国　际　电　信　联　盟

ITU-T P.800.2

国际电信联盟
电信标准化部门

(07/2016)

P系列：终端和主观与客观评估方法
话音和视频质量的客观和主观评估方法

平均意见得分说明和报告

ITU-T P.800.2建议书

国际电信联盟

ITU-T P系列建议书

终端和主观与客观评估方法

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
语音终端特性	系列	P.30
		P.300
参考系统	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
话音质量的客观和主观评估方法	系列	P.80
话音和视频质量的客观和主观评估方法	**系列**	**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
涉及车辆的通信	系列	P.1100
流媒体质量评估的模型和工具	系列	P.1200
远程会议评估	系列	P.1300
质量测量的统计分析、评估和报告导则	系列	P.1400
非话音和视频业务质量的客观和主观评估方法	系列	P.1500

欲了解更详细信息，请查阅ITU-T建议书目录。

ITU-T P.800.2建议书

平均意见得分说明和报告

摘要

ITU-T P.800.2建议书介绍了部分较常见类型的平均意见得分（MOS），并描述了能够使这些值得到正确理解所需的最起码的MOS值伴随信息。

历史沿革

版本	建议书	批准日期	研究组	唯一ID*
1.0	ITU-T P.800.2	2013-05-14	12	11.1002/1000/11934
2.0	ITU-T P.800.2	2016-07-29	12	11.1002/1000/12973

关键词

绝对等级评定、ACR、平均意见得分、MOS、客观模型、报告、主观试验。

---

* 为了访问该建议书，在你的网络浏览器的地址部分输入URLhttp://handle.itu.int/，后面接着是建议书的唯一ID，例如， http://handle.itu.int/11.1002/1000/11830-en。

前言

国际电信联盟（ITU）是从事电信、信息和通信技术（ICT）领域工作的联合国专门机构。国际电信联盟电信标准化部门（ITU-T）是国际电联的常设机构，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 ITU-T 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）和国际电工委员会（IEC）合作制定的。

注

本建议书为简要起见而使用的“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，因此大力提倡他们通过下列网址查询电信标准化局（TSB）的专利数据库：http://www.itu.int/ITU-T/ipr/.

目录

ITU-T P.800.2建议书

平均意见得分说明和报告

1 范围

本建议书介绍部分较常见类型的平均意见得分（MOS），并描述了能够使这些值得到正确理解所需的最起码的MOS值伴随信息。

应该注意的是，本案文不旨在提供对主观或客观测试的权威指南。本建议书结尾的参考书目提供更多详细的资料信息。

2 参考

下列ITU-T建议书和其他参考文献的条款，在本建议书中的引用而构成本建议书的条款。在出版时，所指出的版本是有效的。所有的建议书和其他参考文献均会得到修订，建议本建议书的使用者应查证是否有可能使用下列建议书或其他参考文献的最新版本。当前有效的ITU-T建议书清单定期出版。本建议书引用的文件自成一体时不具备建议书的地位。

[ITU-T P.800.1] ITU-T P.800.1建议书（2006年），*平均意见得分（MOS）术语*。

3 定义

3.1 他处定义的术语

无。

3.2 本建议书定义的术语

本建议书定义了如下术语：

3.2.1 条件（condition）：在主观试验中被评估的一组用例中的一个；在视频试验中通常指假设参考电路（HRC）。

3.2.2 子条件（sub-condition）：由用例的特殊特性定义的条件的子集，例如来自特定演讲者的话音材料。

3.2.3 受试者（subject）：主观试验中的参与者。

3.2.4 投票（vote）：受试者对单个测试样本或互动的评分量表中问题的回答。

4 缩略语和首字母缩写词

本建议书采用下列缩略语和首字母缩写词：

ACR 绝对等级评定

DCR 降级等级评定

DMOS 降级平均意见得分

HRC 假设参考电路

MOS 平均意见得分

MUSHRA 含有隐藏参考和基准的多刺激测试

QCIF 四分之一通用中间格式

SSCQE 单一刺激持续质量评估

VGA 视频图形阵列

5 惯例

无。

6 介绍性信息

音频和视频质量是固有的主观量。这意味着音频和视频质量的基线是用户意见。然而，每个人关于“好”的意见都是不同的 – 这种意见无谓对错。

在部署一种新的音频或视频传输技术之前，通过一次或多次主观试验评估传输质量是较好的方法。主观试验的目的是收集多人（“受试者”）关于系统性能的意见，用于许多已定义明确的用例（“条件”）[1]。给定条件下的平均意见得分（MOS）只是为该用例收集的意见（“投票”）的平均值。

客观质量测量算法旨在预测给定输入信号将在主观试验中产生的MOS值。因此，当说明推导出的客观MOS值时，有必要理解所预测试验的基本设计。

MOS值有不同类型，且有许多不同的测试方法产生这些不同的值。本建议书的目的是让读者了解在说明MOS值时应考虑的要点以及报告MOS值时所需的MOS值最起码的伴随信息。

7 主观MOS值

MOS的类型

存在一种普遍的关于MOS值的误解，即其只适用于语音业务，但是，用于话音业务的要求受试者提供质量评估的过程可以一样轻松地应用于视频和一般音频业务。还可以要求受试者评价业务的整体视听质量。除了语音应用，国际电联制定了不同的标准以描述视频和一般音频应用的主观测试的不同方面，这些标准示于参考书目。

主观试验可大致分为两种类型：被动和互动。在被动主观试验中，向受试者呈现表示感兴趣的条件的预先记录的测试样品。受试者被要求被动地收听和/或观看测试资料并使用提供的评分量表给出其意见。在互动试验中，两个或多个受试者使用旨在模拟感兴趣的用例的设备积极地参与对话。受试者通常被给予任务以模拟对话并互动。大多数试验本质上是被动的，然而，用户体验的某些方面，例如延迟和回音的影响，只有在对话场景中才会变得明显。

---

[1] 在视频试验中，条件通常指假设参考电路（HRC）。

测试方法与评分量表

在主观试验中，受试者被要求使用“评分量表”给出自己的意见。量表的目的是将受试者的主观质量评估转化成为一个可以在受试者和其他试验因素之间平均的数值。

通常使用的评分量表有很多种，不同量表的相对优势不在本建议书的范围之内。最常用的量表是5点绝对等级评定（ACR）量表：

很好 5

好 4

一般 3

差 2

很差 1

ACR量表是离散量表，意味着受试者的响应限于上示五个值中的一个。然而，用于组合不同受试者的结果的平均过程意味着MOS值不局限于整数值。一些评分量表有五个以上的离散标签，而另一些则允许受试者在标签之间的点上提供中间反应。

ACR的“绝对”部分涉及这样一个事实，即要求受试者对每个样本进行独立的评分。一些评分量表，诸如降级等级评定（DCR），询问受试者对通过感兴趣的条件处理的样本与同一样本的未处理版本之间的差异的意见。在这样的试验中产生的MOS值通常被称为降级MOS或DOMS。

在大多数试验设计中，受试者被要求对较短的音频或视频样本进行评分。这种样本的持续时间为6至10秒，因为这为受试者提供了足够的时间来形成意见，且不会在样本结束时引入任何偏差。这种持续时间的单独样本很难呈现全部情况，因此受试者通常被要求对同一用例衍生出来的多个样本进行评分。例如，在语音试验中，受测的每种网络条件都可以用来自三位男性和三位女性谈话者的话音样本来表示。这意味着，可以通过对受试者和谈话者的平均，或对特定谈话者或谈话者的性别等子条件的平均，为全部情况生成MOS值。

单一刺激连续质量评估（SSCQE）等测试方法使用更长时间的测试样本，并要求受试者在测试样本播放时不断更新其关于质量的意见。这将导致来自每个受试者关于质量的评分是按时间排序的，而不是一个单一的意见值。

一些测试方法要求受试者回答多个问题。这不仅能产生更多关于受测条件的信息，也是测试设计必要的组成部分。例如，ITU-T P.835建议书测试方法要求受试者在给出整体质量评分前，提供关于样本的话音质量和噪声质量的不同意见。与单一问题ACR测试方法相比，该方法在噪声抑制系统中会得到更稳定的结果。

值得注意的是，一些问题可能不是直接与质量相关的，可能是关于通信的其他方面。例如，[b-ITU-T P.800]为语音试验定义了收听效果量表。同样地，一些会话试验询问受试者在谈话时的体验，而不是在收听时的体验。

8 说明MOS值

下面的讨论最初侧重于语音MOS值；但是，在子小节中所做的许多要点同样适用于视频、音频和音频—视频MOS值。视频的主要区别在以下小节中描述。

一个特定的语音编解码器有一个确切的MOS值是另一种常见的误解。产生这种误解的一个根源是客观质量评估模型（产生了非常可重复的结果）的广泛使用。这种模型设计用于预测或估计主观试验的结果；然而，对于给定比特率的给定编解码器，不同主观试验获得的MOS值差别很大。产生这种现象有一系列原因。

首先，从主观试验特定条件下获得的MOS值可能受到一系列因素的影响，包括但不限于：

– 给予受试者的指导和意见量表的表达措辞；
– 用于呈现材料的设备（手机、耳机、扬声器）；
– 单耳、双耳或立体声呈现；
– 呈现级别；
– 声学环境；
– 受试者的准备；
– 受试者概况，例如年龄和技术暴露；
– 不同文化背景说明和使用评分量表的差异；
– 话音材料（语音内容和谈话者特征）；
– 语言（特定声音和过渡的存在/缺失、普及和重要性）。

其次，从主观试验特定条件下获得的MOS值取决于试验中其他条件的质量。例如，如果大多数其他条件的质量比ITU-T G.729规定的差，则在ACR试验中，ITU-T G.729规定的语音编解码器条件的得分可能超过3.9；反之，如果大多数其他条件的质量比ITU-T G.729规定的好，则ITU-T G.729规定的语音编解码器条件的得分可能显著低于3.9。

此外，如果在不同音频带宽下使用编解码器进行试验，则较高带宽条件的存在将减少为较低音频带宽条件产生的MOS。在语音试验中出现的最高音频带宽通常被称作试验的“背景”。例如，在窄带（300-3700 Hz）ACR试验中，ITU-T G.711规定的语音编解码器条件通常会得到4.0以上的分数；然而，由于存在更高质量的宽带样本，在宽带（50-7000 Hz）ACR试验中更有可能产生3.5-3.7范围内的分数。

最后两点反映了这样一个事实：试验中的受试者倾向于根据试验内容调整自己对于评分量表的使用。确实，精心设计的试验在开始时会包括一个实践阶段，该阶段受试者听到一系列条件的示例，包括最好和最坏的情况。

上述考虑最重要的一个结果是，直接比较单独试验产生的MOS值是没有意义的，除非这些试验是明确设计来进行比较的，即使这样，数据也应该进行统计分析，以确保这种比较是有效的。

9 视频考虑因素

上述许多与语音主观试验相关的考虑也适用于视频试验。试验条件，通常被称作假设参考电路（HRC），通常定义视频编解码器、比特率、帧速率和传输条件的各种组合。从特定条件获得的影响确切的MOS值的因素包括但不限于：

– 用于呈现材料的设备（显示技术、刷新率、对比度等）；

– 观看环境（颜色、温度和对比度）；

– 观看距离（通常表示为观看距离与显示高度的比）；

– 视频内容。

最后一点对于视频试验而言尤为重要。在视频试验中，测试材料的选择比在语音试验中更为重要。这是因为视频序列的内容可能对其编码效率具有非常显着的影响。例如，一个快速移动的运动序列中的信息内容要比一个都是坐着的人的视频会议序列中的信息内容多得多。

对于视频试验，主要背景由视频图像的分辨率决定。通常而言，主观试验不会混合不同的分辨率，因此，视频MOS值属于特定的分辨率，例如，480p或1080p。在分辨率混合的情况下，试验的背景将由具有最大行数的分辨率定义。在这种情况下，重要的是要注意较小的分辨率是原生显示还是调整为试验中的最大分辨率。

10 MOS统计分析

主观MOS值的统计分析不在本建议书的范围之内。但是，MOS值应附有足够的信息，以便进行基本的统计分析。例如，计算每种条件的置信区间。对于任何给定的条件或子条件，该信息包括投票数量、投票的平均数和投票的标准偏差。

11 客观MOS值

客观质量模型的目的是预测主观试验中音频或视频信号可能获得的MOS值。正如上文所讨论，任何给定试验中产生的用于特定编解码器或传输链的确切的MOS值取决于试验设计和执行的多个不同方面。客观模型设计者因此需要预测理想化的试验。这种试验通常是一个根据特定测试方法（通常是ACR）进行的试验，包括在应用领域中遇到的畸变的平衡样本。

例如，[b-ITU-T P.862.1]中定义的映射采用ITU-T P.862建议书规定的客观模型的原始输出，并将其映射到一个范围，该范围由根据[b-ITU T P.800]中所述的ACR方法进行的大量主观试验的输出进行平均确定。ITU-T P.863建议书规定的模型的输出阶段建立了类似的映射。

客观模型的一个优势是结果是可重复的，因此可以直接比较在不同时间和地点进行的测量。然而，仍然应该谨慎，如对测试材料的选择和任何试验前或后处理，仍然可能在结果中引入偏差。

由于现在应该已广泛知晓的原因，不同的客观模型可以在相同的条件产生不同的预测MOS值。例如，ITU-T P.862.1建议书和ITU-T P.863建议书规定的模型不能为ITU-T G.729建议书规定的编码话音生成完全相同的预测MOS值，即使该编解码器在这两个模型的范围

内。出现此现象的部分原因在于：两种模型已经使用不同的主观试验进行了训练和优化。出于此原因，当比较客观MOS预测与阈值时，例如，要监测服务级别协议或发出警报，应该在产生预测模型的背景下选择这样的阈值。

12 报告主观MOS值

表1描述了在报告主观MOS值时必须要提供的信息，以及建议提供的额外信息。

如果试验根据国际电联建议书的规定实施，关于方法的信息通常可以通过简单参考相关标准和所使用的特定方法来表示，但应注意标准程序的变化。

始终提供有关被动试验测试样本的信息非常重要。对于视频样本，提供更详细的信息可能是有用的，例如，特定序列是否包含平移或场景变化。

表1 – 报告主观MOS值所需的最低量信息

信息	试验类型	规定
方法 被动或互动 基于样本或连续评估 样本的绝对或相对评估 给予受试者的指导和问题 评分量表标签 评分量表是离散的还是连续的 样本持续时间 或者 使用的国际电联建议书和方法	全部	强制
测试计划 试验目的 进行测试的日期和场地 处理信息 试验设计，例如，模块化设计 会话数量和持续时间 受试者数量 受试者概况，年龄和性别分布 受邀的受试者类型，例如新人或专家 关于所用设备的信息 呈现环境，即：本底噪声、对比度等	全部	建议
条件/HRC信息 条件数量 条件列表 每种条件的平均投票（MOS） 每种条件的投票标准差 每种条件的投票数量	全部	强制

表1 – 报告主观MOS值所需的最低量信息

信息	试验类型	规定
子条件信息 子条件因素列表 子条件的MOS值 每种子条件的投票数量和方差	全部	可选
音频呈现 音频带宽 音频信道，单声道、立体声等 音频呈现级别 音频呈现方法，例如，扬声器、耳机（单耳、双耳等）	语音、音频、AV	强制
视频呈现　视频图像分辨率（注1和2） 观看距离作为高度的函数，例如3H 设备类型和尺寸，即电视、平板电脑、电话等 应用，即视频电话、视频点播、线性电视等	视频、AV	强制
语言	被动语音、AV	强制
谈话者数量和性别	被动语音、AV	强制
视频材料的类型，例如，运动、头部和躯干	被动视频、AV	强制
音频类型，例如，古典音乐、流行音乐、电影配乐	被动音频、AV	强制
注1 – 必须注意隔行扫描图像的使用。 注2 – 如果试验包含多个图像分辨率，则必须提供有关较小图像分辨率是原生还是放大的信息。		

13　报告客观MOS值

当报告ITU-T客观模型产生的MOS值时，报告所使用的模型和任何非默认设置通常就足够了。对于非标准化模型，必须提供表1中“方法”行中的信息来描述预测的试验设计。还建议提供有关用于测试和/或训练客观模型的试验中使用的测试材料类型的信息。

14　表示法

[ITU-T P.800.1]提供了通用和高级表示法，可用于帮助识别MOS值的来源。虽然P.800.1中的表示法有助于提供生成MOS值的背景概述，但它不能替代根据本建议书提供的详细背景描述，在可能的情况下应始终提供详细背景描述。

参考书目

国际电联已经为不同的应用标准化了许多主观测试方法。下面列出了一些最广泛使用的方法。

[b-ITU-T G.729] ITU-T G.729建议书（2012年），采用共轭结构代数码本激励线性预测（CS-ACELP）的8 kbit/s话音编码。

ITU-T P.800系列包括与话音质量的主观和客观评价相关的许多建议书；特别要注意的是：

[b-ITU-T P.800] ITU-T P.800建议书（1996年），传输质量主观测定方法。

[b-ITU-T P.805] ITU-T P.805建议书（2007年），会话质量的主观评估。

[b-ITU-T P.835] ITU-T P.835建议书（2003年），评估包含噪声抑制算法的话音通信系统的主观测试方法。

[b-ITU-T P.862.1] ITU-T P.862.1建议书（2001年），P.862原始结果得分到MOS-LQO变换的映射函数。

[b-ITU-T P.863] ITU-T P.863建议书（2014年），感知客观收听质量评估。

ITU-T P.900系列包括多媒体评估建议书：

[b-ITU-T P.910] ITU-T P.910建议书（2008年），多媒体应用的主观视频质量评估方法。

[b-ITU-T P.911] ITU-T P.911建议书（1998年），多媒体应用的主观视听质量评估方法。

[b-ITU-T P.912] ITU-T P.912建议书（2016年），识别任务的主观视频质量评估方法。

ITU-R也发布了与音频和视频质量的主观评估相关的建议书：

[b-ITU-R BS.1116-1] ITU-R BS.1116-1建议书（1997年），音频系统（包括多声道声音系统）中轻微损伤的主观评估方法。

[b-ITU-R BS.1534-1] ITU-R BS.1534-1建议书（2003年），编码系统中间质量水平的主观评估方法。

[b-ITU-R BT.500-13] ITU-R BT.500-13建议书（2012年），电视画面质量的主观评估方法。

[b-ITU-R BT.710-4] ITU-R BT.710-4建议书（1998年），高清电视图像质量的主观评估方法。

注 – ITU-R定义的标准化方法并不是所有都能测量平均意见得分。出版的ITU-R建议书为所有相关方法提供了完整的文件和参考。为了获得对于上述测试方法的更好的描述和澄清，读者可以参考ITU-R提供的各个已出版的建议书。

下述手册提供了对主观测试方法和最佳实践的深入处理。

[b-ITU-T handbook] 主观测试实用程序（2011年）。

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、声音节目及其他多媒体信号的传输
K 系列	干扰的防护
L 系列	环境与 ICT、气候变化、电子废物、节能；外部设备的线缆和其他组件的建设、安装和保护
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	测量设备的技术规范
P 系列	**终端和主观与客观评估方法**
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、开放系统通信和安全性
Y 系列	全球信息基础设施、互联网的协议问题、下一代网络、物联网和智慧城市
Z 系列	用于电信系统的语言和一般软件问题

瑞士印刷
2019年，日内瓦