将拍号跟踪模型适配于拉丁美洲音乐

Lucas S. Maia[1] **Martín Rocamora**[2] **Luiz W. P.Biscainho**[1] **Magdalena Fuentes**[3]
PEE/COPPE，里约热内卢联邦大学，巴西[2] FING，共和国大学，乌拉圭[3]
MARL-IDM，纽约大学，美国lucas.maia@smt.ufrj.br

arXiv:2304.07186v1 [cs.SD] 14 Apr 2023

摘要

近年来，随着深度学习方法的引入，节拍和强拍跟踪模型取得了显著进步。然而，尽管有所改进，仍有若干挑战有待解决。尤其是，在音乐信息检索领域，将现有模型适配于代表性不足的音乐传统，通常意味着要收集并标注大量数据，这既不切实际又耗费时间。迁移学习、数据增强和微调技术已在相关任务中取得了相当好的成效，众所周知，这些技术有助于缓解这一瓶颈。此外，在研究这些音乐传统时，模型无需泛化到多种主流音乐流派，而只需在范围更有限、风格更统一的条件下表现良好。本文探讨了将节拍和强拍跟踪模型适配于两种不同拉丁美洲音乐传统的简单而有效的策略，并分析了这些适配在实际应用中的可行性，包括所需的数据和计算资源。与普遍看法相反，我们的研究结果表明，只需花几分钟标注部分数据，并在标准中央处理器设备上训练模型，就能获得良好性能；具体所需资源取决于任务和数据集的复杂程度。

1. 引言

拍号跟踪是指从音频信号中追踪音乐脉动的时间结构，这意味着至少要识别节拍和强拍 [1]。这是音乐信息检索（MIR）中一个长期的研究领域，应用范围从自动 DJ 混音 [2]到音乐学研究 [3]。过去十年间，随着深度学习（DL）技术的引入，拍号跟踪发生了重大转变 [4–7],；这些技术不仅提升了性能，也改变了相关方法的设计范式 [8]。

如今，节拍和强拍跟踪模型主要依赖监督式深度学习 [8], ，因此变得由数据驱动，并需要大量标注数据，才能泛化到不同的歌曲、流派或数据集。

这种对标注数据的依赖给此类模型的广泛应用和普及带来了诸多挑战，尤其是在特定文化音乐传统中 [9–11],，这些传统往往缺乏标注数据，因为生成标注需要具备文化意识的专业知识。因此，现成的通用模型在这些音乐流派中的表现通常不佳，因为这些流派在用于训练的数据集中代表性不足。尽管如此，针对某些拉丁美洲音乐传统的既有研究表明，如果有标注可用，训练统计模型能够取得良好的表现 [12]。

近期的研究开始从不同视角探讨节拍跟踪。与其开发能够在各种音乐流派中都取得同样出色表现的“通用”模型（这需要大量带标签数据），近期的研究转而致力于调整现有模型，使其能在特定目标范围内取得良好表现 [13], ，而目标范围甚至可以小到单首音乐作品 [14, 15]。这一范式与实际应用十分契合，因为用户花少量时间制作几秒钟的标注，以获得良好的表现，是合理可行的。

我们将这一思路用于改进拍号跟踪模型，使其能够在特定音乐流派中表现良好。我们认为，如果某一流派在乐器编制和节拍结构方面具有足够的同质性（许多拉丁美洲音乐传统便是如此），那么只需少量标注数据，就有可能让拍号跟踪模型经过适配后取得显著出色的表现。我们从数据、性能和计算成本三个方面探究这种适配能力。我们聚焦于两种拉丁美洲音乐流派——桑巴和坎东贝，研究如何适配一种最先进的深度学习模型 [16] ，并将其与一个较简单的统计模型 [17]进行比较。我们的贡献如下：1）详细分析在桑巴 和坎东贝中，要达到接近“完整数据集”性能，需要多少标注数据和中央处理器计算时间；分析对象包括从头训练和微调的模型，并将它们与使用西方音乐训练的现成模型进行比较；2）提出初步实验，以了解在何种同质性条件下这种适配能够成功；3）开源我们的实验并提供预训练模型。

© L. S. Maia、M. Rocamora、L.W.P. Biscainho 和 M. Fuentes。根据知识共享署名 4.0 国际许可协议（CC BY 4.0）授权。
署名： L.S. Maia、M. Rocamora、L.W.P. Biscainho 和 M. Fuentes，“将拍号跟踪模型适配于拉丁美洲音乐”，发表于第 *23* 届国际音乐信息检索学会会议论文集，印度班加罗尔，2022年。

1.1 其他适配方法

现有自适应方法通常采用某种形式的迁移学习或微调，使用深度学习模型，并关注节拍跟踪 [13–15] 或起音检测 [18]。Fiocchi 等人 [13] 通过迁移学习，将一个节拍跟踪模型从西方音乐适配到希腊音乐数据集。作者探究了循环神经网络（RNN）和长短期记忆网络（LSTM），并使用动态贝叶斯网络（DBN）进行推断。该模型使用大型训练集进行微调，但成效有限；作者将此归因于该数据集带来的挑战。尽管这是探索如何适配特定音乐流派这一思路的一种有趣方法，但众所周知，RNN 和 LSTM 的计算成本很高，因此在现实场景中的模型适配方面，它们并非理想选择。

另一方面，Pinto 等人 [14] 和 Yamamoto [15] 研究了时间卷积网络（TCN），并着重探索如何将模型适配到用户感兴趣的特定曲目。这些作者证明，只需极少量的数据（以秒计），便能适配 TCN 模型，使其表现良好，尤其适用于在音乐上颇具挑战性的曲目。此外，TCN 模型轻量且计算效率更高。

最后，Fonseca 等人 [18]将类似思路应用于起音检测模型（同样采用 TCN）的适配，使其适用于一种拉丁美洲音乐传统：*Maracatu de baque solto*。作者仅使用几秒钟的人工标注数据，对 TCN 的最后几层进行微调，并展示了乐器专用模型在音乐学研究的起音自动标注方面的优势。

1.2 拉丁美洲音乐传统

坎东贝鼓乐是乌拉圭的一种音乐传统，是当地大众文化和非洲文化遗产的重要组成部分。其节奏以四四拍为基础，人们一边在街道上行进一边演奏，使用三种大小和音高各异的鼓：*chico* 鼓、*repique* 鼓和*piano* 鼓。每种鼓都有独特的节奏型和音乐角色。三种鼓还共用一种称为*clave* 节奏型或*madera* 节奏型的额外时间线节奏型。*chico* 鼓是节拍维持者；它重复一个一拍的节奏型，在整个演奏过程中确立拍点。*repique* 鼓是即兴演奏者；它交替演奏*clave* 节奏型和具有鲜明切分特点的乐句。*piano* 鼓通过独特的单循环节奏型勾勒时间线，偶尔还会插入带装饰音的、类似*repique* 鼓的乐句。这种节奏与非裔大西洋世界的其他音乐传统有许多共同特征。值得注意的是，其中一些节奏型具有鲜明的现象学重音，这些重音相对于节拍结构发生偏移，并以不规则方式划分节奏循环，在拍点上只有少数击奏。

与此同时，桑巴 是一种深深植根于巴西文化、也有非洲起源的巴西音乐流派。“桑巴”一词实际上指的是一系列不同的子流派，其中最著名的或许是恩雷多桑巴、帕蒂杜阿尔图、巴萨诺瓦和帕戈德。与坎东贝类似，桑巴也可以在游行时演奏，而这种形式最常见于狂欢节庆典期间。它也可以在更随意的场合演奏，例如桑巴圆圈舞和酒吧中，甚至可以采用类似室内乐的风格。其节奏通常被感知为二四拍，并由多种打击乐器来呈现——坦博林鼓、手鼓、苏尔多鼓、库伊卡鼓、阿戈戈铃等。每种乐器都有若干独特的节奏型 [19], ，而且不止一种乐器可以担任节拍维持者。由于音色与音高的组合，演奏的织体可能变得十分复杂。桑巴 具有鲜明的特点，例如第二拍上的强重音，以及反节拍结构的发展。

2. 方法

基于我们在第 1.2 节中讨论的、关于坎东贝 和桑巴 具有高度同质性的直觉，我们的目标是了解能否使用少量来自这些音乐传统的拍号标注数据来训练节拍跟踪模型；如果可以，需要多少数据。为此，我们使用逐渐增加的标注数据量训练模型，数据量从不到一分钟到近 40 分钟不等，并比较各配置之间的性能和计算成本。我们对比三种不同的训练策略：1）使用坎东贝 或桑巴 片段从头训练模型；2）对使用来自西方音乐多个数据集的 38 小时数据训练的模型进行微调，使其适用于坎东贝或桑巴；3）与前两种策略相同，但通过数据增强训练模型，以人为扩大“小数据”输入。实验中，我们采用最先进的时序卷积网络模型 [16]，因为它在性能和计算成本之间实现了良好平衡。我们将该模型与使用西方音乐训练的现成模型进行比较。为了解深度学习方法的适应性和计算成本，我们将时序卷积网络与另一种简单而有效的基线——贝叶斯模型（BayesBeat） [17]进行比较。下文将详细介绍我们的方法。

2.1 数据集

我们为实验选取了两种不同的非洲裔拉丁美洲音乐传统的数据集。首先是坎东贝数据集 [12, 20], ，其中包含 35 段坎东贝鼓乐录音，总时长近 2.5 小时。每条音轨都录制了由三至五名鼓手组成的集成演奏，他们使用不同的鼓组合。速度变化很大，而且演奏过程中通常会逐渐加快。为呈现桑巴流派，我们使用 BRID 数据集 [21]中的“声学混音”数据。这些数据由 93 段短音轨组成（每段约 30 秒），记录了音乐家合奏桑巴及其两个子流派（恩雷多桑巴和帕蒂杜阿尔图）中的节奏型。数据集中包含十种不同的乐器类别，每条音轨由两至四名音乐家共同演奏。数据集包含多种速度；不过，每条音轨内的速度相当稳定。为了使用两种音乐的数据量大致相同地训练模型，并保持一致，坎东贝

和桑巴，同时也便于比较两组数据各自的结果，我们将坎东贝音轨分割成互不重叠的 30 秒片段。每次实验重复时，我们抽取 93 段坎东贝片段。

我们还使用六个数据集来训练基线TCN模型：Ballroom、Beatles、GTZAN 和 RWC（古典、流行、爵士）。这些数据集常用于拍号跟踪任务，合计包含超过 38 小时的音频数据。Ballroom 和 GTZAN 数据集涵盖了多种不同的音乐流派（例如华尔兹、探戈、伦巴、摇滚、流行、乡村等）。我们使用了 *mirdata v0.3*。*6* [28], 中的加载器，但 Ballroom 使用了自定义加载器。

2.2 使用小规模数据集

在我们的实验中，所有情况下，我们首先将训练数据和测试数据分开（分别为 93 个片段的 80% 和 20%），以确保能够公平地评估模型。然后，我们将训练数据分成六个子集，分别包含 {4, 9, 18, 37, 55, 74} 段 30 秒的音轨。我们希望确定模型对少量数据的适应方式有何不同，因此采用了与 [14] 类似的方法来确定用于训练的数据量。我们选取音频片段开头处连续 10 秒的时间区域，以及对应的拍和强拍标注，并舍弃其余音频部分。随后，我们将每个区域分成两个相邻的 5 秒部分：第一部分用于训练，第二部分留作TCN模型的验证；或者，我们将整个 10 秒区域用于训练贝叶斯模型，并采用现成的参数。考虑到每个片段仅有 10 秒，这些数据子集分别累计约 40 秒、1.5 分钟、3 分钟、6 分钟、9 分钟和 12 分钟的标注。采用这一策略的理由是，在现实应用中，若面对一组具有此类拉丁美洲音乐传统的录音，让用户标注几秒到几分钟的数据是合理的；当然，所需数据越少越好。

鉴于我们使用的数据点很少来训练模型，模型性能会受到数据采样的显著影响。为减轻这一影响，我们使用不同的随机种子生成随机数据划分，并将所有实验重复 10 次；也就是说，对于每种不同的数据子集大小，模型都训练 10 次。需要注意的是，选择最佳的数据采样策略不在本研究的范围内，将留待未来探讨。测试数据不进行裁切，即使用完整的 30 秒片段，以保持与拍号跟踪中常见模型评估实践的兼容性。

2.3 TCN模型

在我们的实验中，我们使用了 [16], 中介绍的时序卷积网络多任务模型，具体来说，是 [8]的开源实现。本研究专注于拍号跟踪，忽略网络的速度估计头。首先，时序卷积网络估计拍和强拍的似然。然后，我们使用 DBN 的两种不同实现（*DBN*- 拍跟踪器 和 *DBNDownBeatTracker* ，来自*madmomv0.17.dev0* [29]）来推断拍和强拍的最终位置。分别推断二者，而非联合推断，取得了更好的结果。

2.4 训练策略

2.4.1 从头训练（*TCN-FS*）

对于乐器配置、节奏模式和速度方面相似度较高的数据集，我们预计只需使用少量训练样本从头训练模型，就能使其在大多数数据上表现良好。

根据第 2.2 节中的说明，我们针对每个数据子集训练一个模型，并使用随机初始化的权重和随机种子重复此过程 10 次。我们还考虑了全部标注均可用的情况，并分析使用全部 30秒片段进行训练时的模型性能。在这种情况下，我们将 74 个训练片段划分为训练集和验证集（75%/25%）。对于每种策略，我们使用 0.005 的学习率；如果验证损失在 10 个训练轮次后没有改善，则将学习率降低至原来的 0.2 倍。训练最多进行 100 个训练轮次，并在 20 个训练轮次时提前停止。

2.4.2 微调（*TCN-FT*）

我们也从“迁移学习”的视角来处理特定文化场景下的拍号跟踪问题。按照 [13, 14, 18], ，我们调整了一个此前针对不同音乐语境训练的拍号跟踪模型。这里的直觉是，如果模型先在大型数据集上训练，即使它是围绕西方音乐构建的，也能为针对特定训练范围之外的音乐传统进行调优的模型提供良好的起点。这是一种切实可行的方法，因为大多数可用的标注数据和训练好的模型都以西方音乐为基础。为此，我们在 Ballroom、Beatles、GTZAN 和 RWC 数据集上训练了一个基线TCN模型。由于训练数据的特点，这个基线模型必须应对许多不同的节拍、流派和声学条件，因此可作为良好的起点。我们使用此前描述的相同训练流程对其进行微调，将初始学习率降至 0.001，即 FS 方法所用数值的五分之一，如 [14] 所示。

2.4.3 数据增强（*TCN-FTA*，*TCN-FSA*）

数据增强技术有助于人为增加训练样本的数量，这在像我们这样数据量较少或不足的情况下大有裨益。为了评估数据增强对我们模型的影响，我们在 TCN模型的实验中采用了一种受 [14, 16] 启发的简单策略：以不同的帧率计算输入的短时傅里叶变换，也就是改变跳步长度，从而使训练集中的速度分布更加均衡。我们没有围绕标注速度从正态分布中随机采样，而是选取了其取值上下 ±2.5% 和 ±5% 的一组帧率。这使我们的样本量增加了五倍，同时保持标注工作量不变。通过数据增强流程得到的模型分别标记为 TCN-FSA 和 TCN-FTA，对应第 2.4.1 节和第 2.4.2 节所述的训练策略。

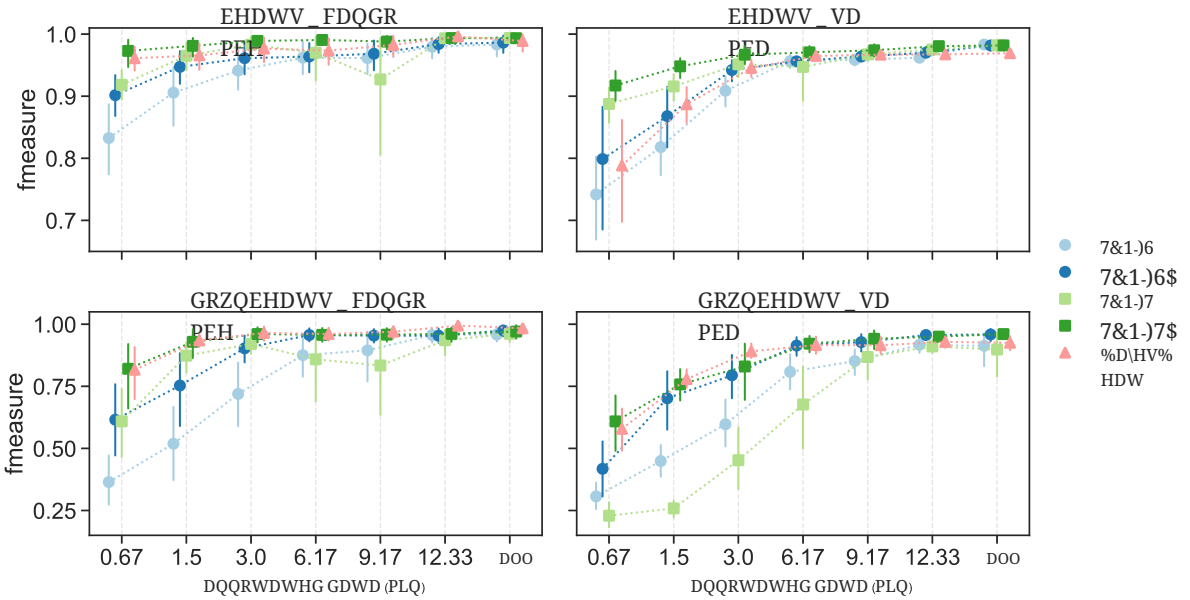


图1：性能　　不同模型和训练配置的表现。“全部”表示完全标注的数据集。

2.5 基线

我们纳入了两种类型的基线。首先，BayesBeat 统计模型 [17] 用作参照，用于评估 TCN 的适配能力和计算成本。它的参数较少，因此训练速度更快。第二种基线包括三个现成模型——一种信号处理技术和两个使用西方音乐训练的神经网络——它们说明了在我们的应用场景中量身定制解决方案／进行适配的必要性。详情如下。

BayesBeat。 该方法基于动态小节指针模型 [30], , 并通过将节拍、强拍、速度、拍号和节奏型表示为隐马尔可夫模型（HMM）中的隐变量，同时对它们进行估计。基于谱通量的观测特征由音频信号计算得出，观测模型则采用高斯混合模型（GMM）；这些模型在训练过程中针对一小节网格中每个格点的特征值进行拟合，从而学习节奏型。模型可以包含多个模式，不过假设整个音频信号中始终保持同一种模式。

BayesBeat 有一些超参数，用户应根据音乐选择。这些参数包括节奏型的数量、所用特征的类型（例如仅使用低频，或同时使用低频和高频）、特征分组方式（例如如何计算节奏型簇）、速度范围以及全音符细分方式。在[17], 中报告称，使用两个独立的频带（$\gtrsim$ 250 Hz）有助于找到正确的节拍层级，并有利于节拍跟踪和强拍跟踪。但考虑使用更多频带并未改善结果 [17]。根据 [31], , 每种节奏类别使用一种节奏型通常就足以获得良好表现，而且在大多数情况下效果最佳。因此，我们使用一种节奏型和两个频带。

现成的基线方法。 我们采用 Böck 等人 [4] 的联合节拍与下拍跟踪模型，其实现见*madmomv0.17.dev0* [29]。该模型由一个长短期记忆网络构成，该网络基于涵盖西方流派以及卡纳提克音乐、克里特音乐和土耳其音乐片段的十个数据集进行训练。我们还纳入了 Ellis [32], 的节拍跟踪器，该跟踪器会估计全局速度，然后使用动态规划找出最能反映该速度的一组节拍。作为最后一个基线方法，我们纳入了使用第 2.1 节中的西方音乐数据集训练的第 2.3 节时序卷积网络（TCN-BL）。

2.6 评估指标

我们采用 F值 [33], 作为主要指标，同时使用基于连续性的指标 [34, 35] CMLt（“正确的节拍层级”），其对应于正确节拍与标注节拍之比；以及 AMLt（“允许的节拍层级”），该指标允许估计中出现半个拍周期的相位误差或八度误差。对于模型的计算成本，我们仅报告使用代码中的内置计时函数对其训练所需的时间。

3. 实验与结果

3.1 模型的表现

图1展示了使用各种训练策略、分别采用不同数据量训练的坎东贝 和桑巴 时序卷积网络模型的 F值结果，以及 BayesBeat 的结果；后者是针对每种模型和数据量组合，使用不同随机种子进行十次实验后计算出的自助法结果（95% 置信度）。

一个引人注目的初步观察是，对于拍和强拍，大多数模型的性能曲线都略微向上倾斜，这意味着只用少量样本训练，就有可能接近模型的最佳性能（要达到这一水平，原本需要使用全部数据集进行训练）。这一点在拍的估计上尤其明显：在两种坎东贝 和桑巴 节奏中，对于所有配置，模型只用 1.5 分钟的数据，F值便迅速超过 80%。这是一个有趣的结果，意味着

模型	坎东贝						桑巴					
	Beat			强拍			Beat			强拍		
	CMLt	AMLt	F	CMLt	AMLt	F	CMLt	AMLt	F	CMLt	AMLt	F
BayesBeat_0.67	95.0	95.0	96.1	82.2	92.0	81.7	70.5	74.2	78.8	57.4	72.9	57.9
BayesBeat_12.33	99.6	99.6	99.6	99.8	99.8	99.4	93.5	96.0	96.7	92.5	94.9	92.9
BayesBeat_全部	98.6	98.6	98.9	98.8	98.8	98.4	94.0	96.0	96.9	92.0	95.3	92.5
TCN-FSA _0.67	88.2	89.1	90.2	56.5	70.4	61.6	74.4	78.2	79.9	24.3	60.1	41.8
TCN-FSA _12.33	97.9	98.0	98.4	95.8	98.0	95.4	94.3	96.3	97.0	92.8	97.5	95.7
TCN-FSA _all	98.4	98.4	98.6	97.4	98.4	97.5	96.0	98.6	98.2	95.0	96.6	95.9
TCN-FTA _0.67	96.7	96.7	97.3	81.8	89.7	82.1	85.3	93.0	91.7	28.5	81.5	60.9
TCN-FTA _12.33	99.4	99.4	99.4	96.5	99.5	96.1	95.8	97.8	98.1	92.0	97.2	95.0
TCN-FTA _all	99.4	99.4	99.4	97.4	99.5	97.1	96.3	98.4	98.2	96.1	97.7	96.0
TCN-BL	11.1	18.7	15.9	14.9	31.9	4.1	46.5	65.6	60.0	5.9	52.5	9.6
Ellis [32]	34.8	38.1	38.0	-	-	-	82.3	87.6	87.1	-	-	-
Böck [4]	11.7	14.4	11.5	26.7	40.3	0.5	46.9	76.0	66.4	5.2	66.6	2.0

表1：两种流派中节拍和强拍跟踪任务的平均 F值（F）及连续性得分（CMLt、AMLt）。

对于这类数据集，增加标注数量预计不会带来太大的性能提升。最终用户只需标注约一分钟的数据，就能获得不错的性能。对于坎东贝中表现最佳的模型，强拍也有同样的情况，但桑巴则不同。对于后者，增加数据显然能带来提升，这与这两种节奏之间的差异有关，以下将对此进行讨论。

坎东贝与**桑巴**之间的差异。观察图1中的结果可以发现，相较于坎东贝，模型在桑巴上通常需要更多数据才能取得更好的表现。我们对此的直觉解释是，正如第 1.2 节所述，由于桑巴的音色和音高组合比坎东贝更加丰富，因此决定标注哪些片段（即采样）对前者而言可能比对后者更为关键，例如确保音色具有代表性。

最佳模型配置。在这两种音乐传统中，节拍和强拍跟踪表现最佳的配置都是采用数据增强的微调后的TCN模型（FTA）。具体而言，数据增强显著提升了桑巴强拍跟踪的表现。有趣的是，对于本文关注的自适应设置，BayesBeat基线与TCN模型相比颇具竞争力，尤其考虑到计算

成本（见第 3.2 节）。

与现成基准的比较。表 1 展示了 TCN 和 BayesBeat 基线在不同数据子集（即最小和最大子集，以及完整数据集）上的表现，也展示了第 2.5 节所述的三个现成基线的表现。与以往研究 [12, 21],一致，使用西方音乐训练的模型（TCN-BL 和 Böck [4]）在坎东贝中的表现很差，在桑巴中的 F值也仅约为 66%，两者都显著低于相同模型在西方音乐流派中的表现。Ellis [32] 的模型得分高得多，但在两个数据集上的表现并不一致。这表明有必要针对这些音乐流派适配拍号跟踪模型，因为即使是使用最小数据子集（0.67 和 1.5 分钟）训练的模型，表现也优于基线。

3.2 模型训练需要多长时间?

我们的分析旨在探讨拍号跟踪模型在真实应用场景中的适配。要让这种适配切实可行，就必须快速完成。为此，我们估算了每种模型配置的训练耗时，并将其与 BayesBeat 基线进行比较。图 2 展示了桑巴的训练时长如何随训练集规模变化（坎东贝也得到了非常相似的结果）。TCN 在桑巴和坎东贝中的耗时大致相同，最小子集的训练耗时约为 100 秒。在各类 TCN 配置中，耗时最长的是使用数据增强的配置。考虑到训练时使用了更多数据，这并不意外。正如预期，BayesBeat 的训练速度显著快于 TCN：使用 0.67 分钟的数据训练平均耗时 1.62 秒；不使用数据增强时，其速度约为 TCN 的 50 至 350 倍。计算时间上的巨大差距，加上图 1 和表 1 的结果，说明 BayesBeat 总体而言是适配这些拉丁美洲音乐拍号跟踪任务的良好替代方案。我们观察到，所有配置的推理耗时大致相同：在完整测试集上约为 20 秒。

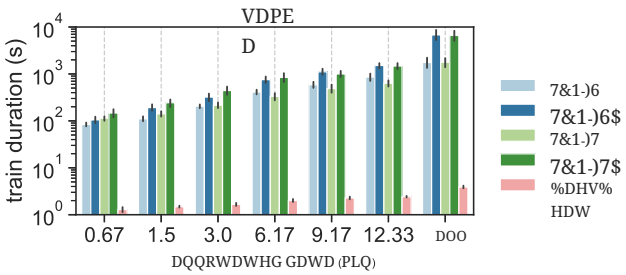


图 2：不同数据量下的训练时间。

3.3 数据量较少时何时可以训练?

我们的直觉是，数据中的变化越多（就拍号、节奏模式和乐器编制而言），模型就越难凭借少量数据进行学习。这与我们对这些方法适用于桑巴 和坎东贝的实验结果一致，也与第 1.2 节的音乐学见解相符。为了更

定量地理解这一点，我们为每种音乐推导出一个小节轮廓。首先，我们利用拍/强拍标注，从每个片段中提取特征图，将局部归一化的起音强度函数 [36] 按细分拍尺度进行时间量化——这是使用*carat* [37] 工具箱完成的，并将细分拍时长设为相邻节拍之间时间跨度的四分之一。然后，对于每个数据集，我们对这些特征图在时间维度上进行汇总，从而得到每个细分拍对应的特征值分布。为了分析这些轮廓在频谱不同区域的差异，我们计算两个频带（20 Hz 至 200 Hz；以及 > 200 Hz）内的起音强度。我们在图 3 中以小提琴图展示坎东贝、桑巴以及 Ballroom 数据集的这些分布。

在图 3 中，我们验证了某些细分拍的强度分布集中在 1 或 0 附近，分别表明小节中该位置具有鲜明的特征性重音，或没有重音。反过来，较高的方差意味着节奏模式“模糊不清”，这或许可以解释为什么难以学习这种节奏，尤其是在数据量较小时。

桑巴每小节有八个细分拍（二四拍），以在第 2 拍上有强烈的节拍重音而闻名；我们可以轻易地在其低频通道的第 5 个细分拍处识别出这一点。第一拍的中位数也较高，但由于方差较大，其表现不那么“确定”。相比之下，坎东贝 的低频轮廓呈现出方差较大的强拍、第 2 拍上没有重音，以及第 3 拍和第 4 拍上的强重音，同时在第 4 个细分拍处还有强烈的反节拍重音。这些特征或许有助于解释为什么预期节拍位置应有重音的现成节拍跟踪模型，在坎东贝上的表现最差。回看桑巴，我们发现第 2 个和第 3 个细分拍的标准差较小，且对应于“弱”细分拍；加上第 2 拍，低频通道中方差极小的细分拍共有八个中的三个。在坎东贝中，除了第 4 个细分拍，第 2、3、7、8、9、14 和 16 个细分拍也呈现出较小的方差。这些“锚点”数量众多，或许可以解释为什么只需少量数据就能完成坎东贝 的适配。

在 Ballroom 中，我们清楚地看到，拍的强度较高且在两个声道中的方差较低，因此十分鲜明；而其余的 tatum 则没有明显趋势。参考点较少，可能会给模型学习带来挑战。此外，节拍模式（拍与拍之间四个 tatum 的组合，包括拍本身）也
彼此无法区分，这可能会加剧这一问题。为检验这些观察结果，我们采用与桑巴 和坎东贝相同的方法，从头开始为 Ballroom 训练了一组模型。结果如图 4 所示。性能结果与我们的直觉一致：Ballroom 是一个更具挑战性的数据集，尤其是在节拍跟踪方面，因为它包含多个流派；此外，要学习节拍和强拍，还需要更多数据。

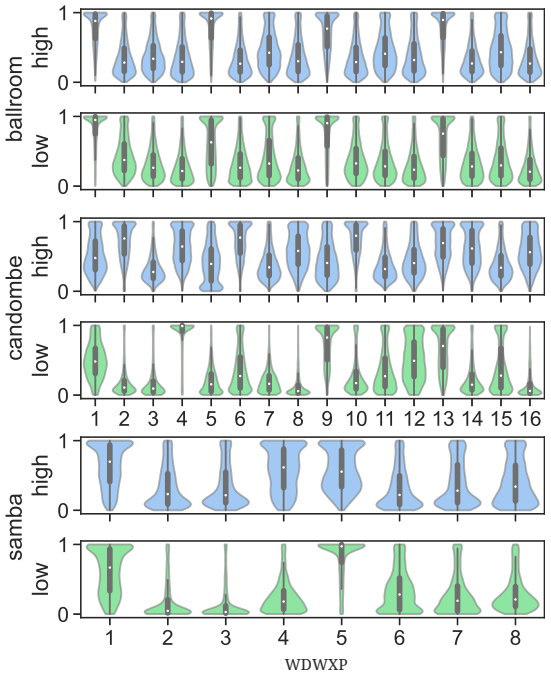


图 3：Ballroom（仅限 4/4 拍曲目）、坎东贝和桑巴各频带的细分拍强度分布。

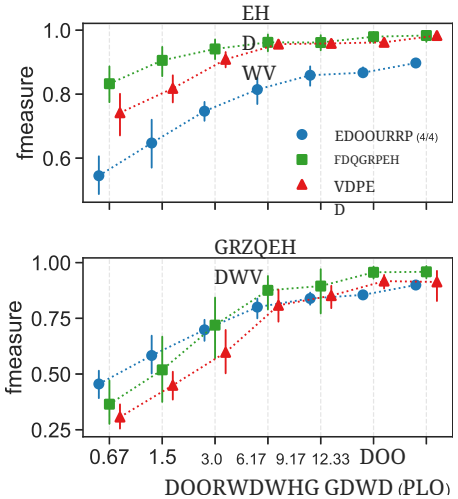


图 4：TCN-FS 在 Ballroom 中的表现。

在满足一定同质性条件的情况下，确实可以仅用几分钟的标注数据和几个训练周期来训练此类模型，并使其性能几乎达到在完整数据集上训练的水平。这一结果对现实应用具有积极意义，因为它让人们有望以适度的标注工作量，将此类模型调整到其他音乐流派。最具竞争力的模型是采用数据增强的微调后 TCN，而在计算成本受限时，BayesBeat 是一个不错的选择。未来，我们将研究可用于预测将拍号跟踪模型适配到特定音乐流派所需标注数据量的节奏复杂度指标。

4. 结论与未来工作

我们利用少量数据调整了一个拍号跟踪模型，使其适用于特定的拉丁美洲音乐传统，即桑巴 和 坎东贝。我们表明，

5. 致谢

本研究部分由巴西高等教育人员提升协调机构（CAPES）资助——财务代码001。

6. 参考文献

[1] A. Klapuri、A. Eronen 和 J. Astola，《声学音乐信号拍号分析》，*IEEE* 音频、语音与语言处理汇刊，第14卷，第1期，第342–355页，2006年。

[2]B.-Y. Chen、W.-H. Hsu、W.-H. Liao、M. A. M. Ramírez、Y. Mitsufuji 和 Y.-H. Yang，《借助可微音频效果和生成对抗网络实现自动 DJ 过渡》，载于 *2022*年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*）论文集，2022年，第 466–470 页。

[3]A. Srinivasamurthy、A. Holzapfel、K. K. Ganguli 和 X. Serra，《印度斯坦音乐中的速度与节奏展开：一项语料库研究》，数字人文前沿，第 4 卷，2017 年。[在线]。获取地址：https://www.frontiersin.org/article/10.3389/fdigh.2017.00020

[4] S. Böck、F. Krebs 和 G. Widmer，《使用循环神经网络联合跟踪拍与强拍》，载于第 *17* 届国际音乐信息检索学会会议（*ISMIR*）论文集，美国纽约，2016 年 8 月，第 255–261 页。

[5] S. Durand、J. P. Bello、B. David 和 G. Richard，《使用卷积网络集成进行稳健的强拍跟踪》，*IEEE*《音频、语音与语言处理汇刊》，第25卷第1期，第76–89页，2017年1月。

[6] M. Fuentes、B. McFee、H. Crayencour、S. Essid 和 J. Bello，《深度学习系统中用于强拍跟踪的常见设计选择分析》，载于第*19*届国际音乐信息检索学会会议论文集（*ISMIR*），法国巴黎，2018年9月，第106–112页。

[7] M. Heydari、F. Cwitkowitz 和 Z. Duan，《Beatnet：用于在线联合节拍、强拍和拍号跟踪的卷积循环神经网络与粒子滤波》，载于第*22*届国际音乐信息检索学会会议论文集（*ISMIR*），在线，2021年11月，第270–277页。

[8]M. E. P. Davies、S. Böck 和 M. Fuentes，速度、节拍与强拍估计。https://tempobeatdownbeat.github.io/tutorial/intro.html，2021年11月。[在线]。可访问：https: // tempobeatdownbeat.github.io/tutorial/intro.html

[9] C. N. Silla Jr.、A. L. Koerich 和 C. A. A. Kaestner，“拉丁音乐数据库”，载于第*9*届国际音乐信息检索会议（*ISMIR*），美国费城，2008年9月，第451–456页。

[10] E. Cano、F. Mora-Ángel、G. A. López Gil、J. R. Zapata、A. Escamilla、J. F. Alzate 和 M. Betancur，“哥伦比亚班布科中的混合拍：节拍与拍号的感知和估计”，载于第*21*届国际音乐信息检索学会会议（*ISMIR*）论文集，加拿大蒙特利尔，2020年10月，第409–415页。

[11] G. M.SarriaM.、J. Diaz 和 C.Arce-Lopera，《分析并扩展萨尔萨音乐数据集》，发表于第二十二届图像、信号处理与人工视觉研讨会（*STSIVA*）论文集，哥伦比亚布卡拉曼加，2019年4月，第 1–5 页。[12] L. Nunes、M. Rocamora、L. Jure 和 L. W. P. Bis- cainho，《基于节奏模式的节拍与强拍跟踪：应用于乌拉圭坎东贝鼓乐》，发表于第十六届国际音乐信息检索学会会议（*ISMIR*）论文集，西班牙马拉加，2015年10月，第 246–270 页。[13] D. Fiocchi、M. Buccoli、M. Zanoni、F. Antonacci 和 A. Sarti，《基于循环神经网络的节拍跟踪：一种迁移学习方法》，发表于第二十六届欧洲信号处理会议（*EUSIPCO*）论文集，意大利罗马，2018年9月，第 1929–1933 页。

[14] A. S. Pinto、S. Böck、J. S. Cardoso 和 M.E.P. Davies，《用于节拍跟踪的用户驱动微调》，《电子学》，第10卷第13期，2021年6月。

[15] K. Yamamoto，《用于交互式音乐节拍跟踪的人在回路自适应》，载于《第*22*届国际音乐信息检索会议（*ISMIR*）论文集》，在线，2021年11月，第794–801页。

[16] S.Böck 和 M.E. P.Davies，《解构、分析、重构：如何改进速度、拍和强拍估计》，载于《第*21*届国际音乐信息检索会议（*ISMIR*）论文集》，加拿大蒙特利尔，2020年10月，第574–582页。

[17] F. Krebs、S. Böck 和 G.Widmer，《从音乐音频中进行拍和强拍跟踪的节奏模式建模》，载于《第*14*届国际音乐信息检索会议（*ISMIR*）论文集》，巴西库里蒂巴，2013年11月，第227–232页。

[18] J. Fonseca、M. Fuentes、F. Bonini Baraldi 和 M. E. P. Davies，“利用自动起音检测分析 Maracatu de baque solto”，载于《音乐、声音与音乐学视角：研究、教育与实践》，L. C. Castilho、R. Dias 和 J. F. Pinho 编。瑞士查姆：施普林格，2021年，第209–225页。

[19] G. Gonçalves 和 O. Costa，卡里奥卡律动：里约热内卢桑巴学校鼓乐组。巴西里约热内卢：Groove，2000年。

[20]M. Rocamora、L. Jure、B. Marenco、M. Fuentes、F. Lanzaro 和 A. Gómez，“用于计算音乐学研究的坎东贝表演视听数据库”，载于第二届国际音乐科学与技术大会（*CICTeM*）论文集，阿根廷布宜诺斯艾利斯，2015年9月，第17–24页。

[21] L. S. Maia、P. D. Tomaz Jr.、M. Fuentes、M. Rocamora、L. W. P. Biscainho、M. V. M. Costa 和 S. Cohen，“巴西节奏乐器的新数据集及计算节奏分析实验”，载于《*2018*年*AES*拉丁美洲音频工程大会》（*AES LAC*）论文集，乌拉圭蒙得维的亚，2018年9月，第53–60页。

[22] F. Gouyon、A. Klapuri、S. Dixon、M. Alonso、G. Tzane-takis、C. Uhle 和 P. Cano，“音频速度诱导算法的实验比较”，《*IEEE*音频、语音与语言处理汇刊》，第14卷，第5期，第1832–1844页，2006年9月。

[23] M. E. P. Davies、N. Degara 和 M. D. Plumbley，《音乐音频节拍跟踪算法的评估指标》，伦敦玛丽女王大学，英国伦敦，技术报告，2009年10月。[24] G. Tzanetakis 和 P. Cook，《音频信号的音乐流派分类》，*IEEE* 音频、语音与语言处理汇刊，第10卷，第5期，第293–302页，2002年7月。[25] U. Marchand 和 G. Peeters，《摇摆比估计》，载于第*18*届数字音频效果国际会议（*DAFx*）论文集，挪威特隆赫姆，2015年12月，第423–428页。[26] M. Goto、H. Hashiguchi、T. Nishimura 和 R. Oka，《RWC音乐数据库：流行、古典和爵士音乐数据库》，载于第*3*届音乐信息检索国际会议（*ISMIR*）论文集，法国巴黎，2002年10月，第287–288页。[27] M. Goto，《RWC音乐数据库开发》，载于第*18*届国际声学大会（*ICA*）论文集，日本京都，2004年4月，第I–553–556页。[28] R. M. Bittner、M. Fuentes、D. Rubinstein、A. Jansson、K. Choi 和 T. Kell，《mirdata：用于可复现使用数据集的软件》，载于第*20*届国际音乐信息检索学会会议（*ISMIR*）论文集，2019年。[29] S. Böck、F. Korzeniowski、J. Schlüter、F. Krebs 和 G. Widmer，《madmom：一个新的 Python 音频与音乐信号处理库》，载于第*24*届 *ACM* 多媒体国际会议论文集，荷兰阿姆斯特丹，2016年10月，第1174–1178页。[30] N. Whiteley、A. Cemgil 和 S. Godsill，《音乐音频时间结构的贝叶斯建模》，载于第*7*届音乐信息检索国际会议（*ISMIR*）论文集，加拿大维多利亚，2006年10月，第29–34页。[31]A. Holzapfel、F. Krebs 和 A. Srinivasamurthy，《追踪“奇数”：文化多样化音乐语料库中的拍号推断》，载于第*15*届国际音乐信息检索学会会议（*ISMIR*）论文集，中国台北，2014年10月，第425–430页。[32] D. P. W. Ellis，《通过动态规划进行节拍跟踪》，《新音乐研究杂志》，第36卷，第1期，第51–60页，2007年3月。[33] S. Dixon，《音频节拍跟踪系统 BeatRoot 的评估》，《新音乐研究杂志》，第36卷，第1期，第39–50页，2007年。[34] S. W. Hainsworth，《音乐音频自动分析技术》，博士学位论文，剑桥大学工程系，2003年。

[35] A. P. Klapuri、A. J. Eronen 和 J. T. Astola，《声学音乐信号的拍号分析》，*IEEE* 音频、语音与语言处理汇刊，第14卷，第1期，第342–355页，2006年1月。[36] M. Rocamora、L. Jure 和 L. W. P. Biscainho，《从坎东贝录音中检测和分类钢琴鼓点模式的工具》，载于第*9*届跨学科音乐学会议（*CIM14*）论文集，德国柏林，2014年12月，第382–387页。[37] M. Rocamora 和 L. Jure，《carat：计算机辅助节奏分析工具箱》，载于世界音乐分析方法会议（*AAWM*）论文集，英国伯明翰，2019年7月。