国 际 电 信 联 盟

ITU-R

国际电联无线电通信部门

ITU-R BS.1116-3建议书

(02/2015)

音频系统中细小损伤的主观评价方法

BS 系列

广播业务(声音)

国际电信联盟

前言

无线电通信部门的职责是确保卫星业务等所有无线电通信业务合理、平等、有效、经济地使用无线电频谱，不受频率范围限制地开展研究并在此基础上通过建议书。

无线电通信部门的规则和政策职能由世界或区域无线电通信大会以及无线电通信全会在研究组的支持下履行。

知识产权政策（IPR）

ITU-R的IPR政策述于ITU-R第1号决议的附件1中所参引的《ITU-T/ITU-R/ISO/IEC的通用专利政策》。专利持有人用于提交专利声明和许可声明的表格可从http://www.itu.int/ITU-R/go/patents/en获得，在此处也可获取《ITU-T/ITU-R/ISO/IEC的通用专利政策实施指南》和ITU-R专利信息数据库。

ITU-R 系列建议书

（也可在线查询 http://www.itu.int/publ/R-REC/en）

系列	标题
BO	卫星传送
BR	用于制作、存档和播出的录制；电视电影
BS	**广播业务（声音）**
BT	广播业务（电视）
F	固定业务
M	移动、无线电定位、业余和相关卫星业务
P	无线电波传播
RA	射电天文
RS	遥感系统
S	卫星固定业务
SA	空间应用和气象
SF	卫星固定业务和固定业务系统间的频率共用和协调
SM	频谱管理
SNG	卫星新闻采集
TF	时间信号和频率标准发射
V	词汇和相关问题

说明：该ITU-R建议书的英文版本根据ITU-R第1号决议详述的程序予以批准。

电子出版
2015年，日内瓦

ITU-R BS.1116-3建议书*,**

音频系统中细小损伤的主观评价方法

（1994-1997-2014-2015年）

范围

本建议书旨在用于以下系统的评价，即系统引入了细小的损伤，但若不严格控制实验条件和进行适当的统计分析，将无法探测到这些细小的损伤。如果用于以下系统，即系统引入了相对较大且易于探测的损伤，那么它将带来时间和精力的过度支出，并可能带来比简单测试更不可靠的结果。本建议书是其他建议书的基本参照，它可包含额外的特殊条件，或者放宽本建议书中所含的要求。

关键字

音频质量；细小损伤；主观评价；测听；音频编码；高质量音频；测听室

国际电联无线电通信大会，

考虑到

a) ITU-R BT.500和ITU-R BT.1284建议书已经建立了若干用于主观评价音频和视频系统质量的方法；

b) 主观测听允许对在声音源头与听者之间传输过程中因期望之信号的任何损伤而引起的听者不快程度进行评价；

c) 经典的客观评价方法可能不适于评价先进的音频编码方案，正在开发用于测试音响系统音质的、感知的客观评价方法；

d) 对测试数据的交换、兼容性和正确评估而言，使用标准化的方法至关重要；

e) 引入新的、利用心理声学属性的高级数字音频系统，尤其对细小损伤，需要在主观评价方法方面有进一步的进展；

f) 引入如ITU-R BS.775建议书所述的、高至3/2声道的多声道立体声音响系统，以及ITU-R BS.2051建议书所述的高级音响系统，不论带还是不带伴随图片，都需要新的主观评价方法，包括实验条件，

建议

1 附件1中给出的测试、评估和报告程序用于音频系统中细小损伤的主观评价，包括多声道音响系统（带或不带图片），

---

* 本建议书应引起国际标准化组织/动态图像专家组（ISO/MPEG）– 音频特别小组的注意。

** 无线电通信第6研究组于2015年7月和2023年3月根据ITU-R第1号决议对该建议书进行了编辑性修正。

进一步建议

1 需要对测听室的特性和高级音响系统的再现设备做进一步研究，当完成这些研究后，应对本建议书进行更新。

附件1

1 概述

1.1 目录

附件1分为11个部分，给出了测试各个方面的详细要求：

1 概述

2 实验设计

3 测听组的选择

4 测试方法

5 属性

6 编排材料

7 再现设备

8 测听条件

9 统计分析

10 统计分析结果的表示

11 测试报告的内容

此外还包括含有关于测听专家的选择指南以及提供给测试者的指令样例的附加材料。

使用了一些常见的、带有技术含义的词汇。这些词汇的汇总表见附件之附件材料4。

2 实验设计

在科学研究领域，为收集可靠信息，可使用许多不同种类的研究策略。在音频系统细小损伤的主观评价中，当使用最正式的实验方法。主观实验的特点将首先通过实际控制和操纵实验条件来体现，其次则通过来自人类观察员的定量数据来体现。

需要仔细做好实验设计和规划，以确保不受控因素不会污染测听测试，从而不会带来模糊。作为一个例子，如果音频条目的实际序列对测听测试中的所有测听者都是相同的，那么无法确定测听者做出的判断是由于该序列的缘故，还是由于所呈现的不同程度的损伤。因此，必须按以下方式来安排测试条件，即应能揭示独立因素的影响，而且只有这些因素。

在可以预期的情况下，潜在的损伤和其他特性将均匀地分布在整个测听测试中，对测试条件的表示可实现真正的随机化。

当预计无法实现均匀分布时，在测试条件的表示中必须考虑到这一点。例如，当待评估材料随难度不同而不同时，激励的表示次序必须随机分布，不论是在完整测试内还是在完整测试之间。

同样，对测听测试也需要做好设计，这样，才不会出现测听者因负担过重而降低判断准确性的现象。除了以下情况，即声音与视觉之间的关系是重要的，对音频系统的评价最好在没有附带图像的情况下进行。

一个主要的考虑因素是要包含适当的控制条件。典型地，控制条件包括未受损伤音频材料的演示，它们按以下方式引入，即对测听者而言，它们是不可预测的。是对这些控制激励的判断与潜在的受损者之间的差别使得能够得出以下结论，即等级评分是对损伤所做的真实评价。

本文稍后将对这些考虑因素中的一些因素进行讨论。应该明白，实验设计、实验实施、统计分析等问题是复杂的，在一个如本建议书的建议书中只能给出最通用的指南。建议在规划测听测试之初，就咨询或引入在实验设计和统计分析方面拥有专业技能的专业人士。

3 测听组的选择

3.1 测听专家

重要的是，来自评价音频系统细小损伤的测听测试中的数据，应完全来自以下测听者，即其在检测这些细小损伤方面拥有专业技能。待测系统想达到的评价质量越高，拥有测听专家就显得越重要。

3.2 测听者的选择准则

利用一组选定的测听者来对音响系统细小损伤进行主观测试，其结果主要不是为了外推给普通公众。通常，其目的是为了调查，在一定条件下，一组测听专家能否感知相对微妙的性能退化，并形成对引入之损伤的定量估计。对测试程序的要求是揭示这些问题，它们可能在不同条件下、在整个测试期间得以揭示，当系统到消费者手中后，这些问题可能在实际的应用过程中出现。

有时需要在真正开始测试之前（预筛选）或之后（后筛选）引入拒绝技术。在某些情况下，可能使用两种类型的拒绝。此处，消除指的是一个过程，在此过程中，忽略来自某个特定测听者的所有判断。

任何未经仔细分析和应用的拒绝技术类型都可能导致存在偏差的结果。因此，非常重要的是，任何时候当消除数据时，测试报告都需清楚描述所用的准则，以便读者可以做出自己的判断。

3.2.1 测听者的预筛选

预筛选程序包括以下方法，如测听测试、根据其之前经验以及在之前测试中的性能选择测听者、根据对预测试的统计分析结果消除测听者。培训程序可用作预筛选的一种工具。

引入预筛选技术的主要理由是提高测听测试的效率。不过，必须做好均衡，以抵消太多限制结果相关性可能带来的风险。

3.2.2 测听者的后筛选

后筛选方法可大致分为至少两类：一类是基于相比平均结果的不一致性；另一类是依赖测听者做出正确鉴别的能力。第一类从来都不是合理的。无论何时用此处建议的测试方法来进行主观测听测试，第二类后筛选所需的信息将自动可用。用于此的一种建议的统计方法如附件材料1所示。

方法主要用于消除那些不能适当做出鉴别的测听者。应用后筛选方法可弄清楚测试结果中的倾向性。不过，应注意测听者对不同内容敏感度的可变性，因此需要格外小心。

3.3 测听组的规模

如果可以估计方差，并且知道实验要求的解决方案，那么可以预计出测听小组的准确大小。

在技术和行为上严格控制测听测试条件的情况下，经验表明，对从测试中获得适当结论来说，来自20个测听者的数据通常已经足够。如果随着测试的进行能完成对数据的分析，那么当已获得能从测试中得到正确结论、适当水平的统计意义时，则无需对更多的测听者进行处理。

如果预计一些在测系统将接近透明，那么需要更多数量的测听者，以确保有足够多数量的测听者通过后筛选测试。

如果出于任何原因，无法实现严格的实验控制，那么可能需要更多数量的测听者，已获得所需的解决方案。

测听小组的规模不仅仅考虑所需的解决方案。原则是，来自本建议书有关的试验类型的结果，仅对测试实际涉及的测听专家小组有关。这样，通过增大测听小组的规模，可以认为结果可适用于一个更广泛的测听专家小组，并因此有时可能认为更有说服力。测听小组的规模也需要增大，以便测听者可对不同的内容改变其敏感度。

4 测试方法

为了在系统产生细小损伤的情况下进行主观评价，有必要选择一种适当的方法。“双盲三激励隐藏参考”方法认为尤其敏感、稳定，使得能够准确检测细小损伤。因此，它应用于这种类型的测试。

在这种方法的首选和最敏感的形式中，涉及一次一个测听者的问题，由该测听者自行判断选择三个激励中的一个（“A”、“B”“C”）。已知的参考总是可用，如激励“A”。隐藏参考和客体同时可用，但根据试验情况“随机地”分配给“B”和“C”。

要求测听者根据连续五级损伤评分刻度，比较“A”，对“B”的损伤情况做出评价，以及比较“A”，对“C”的损伤情况做出评价。对激励“B”或“C”中的一个，激励“A”应看不见；另一个可揭示损伤。对参考与其他激励之间的任何感知到的差别，都必须被认为是一种损伤。

一旦首选方法中的测听者完成对试验的等级评分，则可能直接继续到下一个试验。素材可重复使用，直至测听者完成一次评估。这样，测听过程可按自己的步调向前推进。

相对ITU-R BS.1284建议书和表1中给出的、源自ITU-R五等级损伤程度的 “锚点”，等级评分将被视为是连续的。

表 1

损伤程度	等级评分
感知不到失真	5.0
能感知到失真，但不恼人	4.0
能感知到失真，稍微恼人	3.0
能感知到失真，恼人	2.0
能感知到失真，非常恼人	1.0

注1 – 已经表明使用预先定义的中间锚点可能会引入偏差[Poulton，1992]。有可能使用没有对锚点进行描述的数字刻度。在这种情况下，必须指出刻度所代表的倾向。这可帮助克服为比较采用不同语言完成的测试中存在的翻译问题。

如果不使用中间的锚点，那么就平均值和标准差对单个测听者的结果进行归一化处理是必要的。以下公式可用于进行此类归一化处理，同时保留最初的刻度：

$$Z_i = \frac{\left(x_i - x_{si}\right)}{s_{si}} \cdot s_s + x_s$$

式中：

Z_i： 归一化结果；

x_i： 测听者 i 的评分；

x_{si}:: 测听者 i 在 s 组完整测试中的平均分；

x_s: 所有测听者在 s 组完整测试中的平均分；

s_s： 所有测听者在 s 组完整测试中的标准方差；

s_{si}： 测听者 i 在 s 组完整测试中的标准方差。

使用没有中间锚点的等级评分也排除了对结果的绝对解释。

建议等级评分准确到小数点后一位。

测试方法包括两部分：熟悉或训练阶段和等级评分阶段。

4.1 熟悉或训练阶段

在正式进行等级评分之前，必须使测听者变得彻底熟悉测试设备、测试环境、等级评分过程、等级评分刻度以及它们的使用方法。测听者也应变得彻底熟悉要研究的内容。对最敏感的测试，测听者应能接触到所有在之后的正式等级评分完整测试中将要做出等级评分的材料。在熟悉或培训期间，测听者最好应成组在一起（即由三个测听者组成），这样，他们就可以对其检测的内容自由地进行相互交流和讨论。

作为一个模型，在附加材料“给测听者的指令”中给出了一个指令集样例。这些指令包括对有关激励表示的“双盲三激励隐藏参考”技术的描述。适当完成熟悉过程可使一些原来能力较弱的测听者成为测试专家。在熟悉过程结束后，测听者应对评分刻度有一稳定的感觉，将用在熟悉或培训阶段之后的正式等级评分阶段。

4.2 等级评分阶段

在当天第一次正式等级评分完整测试开始之时，应对每一个测听者口头陈述测试指令，最好辅以书面材料。在正式等级评分展示开始之前，可以进行一些说明性的比较。

由于中长期听觉记忆是不可靠的，因此测试程序应完全依赖短期记忆。如果近即时切换（见注1）方法与三激励系统结合使用，如附加材料3所述，那么这样做是最好的。此类切换要求激励间对齐关闭时间。

注1 – 如果连续激励的波形是不同的，那么准确的即时切换可生成内容。例如，对逐渐下降/翻转/逐渐上升，最好是总计约40 ms的近即时切换。

对最关键的评价，一次应只处理一个测听者。只有这样，测听者才有彻底的个人自由来在三激励方法中进行激励切换。此类自由是必要的，这样，测听者可根据其自身判断来在每个试验的激励间进行详细的比较。

最好是，测听者应能在激励间进行切换，而无需视觉上的引导，这样，只要测听者愿意，在最小干扰条件下，为更好地集中精神，测听者可闭着眼睛。不应有任何能发出声响的切换系统内容，原因是，此类内容可严重干扰评价过程。

等级评分完整测试的持续时间不应超过20-30 min，尽管此处提议的试验自学特性将在测听者间引入不受控制的变化。经验表明，每次完整测试安排的试验不应超过10-15个试验，以便获得所需的完整测试长度。测听者疲劳可能成为一个主要的因素，将严重干扰判断的有效性。为了避免这种情况，对每个测听者，在连续的完整测试之间安排的休息时间，其持续时间不应短于完整测试的时间长度。

5 属性

下面列出的是单声道、双声道立体声、多声道立体声（即3/2声道）和高级音响系统评估特定的属性。首选的是，在每种情况下对“基本音频质量”属性进行评估。实验者可以选择定义和评估其他属性。

因测听者试图评估每个试验中的多个属性而可能出现的问题是响应负担之一。如果测听者因试图回答多个有关给定激励事件的问题而出现超负荷或困惑，那么这可能对所有问题产生不可靠的等级评分。

5.1 单声道系统

基本声音质量

– 这一全局属性用于判断任何和所有检测到的、参考与客体之间的差别。

5.2 双声道立体声系统

基本声音质量

– 这一全局属性用于判断任何和所有检测到的、参考与客体之间的差别。

以下附加属性可能是重要的：

立体声声像质量

– 根据声音图像位置以及音频事件深度和真实度的感觉，该属性与参考和对象之间的差别有关。

尽管一些研究已经表明，立体声图像质量可遭到损伤，但还没有开展足够的研究来指明，是否需要对立体声图像质量单独进行有别于基本声音质量的等级评分。

注1 – 到1993年，对双声道立体声系统的绝大多数细小损伤主观评估研究使用的都只是有关基本音频质量的属性。因此，有关立体声图像质量的属性隐性地或显性地包含在基本音频质量内，作为这些研究的一种全局属性。

5.3 多声道立体声系统

基本声音质量

– 这一全局属性用于判断任何和所有检测到的、参考与客体之间的差别。

以下附加属性可能是重要的：

前方声像质量

– 该属性与前方声源的定位有关，包括立体声图像质量和清晰度损失。

环绕印象质量

– 该属性与空间印象、氛围或特殊方向环绕效果有关。

5.4 高级音响系统

基本声音质量

– 这一全局属性用于判断任何和所有检测到的、参考与客体之间的差别。考虑高级音响系统的属性应包括有关多声道系统的属性。

此外，以下属性可能是重要的：

音色质量 – 已发现该属性具有特殊的意义

– 音色质量属性可通过两组特性来描述：

第一组音色特性与声音色彩有关，如亮度、音调色彩、着色、清晰度、硬度、均衡度或丰满度。

第二组音色特性与声音均匀性有关，如稳定性、尖锐度、真实性、逼真度和动态性。这些特性可以是对声音音色的描述，但也可以是对声音其他特性的描述。

定位质量

– 该属性与所有定向声源的定位有关，包括立体声图像质量和清晰度损失。该属性可以分为水平定位质量、垂直定位质量和远程定位质量。在带有附带图像的测试的情况下，这些属性也可以分为显示器上的定位质量和围绕测听者的定位质量。

环境质量 – 这扩展了环绕质量属性

– 该属性与空间印象、环绕感、临场感、扩散性或空间定向环绕效果有关。该属性可以分为水平环境质量、垂直环境质量和远程环境质量。

6 编排材料

为了揭示在测系统之间的差异，只使用关键材料。关键材料指的是对在测系统起重要作用的材料。没有普遍“适用的”编排材料可用来评价在所有情况下的所有系统。因此，对每个实验中待测的每个系统，都必须明确地寻找关键编排材料。寻找好的材料通常是耗时的；不过，除非为每个系统找到真正的关键材料，否则实验将无法揭示系统之间的差异，也将无法确定结论。

在一个“空的”发现结果可被认为是有效的之前，在经验上和统计上都必须指明，未能找到系统之间的差异不是因为因音频材料可选余地少得可怜而造成的实验不敏感性，或者因为任何其他的实验薄弱环节。在极端情况下，如果发现若干系统或所有系统都是完全透明的，那么为了明确检查测听者的专业技能，可能需要以较低或中等的锚点来编排特殊的试验（见附加材料1）。

必须知道这些锚点（例如，基于先前的研究）对测听专家而非测听生手而言是可探知的。这些锚点作为测试条目引入，不仅用于检查测听者的专业技能，而且用于检查实验所有其他方面的敏感性问题。

如果通过应用附件材料1中所述的统计方面的考虑，在标准测试方法中（见本附件§ 3）的所有测听者都能正确地确定这些锚点（要么不可预知地嵌入显然透明的条目中，要么嵌入单独的测试中），那么这可用作以下证据，即测听者的专业技能是可接受的，并在其他的实验状况方面不存在任何敏感性问题。那么在这种情况下，对这些测听者无法区分编码和未编码版本的条目或系统来说，这些测听者发现的明显透明度就是“真正透明度”的证据。

另一方面，如果任何测听者都无法正确鉴别这些锚点，那么这表明这些测听者缺乏足够的专业技能，或者在该情况中存在敏感性缺陷，或者两者兼而有之。在这种情况下，系统的明显透明度不能得到正确解释，将需要用新的测听者替代无法完成此附加测试的测听者再次进行实验，其他任何变化都可能会增加实验的灵敏度。

在寻找关键材料中，任何激励都可被认为是允许的潜在广播材料。特意设计用来突破某个特定系统的合成信号都不应包括在内。编排序列的艺术性内容或知识性内容不应太有吸引力，也不应太令人讨厌或令人乏味，以致于测听者无法专注对损伤的检测。应考虑到在实际广播中每种类型编排材料的预期发生频率。不过，应该理解，广播材料的性质可能会随未来音乐风格和喜好的变化而随时发生变化。未来，客观感知模型有望在关键材料的选择上提供帮助。

在选择编排材料时，重要的是要准确定义待评价的属性。选择材料的责任将委派给一组基本了解预期损伤的专业测听者。其出发点将基于一个非常广泛的材料范围。通过专用记录可对范围进行扩展。

出于准备主观比较测试录音带的目的，在测试媒介上进行记录之前，专业测听者需主观地对每个素材的响度进行调整。这将允许后续在为所有编排条目设定的固定增益上使用测试媒介。

因此，对所有测试序列，专业测听者小组将就单个测试素材的相对噪音电平召开会议并达成共识。此外，对相对校准电平的整个序列，专家应就绝对再现声压电平达成共识。

在校准信号电平上的音调猝发（例如1 kHz，300 ms，–18 dBFS）（FS：全刻度）应包括在每个记录的开头，以便使其输出校准电平能够调整为再现声道要求的输入校准电平（见§ 8.4.1）。对数字记录的测试材料，关于数字系统的最大可能编码电平[EBU，1992年]，校准电平应对应–18 dB。应对声音编排信号做好控制，使得峰值振幅仅在很小的可能性上会超过允许之最大信号的峰值振幅（在ITU-R BS.645建议书中定义，即一个比校准电平高9 dB的

正弦波）。注意，在这些条件下，峰值编排表将指明电平不能超过允许之最大信号的电平。对参考和测试激励的时间校准，音调猝发也可能是有用的。

一个测试中可包括的可行的素材数量是变化的：对每个客体，它将相同。对最小数量5个素材，一个合理的估计是1.5（客体数量）。典型地，音频素材将为10 – 25 s长。出于任务的复杂性，客体应是可用的。只有当确定一个合适的时间表时，才能成功做出一个选择。

对单声道和立体声系统的评价，如果素材选自容易获得的来源，那么是一种有利的情况，这样，可以随时对准备好的测试录音带进行检查，如有必要，依据最初来源进行检查。SQAM光盘是此类来源的一个例子。不过，更重要的是，即使这些素材来自不那么容易获得的来源，也应使用真正关键的素材。

在双声道立体声回放的条件下，将利用参考低音混合对多声道系统的性能进行测试。尽管在某些情况下使用固定的低音混合可以被认为是受到限制的，但对广播公司长期使用来说，这无疑是一种最明智的选择。有关参考低音混合的方程式为（见ITU-R BS.775建议书）：

$$L_0 = 1.00\,L + 0.71\,C + 0.71\,L_s$$

$$R_0 = 1.00\,R + 0.71\,C + 0.71\,R_s$$

对以下条件，即当对一个高级音响系统进行测试时，应在测试报告中对用于从高级音响系统到双声道或多声道系统低音混合的方程式进行描述，或者当重新进行呈现时，对重新呈现过程进行描述。

预选适当的测试素材，用于参考双声道低音混合性能的关键评估，这应基于对双声道低音混合编排材料的再现。

7 再现设备

7.1 概述

参考监控扬声器或耳机应依据以下目标来选择，即所有声音项目信号或其他测试信号都可以最佳的方式进行再现；也就是说，它们应为任何类型的再现提供中立的声音，并应对单声道评价以及两声道或多声道立体声音响系统有用。

在耳机再现的情况下，某些质量缺陷可更清晰地被感知到；不过，在扬声器再现的情况下，其他的质量缺陷可更清晰地被感知到。因此，需要通过主观预测试来确定适当类型的再现设备。

尤其在以下情况下，即缺陷将影响立体声声音图像的特点，应使用扬声器再现。

为评价双声道立体声音响系统，既使用立体声扬声器也使用耳机可能是必要的。为评价单声道音响系统，可使用一个中心扬声器与/或耳机。

为单个试验或成组试验选择扬声器或耳机，将使效果的可听性与使用的传感器关联起来，但测听者的有效数量将被减少。或者，如果测听者能够根据意愿在扬声器和耳机之间切换，那么将不可能把效果的可听性与使用的传感器关联起来。

为评价有或没有附带图像的多声道音响系统和高级音响系统，如果需要评价对同时作用之所有再现声道的影响，那么必须使用扩音器。

在所有情况下，每个扬声器都必须在声学上匹配相关的频率范围，这样，当中才存在最小内在音色差别。

7.2 参考监控扬声器

7.2.1 概述

“参考监控扬声器”指的是高质量的工作室测听设备，由一个扬声器系统的集成单元组成，安装在特别尺寸的机壳中，加上特殊的均衡、高质量的功率放大器和适当的交叉网络。

电声特性应满足以下在自由场条件下测得的最低要求。绝对声音电平值参照的是距声音中心1 m的测量距离，除非另有说明。

7.2.2 电声要求

7.2.2.1 幅频响应

对预选的扬声器，40 Hz-16 kHz频率范围上的频率响应曲线（在主轴上（方向角 = 0°），使用粉红噪声，在三分之一八度音阶频段上测得）最好应落在4 dB的公差频段内。在方向角±10°上测得的频率响应曲线不应有别于主轴频率响应3 dB以上，在方向角±30°（仅在水平面上）不应有别于主轴频率响应4 dB以上。

应匹配不同扬声器的频率响应。在至少250 Hz-2 kHz的频率范围内，差别最好应不超过1.0 dB。

注1 – §8.3.4中提到的操作房间响应曲线描述了测听房间中声场内的频率特性。

7.2.2.2 指向性指数

在500 Hz - 10 kHz频率范围上、以三分之一八度音阶频段噪声测得的方向性指数C应在以下限值内：

$$6\ \text{dB} \le C \le 12\ \text{dB}$$

方向性指数应随频率平滑增加。

7.2.2.3 非线性失真

产生90 dB平均声压电平（SPL）的恒定电压输入信号提供给扬声器。与该SPL相关，在基本频率范围40 Hz - 16 kHz中任何谐波失真部件都不得超过以下值：

$$-30\ \text{dB}\ (3\%) \qquad \text{for } f < 250\ \text{Hz}$$

$$-40\ \text{dB}\ (1\%) \qquad \text{for } f \ge 250\ \text{Hz}$$

7.2.2.4 瞬态保真

示波器上测得的衰减时间应为（测量电平为原始电平的1/e（约0.37），仅在主轴上）：

$$t_s < 5 / f$$

式中，f指的是频率。

这意味着正弦猝发音的衰减时间不得超过相应正弦波的5倍。

7.2.2.5 时间延迟

立体声或多声道系统的声道之间的延时差别不应超过100 μs。

注1 – 这不包括从扬声器到测听位置的延时。

在系统带有附带图像的情况下，在测系统中参考监控耳机总的延时不应超过ITU-R BS.775建议书中设定的限值。

7.2.2.6 动态范围

对至少10 min的时间，扬声器能产生的最大工作声音电平应为（无热或机械损伤以及无过载电路被激活，以项目模拟噪声信号测得（根据国际电工委员会（IEC）出版物268-1c））：

$$L_{eff\ max} > 108 \text{ dB}$$

通过使用设为平坦响应和r.m.s.（慢）的声音电平计测得。

单个参考监控扬声器和相关放大器产生的等效噪声电平应为（参照自声学中心的距离1m（注1））：

$$L_{noise} < 10 \text{ dBA}$$

注1 – 声学中心为出于测量目的的参考点。它通常对应表面的几何中点，它辐射扬声器的最高频率。这应由制造商来指明。

7.3 参考监控耳机

7.3.1 概述

参考监控耳机指的是高质量的房间测听设备，等同扩散场响应。

7.3.2 电声要求

7.3.2.1 频率响应

ITU-R BS.708建议书对工作室监控耳机的扩散场频率响应提出了建议。

7.3.2.2 时间延迟

立体声系统声道之间的延时差异不应超过20 μs。

在系统带有附带图像的情况下，在测系统中参考监控耳机总的延时不应超过ITU-R BS.775建议书中设定的限值。

8 测听条件

8.1 概述

对扬声器再现的声音而言，术语“测听条件”描述了参考声场复杂的声学要求，参考声场会影响参考测听点处测听房间中的测听者。这包括：

– 测听房间的声学特征；

– 测听房间中扬声器的配置方案；

– 参考测听点的位置或区域；

形成该测听点或区域最终的声场特性。

由于工艺水平尚不允许只用声学参数来全面、唯一地描述参考声场，因此提供了一些有关参考测听房间的几何和房间声学要求，以确保所述测听条件的可行性。

8.2 参考测听房间

8.2.1 概述

在扬声器再现的情况下，对主观测试而言，应注意观察以下要求。参考测听房间的最低要求如下所述。

在耳机再现的情况下，测听房间至少应满足有关背景噪音电平的要求。

8.2.2 集合属性

以下值用于描述参考测听房间的、适当的净尺寸。如果测试房间不能满足这些尺寸，那么至少应满足在后续章节中提到的、有关声场条件和扬声器配置方案的要求。

8.2.2.1 房间大小（房屋面积）

– 对单声道或双声道立体声再现：20 – 60 m^2。

– 对多声道立体声或高级音响系统再现：30 – 70 m^2。

注1 – 较小尺寸的空间将限制同一时间可容纳的、最大的测听者数量。

注2 – 还需进一步研究确定有关高级音响系统测听房间的最优特征。房间大小、形状、比例和声学特性都应写入测试报告。

8.2.2.2 房间形状

相对立体基线中间垂线上的垂直面，房间应是对称的。底面积最好应该呈矩形或梯形。

8.2.2.3 房间比例

应注意以下尺寸比例，以确保房间低频本征音的合理均匀分布：

$$1.1\,w/h \le l/h \le 4.5\,w/h - 4$$

式中：

l: 长度；

w: 宽度；

h: 高度。

此外，应用到条件：$l/h < 3$ 和 $w/h < 3$。

8.2.3 房间声学属性

8.2.3.1 混响时间

在频率范围200 Hz - 4 kHz上测量的混响时间平均值T_m应为：

$$T_m = 0{,}25\,(V/V_0)^{1/3} \quad \text{s}$$

式中：

V: 房间体积；

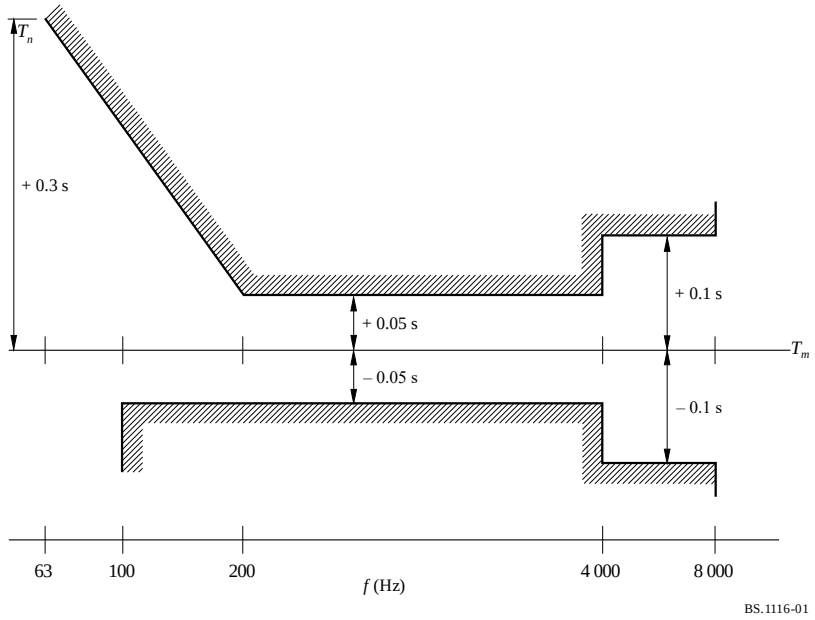
V_0: 参考体积为100 m^3。

图1给出了在频率范围63 Hz（见注1） - 8 kHz上用于T_m的容差。

注1 – 测量低频上小的混响时间值是有难度的。

图1

混响时间的公差范围，相对于平均值 T_m



8.3 参考声场条件

8.3.1 概述

对主观感知或质量评估听觉事件及其在其他地方或房间的再现性而言，测听区域的声场特性最为重要。这些特性源自扬声器与测听房间的互动，所用的测听配置方案将参考之（见§8.5）。

目前可描述以下特性。

8.3.2 直达声

8.3.2.1 监控扬声器的频率响应

在自由场条件下测得的扬声器频率响应应满足§7.2.2中所述的要求。

8.3.3 反射声

8.3.3.1 前期反射

测听房间边界表面引起的前期反射，它将在直接声音后最多15 ms的时间间隔内到达测听区域，相对直接声音，在1 - 8 kHz范围内，它应至少衰减10 dB。

8.3.3.2 后期能量

除了对前期反射和回响规定的要求（见§8.2.3），需要避免在声场出现其他重大的异常，如颤动回声、音调色彩等。

8.3.3.3 回响时间

（见§8.2.3.1）

8.3.3.4 脉冲响应

在所有评估者测听位置测得的、来自每一个扬声器的脉冲响应，测听房间按即将用于测试的方式进行设置（包括室内陈设），在测试报告中，应显示在时间域中。这有助于验证扬声器与房间音响效果的结合程度，以满足前期反射、后期能量以及回响的要求。

8.3.4 稳态声场

8.3.4.1 操作间响应曲线

操作房间响应曲线被定义为参考测听位置处每个监控扬声器产生之声压电平的三分之一八度音阶频率响应，使用的是50 Hz-16 kHz频率范围上的粉红噪声。测得的操作房间响应曲线将落在图2中给出的公差限值范围内。

参考测听位置处每个监控扬声器产生的操作房间响应曲线之间的差异，在整个频率范围内最好不超过2 dB目标容限值；前（±60度的方位角）扬声器（特别是中水平层的扬声器）之间的匹配最为重要。测得的操作房间响应曲线应包含在测试报告中。这些规范可以通过纳入均衡来获得。如果包含均衡，那么在测试报告中应包括对这种包含的认可，以及所采用均衡的细节。

8.3.4.2 背景噪声

在测听区域、在标称的坐着状态下测听者耳朵高度处测得的连续背景噪声（由一个空调系统、内部设备或其他外部来源产生），最好不应超过NR 10（见图3和图4）。

在任何情况下，背景噪声都不应超过NR 15。

在特性上，背景噪音不应能感知到脉动、周期或音调。

8.4 测听电平

8.4.1 扬声器再现

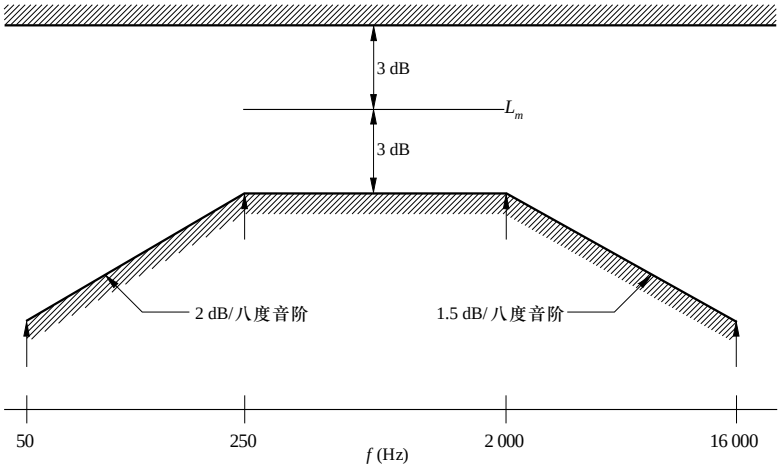
8.4.1.1 操作声压电平（参考测听电平）

参考测评电平定义为首选的听力电平，在参考测听点以某个给定的测量信号生成。它表征再现声道的声增益，以确保在不同的测听房间测听相同的素材时有相同的声压电平。

必须利用粉红噪声完成对测评配置方案中各扬声器的电平校准。

图2

操作间响应曲线的公差范围

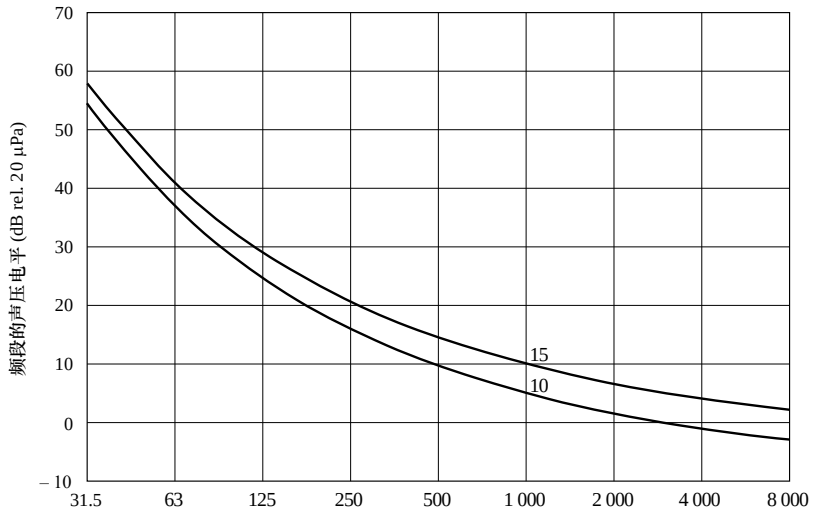


L_m：声压电平的平均值

BS.1116-02

图3

三分之一八度音阶频段背景噪音电平限值噪声等级评定曲线，基于之前的ISO NR曲线，ISO R1996建议书（1972年）

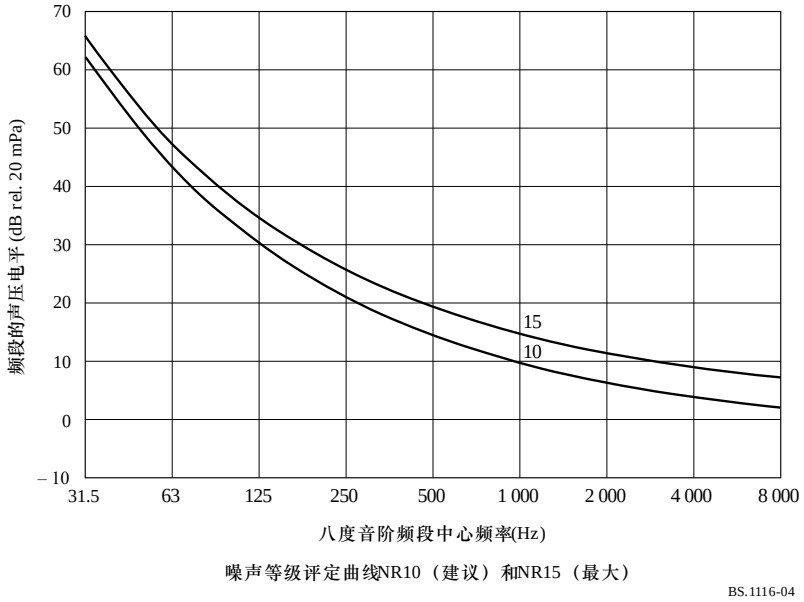


噪声等级评定曲线NR10（建议）和NR15（最大）

BS.1116-03

图4

八度音阶频段背景噪音电平限值噪声等级评定曲线，
基于之前的ISO NR曲线，ISO R1996建议书（1972年）



对依次输入各个再现声道（即功率放大器及其相关扬声器）的、均方根（r.m.s.）电压等于“校准信号电平”（根据ITU-R BS.645建议书，为0 dBμ0s；根据[EBU，1992]，为数字磁带录音限幅电平以下–18 dB）的测量信号，对放大器的增益应调整给参考声压电平（IEC/A-加权，缓慢）。

$$L_{ref} = 78 \pm 0.25 \quad \text{dBA}$$

注 1 – 高级音响系统声音参数的测量可能比早期多声道音频系统的情况复杂很多。必须慎重选择测量麦克风及其测量时的朝向。

（从之前的测试序列已经注意到，单个测听者可能倾向于不同的绝对测听电平。这不是一个首选的选项，但并不总有可能防止测听者要求这种程度的灵活性。目前尚不清楚这是否会影响某些正在接受评价之作品的可听性。因此，如果测听者确实调整了系统的增益，那么在测试结果中应注意到这一事实。）

8.4.2 耳机再现

应按以下方式对电平进行调整，即获得等同扬声器产生之参考声场的响度。为确定相同的响度，测听者应位于参考测听点。

8.5 测听配置方案

8.5.1 概述

测听配置方案描述了在测听房间中的扬声器位置和测听地点（测听面积）。

测听试验通常将在参考测听位置和其他的推荐测听位置进行。不过，也需要对因重大异常测听而引起的任何影响做出评估。出于该理由，应纳入在“最坏情况”下的测听位置。

8.5.1.1 监控扬声器的高度和方向

在方位平面中的所有扬声器的高度，测量至每个扬声器的声学中心，应位于坐着的测听者的耳朵高度。扬声器的方向应是这样的：其参考轴应通过测听者耳朵高度所在的参考位置。如果高级音响系统包括放置在不同高度位置的扬声器，那么需要记录和描述所有扬声器在水平和垂直维度上相对房间大小和测听位置的位置。

8.5.1.2 至墙壁的距离

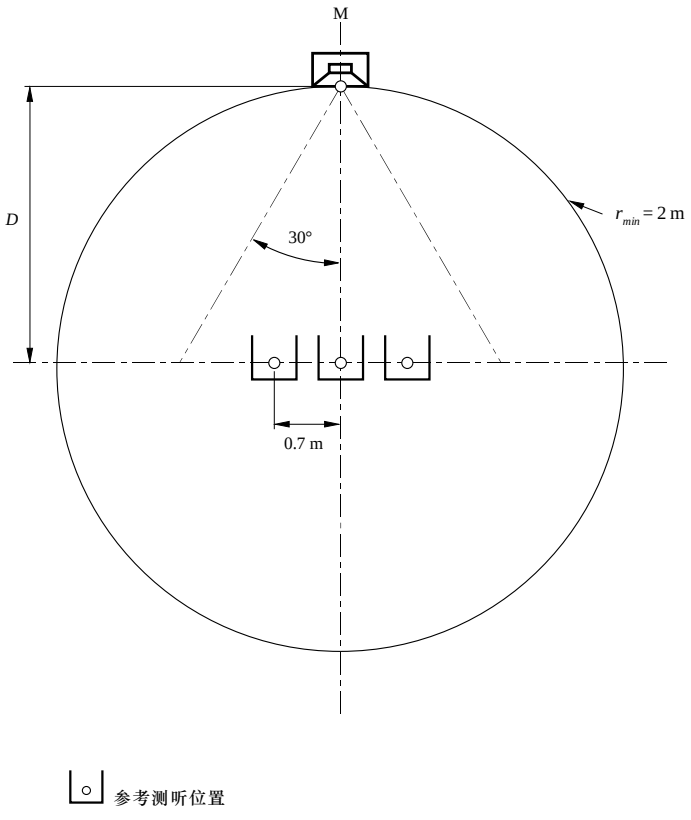
对自由站立的扬声器，自周围反射表面的扬声器声学中心距离应至少为1 m。即使因房间大小而使这变得不可能，本建议的方法仍可使用，但要求测试报告说明有关墙壁距离的标准要求未得到满足。而后应以某种其他方式对早期反射进行控制，以便满足§8.3.3.1中给出的要求，并应在测试报告对方法做出说明。

8.5.2 单声道再现

对单声道信号的再现，必须使用一个单个的扬声器。最小测听距离应为2 m，且所有的测听位置都应在距扬声器轴±30°角的范围内（见图5）。

图5

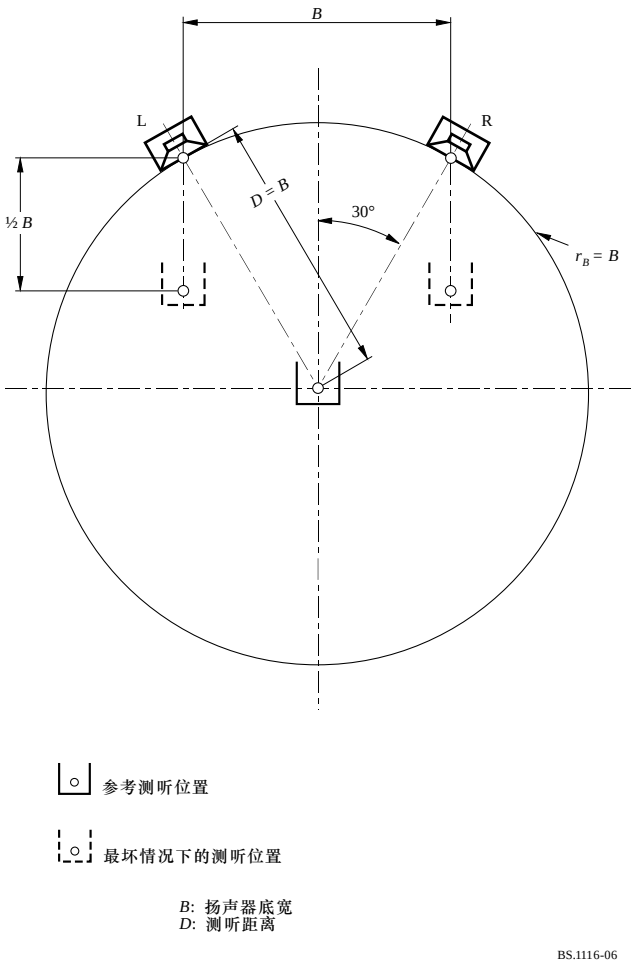
带扬声器M的参考测听配置方案以及
单声道音响系统允许的测听区域



8.5.3 双信道立体声再现

图6

针对带有细小损伤的立体声音响系统的、带扬声器L和R的测听配置方案



8.5.3.1 底宽*B*

首选的限值为*B* = 2-3 m。在设计适当的房间中，高至4 m的*B*值是可接受的。

8.5.3.2 测听距离*D*（扬声器与测听者之间的距离）

测听距离的限值为$D = 2 - 1.7\ B$ (m)。

8.5.3.3 测听位置

所谓的参考测听点通过60°的测听角来定义。

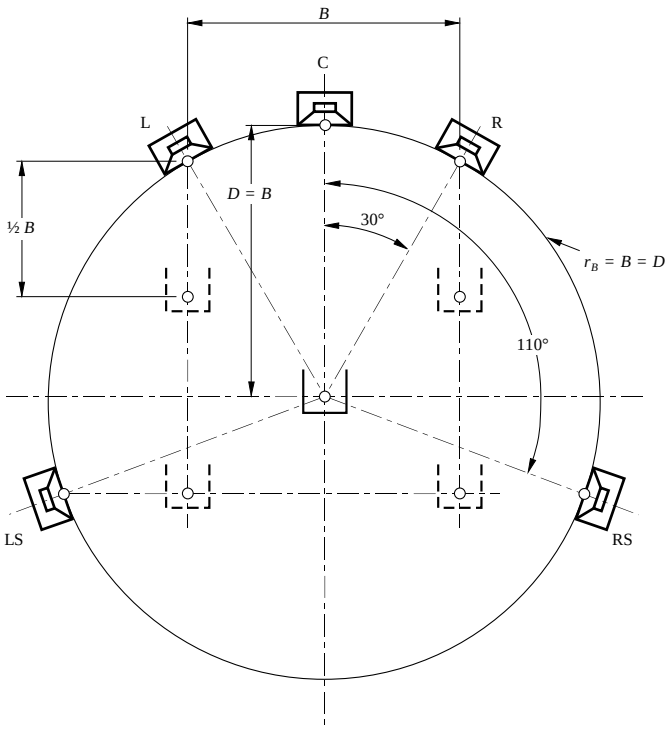
建议的测听区域不应超出以参考测听点为圆心、半径为0.7 m的圆形区域。额外的、“最坏情况下的”测听位置如图6所示。

8.5.4 多声道立体声再现

测听配置方案原则上应对应3/2多声道声音布局，如ITU-R BS.775建议书中图1所规定的那样：带扬声器L/C/R和LS/RS的参考扬声器配置方案。

图7

针对带有细小损伤的多声道音响系统的、带扬声器L/C/R和LS/RS的测听配置方案



8.5.4.1 底宽

首选的限值为B = 2 - 3 m。在设计适当的房间中，高至5 m的B值是可接受的。

8.5.4.2 测听距离和底角

参考测听距离将为B，因此参考底角等于60°。

8.5.4.3 测听位置

如上所述，所谓的参考测听点通过60°的测听角来定义。额外的、“最坏情况下的”测听位置如图7所示。

8.5.5 高级音响系统再现

为了阐明实验条件，测试中使用的所有扬声器的位置（距离和角度）及其有关测听位置的相对位置，都必须在测试报告中进行详细描述。该描述必须遵循与ITU-R BS.775建议书中所述之扬声器布局和测听位置相称的形式和内容细节。它还需要确定和描述高级音响系统布局垂直维度上的所有扬声器的位置，包括在不同高度位置上的扬声器。ITU-R BS.2051建议书包括在本文中也有用的信息。

9 统计分析

对测试结果进行统计分析的基本目的是准确确定每个被测系统的平均性能以及这些平均性能数据间任何差异的可靠性。后者方面需要对结果的变化或方差做出估计。

如果测试是根据当前文档其他章节中所述的程序进行的，那么刻度很可能将是刻度形式的，即评分刻度的每一步都将大致等同任何其他步的刻度。不过，实现的刻度属性将既不禁止也不规定任何特定的统计方法。

只要能够合理满足用于基础参数统计的假定，那么这种方法将最敏感、最强大，因此建议采用。只有当数据的重要属性显示严重偏离用于方差分析（ANOVA）的假定时，才考虑采用替代的分析方法（如非参数的方法）。具体而言，建议第一阶段在做主要分析时，采用一个方差分析模型。随后，可采用使用ANOVA提供之方差估计的其他方法（如t-测试、Neuman-Keuls、Scheffe等），来开展更加详细的研究，当中有待发现由ANOVA来揭示的、大的总体影响（如果有的话）。

某个特定的假设通常可以通过几种不同的统计方法来验证。如果发现某个特定的假设也可利用一种替代的统计方法来进行验证，那么做出决定的依据可得到增强。因此建议采用补充数据分析方法（如Wilcoxon等）。

同样重要的是，在某个阶段也需要考虑心理学方面的问题。这些当然会对从一个非物理的刻度能得到什么样类型的有意义结论产生影响。

应该注意的是，除非可以证明评分刻度是线性的，否则只能基于等级次序，对不同的评分进行比较。

10 统计分析结果的表示

10.1 概述

应对统计分析结果进行适当表示，以便一个幼稚的读者和一个专家都能评估相关的信息。最初，任何读者都希望能够看到整个实验结果，最好以某种图形化的形式，这样一种表示方式，可以通过更加详细的定量信息来支持，虽然完整、详细的数值分析应放在附录中。

10.2 绝对等级评分

分别针对客体和隐藏参考的绝对平均等级评分表示，可以形成对数据的一个很好的初始印象。

但应记住，对任何详细的统计分析而言，这不是一个适当的基础。这是由于以下事实：当使用推荐的测试方法时，测听者清楚了解成对比较中的一个来源等同于参照物。因此，观察将变得不独立，这些绝对等级评分的统计分析结果将不会带来有意义的信息，因此不应该这样做。

10.3 差别等级评分

对统计分析而言，提供给隐藏参考和客体的、等级评分之间的差别是适当的输入。图形化表示清楚地揭示了至透明度的实际距离，这通常是关心的主要问题之所在。

10.4 等级评分显著性水平和置信区间

测试报告应向读者提供有关所有主观数据内在统计特性的信息。应对显著性水平做出说明，并对统计方法和结果的其他细节做出说明，以便于读者理解。这些细节可包括置信区间或错误柱状图。

当然没有任何“正确的”显著性水平。不过，值0.05是传统的选择。原则上，依据正在接受测试的假设，有可能使用一个单尾的或双尾的测试。

11 测试报告的内容

测试报告应尽可能清晰地传达研究的基本原理、所用方法和所获结论。应充分陈述细节，这样，原则上，为根据经验对研究结果进行检查，一个知识渊博的人就可以复现研究工作。一个见多识广的读者应能理解并对测试的主要细节做出评判，如研究的基本理由、实验设计方法和事实、分析和结论。

应特别注意以下几点：

- 有关测听者和素材的规格说明和选择；
- 测听环境和设备的物理细节，包括房间的尺寸和声学特性、传感器类型和布置、电气设备规格；
- 确定和描述所测声道配置是否在ITU-R BS.775建议书或ITU-R BS.2051建议书中进行了描述；

  如果未在ITU-R BS.775建议书中对进行说明，那么所测音响系统的所有扬声器的位置都必须如ITU-R BS.775建议书中所规定的那样，以可比的细节予以记录，以允许外部可重复性。相对所测音响系统相关的扬声器位置，对参考测听位置，也必须予以记录（见§§8.5.4和§§8.5.5）；
- 是否满足§8.5.1.2中确定的距离要求；如果不满足，那么必须引起注意；
- 如果不满足§8.5.1.2中确定的距离要求，那么应对用于控制早期反射并满足§8.3.3.1中给出的要求的各方法予以描述；
- 测量操作空间对所有扬声器的响应；如果对均衡进行处理，那么应提到对该过程所的确认，以及均衡使用的方法；
- 对本文档中所述之声学和物理房间要求的任何偏差都应予以报告。这些偏差包括：容忍的操作房间声学测量和响应，如§8.3中所规定的那样；所有扬声器行为响应性能指标，如§8.4中所指出的那样；所有的物理距离要求，如§8.5中所指出的那样；

– 在评估者测听位置测得的、来自每一个扬声器的脉冲响应，测听房间按即将用于测试的方式进行设置（包括室内陈设），显示在时间域中；
– 实验设计、培训、指导、实验序列、测试程序、数据生成；
– 处理数据，包括详细的描述性和分析性推理统计；
– 获得的所有结果的详细依据。

附件1之
附加材料1

对测听者进行后筛选的统计方面考虑

1 评估测听者的专业技能

双盲三激励隐藏参考方法为每个试验提供了两个等级评分，使得基于逐个测听者，对这两个等级评分直接进行比较并逐个对所有试验的比较结果进行检查成为可能。对每个试验，可以考虑每个试验两个等级评分之间的代数差，当然，总可在相同方向上相减。假设从客体的等级评分中减去隐藏参考的等级评分。

如果测听者整体上未能正确确定隐藏参考和客体，那么相对测听测试中测听者的所有差别等级评分的平均值都将为零或接近零，原因是存在倾向于相互抵销的正等级评分和负等级评分。如果测听者整体上能够正确检测哪个是隐藏参考、哪个是客体，那么差别等级评分的平均值将在负方向上偏离零，原因是相比正的等级评分，将有更多的负的等级评分。

因此而获得的数据将接受单方面t-测试，以评估以下可能性，即每个测听者的分布的平均值为零。如果对某个给定测听者拒绝该无效假设，那么可以得到以下结论：该测听者的数据源自以下分布，即在某个给定的置信度水平上，在负方向上，分布的平均值大于零。而后可得到以下结论：对情况如上所述的每一个测听者，显示他或她整体上不只是在猜，而是可认为这些测听者已显示具有足够的专业技能来分析和评判包括数据在内的最终实验结果。其他测听者的数据－－据此统计标准评判整体上是在猜－－可拒绝其参加进一步的分析工作。

应记住，关于本文的各建议只涉及细小损伤。如果事实证明，不管出于什么原因，在测试中包括了许多“大的”损伤，而不只是“小的”损伤，那么如上所述幼稚地应用的后筛选方法可能导致虚假的或不当的结论。一个“大的”损伤在此指的是一个比较容易检测的损伤，即使是“非专家”测听者也可检测到。显然，一些嵌在上下文中的、真正的“小的”（很难检测的）损伤，在如上所述的t-测试中所占份量是很轻的，当中，绝大多数损伤是“大的”（容易检测的）损伤。因此，可对细小损伤条目正确做出判断的专家在整体性能上可能无法区别于非专家，非专家对这些细小损伤条目的判断处在“猜”的水平上。这将是真实的，原因是，在t-测试评价中，有关细小损伤条目的性能可能在统计噪声上失去，原因是，t重要性的最大权重将由大的损伤条目来给出。

即使在最好的“细小损伤”测试中，一些大的损伤条目或者容易的条目也几乎都不可避免地被发现，即使这些条目通常远短于大多数条目。考虑到这点，那么建议：出于排他性的、非常严格的后筛选t-测试目的，所有“容易的”或大的损伤条目通常都应被排除在有关评价测听者专业技能的t-测试程序之外。这些可能是在所有测听者中得到较低平均等级评分的所有条目，也就是说，差别等级评分在–2.0和–4.0之间。对此类条目，大多数的测听者都将能正确地将客体从隐藏参考中辨别出来，在t-测试中纳入此类条目将给差别测听者专业技能的评价带来模糊，而非便利。将大的损伤条目排除在t-测试分析之外的效果是夸大或高估测听者的表象专业技能。

在相反的情况下，可能会有太多的“真正透明的”条目被引入本建议书的§5中。在这种情况下，显然这是透明的（“太难的”）条目，有可能在后筛选t-测试中被省略掉。然后，因其已知的影响力而引入的特殊条目，将按预期在t-测试中拥有更大权重。将显然透明的条目在t-测试中省略掉，这对t-测试的影响将是低估测听者的专业技能。

一般来说，一贯“太难的”或“太简单的”条目对区分适当的专家和不适当的专家而言是没有区别的。

适当应用后筛选t-测试的独特优势在于，可通过实验中的性能，来评估针对某个特定实验的专业技能是否足够。在一系列涉及不同实验中相同测听者的实验中，可发现，在所有测听者成功通过预筛选的同时，这些测听者中的某些测听者可能是有关实验某个子集的适当专家，但后筛选可能证明，他们并不是有关所有实验的适当专家。那么在这种情况下，对特定的测试结果而言，某个特定测听者的数据可能被接受或者被拒绝。实际上，这是对“专业技能”概念所做的一个微调，超出了可能与单纯依赖预筛选有关的概念。

有一点需要引起注意。一个不适当的测听专家不能提供良好的数据。因此，拒绝因不足的专业技能而得到的数据（通过严格的后筛选来客观地确定）是正当的。另一方面，不能保证来自某恰当通过t-测试后筛选的测听者的数据就一定是好的数据。作为一个极端的例子，一个测听者可能在实验中100%地正确区分客体和隐藏参考，但数据可能显示，他或她对所有试验中的所有客体都给了1.0的等级评分。换句话说，对所有的试验而言，来自该测听者的所有数据集，可以换为一个不同的等级评分–4.0。

假设在所有试验中，该实验中的所有其他测听者显示出一个“更平常的”等级评分分布，那么来自某个测听者的、非常奇怪响应模式（所有都为“–4.0”差别等级评分），可能会导致其主张拒绝该数据。不过，或许除了在一个明显出现偏差的单个例子中之外，如此处所述，对数据的适合性将很难应用此类事后分析准则。这相当于根据实验者的偏见刻意塑造数据，而不是接受实际结果的经验性证据。

此类事后分析方法不可以使用。只要在一个实验中测听者的总数是足够的，那么即使测听专家的数据出现明显偏差，也将对总的数据集产生很少的扭曲影响。即使除了测听专家之外，还包括其他偏差，敏感的实验中出现重大的、可复制的结果也是相当平常的。在完成一个实验后，如果对数据的“良好性”存在负面的怀疑，那么唯一的办法是从头开始，使用一组全新的测听者，重新做过整个实验，并努力纠正在之前所用的实验程序中存在的任何可疑的缺陷。

2 进一步评估测听者的专业技能

随着基于感知质量的有损编码解码器的增加，不可避免地，拥有适当专业技能、足以辨别剩余编码质量的测听者将越来越少。对过去的测试（包含比较“容易听得见的”声音内容）而言拥有适当专业技能的测听者，在一个没有太多“容易听得见的”声音内容的测试中，可能就不是一个适当的专家了。此外，尽管一个测听者的t-分数可整体表明其对实验拥有适当的专业技能，但测听者可能没有足够的专业技能来辨别出参考信号与高质量编码信号之间存在的差别。在这种情况下，对总的测听者数据可能会增加“统计噪声”数据，从而掩盖被其他测听者感知到的真正差异。

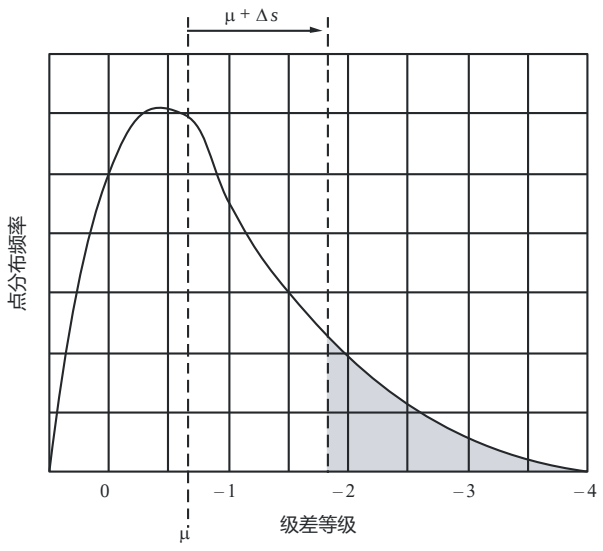
附件1之
附加材料2

评估测听者的专业技能水平

目前，在某个给定测试中的测听者的所有数据都被用于评估其t-评分。而后，来自所有具有足够高t-评分的测听者的数据都将被纳入到方差分析（ANOVA）中。

图8

在t-测试之前丢弃数据点的方法



BS.1116-08

在目前的提议中，我们建议就每个测听者的数据子集，对t-测试进行若干次迭代。对每次迭代，用于评估某个测听者专业技能水平的准则将变得更加严格。

将对测听者的专业技能水平重新进行评估，如果他表现出足够的专业技能水平，那么他的数据将被包括在随后的方差分析（ANOVA）中。因此，通过每次迭代，有关何为恰当的专业技能的准则将提高，并利用来自剩余测听者的数据进行方差分析（ANOVA）。提议的、用于评估专业技能的准则如下所述。

对一个假设的数据集而言，过程如图8所示。首先，计算有关测听者数据的平均值和标准偏差。而后这将用来确定该测听者数据对应的z-分数（见注1）。自此，一个测听者的所有数据点（跌穿某个准则（$\mu + \Delta 1\,s$））将被丢弃，并对剩余的数据点进行一次新的t-测试。如图所示，跌穿$\mu + \Delta 1$（阴影区域）的那些数据点将被丢弃，其余的数据点（非阴影区域）将被用于后续的t-测试。如果对剩余的数据点，t-测试仍显示测听者拥有足够的专业技能，那么该测听者的所有数据都将包含在后续的方差分析（ANOVA）中。如果该测听者在t-测试中未能显示拥有足够的专业技能，那么该测听者的数据将被从所有后续的方差分析（ANOVA）中彻底消除。而后以更加严格的专业技能准则（$\mu + \Delta 2\,s$），重复此过程。以准则（$\mu + \Delta i\,s$，$i = 0$，1，..，N）重复此过程N次。现在，利用CRC（加拿大通信研究中心）在之前的研究中得到的数据，对适当的$\Delta i\,s$和N值进行分析研究。

注1 – z-评分表示对平均值为0、标准差为1的分布进行归一化后得到的评分。它定义为 $z = \frac{x - \mu}{s}$，式中，x 为一个数据点，μ 为样本平均值，s 为样本标准差：

$$s = \sqrt{\frac{N\sum x^2 - \left(\sum x\right)^2}{N\,(N-1)}}$$

附件1之
附加材料3

给测听者的指令样例

在这些指令中使用的术语并不严格遵守词汇表中的定义。

1 熟悉或训练阶段

训练阶段的目的是让测听者识别和熟悉潜在的失真以及测试系统产生的内容。训练后，你应该知道“听什么”。今天下午，将要求你“盲测”为今天上午你将要听到的所有音频材料评定分数。在训练阶段，你也将要熟悉测试程序。

你将听到有关音频材料每个条目的参考版本（原件）和经过处理的版本。在视频监控屏幕上，将用字母“A”来指明信号的参考版本，用字母“B”来指明信号经过处理的版本，用字母“C”来指明信号的“隐藏参考”版本。在演示期间，在任何时候，你都可以自由地在“A”、“B”、“C”之间进行自由切换。这应允许细致地、详细地对“A”、“B”、“C”之间进行比较。这是“A”与“B”之间以及“A”与“C”之间的差别，将给出其等级评分。典型地，音频序列将为10 ~ 25 s 长，只要你想要，就可以反复播放。在训练期间，

你可以自由地使用扬声器、耳机或两者。对所有条目，你至多有三个小时的训练时间，今天下午，在“盲测”评分阶段，你将为这些条目正式确定等级评分。

在下午测试期间，将要求你依据表2中的评分刻度，为展示的内容评定分数：

表2

损伤程度	等级评分
感知不到失真	5.0
能感知到失真，但不恼人	4.0
能感知到失真，稍微恼人	3.0
能感知到失真，恼人	2.0
能感知到失真，非常恼人	1.0

将向测听者描述等级评分的含义。需要强调的是，等级评分将被认为是一个连续的、等间隔的刻度，在特定的值上定义锚点。

由于下午的每个试验都包含一个隐藏参考（即参考的一个完美的复制品），对每个试验，预计至少给出一个5.0的等级评分（但只能有一个（见注1））。如果你发现“B”或“C”优于参考，那么这意味着，发现了一个“能感知到失真，但不恼人”的差别，依据检测到的差别，可为之给出一个4.0 ~ 4.9之间的等级评分。

同时你应考虑在训练阶段，如何作为一个个体，就等级评分而言，对声音损伤做出解释，重要的是，任何时候，你不应与其他测听者对这种个人解释进行讨论。

注1 – 建议之变化的目的是迫使测听者做一次“最佳的猜测”，针对该激励的是经过编码的材料。我们认为，某些测听者实际上是能够检测到非常细小的内容的，但由于其保守的方法，将给出两个5.0的等级评分，而不是更加明确地表明其态度。建议之变化将解决该问题。

2 训练阶段内容的样例

主要的培训，持续至多三个小时，在第一天上午，应该由约四个测听者组成一组，以组的形式进行。应提前向测听者发出书面指令。

训练的完整测试应包括以下几点：

– 简要介绍测试的目的和目标；
– 回放选中的测试素材，以使测试的测听者熟悉声音内容并了解编排材料，以便之后进行评价；
– 简要说明被测系统，口语解说预选小组建立的的损伤类别；
– 使用一些最受损伤的条目，对损伤进行示范；
– 解释待分级的属性；
– 解释五级损伤等级评分；
– 培训切换和等级评定。

在后续的测试天数里，应提请测听者注意在主训练完整测试中覆盖的各点。这可能包括在进行正式的测试之前，再次测听测试条目。

3 盲测评分阶段

盲测的目的是评定今天早上在训练阶段你听到的各种各样音频材料的等级。

在每个试验中，你将试听给定音频材料的三个版本。对这些，将在视频监控屏幕上贴上“A”、“B”、“C”。“A”总为参考版本（原件），依据它，对“B”和“C”进行比较，并评定等级。“B”或“C”之一为经过处理的版本，另一个为隐藏参考（等同参考）。将不告诉你“B”和“C”哪个是经过处理的版本、哪个是隐藏参考，因此，“盲测”针对的就是该等级评定阶段。你可以在任何时候在“A”、“B”、“C”之间自由地进行切换。可以反复播放音频序列，直到你确信你做出的评估。由你来决定，当对某个给定的试验所做的评估感到满意后，你可以继续进行下一个试验。

在每个试验中，一方面，要求你使用表3中所示的五级等级评分刻度，对在“B”和“A”之间感知到的差异（如果有的话）做出评价，另一方面，要求你使用表3中所示的五级等级评分刻度，对在“C”和“A”之间的差异做出评价。因此，在每个试验中，必须给出两个等级评分，一个针对“B”，一个针对“C”。对每个试验，预计至少给出一个5.0的等级评分（但只能有一个（见本附加材料§1中的注1））。在每个试验结束时，请在计算机中输入你的等级评分。

而不是将等级评分输入到计算机中，可以使用一个纸质的等级评分表。

而后将向测听者展示表3，并将在整个盲测等级评分完整测试中使用表3的一个副本。

将向测听者描述等级评分的含义。需要强调的是，等级评分将被认为是一个连续的、等间隔的刻度，在特定的值上定义锚点。

表3

损伤程度	等级评分
感知不到失真	5.0
能感知到失真，但不恼人	4.0
能感知到失真，稍微恼人	3.0
能感知到失真，恼人	2.0
能感知到失真，非常恼人	1.0

附件1之
附加材料4

主观评价：词汇表

为清楚起见，在此处定义了以下在本建议书中使用的术语。也可参见图9，它说明了其中一些术语间的相互关系。

属性（Attribute）

根据给定的口头或书面定义，测听活动可感知的特征。

盲测（Blind Test）

一种测试方法，在该测试中，激励是为测听者提供的唯一信息源。

双盲测试（Double Blind Test）

一种盲测方法，在该盲测中，试验者和测听测试之间不存在任何不受控制的交互的可能性。

素材（Excerpt）

适合于评价被测给定系统声音质量的个性特征或参数的一段音乐、语音或其他声音信号样本。

测试素材通常为声音录音（CD、R-DAT或其他录音或源格式）。

等级评分（Grade）

根据给定标准，一个属性量级的数字表示。

隐藏参考（Hidden Reference）

未向测听者标识的参考

条目（Item）

由被测系统处理过的一段素材。

测听组（Listening Panel）

为测听测试生成数据的所有测听者。

位置（Location）

测听测试进行的位置。可能是测听室的地理位置或测听者位置。它可以是测试要素之一。

客体（Object）

用被测系统处理过的一些素材表示的被测系统。

参考（Reference）

测试素材，未经测试客体处理而生成的测试素材，用作损伤测试的比较基础。

完整测试（Session）

在某一持续时间内，有待一个测听者或一个测听组进行评估的所有试验。

激励（Stimulus）

客体或隐藏参考或者参考和部分或全部素材的组合。

测听者（Subject）

在一个测听测试中负责评估激励的测试人。

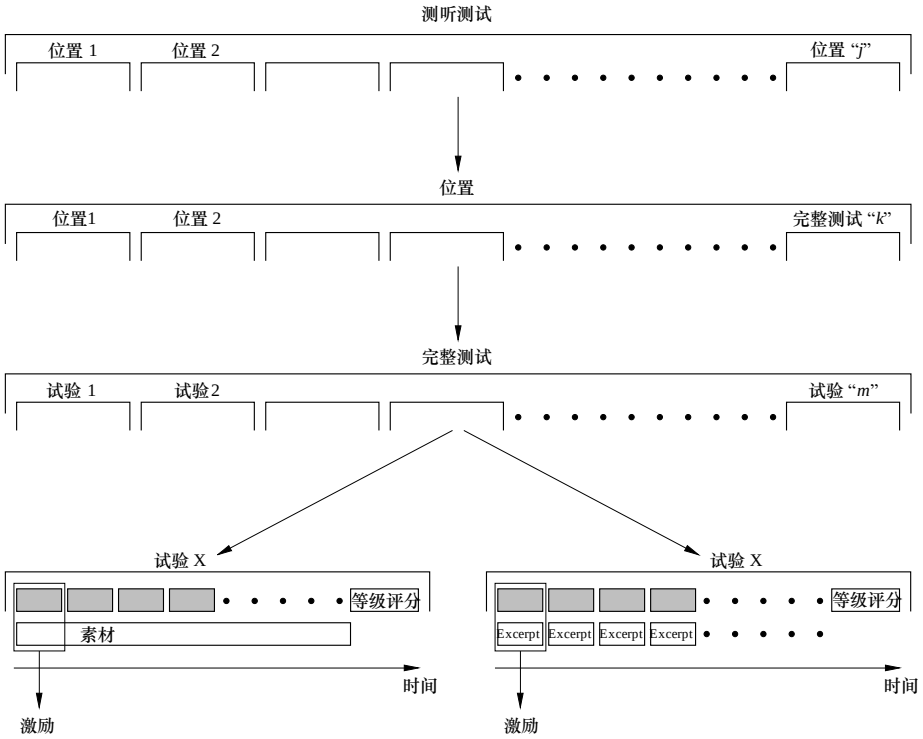
试验（Trial）

一个完整测试的子集，该子集以陈述一组激励开始，并以完成对其等级评分结束。

测听测试为针对某一被测声音系统所进行的一次主观测试活动。该活动涉及到上面所介绍的一些术语，这些术语之间的关系如图 1所示。

图 9

词汇表中使用的一些术语之间的相互关系示意图



所示的两个试验表示一系列可能的配置方案的结束点。

BS.1116-09