сформулирована в [16]. Динамический детектор представляет собой некоторую функцию от частоты и времени $A(\omega, t)$, которая вычисляется как логарифм отношения мгновенных спектров речевого сигнала $S(\omega, t)$, сглаженных с разными параметрами по времени τ и частоте θ, и сдвинутыми относительно друг друга по времени или частоте на некоторый интервал ΔT или $\Delta\Omega$:

$$A(\omega, t) = \lg \frac{S(\omega + \Delta\Omega, \theta_1, t \pm \Delta T_1, \tau_1) + C}{S(\omega - \Delta\Omega, \theta_2, t \mp \Delta T_2, \tau_2) + C},$$

где C - некоторая константа.

Динамические детекторы могут быть настроены таким образом, что они оценивают относительную скорость возрастания или спада энергии в различных частотных областях для переходных процессов различной длительности. Это свойство позволяет использовать их для определения начала или конца речевого сегмента определенного типа. В частности, отсчет формантных частот выполняется в моменты максимальной скорости нарастания или спада энергии на переходных процессах между гласными и согласными звуками, что добавляет два формантных вектора на этих переходных процессах: $\{F_1^+, F_2^+, F_3^+\}$ и $\{F_1^-, F_2^-, F_3^-\}$.

Динамические детекторы используются также для определения длительности различных участков речевого сигнала. В нормальных условиях темп речи, т.е. число фонетических сегментов в единицу времени, у разных дикторов может отличаться более чем в два раза [17]. При этом и длительность сегментов, входящих в одно и то же слово, различается у разных дикторов. Это свойство системы управления артикуляцией может использоваться для верификации дикторов [18]. В речевом сигнале, однако, не всегда удается автоматически установить границу между сегментами с целью измерения их длительности. Поэтому в задачах верификации с фиксированным словарем для каждого слова нужно выбрать такие участки речевого сигнала, границы которых определяются достаточно устойчиво. Признаками временной структуры слова служат как абсолютные, так и относительные значения длительности сегментов. Интервалы измерения длительности сегментов для русских числительных от 0 до 9, которые оказались наиболее устойчивыми и информативными в задаче верификации диктора, приводятся в Табл. 2.

Табл. 2. Временные интервалы словаря числительных.

	D_1	D_2	D_3	D_4
ноль	начало «Н» – конец «Л»	«О»	«Н»	
один	начало «а» – конец «Н»	«И» от взрыва	Смычка «Д»*	«Н»
два	взрыв «Д» – конец «А»	«А»		
три	взрыв «Т» – конец «И»	«И»		
четыре	начало «Ч» – конец «и»	Смычка «Т»*	«ТЫ»	
пять	взрыв «П» – взрыв «Т»	Смычка «Т»*		
шесть	начало «Ш» – взрыв «Т»	«Э»	Смычка «Т»*	
семь	начало «С» – конец «М»	«М»	«ЕМ»	
восемь	начало «В» – конец «М»	«О»	«С»	
девять	взрыв «Д» – взрыв «Т»	Смычка «Т»*		

* В словах /*пять, шесть, девять*/ иногда отсутствует взрыв последнего согласного /*т*/, и в этом случае длительность глухой смычки не может быть измерена.

Часть речевой базы данных была подвергнута ручной разметке. При разметке различалось 127 типов артикуляторно-акустических элементов (Табл. 3).

Таблица 3. Артикуляторные и фонетические элементы алфавита разметки.

Тип сегмента	Символ
Гласные ударные	А, Э, О, У, Ы, И
Гласные предударные	а, э, о, у, ы, и
Гласные безударные (редуцированные)	ъ, ь, уу
Сегмент последнего безударного гласного	ъ_, ь_, уу_
Сегмент последнего ударного гласного	А_, Э_, О_, У_, Ы_, И_, Я_, Е_, Ё_, Ю_
Огласованный сегмент	ъ/, ь/, у/
Полугласный язычный звонкий	Й
Полугласный язычный глухой	й
Дифтонги ударные	Я, Е, Ё, Ю
Дифтонги безударные	я, е, ё, ю
Согласные твердые	Б, В, Г, Д, Ж, З, К, Л, М, Н, П, Р, С, Т, Ф, Х, Хг, Ц, Ш
Согласные мягкие	Б', В', Г', Д', Ж', З', К', Л', М', Н', П', Р', С', Т', Ф', Х', Ц', Ч, Ш', Хг'
Смычка аспиративная твердая	Бв, Дз, Дж, Пф, Кх, Рш, Тс
Смычка аспиративная мягкая	Бв', Дз', Дж', Пф', Кх', Рш', Тс'
Взрыв твердый	Б!, Д!, Г!, П!, Т!, К!, Л!, Р!, М!, Н!
Взрыв мягкий	Б'!,Д'!,Г'!,П'!,Т'!,К'!, Л'!, Р'!, М'!, Н'!
Взрыв аспиративный твердый	Б!h, Д!h, Г!h, П!h, Т!h, К!h, Л!h, Р!h
Взрыв аспиративный мягкий	Бh', Дh', Гh', Пh', Тh', Кh', Лh', Рh'
Короткая пауза (провал после фрикативных, взрывов и т.д. "epenthetic")	\/
Фарингальный взрыв	gb; перед начальными гласными
Аспиративный сегмент	vh; придыхание перед началом слова или по окончании слова
Неречевые сегменты	
Не речь (шум канала)	h#
Пауза (между раздельными словами)	z#
Короткая пауза ("провал") между словами в слитном тексте	][
Неизвестный	?
Чмок	ch

В Табл. 4 представлено 45 спектральных признаков числительных, что вместе с 4 признаками длительности составляет полный алфавит из 49 признаков. При этом размерность пространства для каждого числительного находится в диапазоне от 17 (для слова /*ноль*/) до 37 (для слова /*девять*/.

Таблица 4. Полная таблица признаков для всех слов. Фон указывает на объединение в пространства.

Признак	ноль	один	два	три	четыре	пять	шесть	семь	восемь	девять
T_0 ядра ударного гласного	+	+	+	+	+	+	+	+	+	+
$F_1^{уг}$	+	+	+	+	+	+	+	+	+	+
$F_2^{уг}$	+	+	+	+	+	+	+	+	+	+
$F_3^{уг}$	+	+	+	+	+	+	+	+	+	+
F_1^+ ударного гласного	+	+	+	+	+	+	+	+	+	+
F_2^+ ударного гласного	+	+	+	+	+	+	+	+	+	+
F_3^+ ударного гласного	+	+	+	+	+	+	+	+	+	+
F_1^- ударного гласного	+	+			+	+	+	+	+	+
F_2^- ударного гласного	+	+			+	+	+	+	+	+
F_3^- ударного гласного	+	+			+	+	+	+	+	+
$F_{i\div 1}^{\delta}$		Дh'			Ч	Th'	Ш	С'	С’	Дh'
$F_{цт1}$		Дh'			Ч	Th'	Ш	С'	С’	Дh'
$s_{ф1}$		Дh'			Ч	Th'	Ш	С'	С’	Дh'
$F_{i\div 2}^{\delta}$							С’			Th'
$F_{цт2}$							С’			Th'
$s_{ф2}$							С’			Th'
$F_{i\div 3}^{\delta}$										Th'
$F_{цт3}$										Th'
$s_{ф3}$										Th'
T_0 назального	+	+						+	+	
$F_1^{н}$	+	+						+	+	
$F_2^{н}$	+	+						+	+	
$F_3^{н}$	+	+						+	+	
T_0 безударного гласного		а	ъ/	ъ/	и				ь	ь
$F_1^{бг}$		а	ъ/	ъ/	и				ь	ь
$F_2^{бг}$		а	ъ/	ъ/	и				ь	ь
$F_3^{бг}$		а	ъ/	ъ/	и				ь	ь
F_1^+ безударного гласного		а	ъ/	ъ/	и					ь
F_2^+ безударного гласного		а	ъ/	ъ/	и					ь
F_3^+ безударного гласного		а	ъ/	ъ/	и					ь
F_1^- безударного гласного		а	ъ/	ъ/	и					ь
F_2^- безударного гласного		а	ъ/	ъ/	и					ь
F_3^- безударного гласного		а	ъ/	ъ/	и					ь
$F_{д1}^{ф+}$		Дh'			Ч		Ш	С'	С'	Дh'
$A_{д1}^{ф+}$		Дh'			Ч		Ш	С'	С'	Дh'
$F_{д1}^{ф-}$		Дh'			Ч	Th'	Ш	С'	С'	Дh'
$A_{д1}^{ф-}$		Дh'			Ч	Th'	Ш	С'	С'	Дh'
$F_{д2}^{ф+}$							С'			
$A_{д2}^{ф+}$							С'			
$F_{д2}^{ф-}$							С'			
$A_{д2}^{ф-}$							С'			
$F_{д3}^{ф+}$							Th'			
$A_{д3}^{ф+}$							Th'			
$F_{д3}^{ф-}$							Th'			
$A_{д3}^{ф-}$							Th'			

Варианты произнесения слов терминах разметки представлены в Табл. 5.

Табл. 5. Фонетические транскрипции числительных.

ноль	Н_О_Л'
один	а_Д'_ Дh'_И_Н, а_Д'_ Д'!_И_Н
два	Д_В_А, Д_ъ/_В_А, Д_ъ/_В_А_А_
три	Т_ъ/_ Р'_И
четыре	Ч_и_Т_ Т!h_Ы_ Р'_ь, Ч_и_Т_ Т!h_Ы_ Р'_и
пять	П'!_Я_ Т'_ Th', П'!_Я_ Т'
шесть	Ш_Э_ С'_ Т'_ Th', Ш_Э_ С'_ Т'
семь	С'_Е_ М'
восемь	В_О_ С'_ь_ М', В_О_ С'_и_ М'
девять	Д'_Е_ В'_ь_ Т'_ Th', Д'_Е_ В'_ь_ Т'

Формантный анализ.

Определение формантных частот по речевому сигналу в силу ряда причин наталкивается на серьезные трудности. Спектр речевого сигнала искажается эффектами реверберации, амплитудно-частотными характеристиками приемника звука и внешними помехами. Степень выраженности пиков спектра, ассоциирующихся с формантами, зависит от расстояния между резонансными частотами речевого тракта и ширины этих резонансов. Колебания на резонансных частотах подсвязочной области могут проникать в ротовую полость, создавая дополнительные пики. На интервале открытой голосовой щели спектральные характеристики речевого тракта искажаются как из-за влияния граничных условий, так и вследствие возмущений, создаваемых голосовым источником [19]. Эти факторы приводят к неустойчивости и неоднозначности определения формантных частот речевого тракта [20]. Для того, чтобы уменьшить погрешность вычисления формантных частот необходимо использовать дополнительную информацию о временных и частотных свойствах акустических колебаний в речевом тракте.

Один из способов повышения стабильности и точности формантного анализа состоит в оценке формантных частот на интервале закрытой голосовой щели, т.е. анализа, синхронного с периодом основного тона. С этой целью выполняется обратная фильтрация речевого сигнала методом линейного предсказания, и начало интервала закрытой голосовой щели определяется с некоторым сдвигом относительно пика сигнала-остатка. Способ определения периода основного тона описан в [21].

Другой способ состоит в использовании информации о совместном распределении формантных частот для различных фонетических элементов. Если, например, известен тип гласного, на интервале которого выполняется формантный анализ, то можно сформировать метод выбора вектора формантных частот. В таком методе на каждом периоде основного тона перебираются все возможные сочетания кандидатов на формантный вектор и вычисляется апостериорная вероятность принадлежности к ранее найденному распределению частот для множества дикторов. В критерий выбора входит также и амплитуда спектра на частотах кандидатов. На квази-стационарном участке гласного можно использовать критерий наименьшей вариации трека формантой частоты, тогда как на переходных участках могут наблюдаться разрывы в этих треках, и такой критерий только ухудшает стабильность оценки частот [19].

Сегментация.

Самый общий случай сегментации состоит в определении начала и конца речевого сигнала. В задачах верификации неопределенность начала сигнала снижается, поскольку речевой сигнал ожидается только после того, как пользователь заявил о своем присутствии путем ввода своего идентификатора. Если словарь ограничен, и система верификации сообщает пользователю о том, какое слово он должен произнести, то детектирование начала-конца слова значительно упрощается.

В описываемой системе детектор "речь/не речь" работает в два этапа. На первом этапе вычисляется превышение абсолютной энергии речевого сигнала некоторого порога, который определяется в процессе обучения системы на параметры данного диктора. Этот порог устанавливается таким образом, чтобы свести к минимуму вероятность пропуска сигнала. На втором этапе в окрестности предполагаемого начала слова с порогом сравнивается энергия

динамического детектора, настроенного на переходы от шума канала к возможным сегментам в начале данного слова. Аналогично оценивается и конец слова, который ожидается на некотором интервале времени после начала слова.

Сегментация слова на фонетические элементы также выполняется в два этапа. На первом этапе возможная граница между сегментами устанавливается путем анализа экстремумов сегментирующей функции R, оценивающей меру близости между спектрально-временными характеристиками речевого сигнала в некоторые моменты времени. Компонентами функции R являются мера сходства R_1 между логарифмическими спектрами, вычисленными на интервале закрытой голосовой щели, мера сходства R_2 между амплитудно-частотными характеристиками динамического детектора, вычисленными на некотором расстоянии по времени, и вероятность p_{ij} перехода от i-го сегмента к j-му в данном слове. В простейшем случае функция R определяется как взвешенная сумма компонент: $R = a_1R_1 + a_2R_2 + a_3p_{ij}$, а более сложный вариант состоит в представлении этой функции как функции трех аргументов $R(R_1, R_2, p_{ij})$. И в том, и в другом случае порог принятия решения о возможной границе между сегментами устанавливается на основе обучения, в идеале, по базе данных для данного диктора, или по общей базе данных для данного словаря.

$$R_1(t) = 1 - \frac{\int_0^{\Theta} \lg S(\omega,t) \lg S(\omega,t+kT_0)\,d\omega}{\left[\int_0^{\Theta} \lg S^2(\omega,t)\,d\omega \int_0^{\Theta} \lg S^2(\omega,t+kT_0)\,d\omega\right]^{1/2}},$$

где Θ - максимальная частота в спектре, T_0 – период основного тона, коэффициент k = 1, 2, …, определяет ширину временного окна.

$$R_2(t) = 1 - \frac{\int_0^{\Theta} A(\omega,t) A(\omega,t+\Delta t)\,d\omega}{\left[\int_0^{\Theta} A^2(\omega,t)\,d\omega \int_0^{\Theta} A^2(\omega,t+\Delta t)\,d\omega\right]^{1/2}},$$

где $A(\omega,t)$ – динамический детектор, настроенный на искомый переход между сегментами, Δt – ширина временного окна.

Стабилизация оценок формантных частот выполняется путем усреднения на "ядре" гласного, которое определяется как квази-стационарный участок, причем под стационарностью понимается стабильность нормированного к общей энергии логарифмического спектра $\bar{S}(\omega,t)$. Поиск ядра начинается на границе прихода к гласному, а затем оценивается разница между текущим нормированным спектром $\bar{S}(\omega,t)$ и накопленным средним спектром $\bar{S}_i(\omega)$, где i – число импульсов основного тона, которое отсчитывается от начала перехода к гласному.

Средний спектр текущего периода основного тона $\bar{S}_i(\omega)$ рекуррентно вычисляется как

$$\bar{S}_i(\omega) = \frac{\bar{S}_{i-1}(\omega)(i-1) + \bar{S}(\omega,t)}{i}, \quad \bar{S}_0(\omega) = \bar{S}(\omega,t_0),$$

t_0 - момент начала перехода к гласному. Момент времени t является концом ядра гласного, если невязка в Евклидовой метрике между $\bar{S}_i(\omega)$ и $\bar{S}(\omega,t)$ превышает некоторый порог.

На этом этапе средняя вероятность пропуска границы между сегментами 0.66%, тогда как появляется примерно 26% лишних границ. Уточнение числа и положения границ между сегментами выполняется на втором этапе сегментации. С этой целью создается база эталонных произнесений, основанная на ручной разметке.

При произнесении заданного слова диктором в процессе верификации выполняется сравнение динамической сонограммы этого слова с сонограммами размеченных эталонов с использованием динамической трансформации временной оси. Разметка того эталона, расстояние до которого оказалось наименьшим, используется для коррекции границ между сегментами, полученными на первом этапе сегментации. Эталоны были получены путем кластеризации произнесений слов более чем 400 дикторами, из которых было отобрано 49 мужчин и 37 женщин в качестве представителей базы данных.

Погрешность определения границ в такой двухступенчатой процедуре составила около 5 мс.

Решающие правила.

Как видно из Табл. 2 и 4, минимальная размерность пространства признаков равна 17, что делает совершенно безнадежным принятие решения о подтверждении или отказе в подтверждении личности в таком пространстве в силу того, что размер обучающей выборки всегда будет непредставительным. Ситуация с размерностью могла бы облегчиться, если бы удалось разделить полное пространство на независимые подпространства малой размерности. Для исследуемого словаря числительных такое разбиение оказалось возможным. В Табл. 4 белым и серым фоном отмечены более или менее независимые двумерные и трехмерные подпространства. Однако и в пространствах такой размерности представительность обучающей выборки может оказаться недостаточной.

Можно несколько увеличить размер обучающей выборки, если предположить, что в малой окрестности каждого вектора признаков должны находиться и другие вектора, принадлежащие этой выборке. Вероятность появления дополнительных векторов должна падать по мере увеличения расстояния от данного вектора, например, по нормальному закону распределения вероятностей. При этом необходимо позаботиться о том, чтобы дисперсия σ этого распределения не была бы слишком большой, и соответствовала свойствам исходного распределения.

Контролируя величину ошибки распознавания диктора, удалось сконструировать алгоритм выбора дисперсии σ. Ее можно вычислить как произведение среднего минимального расстояния между векторами данного диктора на дисперсию данного пространства, вычисленную по всем дикторам. Среднее минимальное расстояние находится по матрице взаимных расстояний между векторами, где сначала определяются минимальные расстояния по каждому столбцу матрицы, а затем эти расстояния усредняются. При размножении исходной выборки ширина нормального распределения для каждого исходного вектора ограничивалась на уровне 3σ. Это уменьшает вероятность чрезмерного размытия исходного распределения.

Этот метод хорошо работает в одномерных пространствах, тогда как в многомерных случаях необходимо применять другой подход. Наиболее популярно использование смесей распределений:

$$p(x)=\sum_{j=1}^{k} w_j p_j(x),\ \sum_{j=1}^{k} w_j = 1,$$

где $p_j(x)$ – функция распределения многомерного аргумента x, w_j – её вес. Предполагается, что функции распределения принадлежат к некому параметрическому семейству распределений $\varphi(x;\theta)$ и отличаются только значениями параметра, $p_j(x)=\varphi(x;\theta_j)$. Задача разделения смеси заключается в том, чтобы, зная выборку $X^m=\{x_1,\ldots,x_m\}$, число k и семейство $\varphi(x;\theta)$, оценить вектор параметров $\Theta=(w_1,\ldots,w_k,\theta_1,\ldots,\theta_k)$. Для её решения используется метод максимума правдоподобия, который позволяет оценить неизвестные параметры плотности распределения по случайной, независимой, одинаково распределённой выборке наблюдений X^m. Метод состоит в том, чтобы найти значение вектора параметров, при котором наблюдаемая выборка наиболее вероятна.

Попытка разделить смесь распределений, используя принцип максимума правдоподобия «в лоб», приводит к слишком громоздкой оптимизационной задаче. Обойти эту трудность позволяет алгоритм EM (expectation-maximization). Идея алгоритма заключается в следующем. Искусственно вводится вспомогательный вектор скрытых переменных G, обладающий двумя

замечательными свойствами. С одной стороны, он может быть вычислен, если известны значения вектора параметров Θ. С другой стороны, поиск максимума правдоподобия сильно упрощается, если известны значения скрытых переменных.

Классический EM-алгоритм [22] состоит из итерационного повторения двух шагов. На E-шаге вычисляется ожидаемое значение (expectation) вектора скрытых переменных G по текущему приближению вектора параметров Θ. На M-шаге решается задача максимизации правдоподобия (maximization) и находится следующее приближение вектора Θ по текущим значениям векторов G и Θ.

Как и во всякой задаче оптимизации, результат и скорость сходимости могут существенно зависеть от начального приближения. Сходимость ухудшается в тех случаях, когда делается попытка поместить несколько центров в один фактический сгусток распределения, либо поместить центр компоненты посередине между сгустками. Кроме того, в классическом EM-алгоритме необходимо задавать число компонент *k*.

Для устранения этих недостатков был разработан алгоритм последовательного разделения и добавления компонент, что позволило решить все три проблемы: поиска локального максимума правдоподобия, выбора начальных приближений и определения числа компонент смеси. На начальном шаге алгоритма выборка аппроксимируется однокомпонентной смесью, чьи параметры находятся однозначно. Это снимает проблему начального приближения. Затем компоненты поочередно добавляются.

Вводится критерий $C(\Theta_k)$ эффективности описания выборки смесью из k компонент, включающий в себя штраф на число компонент. В условиях рассматриваемой задачи наибольшую эффективность показал критерий ICL-BIC [23]:

$$C(\Theta_k) = -2L(\Theta_k) + 2E(\Theta_k) + \nu(\Theta_k)\log m .$$

Здесь $L(\Theta_k)$ – логарифм функции правдоподобия, $E(\Theta_k)$ – энтропия, $\nu(\Theta_k)$ – число свободных параметров в смеси Θ_k, m – число элементов в выборке. Логарифм функции правдоподобия $L(\Theta_k)$ определяется как

$$L(\Theta_k) = \sum_{i=1}^{m} \ln \sum_{j=1}^{k} w_j \varphi(x_i; \theta_j) .$$

Энтропия $E(\Theta_k)$ записывается формулой

$$E(\Theta_k) = -\sum_{i=1}^{m}\sum_{j=1}^{k} g_{ij} \log(g_{ij}) \geq 0 .$$

Здесь $g_{ij} \equiv \mathrm{P}(\theta_j \mid x_i)$ – апостериорная вероятность того, что обучающий объект x_i был сгенерирован *j*-й компонентой смеси.

В алгоритме также используется так называемый критерий разделимости $S(j;\Theta_k)$, характеризующий качество описания j-й компонентой смеси принадлежащих ей объектов [24]:

$$S(j;\Theta_k) = \int f_j(x;\Theta_k) \ln \frac{f_j(x;\Theta_k)}{\varphi(x;\theta_j)} dx ,$$

где $f_j(x;\Theta_k)$ – локальная плотность выборки для j-го распределения. Распределение с наибольшим $S(j;\Theta_k)$ имеет наихудшую оценку локальной плотности и, следовательно, является первым кандидатом на разделение.

В результате работы этого алгоритма выборка данных апрроксимируется смесью многомерных нормальных распределений, параметры которых определяются етодом максимального правдоподобия. Восстановленные таким образом плотности вероятности используются для вычисления правдоподобия принадлежности некоторого вектра акустических параметров голосу диктора, подлежащего верификации.

Тестирование системы верификации показало, что минимальный объем обучающей выборки должен содержать не менее 40 произнесений каждого слова. Обучение системы происходит сессиями, в каждой из которых диктору в случайном порядке предъявляются к произнесению по 10 раз каждое слово из словаря числительных. Желательно, чтобы между сессиями прошло время в несколько часов или дней. Это обеспечивает разумную представительность выборки для аппроксимации плотности вероятности в пространствах признаков.

Поскольку характеристики голоса подвержены непрерывным изменениям, то все известные системы верификации требуют переобучения каждые 2 – 3 недели. В описываемой системе этот недостаток преодолевается скрытым дообучением, при котором статистические характеристики диктора пересчитываются каждый раз, когда произошла его успешная верификация. Для того, чтобы избежать чрезмерного размытия плотностей вероятности признаков, по мере накопления новых данных из статистики изымаются старые данные. Тем самым обеспечивается непрерывное отслеживание изменения характеристик голоса без риска увеличения ошибки пропуска самозванца.

При таких предположениях апостериорная вероятность принадлежности произнесенного слова голосу данного диктора может быть найдена как

$$P = \frac{\prod_{i=1}^{I} W_i}{1 + \prod_{i=1}^{I} W_i},$$

где W_i – отношение правдоподобия принадлежности к голосу данного диктора в i-м подпространстве этого слова, I – число подпространств. Отношение правдоподобия W_i вычисляется поочередно для каждого диктора из референтной базы данных, и апостериорная вероятность P определяется как минимальное значение по всем референтным дикторам, характеризуя меру отличия от ближайшего референтного диктора.

Разбиение пространства признаков на подпространства малой размерности имеет еще одно преимущество. Если по каким-то причинам в некотором подпространстве признаки не могут быть измерены, то это подпространство исключается из вычисления апостериорной вероятности, и решение принимается по остальным подпространствам. Таким образом, отсутствующие признаки не приводят к отказу от вычисления принадлежности слова к голосу диктора.

Верификация по одному слову не может обеспечить высокую надежность, хотя и удобна для пользователя в процессе эксплуатации. Пароль должен состоять из нескольких слов, причем наилучшая эффективность системы верификации достигается в том случае, когда диктору сообщается, какое слово он должен произнести. Тогда апостериорная вероятность принадлежности произнесенной последовательности слов голосу данного диктора вычисляется как

$$P = \frac{\prod_{k=1}^{K} W_k}{1 + \prod_{k=1}^{K} W_k},$$

где K – число слов в парольной фразе, W_k – правдоподобие принадлежности k-го слова голосу данного диктора. Вычисленная таким образом апостериорная вероятность сравнивается с порогом, который индивидуален для каждого диктора и определяется по обучающей выборке как наименьшая величина P. Этот порог в данном исследовании определялся из условия минимума суммарной ошибки пропуска самозванца и ложного отказа.

Последовательно предъявляя диктору слова, которые он должен произнести, система верификации имеет возможность компенсации внезапных помех, ошибок произнесения, а также противодействовать попыткам вторжения с помощью заранее записанных слов.

Отношение сигнал/шум оценивается как отношение среднего уровня сигнала в канале на интервале паузы до начала произнесения слова к среднему уровню речевого сигнала. Если это

отношение меньше заданного порога, например +12 дБ, то анализ произнесенного слова не производится, и осуществляется переспрос. Таким образом, избегаются ошибки, возникающие при слишком тихом произнесении. Система верификации сигнализирует о низком отношении сигнал/шум, и необходимости говорить либо громче, либо ближе к микрофону. Если интерфейс общения с системой верификации включает монитор, то при малом отношении сигнал/шум, например, не изменяется цвет слова, которое диктор должен произнести. Это означает, что диктор должен повторить это слово. Если же слово произнесено слишком громко, что чревато перегрузкой АЦП, то включается индикатор перегрузки, также означающий необходимость повторения этого слова несколько тише.

В описываемой системе верификации методом спектрального вычитания выполняется компенсация амплитудно-частотных искажений сигнала вследствие присутствия внешних шумов. Внезапные помехи высокого уровня и неизвестного спектрального состава таким способом компенсировать нельзя. Если в результате действия таких помех апостериорная вероятность P_k принадлежности произнесенного слова к голосу данного диктора оказывается ниже некоторого порога, вычисленного в процессе обучения, то P_k не принимается к вычислению накопленной апостериорной вероятности парольной фразы, а это слово вновь предъявляется диктору через некоторое время. Аналогично, k-е слово вновь предъявляется диктору, если в процессе сегментации наилучшая мера сходства с эталоном из референтной базы данных окажется ниже некоторого порога.

В результате система верификации оказывается устойчивой к посторонним разговорам, музыке, телефонным звонкам и другим помехам. Она не требует, как подавляющее большинство известных систем, полной тишины при общении с пользователем. Так она удовлетворяет реальным условиям эксплуатации.

Если последовательность слов в пароле не фиксирована, и система верификации генерирует ее каждый раз в случайном порядке, то тем самым исключается возможность злонамеренного вторжения путем воспроизведения записанного путем подслушивания пароля. Злоумышленник может вырезать из записанной парольной фразы отдельные слова, и воспроизводить их в ответ на запрос системы верификации. Для этого нужно, однако, чтобы в такой записи присутствовали все слова из заданного словаря. Простейший способ противодействия такой тактике состоит в повторном запросе на произнесение нескольких слов из уже произнесенных. Проверка на тождественность речевых сигналов может показать, что они принадлежат одной записи, тогда как повторное произнесение одного и того же слова человеком всегда несколько отличается от предыдущего произнесения.

Если верификация выполняется при передаче речевого сигнала по цифровому каналу связи, то противодействие воспроизведению перехваченных слов может быть реализовано путем вставки в речевой сигнал "водяных знаков" в виде, например текущей даты и времени произнесения с последующей проверкой на совпадение на приемном конце.

В результате принятых мер противодействия максимально затрудняется злонамеренное вторжение путем воспроизведения подслушанных слов. И хотя нельзя гарантировать 100% защиту от такого вторжения, может оказаться, что технические трудности преодоления принятых мер в ряде случаев приведут к отказу от попыток вторжения.

В референтную базу дикторов могут быть включены и вероятностные распределения коллектива дикторов, имеющих доступ к определенным ресурсам. В описываемой системе верификации автоматически определяется диктор, ближайший по вероятности к параметрам данного пользователя. Если произошел отказ от верификации, то можно установить, не пытался ли кто-нибудь из коллектива дикторов незаконно получить доступ ресурсам данного диктора.

Важной характеристикой системы верификации является ее реакция на многократный отказ подтвердить личность пользователя. Окончательный отказ, например, в допуске к информационным ресурсам, финансовым операциям или запуске двигателя автомобиля, независимо от его причины, может причинить серьезные неудобства пользователю. В зависимости от условий эксплуатации системы верификации, могут использоваться различные меры по выяснению причин такого отказа. Если, например, верификация используется для санкционирования доступа к операционной системе компьютера в какой-то организации, то такой отказ, прежде всего, является поводом для персональной проверки личности пользователя. Если этот человек действительно имеет право доступа, то отказ системы верификации может быть преодолен с помощью кода, который известен только представителю службы безопасности.

Конечно, в таком случае возникает вопрос о возможности утечки информации о таком коде, но эта проблема решается уже другими средствами.

Если верификация выполняется по каналу связи, то в случае отказа признать голос пользователя инициируются стандартные методы проверки личности, использующие, например, систему вопросов со стороны оператора.

Блок-схема.

На Рис.1 показана структура процессов верификации.

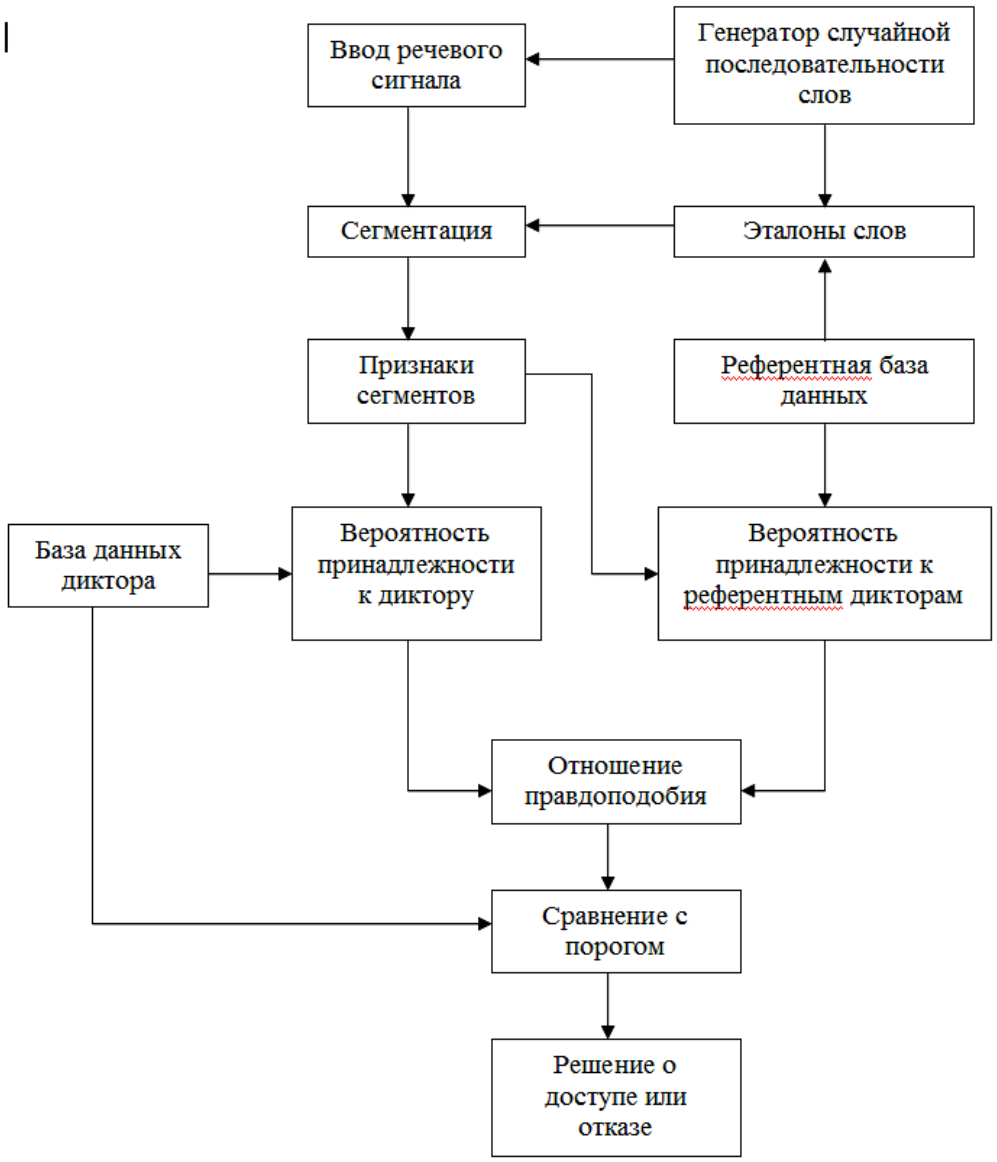

Рис. 1. Блок-схема системы верификации.

Результаты тестирования.

Доверие к показателям эффективности, декларируемым разработчиками систем верификации диктора, зависит от объема и типа базы речевых данных, на которой проводилось испытание системы. Объем базы данных и число проведенных испытаний определяют доверительный интервал вероятностей ошибок. Поэтому, если такие сведения не приводятся (а они почти никогда не приводятся), то невозможно оценить достоверность опубликованных характеристик. Примечательно, что иногда не сообщаются и сами вероятности ошибок первого и второго рода.

Состав базы данных определяет возможность распространения декларируемых показателей на реальные условия эксплуатации. Если база данных сформирована для одних и тех же акустических условий, типа микрофона и расстояния до диктора, то другие условия могут катастрофически ухудшить эти показатели. Это типичная ситуация для распознавания речи и диктора, поскольку стандартные математические методы обучения и распознавания, такие, как скрытые Марковские модели, весьма чувствительны к различию в условиях обучения и распознавания.

Разработка и тестирование описываемой системы верификации производились на неоднородной базе данных, содержащей произнесения более 400 дикторов (243 мужчины и 186 женщин). Для разных групп дикторов использовались приемники звука разного типа и разных производителей, с разными расстояниями до диктора, в разных помещениях и с разными звуковыми картами. Всего применялось 5 типов микрофонов и 2 типа телефонных трубок. Среднее число произнесений для каждого диктора составляло около 400. Такая структура базы позволяет рассчитывать на сохранение в реальных условиях показателей, полученных при тестировании.

В процессе тестирования каждый диктор из базы данных по очереди назначался целевым диктором, и ошибки первого и второго рода определялись в два этапа. При оценке вероятности отказа от подтверждения личности, из обучающей выборки этого диктора поочередно изымались произнесенные им слова в количестве от одного до десяти, перестраивалась плотность вероятности в пространствах признаков, и вычислялась вероятность принадлежности изъятой группы слов к голосу целевого диктора. Затем изъятые слова возвращались в обучающую выборку, в случайном порядке формировалась и изымалась другая группа слов, и вновь повторялся процесс распознавания. Это так называемый скользящий алгоритм распознавания ("*jack knife*"), который позволяет довольно корректно использовать одну и ту же выборку данных как для обучения, так и для распознавания.

При оценке вероятности пропуска самозванца из выборки для каждого референтного диктора в случайном порядке формировалась последовательность слов, также содержащая от одного до десяти слов. Признаки этих слов использовались для оценки правдоподобия принадлежности к голосу целевого диктора.

Всего было произведено более 30 миллионов тестов, что соответствует доверительному интервалу менее ±0.001%. Поскольку характеристики мужских и женских голосов различаются настолько, что не имеет смысла их сравнивать, оценки вероятности ошибок выполнялись отдельно для мужчин и женщин. Кстати, если в сообщениях о характеристиках системы верификации не указываются раздельные показатели для мужчин и женщин, то, скорее всего, приводятся средние показатели.

В Табл. 6 показана вероятность (в процентах) суммарной ошибки ложного пропуска и ложного отказа, определенная при работе системы с критерием минимума этой ошибки.

Таблица 6. Суммарная ошибка в зависимости от числа слов в пароле (%).

число цифр в пароле	1	2	3	4	5	6	7	8	9	10
ошибка, %, мужчины	8.49	2.58	0.97	0.43	0.18	0.10	0.054	0.021	0.013	0.006
ошибка, %, женщины	10.92	3.70	1.52	0.66	0.35	0.20	0.097	0.060	0.041	0.025

Таким образом, достоверные оценки суммарной ошибки при длине пароля в 10 слов для мужских голосов составляют около 0.006%, а для женских голосов – около 0.025% , т.е. вероятность суммарной ошибки есть $6 \cdot 10^{-5}$ и $2.5 \cdot 10^{-4}$, соответственно. Эти же данные представлены на Рис. 2 для того, чтобы наглядно показать скорость падения ошибки в зависимости от числа слов в пароле. Как видно, суммарная ошибка для женщин в среднем заметно больше, чем для мужчин, и эта ошибка уменьшается медленнее в зависимости от числа слов в пароле.

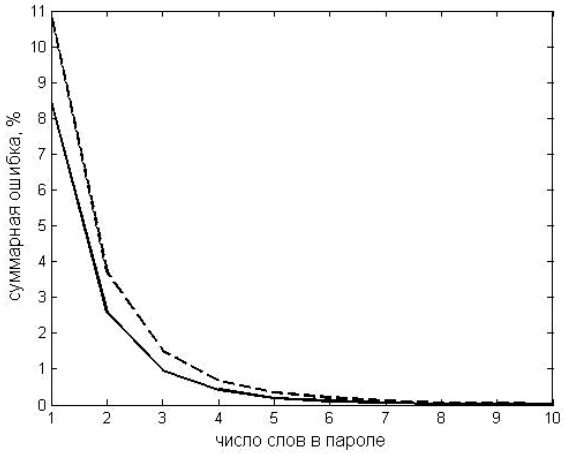

Рис. 2. Зависимость средней суммарной ошибки верификации от числа слов в пароле.

Информативность каждого числительного зависит от его фонетического состава. Разница в ошибках верификации отдельных слов находится в диапазоне от 5.3% до 12%.

Оценка вероятности ошибок в среднем по всем дикторам, однако, не полностью характеризует систему верификации. Известно, что голоса дикторов сильно различаются по индивидуальности. В Табл. 7 представлено распределение дикторов для определенного уровня ошибок. Видно, что подавляющее большинство дикторов распознается с чрезвычайно малой ошибкой, тогда как только 3 – 4% дикторов распознается с заметной ошибкой, и именно такие дикторы ухудшают общую статистику.

Таблица 7. Доля дикторов с заданной максимальной ошибкой. Длина пароля – 10 слов.

максимальное значение суммарной ошибки (%)	0.001	0.01	0.05	0.25	0.50
мужчины (%)	73	17	7	3	0
женщины (%)	72	11	11	3	3

Соотношение между вероятностью пропуска самозванца и ложного отказа показано в Табл. 8, в которой видно, что при критерии минимума суммарной ошибки вплоть до длины пароля в 9 слов вероятность пропуска самозванца заметно превышает вероятность ложного отказа. Для паролей длиной в 9 и 10 слов суммарная ошибка примерно поровну делится между ошибками ложного пропуска и ложного отказа, и она настолько мала, что соотношение между этими ошибками уже не имеет значения.

Таблица 8. Соотношение между ошибками первого и второго рода для мужских голосов.

число цифр в пароле	1	2	3	4	5	6	7	8	9	10
ложные отказы, %	3.54	1.067	0.347	0.168	0.063	0.038	0.015	0.006	0.006	0.003
ложные пропуски, %	4.496	1.513	0.628	0.267	0.121	0.062	0.037	0.015	0.007	0.003

Некоторые возможные приложения системы верификации предусматривают передачу речевого сигнала по мобильному каналу связи. В таких каналах используется сжатие речевого сигнала, сопровождающееся искажением его характеристик и заметной потерей индивидуального качества голоса. Поэтому для верификации нельзя использовать параметры системы, полученные при обучении на неискаженных речевых сигналах. Поскольку все системы сжатия эксплуатируют способность слуха к компенсации определенных искажений, то можно ожидать существенное ухудшение показателей системы верификации, обученной на сигналах со сжатием, по сравнению с системой, использующей натуральный речевой сигнал.

Степень ухудшения качества верификации была определена в экспериментах, в которых вся исходная база данных была преобразована с помощью кодека на 9.6 кбит/с. Затем было выполнено тестирование по схеме, описанной выше. Как и ожидалось, ошибки верификации заметно увеличились (Табл. 9), но в целом метод сохранил свою эффективность.

Таблица 9. Суммарная ошибка в зависимости от числа слов в пароле для кодека (%).

число цифр в пароле	1	2	3	4	5	6	7	8	9	10
ошибка, %, мужчины	12.84	4.69	2.11	1.071	0.6	0.34	0.2	0.13	0.08	0.05
ошибка, %, женщины	16.06	6.44	3.1	1.68	0.98	0.57	0.34	0.22	0.14	0.09

Обсуждение.

Как и любая задача распознавания в речевых технологиях, задача верификации диктора является обратной задачей, поскольку требуется определить индивидуальные характеристики диктора по параметрам речевого сигнала. Измерение этих параметров выполняется с ошибкой, поэтому для достижения желаемой точности верификации нужно использовать ограничения на условия верификации. В эти ограничения входит обучение на основе конкретного языка, фиксированный словарь и генерирование системой верификации последовательности слов, которые должен произнести пользователь. Кроме того, база речевых данных должна быть неоднородной относительно условий записи речевых сигналов, часть ее должна быть вручную размечена на артикуляторно-акустические сегменты, отбор информативных признаков для каждого слова из заданного словаря должен выполняться индивидуально.

При этих условиях в данном исследовании получены результаты, многократно превосходящие все известные показатели систем верификации. Эти результаты служат ориентиром в качестве потенциально достижимых ошибок верификации.

Существуют различные постановки задачи верификации, где в критерий эффективности в разной пропорции входят удобство для пользователя и степень защиты. Удобство для пользователя определяется, прежде всего, вероятностью отказа. Обычно считается, что один отказ на 100 актов верификации приемлем для большинства пользователей. Это означает, что вероятность отказа должна быть не более 1%. В понятие удобства также входит длительность обучения, длительность процесса верификации и необходимость помнить пароль. Может показаться, что пользователю удобно не запоминать пароль, а произносить каждый раз разные слова. Но при этом должна использоваться большая статистика его речи, что требует длительного обучения. Однако и при этом максимальная эффективность не может быть достигнута, поскольку система верификации опирается только на средне-статистические характеристики, а не на конкретные параметры произнесенных слов. К тому же совсем не обязательно, что необходимость произносить каждый раз разные слова удобна для пользователя, поскольку заставляет его придумывать новую фразу.

С другой стороны, использование фиксированного пароля существенно снижает устойчивость системы к злонамеренному вторжению путем записи и воспроизведения речевого сигнала. Поэтому с точки зрения удобства и надежности системы верификации представляется оптимальным использование фиксированного словаря из хорошо знакомых слов, которые в случайном порядке предъявляются пользователю для произнесения. Такая подсказка может быть реализована визуальными или звуковыми средствами.

С точки зрения удобства эксплуатации желательно, чтобы длина пароля была как можно меньше. Из наших исследований следует, что пароль, содержащий лишь одно слово, даже в наилучших условиях не обеспечивает сколько-нибудь приемлемой ошибки верификации. Похоже, что в условиях, когда защита выполняется только с помощью верификации по голосу, базовая минимальная длина пароля должна быть не меньше 4 слов, причем система может добавить несколько слов для компенсации помех. При этом можно существенно снизить вероятность пропуска самозванца за счет увеличения вероятности отказа путем смещения порога принятия решения. Возможно, что при малом риске вторжения самозванца базовая длина пароля может быть снижена до 3 слов. Оценка такого риска, конечно, должна принадлежать пользователю, а система верификации лишь предоставляет возможность управления длительностью пароля, как это сделано в описываемой системе.

Заключение.

Контекстно-независимая система верификации для фиксированного словаря числительных русского языка с базовой длиной парольной фразы в 10 слов и подсказкой текущего слова обеспечивает вероятность суммарной ошибки пропуска самозванца и ложного отказа $6 \cdot 10^{-5}$ для мужских голосов и $2.5 \cdot 10^{-4}$ - для женских. Эти показатели получены в результате отбора спектрально-временных признаков для каждого числительного. Они на два порядка лучше известных по литературным источникам, и могут служить референтными показателями в качестве потенциально достижимых. Специальные алгоритмы взаимодействия с пользователем компенсируют динамические помехи, в том числе посторонние разговоры и музыку.

СПИСОК ЛИТЕРАТУРЫ

1. Jain A., Ross A., Pankanti S. (2006). Biometrics: A Tool for Information Security. IEEE Transactions on Information Forensics and Security, vol. 1, No. 2.
2. http://www.nist.gov/speech/tests/sre/2008/official_results/index.html.
3. Рамишвили Г.С. (1981). *Автоматическое распознавание говорящего по голосу*. М.: Радио и связь, 221 с.
4. Kersta L.G. (1962). Voiceprint identification. Nature, v. 196, N4861.
5. Furui S. (1981). Cepstral analysis techniques for automatic speaker verification. IEEE Tran. Acoust., Speech, Signal Processing, v. 27, 254-277.
6. Rabiner L.R., Juang B.-H. (1993). *Fundamentals of speech recognition*. PTR Prentice Hall, Englewood Cliffs, N.Y.
7. Reynolds D.A. (1995). Speaker identification and verification using Gaussian mixture speaker models. Speech Communication, v. 17, 91-108.
8. Reynolds D.A., Quatieri T.F., Dunn R.B. (2000). Speaker verification using adapted Gaussian mixture models. Digital Signal Process., v. 10, 19-41.
9. Furui S. (1996). An overview of speaker recognition technology. In: Lee C.H, Soong F.K., Paliwal K.K. (eds). *Automatic speech and speaker recognition.* Kluwer Academic, Boston, Chapter 2.
10. Yegnanaryana B. (1999). *Artificial neural networks*. Prentice Hall., New Deli, India.
11. Сорокин В.Н., Цыплихин А.И. Способ верификации пользователя в системах санкционирования доступа. Патент № 2351023, приоритет от 2 мая 2007.
12. Lavner Y., Gath I., Rosenhouse J. (2000). The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. Speech Communication, v. 30, 9-26.
13. Сорокин В.Н.. *Теория речеобразования*. (1985). Радио и связь, М., 313 с.
14. Леонов А.С., Макаров И.С., Сорокин В.Н., Цыплихин А.И. (2004). Артикуляторный ресинтез фрикативных. Информационные процессы, т.4, №2, 141-159. *www.jpg.ru*.
15. Цыплихин А.И., Сорокин В.Н. (2006). Сегментация речи на кардинальные элементы. Информационные процессы, т. 6, №3, 177-207. *www.jpg.ru*.
16. Сорокин В.Н., Чепелев Д.Н. (2005). Первичный анализ речевых сигналов, Акустический ж., т. 51, №4, 536-542.
17. Сорокин В.Н.. *Синтез речи*. (1992). Наука, М., 392 с.
18. Navratil J., Jin Q., Andrews W., Campbell J. (2003). Phonetic speaker recognition using maximum likelihood binary decision tree models. Proc. ICASSP, v. 4, 769-799.
19. Леонов А.С., Макаров И.С., Сорокин В.Н. (2009). Частотные модуляции в речевом сигнале. Акустический ж., № т. 55, №6, 809-821.
20. Леонов А.С., Макаров И.С., Сорокин В.Н. (2009). Устойчивость оценок формантных частот. Речевые технологии, № 1, 3-18.
21. Цыплихин А.И. (2007). Анализ импульсов голосового источника. Акустический журнал, т. 53, №1, 119-133.
22. Dempster A. P., Laird N. M., Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J. of the Royal Statistical Society, Series B, , N. 34, 1–38.
23. McLachlan G., Peel D. (2000). Finite Mixture Models. New York: John Wiley & Sons Inc.
24. Naonori U., Ryohei N., Ghahramani Z., Hinton G.E. (2000). SMEM Algorithm for Mixture Models. Neural Computation, v. 12, N. 9, 2109-2128.