古琴数据集：一个古琴曲的符号化音乐数据集*

吴雨松[1]，李圣辰[1]

（[1] 北京邮电大学，北京 **100876**）

摘 要：随着机器学习与深度学习等数据驱动的算法在符号化音乐音乐分析与自动作曲上的大量使用，越来越多的数据需要输入这些算法进行训练。因此，数据集的收集与整理变得至关重要。本文介绍了古琴数据集，一个中国古琴曲的符号化数据集。古琴数据集包含 71 首古琴曲，共包含 408 个段落，9860 个小节，以 MusicXML 格文件格式储存。本文首先介绍古琴数据集的组成与其建立方法；之后，介绍了古琴、古琴音乐与其记谱方式；最后，讨论了古琴数据集的可能应用并介绍了一个应用实例。我们希望本数据集可以在现有的符号化音乐数据集基础上贡献更多实用的符号化音乐数据，并助力古琴音乐的分析与研究。

关键词：符号化音乐，数据集，古琴

中图分类号： TBC　　**文献标志码**： A

1 概述

如今，越来越多的研究使用机器学习与深度学习方法分析、建模与生成符号化音乐（有别于音频，使用符号表示的音乐，例如 MIDI）。在此过程中，大量的数据需要输入这些数据驱动的方法以进行训练。然而，有着较大的知识鸿沟与获取数据的相对困难，只有很少的符号化音乐数据被收集与整理。在现有的符号化音乐数据集中，有相当大一部分是西方音乐，缺乏世界上其他文化的音乐尤其是中国音乐。本文中介绍了古琴数据集，一个精心整理与清洗的符号化数据集，其中包含来自中国古琴的琴曲。古琴数据集旨在添加更多的数据于现有的符号化数据集中并同时带来跨文化的音乐数据。古琴数据集可以于如下链接中公开访问：`https://github.com/lukewys/Guqin-Dataset`。

本文首先介绍古琴数据集的组成与其建立方法；之后，介绍了古琴、古琴音乐与其记谱方式；最后，讨论了古琴数据集的可能应用并介绍了一个应用实例。

2 相关工作

尽管数字形式的符号化音乐，如 MIDI 文件，随着计算机与互联网的发展得到了广泛使用，但整理完好的、具有元信息标注的数据集寥寥无几。自 2013 年开始，Colin Raffel 收集并建立了 Lakh MIDI 数据集 [1]，其中包括 176,581 首不重复的 MIDI 文件，其中一部分得到了元信息标注，并与 Million Song Dataset[2] 中的音乐条目对齐。SymbolicMusicMidiData[3] 则收录了 20,006 首古典乐 MIDI 文件。另有一部分符号化音乐数据集为较小型的集中于某一风格的数据集。例如 Bach Chorales Dataset[4] 中收集了超过 400 首包含四个声部的巴赫众赞歌音乐；Nottingham Dataset[5] 则收集了超过 1000 首包含旋律与和弦的英美民歌。可以看出，现有的数据集中的音乐基本为西方音乐，缺乏其他文化的音乐数据，如中国音乐。同时，一些大型数据集缺乏元信息标注，音乐内容风格不集中。

3 古琴数据集

3.1 古琴数据集的组成

古琴数据集由 71 首古琴曲组成，这些古琴曲则由我们收集的琴谱转录而成。我们收集的琴谱出自近期出版的带有简谱形式的古琴曲集 [6–11]，这些曲集中的古琴曲则源自古谱与古琴演奏家的演奏流传。

*收稿日期：2020 年 1 月 4 日

基金项目：2019 年度上海市音乐声学艺术重点实验室委托科研项目（编号 SKLMA-2019-03）研究成果

作者简介：吴雨松（1998–），男，本科，wuyusongwys@gmail.com

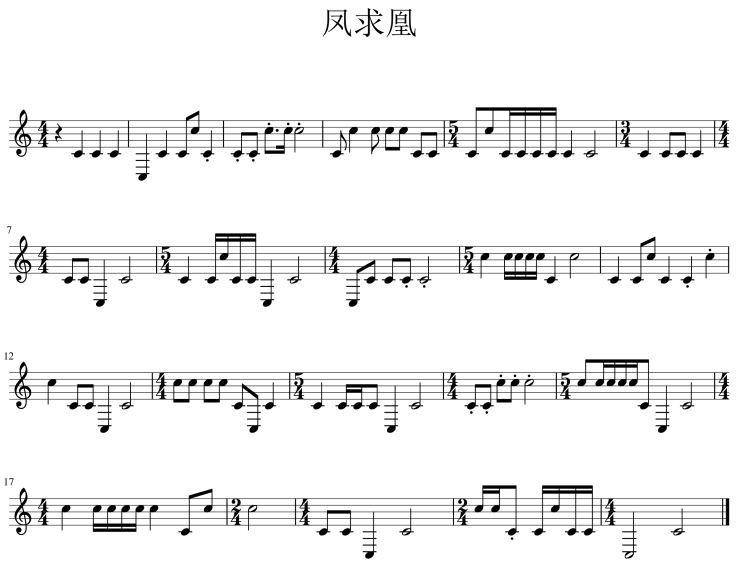


图 1: 琴曲《凤求凰》在古琴数据集中以 MusicXML 格式储存的五线谱。其中，使用跳音记号表示泛音记号。

Figure 1: The Guqin piece Feng Qiu Huang (Male Phoenix Pursui) in MusicXML format. In here, the staccato notation is used to annotate the harmonic notation.

古琴数据集可以于如下链接中公开访问：`https://github.com/lukewys/Guqin-Dataset`。古琴数据集中的数据为 MusicXML 文件格式，以五线谱形式储存。古琴数据集中的内容为琴谱中的简谱旋律与其中的泛音标记[1]。在简谱中，音高以首调唱名的方式书写（即以当前调式的一级音为 C 或 1），而琴谱中音的绝对音高则须由定弦音高共同确定。一首古琴曲通常含有多个段落，在古琴数据集中，每个段落为单独的文件，使用“琴曲名-下划线-段落编号”形式的文件名储存。数据集共包含 408 段、9860 小节的数据，其中的一个谱例如图 1所示。

我们对古琴数据集中的琴谱进行了元信息标注，标注内容包括曲谱名称、定弦、琴谱来源、琴曲来源、演奏者与打谱或记谱者。标注信息与示例详见表 1。完整的元信息标注可以在如下链接获得：`https://github.com/lukewys/Guqin-Dataset/blob/master/Guqin_Dataset_v1/reference.csv`。因为我们获得的琴谱中不是所有琴谱都写明了琴谱来源、琴曲来源、演奏者、打谱或记谱者，因此在元信息中这四项并一定皆有标注。

尽管古琴曲长短不一，在我们收集的古琴谱中，绝大部分短于 200 小节。我们收集的谱子长度的直方图如图 3a所示。同时，我们收集的谱子通常包含分段记号并且分为多个段落。这些段落中绝大多数都相对较短，长度短于 50 小节。分段长度的直方图如图 3b所示。直方图中有少数段落的长度很长，这是由于这些谱子中无段落标记而整首曲子被视为一个段落。

在我们的数据集中，5% 的音符为和弦，在和弦之中 95% 的和弦包含两个音高。值得被注意的是，由于我们避免了一些不易录入的含有较多和弦的谱子古琴曲中总体的和弦比例可能会稍高于本数据集。

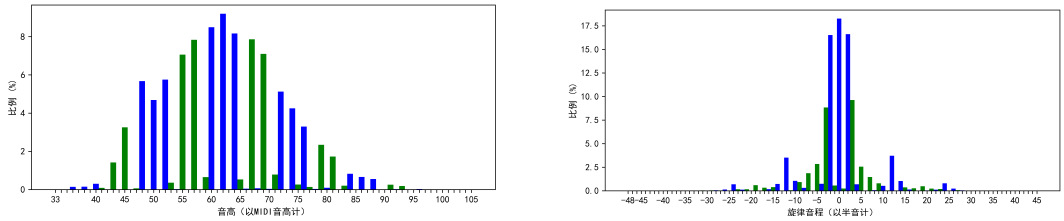
为了了解古琴音乐中的音高特性，我们使用古琴数据集制作了音高与旋律音程的柱状图。对于少数存在的和弦，我们计算旋律音程的方法为计入从一个和弦到下一个音符或和弦的所有可能音高转移。古琴数据集中古琴谱（首调）的音高柱状图如图 2a所示。音高柱状图中音高使用 MIDI 音高表示，最高音与最低音分别为 33 与 105。由图 2a中可以发现，古琴音乐的音域很广，横跨六个八度。使用比例最大的音高为 D，音高分布的包络线呈正态分布形状。同时，从音高分布中可以发现，古琴音乐只有少数四、七级音的使用并鲜有变化音的使用。从音高使用比例的分布中即可以体现出古琴音乐的五声调式特性。

[1]我们发现现有的大部分打谱软件都无法正确的显示 MusicXML 格式曲谱的泛音标记。因此，为了顾及数据观看效果，我们暂使用跳音记号（实心圆点）来表示泛音记号。我们已针对 MuseScore 软件提交了相关的 bug 修复，详情见`https://musescore.org/en/node/294628`。该 bug 已被修复，但截止本文交稿，该修复还未被加入发行版中。一旦该修复加入发行版，我们将对于数据集进行更新。

表 1：元信息标注

Table 1: Metadata Annotation

元信息	样例
文件名称	樵歌 _1.xml
曲谱名称	樵歌
定弦	正调定弦 1=F
琴谱来源	古琴考级经典作品示范
琴曲来源	蕉庵琴谱
演奏者	刘少椿
打谱/记谱者	许健



(a) (b)

图 2：古琴数据集中古琴音乐的音高柱状图（a）与旋律音程柱状图（b）。图中列举了所有出现的音程与音高，每个柱代表一个旋律音程（以半音大小计）或音高（以 MIDI 音高计）。奇数大小的音高或旋律音程使用绿色的柱表示，偶数大小的音高或旋律音程使用蓝色表示。

Figure 2: The pitch histogram (a) and melodic interval histogram (b) of the Guqin Dataset. All the appeared pitch and melodic interval are all shown in the figure, and each bar represents melodic interval (in semitone) or note pitch (in MIDI pitch). The odd number is representing in green, while the even number is representing in blue.

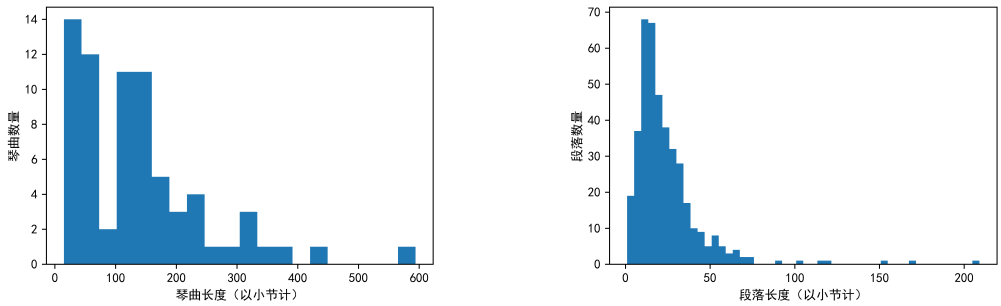
古琴数据集中古琴音乐的旋律音程柱状图如图 2b所示。旋律音程柱状图中的音程为半音大小，出现的最大旋律音程与最小旋律音程分别为 45 个半音与-48 个半音。大跨度旋律音程的使用体现了古琴音乐中跳跃的旋律。使用比例最大的旋律音程为同音音程，并且旋律音程的分布以 2、3、5 个半音为主，也体现出古琴音乐的五声调式特性。值得注意的是，在上、下行八度音程的比例上有一个高峰，这表明了八度音程与八度并行旋律在古琴音乐中的广泛使用。

3.2 古琴数据集的构建

我们使用了一种自定义文本标记格式以将古琴谱中的简谱快速转录至文本，并将文本转换至 MusicXML 文件。使用我们自定义的录入方法转录一页琴谱仅需 3-5 分钟。

值得指出的是，在构建古琴数据集中，我们没有转录琴谱中的所有信息。由于古琴数据集旨在提供进行音乐分析的数据，对于减字谱中的指法以及演奏方法我们并未进行转录。我们收集到的古琴谱中包含有装饰音、连线以及表情记号，由于这些相对稀少，这些记号在转录中都被忽略。在相对罕见的情况下，我们收集到的古琴谱包含多个独立的旋律线（如前述图 6b所示）。对于这些情况，只有一条主要的旋律（在图 6b所示的例子中为上方的旋律）被转录。但是，我们记录了古琴谱中的泛音记号，以便研究者进行古琴作曲中的泛音使用分析。

尽管我们在从文本记录到 MusicXML 文件的转换过程中使用了自动检查措施以修正可能的转录错误，但我们的数据集中会不可避免地包含一些错误。我们同样也欢迎大家指出这些错误。



(a) (b)

图 3: 古琴数据集中的谱子长度直方图 (a) 与段落长度直方图 (b)。在图中，横轴为长度，以小节计；纵轴为该长度区间中谱子或段落的数量。

Figure 3: The score length histogram (a) and phrase length histogram (b). The horizontal axis represents length in bar, and the vertical axis represents number of pieces or phrases in the corresponding length.

3.3 古琴数据集的贡献

古琴数据集提供了一个经过清洗并带有元信息标注的中国音乐数据集。对于计算音乐学分析，古琴数据集提供了一个不同风格、文化的音乐数据集。对于古琴曲分析，古琴数据集提供了一个易于获得、分析与统计的古琴曲形式。本数据集将极大地方便古琴曲的分析，有助于古琴曲的分析研究。

4 古琴与古琴音乐

4.1 中国古琴

中国古琴，有时被叫做“琴”或“七弦琴”，是一种源自中国的古老乐器。中国古琴有着细长的木制琴身与横向固定在琴身上的七根弦，琴弦上方有着指示音位的标记，被称作“徽位”。中国古琴是一种弹拨乐器，通过弹拨琴弦来发出声音。一个古琴的图片如图 4所示。

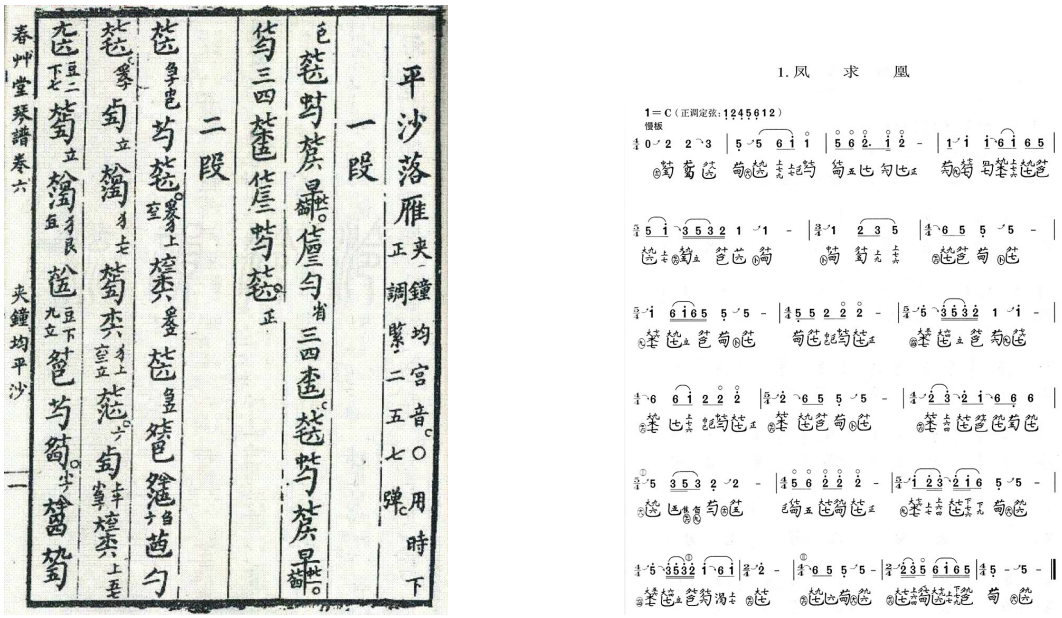
古琴有着悠久的历史，它是中国现在仍在演奏的最古老的乐器。古琴的传说可以追溯至中国的伏羲神农时期。现存最古老的古琴距今已有 2700 年的历史，2016 年于湖北省枣阳市的一座墓中出土 [12]。

图 4: 一副古琴的图片。在图片中，七根弦被横向固定在木制的琴身上，琴弦上方的白点为徽位标记，用来指示弦上音的位置。

Figure 4: A picture of Guqin. In the picture, seven strings are horizontally installed on a wooden body, and the white dot above the strings are called “hui”, for indicating the position of each pitch on the string.

4.2 古琴曲

现存最早的古琴谱《碣石调 • 幽兰》至今已有 1400 余年历史 [13, p. 19]。在千余年的发展中，古琴曲形成了形态各异的风格。相同的古琴曲在流传中也演化出了多种版本，形成不同的流派 [14, p. 157]。值



(a) (b)

图 5: 减字谱 (a) 与带有简谱的减字谱 (b)。在减字谱中，谱面由上至下书写，而在带有简谱的减字谱中，谱面从左至右书写，简谱在上，减字谱在下。在带有简谱的减字谱中，减字谱字符与简谱音符对应。

Figure 5: The abbreviated character notation (a) and abbreviated character notation with numbered notation (b). In (a), the score is written vertically from top to bottom, whereas in (b), the score is written horizontally from left to right. Note that in (b), the numbered notation is written above the abbreviated character notation, and the abbreviated character is rearranged to written from left to right.

得指出的是，古琴曲在中国独立发展，与西方音乐没有联系，也并未受到西方音乐或西方文化的影响 [14, p. 3]。

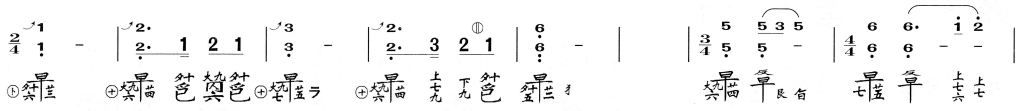
4.3 古琴音乐的记谱方式

尽管在历史上古琴谱曾使用包括文字谱、工尺谱、减字谱等多种记谱方式，其使用时间最久、使用范围最广的古琴谱记谱方式为减字谱。减字谱将汉字的偏旁部首以一定方式结合，以指示指法，弦号，徽位与弹奏方式等信息。有了上述信息，进而也可以根据弹奏的弦与徽位转换出每一个减字谱字符对应的音高。一页减字谱形式的古琴谱如图 5a所示。在减字谱中，文字以从上到下的方向书写，与传统中国纵排排版相同。

然而，传统的减字谱没有精确地记录琴谱的演奏方法。减字谱只是记录了指法或音高信息，而几乎无节奏信息的记录 [15, p. 110]。对此，现代琴谱的记录需要使用精确的记谱方式。从二十世纪中期开始，琴家开始提出不同的古琴谱打谱方式以精确记录古琴音乐 [13, p. 22]。五线谱或简谱开始与传统的减字谱结合，共同表示古琴音乐的弹奏方法、音高与节奏。多数近年出版的琴谱使用简谱与减字谱结合的记谱方式，因此我们转录使用的琴谱皆使用带有简谱的减字谱。一页带有简谱的减字谱形式的古琴谱如图 5b所示。在图 5b中，减字谱书写在简谱下方，并且简谱与减字谱皆从左至右书写。减字谱中的字符与简谱中的音符相对应，以指示弹奏方法。

4.4 古琴音乐的调式

古琴音乐中的调式多采用五度相生律。不同于西方音乐中广泛使用的基于十二平均律的调式，中国民族调式中的五度相生律，以五声调式为主，并有以此衍生出六声调式与七声调式。与西方音乐系统类似，古琴音乐中的调式也有多种调式种类及变化。尽管五声调式与西方调式中的音有类比关系，但是它们的绝对



(a) (b)

图 6: 我们数据集收录的琴谱中八度并行旋律的谱例 (a) 与独立旋律的谱例 (b)。八度并行旋律的谱例来自《释谈章》，独立旋律的谱例取自《醉渔唱晚》。需要指出的是，后者相对前者更少出现。

Figure 6: Example of parallel melody in one octave apart (a) and independent melody line (b) in Guqin music. The latter are rarer than the former.

音高频率有所区别。以宫调为例，“宫、商、角、徵、羽”在首调中可对应“1、2、3、5、6”，并可对应“C、D、E、G、A”[p. 192-196][16]。

然而，古琴音乐中的调式不仅限于五度相生律，还包括纯律的调式 [p. 170][16]。古琴中的调式远比上文介绍的复杂，但其详细的内容不在本文的介绍范围内。我们建议读者查阅《中国乐理》[16] 以了解中国民族调式。

4.5 古琴音乐的旋律特性

古琴有着宽广的音域，其在超过四个八度范围的声音都较好，而古琴曲的旋律也相当跳跃，超过一个八度的旋律音程在古琴曲中相当普遍。依据王震亚 [14, p. 143] 的观点，琴曲曲调中常将相距八度（甚至两个八度）的音作同音使用，经八度音作同音使用后曲调显得变化丰富而有活力。因此，在古琴曲中有大量八度以上的旋律音程。

4.6 古琴音乐中的多音元素

古琴音乐中包含有多声或复调的元素。古琴音乐通常使用同度或八度的双音和弦以强调或突出旋律中的某些部分 [14, p. 150]。因此，古琴音乐中会经常出现八度平行的旋律。少数古琴音乐也会将八度平行旋律的高音旋律进行变化，形成两条独立的旋律线。八度并行旋律的谱例如图 6a所示，独立旋律的谱例如图 6b所示。需要指出的是，在我们收集到的琴谱中，多音的情况只占很少一部分。在我们的数据集中，和弦只占总音符的约 5%，其中绝大部分和弦只包含两个音高。因此，古琴音乐总体上是单声音乐，同时基本无平行旋律出现，从而可以被看做由单音与和弦组成的音乐序列。

4.7 古琴音乐的节奏特性

原始的减字谱形式的琴谱没有包含节奏、节拍标记，而仅模糊的指示了节奏 [15, p. 110]。因此，琴家通常经过打谱“打谱”，对古谱进行整理、译谱、考证、鉴别，加上表演性的再创作以转换成五线谱或简谱形式的精确记谱方式记录 [13, p.]。同时，不同演奏流派的琴家对琴曲有不同的演绎方式，因此，同一份古谱可以有多种演奏方式。不同版本的琴曲通产具有相似的旋律，但在节奏上差别很大。在本数据集中，一首琴曲仅选择了一个版本，并标明了其出处与撰谱人。

除此之外，减字谱形式的琴谱中也没有小节线与拍号。在近年出版的带有简谱的减字谱琴谱中，小节线通常被划定并且少数谱子中标明了拍号（如图 5b所示）大多数我们收集的古琴谱小节中的节拍以四分音符为一拍，但是在以四分音符为一拍的琴谱中间会有少数以八分音符甚至十六分音符为一拍的小节出现。因此，我们收集的古琴谱中拍号不固定且多变（同样可以在图 5b看到）。

在一些古琴作品中，一个通常很长的“散板”小节会出现在段首。在这些“散板”小节中，节拍自由且速度通常缓慢。一个“散板”小节的谱例如图 7所示，其中左端类似草字头的汉字偏旁为“散板”的符号。在图 7的谱例中，也同时出现了八度并行的旋律。

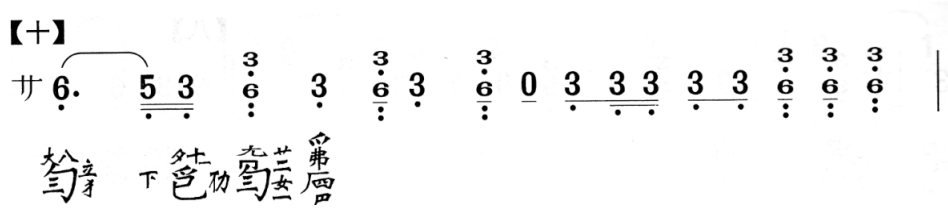


图 7: 琴曲《雉朝飞》中“散板”小节的谱例，左端类似草字头的汉字偏旁为“散板”的符号。在这个谱例中，也同时出现了八度并行的旋律。

Figure 7: The cadenza measure at the beginning of phrase ten in Guqin piece《Zhi Chao Fei》, the Chinese radical at the left of the measure is the symbol in abbreviated notation for “San Ban”. In this piece, parallel melody consist by chords in one octave apart can be seen.

5 潜在的应用及讨论

我们提供了一个易于传播、保存与分析的古琴音乐数据集。在本节中，我们将介绍几种本数据集的可能应用并讨论其中存在的困难。

- **古琴与古琴音乐的研究**

  尽管本数据集中只包含音乐数据而缺乏指法、演奏方法及表情记号等数据，但是我们希望本数据集可以有助于古琴古琴曲的研究，同时促进古琴文化的推广，使世界上不熟悉古琴的研究者及其他文化的研究者参与到古琴与古琴曲的研究当中。古琴音乐是中国传统音乐的重要组成部分，我们希望本数据集可以有利于中国传统音乐的研究。

- **音乐建模与音乐生成的研究**

  有了音乐数据，我们可以对古琴音乐进行音乐建模与生成。然而，由于古琴音乐有着有别于西方音乐的特殊的调式系统、多变的节拍与拍号，再加之古琴音乐表演中有着丰富的表情表演，使用常用的适合对西方音乐进行建模的方式也许不适合。并且，由于古琴谱，尤其是完成打谱的古琴谱的稀少，我们的古琴数据集并不是一个大型的数据集。所有上述这些都因素为使用古琴数据集进行音乐建模与音乐生成的研究带来挑战。

- **表情演奏（expressive performance）研究**

  古琴是一种独奏乐器，同时古琴表演中的表情演奏十分丰富。在此，表情演奏指的是演奏者不严格按照谱子中的节奏进行演奏并加入一些富有表情的变化。古琴演奏中的表情因素在节奏与音高中都有体现。在节奏方面，表情节拍（expressive timing，指表演者在演奏的固定节奏中引入细微的时间偏差）在古琴表演中的运用相当丰富，并且演奏的速度不固定，有时变化很大。在音高方面，在前一个音的演奏与后一个音的转换中，古琴演奏家通常会使用滑音技法。因此，对于古琴演奏的表情演奏研究是相当有意义的。若将符号化音乐数据与古琴演奏的音频对齐，可以进行古琴音乐中的表情节拍的研究。

6 一种基于统计模型的巴赫音乐与古琴音乐风格区分方法

本节介绍一种基于统计模型的音乐风格区分方法以展示古琴数据集的一个应用实例。给定一个音乐样本，本方法可以衡量该样本的风格对于古琴与巴赫风格的相似度。本方法首先提取古琴音乐与巴赫音乐风格的特征。对于给定音乐，同样提取该音乐的特征并向两种风格的特征分别进行相似度衡量。本方法使用旋律音程对符号化数据进行表示，并尝试建立旋律音程直方图与旋律音程马尔科夫链两种方法进行特征的提取。对于提取的特征，本方法使用 Kullback-Leibler 散度进行相似度衡量，并使用五折交叉验证与配对 t 校验以验证方法的有效性。结果显示，在巴赫音乐与古琴音乐上使用本方法提取的风格特征差异明显，同时测试样本对两个风格的相似度度量有显著差异。该结果表明，我们的方法可以区分古琴音乐风格与巴赫音乐风格。

6.1 数据集与数据表示

本方法使用的数据集为巴赫众赞歌数据集（Bach Chorales Dataset）[4] 与本文介绍的古琴数据集中的一部分。

巴赫众赞歌是一个多声部数据集，每个声部都是单声（monophonic）的。在巴赫众赞歌数据集中，每个谱子中至少包含四个人声声部（高音、中音、次中音、低音）。有些谱子中包含如小号、小提琴等伴奏声部，但由于这些声部中音乐数据的稀疏性，我们在处理中忽略了这些声部，从而只保留四个人声声部。在巴赫众赞歌数据集中，我们将四个人声声部中每一个声部都看做单独的旋律处理。

在本方法中音乐数据使用旋律音程的数据表示，将旋律表示为它们的旋律音程序列。在此，旋律音程指旋律中后一个音与前一个音的音高半音差。为了降低数据表示的维度，我们将不为 12 倍数的音程模 12，而对于 12 倍数的音程则取为 12，以保留八度音程。因此，旋律音程共有 13 种表示，大小为 [0-12]。此外，由于古琴数据集中包含少量的和弦，我们计算旋律音程时将计算从一个和弦到下一个音符或和弦的所有可能音高转移。

一个旋律音程表示的实例如图 8所示。该旋律使用 MIDI 音高表示为：{[60], [57], [57, 45], [64], [62], [60], [62], [64, 52]}，其中方括号为单个音或和弦，方括号中的数字代表音高；该旋律使用本方法中的旋律音程表示为：{[3], [0, 12], [7, 7], [2], [2], [2], [2, 10]}，其中方括号为旋律音程中的单个时间点，方括号中的数字代表以半音计的音程大小。

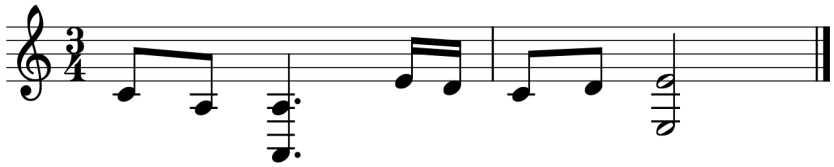


图 8：数据集中的一个旋律片段。在此片段中的旋律音程表示为：{[3], [0, 12], [7, 7], [2], [2], [2], [2, 10]}。其中方括号为旋律音程中的单个时间点，方括号中的数字代表以半音计的音程大小。

Figure 8: One music segment example for data representation. The melodic interval sequence for this music segment example would be: {[3], [0, 12], [7, 7], [2], [2], [2], [2, 10]}. Each square bracket represents an interval set and each number in square bracket represents the interval in the interval set. The value of the number denotes the interval value in semitone.

6.2 特征提取

我们在对风格的音乐数据建立音程直方图与马尔科夫链以提取该风格的特征。

在风格 G 上建立的旋律音程直方图 IH 定义如下：:

$$\mathrm{IH}^G\{x=i\} = \frac{\mathrm{count}(x=i)}{N_I} \tag{1}$$

上式中，N_I 是计入旋律音程的总和，$\mathrm{count}(x=i)$ 表示音程 i 计入的数量，$i = 0, 1, 2 \ldots 12$。

但是，旋律音程直方图只能反映音乐中旋律音程的分布，而不能反映旋律音程的顺序。对此，在音乐序列上训练一个马尔科夫链有助于提取进一步的时序信息。马尔科夫链描述了马尔科夫过程（下一个状态只取决于当前状态的过程）中状态转移的概率。马克科夫链训练完成后将会得到状态转移矩阵 $\mathbf{M} = (p_{ij})$。其中 p_{ij} 为状态 i 转移到状态 j 的概率：

$$p_{ij} = \Pr\left(X_1 = j | X_0 = i\right) \tag{2}$$

通过计算每种音程转移的数量并进行归一化，就可以得出状态转移矩阵。在本方法中，马尔科夫链直接使用数据表示中的旋律音程作为状态。由于旋律音程共有从 0 到 12 共 13 种，状态转移矩阵 $\mathbf{M}$ 的维度也相应的为 13×13。对于一个时间点包含多个旋律音程的情况，我们计入所有可能的旋律音程转移组合。

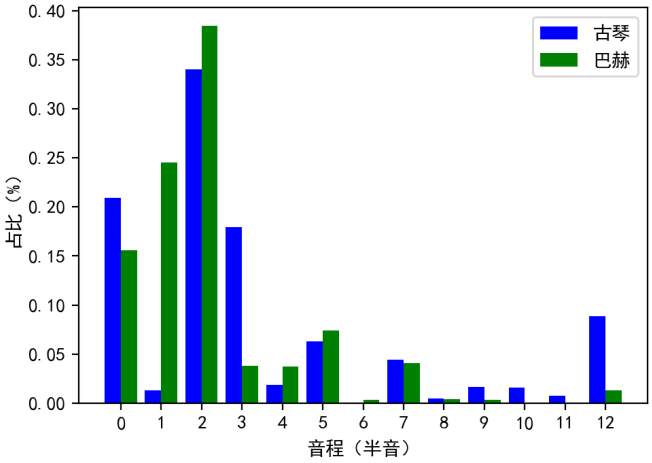


图 9: 中国古琴音乐的旋律音程直方图（蓝）与巴赫音乐的旋律音程直方图（绿）。

Figure 9: The melodic interval histogram of Guqin music (blue) and Bach music (green).

6.3 相似度度量

在建立统计模型并提取特征后，本方法将衡量特征间的相似度以评估风格的相似度。其中，特征矩阵的相似度度量使用 Kullback-Leibler 散度计算。Kullback-Leibler 散度是一种直观的非对称度量方式，用以衡量两个矩阵的相似度。两个矩阵越相似，Kullback-Leibler 散度的计算结果就越小，若两个矩阵完全一致，则 Kullback-Leibler 散度计算结果为 0。

设 $P(x)$ 与 $Q(x)$ 为离散随机变量 x 的两个概率分布，则 P 与 Q 间的 Kullback-Leibler 散度 $D_{\mathrm{KL}}(P\|Q)$ 定义如下：

$$D_{\mathrm{KL}}(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right) \tag{3}$$

对于在某一风格 G 上提取的某一特征矩阵 D^G，与在任意测试样本 S 上提取的同一种特殊矩阵 D^S，两个矩阵的相似度使用 Kullback-Leibler 散度计算如下：

$$D_{\mathrm{KL}}(D^G\|D^S) = \sum_{x\in\mathcal{X}} D^G(x)\log\left(\frac{D^G(x)}{D^S(x)}\right) \tag{4}$$

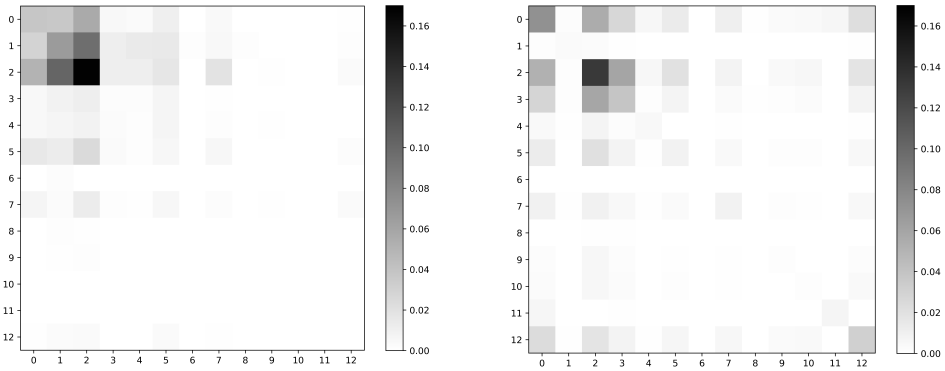
在对两个风格的音乐数据建立旋律音程直方图与马尔科夫链后，我们将得到这两种统计模型提取的特征矩阵：IH^{Bach}、$\mathbf{M}^{Bach}$、IH^{Guqin} 与 $\mathbf{M}^{Guqin}$，其中矩阵的上标代表了该矩阵代表的风格。给定一个新的音乐样本 S 我们同样计算出旋律音程直方图 IH^S 与旋律音程转移矩阵 $\mathbf{M}^S$ 并将对应种类的矩阵计算 Kullback-Leibler 散度，以获得相似度。

在实现中，我们在 Kullback-Leibler 散度计算之前对所有矩阵的元素都加上大小为 $1e-5$ 的微小常量，以避免 Kullback-Leibler 散度在 0 概率上的计算。

6.4 实验结果与分析

在两种风格的数据集上建立的旋律音程直方图如图 9所示。从旋律音程直方图中可以看出，巴赫音乐与古琴音乐最大的区别在于小二度的使用。巴赫音乐中使用了大量的小二度，而古琴音乐中包含更多的小三度。这种使用的差异反映了两种音乐调式的差异。我们对巴赫音乐的旋律音程统计结果与 Knopoff 和 Hutchinson[17] 统计结论相似，后者在巴赫的赋格作品中统计的结果表明大二度与小二度共占旋律音程的约 70%。在古琴的旋律音程统计中，我们发现其在八度音程的比例较大，这也反映了古琴音乐对八度跳跃音程的偏好。

在两种风格的数据集上建立的旋律音程转移矩阵如图 10所示。在图 10中可以看出，巴赫的旋律音程转移矩阵集中在左上角，而古琴的旋律音程转移矩阵则更为分散。这种稀疏性差异进一步体现了调式的差异，同时也表现出巴赫音乐的旋律更连续，古琴音乐旋律跳跃的特点。有趣的是，两个矩阵中的最大值皆为小二度到小二度的转移。



(a) 巴赫　　(b) 古琴

图 10: 数据集中巴赫音乐与古琴音乐的旋律音程转移矩阵。横轴与纵轴的数字代表半音大小的音程，矩阵中元素的灰度代表了其相对占比。

Figure 10: The melodic interval transition matrix of Bach and Guqin music. The numbers on the axis represents melodic interval in semitone, and the gray scales of the matrix represent its proportion.

本方法使用五折交叉验证以验证正确性。对于两个风格的数据集，均将其按小节长度均分为五份。五折交叉验证中，数据将被计算五次，在每次中，将五等分的数据集其中一份作为测试集，其余四份作为训练集。统计模型使用训练集中的数据进行训练，并使用测试集中的数据进行相似度计算。在计算测试集样本对两个风格的相似度后，本方法使用配对 t 校验以验证两个相似度差异的显著性。

配对 t 校验的计算方法如下。给定 n 对实数对 X_i 与 Y_i，先计算其差的平均值 $\bar{d} = Y_i - X_i$；之后，计算实数对差的标准差 s_d；最后，计算 t 检验统计量 $t = \overline{d}/(s_d \times \sqrt{n})$。计算得出的 t 检验统计量服从自由度为 $df = n - 1$ 的 t_{n-1} 分布，因此将 t 检验统计量带入 t_{n-1} 分布，即可得出 p 值。p 值是当原假设为真时所得到的样本观察结果或更极端结果出现的概率，在此，取 $p = 0.01$ 为阈值，低于该阈值则拒绝假设检验。此处，假设检验为两个相似度无显著差异。

在五折交叉验证上进行的假设配对 t 校验的结果如表 2b所示。表中列出了差的平均值（$\bar{d}$）, 自由度（df）, t 检验统计量（t）和单边 p 值（p）。第一列的文字为测试样本的风格，风格后的数字代表了五折交叉验证的次数。

两个表中的 p 值均小于 0.01，这证明计算得出的相对两个风格的相似度在统计意义上有显著差异。该结论说明使用的两种方法皆可以区分巴赫音乐与古琴音乐。同时，使用旋律音程转移矩阵计算的配对 t 检验结果中 p 值普遍小于使用旋律音程直方图计算的配对 t 检验结果，这表明建立马尔科夫链可以更有效的区分巴赫与古琴音乐。

(a) 旋律音程直方图					(b) 旋律音程转移矩阵				
	$\bar{d}$	df	t	p		$\bar{d}$	df	t	p
古琴 1	-1.72	47	-13.52	2.03e-18	古琴 1	-1.84	47	-13.25	4.32e-18
古琴 2	-1.88	48	-20.44	6.27e-26	古琴 2	-2.21	48	-26.47	6.90e-31
古琴 3	-1.81	48	-16.15	1.22e-21	古琴 3	-2.1	48	-17.8	2.22e-23
古琴 4	-1.67	49	-14.43	7.10e-20	古琴 4	-2.07	49	-19.05	6.70e-25
古琴 5	-1.79	56	-16.36	1.37e-23	古琴 5	-2.22	56	-21.19	5.13e-29
巴赫 1	0.66	79	27.99	2.31e-43	巴赫 1	1.98	79	51.39	4.42e-63
巴赫 2	0.64	85	44.52	2.77e-61	巴赫 2	2.06	85	54.67	1.31e-68
巴赫 3	0.66	77	29.65	1.81e-44	巴赫 3	2.03	77	48.79	2.88e-60
巴赫 4	0.68	80	36.74	1.87e-52	巴赫 4	2.06	80	69.22	9.27e-74
巴赫 5	0.64	81	37.24	2.53e-53	巴赫 5	1.95	81	51.99	1.38e-64

表 2: Kullback-Leibler 散度计算值的配对 t 校验结果。其中差的平均值 ($\bar{d}$), 自由度（df）, t 检验统计量（t）和单边 p 值（p）如表所示。第一列的文字为测试样本的风格，风格后的数字代表了五折交叉验证的次数。

Table 2: The paired t-test result for Kullback-Leibler results of melodic interval distribution (a) and transition matrix (b). The mean-difference ($\bar{d}$), degrees-of-freedom (df), statistic (t) and the one-tail p-value (p) are presented. The number followed the genre represents the number of the five-fold cross-validation.

7 结语

本文提出了一个古琴曲的符号化数据集—古琴数据集。该数据集包含 71 首古琴曲，共包含 408 个段落，9860 个小节，以 MusicXML 格式储存古琴音乐，并包含元信息标注。本文介绍了古琴数据集、古琴与古琴音乐，讨论了可能的应用并介绍了一个应用实例。古琴数据集为现有的符号化音乐数据集中添加了来自中国传统音乐的实用音乐数据，并可以便利古琴曲的研究。未来，我们将继续扩充古琴数据集，收录更多的琴曲。

8 致谢

我们衷心地感谢张子谦、张逸嘉、许阳、苗天辰同学在转录谱子中做出的贡献，并感谢所有为本工作提供帮助的朋友们。

参考文献

[1] Lakh MIDI Dataset[M/OL]. May 2019. https://colinraffel.com/projects/lmd/.

[2] Bertin-Mahieux T, Ellis D P, Whitman B, et al. The million song dataset[C]// University of Miami. ISMIR 2011: Proceedings of the 12th International Society for Music Information Retrieval Conference, October 24-28, 2011, Miami, Florida. [S.l.]: University of Miami, 2011: 591–596.

[3] Nottingham Dataset[M/OL]. May 2019. http://abc.sourceforge.net/NMD/.

[4] SymbolicMusicMidiData[M/OL]. May 2019. http://users.cecs.anu.edu.au/~u1018264/data.html.

[5] Bach Chorales Dataset[M/OL]. May 2019. https://github.com/cuthbertLab/music21/tree/master/music21/corpus/bach.

[6] 许光毅. 古琴秘谱遗存第 1 卷 [M]. 北京: 上海三联书店, 2015.

[7] 朴东生，张殿英. 古琴考级曲集 1[M]. 北京: 人民音乐出版社, 2010.

[8] 李祥霆，龚一. 古琴考级曲集 2[M]. 北京: 人民音乐出版社, 2010.

[9] 朴东生，张殿英. 古琴考级曲集 3[M]. 北京: 人民音乐出版社, 2010.

[10] 嗨的国风音乐. 古琴名师琴谱集录 [M]. 北京: 人民邮电出版社, 2018.

[11] 杨青. 古琴考级经典作品示范 [M]. 北京: 人民音乐出版社, 2016.

[12] 中国社会科学网. 我国迄今最早琴瑟组合出土 [M/OL]. May 2016. http://www.cssn.cn/sjs/sjs_hqzs/201605/t20160505_2996265.shtml.

[13] 杨青. 古琴艺术知识 200 问 [M]. 北京: 人民音乐出版社, 2011.

[14] 王震亚. 古琴曲分析 [M]. 北京: 中央音乐学院出版社, 2005.

[15] 卢静云. 古琴 [M]. 北京: 西南师范大学出版社, 2015.

[16] 杜亚雄，秦德祥. 中国乐理 [M]. 北京: 上海音乐学院出版社, 2007.

[17] Knopoff L, Hutchinson W. An index of melodic activity[J]. Interface. dec 1978, 7(4):205–229. DOI: 10.1080/09298217808570260.

[18] Van Gulik R H. The lore of the Chinese lute: an essay in the ideology of the Ch in[M]. [S.l.]: Tuttle Pub, 1969.

[19] 章华英. 古琴 [M]. 北京: 浙江人民出版社, 2005.

Guqin Dataset: A symbolic music dataset of Chinese Guqin collection

Yusong Wu[1], Shengchen Li[1]

([1] *Beijing University of Posts and Telecommunications, Beijing 100876*)

Abstract: With the extensive use of data-driven algorithms on symbolic music analysis and automatic composition, increasing amount of data is needed. In this paper, we introduce Guqin Dataset, a symbolic music dataset of Chinese Guqin pieces. Guqin dataset contains 71 Guqin pieces, 408 phrases, with 9860 measures in total, formatted in MusicXML files. We first presented the composition of Guqin dataset and its construction. Then, Guqin, Guqin music and its notations are introduced. Last, possible applications of Guqin dataset are discussed and a application instance of Guqin dataset is presented. We hope this dataset can bring more useful data to symbolic music dataset and help research on musicology analysis of Guqin music.

Key Words: Symbolic music, Dataset, Guqin