表现性钢琴演奏中意图情绪的解码与可视化

ShreyanChowdhury[1] **Gerhard Widmer**[1,2][1]计算感知研究所，林茨约翰·开普勒大学，奥地利[2]LIT人工智能实验室，林茨理工学院，奥地利

firstname.lastname@jku.at

arXiv:2303.01875v1 [cs.SD] 3 Mar 2023

摘要

专业音乐家能够塑造音乐作品，以传达他们意图交流的特定情绪。本文将一种基于中层特征的音乐情绪模型置于演奏者至听众的交流场景中，并通过一个小型可视化演示实时音乐情绪解码。我们还借鉴感知速度和感知力度的对应指标，扩展了现有的中层特征集合。

1. 引言

音乐情感识别旨在利用计算机系统识别音乐中的情绪内容，通常关注感知情绪，即人类听众在听一首歌曲时可能识别出的情绪（这可能不同于作曲者试图表达的情绪，也不同于听众对此作出的情绪反应）。然而，在意图情绪（作曲者或表演者旨在传达的情绪）已知的情况下（例如 [1] 的实验），我们可以从不同角度看待情绪识别模型。模型预测结果不必与听众对感知情绪的评价进行比较，也可以直接与意图情绪进行比较。

本文关注的是在类似情境中使用基于中层特征的音乐情感识别模型——实时解码音乐情绪，并与表演者的预期情绪相比较。（我们在此前的研究 [2] 中已表明，中层感知特征（例如节奏复杂度或感知到的大调／小调和声特征）能够有效地对音乐情绪进行建模。）此外，除了 [3] 和 [2], 中原有的七种中层特征外，我们还研究了另外两种中层特征的近似指标——感知速度和感知力度。

在第 2 节中，我们介绍原有的七个中层特征以及新增的两个特征。在第 3 节中，我们比较了采用原有（7）个中层特征集与扩充后的（9）个中层特征集进行情绪建模的效果，并发现新增的两个特征显著提升了情绪建模的效果。最后，在第 4 节中，我们展示了新的（9）个中层特征模型如何实时有效地解读音乐家（多次荣获格莱美奖的 Jacob Collier）在一场他与全世界分享于 YouTube 上的即兴“情绪传达实验”中明确试图传达的特定情绪 [4]。

Chowdhury 和 G. Widmer。依据知识共享署名 4.0 国际许可协议（CC BY 4.0）授权。**署名：** S. Chowdhury 和 G. Widmer, 《表现性钢琴演奏中意图情绪的解码与可视化》，载于第 *23* 届国际音乐信息检索学会会议晚间特别演示环节扩展摘要集，印度班加罗尔，2022年。

2. 特征

在以往的研究中， [2], 我们已经表明，与基于传统低层特征的模型，或直接从频谱图以端到端方式预测情绪的模型相比，基于中层感知特征训练情绪模型能够提高模型的准确率和稳健性。本文采用基于中层特征的方法，同时还考察了另外两项以往中层特征集未包含的特征——感知速度和感知力度。

2.1 中层特征

中层特征是大多数听众无需音乐理论知识便能直观识别的音乐特质。在 [5] 和 [2], 中，我们使用了基于来自 [3] 的人工标注数据训练的中层特征预测器，使采用这些预测器的情绪识别模型符合人类对中层特质的感知方式；我们还表明，基于这些特征的情绪识别不仅具有良好的识别准确率，而且能够建立音乐上可解释的模型。这些研究所使用的七种中层特征包括：旋律性、发音法、节奏复杂度、节奏稳定性、不协和、调性稳定性以及小调性（或调式）。

2.2 感知速度与动态

虽然使用 [3], 中定义的原有七种中层特征能够较好地对音乐情绪进行建模，但该特征集显然缺少两项对情感表达至关重要的特征：感知速度和力度变化 [6, 7]。先前研究 [7] 表明，起音速率和演奏力度变化等音乐线索在音乐情绪表达中发挥着重要作用。我们的假设是，在中层特征空间中加入与感知速度和力度变化类似的特征，应当

显著改善情绪建模。由于缺乏适当的人工标注数据集，我们不依赖实证的听众感知数据，而是基于音乐直觉，以更直接的方式对这些特征进行建模。

2.2.1 感知速度

感知速度表示音乐的整体快慢，不涉及速度等更深入的分析，并且音乐家和非音乐家都很容易理解这一概念 [8]。速度通常根据最显著拍级（拍点）的出现频率来计算，而感知速度也会受到较低或较高拍级的影响。我们参照 [9, 10] 的观察结果，使用起音密度来近似衡量感知速度。起音密度（或事件密度）指单位时间内音乐起音的数量（即音符或其他声音事件的起始时刻）。

实际应用中已有多种基于信号处理的起音密度检测器（例如在 madmom 库[12]中实现的 SuperFlux 算法 [11] ），但我们针对当前情况采用了不同的方法。由于我们在此只关注钢琴独奏演奏，因此使用自动复调钢琴音符转录算法（madmom 的 RNNPianoNoteProcessor [12]）对音频进行转录并提取钢琴音符，再据此推断起音（将转录音符的起始位置视为起音）。我们观察到，对于不同录音中的钢琴独奏音乐，这种方法能够更准确地估计起音。

2.2.2 感知力度

感知力度表示演奏时的力度等级，而不考虑聆听音量（据推测与演奏者所估计的用力程度有关）[8]。对于我们的钢琴独奏音乐，研究发现，音频信号的均方根（RMS，即 Root-Mean-Squared）振幅可以很好地近似表示演奏力度。我们使用 Librosa 的 RMS 函数 [13]来计算这一特征。对于输入音频信号 x，均方根振幅定义为 $\mathrm{RMS}_k = \sqrt{\mathrm{mean}(w_\tau(x)_k^2)} k = 1 \ldots N$，其中，$w_\tau(\cdot)_k$ 是一个矩形窗函数，用于将输入序列划分为多个帧，并返回长度为 τ 的第 k 帧； N 是帧的总数。

3. 预测情绪

在先前的一篇出版物中 [2], ，我们报告了使用不同特征集（包括原始的（7）个中层特征集）在一个包含巴赫《平均律钢琴曲集》（WTC）第一册录音的数据集上进行音乐情绪预测的性能。在本文中，我们再次使用 WTC 数据集[1]，评估原始中层特征集和增强型中层特征集的预测性能。我们通过将不同的特征集拟合到唤醒度/效价评分上，进行基于回归的分析。

首先，我们使用预训练的中层特征模型（在中层特征数据集上训练）预测这 288 段录音的原始（7）项中层特征。该模型的架构是一种经过修改的感受野正则化 ResNet（见 Koutini 等人 [14]），并根据 Chowdhury 和 Widmer [15] 的描述针对钢琴独奏音频进行了领域自适应。由于在第 4 节中，我们的目标是使用滑动窗口连续预测情绪，因此希望输入窗口大小尽可能小，同时不影响模型性能。通过 Chowdhury [16], 中详述的实验，我们发现 5 秒的输入窗口长度最适合我们的目标。

特征集	调整后的 R^2 唤醒度	效价
（7）中层特征集	0.68	0.63
起音密度和均方根振幅	0.74	0.39
（9）个中层特征集	**0.79**	**0.65**

表1：不同特征集在对巴赫《平均律键盘曲集》第一卷数据集的唤醒度和效价进行建模时的表现。

随后，我们采用第2.2节所述的近似方法，计算每段录音的平均起音密度和平均均方根振幅，从而得到巴赫数据集的（9）个中层特征集。接着，在数据上拟合一个具有九个输入和两个输出（唤醒度和效价）的多元线性回归模型，并报告调整后的 R^2评分。我们将其与仅使用原有（7）个中层特征的情况，以及仅使用两个新增特征的情况进行比较。结果列于表1。

此外，我们使用T统计量（或T值）考察用于预测唤醒度和效价的特征重要性，如图1所示。回归模型中某一特征的T统计量定义为其估计权重按其标准误进行缩放后的值。

我们注意到，对于唤醒度和效价，使用组合的（9）个中层特征集能取得最佳结果。我们的发现指明了未来的研究方向：根据实际的人类评分来学习感知速度和感知力度，从而改进以人类感知为基础的音乐情绪识别算法。

4. 可视化

在一段 YouTube 视频中， [4], Jacob Collier 在钢琴上演奏了《Danny Boy》，并根据演奏过程中向他展示的不同情绪对乐曲进行改编。我们使用基于中层特征的情感模型，对从视频片段（0:00 至 2:45 分钟，内容涉及 Jacob 所称的“基本”或“一级情感”）中提取的音频持续预测情绪（跳跃长度为 = 1 秒；窗口长度为 = 5 秒），并将预测结果映射到 Russell 的环形模型上 [17][2]。

预测的情绪值与原始视频一同呈现。为了实现平滑动画，我们采用指数插值法，计算每隔1秒获得的实际预测值之间的数值。我们的可视化视频可通过 https://youtu.be/ 1Y8wDZfuaKU 观看。

[1]WTC 数据集包含 6 位不同钢琴家演奏的巴赫《平均律键盘曲集》第一册中 48 首作品的录音，以及听众对每段录音的唤醒度和效价的多项评价。

[2]请注意，这只是对我们模型的演示，并非对模型性能的详细分析。

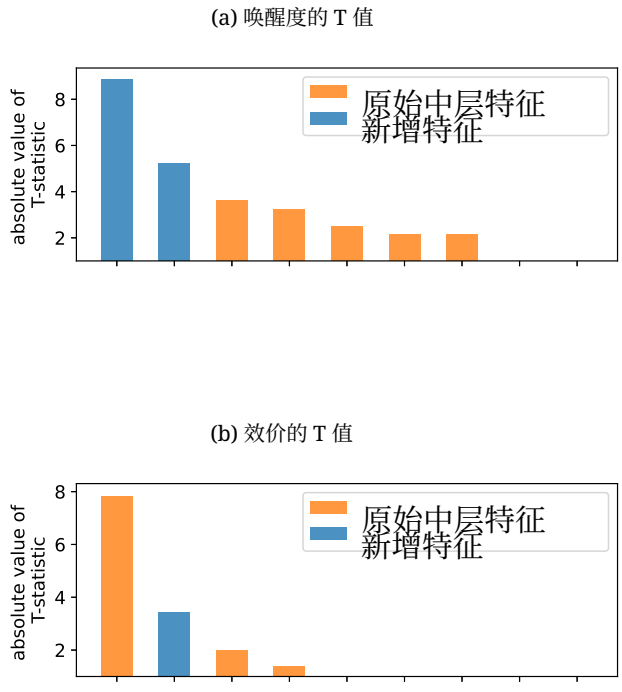


图1：使用 T统计量分析增强型中层特征集的特征重要性值。对于线性回归模型中的每个特征，T统计量定义为其估计权重除以标准误。

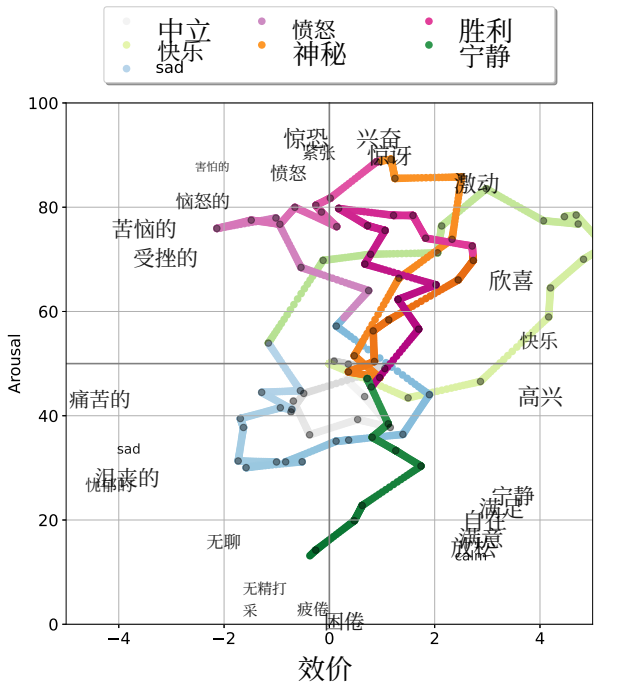


图2：动态情绪预测的完整轨迹。Jacob 的预期情绪（根据原始视频中标注的情绪）以不同颜色表示，时间的推移则以颜色深浅表示——从最浅到最深。

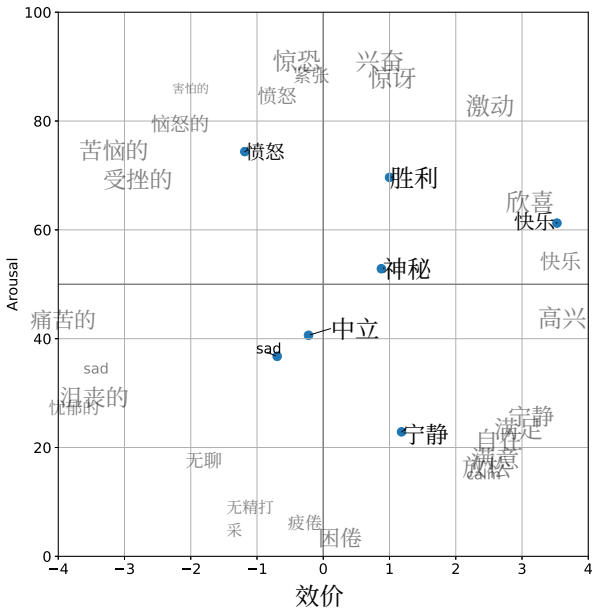


图3：静态情绪预测。预测出的情绪以深色文字标示，罗素环形情绪模型中的情绪词则以浅色文字标示。

5. 结果与讨论

图2展示了预测点的完整轨迹，而图4则展示了视频中的三个画面，并叠加了相应的预测情绪。可以看出，预测的情绪与意图情绪（“Jacob的情绪”）高度吻合。

我们还获得了静态情绪——提取分别对应七种情绪的音频片段，并将其作为独立的音频输入。在这种情况下，我们使用标准输入长度为15秒的中层特征模型；若输入音频不足15秒，则循环播放；若超过15秒，则以5 秒为步长对连续窗口进行预测，并对预测结果取平均。图3展示了这些静态情绪预测结果，其中预测点均标注了各自对应的意图情绪。

这项可视化实验为进一步研究如何利用计算机系统解码意图情绪提供了概念验证。有趣的是，我们观察到，一个仅以少数中层特征（7个原有特征加2个新特征）为输入、在包含288个巴赫《平均律键盘曲集》示例的小型数据集上训练的简单线性回归模型，能够以令人满意的方式，成功预测一组截然不同的演奏所表达的意图情绪。这表明这9个中层特征以及我们的7个中层特征模型具有稳健性，也凸显了这些特征反映编码音乐情绪的出色能力。

总之，本文提供了令人鼓舞的证据，表明中层特征能够有效捕捉音乐演奏所传达的情感，并鼓励进一步探索，以充分了解中层特征在建模更复杂情感方面的潜力。

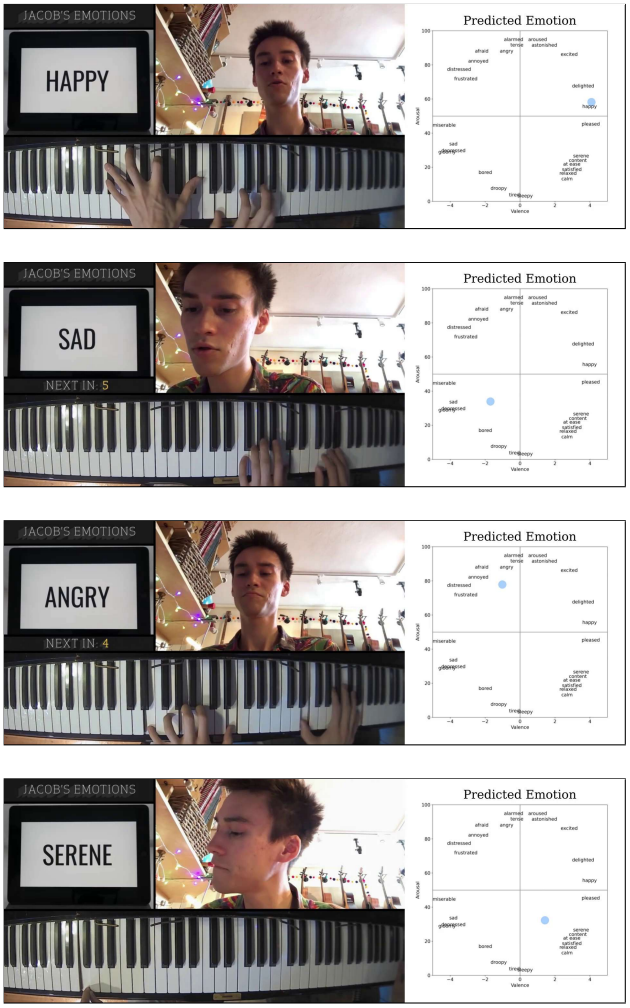


图4：Jacob Collier 表演视频在不同时刻的截图，并叠加相应的预测情绪。

6. 致谢

本研究得到欧洲研究委员会（ERC）通过欧盟地平线2020研究与创新计划资助，资助协议编号为670035（“Con Espressione”）和101019375（“Whither Music?”）。

7. 参考文献

[1] A. Gabrielsson 和 P. N. Juslin，“音乐表演中的情绪表达：表演者意图与听众体验之间的关系”，音乐心理学，第24卷，第1期，第68–91页，1996年。

[2] S. Chowdhury 和 G. Widmer，“表现性钢琴演奏中的感知情绪：进一步的实验性证据，证明中层感知特征的重要性”，载于第*22*届国际音乐信息检索学会年会论文集，*ISMIR 2021*，2021年。

[3] A. Aljanaki 和 M. Soleymani，《一种数据驱动的中层音乐感知特征建模方法——》，载于第*19*届国际音乐信息检索学会会议论文集，*ISMIR 2018*，法国巴黎，2018年，第615–621页。

[4] WIRED.com。（2020）Jacob Collier 用18种逐渐复杂的情绪演奏同一首歌 | WIRED。[在线]。可访问：https://www.youtube.com/watch? v=EWHpdmDHrn8

[5] S. Chowdhury、A. Vall、V. Haunschmid 和 G. Wid- mer，《迈向可解释的音乐情绪识别：通过中层特征实现》，载于第*20*届国际音乐信息检索学会会议论文集，*ISMIR 2019*，2019年。

[6] R. Bresin 和 A. Friberg，《音乐中的情绪呈现：七种音乐变量的范围与特征值》，《*Cortex*》，第47卷，第9期，第1068–1081页，2011年。

[7] T. Eerola、A. Friberg 和 R. Bresin，《音乐中的情绪表达：主要音乐线索的贡献、线性关系与可加性》，《心理学前沿》，第4卷，第487页，2013年。

[8] A. Friberg、E. Schoonderwaldt、A. Hedblad、M. Fabiani 和 A. Elowsson，《将基于听众的感知特征用作音乐信息检索中的中间表征》，《美国声学学会杂志》，第136卷，第4期，第1951–1963页，2014年。[在线]。获取地址：https://doi.org/10.1121/1.4892767

[9] G. Madison 和 J. Paulin，《真实音乐中速度评分与原始及操纵后的拍点速度之间的关系》，美国声学学会期刊 *f* 美国，第 128 卷，第 5 期，第 3032–3040 页，2010 年。

[10] A. Elowsson 和 A.Friberg，《音乐音频中速度感知的建模》，载于声音与音乐计算会议论文集。Citeseer，2013 年，第 735–741 页。

[11] S. Böck 和 G. Widmer，《用于起音检测的最大值滤波器颤音抑制》，载于第十六届数字音频效果国际会议论文集（*DAFx*）。爱尔兰梅努斯（*2013* 年 *9* 月），第 7 卷，2013 年，第 4 页。

[12] S. Böck、F. Korzeniowski、J. Schlüter、F. Krebs 和 G. Widmer，《madmom：一种新型 Python 音频与音乐信号处理库》，载于《第*24*届*ACM*国际多媒体会议论文集》，阿姆斯特丹，荷兰，2016年10月，第1174–1178页。[13] B. McFee、C. Raffel、D. Liang、D. P. Ellis、M. McVicar、E. Battenberg 和 O. Nieto，《librosa：用于 Python 的音频与音乐信号分析》，载于《第*14*届*Python*科学会议论文集》，第8卷，2015年，第18–25页。[14] K. Koutini、H. Eghbal-Zadeh、M. Dorfer 和 G. Widmer，《感受野作为深度卷积神经网络声学场景分类的正则化器》，载于《*2019*年第*27*届欧洲信号处理会议（*EUSIPCO*）》。IEEE，2019年，第1–5页。

[15] S. Chowdhury 和 G. Widmer，《面向钢琴录音中表现力特质的解释：通过声学领域自适应迁移解释性特征》，载于《*ICASSP 2021——2021*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*）》，2021年，第561–565页。[16] S. Chowdhury，《使用可解释且可迁移的感知特征对音乐中的情感表达建模》，博士论文，奥地利林茨约翰内斯·开普勒大学，2022年。[17] J. A. Russell，《情感环形模型》。《人格与社会心理学杂志》，第39卷，第6期，第1161页，1980年。