



МЕЖДУНАРОДНЫЙ СОЮЗ ЭЛЕКТРОСВЯЗИ

**МСЭ-Т**

СЕКТОР СТАНДАРТИЗАЦИИ  
ЭЛЕКТРОСВЯЗИ МСЭ

**Серия Р**

**Добавление 24**

(10/2005)

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

---

**Параметры, описывающие взаимодействие  
с переговорными диалоговыми системами**

Рекомендации МСЭ-Т серии Р – Добавление 24

---

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
Характеристики передачи и аспекты КО конечной точки в IP-сети	серия	P.1000

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

## **Добавление 24 к Рекомендациям МСЭ-Т серии Р**

### **Параметры, описывающие взаимодействие с переговорными диалоговыми системами**

#### **Резюме**

В настоящем Добавлении приводятся определения для набора параметров, которые относятся к службам, основанным на переговорных диалоговых системах. Параметры могут быть извлечены из регистрируемых (тестовых) взаимодействий пользователей с исследуемой службой. Они позволяют измерить ход взаимодействия, определить поведение пользователя и системы, а также характеристики устройств с технологией распознавания речи, участвующих в процессе взаимодействия. Они предоставляют полезную информацию, необходимую для развития системы, ее оптимизации и технического обслуживания, и являются добавочными к субъективным оценкам по качеству, получаемым согласно Рекомендации МСЭ-Т Р.851.

#### **Источник**

Добавление 24 к Рекомендациям МСЭ-Т серии Р было утверждено 21 октября 2005 года 12-й Исследовательской комиссией МСЭ-Т (2005–2008 гг.).

#### **Ключевые слова**

Оценка, автоматическое распознавание речи, автоматическое понимание речи, управление диалогом, параметр взаимодействия, генерация речи, речевая технология, переговорная диалоговая система.

## 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.

На Всемирной ассамблее по стандартизации электросвязи (ВАСЭ), которая проводится каждые четыре года, определяются 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.

Соблюдение положений данной Рекомендации носит добровольный характер. Однако в Рекомендации могут содержаться определенные обязательные положения (например, для обеспечения возможности взаимодействия или применимости), и соблюдение положений данной Рекомендации достигается в случае выполнения всех этих обязательных положений. Для выражения необходимости выполнения требований используется синтаксис долженствования и соответствующие слова (такие, как "должен" и т. п.), а также их отрицательные эквиваленты. Использование этих слов не предполагает, что соблюдение положений данной Рекомендации является обязательным для какой-либо из сторон.

## ПРАВА ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

МСЭ обращает внимание на вероятность того, 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 извещение об интеллектуальной собственности, защищенной патентами, которые могут потребоваться для выполнения 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 2006

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

## СОДЕРЖАНИЕ

	<b>Стр.</b>
1 Сфера применения .....	1
2 Справочные документы .....	1
3 Определения .....	1
4 Сокращения .....	2
5 Введение.....	2
6 Характеристики параметров взаимодействия .....	3
7 Обзор параметров взаимодействия .....	4
8 Интерпретация значений параметров взаимодействия .....	14
БИБЛИОГРАФИЯ .....	16



## Добавление 24 к Рекомендациям МСЭ-Т серии Р

### Параметры, описывающие взаимодействие с переговорными диалоговыми системами

#### 1 Сфера применения

В настоящем Добавлении описываются параметры, видимые системным разработчиком и оператором службы и предоставляющие информацию о взаимодействии со службами, основанными на переговорных диалоговых системах. Переговорные диалоговые системы, рассматриваемые в настоящем Добавлении, обеспечивают разговорное языковое взаимодействие с пользователем посредством телефонной связи, основанное на принципе поочередности, функционируют, используя системы автоматического распознавания и понимания речи, управления диалогом, генерации отклика и характеризуются наличием возможностей речевого вывода. Системы позволяют обеспечить доступ к информации, хранящейся в базе данных, или позволяют выполнять различные типы операций.

Параметры, определенные в данном тексте, измеряют поток взаимодействия, поведение пользователя и системы, а также характеристики устройств с технологией распознавания речи, участвующих в процессе взаимодействия. Для извлечения всех параметров переговорная диалоговая система должна быть доступной как "прозрачный ящик"; однако некоторые параметры могут быть получены с использованием подхода "черный ящик", т. е. без доступа к отдельным компонентам системы. Извлечение может быть выполнено частично автоматически и частично с помощью метода экспертного преобразования и файлов аннотаций журнала взаимодействия. Параметры описывают характеристики работы системы с точки зрения системного разработчика, следовательно, они предоставляют дополнительную информацию к субъективным оценкам, получаемым по результатам экспериментов с переговорными диалоговыми системами, рекомендации для которых даны в Рекомендации МСЭ-Т Р.851. Более подробное руководство по методам субъективной оценки в целом и оценке устройств речевого вывода содержится в Рекомендациях МСЭ-Т Р.800 и Р.85 и в Справочнике по телефонометрии. Параметры, перечисленные в данном Добавлении, не связаны конкретно с возможными ухудшениями в работе, происходящими в канале передачи. Их влияние является объектом дальнейшего изучения ИК12 МСЭ-Т.

#### 2 Справочные документы

- Рекомендация МСЭ-Т Р.85 (1994 г.), *Метод субъективной оценки характеристик качества устройств речевого вывода.*
- Рекомендация МСЭ-Т Р.800 (1996 г.), *Методы субъективного определения качества передачи.*
- Рекомендация МСЭ-Т Р.851 (2003 г.), *Субъективная оценка качества телефонной службы, основанной на переговорных диалоговых системах.*
- МСЭ-Т Справочник по телефонометрии (1992 г.).

#### 3 Определения

Для получения информации по определениям, не рассмотренным здесь, просьба обращаться к Рек. МСЭ-Т Р.10.

**3.1 вмешательство в разговор:** Способность человека говорить по системной подсказке или по выходной характеристике системы [10].

**3.2 диалог:** Разговор или обмен информацией. Как единица оценки: одна из нескольких возможных линий диалоговой структуры.

**3.3 производительность:** Меры точности и полноты задач системы относительно ресурсов (например, время, людские усилия), используемых для выполнения определенных задач системы.

**3.4 обмен:** Пара следующих друг за другом связанных высказываний, произносимых каждым участником диалога [8].

**3.5 функциональность:** Возможности системы обеспечивать функции, соответствующие установленным и предполагаемым потребностям, при использовании системы в определенных условиях.

**3.6 метаобщение:** Общение по поводу общения, например, для разрешения ситуаций неправильного понимания ("Я правильно тебя понял?") или для достижения соглашения по поводу употребления языка.

**3.7 эксплуатационные качества:** Возможность блока выполнять функцию, для реализации которой он предназначен.

**3.8 речевая технология:** Область знаний, связанная с исследованием и разработкой систем ввода и вывода разговорной речи, использующая научные достижения смежных областей: акустики, электротехники, статистики, фонетики, обработки естественного языка, и включающая спецификацию системных требований, проектирование, реализацию и оценку результатов, обработку базы и лингвистических ресурсов, а также анализ продукции, ориентированной на потребителя [10].

**3.9 переговорная диалоговая система:** Компьютерная система, с помощью которой пользователь на разговорном языке взаимодействует с системой по принципу поочередности.

**3.10 задача:** Все виды действий, разработанных пользователем с целью достижения поставленной задачи в каком-либо домене.

**3.11 целенаправленный диалог:** Диалог, касающийся определенного предмета, имеющий конкретную цель (такую как разрешение проблемы или получение определенной информации) [8].

**3.12 транзакция:** Часть диалога, посвященная отдельной задаче высокого уровня (например, бронирование билетов, проверка баланса банковского счета). Транзакция может быть одинаковой по временной протяженности с диалогом, либо диалог может состоять из более чем одной транзакции [8].

**3.13 реплика:** Высказывание. Отрывок речи, воспроизводимый одним из участников диалога начиная с того момента, когда этот участник начинает разговор до момента вступления в разговор другого участника [1].

**3.14 высказывание:** См. реплика.

## 4 Сокращения

ASR	Автоматическое распознавание речи
AVM	Матрица значений параметров
AVP	Пара значений параметров
DARPA	Управление перспективных научно-исследовательских проектов МО США
DP	Динамическое программирование
DTMF	Двухтональные многочастотные сигналы
IVR	Интерактивный речевой ответ
MOS	Низкая оценка
SDS	Переговорная диалоговая система
WoZ	Мастер Оз (программа имитации)

## 5 Введение

Переговорные диалоговые системы (SDS), т. е. компьютерные системы, с помощью которых пользователи взаимодействуют на разговорном языке по принципу поочередности с системой, могут быть частью современных телефонных сетей. Такие системы обеспечивают доступ к базам данных и транзакциям посредством телефона, например, с целью получения информации о расписании движения поездов, авиарейсов, о ставках фондовых бирж, туристической информации, или для осуществления операций с банковским счетом, а также резервирования гостиничных номеров. В



отличие от простых систем интерактивного речевого ответа (IVR) с вводом DTMF, SDS обладают полным спектром возможностей речевого взаимодействия, включая распознавание речи пользователя, присвоение значений распознанным словам, принятие решений относительно способа продолжения разговора, составление лингвистического ответа, а также генерация речевого вывода для пользователя. Это позволяет организовать более или менее "натуральное" взаимодействие пользователя с системой.

Для проведения оценки качества служб, основанных на SDS, с точки зрения пользователя, в 2003 году ИК12 МСЭ-Т выпустила Рек. МСЭ-Т Р.851. Данная Рекомендация описывает методы проведения субъективных оценочных экспериментов с целью определения качества *с точки зрения пользователя*, рассматривая SDS как черный ящик. Проведенные в соответствии с Рек. МСЭ-Т Р.851 эксперименты способствуют получению ценной информации по качеству, как ее представляет пользователь. Однако могут возникнуть сложности с определением вклада отдельного компонента системы в общее испытываемое пользователем качество, например, с определением конкретного компонента, нуждающегося в совершенствовании, в случае проблем во взаимодействии в целом. Следовательно, оценка должна быть дополнена информацией, касающейся характеристик системы *с точки зрения системного разработчика и оператора службы*.

Системная информация может быть описана в терминах так называемых *параметров взаимодействия*. Такие параметры позволяют рассчитать поток взаимодействия, поведение пользователя и системы, а также характеристики устройств с технологией распознавания речи, участвующих в процессе взаимодействия. Параметры описывают характеристики работы системы с точки зрения системного разработчика и оператора службы и, следовательно, предоставляют дополнительную информацию к данным субъективной оценки. Для извлечения некоторых параметров переговорная диалоговая система должна быть доступной, как "прозрачный ящик"; другие параметры также могут быть получены с использованием подхода "черный ящик", т. е. без доступа к отдельным компонентам системы.

Настоящее Добавление предоставляет совокупность параметров взаимодействия, которые используются последние 15 лет для оценки SDS. Перечисленные параметры относятся к общей передаче информации между пользователем и системой, метаобщениям в случае неправильных толкований, кооперативности системы, задаче, выполняемой с помощью системы, а также к системным возможностям речевого ввода. Тем не менее, для качества речевого вывода пока не доступно ни одно параметрическое описание (например, что касается качества синтезированной речи). Эта совокупность основана на теоретической работе, которая описана в [17].

Не все параметры взаимодействия находятся в прямой связи с воспринимаемым качеством основанных на SDS служб. В действительности, взаимозависимость отдельных параметров и пользовательских оценок качества, как правило, достаточно небольшая. Все же будет лучше избавиться от большого набора параметров, описывающих взаимодействие пользователя с системой, фиксируя таким образом большую часть информации, потенциально соответствующей воспринимаемому с точки зрения системного разработчика качеству. Такие параметры предоставляют полезную информацию для разработки системы, ее оптимизации и технического обслуживания.

Будучи однажды определенными и использованными в оценочных экспериментах на различных участках испытания, параметры способствуют проведению оценки их влияния на воспринимаемое качество для широкого ряда систем и служб. Это делает возможным разработку алгоритмов прогнозирования качества на базе параметров взаимодействия. Работа в этом направлении продолжает проводиться ИК12 МСЭ-Т и другими.

## **6 Характеристики параметров взаимодействия**

Параметры взаимодействия могут быть извлечены во время взаимодействия реального или тестового пользователя со службой. Извлечение может быть выполнено частично автоматически и частично с помощью файлов журнала, преобразуемых и аннотируемых специалистом. Простые параметры, такие как продолжительность взаимодействия или одного выражения, обычно определяются полностью автоматически на основе соответствующих алгоритмов. С другой стороны, пользовательские преобразования и аннотации необходимы в случае обработки не только

поверхностной формы (речевых сигналов), но и содержания и значения системных и пользовательских выражений (например, при определении точности слова или концепции).

SDS являются системами такого высокого уровня сложности, что описание поведения системы или сравнение между системами или версиями систем необходимо осуществлять, принимая во внимание множество различных параметров [24]. Вследствие этого для получения наибольшего объема информации следует применять оба способа сбора параметров взаимодействия (автоматический и экспертный). На основе собранной информации можно достаточно эффективно оптимизировать и проводить техническое обслуживание переговорных диалоговых систем.

Так как параметры взаимодействия основываются на данных, собираемых в процессе взаимодействия между пользователем и системой, они зависят от характеристик системы, пользователя и взаимодействия между последними. Такие виды влияния обычно нельзя определить, так как поведение пользователя находится в большой зависимости от работы системы (например, от вопросов, задаваемых системой), и наоборот (например, тип разговорной речи и словарный состав, применяемые пользователем влияют на точность распознавания и понимания системы). Следовательно, параметры взаимодействия четко отражают характеристики группы пользователей, которые осуществляли их выборку.

Параметры взаимодействия могут быть установлены как посредством лабораторных тестов в контролируемых условиях, так и с помощью испытаний в рабочих условиях. В последнем случае извлечь все параметры нельзя ввиду невозможности получения достаточного объема необходимой информации. Например, если необходимо определить успех проблемно-ориентированного взаимодействия (например, по поводу расписания движения поездов), то важно иметь четкое представление о целях пользователя. Такая информация может быть получена только в лабораторных условиях, например, способами, описанными в Рек. МСЭ-Т Р.851. В случае, если отсутствует полностью собранная система, возможно получить параметры с помощью так называемой имитации "Мастера Оз", когда экспериментатор заменяет отсутствующие части системы на тесты. Характеристики такой имитации должны учитываться при толковании полученных параметров.

Параметры взаимодействия могут быть вычислены на уровне слова, на уровне предложения или высказывания, или на уровне полного процесса взаимодействия или диалога. В случае параметров уровня слова или выражения средние значения обычно рассчитываются для каждого диалога. Параметры, собираемые определенной группой пользователей, могут быть проанализированы относительно влияния системы (версии), пользовательской группы и экспериментальных установок (сценариев, условий и режимов испытаний, и т.д.), посредством стандартных статистических методов. Описание таких влияний содержится в Рек. МСЭ-Т Р.851.

## **7 Обзор параметров взаимодействия**

На основе обширного литературного обзора были определены параметры, которые использовались в различных оценках и аналитических экспериментах в последние 15 лет. Соответствующая литература приводится в [2] [3] [4] [6] [7] [8] [9] [11] [12] [14] [16] [21] [22] [23] [24] [25] [26] [27] [28] [30] [31] [32], а итоговые данные по параметрам приведены в [17]. Укрупненно параметры могут быть разделены на следующие классы:

- Параметры, относящиеся к диалогу и коммуникации;
- Параметры, относящиеся к метаобщению;
- Параметры, относящиеся к кооперативной связи;
- Целенаправленные параметры;
- Параметры, связанные с речевым входом.

Эти категории будут кратко описаны в следующих разделах. Для каждой категории будет приведен список соответствующих параметров с определениями, уровнем взаимодействия, который описывает параметр (слово, высказывание или диалог), а также метод измерения (автоматический или основанный на аннотации специалиста – экспертный).

## 7.1 Параметры, относящиеся к диалогу и коммуникации

Параметры, относящиеся к диалогу в целом и передаче информации, дают очень приближенное представление о ходе взаимодействия. Они подробно не специфицируют коммуникативную функцию конкретного выражения. Параметры, относящиеся к этой категории, приведены в таблице 1 и включают параметры продолжительности (продолжительность диалога в целом, продолжительность реплики системы и пользователя, задержки ответа системы и пользователя), а также параметры слова и реплики (среднее количество реплик системы и пользователя, среднее количество слов на реплику системы и пользователя, количество вопросов системы и пользователя).

Заслуживают внимание два параметра, предложенные в [11]: *частота запросов* показывает, насколько эффективно пользователь может обеспечивать систему новой информацией, и *эффективность концепции* описывает, насколько эффективно система может воспринимать эту информацию от пользователя. Эти параметры также относятся к возможностям системы понимать язык, однако они включены в данный раздел, так как являются следствием возможностей взаимодействия системы в целом, а не только возможностей понимания языка.

Все параметры этой категории являются глобальными и относятся к диалогу в целом, однако могут быть определены частично на уровне выражения. Глобальные параметры иногда являются проблемными, так как индивидуальные различия в познавательных навыках могут быть большими по отношению к различиям, вызванным системой, а также вследствие необходимости субъекта обучаться стратегиям решения задач, имеющим значительное влияние на глобальные параметры.

Таблица 1 – Параметры взаимодействия, относящиеся к диалогу и коммуникации

Сокр.	Наименование	Определение	Уровень взаимодействия	Метод измерения
<i>DD</i>	продолжительность диалога	Общая продолжительность диалога в [мс], см. [8][6][12][21]	диал.	авт.
<i>STD</i>	продолжительность реплики системы	Средняя продолжительность реплики системы, рассчитанная с момента начала разговора системы до момента остановки разговора, в [мс]. Реплика – это выражение, т. е. отрывок речи, воспроизводимый одной из сторон диалога. [8]	высказывание	авт.
<i>UTD</i>	продолжительность реплики пользователя	Средняя продолжительность реплики пользователя, рассчитанная с момента начала разговора пользователя до момента остановки разговора, в [мс]. [8]	высказывание	авт.
<i>SRD</i>	задержка ответа системы	Средняя задержка ответа системы, рассчитанная с момента остановки разговора пользователя до начала разговора системы, в [мс]. [22]	высказывание	авт.
<i>URD</i>	задержка ответа пользователя	Средняя задержка ответа пользователя, рассчитанная с момента остановки разговора системы до начала разговора пользователя, в [мс]. [22]	высказывание	авт.
<i># turns</i>	кол-во реплик	Общее кол-во реплик, воспроизводимых в диалоге. [30]	диал.	авт./экспер.
<i># system turns</i>	кол-во реплик системы	Общее кол-во реплик системы, воспроизводимых в диалоге. [30]	диал.	авт./экспер.
<i># user turns</i>	кол-во реплик пользователя	Общее кол-во реплик пользователя, воспроизводимых в диалоге. [30]	диал.	авт./экспер.
<i>WPST</i>	кол-во слов на реплику системы	Среднее кол-во слов на реплику системы в диалоге. [6]	высказывание	авт./экспер.
<i>WPUT</i>	кол-во слов на реплику пользователя	Среднее кол-во слов на реплику пользователя в диалоге [6]	высказывание	авт./экспер.

Таблица 1 – Параметры взаимодействия, относящиеся к диалогу и коммуникации

Сокр.	Наименование	Определение	Уровень взаимодействия	Метод измерения
# system questions	кол-во вопросов системы	Общее кол-во вопросов от системы в диалоге.	диал.	экспер.
# user questions	кол-во вопросов пользователя	Общее кол-во вопросов от пользователя в диалоге. [12][21]	диал.	экспер.
QD	частота запросов	Среднее число новых концепций (слотов, см. 7.4.), представленных в запросе пользователя. Если $n_d$ – кол-во диалогов, $n_q(i)$ – общее кол-во запросов пользователя в $i$ -том диалоге, и $n_u(i)$ – кол-во уникальных концепций, "правильно воспринятых" системой в $i$ -м диалоге, тогда $QD = \frac{1}{n_d} \sum_{i=1}^{n_d} \frac{n_u(i)}{n_q(i)}$ Показатель $n_u(i)$ не включает концепцию, ранее воспринятую системой в одном из предыдущих выражений. [11]	набор диал.	экспер.
CE	эффективность концепции	Среднее кол-во реплик, необходимых для системы чтобы "воспринять" концепцию. Если $n_d$ – кол-во диалогов, $n_u(i)$ – кол-во уникальных концепций, "правильно воспринятых" системой в $i$ -м диалоге, и $n_c(i)$ – общее кол-во концепций в $i$ -м диалоге, тогда $CE = \frac{1}{n_d} \sum_{i=1}^{N_d} \frac{n_u(i)}{n_c(i)}$ Концепция засчитывается всякий раз, когда воспроизводится пользователем и не попадает во множество ранее воспринятых системой концепций. [11]	набор диал.	экспер.

## 7.2 Параметры, относящиеся к метаобщению

Метаобщение, т. е. общение по поводу общения, особенно важно для разговорного взаимодействия с системами, имеющими ограниченные возможности распознавания, понимания и осмысления. В этом случае для разрешения проблем непонимания используются корректирующие и поясняющие выражения или субдиалоги.

Параметры, относящиеся к этой группе, определяют количество выражений системы и пользователя, являющихся частью метаобщения. Большинство параметров выражаются в абсолютных количествах выражений в диалоге, связанном с определенной проблемой взаимодействия, а затем усредняются по отношению к набору диалогов. Они включают количество запросов справочной информации от пользователя, превышений выдачи подсказки системой, выражений пользователя, отклоненных системой при невозможности извлечения семантического содержания (ASR-отклонения), сообщений об ошибке диагностической системы, попыток пользователя вмешаться, а также попыток пользователя отменить предыдущее действие.

Способность системы (и пользователя) восстановиться после проблем взаимодействия может быть описана двумя способами: либо в явном виде с помощью процента коррекции, т. е. процента всех (системных и пользовательских) реплик, которые связаны главным образом с устранением проблем взаимодействия, либо неявно с помощью параметра *неявного восстановления*, определяющего возможность системы восстановить выражения, частично не распознанные и не воспринятые.

В отличие от общих показателей, большинство параметров, относящихся к метаобщению, описывают работу системы и высказывания пользователей в процессе связи. Таким образом, большинство параметров должны быть определены при помощи эксперта – мастера комментариев. Параметры приведены в таблице 2.

**Таблица 2 – Параметры взаимодействия, относящиеся к метаобщению**

Сокр.	Наименование	Определение	Уровень взаимодействия	Метод измерения
# <i>help request</i>	количество справочных запросов от пользователей	Общее количество справочных запросов пользователей в диалоге. Справочный запрос пользователя помечен мастером комментариев, если пользователь явно запрашивает помощь. Этот запрос должен быть сформулирован как вопрос (например, "Что такое доступные опции?") или как предложение ("Предоставьте мне доступные опции!"). [30]	высказывание	эксперт
# <i>system help</i>	количество диагностич. системных справочных сообщений	Общее количество справочных сообщений, созданных системой в диалоге. Справочное сообщение – это система высказываний, в которых пользователь информируется о доступных опциях в определенных местах диалога.	высказывание	инстр./эксперт
# <i>time-out</i>	количество сообщений о тайм-ауте	Общее количество экстренных сообщений о тайм-ауте в диалоге из-за отсутствия ответа от пользователя. [30]	высказывание	инстр.
# <i>ASR rejection</i>	количество ослаблений ASR	Общее количество ослаблений ASR в диалоге. Ослабление ASR определяется как системное экстренное сообщение о том, что в системе невозможно "слышать" или "понимать" пользователя, например, что в системе невозможно понять какой-либо смысл в высказываниях пользователя. [30]	высказывание	инстр.
# <i>system error</i>	количество диагностич. системных ошибочных сообщений	Общее количество диагностических ошибочных сообщений от системы в диалоге. Диагностическое ошибочное сообщение определяется как системное высказывание, в котором в системе указывается, что невозможно выполнять определенную задачу или обеспечивать определенную информацию. [22]	высказывание	инстр./эксперт
# <i>barge-in</i>	количество попыток пользователя вмешиваться в разговор	Общее количество попыток пользователя вмешиваться в разговор в диалоге. Попыткой пользователя вмешиваться в разговор считается, когда пользователь намеренно обращается к системе, в то время как в системе идет разговор. В этом определении высказывания пользователя, которые не предназначены влиять на течение диалога (смех, выражения гнева, раздражение или невежливые выражения) не считаются вторжением. [30]	высказывание	эксперт
# <i>cancel</i>	Количество отмененных попыток пользователя	Общее количество отмененных попыток пользователя в диалоге. Отмена пользователя классифицируется как отмененная попытка, если пользователь пытается восстановить диалог от начала или если он/она явно хочет пройти один или несколько шагов в обратном направлении в структуре диалога. [16][23]	высказывание	эксперт

Таблица 2 – Параметры взаимодействия, относящиеся к метаобщению

Сокр.	Наименование	Определение	Уровень взаимодействия	Метод измерения
<i>SCT, SCR</i>	количество системных поправочных отклонений, системная поправочная интенсивность	Общее количество системных поправочных отклонений (SCT) или процент системной поправочной интенсивности (SCR) всех системных отклонений в диалоге, которые в основном связаны с исправляемыми "проблемами", не вносящими таким образом новых предлагаемых значений и не прерывающими диалог. "Проблема" может быть вызвана распознаванием речи, пониманием ошибок или нелогичными противоречивыми неопределяемыми фрагментами речи пользователя. В случае, если пользователь не дает ответа на вопрос системы, соответствующий системный ответ помечается как системное поправочное отклонение, за исключением случаев, когда пользователь запрашивает информацию или действие, которое не поддерживается текущей функциональностью системы. [8][24][9][7]	высказывание	эксперт
<i>UCT, UCR</i>	количество поправочных отмен пользователя, поправочная интенсивность пользователя	Общее количество поправочных отмен и процент поправочной интенсивности всех отмен пользователя в диалоге, которые в основном связаны с исправляемыми "проблемами", таким образом не вносящими новых предлагаемых значений и не прерывающими диалог (см. SCT, SCR). [8][24][9][7]	высказывание	эксперт
<i>IR</i>	неявное восстановление	Способность системы восстанавливать из высказываний пользователя, для которых распознавание речи или понимание процесса частично повреждено. Определенные метками и частично проанализированные высказывания (см. определение PA:PA в 7.5), как, например, был ли системный ответ "правильным" или нет: $IR = \frac{\# \text{фрагменты речи с ответом}}{PA : PA}$ Для определения "правильности" см. 7.3. [7]	высказывание	эксперт

### 7.3 Параметры, относящиеся к кооперативной связи

Кооперативность определена, как ключевой аспект для успешного взаимодействия с разговорными диалоговыми системами [1]. К сожалению, трудно определить количественно, работает ли система кооперативно или нет. Несколько параметров диалоговой связи и метаобщения тем или иным образом являются связанными с системной кооперативностью, но не было попыток определить их количественно в этом аспекте.

Прямые показатели кооперативности – это контекстные правомерные параметры, представленные Simpson и Fraser [24]. Каждое системное высказывание должно быть оценено экспертами, как например, не нарушает ли оно один или более принципов Grice для кооперативности, см. [13]:

- *Количество информации:* Делайте ваше сотрудничество таким информативным, как требуется (для данной цели обмена); не делайте ваше сотрудничество более информативным, чем это необходимо.
- *Качество:* Старайтесь сделать ваше сотрудничество истинным; не говорите, что вы полагаете, что оно может быть ошибочным; не говорите, что вы нуждаетесь в адекватном доказательстве.
- *Отношение:* Соответствующее.
- *Образ действий:* Ясно выражайте мысли; избегайте непонятных выражений; избегайте неточностей; будьте краткими (избегайте ненужного многословия); будьте спокойными и организованными.

Эти принципы сформулированы более точно Bernsen и Dybkjær [1] по отношению к разговорным диалоговым системам.

Высказывания классифицируются в категории соответствующие (не нарушающие принципы Grice), несоответствующие (нарушающие один или больше принципов), соответствующие/ несоответствующие (в этой классификации эксперты не могут достигнуть согласия), малопонятные (содержание высказываний не может быть понято в контексте диалога) или полностью ошибочные (нет лингвистического соответствия в системе). Необходимо отметить, что классификация не всегда проста и что может быть необходима интерпретация принципов.

**Таблица 3 – Параметры взаимодействия, относящиеся кооперативной связи**

Сокр.	Наименование	Определение	Уровень взаимодействия	Метод измерения
<i>CA:AP</i> , <i>CA:IA</i> , <i>CA:TF</i> , <i>CA:IC</i> , <i>%CA:AP</i> , <i>%CA:IA</i> , <i>%CA:TF</i> , <i>%CA:IC</i>	контекстная правильность	Общее количество или процент системных высказываний, которые считаются подходящими в непосредственном диалоговом контексте. Определенные отмечаемыми фрагментами речи согласно тому, нарушают ли они один или более Grice's maxims для кооперативности: <ul style="list-style-type: none"> <li><i>CA:AP</i>: Соответствующие, не нарушающие Grice's maxims, не неожиданно замеченные или отмеченные до некоторой степени.</li> <li><i>CA:IA</i>: Несоответствующие, нарушающие один или больше Grice's maxims.</li> <li><i>CA:TF</i>: Полностью ошибочные, нет лингвистического соответствия.</li> <li><i>CA:IC</i>: Малопонятные, содержание не может быть понято экспертом.</li> </ul> <p>Для более подробной информации см. [24][8][9]; классификация подобна принятой в [14].</p>	высказывание	эксперт

#### 7.4 Целенаправленные параметры

Современное состояние услуг включает целенаправленные взаимодействия между системой и пользователем. Успешное решение этой задачи – это ключевой вопрос пригодности предоставления услуг. Успешное решение задачи может быть лучше всего определено в лабораторной ситуации, где явно заданные задачи берутся как тестируемые, см. Рек. МСЭ-Т Р.851. Однако практические показатели успешного решения задачи должны принимать во внимание потенциальные отклонения от сценария пользователя, потому что он/она не обращали внимание на инструкции в сценарии, из-за его/ее невнимательности к системным высказываниям, или потому что задача была нерешаемой и должна была быть модифицирована в ходе диалога.

Модификация экспериментальной задачи рассматривается в большинстве определений успешного решения, которые представлены в литературе. Успех может быть достигнут простым предоставлением правильного ответа на ограничения в инструкциях, ограничительной релаксацией от системы или от пользователя (или от того и другого), или сообщениями, которые не имеют решения для определенной задачи. Сбой задачи может быть экспериментально отнесен к работе системы или пользователя, однако последний находится под влиянием системы.

Другим подходом в определении успеха решения задачи является коэффициент ( $k$ ). В нем принимается метод понимания речи, который базируется на атрибутах (концепциях, слотах), для которых допустимые значения должны быть назначены в ходе диалога между системой и пользователем. Пары атрибутов и назначенные значения называются атрибутивно-значимыми парами (AVP). Комплект доступных атрибутов вместе со значениями, определенными задачей (так называемая атрибутивно-значимая матрица (AVM)), полностью характеризует задачу, которая может быть выполнена при помощи системы. Для того чтобы определить коэффициент  $k$ , поврежденная матрица  $M(i,j)$  восстанавливается для атрибутов в ключе (сценарном определении) и в сообщенном

решении (журнале регистрации диалога). Затем согласование между ключом и решением  $P(A)$  и измененное согласование  $P(E)$  может быть рассчитано из этой матрицы, см. таблицу 4.  $M(i,j)$  может быть рассчитана для индивидуального диалога или для группы диалогов, которые являются частью особой системы или системной конфигурации.

Коэффициент  $\kappa$  зависит от наличия простой схемы кодирования задачи, а именно в терминах AVM. Однако некоторые задачи не могут быть так легко определены. В этом случае необходимы более совершенные подходы для успешного решения задачи, которые обычно зависят от типа рассматриваемых задач.

**Таблица 4 – Параметры целенаправленного взаимодействия**

Сокр.	Наименование	Определение	Уровень взаимодействия	Метод измерения
<i>TS</i>	успех задачи	<p>Определение успеха задачи в соответствии с тем, достиг ли пользователь своей цели к концу диалога в том случае, если эта цель была достигнута при помощи системы. Определения показывают, была ли достигнута цель или нет, и предполагаемые источники проблем:</p> <ul style="list-style-type: none"> <li>• <i>TS:S</i>: Успешная (задача, для которой существуют решения)</li> <li>• <i>TS:SCs</i>: Успешная с ограниченной релаксацией от системы</li> <li>• <i>TS:SCu</i>: Успешная с ограниченной релаксацией от пользователя</li> <li>• <i>TS:SCsCu</i>: Успешная с ограниченной релаксацией от системы и от пользователя</li> <li>• <i>TS:SN</i>: Успешная в сообщении, что нет решений</li> <li>• <i>TS:F<sub>s</sub></i>: Неудачная из-за работы системы в результате адекватности системы</li> <li>• <i>TS:F<sub>u</sub></i>: Неудачная из-за работы пользователя в результате некооперативной работы пользователя</li> </ul> <p>См. также [8][7][24].</p>	диалог	эксперт
$\kappa$	коэффициент каппа	<p>Процент завершения задачи согласно каппа-статистике. Устанавливается на основе корректности результата AVM, достигнутого к концу диалога, по отношению к сценарию (ключу) AVM. В результате поврежденная матрица <math>M(i,j)</math> восстанавливается для атрибутов и в ключе, с <math>T</math> – количеством подсчетов в <math>M</math>, и <math>t_i</math> – суммой подсчетов в столбце <math>i</math> в <math>M</math>. Тогда</p> $\kappa = \frac{P(A) - P(E)}{1 - P(E)}$ <p>где <math>P(A)</math> – отношение времен соответствий AVM действительного диалога и ключа, <math>P(A) = \sum_{i=1}^n \frac{M(i,i)}{T}</math>.</p> <p><math>P(E)</math> может быть подсчитано из отношения, когда ожидается, что они согласовываются случайно, <math>P(E) = \sum_{i=1}^n \left(\frac{t_i}{T}\right)^2</math>. [31][4]</p>	диалог или набор диалогов	эксперт



## 7.5 Параметры, связанные с речевым входом

Возможности речевого входа разговорных диалоговых систем определены их возможностями распознавать слова и высказывания, извлекать смысл из распознанных строк (так называемое "речевое понимание"). Для автоматического речевого распознавания должны быть определены два метода: распознавание слова, когда извлекаются отдельные слова из речи пользователя, когда слова сказаны отдельно (изолированное распознавание слов), или непрерывное (ключевые слова определяемы). С другой стороны, непрерывное распознавание речи в состоянии распознать целые предложения или высказывания. Речевое понимание часто выполняется на базе атрибутно-значимых пар, см. раздел 7.4. Параметры, представленные в следующем пункте, относятся к распознаванию речи и речевому пониманию.

Как правило, непрерывное распознавание речи предусматривает как результат словарное строчное предположение. Для того, чтобы оценить, правильно ли строка отображает то, что было сказано, справочная транскрипция должна быть предусмотрена экспертом. Для каждого высказывания предполагаемая и ссылочная строка вначале ориентирована на уровень слов, с использованием согласующего алгоритма Динамического программирования (DP) [19] [20]. На базе выверки подсчитывается количество правильно определенных слов  $c_w$ , замен  $s_w$ , вставок  $i_w$ , и удалений  $d_w$ . Эти подсчеты могут быть связанными с общим количеством слов в ссылке  $n_w$ , подводящей итог в двух альтернативных показателях распознавательной работы, частотой ошибок слов  $WER$  и правильностью слов  $WA$ , см. таблицу 5.

Дополняющие показатели работы могут быть определены на уровне предложения, в терминах правильности предложений  $SA$ , или частоте ошибок предложений  $SER$ , см. таблицу 5. В общем,  $SA$  ниже, чем  $WA$ , потому что единственное неправильно распознанное слово в предложении сильно воздействует на  $SA$  параметр. Тем не менее, он может стать выше, чем правильность слова, особенно когда правильно распознано много предложений с единственными словами. Тот факт, что  $SER$  и  $SA$  ставят в трудное положение целые высказывания, когда встречается одиночное нераспознанное слово, был указан Strik и др. [26] [27]; проблему можно обойти с помощью параметров  $NES$  и  $WES$ , см. таблицу 5. Когда высказывания не разделены в предложениях, все метрики, относящиеся к предложению, могут также быть подсчитаны в высказываниях вместо уровня предложения.

Устройства распознавания отдельных слов осуществляют на выходе трактовку каждого входящего слова или высказывания. Слова на входе и выходе могут прямо сравниваться, и подобные показатели качества, как в непрерывном распознавательном случае, могут быть определены как пропуск вставки. Вместо вставок может быть подсчитано количество ошибочных звуковых предупреждений во временном периоде, см. van Leeuwen и Steeneken [28].  $WA$  и  $WER$  могут также быть определены для ключевых слов только в случае, когда распознаватель работает в режиме опознавания ключевых слов.

Для оценки понимания речи должны быть рассмотрены два общих подхода. Первый базируется на классификации системных ответов на вопросы пользователя в категориях правильно отвечаемых, частично правильно отвечаемых, неправильно отвечаемых или несостоявшихся ответов. Индивидуальные ответные категории могут быть комбинированы в показатели, которые использовались в программе US DARPA, см. таблицу 5. Второй способ базируется на классификации аналитических возможностей системы или в терминах правильного разбора высказываний либо правильно распознанных AVP. На базе распознанных AVP могут быть подсчитаны глобальные показатели, такие как концепция правильности,  $CA$ , концепция частоты ошибок,  $CER$  или точность понимания,  $UA$ . Все параметры перечислены в таблице 5.

Таблица 5 – Параметры взаимодействия, связанные с речевым входом

Сокр.	Наименование	Определение	Уровень взаимодействия	Метод измерения
<i>WER, WA</i>	частота ошибок слов, правильность слов	<p>Процент слов, которые распознаны правильно на основе орфографических форм, предполагаемых и (транскрибированных) рекомендованных высказываний. Выравнивание, доведенное до конца при помощи алгоритма "sclite", см. [18]. Определение <math>n_w</math> общего количества слов из всех высказываний в диалоге пользователя, <math>s_w</math>, <math>d_w</math> и <math>i_w</math> количества замененных, удаленных и вставленных слов, затем оценка ошибочных слов и неправильности слов могут быть определены следующим образом:</p> $WER = \frac{s_w + i_w + d_w}{n_w}$ $WA = 1 - \frac{s_w + i_w + d_w}{n_w} = 1 - WER$ <p>См. [24]; подробное описание того, как эти параметры могут быть вычислены в случае распознавания отдельных слов, даются в [28].</p>	слово	инст./эксперт
<i>SER, SA</i>	частота ошибок предложений, правильность предложений	<p>Процент полных предложений, которые правильно распознаны. <math>n_s</math> означает общее количество предложений, <math>s_s</math>, <math>i_s</math> и <math>d_s</math> – количество замененных, вставленных и удаленных предложений, соответственно, затем:</p> $SER = \frac{s_s + i_s + d_s}{n_s}$ $SA = 1 - \frac{s_s + i_s + d_s}{n_s} = 1 - SER$ <p>[24]</p>	высказывание	инст./эксперт
<i>NES</i>	количество ошибок на предложение	<p>Среднее количество распознанных ошибок в предложении. Где <math>s_w(k)</math>, <math>i_w(k)</math> и <math>d_w(k)</math> – количество замененных, вставленных и удаленных слов в предложении <math>k</math>, затем: <math>NES(k) = s_w(k) + i_w(k) + d_w(k)</math></p> <p>Среднее <i>NES</i> может быть вычислено следующим образом:</p> $NES = \frac{\sum_{k=1}^{\# \text{ реплики пользов.}} NES(k)}{\# \text{ реплики пользов.}} = \frac{WER \cdot \# \text{ слова пользов.}}{\# \text{ реплики пользов.}}$ <p>[26]</p>	высказывание	инст./эксперт

Таблица 5 – Параметры взаимодействия, связанные с речевым входом

Сокр.	Наименование	Определение	Уровень взаимодействия	Метод измерения
<i>WES</i>	количество ошибок на предложение	Связанные с <i>NES</i> , но урегулированные с количеством слов в предложении <i>k</i> , <i>w(k)</i> : $WES(k) = \frac{NES(k)}{w(k)}$ Среднее <i>WES</i> может быть вычислено следующим образом: $WES = \frac{\sum_{k=1}^{\#отмены\ пользов.} WES(k)}{\# отмены\ пользов.}$ [26]	слово	инст./эксперт
<i>AN:CO</i> , <i>AN:IN</i> , <i>AN:PA</i> , <i>AN:FA</i> , <i>%AN:CO</i> , <i>%AN:IN</i> , <i>%AN:PA</i> , <i>%AN:FA</i>	количество или процент правильных/неправильных/частично правильных/несостоявшихся системных ответов	Общее количество или процент вопросов от пользователя, которые: • правильные ( <i>AN:CO</i> ); • неправильные ( <i>AN:IC</i> ); • частично правильные ( <i>AN:PA</i> ); • несостоявшиеся ( <i>AN:FA</i> ), на которые был получен ответы системы, по диалогу, см. [21][12][14].	высказывание	эксперт
<i>DARPA<sub>s</sub></i> , <i>DARPA<sub>me</sub></i>	<i>DARPA</i> счет, <i>DARPA</i> исправленная ошибка	Показатели способности речевого понимания согласно <i>DARPA</i> , модифицированные Skowronek [25] [17], для того чтобы считать частично правильные ответы: $DARPA_s = \frac{AN : CO - AN : IC}{\# вопросы\ пользов.}$ $DARPA_{me} = \frac{AN : FA + 2 \cdot (AN : IC + AN : PA)}{\# вопросы\ пользов.}$ [21][12][25]	высказывание	эксперт
<i>PA:CO</i> , <i>PA:PA</i> , <i>PA:IC</i> , <i>%PA:CO</i> , <i>%PA:PA</i> , <i>%PA:IC</i>	количество правильных/частично правильных/неправильных проанализированных высказываний пользователя	Определение количества концепций (атрибутивно-значимых пар, AVP) в высказываниях, которые получены системой: • <i>PA:CO</i> : Все концепции высказываний пользователя правильно поняты системой. • <i>PA:PA</i> : Не все, но по крайней мере одна концепция высказываний пользователя правильно понята системой. • <i>PA:IC</i> : Нет концепций высказываний пользователя, правильно понятых системой. Отображенный как общее количество или процент высказываний пользователя в диалоге, которые проанализированы правильно/частично правильно/неправильно. [7]	высказывание	эксперт

**Таблица 5 – Параметры взаимодействия, связанные с речевым входом**

Сокр.	Наименование	Определение	Уровень взаимодействия	Метод измерения
<i>CA, CER</i>	концептуальная правильность, частота концептуальных ошибок	<p>Процент правильно понятых семантических элементов по диалогу. Концепции определены, как атрибутивно-значимые пары (AVP), с <math>n_{AVP}</math> – общим числом AVP, <math>s_{AVP}</math>, <math>i_{AVP}</math> и <math>d_{AVP}</math> – количество замененных, вставленных и удаленных AVP. Концептуальная правильность и частота концептуальных ошибок может быть определена следующим образом:</p> $CA = 1 - \frac{s_{AVP} + i_{AVP} + d_{AVP}}{n_{AVP}}$ $CER = \frac{s_{AVP} + i_{AVP} + d_{AVP}}{n_{AVP}}$ <p>[9][24][3][2]</p>	высказывание	эксперт
<i>UA</i>	правильность понимания	<p>Процент высказываний пользователя, в котором все семантические элементы (AVP) правильно получены:</p> $UA = \frac{PA : CO}{\# \text{отмены пользов.}}$ <p>[32]</p>	высказывание	эксперт

## 7.6 Дополнительные параметры

Большинство интерактивных параметров, перечисленных в таблицах, описывают работу системы, что очевидно, потому что именно система и качество услуг представляют интерес. В дополнение к этому, могут быть определены параметры, связанные с пользователем. Они являются специфическими для тестируемой группы пользователей, но тем не менее могут быть близко связаны с качественными возможностями, получаемыми пользователем.

Когда качество услуг, основанных на SDS, выделяется в качественные аспекты способом, который указан в 5.3/P.851, можно заметить, что несколько аспектов качества не затрагивается интерактивными параметрами. Нет параметров, прямо связанных с удобством работы, удовлетворенностью пользователя, доступностью или качеством выходящей речи. До настоящего времени было создано очень мало методов, которые определяют качество выходящей речи (связанной или синтезированной) параметрическим способом. Инструментальные показатели, связанные с понятностью речи, определены, например, в МЭК 60268-16 [15], но они не разработаны для телефонного оборудования. Были предложены объединенные синтезированные стоимостные показатели, которые могут быть вычислены из входящих текстов и речевых баз данных объединенных синтезирующих систем [5]. Несмотря на то, что они иногда показывают высокую корреляцию с множеством MOS, полученным в слуховых экспериментах, такие измерения очень специфичны для речевого синтезатора и объединенных баз синтезаторов.

## 8 Интерпретация значений параметров взаимодействия

Несмотря на то, что интерактивные параметры, как это определено в настоящем Добавлении, являются важными для проектирования систем, оптимизации и технического обслуживания, они не прямо связаны с качеством, которое воспринимается пользователем – человеком. Следовательно, совокупность интерактивных параметров следует дополнять сбором мнений пользователя о различных аспектах качества, как описано в Рек. МСЭ-Т Р.851. Только таким образом может быть получена достоверная информация о качестве услуг, которые основаны на разговорных диалоговых системах.

Интерпретация значений параметров взаимодействия может основываться на экспериментальных поисках, которые, однако, часто являются специфичными для рассматриваемой системы или для услуг. Например, возрастающее количество тайм-аут подсказок может указывать, что пользователь не знает, что говорить в особых местах диалога или что он/она запутался в действиях системы [29]. Возрастающее число попыток вмешаться в разговор может просто отображать, что пользователь изучал возможности прерывать систему. Напротив, уменьшающееся количество может также отображать, что пользователь не знает, что говорить системе. Многословная речь пользователя может следовать из большого количества инициатив, свойственных пользователю. Можно ожидать уменьшения значений параметра метаобщения (особенно метаобщения начинающего пользователя) для того, чтобы увеличить устойчивость системы, плавность диалогов и эффективность общения [1].

## БИБЛИОГРАФИЯ

- [1] BERNSEN, N.O., DYBKJÆR, H., DYBKJÆR, L.: *Designing interactive speech systems: From first ideas to user testing*, Springer, DE-Berlin, 1998.
- [2] BILLI, R., CASTAGNERI, G., DANIELI, M.: Field trial evaluations of two different information inquiry systems, *Proc. 3rd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA'96)*, US-Basking Ridge NJ, pp. 129-134, 1996.
- [3] BOROS, M., ECKERT, W., GALLWITZ, F., GORZ, G., HANRIEDER, G., NIEMANN, H.: Towards understanding spontaneous speech: Word accuracy vs. concept accuracy, *Proc. 4th Int. Conf. on Spoken Language Processing (ICSLP'96)*, IEEE, US-Piscataway NJ, 2, pp. 1009-1012, 1996.
- [4] CARLETTA, J.: Assessing agreement of classification tasks: The kappa statistics, *Computational Linguistics*, Vol. 22(2), pp. 249-254, 1996.
- [5] CHU, M., PENG, H.: An objective measure for estimating MOS of synthesized speech, *Proc. 7th Europ. Conf. on Speech Communication and Technology (Eurospeech 2001 – Scandinavia)*, DK-Aalborg, 3, pp. 2087-2090, 2001.
- [6] COOKSON, S.: Final evaluation of VODIS – Voice operated data inquiry system, *Proc. of Speech'88, 7th FASE Symposium*, UK-Edinburgh, 4, pp. 1311-1320, 1988.
- [7] DANIELI, M., GERBINO, E.: Metrics for evaluating dialogue strategies in a spoken language system, *Empirical Methods in Discourse Interpretation and Generation. Papers from the 1995 AAI Symposium*, US-Stanford CA, AAAI Press, US-Menlo Park CA, pp. 34-39, 1995.
- [8] FRASER, N.: Assessment of interactive systems, *Handbook on Standards and Resources for Spoken Language Systems* (D. Gibbon, R. Moore and R. Winski, eds.), Mouton de Gruyter, DE-Berlin, pp. 564-615, 1997.
- [9] GERBINO, E., BAGGIA, P., CIARAMELLA, A., RULLENT, C.: Test and evaluation of a spoken dialogue system, *Proc. Int. Conf. Acoustics Speech and Signal Processing (ICASSP'93)*, IEEE, US-Piscataway NJ, 2, pp. 135-138, 1993.
- [10] GIBBON, D., MOORE, R., WINSKI, R., Eds.: *Handbook on Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, DE-Berlin, 2000.
- [11] GLASS, J., POLIFRONI, J., SENEFF, S., ZUE, V.: Data collection and performance evaluation of spoken dialogue systems: The MIT experience, *Proc. 6th Int. Conf. on Spoken Language Processing (ICSLP 2000)*, CN-Beijing, 4, pp. 1-4., 2000.
- [12] GOODINE, D., HIRSCHMAN, L., POLIFRONI, J., SENEFF, S., ZUE, V.: Evaluating interactive spoken language systems, *Proc. 2nd Int. Conf. on Spoken Language Processing (ICSLP'92)*, CA-Banff, 1, pp. 201-204, 1992.
- [13] GRICE, H.P.: Logic and conversation, *Syntax and Semantics, Vol. 3: Speech Acts* (P. Cole and J.L. Morgan, eds.), Academic Press, US-New York NY, pp. 41-58, 1975.
- [14] HIRSCHMAN, L., PAO, C.: The cost of errors in a spoken language system, *Proc. 3rd Europ. Conf. on Speech Communication and Technology (Eurospeech'93)*, DE-Berlin, 2, pp. 1419-1422, 1993.
- [15] IEC 60268-16 (2003), *Sound system equipment – Part 16: Objective rating of speech intelligibility by speech transmission index*. International Electrotechnical Commission, CH-Geneva.

- [16] KAMM, C.A., LITMAN, D.J., WALKER, M.A.: From novice to expert: The effect of tutorials on user expertise with spoken dialogue systems, *Proc. 5th Int. Conf. on Spoken Language Processing (ICSLP'98)*, AU-Sydney, 4, pp. 1211-1214, 1998.
- [17] MÖLLER, S.: *Quality of telephone-based spoken dialogue systems*. Springer, US-New York NY, 2005.
- [18] NIST Speech Recognition Scoring Toolkit, *Speech recognition scoring toolkit*, National Institute of Standards and technology, <http://www.nist.gov/speech/tools>, US-Gaithersburg MD, 2001.
- [19] PICONE, J., DODDINGTON, G.R., PALLETT, D.S.: Phone-mediated word alignment for speech recognition evaluation, *IEEE Trans. Acoustics Speech and Signal Processing*, Vol. 38(3), pp. 559-562, 1990.
- [20] PICONE, J., GOUDIE-MARSHALL, K.M., DODDINGTON, G.R., FISHER, W.: Automatic text alignment for speech system evaluation, *IEEE Trans. Acoustics Speech and Signal Processing*, Vol. 34(4), pp. 780-784, 1986.
- [21] POLIFRONI, J., HIRSCHMAN, L., SENEFF, S., ZUE, V.: Experiments in evaluating interactive spoken language systems, *Proc. DARPA Speech and Natural Language Workshop*, US-Harriman CA, pp. 28-33, 1992.
- [22] PRICE, P.J., HIRSCHMAN, L., SHRIBERG, E., WADE, E.: Subject-based evaluation measures for interactive spoken language systems, *Proc. DARPA Speech and Natural Language Workshop*, US-Harriman CA, pp. 34-39, 1992.
- [23] SAN-SEGUNDO, R., MONTERO, J.M., COLÁS, J., GUTIÉRREZ, J., RAMOS, J.M., PARDO, J.M.: Methodology for dialogue design in telephone-based spoken dialogue systems: A Spanish train information system, *Proc. 7th Europ. Conf. on Speech Communication and Technology (Eurospeech 2001 – Scandinavia)*, DK-Aalborg, 3, pp. 2165-2168, 2001.
- [24] SIMPSON, A., FRASER, N.M.: Black box and glass box evaluation of the SUNDIAL system, *Proc. 3rd Europ. Conf. on Speech Communication and Technology (Eurospeech'93)*, DE-Berlin, 2, pp. 1423-1426, 1993.
- [25] SKOWRONEK, J.: *Entwicklung von Modellierungsansätzen zur Vorhersage der Dienstqualität bei der Interaktion mit einem natürlichsprachlichen Dialogsystem*. Diploma thesis (unpublished), Institut für Kommunikationsakustik, Ruhr-Universität, DE-Bochum, 2002.
- [26] STRIK, H., CUCCHIARINI, C., KESSENS, J.M.: Comparing the performance of two CSRs: How to determine the significance level of the differences, *Proc. 7th Europ. Conf. on Speech Communication and Technology (Eurospeech 2001 – Scandinavia)*, DK-Aalborg, 3, pp. 2091-2094, 2001.
- [27] STRIK, H., CUCCHIARINI, C., KESSENS, J.M.: Comparing the recognition performance of CSRs: In search of an adequate metric and statistical significance test, *Proc. 6th Int. Conf. on Spoken Language Processing (ICSLP 2000)*, CN-Beijing, 4, pp. 740-743, 2000.
- [28] VAN LEEUWEN, D., STEENEKEN, H.: Assessment of recognition systems, *Handbook on Standards and Resources for Spoken Language Systems* (D. Gibbon, R. Moore and R. Winski, eds.), Mouton de Gruyter, DE-Berlin, pp. 381-407, 1997.
- [29] 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 Conference Proc.*, US-Los Angeles CA, ACM, US-New York NY, pp. 582-589, 1998.

- [30] WALKER, M.A., LITMAN, D.J., KAMM, C.A., ABELLA, A.: Evaluating spoken dialogue agents with PARADISE: Two case studies, *Computer Speech and Language*, Vol. 12(3), pp. 317-347, 1998.
- [31] WALKER, M.A., LITMAN, D.J., KAMM, C.A., ABELLA, A.: PARADISE: A framework for evaluating spoken dialogue agents, *Proc. of the 35th Ann. Meeting of the Assoc. for Computational Linguistics*, ES-Madrid, pp. 271-280, 1997.
- [32] ZUE, V., SENEFF, S., GLASS, J.R., POLIFRONI, J., PAO, C., HAZEN, T.J., HETHERINGTON, L.: JUPITER: A telephone-based conversational interface for weather information, *IEEE Trans. Speech and Audio Processing*, Vol. 8(1), pp. 85-96, 2000





## СЕРИИ РЕКОМЕНДАЦИЙ МСЭ-Т

Серия А	Организация работы МСЭ-Т
Серия D	Общие принципы тарификации
Серия E	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 служб
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	Управление электросвязью, включая СУЭ и техническое обслуживание сетей
Серия N	Техническое обслуживание: международные каналы передачи звуковых и телевизионных программ
Серия O	Требования к измерительной аппаратуре
<b>Серия P</b>	<b>Качество телефонной передачи, телефонные установки, сети местных линий</b>
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, взаимосвязь открытых систем и безопасность
Серия Y	Глобальная информационная инфраструктура, аспекты межсетевых протоколов и сети последующих поколений
Серия Z	Языки и общие аспекты программного обеспечения для систем электросвязи