ФЕДЕРАЛЬНОЕ АГЕНТСТВО

ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСТ Р 53452— 2009/ISO/TR 19358:2002

Эргономика

ПРОЕКТИРОВАНИЕ И ПРИМЕНЕНИЕ ИСПЫТАНИЙ РЕЧЕВЫХ ТЕХНОЛОГИЙ

ISO/TR 19358:2002
Ergonomics — Construction and application of tests for speech technology (IDT)

Издание официальное

БЗ 9—2009/561

Москва
Стандартинформ
2010

gost.e-ecolog.ru

Предисловие

Цели и принципы стандартизации в Российской Федерации установлены Федеральным законом от 27 декабря 2002 г. № 184-ФЗ «О техническом регулировании», а правила применения национальных стандартов Российской Федерации — ГОСТ Р 1.0—2004 «Стандартизация в Российской Федерации. Основные положения»

Сведения о стандарте

1 ПОДГОТОВЛЕН Автономной некоммерческой организацией «Научно-исследовательский центр контроля и диагностики технических систем» (АНО «НИЦ КД») на основе собственного аутентичного перевода стандарта, указанного в пункте 4

2 ВНЕСЕН Техническим комитетом по стандартизации ТК 201 «Эргономика»

3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Приказом Федерального агентства по техническому регулированию и метрологии от 7 декабря 2009 г. № 572-ст

4 Настоящий стандарт идентичен международному стандарту ISO/TR 19358:2002 «Эргономика. Проектирование и применение испытаний речевых технологий» (ISO/TR 19358:2002 «Ergonomics — Construction and application of tests for speech technology»)

5 ВВЕДЕН ВПЕРВЫЕ

Информация об изменениях к настоящему стандарту публикуется в ежегодно издаваемом информационном указателе «Национальные стандарты», а текст изменений и поправок — в ежемесячно издаваемых информационных указателях «Национальные стандарты». В случае пересмотра (замены) или отмены настоящего стандарта соответствующее уведомление будет опубликовано в ежемесячно издаваемом информационном указателе «Национальные стандарты». Соответствующая информация, уведомление и тексты размещаются также в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет

© Стандартинформ, 2010

Настоящий стандарт не может быть полностью или частично воспроизведен, тиражирован и распространен в качестве официального издания без разрешения Федерального агентства по техническому регулированию и метрологии

Содержание

1 Область применения 1
2 Термины и определения 1
3 Описание речевых технологий 2
4 Описание важных переменных речевых технологий 4
5 Методы оценки 5
Приложение А (справочное) Пример оценки 9
Приложение В (справочное) Критерии качества работы 11
Библиография 11

Введение

В настоящем стандарте установлены методы определения систем речевых технологий (программ автоматического распознавания речи, систем преобразования текста в речь и устройств, использующих речевой сигнал) и выбора соответствующих процедур испытаний. Речевое общение человека с человеком в настоящем стандарте не рассматривается (см. ИСО 9921:2003 «Эргономика. Оценка речевой связи»).

Международный стандарт, на основе которого подготовлен настоящий стандарт, разработан техническим комитетом ИСО/ТС 159 «Эргономика».

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

Эргономика

ПРОЕКТИРОВАНИЕ И ПРИМЕНЕНИЕ ИСПЫТАНИЙ РЕЧЕВЫХ ТЕХНОЛОГИЙ

Ergonomics.
Construction and application of tests for speech technology

Дата введения — 2010—12—01

1 Область применения

В настоящем стандарте установлены методы испытаний и оценки товаров и услуг, связанных с речевыми технологиями. Стандарт предназначен для специалистов в области речевых технологий, а также покупателей и пользователей таких систем.

2 Термины и определения

В настоящем стандарте использованы следующие термины с соответствующими определениями:

2.1 **автоматическое распознавание речи** (automatic speech recognition; ASR): Способность системы принимать входную информацию в виде человеческой речи.

2.2 **диалог** (dialogue): Интерактивный обмен информацией между речевой системой и говорящим человеком (пользователем).

2.3 **управление диалогом** (dialogue management): Управление речевым диалогом между системой и человеком.

2.4 **обработка естественного языка** (natural language processing; NLP): Автоматическая обработка текста, создаваемого человеком.

2.5 **объективная оценка** (objective assessment): Оценка, обычно полученная без прямого участия человека в процессе измерений на основе предварительно записанной речи.

2.6 **критерии качества работы** (performance measures): Способы оценки работоспособности системы, обычно с использованием методов диагностики или оценки относительной эффективности.

2.7 **обучаемая система (распознавания речи)** (speaker-dependent system): Система распознавания речи, для работы которой необходимо обучение системы на речи конкретного пользователя.

2.8 **идентификация говорящего** (speaker identification): Идентификация конкретного пользователя среди ограниченного набора возможных пользователей.

2.9 **система распознавания речи** (speaker-independent system): Система, не требующая обучения на речи конкретного пользователя, пригодная для любого пользователя из выбранной группы (носителей языка, взрослых и т. д.).

2.10 **распознавание речи пользователя** (speaker recognition): Основной элемент технологий, идентифицирующих или верифицирующих идентификацию пользователя.

2.11 **верификация идентификации пользователя** (speaker verification): Проверка идентификации пользователя с помощью оценки особенностей речи.

2.12 **манера речи** (speaking style): Особенности речи, такие как прерывистость или непрерывность, чтение напечатанного текста, импровизация и т. п.

2.13 **речевое взаимодействие** (speech communication): Обмен информацией с помощью возможностей речи (тон и тембр, модуляции голоса).

Примечание — Речевое взаимодействие предусматривает краткие тексты, предложения, группы слов, отдельные слова, речь с запинанием, мямленьем и части слов.

Издание официальное

2.14 **программа распознавания речи** (speech recognizer): Программное средство, обеспечивающее распознавание речи.

Примечание — Это процесс, с помощью которого компьютер трансформирует акустический речевой сигнал в текст.

2.15 **синтез речи** (speech synthesis): Генерация речи на основе данных.

2.16 **понимание речи** (speech understanding): Технология извлечения семантического содержания речи.

2.17 **субъективная оценка** (subjective assessment): Оценка, обычно полученная с непосредственным участием людей в процессе измерений.

2.18 **синтез речи по тексту** (text-to-speech synthesis): Генерация слышимой речи на основе текста.

2.19 **словарь** (vocabulary): Набор слов, используемых в контексте.

2.20 **объем словаря** (vocabulary size): Количество слов в словаре программы распознавания речи.

3 Описание речевых технологий

3.1 Введение

Речевые технологии предусматривают автоматическое распознавание речи, говорящего пользователя, а также синтез речи и т. п. Обработка естественного языка (NLP) включает в себя понимание элементов текста и управление диалогом между пользователем и машиной. Современные технологии, по большей части, основаны на алгоритмах, которые используют обработку цифрового сигнала с помощью процессора цифровых сигналов или (персональной) компьютерной системы. Алгоритмы формируют ответы практически в реальном времени. Производительность зависит от применения. Например, система распознавания речи с небольшим объемом словаря, обученная для работы с речью одного пользователя (например, при управлении личным портативным телефоном), намного удобнее (для этого пользователя), чем система, имеющая большой объем словаря и разработанная для большой группы неизвестных пользователей (например, система информационных услуг общественной телефонной сети).

Для товаров и услуг, использующих речевые технологии, можно выделить 4 главных направления применения:

a) управление и контроль. Взаимодействие между пользователем и системой осуществляется с помощью автоматического распознавания речи (ASR). ASR обычно используют при мультимодальном проектировании, в котором речевое управление системой является одним из возможных способов управления (т. е. клавиатура, мышь, сенсорный экран и т. д. могут быть альтернативными средствами). Управление при помощи системы ASR характерно для ситуаций, когда руки оператора заняты.

b) услуги и телефонные приложения. Услуги, такие как информационный киоск, обычно требуют сочетания управления распознаванием, пониманием, синтезом речи и диалогом для управления неконтролируемым диалогом пользователя с системой. Существующие современные системы охватывают относительно простые структуры диалога, такие как туристические информационные системы (день, время, «откуда—куда») и телефонные центры (выбор требуемой информации).

c) генерация документов. Системы речевого ввода текста, обученные для работы на нескольких языках, уже присутствуют на рынке. Эти системы могут использовать стандартные системы обработки текста. Простые применения предусматривают ввод данных установленного вида (например, медицинские отчеты), более сложные системы позволяют диктовать целые документы и управлять системой обработки текста. Эти более сложные системы обычно обучены для работы со словарем большого объема и зависят от особенностей речи пользователя. Однако для обеспечения приемлемой производительности система должна быть знакома с пользователем и областью использования. Обычно это выполняют в два этапа: с помощью адаптивного акустического сеанса обучения, в котором пользователь должен читать установленный текст, и представлением нескольких документов, написанных для пользователя, которые предназначены для расширения словаря и корректировки языковой модели.

d) поиск документов. Поиск готовых документов (в архиве речевых документов), поиск информации или определенных отрывков документов или высказываний определенного пользователя представляет интерес для архивного документирования и управления, а также для компиляции кратких обзоров. Различные технологии используют для маркировки произношения, например в ASR, определения слова и распознавания говорящего. Определенные алгоритмы поиска используют для восстановления запрошенной информации.

3.2 Доступные технологии

3.2.1 Распознавание речи

Системы автоматического распознавания речи способны производить транскрипцию текстовой строки речи. Для этой цели используют обученные системы. Современные системы, использующие словарь большого объема, извлекают из речи установленные спектральные параметры, которые идентифицируют ее подэлементы (фонемы). При этом слова описывают в виде цепочки этих фонем. Схема распознавания может использовать фонемы различных уровней, относящихся к фонетическим моделям, словам (словарю) и статистическому описанию словесных комбинаций (языковой модели). Обучение фонетическим моделям необходимо для работы с большим количеством пользователей, что приводит к основанному на статистике представлению. Статистический подход обычно основан на скрытом марковском моделировании (HMM[1)]) или нейронной сети (NN[2)]). Для составления словаря и языковой модели обычно используют доступный текст в цифровой форме, который является репрезентативным для области применения.

3.2.2 Идентификация и верификация пользователя

Автоматическая идентификация пользователя — это способность системы распознать пользователя в группе известных пользователей. Она отвечает на вопрос: «Кому принадлежит данный образец речи?». Метод включает в себя два этапа: моделирование речи популяции пользователей (обучение) и сравнение неизвестной речи с моделями речи всех пользователей (тестирование).

Верификация пользователя — это метод проверки того, что говорящий является тем, за которого он себя выдает. Основой системы верификации пользователя является алгоритм, сопоставляющий высказывание пользователя с моделью, построенной в процессе обучения на основе авторизованных зарегистрированных высказываний пользователя. Если речь соответствует модели в пределах допустимых отклонений, то система признает пользователя соответствующим заявленной личности. Для защиты от самозванцев, пытающихся обмануть систему, используют запись голоса зарегистрированного пользователя. При этом для верификации система обычно требует от пользователя произнести установленную фразу, например последовательность чисел, выбранных случайным образом каждый раз, когда пользователь пытается получить доступ. Для обеспечения достоверности верификации система верификации обращается к системе распознавания.

3.2.3 Синтез речи

Для синтеза речи используют два метода: первый, обычно называемый «консервированной речью», генерирует речь на основе сохраняемых сообщений. Чтобы сохранить пространство памяти, обычно используют методы кодирования для сжатия сообщений. Такой метод синтеза позволяет получить высококачественную речь, особенно в приложениях с быстрым откликом, где используют набор стандартных ответов. Второй метод — это «синтез речи по тексту». Он позволяет генерировать сообщение по написанному тексту. Обычно он включает в себя первый этап лингвистической обработки, на котором исходный текст преобразуется во внутреннее представление с помощью фонемических и интонационных маркеров, и второй этап генерации звука на основе этого представления. Генерация звука может быть выполнена либо полностью по правилам сложных моделей, обычно используемых для описания речи (форматный синтез, интонация), либо с помощью соединения коротких, предварительно сохраненных элементов речи (соединительный синтез). Качество речи, полученной при соединительном синтезе, обычно более высокое.

3.2.4 Понимание речи

Системы понимания речи могут быть отнесены к одному из двух видов. Первый вид рассматривает взаимодействие человека с машиной. В этом случае человек и машина работают совместно над решением конкретной проблемы. Интерактивная природа задачи дает возможность машине задать вопрос в случае, когда она не понимает намерений пользователя. В свою очередь пользователь может перефразировать запрос или команду. Системы второго вида предназначены для извлечения необходимой информации из речи, без возможности обратной связи или взаимодействия (например, при резюмировании разговорной документации).

3.2.5 Управление диалогом

Диалогом обычно считают взаимодействие двух партнеров, во время которого некоторая информация поступает от одного к другому. Более полезно рассматривать диалог как начало одним из партнеров обмена информацией для достижения определенной цели. Поэтому партнеров в диалоге следует рассматривать асимметрично: одного — как инициатора диалога, другого — как получателя информации.

[1)] HMM — Hidden Markov Model.

[2)] NN — Neural Network.

При этом диалог успешно завершен, если инициатор считает, что получатель находится в состоянии, для достижения которого предназначался диалог. Целевое состояние может состоять в том, что получатель имеет некоторую информацию или выполняет задание в интересах инициатора. Возможно, что единственное сообщение прошло от инициатора к получателю и имело желаемые последствия, наблюдаемые инициатором.

4 Описание важных переменных речевых технологий

4.1 Введение

На пригодность речевых и лингвистических систем влияют различные факторы. Поэтому оптимальное использование системы может быть связано с условиями ее применения. Для оптимизации использования системы необходимо определить связанные с заданием характеристики и требования к производительности системы до ее проектирования. К важным характеристикам относятся требования к типу речи пользователя, производственному заданию, обучению, окружающей среде, устройствам ввода и системе.

4.2 Тип речи

Отдельные слова: ряд слов, произносимых отдельно; часто используется для задач управления, контроля или ввода данных. Краткие паузы указывают границы слов.

Слитно произносимые слова: ряд слов, произносимых без пауз; часто используется для управления, контроля или ввода данных, таких как ряды чисел. Такие системы обычно проходят обучение на отдельных словах.

Чтение текста: речь, читаемая непрерывно, например, чтение книги без пауз.

Диктовка: речь, читаемая непрерывно, но с управляемой скоростью и особым вниманием к правильности произношения. Пользователь осведомлен о работе с автоматическим распознаванием речи.

Произвольная речь: разговорная речь, включающая все виды прерываний, таких как кашель, неуверенность, замедление и т. д. Обычно это ситуации, когда пользователь не осведомлен о распознавании речи.

4.3 Аспекты, зависящие от пользователя

Зависимость от пользователя: зависимость от пользователя имеет значение для системы, предназначенной и обученной для работы с одним пользователем или небольшой группой пользователей. Для системы, предназначенной и обученной для работы со многими пользователями, в том числе с теми, которых не было при обучении системы, имеет значение независимость от пользователя.

Пол: речь мужчин и женщин обычно отличается по основной частоте (высоте тона) и спектру. Это может оказывать влияние на распознавание, если система не обучена для работы с пользователем данного пола.

Возраст: возраст пользователя, так же, как и пол, оказывает влияние на высоту тона и спектр голоса. Группировка пользователей по возрасту может охватывать следующие периоды: 12—18, 19—22, 22—65 лет. Однако в пределах каждой группы возможны существенные отклонения по характеристикам речи. Для пользователей в возрасте моложе 12 и старше 65 лет могут иметь место очень большие индивидуальные вариации.

Голосовое усилие: уровень речевого сигнала зависит от голосового усилия пользователя. Голосовое усилие определяется измерением эквивалентного непрерывного уровня звукового давления речи, измеряемого на расстоянии одного метра напротив рта.

Темп речи: количество элементов речи, произносимых за установленный промежуток времени, количество слов в минуту или слогов в секунду. Нормальный темп составляет 3—5 слогов в секунду.

Родной язык, акцент: как правило, уровень распознавания ниже для пользователей, говорящих на неродном языке, и пользователей с сильным акцентом.

4.4 Задание (специализированное описание важных параметров распознавания)

Объем словаря: объем словаря зависит от решаемой задачи. Для выполнения задач управления и контроля может быть достаточно от 15 до 100 слов. Для распознавания речи с большим словарным запасом может потребоваться 50000 слов и более. В последнем случае возможно использование слов, не представленных в словаре (OOV[1)]).

Сложность синтаксиса: для древовидной структуры команд с вложенным меню достаточно ограниченного набора слов. Количество возможных альтернатив на каждом уровне должно соответствовать сложности задачи.

[1)] OOV — Out-Of-Vocabulary words.

Структура диалога: следует определить начало диалога и его порядок. В случае ошибки распознавания система может впасть в непредвиденное состояние. Возврат к предыдущему состоянию требует осведомленности (неподготовленного) пользователя о возможности такой ситуации.

Управление корректировками: в случае ошибок (допущенных пользователем или системой) должна быть возможность их исправления. Это может быть достигнуто простой корректировкой (использование «команд корректировки») или сложной (восстановление диалога из непредвиденного состояния).

4.5 Обучение, связанное с заданием

Зависимость от пользователя: система обучена для работы с одним пользователем или ограниченной группой пользователей. Для программы распознавания слов обычно проводят индивидуальное обучение для работы с каждым пользователем.

Независимость от пользователя: система обучена работе с большой речевой базой данных. База данных состоит из образцов речи многих пользователей (до 50—100 часов речи). Обычно обучение выполняется на предприятии.

Адаптация к пользователю: система настроена под конкретного пользователя. Обычно система сначала является независимой от пользователя, а затем адаптируется к определенному пользователю в процессе обучения. Эту особенность часто используют в системах речевого ввода текста.

Тип речи: в зависимости от применения тип речи может охватывать отдельные слова, связанные слова, непрерывную речь или произвольную речь.

4.6 Окружающая среда (требование к качеству речи в особых условиях для ввода и вывода)

Шум: Окружающий шум может искажать речевой сигнал. Для устройства автоматического распознавания речи влияние шума на качество распознавания намного выше, чем для слушателей-людей. Уровень и спектр шума должны быть определены. При синтезе речи понимание речи зависит от способностей слушателя.

Эхо: эхо нарушает звуковой сигнал и ухудшает распознавание. В большинстве случаев для приемлемого уровня распознавания речи системой требуется микрофон с шумоподавлением, расположенный в оптимальном положении у рта говорящего.

Помехи внутри канала: помехи от других речевых сигналов обычно мешают сильнее, чем стационарный шум, так как алгоритм распознавания не может отличить основной речевой сигнал от помехи.

4.7 Ввод (требования к передаче речевого сигнала с микрофона для распознавания)

Микрофон: микрофон может оказывать большое влияние на качество сигнала. Для систем, управляемых через телефонную сеть, качество микрофона со стороны пользователя не определено. Рекомендуется выполнять обучение и испытание системы с одним и тем же типом микрофона, если это возможно. Важным параметром является правильное размещение микрофона.

Искажение: если система интегрирована в сеть, то могут появляться различные искажения. Для телефонной сети обычно устанавливают ограничение по диапазону частот (от 300 до 3400 Гц), что ухудшает качество работы системы при использовании портативных телефонов. Ограничения полосы пропускания, реакция на перегрузку, эхо и шумы в системе являются главными проблемами.

4.8 Требования к модулям речевых технологий

Распознавание: параметры системы распознавания обычно предварительно установлены. В большинстве случаев параметров так много, что невозможно настроить их для оптимальной работы. Важно установить модель словаря и лингвистическую модель. Если используется адаптивная система, то ее производительность может изменяться в процессе использования или испытаний. Поэтому важно сохранять значения значимых параметров при использовании. Если это невозможно, может потребоваться использование системы загрузки начальных параметров.

Управление диалогом: для оценки влияния ошибок системы или пользователя на выполнение задания или исправление ошибок необходимо точное описание структуры диалога.

Вывод речи: параметры системы синтеза речи по тексту, как и для распознавания, настраивают на предприятии. Иногда система предоставляет опции, улучшающие качество речи при произнесении, для имен, адресов и т. д.

5 Методы оценки

5.1 Общие положения

Выполнение методов, использующих речевые технологии, зависит от многих переменных. Некоторые из них являются управляемыми, а другие находятся под воздействием неконтролируемых явлений. Требования к выполнению метода или функционированию системы обычно включают в себя набор переменных с фиксированными значениями параметров. Для оценки системы в конкретном применении необходима процедура оценки характеристик применения.

Спектр стратегий оценки и соответствующих испытаний крайне неоднороден. Прежде всего, термины в области оценки систем с использованием речевых технологий сильно различаются. Различия касаются следующих терминов: оценка (assessment) и сравнительная оценка (evaluation); лабораторные и эксплуатационные методы; прозрачность системы (метод черного и прозрачного ящика, иногда белого или серого ящика); субъективные и объективные испытания. Эти термины не являются полностью независимыми.

Важным источником неоднородности в области стратегий оценки и испытаний является разнообразие областей применения. Динамика текущих исследований, разработок и маркетинга товаров, а также возрастающее разнообразие устройств, связанных с речевыми технологиями, означает, что отдельный продукт зачастую требует новой индивидуальной стратегии оценки с соответствующими испытаниями. Натуральность синтезированной речи требуется, например, в индустрии электронных развлечений или образовательной сфере. Не сложно придумать сценарии, в которых натуральность не является главным критерием и система должна звучать как искусственная система. В таких случаях более важен критерий разборчивости звучания независимо от того, нравится ли маркетологам товар с голосом, похожим на человеческий.

Другой источник неоднородности кроется в увеличивающемся использовании встроенных систем с разговорными устройствами ввода/вывода, иногда с критическими для безопасности функциями. Это подразумевает быстрый рост сложности человеко-машинных интерфейсов, с которыми многие существующие виды оценки не могут справляться и которые требуют большой осторожности при использовании и привлечении экспертов для определения границ применения речевых технологий. Примером в этой области является голосовое управление в реальном времени системами критериев безопасности и автоматизированными системами предупреждения об опасности.

Выбор метода оценки зависит от цели оценки, состоящей в:

a) сравнении различных систем или разных версий одной системы;

b) валидации использования системы для поставленной задачи или установленного критерия;

c) диагностике дисфункций и их происхождения;

d) прогнозе поведения системы в заданных условиях.

5.2 Сравнительная оценка в лабораторных условиях и в условиях эксплуатации

В лабораторных условиях формируют лишь некоторые показатели окружающей среды, в которой используют систему, и не учитывают и не оценивают воздействие остальных показателей, тогда как оценка в условиях эксплуатации исследует фактические показатели системы в ее области применения, окружающей среде, для которой предназначена система. Поэтому система может показать хорошие результаты в лабораторных условиях, но не сможет достичь их в условиях эксплуатации. Ключевой проблемой при переходе от лабораторной оценки к оценке в условиях эксплуатации является устранение шума, возникающего при измерениях в конкретной среде. При этом есть реальные проблемы, возникающие в работе, например, при оценке соответствия системы ее предполагаемому использованию. Оценка в условиях эксплуатации позволяет учесть показатели системы, важные для пригодности ее использования, но не обязательно напрямую связанные с ее функционированием (если эти показатели лежат за порогом приемлемости). Следовательно, некоторые измерения становятся не важны для оценки показателей собственно метода, но больше связаны с эргономикой или даже возможностью реализации.

При выборе условий оценки следует учитывать свойства оценки в условиях эксплуатации и лабораторной оценки:

Оценка в условиях эксплуатации	Лабораторная оценка
Реальное применение	Лабораторное применение
Неконтролируемые условия	Воспроизводимые условия
Дорого	Недорого
Большой набор переменных	Маленький набор переменных
Испытания на пригодность использования	Испытания для оценки параметров технологии
Внешние критерии	Внутренние критерии

Для оценки систем с использованием речевых технологий может быть применено сочетание обоих методов. С помощью калибровки базы репрезентативных данных (например, записанных в ходе испытаний в условиях эксплуатации) могут быть определены значения параметров, которые могут быть использованы при проведении лабораторных экспериментов.

Поскольку естественный язык связан с человеческой психикой, поведение пользователей и их реакция на речевую технологию оказывают значительное влияние на измеряемые показатели в реальных условиях эксплуатации. Например, при тестировании телефонного сервера для заказа железнодорожных билетов в лабораторных условиях было обнаружено, что измеренная интенсивность успешной

транзакции и ее средняя длина значительно выше, чем таковые в условиях эксплуатации, так как в первом случае испытателям оплачено взаимодействие с системой и они не всегда реагировали на несогласованность и повторения в диалоге, тогда как реальные пользователи застревали при первой пропущенной команде в процедуре бронирования билетов. Оценки по результатам лабораторных и эксплуатационных испытаний должны быть коррелированны. Это позволяет оценить качество процедуры оценки систем.

5.3 Прозрачность системы

Прозрачность системы может быть любой: от метода белого ящика до метода черного ящика. При оценке системы методом белого ящика исследователь имеет полный доступ к внутренней работе системы и документации (когда документация недоступна, метод часто называют методом стеклянного ящика). Исследователь имеет возможность выбрать точки для измерений, т. е. точки, между которыми он будет выполнять измерения выбранного репрезентативного параметра функционирования системы. При оценке системы методом черного ящика исследователь рассматривает только взаимосвязь входных и выходных данных системы без учета механизма их связи.

На практике чаще всего исследователь имеет мало возможностей для контроля прозрачности системы, а применяемый метод оценки использует возможности, предлагаемые системой. В некоторых случаях возможно использование метода серого ящика, если в системе предусмотрены точки получения информации, например, когда доступны функции прослеживания или устранения ошибок или могут быть просмотрены модули многократного использования. В этом случае, возможно, исследователю придется строить гипотезу о выполнении системой функции между точками измерений, так как он может иметь лишь ее частичное описание. Стоит отметить, что нет обязательного точного соответствия между действующими модулями, образующими систему, и набором функций, которые могут быть подвергнуты оценке. Например, в любой диалоговой системе управление диалогом является важной функцией, в выполнении которой могут быть задействованы на различных этапах обработки входной информации различные модули.

5.4 Сравнение субъективных и объективных методов

Методы оценки разделяются на субъективные (оценка с прямым участием в измерениях людей) и объективные (оценка без прямого участия в измерениях людей, обычно с использованием предварительно записанной речи). Существуют также методы, представляющие собой их комбинацию. Преимуществом объективных методов является получение воспроизводимых результатов, а также и то, что они автоматизированы по своей природе и, следовательно, более дешевы. Недостаток объективных методов при оценке речевых и языковых применений состоит в том, что они не всегда подходят для понимания естественного языка и речевого взаимодействия. Субъективные методы больше подходят для оценки применений систем с более высоким семантическим или диалоговым содержимым. Недостатком субъективных методов является то, что человек не может выполнять измерения с высокой воспроизводимостью и не может работать с мелкоструктурными шкалами измерений (в среднем человек использует шкалы с градаций не более чем 5—10 уровней). Использование сглаживающих статистических методов, таких как каппа-статистика, для оценки согласованности между экспертами может помочь, но они не устраняют этот недостаток. Кроме того, их использование обычно требует привлечения большего количества испытателей, тем самым увеличивая стоимость оценки.

5.5 Системы распознавания речи

Существует много параметров, характеризующих системы распознавания речи. Однако так как устройства ввода разговорной речи управляются на основе обучения с последующей статистической обработкой, объективные испытания многих видов систем требуют наличия предварительно записанного, четко определенного набора данных, который делят на обучающий набор и испытательный набор, с пропорцией обучающей части к испытательной 9:1, часто с использованием многочисленных испытаний на основе различных сегментов общего набора данных. Очевидно, что испытания системы на основе испытательной части данных демонстрируют верхний предел производительности, который не встречается на практике. До испытаний должно быть проведено достаточное обучение. Не существует общего правила определения «достаточности» обучения. Процедуру обучения определяет изготовитель. Обучение некоторых систем производят не с помощью предварительной записи речи, а с помощью прямого микрофонного ввода. В особых случаях к речи может быть добавлен шум — определенные шумовые сигналы или шум окружающей среды (в офисе или автомобиле). Критическими параметрами, которые оказывают сильное влияние на результаты, являются не только компоненты акустического декодера, но также лингвистические факторы, такие как объем словаря и лингвистическая модель, используемая в системе. Во встроенных системах, таких как программное обеспечение для обработки диктовки, существует много других параметров, включая исправление ошибок, которые не могут быть проверены полностью при испытаниях.

5.6 Системы синтеза речи

Широкий спектр областей применения, упомянутых выше, также касается систем вывода разговорного языка. В процессе разработки могут быть применены объективные испытания, но обычно для встроенных систем применяют субъективные оценки (например, натуральности, приятности, правильности речи) и проверки функционирования (например, разборчивости или распознаваемости звуков). Существуют виды субъективных испытаний, которые предусматривают участие людей. Так же, как для систем речевого ввода, виды и уровни шума являются важными факторами, и голосовая адаптация к специфическому заданию требует аккуратности и внимания: предупреждение, сказанное мягким приятным голосом, может быть не только неуместным, но и неэффективным. Например, в случае, подобном этому, не только вразумительность должна быть проверена, но также и реакция слушателей, которую сложно смоделировать.

5.7 Идентификация и верификация пользователя

Для систем идентификации и верификации пользователя, которые могут быть отнесены к биометрическим системам, главными параметрами являются виды ошибок, т. е. ошибочный прием или отклонение пользователя и действия с ролями пользователя: претендент (предоставленный пользователь), зарегистрированный пользователь (авторизованный пользователь), подлинный пользователь (претендент, который является зарегистрированным пользователем), самозванец (претендент, который не является зарегистрированным пользователем). При ошибочном отклонении подлинного пользователя лишают доступа, а при ошибочном приеме самозванца принимают за истинного пользователя. Количество пользователей и окружающая среда являются критическими факторами как при обучении, так и при испытаниях. В отличие от устройств ввода/вывода разговорного языка система должна уметь забывать, например, если регистрацию пользователя аннулируют. Применение идентификации и верификации пользователей как в реальном времени, так и вне его обычно очень чувствительно к защите. Технологии биометрических систем в настоящее время быстро развиваются, поступают на рынок и становятся очень сложными. При определении процедур испытаний следует использовать ссылки на соответствующие стандарты.

5.8 Совокупность данных

Три вида речевых и языковых совокупностей представляют интерес:

- аналитико-диагностический материал, который, в основном, имеет значение для развития науки и специально разработан для демонстрации специфических фонетических и лингвистических свойств;
- материал «общих целей», который включает в себя наборы слов, являющихся типичными для широкого ряда применений (например, буквенно-цифровые слова или стандартные термины управления);
- материал, ориентированный на конкретную задачу, который отражает различные уровни формализованного разговорного монолога/диалога в пределах ограниченных областей дискурса (речи).

Очевидно, что материал общих целей легко собрать и он полезен для общего понимания, но имеет ограниченное практическое значение. С другой стороны, несмотря на то, что сбор материала, ориентированного на конкретную задачу, является более трудоемким, этот материал имеет значение только для специфической области: такой материал очень полезен для исследовательских целей.

Наличие стандартных совокупностей материала имеет большое значение. Ответственность за координацию, распространение и выпуск соответствующих баз данных должна быть возложена на установленные организации.

5.9 Источники информации

Полезным источником информации является серия справочников по речевым и лингвистическим стандартам, выпущенная экспертной консультативной группой по разработке лингвистических стандартов Европейского союза. Эти справочники охватывают широкую область вопросов, включая создание и обмен электронными лингвистическими ресурсами, такими как речевые и текстовые наборы, формализованные математические модели, лексики и грамматики, а также количественную и качественную оценку систем лингвистической обработки и их компонентов.

В нормативных справочниках, как правило, указаны также подходящие методы, в том числе обзор методов статистической обработки измерений, исследование методов оценки и испытаний. Для многих целей достаточно базовой статистики (среднее стандартное отклонение, стандартная ошибка) и дисперсионного анализа (ANOVA[1)]) в некоторых случаях. Также распространены показатели корреляции и ошибок. Одной из наиболее существенных ошибок статистической обработки является слишком сложная интерпретация или применение неподходящего метода, например, когда смешивают качественные и числовые данные. По сложным вопросам следует консультироваться с экспертами в области математической статистики, так как краткие обзоры таких вопросов, вероятно, будут вводить в заблуждение.

[1)] ANOVA — Analysis Of Variance (дисперсионный анализ).

Приложение А
(справочное)

Пример оценки

А.1 Управление и контроль: управляемый голосом вызов в сети GSM[1)]

Проводится сравнение двух управляемых голосом GSM-телефонов с голосовой системой вызова номера, предназначенных для автомобиля. Для этой цели использована модель автомобиля, в которой установлена система голосового вызова[2)] для GSM-телефона. Акустическое окружение включало в себя фоновый шум с частотным спектром, имитирующим шум в салоне автомобиля. Система голосового вызова оснащена специальным микрофоном с шумоподавлением, который расположен в типичной позиции (50 см от рта испытателя). Телефонная система автомобиля подсоединена к реальной телефонной сети, которая поддерживает услугу голосового вызова. По результатам испытаний определены и сопоставлены показатели голосового ввода по исследуемым моделям GSM-телефонов с применением двух систем голосового вызова. Чтобы исключить нежелательное взаимодействие между сетью и GSM-телефоном, в испытаниях использованы два различных GSM-телефона от двух разных производителей.

В испытаниях участвовали 20 испытателей, не имевших опыта использования систем голосового вызова. Выбор испытателей был сбалансирован с учетом пола и возраста (возраст от 18 до 60 лет).

До начала испытаний испытатели были ознакомлены с руководством пользователя сетевого оператора, предоставляющего услуги голосового вызова. Испытатели должны были читать руководство в течение 10 минут. После обучения испытателей просили выбрать 5 человек среди своих знакомых. Имена этих людей использовались для управляемого голосом вызова. Преимуществом этой процедуры является то, что испытатели произносят имена без ошибок и заминок, кроме того, выборка имен является репрезентативной, поскольку соответствует реальным пользователям. Каждое испытание выполнялось в два этапа:

a) обучение системы по пяти выбранным именам в соответствии с указаниями в руководстве пользователя;

b) проведение испытаний, включающих в себя последовательное произнесение всех пяти имен в случайном порядке.

В процессе испытаний каждый испытатель использовал:

- две системы голосового вызова;
- два GSM-телефона;
- два вида автомобильного шума (80 и 110 км/ч);
- две группы по десять испытателей (мужчины, женщины).

Последовательность изменений условий испытаний подобрана таким образом (система вызова, GSM-телефон, шумовые условия), чтобы избежать привыкания испытателя при сравнении систем вызова.

В процессе испытаний фиксировались время реакции системы с момента начала вызова выбранного имени до момента соединения, а также количество и виды ошибок соединения. Процедура подсчета была основана на штрафной системе. При каждом последовательном вызове не было штрафа, если требуемое соединение удавалось после произнесения нужной команды. Если для соединения требовалось дополнительное взаимодействие с пользователем, были использованы следующие штрафные баллы: подтверждение имени — 1, удаление имени — 2, замена — 5, ошибка обучения — 15. Среднее значение штрафных баллов было вычислено для испытателя и условий испытаний.

Результаты испытаний приведены в таблице А.1. Основной интерес представляет сравнение показателей двух систем голосового вызова. Поскольку среднее значение штрафных баллов для системы А составило 3,1, для системы Б — 5,1, то можно сделать вывод, что система А функционирует лучше. Однако для доказательства этого был проведен дисперсионный анализ (ANOVA), позволяющий оценить значимость различий этих двух результатов. Результат дисперсионного анализа показал, что по предоставленному количеству соединений две системы значимо отличаются по количеству успешных соединений с вероятностью ошибочного решения $p = 0{,}03$.

Следующим показателем функционирования системы является время полного соединения. Среднее время соединения составило 27,7 и 17,4 с для систем А и Б соответственно. Исследование ANOVA показало, что эта разница является существенной с вероятностью ошибочного решения $p = 0{,}001$. Таким образом, по этому показателю система Б работает лучше.

Кроме того, было исследовано влияние на работу системы всех независимых переменных (пол испытателя, комплект GSM и уровень шума). Эти исследования не приведены в настоящем стандарте.

[1)] GSM — Global System for Mobile Communications (Глобальная система мобильной связи).

[2)] Система «свободные руки».

Т а б л и ц а А.1 — Время соединения и показатель ошибок соединения управляемого голосом вызова для двух систем

Система	Пол испытателя	Устройство GSM	Уровень шума	Время соединения, с	Средний штрафной балл	Уровень значимости (значение p)
А	—	—	—	24,7	3,1	0,03
Б	—	—	—	17,4	5,1	
—	мужской	—	—	19,9	3,3	0,08
—	женский	—	—	22,2	4,9	
—	—	А	—	21,2	4,0	0,84
—	—	Б	—	20,9	4,2	
—	—	—	80 км/ч	20,6	3,5	0,07
—	—	—	110 км/ч	21,4	4,7	

А.2 Диктовка. Многоязычное сравнение системы речевого ввода текста

Данный пример касается сравнения системы речевого ввода текста для применений с большим словарем и для отдельных слов. Система разработана для пяти языков (немецкий, испанский, итальянский, французский и английский). Испытания проводились изготовителем. Результаты опубликованы в открытой печати.

Исследуемые системы речевого ввода текста состояли из одинакового базового программного обеспечения, но различных словарей и языковых моделей. Всегда сложно сравнивать системы, основанные на различных языках, так как связанными с языком переменными сложно управлять. Особенности языка и различные обучающие материалы могут влиять на показатели системы.

При испытаниях системы был использован один и тот же текст, переведенный на 5 различных языков. Дополнительно был включен раздел из руководства пользователя системы речевого ввода текста.

Для испытаний на каждом языке были привлечены четыре пользователя, говорящие на родном языке (двое мужчин и две женщины). Продиктованные речевые сигналы были записаны и сохранены с рекомендуемой транскрипцией для выполнения испытаний по автоматической управляемой компьютером программе испытаний. Это позволяло повторять испытание с одним и тем же устройством распознавания, но с различными настройками. Испытания проводились с возможностью адаптации системы к пользователю. Адаптация — это свойство устройства распознавания обучаться распознаванию речи конкретного пользователя. Система была сделана частично зависимой от говорящего пользователя. Испытания включали в себя два режима: с адаптацией и без нее. При условии отсутствия адаптации испытания проводились с исходными заводскими регулировками системы для каждого языка и четверых испытателей, говоривших на родном языке. В режиме адаптации для устройства распознавания было проведено обучение по четырем текстам четырех различных авторов, после чего было проведено распознавание пятого автора. Это было повторено для каждого автора.

Подсчет был выполнен для четырех условий: без адаптации, с адаптацией, для слов, включенных в словарь, и для вероятности омофонической ошибки. Очевидно, что адаптация должна улучшать распознавание текста, однако она требует некоторых усилий для обучения системы работе с конкретным человеком. Также представляет интерес определение показателя работы для слов, включенных в словарь. Этот показатель характеризует работу системы с изученным материалом.

Омофоны (слова, которые имеют одну фонологическую, но разную орфографическую форму, имеют похожее звучание, но пишутся по-разному) должны быть выявлены языковой моделью.

Некоторые результаты этого исследования приведены в таблице А.2. Очевидно, что возможность адаптации дает существенное увеличение производительности системы. Влияние языка на интенсивность появления ошибок распознавания слов также является значимым ($p = 0,01$).

Немецкий язык намного сложнее поддается распознаванию, чем итальянский или английский. Для немецкого языка пределы распознавания с одним и тем же размером словаря намного меньше, чем для английского, из-за наличия множества применяемых форм слов в немецком языке.

Т а б л и ц а А.2 — Интенсивность появления ошибок слов системы речевого ввода текста для некоторых условий эксперимента в зависимости от языка и наличия обучения

Условия эксперимента	Язык				
	Немецкий	Испанский	Итальянский	Французский	Английский
С адаптацией	82	86	89	84	87
Без адаптации	87	89	92	87	91
С адаптацией, слова из словаря	91	92	94	88	91
Омофонические ошибки	22	25	17	73	25

Приложение В
(справочное)

Критерии качества работы

Система распознавания выражений включает в себя системы, разработанные для слов, устных команд, текстовых строк, пользователей и языков. Техническая оценка (т. е. лабораторная оценка) систем распознавания речи обычно использует интенсивность распознавания как показатель качества работы. В связи с этим используют интенсивность ошибок. Под точностью понимают количество ошибок каждого типа (отклонения, ложный ввод, ложная тревога). Вместо общего показателя качества работы системы распознавания может быть использовано (для системы распознавания речи) отклонение несловарных слов. Несловарное слово (OOV) — это слово, сказанное пользователем, но не включенное в словарь системы. Следовательно, OOV не может быть распознано правильно.

При проблемно-ориентированной оценке готовой системы потенциальными пользователями критерии качества работы системы обычно связаны с заданием, т. е. количеством успешных соединений, временем соединения и исправления ошибок. Системы с древовидной структурой ввода могут вызвать дезориентацию пользователя относительно прогноза о завершении задания. Пользователь должен быть осведомлен о состоянии системы. Ситуационная осведомленность играет важную роль для успешного завершения задания или, в случае ошибок, для исправления ошибок.

Интенсивность ошибок распознавания слов находят по формуле

$$w = \frac{i + d + s}{N}, \qquad \text{(B.1)}$$

где w — интенсивность ошибок распознавания слов;
i — количество введенных символов (слов);
d — количество удалений;
s — количество замен;
N — количество слов.

Интенсивность ошибок распознавания слов может быть выражена также в процентах. Стандартное отклонение w (s_w) рассчитывают по формуле

$$s_w = \sqrt{\frac{w(1-w)}{N}}. \qquad \text{(B.2)}$$

Библиография

ISO 9921, Ergonomics — Assessment of speech communication
Cohen J., A coefficient of agreement for nominal scales, Educational and Psychological Measurement, 20, pp. 37—46, 1960
Cohen J., Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit, Psychological Bulletin, (70)4, pp. 213—220
Barnett, J., Bamberg, P., Held, M., Huerta, J., Manganaro, L. and Weiss, A. (1995), Comparative performance in large vocabulary isolated word recognition in five European languages. Proc. Eurospeech '95 Madrid, Spain, pp. 189—192
ELRA (European Linguistic Resources Association), ELRA/ELDA, «gost.e-ecolog.ru
Gibbon, Dafydd, Inge Mertins & Roger Moore, eds. (2000). Handbook of Multimodal and Spoken Language Systems: Resources, Terminology and Product Evaluation. Boston, Dordrecht, London: Kluwer Academic Publishers
Gibbon, Dafydd, Roger Moore & Richard Winski, eds. (1997). Handbook of Standards and Resources for Spoken Language Systems. Berlin: Mouton de Gruyter
King, M. et a l., Evaluation of Natural Language Processing Systems — EAGLES Final Report, EAG-WEG-PR.2, (October 1996), ISBN-87-90708-00-8
Krippendorf, K., Content Analysis: An Introduction to Its Methodology, Sage Publications, Beverly Hills, CA, 1980
LDC (Linguistic Data Consortium), «gost.e-ecolog.ru
Leeuwen, D.A. van, and Steeneken, H.J.M., Handbook of Standards and Resources for Spoken Language Systems, Chapter Assessment of recognition systems, pp. 381—407. Mouton de Gruyter, Berlin, New-York (1997)

Leeuwen, D.A. van, and Steeneken, H.J.M., Handbook of Multimodel and Spoken Dialogue Systems, Chapter: Consumer off-the-shelf (COTS) speech technology product and service evaluation, pp. 204—239. Kluwer academic publisher. Berlin, New-York (2000), ISBN 0-7923-7904-7
Sparck Jones, K., Galliers, J. R, Evaluating Natural Language Processing Systems, Springer-Verlag (1995), ISBN-3-540-61309-9
Steeneken, H.J.M. Digital Speech Processing, Chapter 6, Quality evaluation of speech processing systems. Kluwer Academic Publishers Boston/Dordrecht/London (1992)
Walker, M., Kamm, C and Litman, D., Towards Developing General Models of Usability with PARADISE, Natural Language Engineering, Best Practice in Spoken Language Dialogue System Engineering, Special Issue, Volume 6, Part 3, October 2000
Potentials of speech and language technology systems for military use: an application and technology-oriented survey. Ed. H.J.M. Steeneken, NATO-RTO, Neuillysur Seine, (1996)

УДК 331.433:006.354 ОКС 13.180 Э65

Ключевые слова: речевые технологии, системы речевых технологий, оценка речевых технологий, критерии производительности

Редактор *И.В. Меньших*
Технический редактор *В.Н. Прусакова*
Корректор *В.И. Варенцова*
Компьютерная верстка *И.А. Налейкиной*

Сдано в набор 22.10.2010. Подписано в печать 11.11.2010. Формат 60 × 84 1/8. Бумага офсетная. Гарнитура Ариал. Печать офсетная. Усл. печ. л. 1,86. Уч.-изд. л. 1,60. Тираж 99 экз. Зак. 908.

ФГУП «СТАНДАРТИНФОРМ», 123995 Москва, Гранатный пер., 4.
gost.e-ecolog.ru gost.e-ecolog.ru
Набрано во ФГУП «СТАНДАРТИНФОРМ» на ПЭВМ.
Отпечатано в филиале ФГУП «СТАНДАРТИНФОРМ» — тип. «Московский печатник», 105062 Москва, Лялин пер., 6.

gost.e-ecolog.ru