国 际 电 信 联 盟

ITU-T

国际电信联盟
电信标准化部门

P系列

增补 24
(10/2005)

P系列：电话传输质量、电话装置、本地线路网络

描述与语音对话系统交互的参数

ITU-T P系列建议书 – 增补24

国际电信联盟

ITU-T P系列建议书

电话传输质量、电话装置和本地线路网络

名词术语和传输参数对用户传输质量意见的影响	系列 P.10
用户线和话机	系列 P.30 P.300
传输标准	系列 P.40
客观测量装置	系列 P.50 P.500
客观电声测量	系列 P.60
与话音响度有关的测量	系列 P.70
质量的客观和主观评定方法	系列 P.80 P.800
多媒体业务的音视频质量	系列 P.900
IP端点的传输性能和业务质量问题	系列 P.1000

欲了解更详细信息，请查阅*ITU-T*建议书目录。

ITU-T P系列建议书增补24

描述与语音对话系统交互的参数

摘 要

本增补定义了一系列能够从依托语音对话系统的服务中获取的参数。记录（测试）用户与所研究的服务之间的交互，能够从中获取这些参数。这些参数量化了交互的流程、用户和系统的行为，以及交互中采用话音技术设备的性能，为系统的开发、优化和维护提供了有用的信息，是对符合ITU-T P.851建议书的主观质量判定方法的补充。

来 源

ITU-T 第12研究组（2005-2008）于2005年10月21日批准了ITU-T P系列建议书增补24。

关键词

评定、自动话音识别、自动话音理解、对话管理、交互参数、话音产生、话音技术、语音对话系统。

前　言

国际电信联盟（ITU）是从事电信领域工作的联合国专门机构。ITU-T（国际电信联盟电信标准化部门）是国际电信联盟的常设机构，负责研究技术、操作和资费问题，并且为在世界范围内实现电信标准化，发表有关上述研究项目的建议书。

每四年一届的世界电信标准化全会（WTSA）确定ITU-T各研究组的研究课题，再由各研究组制定有关这些课题的建议书。

WTSA第1号决议规定了批准建议书须遵循的程序。

属ITU-T研究范围的某些信息技术领域的必要标准，是与国际标准化组织（ISO）和国际电工技术委员会（IEC）合作制定的。

注

本建议书为简要而使用的“主管部门”一词，既指电信主管部门，又指经认可的运营机构。

遵守本建议书的规定是以自愿为基础的，但建议书可能包含某些强制性条款（以确保例如互操作性或适用性等），只有满足所有强制性条款的规定，才能达到遵守建议书的目的。“应该”或“必须”等其他一些强制性用语及其否定形式被用于表达特定要求。使用此类用语不表示要求任何一方遵守本建议书。

知识产权

国际电联提请注意：本建议书的应用或实施可能涉及使用已申报的知识产权。国际电联对无论是其成员还是建议书制定程序之外的其他机构提出的有关已申报的知识产权的证据、有效性或适用性不表示意见。

至本建议书批准之日止，国际电联尚未收到实施本建议书可能需要的受专利保护的知识产权的通知。但需要提醒实施者注意的是，这可能不是最新信息，因此大力提倡他们查询电信标准化局（TSB）的专利数据库。

目　录

ITU-T P系列建议书增补24

描述与语音对话系统交互的参数

1 范围

如同系统开发者和服务运营者所看到的一样，本增补描述了一些参数，这些参数提供了关于与基于语音对话系统的服务交互的信息。本增补所称的语音对话系统可以通过电话网络与人进行轮流的口头语言交互，具有自动话音识别、话音理解、对话管理、响应产生和话音输出能力，它们可以访问存储在数据库中的信息，或者处理不同类型的事务。

这里定义的参数量化了交互的流程、用户和系统的行为，以及交互采用的话音技术设备的性能。为了获取所有的参数，语音对话系统必须像一只玻璃盒一样是可以访问的；然而，有些参数也可以采用黑盒的方法获得，即无需访问单个系统的组件。有些参数可以自动地获取，有些参数则要依赖于人类专家对交互日志文件的抄录和注解。从系统开发者的角度来看，这些参数说明了系统的性能；因此，它们为语音对话系统的主观评估实验提供了补充信息，ITU-T P.851建议书给出了关于主观评估实验的建议。关于主观评估通用方法和话音输出设备评定的更多指导，参见ITU-T P.800和P.85建议书以及通话计时手册。本增补所列举的参数没有具体地涉及由传输信道产生的可能衰减，这些影响是ITU-T第12研究组下一步的研究项目。

2 参考文献

- 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 Recommendation P.851 (2003), *Subjective quality evaluation of telephone services based on spoken dialogue systems.*
- ITU-T *Handbook on Telephonometry* (1992).

3 定义

这里没有列出的定义，请参考ITU-T P.10建议书。

3.1 barge-in插话：人越过系统的提示或系统的输出，讲话的能力[10]。

3.2 dialogue对话：会话或者信息交流。作为一个评估单元：几种可能通过对话结构的途径之一。

3.3 efficiency效率：度量方法的准确性和系统任务的完整性，与为了完成特定的系统任务所用的资源（如，时间、人的努力）有关。

3.4 exchange交流：两次连续的、有关联的轮流讲话，在对话中各方讲话[8]。

3.5 functionality功能性：系统在特定的条件下使用，提供功能来满足确定的和隐含的需要的能力。

3.6 meta-communication元通信：关于沟通上的交流，如，为了消除误解（“我正确理解你了吗？”）或者在措辞方面达成共识。

3.7 performance性能：设备提供其设计功能的能力。

3.8 speech technology话音技术：该学科与口头语言输入输出系统的研究和开发有关，运用了声学、电机工程、统计学、语音学、自然语言处理等相邻学科的成果，包括系统的要求说明、设计、实现和评估，语言资料集和语言资源处理，以及面向用户的产品评估[10]。

3.9 spoken dialogue system语音对话系统：人们可以与其进行轮流口头语言交互的一种计算机系统。

3.10 task任务：为了达到某一领域的确定目标，用户必须完成的所有工作。

3.11 task-oriented dialogue面向任务的对话：与特定主题有关的对话，对话有一个明确的目标（如解决一个问题或者获得特定的信息）[8]。

3.12 transaction事务：专注于单一高级任务（如，进行旅游预约或者核查银行账户的余额）对话的一部分。事务可以与对话持续的时间一样长，或者一个对话可以包含一个以上的事务[8]。

3.13 turn轮流讲话：话语。一段时间的话音，在对话中某一方讲话，从这方开始讲话直到另一方明确地接替讲话为止[1]。

3.14 utterance话语：参见轮流讲话。

4 缩写

ASR	自动话音识别
AVM	属性值矩阵
AVP	属性值对
DARPA	国防高级研究项目署
DP	动态编程
DTMF	双音多频
IVR	交互式的声音响应
MOS	平均主观得分
SDS	语音对话系统
WoZ	阿兹国的魔术师

5 引言

语音对话系统（SDS），即人们可以与其进行轮流口头语言交互的计算机系统，是现代电话网络的一部分。系统实现了通过电话访问数据库和事务，例如，为了获得火车或航班时刻表、股票汇率、旅游信息，进行银行账户操作或者旅馆预订。与采用DTMF输入的简单交互式的声音响应(IVR)系统相比，SDS具有全程的语音交互能力，包括识别用户的语音、已识别的单词中指定的意思、关于如何继续对话的决策、语音响应的简洁表达和产生输出给用户的语音，从而，实现了用户与系统之间或多或少“自然”的语音交互。

为了从用户的角度来评估SDS服务的质量，ITU-T第12研究组于2003年提出了ITU-T P.851建议书。为了从用户的角度来评定其质量，这份建议书描述了将SDS作为黑盒进行主观评估实验的方法。如同用户看到的一样，在按照ITU-T P.851建议书进行实验的帮助下，可以获得质量方面的有用信息，然而，很难确定单个系统的组件对于用户体验到的整体质量的影响，例如，出现交互问题时，确定哪个组件需要改进。这样，从系统设计者和服务运营者的角度说明系统性能的信息可以作为评估的补充。

与系统有关的信息可以用通常所说的交互参数来描述。这些参数有助于量化交互的流程、用户和系统的行为和交互中采用话音技术设备的性能。它们从系统开发者和服务运营者的角度说明了系统的性能，从而为主观评估的数据提供了补充信息。为了获取其中一些参数，语音对话系统必须能被看作一个可以访问的玻璃盒；另外的一些参数也可以采用黑盒的方法获得，即无需访问单个系统的组件。

本增补提供了一个交互参数集，在过去的15年中这些参数已经被用来评估SDS。所列举的参数均与用户和系统之间整个的信息交流有关：出现误解时的元通信、系统的协同性、在系统的帮助下能够完成的任务和系统的语音输入能力。至今仍没有关于语音输出质量的参数化描述（如，关于合成语音质量）。本参数集的理论基础在参考资料[17]中有描述。

并不是所有的交互参数都与感觉到的、SDS服务的质量有着直接的关联。实际上，单个参数与用户对质量的判断之间的相关性通常是十分有限的。尽管如此，提出大量的、描述用户与系统之间交互的参数是有益的，通过这种方法，获得了从系统设计者的角度、与感觉到的质量有着潜在相关性的绝大多数信息。这些参数为系统的开发、优化和维护提供了有用的信息。

这些参数一旦被明确并在不同的试验地点用于评估实验，就可为各种系统和服务方便地评估它们对感觉到的质量的影响。这样，以这些交互参数作为基础，就有可能开发出预测质量的算法。 ITU-T第12研究组和其它机构仍在对此方向进行研究。

6 交互参数的特性

当实际的或测试的用户与服务交互时，能够获取交互参数。有些参数可以通过仪器获得，有些参数则需借助人类专家对于日志文件的抄录和注解。简单的参数，如交互或单段话语的持续时间，通常可采用适当的算法、完全通过仪器测量得到。另一方面，当不仅表达出了表面上的东西（语音信号），而且还有系统或用户话语的内容和含义时，人的抄录和注解是必需的（如，确定一个单词或概念的准确性）。

SDS是如此的复杂，以至于描述系统的状态和对比不同的系统或者系统的不同版本，都需要以大量不同的参数作为基础[24]。因此，为了得到尽可能多的信息，应当采取两种获得交互参数的方法（基于仪器和专家）。以获得的信息为基础，才能对语音对话系统进行非常高效地优化和维护。

由于交互参数是以从用户与系统的交互中获得的数据作为基础，它们会受到系统的、用户的以及双方之间交互的特性的影响。这些影响往往不是孤立的，原因是用户行为会受到系统行为强烈的影响（如，系统提出的问题），反之亦然（如，用户的词汇和语言风格会影响系统识别和理解的准确性）。因而，交互参数会明显地反映用来获取参数的用户群的特性。

交互参数既可以在受控的条件下从实验室测试环境中得到，也可以从现场测试中得到。在后一种情况下，由于收集不到所有的必要信息，不大可能获取所有的参数。例如，如果面向任务的交互（如，获得火车时刻表）确定取得成功，则必须了解用户的确切目的。这样的信息只能从实验环境中获得，例如，ITU-T P.851建议书对这种方式有描述。如果系统并不是十分完整，可以从一个所谓的“阿兹国的魔术师”（WoZ）仿真中获得参数，在仿真中，用实验人员来代替被测系统中缺少的部分。在说明所获得的参数时，必须考虑这个仿真的特性。

交互参数可以按单词级、句子级或者话语级，或者整个交互或对话级来计算。计算单词或话语级的参数时，通常会计算每次对话的平均值。从一个特定用户群获得的参数，要就系统（版本）、用户群和实验环境（设想、测试环境等）的影响，运用标准统计的方法来进行分析。ITU-T P.851建议书描述了这些影响的特性化。

7 交互参数回顾

通过广泛地查阅文献，可以确定在过去的15年里已被用于各种评定和评估实验的参数。相应文献见 [2][3][4][6][7][8][9][11][12][14][16][21][22][23][24][25][26][27][28][30][31][32]，参考资料[17]已经对这些参数进行了总结。可以粗略地将这些参数分为以下几类：

- 与对话和交流有关的参数；
- 与元通信有关的参数；
- 与协同性有关的参数；
- 与任务有关的参数；
- 与语音输入有关的参数。

下面将简要地讨论这些类别。对于每一类，将列出相应的参数，连同定义、参数表示的交互级别（单词、话语或对话），以及测量的方法（仪器或者基于专家的注解）。

7.1 与对话和交流有关的参数

与整个对话和信息交流有关的参数对交互是如何发生的给出了一个非常粗略的说明，它们并没有具体地明确每一段话语的信息传递功能。本类参数如表1所示，包括与持续时间有关的参数（整个对话持续时间，系统和用户轮次的持续时间，系统和用户响应延时），与单词和轮次有关的参数（系统和用户轮次的平均数量，每轮系统和每轮用户讲话的平均单词数，系统和用户提出问题的数量）。

参考资料[11]提出的两个参数值得关注：*询问密度*表明用户能以怎样的效率向系统提供新信息，*概念效率*描述的是系统能以怎样的效率吸取用户信息。这些参数也与系统的语言理解能力有关，但是之所以把它们包含在本节中，是因为它们是将系统的交互能力作为一个整体而得出的参数，而不是纯粹地由语言理解能力得出来的。

本类中所有的参数都有全局特性，是将对话作为一个整体，尽管这些参数在话语级是从局部来进行计算的。全局参数有时很难获得，原因是认知技巧的个体差异可能与源自系统的不同有很大的关联，研究对象为了完成任务可能会学习策略，这对全局参数有重大的影响。

表1–与对话和交流有关的交互参数

缩写	名称	定义	交互级别	测量方法
DD	对话持续时间	整个对话持续时间，以ms计，参见[8][6][12][21]。	对话	仪器
STD	系统轮次持续时间	轮到系统讲话的平均持续时间，从系统开始讲话到系统停止讲话，以ms计。一个轮次是一段话语，即对话中一方讲话的一段时间。[8]	话语	仪器
UTD	用户轮次持续时间	轮到用户讲话的平均持续时间，从用户开始讲话到用户停止讲话，以ms计。[8]	话语	仪器
SRD	系统响应延时	系统响应的平均延时，从用户停止讲话到系统开始讲话，以ms计。[22]	话语	仪器
URD	用户响应延时	用户响应的平均延时，从系统停止讲话到用户开始讲话，以ms计。[22]	话语	仪器
#轮次	轮次数量	对话中轮流讲话的总次数。[30]	对话	仪器/专家
#系统轮次	系统轮次数量	对话中轮到系统讲话的总次数。[30]	对话	仪器/专家
#用户轮次	用户轮次数量	对话中轮到用户讲话的总次数。[30]	对话	仪器/专家
WPST	每个系统轮次单词数	对话中每个系统轮次的平均单词数。[6]	话语	仪器/专家
WPUT	每个用户轮次单词数	对话中每个用户轮次的平均单词数。[6]	话语	仪器/专家

表1–与对话和交流有关的交互参数

缩写	名称	定义	交互级别	测量方法
#系统问题	系统提出问题数	每次对话系统提出问题的总数。	对话	专家
#用户问题	用户提出问题数	每次对话用户提出问题的总数。 [12][21]	对话	专家
QD	询问密度	每次用户询问引入新概念（位置，见7.4节）的平均数。对话数记作n_d，在第*i*次对话中用户询问的总次数记作$n_q(i)$，在第*i*次对话中系统正确“理解”且前面没有的概念数记作$n_u(i)$，则 $$QD = \frac{1}{n_d}\sum_{i=1}^{n_d}\frac{n_u(i)}{n_q(i)}$$ 前面某一段话语中系统已经理解了的概念不计入$n_u(i)$。[11]	对话组	专家
CE	概念效率	每使系统“理解”一个概念必需的平均轮次数。对话数记作n_d，在第*i*次对话中系统正确“理解”且前面没有的概念数记作$n_u(i)$，在第*i*次对话中总的概念数记作$n_c(i)$，则 $$CE = \frac{1}{n_d}\sum_{i=1}^{N_d}\frac{n_u(i)}{n_c(i)}$$ 只要用户说出的且尚未被系统理解的概念均计算在内。[11]	对话组	专家

7.2 与元通信有关的参数

元通信，即关于沟通上的交流，对于与识别、理解和推理能力受限的系统的语音交互是十分重要的。在这种情况下，为了消除误解，纠错和澄清话语或者甚至辅助对话都是必需的。

系统的和用户的话语数量是元通信的一部分，本类参数对它们进行了量化。绝大多数参数都是计算在一次关于特定交互问题的对话中，话语的绝对数量，然后求出一组对话的平均值，包括用户求助请求的数量、系统超时提示的数量、当用户的话语中没有语义内容可以获取时被系统拒绝的数量（ASR拒绝）、诊断系统错误信息数、用户插话尝试的次数和用户尝试取消以前动作的次数。

有两种方法可以描述系统（和用户）从交互问题中恢复的能力：既可以明确地用纠错率，即与纠正一个交互问题有主要关系的所有（系统或用户）轮次的百分比，又可以隐含地用隐式恢复参数，这个参数量化了系统重新识别或理解那些没有完全识别或理解的话语的能力。

与全局的度量方法相比，绝大多数与元通信有关的参数描述的是在交流过程中系统和用户话语的功能。因此，绝大多数参数必须要在注释专家的帮助下得以确定，参数如表2所示。

表 2 – 与元通信有关的交互参数

缩写	名称	定义	交互级别	测量方法
#求助请求	用户求助请求数	在一次对话中用户求助请求的总数。如果用户明确地请求帮助，用户的求助请求会被注释专家进行标注。请求可以简洁地表达为一个问题（如，“什么是有效的选择？”）或一个陈述（“告诉我有效的选择！”）。[30]	话语	专家
#系统帮助	诊断系统帮助信息数	在对话中系统产生的帮助信息的总数。帮助信息是系统在对话中发出的、告诉用户在某一时刻如何进行有效选择的话语。	话语	仪器/专家
#超时	超时提示的次数	在对话中，由于用户没有响应，超时提示的总数。[30]	话语	仪器
#ASR拒绝	ASR拒绝的次数	对话中ASR拒绝的总数。ASR拒绝定义为一种系统提示，表明系统不能“听见”或“理解”用户，即系统不能从用户的话语中获取任何含义。[30]	话语	仪器
#系统错误	诊断系统错误信息数	对话中系统发出的诊断错误信息的总数。诊断错误信息定义为系统发出的说明系统不能执行某项任务或者提供某一条信息的话语。[22]	话语	仪器/专家
#插话	用户插话尝试的次数	对话中用户插话尝试的总数。当系统正在讲话时，用户有意讲话记作一次用户插话尝试。在本定义中，不是故意影响对话进程的用户话语（笑，愤怒或礼貌的话）不记为插话。[30]	话语	专家
#取消	用户取消尝试的次数	对话中用户取消尝试的总数。如果用户企图从头重新开始对话，或者他 / 她明确地想在对话系统中后退一级或几级时，用户这次话语就被当作一次取消尝试。[16][23]	话语	专家

表 2 – 与元通信有关的交互参数

缩写	名称	定义	交互级别	测量方法
SCT, SCR	系统纠错轮次数，系统纠错率	在一次主要与纠正“问题”有关的对话中，所有系统轮次的总数（SCT）或百分比（SCR），这次对话不会提供新的提议性的内容而且会中断对话流程。“问题”可能是由话音识别或理解错误引起的，或者是由不合逻辑的、自相矛盾的或者不明确的用户话语引起的。当用户对系统的提问不作回答时，相应的系统回答会被当作一个系统纠错轮次，除非系统对于用户要求的信息或动作从功能上不予以支持。[8][24][9][7]	话语	专家
UCT, UCR	用户纠错轮次数，用户纠错率	在一次主要与纠正“问题”有关的对话中，所有用户轮次的总数（UCT）或百分比（UCR），这次对话不会提供新的提议性的内容且会中断对话流程（参见SCT，SCR）。[8][24][9][7]	话语	专家
IR	隐式恢复	系统从那些没有完全识别或理解的用户话语中重新获得信息的能力。系统的响应是否“适当”取决于加了标记的部分解析的话语（见7.5节中PA：PA的定义）： $IR = \frac{\#\ \text{有着适当系统响应的话语}}{PA : PA}$ “适当性”的定义见7.3节。[7]	话语	专家

7.3 与协同性有关的参数

协同性已经被当作是与语音对话系统成功交互的关键[1]。不幸的是，量化系统运行是否协调比较困难。与对话、元通信有关的一些参数和系统的协同性存在着某种的关联，但它们并不能量化协同性。

Simpson和Fraser [24]提出的语境适合参数可直接度量协同性。系统的每一段话语是否违反了一条或多条关于协同性的Grice准则必须由众多的专家来判断，见参考资料[13]：

- 信息的量化：只提供需要的信息（为了当前交流的目的）；不提供需要的多余的信息。
- 质量：设法让你提供的信息是真实的；不说你认为是虚假的信息；不说你缺少足够证据的信息。
- 关系：是相关的。
- 风格：清晰明了；避免晦涩的表达；避免含糊；简洁（避免不必要的罗嗦）；有条理。

关于语音对话系统的这些准则，Bernsen和Dybkjær [1] 已经作了更为准确地阐述。

话语可分为几类：适当的（不违反Grice准则）、不适当的（违反了一条或多条准则）、介于适当与不适当之间的（专家在它们的分类上不能达成共识）、不能理解的（在对话的语境中，不能理解话语的内容）和完全失败（系统不能做出语言响应）。必须注意分类并不总是一目了然的，分类原则的说明可能是必需的。

表 3 – 与协同性有关的交互参数

缩写	名称	定义	交互级别	测量方法
CA:AP, CA:IA, CA:TF, CA:IC, %CA:AP, %CA:IA, %CA:TF, %CA:IC	语境适当性	由紧接着的对话语境判定为适当的系统话语的总数或百分比。通过标注话语是否违反了一条或多条协同性方面的Grice准则来确定： • *CA:AP*：适当的，没有违反Grice准则，在某些方面没有格外突出地或明显地违反Grice准则。 • *CA:IA*：不适当的，违反了一条或多条Grice准则。 • *CA:TF*：完全失败，没有语言响应。 • *CA:IC*：不能理解的，注释专家不能理解其内容。 更多的详细资料见参考资料[24][8][9]；分类方法与参考资料[14]收录的分类方法相似。	话语	专家

7.4 与任务有关的参数

当前最高技术水平的服务能够实现系统和用户之间面向任务的交互，任务的成功是服务有效性的关键。在实验室环境里，会赋予测试对象一个明确的任务，因而实验室环境最适宜判定任务的成功与否，见ITU-T P.851建议书。然而，在实际度量任务成功与否时必须考虑由用户引起的、与设想之间可能的偏差，或者因为他/她不重视设想中给出的说明，或者因为他/她疏忽了系统的话语，或者因为任务无法完成、不得不在对话过程中修改任务。

在文献中记载的绝大多数任务成功界定都考虑了试验性任务的修改。通过放宽系统的或者用户的（或者两者的）限制，或者发现不存在完成规定任务的办法，对于说明中的约束条件简单地给出正确的解答就可取得成功。任务失败可能会不确定是由于系统的还是用户的行为造成的，然而后者要受到系统行为的影响。

评定任务成功的一个不同的方法是κ系数，它采取的是一种基于属性（概念，位置）的话音理解方法，在系统和用户对话的过程中，必须为这种属性赋予容许值。成对的属性和赋值被称作为属性值对（AVP）。一个由所有有用的属性组成的集合，和任务赋予的数值（通常称作属性值矩阵（AVM）），可以完整地描述一个能在系统的帮助下完成的任务。为了确定κ系数，为答案（设想确定）和记录下的解答（对话日志文件）中的属性建立一个模糊矩阵$M(i,j)$，这样，通过这个矩阵就可以计算出答案和解答之间的一致性$P(A)$和达成一致的可能性$P(E)$，见表4。可以计算单个对话的$M(i,j)$，也可以计算属于一个特定系统或者系统配置的一组对话的$M(i,j)$。

κ系数依赖于简单任务编码方案的可用性，也就是AVM。然而，一些任务不容易被特性化。在这种情况下，就需要更加精心设计的方法来评定任务成功，而这些方法通常会取决于所研究任务的类型。

表 4 –与任务有关的交互参数

缩写.	名称	定义	交互级别	测量方法
TS	任务成功	根据到对话结束为止用户是否达到了他/她的目的来标记任务成功，假设在系统的帮助下能够达到这一目的。用标记来表示是否达到了目的以及问题假定的来源： • *TS:S*：成功（存在解决办法的任务） • *TS:SCs*：由于系统限制放宽而成功 • *TS:Scu*：由于用户限制放宽而成功 • *TS:ScsCu*：由于系统和用户限制均放宽而成功 • *TS:SN*：成功发现不存在解决办法 • *TS:Fs*：由于系统的行为而造成的失败，归因于系统的充分性 • *TS:Fu*：由于用户的行为而造成的失败，归因于用户行为的不合作 也见参考资料 [8][7][24]。	对话	专家
κ	*Kappa*系数	按照kappa统计法，任务完成的百分比。以在对话结束时得到的结果AVM相对于设想（答案）AVM的正确性为基础来评定。为结果和答案中的属性建立一个模糊矩阵*M(i,j)*，*T*是*M*中总的计数，t_i是*M*第*i*列中计数之和，则 $$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$ *P(A)* 是实际对话的AVM和答案一致次数的比例，$P(A) = \sum_{i=1}^{n} \frac{M(i,i)}{T}$。*P(E)* 可从预计的它们会偶然一致次数的比例里估算出来，$P(E) = \sum_{i=1}^{n} (\frac{t_i}{T})^2$。 [31][4]	对话或 对话组	专家

7.5 与话音输入有关的参数

语音对话系统的话音输入能力取决于它识别单词和话语、并从已识别的单词串中获取含义（通常称为“话音理解”）的能力。对于自动话音识别，必须辨别两种方法：当语音是孤立的（单个词识别）或者连续的（关键字识别）时，单词识别器能从用户话语中获取单个单词；另一方面，连续话音识别器能够识别整个句子或话语。进行话音理解经常要以属性值对作为基础，见7.4节。下面段落中描述的参数既可表示话音识别又可表示话音理解。

连续话音识别通常提供一个单词串假定作为输出。为了判断这个单词串是否正确表示了刚才说的话，必须要由抄录专家提供一份参考副本。对于每段话语，首先在单词级采用动态编程(DP)匹配算法来校准假设的单词串和参考的单词串[19] [20]。在校准的基础上，计算正确确定单词的数量c_w、替代词的数量s_w、插入词的数量i_w和删除词的数量d_w。这些数与参考的单词总数n_w相关，可得到两种可选择的度量识别性能的方法：单词错误率*WER*和单词准确率*WA*，见表5。

在句子级，可以用句准确率*SA*或者句错误率*SER*作为补充的性能度量方法，见表5。一般来说，由于句子中单个误识的单词会影响*SA*参数，*SA*要低于*WA*。然而，*SA*可能会高于单词准确率，特别是当许多单个词的句子已经被正确识别时。当发生单个单词误识时，*SER*和*SA*会恶化整段话语的性能，Strik et al. 已经指出了这一情况[26] [27]；参数*NES*和*WES*可以避开这一问题，见表5。当话语没有分解成句子时，所有与句子有关的度量方法也可以在话语级进行计算，而不是在句子级。

孤立单词识别器为每个输入的单词或话语提供一个输出假定。输入的单词和输出的单词可以直接进行比较，与连续话音识别的情况一样，可以规定类似度量性能的方法，只是省去插入词的数量。用一个时间周期内错误告警的数量来代替插入词的数量，见van Leeuwen和Steeneken [28]。只有当识别器工作在关键字识别模式时，也可用*WA*和*WER*来判定关键字。

为了评定话音理解能力，必须辨别两种通用的方法。第一种方法是基于将系统对用户的回答分为正确的回答、部分正确的回答、不正确的回答或不回答几类，各个回答类别可以结合成已在美国国防高级研究项目署（DARPA）项目中采用的度量方法，见表5。第二种方法是将系统的解析能力进行分类，既可以用正确解析的话语，也可以用正确识别的AVP。以已识别的AVP为基础，计算全局度量参数，比如概念准确率*CA*、概念错误率*CER*，或者理解准确率*UA*。所有参数如表5所示。

表 5 – 与话音输入有关的交互参数

缩写	名称	定义	交互级别	测量方法
WER, WA	单词错误率，单词准确率	已被正确识别单词的百分比，以正确拼写假设的和（抄录的）参考的话语为基础，在“sclite”算法的帮助下进行校准，见参考资料[18]。对话中所有用户话语的单词总数记作n_w，替代词、删除词和插入词的数量分别记作s_w、d_w和i_w，则单词错误率和单词准确率可由下面的公式计算得到： $WER = \frac{s_w + i_w + d_w}{n_w}$ $WA = 1 - \frac{s_w + i_w + d_w}{n_w} = 1 - WER$ 见参考资料[24]；在孤立单词识别的情况下，关于如何计算这些参数的详细资料可查阅参考资料[28]。	单词	仪器/专家
SER, SA	句子错误率，句子准确率	已被正确识别整个句子的百分比，以n_s表示句子的总数，s_s、i_s和d_s分别表示替代词、插入词和删除词的数量，则： $SER = \frac{s_s + i_s + d_s}{n_s}$ $SA = 1 - \frac{s_s + i_s + d_s}{n_s} = 1 - SER$ [24]	话语	仪器/专家
NES	每句错误数	一个句子中识别错误的平均数。在句子k中替代词、插入词和删除词的数量分别记作$s_w(k)$、$i_w(k)$和$d_w(k)$，则 $NES(k) = s_w(k) + i_w(k) + d_w(k)$ 按下式计算*NES*的平均值： $NES = \frac{\sum_{k=1}^{\#\ user\ turns} NES(k)}{\#\ user\ turns} = \frac{WER \cdot \#\ user\ words}{\#\ user\ turns}$ [26]	话语	仪器/专家

表 5 – 与话音输入有关的交互参数

缩写	名称	定义	交互级别	测量方法
WES	每句单词错误数	与*NES*有关，但要用句子*k*中的单词数*w(k)*进行归一化： $WES(k) = \frac{NES(k)}{w(k)}$ 按下式计算*WES*的平均值： $WES = \frac{\sum_{k=1}^{\#\ user\ turns} WES(k)}{\#\ user\ turns}$ [26]	单词	仪器/专家
AN:CO, *AN:IN*, *AN:PA*, *AN:FA*, *%AN:CO*, *%AN:IN*, *%AN:PA*, *%AN:FA*	系统回答正确/错误/部分正确/没有回答的数量或百分比	每次对话中，对于用户提出问题，各种情况的系统回答的总数或百分比，包括： • 正确的(*AN:CO*)； • 不正确的(*AN:IC*)； • 部分正确的(*AN:PA*)； • 根本没有回答(*AN:FA*)。 见参考资料 [21][12][14]。	话语	专家
$DARPA_s$, $DARPA_{me}$	DARPA分数, DARPA改进的错误	按DARPA话音理解而确定的度量方法，由Skowronek [25] [17]首先提出并改进，用来表示部分正确的回答： $DARPA_s = \frac{AN:CO - AN:IC}{\#\ user\ questions}$ $DARPA_{me} = \frac{AN:FA + 2 \cdot (AN:IC + AN:PA)}{\#\ user\ questions}$ [21][12][25]	话语	专家
PA:CO, *PA:PA*, *PA:IC*, *%PA:CO*, *%PA:PA*, *%PA:IC*	正确/部分正确/不正确地解析用户话语的数量	用来评估系统从一段话语中获取概念的数量（属性值对，AVP）： • *PA:CO*：用户话语中的所有概念均被系统正确理解。 • *PA:PA*：不是所有的但至少有一个用户话语中的概念已经被系统正确地理解。 • *PA:IC*：用户话语中的概念均没有被系统正确地理解。 可以表示对话中被正确地/部分正确地/不正确地解析的用户话语总数或百分比。[7]	话语	专家

表 5 – 与话音输入有关的交互参数

缩写	名称	定义	交互级别	测量方法
CA, CER	概念准确率,概念错误率	每次对话中正确理解语义单元的百分比。将概念定义为属性值对（AVP），AVP的总数记为n_{AVP}，替代的、插入的和删除的AVP数量记为s_{AVP}、i_{AVP}和d_{AVP}。可由下式得到概念准确率和概念错误率： $CA = 1 - \frac{s_{AVP} + i_{AVP} + d_{AVP}}{n_{AVP}}$ $CER = \frac{s_{AVP} + i_{AVP} + d_{AVP}}{n_{AVP}}$ [9][24][3][2]	话语	专家
UA	理解准确率	在用户话语中已经被正确获取所有语义单元（AVP）的百分比： $UA = \frac{PA:CO}{\#\ user\ turns}$ [32]	话语	专家

7.6 更多的参数

表中所列的大部分交互参数描述了系统的行为，这一点是显然的，因为系统的行为正是系统和服务的质量所关注的。除此之外，还定义了一些与用户有关的参数，虽然对测试的用户群而言，这些参数是特定的，然而它们也与用户所感觉到的质量特性密切相关。

当采用5.3/P.851中描述的方法，把基于SDS服务的质量分解成质量问题的多个方面时，可以看到，对于质量问题的有些方面，交互参数并没有表示出来。没有参数直接与可用性、用户满意度、可接受度或话音输出质量有关。迄今为止，只有极少的方法可以用参数的形式来表示输出话音（被串联或合成）的质量。例如，在IEC 60268-16 [15]中规定了与话音可懂度有关的仪器度量方法，但是它们并不是为电话环境设计的。已经提出了可从串联合成系统的输入文本和话音数据库中计算串联成本的度量方法[5]。虽然这些度量方法在听觉试验中有时表现出与MOS分数很高的相关性，但它们对于话音合成器和它的串联容量是非常特定的方法。

8 交互参数数值的说明

虽然本增补中定义的交互参数对于系统的设计、优化和维护是重要的，但它们与用户感觉到的质量并没有直接的关联。因此，在质量问题的不同方面，用户判定集应作为交互参数集的补充，参见ITU-T P.851建议书。只有这样，才能获得语音对话系统服务质量的有效信息。

说明交互参数的数值应以实验结论作为基础，然而，这些实验结论对于考察的系统或服务而言往往是特定的。举一个例子，对话中超时提示次数的增加可能说明用户在特定的时刻不知道说什么，或者他/她对系统的行为感到困惑[29]。插话尝试次数的增加可以简单地反映出用户认识到有可能中断系统。相反地，插话尝试次数减少可能等同于说明用户不知道对系统说什么。用户话语过长可能源于用户相当大的主动性。与元通信有关参数值（尤其是用户自发的元通信）的减少预计会提高系统的稳健性、对话平稳度和通信效率[1]。

参考资料

[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 AAAI 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 Dienstequalität 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.

ITU-T 系列建议书

A系列	ITU-T工作的组织
D系列	一般资费原则
E系列	综合网络运行、电话业务、业务运行和人为因素
F系列	非话电信业务
G系列	传输系统和媒质、数字系统和网络
H系列	视听和多媒体系统
I系列	综合业务数字网
J系列	有线网和电视、声音节目和其他多媒体信号的传输
K系列	干扰的防护
L系列	线缆的构成、安装和保护及外部设备的其他组件
M系列	电信管理，包括TMN和网络维护
N系列	维护：国际声音节目和电视传输电路
O系列	测量设备技术规程
P系列	**电话传输质量、电话装置、本地线路网络**
Q系列	交换和信令
R系列	电报传输
S系列	电报业务终端设备
T系列	远程信息处理业务的终端设备
U系列	电报交换
V系列	电话网上的数据通信
X系列	数据网和开放系统通信及安全
Y系列	全球信息基础设施、互联网的协议问题和下一代网络
Z系列	用于电信系统的语言和一般软件问题

瑞士印刷
2005年，日内瓦