arXiv:2304.11521v1 [cs.SD] 23 Apr 2023

一种用于同音音乐演奏美学质量评估的秩序—复杂性模型

辛鑫[1,2], 吴洲[1], 王金玉[1],徐多[3], [4*]、荣怡晴[1] 和孙嘉林 [1]

[1]北京电子科技学院，100070，中国北京，many33@126.com[2]北京通用人工智能研究院（BIGAI），100085，中国北京[3]同济大学，200082，中国上海 [4]天津音乐学院，300171，中国天津

摘要

尽管计算美学评估已在许多领域取得一定成果，但音乐演奏方面的研究仍有待探索。目前，主观评估仍是音乐审美研究的终极方法，但这会耗费大量人力和物力。此外，人工智能生成的音乐演奏仍显得机械、单调且缺乏美感。为指导人工智能音乐演奏的生成任务，并提升人类表演者的演奏效果，本文采用伯克霍夫审美度量，提出一种客观的美感测量方法。本文的主要贡献如下：首先，我们提出一种客观的美学评估方法，用于衡量音乐演奏的审美性；其次，我们提出了10项基本音乐特征和4项音乐审美特征。实验表明，我们的方法在演奏评估方面表现良好。

索引词— 计算美学、音乐演奏评估、伯克霍夫度量、音乐特征

1. 引言

计算美学评估 [1]使计算机能够对艺术作品进行定量美学判断，艺术作品通常包括建筑、绘画和音乐。它可用于比较不同艺术作品的审美感受，以进行美学质量评估 [2]。

音乐审美一直是一个难题，从毕达哥拉斯时代起，数学家便尝试提出并解决这一问题；此外，还有将音乐视为情感表达 [3]或语言 [4]的观点。本研究认为，有必要建立一种客观方法来量化音乐的审美感受，而且音乐的审美感受并非完全主观。

如今，与人类表演者相比，人工智能生成的演奏听起来单调，缺乏灵性。目前，机器可能只学到了音乐演奏属性（如力度和）的分布

速度，这或许是机器演奏显得非常机械、笨拙的原因之一。

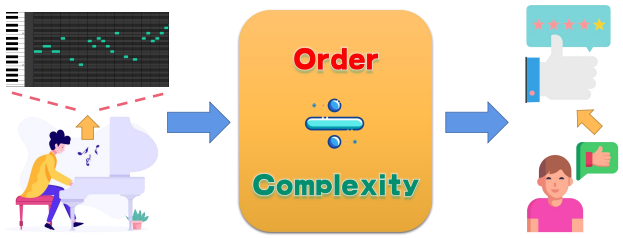

图 1：音乐演奏的质量可通过演奏美学评估模型进行评估。

目前的音乐演奏数据尚未像图像美学评价领域中的AVA [5] 那样标注美学评分。由于缺少带有主观评分标签的数据集，我们采用传统客观美学度量。

本文选择了伯克霍夫方法 [6] 来开展美学质量评估研究。选择该方法的原因在于，它对美具有很强的可解释性。伯克霍夫将对象的美学度量形式化为秩序与复杂性的比值：

$$M = \frac{O}{C} \tag{1}$$

图 1 展示了本文的主要任务。我们的主要贡献如下：

- 我们提出了一种用于评估音乐演奏美学的评价方法。
- 我们提出了10种基本的音乐演奏特征和4种音乐美学特征，以便开展后续的音乐信息检索研究任务。

2. 相关工作

2.1. 音乐演奏生成

在本研究中，我们聚焦于与音乐演奏生成相关的任务。Maezawa等人使用基于LSTM的变分自编码器[7]，通过提供条件乐谱来生成具有表现力的演奏。

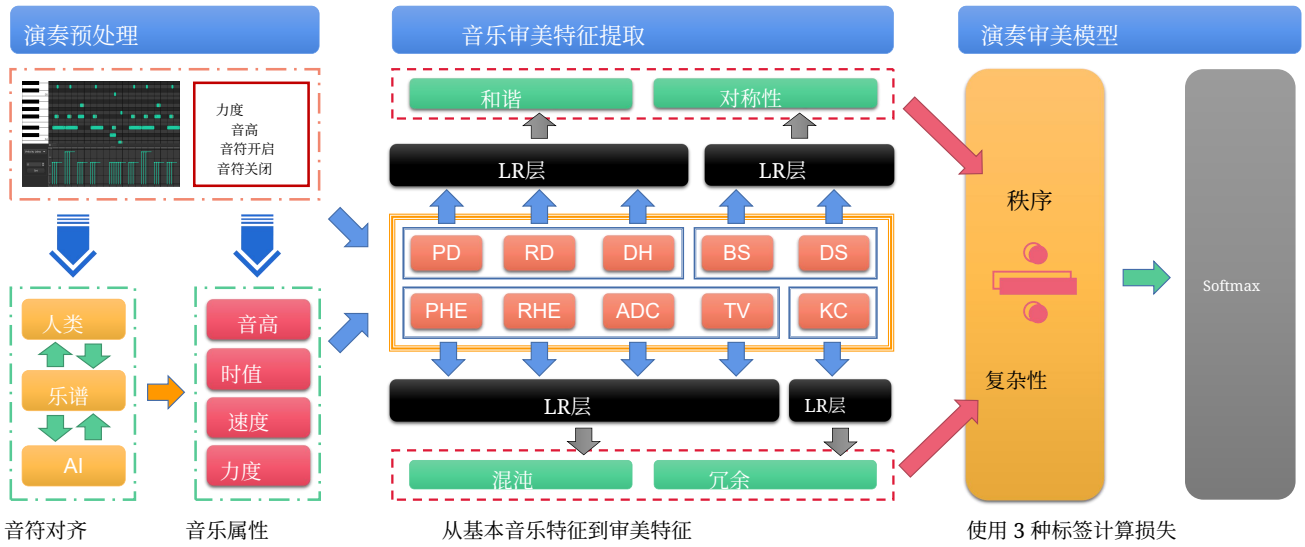

图 2：我们方法的概述。首先，我们对齐乐谱和演奏 MIDI 中的音符。然后，在原始演奏和对齐之后，提取每个音符的音乐属性。接下来，我们使用第 3 节中的公式处理音乐属性，并计算 10 个基本音乐特征。之后，我们将这 10 个基本音乐特征输入相应的 4 个逻辑回归层（图中的 4 个黑色方块），得到 4 个审美特征。最后，我们将这 4 个审美特征输入审美评估模型，通过 Softmax 建立损失，以确定权重。

该乐谱生成的演奏包含诠释上的变化。VirtuosoNet [8] 是一项相对成熟的工作。它考虑了速度、力度和发音方式在演奏中的重要作用，这被认为是一个关键问题。此外，还有DeepJ [9] 和MIDI-VAE[10]等工作。它们也考虑了力度，但仍聚焦于乐谱生成任务，因此此处不再赘述。

2.2. 音乐评估

客观评估指标旨在统一不同模型的结果，以便比较不同模型的演奏表现。这些指标通常是统计量，主要包括四类：与音高、节奏、和声及风格迁移相关的指标。这些类别的指标通常用于 AIGC 模型。

主观评估方法主要包括聆听测试和视觉分析。聆听测试的主要方法包括图灵测试 [11], 并排评分 [12], 等。这些方法用于主观评估音乐是由机器还是人创作的，并评估音乐的各个方面，例如和声竞争、节奏、结构、连贯性和总体评分。

目前，关于音乐审美评估的研究还不多。Audio Oracle [13](AO) 使用信息率 (IR) 作为美学度量。然而，它无法阐明该系统具体具有什么样的审美意图。此外，还有一种利用齐夫定律来衡量音乐之美的方法

定律 [14],，它将词频规则应用于音乐。

3. 我们的审美评估模型

3.1. 形式化

基于伯克霍夫度量，我们提出了四种审美特征：和谐、对称性、混沌和冗余。我们将分子的有序度量与分母的复杂度度量进行线性组合。详细的度量说明将在第 3.2、3.3、3.4 和 3.5 节中介绍。图 2 展示了我们的工作流程。音乐审美度量公式如下：

$$Aesthetic\ Measure = \frac{\omega_1 H + \omega_2 S + \theta_1}{\omega_3 C + \omega_4 R + \theta_2} \tag{2}$$

H 是和谐， S 是对称性， C 是混沌， R 是冗余。 ω 是权重， θ 是常数。

3.2. 和谐

3.2.1. 音高偏差与节奏偏差

使用音高偏差和节奏偏差来判断表演者演奏的音高和节奏准确率。这两个特征分别比较演奏中每个音符的音高属性和节奏属性与原始乐谱中每个音符的音高属性和节奏属性，从而计算演奏与

原始乐谱之间的音高偏差和节奏偏差。表演者必须正确演奏音乐中的音符。偏差的计算公式如下:

$$Deviation = \frac{\sum_{i=1}^{n} |X_i - T_i|}{\sum_{i=1}^{n} |T_i|} \quad (3)$$

其中， i 表示音乐中的第 i 个音符， n 表示对齐音符的总数， X_i 表示演奏中第 i 个音符的属性值， T_i 表示乐谱中第 i 个音符的属性值。

3.2.2. 力度和谐度

根据GTTM理论 [4], 给定一段音乐，可以获得该音乐的节拍结构。节拍结构是音乐片段中的节奏结构，能够很好地描述音乐中的强弱结构。

我们使用余弦相似度衡量力度值与节拍结构值之间的匹配程度。力度和谐度的公式如下:

$$Dynamic\ Harmony = \frac{\sum_{i=1}^{n} D_i \times M_i}{\sqrt{\sum_{i=1}^{n} (D_i)^2 \times \sum_{i=1}^{n} (M_i)^2}} \quad (4)$$

其中， i 表示小节开头的第 i 个音符，n 表示小节开头的对齐音符总数， D 表示力度值向量， M 表示节拍结构值向量。

3.3. 对称性

3.3.1. 节拍偏度 &力度偏度

偏度用于衡量分布的不对称程度，是样本的三阶标准化矩。在演奏中，两个较为重要的因素是力度和节拍。我们计算它们在对称情况下的偏度，并取其绝对值。公式如下:

$$Skewness = E\left[\left(\frac{X - \mu}{\sigma}\right)^3\right] \quad (5)$$

其中， E 表示数学期望， μ 表示均值， σ 表示标准差。

3.4. 混沌

3.4.1. 音高 &节奏直方图熵

我们使用香农熵，直方图熵的定义如下:

$$Histogram\ Entropy = -\sum_{x \in \Omega} p(x) \log(x) \quad (6)$$

它通常用于衡量系统内部状态中混沌和不确定性的程度。

3.4.2. 平均力度变化

乐谱中有强音（f）和弱音（p）符号，用于表示演奏时的力度。平均力度变化的计算方法如下:

$$Average\ Dynamic\ Changes = \frac{\sum_{i=2}^{n} |D_i - D_{i-1}|}{n-1} \quad (7)$$

其中， i 表示音乐中的第 i 个音符， n 表示对齐音符的总数。从第二个音符开始，计算每个音符D_i 的力度与前一个音符 D_{i-1}的力度之差，将这些差值相加后除以 n-1。

3.4.3. 速度变化性

我们设计了速度变化性这一指标，用于衡量速度的混沌程度。速度变化性的定义如下:

$$Tempo\ Variability = \sqrt{\frac{\sum_{i=1}^{n} (t_i - \bar{t})^2}{n-1}} \quad (8)$$

速度变化性用于衡量以每分钟节拍数表示的速度的标准差。其中， t_i 表示第 i 个速度样本值， t 表示速度的平均值。

3.5. 冗余

*3.5.1. Kolmogorov*复杂度

对于字符串 s，字符串 s 的 Kolmogorov复杂度 $K(s)$ 是指在计算机上计算字符串 s 所需的最短程序。本质上，字符串的 Kolmogorov复杂度就是该字符串最终压缩版本的长度。

音乐的无损压缩是衡量音乐冗余度的一般指标。我们用音乐的无损压缩来描述它。它可以形式化为以下公式:

$$Kolmogorov\ Complexity = 1 - \frac{K}{I_m} \quad (9)$$

其中， K 是音乐经无损压缩后的信息量， I_m 是音乐原始信息量。

4. 实现

4.1. 数据集

目前只有少量乐谱与演奏对齐的数据集，我们最终选择 ASAP [15]作为我们的数据集。

如表1所示，我们计划选取 ASAP 中由机器演奏的乐谱作为负样本（不美观的演奏），然后选取 ASAP 中与该乐谱对应的演奏作为正样本（美观的演奏）。接着，为了获得一个中间值，

乐谱	人工智能（VirtuosoNet）	人类
234	234	1060

表1：每种正值、负值和中间值样本的数量。

我们使用 VirtuosoNet 的预训练模型 [8],，以 ASAP 乐谱作为输入，并生成相应数量的演奏作为中间值（一般的演奏）。

通过这种方式，我们获得了三类样本：由乐谱直接渲染的演奏、由人工智能生成的演奏，以及人类演奏。

4.2. 音符对齐

我们需要将乐谱中每个音符的属性作为参照值，以确保演奏中的“即兴发挥”同样有据可依。要计算偏差等特征，需要对齐乐谱与演奏中的音符。

我们采用 Nakamura 等人提出的乐谱到演奏的对齐算法 [16]。该算法能够处理异步演奏的音符，以及演奏中缺失和额外的音符，并返回音符之间的匹配列表。

虽然使用了音符对齐算法，但结果表明我们的对齐仍存在误差。因此，我们采用 VirtuosoNet [8] 中的细化算法来改进音符对齐。

4.3. 基本特征计算

为了计算基本音乐特征，我们必须获取音符的属性。我们使用 music21[1] 和 jSymbolic [17] 获取音符的音高、时值、速度和力度。

我们使用绝对音高进行后续计算。根据 3.2.1 节中的公式，我们计算对齐音符的偏差，以得到 PD。随后计算绝对音高直方图，并使用熵公式计算 PHE。至此，音高相关特征均已计算完成。

对于 RD，我们可以直接使用 music21 获取音符的时值属性，并直接计算节奏偏差。对于 RHE 的计算方法，我们采用 jSymbolic 中的一种节奏编码形式，将节奏编码为 12 种不同的时值，以消除演奏与乐谱之间的差异，从而得到归一化的节奏直方图。

我们将力度（力度标记）平均划分为五个等级：0–24 为 0 级，25–50 为 1 级，51–76 为 2 级，76–101 为 3 级，102–127 为 4 级。由于参考力度值按小节标注，我们取每小节所有音符力度值的平均值，构成一个 n 维向量，其中 n 为小节数。参考力度同样是一个 n 维向量。然后，我们计算参考力度值与演奏中的力度之间的余弦相似度

对演奏进行分析以获得力度和谐度（DH）。同样，我们可以根据前述公式计算 DS 和 ADC 的值。图 3 简要展示了我们的方法。

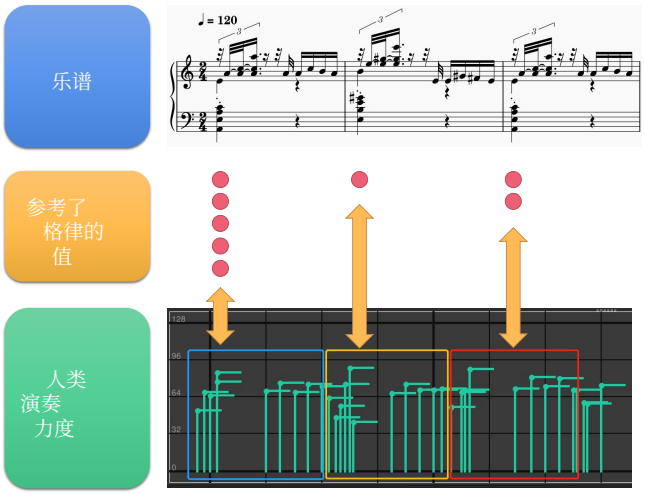

图3：计算每个小节中力度的均值，以匹配参考的节拍值。

对于速度，我们采用 jSymbolic 中的节拍概念来描述音乐的速度，这本质上是描述节奏的另一种方式。我们使用节拍直方图计算 BS。对于速度变化这一特征，我们直接采用 jSymbolic 的实现来获取 TV 的值。

最后，只剩 KC 尚未计算。我们使用 mus-escore3 渲染 MIDI 文件，将其转换为 WAV 音频文件，并以 44100Hz 的频率采样，以确保音乐信息无损。然后，我们参考 Mon-key’sAudio 的[2] 无损压缩方法，将音乐压缩为 ape 格式，从而计算 KC。

4.4. 审美特征与损失优化

我们将基本音乐特征结合到逻辑回归模型中，以得到四个审美特征。我们对基本音乐特征的值进行归一化，然后将所有样本输入逻辑回归模型进行训练。

然后，我们将审美特征的原始值代入 3.1 节中的公式，并计算其 Softmax。我们选择交叉熵损失函数，并使用梯度下降法最小化损失函数。将学习率设为 0.01 后，经过 1000 次迭代，损失函数收敛，模型参数也随之求得。

5. 实验

5.1. 主要结果与消融研究

主要结果分为两部分。一部分介绍特征分布，另一部分使用第 5.1 节中的指标来评估模型的演奏效果。

[1]https://web.mit.edu/music21/ [2] https://www.monkeysaudio.com/

数据集	PD↓	RD↓	DH↑	BS↓	DS↓	PHE↓	RHE↓	ADC↑	TV↑	KC↓	H↑	S↑	C↑	R↑
乐谱	**0.09**	0.32	0.19	0.76	0.65	3.30	1.62	0.87	3.30	0.77	-0.77	2.71	9.40	-0.48
AI	0.12	**0.22**	0.56	0.59	0.41	**3.29**	1.44	5.25	3.20	0.75	0.56	0.91	5.15	-0.46
人类	0.13	0.29	**0.78**	**0.33**	**0.20**	3.63	**1.01**	**10.4**	**7.26**	**0.72**	**2.76**	**3.27**	**11.65**	**1.01**

表 2：以上是三个数据集的实验数据，按十种基本音乐特征和四种审美特征的均值展示。**表中的加粗数据具有更高的审美意义**。显而易见，与乐谱和人工智能生成的演奏相比，人类演奏具有更高的审美意义。

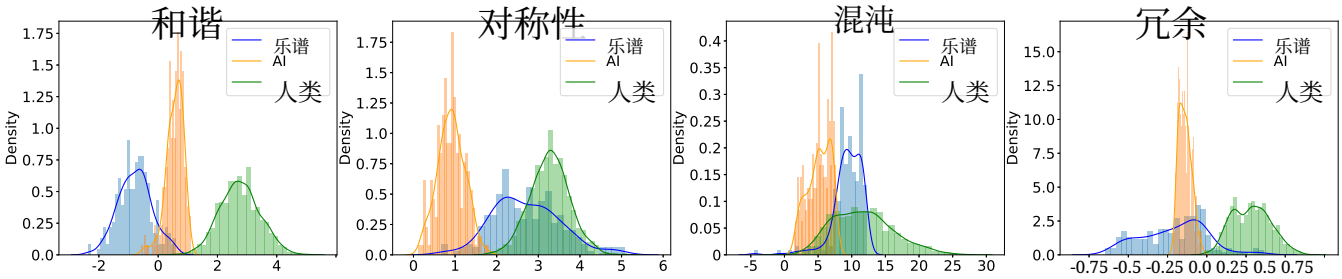

图 4：通过逻辑回归获得的四种审美特征分布。其中，和谐的区分度较高，而对称性、混沌和冗余的区分度较低，表明使用 O/C 模型的必要性。

在表 2 中，我们计算了三个数据集上 10 个基本音乐特征和 4 个审美特征的均值，使“美感”具有可解释性。下面逐一详细分析各项基本特征：

乐谱的 PD 最低，而人类的 PD 最高。常识认为，人类演奏得“不太准确”，但这恰恰说明乐谱演奏非常机械；而具备一定演奏技巧的人类，演奏则可以更加灵活。RD 也是如此。DH 是我们研究的重点。人类较高的 DH 值表明，力度演奏对于符合节拍结构十分重要。人类演奏中的 BS 和 DS 值显著较低，说明与人工智能和乐谱相比，人类演奏的节拍和力度更加对称、更具审美性。人类演奏的 PHE 较高、RHE 较低，这表明人们可能更倾向于加入一些自己的演奏技巧，例如装饰音。较高的 ADC 和 TV 值也表明人类演奏具有多样性，而较低的 KC 值则表明人类演奏的冗余度较低。

图 4 展示了四种审美特征的分布。不同于直接使用深度学习或利用神经网络进行分类，我们可以参考每个步骤生成的数值，从而解释“美”从何而来。图 5 展示了三个数据集上的审美评分分布。如图所示，各分布的交集较小，证明我们的模型表现非常出色。

我们还分别进行消融实验，移除和声、对称性、混沌和冗余，以训练
四种不同的模型。我们将它们与完整模型进行比较。

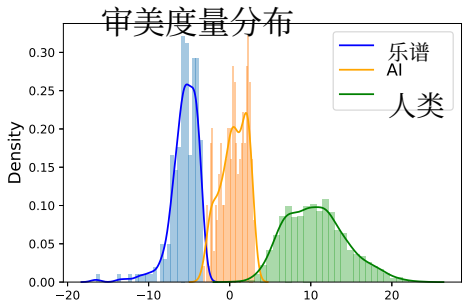

图 5：伯克霍夫审美度量的 3 个分布。

我们选择准确率和Kappa系数作为评估指标。结果如表 3 所示。

5.2. 主观评估

主观评估仍然是检验审美评分有效性的标准。我们从乐谱、人工智能和人类数据集中随机选取了15（5 * 3）段音乐，每段时长约为15秒。参与主观实验的志愿者需要在每组（共5组）中选出他们认为最好的一段。

我们共招募了31名志愿者，收集到155个样本，每位志愿者都聆听了所有曲目。其中，32个样本（20.6%）偏好乐谱表现，39个样本（25.2%）偏好人工智能演奏，84个样本（54.2%）偏好人类演奏。

指标	不含和谐	不含对称性	不含混沌	不含冗余	我们的完整模型
准确率	77.6%	88.7%	82.4%	86.5%	**92.3%**
Kappa系数	0.579	0.791	0.653	0.726	**0.874**

表3：Kappa系数的取值范围为−1至1。Kappa系数越大，模型的演奏性能越好。结果表明，我们的完整模型具有最佳性能。

6. 结论

总之，我们提出了一种基于伯克霍夫秩序-复杂性度量的审美模型，用于评估复调音乐演奏的美感。针对人工智能音乐生成任务，我们提出了10种基本音乐特征和4种审美特征，用于评估生成音乐的美感，这有助于提升人工智能音乐生成任务的质量。不过，我们也必须承认，该模型存在缺陷，例如对称性这一概念的处理，而且模型未考虑“创造力”。此外，音乐音频质量评估的审美研究值得在未来进一步探索。

致谢

我们感谢BI-GAI的AC、审稿人和人工智能音乐团队。本研究得到中国国家自然科学基金（62072014和62106118）、中国教育部人文社会科学研究项目（20YJC760115）以及复杂系统管理与控制国家重点实验室开放基金项目（2022111）的部分资助，

7. 参考文献

[1] Philip Galanter，《计算美学评估：迈向机器创造力》，载于*ACMSIGGRAPH2012 Courses*，第 1–162 页，2012 年。[2] Hancheng Zhu、Leida Li、Jinjian Wu、Weisheng Dong 和 Guangming Shi，《Metaiqa：用于无参考图像质量评估的深度元学习》，载于*Proceedingsof the IEEE/CVF Conference on Computer Vision andPatternRecognition*，2020 年，第 14143–14152 页。[3] Peter Kivy，《style id='8'>声音情感：关于音乐情感的论文，包括《有声贝壳》的完整文本，Temple University Press，1989 年。[4] Fred Lerdahl 和 Ray S Jackendoff，《调性音乐生成理论：再版及新版序言》，MIT press，1996 年。[5] Naila Murray、Luca Marchesotti 和 Florent Perronnin，《Ava：用于美学视觉分析的大规模数据库》，载于*2012 IEEE conference on computer vision andpattern recognition*，IEEE，2012 年，第 2408–2415 页。[6] George David Birkhoff，《美学度量》，载于美学度量，Harvard University Press，2013 年。

[7] Akira Maezawa，《基于条件 VAE 的深度钢琴演奏渲染》，载于第 *19* 届国际音乐信息检索学会会议（*ISMIR*）最新研究及演示论文，2018 年。[8] Dasaem Jeong、Taegyun Kwon、Yoojin Kim、Kyogu Lee 和 Juhan Nam，《VirtuosoNet：用于建模表现力钢琴演奏的分层 RNN 系统》，载于*ISMIR*，2019 年，第 908–915 页。[9] Huanru Henry Mao、Taylor Shin 和 Garrison Cottrell，《DeepJ：特定风格的音乐生成》，载于*2018 IEEE* 第 *12* 届语义计算国际会议（*ICSC*），IEEE，2018 年，第 377–382 页。[10] Gino Brunner、Andres Konrad、Yuyi Wang 和 Roger Wattenhofer，《MIDI-VAE：音乐力度与配器建模及其风格迁移应用》，*arXiv* 预印本 *arXiv:1809.07600*，2018 年。[11] Ga¨etan Hadjeres、Franc¸ois Pachet 和 Frank Nielsen，《DeepBach：可控的巴赫合唱生成模型》，载于国际机器学习会议，PMLR，2017 年，第 1362–1371 页。[12] Albert Haque、Michelle Guo 和 Prateek Verma，《条件端到端音频变换》，*arXiv* 预印本 *arXiv:1804.00047*，2018 年。[13] Shlomo Dubnov、G´erard Assayag 和 Arshia Cont，《Audio Oracle 音乐信息率分析》，载于*2011 IEEE* 第五届语义计算国际会议，IEEE，2011 年，第 567–571 页。[14] Bill Manaris、Charles McCormick 和 Tarsem Purewal，《计算机能够识别美丽的音乐吗》，技术报告，Technical Report CoC/CS，2002 年。[15] Francesco Foscarin、Andrew Mcleod、Philippe Rigaux、Florent Jacquemard 和 Masahiko Sakai，《ASAP：用于钢琴转录的对齐乐谱与演奏数据集》，载于国际音乐信息检索学会会议，2020 年，会议编号 CONF，第 534–541 页。[16]Eita Nakamura、Kazuyoshi Yoshii 和 Haruhiro Katayose，《快速且准确的符号音乐对齐中的演奏错误检测与后处理》，载于*ISMIR*，2017 年，第 347–353 页。[17]Cory McKay、Julie Cumming 和 Ichiro Fujinaga，《JSymbolic 2.2：从符号音乐中提取特征，用于音乐学与音乐信息检索研究》，载于*ISMIR*，2018 年，第 348–354 页。