

МСЭ-Т

СЕКТОР СТАНДАРТИЗАЦИИ
ЭЛЕКТРОСВЯЗИ МСЭ

P.851

(11/2003)

СЕРИЯ Р: КАЧЕСТВО ТЕЛЕФОННОЙ ПЕРЕДАЧИ,
ТЕЛЕФОННЫЕ УСТАНОВКИ, СЕТИ МЕСТНЫХ
ЛИНИЙ

Методы объективной и субъективной оценки качества

**Субъективная оценка качества телефонных
услуг на основе речевых диалоговых систем**

Рекомендация МСЭ-Т Р.851

РЕКОМЕНДАЦИИ МСЭ-Т СЕРИИ Р

КАЧЕСТВО ТЕЛЕФОННОЙ ПЕРЕДАЧИ, ТЕЛЕФОННЫЕ УСТАНОВКИ, СЕТИ МЕСТНЫХ ЛИНИЙ

Словарь и воздействие параметров передачи на мнение клиента о качестве передачи	Серия	P.10
Абонентские линии и аппараты	Серия	P.30 P.300
Стандарты передачи	Серия	P.40
Аппарат объективного измерения	Серия	P.50 P.500
Объективные электроакустические измерения	Серия	P.60
Измерения, относящиеся к громкости речи	Серия	P.70
Методы объективной и субъективной оценки качества	Серия	P.80 P.800
Аудиовизуальное качество в мультимедийных услугах	Серия	P.900

Для получения более подробной информации просьба обращаться к перечню Рекомендаций МСЭ-Т.

Рекомендация МСЭ-Т Р.851

Субъективная оценка качества телефонных услуг на основе речевых диалоговых систем

Резюме

В настоящей Рекомендации описываются методы и процедуры для проведения экспериментов по субъективной оценке качества телефонных услуг, которые основаны на речевых диалоговых системах. Эти системы позволяют осуществлять естественное общение путем разговора с обеспечением возможностей распознавания и интерпретации речи, управления диалогом и воспроизведения речи. Дается описание организации и проведения соответствующих экспериментов по взаимодействию, и приводятся вопросники для количественного определения уровня качества, воспринимаемого пользователем.

Источник

Рекомендация МСЭ-Т Р.851 утверждена 13 ноября 2003 года 12-й Исследовательской комиссией МСЭ-Т (2001–2004 гг.) в соответствии с процедурой, изложенной в Рекомендации МСЭ-Т А.8.

Ключевые слова

Параметр взаимодействия, понимание речи, распознавание речи, речевая диалоговая система, субъективная оценка, управление диалогом, формирование речи.

ПРЕДИСЛОВИЕ

Международный союз электросвязи (МСЭ) является специализированным учреждением Организации Объединенных Наций в области электросвязи. Сектор стандартизации электросвязи МСЭ (МСЭ-Т) – постоянный орган МСЭ. МСЭ-Т отвечает за изучение технических, эксплуатационных и тарифных вопросов и за выпуск Рекомендаций по ним с целью стандартизации электросвязи на всемирной основе.

Всемирная ассамблея по стандартизации электросвязи (ВАСЭ), которая проводится каждые четыре года, определяет темы для изучения Исследовательскими комиссиями МСЭ-Т, которые, в свою очередь, вырабатывают Рекомендации по этим темам.

Утверждение Рекомендаций МСЭ-Т осуществляется в соответствии с процедурой, изложенной в Резолюции 1 ВАСЭ.

В некоторых областях информационных технологий, которые входят в компетенцию МСЭ-Т, необходимые стандарты разрабатываются на основе сотрудничества с ИСО и МЭК.

ПРИМЕЧАНИЕ

В настоящей Рекомендации термин "администрация" используется для краткости и обозначает как администрацию электросвязи, так и признанную эксплуатационную организацию.

Соответствие положениям данной Рекомендации является добровольным делом. Однако в Рекомендации могут содержаться определенные обязательные положения (для обеспечения, например, возможности взаимодействия или применимости), и тогда соответствие данной Рекомендации достигается в том случае, если выполняются все эти обязательные положения. Для выражения требований используются слова "shall" ("следует", "обязан") или некоторые другие обязывающие термины, такие как "must" ("должен"), а также их отрицательные эквиваленты. Использование таких слов не предполагает, что соответствие данной Рекомендации требуется от каждой стороны.

ПРАВА ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

МСЭ обращает внимание на то, что практическое применение или реализация этой Рекомендации может включать использование заявленного права интеллектуальной собственности. МСЭ не занимает какую бы то ни было позицию относительно подтверждения, обоснованности или применимости заявленных прав интеллектуальной собственности, независимо от того, отстаиваются ли они членами МСЭ или другими сторонами вне процесса подготовки Рекомендации.

На момент утверждения настоящей Рекомендации МСЭ не получил извещения об интеллектуальной собственности, защищенной патентами, которые могут потребоваться для реализации этой Рекомендации. Однако те, кто будет применять Рекомендацию, должны иметь в виду, что это может не отражать самую последнюю информацию, и поэтому им настоятельно рекомендуется обращаться к патентной базе данных БСЭ.

© ITU 2005

Все права сохранены. Никакая часть данной публикации не может быть воспроизведена с помощью каких-либо средств без предварительного письменного разрешения МСЭ.

СОДЕРЖАНИЕ

	Стр.
1 Сфера применения	1
2 Ссылки	1
3 Сокращения	2
4 Введение.....	2
4.1 Задачи и компоненты речевой диалоговой системы	2
4.2 Взаимодействие с речевой диалоговой системой через телефонную сеть	3
4.3 Аспекты качества и влияющие факторы	4
4.4 Методы субъективной оценки	7
5 Характеристики речевой диалоговой системы	8
5.1 Факторы агента	8
5.2 Факторы задачи.....	10
5.3 Факторы пользователя.....	11
5.4 Факторы среды.....	11
5.5 Контекстуальные факторы.....	11
6 Организация эксперимента	12
6.1 Подготовка системы к работе и имитация с помощью "Wizard-of-Oz"	12
6.2 Сценарии испытаний	13
6.3 Субъекты испытаний.....	14
7 Вопросники.....	15
7.1 Вопросы, относящиеся к квалификации пользователя	16
7.2 Вопросы, относящиеся к отдельному взаимодействию	18
7.3 Вопросы, относящиеся к общему впечатлению пользователя от системы.....	20
8 Оценка применимости	22
9 Анализ и интерпретация собранной информации	23
Добавление I – Примеры сценариев	24
ЛИТЕРАТУРА	26

Субъективная оценка качества телефонных услуг на основе речевых диалоговых систем

1 Сфера применения

В настоящей Рекомендации описываются методы субъективной оценки, обеспечивающие получение информации о качестве телефонных услуг, основанных на речевых диалоговых системах, как оно воспринимается пользователями таких услуг. Речевые диалоговые системы, которым посвящена Рекомендация, позволяют осуществлять общение путем разговора с человеком-пользователем через телефонную сеть поочередно в двух направлениях и имеют возможности распознавания речи, понимания речи, управления диалогом, формирования ответа и вывода речи. Они могут обеспечить доступ к информации, хранимой в базе данных, или позволить выполнить различные виды обработки запроса.

Описанные здесь методы оценки посвящены различным аспектам качества с точки зрения пользователя, воспринимающего речевую диалоговую систему как черный ящик. Важными аспектами качества являются применимость услуги, эффективность связи, эффективность задачи и услуги, удовлетворенность пользователя, воспринимаемое качество ввода и вывода речи, совместимость системы, симметрия взаимодействия и воспринимаемая плавность взаимодействия. Методы основаны на лабораторных экспериментах, в которых субъекты взаимодействуют с речевой диалоговой системой, чтобы выполнить заранее определенную реалистическую задачу. Мнение субъектов о воспринимаемом уровне качества может быть оценено управляемым и неуправляемым способом по вопросам, которые выдаются им после эксперимента, или с помощью других методов оценки удобства использования. В настоящей Рекомендации описываются организация и выполнение соответствующих экспериментов взаимодействия, относящихся к уровню качества, воспринимаемому пользователем, и методика, которая обеспечивает получение информации об этом уровне качества. Дальнейшее руководство по методам субъективной оценки в целом и оценке устройств воспроизведения речи можно найти в Рекомендациях МСЭ-Т Р.800 и Р.85 и в Справочнике по телефонометрии МСЭ-Т.

2 Ссылки

Указанные ниже Рекомендации МСЭ-Т и другие источники содержат положения, которые путем ссылки на них в данном тексте составляют положения настоящей Рекомендации. На момент публикации указанные издания были действующими. Все Рекомендации и другие источники могут подвергаться пересмотру; поэтому всем пользователям данной Рекомендации предлагается изучить возможность применения последнего издания Рекомендаций и других источников, перечисленных ниже. Список действующих в настоящее время Рекомендаций МСЭ-Т регулярно публикуется. Ссылка на документ в данной Рекомендации не придает ему как отдельному документу статус Рекомендации.

- ITU-T Recommendation E.800 (1994), *Terms and definitions related to quality of service and network performance including dependability*.
- ITU-T Recommendation G.107 (2003), *The E-Model, a computational model for use in transmission planning*.
- ITU-T Recommendation G.1000 (2001), *Communications Quality of Service: A framework and definitions*.
- ITU-T Recommendation P.85 (1994), *A method for subjective performance assessment of the quality of speech voice output devices*.
- ITU-T Recommendation P.800 (1996), *Methods for subjective determination of transmission quality*.
- ITU-T Handbook on Telephony (1992).

3 Сокращения

В настоящей Рекомендации используются следующие сокращения:

ACR	Absolute Category Rating	Оценка по абсолютным категориям
ANOVA	Analysis of Variance	Анализ дисперсии
ASR	Automatic Speech Recognition	Автоматическое распознавание речи
CCR	Comparison Category Rating	Сравнительная оценка по категориям
DARPA	Defense Advanced Research Projects Agency	Управление перспективных оборонных исследовательских проектов
DCR	Degradation Category Rating	Оценка по категориям ухудшения качества
DTMF	Dual Tone Multiple Frequency	Двухтональная многочастотная сигнализация
HMM	Hidden Markov Model	Скрытая марковская модель
HSD	Honestly Significant Difference	Истинно значащая разность
MOS	Mean Opinion Score	Средняя оценка мнений
MLP	Multi-Layer Perceptron	Многоуровневый перцептрон
PARADISE	PARAdigm for Dialogue System Evaluation	Образец для оценки диалоговой системы
QoS	Quality of Service	Качество обслуживания
SDS	Spoken Dialogue System	Речевая диалоговая система
WoZ	Wizard-of-Oz	Оперативный консультант (мастер) "Wizard-of-Oz"

4 Введение

Речевые диалоговые системы (SDS), то есть компьютерные системы, с которыми человек-пользователь взаимодействует посредством разговора поочередно в двух направлениях, могут быть частью современных телефонных сетей. Они предоставляют возможность доступа к базам данных и осуществления операций с помощью телефонной связи, например, для получения информации о расписании поездов и авиарейсов, биржевых курсах, туристической информации, или для выполнения банковских расчетных операций, или для резервирования мест в гостиницах и т. д. В отличие от простых систем DTMF, речевые диалоговые системы обладают возможностями автоматического распознавания и понимания речи (то есть синтаксического/семантического/практического анализа и тем самым интерпретации речи), а также модулем управления диалогом, что обеспечивает плавное и естественное выполнение речевого взаимодействия между пользователем и системой. В результате взаимодействие становится более похожим на общение между людьми, и услуга, предоставляемая такими системами, может привлечь более широкий круг потенциальных пользователей. Часто различные типы систем, основанных на DTMF и речевых диалогах, реализуются интегрированно, и часть соответствующих аспектов качества будет идентична для обоих типов систем. Иногда системы на основе речевых диалогов используют структуры и протоколы интерфейса, применяемые в прикладной среде Web, и строятся тем же способом, что и Web-интерфейсы; таким образом, Web-интерфейсы могут служить образцом для получения тех же функциональных возможностей.

4.1 Задачи и компоненты речевой диалоговой системы

С технической точки зрения компоненты речевой диалоговой системы, работающей через телефонную сеть, могут наилучшим образом быть отображены в структуре следования. Пример такой структуры показан на рисунке 1. Она состоит из шести основных компонентов, к которым имеется доступ пользователя через интерфейс телефонного сервера. Речевой сигнал от пользователя сначала обрабатывается устройством распознавания речи. В процесс распознавания этот сигнал преобразуется в последовательность слов или граф гипотезы о распознанном слове, который затем подвергается семантическому анализу. Результатом является семантическая структура, представляющая то, что было "понято" из высказывания пользователя. Задача средства управления диалогом состоит в интерпретации семантической структуры в контексте диалога и задачи и в

отслеживании истории диалога. Когда от пользователя получена вся соответствующая информация, может быть начат запрос основного приложения (в этом примере базы данных). Информация, исходящая от прикладной программы, так же как другие коммуникативные задачи средства управления диалогом, должна быть преобразована в ответ для пользователя. Это задача модуля формирования ответа. Он формирует ответ в текстовой форме, который затем преобразуется в синтезаторе речи в речевой сигнал, передаваемый человеку-пользователю. Иногда формирование ответа и синтез речи выполняются в одном модуле (без перехода к текстовому представлению), и вместо синтезированной речи используются заранее записанные сообщения.



Рисунок 1/Р.851 – Структура следования речевой диалоговой системы на основе телефонной сети [27], [33]

Эту принципиальную структуру можно реализовать различными способами. Примеры можно найти в [4]. Одним из популярных способов является так называемая "архитектура с концентратором" [37], [45], которая используется в проекте коммуникатора DARPA. Другие структуры основаны на асинхронных рабочих модулях для интерпретации, поведения (логических выводов и выполнения действий) и формирования (см. [1]).

4.2 Взаимодействие с речевой диалоговой системой через телефонную сеть

Взаимодействие с речевой диалоговой системой имеет место через сеть электросвязи определенного типа. Эта сеть будет вносить в передачу ряд ухудшений, оказывающих влияние на качество передаваемой речи и, как следствие, на показатели устройства распознавания речи и последующие компоненты технологии речи и естественного языка в речевых диалоговых системах. На пути обратно к человеку-пользователю канал передачи ухудшит сигнал, генерируемый диалоговой системой. Поскольку сети электросвязи будут сталкиваться как со сценариями общения между людьми, так и со сценариями взаимодействия человек–машина, важно рассмотреть требования как к человеку-пользователю, так и к технологическому речевому устройству. Очевидно, что эти требования отличаются, так как особенности восприятия, влияющие на оценку пользователем качества, не идентичны характеристикам технологического речевого устройства, например автоматического устройства распознавания речи (ASR).

Человек-пользователь осуществляет взаимодействие через некоторый интерфейс пользователя, например микротелефонную трубку, терминал без снятия телефонной трубки или наушники. Акустические характеристики упомянутых интерфейсов очень разнообразны, так же как и их чувствительность к акустическим явлениям, возникающим в комнате вокруг пользователя, пока он говорит или слушает. Например, окружающий шум может значительно повлиять на разборчивость речевых сигналов, передаваемых через терминал без снятия телефонной трубки, а также сказывается на поведении говорящего пользователя. В результате для определения общего качества взаимодействия следует принимать во внимание весь сценарий взаимодействия, включая речевую диалоговую систему, канал передачи и интерфейс пользователя.

4.3 Аспекты качества и влияющие факторы

Пользователями услуг на основе SDS, которые предоставляются по телефону, являются люди. Таким образом при определении функций системы/услуги и степени их выполнения должен быть принят во внимание человеческий фактор. Качество обслуживания – это восприятие пользователем полученной услуги относительно того, что он ожидает или желает получить от нее. Исходя из определения качества, изложенного в [24], *качество обслуживания речевой диалоговой системы* является результатом оценки воспринимаемых составляющих услуги по отношению к требуемым. Таким образом, качество, воспринимаемое пользователем, является компромиссом между тем, что он/она ожидает или желает получить, и характеристиками, которые он/она воспринимает, пользуясь услугой. Это в значительной мере зависит от ситуации, в которой имеют место восприятие и вынесение оценки качества. Данный факт нужно учитывать при проведении экспериментов по субъективной оценке качества, а именно создавать более или менее естественную обстановку для испытаний и обеспечивать реалистическую мотивацию пользователя при испытаниях.

В отличие от понятия качества передачи речи в сценариях общения между людьми, пользователь услуги на основе речевой диалоговой системы принимает активное участие в формировании речи и ходе диалога. Таким образом, характеристики и поведение пользователя могут иметь решающее значение для выполнения необходимой задачи. Следовательно, пользователь будет оказывать сильное влияние на характеристики системы и услуги. Чтобы описать поведение системы и пользователя упрощенным способом, можно регистрировать параметры во время взаимодействия. Такие параметры взаимодействия могут быть, но не обязательно, измерены инструментально. Примеры включают ряд высказываний или часть диалога, которые могут быть инструментально измерены, или коэффициент ошибок по словам либо показатель успешного выполнения задачи, который может быть определен только с помощью экспертов-людей. Обзор параметров взаимодействия можно найти в [18] и [33].

В принципе качество обслуживания речевой диалоговой системы можно рассматривать с двух различных точек зрения – поставщика услуги и пользователя¹. Поставщик услуги интересуется главным образом влиянием отдельных элементов услуги и тем, как они соотносятся со степенью удовлетворенности пользователя обслуживанием или приемлемостью. Поставщики услуг руководствуются определением качества обслуживания (QoS), приведенным в Рекомендации МСЭ-Т E.800. Пользователь воспринимает и оценивает воспринимаемые характеристики (свойства) услуги, сравнивает восприятие с некоторым внутренним эталоном и судит о них в соответствии с тем, оправдали ли они его/ее ожидания или желания. При изучении качества обслуживания важно учитывать обе точки зрения. Методы субъективной оценки, такие как описываемые в настоящей Рекомендации, концентрируются на точке зрения пользователя. Однако они будут также полезны для поставщиков услуг, поскольку показывают характеристики, которые нужно улучшать.

Как *действенность*, так и *эффективность* относятся к показателям достижения заданной цели, для которой создана конкретная услуга. Действенность является абсолютным показателем, описывающим, до какой степени достигнута цель, с точки зрения точности и полноты достижения цели (см., например, [14]):

"Действенность: Точность и полнота, с которыми определенные пользователи могут достичь заданных целей в определенной обстановке".

Критериями действенности, о которых сообщается в литературе, являются, например, успех выполнения задачи или "P_i"-показатели [33]. Эффективность, с другой стороны, является относительным показателем достижения цели по отношению к используемым ресурсам [14]:

"Эффективность: Расход ресурсов относительно точности и полноты достигнутых целей".

¹ В Рекомендации МСЭ-Т G.1000 определено даже четыре точки зрения: требования потребителя к QoS, предложения поставщика услуги по QoS, достижимое или обеспечиваемое поставщиком QoS и QoS, воспринимаемое потребителем.

К обычно используемым показателям относятся, например, длительность диалога или число смен направления передачи, произведенных системой или пользователем.

Эффективность и когнитивная потребность являются критериями, характеризующими систему, с помощью которой пользователь способен достичь своих целей в рамках конкретной задачи. *Применимость*, однако, обычно определяется в более широком смысле и описывает возможность для определенных пользователей понять услугу, научиться пользоваться ею и использовать ее при заданных условиях. Она показывает пригодность услуги для выполнения потребностей пользователя, включает действенность и эффективность системы приводит к удовлетворенности пользователя [35]. *Удовлетворенность пользователя* является индикатором воспринимаемой полезности услуги и ее применимости для адресуемой группы пользователей. Она заключается, в частности, в том, получает ли пользователь информацию, которая ему/ей нужна, доволен ли он этой услугой и получает ли он информацию в пределах допустимого времени [31].

Описанные понятия качества речевой диалоговой услуги на основе телефонной сети можно проиллюстрировать с помощью диаграммы, как это предложено в [34] (см. рисунок 2). Помимо упомянутых выше факторов пользователя, восприятию пользователем качества способствуют четыре типа факторов: факторы агента (относящиеся главным образом к диалогу и самой системе), факторы задачи (относящиеся к тому, как речевая диалоговая система выполняет задачу, для которой она была разработана), факторы среды (например, факторы, относящиеся к акустической среде и каналу передачи) и контекстуальные факторы, такие как затраты, вид доступа или готовность. Аспекты качества, воспринимаемые пользователем, представлены в нижней части диаграммы.

Факторы среды, агента и задачи оказывают влияние на *качество ввода и вывода речи*, *согласованность* поведения системы и на *симметрию* взаимодействия в режиме диалога. Качество ввода и вывода речи включает такие аспекты, как разборчивость, естественность, напряжение при прослушивании для понимания сообщения системы или воспринимаемая способность системы к пониманию. *Согласованность* определяется здесь в смысле отсутствия нарушений принципов совместного поведения при диалоге, как определено Грайсом [20]. Она включает аспекты информативности, правдивости и очевидности, уместности, манеры разговора, предварительных знаний и обработки метаданных (то есть подтверждение, уточнение, исправление и восстановление от ошибок передачи данных) (см. [6]). Аспект асимметрии партнеров (к асимметрии партнеров по диалогу следует отнести разницу в интерактивном поведении) охватывается категорией, называемой симметрией в диалоге. Эта категория включает также влияние возможностей инициативы в диалоге и управления взаимодействием.

Упомянутые аспекты качества приводят к (более или менее) эффективной связи (взаимодействию) и эффективному решению подлежащей выполнению задачи. *Эффективность связи* относится к скорости или темпу взаимодействия, краткости и плавности диалога. *Эффективность задачи*, с другой стороны, связана с успешным выполнением задачи и простотой ее выполнения. Важными являются дополнительные аспекты качества: "личные качества" машинного агента (вежливость, дружелюбие, естественность поведения) и усилия, требуемые от человека-пользователя при взаимодействии (простота общения, стресс/волнение и т. д.). Эти аспекты отнесены к понятию *комфорта*.

Эффективность связи, эффективность задачи и комфорт – все это должно способствовать применимости услуги, индикатором чего можно считать удовлетворенность пользователя. С другой стороны, на *эффективность услуги* влияют эффективность задачи и контекстуальные факторы. Это важно для применимости услуги (для выполнения требуемой задачи) и для дополнительного значения, приписываемого услуге (например, в сравнении с аналогичными методами для получения такой же информации, как Web-интерфейс или вновь объявленный метод). Следствием применимости, эффективности услуги и *экономической выгоды* от нее являются *полезность* услуги и, наконец, ее *приемлемость*.

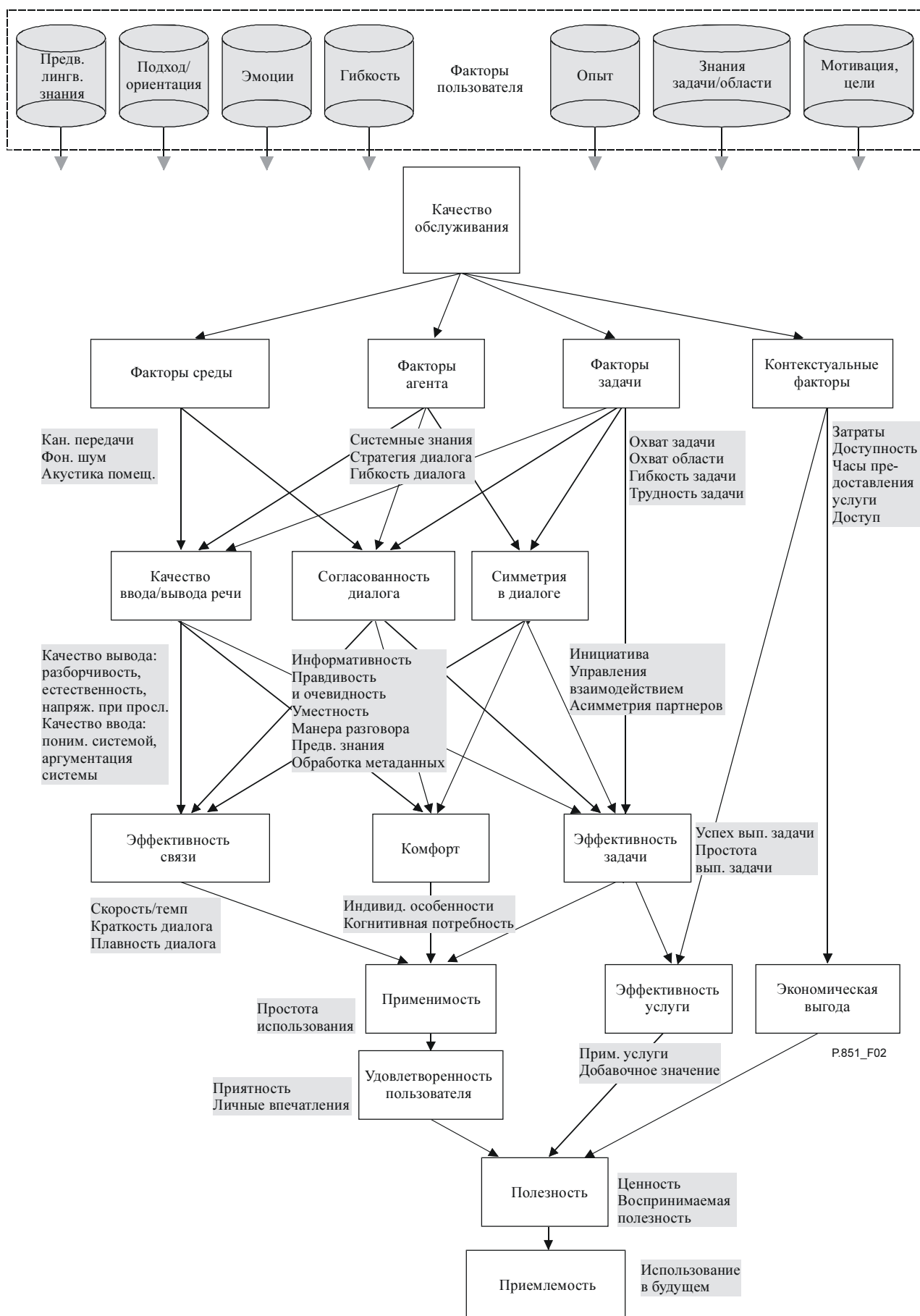


Рисунок 2/Р.851 – Аспекты качества и влияющие факторы (см. [34])

4.4 Методы субъективной оценки

Речевые диалоговые системы могут быть оценены на уровне компонентов (например, по устройству распознавания речи, компоненту понимания речи или компоненту вывода речи) или на уровне всей (интегрированной) системы. Аналитическая оценка на уровне компонентов является ценным источником информации при определении того, как отдельные части системы выполняют свою задачу. При этом, однако, иногда можно пропустить соответствующие компоненты, вносящие вклад в общее качество предоставляемой услуги, как оно воспринимается пользователем. Например, ошибочное распознавание или понимание речи может быть компенсировано компонентом обработки речевого материала без влияния на общее качество системы. По этой причине при необходимости определить качество услуг речевой диалоговой системы незаменимы эксперименты по субъективной оценке качества при взаимодействии реальных пользователей или пользователей – участников испытаний со всей системой.

Чтобы оценить различные аспекты качества услуг речевой диалоговой системы, должны проводиться эксперименты по субъективной оценке с участием человека-пользователя. Эти эксперименты служат двум главным целям:

- 1) В ходе взаимодействия собираются данные по инструментально измеряемым параметрам системы и регистрируются высказывания системы и пользователя. Файлы регистрации подвергаются экспертной оценке, результатом которой является совокупность параметров, характеризующих специфические аспекты взаимодействия человек–машина на уровне высказывания, диалога и задачи, с точки зрения разработчика системы.
- 2) После взаимодействия субъектам испытания выдается вопросник, целью которого является сбор информации о воспринимаемых свойствах качества, относящихся к формированию впечатления человека-пользователя о качестве. Такие эксперименты могут быть выполнены с помощью полнофункциональных систем или систем, которые еще находятся на стадии разработки и части модулей которых должны имитироваться. Подробная информация об организации экспериментов, вопросниках и методах оценки применимости приводится в разделах 6–8.

При лабораторных экспериментах оба типа информации могут быть получены параллельно. В условиях эксплуатационных испытаний с участием реальных пользователей, однако, инструментально зарегистрированные параметры взаимодействия часто являются единственным источником информации для поставщика услуг, чтобы он мог проконтролировать качество системы. Объем данных, которые можно собрать от действующей службы, может быть очень большим. В этом случае важно определить базовый набор результатов оценки, которые определяют качественные показатели системы, и иметь в наличии инструменты, которые автоматизируют значительную часть процесса анализа данных. Задачей человека, осуществляющего оценку, тогда являются анализ и интерпретация этих данных и определение влияния собранных результатов оценки качественных показателей на качество, которое может восприниматься пользователем (прототипом). Некоторые общие рассуждения об анализе и интерпретации результатов испытаний приводятся в разделе 9. При наличии обоих типов информации может быть установлена взаимосвязь между параметрами взаимодействия и субъективными оценками. Такие модели прогнозирования качества для речевых диалоговых систем на основе телефонной сети еще находятся в стадии изучения, они кратко рассматриваются в разделе 9.

В соответствии с общепринятой практикой проведения экспериментов по субъективной оценке следует четко определить цель и условия оценки или эксперимента по оценке и задокументировать их. В Европейском проекте DISC для этой цели был разработан шаблон [5]. На основе этого шаблона и классификации методов, приведенных в [33], можно определить следующие критерии:

- Мотивация оценки/определения качества (например, подробный анализ механизмов восстановления системы или оцениваемая удовлетворенность будущих пользователей).
- Объект оценки/определения качества (например, устройство распознавания речи, средство управления диалогом или вся система).
- Среда оценки/определения качества (например, управляемый лабораторный эксперимент или эксплуатационные испытания).
- Тип методов измерения (например, посредством инструментального измерения параметров взаимодействия или посредством открытых или закрытых оценок качества, полученных от пользователей).

- Признаки, подлежащие поиску (например, вопросы пользователя для разъяснения или отказы ASR).
- Стадия жизненного цикла, на которой имеет место оценка/определение качества (например, во время имитации, для версии прототипа или полностью действующей системы).
- Доступность системы и ее компонентов (например, подход "стеклянного ящика" или "черного ящика").
- Эталон, используемый для измерений (например, качественная оценка абсолютных показателей системы или количественные значения по отношению к измеряемому эталону или эталонному тексту).
- Инструменты поддержки, которые имеются для оценки/определения качества.

Эти критерии формируют основной комплект документации, который нужно обеспечить при эксперименте по оценке/определению качества. Документация может быть выполнена в виде приведенного здесь перечня вопросов или подробного описания эксперимента.

5 Характеристики речевой диалоговой системы

В соответствии со схематическим изображением аспектов качества на рисунке 2 на взаимодействие со службами на основе диалоговых речевых систем оказывают влияние пять факторов: *факторы агента, факторы задачи, факторы пользователя, факторы среды и контекстуальные факторы*. Эти факторы определяют качественные показатели системы (компонентов) и качество, воспринимаемое пользователем. Таким образом их следуют учитывать при проведении и документировании экспериментов по субъективной оценке.

5.1 Факторы агента

Систему как партнера взаимодействия можно охарактеризовать технически, то есть путем определения характеристик отдельных компонентов системы и их соединения в структуре следования, показанной на рисунке 1, или путем установления рабочих функций агента. Наиболее важными подлежащими определению функциями агента являются: способность распознавания речи, способность понимания естественного языка, способность управления диалогом, способность формирования ответа и способность воспроизведения речи. Компоненты способности понимания естественного языка и формирования ответа тесно связаны со смежными компонентами, а именно со средством управления диалогом, с одной стороны, и устройством распознавания речи или синтезатором речи – с другой. Следовательно, должны быть точно описаны интерфейсы этих компонентов.

5.1.1 Характеристики устройства распознавания речи

С функциональной точки зрения устройства распознавания речи можно классифицировать в соответствии со следующими параметрами [41]:

- Объем словаря, например, устройства распознавания речи с малым, средним или большим словарем.
- Сложность словаря, например, по отношению к возможности перепутать слова.
- Тип речи, например, отдельные слова, соединенные слова, непрерывная речь, непринужденная речь, включающая такие прерывания, как кашель, заикание, перерывы, повторения и т. д.
- Язык: одноязыковые или многоязыковые устройства распознавания, результаты распознавания, зависящие от языка, портативность языка.
- Зависимость от говорящего, например устройства распознавания, зависимые от говорящего, не зависимые от него или адаптируемые к нему.
- Тип и сложность грамматики. Сложность грамматики можно определить с точки зрения ее неоднозначности, которая является критерием того, как хорошо можно прогнозировать последовательность слов с помощью модели языка.
- Метод обучения, например обучение многократно четко высказанным отдельным словам или вложенное обучение последовательностям слов, в которых начальная и конечная точки не определены.

С другой стороны, компоненты устройств распознавания речи можно описать с точки зрения общих технических характеристик, которые в отдельных системах [27] могут быть реализованы по-разному. Следующие технические характеристики частично были использованы в проекте DISC:

- Получение сигнала: частота дискретизации, полоса частот сигнала, квантование, кадрирование.
- Анализ особенностей, например кепстральные коэффициенты в нелинейном частотном масштабе, энергия и производные слова первого или второго порядка.
- Основные единицы речи, например модели звуков или модели слов, моделирование молчания или других неречевых звуков.
- Словарь: число статей каждого слова с одним или несколькими вариантами произношения; формируемых либо из словарей, либо из графемо-фонемных преобразователей; дополнительные статьи для слов-заполнителей и шумов; ожидаемый охват словаря по отношению к целевому словарю.
- Акустическая модель: тип модели, например сети многоуровневого перцептрона (MLP) или скрытые марковские модели (НММ); учебные данные или параметры; постобработка модели.
- Языковая модель: тип модели, например статистическая языковая модель с N-граммовой задержкой или контекстно-независимая грамматика; учебный материал, например большая совокупность учебных материалов общецелевого назначения или данные, собранные в ходе ограниченного эксперимента; моделирование отдельных слов или классы для особых категорий (например, даты или имена); независимые или зависимые от состояния диалога модели.
- Тип декодера, например декодер на основе НММ.
- Степень использования просодической информации.

5.1.2 Характеристики понимания языка

Для способности понимания языка со стороны системы важными являются следующие характеристики:

- Семантическое описание задачи, например, посредством интервалов (пар значений атрибутов).
- Синтаксически-семантический анализ: возможность общего грамматического разбора, например полный грамматический разбор или оптимальный частичный грамматический разбор; число и сложность допустимого синтаксиса, например число альтернативных вариантов, имеющих на данном уровне.
- Контекстуальный анализ: число и сложность правил.
- Взаимодействие с модулями распознавания речи и управления диалогом: тип и количество вводимой и выводимой информации (одна гипотеза, перечни по категориям и т. д.), зависимость от синтаксически-семантической и контекстуальной интерпретации состояния диалога.

5.1.3 Характеристики средства управления диалогом

Подход, принятый для управления диалогом, может быть определен с технической точки зрения, например, как грамматика диалога, плановый подход или совместный подход [8], [32]. Наиболее важными характеристиками средства управления диалогом являются вид и количество знаний, используемых в устройстве, распределение инициативы между системой и пользователем и стратегия метасвязи системы (подтверждение, уточнение, исправление и восстановление):

- Знания средства управления диалогом: модель истории диалога (информация, которой обменивались при диалоге), модели задачи и области применения (сценарий, планы, цели и вспомогательные цели, объекты и их характеристики), модель общеобразовательных знаний, разговорная модель и модель пользователя.
- Инициатива: инициатива системы, смешанная инициатива или инициатива пользователя.
- Стратегия подтверждения: явное подтверждение, подразумеваемое подтверждение, "эхо-подтверждение", суммирующее подтверждение.
- Стратегии исправления, уточнения и восстановления.

- Адаптивность средства управления диалогом: конструктивные средства управления, которые должны узнавать новые понятия при обычной работе или адаптивные средства управления, которые могут включать динамическую модель пользователя и могут быть способны к обучению стратегиям общения пользователя.

Кроме взаимодействия с пользователем должно быть определено взаимодействие с прикладной системой, включая интерфейс (язык программирования), и потенциальные механизмы управления при обработке динамики прикладной системы.

5.1.4 Характеристики формирования речи

Формирование речи включает возможное формирование текстового ответа и перевод его в речевую форму. В большинстве систем используется один из трех типов формирования речи: заранее записанная речь, шаблонные предложения или преобразование текста в речь. Необходимо определить следующие характеристики:

- Взаимодействие со средством управления диалогом: тип и объем вводимой информации, обеспечиваемой средством управления диалогом, например письменный или аннотированный текст, концентрированная или просодическая информация и т. д.
- Формирование ответа: стратегия (например, формальная грамматика или простые шаблоны), гибкость (заранее определенный словарь или "открытый" словарь), тип и объем информации, подлежащей включению в каждое высказывание, форма сообщения (синтаксис, выбор слов).
- Голос системы: число голосов, пол, профессионализм, обучение, просодическое качество, условия записи, адаптивность.
- Язык: одноязычные или многоязычные синтезаторы, возможность идентификации языка, портативность языка.
- Тип формирования речи: заранее записанные сообщения, шаблонные предложения, преобразование текста в речь или понятий в речь.
- Характеристики преобразования текстовой информации в речь: стратегия (например, на основе модели или на основе свода законов языка), возможности предварительной обработки текста, параметры модели, характеристики совокупности единиц (типы и длина единиц, охват целевого словаря и т. д.), алгоритмы конкатенации и/или выбора, просодическая стратегия формирования (основная частота, длительность, интенсивность) и т. д.
- Контекстуальные характеристики: манера разговора, темп речи, контекстуальная адаптивность.

5.2 Факторы задачи

Задача, подлежащая выполнению пользователем, является определяющим фактором взаимодействия. Она может быть охарактеризована по типу задачи, области ее применения, сложности, частоты повторения, важности и портативности:

- Тип задачи: может быть проведено различие согласно [6] между:
 - хорошо структурированными задачами, имеющими стереотипную структуру, которая предписывает, какой частью информации необходимо обмениваться и зачастую – в каком естественном порядке; и
 - плохо структурированными задачами, содержащими большое число дополнительных подзадач, характер и порядок которых трудно предсказать,
 а также между:
 - однородными задачами; и
 - разнородными задачами, что означает *неотъемлемую* комбинацию нескольких различных задач, которые различаются по своей природе (например, упорядочение плюс информация плюс управление устройством).
- Область применения задачи: ценность, скалярность, число пользователей, которые знакомы с областью применения, полезность для области применения, возможность обобщений и т. д.
- Сложность задачи: число охватываемых сценариев, максимальное число вспомогательных целей, число подзадач, которые могут быть решены параллельно, минимальное число обменов, необходимых для разрешения проблемы, ожидаемая сложность синтаксиса/словаря и т. д.

- Частота повторения задачи, то есть частота, с которой, как ожидается, пользователь будет обращаться к системе для решения данной задачи. Системы для маршрутизации вызовов или потоков информации (так называемые системы "walk-up-and-use") будут использоваться с относительно низкой частотой, так что нельзя ожидать, что потенциальные пользователи будут иметь знания о системе или демонстрировать эффекты обучения (помнить поведение из предыдущих вызовов) или согласятся на обучение.
- Важность задачи, например вопросы безопасности.
- Портативность задачи.

5.3 Факторы пользователя

В большинстве случаев определение характеристик пользователя ограничивается широкой категоризацией по стоящей перед ним/ней задаче, области применения и жизненному опыту, поскольку не может быть получено точное определение факторов, важных для отдельного пользователя (подход; мотивация; эмоции; гибкость). В протоколах оценки часто приводятся следующие характеристики:

- Число пользователей.
- Возраст и пол: Ожидается, что они оказывают воздействие на основную частоту и спектр речи, а также на взаимодействие при диалоге.
- Уровень опыта: от начинающего пользователя до опытного, от случайного пользователя до регулярного, от подготовленного пользователя до неподготовленного.
- Уровень компетенции в области применения: от профессиональных пользователей до частных.
- Четкая мотивация при пользовании услугой.
- Физическое состояние, напряжение голоса, темп речи и т. д.
- Родной язык, акцент, диалект и т. д.

Для специальных применений может оказаться необходимым более четко определить опыт и компетенцию, например, по знаниям целей, способности разрабатывать стратегию для оптимизации показателей задачи и способности использовать устройства, необходимые для выполнения задачи [29].

5.4 Факторы среды

Окружающая среда содержит весь материальный контекст взаимодействия. Полностью охарактеризовать ее обычно не возможно, и следует описать только факторы, которые непосредственно влияют на речевой сигнал, а именно:

- Тип и акустические свойства интерфейса пользователя.
- Телефонный канал передачи: Описание может быть выполнено на различных уровнях, например, по способу передачи, коммутации и окончному оборудованию, используемых при соединении, или по параметрам эталонной конфигурации при планировании сети; см. рекомендацию МСЭ-Т G.107.
- Акустическое состояние помещения: включает реверберацию, окраску звука, уровень и спектр окружающего шума, одновременно говорящих и пр.

5.5 Контекстуальные факторы

Они являются нефизическими факторами, характеризующими контекст использования рассматриваемой услуги. Типичные факторы включают:

- Средство доступа: наличие телефонных номеров, линий к другим службам и от них и т. д.
- Доступность услуги: часы предоставления услуги, потенциальные ограничения доступа.
- Затраты: фиксированная или зависящая от времени стоимость взаимодействия, особые условия учета и т. д.
- Услуги с аналогичными функциональными возможностями: подлежат сравнению со всеми другими контекстуальными факторами.

6 Организация эксперимента

Эксперименты по субъективной оценке взаимодействия с речевой диалоговой системой следует организовывать в соответствии с общими правилами для испытаний с оценкой мнений, описание которых приводится в Рекомендации МСЭ-Т Р.800. Более подробное описание практических вопросов можно найти в Справочнике по телефонометрии МСЭ-Т. Этот принцип применяется к физическим условиям помещения для испытаний, к характеристикам окружающего шума, к планированию эксперимента и к общим правилам для анализа данных. В следующих разделах описываются только те вопросы, которые являются специфическими для взаимодействия с речевыми диалоговыми системами, а именно подготовка системы к работе, сценарии испытаний и субъекты испытаний.

Эксперименты по субъективной оценке могут быть выполнены либо с использованием полностью работающих систем, либо с помощью человека-экспериментатора, имитирующего отсутствующие части системы или систему в целом (так называемая "имитация Wizard-of-Oz"). Чтобы получить правильные и надежные результаты, (имитируемая) система, участвующие в испытании пользователи и экспериментальная задача должны выполнить несколько требований (см. подразделы 6.1–6.3). Каждое взаимодействие обычно регистрируется и аннотируется человеком-экспертом, так что параметры взаимодействия могут быть вычислены. После каждого взаимодействия и после всего сеанса испытаний субъектами испытаний должны быть заполнены вопросники. Эти вопросники позволяют определить различные аспекты качества услуги на основе речевой диалоговой системы. Назначение таких вопросников рассматривается в разделе 7. В разделе 8 приводится краткий обзор методов оценки с целью определения применимости услуг.

6.1 Подготовка системы к работе и имитация с помощью "Wizard-of-Oz"

Для выполнения экспериментов по взаимодействию с помощью человека-пользователя следует провести подготовку к работе, обеспечивающую полную функциональность системы. Точный характер подготовки к работе будет зависеть от готовности компонентов системы, а следовательно, от стадии разработки системы. Если компоненты системы еще не реализованы или их реализация будет невозможной (например, из-за отсутствия данных) или неэкономичной, требуется имитация соответствующих компонентов системы или системы в целом.

Имитация интерактивной системы с помощью участия человека (называемая "мастером"), то есть имитация "Wizard-of-Oz" (WoZ), является широко распространенным методом на стадии разработки системы. В то же время она служит инструментом для оценки системы по шлейфу или бионической системы (полусистема, полумастер). Замысел состоит в том, чтобы имитировать систему, в которую вводится речь и которая обрабатывает ее некоторым *принципиальным* способом и формирует речевые ответы пользователю. Чтобы обеспечить реалистическую ситуацию предоставления телефонной услуги, ввод и вывод речи для пользователя следует обеспечить через имитируемое или реальное телефонное соединение, используя стандартный интерфейс пользователя. Подробные описания организации экспериментов с помощью WoZ можно найти в [16], [6], [3] и [9].

Имитацию WoZ можно с пользой применять в случаях, когда человеческие способности выше способностей компьютера, что в настоящее время имеет место в отношении понимания или воспроизведения речи. Поскольку систему можно оценить до того, как она будет полностью готова к работе, характеристики определенных компонентов системы можно имитировать до степени, превышающей современный уровень развития техники. Таким образом, становится возможной экстраполяция до технологий, которые будут доступны в будущем [23]. Имитация WoZ позволяет относительно экономичным способом проверить осуществимость, охват и адекватность системы до ее реализации. Модели сложного взаимодействия и с высокой степенью новизны, возможно, проще имитировать в WoZ, чем реализовать, используя подход реализация–испытания–переработка. Однако последний, вероятно, завоюет позиции по мере появления стандартного программного обеспечения и прототипов инструментов, а также в промышленных установках, где в широком масштабе используются платформы. WoZ, тем не менее, целесообразен, если применение весьма рискованно, а затраты на то, чтобы перестроить систему, достаточно высоки [6].

На взаимодействие между человеком-пользователем и системой или мастером значительное влияние оказывают пять типов факторов, описанных в разделе 5. С точки зрения экспериментатора эти факторы формируют переменные экспериментальной установки. Переменные либо находятся под управлением экспериментатора (управляемые переменные), доступны или измеряемы им (ответные

переменные), либо являются неожиданными факторами, которые не интересуют экспериментатора или не управляются им. Неожиданные факторы могут быть ограничены с помощью процедур тщательного планирования экспериментов, а именно при полном или частичном планировании с участием субъекта.

Основной характеристикой имитации WoZ является то, что субъекты испытания не представляют себе, что система, с которой они взаимодействуют, имитируется. Данные, приведенные в [16] и [9], показывают, что этой цели можно достичь почти в 100% случаев, если имитация тщательно спланирована. Наиболее важным аспектом для иллюзий субъекта является способность системы по вводу и выводу речи. Некоторые авторы, например [17] и [2], подчеркивают, что иллюзия диалога с компьютером должна поддерживаться искажением речи. Однако другие параметры системы могут вызвать тот же эффект, например направленность системы.

Имитация WoZ должна обеспечивать реалистическую имитацию функциональных возможностей системы. Поэтому, до того как можно будет организовать имитацию WoZ, необходимо иметь точное описание функциональных возможностей и поведения системы. Важно, чтобы мастер точно придерживался этого описания и избегал всего того, что он знает и умеет делать лучше, чем испытываемая система. Это требует значительного обучения и поддержки для мастера. Поскольку человек может интуитивно использовать свои превосходящие умения и навыки, работу мастера следует автоматизировать, насколько это возможно. Для этой цели разработан ряд инструментов. Они обычно состоят из представления модели взаимодействия, например, в виде визуального графика или программного инструмента быстрого создания прототипа, системных фильтров для канала ввода и вывода (например, структурированные имитаторы воспроизведения звука, маскировки голоса и распознавания), а также из других инструментов поддержки, таких как регистрация взаимодействия (аудио, текст, видео) и поддержка области применения (например, расписаний). Типичные примеры описаны в [23], [15], [9], [6] и [33].

6.2 Сценарии испытаний

Из-за отсутствия реальной мотивации при лабораторных испытаниях ставятся экспериментальные задачи, которые субъекты должны выполнять. Экспериментальная задача имеет четкую цель, но ее не следует путать с целью, которую пользователь хотел бы достичь в реальной жизненной ситуации. Из-за этого противоречия на лабораторной испытательной установке не просто получить правильные суждения пользователя о полезности и приемлемости системы.

При лабораторном испытании экспериментальная задача определяется описанием сценария. Сценарий описывает определенную задачу, которую субъект должен выполнить путем взаимодействия с системой, например собрать информацию о конкретном железнодорожном направлении или найти конкретный ресторан [6]. Примеры таких сценариев для услуги, связанной с поиском информации о ресторане, приводятся в Добавлении I. Использование заранее определенного сценария дает возможность максимального контроля над задачей, выполняемой субъектами испытаний, охватывая при этом широкий диапазон возможных ситуаций (и возможных проблем) взаимодействия. Сценарий может быть специально разработан для испытания конкретных функциональных возможностей системы (так называемый сценарий разработки) или охватывать широкий диапазон потенциальных интерактивных ситуаций, которые желательно подвергнуть оценке. Таким образом, сценарии разработки обычно отличаются от сценариев оценки.

Сценарии помогают выявить различные недостатки в диалоге и таким образом повысить полезность и приемлемость окончательной системы. В них определяются цели пользователя в виде задачи и подобласти применения, для которых предназначен диалог, и они являются необходимым условием для определения того, достиг ли пользователь своей цели. Без заранее определенного сценария было бы чрезвычайно трудно сравнить результаты, полученные в ходе различных диалогов, поскольку запросы пользователей могут отличаться и выходить за пределы знаний области применения системы. Если фактором, который необходимо исследовать в эксперименте, является воздействие экспериментатору, нужно обеспечить, чтобы все пользователи выполняли одни и те же задачи. Это может быть достигнуто только с помощью заранее определенных сценариев.

К сожалению, заранее определенные сценарии могут оказывать некоторое негативное влияние на поведение пользователя. Хотя сценарии не обеспечивают реальных жизненных целей для субъектов испытаний, они инструктируют пользователя, как нужно взаимодействовать с системой. Письменные сценарии могут предлагать субъектам испытания имитировать речь, предусмотренную в сценарии, что приводит к чтению вслух вместо спонтанной речи. Было установлено, что выбор сценария может также повлиять на выбор стратегии решения, которая наиболее эффективна для решения задачи [43].

Субъектов испытаний, выполняющих заранее определенные сценарии, обычно не особенно интересуют ответы системы, поскольку в действительности они не нуждаются в этой информации. В результате успех задачи не может оказать значительного влияния на даваемую субъектами испытания оценку полезности. Кроме того, как сообщается, субъекты испытания не всегда внимательно читают инструкции и могут игнорировать или неправильно интерпретировать ключевые ограничения сценария.

Влияние предварительного инструктирования на речь пользователя можно уменьшить с помощью графического описания сценария; см. примеры в Добавлении I. Сравнение письменных и графических сценариев в [6], [13] показало, что массированного влияния инструктирования с помощью письменных сценариев можно почти полностью избежать при графическом представлении, но что многообразие лингвистических элементов (общее число слов, число слов, которых нет в словаре) аналогично в обоих случаях. Таким образом, многообразие языка все же необходимо обеспечивать путем сбора высказываний от достаточно большого числа различных пользователей, например, во время эксплуатационных испытаний. Другая возможность состоит в том, чтобы передать субъектам испытаний записанные на пленку речевые описания задач и посоветовать им составить конспекты [42]. Таким образом предполагается, что вовлеченные в эту работу процессы понимания и памяти заставят субъекты сосредоточиться на закодированном смысле описания задачи, а не на представлении внешней формы. Эмпирических доказательств этого предположения, однако, еще не дано.

6.3 Субъекты испытаний

Общее правило при проведении экспериментов по оценке качества состоит в том, что при выборе субъектов испытаний следует руководствоваться целью испытаний. Например, аналитическую оценку конкретных характеристик системы могут дать только подготовленные субъекты испытаний, которые являются специалистами по рассматриваемой системе. Однако эта группа не способна судить об общих аспектах качества системы так, чтобы на их оценку не влияло знание системы. Правильной общей оценки качества можно ожидать только от субъектов испытаний, которые как можно больше соответствовали бы группе будущих пользователей услуги. Общие рекомендации по подбору субъектов испытаний, приведенные в Рекомендации МСЭ-Т Р.800, следует также соблюдать при проведении экспериментов по субъективной оценке взаимодействия с использованием услуг на основе речевых диалоговых систем.

Обзор факторов пользователя приводится в подразделе 5.3. Некоторые из этих факторов важны для акустических и лингвистических характеристик речи пользователя, а именно возраст и пол, физическое состояние, темп речи, напряжение голоса, родной язык, диалект или акцент. Поскольку эти факторы могут быть крайне важными для показателей распознавания и понимания речи, оценки качества, полученные от группы пользователей, отличающихся по акустическим и языковым характеристикам, могут не отражать качества, которое можно ожидать для целевой группы пользователей. Группы пользователей, однако, являются изменяемыми и плохо определяемыми. Услуга, которая открыта для общего пользования, рано или поздно столкнется с широким диапазоном различных пользователей. Поэтому проведение испытаний с участием определенных пользователей за пределами целевой группы пользователей обеспечит показатель устойчивости системы относительно характеристик пользователя.

Вторая группа факторов пользователя относится к опыту и компетенции в отношении системы, задачи и области применения. Некоторые исследования показывают, что опыт пользователя влияет на очень широкий диапазон характеристик речи и диалога. Например, сообщается, что пользователи склонны решать больше проблем за один вызов, когда они привыкают к системе, и что взаимодействие становится короче [10]. Другие исследования показали, что число выражений, имеющих в словаре, увеличивается при ознакомлении пользователей с системой. В то же время повышается скорость выполнения задачи [25]. Знакомство с системой может также привести к уменьшению числа обращений пользователей к системе и справочных сообщений, а также к уменьшению времени обработки запроса [26], [28].

Представляется, что пользователи разработают определенные образцы взаимодействия, когда они ознакомятся с системой. Как постулируется, такой образец представляет собой воспринимаемый оптимальный баланс между усилиями, которые каждый отдельный пользователь вкладывает во взаимодействие с системой, и эффективностью такого взаимодействия [39]. Видимо, почти все пользователи выработают устойчивые образцы взаимодействия с системой, но образцы не идентичны для всех пользователей. Образец взаимодействия, который будет разработан пользователем, может

отражать также его убеждения машинного агента в том смысле, что пользователь может иметь "познавательную модель" системы, которая отражает то, что рассматривается как убеждение текущей системы [38]. Такая модель частично определяется высказываниями, выдаваемыми системе, и частично высказываниями, поступающими от системы. Пользователь обычно предполагает, что его/ее высказывания хорошо понимаются системой. В случае неправильного его понимания пользователь придет в замешательство, и скорее всего возникнут проблемы с плавностью диалога. Другим источником расхождения между познавательной моделью пользователя и убеждениями системы является то, что система имеет доступ к вторичным источникам информации, таким как база данных приложения. Пользователь может быть удивлен, если столкнется с информацией, которую он/она не предоставляли системе.

7 Вопросники

Для получения информации о параметрах качества, воспринимаемых пользователем, должны быть собраны субъективные оценки. При сборе оценок можно применять два различных принципа: либо определять соответствующие параметры качества более или менее неуправляемым способом, либо давать количественную оценку заранее определенным аспектам качества как ответы на закрытые вопросы или задачи определения оценок с помощью шкал. Оба способа имеют преимущества и недостатки: открытые запросы помогают установить уровень качества, который в противном случае остается невыявленным, и определить аспекты качества, которые с точки зрения пользователя наиболее относятся к делу. Таким образом можно облегчить интерпретацию закрытых количественных оценок. Закрытые вопросы или задачи определения оценок с помощью шкал облегчают сравнение оценок между субъектами и экспериментами и предоставляют точное средство для количественного определения восприятий пользователя. Они могут выполняться относительно легко, и неподготовленные субъекты часто предпочитают этот метод оценки.

Определение оценок с помощью шкал дает правильные и надежные результаты при соблюдении двух основных требований: позиции, подлежащие оценке, должны быть выбраны компетентно и осмысленно, и оценки с помощью шкал должны определяться в соответствии с хорошо установленными правилами. Методы определения оценок с помощью шкал подробно описаны в психометрической литературе, например в [21], [12] или [7]. Для оценки качества передачи МСЭ-Т рекомендует методы оценки по абсолютным категориям (ACR), оценки по категориям ухудшения качества (DCR) и сравнительной оценки по категориям (CCR) (см. рекомендацию МСЭ-Т Р.800). Для оценки качества услуг на основе речевой диалоговой системы субъектов испытаний обычно просят дать оценки по непрерывной шкале оценок или по различным шкалам ACR (оценка по абсолютным категориям). Шкала ACR состоит из ряда дискретных категорий, одна из которых должна быть выбрана субъектом испытаний. Категории отображаются визуально и могут быть отмечены атрибутами для каждой категории или только для крайних (левой крайней и правой крайней) категорий. Примеры непрерывных шкал оценок приведены в следующих подразделах. Хотя шкала "общего впечатления" аналогична соответствующей шкале ACR для оценки качества передачи (см. Рекомендацию МСЭ-Т Р.800), непосредственной связи между оценками нет, и таким образом нет закона трансформации, связывающего средние оценки, полученные по одной из непрерывных шкал, с оценками MOS, используемыми для описания общих оценок качества передачи.

Задачи определения оценок по обеим шкалам – непрерывной и с категориями – часто описываются в виде утверждения (например, "Систему было легко понять."), и субъекты испытаний должны выразить свое согласие с утверждением, ставя соответствующую "галочку" или отмечая категорию на шкале. Этот метод основан на давних предложениях, сделанных Лайкертом [30], и пример шкалы приведен на рисунке 3. Числа относятся к категориям или к позициям шкалы в зависимости от того, является ли утверждение положительным (от 1 для "решительно не согласен" до 5 для "решительно согласен") или отрицательным (от 5 для "решительно не согласен" до 1 для "решительно согласен"), и индивидуальные оценки суммируются для всех субъектов. Другой возможностью является определение информативных меток для каждой категории, как это предлагается, например, МСЭ-Т для экспериментов по определению качества передачи речи (см. Рекомендацию МСЭ-Т Р.800).

	Решительно не согласен	Не согласен	Ни согласен, ни не согласен	Согласен	Решительно согласен
Систему было легко понять	☐	☐	☐	☐	☐
					P.851_F03

Рисунок 3/Р.851 – Оценка состояния способом, как было предложено Лайкертом [30]

Хорошо разработанные шкалы не обеспечивают достоверную информацию, когда плохо определен параметр качества, подлежащий оценке, или когда он неправильно выбран для рассматриваемой услуги. В последующих разделах приводится примерный выбор аспектов качества, каждому из которых может быть предназначен конкретный вопрос, подлежащий оценке субъектами испытаний. Даются также примеры сформулированных вопросов или утверждений. Выбор вопросов для конкретной услуги будет зависеть от типа услуги, задач, которые могут быть выполнены, конкретного интерактивного поведения службы, группы субъектов испытаний, а также от конкретной цели, для которой проводится эксперимент по оценке. Обычно число позиций, подлежащих оценке в одном вопроснике, должно быть ограничено числом от 15 до 20, чтобы субъекты испытаний были способны различить и отразить отношение к отдельным позициям.

При лабораторном эксперименте вопросник может быть выдан субъектам испытаний непосредственно после выполнения взаимодействия (потенциально отражая впечатление после этого взаимодействия) и/или после ряда взаимодействий (обеспечивая некоторую интеграцию предыдущего опыта). При эксплуатационных испытаниях заполнение вопросников не может быть строго проконтролировано, и оценка обычно относится к ряду взаимодействий, выполненных в течение широко определенного периода времени. Может оказаться, что отрицательный опыт станет более запоминающимся и окажет более сильное влияние на процесс интеграции во времени, чем положительный [11].

7.1 Вопросы, относящиеся к квалификации пользователя

На ряд вопросов следует получить ответы в начале сеанса испытаний, чтобы определить пользователя и его квалификацию, которая относится к эксперименту. Эти вопросы касаются следующего:

- Персональная информация: возраст, пол, профессия, место и дата рождения, место жительства в настоящее время, языковая подготовка.
- Информация, относящаяся к задаче: частота выполнения задачи, обычный подход при решении задачи (альтернативные интерфейсы), мотивация, другие важные аспекты, относящиеся к задаче и области применения.
- Информация, относящаяся к системе: опыт пользования услугами на основе DTMF или на основе речевых диалоговых систем, опыт работы с устройствами речевой технологии (распознавание речи, синтезированная речь и т. д.).

В следующем перечне содержатся примеры вопросов, которые могут быть заданы субъектам испытаний. Они относятся к услуге, связанной с информацией о ресторанах, но могут быть легко приспособлены для других задач и услуг.

Вопросы, относящиеся к квалификации пользователя

Персональные данные

Пол: ☐ Женский ☐ Мужской

Возраст: _____ лет

Профессия/Образование: _____

Область/Город рождения: _____

Место жительства в настоящее время: _____

1 Как часто в среднем вы едите вне дома?

_____ раз в неделю _____ раз в месяц _____ раз в год

2 Как вы ищете ресторан, находясь за границей (возможен выбор нескольких вариантов)?

- | | | | |
|----------------------------|--------------------------|-----------------------------|--------------------------|
| 2.1 Журналы | ☐ | 2.6 Советы друзей | ☐ |
| 2.2 Коммерческие афиши | ☐ | 2.7 Звонок в автоматическую | |
| 2.3 Путеводитель по городу | ☐ | речевую систему | ☐ |
| 2.4 Телефонная книга | ☐ | 2.8 Другое: _____ | ☐ |
| 2.5 Интернет | ☐ | | |

3 Что для вас важно при принятии решения о ресторане (возможен выбор нескольких вариантов)?

- | | | | |
|--------------------|--------------------------|---------------------------------|--------------------------|
| 3.1 Цена | ☐ | 3.6 Обстановка | ☐ |
| 3.2 Вид пищи | ☐ | 3.7 Часы работы | ☐ |
| 3.3 Качество пищи | ☐ | 3.8 Быстрота обслуживания | ☐ |
| 3.4 Разнообразие | | 3.9 Дружелюбие при обслуживании | ☐ |
| предложенной пищи | ☐ | 3.10 Другое: _____ | ☐ |
| 3.5 Местоположение | ☐ | | |

4 Пользовались ли вы когда-либо автоматической информационной речевой системой?

☐ Да ☐ Нет

4.1 Если да, то по какому поводу?

4.1.1 Как бы вы охарактеризовали опыт работы с ней?

Очень Плохо Посредст- Удовлетвори- Хорошо Превосходно Идеально
плохо венно тельно

5 Имеете ли вы опыт общения с системой, понимающей речь?

☐ Да ☐ Нет

5.1 Если да, какого рода эта система?

6 Имели ли вы когда-либо дело с синтезированной речью?

☐ Да ☐ Нет

6.1 Если да, то по какому поводу?

7 Какую информацию о ресторане вы хотите получить от информационной системы?

7.2 Вопросы, относящиеся к отдельному взаимодействию

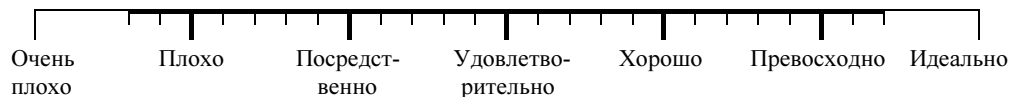
После каждого взаимодействия с (имитированной) услугой субъекты испытаний должны заполнить вопросник с рядом позиций, относящихся к индивидуальному опыту взаимодействия. Эти вопросы могут затрагивать следующие аспекты:

- Информация, полученная от системы: доступность, точность, полнота, последовательность, надежность, ясность и истинность полученной информации и т. д.
- Способность ввода/вывода речи: воспринимаемое понимание системы, частота ошибок системы, воспринимаемая аргументация системы, напряжение при прослушивании, требуемое для понимания сообщений системы, воспринимаемая разборчивость, воспринимаемая удобопонятность и т. д.
- Интерактивное поведение системы: прозрачность взаимодействия, согласованность с ожиданиями пользователя, гибкость взаимодействия, воспринимаемая надежность работы системы, распределение инициативы, способность управления взаимодействием, возможность подтверждения и коррекции, восстановление после проблем взаимодействия, естественность общения, длительность диалога, воспринимаемая скорость системы, плавность диалога и т. д.
- Воспринимаемые личные качества системы: дружелюбие, вежливость и т. д.
- Впечатление пользователя: ощущаемая пользователем естественность собственного поведения, приятность, когнитивные требования, предъявляемые к пользователю, стресс, волнение и т. д.
- Воспринимаемое выполнение задачи: успешность выполнения задачи, надежность полученных результатов.

Типичные вопросы, которые затрагивают эти аспекты, приводятся ниже. Экспериментатор может выбрать наиболее подходящие из них для исследуемой услуги.

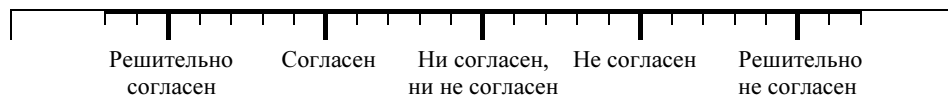
Вопросы, относящиеся к отдельному взаимодействию

Общее впечатление:

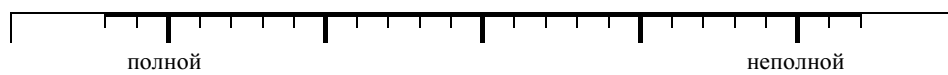


Информация, полученная от системы

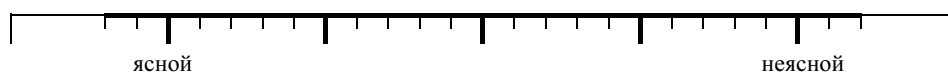
1 Система предоставила необходимую информацию.



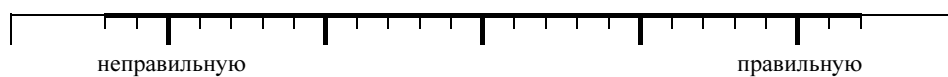
2 Предоставленная информация была ...



3 Информация была ...

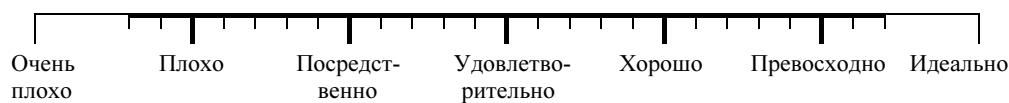


4 Вы оцениваете информацию как ...

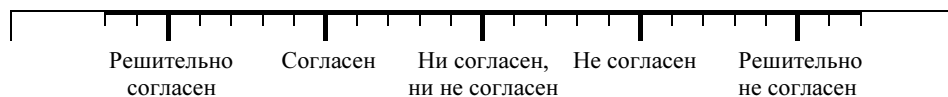


Связь с системой

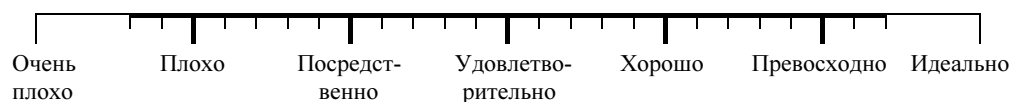
5 Насколько хорошо система понимала вас, по вашему ощущению?



6 Вам пришлось сосредоточиться, чтобы понять, что система ожидает от вас?

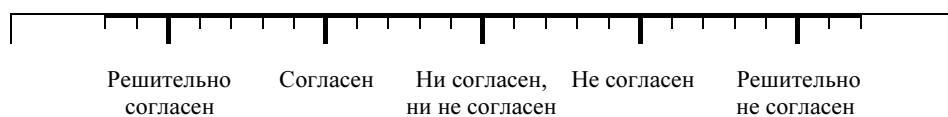


7 Насколько система была доступна для понимания акустически?

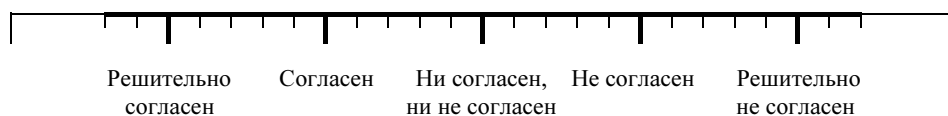


Поведение системы

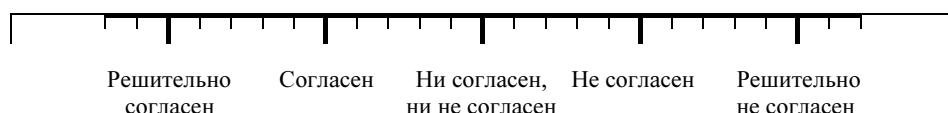
8 В каждый момент диалога вы знали, что от вас ожидает система.



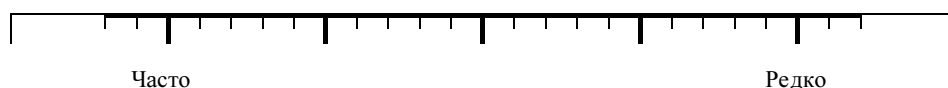
9 По вашему мнению, система обрабатывала ваши задания правильно.



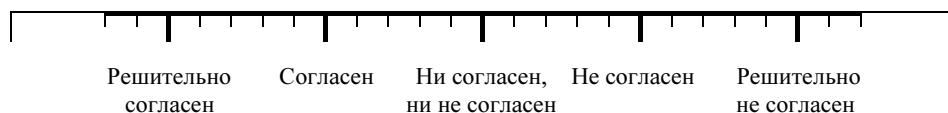
10 Поведение системы всегда соответствовало ожидаемому.



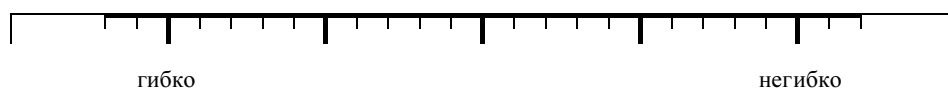
11 Как часто система делала ошибки?



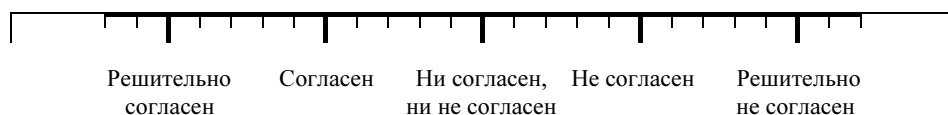
12 Система реагировала так же, как люди.



13 Система реагировала ...



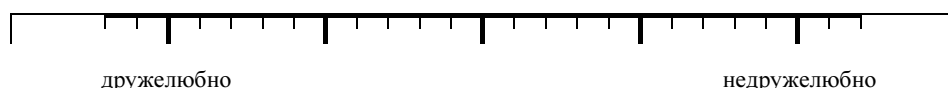
14 Вы были способны управлять диалогом, как хотели.



15 Система реагировала ...



16 Система реагировала ...

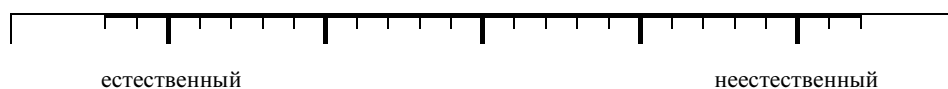


Диалог

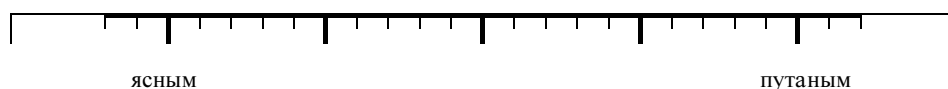
17 Высказывания системы были ...



18 Вы воспринимали диалог как ...



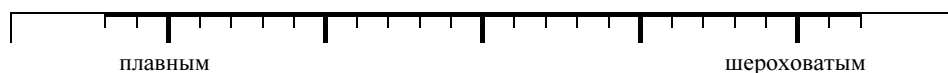
19 Ход диалога был ...



20 Диалог был ...



21 Ход диалога был ...



22 Голос системы был ...

естественным неестественным

23 В общем, вы удовлетворены диалогом.

Решительно согласен Согласен Ни согласен, ни не согласен Не согласен Решительно не согласен

Личное впечатление

24 Вы воспринимали диалог как ...

приятный неприятный

25 Во время диалога вы чувствовали себя

расслабленным напряженным

После того как все взаимодействия с системой были завершены, субъекты испытаний должны ответить на ряд дополнительных вопросов, на этот раз обращаясь к своему общему приобретенному опыту работы с системой. В такой вопросник могут быть включены следующие пункты:

- Пример соответствующего вопросника для информационной услуги приводится ниже. Его можно адаптировать и расширить в соответствии с рассматриваемой услугой, выполняемой задачей, а также целями эксперимента.

Вопросы, относящиеся к общему впечатлению пользователя от системы

1	Общее впечатление.	
2	Способ выражения системы был ...	
3	Система реагировала ...	
4	Вы ожидали больше помощи от системы.	
5	Система была способна отвечать на все ваши вопросы.	
6	Неправильное понимание можно было легко устранить.	
7	Система управляла плавностью диалога.	
8	Вы были способны обращаться с системой без каких-либо проблем.	
9	Что касается диалогов, они ...	
10	Вы получили удовольствие от диалогов.	
11	Вы чувствуете себя достаточно информированным о возможностях системы.	
12	Телефонные вызовы с помощью системы были стоящими.	
13	Вы восприняли эту возможность получения информации как ...	
14	Вы оцениваете систему как ...	

15 Вы предпочитаете использовать другой источник информации.

16 Обращение с системой было

17 Вы предпочитаете человека-оператора.

18 Вы стали ли бы в будущем снова пользоваться системой.

19 Какие характеристики системы вам понравились больше всего?

20 Какие характеристики системы доставили вам больше всего беспокойства?

21 Есть ли у вас какие-либо предложения по улучшению системы?

8 Оценка применимости

Помимо рассмотрения отдельных аспектов применимости с помощью вопросников, имеются специальные методы оценки применимости. Применимость можно оценить либо с помощью реальных пользователей, выполняющих конкретные испытания, либо с помощью методов контроля применимости при участии экспертов по оценке качества. Оба метода дополняют друг друга в том, что методы контроля применимости могут выявить проблемы, которые остаются незамеченными при выполнении испытаний пользователями, и наоборот [36]. Фактически была отмечена большая степень непересекающихся возможностей этих методов. Таким образом при оценке применимости следует сочетать эмпирические испытания и контроль применимости.

Цели методов контроля применимости состоят в выявлении проблем применимости в существующем проекте интерфейса пользователя, потенциальной оценке их серьезности, выдаче рекомендаций по их разрешению и улучшению тем самым применимости системы. Такие методы позволяют легко применять знания и опыт разработчиков интерфейса пользователя при оптимизации новых систем. Важной частью контроля применимости являются подсчет и классификация проблем применимости, которые наблюдаются при взаимодействии человек–машина. Контроль применимости, однако, должен быть эффективным не только в плане выявления проблем, но и в присвоении им веса в соответствии с их серьезностью (нет никакой пользы в разрешении второстепенных проблем) и особенно в предложении изменений и усовершенствований проекта. Так как многие методы контроля опираются на требования к проекту, а не на реализацию проекта, они могут применяться на относительно ранних стадиях процесса разработки проекта системы.

Различаются следующие восемь типов методов контроля применимости [36]:

- *Эвристическая оценка:* Этот неформальный метод предполагает участие специалистов по вопросам применимости, которые выносят суждения о том, соответствует ли элемент диалога установленным принципам применимости, называемым эвристическими.

- *Обзор руководств:* Контроль, при котором услуга на основе речевых диалоговых систем проверяется на соответствие обширному перечню руководств по применимости. Так как общее число руководств может быть весьма велико, этот подход требует высокого уровня компетенции.
- *Плюралистический критический анализ:* Собrania, на которых пользователи, разработчики и эксперты по вопросам человеческого фактора вместе шаг за шагом рассматривают сценарий, обсуждая вопросы применимости, связанные с элементами диалога, которые входят в каждый шаг сценария.
- *Проверка согласованности:* Интерфейс подвергается проверке несколькими разработчиками, представляющими несколько аспектов проекта, и затем оценивается, согласуется ли он со всеми сторонами проекта.
- *Проверка на соответствие стандартам:* Эксперт исследует конкретный интерфейс на соответствие определенному стандарту.
- *Когнитивный критический анализ:* Имитация процесса разрешения проблемы пользователя на каждом шаге взаимодействия и проверка того, можно ли считать, что цели пользователя и память о действии приводят к следующему правильному действию. Обычно проводится в форме вопросов о взаимосвязи между целями задачи, относящимися к пользователю, и действиями системы, необходимыми для их выполнения.
- *Формальный контроль применимости:* Формализованный метод с участием группы экспертов по применимости. Каждый член группы в процессе контроля имеет особую задачу, например модератор, собственник проекта или инспектор. Для подготовки и проведения контроля и анализа результатов организуются собрания.
- *Контроль свойств:* Внимание концентрируется на оперативных функциях интерфейса пользователя и на том, удовлетворяют ли предусмотренные функции требованиям адресуемых конечных пользователей.

Большая часть этих методов подробно рассматривается в соответствующей литературе по применимости [36]. Выбор правильного метода зависит от целей оценки, наличия руководств, опыта оценивающего лица и ограничений во времени и в денежных средствах.

Оценка применимости с помощью контролируемых пользователем экспериментов является второй альтернативой. Такие эксперименты могут проводиться либо "объективным неинтрузивным" способом, либо "субъективным интрузивным" способом [19]. Неинтрузивные методы пытаются получить сведения о поведении человека-пользователя естественным образом, не мешая ему, как, например, путем наблюдения за ним с помощью аудиовизуальной аппаратуры или путем регистрации с помощью записывающего устройства. Интрузивные методы требуют активного участия пользователей, например, путем ответов на вопросники или вопросы интервьюеров (см. последний раздел), путем групповых обсуждений или способом самозаписи, то есть запрашиванием подробного протокола, который отражает соображения или мнения пользователя во время или после взаимодействия. Эти методы описываются более подробно в литературе по оценке применимости [14].

9 Анализ и интерпретация собранной информации

Суждения, которые получены на шкалах закрытой оценки, могут быть проанализированы с помощью гистограмм или интегральных функций распределения. Хотя распределение не обязательно гауссово, обычной практикой является вычисление среднеарифметических значений (а не медиан) для всех оценок, полученных при заданной конфигурации системы (см. Рекомендацию МСЭ-Т Р.800 и Справочник по телефонии МСЭ-Т). Для средних значений оцениваются доверительные границы и определяются критерии значимости с помощью традиционного анализа дисперсии (ANOVA). Допущения, лежащие в основе анализа дисперсии (гауссово распределение и однородность дисперсии), не всегда соблюдаются; тем не менее этот метод представляется достаточно надежным, чтобы обеспечить приемлемые результаты и в случае отклонений от статистически идеальных условий. В случае статистически значимого влияния одной из переменных (конфигурация системы и/или голос, субъект испытаний, сценарий, порядок условий в эксперименте, сеанс испытаний и т. д.) можно использовать дополнительное специальное испытание для выполнения попарных сравнений средних значений и определения источников расхождений. Для этой цели рекомендуются испытания на истинно значащую разность (HSD) Такея [40]. Когда допущения, лежащие в основе параметрической статистики, не соблюдаются, целесообразно дополнительно суммировать результаты по медиане или моде и использовать для сравнения непараметрические испытания, подобно испытаниям согласно Краскелу и Уоллису.

Когда эксперименты по взаимодействию выполняются в контролируемых (лабораторных) условиях, можно получить параметры взаимодействия и субъективные оценки пользователя для одного и того же взаимодействия. В этом случае взаимосвязь между обоими типами показателей может быть определена количественно. На втором этапе можно попытаться прогнозировать суждения пользователя об отдельных аспектах качества или об общем качестве либо степени удовлетворенности качеством на основе измеряемых параметров взаимодействия. Однако все еще не существует универсального (независимого от задачи) подхода к моделированию качества, который позволил бы охватить большинство расхождений в суждениях пользователей в их прогнозах. Общая структура для прогнозирования качества речевых диалоговых систем была предложена Уолкером в 1997 году (образец для оценки диалоговой системы, PARADISE) на основе многомерного линейного регрессионного анализа [44]. Структура в принципе является независимой от задачи, но параметры функции прогнозирования качества должны определяться заново для каждой конкретной системы на основе субъективных экспериментов по взаимодействию. Кроме того, ее способность прогнозирования все еще очень ограничена (обычно в пределах 40–50% охватываемой дисперсии). Следовательно, параметры взаимодействия (как измеряемые инструментально, так и определяемые на экспертной основе) и суждения пользователей о качестве остаются главным источником информации для определения качества взаимодействия с речевой диалоговой системой.

Добавление I

Примеры сценариев

Сценарий № 1

Вы хотели бы узнать, где можно поесть утки. Пожалуйста, запросите у системы.

Название ресторана(ов): _____

Сценарий № 2

Вы планируете пообедать в ресторане греческой кухни во вторник вечером в XXX.

Цена: - | x | | | | +

Название ресторана(ов): _____

Если система не может указать ресторан, пожалуйста, измените запрос следующим образом:

Вы хотите пообедать в YYY.

Название ресторана(ов): _____

Сценарий № 3

Вы планируете ланч в китайском ресторане в центре города.

Цена: - | | | | x | +

Название ресторана(ов): _____

Сценарий № 4

Вы планируете поехать в XXX. Поскольку ваш любимый ресторан закрыт на праздничные дни, запросите систему о ресторане.

Пожалуйста, запишите сначала, какие указания вы хотите дать системе.

Если система не сможет найти подходящий ресторан, пожалуйста, поищите альтернативу, пока система не покажет по крайней мере один ресторан.

Название ресторана(ов): _____

Сценарий № 5

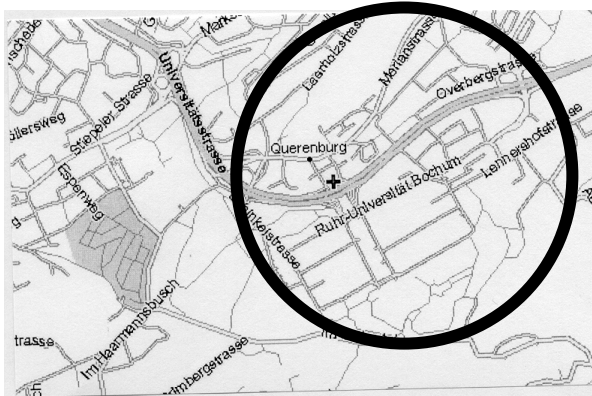
Пожалуйста, соберите информацию на основе следующих подсказок:

Цена: $-|---|---x---|---|+$

Вид пищи:



Местоположение:



Название ресторана(ов): _____

ЛИТЕРАТУРА

- [1] ALLEN (J.), FERGUSON (G.), STENT (A.): An Architecture for More Realistic Conversational Systems, *Proc. of Intelligent User Interfaces 2001 (IUI-01)*, 1-8 Santa Fe NM (2001).
- [2] AMALBERTI (R.), CARBONELL (N.), FALZON (P.): User Representations of Computer Systems in Human-Computer Speech Interaction, *Int. Journal on Man-Machine Studies*, 38, 547-566 (1993).
- [3] ANDERNACH (T.), DEVILLE (G.), MORTIER (L.): The Design of a Real World Wizard of Oz Experiment for a Speech Driven Telephone Directory Information Service, *Proc. 3rd Europ. Conf. on Speech Communication and Technology (EUROSPEECH'93)*, 2, 1165-1168, Berlin (1993).
- [4] ANTONIOL (G.), FIUTEM (R.), LAZZARI (G.), DE MORI, (R.): System Architectures and Applications, *Spoken Dialogues with Computers*, R. de Mori, ed., 583-609, Academic Press, London (1998).
- [5] BERNSEN (N.O.), DYBKJÆR (L.): A Methodology for Evaluating Spoken Dialogue Systems and Their Components, *Proc. 2nd Int. Conf. on Language Resources and Evaluation (LREC 2000)*, 2, 183-188, Athens (2000).
- [6] BERNSEN (N.O.), DYBKJÆR (H.), DYBKJÆR (L.): Designing Interactive Speech Systems: From First Ideas to User Testing, *Springer*, Berlin (1998).
- [7] BORG (I.), STAUFENBIEL (T.): Theorien und Methoden der Skalierung: Eine Einführung, *Verlag Hans Huber*, Bern (1993).
- [8] CHURCHER (G.E.), ATWELL (E.S.), SOUTER (C.): Dialogue Management Systems: A Survey and Overview, *Report 97.06, School of Computer Studies, University of Leeds*, Leeds (1997).
- [9] DAHLBÄCK (N.), JÖNSSON (A.), AHRENBERG (L.): Wizard of Oz Studies – Why and How? *Knowledge-Based Systems*, 6(4), 258-266 (1993).
- [10] DELOGU (C.), DI CARLO (A.), SEMENTINA (C.), STECCONI (S.): A Methodology for Evaluating Human-Machine Spoken Language Interaction, *Proc. 3rd Europ. Conf. on Speech Communication and Technology (EUROSPEECH'93)*, 2, 1427-1430, Berlin (1993).
- [11] DUNCANSON (J.P.): The Average Telephone Call Is Better Than the Average Telephone Call, *The Public Opinion Quarterly*, 33(1), 112-116 (1969).
- [12] DUNN-RANKIN (P.): Scaling Methods, *Lawrence Erlbaum Assoc.*, Hillsdale NJ (1983).
- [13] DYBKJÆR (L.), BERNSEN (N.O.), DYBKJÆR (H.): Scenario Design for Spoken Language Dialogue Systems Development, *Proc. ESCA Workshop on Spoken Dialogue Systems*, P. Dalsgaard, L.B. Larsen, L. Boves and I. Thomsen, eds., 93-96, Vigsø (1995).
- [14] ETSI Technical Report ETR 095: Human Factors (HF); Guide for Usability Evaluations of Telecommunication Systems and Services, *European Telecommunications Standards Institute*, Sophia Antipolis (1993).
- [15] FOSTER (J.C.), DUTTON (R.), JACK (M.A.), LOVE (S.), NAIRN (I.A.), VERGEYNST (N.), STENTIFORD (F.W.M.): Intelligent Dialogues in Automated Telephone Services, *Interactive Speech Technology: Human Factor Issues in the Application of Speech Input/Output to Computers*, C. Baber and J.M. Noyes, eds., 167-175, Taylor and Francis, London (1993).
- [16] FRASER (N.M.), GILBERT (G.N.): Simulating Speech Systems, *Computer Speech and Language*, 5, 81-99 (1991).

- [17] FRASER (N.M.), GILBERT (G.N.): Effects of System Voice Quality on User Utterances in Speech Dialogue Systems, *Proc. 2nd Europ. Conf. on Speech Communication and Technology (EUROSPEECH'91)*, 1, 57-60, Genova (1991).
- [18] GIBBON (D.), MOORE (R.), WINSKY (R.), eds.: Handbook on Standards and Resources for Spoken Language Systems, *Mouton de Gruyter*, Berlin (1997).
- [19] GLEISS (N.): Usability – Concepts and Evaluation, *TELE (English Edition)*, 2/92, 24-30, Swedish Telecommunications Administration, Stockholm (1992).
- [20] GRICE (H.P.): Logic and Conversation, *Syntax and Semantics, Vol. 3: Speech Acts*, P. Cole and J.L. Morgan, eds., 41-58, Academic Press, New York NY (1975).
- [21] GUILFORD (J.P.): Psychometric Methods, *McGraw-Hill Book Company*, New York NY (1954).
- [22] HONE (K.S.), GRAHAM (R.): Towards a Tool for Subjective Assessment of Speech System Interfaces (SASSI), *Natural Language Engineering*, 6(3-4), 287-303 (2000).
- [23] JACK (M.A.), FOSTER (J.C.), STENTIFORD (F.W.M.): Intelligent Dialogues in Automated Telephone Services, *Proc. 2nd Int. Conf. on Spoken Language Processing (ICSLP'91)*, 1, 715-718, Banff (1992).
- [24] JEKOSCH (U.): *Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung*, Habilitation thesis (unpublished), University/GH Essen, Essen (2000).
- [25] KAMM (C.), NARAYANAN (S.), DUTTON (D.), RITENOUR (R.): Evaluating Spoken Dialogue Systems for Telecommunication Services, *Proc. 5th Europ. Conf. on Speech Communication and Technology (EUROSPEECH'97)*, 4, 2203-2206, Rhodes (1997).
- [26] LAMEL (L.), BENNACEF (S.), GAUVAIN (J.L.), DARTIGUES (H.), TEMEM (J.N.): User Evaluation of the MASK Kiosk, *Speech Communication*, 38, 131-139 (2002).
- [27] LAMEL (L.), MINKER (W.), PAROUBEK (P.): Towards Best Practice in the Development and Evaluation of Speech Recognition Components for a Spoken Language Dialogue System, *Natural Language Engineering*, 6(3-4), 305-322 (2000).
- [28] LAMEL (L.), BENNACEF (S.), GAUVAIN (J.L.), DARTIGUES (H.), TEMEM (J.N.): User Evaluation of the MASK Kiosk, *Proc. 5th Int. Conf. on Spoken Language Processing (ICSLP'98)*, 7, 2875-2878, Sydney (1998).
- [29] LIFE (M.A.), LEE (B.P.), LONG (J.B.): Assessing the Usability of Future Speech Technology: Towards a Method, *Proc. of SPEECH'88*, 7th FASE Symposium, 4, 1297-1304, Edinburgh (1988).
- [30] LIKERT (R.): A Technique for the Measurement of Attitudes, *Archives of Psychology*, 140, 1-55 (1932).
- [31] MAIER (E.), MAST (A.), LUPERFOY (S.): Overview. Dialogue Processing in Spoken Language Systems, *Proc. of the ECAI'96 Workshop*, Budapest, E. Maier, M. Mast and S. LuperFoy, eds., Lecture Notes in Artificial Intelligence No. 1236, 1-13, Springer, Berlin (1997).
- [32] McTEAR (M.F.): Spoken Dialogue Technology: Enabling the Conversational Interface, *ACM Computing Surveys*, 34(1), 90-169 (2002).
- [33] MÖLLER (S.): Quality of Telephone-Based Spoken Dialogue Systems, *Habilitation thesis, Institute of Communication Acoustics, Ruhr-University, Bochum (to appear)* (2003).

- [34] MÖLLER (S.): A New Taxonomy for the Quality of Telephone Services Based on Spoken Dialogue Systems, *Proc. 3rd SIGdial Workshop on Discourse and Dialogue*, 142-153, Philadelphia PA (2002).
- [35] MÖLLER (S.): Assessment and Prediction of Speech Quality in Telecommunications, *Kluwer Academic Publ.*, Boston MA (2000).
- [36] NIELSEN (J.), MACK (R.L.), eds.: Usability Inspection Methods, *John Wiley & Sons*, New York NY (1994).
- [37] SENEFF (S.): Galaxy-II: A Reference Architecture for Conversational System Development, *Proc. 5th Int. Conf. on Spoken Language Processing (ICSLP'98)*, 3, 931-934, Sydney (1998).
- [38] SOUVIGNIER (B.), KELLNER (A.), RUEBER (B.), SCHRAMM (H.), SEIDE (F.): The Thoughtful Elephant: Strategies for Spoken Dialog Systems, *IEEE Trans. Speech and Audio Processing*, 8(1), 51-62 (2000).
- [39] STURM (J.), BAKX (I.), CRANEN (B.), TERKEN (J.), WANG (F.): The Effect of Prolonged Use of Multimodal Interaction, *Proc. ISCA Workshop on Multi-Modal Dialogue in Mobile Environments*, L. Dybkjær, E. André, W. Minker and P. Heisterkamp, eds., 1-15, Kloster Irsee (2002).
- [40] TUKEY (J.W.): Exploratory Data Analysis, *Addison-Wesley*, Reading MA (1997).
- [41] VAN LEEUWEN (D.), STEENEKEN (H.): Assessment of Recognition Systems, *Handbook on Standards and Resources for Spoken Language Systems*, D. Gibbon, R. Moore and R. Winsky, eds., 381-407, Mouton de Gruyter, Berlin (1997).
- [42] WALKER (M.A.), RUDNICKY (A.), PRASAD (R.), ABERDEEN (J.), BRATT (E.O.), GAROFOLO (J.), HASTIE (H.), LE (A.), PELLOM (B.), POTAMIANOS (A.), PASSONNEAU (R.), ROUKOS (S.), SANDERS (G.), SENEFF (S.), STALLARD (D.): DARPA Communicator: Cross System Results for the 2001 Evaluation, *Proc. 7th Int. Conf. on Spoken Language Processing (ICSLP 2002)*, 1, 269-272, Denver CO (2002).
- [43] WALKER (M.A.), FROMER (J.), DI FABBRIZIO (G.), MESTEL (C.), HINDLE (D.): What Can I Say? Evaluating a Spoken Language Interface to Email, *Human Factors in Computing Systems. CHI'98 Conf. Proc.*, Los Angeles CA, 582-589, Assoc. for Computing Machinery (ACM), New York NY (1998).
- [44] WALKER (M.A.), LITMAN (D.J.), KAMM (C.A.), ABELLA (A.): PARADISE: A Framework for Evaluating Spoken Dialogue Agents, *Proc. of the ACL/EACL 35th Ann. Meeting of the Assoc. for Computational Linguistics*, 271-280 (1997).
- [45] ZUE (V.), SENEFF (S.), GLASS (J.R.), POLIFRONI (J.), PAO (C.), HAZEN (T.J.), HETHERINGTON (L.): JUPITER: A Telephone-Based Conversational Interface to Weather Information, *IEEE Trans. Speech and Audio Processing*, 8(1), 85-96 (2000).

СЕРИИ РЕКОМЕНДАЦИЙ МСЭ-Т

Серия А	Организация работы МСЭ-Т
Серия В	Средства выражения, определения терминов, символы, классификация
Серия С	Общая статистика электросвязи
Серия D	Общие принципы тарификации
Серия Е	Общая эксплуатация сети, телефонная служба, функционирование служб и человеческие факторы
Серия F	Нетелефонные службы электросвязи
Серия G	Системы и среда передачи, цифровые системы и сети
Серия H	Аудиовизуальные и мультимедийные системы
Серия I	Цифровая сеть с интеграцией службы
Серия J	Кабельные сети и передача сигналов телевизионных и звуковых программ и других мультимедийных сигналов
Серия K	Защита от помех
Серия L	Конструкция, прокладка и защита кабелей и других элементов линейно-кабельных сооружений
Серия M	TMN и техническое обслуживание сетей: международные системы передачи, телефонные, телеграфные, факсимильные и арендованные каналы
Серия N	Техническое обслуживание: международные каналы передачи звуковых и телевизионных программ
Серия O	Требования к измерительной аппаратуре
Серия P	Качество телефонной передачи, телефонные установки, сети местных линий
Серия Q	Коммутация и сигнализация
Серия R	Телеграфная передача
Серия S	Оконечное оборудование для телеграфных служб
Серия T	Оконечное оборудование для телематических служб
Серия U	Телеграфная коммутация
Серия V	Передача данных по телефонной сети
Серия X	Сети передачи данных и взаимосвязь открытых систем
Серия Y	Глобальная информационная инфраструктура, аспекты межсетевого протокола (IP) и сети следующего поколения
Серия Z	Языки и общие аспекты программного обеспечения для систем электросвязи

