第1章

多媒体体验质量评价概述

1.1 背景

根据 Virtual Network Index 报告，2023 年全球互联网用户规模可达到 53 亿；随着技术的不断成熟，2023 年全球连接至互联网的设备可达 293 亿，而全球互联网流量中有 82%是视频流量，因此多媒体业务将是互联网服务的主要关注点。十年前我们的信息传播与交流沟通主要依赖文字，近年来我们更频繁地用视频、图片等传递情感，而在不远的将来，基于多媒体通信服务毫无疑问将成为人类传递与表达的常用方式。

多媒体业务质量提升与分辨率、对比度、色域与色位、帧率等指标紧密相关，而视频分辨率的提升有目共睹。从 20 世纪的 CIF、后续的 4CIF 到十多年前的 720P 与 1080P、2K，再到日趋普及的 4K 与逐渐成熟的 8K，分辨率的提升意味着画面的极致体验，同时也带来了数据量的成倍增长；除了视频分辨率、画质的不断提升，传统的 8-bit RGB 已无法满足需求，诸如 10-bit RGB、12-bit RGB、16bit-RGB 等可保存更多色彩信息的色位方案不断涌现，随之而来的数据量在分辨率提升的基础上进一步增长；高帧率更是绝佳视频体验的

必备元素，传统电影大多使用 24 帧拍摄完成，电视则多使用 30 帧；而在未来诸如 60 帧、120 帧甚至 160 帧等会成为媒体行业的标准配置。

1.2 多媒体质量评价的需求

多媒体业务的不断发展也对通信网络及终端设备提出了越来越高的需求。无线网络从 3G 到 4G，再到 5G，6G。如图 1.1 所示，随着通信技术的演进、硬件设备的更迭，网络的峰值传输速率不断提升，需要提供更大的带宽满足海量数据传输需求。此外引入了网络切片概念，网络切片是指在网络硬件条件下，切出不同的功能网络，应用于不同的场景，它们之间又是相互独立的。网络切片根据每个客户的请求进行差异化处理。通过切片，网络运营商可以将客户视为不同的用户类型，每个用户具有不同的服务请求，根据服务等级协议管理每个用户使用的切片类型和业务，从而满足用户多元化的应用需求。终端设备也在飞速发展，通过更高速的中央处理器、更大的存储空间、更优化的操作系统，尽量满足不同业务类型的使用需求。

总而言之，优异的音视频服务体验正在推动业务、网络、终端的快速发展，并且已成为海量数据传输的基础保障。

在这些技术的发展过程中，对于多媒体质量评价的需求应运而生。基于多媒体服务质量的评价结果，有依据地进行系统升级、网络建设等活动，在避免资源过度浪费的同时保证了提供的服务质量。然而，随着质量评价技术的发展，传统质量评价方法的问题日益凸显：

（1）质量评价是割离进行的，如业务评价仅对多媒体指标进行采集及评价，而忽略了网络的影响；

（2）质量评价停留在服务提供侧，如网络质量评价是基于网络侧获取的指标参数进行的，而不考虑用户侧实际的感受；

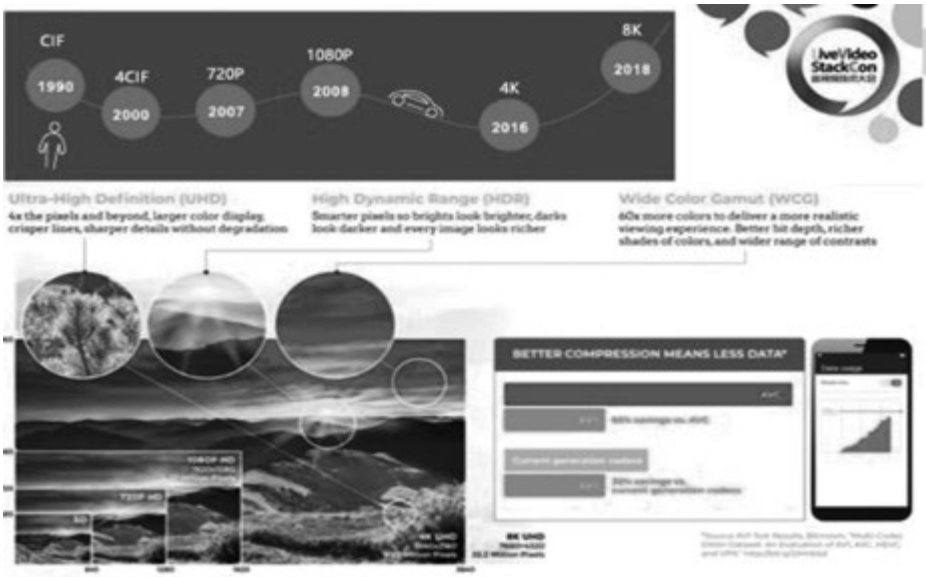

图 1.1　多媒体服务质量需求逐年提高

（3）质量评价属于静态方法，如上述基于网络侧指标参数的网络质量评价，拟合的质量评价体系基于历史统计结果，无法或者难以反映服务需求的动态变化；

（4）落后的用户体验采集方法，如通过传统的问卷调查方式采集用户评价，由于人群的离散性难以获得真实且收敛的调查结果。

基于上述传统质量评价方法的问题，我们可以看到有多个因素影响用户对于服务的体验，分为三大类：用户状态、系统能力和环境因素。因此，工业界提出以用户为中心、以用户获得的主观体验进行质量评价的方法。这样上述几个因素均可以对评价结果产生影响。对于如何获取用户的真实感知、保存量、提增量，亟须引入新方法，从而为多媒体质量提升提供更有价值的数据。此外，应在前期完成平台建设、模块植入，实现大数据分析基本模型的建立，提高用户满意度。

1.3 研究现状

用户体验质量最早被广泛提出是在20世纪90年代中期，由用户体验设计师唐纳德·诺曼(Donald Norman)提出和推广。ISO 9241—210标准将用户体验定义为“人们对于针对使用或期望使用的产品、系统或者服务的认知印象和回应”。因此，用户体验是主观的，且注重实际应用时产生的效果。

多媒体质量评价体系的业务类型可以包括音频、视频/图像、数据，以及新兴的媒体，这些领域都有质量评价的研究。这里挑选当前主流的音频和视频/图像领域的质量评价技术的现状进行介绍。

1.3.1 音频质量评价

音频质量评价方法可以分为主观评价和客观评价两种，早期的音频质量评价采用人工主观打分的方法得到平均意见分(mean opinion score，MOS)，后逐渐发展到基于人耳听觉模型的客观质量评价。

1. 主观评价

• 以人为主体进行音频质量评价，由参与测试的受测试者根据既定的评价准则对音频质量进行打分，它反映了人对音频质量的一种主观的评判。主观评价方法受人的个体影响比较大，为了排除偶然因素，减少评价结果的波动，需要参与评价的测试者数量较多(一般40人以上)，耗时、昂贵，可重复性低。但是由于人是音频的最终接收者，这种评价方法是人对音频质量感受的真实反映。

• 目前使用较多的主观评价方法包括：韵母可懂度测量(diagnostic

rhyme tests，DRT）、MOS、失真平均意见分（degradation mean opinion score，DMOS）、满意度测量（diagnostic acceptability measure，DAM）等。

• 在实际音频质量评价应用中，MOS 方法最常见。ITU P. 800 标准定义了 MOS 评价方法。参加评价的测试者会从表 1. 1 中的五个等级中选择一个作为他对被测试音频质量的评价结果。所有测试者的平均分结果就是被测音频的质量 MOS 值。

表 1. 1　平均主观得分概要

MOS	质　　量	观　　感
5	优异的	无法感知区别
4	好的	可以感知区别但不影响观看
3	一般的	轻微影响观看
2	较差的	影响观看
1	非常差的	严重影响观看

2. 客观评价

• 以机器处理自动判别音频质量。从操作方式上又可分为两类：基于输入输出方式的主动式评价和基于输出方式的被动式评价。主动式评价是基于原始音频信号和劣化音频信号的误差对比，一般采用数值误差距离或者描述人耳听觉系统对声音的感知听觉模型来量化音频的质量；被动式评价仅以音频经过被测系统后的输出信号来评价劣化音频的质量。

• 主动式评价方法主要由 ITU 标准组织定义。如图 1. 2 所示，其中 PESQ 和 POLQA 是目前广泛使用的音频质量评价方法。PSQM 由于种种缺陷，目前在实际中已经很少采用。从主动式评价方法的发展历程及各种方法的性能来看，一些方法与主观评价的相关度已达到了 0. 95 左右，如 POLQA，所以可以比较真实地评价音频经过被测系统后的体验质量。

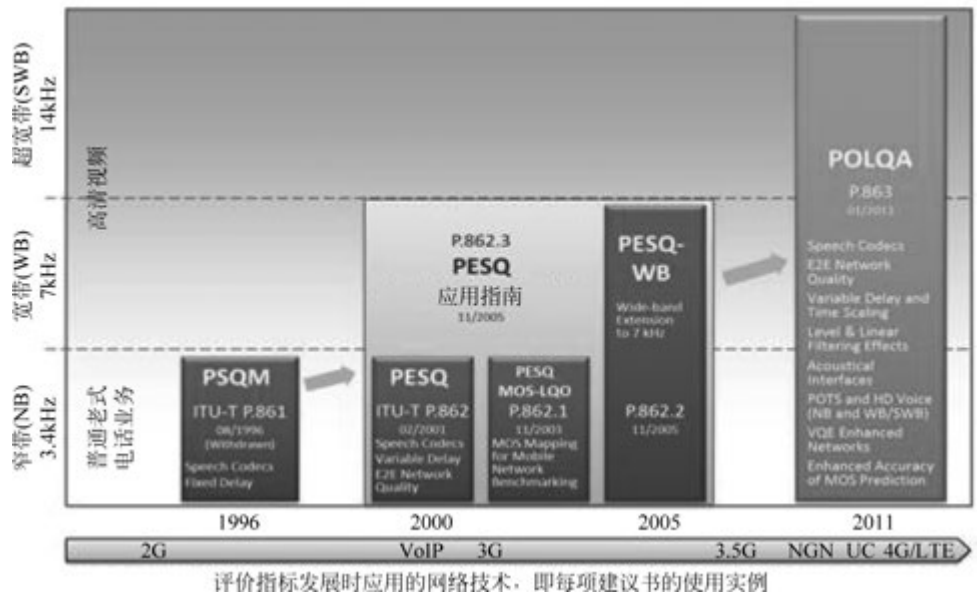

图 1.2 感知语音质量评价标准演进

• 图 1.2 中的 PESQ(perceptual evaluation of speech quality)由英国电信和 KPN 共同开发，并在 2001 年被 ITU 采纳为 P.862 规范。它比较参考音频信号和劣化信号并基于人耳听力模型给出一个评价的 MOS 值。PESQ 既能测试解码器这样的信号处理单元的性能，也能测量端到端的声音传输质量。针对不同的信号劣化原因，如编解码失真、误码、丢包、延时、抖动和滤波，给出一个－0.5～4.5 范围的值。多数情况下正常的分值范围在 2～4.5。PESQ 的思路是对参考音频信号和经过被测系统的劣化音频信号进行电平调整到标准听觉电平，再用输入滤波器模拟标准电话听筒进行滤波。对通过电平调整和滤波后的两个信号进行时间同步，并进行听觉变换，包括对系统中线性滤波和增益变化的补偿和均衡。听觉变换后的信号之间的误差作为扰动，分析扰动曲面并提取出两个失真参数，在频率和时间维度进行累积，然后映射到主观平均意见分的预测值。PESQ 算法将话音的频率、响度等物理特性与人主观的感知特性的对应关系用客观的数学模型来表示。算法中采用时频映射、频

率弯折和响度弯折等方法，尽可能将音频中可以被人耳感知的特性在数学上充分地表示。在 PESQ 模型中，提取出的语音特征都是与人的主观感觉直接相关的。当然，PESQ 算法也有一些缺点，如处理 CDMA 编码时结果不够准确，在特定的 GSM/WCDMA 网络制式下比较敏感；此外由于 PESQ 提出的比较早，所以它不能处理超宽带的音频信号。

• POLQA（perceptual objective listening quality analysis）是在 PESQ 之后的新一代音频质量评价标准，适用于固网、移动通信网络和 IP 网络中的语音质量评价。POLQA 的中文含义是“客观听觉品质感知评价”。它被 ITU-T 确定为推荐规范 P.863，可支持高清语音、3G、4G/VoLTE、5G 网络语音质量评价。POLQA 标准的开发始于 2006 年，并于 2011 年正式发布，由 ITU-T 第 SG 12 工作组组织研究，用以替代和升级 PESQ。从应用测试可以看到，POLQA 系统具有非常高的准确性，特别是 POLQA 增加了对宽带和超宽带音频信号的质量评价的能力，同时它可以支持最新的一些语音编码和 VoIP 传输技术，因此 POLQA 可以作为针对下一代网络质量评价、优化和监控解决方案的重要技术之一。POLQA 由 OPTICOM、SwissQual 和 TNO 联合研发，由 OPTICOM 及其合作伙伴负责在全球进行销售和技术支持。

• 被动式评价方法主要有两种。一种是基于相关的网络损伤参数（如丢包、抖动和延迟）预测传输的音频质量；另一种是根据音频信号测量语音流量的语音质量（如编解码器、回声、语言/讲话人），如 ITU P.563 标准。被动式评价方法的原理是建立音频质量和网络或音频流相关影响参数的关系，从而通过这些参数的动态状态估计被测音频的质量。被动式评价不是直接测量用户体验，而是使用相关的网络参数或语音参数通过预先训练的数学模型预测音频的质量。

1.3.2 视频/图像质量评价

视频/图像质量评价对视频/图像经过系统后的变化与劣化进行衡量与评价。视频/图像质量评价主要包括逼真度与可懂度，逼真度是指被评价图像与标准图像的偏离程度，可懂度是指图像能向人提供正确信息的能力。评价方法分为主观评价方法和客观评价方法两种。

主观评价一般用 MOS 或平均主观得分差（difference mean opinion score，DMOS）来表示。MOS 是通过测试者对于一个视频/图像质量的主观打分的平均分值来评价，DMOS 是通过测试者对原始参考视频/图像和经过被测系统后的劣化视频/图像的评分差异再归一化的分值来评价。

客观评价指机器根据算法计算出视频/图像的质量指标，让机器从人的主观视角出发来预测视频/图像的评分。根据主观感受的标尺，不同客观评价指标的优劣基于预测的准确性、一致性、稳定性和单调性来衡量。准确性是指主观评价打分和客观评价指标分数的相似性；一致性描述指标的泛化性，应该对所有类型的视频/图像都可以表现良好；稳定性是指对同一视频/图像每次评价的结果数值的偏差在可接收的范围内；单调性是指评价分数应该随 MOS 的增减呈现相应的增减。衡量客观评价方法的指标是通过客观评价模型输出 QR 与主观 MOS 的非线性拟合后变化为 MOS_P，准确性体现在 MOS 与 MOS_P 的 Pearson 线性相关系数，一致性体现在 MOS_P 的离群率，稳定性体现在每次相同输入后输出非线性拟合得到的 MOS_P 间误差，单调性体现在 MOS 与 MOS_P 之间的 Spearman 阶相关系数。客观评价方法的应用场景也比较丰富，可以衡量编解码算法及其软硬件实现的优劣；可以评价视频/图像经过通信系统后的损伤来衡量通信系统的优劣；可以衡量图像增强、图像重建算法的优劣；评价方法的结果还可以反馈给信源端的编码器，为下一步的参数设置提供依据，进而有针对性地对编解码损耗、通信传输过程

中的损伤进行参数优化与重新配置，从而在质量评价和参数优化形成闭环。

客观评价指标分为三类：基于误差的评价指标、基于感知模型与图像结构信息的评价指标及基于机器学习的评价指标。基于误差的评价指标将压缩图像和原始图像进行对比，计算两个图像之间的差异，代表指标是均方误差（mean square error，MSE）、峰值信噪比（peak signal noise ratio，PSNR）。基于感知模型与图像结构的评价指标通过引入人类视觉模型（human visual system，HVS）将图像质量劣化转为感知结构信息的变化和一些感知现象（亮度、对比度、观看距离等）的变化，对人类如何感知这些误差进行人眼视觉建模，代表指标是结构相似度指数（structure similarity index measure，SSIM）、恰可识别阈值（just noticeable difference，JND）。基于机器学习的评价指标一般用来度量长视频，从可训练的智能模型开始，将基于误差的评价指标或基于感知的评价指标结果与主观 MOS 进行比较，通过模型训练使其随时间推移而改善，代表性的评价指标是视频多评价方法融合（visual multimethod assessment fusion，VMAF）。此外，从对参考视频/图像的依赖上，还可以分为全参考方法（full reference，FR）、半参考/部分参考方法（reduced reference，RR）和无参考方法（no reference，NR）三类。全参考方法需要提供原始图像，经过对二者差异的学习得到对经过被测系统的劣化图像的评价，上文所提的 MSE、PSNR、SSIM 及视觉信息保真度（visual information fidelity，VIF）、视觉信噪比（visual signal to noise ratio，VSPR）、最显著失真（most apparent distortion，MAD）、图像差异预测（image difference prediction，IDP）等都是全参考方法。半参考/部分参考方法是指参考的不是原视频/图像，而是从原视频/图像中提取的某些特征或添加的信息，并通过无损伤的辅助信道传至信宿，对经过被测系统的视频/图像进行特征提取，分析这些特征信息的损耗程度，进而反映视频/图像的质量。典型的评价方法有：基于特征提取的方法、基于谐波强度的方法及基于小波域统计模型的方法。无参考方法指完全没有原视频/图像，将质量因素分解为某类失真或噪声，然后建立相应的评价模型。

由于视频应用的爆发增长及其参考源难以获取的特点，无参考质量评价方法成为近年来的研究热点，时域 ITU-P. 910 和空间域 ITU-P. 910 是无参考方法。

1. 主观指标

1) ITU-R BT. 500

(1) 双刺激损伤尺度(double stimulus impairment scale，DSIS)：测试者观看多个原始参考视频和劣化视频组成的“视频对”，并且每次总是先观看原始参考视频，然后观看劣化视频。测试者对视频的质量印象进行评分，评分采用 5 分制失真测度。

(2) 双刺激连续质量尺度(double stimulus continuous quality scale，DSCQS)：测试者观看多个原始参考视频和劣化视频组成的“视频对”，但与 DSIS 不同的是，原始参考视频和劣化视频的显示顺序是随机的，并且测试者对每个“视频对”中两幅视频的质量都进行打分。为了避免量化误差，该方法提供了一个连续的评分测度，但是为了与 5 分制的评分标准一致，也被等分成 5 份。测试过程中，首先将测试视频对显示一次或多次，使测试者建立对视频材料的主观认识，然后再一次或多次显示视频对以记录评分。

(3) 单刺激(single stimulus，SS)：以随机的形式显示多个测试视频，并且对于不同测试者，视频序列的随机显示顺序不同。测试者只观看测试视频，对其质量进行打分。具体实现方式有两种：一种是 SS(single stimulus)，即不重复放映测试序列；另外一种是 SSMR(single stimulus with multiple repetition)，即把测试序列重复放映多次。最常用的质量评分测度是 5 分制，除此之外还有 9 分制和 11 分制，它们是 5 分制的扩展，可以提高评分的精度。

(4) 单刺激连续质量评价(single stimulus continuous quality evaluation，SSCQE)：只显示测试序列，与上述几种采用较短独立序列段进行测试的方法不同，该方法选择的序列段持续的时间较长。测试者持续对观测序列进行评