GY

中华人民共和国广播电视和网络视听行业标准

GY/T 340—2020

超高清晰度电视图像质量主观评价方法 双刺激连续质量标度法

Subjective assessment methods for image quality of ultra high-definition television——Double-stimulus continuous quality-scale

2020－12－31 发布　　2020－12－31 实施

国家广播电视总局　　发布

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》的规定起草。

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

本文件由全国广播电影电视标准化技术委员会（SAC/TC 239）归口。

本文件起草单位：国家广播电视总局广播电视规划院、中央广播电视总台、广东广播电视台、江西广播电视台、深圳广播电影电视集团、广州市广播电视台、超高清视频（北京）制作技术协同中心。

本文件主要起草人：张乾、王惠明、李岩、范创奇、范晓琳、封连伟、滕建新、鲁泳、邓向冬、何杰锋、宁金辉、林小海、彭子舟、周立、何向晖、曾靓、李光辉、王丛璐。

超高清晰度电视图像质量主观评价方法　双刺激连续质量标度法

1　范围

本文件规定了实验室环境下对超高清晰度电视图像质量进行双刺激连续质量标度主观评价的方法（简称双刺激连续质量标度法）。

本文件适用于对超高清晰度电视系统和设备的图像质量进行主观评价。

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。

GY/T 307—2017　超高清晰度电视系统节目制作和交换参数值

GY/T 315—2018　高动态范围电视节目制作和交换图像参数值

ITU-R BT.500-14　电视图像质量主观评价方法（Methodology for the subjective assessment of the quality of television pictures）

3　术语和定义

本文件没有需要界定的术语和定义。

4　缩略语

下列缩略语适用于本文件。

HDR　高动态范围（High Dynamic Range）

SDR　标准动态范围（Standard Dynamic Range）

5　双刺激连续质量标度法

5.1　概述

双刺激连续质量标度法一般用于评价被测系统的图像质量或传输系统对图像质量的影响。尤其对于考察被测系统整体质量的情况，双刺激连续质量标度法（符合 ITU-R BT.500）特别有效。采用该评价方法时，观看员观看一对图像，这两个图像均来自同一个信号源，其中一个是信号源直接输出的源图像，另一个是信号源经过被测系统以后的图像，称为被测图像，源图像和被测图像按伪随机顺序进行排列。观看员需对二者的图像质量进行评价。

评价周期最长时间为30min，评价结束后，对源图像和被测图像的评分进行计算。

5.2　实验室主观评价条件、显示器技术要求及参数值

实验室主观评价条件要求应符合表1的规定，显示器技术要求及参数值应符合表2的规定。

表1 实验室主观评价条件

序号	项目		技术要求
1	测试观看距离[a]	7680×4320 图像	0.8 倍图像显示高度
		3840×2160 图像	1.6 倍图像显示高度
2	水平方向测试观看角度	7680×4320 图像	±96°
		3840×2160 图像	±58°
3	显示器后的背景亮度与图像峰值亮度的比值	SDR	约 0.15
		HDR	≤0.005
4	环境亮度		≤5cd/m^2
5	背景色温		D_{65}
[a] 如果对被测系统的图像清晰度进行评价，则应选择测试观看距离；如果对被测系统的图像评价要素不包括清晰度，对于 3840×2160 图像观看距离可以选择 1.6～3.2 倍图像高度，7680×4320 图像观看距离可以选择 0.8～3.2 倍图像高度。			

表2 显示器技术要求及参数值

序号	项目		技术要求及参数值
1	显示器尺寸	7680×4320 图像	对角线的尺寸宜大于等于 1.78m（70in）
		3840×2160 图像	对角线的尺寸应不小于 1.40m（55in），宜大于 1.78m（70in）
2	显示器物理分辨率	7680×4320 图像	≥7680×4320
		3840×2160 图像	≥3840×2160
3	显示器色域		支持 BT.2020 色域
4	显示器峰值亮度(cd/m^2)[a]	SDR	150～300
		HDR	宜大于等于 1000
5	显示器对比度[b]	SDR	≤0.02
		HDR	≤0.000005
[a] 峰值亮度是指 100%峰值视频电平下的亮度。 [b] 该值为显示器黑场亮度与峰值亮度之比，会受到环境光的影响。			

5.3 测试图像

测试图像的格式应符合GY/T 307—2017和/或GY/T 315—2018的要求。

一套测试图像应包含至少4个图像，这些图像可以是静止图像或具有运动特性的图像序列，每个运动图像序列大约持续10s～15s。对被测系统而言，测试图像应具有最佳的图像质量。

测试图像应是“严格的，但又不过分”的通用测试图像，既包含对各种评价因素比较敏感的内容，同时又能代表电视节目的典型内容。图像质量的评价要素包括：清晰度、动态清晰度、亮度层次（高亮度部分层次重现、低亮度部分层次重现）、对比度、彩色还原和运动特性等。

5.4 观看员

观看员即应邀参加主观评价的评分人员，通常有两种类型，即专业观看员和非专业观看员。一般由非专业观看员来进行主观评价，当需要精确判断时，可由受过专业训练的专业观看员来进行评价和分析。

观看员应具有代表性，即应包括不同性别、年龄、文化层次的观众；应具有正常的视力（含校正视力）和色觉；应具有一定分析判断能力；应能较快地接受和掌握评价方法和要求。

主观评价所需观看员的人数应大于等于15人。

5.5 评价测试阶段

在每个评价周期开始时，应向观看员详细、准确地介绍评价方法、质量要素或可能出现的损伤类型、评分标度、测试图像和评价时间长度，并进行评价示范显示。示范显示应使用不同于正式测试的图像或序列，但示范演示图像或序列应代表将要评价的被测系统损伤的类型和损伤的程度范围，且与正式测试中使用的图像或序列具有可比性。

一个评价周期包括示范说明在内应不超过30min。在正式测试开始前，应引入3至5个评价序列来稳定观看员的判别力，其结果数据不纳入测试结果的统计中。不同测试图像的显示顺序采用伪随机方式。为了检测相关性，有些测试可以重复进行，但要避免相同测试图像在相继的评价序列中出现。

评价周期的显示流程见图1。

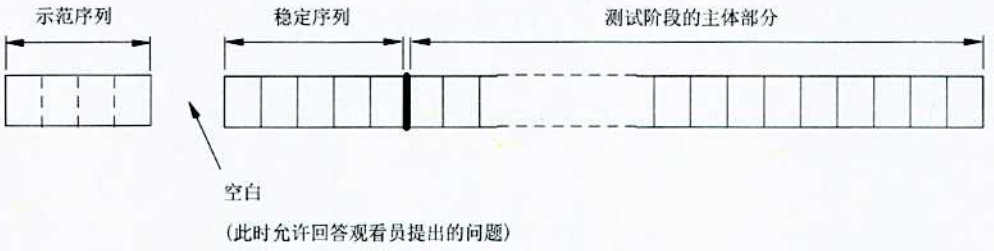
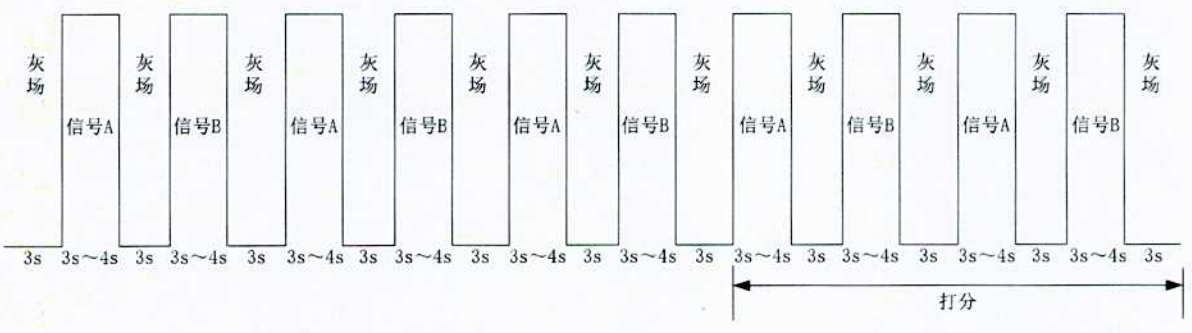


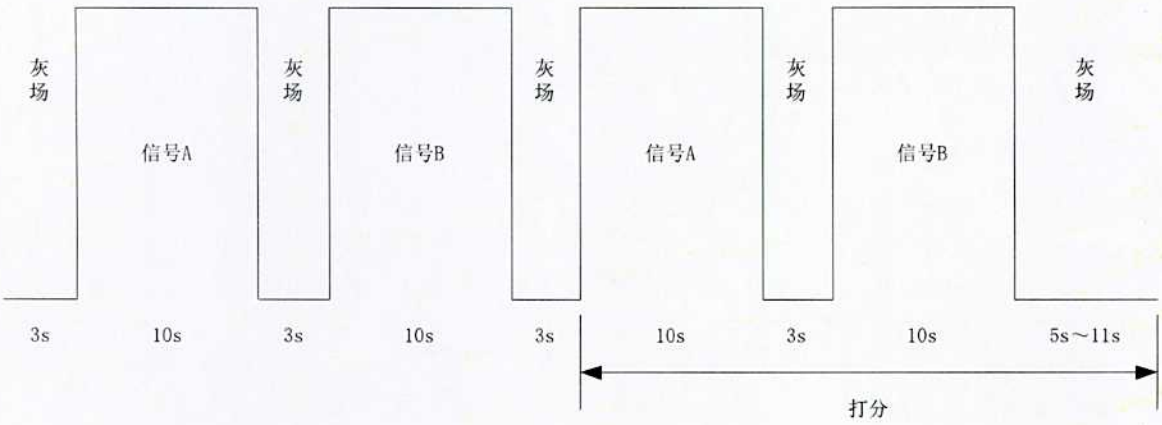
图1 评价周期的显示流程

5.6 测试图像的显示

一个评价周期由多次播放显示组成。在每一次播放显示时，首先要显示一次或多次信号A和信号B，每次持续时间相同，以便使观看员得出相应的判断，然后再显示一次或多次这两个信号，观看员进行评分。重复的次数取决于测试序列的长度。对于静止图像，使用3s～4s的序列并重复5次（在最后2次显示图像期间评分）。对于活动图像，使用10s的序列并重复两次（在第二次重复期间评分）。图2表示了播放显示顺序。其中，信号A和信号B均有可能是源图像或被测图像，且不告知观看员哪一个是源图像和哪一个是被测图像。



a） 静止图像



b） 活动图像

注：对于SDR信号的评价，灰场是电平为200mV的中灰视频信号；对于HDR信号的评价，灰场是30%HDR电平的中灰视频信号。

图2 双刺激连续质量标度法的测试图像演示顺序

5.7 评分标度

评价时，观看员应在垂直标尺上标出记号来确定每次演示图像的总体质量。垂直标尺是成对的，对应每个测试图像的两次演示。为了防止量化误差，标尺提供了连续的评分机制，并分成了长度相等的5段，对应优、良、中、差、劣5个等级。图3给出了典型评分标度。为了防止在标尺与测试结果之间出现混淆，标尺用黑色印刷，评分结果用红色记录。

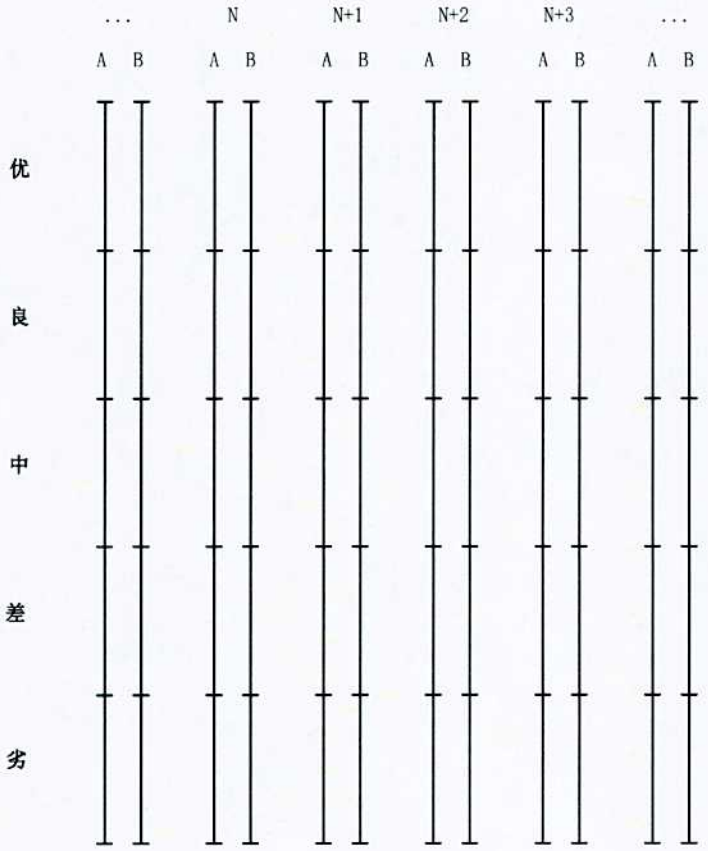


图3 典型评分标度

5.8 结果分析

5.8.1 评分量化

将每个测试条件下的源图像和被测图像从评分标度上的度量长度转换为0至100范围内的评分，然后计算源图像与被测图像之间的差值。

5.8.2 平均分计算

对评分结果进行分析的第一步是计算每一显示片段的平均评分$\bar{u}_{jkr}$，见公式（1）。

$$\bar{u}_{jkr} = \frac{1}{N}\sum_{i=1}^{N} u_{ijkr} \quad \cdots\cdots (1)$$

式中：

u_{ijkr}——观看员i在测试条件j、测试序列/测试图像k、重复r次情况下的评分；

N ——观看员数量。

类似的，可计算出每一测试条件和每一测试序列/测试图像的总平均评分$\bar{u}_j$和$\bar{u}_k$。

5.8.3 置信区间计算

在给出某一显示片段所有评分（即一个样本）的平均值时，也应给出其相应的95%置信区间。置信区间与样本的标准偏差和大小有关。

样本的95%置信区间如下：

$$[\bar{u}_{jkr}-\delta_{jkr},\bar{u}_{jkr}+\delta_{jkr}]$$

其中：

$$\delta_{jkr}=1.96\frac{S_{jkr}}{\sqrt{N}} \quad (2)$$

每一显示片段的标准偏差S_{jkr}由公式（3）给出。

$$S_{jkr}=\sqrt{\sum_{i=1}^{N}\frac{(\bar{u}_{jkr}-u_{ijkr})^2}{(N-1)}} \quad (3)$$

在所有评分满足正态分布的条件下，测试获得的平均值和平均值的真值（即通过相当多的观察者获得的评分）的差值绝对值小于置信度间隔（公式（2）给出）的概率是95%。

同样，能够计算出每一测试条件下的标准差S_j。值得注意的是，当采用较少测试序列/测试图像的情况下，相对于观看员之间的评分差别而言，所用测试序列之间的差别对标准差的影响更大。

5.8.4 观看员筛选

如果测试中观看员数量较少且这些观看员均为非专家时，可对观看员进行筛选。

计算每次评价显示的均值$\bar{u}_{jkr}$、标准偏差S_{jkr}和峰态系数β_{2jkr}，其中β_{2jkr}由公式（4）给出。

$$\beta_{2jkr}=\frac{m_4}{(m_2)^2} \quad (4)$$

式中：

$$m_x=\frac{\sum_{i=1}^{N}(u_{ijkr}-\bar{u}_{jkr})^x}{N}$$

对于每一个观看员 i，计算其P_i和Q_i，P_i和Q_i的初始值为 0。

对于j，k，r=1，1，1 至J，K，R，

若 2≤β_{2jkr}≤4，则：

若$u_{ijkr}\geq\bar{u}_{jkr}+2S_{jkr}$　　则$P_i=P_i+1$

若$u_{ijkr}\leq\bar{u}_{jkr}-2S_{jkr}$　　则$Q_i=Q_i+1$

否则：

若$u_{ijkr}\geq\bar{u}_{jkr}+\sqrt{20}S_{jkr}$　　则$P_i=P_i+1$

若$u_{ijkr}\leq\bar{u}_{jkr}-\sqrt{20}S_{jkr}$　　则$Q_i=Q_i+1$

若$\frac{P_i+Q_i}{J\times K\times R}>0.05$且$\left|\frac{P_i-Q_i}{P_i+Q_i}\right|<0.3$，则删除该观看员$i$。

式中：

N——观看员数量；

J——测试条件的数量，包括基准在内；

K——测试图像或序列的数量；

R——重复次数；

L——测试演示的次数（在大多数情况下，演示的次数等于$J\times K\times R$，不过有些评价对每一测试条件都采用数目不等的序列）。

对于某次评价获得的评分数据，筛选数据采用以上的方法只能进行一次。

5.9 结果说明

在使用双刺激连续质量标度法时，不应将双刺激连续质量标度数值与其他测试方法所用的属性词（例如双刺激损伤标度法中的不可察觉、可察觉但不讨厌等）相关联并描述被测系统图像质量。

用双刺激连续质量标度法得到的结果是源图像评分与被测图像评分之间的差值，因此，将结果和说明质量的描述语关联是不正确的，即便是与双刺激连续质量标度法本身所用的描述语（例如优、良、中等）相关联也是不正确的。

参 考 文 献

[1] GY/T 134—1998 数字电视图象质量主观评价方法
[2] T/CSMPTE 3—2018 超高清电视图像质量主观评价方法