通常，语音质量的评价标准可分为两大类：主观测量和客观测量。前者是建立在人的主观感受上的；而后者主要包括一些客观的物理量，如信噪比等。

**一、 主观评价**

主观评价是以人为主体来评价语音的质量，是人对语音质量的真实反映。语音主观评价方法种类很多，主要指标包括清晰度或可懂度和音质两类。清晰度一般是针对音节以下（如音素，声母、韵母）语音测试单元，可懂度则是针对音节以上（如词，句）语音测试单元的；音质则是指语音听起来的自然度。前者是衡量语音中的字、单词和句的可懂程度，而后者则是对讲话人的辨识水平。这两种不是完全独立的两个概念。一个编码器有可能生成高清晰度的语音但音质很差，声音听起来就像是机器发生的，无法辨别出说话者。当然，一个不清晰的语音是不可能成为高音质的。此外，很悦耳的声音也有可能听起来很模糊。

无论哪种主观测试都是建立在人的感觉基础上的，测试结果很可能因人而异。因此，主观测试的方案设计必须十分周密。同时，为了消除个体的差异性，测试环境应尽可能相同，测试语音的样本也要尽量丰富。每种语音的测试都必须仔细地选择发音，以保证所选的样本具有代表性，同时还要保证能够覆盖各种类型的话音。在选择测试者时，不仅应该包括女声，男声，同时还应根据年龄（包括老人，青年和儿童）选择不同语音。主观评价的优点是直接易于理解，能真实反映人对语音质量的实际感觉，缺点是需要大量的测试者，实施起来比较麻烦，耗时耗力，灵活性差。

1、可懂度评价（Diagnostic Rhyme Test，简称为DRT得分）

DRT是衡量通信系统可懂度的ANSI标准之一，它主要用于低速率语音编码的质量测试。这种测试方法使用若干对（通常96对）同韵母单字或单音节词进行测试，例如中文的“为”和“费”，英文的“veal”和“feel”等。测试中，评听人每次听一对韵字中的某个音，然后判断所听到的音是哪个字，全体评听人判断正确的百分比就是DRT得分。通常认为DRT为95%以上时清晰度为优，85%-94%为良，75%-84%为中，65%-75%为差，而65%以下为不可接受。在实际通信中，清晰度为50%时，整句的可懂度大约为80%，这是因为整句中具有较高的冗余度，即使个别字听不清楚，人们也能理解整句话的意思。当清晰度为90%时，整句话的可懂度已经接近100%。

在DRT测试中，一个重要问题是发音者。众所周知，男性和女性的发音是不同的，一般来说后者要清晰一些。此外，从实际耗费的角度出发，发音者不能太多。根据经验，一般情况下，DRT测试要求三位男性和三位女性。国外著名的Dynastant公司专门从事语音测试，英语测试中最常用的96对测试表就是由该公司提出的。由专业公司测试的好处就在于他们能够提供经过训练的专业测试者，专业的测试环境，从而获得较为准确和公正的测试结果。

但是，DRT也有局限性，因为其只测试第一辅音，并且每次的选择只有两个。在这种情况下，Dynastant公司提出了更为复杂的改进型韵字测试MRT（Modified Rhyme Test）。MRT的基本测试方法和DRT一样，但是其测试语音样本中，不同的辅音不仅可能出现在第一位，也可能在最后一位；而且，每次的选择增加到六个。

2、音质评价

（1）平均意见得分（Mean Opinion Score，简称为MOS得分）

MOS得分法是从绝对等级评价法发展而来的，用于对语音整体满意度或语音通信系统质量进行评价。MOS得分法一般采用5级评分标准，包括优、良、中、差和劣。参加测试的评听人在听完受测语音后，从这5个等级中选择一级作为所测语音的MOS得分。由于主观上和客观上的种种原因，每次测试得到的MOS得分大都会有波动，为了减小波动的方差，除了参加测试的评听人要足够多之外（一般至少40人），所测语音材料也应足够丰富，测试环境也要尽量保持相同。在数字通信系统中，通常认为MOS得分在4.0～4.5分为高质量数字化语音，达到长途电话网的质量要求，接近于透明信道编码。MOS得分在3.5左右称作通信质量，此时重建话音质量下降，但不妨碍正常通话，可以满足话音系统使用要求。MOS得分在3.0以下常称合成语音质量，它一般具有足够的可懂度，但自然度和讲话人的确认等方面不够好。表5-1表示MOS分制的评分标准。极好的语音音质表示所测信号与原始语音相近，没有感知噪声；相反，极差音质表示有非常厌烦的噪声且所测信号有人为噪声。

表5-1 MOS分制的评分标准

|  |  |  |
| --- | --- | --- |
| 得分 | 质量级别 | 失真级别 |
| 5 | 优（excellent） | 不察觉 |
| 4 | 良（good） | 刚有察觉，但不可厌 |
| 3 | 中（fair） | 有察觉且稍觉可厌 |
| 2 | 差（poor） | 明显察觉且可厌但可忍受 |
| 1 | 劣（bad） | 非常可厌，不可忍受 |

（2）判断满意度测量（Diagnostic Acceptability Measure，简称为DAM得分）

DAM方法是由Dynastat公司推出的一种评价语音通信系统和通信连接的主观语音质量和满意度的评测方法，其将直接途径与间接途径结合在一起进行主观质量评价。评听人既有机会表达个人主观喜好，又能依标准对每项指标进行评测。另外，DAM方法要求评听人分别对语音样本本身、背景和其它因素进行评价。一个评听人可将评价过程划分为21个等级，其中10个等级是信号的感觉质量，8个等级是背景情况，另外3级是可懂度、清晰度和总体满意度。总之，DAM是对话音质量的综合评价，是在多种条件下对话音质量可接受程度的一种度量，以百分比评分。

总结来说，无论哪种主观测试，都需要遵循三个原则：第一，要保证足够的说话者，要求其声音特征非常丰富，能够代表实际用户中的绝大部分；第二，要求有足够多的数据。理论上，人数和数据越多越好，可以用方差作为判断样本数的尺度；第三，对于大部分编码器来说，清晰度和品质测试应该都做。但很悦耳的质量较好的语音也可以不做清晰度测试。

二、客观评价

针对主观评价方法的不足，基于客观测度的语音客观评价方法相继被提出。客观评价必然要借鉴主观评价的那种高度智能和人性化的过程，但是不可能找到一个绝对完善的测度和十分理想的测试方法，只能尽量利用所获信息做出基本正确的评价。一般地，一种客观测度的优劣取决于它与主观评价结果的统计意义上的相关程度。目前所用的客观测度分为时域测度、频域测度和在两者基础上发展起来的其它测度。主要的客观评价方法有：基于信噪比的评价方法，如信噪比（signal-to-noise，SNR）、分段信噪比（segmental SNR，segSNR）等，把信噪比作为评价语音质量的指标；基于谱距离的评价方法，如加权谱斜率测度（weighted spectral slope measure，WSS），主要比较语音信号之间的平滑谱；基于听觉模型的评价方法，如语音质量感知评价方法（perceptual evaluation of speech quality，PESQ），以人对语音的感知特性为基础。

（1）信噪比SNR

信噪比计算简单，是一种应用广泛的客观评价方法。假设为带噪语音信号，为其中的纯净语音信号，为经处理后的语音信号，则信噪比定义为：

 (5-2)

由于计算时需要纯净的语音信号，而实际环境中难以获得纯净的语音信号，因此信噪比主要用在纯净语音信号已知的实验仿真中。

（2）分段信噪比

经典形式的信噪比同等对待时域波形中的所有误差，不能很好地反映语音质量的属性。由于语音信号的时变特性，不同时间段上的信噪比应该是不一样的。由此，出现了分段信噪比。它的定义如下：

 (5-3)

其中，表示语音的帧数，是语音帧长度，表示语音帧的起始点。从公式（5-3）可以看出，分段信噪比先计算每一帧的信噪比，再对所有帧的信噪比取平均。为了减小没有语音的帧和信噪比过高的帧对信噪比带来的影响，一般设置两个门限值，如高低门限分别设为35dB和0dB，不在此范围内的信噪比都置为门限值。

（3）加权谱斜率测度WSS

WSS使用36个临界频带滤波器来计算，反映纯净语音和处理后语音的频带谱斜率间的加权差距，WSS距离越小，表示两者之间的差距越小，语音质量越好。

令、分别表示纯净语音和处理后语音的谱斜率，其定义为

 (5-4)

其中，、分别表示纯净语音和处理后语音的第个临界频带谱。

令表示权重，其定义为：

 (5-5)

其中，为所有频带中最大的对数谱幅度，为最靠近第个频带的峰值，、为常数，用来使主观测试和客观指标有最大的相关性，根据经验分别取值为20和1。

最后，WSS距离的计算公式如下：

 (5-6)

（4）语音质量感知评价方法PESQ

PESQ方法是国际电信联盟ITU在2001年提出的一种新的语音质量评价方法，是目前与MOS评分相关度最高的客观语音质量评价算法，相关度系数达到0.97。该算法将参考语音信号和失真语音信号进行电平调整、输入滤波器滤波、时间对准和补偿、听觉变换之后，分别提取两路信号的参数，综合其时频特性，得到PESQ分数，最终将这个分数映射到主观平均意见分上。PESQ得分范围在-0.5—4.5之间，得分越高表示语音质量越好，PESQ方法模型如图5-1所示：



图5-1 PESQ方法模型图

客观评定方法的特点是计算简单，缺点是客观参数对增益和延迟都比较敏感，而且最重要的是，客观参数没有考虑人耳的听觉特性，因此客观评定方法主要适用于速率较高的波形编码类型的算法。而对于低于16kbps的语音编码质量的评价通常采用主观评定的方法，因为主观评定方法符合人类听话时对语音质量的感觉，因此主观评估参数就显得非常重要，特别是许多低码率算法都是基于人耳的感知标准设计的，故而应用较广。总结起来，语音主观评价和客观评价各有其优缺点。通常这两种方法应该结合起来使用。一般的原则是，客观评价用于系统的设计阶段，以提供参数调整方面的信息，主观评价用于实际听觉效果的检验。