引用格式：黄心仪，谢凌云，王鑫.三维声双耳渲染算法音质主客观评价分析[J].中国传媒大学学报（自然科学版），2023，30(04)：62-68.
文章编号：1673-4793(2023)04-0062-07

三维声双耳渲染算法音质主客观评价分析

黄心仪[1]，谢凌云[2]，王鑫[1*]
（1.中国传媒大学音乐与录音艺术学院，北京 100024；2.中国传媒大学信息与通信工程学院，北京 100024）

摘要：随着三维声的应用逐渐广泛，对三维声进行双耳渲染成为了新的技术热点，如何有效地评价三维声双耳渲染算法成为关键问题。本文针对6种三维声双耳渲染算法进行了音质维度的主观评价实验，对实验数据进行方差分析和回归分析。通过对双耳录音的实验素材进行客观特征的提取和筛选，与主观评价结果进行偏最小二乘回归分析，建立了总体音质评价维度的客观评测模型，并探究了主观感知与客观特征之间的关联。主观实验结果表明，进行双耳渲染算法处理会对音质造成损伤，但对音质进行算法补偿，可以在一定程度上弥补渲染算法造成的音质损伤。客观预测模型表明音质与2560～5120Hz和40～320Hz这两个频段的时频特征高度相关，例如谱通量和谱滚降等。低频段的双耳互相关系数和侧向声能比也是影响音质维度的重要特征。
关键词：三维声；双耳渲染算法；主观评价；客观评测模型；音质
中图分类号：TN912.2　**文献标识码：**A

Analysis of subjective and objective evaluation of sound quality by three-dimensional sound binaural rendering algorithm

HUANG Xinyi[1], XIE Lingyun[2], WANG Xin[1*]
(1. School of Music and Recording Arts, Communication University of China, Beijing 100024, China; 2. School of Information and Communication Engineering, Communication University of China, Beijing 100024, China)

Abstract: As the application of 3D sound becomes increasingly widespread, binaural rendering of 3D sound has emerged as a new technological focus. The effective evaluation of binaural rendering algorithms for 3D sound has become a key issue. In this paper subjective quality assessment experiments on six different binaural rendering algorithms for 3D sound were conducted, followed by variance analysis and regression analysis of the experimental data. Objective features were extracted and selected from binaural recordings, and a partial least squares regression analysis was performed to establish an objective evaluation model for overall sound quality dimensions. The relationship between subjective perception and objective features was also explored. The subjective experimental results indicate that the binaural rendering algorithm processing can have a negative impact on sound quality. However, compensating for sound quality using algorithmic adjustments can partially mitigate the sound quality degradation caused by the rendering algorithm. The objective prediction model reveals that sound quality is highly correlated with time-frequency features in the frequency ranges of 2560-5120Hz and 40-320Hz, such as spectral flux and spectral rolloff. Additionally, the interaural cross-correlation coefficient and

作者简介(*为通讯作者)：黄心仪(2000-)，女，硕士研究生，主要从事音乐感知与空间音频研究。E-mail：rubydiva@163.com；谢凌云(1977-)，男，博士，副研究员，主要从事音频信号处理与心理声学研究。E-mail：xiely@cuc.edu.cn；王鑫 (1978-)，女，博士，教授，主要从事音乐感知与音乐声学研究。E-mail：metero_wx@cuc.edu.cn

lateral sound energy ratio in the low-frequency range are important features influencing sound quality dimensions.

Keywords: three-dimensional sound; binaural rendering algorithm; subjective evaluation; objective evaluation model; sound quality

1 引言

环绕声系统经历几十年的发展，观众的观影体验逐渐被改变，人们在此基础上增加了高度声道，声音由此可以进行精准定位，在三维空间中真实地呈现。尽管Dolby Atoms、DTS:X等重放系统技术已被广泛应用，但其复杂的扬声器阵列在日常生活中难以实现，因此对三维声进行双耳可听化处理日益具有实用价值，且三维声双耳渲染效果的评价也成为了人们关注的问题。

音频评价通常有两种方法，即主观评价和客观评测。主观评价是以人为主体评价音频的听感，结果往往准确且令人信服，但被听音环境等种种条件所限制，费时费力且成本较高。客观评测指采用计算机信息处理技术来判断音频的质量，相对高效便捷，但也存在模型不匹配、结果不够准确等问题。因此若能将主观评价与客观评测有机结合，将能更加全面地对音频进行评测。

三维声双耳渲染算法，是将已经制作好的多声道节目进行渲染处理，使其达到可以用耳机还原三维声听感效果的目的。随着三维声的应用逐渐广泛，许多学者开始讨论如何对三维声进行有效的听感评价。对于用耳机重放的双耳声信号的评价，Lorho提出16个评价参数，主要包含定位、空间、音质三个维度属性[1]。Rumsey等人采用直接下变换等算法后，使用音质有些裂变的声音素材为研究对象，以变换前的原始信号为参考信号，考察总体音质与子评价维度的关系，并提出使用偏最小二乘回归统计方法（Partial Least-Squares Regression, PLSR）[2]。Lindau等人对比真实房间扬声器重放和使用个性化房间双耳脉冲响应(Binaural Room Impulse Response, BRIR）以及非个性化BRIR的双耳重放系统，采用空间音频质量列表(Spatial Audio Quality Inventory, SAQI）方法进行个性化动态双耳渲染听感实验[3]。Reardon等人对六种双耳渲染器进行了较为全面的评测，并将评测分为了定量特征、定性特征及总体偏好三个部分[4]。范欣欣等人针对三维声双耳渲染算法设计了主观评价实验，利用方差分析和回归分析，对比不同渲染算法的特点和信号适用性，以及探究总体评价与音质、定位、空间之间的关联[5]。

随着近年来通信系统的快速发展，音频客观评测方法的研究也取得了较快的进展。近年来对于音频信号的客观评测大多都为基于有参考信号的客观评测方法，但就目前发展技术而言，其客观评测的结果与主观评价之间的关联性较低，在理论方法和技术层面还有待突破性进展。目前国际上唯一的数字音频质量客观评测标准ITU-R BS.1387就是此类基于有参考信号的客观评测方法，其采用了音频质量感知评价模型(Perceptual Evaluation of Audio Quality, PEAQ），通过模拟人耳听觉特性，可以较准确地得到待测音频质量的得分，随后也有大量基于PEAQ的改进模型出现，但这些模型大多针对独立声道，无法考察声道之间的相关信息，且三维声不同于传统的音频评测，三维声的空间属性以及其涉及的主观评价维度更为复杂，国际上目前还没有针对三维声的客观评测标准。覃龙靖等人在范欣欣的工作基础上对五种双耳渲染算法的渲染效果进行了客观评测，对双耳信号提取双耳特征和单耳特征，并进行特征选择和降维，最终选择了岭回归建立了总体评价和其二级维度，即音质、定位和空间感的客观评测模型，并探究了主观感知和客观参数之间的关系[6]。

综上所述，虽然目前已存在一些学者进行了针对三维声的主观评价实验，并探讨了相关的主观评价术语，但这些实验仍然存在一些问题。范欣欣等人的工作中主观评价实验在一个信号的基础上同时听评多个维度，可能导致各维度得分之间存在相互关联。在客观评测方面，覃龙靖等人的工作中基于评测准确率建立的客观模型，在优选特征的可解释方面较为困难，很难去解释主观感知的机理。因此，本文围绕音质维度进行三维声主观评价实验，深入探讨了总体音质与双耳渲染算法之间的关联，并从探究主观感知的角度出发，基于偏最小二乘回归分析（PLSR）建立了总体音质维度的客观评测模型。

2 主观评价实验

2.1 实验素材及渲染算法

实验素材首先在素材库中进行选取。对筛选的实验素材进行预实验,所有信号进行不同的双耳渲染算法处理,让被试对不同双耳声信号进行总体音质和4个二级维度进行评价打分,实验素材包含了影视声音、环境声音、流行音乐、交响乐、管乐合奏、弦乐合奏、阿卡贝拉、电子音乐等各种类型。预实验结束后整理数据,选取了在音质组维度中最易听辨的五个信号素材,具体素材描述见表1。

表1 三维声双耳渲染算法主观评价实验素材

序号	信号类型	时长	信号特点
1	真人电影	26s	包含对白,定位和距离不同的音效声,交响乐的背景音乐,不同声音空间转换,声音元素丰富
2	流行音乐	32s	定位清晰,常规流行音乐配器
3	弦乐四重奏	22s	音乐厅录制的弦乐四重奏,音质好,声像深度、宽度明显,音乐厅厅堂感
4	阿卡贝拉	20s	教堂录制的阿卡贝拉,音质好,高中低频不同声部,教堂厅堂感
5	电子音效	29s	包含高、中、低频的电子乐器,不同乐器定位在不同的空间位置,存在移动声源,定位清晰,定位变化明显

原始信号素材均为5.1.4或5.0.4的三维声信号,经过6种不同的双耳渲染算法,最终形成30个双耳渲染实验信号。6种算法均为目前国内外较成熟的公司或科研机构研发,渲染算法的选取原则是尽可能地涵盖目前国内外比较全且比较主流的双耳渲染算法类别,同时主观听感上各算法之间具有一定差异。本次选取的6种渲染算法如表2所示。

2.2 实验环境及设备

本次音质主观评价实验在中国传媒大学三维声混录棚内进行,其房间声学参数、音响系统、前期系统校准过程与范欣欣论文中所述一致[5]。响度校准后,利用人工头RS Tech Head001连接RME Fireface UCX声卡,进行了标准双耳声信号的录制。录制的信号包括5个素材的扬声器三维声参考信号以及5个素材对应的6个双耳渲染音频,共35个信号。录制过程中,所有信号连贯播放,中途未更改任何参数,录制完毕的音频信号将预备后续客观评测部分的音频特征提取。

表2 三维声双耳渲染算法

序号	算法特点
A	将多声道信号通过加权合并直接线性变换得到立体声信号
B	基于HRTF卷积,音质劣化不明显
C	基于HRTF卷积,声音外化效果明显
D	经过Ambisonic编解码,对Ambisonic信号计算渲染矩阵函数,基于HRTF卷积[7]
E	经过Ambisonic编解码,对Ambisonic信号计算渲染矩阵函数,再采用虚拟扬声器方法进行渲染,音色较为明亮[8]
F	基于虚拟半球幅度平移技术[9]

2.3 被试人员及流程

本次实验共招募42名被试,平均年龄在18~24岁之间,均有着6年以上的音乐及乐器学习经验以及1年以上的录音混音学习经验,学习专业以录音艺术为主,均能较好地理解评价维度及其含义。总体音质的含义即音质的总体评价,从声音是否音色均衡,浑厚、清晰可辨、有力度等方面进行音质的总体评价(劣-优)。

本次实验采用MUSHRA的方法进行打分,被试每次只针对其中一个评价术语进行实验,依次听辨所有实验素材,并对每个素材中的各个算法进行打分。本次实验采取了双盲模式,实验页面中的算法均由A~F六个字母进行表示,且每次打开新的页面时,6个渲染音频将被随机填充至A~F六个位置处。实验页面中的6个音频播放位置均设置了进度条,被试可以根据自己情况选

取播放进度，并随时选择暂停或播放。实验过程中，每个算法的打分均需以扬声器三维声信号作为参考，而非算法之间的横向比较。之所以选择这种听辨方式，是因为预实验中发现被试间的一致性较好。每个素材打分结束后，被试需要填写最影响自己判断该评价维度的声音元素，主试将根据被试填写的元素判断其数据可靠性，并在了解大部分被试的整体关注声音元素的基础上，在后面的被试无法听辨合适元素时适时进行引导。每位被试实验时长为1小时左右，并中途设置休息时间，填写问卷调查。实验打分页面如图1所示。

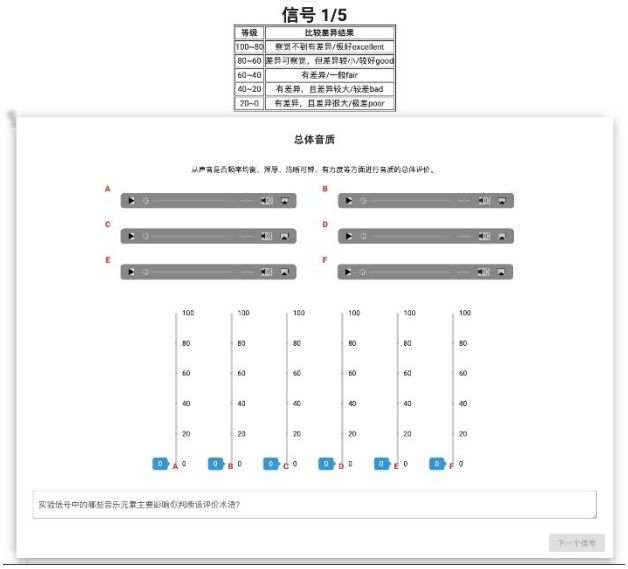

图1 主观评价实验打分页面

2.4 实验结果及分析

利用一致性系数进行被试间信度检验。为了消除被试之间打分的差异，首先将被试的打分进行了归一化处理，转化成Z分数[10]，随后将被试的Z分数进行一致性检验，并剔除了少量不可靠的被试数据使得所有评价维度的克朗巴哈系数均为0.7以上。

本次实验采用实验信号(5水平)*渲染算法(6水平)双因素方差分析对实验结果进行分析讨论，所有统计分析都采用双侧检验，且显著性水平为0.05，用η_p^2估算效应量。总体音质维度内，不同渲染算法对于各子维度的主效应结果，以及实验信号与渲染算法的交互效应结果如表3所示。

表3 总体音质渲染算法的主效应及渲染算法和实验信号的交互效应

方差分析	F	p	df	η_p^2
主效应	11.65	<.001	5, 65	0.47
交互效应	1.947	0.010	20, 260	0.13

不同渲染算法在总体音质上的主效应结果如图2所示，用标准误表征数据的离散程度。从图中可以看出算法A除了在总体音质维度显著高于算法B，在所有二级维度评价中算法A和算法B差异不显著，这说明算法B在进行了音质补偿后，可以做到在各个子维度接近直接下变换算法的听感效果。其次得分表现位于中间的是算法C，算法D和F的表现较差。值得注意的是算法D与算法E渲染原理相似，都是基于Ambisonic解编码，但二者的均值与标准差都有着较大差异。

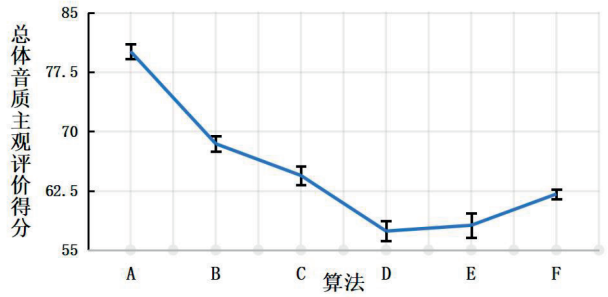

图2 渲染算法在总体音质维度上主效应结果图

图3展示了实验信号与渲染算法在总体音质上的交互效应结果。从图中可以看出算法A在各个信号上得分都非常高，且远高于其他信号，这说明现有的双耳渲染算法在音质方面仍然存在明显的损伤。此外，算法B相比于其他算法，整体分值较高，说明对不同声音类别的适用性较好。

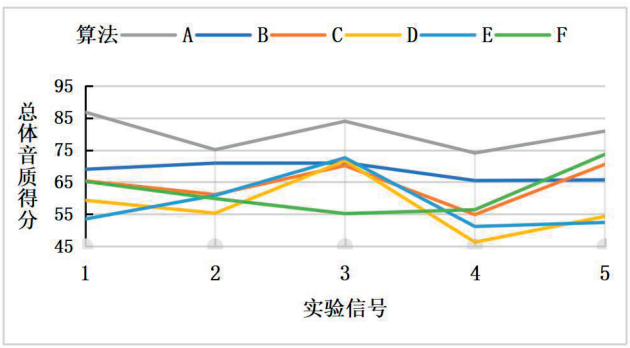

图3 总体音质的交互效应结果

3 客观评测分析

3.1 特征提取与筛选

本次实验主要评测了与音质相关的维度，因此在客观特征的选取上尽量包含表征音质的特征以及影响音质感知的双耳特征，所有特征的物理含义明显且

易于解释。本文共选取了16个客观特征,具体如表4所示。

所有客观特征的提取均基于Mir toolbox工具包获得。在提取过程中,先对所有信号分帧处理,选用50ms的帧长以及50%的帧移。考虑到部分音频特征在不同频段内有着不同的物理含义,因此对于所有双耳特征以及除明亮度、粗糙度、频谱不规则度以外的时频特征,在提取时将音频素材先按照一个倍频程的方式分成10个频段,在每一频段内逐一提取相关特征,再统计每个特征的均值和方差。所有时频特征也同时计算了全频段的分帧后结果,并进行均值和标准差的计算。对于时频特征的提取,由于合并声道会出现相位抵消等各种问题,因此时频特征将分别对左、右声道进行计算,再对其左右耳特征的统计结果做均值处理。在后续的计算过程中,其中均值均用m表示,标准差均用std表示。

表4 所提取的客观特征列表

特征名	特征中文名称	是否分频	是否提取全频段特征	统计特征维数	合计维数
IACC	双耳互相关系数	是	否	20	288
LF	侧向声能比	是	否	20	
RMS	均方根	是	是	22	
ZeroCross	短时过零率	是	是	22	
Roughness	粗糙度	否	是	2	
Spectral Brightness	明亮度	否	是	2	
Spectral Irregularity	频谱不规则度	否	是	2	
Spectral RollOff	频谱滚降	是	是	22	
Spectral Flux	频谱通量	是	是	22	
Spectral Centroid	频谱质心	是	是	22	
Spectral Spread	频谱扩散度	是	是	22	
Spectral Skewness	频谱偏态	是	是	22	
Spectral Kurtosis	频谱峰度	是	是	22	
Spectral Flatness	频谱平滑度	是	是	22	
Spectral Entropy	频谱信息熵	是	是	22	
Spectral Energy	频域能量	是	是	22	

特征筛选的流程如图4所示。首先对客观特征进行预处理,保证量纲的一致;然后将客观特征与5个评价维度的主观结果进行f检验,保留影响最显著的80维客观特征;最后将各个客观特征之间进行相关性筛查,剔除相关性超过0.8的特征,保证筛选后客观特征之间的独立性。最终不同评价维度保留了约30维的客观特征,用于后续的回归分析。客观特征的表示方式为“特征名称_数字m/std”,其中数字表示频段数,如果没有数字表示全频带结果。

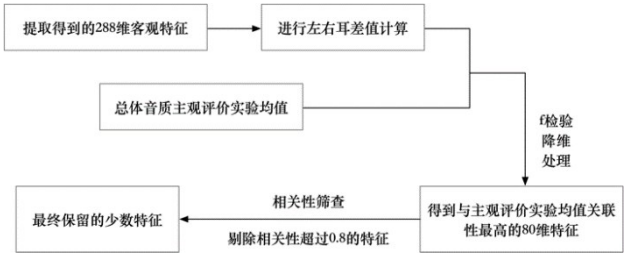

图4 客观特征筛选及降维流程

3.2 基于PLSR的回归模型

本文将对总体音质分别以五个评价维度的主观结果为因变量,筛选的客观特征为自变量,进行PLSR分析。模型的性能用R^2和Q^2进行描述,R^2表示拟合精度,Q^2表示模型的预测能力。PLSR使用VIP(Variable Importance in Prediction)分值来评估自变量的重要性,通常VIP分值大于1的为重要变量。本文数据基于Simca软件进行计算,载荷图由Matlab绘制完成。总体音质不同评价维度的回归模型结果如表5所示。

表 5 总体音质维度的R^2、Q^2及主成分方差贡献率

因变量名称	R^2	Q^2	主成分1方差贡献率	主成分2方差贡献率
总体音质	0.71	0.46	0.54	0.16

图5绘制了总体音质的35个信号的分布及客观特征载荷图,其中两个坐标轴分别对应两个主成分,图中

各点代表该实验信号的主成分得分，向量在坐标轴的投影长度代表了各个客观特征对于主成分的方差贡献率。总体音质PLSR模型中，主成分1解释了54%的方差贡献率，主要与LF_3_m、IACC_4_m等低频段内双耳特征和Spread_2_m、Flux_10_m等时频特征相关；主成分2解释了16%的方差贡献率，主要与IACC_2_std、IACC_2_m等低频段内双耳特征和Skewness_6_std、Spread_2_m等中低频段内时频特征相关。

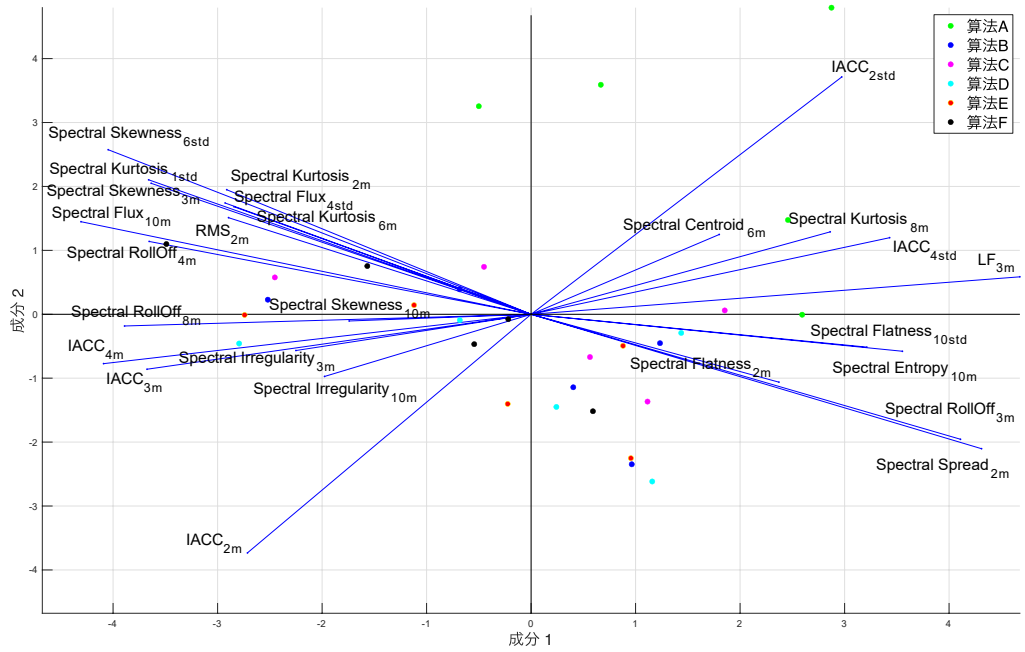

图5 总体音质信号主成分得分及客观特征载荷图

为了进一步探究哪些特征对于总体音质的影响较大，将总体音质PLSR模型中各客观特征的VIP分值进行排序，并将VIP分值大于1的客观特征展示在图6中。从图中可见中低频段内的双耳特征以及中高频段的频谱能量分布及谱结构相关特征对总体音质感知尤为重要。

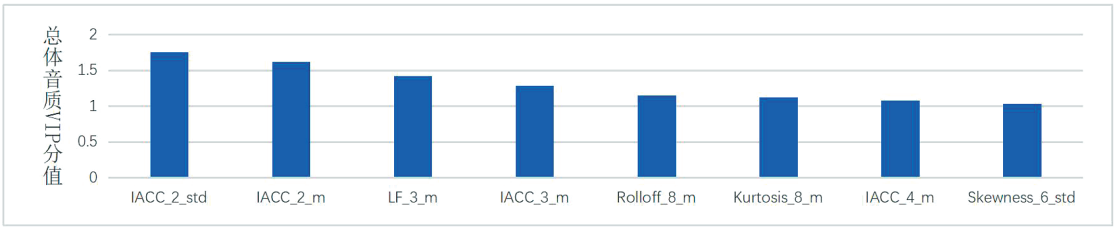

图6 总体音质PLSR模型VIP分值图

4 分析与讨论

4.1 不同双耳渲染算法的对比分析

从主观实验结果可以看出，直接进行线性变换的算法A相比于其他渲染算法在音质组各评价维度上普遍取得了较高的分数，证实了现有的所有双耳渲染处理会影响音质。但值得注意的是，算法B在双耳渲染处理前对头部相关脉冲响应（Head Related Impulse Response，HRIR）进行了音质补偿，该算法除了在总体音质外，在其他维度上与算法A得分较为接近，可见后期对音质进行相应补偿也能极大改善音质。

经过Ambisonic编解码技术的算法D与算法E在各维度上得分差异很大，可见基于Ambisonic编解码原理的具体实现算法会对音质造成很大的影响。虽然两种算法都是基于幅度最小二乘法计算渲染矩阵函数，但是算法E还加入扩散场协方差矩阵约束及分段均衡处理等方法[8]，导致二者在音质维度的听感上产生很大差异。但是这两种算法与经过基于HRTF卷积的算法比较来看，除了算法E在清晰度和力度上存在一定优势，在其他维度上并没有起到优化作用。

从交互效应结果图中,发现信号1在区分不同渲染算法上有着较好的区分度。信号1是包含背景音乐的真人电影片段,频率响应宽且较为平直,没有某个声音元素从头到尾存在,被试在实验过程评分中更加容易倾向于对所有声音元素全局考量,是较为合适的评测素材,这与Olive等人的研究结果相吻合[11]。

4.2 重要声学特征的分析

观察总体音质的客观特征载荷图及VIP分值图,可以发现在40~320Hz低频段内的双耳互相关系数IACC具有较高的贡献率,可见当低频部分双耳信号的差异越大且差异波动情况越大,越能引起被试对于低频的感知,从而影响对总体音质的判断。此外,80~160Hz频段内侧向声能比LF也具有着较高的方差贡献率,且与总体音质主观得分成正相关,Barron等人指出侧向到达的早期反射声可以有效提升视在声源宽度,是与主观听感高度相关且极其重要的声学特征之一[12]。本文的结果表明提升80~160Hz频段内侧向声能比LF也有利于提高主观总体音质的感知。

在时频特征中,2560~5120Hz频段内的频谱滚降Rolloff和频谱峰度Kurtosis也非常重要,频谱滚降Rolloff与主观结果呈负相关,频谱峰度Kurtosis呈正相关。由此可以看出这个频段频谱能量的分布是被试判断总体音质的重要因素。

综合来看,可发现频段范围2560~5120Hz(倍频程8)以及频段范围40~320Hz(倍频程2~4)的时频特征和双耳特征在总体音质PLSR模型中极其重要。从人耳等响曲线可知人耳对于2k~4kHz的声音最为敏感,倍频程8恰好覆盖了该敏感频段,较为显著的影响人耳对音色明亮度的感知;40~320Hz为人声及乐器基频所在的频段区间,且人耳对于低频的感知几乎全部来自这三个倍频程,因此在客观模型中也具有重要作用。

5 结论

本文以三维声双耳渲染算法为研究对象,针对总体音质维度进行了主观和客观评测的研究。本文的主要结论如下:

(1)与线性变换的算法相比,双耳渲染算法确实会对音质的不同维度造成音质损伤,而且不同的算法在各个维度的表现也存在一定差异。算法B经过音质补偿处理,在主观得分上与线性变换算法最为接近,因此对音质进行有效补偿十分必要。

(2)在双耳特征中,低频段的IACC及侧向声能比是影响总体音质较为重要的特征;而对于时频特征而言,2560~5120Hz和40~320Hz这两个频段的时频特征是需要特别关注的特征,例如频谱滚降、频谱峰度等等。

参考文献(References):

[1] Lorho G. Evaluation of spatial enhancement systems for stereo headphone reproduction by preference and attribute rating[C]//Proceeding of the 118th Audio Engineering Society Convention, 2005:6514.

[2] Rumsey F, Zieliński S, Kassier R, et al. On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality[J]. The Journal of the Acoustical Society of America, 2005, 118(2): 968-976.

[3] Lindau A, Erbes V, Lepa S, et al. A spatial audio quality inventory (SAQI)[J]. Acta Acustica United with Acustica, 2014, 100(5): 984-994.

[4] Reardon G, Roginska A, Flanagan P, et al. Evaluation of binaural renderers: a methodology[C]//Proceeding of the 143th Audio Engineering Society Convention,2017: 359.

[5] 范欣欣,祁乐融,杨钊阳,等. 三维声双耳渲染算法的主观评价分析[J]. 复旦学报(自然科学版), 2022, 61(5): 527-535.

[6] 覃龙靖,王鑫,谢凌云. 三维声双耳渲染算法的客观评测和分析[J]. 复旦学报(自然科学版), 2023, 62(1):53-62.

[7] Schörkhuber C, Zaunschirm M, Höldrich R. Binaural rendering of ambisonic signals via magnitude least squares [C]//Proceedings of Fortschritter der Akustik (DAGA), 2018: 339-342.

[8] McCormack L, Politis A. SPARTA & COMPASS: real-time implementations of linear and parametric spatial audio reproduction and processing methods [C]//AES International Conference on Immersive and Interactive Audio, 2019:111.

[9] Lee H, Johnson D, Mironovs M. Virtual hemispherical amplitude panning (VHAP): a method for 3D panning without elevated loudspeakers[C]//Proceeding of the 144th Audio Engineering Society Convention, 2018: 9965.

[10] ITU-R. BS.1284-2:General methods for the subjective assessment of sound quality [S/OL]. [2019-01-21]. https://www.itu.int/rec/R-REC-BS.1284-2-201901-I/en.

[11] Olive, Sean E. A method for training listeners and selecting program material for listening tests[C]//Proceeding of the 97th Audio Engineering Society Convention, 1994: 3893.

[12] Barron M, Marshall A H. Spatial impression due to early lateral reflections in concert halls: the derivation of a physical measure[J]. Journal of Sound and Vibration, 1981, 77(2): 211-232.

编辑:赵志军,龙学锋