ICS 33.160.99
CCS M74

UHD World Association
世界超高清视频产业联盟

世 界 超 高 清 视 频 产 业 联 盟 标 准

T/UWA XXX-2025

三维声质量主观评价通用方法

General Methods for the Subjective Assessment of 3D Audio Sound Quality

（ 征求意见稿 ）

2025-xx-xx 发布　　2025-xx-xx 实施

世界超高清视频产业联盟

目　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》给出的规则起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本文件由世界超高清视频产业联盟提出并归口。

本文件主要起草单位：

本文件主要起草人：

三维声质量主观评价通用方法

1 范围

本文件规定了广播电视和网络视听三维声音频系统和设备的主观评价通用方法。

本文件适用于广播电视和网络视听中三维声音频系统建设及设备的设计、生产、验收、运行和维护。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 35784 视听设备 音频系统小损伤的主观评价方法

ITU-R BS.1534-3 Method for the subjective assessment of intermediate quality level of audio systems

ITU-R BS.2126 Methods for the subjective assessment of sound systems with accompanying picture

Report ITU-R BS.2300 Methods for Assessor Screening

T/CSMPTE 5-2018 基于 5K 超高清图像和环绕声/三维声的家庭影院配置规范

3 术语和定义

下列术语以及定义适用于本文件。

3.1

主观评价 subjective assessment

评价员在规范的主观评价环境下，根据对节目声音的主观感受来评价其质量优劣的一种方法。

3.2

属性 attribute

根据给定的口头或书面定义，主观评价活动中可感知的特征。

3.3

序列 excerpt

适于评价给定被测系统声音属性质量的一段音乐、语音或其他声音信号。

3.4

被测对象 object

被测系统，通常以经过该系统处理后的一些测试序列来代表。

3.5

参考 reference

序列，用作声音质量对比的基准。

3.6

隐藏参考　hidden reference
未向评价员明示的参考。

3.7

刺激　stimulus
在主观评价中，评价员听到的一段声音，该段声音可以是被测对象、参考或隐藏参考。

3.8

带参考的主观评价方法　subjective assessment methods with reference
一轮评价的刺激中包含参考。

3.9

评价员　subject
在听音测试中评价刺激的测试人员。

3.10

评价小组　listening pannel
在一个听音测试中，给出听音测试数据的评价员的整体。

3.11

盲测　blind test
一种测试方法，在该种测试中，刺激是向评价员提供的唯一信息源。

3.12

双盲测试　double blind test
盲测的一种，在该种盲测中，听音测试的组织者和听音测试之间没有不受控制的交互可能。

3.13

声像　sound image
声音经过录音和重放后，声源被感知到的空间特性，包括定位、纵深、宽窄。

4 缩略语

下列缩略语适用于本文件。
IQR　四分位距（Inter-Quartile Range）
HRTF 头部相关传输函数（Head-Related Transfer Function）
oct　倍频程 （octave）

DTF 方向性传递函数（Directional Transfer Function）

5 概述

主观评价是评估三维声系统或设备性能的不可替代的方法。一般而言，主观评价方法分为两类：第一类是评估系统在最佳条件下的性能，通常称为“质量评价”；第二类是评估系统在与传输或发射相关的非最佳条件下保持质量的能力，通常称为“损伤评价”。

三维声设备的应用涵盖从信号采集到接收渲染等多个领域，不存在适用于有效评估所有设备的统一的评价方法，第 6 章中列出通用主观评价要素，第 7 章列出通用主观评价方法，附录 A 至附录 G 列出了根据评价需求和目标选择合适的评价要素和评价方法，并对评价要素进行限定和组合形成的贴切各类评价需求的主观评价条件，附录 H 给出了半专家评价员的训练和筛选过程。

6 通用主观评价要素

6.1 评价环境

分为实验室评价环境和消费端评价环境。

实验室评价环境旨在提供对系统和设备进行评价的严格条件，包括主观评价重放设备的技术指标要求和评价环境的声学指标要求。

消费端评价环境旨在为三维声链的消费端侧提供质量评价的手段，重放设备和声学环境可局部调整为更加贴近用户实际使用场景，但仍需遵循对评价结果可靠性和可重复性的基本要求。

由于主观评价中提供给评价员的信息量影响评价结果，过载的信息量可能降低评价员判断的准确性，因此，除非声音和图像间关系十分重要，声音主观评价首选无伴随图像的方式。与视频相关的属性评价中，需考虑屏幕摆放位置及听距、视距、视角、屏幕尺寸之间的关系。

6.2 评价员和评价小组

6.2.1 评价小组构成

实验室评价环境下应采用专家评价员。专家评价员是指具有音频工程、录音工程、心理声学或相关领域的专业知识，能够敏锐地感知和描述声音的细微差异，具备声音主观评价经验和能力的人员。

消费端评价环境下可采用专家评价员，也可采用半专家评价员和普通评价员。半专家评价员是指由非专业人士（对音色、空间音频具有基础感知分辨能力，但没有丰富的听音训练、评测经验）逐步训练和筛选而来，训练和筛选过程见附录H。普通评价员是指普通听众，更适用于听觉体验和偏好评价。普通评价员的评价结果个体变异大，需要更多数据样本以抵消随机误差，获得稳定的统计结果。

6.2.2 评价员的测后筛除

为保证评价结果的可靠性，宜对评价员进行测后筛除。

测后筛除是指根据评价数据，依靠统计学的方法对评价员的评价的能力进行检验，并进一步筛除未通过检验的评价员的全部数据的过程。一些主观评价方法自身即可构成进行测后筛除的数据条件，如带隐藏参考的评价方法。一些评价方法需要通过其他辅助手段来构建进行测试筛除的数据条件，本条给出了通过重复评价构建测后筛除数据条件的示例。

通过重复评价，可检验评价员在主观评价中应具备的辨别能力和可重复能力：辨别能力指评价员在盲测中正确区分被测对象的能力；可重复能力是指在两次或多次重复评价中，评价员对相同被测对象的同一个序列，给出相同或相近评分的能力。

测后筛除需要至少1次的重复评价，可通过预测试评价进行，也可贯穿正式评价。预测试评价是指正式评价前的先行测试，用于评价员的能力筛除，通过预测试考核的评价员方可进入正式评价。预测试应包含所有被测对象，包含待评价序列的有代表性的子集，覆盖待评价质量的全范围。

测后筛除的具体方法见ITU-R BS.2300。

6.3 评价序列

评价序列通常对评价结果产生直接影响，需谨慎选取。评价序列选取中的考虑因素及选取原则见表1。

表1 评价序列选取考虑因素及选取原则

考虑因素	选取原则
暴露被测对象缺陷及潜在问题	苛刻性，能够对被测对象施压，但不应包括针对特定被测对象精心设计的人工信号
不同被测对象缺陷的多样性	匹配目标使用场景，覆盖多样化的节目类型
避免分散评价员专注力	避免使用过于吸引注意力的序列（如包含特别吸引人或令人厌烦的内容），以及可能引发评价员偏见的序列（如包含过于熟悉或情感色彩强烈的内容）
评价结果的广泛认知性	使用公认高质量或标准化的测试序列

6.4 评价属性

6.4.1 选取与分类

属性反映被测对象在某方面的可感知特征，通常以描述词的形式出现。

应用中，如果需要对单独属性进行评价，组织者需要在规划主观评价时选取待评价的属性，并在训练过程中使评价员充分理解待评价属性并与组织者达成共识。

评价属性选取中的考虑因素及选取原则如下：

- ✓ 区分性强，对不同的刺激产生良好的区分能力。
- ✓ 稳定性好，评价员能够给出可重复的评分。
- ✓ 共识性好，属性定义明确，易形成良好共识。
- ✓ 独立性强，与其他待评价属性的冗余度和相关性低。
- ✓ 可指定标度，可以文本或参考声音样本的方式指定属性的评分标度。

与三维声相关的典型属性包括但不限于：音色、均匀性、定位和声像、空间感、与视觉相关属性。

6.4.2 音色

指与频谱特征相关的听觉感受，包括明亮度、清晰度、丰满度、声染色等。

6.4.3 均匀性

指声音在时间或频谱上的一致性或平滑性，包括稳定度、均衡度、动态感等。

6.4.4 定位和声像

指与声音方向、距离和空间分布相关的听觉感受，包括水平方向定位、垂直方向定位、声源距离感、声像宽度、声像深度、声像高度、声像稳定性等。

6.4.5 空间感

指与声音整体空间效果相关的听觉感受，包括包围感、真实感、沉浸感、声场扩散度等。

6.4.6 与视觉相关属性

指那些通过视觉信息对声音感知产生影响的属性，包括声画同步、声音和图像空间感和谐性等。

6.5 锚点

锚点是指在声音主观评价中引入的对照信号或对照样本，通常是一个已知质量的音频序列。

锚点的作用包括：

- ✓ 建立评分对照点，校准评分尺度：为评价员提供一个明确的对照点，帮助评价员在评价过程中更好地理解声音质量的范围，校准评分尺度，避免评分过于集中或分散。
- ✓ 减少主观偏差，提高评价结果的一致性：评价员通过对比锚点和被测对象的质量差异调整自己的评分，从而减少评价员因个人偏好或环境差异导致的主观偏差，提高不同评价员或不同测试环境下的评价结果的可比性。
- ✓ 验证主观评价条件：通过评价员对锚点的评分验重放设备、评价环境、评价方法中是否存在不可控因素。
- ✓ 用于评价员能力的筛除，通过评价员对已知质量锚点的评分判断评价员的评价能力。

组织者应根据评价目的、被测对象的质量范围、评价环境等确定是否需要引入锚点，锚点的引入应遵循以下原则：

- ✓ 质量已知，与待测属性有相关性；
- ✓ 能代表测试序列的典型特征，与其他被测信号之间的感知差异来源于质量而非内容的不同；
- ✓ 引入锚点前后，对多个被测系统的质量高低排序无影响；
- ✓ 锚点的制作过程和结果可重复。

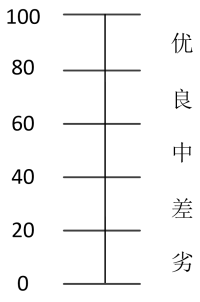
注：按照如上原则确定的锚点不仅仅限制于常用的音质损伤评价中的3.5kHz/7kHz低通滤波序列，还可以按照应用和属性确定适宜的锚点。按应用举例，在双耳渲效果染评价中，由多声道直接下混生成的立体声序列、参考序列与常见HRTF库卷积生成的序列可作为低、中等质量的锚点；按属性举例，在声像定位属性的评价中，假设在一定角度范围内水平方向声像定位分辨率为φ°，可以2～3倍系数与φ相乘形成的角度偏差、1.5×φ形成的角度偏差作为低、中等质量的锚点。

6.6 评分标度

6.6.1 5级质量标度和百分制连续质量标度

5级质量标度适用于快速、标准化的评估，百分制连续质量标度适用于高精度和区分度的质量评估。5级质量标度和百分制连续质量标度见表2。

表2　5级质量标度、百分制连续质量标度

5级质量标度	百分制连续质量标度
5 优 4 良 3 中 2 差 1 劣	

6.6.2　5 级损伤标度

5级损伤标度适用于对损伤进行评价，5级损伤标度见表3。

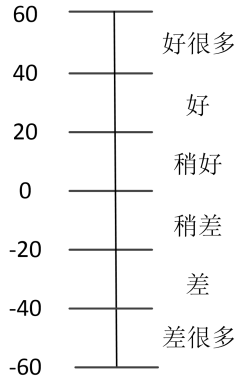
表3 5级损伤标度

损伤程度	评分等级
损伤不可觉察	5.0
损伤可察觉，但不至引起不悦	4.0
损伤稍令人不悦	3.0
损伤令人不悦	2.0
损伤令人非常不悦	1.0

6.6.3　双向离散 7 级标度和±60 分制对比标度

双向离散7级标度和±60分制对比标度适用于多个被测的直接对比评估。双向离散7级标度和±60分制对比标度见表4。

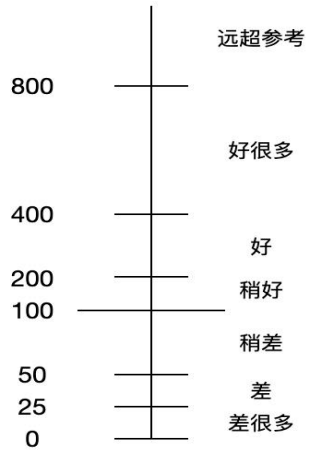
表4 双向离散7级标度、±60分制对比标度

双向离散7级标度	±60分制对比标度
3 好很多 2 好 1 稍好 0 质量相同 -1 稍差 -2 差 -3 差很多	

6.6.4　无界质量标度

针对幅度估计方法的无界质量标度示例见表5。

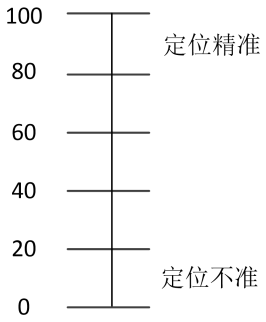
表5 幅度估计方法无界标度

无界标度（示例）


6.6.5 针对具体属性评分的 5 级质量标度和百分制连续质量标度

针对具体属性评分的5级质量标度和百分制连续质量标度示例见表6。

表6 属性评价-5级质量标度、百分制连续质量标度示例

5级质量标度（示例，声像定位）	百分制连续质量标度（示例，声像定位）
5 优 4 良 3 中 2 差 1 劣	

6.7 评价过程

评价过程包括训练过程、预测试过程和正式评价过程。

训练过程的目的是：

- ✓ 使评价员逐步熟悉评价环境和评价方法；
- ✓ 使评价员了解被测对象的质量范围或损伤程度；
- ✓ 使评价员的评分尺度逐步趋于稳定。

预测试包括为筛选评价员进行的预测试，为确定评价方法进行的预测试等，根据需求选用。

正式评价中应避免不可控因素对评价结果的影响，比如在评价中，如果音频序列的顺序或被测对象的顺序对所有评价员都相同，则无法确定评价员所给出的判断是出于播放顺序还是出于不同的质量或损伤等级。因此，必须以揭示独立因素且只包括这些因素影响效果的方式安排测试条件。

7 评价方法

7.1 带隐藏参考的双盲三刺激方法（GB/T 35784 方法）

适用于参考可获得且需要对小损伤进行准确探察的应用场景，典型设备包括编解码器、音频处理器等。

带隐藏参考的双盲三刺激是指对每个测试序列，评价员听到“A”，“B”，“C”三个刺激，其中“A”为已知参考，“B”和“C”为隐藏参考和被测对象的随机分配。

7.2 带隐藏参考和锚点的多刺激方法（ITU-R BS.1534 方法）

适用于参考可获得且需要对多个中等音频质量被测对象进行对比的应用场景，典型设备包括编解码器、音频处理器等。

带隐藏参考和锚点的多刺激方法是指对每个测试序列，评价员听到多个刺激，这些刺激包含一个已知参考，顺序随机的隐藏参考（已知参考的拷贝）、锚点和多个被测对象。常用的低质量锚点为参考序列的3.5kHz低通滤波版本，中等质量锚点为7kHz低通滤波版本。

7.3 成对比较法

适用于参考不可获得且需要对两个被测对象进行直接比较的应用场景，典型设备包括传声器、扬声器、渲染器等。

对每个测试序列，评价员听到一对刺激，这一对刺激是两个被测对象的随机分配。

7.4 多刺激法

适用于参考不可获得且需要对两个以上被测对象进行直接比较的应用场景，典型设备包括传声器、扬声器、渲染器等。

对每个测试序列，评价员听到多个刺激，这多个刺激是多个被测对象的随机分配。

7.5 ABX 方法

适用于参考可获得或不可获得，需要快速判断两个被测对象间是否存在差异，但无需仔细斟酌差异程度的应用场景，如编码器研发中的音质优化判断。

ABX是一种双盲测试方法，该方法中，

-“A”，第一个被比较对象，也可以是未经处理的参考序列。

-“B”，第二个被比较对象，通常是经过处理的序列。

-“X”：待测试样本，随机选自A或B。评价员的任务是判断x更接近A还是B。

7.6 幅度估计方法

一种心理物理学方法，用于对参考可获得或不可获得的情况下，量化评价员对多个被测对象总体质量或某一属性的主观感知强度。该方法中的参考不一定是未经处理的无损序列，可以是某一个被测对象，也可以是按6.5确定的锚点。评价员听到参考刺激、多个被测对象刺激和一个隐藏参考（已知参考的拷贝），根据感知到的参考刺激，给出评分基准值（如为100），再与参考刺激对比，对被测对象刺激的感知强度进行相对评分（如，若感知强度是参考刺激强度的两倍，则标为200；若为一半，则标为50）。

7.7 单刺激法

适用于参考不可获得的情况下，对成品节目的声音质量评价的场景。

对每个节目，评价员直接对刺激给出评分。

附 录 A
（规范性）
小损伤音频系统主观评价方法

A.1 适用范围

该方法适用于在参考可获得的情况下评价高质量的音频系统或设备的损伤。

A.2 评价方法

应采用带隐藏参考的双盲三刺激方法。

A.3 评价条件

应采用实验室评价环境，实验室重放设备应符合GB/T 35784第9章的要求，声学环境应符合GB/T 35784第10章的要求。

应采用专家评价员，评价小组人数应不少于20人。使用带隐藏参考的双盲三刺激方法形成的评价数据即可进行评价员的测后筛除，测后筛除方法见GB/T 35784附录B。

测试序列的选取原则和相关要求见6.3。音频序列的典型时长为10秒至25秒。测试序列的数量最小为5，合理估算值为被测对象数目的1.5倍。

除整体损伤外，如需对单独属性进行损伤评价，属性的选取原则和常见属性见6.4。

不宜使用锚点。

评分标度采用5级损伤标度，见表3。

评价过程包括训练过程和正式评价过程，见6.7。应将连续评价时间控制在30分钟内，以避免评价员因疲劳而影响判断。

A.4 统计分析

应给出评分均值及在一定置信水平下的置信区间、被测对象与隐藏参考的分差均值及置信区间。

附 录 B
（规范性）
带隐藏参考和锚点的多刺激方法

B.1 适用范围

该方法适用于在参考可获得的情况下评价中等质量的音频系统或设备。

B.2 评价方法

应采用带隐藏参考和锚点的的多刺激方法。

B.3 评价条件

应采用实验室评价环境，实验室重放设备应符合GB/T 35784第9章的要求，声学环境应符合GB/T 35784第10章的要求。

应采用专家评价员，评价小组人数应不少于20人。使用带隐藏参考和锚点的多刺激方法形成的评价数据即可进行评价员的测后筛除，测后筛除方法见ITU-R BS.1534-3 4.1.2。

测试序列的选取原则和相关要求见6.3。音频序列的时长不宜超过12秒。

除整体损伤外，如需对单独属性进行损伤评价，属性的选取原则和常见属性见6.4。

锚点的确定原则见6.5。

评分标度采用百分制连续质量标度，见表2。

评价过程包括训练过程和正式评价过程，见6.7。应将连续评价时间控制在30分钟内，以避免评价员因疲劳而影响判断。

B.4 统计分析

应给出每个被测对象的评分均值及在一定置信水平下的置信区间、中位数及IQR区间，以数值表格和图示表示。

对多个被测对象，宜根据ITU-R BS.1534-3 9.3规定的方法进行方差分析及方差分析后的多重比较。

附 录 C
（规范性）
无参考成对设备的对比评价方法

C.1 适用范围

该方法适用于参考不可获得的情况下对成对设备的声音质量进行对比评价。

C.2 评价方法

应采用成对比较法。

C.3 评价条件

专业设备评价应采用实验室评价环境，实验室重放设备应符合GB/T 35784第9章的要求，声学环境应符合GB/T 35784第10章的要求。

消费端设备评价可采用实验室评价环境，也可采用6.1描述的消费端评价环境，当采用家庭影院环境时，家庭影院重放设备应符合T/CSMPTE 5第7章的要求，评价环境应符合T/CSMPTE 5表1的要求。

专业设备评价应采用专家评价员，评价小组人数应不少于20人。消费端设备评价可采用专家评价员，也可采用6.2.1描述的普通评价员，普通评价员人数宜不少于30人。应考虑6.2.2描述的测后筛除方法或其他方法以保障测试结果的可靠性。

测试序列的选取原则和相关要求见6.3。

除整体质量外，如需对单独属性进行评价，属性的选取原则和常见属性见6.4。

锚点的确定原则见6.5。

评分标度采用双向离散7级标度或±60分制对比标度，见表4。

评价过程包括训练过程和正式评价过程，见6.7。

C.4 统计分析

应给出一个被测对象相对于另一个被测对象的评分均值及在一定置信水平下的置信区间、中位数及IQR区间，以数值表格和图示表示。

应进行分布验证并根据验证结果采用 t 检验（正态分布）或非参数检验方法（如 Wilcoxon 符号秩检验，非正态分布）进行假设检验。

附 录 D
（规范性）
无参考多设备的对比评价方法

D.1 适用范围

该方法适用于参考不可获得的情况下对多个被测对象的声音质量进行对比评价。

D.2 评价方法

应采用多刺激法。

D.3 评价条件

专业设备评价应采用实验室评价环境，实验室重放设备应符合GB/T 35784第9章的要求，声学环境应符合GB/T 35784第10章的要求。

消费端设备评价可采用实验室评价环境，也可采用6.1描述的消费端评价环境，当采用家庭影院环境时，家庭影院重放设备应符合T/CSMPTE 5第7章的要求，评价环境应符合T/CSMPTE 5表1的要求。

专业设备评价应采用专家评价员，评价小组人数应不少于20人。消费端设备评价可采用专家评价员，也可采用6.2.1描述的普通评价员，普通评价员人数宜不少于30人。应考虑6.2.2描述的测后筛除方法或其他方法以保障测试结果的可靠性。

测试序列的选取原则和相关要求见6.3。

除整体质量外，如需对单独属性进行评价，属性的选取原则和常见属性见6.4。

锚点的确定原则见6.5。

评分标度采用百分制连续质量标度或5级质量标度，见表2。

评价过程包括训练过程和正式评价过程，见6.7。

D.4 统计分析

应给出每个被测对象的评分均值及在一定置信水平下的置信区间、中位数及IQR区间，以数值和图示表示。

对多个被测对象，宜根据ITU-R BS.1534-3 9.3规定的方法进行方差分析及方差分析后的多重比较。

附 录 E
（规范性）
无参考成对设备的快速对比评价方法

E.1 适用范围

该方法适用于参考可获得或不可获得的情况下对两个被测对象声音质量的快速对比评价。

E.2 评价方法

应采用ABX方法。

E.3 评价条件

专业设备评价应采用实验室评价环境，实验室重放设备应符合GB/T 35784第9章的要求，声学环境应符合GB/T 35784第10章的要求。

消费端设备评价可采用实验室评价环境，也可采用6.1描述的消费端评价环境，当采用家庭影院环境时，家庭影院重放设备应符合T/CSMPTE 5第7章的要求，评价环境应符合T/CSMPTE 5表1的要求。

专业设备评价应采用专家评价员，评价小组人数应不少于20人。消费端设备评价可采用专家评价员，也可采用6.2.1描述的普通评价员，普通评价员人数宜不少于30人。锚点的确定原则见6.5，宜通过已知质量的锚点来检验评价员的评价能力，并按照对已知质量锚点的判断正确率不低于85%的原则进行评价员的测后筛除。

测试序列的选取原则和相关要求见6.3。

除整体质量外，如需对单独属性进行评价，属性的选取原则和常见属性见6.4。

评分标度采用5级质量标度，见表2。

评价过程包括训练过程和正式评价过程，见6.7。

E.4 统计分析

按公式（E.1）计算“ X”的判断正确率：

$$\text{正确率} = \frac{\text{正确判断的次数}}{\text{总测试次数}} \times 100\% \qquad \text{（E.1）}$$

应根据二项检验（Binomial Test）（评价员数<30）或卡方检验（Chi-Square Test）（评价员数>30）的分析结果来判断被试者的正确率是否显著高于随机概率。

附 录 F
（规范性）
无参考多设备的快速对比评价方法

F.1 适用范围

该方法适用于参考可获得或不可获得的情况下对多个被测对象的快速对比评价。

F.2 评价方法

应采用幅度估计方法。

F.3 评价条件

专业设备评价应采用实验室评价环境，实验室重放设备应符合GB/T 35784第9章的要求，声学环境应符合GB/T 35784第10章的要求。

消费端设备评价可采用实验室评价环境，也可采用6.1描述的消费端评价环境，当采用家庭影院环境时，家庭影院重放设备应符合T/CSMPTE 5第7章的要求，评价环境应符合T/CSMPTE 5表1的要求。

专业设备评价应采用专家评价员，评价小组人数应不少于20人。消费端设备评价可采用半专家评价员，半专家评价员人数宜不少于25人。

锚点的确定原则见6.5，宜通过已知质量的锚点来检验评价员的评价能力，并按照对已知质量锚点的判断正确率不低于85%的原则进行评价员的测后筛除。

测试序列的选取原则和相关要求见6.3。

除整体质量外，如需对单独属性进行评价，属性的选取原则和常见属性见6.4。

评分标度采用无界标度，见表5。

评价过程包括训练过程和正式评价过程，见6.7。

F.4 统计分析

数据处理的方法和步骤如下：

1) 内部标准归一化和对数变换

将评价员对某条测试序列的被测对象评分除以参考刺激，并取以 2 为底的对数。

例如，评价员 A 对测试序列 trail1 的评价中，对参考刺激的评分 X 为 50 分，对被测对象刺激的评分 D 为 75 分。以对参考刺激的归一化评分记为 0，计算被测对象评分的内部标准归一化评分数值 G：

$$G = \log_2 \frac{D(A,trail1)}{X(A,trail1)} = \log_2 \frac{75}{50} = 0.58 \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (F.1)$$

2) 评分尺度归一化

由于不同评价员对不同测试序列的感知程度不同，评分尺度差异较大，需要对评分尺度进行归一化。尺度归一化计算方法如下：

$$G'(i,j) = \frac{G(i,j)}{max_j|G(i,j)|} \cdots\cdots\cdots\cdots\cdots\cdots (F.2)$$

其中，i 是评价员编号， j 是测试序列编号。

3) 分值标准化

将所有评分标准化到 0~100 之间，便于展示和比较：

$$G''(i,j) = G'(i,j) * 50 + 50 \cdots\cdots\cdots\cdots\cdots (F.3)$$

4) 方差分析

应给出每个被测对象的评分均值及在一定置信水平下的置信区间、中位数及IQR区间，以数值和图示表示。

对多个被测对象，宜根据ITU-R BS.1534-3 9.3规定的方法进行方差分析及方差分析后的多重比较。

附 录 G
（规范性）
成品节目评价方法

G.1 适用范围

该方法适用于对成品节目节目的声音质量进行评价。

G.2 评价方法

应采用单刺激法。

G.3 评价条件

广播和电视节目评价应采用实验室评价环境，实验室重放设备应符合GB/T 35784第9章的要求，声学环境应符合GB/T 35784第10章的要求。在进行电视节目的声音质量评价时，视距、听距、屏幕尺寸的关系应符合ITU-R BS.2126。应采用专家评价员，评价小组人数为7~15人。

网络视听节目在消费端的效果评价，可采用实验室评价环境，也可采用6.1描述的消费端评价环境，模拟消费端的实际应用场景。可采用专家评价员，也可采用半专家评价员或普通评价员。普通评价员的人数不宜少于30人。

评分标度采用百分制连续质量标度，见表2。

除整体质量外，如需对单独属性进行评价，属性的选取原则和常见属性见6.4。

评价过程包括训练过程和正式评价过程，见6.7。

G.4 统计分析

应给出每个节目的评分均值及在一定置信水平下的置信区间、中位数及IQR区间，以数值和图示表示。

附 录 H
（资料性）
半专家评价员的训练和筛选过程

H.1 评测方法

评估参与者对音频的基本属性的感知能力。评估方式采用AB对照方式，固定参考项，通过逐步加大测试项与参考项之间的属性差异的方式评测参与者的感知阈值。AB对照组中，一例为隐藏参考，另一例为加入了感知偏差的测试项，两者顺序随机。

H.2 评测维度

H.2.1 音频均衡感知能力

评估参与者对音频低频（250Hz以下）和高频（4000Hz以上）频率的敏感程度。评测过程中，将测试项顺序随机排布，使参与者选择主观听感低频/高频缺失更严重的选项。将参与者无法正确分辨的最大低频/高频衰减值作为该参与者对低频/高频成分的分辨阈值。测试参考项以及由对参考项进行处理形成各类刺激的具体内容见表G.1。

表H.1 音频均衡感知能力评测-各类刺激

低频测试音源：低频丰富音乐片段，时长 15 秒，48kHz/16bit		
刺激	处理方式	备注
参考	响度归一化	
隐藏参考	响度归一化	
低频衰减 3/5/7/9/11dB	Low Shelving Filter: 截止频率 250Hz，Slope 72dB/oct，增益-3/-5/-7/-9/-11dB 响度归一化	共 5 项，对应 5 种程度的低频损失
高频测试音源：高频丰富音乐片段，时长 17 秒，48kHz/16bit		
测试项	处理方式	备注
参考	响度归一化	
隐藏参考	响度归一化	
高频衰减 3/5/7/9/11dB	High Shelfing Filter: 截止频率 4000Hz，Slope 72dB/oct，增益-3/-5/-7/-9/-11dB 响度归一化	共 5 项，对应 5 种程度的高频损失

H.2.2 空间定位感知能力

评估参与者对正前方和正侧方声像的双耳定位能力。每组测试使参与者对比特定角度偏移的音频与参考音频，判断两者是否来自同一方向，取参与者无法辨别为不同方向的最大角度作为其最小可辨别角度。考虑到人耳在不同水平方向上感知敏感度存在差异，分别在前方与侧方采用不同颗粒度的角度步进进行测试。测试用各类刺激的具体内容见表G.2。

表H.2　空间定位感知能力评测-各类刺激

测试音源：单声道白噪声，时长 5 秒，48kHz/16bit		
刺激	处理方式	备注
前方参考	0 度 DTF 处理 响度归一化	使用 48kHz 的 HRTF 数据集，并进行漫射场均衡得到 DTF
前方隐藏参考	前方参考的拷贝	与前方参考对比
2/4/6/8/10/12/14 度声源	2/4/6/8/10/12/14 度 DTF 处理 响度归一化	分别与前方参考对比
侧方参考	90 度 DTF 处理 响度归一化	
侧方隐藏参考	侧方参考的拷贝	与侧方参考对比
85/80/75/70/65/60 度声源	85/80/75/70/65/60 度 响度归一化处理	与侧方参考对比

H.2.3　动态范围感知能力

评估参与者对音频动态范围的敏感程度。测试以未经任何处理具有较大动态范围的音频作为参考，对参考施加不同程度动态压缩效果作为测试项，评估参与者能感知到的最小动态压缩程度。参与者将经过不同程度的压缩处理后的音频与参考音频进行对比，将参与者无法正确分辨的最大压缩比例作为其动态范围的感知阈值。测试参考项以及由对参考项进行处理形成各类刺激的具体内容见表G.3。

表H.3　动态范围感知能力评测-各类刺激

测试音源：大动态电影片段，时长 15 秒，48kHz/16bit		
测试项	处理方式	备注
参考	响度归一化	例如音源中包含雨声背景，瞬态撞击声等，动态范围大
隐藏参考	响度归一化	
5 种不同程度动态压缩	动态范围压缩器： Threshold：20dB Ratio：1.5/2/2.5/3/4 Attack：1ms Release：5ms 响度归一化	

H.2.4　混响感知能力

评估参与者对混响的感知能力。测试以未经任何处理的干声音乐信号作为参考，对参考项施加不同干湿比的房间混响作为测试项，评估参与者能感知到的最小混响程度。评价员对不同干湿比混合后的信号与参考信号进行对比，分辨混响感更重的信号，评价员无法正确分辨的最大干湿比比例作为其混响感知的评测阈值。测试参考项以及由对参考项进行处理形成各类刺激的具体内容见表G.4。

表H.4　混响感知能力评测-各类刺激

测试音源：消声室采集的音乐片段，时长 15 秒，48kHz/16bit		
测试项	处理方式	备注
参考	响度归一化	
隐藏参考	响度归一化	
6 种不同干湿比混合信号	Fabfilter R 混响器： 房间类型：Medium Chamber Mix ： 5%/10%/15%/20%/25%/30% 响度归一化	与参考信号对比

H.3 评测筛选

综合考虑报名人数、目标评价小组人数以及评测结果，对参与者进行筛选。由于筛选出的具有基础听音能力的人员，通过后续持续的评测与培训有可能使听音能力进一步提升从而接近或达到专家评价员标准，因此筛选过程不设定严格的感知阈值，而是通过得分排序的方式进行筛选。

筛选步骤包括：

步骤一：单项打分。对每一个评测维度进行打分。采用5分制，将离散排序的不同感知等级缩放至0～5区间内。对评测者i在评测维度j的第k个听感差异等级评测结果计为X_{ijk}，假设该评测维度共设有K个听感差异等级，于是得到K个评估结果。如G.2.1所述，听感差异等级是测试项与参考项逐次加大的，即$X_{ijk}(k=0,1,\dots K-1)$按听感差异等级从小到大排序。X_{ijk}的取值定义为：

0	隐藏参考感知更优
1	隐藏参考与测试项没有明显感知差异
2	隐藏参考感知更差

评测者 i 对评测维度 j 的得分G_{ij}为：

$$G_{ij}=\frac{K-(\arg min_k X_{ijk}==0)}{K}\cdot 5 \quad \text{.....(G.1)}$$

其中，$\arg min_k X_{ijk}==0$为所有取值为0的X_{ijk}最小k序号。特别地，当前测试维度j中无任意一次打分X_{ijk}为0时，则判定该维度j得分G_{ij}为0。

步骤二：汇总及排序。对所有评测维度进行平均，汇总得到该测试者的总分。对所有测试者进行排序。评测者 i 的各个维度总分为：

$$G_i=\frac{\sum_j G_{ij}}{J} \quad \text{.....(G.2)}$$

步骤三：筛选。根据总参与人数，对目标人数进行按序筛选。