ITU出版物
建议书

国际电信联盟
无线电通信部门

ITU-R BS.1387-2 建议书

(05/2023)

BS系列：广播业务（声音）

感知音频质量的客观测量方法

前言

无线电通信部门的作用是确保所有无线电通信业务，包括卫星业务，合理、公平、有效和经济地使用无线电频谱，并开展没有频率范围限制的研究，在此基础上通过建议书。

无线电通信部门制定规章制度和政策的职能由世界和区域无线电通信大会以及无线电通信全会完成，并得到各研究组的支持。

知识产权政策（IPR）

国际电联无线电通信部门（ITU-R）的IPR政策述于ITU-R第1号决议所参引的《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/zh 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 建议书系列

（可同时在以下网址获得：http://www.itu.int/publ/R-REC/zh）

系列	标题
BO	卫星传输
BR	用于制作、存档和播放的记录；用于电视的胶片
BS	**广播业务（声音）**
BT	广播业务（电视）
F	固定业务
M	移动、无线电测定、业余无线电以及相关卫星业务
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	卫星固定和固定业务系统之间频率共用和协调
SM	频谱管理
SNG	卫星新闻采集
TF	时间信号和标准频率发射
V	词汇和相关课题

注：本ITU-R建议书英文版已按ITU-R第1号决议规定的程序批准。

电子出版物
2024年，日内瓦

ITU-R BS.1387-2建议书

感知音频质量的客观测量方法

（1998-2001-2023年）

范围

本建议书规定了感知音频质量的客观测量方法。

关键词

感知音频质量、客观测量、低比特率编码

国际电联无线电通信全会，

考虑到

a) 传统的客观方法（例如用于测量信噪比和失真）不再适合于测量使用低比特率编码方案或采用模拟或数字信号处理的系统的感知音频质量；

b) 低比特率编码方案正在迅速部署；

c) 并非所有符合规范或标准的实施均保证该规范或标准可实现的最佳质量；

d) 正式的主观评估方法不适合连续监控音频质量，例如在操作条件下；

e) 感知音频质量的客观测量最终可以在所有测量领域补充或取代传统的客观测试方法；

f) 感知音频质量的客观测量可以有效地补充主观评估方法；

g) 对于一些应用来说，可以实时实施的方法是必要的，

做出建议

1 对于附件1中列出的每个应用，将附件2中给出的方法用于感知音频质量的客观测量。

前言

本建议书规定了待测设备（如低比特率编解码器）感知音频质量的客观测量方法，其中包括两个附件。附件1为用户提供了该方法的概述，并包括四份后附资料。后附资料1描述了应用和测试信号。后附资料2列出了模型输出变量，并讨论了使用和精度方面的限制。后附资料3给出了模型的概要，而后附资料4描述了客观感知音频质量测量方法的原则和特性。

附件2为实施者提供了使用两个版本的心理声学模型的方法的详细说明，这两个版本是在组合六个模型的集成阶段开发的。附件2的后附资料1中描述了客观测量方法的验证过程。附件2的后附资料2概述了方法开发和验证中使用的所有数据库。

目录

附件1

概述

1 引言

音频质量是设计数字广播系统的关键因素之一。各种比特率降低方案的快速引入推动了在建立和完善主观评估程序方面的重大工作，而这仅仅是因为正式的听力测试一直是判断音频质量的唯一相关方法。在此方面所获得的经验为ITU-R BS.1116建议书奠定了基础，而该建议书在此后成为大多数此类听力测试的依据之所在。

由于主观质量评估既耗时又昂贵，因此最好开发一种客观测量方法来对音频质量进行估算。如信噪比（S/N）或总谐波失真（THD）等传统的客观测量方法从未真正被证明与感知的音频质量可靠相关。当这些方法应用于非线性和非平稳的现代编解码器时，问题变得更加明显。

在过去的十年中，已经引入了许多对感知音频质量进行客观感知测量的方法。但是，这些方法均未经过彻底的验证，因此既没有标准化亦没有被广泛接受。1994年，国际电联无线电通信部门（ITU-R）明确表示亟需在此领域制定标准，并启动了相关工作。为此，发出了公开征集建议的呼吁，并收到了以下六种备选测量方法：干扰指数（DIX）、噪声掩蔽比（NMR）、感知音频质量测量（PAQM）、感知评估（PERCEVAL）、感知客观测量（POM）和工具箱方法。附件1的后附资料4中描述了这些方法。

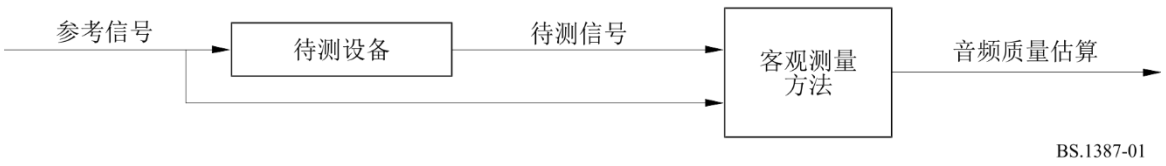
本建议书中的测量方法是一个过程的结果，在此过程中，对上述六种方法的性能进行了研究，并提取了最有前途的工具，以将其整合为一种方法。所所建议的方法已经在多个测试点得到了仔细验证。事实证明，此方法可以为多种应用生成可靠而有用的信息。不过，必须牢记，本建议书中的客观测量方法通常不能替代正式听力测试。

2 应用

使用建议方法进行客观测量的基本概念如下图1所示。

图1

进行客观测量的基本概念



本建议书中的测量方法适用于大多数类型的数字和模拟音频信号处理设备。不过，预计许多应用将侧重采用音频编解码器。

已经确定了以下八类应用：

表1

应用

	应用	简述	版本
1	实施情况评估	描述音频处理设备（在许多情况下是音频编解码器）不同实施的程序	基本/高级
2	感知质量排列	在设备或电路投入使用之前进行的快速程序	基本
3	在线监控	监控业务中音频传输的连续程序	基本
4	设备或连接状态	对设备或电路的详细分析	高级
5	编解码器识别	识别特定编解码器的类型和实施的程序	高级
6	编解码器开发	尽可能详细地描述编解码器性能的程序	基本/高级
7	网络规划	在给定约束条件下优化传输网络成本和性能的程序	基本/高级
8	主观评估辅助	一种筛选听力测试中关键材料的工具	基本/高级

3 版本

为了达到不同成本和性能要求的最佳匹配，本建议书中建议的客观测量方法有两个版本。基本版本旨在实现经济高效的实时实施，而高级版本则侧重于实现尽可能高的精度。与基本版本相比，高级版本的额外精度使复杂性增加了大约四倍，这具体取决于实施方式。

表1针对就每个应用采用何种版本给出了一些指南。

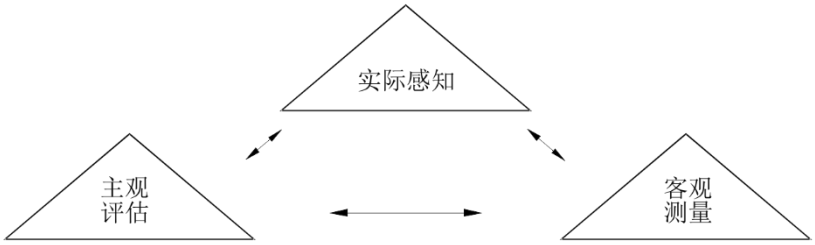
4 主观域

正式的主观听觉测试（例如基于ITU-R BS.1116建议书的测试）经过了精心设计，以尽可能接近对音频质量判断的可靠估算。不过，不能指望主观听力测试的结果来完全反映实际感受。图2说明了主观和客观领域中隐含的缺陷。

显然，不可能对一种客观方法进行直接验证。相反，客观测量方法是根据主观听力测试进行验证的。

图2

验证概念



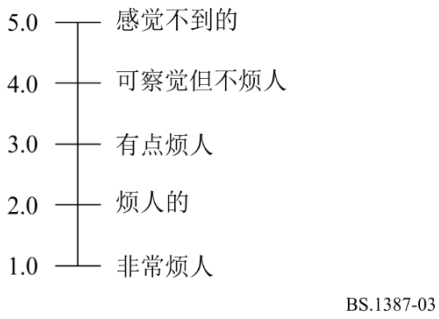
本建议书中的客观测量方法主要针对通常在主观域通过应用ITU-R BS.1116建议书进行评估的应用。该特定测试方法的基本原则可简述如下：听者可在三个来源之间进行选择（“A”、“B”和“C”）。已知的参考信号始终作为源“A”可用。隐藏参考信号和待测信号同时可用，但根据试验“随机”分配给“B”和“C”。

听者被要求根据连续五级损伤量表来评估“B”相对于“A”的损伤以及“C”相对于“A”的损伤。其中一个源“B”或“C”应该与源“A”区分开来；另一个源则可能显示损伤。参考数据和其他来源之间的任何感知差异均须解释为损伤。通常，只使用一个属性“基本音频质量”。它被定义为一个全局属性，其中包括参考信号和待测信号之间的任何及所有检测到的差异。

分级量表应被视为连续的“锚”，该锚源自ITU-R BS.1284建议书中给出的ITU-R五级损伤量表，如下所示。

图3

ITU-R五级损伤量表



主观听力测试结果的分析通常基于主观差异等级（SDG），该等级被定义为：

$$SDG = 等级_{待测信号} - 等级_{参考信号}$$

理想情况下，SDG值的范围应为0至−4，其中0对应于不易察觉的损伤，而–4对应于被判断为非常恼人的损伤。

5 分辨率和精度

客观差异等级（ODG）是客观测量方法的输出变量，与主观域的SDG相对应。ODG的分辨率被限制为一位小数。但是，对此应持谨慎态度，不要泛泛地认为任何一对ODG之间十分之一等级的差异会很大。当审视主观听力测试的结果时，上述评论亦有效。

并无任何一个单一的数字可以完全描述客观测量方法的精度。相反，必须考虑诸多不同的品质因数。其中之一是SDG和ODG之间的相关性。必须指出的是，不能保证相关性会超过预定义的值。例如，测量方法的性能很可能会随着引入的劣化的类型和水平而发生变化。

另一个相关品质因数是异常值的数量。异常值是指不符合预定义容限方案的测量值。根据用户要求，测量方法应为分级量表的上限（即高音频质量）提供尽可能高的精度。因此，在分级量表的中低范围内，所获得的精度可相对较低。

尽管相关性通常可以很好地估算客观测量方法的精度，但必需牢记，即使相对较高的相关性数字亦可能隐含着测量方法的不可接受的性能（从异常值的角度来看）。

在验证过程中使用的第三个品质因数是绝对误差分数（AES），它反映了SDG置信区间的大小与SDG和ODG间距之间的平均关系。

有关测量方法的预期性能以及验证过程中性能方面的更多详情，请参见附件2的后附资料1。

6 要求和限制

在整个测量间隔期间，来自待测设备的信号和参考信号必须以24个样本的精度进行时间对齐。同步机制不是本建议书的一部分，预计在不同的实施中会有所不同。

附件1的
后附资料1

应用

1 综述

本后附资料提供了主要应用的定义和具体要求，所建议的感知音频质量客观测量方法适用于这些应用。

一些应用需要客观测量方法的实时实施，而对于其他应用而言，非实时测量即已足够。对于实时实施，建议通过测量设备的最大延迟不超过200毫秒，超过1秒是不可接受的。

此外，必须区分在线和离线测量。在离线测量中，测量程序可以完全访问设备或连接，而在线测量意味着程序正在运行，不得被测量中断。

2 主要应用

2.1 实施情况评估

当选择要购买的设备或进行验收测试时，广播公司、网络运营商和其他各方需要评估设备的不同实施方式，尤其是音频编解码器。

对于此类应用，特别需要高精度来对小损伤加以评估，并对不同的实施进行正确排列。关于输出变量，一个简单的输出（如ODG）对用户来说即已足够，但音频编解码器的开发人员可以通过使用一组合适的模型输出变量（MOV）来进行更全面的分析。

两种模型版本均可使用，但建议使用高级版本。

2.2 感知质量排列

这是一个快速程序，发生在设备或电路投入使用之前。目的是检查功能和质量。测量设备将由操作人员操作。可能存在各种失真。

需要实时测量。可以使用测试信号或预定义的音频信号。ODG应正确显示，并应至少每秒显示两次，或者如果使用特殊测试信号，则应在测试信号结束后立即显示。

使用基本版本即已足够。

2.3 在线监控

这是一个持续的过程，发生在正在进行的音频传输期间。测量程序不得中断节目的传输。因此，必须使用节目信号本身或预定义的音频片段进行测量。后者可以是车站信号或叮当声。测量设备将由操作人员操作。

需要实时测量。ODG必须正确显示，应至少每秒显示两次或在预定义信号结束后立即显示。最好不要显示MOV。

使用基本版本即已足够。

2.4 设备或连接状态

为了确保音频连接或设备的功能，需要不时进行全面的质量检查。与在线监控或感知排列相比，此应用需要检查几个技术参数。

除了ODG之外，测量系统还应通过显示整套MOV来提供有关设备或连接状态对感知音频质量的影响的详情。不需要实时测量。

建议使用高级版本。

2.5 编解码器识别

为了识别编解码器（不同的算法或相同算法的不同实施），测量系统必须能够存储、检索和比较特性模式。模式之间的相似性可以作为不同编解码器实施方式的相似性指标。此程序用于识别特定编解码器的类型和实施方式。

测量系统必须记录尽可能多的模式信息。仅考虑ODG可能无法提供足够的信息。

使用基本版本即已足够，即使无需实时测量时亦是如此。

注 – 对所建议的方法只有很少的经验。此外，尚未定义模式之间相似性的单一指标。

2.6 编解码器的开发

对于此应用，测量方法必须尽可能准确且详细地描述待测编解码器的性能特征，对较小的损伤尤应如此。

连续监控测试需要实时处理，而高级版本不一定支持实时处理。不过，小的劣化和相关详情将需要高级版本。测量系统必须能够以与计算速率相同的速率显示输出。最好能够直接访问4秒内的输出历史记录。

建议使用高级版本。不过，对于实时测量，基本版本即已足够。需要进行实时、非实时和逐帧分析。必须标示各类严重的失真，例如通过峰值显示。最好能访问整套MOV。

2.7 网络规划

网络规划需要在规划过程的各个阶段评估预期质量。网络组件的软件模拟允许组合不同的音频处理阶段，并可用于检查不同的配置，以优化音频质量。在稍后阶段，可以在选定的配置中测试实际的音频处理组件。

网络规划由系统工程师完成，系统工程师应检索有关网络特性对音频质量影响的详情。不同的可能网络配置的分级应基于一组合适的MOV，具体取决于网络的具体应用。因此，仅显示ODG是不够的。在此应用中，评估不需要实时测量。

两种模型版本均可使用，但建议使用高级版本。

2.8 主观评估辅助

客观测量方法为筛选主观听力测试中使用的关键音频材料提供了工具。整套MOV可用于关键材料的分类。

须提供尽可能高的精度，建议使用高级版本。不过，为了减少选择关键材料所需的时间，最好进行实时测量。

2.9 应用概要

表2总结了主要应用对测量方法的要求。

表2

对测量方法的要求

	应用	类别	实时	最小ROV(1)（赫兹）	在线/离线	模型版本
1	实施情况评估	诊断	否	–	离线	两者
2	感知质量排列	操作	是/否	2	离线	基本
3	在线监控	操作	是	2	在线	基本
4	设备或连接状态	诊断	是/否	–	在线/离线	高级
5	编解码器识别	诊断	否	–	离线	两者
6	编解码器开发	开发	是/否	–	离线	两者
7	网络规划	开发	是/否	–	离线	两者
8	主观评估辅助	开发	是/否	–	离线	高级

(1) 输出值的速率（每秒）。

3 测试信号

测试信号可以分为两组：自然信号与合成信号。此处提供的自然测试信号列表包含ITU-R和其他组织为评估音频质量而执行的听力测试中已经使用的关键音频序列。信号必须在发射地点和测量地点均可用。因此，需要测量设备中的存储器。

合成信号是数学定义的，并且可以以受控的方式发生变化。此类信号可以在发射和测量地点产生。测量设备中不需要额外的存储器。由于此类信号的性质，很难（如果不是不可能的话）对其进行主观评级。因此，测量方法尚未针对此类信号的主观结果进行验证。

3.1 自然测试信号的选择

表3列出了本建议书验证程序中使用的测试信号子集，并列出了此类信号通常由于低比特率编码而暴露的伪像类型。

表3

包含测试信号子集的列表

编号	项目	文件名	评论
1	响板	cas	(1)
2	单簧管	cla	(2)
3	响棒	clv	(1)
4	长笛	flu	(2)
5	钟琴	glo	(1), (2), (5)
6	竖琴	hrp	(1), (2), (4)
7	定音鼓	ket	(1)
8	马林巴琴	mar	(1)
9	钢琴（舒伯特）	pia	(2)
10	律管	pip	(4)
11	莱 • 库德（音乐人）	ryc	(2), (4)
12	萨克斯管	sax	(2)
13	风笛	sb1	(2), (4), (5)
14	女性英语演讲	sfe	(3)
15	男性英语演讲	sme	(3)
16	男性德语演讲	smg	(3)
17	小军鼓	sna	(1)
18	女高音（莫扎特）	sop	(4)
19	铃鼓	tam	(1)
20	喇叭	tpt	(2)
21	三角琴	tri	(1), (2), (5)
22	大号	tub	(2)
23	苏珊 • 薇格（音乐人）	veg	(3), (4)
24	木琴	xyl	(1), (2)

(1) 瞬变：预回声敏感，时域中的噪声模糊。

(2) 音调结构：噪声敏感，粗糙。

(3) 自然语音（音调部分和攻击的关键组合）：失真敏感，攻击的模糊。

(4) 复杂声音：对待测设备施加压力。

(5) 高带宽：对待测设备、高频损耗、节目调制高频噪声施加压力。

3.2 持续时间

自然测试信号的持续时间应该与听力测试中使用的信号相同。持续时间通常在10到20秒的量级。很可能暴露大部分伪像的测试信号的关键部分仅局限于持续时间的较短部分。

合成测试信号的持续时间应该足够长，以对待测编解码器施加压力，该编解码器可能包含编码音频信号的缓冲区。考虑到测量方法中存在的这些缓冲长度和时间常数，序列中每个单个测试项目的持续时间应大于500毫秒。之所以可将持续时间限制为如此短的值，是因为预计此类信号不会用于主观听力测试。

4 同步

对于测量程序，待测信号和参考信号应在时间上同步。这适用于自然及合成测试信号。

5 版权问题

表3中给出的测试信号可以与本建议书附件2中描述的客观测量方法一起免费用于测量目的。

注 – 所有序列均须获得版权许可，主要来自欧洲广播联盟（EBU）（EBU声音质量评估材料（SQAM）光盘）。

附件1的
后附资料2

输出变量

1 引言

本建议书中描述的客观测量方法测量音频质量，并输出与感知音频质量相对应的值。该测量方法模拟了听觉系统的基本特性。几个中间阶段模拟生理和心理声学效应。

这些中间输出可用于表征伪像。这些参数被称为模型输出变量（MOV）。测量模型的最后阶段将MOV值组合在一起，生成一个输出值，该值直接对应于主观质量评估的预期结果。

2 模型输出变量

表4描述了用于预测客观差异等级的MOV。下标$_A$来自模型的滤波器组部分，下标$_B$来自模型的快速傅立叶变换（FFT）部分。可以仅从FFT部分（基本版本）或从FFT和滤波器组部分的组合（高级版本）来预测客观差异等级。平均总是随着时间进行。

3 基本音频质量

主观听力测试中最广为人知的参数是基本音频质量（BAQ）。BAQ以主观差异等级（SDG）来衡量，其计算方法是从主观测试中测试信号的等级中减去参考信号的等级[1]。SDG通常为负值。模型的相应输出参数称为客观差异等级（ODG）。MOV到ODG的映射基于大量可靠的测试项目，见附件2的后附资料2。

表4

模型输出变量的描述

模型输出变量	描述
$WinModDiff_B$	参考信号和待测信号之间调制（包络）的窗口平均差异
$AvgModDiff1_B$	平均调制差
$AvgModDiff2_B$	平均调制差，重点是引入的调制和调制变化，其中参考包含很少调制或不包含调制
$RmsModDiff_A$	调制差的均方根值
$RmsMissingComponents_A$	丢失频率分量的噪声响度的均方根值（用于RmsNoiseLoudAsymA）
$RmsNoiseLoud_B$	强调引入分量的平均噪声响度的均方根值
$RmsNoiseLoudAsym_A$	$RmsNoiseLoud_A + 0.5RmsMissingComponents_A$
$AvgLinDist_A$	平均线性失真指标
$BandwidthRef_B$	参考信号的带宽
$BandwidthTest_B$	待测设备输出信号的带宽
$TotNMR_B$	平均总噪声与掩蔽比的对数
$RelDistFrames_B$	至少一个频带包含显著噪声分量的帧的相对比例
$AvgSegmNMR_B$	噪声掩蔽比的分段平均对数
$MFPD_B$	低通滤波后的最大检测概率
ADB_B	平均失真块（=帧），取总失真与严重失真帧总数之比的对数
EHS_B	误差随时间变化的谐波结构

ODG是对应于主观感知质量的客观测量参数。由于听力测试中听者的任务是评估测试项目的BAQ，因此ODG亦是BAQ的一种测量方法。

[1] 见ITU-R BS.1116建议书。

4 编码余量

有望被证明有价值的另一个参数是编码余量（CM），这是一种描述无声伪像的方法。主观编码余量（SCM）可以通过放大伪像直到测试人员可以听到它们来评估。SCM描述了伪像可听度阈值的余量。

为了找到阈值，在听力测试期间必须放大或衰减伪像。一种合适的方法是差分法。时间同步原始信号和编码信号的差分信号被放大并加到原始信号上。可听度阈值的检测最好用强制选择方法来执行。SCM是通过对从测试人员处获得的放大或衰减阈值进行平均而获得的。负的CM值代表可听的人工噪声，而正的CM值代表不可听的人工噪声。与BAQ不同，编码余量是有关何时（在何种水平上）可听见伪像的指标，而非有关伪像恼人程度的指标。有关SCM测量方法的定义和验证见[Feiten，1997年]中的描述。

客观编码余量（OCM）亦来自MOV。目前，仅对几个测试项目的主观编码余量进行了评估。尚未针对本建议书中模型到OCM的映射展开研究。

5 用户要求

用户对测量方法输出变量的要求因应用而异。对于某些应用，例如编号2和3（见附件1的后附资料1），测量是操作程序的一部分。在这些情况下，对于没有深入了解测量技术的人来说，该方法的输出必需既易于读取又易于解释。实现此目的的最佳场景是：该方法只输出一个对应于感知音频质量的**单个值**。

这同样适用于其他应用，例如应用1和4。不过，对于此类应用以及应用5-8，更复杂的输出变量可能有利于对测量方法中的机制有更深入了解的用户。

附件1的
后附资料3

模型概述

根据ITU-R BS.1116建议书，在听力测试中获得音频测试项目的SDG，多个听者的平均SDG代表项目的主观质量。该项目可能包含不同类型的音频失真，因此质量变化会随着时间的推移而累积。因此，基于物理测量的SDG预测需要外周听觉系统的精确模型以及音频质量判断的认知方面。

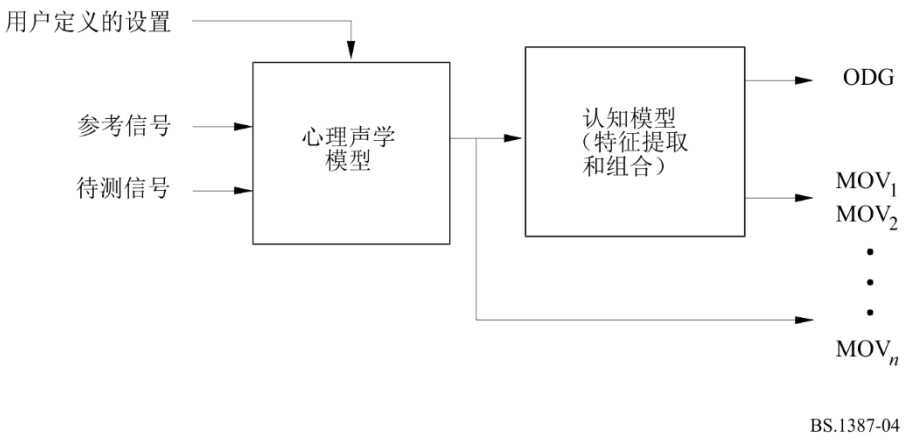
基于参考信号和待测信号之间的比较，用于客观测量的建议模型产生许多模型输出变量（MOV）。利用一种优化技术，这些MOV被映射到ODG，对于足够大的数据集，此优化技术可使ODG分布和相应的平均SDG分布之间的平方差最小化。

兹描述该模型的两种变体 — 可用于实时监控的基于DFT的版本，以及基于滤波器组和DFT的另一版本，该版本有望提供更准确的结果。基于DFT的版本称为基本版本，而组合版本称为高级版本。

基本版本和高级版本的高级结构如图4所示。

图4

模型中实施的处理阶段



1 音频处理

与主观听力测试一样，测试信号的质量是相对于参考信号来判断的。参考信号和待测信号（单声道或立体声信号）均被转换成心理声学表示。对这些表示进行比较，以得出ODG。这些操作由图4所示的处理阶段执行。

1.1 用户定义的设置

测量方法需要假设的听力水平作为参数。因此，用户必须提供由1 019.5赫兹的满量程正弦波产生的声压级（单位为dB SPL）。如果不知道确切的听力水平，则建议假设听力水平为92 dB SPL。

1.2 心理声学模型

心理声学模型将时域信号的连续帧转换为基膜表示。这个过程从使用DFT和滤波器组开始。DFT将数据转换到频域，并将结果从频率标度映射到音高标度，即心理声学的频率等效标度。在模型的滤波器组部分中，带通滤波器的带宽和间距直接考虑了频率到音高的映射。

使用两种不同的概念来实现同时掩蔽。一些MOV是使用*掩蔽阈值概念*计算的，而另一些则是基于*内部表示的比较*。第一个概念使用心理-物理掩蔽函数直接计算掩蔽阈值。模型输出变量基于物理误差信号到该掩蔽阈值的距离。在内部表示的比较中，待测信号（SUT）和参考信号的能量均被扩展到相邻的音调区域，以便获得激发模式。模型输出变量基于这些激发模式之间的比较。非同时掩蔽是通过随时间将信号表示加以模糊来实施的。

绝对阈值部分通过应用依赖于频率的加权函数来建模，部分通过向激发模式添加依赖于频率的偏移来建模。该阈值是最小可听压力的近似值[ISO 389-7，声学 — 听力设备校准的参考零点 — 第7部分：自由场和扩散场听力条件下的听力参考阈值，1996年]。

心理声学模型的主要输出是作为时间和频率函数的激发和掩蔽阈值。模型在几个级别上的输出可用于进一步处理。

1.3 认知模型

认知模型从心理声学模型产生的一系列帧中浓缩信息。进行质量测量的最重要的信息源是参考信号和待测信号在频域和音高域中的差异。在频域中，测量两个信号的频谱带宽以及误差中的谐波结构。在音高域中，误差测量从激发包络调制和激发幅度中导出。

对计算出的特征进行加权，使得它们的组合产生一个ODG，该ODG足够接近相关特定音频失真的SDG。基本版本使用11个特征来产生ODG，而高级版本则使用5个特征。使用反向传播神经网络学习算法进行优化（见附件2，第6节）。训练数据由数据库1和2的全部以及数据库3的一部分组成。泛化测试数据从数据库3的其余部分和所有的CRC97数据集获得（见附件2的后附资料2）。

附件1的
后附资料4

客观感知音频质量测量方法的原则和特性

1 引言和历史

音频信号的数字传输和存储正日益基于数据简化算法，该算法适应人类听觉系统的特性，尤其依赖于掩蔽效应。此类算法的主要目的不是使失真最小化，而是试图以尽可能少被察觉的方式处理这些失真。这些感知编码器的质量不能再通过传统的测量方法来加以评估，传统的测量方法通常确定失真的总体值。说明这些限制的常见例子之一是所谓的13 dB奇迹：频谱结构与音频信号相适应的叠加噪声几乎听不见，即使最终的未加权信噪比下降到13 dB亦是如此。

因此，为了对感知编解码器进行评估，需要进行听力测试，以评估音频质量。听力测试的充分可靠性和可重复性亦需要大量的时间和工作。

结合人类听觉系统特性的客观测量方案有助于克服这些问题。这一观点最早由[Schroeder等人，1979年]发表。在这篇主要关于语音编码的文章中，描述了测量方案“噪声响度（NL）”。

在上述文章中，对大约20毫秒的每个时间帧估算语音编解码器的噪声信号的感知响度，即其输入和输出信号之间的差异。如果噪声信号被完全掩蔽，则感知响度为零。部分掩蔽降低了未掩蔽噪声信号的响度。所使用的掩蔽阈值针对音调掩蔽噪声进行了优化，并针对每一帧计算最终的语音劣化。不计算语音样本总质量的汇总。

1985年，Karjalainen发表了测量方案“听觉频谱差异（ASD）”[Karjalainen，1985年]。他从Schroeder、Atal和Hall的几个想法开始，但用带重叠滤波器的滤波器组取代了基于帧的分析，改变了包含绝对阈值的方式，并添加了时间掩蔽模型。测量方案的两个输入信号以完全相同的方式进行处理，并产生了一种内部表示。对这些内部表示加以相互比较，以解释语音编码方案的输入和输出信号之间的感知差异。不计算语音样本总质量的汇总。ASD的时间分辨率更好地适应人类听觉系统的特性，但是增加了算法的复杂性。

1987年，Brandenburg发表了测量方案“噪声掩蔽比（NMR）”[Brandenburg，1987年]，该方案旨在被用作开发音频编码方案的工具。通过使用被设计为最坏情况曲线的扩展函数来计算感知频带上的扩展，与噪声响度相比，该方案的复杂度降低了。所使用的掩蔽阈值针对噪声掩蔽音进行了优化。添加了一个简单的后期掩蔽建模方案和多种评估较长音频片段感知质量的方法。这是第一个在实时硬件中实施的方案。

1989年，Moore和Glasberg[Moore，1989年]提出了一个感知模型，但没有提出判断受损音频信号的感知质量的方法。

2 客观感知音频质量测量方法的一般结构

所有感知测量方案均使用两个输入信号：一个被称为参考信号（REF），另一个被称为待测信号（SUT）。在参考信号无法传输到测量设备但信号已知的情况下，参考信号可以是存储在测量设备自身中的内部参考信号。输入信号必须是时间对齐的。

可以通过两种不同的方式将心理声学融入测量方案。第一种可能性与音频编码方案的结构非常相似：参考信号用于估算实际掩蔽阈值（见下文）。将待测信号和参考信号之间的差异与该掩蔽阈值进行比较。这种方法被称为“掩蔽阈值概念”，用于噪声响度和NMR。输入信号之间的差异可以在时域中计算，或者作为短时能谱之间的差异来计算。后者针对时间对齐误差提供了更好的稳健性，但是降低了时间分辨率。时域中的差异通常对相位失真过于敏感，因此不再使用。

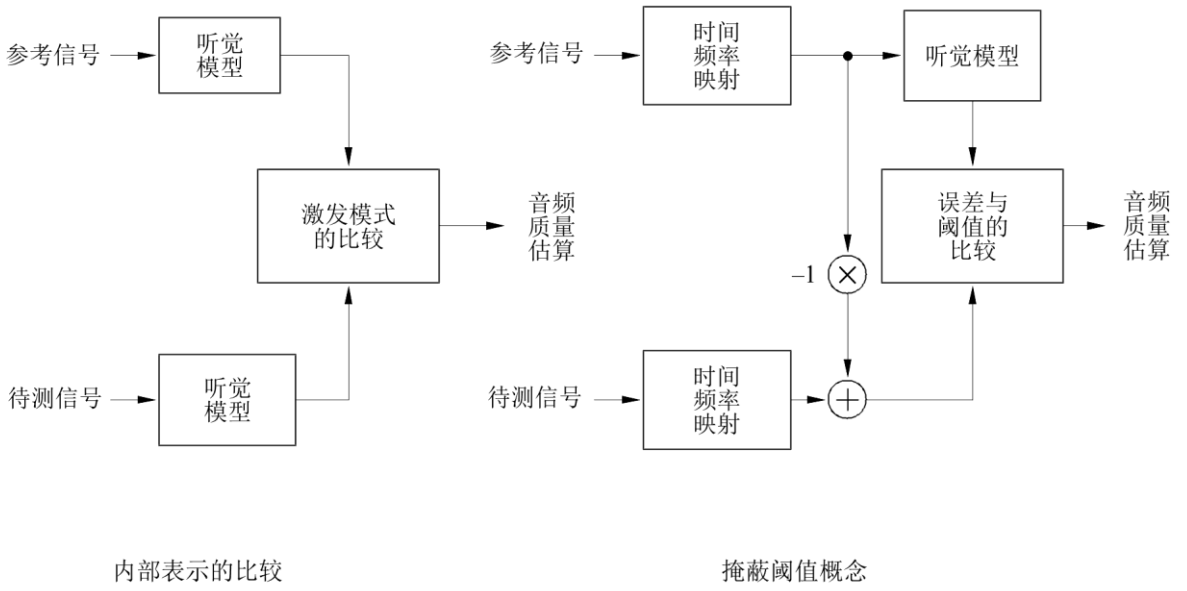
第二种方法更接近人类听觉系统的生理过程：计算参考信号和待测信号的所谓内部表示。这种内部表示是对人脑可用于信号比较的信息的估算。这种方法被称为“内部表示的比较”，在ASD中使用。

3 心理声学和认知基础

本节讨论人类听觉系统的特性，这些特性在音频信号感知质量的评估中最为突出。重点在于如何对这些特性建模。

图5

感知测量方案中不同方法使用的心理声学概念



3.1 外耳和中耳传递特性

一般来说，声音信号必须通过外耳和中耳，随后它们到达内耳，并在内耳中执行声音检测和分析过程。外耳和中耳对输入信号执行带通滤波操作。听觉神经中存在的噪声以及由血流引起的噪声被添加到输入信号中。这种噪声的幅度随着频率的降低而增大。外耳和中耳传递函数以及内部噪声限制了检测小音频信号的能力，并且对听觉的绝对阈值影响最大。

3.2 感知频率标度

人耳中的声压感受器是毛细胞，它们位于内耳，更准确地说是位于耳蜗。在耳蜗中，执行频率到位置的转换。最大激发的位置取决于输入信号的频率。耳蜗上给定位置的每个毛细胞负责频率标度上的重叠范围。音高的感知印象与毛细胞的恒定距离相关。

根据所使用的心理声学实验，已经发现了从频率到音高的不同变换函数：

在[Zwicker和Feldtkeller，1967年]中，给出了一个表格，该表格将频率范围（单位为赫兹）划分为24个非重叠频带，即所谓的临界频带。表5给出了这些频带的截止频率上限。该表还包含巴克标度的定义：1巴克对应于100赫兹，24巴克对应于15 500赫兹。

表5

Zwicker定义的临界频带范围

临界频带	1	2	3	4	5	6	7	8	9	10	11	12
上限截止频率（赫兹）	100	200	300	400	510	630	770	920	1 080	1 270	1 480	1 720

临界频带	13	14	15	16	17	18	19	20	21	22	23	24
上限截止频率（赫兹）	2 000	2 320	2 700	3 150	3 700	4 400	5 300	6 400	7 700	9 500	12 000	15 500

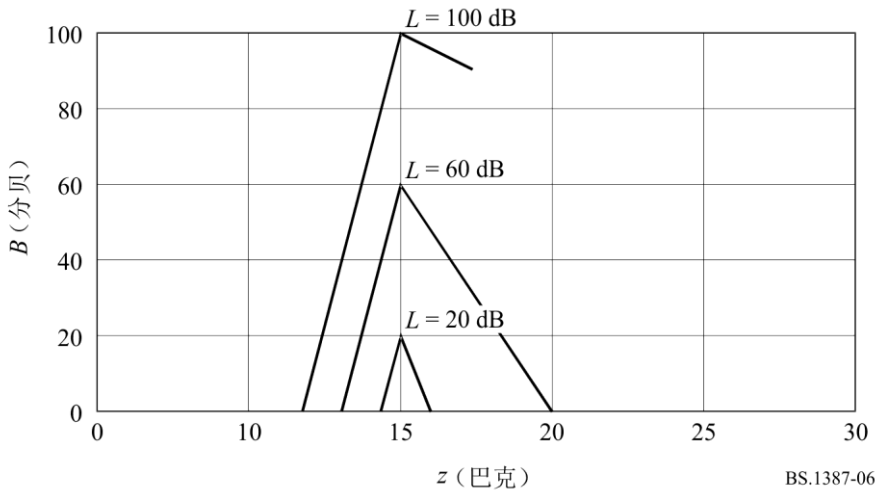
过去发现了几种近似的巴克标度。有关不同标度的详细讨论可参见[Cohen和Fielder，1992年]。在感知音频质量客观测量的背景下，使用巴克标度可以获得最佳结果。

3.3 激发

每个毛细胞对一系列频率做出反应，这些频率可以通过滤波器特性来描述。如上所述，滤波器的斜率可以在感知量表上得到最佳表达。这种量表的滤波器形状几乎与中心频率无关。激发的较低斜率与输入信号的电平L无关（约27分贝/巴克）。输入信号较低电平时的上斜率比较高电平时更陡（−5至−30分贝/巴克）。这种陡峭的特性是由两种不同毛细胞之间的反馈机制引起的，需要一些时间来加以稳定。因此，对于信号开始后几毫秒的稳定信号，可以获得最佳的听觉频率分辨率。由几个分量组成的信号的激发模式以非线性方式相加。

图6

基于Terhardt[1979年]的激发水平依赖性



暴露于信号后，毛细胞和神经处理需要一些时间来恢复，直到再次达到完全敏感。恢复过程的持续时间取决于信号的电平和持续时间，且可持续数百毫秒。在毛细胞和大脑之间的通路上，高电平信号比低电平信号处理得更快。因此，强信号的出现会掩蔽之前较弱的信号。

另一种模型激发方法基于ERB标度[Moore，1986年]。这种方法使用所谓的ROEX滤波器[Moore，1986年]。在感知音频质量客观测量的背景下，基于[Zwicker和Feldtkeller，1967年]和[Terhardt，1979年]的模型可以获得更好的结果。

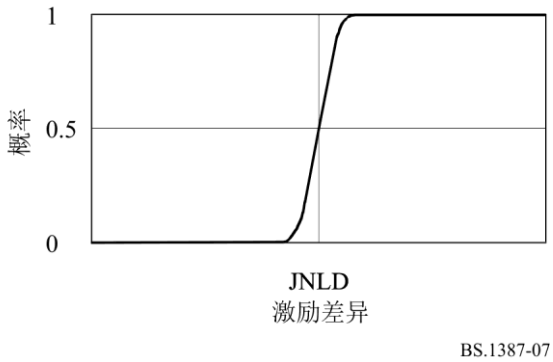
3.4 检测

不同音频信号的激发被传递到人脑。在此方面，存在三种不同的记忆：长期记忆、短期记忆和超短期记忆，其详细程度和信息存在的持续时间亦有所不同。在听力测试中，超短期记忆的作用最为突出。如果音频节录的持续时间少于五到八秒（取决于听者和音频节录），则信号的大部分细节均会保留。在ITU-R BS.1116建议书中定义的评估程序中考虑到了这一点，其中允许受试者选择音频节录的非常短的部分来更仔细地聆听。在检测阈值处，检测概率为50%。在阈值附近，检测到差异的概率从0%平滑增加到100%。

最小可觉电平差（JNLD）是电平差的检测阈值。JNLD受输入信号电平的影响。对于小信号，检测需要较大的差异（电平：20 dBSPL，JNLD：0.75 dB）。对于强信号，对微小差异的灵敏度要高得多（电平：80 dBSPL，JNLD：0.2 dB）。这些数字是基于调幅实验得出的。

图7

检测概率原则



BS.1387-07

3.5 掩蔽

一个单独出现时清晰可闻的信号在另一个信号（掩蔽信号）出现时可能完全听不见。这种效应称为掩蔽效应，被掩蔽的信号称为被掩蔽声（maskee）。必须区分两种情况：

– 同时掩蔽

在此情况下，掩蔽声（masker）和被掩蔽声同时出现，并且是准静态的。如果掩蔽声具有离散带宽，则即使对于低于或高于掩蔽声的频率，听觉阈值亦会提高。掩蔽的量取决于掩蔽声和被掩蔽声的结构。在类噪声信号掩蔽音调信号的情况下，掩蔽量几乎与频率无关。如果被掩蔽声的声压级比掩蔽声的声压级低5分贝左右，则被掩蔽声就听不见了。在音调信号掩蔽类噪声信号的情况下，掩蔽量取决于掩蔽声的频率。通过公式$\left(15.5+\frac{z}{Bark}\right)$dB, 可以对掩蔽量进行估算，其中$z$是掩蔽声的临界频带速率。此外，在高信号电平情况下，非线性效应降低了掩蔽声附近的掩蔽阈值。音调掩蔽音亦有类似的效果。几个信号的掩蔽阈值以非线性方式相加。一般来说，得到的掩蔽阈值高于每个单独信号产生的掩蔽阈值。

– 时间掩蔽

在此情况下，掩蔽声和被掩蔽声出现在不同的时间。在掩蔽声衰减后不久，掩蔽阈值更接近于该掩蔽声的同时掩蔽，而非绝对阈值。根据掩蔽的持续时间，阈值的衰减时间可以在5毫秒（掩蔽声：持续时间约为0.05毫秒的高斯脉冲）和150毫秒以上（掩蔽声：持续时间为1秒的粉红噪声）之间。恰处于较大信号之前的微弱信号被掩蔽。这种反向掩蔽效应的持续时间约为5毫秒。如果被掩蔽声恰好高于阈值，则其在掩蔽声之前不会被感知到，而是被感知为掩蔽声的变化。反向掩蔽显示了不同听者之间的较大偏差。

3.6 响度和部分掩蔽

音频信号的感知响度取决于其频率、持续时间和声压级。由于自动掩蔽，复合信号的响度小于其所有分量的响度之和。在音频质量测量的背景下，添加到参考信号的无用失真的响度（噪声响度）通过参考信号引起的部分掩蔽而降低。

3.7 锐度

锐度是感觉的基本值之一，与音色有关。如果声音主要包含高频成分，则会被认为是尖锐的。例如，高频的正弦音或带限噪声，或者截止频率在约3 kHz以上的频率范围内的高通噪声被称为锐噪声。不过，音频信号的详细频率结构对锐度没有重大影响。Von Bismarck[1974年]进行了有关锐度的基础研究。

[Aures，1984年]进行了有关锐度的其他研究。这些研究的结果是一个与Bismarck定义的加权函数相比略有修改的加权函数。在非常低和非常高的临界频带速率下，其对锐度属性的贡献较小，而在14和20巴克之间的临界频带速率下贡献较大。此外，这些研究表明，具有音频信号声压水平的高变化和强高频内容的音频信号的锐度不能仅基于整体响度，而是基于取决于整体响度的加权函数。

3.8 认知处理

很明显，感知音频质量受到认知效应的强烈影响。这可以通过一个简单的实验来证明。

具有清晰可闻的背景噪声的参考信号由一些不能传输该背景噪声的音频设备进行处理。由于噪声是一种无用失真，在收听测试中，参考信号将被评定为比处理后的信号更差。另一方面，如果参考信号中最重要的部分是软背景噪声，则同一经处理信号的得分会更低。

列出所有可能的认知效应超出了本后附资料的范围，但在此列出一些示例：

示例1：线性失真和非线性失真的分离

线性失真比非线性失真更不令人反感。通过使用输出信号的自适应逆滤波，可以相当容易地实现线性失真和非线性失真的分离。本建议书中规定的方法使用线性和非线性失真的分离。

示例2：听觉场景分析

听觉场景分析[Bregman，1990年]是一种认知过程，它允许听者将不同的听觉事件分开并将其分组到不同的对象中。[Beerends和Stemerdink，1994年]中给出的实用方法对于量化听觉场景分析效果非常有用。如果时间-频率分量没有被编解码器编码，则剩余的信号仍然形成一个连贯的听觉场景，而新的不相关时间-频率分量的引入则导致两种不同的感知。由于分成两种不同的感觉，失真将比根据新引入的失真分量的响度所预期的情况更令人反感。这

导致由未编码时间-频率分量引起的失真的感知干扰与由引入新的时间-频率分量引起的干扰之间出现不对称的问题。

示例3：信息掩蔽

信息掩蔽可以通过定义类似熵的频谱-时间复杂度指标来建模。此类效应很可能取决于受试者在进行主观评估之前接受的训练量。在[Beerends等人，1996年]中介绍了对此效应建模的第一次尝试，其中计算了大约100毫秒的时间窗口内的本地复杂度估算。如果这个本地复杂度很高，那么在这个时间窗口内的失真比本地复杂度低时更难听到。训练可以将掩蔽阈值降低几个10分贝[Leek和Watson，1984年]。

示例4：频谱-时间加权

音频信号中的一些频谱-时间区域携带更多信息，因此可能比其他区域更重要。频谱-时间加权在语音编解码器的质量判断中十分重要。在语音中，一些频谱-时间成分（如共振峰）显然比其他成分携带更多的信息[Beerends和Stemerdink，1994年]。不过，在音乐中，信号中所有的频谱-时间成分均可能携带信息，甚至在静音中亦会如此。

4 模型合并

4.1 干扰指数（DIX）

感知测量方法DIX（干扰指数）[Thiede和Kabot，1996年]基于听觉滤波器组，该滤波器组产生高时间分辨率，从而允许（与基于FFT的方法相比）对时间效应（如预掩蔽和后掩蔽）进行更精确的建模。每个听觉滤波器处的包络的时间精细结构被保留，并用于获得有关信号和所引入失真的附加信息。

各个滤波器的中心频率在感知音高范围内均匀分布。滤波器形状的顶部略呈圆形，以确保所选数量的滤波器覆盖整个频率范围，而不会在整体频率响应中产生纹波。为了对掩蔽阈值建模，滤波器斜率在巴克标度上呈指数下降。滤波器斜率的陡度取决于输入信号的电平。在DIX的第一个版本中，可听频率范围由80个滤波器覆盖，后来减少到40个滤波器，即频率分辨率对应于大约0.6巴克。与具有单个滤波器的其他滤波器组相比，滤波器组算法相当快，但仍比基于块的变换（如FFT和小波包变换）耗时得多。

DIX动态调整待测信号和参考信号之间的电平和频谱，以区分线性失真和非线性失真。它评估滤波器输出端的时间包络结构，以模拟与纯音相比由调制和类噪声掩蔽器引起的掩蔽量增加。

通过比较待测信号和参考信号的内部表示，计算出许多输出参数，其中包括非线性失真的部分响度、线性失真量指标以及时间和双耳效应指标。不过，基本音频质量的良好估算可以通过仅使用两个输出参数来实现：非线性失真的部分响度以及线性失真量的指标之一被映射到待测信号的预期基本音频质量的估算。

4.2 噪声掩蔽比（NMR）

测量方案NMR（噪声掩蔽比）[Brandenburg，1987年]评估掩蔽阈值和噪声信号之间的电平差。具有约20毫秒汉宁窗的DFT用于分析信号的频率成分。根据巴克标度将变换系数组合成频带。为每个频带估算掩蔽阈值。掩蔽阈值的斜率是使用最坏情况方法导出的，该方法考虑了以下事实：对于弱信号，斜率更陡，但在较高电平时会达到绝对阈值。绝对阈值适应输入信号的分辨率（通常为16位），但不适应心理声学需求。由于上述事实，NMR对于再现电平的变化是稳健的。音高标度分辨率约为1巴克。由于所需的计算能力较低，因此有可能在其开发的早期阶段将NMR实施为实时系统。

该模型自1987年以来一直在使用，且已证明其基本可靠性。

NMR最重要的输出值是掩蔽标志率（给出听觉失真帧的百分比）以及总NMR和平均NMR，它们是平均误差能量和掩蔽阈值之间距离的不同方法。

4.3 客观音频信号评估（OASE）

测量方案OASE（客观音频信号评估）[Sporer，1997年]使用具有241个滤波器的滤波器组来分析输入信号。中心频率在巴克标度上等距分布，距离为0.1巴克。滤波器相互重叠。每个滤波器均适用于基膜上一点的频率响应。斜率的水平相关性通过NMR中的最坏情况方法纳入。低中心频率的滤波器需要以全采样率进行计算，而高中心频率的滤波器可以降低的采样率进行计算。在滤波器之后将计算人类听觉系统的时间效应模型，如在ASD中所做的那样。在此步骤之后，所有滤波器频带中的采样率均有可能降低。这导致滤波器组在48 kHz采样速率下的时间分辨率为0.66毫秒。将参考信号和待测信号的匹配滤波器的输出与检测概率函数进行比较。该函数使用输入信号的响度作为输入来计算JNLD。检测的总概率是从每个频带的检测概率中导出的。对两个输入声道和所谓的中心声道均进行该操作。每个频带在中心声道中的检测概率是左声道和右声道检测概率的最坏情况。对于0.66毫秒的每一帧，还会计算阈值以上步长的总和。

使用检测概率和阈值以上步长的几种时间平均方法：

– 检测概率的时间平均值；
– 检测概率高于0.5的帧的频率；
– 低通滤波检测概率的最大值；
– 遗忘检测的低通滤波概率的最大值；
– 检测概率高于0.5的帧高于阈值的平均步数；
– 阈值以上的平均步数；
– 阈值以上的最大步数；
– 超过10%最差帧阈值的平均步数。

4.4 感知音频质量指标（PAQM）

PAQM的基本原则[Beerends和Stemerdink，1992年]是减去参考信号和劣化信号的内部表示（受试者头部内部的表示），并用认知映射将差异映射到主观感知的音频质量。从物理外部领域到心理-物理内部领域的转换通过四种操作来执行：

– 通过具有大约40毫秒持续时间的汉宁窗的DFT进行的时间-频率映射；
– 使用巴克标度的频率弯曲；
– 时频扩展（非线性卷积）；
– 强度弯曲（压缩）。

模糊和压缩的结合允许在掩蔽阈值及以上模拟人类听觉系统的掩蔽行为。通过使用第一个MPEG音频编解码器评估的主观结果来执行压缩的优化[ISO/IEC JTC 1/SC 2/WG 11 MPEG/Audio测试报告，文件MPEG90/N0030，1990年10月][ISO/IEC JTC 1/SC 2/WG 11 MPEG/Audio测试报告，文件MPEG91/N0010，1991年6月]。内部表示的差异用噪声干扰来表示。在提交给ITU-R的最新PAQM版本中，从噪声干扰到主观质量的映射中包括两种认知效应：感知流[Beerends和Stemerdink，1994年]和信息掩蔽[Beerends等人，1996年]。

PSQM[Beerends和Stemerdink，1994年]的PAQM（感知语音质量测量）的简化版本是使用[Beerends和Stemerdink，1994年]中提出的认知模型开发的，但扩展了无声区间的权重。在PSQM的开发过程中发现，在判断电话环境中的语音质量时，无声区间中出现的噪声不如语音活动区间中出现的噪声重要。在ITU-T的基准测试中，PSQM建议显示了客观和主观质量之间的最高相关性（电信标准化第12研究组，COM 12-74 — 客观语音质量指标验证测试回顾）。ITU-T P.862建议书规定了语音质量感知评估（PESQ），其中包含一种改进的语音质量评估算法。

4.5 感知评估（PERCEVAL）

PERCEVAL（感知评估）[Paillard等人，1992年]对中耳和内耳的传递特性建模，以形成信号的内部表示。使用DFT将输入信号分解为时频表示。通常，将约40毫秒的汉宁窗应用于输入数据，连续窗口之间有50%的重叠。能谱乘以频率相关函数，该函数对耳道和中耳的效应建模。衰减的频谱能量值从频率标度映射到音高标度，该音高标度相对于内耳的物理特性和观察到的心理-物理效应更加线性。然后，将转换后的能量分量与扩展函数进行卷积，以模拟能量沿基膜的扩散。最后，一个内在的频率相关的能量被添加到每个音高成分，以说明听觉的绝对阈值。将能量转换为分贝会产生信号的基底膜表示。

在听觉掩蔽实验的模拟中，为每个刺激形成一个基底膜表示，表示之间的差异是可用于执行任务的信息。一种表示是掩蔽信号本身，另一种表示则是掩蔽信号和测试信号的组合。它们的差值代表信号中未被掩蔽的成分。PERCEVAL计算出检测到这种差异的概率。使用S形概率函数来估算每个检测器沿模拟基底膜的差异的未检测概率。假设检测器在统计上是独立的，则整组检测器的全局检测概率被计算为各个未检测概率的乘积的补码。使用这种方法成功模拟了几个掩蔽实验，并使用该模型评估了对单个听者建模的可行性[Treurniet，1996年]。

作为估算音频质量的工具，PERCEVAL计算参考信号和测试信号之间的差异。通过应用关于更高级感知和认知过程的合理假设，许多感知相关变量被计算并映射到待测信号的客观质量指标。通过最小化客观质量分布和可用数据集的平均主观质量评级的相应分布之间的差异，相关映射得以优化。

4.6 感知客观测量（POM）

感知客观测量（POM）[Colomes等人，1995年]旨在量化参考信号与其“劣化”版本之间可能发生的一定量的劣化。无论劣化的产生出于何种原因，POM是通过比较两个信号的内部基底表示来实现的。基底表示模拟了音频信号通过人耳时所经历的不同过程。因此，POM的第一阶段是计算音频信号的内部表示。分布在基底膜上的激发模式（以分贝为单位）被用来模拟基底膜上神经元的放电频率。

计算激发模式的过程被称为人工耳。然后，一旦要对两个信号的内部表示进行相互比较，那么POM必须说明它们的内部表示之间的差异是否听得见，如果听得见，又须说明以何种方式听得见。这就是所谓的检测过程。

POM使用具有大约40毫秒持续时间的汉宁窗的DFT（两个汉宁窗之间有50%的重叠）。分析基底通道的数量为620。听觉模型的其余部分与PAQM和PERCEVAL使用的几乎相同。

扩展函数由更确切的近似值进行非常准确的描述，该近似值考虑了基于[Terhardt，1979年]的水平依赖性和基于[Schroeder等人，1979年]的圆形形状。

该模型输出检测到所比较的两个信号之间出现失真的概率，以及表示所比较的两个激发之间感知差距的所谓基底距离。

4.7 工具箱方法

工具箱使用三步方法来测量音频测试信号的音频质量相对于音频参考信号的感知距离，从而说明测试信号的总体主观音频质量水平。该方法基于众所周知的感知模型，而该感知模型用于描述两个音频信号之间差异的感知表示。此外，它还包括立体声测试信号的感知音频质量的加权过程，同时亦考虑左声道和右声道的结果。不需要在逐个样本的基础上对参考信号和待测音频信号进行严格的相关。

工具箱第一步的主要功能是基于特定响度的计算，根据[Zwicker和Feldtkeller，1967年]进行计算，使用2 048点的FFT，并通过汉宁窗进行窗口化，这相当于约40毫秒的持续时间。整个窗口以10毫秒的增量进行移动。此外，还应用了时间掩蔽效应，如Zwicker所说的后掩蔽和前掩蔽。利用上述感觉的基本值，根据[von Bismarck，1974年]和[Aures，1984年]计算其他感知参数，如综合响度、部分掩蔽响度、锐度以及预回声量，并将其作为下一步预处理阶段的结果。

工具箱的第二步包括加权过程，该过程主要取决于响度的感知差异量和响度随时间的变化。

工具箱的第三步包括生成一组中间工具箱输出值，这些输出值基于对第一步和第二步中获得的值的统计分析。该统计分析的输出包括平均值、最大值和均方根值，以及平均值的标准偏差。这些中间工具箱输出值的加权被用于待测信号和参考信号之间的感知距离的最终拟合。如有必要，该单个输出值可以与主观差异等级相匹配，在主观听力测试中，主观差异等级通常通过使用线性或高阶多项式函数为每个时间增量顺序拟合输出数据来获得。

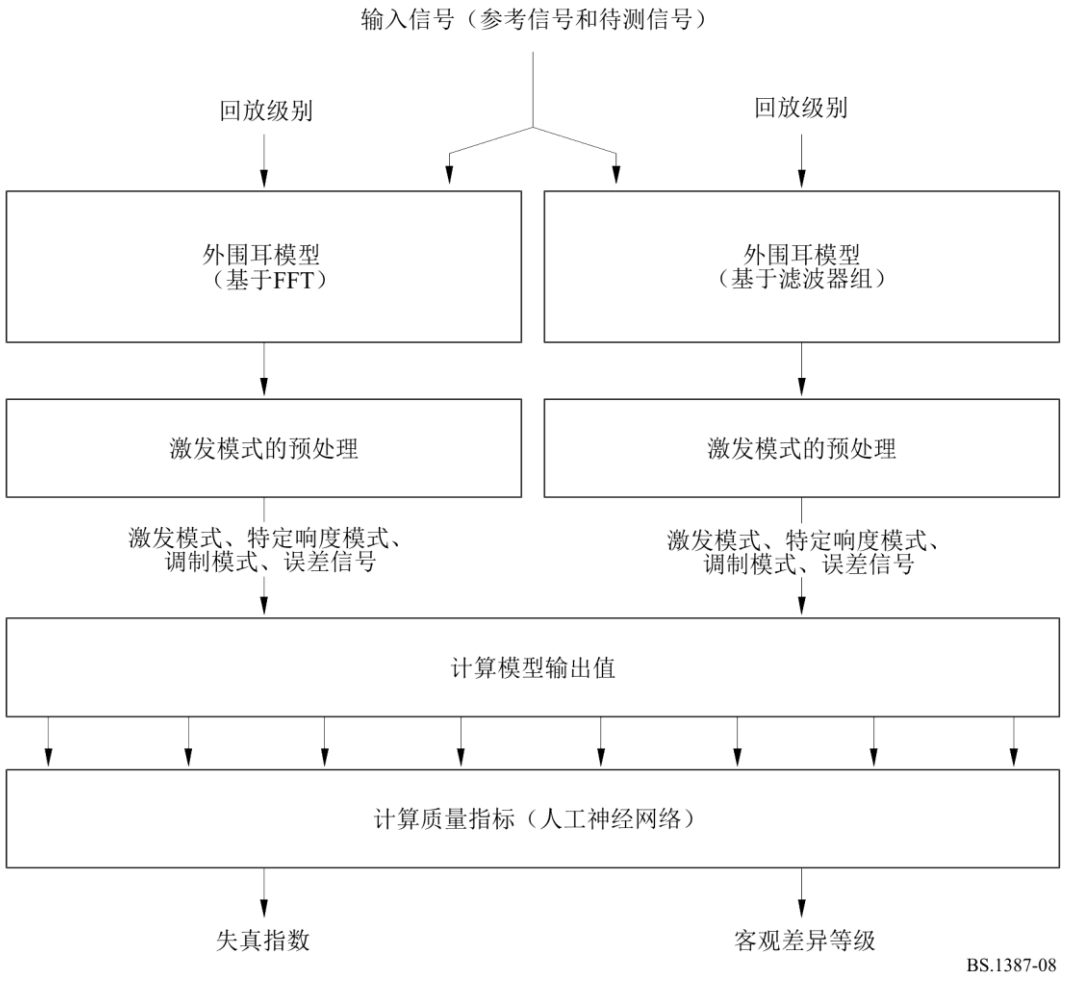
附件2

模型描述[2]

1 概述

图8

测量方案的通用框图



[2] 本建议书中所述技术的支持者已经提交了符合ITU-R第1号决议附件1的专利声明。本建议书中描述的技术受国际专利保护，并且像所有国际电联建议书一样受到版权保护。要使用这项技术，必须事先征得所有者的许可。要获得有关该技术许可的更多信息，请查阅ITU-R的专利数据库或向无线电通信局（BR）秘书处垂询。

所建议的用于客观测量感知音频质量的方法包括外围耳模型、几个中间步骤（此处称为“激发模式的预处理”）、（主要）基于心理声学的模型输出变量（MOV）的计算以及从一组模型输出变量到表示待测信号的基本音频质量的单个值的映射。此方法包括两个外围耳模型，一个基于FFT，另一个基于滤波器组。除了误差信号的计算（仅用于人耳模型的基于FFT的部分）之外，两种外围耳模型的总体结构是相同的。

MOV计算的输入为：

– 测试和参考信号的激发模式。
– 测试和参考信号的频谱适应激发模式。
– 测试信号和参考信号的特定响度模式。
– 测试信号和参考信号的调制模式。
– 作为测试和参考信号之间的频谱差异计算的误差信号（仅适用于基于FFT的人耳模型）。

如果未做其他说明，在立体声信号的情况下，所有计算均为独立进行的，并且对于左声道和右声道而言均为以相同方式进行的。

该描述定义了两种设置，一种称为“基本版本”，另一种称为“高级版本”。

在所有给定的等式中，指数“Ref”代表从参考信号计算的所有模式，指数“Test”代表从待测信号计算的所有模式。指数“k”代表离散频率变量（即频带），而“n”则代表离散时间变量（即帧计数器或样本计数器）。如果k或n的值没有明确定义，则计算将针对k和n的所有可能值进行。所有其他缩写将在其出现时加以解释。

在模型输出变量的名称中，指数“A”代表使用基于滤波器组的人耳模型部分计算的所有变量，指数“B”代表使用基于FFT的人耳模型部分计算的所有变量。

1.1 基本版本

基本版本仅包括根据基于FFT的人耳模型计算的MOV。不使用模型的基于滤波器组的部分。基本版本总共使用11个MOV来预测感知的基本音频质量。

1.2 高级版本

高级版本包括根据基于滤波器组的人耳模型计算的MOV以及根据基于FFT的人耳模型计算的MOV。频谱适应的激发模式和调制模式仅从模型的基于滤波器组的部分计算。高级版本总共使用5个MOV来预测感知的基本音频质量。

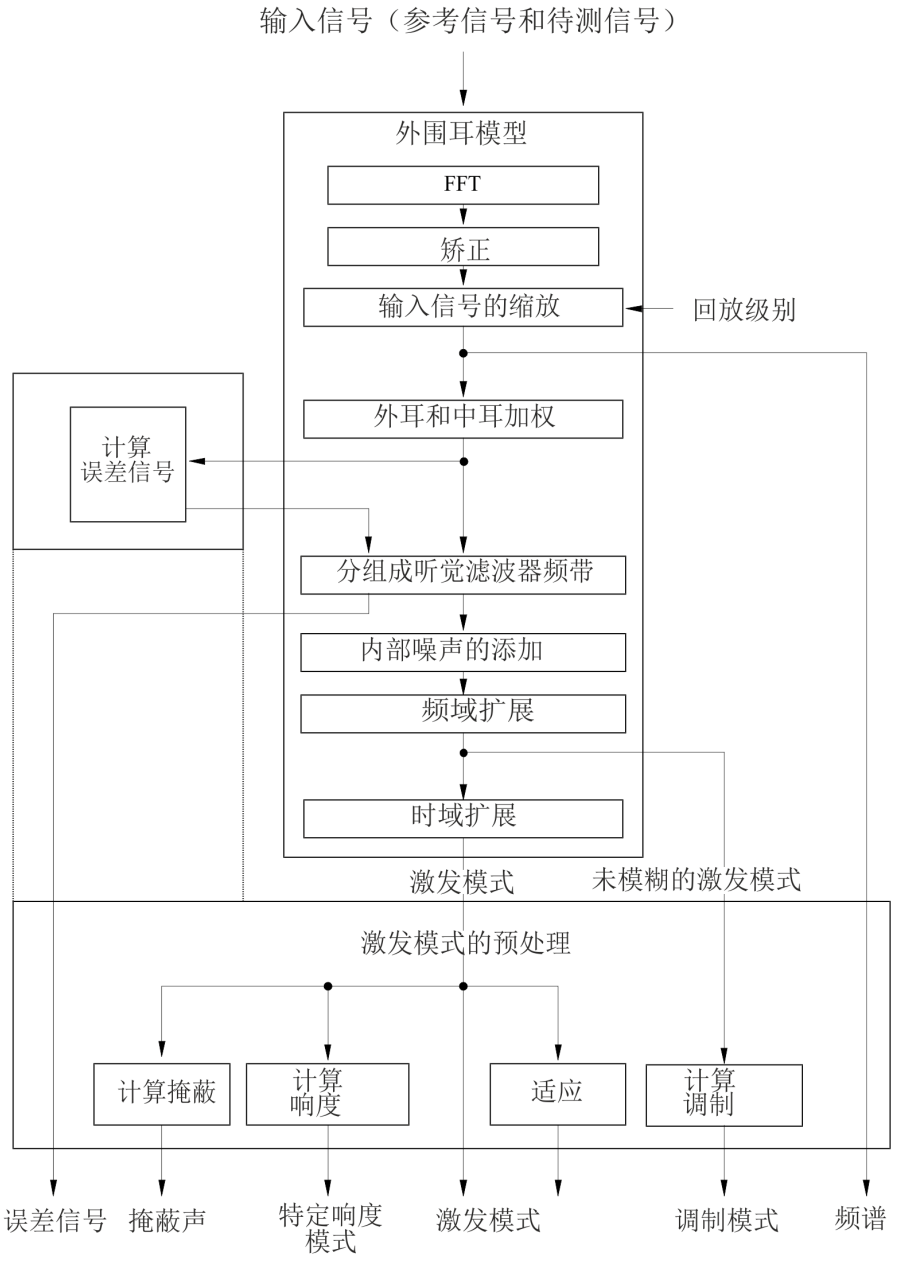
2 外围耳模型

2.1 基于FFT的人耳模型

2.1.1 概述

图9

外围耳模型和该模型基于FFT部分的激发模式的预处理



BS.1387-09

基于FFT的人耳模型的输入（48 kHz采样时间对齐的参考和测试信号）被切割成约0.042秒的帧，且具有50%的重叠。使用汉宁窗和短期FFT将每一帧变换到频域，并缩放到回放级别。对频谱系数应用加权函数，该函数模拟外耳和中耳的频率响应。通过将加权频谱系数分组到临界频带中来完成到音高表示的变换。添加频率相关偏移来模拟听觉系统中的内部噪声。使用电平相关扩展函数在频域中为频谱听觉滤波器建模。它遵循考虑正向掩蔽效应的时域扩展。

此时获得的激发模式用于计算特定响度模式和掩蔽模式。最终时域扩展之前的模式（“未模糊的激发模式”）用于计算调制模式。

为了对误差信号建模，外耳和中耳滤波器输出的参考和测试信号模式被组合在一起，并通过分组到临界频带而映射到音高标度。

这些输出与激发模式一起用于计算模型输出变量的值。

2.1.2 时间处理

基于FFT的人耳模型、测试和参考信号的输入被切割成具有1 024个样本重叠的2 048个样本的帧：

$$t_n[k_t, n] = t[1\,024 \cdot n + k_t] \quad n = 0, 1, 2 \ldots k_t = 0..2047 \tag{1}$$

其中n是时间帧的编号，k_t是帧内的时间计数器。

2.1.3 快速傅里叶变换（FFT）

使用汉宁窗完成从时域到频域的映射：

$$h_w[k] = \frac{1}{2}\sqrt{\frac{8}{3}}\left[1 - cos\left(2\pi\frac{k}{N-1}\right)\right] \Bigg| \quad N = 2048 \tag{2}$$

$$t_w[k_t, n] = h_w[k_t] \cdot t_n[k_t, n] \tag{3}$$

随后是短期傅立叶变换：

$$F_f[k_f, n] = \frac{1}{2048}\sum_{k_t=0}^{2047} t_w[k_t, n]\mathrm{e}^{-j\frac{2\pi}{2048}k_f k_t} \tag{4}$$

FFT的比例因子通过满量程正弦波的假设声压级Lp计算得出：

$$fac = \frac{10^{\frac{Lp}{20}}}{Norm} \tag{5}$$

$$F[k_f, n] = fac \cdot F_f[k_f, n] \tag{6}$$

其中归一化因数$Norm$的计算方法是：将1 019.5赫兹、0 dB满量程的正弦波作为输入信号，并计算10帧内频谱系数的最大绝对值。

如果声压级未知，建议将Lp设置为92 dB_{SPL}。

2.1.4 外耳和中耳

外耳和中耳的频率响应由频率相关加权函数建模：

$$W[k]/\mathrm{dB} = -0.6 \cdot 3.64 \cdot \left(\frac{f[k]}{\mathrm{kHz}}\right)^{-0.8} + 6.5 \cdot \mathrm{e}^{-0.6 \cdot \left(\frac{f[k]}{\mathrm{kHz}} - 3.3\right)^2} - 10^{-3} \cdot \left(\frac{f[k]}{\mathrm{kHz}}\right)^{3.6} \quad (7)$$

其中：

$$f[k]/\mathrm{Hz} \;=\; k \;\cdot\; 23.4375 \quad (8)$$

是应用于FFT输出的k线频率表示（等式（9））。

$$F_e[k_f, n] \;=\; \left|F[k_f, n]\right| \;\cdot\; 10^{\frac{W[k_f]}{20}} \quad (9)$$

$F_e[k_f]$被称为“外耳加权FFT输出”。

2.1.5 分组到临界频带

听觉音高标度是根据[Schroeder等人，1979年]给出的近似值计算的。

$$z/Bark \;=\; 7 \;\cdot\; \mathrm{arsinh}\left(\frac{f/\mathrm{Hz}}{650}\right) \quad (10)$$

音高单位被命名为“巴克”（尽管该标度并不完全代表[Zwicker和Feldtkeller，1967年]定义的巴克标度）。

滤波器的频率范围从80赫兹到18 000赫兹。滤波带的宽度和间距对应于基本版本的分辨率为$res = 0.25$巴克，高级版本的分辨率为$res = 0.5$巴克。

这导致基本版本的频带数量为$Z = 109$，高级版本的频带数量为$Z = 55$。

表6

基本版本中使用的基于FFT的人耳模型的频带

组	较低频率/赫兹	中心频率/赫兹	较高频率/赫兹	频率宽度/赫兹
k	$f_l[k]$	$f_c[k]$	$f_u[k]$	$f_w[k]$
0	80	91.708	103.445	23.445
1	103.445	115.216	127.023	23.577
2	127.023	138.87	150.762	23.739
3	150.762	162.702	174.694	23.932
4	174.694	186.742	198.849	24.155
5	198.849	211.019	223.257	24.408
6	223.257	235.566	247.95	24.693
7	247.95	260.413	272.959	25.009
8	272.959	285.593	298.317	25.358
9	298.317	311.136	324.055	25.738
10	324.055	337.077	350.207	26.151
11	350.207	363.448	376.805	26.598
12	376.805	390.282	403.884	27.079
13	403.884	417.614	431.478	27.594
14	431.478	445.479	459.622	28.145
15	459.622	473.912	488.353	28.731
16	488.353	502.95	517.707	29.354
17	517.707	532.629	547.721	30.014
18	547.721	562.988	578.434	30.713
19	578.434	594.065	609.885	31.451
20	609.885	625.899	642.114	32.229
21	642.114	658.533	675.161	33.048
22	675.161	692.006	709.071	33.909
23	709.071	726.362	743.884	34.814
24	743.884	761.644	779.647	35.763
25	779.647	797.898	816.404	36.757
26	816.404	835.17	854.203	37.799
27	854.203	873.508	893.091	38.888
28	893.091	912.959	933.119	40.028
29	933.119	953.576	974.336	41.218
30	974.336	995.408	1016.797	42.461
31	1016.797	1038.511	1060.555	43.758
32	1060.555	1082.938	1105.666	45.111

表6（续）

组	较低频率/赫兹	中心频率/赫兹	较高频率/赫兹	频率宽度/赫兹
k	$f_l[k]$	$f_c[k]$	$f_u[k]$	$f_w[k]$
33	1105.666	1128.746	1152.187	46.521
34	1152.187	1175.995	1200.178	47.991
35	1200.178	1224.744	1249.7	49.522
36	1249.7	1275.055	1300.816	51.116
37	1300.816	1326.992	1353.592	52.776
38	1353.592	1380.623	1408.094	54.502
39	1408.094	1436.014	1464.392	56.298
40	1464.392	1493.237	1522.559	58.167
41	1522.559	1552.366	1582.668	60.109
42	1582.668	1613.474	1644.795	62.128
43	1644.795	1676.641	1709.021	64.226
44	1709.021	1741.946	1775.427	66.406
45	1775.427	1809.474	1844.098	68.671
46	1844.098	1879.31	1915.121	71.023
47	1915.121	1951.543	1988.587	73.466
48	1988.587	2026.266	2064.59	76.003
49	2064.59	2103.573	2143.227	78.637
50	2143.227	2183.564	2224.597	81.371
51	2224.597	2266.34	2308.806	84.208
52	2308.806	2352.008	2395.959	87.154
53	2395.959	2440.675	2486.169	90.21
54	2486.169	2532.456	2579.551	93.382
55	2579.551	2627.468	2676.223	96.672
56	2676.223	2725.832	2776.309	100.086
57	2776.309	2827.672	2879.937	103.627
58	2879.937	2933.12	2987.238	107.302
59	2987.238	3042.309	3098.35	111.112
60	3098.35	3155.379	3213.415	115.065
61	3213.415	3272.475	3332.579	119.164
62	3332.579	3393.745	3455.993	123.415
63	3455.993	3519.344	3583.817	127.823
64	3583.817	3649.432	3716.212	132.395
65	3716.212	3784.176	3853.348	137.136
66	3853.348	3923.748	3995.399	142.051
67	3995.399	4068.324	4142.547	147.148
68	4142.547	4218.09	4294.979	152.432
69	4294.979	4373.237	4452.89	157.911
70	4452.89	4533.963	4616.482	163.592
71	4616.482	4700.473	4785.962	169.48
72	4785.962	4872.978	4961.548	175.585
73	4961.548	5051.7	5143.463	181.915

表6（完）

组	较低频率/赫兹	中心频率/赫兹	较高频率/赫兹	频率宽度/赫兹
k	$f_l[k]$	$f_c[k]$	$f_u[k]$	$f_w[k]$
74	5143.463	5236.866	5331.939	188.476
75	5331.939	5428.712	5527.217	195.278
76	5527.217	5627.484	5729.545	202.329
77	5729.545	5833.434	5939.183	209.637
78	5939.183	6046.825	6156.396	217.214
79	6156.396	6267.931	6381.463	225.067
80	6381.463	6497.031	6614.671	233.208
81	6614.671	6734.42	6856.316	241.646
82	6856.316	6980.399	7106.708	250.392
83	7106.708	7235.284	7366.166	259.458
84	7366.166	7499.397	7635.02	268.854
85	7635.02	7773.077	7913.614	278.594
86	7913.614	8056.673	8202.302	288.688
87	8202.302	8350.547	8501.454	299.152
88	8501.454	8655.072	8811.45	309.996
89	8811.45	8970.639	9132.688	321.237
90	9132.688	9297.648	9465.574	332.887
91	9465.574	9636.52	9810.536	344.962
92	9810.536	9987.683	10168.013	357.477
93	10168.013	10351.586	10538.46	370.447
94	10538.46	10728.695	10922.351	383.891
95	10922.351	11119.49	11320.175	397.824
96	11320.175	11524.47	11732.438	412.264
97	11732.438	11944.149	12159.67	427.231
98	12159.67	12379.066	12602.412	442.742
99	12602.412	12829.775	13061.229	458.817
100	13061.229	13296.85	13536.71	475.48
101	13536.71	13780.887	14029.458	492.748
102	14029.458	14282.503	14540.103	510.645
103	14540.103	14802.338	15069.295	529.192
104	15069.295	15341.057	15617.71	548.415
105	15617.71	15899.345	16186.049	568.339
106	16186.049	16477.914	16775.035	588.986
107	16775.035	17077.504	17385.42	610.385
108	17385.42	17690.045	18000	614.58

表7

高级版本中使用的基于FFT的人耳模型的频带

组	较低频率/赫兹	中心频率/赫兹	较高频率/赫兹	频率宽度/赫兹
k	$f_l[k]$	$f_c[k]$	$f_u[k]$	$f_w[k]$
0	80	103.445	127.023	47.023
1	127.023	150.762	174.694	47.671
2	174.694	198.849	223.257	48.563
3	223.257	247.95	272.959	49.702
4	272.959	298.317	324.055	51.096
5	324.055	350.207	376.805	52.75
6	376.805	403.884	431.478	54.673
7	431.478	459.622	488.353	56.875
8	488.353	517.707	547.721	59.368
9	547.721	578.434	609.885	62.164
10	609.885	642.114	675.161	65.277
11	675.161	709.071	743.884	68.723
12	743.884	779.647	816.404	72.52
13	816.404	854.203	893.091	76.687
14	893.091	933.119	974.336	81.245
15	974.336	1016.797	1060.555	86.219
16	1060.555	1105.666	1152.187	91.632
17	1152.187	1200.178	1249.7	97.513
18	1249.7	1300.816	1353.592	103.892
19	1353.592	1408.094	1464.392	110.801
20	1464.392	1522.559	1582.668	118.275
21	1582.668	1644.795	1709.021	126.354
22	1709.021	1775.427	1844.098	135.077
23	1844.098	1915.121	1988.587	144.489
24	1988.587	2064.59	2143.227	154.64
25	2143.227	2224.597	2308.806	165.579
26	2308.806	2395.959	2486.169	177.364
27	2486.169	2579.551	2676.223	190.054
28	2676.223	2776.309	2879.937	203.713
29	2879.937	2987.238	3098.35	218.414
30	3098.35	3213.415	3332.579	234.229
31	3332.579	3455.993	3583.817	251.238
32	3583.817	3716.212	3853.348	269.531
33	3853.348	3995.399	4142.547	289.199

表7（完）

组	较低频率/赫兹	中心频率/赫兹	较高频率/赫兹	频率宽度/赫兹
k	$f_l[k]$	$f_c[k]$	$f_u[k]$	$f_w[k]$
34	4142.547	4294.979	4452.89	310.343
35	4452.89	4616.482	4785.962	333.072
36	4785.962	4961.548	5143.463	357.5
37	5143.463	5331.939	5527.217	383.754
38	5527.217	5729.545	5939.183	411.966
39	5939.183	6156.396	6381.463	442.281
40	6381.463	6614.671	6856.316	474.853
41	6856.316	7106.708	7366.166	509.85
42	7366.166	7635.02	7913.614	547.448
43	7913.614	8202.302	8501.454	587.84
44	8501.454	8811.45	9132.688	631.233
45	9132.688	9465.574	9810.536	677.849
46	9810.536	10168.013	10538.46	727.924
47	10538.46	10922.351	11320.175	781.715
48	11320.175	11732.438	12159.67	839.495
49	12159.67	12602.412	13061.229	901.56
50	13061.229	13536.71	14029.458	968.229
51	14029.458	14540.103	15069.295	1039.837
52	15069.295	15617.71	16186.049	1116.754
53	16186.049	16775.035	17385.42	1199.371
54	17385.42	17690.045	18000	614.58

频率到音高的映射通过以下小节中描述的算法完成，其中$Fsp[k_f]$分别是“外耳加权FFT输出”的能量表示：

$$Fsp[k_f, n] = \left|F_e[k_f, n]\right|^2 \quad (11)$$

或者误差信号的能量表示

$$Fsp[k_f, n] = \left|F_{noise}[k_f, n]\right|^2 \quad (12)$$

误差信号的计算见第3.4节。

这个处理阶段的输出是频率组的能量$P_e[k, n]$。

2.1.5.1 伪代码

/* 输入 */

Fsp[ ] 输入能量

/* 输出 */

Pe[ ]	：	音高映射能量
/* 中间值 */		
i	：	频率组索引
k	：	\ fft线路索引
Z		频率组数量： 基本版本为109 高级版本为55
fl[]	：	频率组的较低频率
fu[]	：	频率组的较高频率
Fres	：	频率分辨率常数

```
Fres = 48000/2048;
对于(i=0; i<Z; i++ )
{
 Pe[i]=0;
对于(k=0;k<1024;k++)
 {
  /* 频率组内的线路 */
  如果( (( k-0.5)*Fres >= fl[i])  && ((k+0.5)*Fres <= fu[i]))
  {
   Pe[i] += Fsp[k];
  }
  /* 内部频率组*/
  否则，如果( (( k-0.5)*Fres < fl[i])  && ((k+0.5)*Fres > fu[i]))
  {
   Pe[i] += Fsp[k]*(fu[i]-fl[i])/Fres;
  }
  /*左边框 */
  否则，如果( ((k-0.5)*Fres < fl[i]) && ((k+0.5)*Fres > fl[i]))
  {
   Pe[i] += Fsp[k]*( (k+0.5)*Fres - fl[i])/Fres;
  }
  /* 右边框
  否则，如果( ((k-0.5)*Fres < fu[i]) && ((k+0.5)*Fres > fu[i]);
```

```
    {
     Pe[i] += Fsp[k]*(fu[i]- (k-0.5)*Fres)/Fres;
    }
    /* 频率组外的线路 */
    否则
    {
     Pe[i] += 0;
    }
  }

  /* 限制结果 */
   Pe[i]=max(Pe[i],0.000000000001);
}
```

2.1.6 增加内部噪声

频率相关偏移P_{Thres}被添加到每个频率组中的能量：

$$P_{Thres}[k] = 10^{0.4\cdot 0.364\cdot\left(\frac{f_c[k]}{\text{kHz}}\right)^{-0.8}} \quad (13)$$

$$P_p[k,n] = P_e[k,n] + P_{Thres}[k] \quad (14)$$

此处理阶段的输出$P_p[k,n]$被称为“音高模式”。

2.1.7 扩展

音高模式$P_p[k,n]$使用电平相关扩展函数在频率上进行模糊。扩展函数是双边指数函数。较低的斜率始终为27分贝/巴克，较高的斜率取决于频率和能量。

斜率根据以下公式计算：

$$\frac{S_u[k,L[k,n]]}{\text{dB}/Bark} = -24 - \frac{230\ \text{Hz}}{f_c[k]} + 0.2 \cdot L[k,n]/\text{dB} \quad (15)$$

$$S_l[k,L[k,n]] = 27\ \frac{\text{dB}}{Bark} \quad (16)$$

使用：

$$L[k,n] = 10 \cdot \log_{10}\left(P_p[k,n]\right)$$

对每个频率组k独立地进行扩展：

$$E_2[k,n] = \frac{1}{Norm_{SP}[k]}\left(\sum_{j=0}^{Z-1} E_{line}[j,k,n]^{0.4}\right)^{\frac{1}{0.4}} \quad (17)$$

其中E_{line}由下式给出：

$$E_{line}[j,k,n]=\begin{cases}\dfrac{10^{\frac{L[j,n]}{10}}\cdot 10^{\frac{-res\cdot(j-k)\cdot s_l[j,L[j,n]]}{10}}}{\sum_{\mu=0}^{j-1}10^{\frac{-res\cdot(j-\mu)\cdot s_l[j,L[j,n]]}{10}}+\sum_{\mu=j}^{Z-1}10^{\frac{res\cdot(\mu-j)\cdot s_u[j,L[j,n]]}{10}}} & \text{if } k<j\\ \dfrac{10^{\frac{L[j,n]}{10}}\cdot 10^{\frac{res\cdot(k-j)\cdot s_u[j,L[j,n]]}{10}}}{\sum_{\mu=0}^{j-1}10^{\frac{-res\cdot(j-\mu)\cdot s_l[j,L[j,n]]}{10}}+\sum_{\mu=j}^{Z-1}10^{\frac{res\cdot(\mu-j)\cdot s_u[j,L[j,n]]}{10}}} & \text{if } k\geq j\end{cases} \quad (18)$$

$Norm_{SP}[k]$的计算公式如下：

$$Norm_{SP}[k]=\left(\sum_{j=0}^{Z-1}\tilde{E}_{line}[j,k]^{0.4}\right)^{\frac{1}{0.4}} \quad (19)$$

使用：

$$\tilde{E}_{line}[j,k]=\begin{cases}\dfrac{10^{\frac{-res\cdot(j-k)\cdot s_l[j,0]}{10}}}{\sum_{\mu=0}^{j-1}10^{\frac{-res\cdot(j-\mu)\cdot s_l[j,0]}{10}}+\sum_{\mu=j}^{Z-1}10^{\frac{res\cdot(\mu-j)\cdot s_u[j,0]}{10}}} & \text{if } k<j\\ \dfrac{10^{\frac{res\cdot(k-j)\cdot s_u[j,0]}{10}}}{\sum_{\mu=0}^{j-1}10^{\frac{-res\cdot(j-\mu)\cdot s_l[j,0]}{10}}+\sum_{\mu=j}^{Z-1}10^{\frac{res\cdot(\mu-j)\cdot s_u[j,0]}{10}}} & \text{if } k\geq j\end{cases} \quad (20)$$

*res*是音高标度的分辨率，单位为巴克（基本版本为0.25，高级版本为0.5）。

该处理阶段的模式$E_2[k,n]$稍后用于计算调制模式，并被称为“未模糊的激发模式”。

2.1.8 时域扩展

为了模拟正向掩蔽，每个频率组中的能量随着时间的推移被一阶低通滤波器抹平。时间常数取决于各组的中心频率（如等式（10）和表6所示），并根据以下公式计算：

$$\tau=\tau_{min}+\frac{100\text{ Hz}}{f_c[k]}\cdot(\tau_{100}-\tau_{min}) \quad \left|\begin{array}{l}\tau_{100}=0.030\text{ s}\\ \tau_{min}=0.008\text{ s}\end{array}\right. \quad (21)$$

一阶低通滤波器的计算公式如下：

$$E_f[k,n] = a \cdot E_f[k,n-1] + (1-a) \cdot E_2[k,n] \quad (22)$$

$$E[k,n] = \max\left(E_f(k,n), E_2(k,n)\right) \quad (23)$$

其中a由上述时间常数通过下式计算得出：

$$a = \mathrm{e}^{-\frac{4}{187.5} \cdot \frac{1}{\tau}} \quad (24)$$

n是实际帧数，k是组索引，$E_f[k,0] = 0$。

在这个处理阶段的模式$E[k,n]$被称为“激发模式”。

2.1.9 掩蔽阈值

掩蔽描述了当相应的较大信号出现时，较弱但清晰可闻的信号变得听不见的效果。这个阈值是通过用加权函数$m[k]$对激发模式进行加权来计算的。

$$m[k] = \begin{cases} 3.0 & \text{对于} \quad k \cdot \mathrm{res} \leq 12 \\ 0.25 \cdot k \cdot \mathrm{res} & \text{对于} \quad k \cdot \mathrm{res} > 12 \end{cases} \quad (25)$$

$$M[k,n] = \frac{E[k,n]}{10^{\frac{m[k]}{10}}} \quad (26)$$

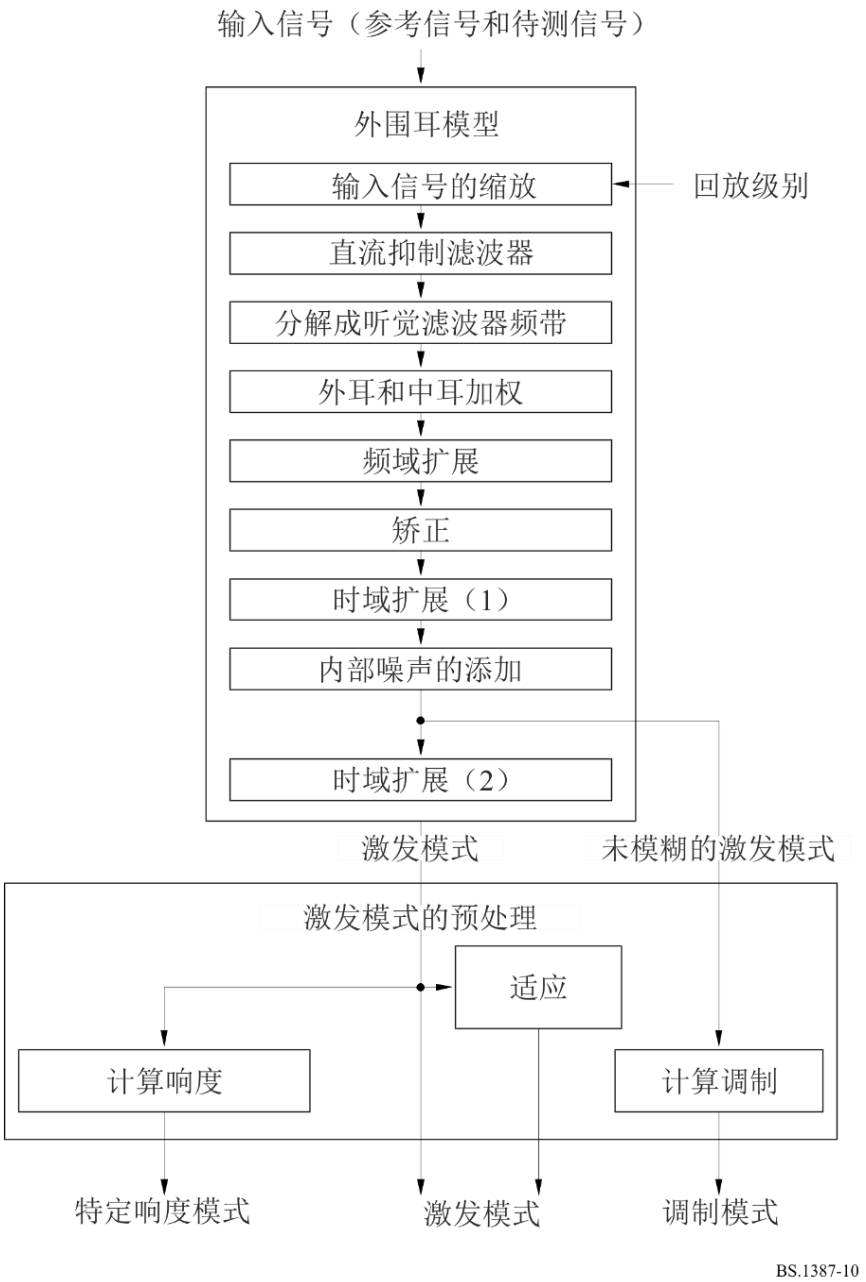
在这个处理阶段的模式$M[k,n]$被称为“掩蔽模式”。

2.2 基于滤波器组的人耳模型

2.2.1 概述

图10

外围耳模型和该模型基于滤波器组的部分的激发模式的预处理



在基于滤波器组的人耳模型的输入端，待测信号和参考信号被调整到假设的回放水平，并通过高通滤波器发送，以去除信号的DC和亚音速分量。然后，信号被线性相位滤波器分解为带通信号，该线性相位滤波器在感知音高范围内均匀分布。对带通信号进行频率相关加权，以模拟外耳和中耳的频谱特性。听觉滤波器的电平相关频谱分辨率通过具有电平相关扩展函数的输出的频域卷积来建模。

使用带通信号的希尔伯特变换（“校正”）来计算信号的包络，并且应用具有窗口函数的时域卷积，以便对反向掩蔽建模。然后，添加与频率相关的偏移，该偏移考虑了听觉系统中的内部噪声，并对安静时的阈值建模。最后，使用说明正向掩蔽的指数扩展函数进行第二次时域卷积。

此时获得的激发模式用于计算特定响度模式，并且最终时域扩展之前的模式（“未模糊的激发模式”）用于计算调制模式。上述模式以及激发模式本身是计算模型值的基础。为了将待测设备的稳态频率响应的影响与其他失真分开，待测信号和参考信号的激发模式亦在频谱上相互适应（“适应”）。调制模式和特定响度模式是根据适应的和非适应的激发模式计算的。

2.2.2 二次采样

在滤波器组的输出端，信号以因数32进行向下采样，并且在第一时域扩展之后，信号以因数6进行向下采样（见图11）。

2.2.3 播放级别的设置

输入的比例因子通过满量程输入信号的假设回放电平计算得出：

$$fac = \frac{10^{Lmax/20}}{32767} \quad (27)$$

如果不知道确切的回放电平，建议将L_{max}设置为92 dB_{SPL}。

2.2.4 DC抑制滤波器

由于滤波器组对输入信号中的次声敏感，因此对输入信号应用了DC抑制滤波器。使用截止频率为20赫兹的四阶巴特沃兹高通滤波器。该滤波器由两个二阶IIR滤波器级联而成。

$$y_n = x_n - 2x_{n-1} + x_{n-2} + b_1 y_{n-1} + b_2 y_{n-2} \quad (28)$$

其中第一块的系数为：

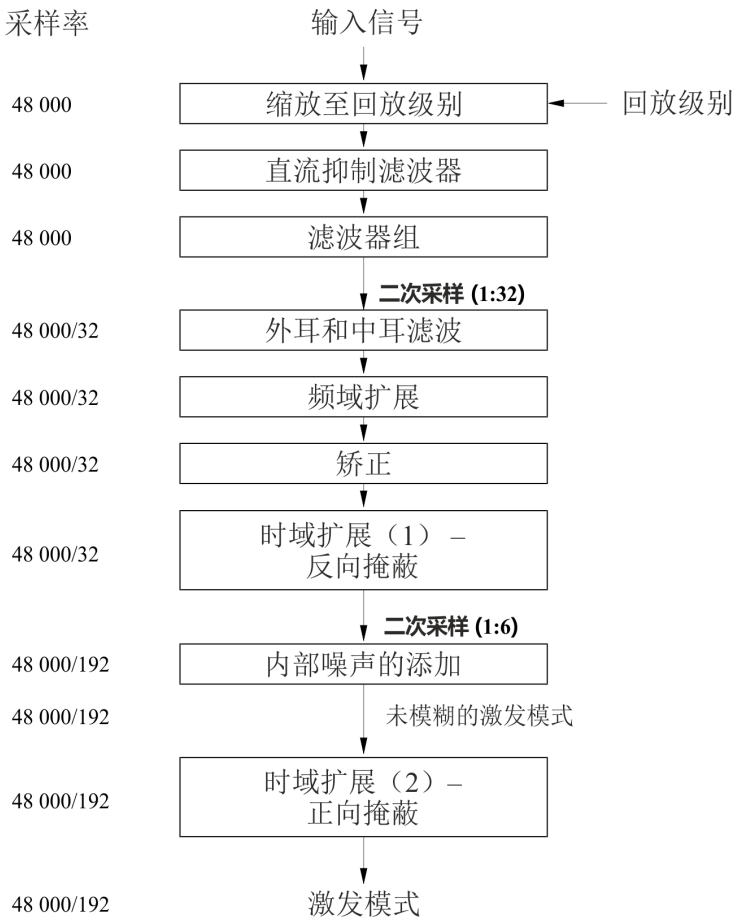
$$b_{1,2} = 1.99517, -0.995174$$

第二块的系数为：

$$b_{1,2} = 1.99799, -0.997998$$

图11

基于滤波器组的外围耳模型中的二次采样



2.2.5 滤波器组

滤波器组由40个滤波器对组成，用于待测信号和参考信号的每个通道。当与听觉音高标度相关时，滤波器是等间距的，并且具有恒定的绝对带宽。每对滤波器由两个频率响应相同但相位响应相差90°的滤波器组成。因此，第二滤波器的输出代表第一滤波器输出的希尔伯特变换（或者虚部，前提是假设第一滤波器代表复信号的实部）。其脉冲响应的包络具有$\cos^2$形状。滤波器由表8和等式（29）定义（其中k是滤波器的索引，n是时间样本的索引，T是两个样本之间的时间：$T = 1/48\,000$）。它们可以使用值$h_{re}(k,n)$和$h_{im}(k,n)$作为系数来实现为FIR滤波器。当输入信号受时间限制时，滤波器输出亦可以通过非常快速的递归算法来计算。

$$\left.\begin{aligned} h_{re}(k,n) &= \frac{4}{N[k]} \cdot \sin^2\left(\pi \cdot \frac{n}{N[k]}\right) \cdot \cos\left(2\pi \cdot f_c[k] \cdot \left(n - \frac{N[k]}{2}\right) \cdot T\right) \\ h_{im}(k,n) &= \frac{4}{N[k]} \cdot \sin^2\left(\pi \cdot \frac{n}{N[k]}\right) \cdot \sin\left(2\pi \cdot f_c[k] \cdot \left(n - \frac{N[k]}{2}\right) \cdot T\right) \end{aligned}\right| \; 0 \le n < N[k]$$

$$h_{re}(k,n) = h_{im}(k,n) = 0 \quad \left| \begin{array}{l} n < 0 \\ n \ge N[k] \end{array} \right. \qquad (29)$$

表8

每个滤波器的中心频率、脉冲响应长度和附加延迟

滤波器索引	中心频率/赫兹	脉冲响应/样本长度	额外延迟/样本
k	$f_C[k]$	*N[k]*	*D[k]*
0	50.00	1456	1
1	116.19	1438	10
2	183.57	1406	26
3	252.82	1362	48
4	324.64	1308	75
5	399.79	1244	107
6	479.01	1176	141
7	563.11	1104	177
8	652.97	1030	214
9	749.48	956	251
10	853.65	884	287
11	966.52	814	322
12	1089.25	748	355
13	1223.10	686	386
14	1369.43	626	416
15	1529.73	570	444
16	1705.64	520	469
17	1898.95	472	493
18	2111.64	430	514
19	2345.88	390	534
20	2604.05	354	552
21	2888.79	320	569
22	3203.01	290	584
23	3549.90	262	598
24	3933.02	238	610
25	4356.27	214	622
26	4823.97	194	632
27	5340.88	176	641
28	5912.30	158	650
29	6544.03	144	657
30	7242.54	130	664
31	8014.95	118	670
32	8869.13	106	676
33	9813.82	96	681
34	10858.63	86	686

表8（完）

滤波器索引	中心频率/赫兹	脉冲响应/样本长度	额外延迟/样本
k	$f_C[k]$	*N[k]*	*D[k]*
35	12014.24	78	690
36	13292.44	70	694
37	14706.26	64	697
38	16270.13	58	700
39	18000.02	52	703

中心频率范围从50赫兹到18 000赫兹。听觉音高标度是根据[Schroeder等人，1979年]给出的近似值计算的：

$$z/Bark \;=\; 7 \cdot \operatorname{arsinh}\left(\frac{f/\mathrm{Hz}}{650}\right) \qquad (30)$$

音高单位被命名为“巴克”（尽管该标度并不完全代表[Zwicker和Feldtkeller，1967年]定义的巴克标度）。

为了使所有滤波器具有相等的延迟，每个滤波器的输入延迟*D*个样本，其中*D*是其脉冲响应长度与具有最长脉冲响应的滤波器的脉冲响应长度之差的一半[3]：

$$D[k] \;=\; 1 \;+\; \frac{1}{2}(N[0] \;-\; N[k]) \qquad (31)$$

滤波器输出以系数32进行二次采样，即所有滤波器的每第32个输入样本计算一次输出值[4]。

2.2.6 外耳和中耳滤波

外耳和中耳的频率响应由应用于滤波器输出的频率相关加权函数建模（等式（32））。

$$W[k]/\mathrm{dB} = -0.6 \cdot 3.64 \cdot \left(\frac{f_c[k]}{\mathrm{kHz}}\right)^{-0.8} + 6.5 \cdot \mathrm{e}^{-0.6 \cdot \left(\frac{f_c[k]}{\mathrm{kHz}} - 3.3\right)^2} - 10^{-3} \cdot \left(\frac{f_c[k]}{\mathrm{kHz}}\right)^{3.6} \qquad (32)$$

2.2.6.1 伪代码

/* 输入 */

out_re,out_im : 滤波器组输出（实部和虚部）

W : 加权函数（见等式（32））

/* 输出 */

out_re,out_im : 滤波器组输出

/* 中间值 */

k : 滤波器的索引

---

[3] 一个样本的额外延迟对于实施是不必要的。请注意，用于合规性测试的参考实施包括这一额外延迟。

[4] 实际上，高频带内滤波器的包络不一定满足采样定理。尽管错认假频伪信号（aliasing）只会在非常特殊的条件下发生（即，用大于1.5 kHz的频率调制的高频分量），并且在已知的数据库中从未发生过与此类效应相关的问题，但应该指出的是，错认假频伪信号问题可能会发生，尤其是在使用人工测试信号时。

Wt	:	权重因数

```
/* 外耳和中耳滤波 */
对于(k=0..39)
{
        Wt    = pow(10,W[k]/20)
        out_re[k] *= Wt;
        out_im[k]  *= Wt;
}
```

2.2.7 频域扩展

使用电平相关扩展函数在频率上模糊滤波器组的输出值。扩展函数是双边指数函数。较低的斜率始终为31分贝/巴克，较高的斜率在–24和–4分贝/巴克之间变化。

上斜率$s[k]$的计算公式如下：

$$\frac{s[k]}{\mathrm{dB}/Bark} = \min\left(-4, -24 - \frac{230\ \mathrm{Hz}}{f_c[k]} + 0.2 \cdot L[k]/\mathrm{dB}\right) \quad (33)$$

通过取滤波器输出的绝对值平方并将其转换为分贝标度，可以独立计算每个滤波器通道的电平$L[k]$。中心频率$f_c[k]$取自表8。斜率的线性表示通过时间常数为100毫秒的一阶低通滤波器随时间平滑化。

对代表信号实部的滤波器和代表信号虚部的滤波器独立地进行扩展（等式（29））。首先对（电平相关的）上斜率进行频率扩展，然后使用一阶IIR滤波器算法对下斜率进行频率扩展。

2.2.7.1 伪代码

/* 输入 */		
out_re,out_im	:	滤波器组输出（实部和虚部）
z[ ]	:	滤波器频带中心频率的临界频带速率，单位为巴克（根据表8和等式（30））
/* 输出 */		
A_re,A_im	:	输出模式
/* 中间值 */		
j,k	:	滤波器的索引
a,b	:	时间平滑系数
dist	:	用于计算串扰的常数
L[ ]	:	每个滤波器输出的电平
s[ ]	:	向上扩展的本地陡度
d1,d2	:	缓冲
/* 静态 */		（保留前一帧的值；测量开始时用零来对各值进行初始化）
cl, cu[ ]	:	信号的扩展部分

```
/* 电平相关的向上扩展 */
dist  = pow(0.1,(z[39]-z[0])/(39*20));
```

```
/* (z[39]-z[0])/39是两个相邻滤波频带之间的距离，单位为巴克 */
a    = exp(-32/(48000*0.1));
b    = 1 - a;
对于(k=0..39)
{
    A_re[k]    = out_re[k];
    A_im[k]   = out_im[k];
}
对于(k=0..39)
{
    /* 计算电平相关斜率 */
    L[k]  = 10*log10(out_re[k]* out_re[k] + out_im[k]* out_im[k];
    s[k]   = max(4,(24 + 230/fcentre[k] - 0.2*L[k]));
    /* 计算扩展分数并随时间平滑*/
    cu[k] = a*pow(dist,s[k])+b*cu[k];
    /* 频带k的扩展 */
    d1     = out_re[k]
    d2     = out_im[k]
    对于(j=k+1..39)
    {
     d1   *= cu[k];
     d2   *= cu[k];
     A_re[j]  += d1;
     A_im[j] += d2;
    }
}
/* 向下扩展 */
cl   = pow(dist,31);
d1  = 0;
d2  = 0;
对于(k=39..0)
{
    /* 频带k的扩展 */
    d1    = d1 * cl + A_re[k];
    d2    = d2 * cl + A_im[k];
    A_re[k]    = d1;
    A_im[k]   = d2;
}
```

2.2.8 矫正

滤波器输出端的能量通过将代表信号实部的滤波器平方值与代表信号虚部的滤波器平方值相加来计算。

$$E_0[k,n] = A_{re}[k,n]^2 + A_{im}[k,n]^2 \quad (34)$$

以下所有操作均为在这些能量上进行的。

2.2.9 时域模糊（1）– 反向掩蔽

为了对反向掩蔽建模，滤波器输出端的能量随着时间的推移被具有12抽头的$\cos^2$整形脉冲响应的FIR滤波器抹平（对应于滤波器组输入采样率下384个样本的滤波器响应）。在时间模糊之后，输出以因子6进行向下采样。结果值乘以校准系数$cal_1 = 0.9761$，以获得给定回放电平的适当输出电平。

$$E_1[k,n] = \frac{0.9761}{6} \cdot \sum_{i=0}^{11} E_0[k,6n-i] \cdot \cos^2\left(\pi \cdot \frac{(i-5)}{12}\right) \quad (35)$$

2.2.10 内部噪声的添加

在第一时域扩展之后，频率相关偏移E_{Thres}被添加到每个滤波器通道中的能量。

$$E_{Thres}[k] = 10^{0.4 \cdot 0.364 \cdot \left(\frac{f_c[k]}{\text{kHz}}\right)^{-0.8}} \quad (36)$$

$$E_2[k,n] = E_1[k,n] + E_{Thres}[k,n] \quad (37)$$

该处理阶段的模式$E_2[k,n]$稍后用于计算调制模式，并被称为“*未模糊的激发模式*”。

2.2.11 时域模糊（2）– 正向掩蔽

为了对正向掩蔽建模，每个滤波器通道中的能量随着时间的推移被一阶低通滤波器抹平。时间常数取决于每个滤波器的中心频率（如表5所示），计算公式如下：

$$\tau = \tau_{min} + \frac{100\ \text{Hz}}{f_c[k]} \cdot (\tau_{100} - \tau_{min}) \quad \left| \begin{array}{l} \tau_{100} = 0.020\ \text{s} \\ \tau_{min} = 0.004\ \text{s} \end{array} \right. \quad (38)$$

一阶低通滤波器的计算公式如下：

$$E[k,n] = a \cdot E[k,n-1] + (1-a) \cdot E_2[k,n] \quad (39)$$

其中a由上述时间常数通过下式计算得出：

$$a = \text{e}^{-\frac{192}{48\,000 \cdot \tau}} \quad (40)$$

在这个处理阶段的模式$E[k,n]$被称为“*激发模式*”。

3 激发模式的预处理

本节中描述的大多数计算与基于滤波器组的人耳模型以及基于FFT的人耳模型一起使用。由于两种人耳模型的二次采样因子和频带数量不同，因此使用人耳模型因变量*StepSize*和*Z*来描述取决于该因子的常数。对于基于FFT的人耳模型，*StepSize*值为*1 024*，*Z*为*55*（高级版本）或*109*（基本版本）。对于基于滤波器组的人耳模型，*StepSize*的值为*192*，Z为*40*。如果没有另外给出，所有变量和递归滤波器均被初始化为零。

3.1 电平和模式适应

为了补偿测试信号和参考信号之间的电平差和线性失真，对测试信号和参考信号的平均电平进行相互适应。

在第一步中，通过一阶低通滤波器平滑每个滤波器通道中的能量。时间常数取决于滤波器的中心频率，选择如下：

$$\tau = \tau_{min} + \frac{100\ \text{Hz}}{f_c[k]} \cdot (\tau_{100} - \tau_{min}) \quad \left| \begin{array}{l} \tau_{100} = 0.050\ \text{s} \\ \tau_{min} = 0.008\ \text{s} \end{array} \right. \tag{41}$$

一阶低通滤波器的计算公式如下：

$$P_{Ref}[k,n] = a \cdot P_{Ref}[k,n-1] + (1-a) \cdot E_{Ref}[k,n] \tag{42}$$

$$P_{Test}[k,n] = a \cdot P_{Test}[k,n-1] + (1-a) \cdot E_{Test}[k,n] \tag{43}$$

其中E_{Test}和E_{Ref}是要相互适应的激发模式，a由时间常数通过下式计算：

$$a = \mathrm{e}^{-\frac{StepSize}{48\,000 \cdot \tau}} \tag{44}$$

3.1.1 电平适应

根据低通输入模式P_{Test}和P_{Ref}，瞬时校正系数*LevCorr*计算如下：

$$LevCorr[n] = \left(\frac{\sum_{k=0}^{Z-1} \sqrt{P_{Test}[k,n] \cdot P_{Ref}[k,n]}}{\sum_{k=0}^{Z-1} P_{Test}[k,n]} \right)^2 \tag{45}$$

如果校正系数大于1，则参考信号除以校正系数，否则测试信号乘以校正系数。

$$E_{L,Ref}[k,n] = E_{Ref}[k,n] / LevCorr[n] \ | LevCorr[n] > 1 \tag{46}$$

$$E_{L,Test}[k,n] = E_{Test}[k,n] \cdot LevCorr[n] \ | LevCorr[n] \leq 1 \tag{47}$$

3.1.2 模式适应

通过比较测试和参考信号的滤波器输出的时间包络来计算每个通道的校正因子：

$$R[k,n]=\frac{\sum_{i=0}^{n} a[k]^i \cdot E_{L,Test}[k,n-i]\cdot E_{L,Ref}[k,n-i]}{\sum_{i=0}^{n} a[k]^i \cdot E_{L,Ref}[k,n-i]\cdot E_{L,Ref}[k,n-i]} \quad (48)$$

根据等式（41）给出的时间常数，如上计算a的值（等式（44））。如果$R[k,n]$大于1，则测试信号的校正系数设置为$R[k,n]^{-1}$，参考信号的校正系数设置为1。在相反的情况下，参考信号的校正因子被设置为$R[k,n]$，测试信号的校正因子被设置为1。

$$\begin{aligned} &R_{Test}[k,n]=\frac{1}{R[k,n]}, \quad R_{Ref}[k,n]=1 \quad \big| \; R[k,n]\geq 1 \\ &R_{Test}[k,n]=1, \quad R_{Ref}[k,n]=R[k,n] \quad \big| \; R[k,n]<1 \end{aligned} \quad (49)$$

如果（48）的分母为0（因此$R[k,n]$未定义），并且分子大于0，则将$R_{Test}[k,n]$设置为0，并将$R_{Ref}[k,n]$设置为1。当（48）的分子亦为0时，比率$R_{Test}[k,n]$和$R_{Ref}[k,n]$从下面的频带复制。如果不存在下面的频带（即$k=0$），则比率$R_{Test}[k,n]$和$R_{Ref}[k,n]$设置为1。

校正因子在M个滤波器通道上平均，并使用与上面给出的相同时间常数（等式（41）至（44））随时间平滑（等式（50））。对于基于滤波器组的人耳模型，频率窗口M的宽度是3。对于基于FFT的人耳模型，它分别是4（高级版本）或8（基本版本）。

$$\begin{aligned} &PattCorr_{Test}[k,n]=a\cdot PattCorr_{Test}[k,n-1]+(1-a)\cdot\frac{1}{M}\cdot\sum_{i=-M_1}^{M_2} R_{Test}[k+i,n] \\ &PattCorr_{Ref}[k,n]=a\cdot PattCorr_{Ref}[k,n-1]+(1-a)\cdot\frac{1}{M}\cdot\sum_{i=-M_1}^{M_2} R_{Ref}[k+i,n] \end{aligned} \quad (50)$$

$$\begin{cases} M_1=M_2=\dfrac{M-1}{2} & \big| \; M \text{ odd} \\ M_1=\dfrac{M}{2}-1, \quad M_2=\dfrac{M}{2} & \big| \; M \text{ even} \end{cases}$$

在频率窗口超出滤波器频带范围的频率标度边界处，频率窗口的宽度相应减小：

$$M_1=\min(M_1,k), \qquad M_2=\min(M_2,z-k-1), \qquad M=M_1+M_2+1 \quad (51)$$

用相应的校正因子$PattCorr_{Test/Ref}[k,n]$对电平适应的输入模式进行加权，以便获得频谱适应的模式。

$$E_{P,Ref}[k,n] = E_{L,Ref}[k,n]\cdot PattCorr_{Ref}[k,n] \quad (52)$$

$$E_{P,Test}[k,n] = E_{L,Test}[k,n]\cdot PattCorr_{Test}[k,n] \quad (53)$$

3.2 调制

根据*未模糊的激发模式*$E_2[k,n]$，通过将激发提高到0.3的幂来计算简化的响度。该值及其时间导数的绝对值随着时间的推移而抹平。

$$\bar{E}_{der}[k,n]=a\cdot\bar{E}_{der}[k,n-1]+(1-a)\cdot\frac{48\,000}{StepSize}\cdot\left|E_2[k,n]^{0.3}-E_2[k,n-1]^{0.3}\right| \quad (54)$$

$$\bar{E}[k,n] = a\cdot\bar{E}[k,n-1]+(1-a)\cdot E_2[k,n]^{0.3} \quad (55)$$

a的值按照（44）中的方法由下式给出的时间常数计算：

$$\tau = \tau_0 + \frac{100\ \text{Hz}}{f_c} \cdot (\tau_{100} - \tau_0) \left| \begin{matrix} \tau_{100} = 0.050\ s \\ \tau_0 = 0.008\ s \end{matrix} \right. \quad (56)$$

根据结果值，计算每个滤波器输出的包络调制测量值：$\bar{E}_{der}$和$\bar{E}$,

$$Mod[k,n] = \frac{\bar{E}_{der}[k,n]}{1 + \bar{E}[k,n]/0.3} \quad (57)$$

$\bar{E}$的值还将在以后计算调制差时使用。

3.3 响度

根据以下公式计算待测信号和参考信号的特定响度模式：

$$N[k,n] = const \cdot \left(\frac{1}{s[k]} \cdot \frac{E_{Thres}[k]}{10^4} \right)^{0.23} \cdot \left[\left(1 - s[k] + \frac{s[k] \cdot E[k,n]}{E_{Thres}[k]} \right)^{0.23} - 1 \right] \quad (58)$$

如[Zwicker和Feldtkeller，1967年]所述。待测信号和参考信号的总响度被计算为所有滤波器通道上大于0的所有特定响度值的总和。

$$N_{total}[n] = \frac{24}{Z} \cdot \sum_{k=0}^{Z-1} \max\left(N[k,n], 0\right) \quad (59)$$

对于基于FFT的外围耳模型，缩放常数选择为$const = 1.07664$，对于基于滤波器组的外围耳模型，缩放常数选择为$const = 1.26539$，以便对于1 kHz的40 dB_{SPL}正弦音调给出1宋（sone）的整体响度。阈值指数s和阈值E_{Thres}处的激发分别根据以下公式计算：

$$E_{Thres}[k] = 10^{0.364 \cdot \left(\frac{f}{1\ \text{kHz}}\right)^{-0.8}} \quad (60)$$

和

$$s[k] = 10^{\frac{1}{10}\left(-2 - 2.05 \cdot atn\left(\frac{f}{4\ \text{kHz}}\right) - 0.75 \cdot atn\left(\left(\frac{f}{1\ 600\ \text{Hz}}\right)^2\right)\right)} \quad (61)$$

注 – 由于不同的外围耳模型，此处计算的响度与ISO 532定义的响度不同（声学 – 1975年响度等级计算方法）。

3.4 误差信号的计算

误差信号仅在基于FFT的模型中计算。通过取参考信号和测试信号的外耳和中耳滤波幅度谱之间的差值，在频域中计算误差信号（见第2.1.4节）。

$$F_{noise}[k_f, n] = \left| \left| F_{eref}[k_f, n] \right| - \left| F_{etest}[k_f, n] \right| \right| \quad (62)$$

使用第2.1.5节中描述的算法将F_{noise}映射到音高域。

该算法的输出$P_{noise}[n,k]$被称为“噪声模式”。

4 模型输出变量的计算

4.1 概述

表9

用于预测基本音频质量的模型输出变量概览

模型输出变量（MOV）	计算采用的人耳模型		所用版本	
	快速傅里叶变换（FFT）	滤波器组	基本	高级
$WinModDiff1_B$	是	否	是	否
$AvgModDiff1_B$	是	否	是	否
$AvgModDiff2_B$	是	否	是	否
$RmsModDiff_A$	否	是	否	是
$RmsNoiseLoud_B$	是	否	是	否
$RmsNoiseLoudAsym_A$	否	是	否	是
$AvgLinDist_A$	否	是	否	是
$BandwidthRef_B$	是	否	是	否
$BandwidthTest_B$	是	否	是	否
Total NMR_B	是	否	是	否
$RelDistFrames_B$	是	否	是	否
Segmental NMR_B	是	否	否	是
$MFPD_B$	是	否	是	否
ADB_B	是	否	是	否
EHS_B	是	否	是	是

4.2 调制差

通过计算每个滤波器通道的本地调制差（等式（63））来测量待测信号和参考信号的时间包络的调制差，其中Mod_{test}和Mod_{Ref}是通过将等式（57）应用于参考R_{test}信号而得到的。

$$ModDiff[k,n] = w \cdot \frac{\left|Mod_{test}[k,n] - Mod_{Ref}[k,n]\right|}{offset + Mod_{Ref}[k,n]}$$

$$\begin{cases} w = 1.0 & Mod_{test}[k,n] > Mod_{Ref}[k,n] \\ w = negWt & Mod_{test}[k,n] < Mod_{Ref}[k,n] \end{cases} \qquad (63)$$

瞬时调制差被计算为所有滤波器通道上的本地调制差的平均值（等式（64））。

$$ModDiff[n] = \frac{100}{Z} \sum_{k=0}^{Z-1} ModDiff[k,n] \qquad (64)$$

安静时的阈值通过电平相关加权因子（等式（65））来考虑，该加权因子是根据等式（55）中给出的参考信号的经修改激发模式以及等式（36）中定义的基于滤波器组的人耳模型和等式（13）中定义的基于FFT的人耳模型的内部噪声函数来计算的。

$$TempWt[n] = \sum_{k=0}^{Z-1} \frac{\overline{E}_{ref}[k,n]}{\overline{E}_{ref}[k,n] + levWt \cdot E_{Thres}[k]}^{\wedge 0.3} \qquad (65)$$

第5.2节描述了使用加权因子*TempWt[n]*对瞬时调制差*ModDiff[n]*进行时间平均。表10给出了常量*negWt*、*offset*和*levWt*的值。

表10

估算总体调制差的模型输出变量

MOV (Xxx=Win/Avg/Rms)	negWt	offset	levWt
XxxModDiff1B	1	1	100
XxxModDiff2B	0.1	0.01	100
XxxModDiffA	1	1	1

4.2.1 $RmsModDiff_A$

模型输出变量$RmsModDiff_A$是根据基于滤波器组的人耳模型计算的调制差的平方平均值。时间平均见第5.2.2节，常数见表10。

4.2.2 $WinModDiff1_B$

模型输出变量$WinModDiff1_B$是根据基于FFT的人耳模型计算的调制差的窗口平均值。时间平均见第5.2.3节，常数见表10。等式（65）中给出的时间加权因子不适用于该MOV。

4.2.3 $AvgModDiff1_B$和$AvgModDiff2_B$

模型输出变量$AvgModDiff1_B$和$AvgModDiff2_B$是根据基于FFT的人耳模型计算的调制差的线性平均值。$AvgModDiff2_B$和$AvgModDiff1_B$的区别在于常数的选择不同。时间平均见第5.2.1节，常数见表10。

4.3 噪声响度

这些模型输出变量估算存在掩蔽参考信号时附加失真的部分响度。部分响度的公式（等式（66））被设计为根据[Zwicker和Feldtkeller，1967年]在不存在掩蔽声的情况下产生噪声的特定响度，并且在噪声与掩蔽声相比非常小的情况下产生类似于噪声与掩蔽声之间的比率的值。

部分噪声响度根据以下公式计算：

$$NL[k,n] = \left(\frac{1}{s_{test}} \cdot \frac{E_{Thres}}{E_0}\right)^{0.23} \cdot \left[\left(1 + \frac{max(s_{test} \cdot E_{test} - s_{ref} \cdot E_{ref}, 0)}{E_{Thres} + s_{ref} \cdot E_{ref} \cdot \beta}\right)^{0.23} - 1\right] \qquad (66)$$

其中E_0始终为1，E_{Thres}是（36）中定义的内部噪声函数$E_{Thres}[k]$，s的计算公式如下：

$$s = ThresFac_0 \cdot Mod[k,n] + S_0 \qquad (67)$$

当没有不同描述时，频谱适应激发模式（见第3.1节）用作输入：$E_{Test} = E_{P,Test}[k,n]$和$E_{Ref} = E_{P,Ref}[k,n]$。决定掩蔽量的系数β计算如下：

$$\beta = \exp\left(-\alpha \cdot \frac{E_{test} - E_{ref}}{E_{ref}}\right) \qquad (68)$$

直到左或右音频通道的总响度一度超过测试和参考信号的$N_{Thres} = 0.1\ sone$值后50毫秒，才考虑瞬时噪声响度值（见第5.2.4.2节）。

在频谱平均中，瞬时值通过每个临界频带的滤波器频带数而非滤波器频带总数进行归一化，即频谱平均的结果乘以因数24。

如果瞬时噪声响度低于阈值NL_{min}，则将其设置为0。

表11

估算整体噪声响度的模型输出变量

MOV (Xxx=Win/Avg/Rms)	α	$ThresFac_0$	S_0	NL_{min}
$XxxMissingComponents_B$	1.5	0.15	1	0
$XxxNoiseLoud_B$	1.5	0.15	0.5	0
$XxxMissingComponents_A$	1.5	0.15	1	0
$XxxNoiseLoud_A$	2.5	0.3	1	0.1
$XxxLinDist_A$	1.5	0.15	1	0

4.3.1 $RmsNoiseLoud_A$

模型输出变量$RmsNoiseLoud_A$是根据基于滤波器组的人耳模型计算的噪声响度的平方平均值。时间平均见第5.2.2节，常数见表11。

4.3.2 $RmsMissingComponents_A$

模型输出变量$RmsMissingComponents_A$是根据基于滤波器组的人耳模型计算的噪声响度的平方平均值。它是用互换的测试和参考信号的频谱适应激发模式来计算的，以便产生在测试信号中丢失的参考信号分量的响度。时间平均见第5.2.2节，常数见表11。

4.3.3 $RmsNoiseLoudAsym_A$

模型输出变量$RmsNoiseLoudAsym_A$是噪声响度（见第4.3.1节）和丢失信号分量响度（见第4.3.2节）的平方平均值的加权和，两者均为根据基于滤波器组的人耳模型计算的。

$$RmsNoiseLoudAsym = RmsNoiseLoud + 0.5 \cdot RmsMissingComponents \qquad (69)$$

4.3.4 $AvgLinDist_A$

模型输出变量$AvgLinDist_A$测量在待测信号和参考信号的频谱适应期间丢失的信号分量的响度。它使用参考信号的频谱适应激发作为参考信号，使用参考信号的非适应激发作为测试信号。此MOV是从基于滤波器组的人耳模型中计算出来的。时间平均见第5.2.1节，常数见表11。

4.3.5 $RmsNoiseLoud_B$

模型输出变量$RmsNoiseLoud_B$是根据基于FFT的人耳模型计算的噪声响度的平方平均值。时间平均见第5.2.2节，常数见表11。

4.4 带宽

这些模型输出值估算FFT线路中待测信号和参考信号的平均带宽。

对于每个帧，根据下面的伪代码计算本地带宽$Bw_{Ref}[n]$和$Bw_{Test}[n]$。

4.4.1 伪代码

/* 输入 */

FLevRef[], FlevelTest[] ： FFT输出电平（分贝）

/* 输出 */

BwRef, BwTest ： 输出模式

/* 中间值 */

k ： FFT线路索引

ZeroThreshold （零阈值） ： 带宽阈值

```
ZeroThreshold = FLevelTst(921);
BwRef = BwTst = 0.0;
对于(k=921;k<1024;k++)
{
 ZeroThreshold=max(ZeroThreshold,FLevelTst(k));
}

 对于 (k = 920; k>=0; k--)
 {
   如果 (FLevelRef[k] >= 10.0+ZeroThreshold)
  {
      BwRef = k+1;
      break;
  }
 }
 对于 (k = BwRef-1; k>=0; k--)
 {
   如果(FLeveltest[k] >= 5.0+ZeroThreshold)
  {
      BwTest=k+1;
      break;
  }
 }
```

4.4.2 $BandwidthRef_B$和$BandwidthTest_B$

$BandwidthRef_B$是BwRef的线性平均值，$BandwidthTest_B$是BwTest的线性平均值。求平均值时，仅考虑BwRef > 346的帧。项目开头和结尾的低能量帧被忽略（见第5.2.4.4节）。时间平均见第5.2.1节。

4.5 噪声掩蔽比

以下模型值是根据噪声和掩蔽值计算的。

当前帧n的本地NMR定义为：

$$NMR_{local}[n] = 10 * \log_{10} \frac{1}{Z} \sum_{k=0}^{Z-1} \frac{P_{noise}[k,n]}{M[k,n]} \qquad (70)$$

4.5.1 Total NMR_B（总NMR_B）

模型输出变量$Total\ NMR_B$是噪声掩蔽比的线性平均值，使用

$$NMR_{tot} = 10 * \log_{10} \frac{1}{N} \sum_{n} \left(\frac{1}{Z} \sum_{k=0}^{Z-1} \frac{P_{noise}[k,n]}{M[k,n]} \right) \qquad (71)$$

项目开头和结尾的低能量帧被忽略（见第5.2.4.4节）。

4.5.2 Segmental NMR_B（分段NMR_B）

模型输出变量$Segmental\ NMR_B$是本地NMR的线性平均值。时间平均见第5.2.1节。

项目开头和结尾的低能量帧被忽略（见第5.2.4.4节）。

4.6 Relative Disturbed $Frames_B$（相对干扰帧$_B$）

模型输出变量相对干扰帧$_B$（缩写：$RelDistFrames_B$）表示具有以下特征的帧的数量：

$$\max_{\forall k} \left(10 \cdot \log \left(\frac{P_{noise}[k,n]}{M[k,n]} \right) \right) \geq 1.5 \text{ dB} \qquad k \in [0, Z-1]$$

与项目的总帧数相关。

项目开头和结尾的低能量帧被忽略（见第5.2.4.4节）。

4.7 检测概率

本节定义的模型输出变量基于 $\tilde{E}[k,n]$（k频带，n帧），即以分贝表示的激发模式$E[k,n]$：

$$\tilde{E}[k,n] = 10 \cdot \log_{10}(E[k,n]) \qquad (72)$$

对于每一帧n：

对每个通道c独立执行以下步骤（c的值为左和右）。对数激发模式$\tilde{E}_{ref}[k,n]$和$\tilde{E}_{test}[k,n]$分别用于参考信号和待测信号。

对于每个频带k：

– 计算不对称平均激发。

$$L[k,n] = 0.3 \cdot \max(\tilde{E}_{ref}[k,n], \tilde{E}_{test}[k,n]) + 0.7 \cdot \tilde{E}_{test}[k,n] \quad (73)$$

– 计算有效检测步长s。以下公式是[Zwicker和Fastl，1990年]测量的最小可觉电平差的近似值。

如果$L[k,n] > 0$：

$$s[k,n] = 5.95072 \cdot ((6.39468)/L[k,n])^{1.71332} + 9.01033 \times 10^{-11} \cdot L[k,n]^4 + 5.05622 \times 10^{-6} \times L[k,n]^3 - 0.00102438 \cdot L[k,n]^2 + 0.0550197 \cdot L[k,n] - 0.198719$$

否则

$$s[k,n] = 1.0 \times 10^{30} \quad (74)$$

– 计算带符号误差e：

$$e[k,n] = \tilde{E}_{ref}[k,n] - \tilde{E}_{test}[k,n] \quad (75)$$

– 如果$\tilde{E}_{ref}[k,n] > \tilde{E}_{test}[k,n]$，那么斜率$b$的陡度被设置为4.0，否则被设置为6.0。如此便模拟了以下效果，即：与参考信号相比，待测信号的信号能量的增加比减少更显著。

– 计算比例因子a：

$$a[k,n] = \frac{10^{\frac{\log_{10}(\log_{10}(2.0))}{b}}}{s[k,n]} \quad (76)$$

– 计算被检测到的概率。等式（76）设置比例因子a，使得如果$e[k,n]$等于$s[k,n]$，则$p_c[k,n]$变为*0.5*：

$$p_c[k,n] = 1 - 10^{(-a[k,n] \cdot e[k,n])^b)} \quad (77)$$

– 计算阈值以上的总步数：

$$q_c[k,n] = \frac{|\mathrm{INT}(e[k,n])|}{s[k,n]} \quad (78)$$

– 双耳检测概率为：

$$p_{bin}[k,n] = \max(p_{left}[k,n], p_{right}[k,n]) \quad (79)$$

– 双耳声道高于阈值的步数为：

$$q_{bin}[k,n] = \max(q_{left}[k,n], q_{right}[k,n]) \quad (80)$$

检测到帧n的通道c的总概率为：

$$P_c[n] = 1 - \prod_{\forall k} (1 - p_c[k,n]) \quad (81)$$

其中c可以是左、右或二进制。帧n的通道c高于阈值的总步数为：

$$Q_c[n] = \sum_{\forall k} q_c[k,n] \quad (82)$$

4.7.1 最大滤波检测概率（$MFPD_B$）

计算每个通道c的检测概率的平滑版本：

$$\tilde{P}_c[n] = (1 - c_0) \cdot P_c[n] + c_0 \cdot \tilde{P}_c[n-1] \quad (83)$$

其中$P_c[-1] = 0$。常数c_0取决于步长（*StepSize*）：

$$c_0 = 0.9^{StepSize/1\,024} \quad (84)$$

c_0降低了对极短失真的敏感度。

最大滤波检测概率（缩写：MFPD）计算如下：

$$PM_c[n] = \max\left(PM_c[n-1] \cdot c_1, \tilde{P}_c[n]\right) \quad (85)$$

其中$PM_c[-1]$为0。常数c_1取决于步长：

$$c_1 = 0.99^{StepSize/1\,024} \quad (86)$$

c_1对由于遗忘而导致的音频节录开始处的失真不如节录结束处的失真严重的效应进行建模。请注意，此常数对于模拟听力测试很有用，原因是在听力测试中，受试者不允许选择节录的较短部分。对于根据ITU-R BS.1116建议书使用来自听力测试的数据加以校准的本模型，c_1应为1.0。

模型输出变量*MFPD*是最后一帧的$PM_{bin}[n]$的值。

4.7.2 平均失真块[5]（ADB_B）

对双耳声道检测概率$P_{bin}[n]$高于0.5的有效帧的数量进行计数（$n_{distorted}$）。

对于所有有效帧，计算双耳声道$Q_{bin}[n]$阈值以上的总步长数：

$$Q_{sum} = \sum_{\forall n} Q_{bin}[n]$$

平均失真块ADB的失真计算如下：

– 如果$n_{distorted}$为0 则 $ADB = 0$（听不到失真）；

– 如果$n_{distorted} > 0$ 且 $Q_{sum} > 0$ 则 $ADB = \log_{10}((Q_{sum}) / n_{distorted}))$；

– 如果$n_{distorted} > 0$ 且 Q_{sum}为0 则 $ADB = -0.5$。

4.8 误差的谐波结构

包含强谐波的参考信号（例如低音单簧管、竖琴）的频谱的特性是被深谷分隔的多个规则间隔的波峰。在某些情况下，误差信号可能会继承该结构。例如，与此类信号混合的噪声更有可能保持不被掩蔽，其中信号在频谱波谷中较低。由此产生的误差频谱将包含类似于原始频谱的结构，但在频率上发生偏移，以对应于波谷的位置。这种结构可能会导致音质失真，从而增加误差的显著性。

误差被定义为参考信号和所处理信号的对数频谱的差异，每个信号均由外耳和中耳的频率响应加权（见第2.1.4节，等式（7））。这里不使用来自心理声学模型的激发模式，原因是非线性频率到巴克变换会模糊谐波结构。

---

5 术语“块”在此上下文中等同于“帧”。

4.8.1 EHS_B

通过识别和测量自相关函数频谱中的最大峰值来获得谐波结构幅度。每个相关性根据以下公式计算为两个矢量之间角度的余弦值，其中$\vec{F}_0$是误差矢量，$\vec{F}_t$是滞后一定量的同一个矢量。相关性的长度与最大滞后时间相同（即下例中的256）。

$$C = \frac{\vec{F}_0 \cdot \vec{F}_t}{\left|\vec{F}_0\right| \cdot \left|\vec{F}_t\right|} \tag{87}$$

获得自相关函数的最大滞后是小于对应于18 kHz的FFT频率分量数一半的2的最大幂。

例如，当采样速率为48 kHz且FFT窗口大小为2 048个样本时，对应于18 kHz的FFT分量为$(18/24) \times 1\,024 = 768$。因此，最大延迟为384。滞后的实际数量为256，即小于384的2的最大幂。相关函数的第一个值将通过把$F_t[0]$与$F_0[0]$对齐来获得，最后一个值将通过把$F_t[0]$与$F_0[255]$对齐来获得。

得到的相关向量用归一化的汉宁窗加窗，在通过减去平均值去除DC分量后，用FFT计算功率谱。第一个波谷之后的频谱中的最大峰值标识自相关函数中的主频率。帧上该最大值的平均值乘以1 000.0即为误差谐波结构（EHS）变量。

5 平均

5.1 频谱平均

如果MOV描述中没有不同的说明（见第4节），则在对频带上的本地值求平均值时使用以下算法。

5.1.1 线性平均值

线性平均值的计算方法如下：

$$AvgS = \frac{1}{Z} \cdot \sum_{k=0}^{Z-1} S[k] \tag{88}$$

其中S代表模型输出变量的名称，Z代表频带数。

5.2 时间平均

如果在MOV（第4节）的描述中没有不同的说明，则在对一段时间内的瞬时值进行平均时，将使用以下一种或几种算法。时间加权因子（如果应用的话）由符号W表示，Z是频带的数量。

5.2.1 线性平均值

线性平均值（前缀“*Avg*”）的计算公式如下：

$$AvgX = \frac{1}{N} \cdot \sum_{n=0}^{N-1} X[n] \tag{89}$$

其中X代表MOV的名称，N是计算X瞬时值的时间样本数。

如果应用了时间加权（见第4.2节），则根据以下公式计算线性平均值：

$$AvgX = \frac{\sum_{n=0}^{N-1} W[n] \cdot X[n]}{\sum_{n=0}^{N-1} W[n]} \quad (90)$$

5.2.2 平方平均值

平方平均值（前缀“*Rms*”）的计算公式如下：

$$RmsX = \sqrt{\frac{1}{N} \cdot \sum_{n=0}^{N-1} X[n]^2} \quad (91)$$

其中*X*代表MOV的名称，*N*是计算*X*瞬时值的时间样本数。

当应用了时间加权时（见第4.2节），根据以下公式计算平方平均值：

$$RmsX = \sqrt{Z} \cdot \sqrt{\frac{\sum_{n=0}^{N-1} W[n]^2 \cdot X[n]^2}{\sum_{n=0}^{N-1} W[n]^2}} \quad (92)$$

5.2.3 窗口平均值

窗口平均值（前缀“*Win*”）的计算公式如下：

$$WinX = \sqrt{\frac{1}{N-L+1} \cdot \sum_{n=L-1}^{N-1} \left(\frac{1}{L} \cdot \sum_{i=0}^{L-1} \sqrt{X[n-i]} \right)^4} \quad (93)$$

其中*X*代表MOV的名称，*N*是已经计算出*X*的瞬时值的时间样本的数量，*L*是时间样本中滑动时间窗口的长度。窗口长度约为100毫秒，即：基于FFT的人耳模型的*L*为4，基于滤波器组的人耳模型的*L*为25。

5.2.4 帧的选择

5.2.4.1 延迟平均

对于使用该标准的MOV，在时间平均中不考虑测量前0.5秒期间计算的值。延迟平均用于以下模型输出变量：

WinModDiff1、AvgModDiff1、AvgModDiff2、RmsNoiseLoudness、RmsNoiseLoudAsym、RmsModDiff、AvgLinDist。

5.2.4.2 响度阈值

对于使用该标准的MOV，在时间平均中不考虑所有瞬时值，这些瞬时值是在一个相应音频通道的总响度达到测试和参考信号的N_{Thres}宋值后的50毫秒计算的。响度阈值仅用于第4.3节中描述的MOV。

5.2.4.3 能量阈值

当2 048个样本的帧的最近一半的能量在单声道或参考和测试数据的左右声道中小于8 000*时，该帧被忽略。帧有50%的重叠，并且只评估包含新数据的那一半帧。该标准的应用防止了以非常小的能量处理帧。

该标准仅用于第4.8节中描述的MOV。

5.2.4.4 数据边界

如果经处理的文件在合法参考文件数据之前或之后包含噪声，则相对误差可能非常大，原因是参考电平接近–∞。当该误差被认为是人为误差时，可以通过应用数据边界剔除标准将其忽略。

当文件第一次打开时，参考文件中实际数据的开始和结束位置被识别。数据的开头或结尾被定义为第一个位置，从文件的开头或结尾开始扫描，在其中相应的音频通道之一内，五个连续样本的绝对值之和超过200。完全超出此范围的帧随后会被忽略。

该标准用于MOV。

5.3 音频通道平均

当没有另做说明时，在立体声信号的情况下，左和右声道的MOV在时间平均之后被线性平均。

6 感知基本音频质量的估算

*感知基本音频质量*是通过使用具有一个隐藏层的人工神经网络结构将几个MOV映射到单个数字来估算的。

6.1 人工神经网络

神经网络的激活函数是不对称的S形函数：

$$\mathrm{sig}(x) = \frac{1}{1+\mathrm{e}^{-x}} \qquad (94)$$

网络在隐藏层中使用I个输入和J个节点。该映射由一组输入比例因子$a_{min}[i]$、$a_{max}[i]$、一组输入权重$w_x[i]$、一组输出权重$w_y[j]$和一对输出比例因子b_{min}和b_{max}定义。输入被映射到一个*失真指数*

$$DI = w_y[J] + \sum_{j=0}^{J-1}\left(w_y[j] \cdot \mathrm{sig}\left(w_x[I,j] + \sum_{i=0}^{I-1} w_x[i,j] \cdot \frac{x[i] - a_{min}[i]}{a_{max}[i] - a_{min}[i]} \right) \right) \qquad (95)$$

---

* 该数字是指光盘上使用的16位有符号整数格式的输入数据，范围为－32 768至32 767。

其根据客观差异等级（ODG）与所估算的感知基本音频质量直接相关。失真指数和客观差异等级的关系由下式给出：

$$ODG = b_{min} + (b_{max} - b_{min}) \cdot \mathrm{sig}(DI) \quad (96)$$

6.2 基本版本

基本版本仅使用基于FFT的人耳模型。它使用以下MOV：*BandwidthRef*$_B$、*BandwidthTest*$_B$、*Total NMR*$_B$、*WinModDiff1*$_B$、*ADB*$_B$、*EHS*$_B$、*AvgModDiff1*$_B$、*AvgModDiff2*$_B$、*RmsNoiseLoud*$_B$、*MFPD*$_B$和*RelDistFrames*$_B$。这11个MOV使用第6.1节（人工神经网络）中描述的神经网络映射到单个质量指数，该神经网络在隐藏层中具有三个节点。表12至16给出了映射的参数。

表12

基本版本中使用的MOV

MOV	目的
$WinModDiff1_B$	调制变化（与粗糙度相关）
$AvgModDiff1_B$	
$AvgModDiff2_B$	
$RmsNoiseLoud_B$	失真响度
$BandwidthRef_B$	线性失真（频率响应等）
$BandwidthTest_B$	
$RelDistFrames_B$	听觉失真的频率
Total NMR_B	噪声掩蔽比
$MFPD_B$	检测概率
ADB_B	
EHS_B	误差的谐波结构

表13

基本版本输入的比例因子

索引 (*i*)	MOV (x[i])	$\mathbf{a}_{min}$[i]	$\mathbf{a}_{max}$[i]
0	$BandwidthRef_B$	393.916656	921
1	$BandwidthTest_B$	361.965332	881.131226
2	Total NMR_B	–24.045116	16.212030
3	$WinModDiff1_B$	1.110661	107.137772
4	ADB_B	–0.206623	2.886017
5	EHS_B	0.074318	13.933351

表13（完）

索引 (*i*)	MOV (x[i])	$a_{min}[i]$	$a_{max}[i]$
6	$AvgModDiff1_B$	1.113683	63.257874
7	$AvgModDiff2_B$	0.950345	1145.018555
8	$RmsNoiseLoud_B$	0.029985	14.819740
9	$MFPD_B$	0.000101	1
10	$RelDistFrames_B$	0	1

表14

基本版本输入节点的权重

索引 (*i*)	MOV (x[i])	节点1 ($w_x[i,0]$)	节点2 ($w_x[i,1]$)	节点3 ($w_x[i,2]$)
0	$BandwidthRef_B$	–0.502657	0.436333	1.219602
1	$BandwidthTest_B$	4.307481	3.246017	1.123743
2	Total NMR_B	4.984241	–2.211189	–0.192096
3	$WinModDiff1_B$	0.051056	–1.762424	4.331315
4	ADB_B	2.321580	1.789971	–0.754560
5	EHS_B	–5.303901	–3.452257	–10.814982
6	$AvgModDiff1_B$	2.730991	–6.111805	1.519223
7	$AvgModDiff2_B$	0.624950	–1.331523	–5.955151
8	$RmsNoiseLoud_B$	3.102889	0.871260	–5.922878
9	$MFPD_B$	–1.051468	–0.939882	–0.142913
10	$RelDistFrames_B$	–1.804679	–0.503610	–0.620456
11	bias	–2.518254	0.654841	–2.207228

表15

基本版本输出节点的权重

节点1 ($w_y[0]$)	节点2 ($w_y[1]$)	节点3 ($w_y[2]$)	偏差 ($w_y[3]$)
–3.817048	4.107138	4.629582	– 0.307594

表16

基本版本输出的比例因子

	b_{min}	b_{max}
ODG	–3.98	0.22

6.3 高级版本

高级版本同时使用基于滤波器组的人耳模型和基于FFT的人耳模型。它使用的MOV包括$RmsModDiff_A$、$RmsNoiseLoudAsym_A$、$AvgLinDist_A$、Segmental NMR_B和EHS_B。使用第6.1节中描述的神经网络将这5个MOV映射到单个质量指数，在隐藏层中有5个节点。表17至21给出了映射的参数。

表17

高级版本中使用的MOV

MOV	目的
$RmsNoiseLoudAsym_A$	失真响度
$RmsModDiff_A$	调制变化（与粗糙度相关）
$AvgLinDist_A$	线性失真（频率响应等）
Segmental NMR_B	噪声掩蔽比
EHS_B	误差的谐波结构

表18

高级版本输入节点的比例因子

索引 (i)	MOV (x[i])	$a_{min}[i]$	$a_{max}[i]$
0	$RmsModDiff_A$	13.298751	2166.5
1	$RmsNoiseLoudAsym_A$	0.041073	13.24326
2	Segmental NMR_B	–25.018791	13.46708
3	EHS_B	0.061560	10.226771
4	$AvgLinDist_A$	0.024523	14.224874

表19

高级版本输入的权重

索引 (*i*)	MOV (x[i])	节点1 ($w_x[i,0]$)	节点2 ($w_x[i,1]$)	节点3 ($w_x[i,2]$)	节点4 ($w_x[i,3]$)	节点5 ($w_4[i,4]$)
0	$RmsModDiff_A$	21.211773	–39.913052	–1.382553	–14.545348	–0.320899
1	$RmsNoiseLoudAsym_A$	–8.981803	19.956049	0.935389	–1.686586	–3.238586
2	Segmental NMR_B	1.633830	–2.877505	–7.442935	5.606502	–1.783120
3	EHS_B	6.103821	19.587435	–0.240284	1.088213	–0.511314
4	$AvgLinDist_A$	11.556344	3.892028	9.720441	–3.287205	–11.031250
5	bias	1.330890	2.686103	2.096598	–1.327851	3.087055

表20

高级版本输入的权重

节点 1 ($w_x[i,0]$)	节点 2 ($w_x[i,1]$)	节点 3 ($w_x[i,2]$)	节点 4 ($w_x[i,3]$)	节点 5 ($w_4[i,4]$)	偏差 ($w_y[5]$)
–4.696996	–3.289959	7.004782	6.651897	4.009144	–1.360308

表21

高级版本输出的比例因子

	b_{min}	b_{max}
ODG	–3.98	0.22

7 实施的一致性

7.1 综述

本节提供了一组测试项目来验证方法的实施正确与否。

7.2 选择

测试项目选自数据库3（DB3），该数据库用于验证模型。为了简化测试，选择了DB3的84个项目的子集。该子集由16个项目组成。选择的主要标准是所得的MOV和DI（失真指数）值涵盖较宽的范围。

7.3 一致性测试的设置

测试项目可以下载为WAV文件（微软RIFF格式）。所有项目均以48 kHz、16位PCM进行采样。由国际电联提供的参考和测试信号已经在时间和电平上相互适应，因此不需要额外的增益或延迟补偿。测量算法必须调整到92 dB SPL的收听电平。

7.4 可接受的容限区间

为了符合建议书，所计算的DI值必须再现表22和表23中给出的值，所有测试项目的容限小于±0.02[6]。如果一项实施不能在此容限范围内产生结果，则该实施不符合此建议书。

7.5 测试项目

下表显示了参考和测试项目的名称[7]以及最终DI值。表22与基本版本相关，表23包含高级版本的值。

---

[6] 为了达到这一精度，应使用IEEE浮点算法。

[7] 相应参考项目的名称是通过将测试项目名称中的子字符串“cod”替换为“ref”得出的，例如“bcodtri.wav”的参考项目是“breftri.wav”。

表22

基本版本的测试项目和最终DI值

项目	DI	ODG
acodsna.wav	1.304	–0.676
bcodtri.wav	1.949	–0.304
ccodsax.wav	0.048	–1.829
ecodsmg.wav	1.731	–0.412
fcodsb1.wav	0.677	–1.195
fcodtr1.wav	1.419	–0.598
fcodtr2.wav	–0.045	–1.927
fcodtr3.wav	–0.715	–2.601
gcodcla.wav	1.781	–0.386
icodsna.wav	–3.029	–3.786
kcodsme.wav	3.093	0.038
lcodhrp.wav	1.041	–0.876
lcodpip.wav	1.973	–0.293
mcodcla.wav	–0.436	–2.331
ncodsfe.wav	3.135	0.045
scodclv.wav	1.689	–0.435

表23

高级版本的测试项目和最终DI值

项目	DI	ODG
acodsna.wav	1.632	–0.467
bcodtri.wav	2.000	–0.281
ccodsax.wav	0.567	–1.300
ecodsmg.wav	1.594	–0.489
fcodsb1.wav	1.039	–0.877
fcodtr1.wav	1.555	–0.512
fcodtr2.wav	0.162	–1.711
fcodtr3.wav	–0.783	–2.662
gcodcla.wav	1.457	–0.573
icodsna.wav	–2.510	–3.664
kcodsme.wav	2.765	–0.029
lcodhrp.wav	1.538	–0.523
lcodpip.wav	2.149	–0.219
mcodcla.wav	0.430	–1.435
ncodsfe.wav	3.163	0.050
scodclv.wav	1.972	–0.293

附件2的
后附资料1

验证过程

1 综述

1994年，ITU-R通过了ITU-R第210/10号课题 — 客观感知质量评估方法，并成立了一个任务组。首先采取的行动之一是公开征集建议，并收到了六个模型支持者的回应。

为此，各方花了很多精力来定义验证过程的程序，并发现汇编第一个数据库（称为DB1数据库）是有助益的，该数据库由已进行的听力测试的材料组成。在此过程中，主要关注了中高音频质量，因此只考虑符合ITU-R BS.1116建议书的听力测试结果。这些测试的材料代表了低比特率编解码器的关键广播材料，如MPEG1第二层、MPEG1第三层、杜比AC2、Mini Disc、NICAM等。创建数据库1旨在向模型支持者提供一个通用平台，该平台包含涵盖大量损伤、各种编解码器和级联编解码器劣化的材料。数据库1中汇编的测试的详细说明见附件2的后附资料2。

显然，模仿人类行为的感知音频质量的客观测量方法只能在包含主观测试结果的数据库上进行验证。适当的验证需要基于未知材料的数据库。因此，有必要进行新的听力测试。由于测量方法在理想情况下应该针对可能出现在广播应用中的任何类型的伪像，因此不仅应包括编码伪像，还应包括失真和噪声等更传统的伪像。数据库2和数据库3分别于1996年和1997年建立，以满足上述要求。除了DB1已有的编解码器外，还包括杜比AC-3和AAC。更多详情见附件2的后附资料2。

验证应考虑主观听力测试中固有的不确定性，这通常表现为置信区间。置信区间的大小取决于诸多因素。最重要的是受试者的经历、训练程序和测试项目出现的背景以及受试者人数。

本建议书中给出的客观方法的调整和验证基于“普通专家听者”。主观质量评估的平均值和95%的置信区间用于描述“普通专家听者”的特性。

主观听力测试对影响结果的各种因素均非常敏感。数据库2和数据库3的SDG是在三个不同的测试点产生的，许多研究对数据是否真的可以合并进行了调查。尽管并非所有研究均得出了相同的结论，但合并数据被证明是合理的，合并后的数据库构成了验证的基础。

验证过程分为三个阶段：

– 第一阶段： 竞争阶段
– 第二阶段： 协作阶段
– 第三阶段： 最终选择

这些阶段将在下面的章节中详细描述。

2 竞争阶段

为了对感知音频质量进行客观测量，提出了六种方法（DIX、NMR、PAQM、PERCEVAL、POM、TTA），并决定使用数据库2和数据库1的子集来比较这些方法的性能。数据库2是1996年初建立的。SR（瑞典广播公司）和BBC（英国广播公司）合作选出了最终测试材料。听力测试在挪威的NRK（挪威公共电视台）、丹麦的DR（丹麦广播公司）

和日本的NHK（日本放送协会）进行。德国电信（德国）和Teracom（瑞典）对测试数据进行了统计分析。在第一阶段，客观数据在中立地点（瑞士，Swisscom（瑞士电信））生成。然后，模型支持者收到数据库2的前半部分，并对方法进行了最后调整（第二阶段）。最后，瑞士电信产生了新的客观差异等级。

Teracom（瑞典）和支持者本身对这些方法的性能进行了分析。尽管一些拟议方法的结果显示与SDG高度相关，但各方一致认为没有一种方法能够满足用户的要求。一项单独的研究表明，没有一种方法明显优于其他方法。因此，各方决定开发一种改进的测量方法，并由当前的所有支持者携手开展工作。新方法的性能应与被称为模型B3的一种已确立的方法进行比较。

3 协作阶段

协作阶段的思路是将不同方法的最佳元素整合成一种新方法。为了更好地满足用户的需求，决定开发两种版本的方法。一种适合实时实施，另一种可能需要更高的计算能力来实现更高的精度。

新方法的验证程序与竞争阶段的验证程序在设计上存在相似之处。为此，创建了一个新数据库（DB3）。项目和条件最终于1997年春敲定，并由SR、Swisscom和BBC进行了汇编。附件2的后附资料2对数据库进行了全面描述。主观听力测试在德国电信、NHK和SR三个测试点进行。所有测试点均采用了ITU-R BS.1116建议书中描述的“三重刺激隐藏参考双盲法”。在瑞典收集了听力测试结果。Teracom和其他机构对听力测试结果进行了广泛的统计分析。在上述分析完成后，进一步评估剔除了一些听者。对测试点提供的结果进行了整合，以形成数据库3。

1997年秋，向支持者发放了数据库中的52个项目。新方法适应了新数据。由于有几个参数设置提供了类似的结果，因此尽可能推迟了最终选择的决定时间。最后，在瑞士，剩余的32个项目用于在“未知”数据集上验证新方法。

此外，CRC（加拿大）进行的新听力测试的结果亦被用于验证有关“未知”材料的新方法。以下各节描述了选择和验证过程。

4 验证

对客观测量方法的18个指定版本进行了广泛测试。本节描述了选择标准，并介绍了SDG与18个测量方法版本的结果进行比较之后的结果。其目的是选择和验证将向国际电联建议的最佳版本。

选择标准

主观和客观结果之间的相关性是验证客观方法的最明显标准。此外，验证过程中还引入了另外两个考虑平均值可靠性的标准 — AES和容限方案。

AES是为了将模型的精度与听力测试的精度联系起来而引入的，并根据以下表达式进行定义：

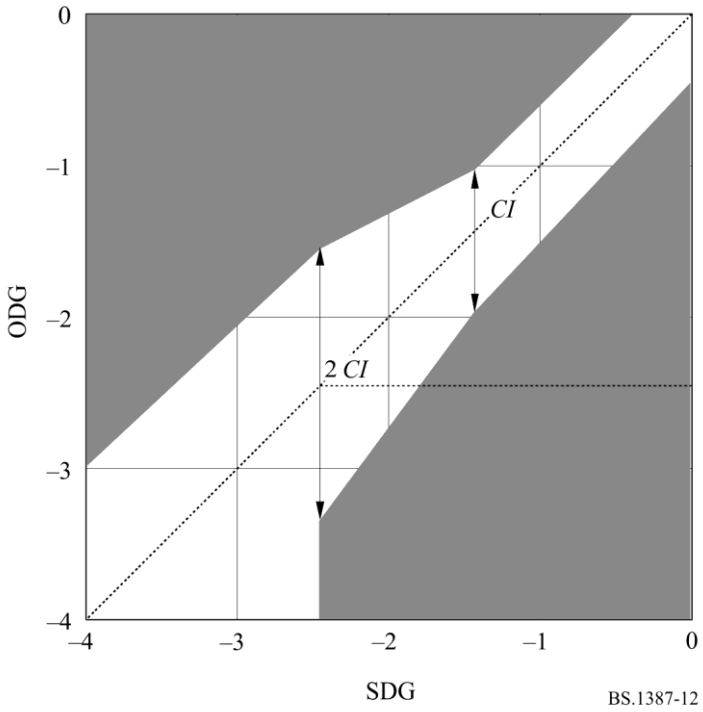
$$AES = 2 * \sqrt{\frac{\sum((ODG - SDG)/CI)^2}{N}} \qquad \text{如果} CI < 0.25\text{，那么} CI = 0.25$$

其中CI是置信区间。

容限方案旨在允许ODG在损伤规模的上限和下限与SDG存在不同的偏差。容限范围与听力测试的置信区间有关。该范围限制在0.25级的最小值。容限方案之外的ODG到容限方案的距离用于评估测量方法的质量。

图12

容限方案，置信区间CI ≥ 0.25



4.1 SDG和ODG值的比较

客观测量分为三个不同的阶段。在第一阶段，除了选择小组之外，所有84个测试项目均为未知的。在第二阶段，公布了52个项目的信息。该信息包含SDG值和实际音频节录。在第三阶段，该知识被用于优化此方法版本的性能。请注意，与第一阶段相比，第三阶段测试了另外四个版本。所呈现的SDG值是根据75名合格受试者生成的数据计算得出的。

为了评估ODG反映SDG的程度，存在诸多不同的方法。遗憾的是，没有一个单一的值真正说明了完整的性能。相反，必须从多个角度来研究此问题。相关性的介绍见第4.2节，AES的介绍见第4.3节。B3模型是ITU-R于1996年测试的模型之一，且已决定将各种新版本与此旧版本进行比较。

4.2 相关性

图13（84项）和图14（32项）绘制了第一阶段和第三阶段的相关图。

图13

SDG和ODG之间的相关性

所有84个项目均包括在内

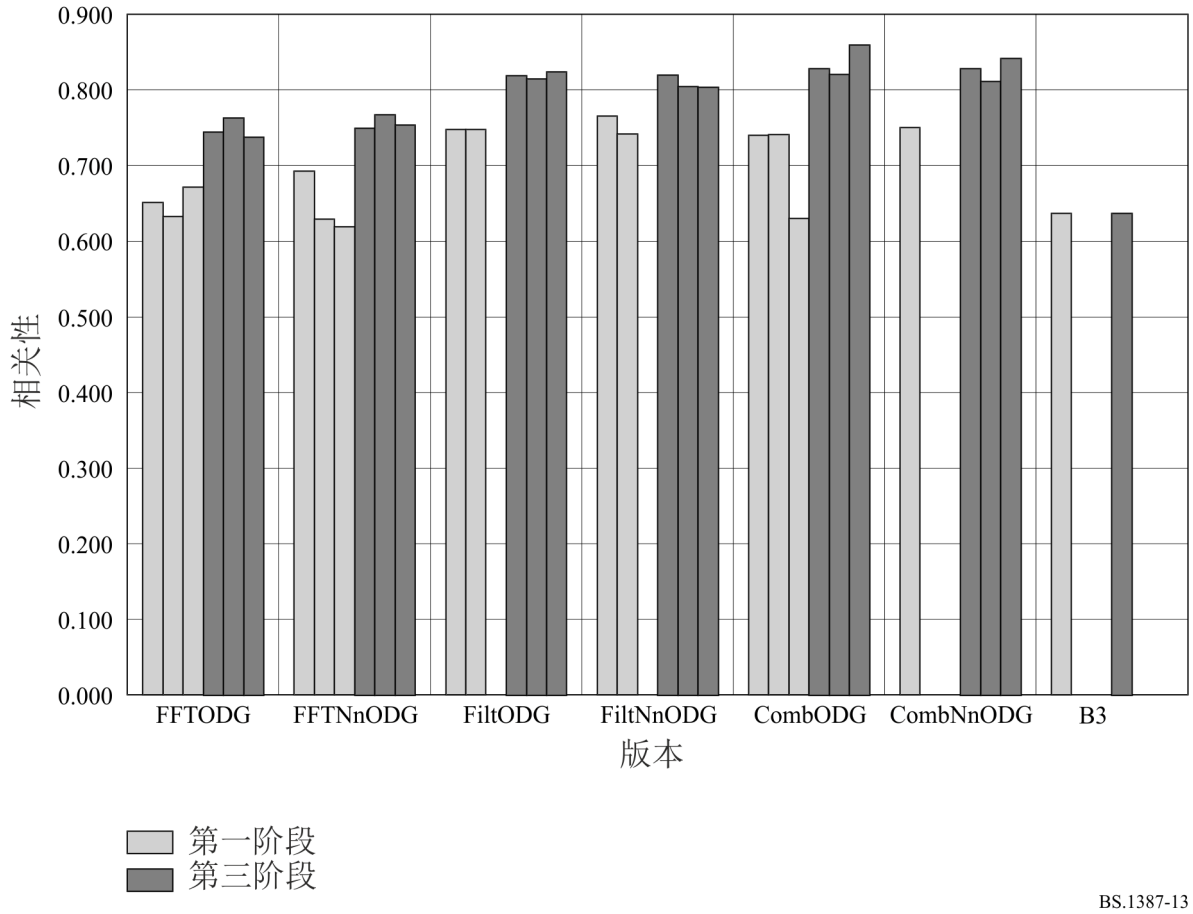
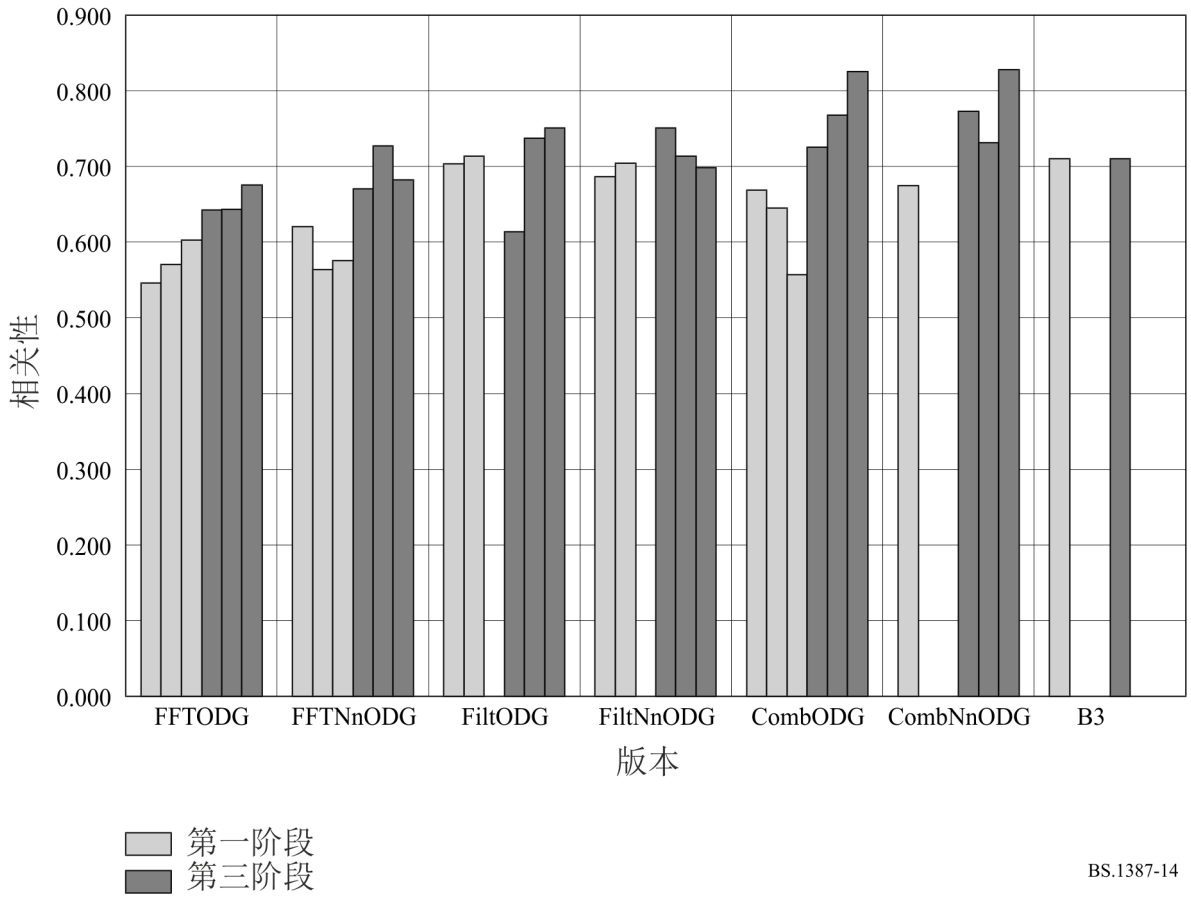


图14

SDG和ODG之间的相关性

32个未发布的项目已包括在内



4.3 绝对误差分数（AES）

平均而言，在SDG置信区间内产生ODG值的模型将获得接近2的AES值。图15至图16概述了AES值。

图15

不同版本的AES

所有84个项目均包括在内

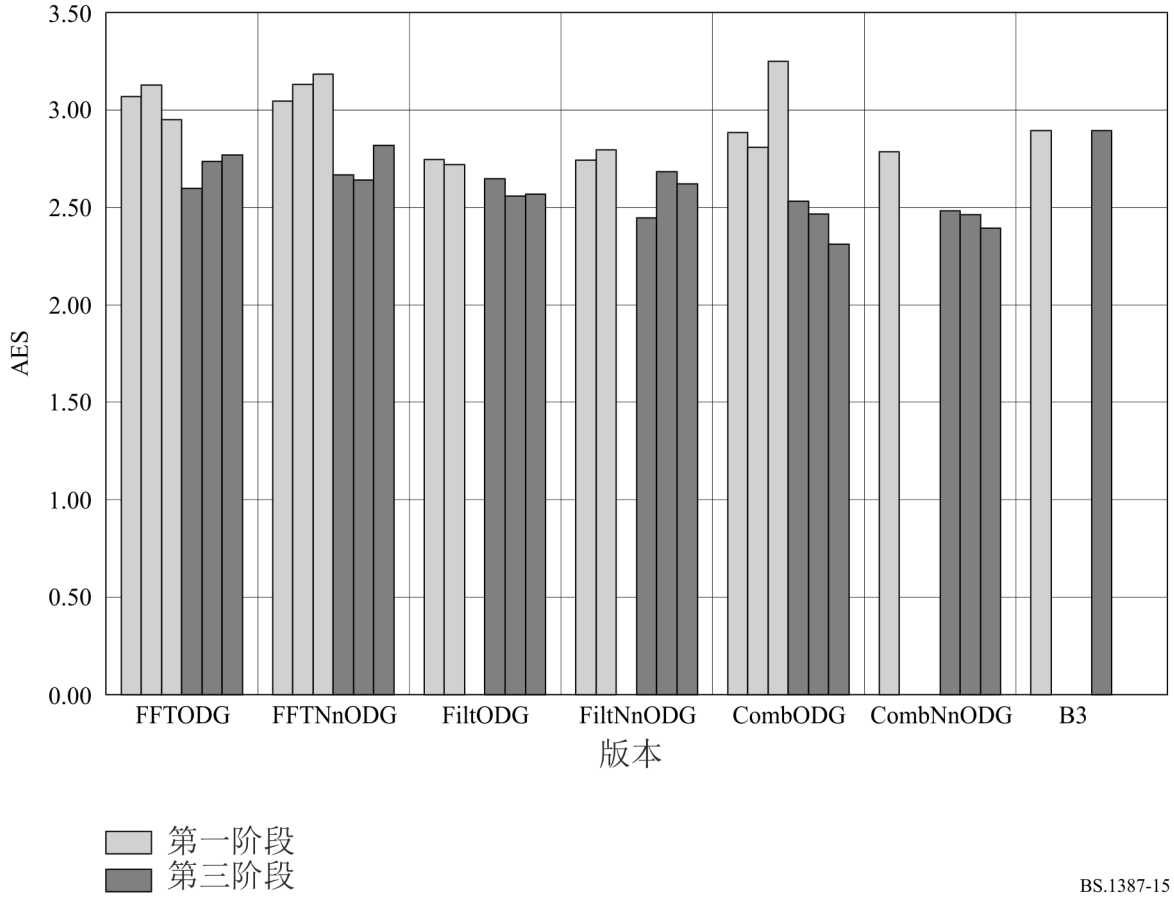
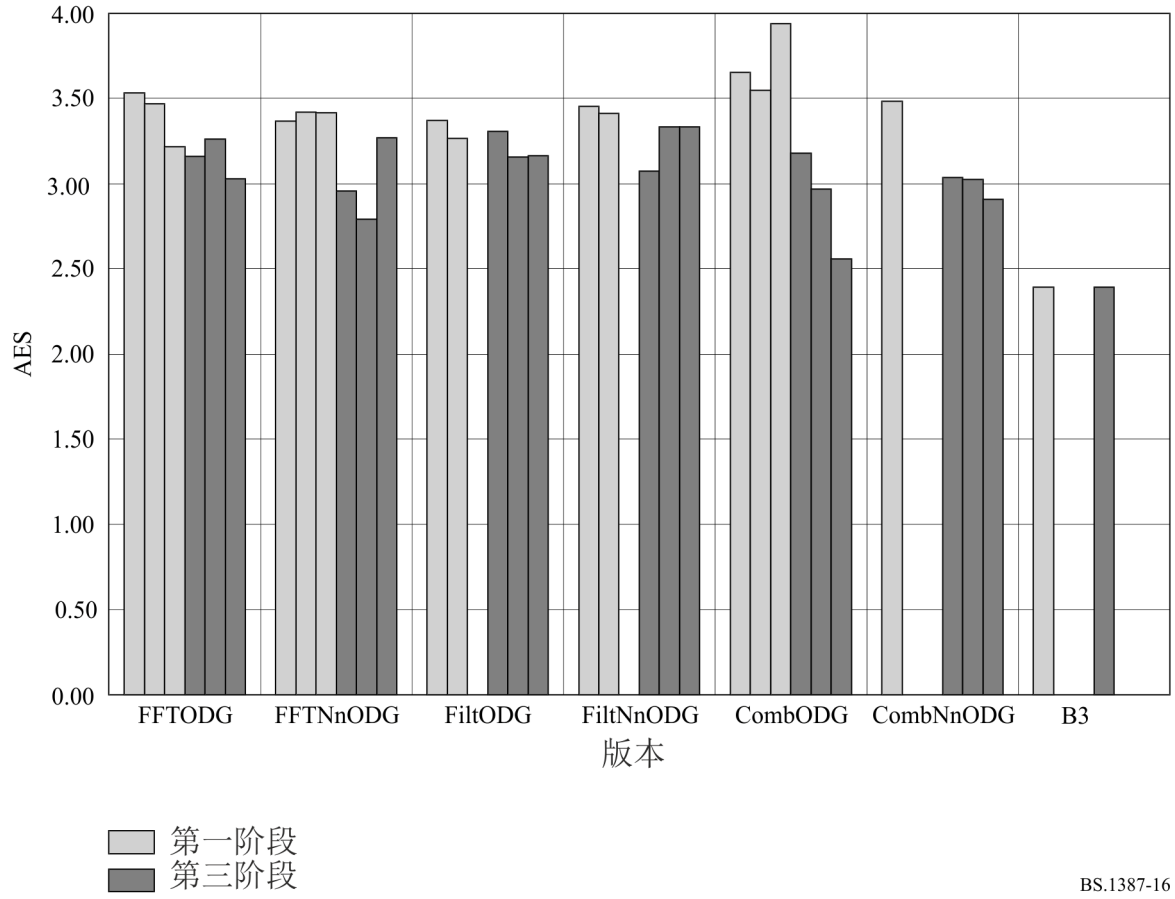


图16

不同版本的AES

32个未发布的项目已包括在内



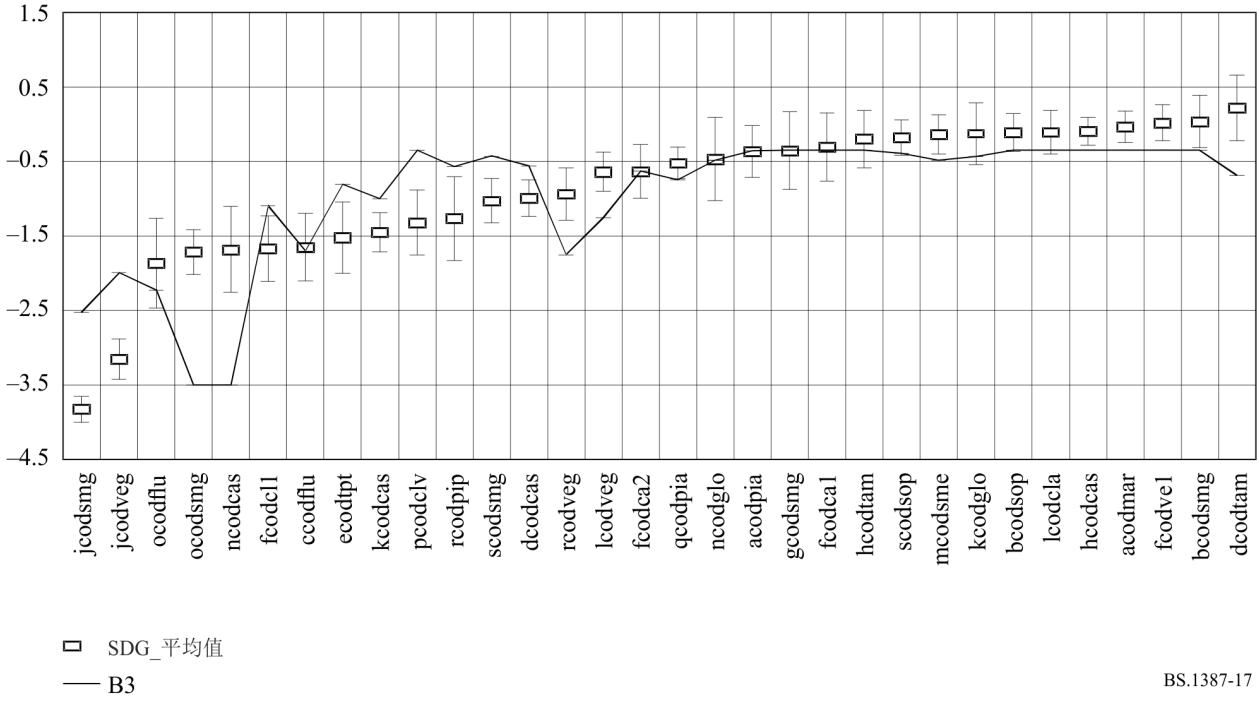
4.4 ODG与置信区间的比较

果不其然，大多数版本的表现存在相似之处。会上展示了诸多图表，但本建议书容量有限无法一一赘述。有关更多详情，请参考完整的验证测试报告。

在图17至图22中，绘制了模型B3以及模型版本FFTNnODG1和CombNnODG3的32个未发布项目的平均SDG、置信区间和ODG。

图17

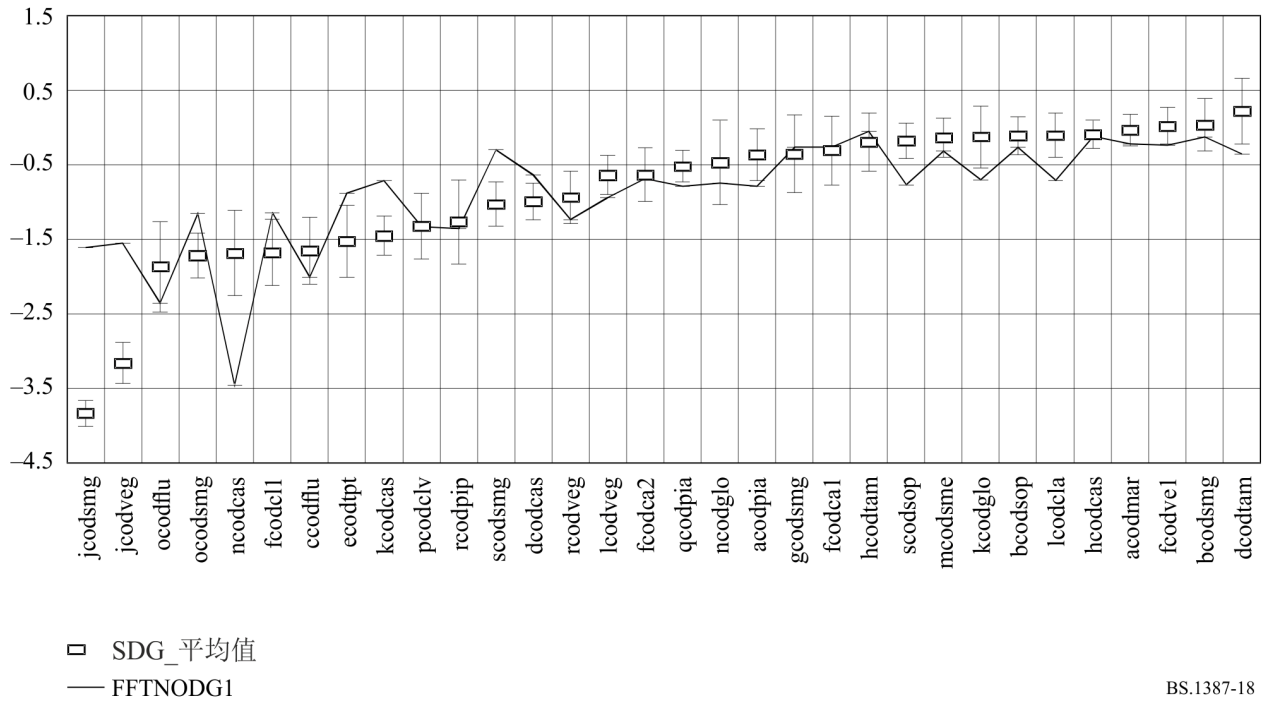
模型B3：32个未发布项目的平均SDG、置信区间和ODG图



BS.1387-17

图18

FFTNnODG1：第三阶段后32个未发布项目的平均SDG、置信区间和ODG图



BS.1387-18

图19

CombNnODG3：第三阶段后32个未发布项目的平均SDG、置信区间和ODG图

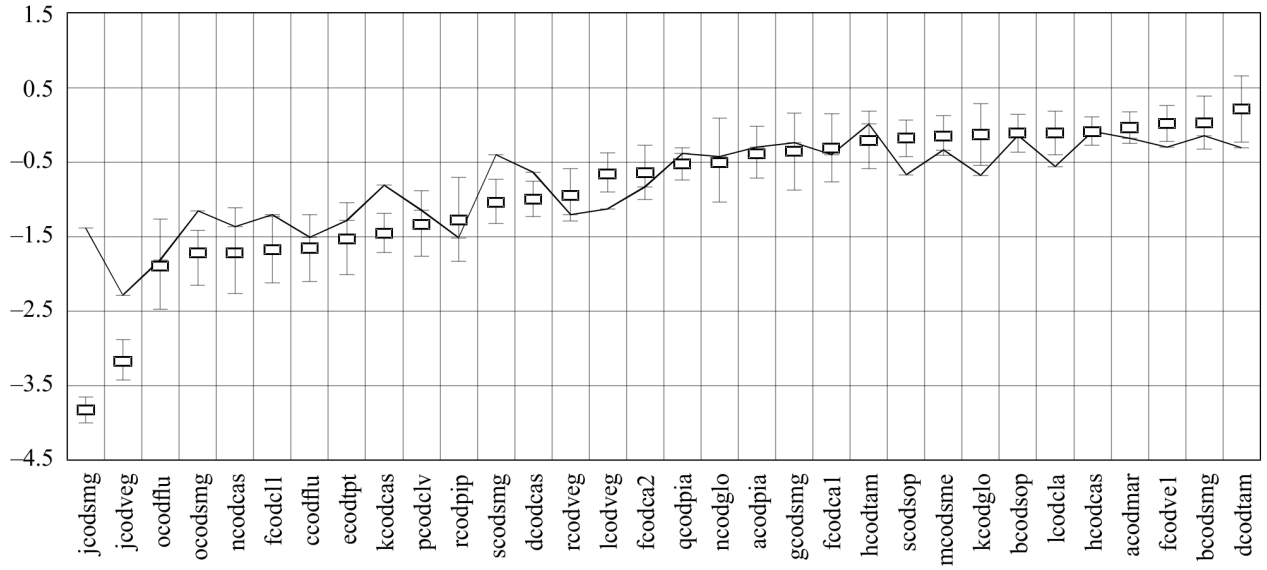


图20和图21给出了类似的曲线，但此情况针对的是第三阶段期间的所有84个项目。此外，图22展示了版本CombNnODG3的性能。

图20

模型B3：所有84个项目的平均SDG、置信区间和ODG图

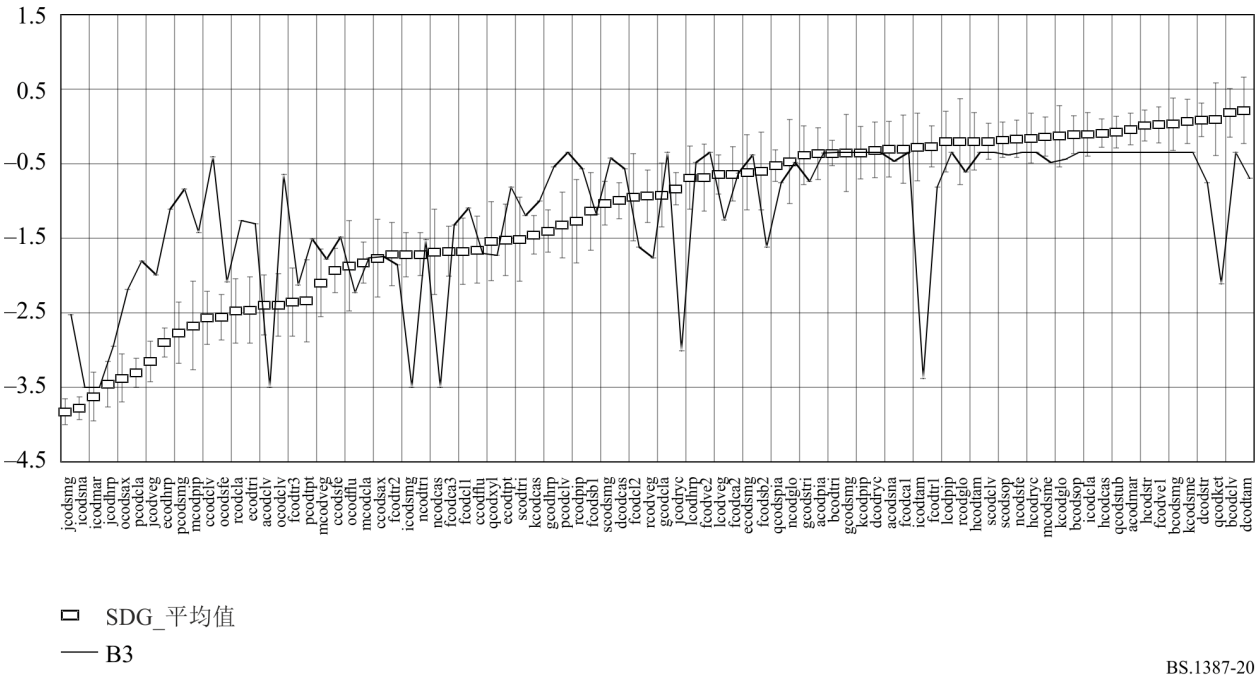
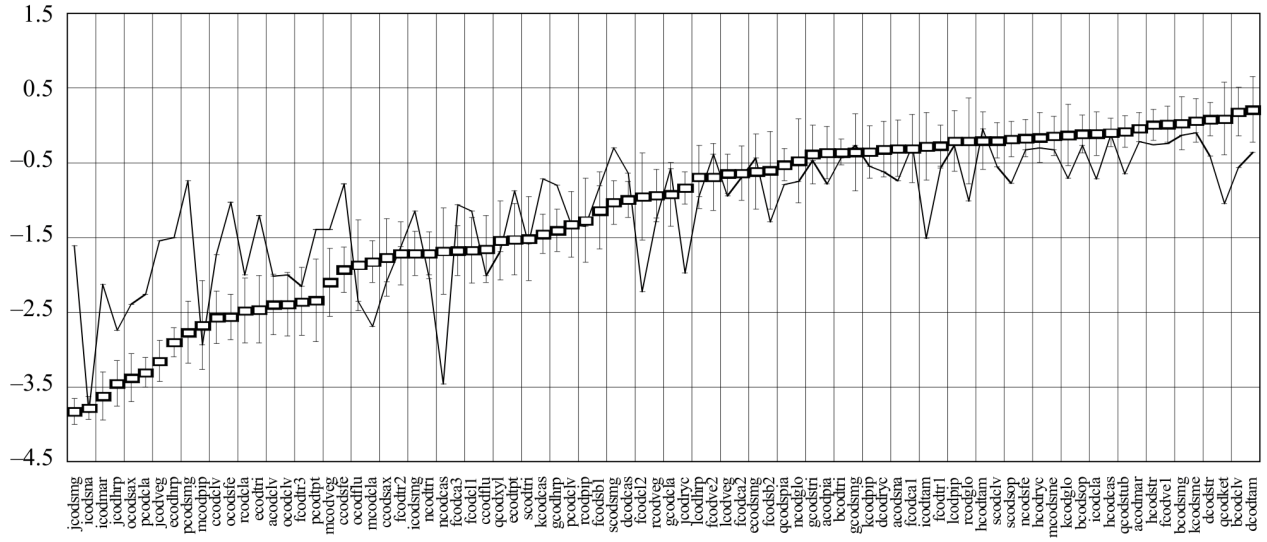


图21

FFTNnODG1：所有84个项目在第三阶段的平均SDG、置信区间和ODG图



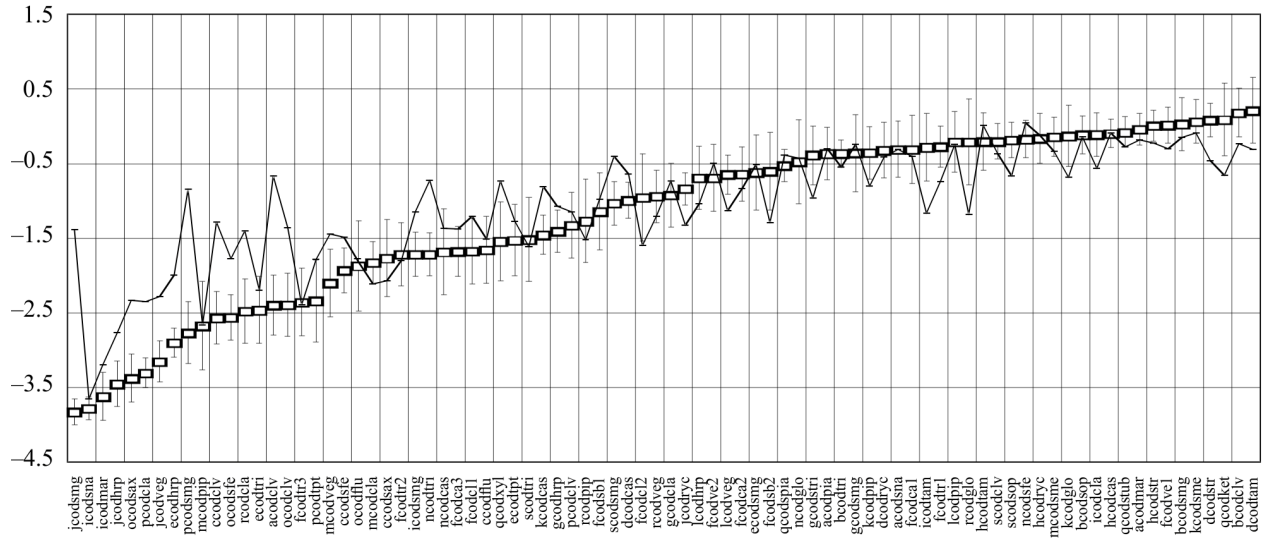
SDG_平均值

FFTNODG1

BS.1387-21

图22

CombNnODG3：所有84个项目在第三阶段的平均SDG、置信区间和ODG图



SDG_平均值

COMNODG3

BS.1387-22

4.5 ODG与容限区间的比较

ITU-R已经定义了一个目标用户要求，该要求可以映射到一个容限区间。对于较高级别的音频质量，目标要求更为严格，而对于较低的音频质量，目标要求则更为宽松。下图显示了模型B3以及模型版本FFTNnODG1和CombNnODG3在第三阶段所有84个项目中的性能。

图23

B3：所有84个项目在第三阶段的平均SDG、容限区间和ODG图

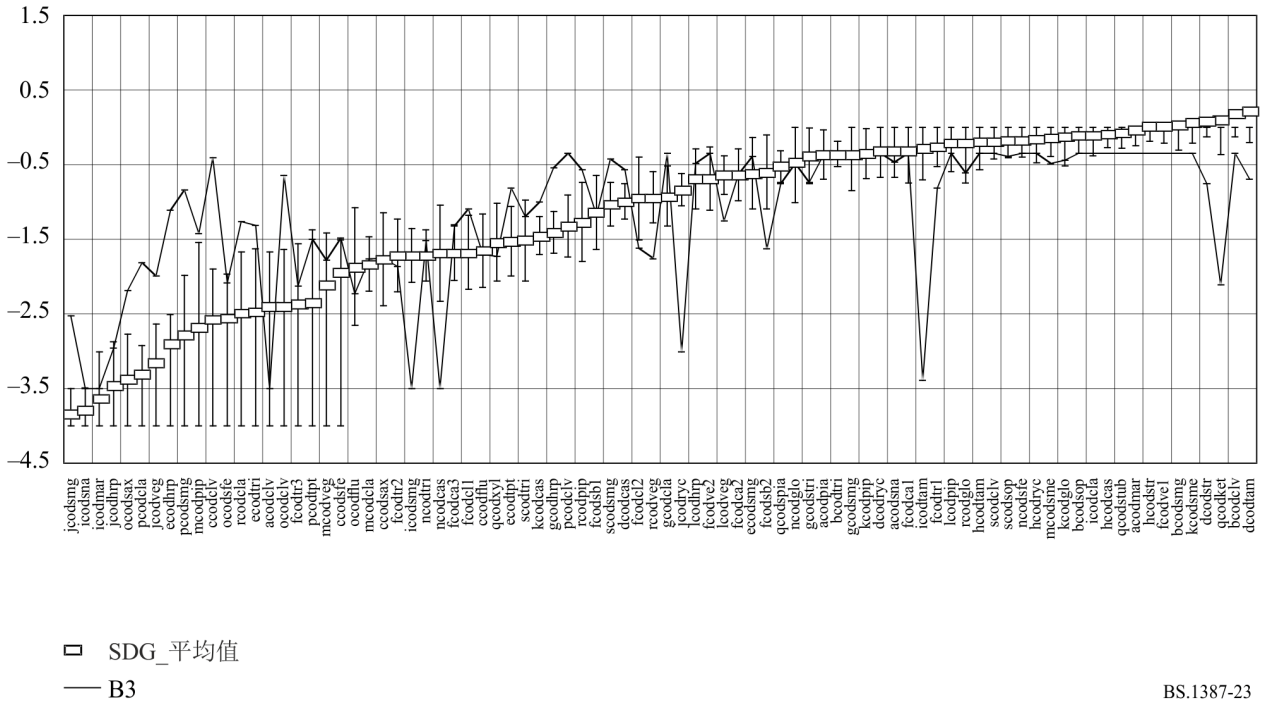


图24

FFTNnODG1：所有84个项目在第三阶段的平均SDG、容限区间和ODG图

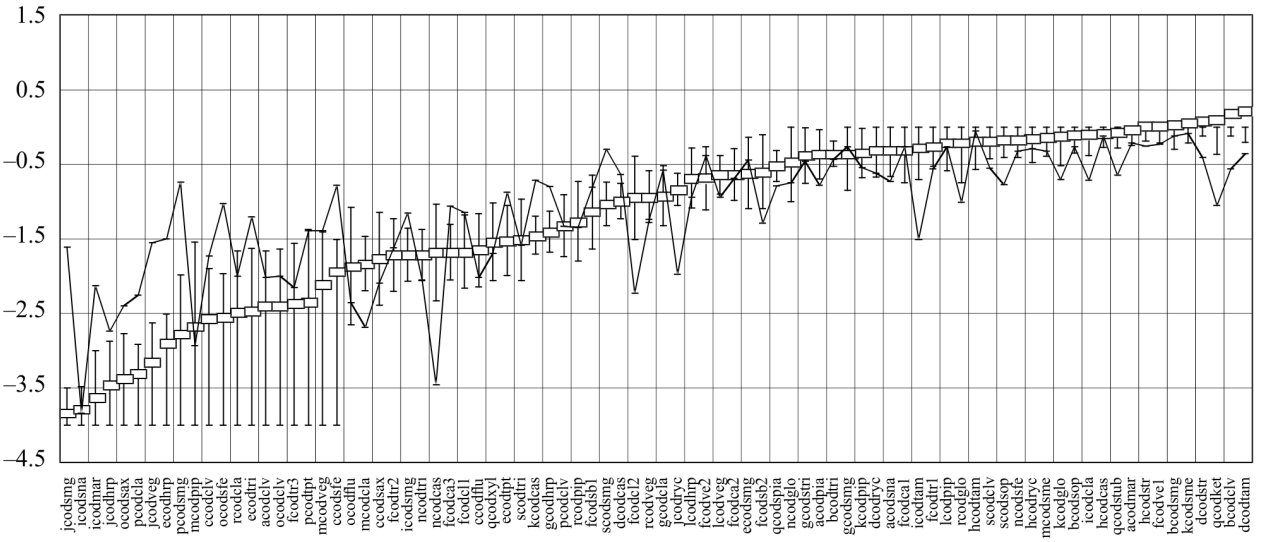
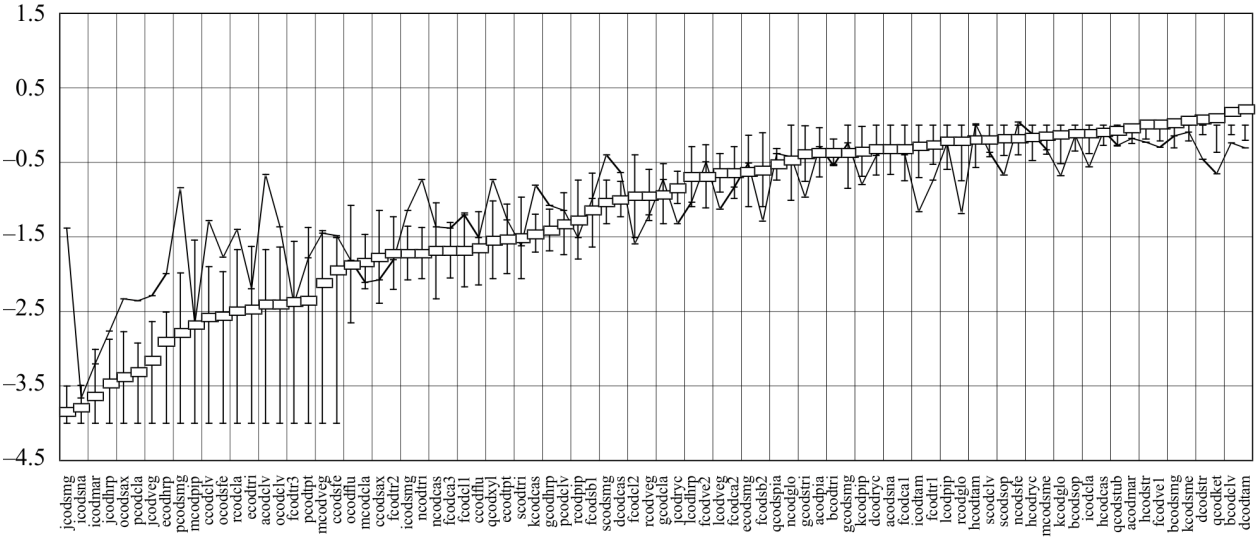


图25

CombNnODG3：所有84个项目在第三阶段的平均SDG、容限区间和ODG图



5 最佳模型版本的选择

向客观测试点提交了18个不同的模型版本：6个版本基于FFT，6个版本使用感知自适应滤波器组，6个版本使用FFT和滤波器组的组合。FFT组旨在用于能够实时实施的测量方案中，其余两组则需要更高的计算复杂度，并有望实现更高的精度。每组中的六个模型版本可以分为两个子组：一个使用神经网络，一个不使用神经网络。此外，亦包括了1996年4月的参考模型（B3）在数据库3上的性能，以供比较。

5.1 基于相关性的预选标准

– 使用两个数据集评估了18个不同模型版本和参考模型B3的性能，这两个数据集基本上并未用于模型训练（第一阶段和第三阶段第二部分以及CRC并未用于训练）。

– 数据库3是专门为评估感知模型而创建的。该数据库的大约一半被用于模型的训练。其余项目的主观结果和客观结果之间的相关性（DB3_2nd）被用于评估客观模型。

– CRC数据库由CRC创建，用于评估感知音频编码方案。该数据库并未用于感知模型的训练。所有项目的主观结果和客观结果之间的相关性被用于客观模型的评估。

数据库3的结果见表24和25（相关性和AES）。

在预选阶段，量化所有模型版本性能的所有输入均被纳入了考虑范围，且DB3的权重很高。在整体比较的基础上，决定对每组中看似最佳的两个模型版本进行进一步比较。表24显示了这六个版本的相关性（3乘以2个模型版本）。

表24

SDG和ODG之间的相关性

	FFTNnODG1	**FFTNnODG2**	**FiltODG2**	**FiltODG3**	**CombNnODG3**	**CombODG3**	**B3**
DB3_2nd	0.671	0.728	0.738	0.751	0.828	0.826	0.710
CRC	0.837	0.779	0.862	0.839	0.851	0.777	0.656

表25

绝对误差分数

	FFTNnODG1	**FFTNnODG2**	**FFTODG2**	**FFTODG3**	**CombNnODG3**	**CombODG3**	**B3**
DB3_2nd	2.96	2.79	3.16	3.16	2.91	2.56	2.39
CRC	1.55	1.85	1.61	1.67	1.61	1.90	2.78

两个基于FFT的模型版本在两个数据库上表现出不同的性能。为了从这两者中选择一个，需要提供进一步的标准。考虑到CRC数据库完全未知，FFTNnODG1的性能似乎稍好一些。

与基于FFT的版本相比，组合模型版本在两个数据库上均给出了更好的结果。不过，与滤波器组版本相比，其在CRC数据库方面的性能却稍逊一筹。滤波器组版本可以被视为组合模型版本的特例，其中基于FFT的版本输出变量的权重为零。因此，首选组合模型版本。

5.2 异常值数量分析

来自听力测试的主观数据的精度通过几个听者平均值周围的95%区间来表示。模型版本的性能亦通过观察异常值来评估。当主观数据和客观数据之间的差异大于置信区间的两倍时，相关项目被视为异常值。

表26显示了DB3的六个模型版本的异常值数量。敏感意味着模型版本指示比主观评估更低的音频质量，不敏感则意味着相反的情况。

表26

异常值

	FFTNnODG1	**FFTNnODG2**	**FiltODG2**	**FiltODG3**	**CombNnODG3**	**CombODG3**
敏感	10	4	4	4	3	5
不敏感	13	13	11	13	12	14
合计	23	17	15	17	15	19

如果观察异常值的数量，模型版本FFTNnODG2的性能似乎优于模型版本FFTNnODG1。模型版本CombNnODG3在更精确的版本中表现最佳。

5.3 异常值的严重性分析

表27

预测和SDG之间偏差超过1.0个差异等级的项目

异常值的数量	14	4	4	12	12	9	6
模型版本	B3	CombODG3	CombNnODG3	FFTNnODG1	FFTNnODG2	FiltODG2	FiltODG3
项目	jcodsmg	jcodsmg	jcodsmg	jcodsmg	jcodsmg	jcodsmg	jcodsmg
	qcodket	qcodket	pcodsmg	qcodket	qcodket	rcodpip	rcodpip
	pcodsmg	pcodsmg	ccodclv	pcodsmg	pcodsmg	qcodket	qcodket
	pcodcla	icodmar	acodclv	ocodsfe	pcodcla	pcodsmg	pcodsmg
	ocodclv			ncodcas	ocodsfe	ocodsax	ocodsax
	ncodcas			mcodcla	ncodglo	kcodpip	fcodsb2
	mcodpip			jcodveg	kcodcas	jcodveg	
	jcodryc			jcodryc	jcodveg	fcodsb2	
	icodtam			icodtam	jcodryc	fcodcl2	
	icodsmg			icodmar	icodtam		
	fcodsb2			fcodcl2	ecodhrp		
	ecodhrp			ecodhrp	ccodclv		
	ccodclv						
	acodclv						

表28

预测和SDG之间偏差超过1.5个差异等级的项目

异常值的数量	8	2	2	3	2	2	2
模型版本	B3	CombODG3	CombNnODG3	FFTNnODG1	FFTNnODG2	FiltODG2	FiltODG3
项目		jcodsmg	jcodsmg	jcodsmg	jcodsmg	jcodsmg	jcodsmg
	qcodket		pcodsmg				
	pcodsmg	pcodsmg		pcodsmg	pcodsmg		
						pcodsmg	pcodsmg
				ncodcas			
	ncodcas						
	jcodryc						
	icodtam						
	icodsmg						
	ccodclv						
	acodclv						

6 结论

根据开发过程中做出的决定，从所建议的18个版本中选择了两个模型版本，以满足客观测量方法的定义应用的要求。定义了一个低复杂度版本和一个高精度版本，前者被设计为允许经济高效的实时实施，后者则不一定实时操作。在选择过程中，应用并分析了上述标准。

基于FFT的模型（称为“FFTNnODG1”）被选为实时版本，原因如下：

就上述标准之一而言，基于FFT的版本均未显示出明显优于其他版本的优势。关于ODG和SDG之间的相关系数，每个已验证版本的异常值的数量和严重程度各有利弊。不过，应该注意的是“FFTNnODG1”在CRC’97数据库上实现了最佳相关性，而该数据库是完全未知的（0.837）。

对于精度更高的版本，优先选择结合FFT和滤波器组的组合版本，原因是这种方法还结合了纯滤波器组模型的子集，因此应该具有更好的性能潜力。总共有六个不同版本的组合模型可供选择。与其他版本相比，选定版本“CombNnODG3”显示出较少的异常值和较高的相关性（对CRC’97而言，r=0.851）。完整数据库3的此版本的相关性与“CombODG3”版本具有相同的数量级，但与其他版本相比显示出更高的相关性。

附件2的
后附资料2

参考数据库的描述

1 引言

在感知音频质量的客观测量方法的开发过程中，使用了诸多数据库进行训练和验证。

列出的一些数据库包含耳机和扬声器数据，而一些数据库只有耳机数据。对于所展示的扬声器和耳机而言，如果其存在单独数据集的数据库，则仅使用耳机数据。

项目被定义为在主观评估中使用的音频片段。条件是指单一的劣化条件。除了DB2和DB3研究之外，所有项目均用于所有实验条件。DB3部分用于训练，部分用于验证（84个项目中的52个用于第二阶段验证的训练）。

训练

- MPEG90
  - 每个项目的平均SDG相当一致地覆盖了从0.0到–4.0的范围。
  - ISO/IEC JTC 1/SC 2/WG 11 MPEG/音频测试报告，文件MPEG90/N0030，1990年10月。
- MPEG91
  - 每个项目至少88%的平均SDG高于–2.0，范围为0.1至–3.8。
  - ISO/IEC JTC 1/SC 2/WG 11 MPEG/音频测试报告，文件MPEG91/N0010，1991年6月。
- ITU92DI
  - 每个项目80%的平均SDG高于–2.0，范围为0.1至–3.4。
- ITU92CO
  - 每个项目至少96%的平均SDG高于–2.0，范围为0.2至–2.4。
- ITU93
  - 每个项目的平均SDG大多高于–2.0，范围为–0.1至–2.3。两个实验室的数据没有显著差异。
  - Grusec等人 [1997年]。
- MPEG95
  - 每个项目至少63%的平均SDG高于–2.0，范围为–0.2至–3.8。
  - Meares和Kim [1995年]。
- EIA95
  - 每个项目至少93%的平均SDG高于–2.0，范围为0.1至–3.7。
  - Grusec等人 [1997年]。
- DB2
  - 并非所有项目均适用于所有条件。

确认

- DB3
  - 并非所有项目均适用于所有条件。
- CRC97
  - 每个项目的平均SDG相当一致地覆盖了从0.1到–3.6的范围。
  - Soulodre等人 [1998年]。

以下各节描述了不同数据库中包含的项目以及所应用的条件。

2 每个数据库的项目

项目	MPEG90	MPEG91	ITU92DI	ITU92CO	ITU93	MPEG95	EIA95	DB2	DB3	CRC97
手风琴/三角琴		*								
阿萨·金德（音乐人）			*	*	*					
风笛						*		*	*	
风笛-2								*		
低音单簧管								*		
低音吉他	*		*	*						
低音合成器	*									
卡门		*								
响板	*		*	*	*	*		*	*	
单簧管					*		*	*	*	*
单簧管-单声道								*		
单簧管 2								*		
响棒									*	
Dalamas 乐队“星期三的游行”			*	*						
恐怖海峡乐队“穿越河流”			*				*			*
双贝斯										*
鼓								*		
焰火	*									
长笛										
乔治·杜克（音乐人）		*								
钟琴	*	*				*	*	*	*	
竖琴			*	*	*	*	*	*	*	*
喇叭								*		
定音鼓									*	
马林巴琴								*	*	
MPE 单声道 = 男性英语演讲单声道								*		
音乐和雨声							*			*
静音小号							*			
奥内特·科尔曼（音乐人）	*	*	*	*						
珍珠果酱乐队							*			
打击乐		*						*		
舒伯特钢琴曲									*	
律管						*			*	*
拉威尔“集市”			*							
莱·库德（音乐人）								*	*	
莱·库德（单声道）										
萨克斯管									*	
小军鼓									*	
莫扎特女高音									*	
女性英语演讲						*			*	
女声德语演讲									*	
男性英语演讲	*	*							*	
男声德语演讲			*	*	*				*	
斯特拉文斯基“八重奏”			*	*						
弦乐								*	*	
弦乐单声道								*		
苏珊·薇格“汤姆斯餐厅”	*	*	*	*				*	*	*
苏珊·薇格以及碎玻璃声							*			
铃鼓		*						*	*	
特雷西·查普曼（音乐人）	*									
三角琴			*	*				*	*	
小号								*	*	*
小号（海顿）	*									
大号								*	*	
德萨尔德·西纳·海曼（小提琴独奏）					*					
水声							*			
管乐合奏								*		
木琴									*	

3 实验条件

对于标示为kbit/s立体声的所有比特率，给出了总比特率，例如，256 kbit/s立体声意味着为立体声信号的两个声道总共分配了256 kbit/s。如果没有特别说明，立体声指的是独立声道编码。

3.1 MPEG90

三种比特率：64 kbit/s单声道、192 kbit/s立体声和256 kbit/s立体声，并非所有材料均可用于此数据库。

– Musicam（掩蔽模式通用子带集成编码和复用）。
– SB-ADPCM（子带 – 自适应差分脉冲编码）。

3.2 MPEG91

三种比特率：64 kbit/s单声道、192 kbit/s和256 kbit/s立体声。

– MPEG1第一层
– MPEG1第二层
– MPEG1第三层
– MUSICAM
– ASPEC（高质量音乐信号自适应谱感知熵编码）
– NICAM（接近即时的缩扩音频多路广播）。

3.3 ITU92DI

五个分发编解码器：240 kbits/s立体声。

每个项目由相同的编解码器处理三次，每次通过前电平下降0.1分贝。

– MPEG1第二层
– MPEG1第三层
– 杜比AC-2
– Aware（环境音）
– NHK

3.4 ITU92CO

六个馈送编解码器：360 kbits/s立体声。每个项目由相同的编解码器处理三次，每次通过前电平下降0.1分贝。

– MPEG1第二层
– MPEG1第三层
– 杜比AC-2。
– 杜比低延迟
– Aware（环境音）

3.5 ITU93

MPEG1第二层串联编解码器配置：

– 256 kbit/s立体声单独发射编解码器
– 192 kbit/s立体声单独发射编解码器（联合立体声编码）
– 八个360 kbit/s的馈送编解码器后跟一个256 kbit/s的发射编解码器，均为立体声
– 八个360 kbit/s的馈送编解码器后跟一个192 kbit/s的发射编解码器，均为立体声
– 五个360 kbit/s的馈送编解码器，然后是三个240 kbit/s的分发编解码器和一个256 kbit/s的发射编解码器，均为立体声
– 五个360 kbit/s的馈送编解码器，然后是三个240 kbit/s的分发编解码器和一个192 kbit/s的发射编解码器，均为立体声

3.6 MPEG95

编解码器的实施（64 kbit/s）：

– 从实施4个低分辨率和17个高分辨率时间/频率模型子集的6个编解码器提供的较大编码方法集中选择了22种编码变体。
– 参与机构包括美国电话电报公司（AT&T）、德国夫琅霍夫研究所（Fraunhofer）、索尼公司、GCL、RAI/阿尔卡特公司和飞利浦公司。
– 所有项目均为单耳录音，并以双耳方式呈现。

3.7 EIA95

– Eureka 147/MPEG1第二层#1 224 kbit/s立体声（联合立体声编码）
– Eureka 147/MPEG1第二层#2 192 kbit/s立体声（联合立体声编码）
– 美国电话电报公司/朗讯 160 kbits/s立体声
– 美国电话电报公司/朗讯/阿玛提#1 128 kbits/s立体声
– 美国电话电报公司/朗讯/阿马蒂#2 160 kbits/s立体声
– 美国之音/JPL 160 kbits/s立体声
– USADR-FM #1 128-256 kbit/s立体声（可变比特率）
– USADR-FM #2 128-256 kbit/s立体声（可变比特率）
– USADR-AM 96 kbits/s立体声

3.8 DB2

– MPEG1第二层 256 kbit/s立体声，1、3、5、7和9级
– 杜比AC2 256 kbit/s立体声，1、3、5、7和9级
– MPEG1第二层 192 kbit/s立体声（联合立体声编码）
– MPEG1第二层 64 kbits/s单声道
– MPEG2第二层 64 kbits/s单声道
– MPEG1第二层 384 kbits/s立体声
– MPEG1第三层 128、160、192 kbits/s，均为立体声。
– APT-X 256和384 kbit/s，均为立体声。

– 量化失真
– 模拟录制1、2、3个阶段
– 剪辑

3.9 DB3

– NICAM
– MiniDisc和MiniDisc+第二层 192 kbit/s，立体声（联合立体声编码）
– 杜比AC2 256 kbit/s立体声，1、3、5、7和9级
– MPEG1第二层 从瑞士电信数据库中选择，>192 kbit/s立体声
– MPEG1第三层 128和160 kbit/s立体声（联合立体声编码）
– MPEG AAC 128 kbit/s立体声（联合立体声编码）
– MPEG第三层 128+第二层，384+第二层，224 kbit/s，均为立体声
– 杜比AC3 256 kbits/s立体声
– 杜比AC3 256+MPEG第二层，224 kbit/s，均为立体声
– 量化失真
– THD（总谐波失真）
– 噪声

3.10 CRC97

– 美国电话电报公司演艺中心（PAC）64、96、128和160 kbit/s，均为立体声
– 杜比AC3 128、160和192 kbit/s，均为立体声
– MPEG1第二层软件 128、160和192 kbit/s，均为立体声
– MPEG1第二层硬件（ITIS） 96、128、160、192 kbits/s，全立体声
– MPEG4 AAC 96和128 kbit/s，均为立体声
– MPEG1第三层 128 kbits/s立体声

4 DB2和DB3每个条件的项目

4.1 DB2

	条件编号	项目
第一测试点，日本NHK		
第二层，256 kbits/s	1个阶段	CO13 CLA、RYC、SB1、STR
	3个阶段 CO11	CLA、RYC、SB1、STR
	5个阶段 CO19	CLA、RYC、SB1、STR
	7个阶段 CO18	CLA、RYC、SB1、STR
	9个阶段 CO15	CLA、RYC、SB1、STR
美国全国广播公司（杜比AC2）	1个阶段 CO1A	CAS、RYC、STR、WIN

3个阶段 CO12	CAS、RYC、STR、WIN
5个阶段 CO17	CAS、RYC、STR、WIN
7个阶段 CO16	CAS、RYC、STR、WIN
9个阶段 CO14	CAS、RYC、STR、WIN

条件编号	项目	

第二测试点，丹麦DR

第二层，256 kbits/s 1个阶段	CO2B	CLA、RYC、SB1、STR
第二层，192 kbit/s联合立体声	CO25	CLA、RYC、SB1、STR
第二层，64 kbit/s单声道	CO27	MLA、MPE、MTR、MYC
美国全国广播公司（杜比AC2）5个阶段	CO29	CAS、RYC、STR、WIN
MPEG2/L2 LSF	CO22	MLA、MPE、MTR、MYC
模拟1	CO23	PER
模拟2	CO2A	PER
模拟3	CO28	PER
误差1	CO24	GLO、HRN、TRI
误差2	CO21	GLO、HRN，TRI
剪报	CO26	BAS、CL2、TUB

第三测试点，挪威NRK

第二层，384 kbits/s	CO34	CLA、RYC、SB1、STR
第二层，256 kbits/s 1个阶段	CO31	CLA、RYC、SB1、STR
美国全国广播公司（杜比AC2）5个阶段	CO3B	CAS、RYC、STR、WIN
第三层（ASPEC3），192 kbits/s	CO32	CLA、STR、TAM、VEG
第三层（ASPEC3），128 kbits/s	CO39	CLA、STR、TAM、VEG
第三层（ASPEC3），160 kbits/s	CO3A	CLA、STR、TAM、VEG
APT-X，256 kbits/s	CO33	HAR、SB2、STR、TPT
APT-X，384 kbits/s	CO36	HAR、SB2、STR、TPT
量化分布1	CO35	DRU
量化分布2	CO37	DRU
量化分布3	CO38	DRU

测试项目

STR	瑞典民间音乐，SR录音，以前用过
SB1	风笛，SR录音
SB2	风笛，SR录音
CLA	单簧管，SQUAM 16/2
TAM	手鼓，SR录音，以前用过
WIN	斯特拉文斯基，管乐合奏，以前用过
TPT	小号，SQUAM 21/2

HAR 英国广播公司丑角剧团录制 G 49/17

VEG 苏珊·薇格，《老师傅》，以前用过

CAS 响板，SQUAM 27

SPE 德语演讲，SQUAM 54

RYC 莱·库德，CD：JAZZ tr 11 （0.25 – 0.47）

PER 日本低音马林巴打击乐，CD：Sony/CBS 32DC 5027

HRN 喇叭，SQUAM 23/2

GLO 钟琴，SQUAM 35/1，以前用过

TRI 三角琴，SQUAM 32/2

DRU 鼓，SQUAM 28

CL2 单簧管，SQUAM 16/2

BAS 低音单簧管，SQUAM 17

TUB 大号，SQUAM 24

MPE SPE的单声道混音

MTR STR的单声道混音

MLA CLA的单声道混音

MYC RYC的单声道混音

测试/项目	单簧管	单簧管单声道	莱库德（音乐人）	莱库德（单声道）	风笛 1	弦乐	弦乐单声道	响板	管乐合奏	MPE 单声道	铃鼓	苏珊·薇格	竖琴	风笛 2	小号	鼓	打击乐	钟琴	喇叭	三角琴	贝斯	单簧管 2	大号
MPEG1 第二层，256 kbit/s，1 个阶段	X		X		X	X																	
MPEG1 第二层，256 kbit/s，3 个阶段	X		X		X	X																	
MPEG1 第二层，256 kbit/s，5 个阶段	X		X		X	X																	
MPEG1 第二层，256 kbit/s，7 个阶段	X		X		X	X																	
MPEG1 第二层，256 kbit/s，9 个阶段	X		X		X	X																	
杜比 AC2，256 kbit/s，1 个阶段			X			X		X	X														
杜比 AC2，256 kbit/s，3 个阶段			X			X		X	X														
杜比 AC2，256 kbit/s，5 个阶段			X			X		X	X														
杜比 AC2，256 kbit/s，7 个阶段			X			X		X	X														
杜比 AC2，256 kbit/s，9 个阶段			X			X		X	X														
MPEG1 第二层，192 kbit/s，联合立体声	X		X		X	X																	
MPEG1 第二层，64 kbit/s，单声道		X		X			X			X													
MPEG2 第二层，64 kbit/s，单声道		X		X			X			X													
MPEG1 第二层，384 kbit/s	X		X		X	X																	
MPEG1 第三层（ASPEC 3），192 kbit/s	X					X					X	X											
MPEG1 第三层（ASPEC 3），128 kbit/s	X					X					X	X											
MPEG1 第三层（ASPEC 3），160 kbit/s	X					X					X	X											
APT-X，256 kbit/s						X							X	X	X								
APT-X，384 kbit/s						X							X	X	X								
量化失真 1																X							
量化失真 2																X							
量化失真 3																X							
模拟录音，1 个阶段																	X						
模拟录音，2 个阶段																	X						
模拟录音，3 个阶段																	X						
比特误差 1																		X	X	X			
比特误差 2																		X	X	X			
剪辑																					X	X	X

4.2 DB3

测试/项目	Name	1	2MD	2MD L2	3.1步	3.3步	3.5步	3.7步	3.9步	4	5	6 低	6 高	7	8	9	10	11	12	13
13(1) 长笛	flu							X	X											
16(2) 单簧管	cla				X		X			X		X		X	X					
20(1) 萨克斯管	sax							X	X											
21(2) 小号	tpt					X	X													
24(2) 大号	tub																	X		
26(1) 响棒	clv		X	X			X	X	X											X
27 响板	cas									X			X			X	X		X	
28 小军鼓	sna	X																		X
30 定音鼓	ket																	X		
32(1/2) 三角琴	tri		X	X	X	X				X									X	
35(1/2) 钟琴	glo											X	X						X	
36(1) 木琴	xyl																	X		
40(1) 竖琴	hrp				X	X					X			X						
49 女性英语演讲	sfe							X	X										X	
54 男性德语演讲	smg	X	X	X	X	X	X				X									
60 舒伯特钢琴曲	pia																	X		X
61 莫扎特女高音	sop		X	X																
53 女性德语演讲	sfg																			
50 男性英语演讲	sme												X		X					
Ref_tam	tam	X														X	X			
Ref_str	str															X	X			
Ref_har	har																			
莱·库德	ryc										X					X	X			
苏珊·薇格	veg									X	X	X		X	X					
律管	pip											X	X	X	X					
马林巴琴	mar	X																		X
风笛	sb1									X										
名称		i	b	s	g	e	p	c	o	f	j	r	k	l	m	h	d	q	n	a

1.) NICAM
2.) MD 和 MD + L2 (192 kbit/s)
3.) AC2 (256 kbit/s), 1, 3, 5, 7, 9
4.) L2 (ST d-b, ³192 kbit/s)
5.) 第二层（256 kbit/s），8 个阶段
6.) 第三层，128 和 256 kbit/s
7.) AAC, 128 kbit/s
8.) L3 (128) + L2 (384) + L2 (224)
9.) AC3 (256)
10.) AC3 (256) + L2 (224)
11.) 量化失真
12.) THD
13.) 噪声

名称示例：
参考: irefflu
测试: icodflu

词汇表

绝对误差分数（AES）

AES源自一个专门为评估从客观感知测量方法获得的结果质量而开发的公式。它考虑了主观听力测试平均值的置信区间。

基本音频质量

基本音频质量被定义为全局主观属性，其中包括参考信号与其处理版本之间的任何和所有检测到的差异。

编码余量

编码余量是一个质量参数，它衡量当听不见的编码伪像变得听得见时，这些伪像相对于阈值的余量。

模型输出变量（MOV）

MOV是感知测量方法的中间输出值。这些变量基于基本的心理声学调查结果，因此可以用于进一步表征编码伪像。

客观差异等级（ODG）

ODG是感知测量方法的主要输出参数。它对应于SDG，是给出全局基本音频质量的测量参数。ODG的范围在0到4之间。

离线测量

不影响正在进行的节目传输的测量程序。

在线测量

依赖于正在进行的节目传输或部分节目传输的测量程序。

主观差异等级（SDG）

在基于ITU-R BS.1116建议书的听力测试中，隐藏参考和参考的经处理版本的基本音频质量按照五级损伤等级进行分级。差异等级被定义为待测信号的等级减去参考信号的等级。在理想情况下，SDG的范围应在0到–4之间。如果未正确识别参考，则SDG为正值。

缩写词

ADB	平均失真块
AES	绝对误差分数
ASD	听觉频谱差异
Avg	平均值（线性）
BAQ	基本音频质量
Bw	带宽
CI	置信区间
CM	编码余量
DBn	数据库 *n*（1、2或3）
DC	直流电
DFT	离散傅里叶变换
DIX	干扰指数
DUT	待测设备
EHS	误差谐波结构
ERB	等效矩形带宽
fac	因子
FFT	快速傅里叶变换
FIR	有限脉冲响应
IIR	无限脉冲响应
ISO	国际标准化组织
JNLD	最小可觉电平差
MFPD	最大滤波检测概率
MOV	模型输出变量
MPEG	活动图像专家组
NL	噪声响度

NMR	噪声掩蔽比
OASE	客观音频信号评估
OCM	客观编码余量
ODG	客观差异等级
PAQM	感知音频质量测量
PEAQ	感知音频质量的客观测量
PERCEVAL	感知评估
POM	感知客观测量
REF	参考信号
res	分辨率
r.m.s.	均方根
ROEX	四舍五入指数
ROV	输出值比率
SCM	主观编码余量
SDG	主观差异等级
SNR	信噪比
SPL	声压级
SUT	待测信号
THD	总谐波失真
Win	窗口平均值

参引

AURAS, W. [1984年9月] 计算任何声音信号的正确性，这有助于与听觉相关的声音分析。德意志联邦共和国慕尼黑工业大学电气技术学院论文。

BEERENDS，J. G.和STEMERDINK，J. A. [1992年12月] 一种基于心理声学声音表示的感知音频质量测量。音频工程学会杂志，第40卷，第963-978页。

BEERENDS，J. G.和STEMERDINK，J. A. [1994年2月] 对音乐编解码器质量测量的认知方面进行建模。第96届AES大会的文稿，预印本 3800。荷兰阿姆斯特丹。

BEERENDS，J. G.和STEMERDINK，J. A. [1994年3月] 一种基于心理声学声音表示的感知语音质量测量。音频工程学会杂志，第42卷，第115-123页。

BEERENDS，J. G.，van den BRINK、W. A. C.和RODGER，B. [1996年5月] 信息掩蔽和感知流在音乐编解码器质量测量中的作用。第100届AES大会的文稿，预印本4176。丹麦哥本哈根。

BRANDENBURG，K. [1987年] 低比特率音频编码质量评估。第82届AES大会的文稿，预印本 2433。英国伦敦。

BREGMAN，A.S. [1990年] *听觉场景分析：声音的感知组织*。美国剑桥麻省理工学院出版社。

COHEN, E. A.和FIELDER, L. D. [1992年5月] 确定录音环境的噪声标准。*音频工程学会杂志*，第40卷，第384-402页。

COLOMES, C.、LEVER, M.、RAULT, J. B.和DEHERY, Y. F. [1995年4月] 一种应用于音频比特率降低的感知模型。*音频工程学会杂志*，第43卷，第233-240页。

FEITEN, B. [1997年3月] 使用差分信号测量感知编解码器的编码余量。第102届AES大会，预印本4417。德意志联邦共和国慕尼黑。

GRUSEC, T.、THIBAULT, L.和SOULODRE, G. [1997年9月] EIA/NRSC DAR系统主观测试。第1部分：音频编解码器质量。*IEEE广播汇刊*第43卷第**3**期。

KARJALAINEN, J. [1985年3月] 音频系统音质评估的新听觉模型。IEEE声学、语音和信号处理国际会议（ICASSP）会议录，第608-611页。美国佛罗里达州坦帕市。

LEEK，M.R.和WATSON，C.S. [1984年] 学习检测听觉模式成分。*美国声学学会杂志*，第76卷，第1037-1044页。

MEARES，D. J.和KIM，S.W. [1995年7月] “NBC时间/频率模块主观测试：总体结果”，ISO/IEC JTC 1/SC 29/WG 11 N0973 MPEG 95/208。

MOORE，B.C. [1986年] *听觉中的频率选择性*。学术出版社，英国伦敦。

MOORE，B.C. [1989年] *听觉心理学导论*。学术出版社，英国伦敦。

PAILLARD，B.，MABILLEAU，P.，MORISETTE，S.和SOUMAGNE，J. [1992年] 感知评估（Perceval）：音频信号质量的感知评估。*音频工程学会杂志*，第40卷，第21-31页。

SCHROEDER，M. R.、ATAL，B. S.和HALL, J. L. [1979年12月] 利用人耳的掩蔽特性优化数字语音编码器。*美国声学学会杂志*，第66卷，第1647-1652页。

SOULODRE，G.、GRUSEC，T.、LAVOIE，M.和THIBAULT，L.，[1998年3月] 对最高级双声道音频编解码器的主观评估。*音频工程学会杂志*。

SPORER，T. [1997年10月] 客观音频信号评估 – 应用心理声学模拟数字音频的感知质量。第103届AES大会，预印本 4512。美国纽约。

TERHARDT，E. [1979年] 计算虚拟音高，听力研究。第1卷，第155-182页。

THIEDE，T.和KABOT，E. [1996年] 比特率缩减音频的新感知质量测量。第100届AES大会的文稿，预印本 4280。丹麦哥本哈根。

TREURNIET，W. C. [1996年] 用听觉模型模拟单个听者。音频工程学会会议录，再版号4154。丹麦哥本哈根。

von Bismarck，G. [1974年] 作为稳定声音音色属性的锐度。《声学期刊》（Acustica），30，第159-172页。

ZWICKER，E.和FASTL，H. 心理声学，事实和模型，德意志联邦共和国柏林海德堡施普林格出版社。

ZWICKER，E.和FELDTKELLER，R. [1967年] 作为信息接收者的耳朵，德意志联邦共和国斯图加特赫泽尔出版社。

参考文献

GRUSEC T.、THIBAULT L.和SOULODRE，G. [1995年] 高质量音频编码系统的主观评估：双声道情况下的方法和结果。预印本 4065（F-5），AES会议录。美国纽约。