个性化 语音通信中的任务负荷预测 N

Robert P. Spang[1], *Karl El Hajal*[2,3], *Sebastian M¨oller*[1,4], *Milos Cernak*[3]

[1]柏林工业大学质量与可用性实验室，德国，[2]´洛桑联邦理工学院（EPFL），瑞士洛桑，[3]罗技欧洲，瑞士洛桑，[4]德国人工智能研究中心（DFKI），德国柏林

arXiv:2303.00630v1 [cs.HC] 1 Mar 2023

摘要

评估远程语音通信的质量是一项复杂的任务，受到说话者、传输信道和听者的影响。例如，传输质量下降可能会增加听者的认知负荷，进而影响对话的整体感知质量。本文提出了一个框架，在模拟对话环境中分离与质量相关的变化，并控制个人偏好等大多数外部影响因素。所进行的统计分析发现，刺激质量与听者的效价及人格（宜人性和开放性）之间存在显著关系；类似地，听力任务期间感知到的任务负荷与听者的人格和挫折不耐受之间也存在显著关系。考虑到听者个体差异后，任务负荷预测机器学习模型的相关系数从 0.48 提高到 0.76。所提出的评估框架和研究结果为个性化音频质量评估铺平了道路，使其能够纳入说话者和听者的个体差异，而不再局限于传统的信道建模。

索引词— 个性化、感知、任务负荷、效价、语音通信

1. 引言

任务负荷是衡量人类表现的一项指标，广义上指个体在执行任务时所遇到的难度水平 [1]。随着近期工作模式转向混合办公，虚拟交互可能会给创意生成带来认知成本 [2]。然而，认知负荷预测仍是一个尚未解决的研究（及工程）问题，尽管近期的音频分析 [3] 和视频质量评估 [4, 5, 6] 研究结果颇具前景。

在研究实时对话的感知（例如质量评估）时，研究人员通常会采用两种方法：事后询问人们对对话的评价或感受（例如，遵循 ITU-T P.800 [7]，用于语音质量评估）。此类主观听音测试可以得到个人评分，但代价是时间分辨率较低，而且只能为整个测试提供一个总体分数。另一种方法是采用基于数千条评分进行预训练的侵入式或非侵入式客观方法。此类系统能够实时评估通信链路，并指出音频失真的原因 [8]，但代价是无法体现个体差异。由于这些模型是基于许多人的平均评分训练的，其输出是平均意见分，或许能代表总体平均水平，却忽略了用户之间的个体差异。

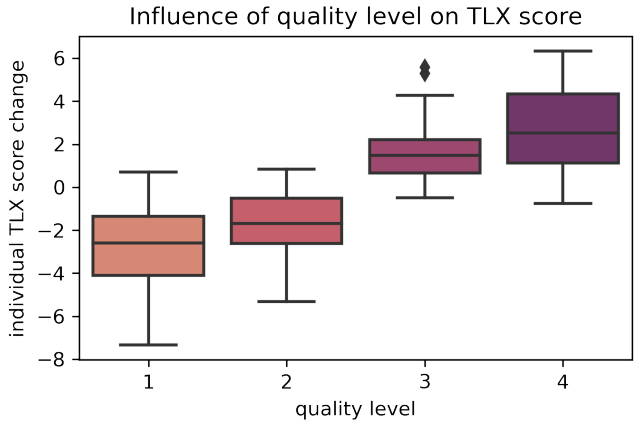

图 1。音频质量与感知任务负荷（任务负荷指数）之间的关系，按用户进行归一化处理。音频质量越低，用户的心理负荷越高。

我们假设，整体质量由说话者（口语内容、说话者熟悉度）、信道（编码失真、传输中断）和听者（人格、情绪）共同决定。例如，传输质量下降会增加听者的认知负荷，进而影响对话的整体质量。图1显示，刺激质量会直接影响听者感知到的任务负荷。这一关系是在 2.1 节所述的用户研究中记录的。随着刺激质量下降（等级 1 最佳，等级 4 最差），听者感知到的心理负荷比刺激质量最佳时更高，

本研究由德国联邦教育与研究部通过 Software Campus 资助项目 01IS17052（KOVOSEQ）提供资金支持

理解刺激内容需要更多心理资源。任务负荷与质量以及其他依赖刺激的用户感知测量指标直接相关。

与其他报告一致，音频质量与任务负荷之间的关系使我们能够利用有关刺激质量的知识，估计用户在例如进行对话时的感知心理负荷 [9, 10, 11, 12, 13]。在上述示例中，这一简单关系解释了相当一部分方差（R^2：0。235）。然而，我们希望探究，关于个体的额外信息是否能够增加可解释的方差。例如，用户的人格以及挫折耐受（不耐受）信息可能有助于更准确地估计某人的感知任务负荷。

因此，在本研究中，我们探索认知任务负荷预测，并认为对其进行个性化处理，即纳入听者的人格、挫折感和情绪水平，能够显著影响其准确性。我们通过严谨的统计分析验证了个性化任务负荷预测，并进一步利用机器学习实验论证了这些发现。侵入式音频质量评估方法在感知层面存在局限 [14]，，因此我们基于非侵入式语音质量评估方法开展实验。

我们假设，了解听者的个性能够提高对其感知的预测能力。本文展示了在模拟视频电话场景中，对主观感知维度进行预测的改进。该模拟使我们能够分离由质量引起的变化，并控制真实对话环境中的大多数外部影响因素，例如个人偏好。我们展示了仅使用（退化的）音频通道作为输入时良好的预测性能，随后通过逐步加入相应听者更多的个性／特质维度，进一步提高预测效果。本文重点关注感知任务负荷，同时讨论其在感知质量、效价或唤醒度等不同维度中的应用。

2. 方法

为了提高对任何感知维度的预测准确率，我们需要理论模型、有效的方法、预测系统以及用于测试的数据。引言部分已经概述了理论模型，下面将介绍我们如何为该实验获取数据、如何通过统计方法检验这一概念，并描述用于预测主观感知任务负荷的机器学习流程。

2.1. 数据收集

我们通过在柏林工业大学开展的一项受控实验室研究收集了数据，共有46名参与者（年龄中位数为30岁，范围为22至57岁），并获得了第四学院电气工程与计算机科学伦理委员会的批准，批准编号：*FR 2022 03*）。参与者接受了 NEO-FFI 问卷 [15]，，这是一项用于评估人格维度的标准化量表，包括宜人性、开放性、尽责性、外向性和神经质。此外，我们采用了挫折不适量表（FDS） [16] 来评估个体接受和应对挫折的能力。这两种量表均用于量化参与者的人格特征。

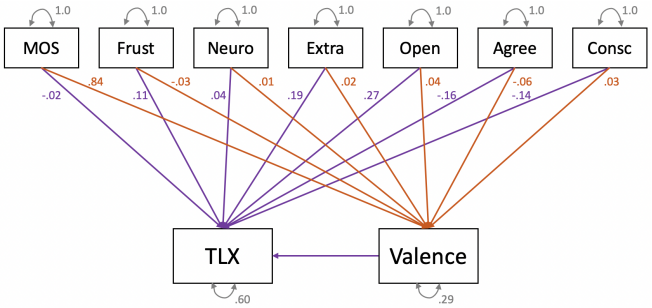

图2。结构方程模型，路径图中的系数为标准化的 β值。

在主要任务中，参与者观看了一系列短视频片段（每段10–12秒）。这些片段呈现了真实的质量失真 [17]; 质量等级为随机选取。参与者观看了三十段视频片段，并在每次试验后填写问卷：使用NASA-TLX问卷 [18], 获得任务负荷评分，并通过情感网格获得效价评分。参与者的时间报酬为每小时15欧元。该流程使每位参与者产生了三十个任务负荷分数。然而，由于一人未填写人格问卷，且存在若干缺失值（按完全随机缺失处理），我们从45名参与者中获得了1,279份完整样本。

2.2. 统计分析

数据集包含TLX分数，范围为 $[0, 20]$, ；效价评分，范围为 $[0, 1]$, ；音频MOS分数（使用POLQA [19]获得），范围为 $[0, 4.79]$；挫折不耐受分数，范围为 $[1, 5]$, ；以及五项人格分数，范围为 $[1, 5]$。所有数值均为连续变量。路径分析是一种不含测量模型的结构方程模型，用于描述一组变量之间的定向依赖关系，可帮助我们考虑效价与任务负荷之间可能存在的中介效应。

我们进行了路径分析，以了解刺激质量（音频MOS）、感知任务负荷（TLX分数）和效价，与参与者个体特征（挫折不耐受和所有人格维度）之间的关系。分析使用 lavaan [20] 在 Jamovi/R [21,22, 23], 中进行，采用稳健最大似然估计量，见图 2。

2.3. 基于机器学习的分析

机器学习实验评估了向模型提供每位个体的信息，是否有助于预测参与者对特定刺激所感知到的任务负荷。因此，我们训练模型，以原始音频信号和听者信息为输入来预测 TLX评分。后者包括挫折不耐受、人格测量结果以及参与者的效价评分。我们尝试了这些维度的所有组合，以确定哪种组合在预测 TLX评分时最具解释力。

用于执行这些实验的模型如图 3 所示。首先，使用 [24],中的 Hybrid BYOL-S/CvT 模型从原始音频中提取特征，该模型已被证明适用于多种语音任务。我们提取话语级嵌入，即将整段原始音频转换为一个维度为 2048 的一维特征向量。Hybrid BYOL-S 特征和选定的个体差异维度随后分别输入多层感知机（MLP），每个感知机都包含一个维度为 512 的隐藏层和一个大小为 128 的输出层。ReLU 层用于实现非线性变换。将两个输出拼接后，输入最终的 MLP；该网络包含一个维度为 512 的隐藏层，并输出预测的任务负荷指数（ TLX ）值。

这些实验对数据进行了归一化。模型使用 ADAM [25]优化器进行训练，批量大小为 8，学习率为 $\alpha = 1$ e-5，损失函数为平均绝对误差（MAE）。我们计算皮尔逊相关系数（PCC）和 MAE 指标来评估模型性能。为应对潜在的过拟合，我们在评估时采用五折交叉验证，并计算所有折的 PCC 和 MAE 均值。

3. 结果

结构模型估计了两个变量：作为中介变量的效价和作为内生变量的TLX。对于效价，我们将挫折不耐受、全部五个人格维度以及音频MOS设定为外生变量。为了估计TLX，我们采用了与效价相同的外生变量，并另外将效价作为一个输入变量。我们还指定了外生变量之间的若干协方差，以优化整体模型拟合度。完整的协方差和相关性表请参见第5节。

最终结构模型的拟合良好， χ^2= 85.049；自由度 = 11；p < 0.001；SRMR = 0.054；比较拟合指数 = 0.977；近似误差均方根 = 0.073。鉴于我们的数据集，这些拟合指标证明了所建模结构总体上有效。因此，该模型是用于检验我们假设的有效工具。效价和TLX的解释方差分别为 $R^2 = 0.71$ 和 $R^2 = 0.404$ 。表1列出了最终的参数估计。

对于效价，七个外生变量中有三个对模型产生显著影响：每个片段的 MOS评分具有

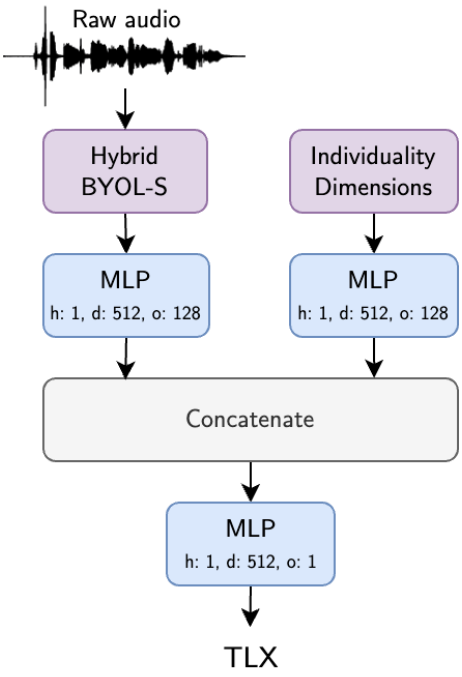

图3。机器学习模型架构。h：隐藏层数量，d：每个隐藏层的维度，o：每个输出层的大小。个体差异维度：五个人格维度+挫折不耐受分数。

最强的影响， β 为 0.839。对于感知到的任务负荷，除两个变量外，其余变量均对最终的回归模型有显著贡献。人格维度神经质仅倾向于影响任务负荷指数，而音频质量对任务负荷指数没有显著贡献。中介变量效价的 β 载荷最大，为 -0.551。

基于机器学习的分析结果如表2所示。将个体差异信息作为输入提供给模型，有助于其比仅音频实验显著更好地预测感知到的任务负荷指数。当仅将一组个体差异维度输入模型时，可以看出，FDS值和人格维度在改善模型结果方面的效果相当。同时，加入效价维度也显示出显著改善。当使用两个个体差异维度作为输入时，结合FDS和人格分数可使模型进一步提升。最后，结合全部三个维度可使模型达到总体最佳性能。

4. 讨论

原始音频质量在很大程度上决定了用户感知到的效价。只有宜人性和开放性这两个人格维度会进一步显著影响效价评分。然而，将这两个人格维度的 β-载荷与 MOS评分进行比较后发现，人格对效价的影响相当有限。

表 1。结构模型的参数估计。缩写释义：val：效价，mos：音频片段的 MOS 分数，agree：宜人性，open：开放性，frust：挫折不耐受得分，consc：尽责性，extra：外向性，neuro：神经质，TLX：任务负荷指数。

			Est	SE	p	β
val	∼	mos	0.156	0.003	$<$ **0.001**	0.839
val	∼	同意	-0.064	0.026	**0.012**	-0.058
val	∼	open	0.043	0.018	**0.014**	0.044
val	∼	挫折感	-0.022	0.033	0.515	-0.033
val	∼	尽责性	0.029	0.022	0.185	0.026
val	∼	额外内容	0.023	0.018	0.204	0.021
val	∼	神经质	0.008	0.015	0.570	0.011
TLX	∼	val	-8.334	0.673	$<$ **0.001**	-0.551
TLX	∼	open	3.962	0.371	$<$ **0.001**	0.269
TLX	∼	额外	3.120	0.378	$<$ **0.001**	0.189
TLX	∼	同意	-2.629	0.420	$<$ **0.001**	-0.156
TLX	∼	意识	-2.427	0.475	$<$ **0.001**	-0.144
TLX	∼	挫折感	1.104	0.310	$<$ **0.001**	0.110
TLX	∼	神经质	0.489	0.278	0.079	0.042
TLX	∼	mos	-0.056	0.124	0.652	-0.020

表 2。结果展示了模型使用不同个体差异维度作为输入来预测任务负荷指数的能力。所用指标为所有折上的平均皮尔逊相关系数（↑）和平均绝对误差（↓）。

输入	PCC	MAE
音频	0.48	0.73
音频 + 效价	0.58	0.66
音频 + 人格	0.68	**0.59**
音频 + FDS	**0.69**	**0.59**
音频 + FDS + 效价	0.69	0.59
音频 + 人格 +效价	0.70	0.57
音频 + FDS + 人格	**0.74**	**0.53**
音频 + FDS + 人格 +效价	**0.76**	**0.51**

另外三个人格维度和挫折不适评分，都无法对效价维度提供具有统计学意义的解释。除其他影响微乎其微的因素外，效价主要由内容质量决定。相比之下，失真连接会降低听者感知到的积极程度，而良好的音频质量则会显著提升参与者感知到的积极程度。

关于任务负荷指数（TLX），人们感知到的效价越高，他们报告的心理负荷就越低。相反，较低的效价评分与更高的心理负荷相关。尽管在不包含效价的模型中，MOS评分是最强的因素，但它对任务负荷估计的贡献并不显著作为中介变量。不过，其大部分解释力似乎都被效价掩盖了。对统计分析进行总结时，关于人格和挫折不耐受的信息有助于解释听者反应中的很大一部分方差。加入关于某人特质的信息，能够更清晰地理解其反应以及他们在听力任务期间的感受。在建模阶段，我们还测试了一个将任务负荷指数（TLX）作为效价中介变量的变体。然而，该模型的拟合指标较差，因此我们最终采用了上述模型。

与我们的统计分析类似，机器学习模型方法的分析表明，与仅使用音频文件预测感知任务负荷相比，加入人格和挫折不耐受维度可带来显著提升。模型输出的相关性增强，同时平均误差也有所降低。加入参与者对各片段的评分信息后，我们便能够生成个性化预测。在实际应用中，这些个性化信号分析可以进行有针对性的估计；而如果输入质量相同，面向所有人的输出则完全一致。

4.1. 结论

我们提出了一个框架，用以展示个体差异指标——此处指人格和挫折不耐受——如何显著改善感知维度的信号分析。我们开展了统计分析和基于机器学习模型的分析，发现了若干显著影响：一方面，感知任务负荷预测的相关系数有所提高（从 0.48 升至 0.74）；另一方面，误差有所降低（从 0.73 分降至 0.53 分）。尽管该数据集及所呈现的分析是展示受试者信息重要性的良好案例研究，但本文首先旨在推动个性化感知估计。

我们的分析适用于探讨信号处理和感知预测中的个体差异。不过，还需要开展更多研究来证实我们的主张，并确定影响最显著的领域。后续研究可以采用众包研究，收集比这项初步探索性研究多一个数量级的答复。由于当前数据集规模较小，过拟合可能是本研究的一项局限。不过，我们将继续探究上述关系，并随着时间推移进一步证实研究结果。

5. 数据可用性与致谢

所述数据集、统计分析文件和机器学习流程均已公开[1]。

我们感谢 Kerstin Pieper 帮助录制视频，并感谢她与 Martin Burghart 和 Leon Schreiber 一起协助开展主观评估研究。

[1]https://osf.io/rxbpv/

6. 参考文献

[1] Molly E. Zimmerman，任务负荷，第 2469–2470 页，Springer NewYork，纽约，纽约州，2011。

[2] Melanie S Brucks 和 Jonathan Levav，“虚拟沟通抑制创意构想的产生”，*Nature*，第 605 卷，第 7908 期，第 108–112 页，2022。

[3] Gasser Elbanna、Alice Biryukov、Neil Scheidwasser-Clow、Lara Orlandic、Pablo Mainar、Mikolaj Kegler、Pierre Beckmann 和 Milos Cernak，“用于分析认知与生理负荷下语音的混合手工设计与可学习音频表征”，载于*Proc.Interspeech 2022*，2022，第 386–390 页。[4] Qi Zheng、Zhengzhong Tu、Yibo Fan、Xiaoyang Zeng 和 Alan C. Bovik，“利用时间带通统计量对可变帧率视频进行无参考质量评价”，载于*ICASSP 2022*，2022，第 1795–1799 页。[5] Shaoguo Wen、Suiyi Ling、Junle Wang、Ximing Chen、Yanqing Jing 和 Patrick Le Callet，“移动游戏视频的主客观质量评价”，载于*ICASSP2022*，2022，第 1810–1814 页。[6] Ao-Xiang Zhang 和 Yuan-Gen Wang，“纹理信息提升视频质量评价”，载于*ICASSP2022*，2022，第 2050–2054 页。[7] ITU-T 建议书 P.800，“质量的客观和主观评价方法”，1996。[8]Karl El Hajal、Milos Cernak 和 Pablo Mainar，“MOSRA：平均意见分与房间声学联合语音质量评估”，载于*Proc. Interspeech2022*，2022，第 3313–3317 页。[9] Maarten Van Segbroeck、Ruchir Travadi、Colin Vaz、Jangwon Kim、Matthew P Black、Alexandros Potamianos 和 Shrikanth S Narayanan，“利用 i-vector 框架从语音中分类认知负荷”，载于*Proc. Interspeech 2014*，2014，第 751–755 页。[10] Avashna Govender 和 Simon King，“利用双任务范式测量合成语音的认知负荷”，载于*Proc. Interspeech 2018*，2018，第 2843–2847 页。[11] Klaus R Scherer、Didier Grandjean、Tom Johnstone、Gudrun Klasmeyer 和 Thomas B¨anziger，“任务负荷和压力的声学相关特征”，载于第七届国际口语处理会议，2002。

[12]Kerttu Huttunen、Heikki Ker¨anen、Eero V¨ayrynen、Rauno P¨a¨akk¨onen 和 Tuomo Leino，“航空中的认知负荷对语音韵律的影响：来自军事模拟器飞行的证据”，《应用人类工程学》，第 42 卷，第 2 期，第 348–357 页，2011。

[13] Bj¨orn Schuller、Stefan Steidl、Anton Batliner、Julien Epps、Florian Eyben、Fabien Ringeval、Erik Marchi 和 Yue Zhang，“INTERSPEECH 2014计算副语言学挑战赛：认知与身体负荷”，载于《*Interspeech 2014*论文集》，2014，第 427–431 页。[14] Pranay Manocha、Zeyu Jin 和 Adam Finkelstein，“音频相似度作为音频质量的替代指标并不可靠”，载于《*Interspeech 2022*论文集》，2022，第 3553–3557 页。[15]Peter Borkenau 和 Fritz Ostendorf，*NEO-F¨Costa* 和 *McCrae* 的人格五因素量表（*NEO-FFI*）：使用手册，1993。[16] Neil Harrington，“挫折不适量表：开发及心理测量学特性”，《临床心理学与心理治疗：理论与实践国际期刊》，第 12 卷，第 5 期，第 374–387 页，2005。[17]Robert P. Spang、Jan-Niklas Voigt-Antons 和 Se-bastian M¨oller，“故事时间数据集：用于质量感知研究的模拟视频电话片段”，载于《*2022*年第*14*届多媒体体验质量国际会议（*QoMEX*）》，2022，第 1–6 页。[18] Sandra G Hart 和 Lowell E Staveland，“NASA任务负荷指数（任务负荷指数）的开发：实证研究与理论研究的成果”，载于《心理学进展》，第 52 卷，第 139–183 页。爱思唯尔，1988。[19] John G Beerends、Christian Schmidmer、Jens Berger、Matthias Obermann、Raphael Ullmann、Joachim Pomy 和 Michael Keyhl，“感知客观听觉质量评估（POLQA）：用于端到端语音质量测量的第三代 ITU-T 标准，第一部分——时间对齐”，《音频工程学会期刊》，第 61 卷，第 6 期，第 366–384 页，2013。[20] Yves Rosseel，“lavaan：用于结构方程建模的 R 软件包”，《统计软件期刊》，第 48 卷，第 1–36 页，2012。[21] Jamovi，“jamovi项目，v.2.3.18”，2020。[22] R核心团队等，“R：一种用于统计计算的语言和环境”，2013。[23] M. Gallucci 和 S. Jentschke，“Semlj：jamovi 结构方程建模分析。 [jamovi模块],”，2021。[24]Gasser Elbanna、Neil Scheidwasser-Clow、Mikolaj Kegler、Pierre Beckmann、Karl El Hajal 和 Milos Cernak，“BYOL-S：通过自举法学习自监督语音表征”，*arXiv*预印本 *arXiv:2206.12038*，2022。[25]Diederik P Kingma 和 Jimmy Ba，“Adam：一种随机优化方法”，*arXiv*预印本 *arXiv:1412.6980*，2014。