柔性动力学 C 用于多音高估计及其他用途的时间扭曲

Michael Krause，Christof Weiß，Meinard M¨uller

埃尔朗根国际音频实验室

arXiv:2304.05032v1 [cs.SD] 11 Apr 2023

摘要

音乐信息检索（MIR）中的许多任务都涉及弱对齐数据，其中确切的时间对应关系是未知的。连接主义时序分类（CTC）损失是一种基于弱对齐训练数据学习特征表示的标准技术。然而，CTC 仅限于离散值目标序列，并且难以扩展到多标签问题。在本文中，我们展示了如何将软动态时间规整（SoftDTW）这一经典动态时间规整的可微变体用作 CTC 的替代方案。以多音高估计为示例场景，我们展示了 SoftDTW能够取得与最先进的 CTC 多标签扩展相当的结果。除了在算法表述方面更加优雅之外，SoftDTW 还能够自然地扩展到实值目标序列。

索引词— 动态时间规整、音乐处理、音乐信息检索、多音高估计、音乐转录

1. 引言

音乐信息检索（MIR）中的许多应用都需要对齐音乐数据序列。通常，给定的序列之间只有弱对齐。例如，在音频转乐谱转录中，音频与乐谱片段很容易找到，但很难确定这些片段之间精确的时间对应关系 [1]。此外，音乐数据序列可能涉及不同的复杂程度。例如，给定一段单乐器单声部音乐录音，单音高估计 [2] 旨在为每个时间步找出一个音高值（另见图 1a）。其他具有离散单标签目标的场景包括转录歌词，或为单人演唱的歌曲进行歌词对齐 [3, 4]。更复杂的序列出现在多音高估计（MPE）中，此时多个音高可能同时发声（图 1b）。最后，一些场景涉及实值序列之间的对齐（图 1c），例如音频间同步 [5, 6]，或多模态对齐问题，例如将舞蹈视频与音乐同步 [7]。

连接主义时序分类(连接时序分类) [8]损失是一种最初为语音识别开发的完全可微损失函数，通常用于从弱对齐的

当目标是有限标签字母表上的序列时的数据。最近，连接时序分类被扩展用于处理多标签学习问题 [9], ，其主要思想是将多标签情形局部转换为单标签情形。然而，除了其复杂的算法表述之外，这种方法不适用于不源自离散词汇表的目标序列。

本研究得到德国研究基金会（DFG MU 2686/7-2）的资助。作者隶属于埃尔朗根国际音频实验室，该实验室是弗里德里希-亚历山大大学埃尔朗根-纽伦堡分校（FAU）与弗劳恩霍夫集成电路研究所 IIS 的联合机构。作者谨此感谢埃尔朗根地区计算中心（RRZE）提供的计算资源与支持。

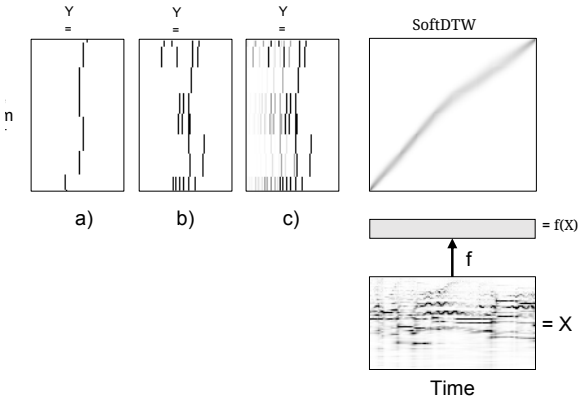

图 1：SoftDTW 用于对齐学习得到的特征序列 $f(X)$ 与目标序列 Y的示意图，其中可以考虑 (a) 单标签、(b) 多标签或 (c) 实值目标。

在音乐信息检索（MIR）中，一种常用的方法是结合手工设计的特征 [10]，通过动态时间规整（DTW）在弱对齐序列之间寻找最优对齐。这种流程能为音频—音频同步等任务提供良好的对齐结果 [6], ，但标准的 DTW 代价函数并非完全可微，因此无法用于端到端深度学习场景。为解决这一问题，Cuturi 和 Blondel [11] 提出了 DTW 的可微变体 SoftDTW，用于近似原始 DTW 代价。近期研究已成功将 SoftDTW 及相关技术用于动作对齐等计算机视觉应用 [12, 13]。据我们所知，先前唯一一项在音乐信息检索场景中应用 SoftDTW 的研究由 Agrawal 等人完成 [17]。

我们的贡献如下：我们展示了 SoftDTW 在 MPE 中的应用。具体而言，我们表明，SoftDTW 的性能与 CTC 的多标签扩展相当，同时在概念上更为简单。此外，如图 1所示，我们展示了SoftDTW方法能够自然地推广到实值目标序列，从而使其适用于广泛的对齐任务。

本文其余部分的结构如下：第 2 节回顾了利用连接时序分类（CTC）从弱对齐数据中进行多音高估计的当前最先进方法。第 3 节对适用于一般序列的 SoftDTW 进行形式化定义，第 4 节则将其应用于多音高估计。第 5 节展示了 SoftDTW 在实值目标学习中的潜力。最后，第 6 节总结全文，并展望未来的应用。

2. 用于 MPE 的弱对齐训练

近年来，自动音乐转录已成为音乐信息检索研究的核心课题，深度学习技术取得了最先进的成果 [14–16]。本文聚焦于多音高估计（MPE），将其视为自动音乐转录的一个子

©2023 IEEE。允许个人使用本材料。对于当前或未来媒体中的所有其他用途，包括为广告或促销目的重印/再版本材料、创作新的合集作品、转售或重新分发至服务器或列表，或在其他作品中重复使用本作品的任何受版权保护的组成部分，均须获得 IEEE 的许可。

问题，其目标是将输入音乐录音 X 转换为所演奏音高的钢琴卷帘表示 Y。具体而言，多个音高可能同时处于激活状态。大多数基于学习的 MPE 方法都需要经过严格对齐的数据进行训练，即为输入录音的每个音频帧标注音高。由于以这种逐帧方式标注数据非常耗时，大多数 MPE 数据集都是通过（半）自动方式生成的，例如使用 MIDI 钢琴，或采用乐谱与音频同步技术（这可能会引入标注错误）。因此，能够利用时间上未对齐的 X 与 Y 配对数据进行学习的技术尤为可取。

正如引言中所讨论的，处理弱对齐学习问题的一种常见技术是连接时序分类 [8]。在这里，目标序列 Y 由离散字母表 L 中的符号组成，其中包括一个用于区分符号重复的特殊空白符号。对于输入序列 X中的每一帧，神经网络都会输出关于 L的概率分布。连接时序分类损失随后对应于在给定这些网络输出的情况下 Y的似然，同时考虑 X 与 Y之间所有可能的对齐。请注意，连接时序分类与 Y中符号的持续时间无关，也就是说，即使有关于符号持续时间的信息，连接时序分类也无法利用这些信息进行对齐。存在一种用于计算连接时序分类损失的高效动态规划算法（时间复杂度$\mathcal{O}(|L|^2 \cdot N)$，其中 N 是 X的长度），但在处理空白符号 [8]时需要特别谨慎。

将 CTC 朴素地扩展到多标签目标序列，会为所有可能的符号组合引入唯一的网络输出，从而导致组合爆炸。相反， [9] 中的作者提出，仅考虑单个训练批次内出现的那些符号组合，将多标签情形局部地简化为单标签情形（称为多标签CTC，即 MCTC）。这定义了一个“依赖于批次的字母表”，从而避免了组合爆炸。该过程的技术细节较为棘手，需要特别注意空白符号的处理。在 [1], 中，这一思想被应用于MPE，将音高视为符号，并将多音高标注视为符-号的组合。这种表述使他们能够在仅弱对齐的 X 和 Y 对上训练用于MPE 的网络，例如， X 是音乐录音，而 Y 是从相应乐谱导出的MIDI 表示。在本文中，我们以来自 [1]的MPE 作为示例应用，展示技术上复杂的 MCTC如何被概念上更优雅的 SoftDTW方法所取代。SoftDTW不需要空白符号；在文本应用中，空白符号可能具有充分的合理性，但在 MPE 等 MIR 问题中可能并不自然。

3. 软动态时间规整

动态时间规整的目标是在两个序列之间找到最优的时间对齐。SoftDTW [11] 是动态时间规整的一种可微近似，允许在对齐过程中传播梯度，从而使 SoftDTW 适用于深度学习。与经典动态时间规整一样，SoftDTW 支持通过高效的动态规划（DP）递归计算最优对齐代价。此外，还存在一种用于高效计算该代价梯度的动态规划算法。本节将简要总结适用于一般序列的SoftDTW问题表述和动态规划递归。随后，我们将在后续章节中将其应用于音乐场景。

考虑两个序列 $X = (x_1, x_2, \ldots, x_N)$ 和 $Y = (y_1, y_2, \ldots, y_M)$，其长度分别为 $N, M \in \mathbb{N}$元素来自某些特征空间 $\mathcal{F}_1, \mathcal{F}_2$（即，$x_n \in \mathcal{F}_1$, $y_m \in \mathcal{F}_2$对所有元素均成立。$n \in [1 : N], m \in [1 : M]$)。给定定义在这些特征空间上的某个可微代价函数$c : \mathcal{F}_1 \times \mathcal{F}_2 \to \mathbb{R}$，我们可以构造一个局部代价矩阵 $C \in \mathbb{R}^{N \times M}$，其中每个条目

$$C(n, m) = c(x_n, y_m)$$

包含局部 对齐 x_n 与 y_m的代价。为了确定序列 X 与 Y之间的最优全局对齐[1]，需要通过递推计算累积代价矩阵 $D^\gamma \in \mathbb{R}^{N \times M}$。

$$\begin{aligned}
D^\gamma(1,1) &= C(1,1), \\
D^\gamma(1,m) &= \sum_{k=1}^{m} C(1,k), \text{ for } m \in [1:M], \\
D^\gamma(n,1) &= \sum_{k=1}^{n} C(k,1), \text{ for } n \in [1:N], \\
D^\gamma(n,m) &= C(n,m) + \mu^\gamma(\{D^\gamma(n-1,m-1), \\
&\qquad D^\gamma(n-1,m), D^\gamma(n,m-1)\}),
\end{aligned}$$

针对 $n \in [2 : N]$， $m \in [2 : M]$。这里， μ^γ 指的是最小值函数的一个可微近似，其定义为

$$\mu^\gamma(S) = -\gamma \log \sum_{s \in S} \exp\left(-\frac{s}{\gamma}\right),$$

其中 S 是一个有限实数集， $\gamma \in \mathbb{R}^{>0}$ 是一个温度参数，用于决定近似的“柔和”程度。可以证明， μ^γ 是最小值函数 [12] 的下界，并且当 $\gamma \to 0$时趋近于真实的最小值。因此，当 $\gamma \to 0$时， D^γ 会变为经典动态时间规整的累积代价矩阵。由此，SoftDTW 在极限情形下会变为 DTW。

计算完 SoftDTW 递归后，条目 DTW$^\gamma(C) = D^\gamma(N, M)$包含对序列 X 和 Y进行对齐的近似最小代价，其中给定的局部代价为 C。类似的递归还可用于计算 DTW$^\gamma(C)$ 相对于任意矩阵系数 $C(n, m)$ 的梯度，其中 $n \in [1 : N]$ 且 $m \in [1 : M]$ [11, 算法 2]。SoftDTW 递归和梯度计算的时间与空间复杂度均为 $\mathcal{O}(N \cdot M)$，足够快，可用于深度学习。

请注意，SoftDTW 不需要预先了解 X 和 Y 之间的对齐关系，因此可以将 DTW$^\gamma(C)$ 用作弱对齐数据学习问题的损失函数。此外，只要能够定义合适的代价函数 c， X 和 Y 就可以来自任意特征空间。

4. 应用于多音高估计

现在，我们将SoftDTW应用于多音高估计。对于给定的乐曲，序列 X 对应于输入录音的某种表示，而 Y对应于所演奏音高的多热编码。注意， Y不需要与 X 在时间上对齐，并且可能例如源自该乐曲的乐谱表示。序列 Y中的元素 y_m 被编码为向量$y_m \in \{0, 1\}^{72}$，而 y_m 的条目对应于从 C1到B6 的 72 个音高。在我们的实验中，我们不直接将 Y与某个固定表示 X进行对齐，而是使用一个以 X 为输入、并为 X中的每个帧输出一个特征向量的神经网络 f。因此，

[1]在一些约束条件下，即两条序列的第一个和最后一个元素彼此对齐（边界约束），不跳过任何元素（步长约束），且对齐是单调的（单调性约束）。

我们获得一个序列 $f(X) = (z_1, \ldots, z_N)$，其长度与 X 相同，即N。我们构造 f，使得 $z_n \in \mathbb{R}^{72}$ 对于 $f(X)$ 的元素 z_n 成立。因此，两个序列 Y 和 $f(X)$都包含来自特征空间 $\mathcal{F}_1 = \mathcal{F}_2 = \mathbb{R}^{72}$ 的元素。随后，我们将 $f(X)$ 与 Y 对齐，如图 1 所示。

据我们所知，SoftDTW 此前尚未用于 MPE，在 MIR 中也很少被探索。 [4] 中的作者将经典的、不可微分的 DTW 递归应用于歌词对齐的注意力机制中，导致训练不稳定。Agrawal 等人的工作 [17] 首次将 SoftDTW 用于 MIR 应用。他们成功地采用 SoftDTW 的一种变体来训练乐谱—音频同步系统。在他们的场景中，SoftDTW 被应用于离散值的一维强对齐序列。相比之下，我们在弱对齐设置下将 SoftDTW 用于多维序列。

4.1. 实现细节和评估指标

由于我们的工作重点是评估SoftDTW在MIR 任务中的效能，并为了保持与 [1], 中所呈现结果的可比性，我们采用相同的训练设置和网络架构。因此，我们使用长度约为十秒的谐波CQT（HCQT， [18]）片段作为输入，并将其输入一个五层卷积神经网络，以获得序列级的逐帧表示 $f(X)$（有关网络架构和 HCQT 表示的详细信息，请参见 [1] ）。

我们通过最小化 SoftDTW对齐代价 动态时间规整$\gamma(C)$来训练我们的网络。[2] 在所有实验中，我们对 c 使用平方欧氏距离，并设置 $\gamma = 10.0$。对于 c 的其他选择，我们没有看到性能提升；而对于 $\gamma \in [0.5, 20.0]$的广泛取值范围，我们获得了相似的结果。此外，我们使用了SoftDTW 递归和梯度计算的快速 GPU 实现，该实现由 [19]实现。

为了将网络预测结果与测试集中的严格对齐音高标注进行比较，我们采用 MPE的常用评估指标，包括预测结果与标注之间的余弦相似度（CS）、精确率-召回率曲线下面积（也称平均精度，AP），以及在 0.4阈值下的 F值和准确率（Acc.，在 [20] 中引入）（该阈值是 MPE 系统中的常见选择，另见 [21]）。

4.2. 与 MCTC 的比较

我们首先将我们的结果与[1], 中报告的主要结果进行比较，这些结果是在舒伯特《冬之旅》数据集（SWD）上获得的[22]。SWD为所有录音提供了高度对齐的标注。因此，可以考虑使用按帧交叉熵损失（CE）在对齐的注-释上训练的基线模型。表1 的第一行展示了这样一个乐观基线的结果（转载自 [1]），其 F值为 0。70，AP为 $= 0.764$。若要改用MCTC训练网络，则必须从标签序列中移除所有关于音符时值的信-息 Y（见图 2b）。以这种方式获得的结果仅略低一些，AP为 $= 0.734$，尽管使用的只是弱对齐标签。使用 SoftDTW（记为$\text{SoftDTW}_{\text{W1}}$）进行相同实验时，我们得到的结果要弱得多，F值为 0。00，AP为 =0。297。[3] 在该实验中，标签序列 Y 可能是无-法显-

[2]请注意，我们根据第一个训练批次的数值对动态时间规整$\gamma(C)$进行归一化。因此，第一个批次的损失恰好为 1，并且无论序列长度 N 和 M 或其他因素如何，其数值范围在不同训练配置之间都保持相近。[3]请注意，通过选择更合适的检测阈值，F值和准确率分数分别可以提高到 0.32 和 0.20。不过，与连接时序分类的结果相比，这些分数仍明显较差。

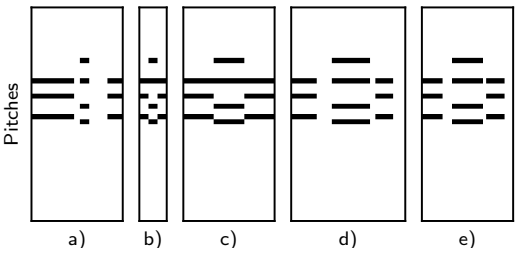

图 2：(a) 音频片段的强对齐音高标注，(b)不含音符时值的标注（MCTC所使用的标注），(c)不含音符时值、拉伸至片段长度的标注，(d) 未与音频片段对齐的乐谱表示，(e) 拉伸至片段长度的乐谱表示

场景	F值	CS	AP	Acc.
CE [1]	0.70	0.759	0.764	0.546
MCTC [1]	0.69	0.744	0.734	0.532
$\text{SoftDTW}_{\text{W1}}$	0.00	0.465	0.297	0.002
$\text{SoftDTW}_{\text{W2}}$	0.69	0.736	0.737	0.529

表1：SoftDTW 与 MCTC 在舒伯特《冬之旅》数据集上的多音高估计结果。

明显短于学习序列 $f(X)$。[4] 我们通过在时间上拉伸序列 Y，使其帧数与 $f(X)$ 相匹配，重复了该实验（如图 2c 所示）。将SoftDTW与这一技巧结合使用时（记为 $\text{SoftDTW}_{\text{W2}}$），结果再次与 MCTC 非常相似（AP $= 0.737$）。因此，在这种情况下，可以使用 SoftDTW 替代 MCTC。

4.3. 纳入音符时值

与MCTC 相比，SoftDTW能够在训练期间纳入（近似的）音符时值。例如，SWD 包含所演奏乐曲的非对齐乐谱表示。现在，我们将这些乐谱表示用作目标序列 Y（记作 $\text{SoftDTW}_{\text{W3}}$，示例见图2d）。表 2 展示了评估结果，与不使用音符时值进行训练相比略有提升（$\text{SoftDTW}_{\text{W2}}$ 的 F值为0.71，而不是0.69；CS $= 0.756$，而不是0.736）。这里，片段与标签序列的长度仅存在中等程度的差异，将标签序列拉伸至输入的长度会得到几乎相同的结果（记作 $\text{SoftDTW}_{\text{W4}}$，见图 2e）。最后，我们还可以使用具有强对齐的标签序列的SoftDTW（记作 $\text{SoftDTW}_{\text{S}}$）。在这一非常乐观的情形下，无需进行对齐，但 SoftDTW 可以弥补数据集标注过程所引入的不准确性。事实上，这种情形取得了最佳结果（F值为 0.72，AP $= 0.769$），甚至略微优于表 1 中的交叉熵基线。

4.4. 跨数据集实验

我们还进行了跨数据集实验（同样遵循 [1]中的设置），在流行的 MAESTRO [23]和 MusicNet [21] 数据集上进行训练。这两个数据集都为训练录音提供了高度对齐的音高标-注，但没有提供乐曲的非- 对齐乐谱表示，因此 $\text{SoftDTW}_{\text{W3}}$ 和

[4]众所周知，序列长度的巨大差异会给经典动态时间规整带来问题。还需进一步研究，以了解这会如何影响使用 SoftDTW 的训练过程。

场景	F值	CS	AP	Acc.
$\text{SoftDTW}_{\text{W3}}$	0.71	0.756	0.755	0.552
$\text{SoftDTW}_{\text{W4}}$	0.71	0.757	0.750	0.555
$\text{SoftDTW}_{\text{S}}$	0.72	0.761	0.769	0.563

表 2：在舒伯特《冬之旅》数据集上的结果，采用 SoftDTW 纳入音符时值。

场景	SWD	AP Bach10	AP TRIOS	AP Phenicx
默认网络架构				
CE [1]	0.684	0.864	0.825	0.829
MCTC [1]	0.666	0.861	0.824	0.833
$\text{SoftDTW}_{\text{W2}}$	0.665	0.835	0.812	0.788
更大的网络架构				
CE [1]	0.701	0.886	0.863	0.846
MCTC [1]	0.677	0.871	0.849	0.850
$\text{SoftDTW}_{\text{W2}}$	0.682	0.896	0.864	0.838

表3：跨数据集实验中的多音高估计结果。此处使用 MAESTRO 和 MusicNet 进行训练，同时使用四个不同的较小数据集进行测试。

$\text{SoftDTW}_{\text{W4}}$ 在此处并不适用。接下来，我们在四个较小的数据集 SWD 、 Bach10 [24], TRIOS [25] 和 Phenicx Ane- choic [26] 上进行评估。请注意，后三个数据集的音频时长均不足十分钟。这是一个具有挑战性的场景，因为测试数据集中的某些风格和乐器在训练期间并未出现。例如，Phenicx Anechoic 包含管弦乐器，而 MAESTRO 和 MusicNet 则包含钢琴和室内乐。

该实验的结果见表3。在这里，MCTC 和交叉熵基线的表现大致相当。SoftDTW 的结果略低，尤其是在 Phenicx 上（AP $= 0.788$，而 MCTC 为 0.833）。鉴于这一评估场景更具挑战性，且训练数据集规模更大，我们还使用更大的网络架构重复了该实验（增加网络中所有卷积层的通道数）。所得架构约有 600 000 个参数，而默认架构有 50 000 个参数。结果见表3下半部分。所有方法和数据集的平均精度分数均稳定提升，例如，较小架构下 Bach10 上 SoftDTW 的 AP $= 0.896$，而较大架构下为 0.835。具体而言，SoftDTW 现在在除 Phenicx 以外的所有测试数据集上均优于 MCTC；在 Phenicx 上，两者的性能差距也显著缩小（AP $= 0.838$，而 MCTC 为 0.850）。

总而言之，我们得出结论：即使是在具有挑战性的跨数据集评估中，MCTC 和 SoftDTW 的结果大致相当。因此，MCTC 可以替换为 SoftDTW，而不会牺牲对齐质量。此外，SoftDTW 还可以推广到其他类型的目标序列，下一节将对此进行讨论。

5. 扩展到实值目标 S

如第 3节所述，用作SoftDTW 输入的两个序列 X 和 Y可以来自任意特征空间。为了说明使用 SoftDTW 从任意序列中学习的潜力，我们现在针对实值目标进行两项实验，即对 Y 的元素 y_n 使用 $y_n \in \mathbb{R}^{72}$。请注意，MCTC 无法处理这种设置。

5.1. 使用泛音模型进行音高估计

首先，我们考虑 MPE的一种直接扩展：根据一个简单的泛音模型添加能量，将 MPE的二值、多热目标向量转换为实值向量，见图 1c。这里，我们为每个活跃音高考虑10 个泛音，第 n 个泛音的振幅为$(1/3)^n$。作为利用强对齐标签的基线，我们将其与一个在每个帧上使用 ℓ_2回归损失训练的模型进行比较（类似于第4节中的交叉熵基线）。为进行评估，我们使用网络输出与标注之间的余弦相似度 CS。请注意，其他 MPE 评估指标不适用于实值向量。

在 SWD 数据集上执行此实验时，我们得到使用强对齐标签进行逐帧训练的 CS $= 0.794$，其高于 SWD 上 MPE 的结果（参见表 1）。在没有强对齐标签的情况下使用 $\text{SoftDTW}_{\text{W2}}$ 训练，仅得到略低的余弦相似度 0。770。这表明 SoftDTW 同样适用于具有实值目标序列的设置。

5.2. 跨版本训练

其次，作为目标序列更贴近实际的情形，我们选择 Y 作为 X 中演奏的乐曲另一版本（即另一场演奏）的 CQT 表示。在这种情况下，两个序列 f（X）和 Y 在时间上并不对应，但可以在训练期间使用 SoftDTW 找到合适的对齐。我们使用 SWD 进行这项实验，该数据集提供同一乐曲的多个版本。具体而言，我们选择一个版本（OL06）作为目标版本，并使用 Soft DTW 训练网络，将其他版本的输入片段与 OL06 的片段对齐。最后，我们将训练期间未见过的版本输入训练好的网络，并使用余弦相似度与 OL06 的片段进行评估。作为无需学习的基线，我们还计算测试录音的原始 CQT 表示与 OL06 表示之间的 CS。为计算测试期间的余弦相似度，我们使用数据集提供的 OL06 与所有其他版本之间的真实对齐，但训练期间不需要真实对齐。

直接比较输入版本与目标版本的 CQT 表示，得到的平均余弦相似度为 0。576。使用 $\text{SoftDTW}_{\text{W3}}$ 进行训练后，在 CS $= 0.720$ 指标上取得了高得多的结果。因此，使用 SoftDTW 训练的网络能够生成与目标版本相似的实值输出。

6. 结论

本文将 SoftDTW 视为解决 MIR 中弱对齐学习问题的工具，尤其用于多音高估计。我们证明，使用 SoftDTW 训练的网络，其性能与使用先进的多标签CTC损失训练的同一网络相当。我们进一步证明，当目标序列包含实值条目时，也可以使用 SoftDTW 学习特征——而使用 CTC 则无法实现这一点。

在未来的工作中，SoftDTW 可应用于更多样化的音乐信息检索任务，例如歌词对齐、音频间同步，或从未对齐的视频-音频对中进行跨模态学习。此外，还可以探索在同一次训练中结合强对齐数据与非对齐数据的可能性。所有这些方案都由同一算法框架支持。

7. 参考文献

[1] Christof Weiß 和 Geoffroy Peeters，“利用多标签 CTC 损失从弱对齐乐谱-音频对中学习多音高估计”，发表于 *IEEE*音频与声学信号处理应用研讨会论文集（*WASPAA*），美国，纽约州纽帕尔茨，2021年，页码 121–125。[2] Juan J. Bosch、Rachel M. Bittner、Justin Salamon 和 Emilia G´omez，“基于源-滤波器建模的旋律提取方法比较”，发表于国际音乐信息检索学会会议论文集（*ISMIR*），美国纽约州纽约市，2016年，页码 571–577。[3] Ye Wang、Min-Yen Kan、Tin Lay Nwe、Arun Shenoy 和 Jun Yin，“Lyrically：声学音乐信号与文本歌词的自动同步”，发表于*ACM*国际多媒体会议论文集，美国纽约州纽约市，2004年，页码 212–219。[4] Kilian Schulze-Forster、Clement S. J. Doire、Ga¨el Richard 和 Roland Badeau，“音素级歌词对齐与文本辅助的歌声分离”，*IEEE/ACM*音频、语音与语言处理汇刊，第 29 卷，2395，2021 2382页 – 。[5] Simon Dixon 和 Gerhard Widmer，“MATCH：音乐对齐工具箱”，发表于国际音乐信息检索学会会议论文集（*ISMIR*），英国伦敦，2005年，页码 492–497。[6] Sebastian Ewert、Meinard M¨uller 和 Peter Grosche，“利用色度起始特征实现高分辨率音频同步”，发表于*IEEE*国际声学、语音与信号处理会议论文集（*ICASSP*），中国台湾台北，2009年4月，页码 1869–1872。[7] Shuhei Tsuchida、Satoru Fukayama、Masahiro Hamasaki 和 Masataka Goto，“AIST舞蹈视频数据库：用于舞蹈信息处理的多流派、多舞者、多摄像机数据库”，发表于国际音乐信息检索学会会议论文集（*ISMIR*），荷兰代尔夫特，2019年，页码 501–510。[8] Alex Graves、Santiago Fern´andez、Faustino J. Gomez 和 J¨urgen Schmidhuber，“连接主义时序分类：使用循环神经网络标注未分段序列数据”，发表于国际机器学习会议论文集（*ICML*），美国宾夕法尼亚州匹兹堡，2006年，页码 369–376。[9] Curtis Wigington、Brian L. Price 和 Scott Cohen，“多标签连接主义时序分类”，发表于国际文档分析与识别会议论文集（*ICDAR*），澳大利亚悉尼，2019年，页码 979–986。[10] Meinard M¨uller，音乐处理基础——使用 *Python* 和 *Jupyter*笔记本，施普林格出版社，第 2 版，2021年。[11] Marco Cuturi 和 Mathieu Blondel，“Soft-DTW：一种适用于时间序列的可微损失函数”，发表于国际机器学习会议论文集（*ICML*），澳大利亚新南威尔士州悉尼，2017年，页码 894–903。[12] Isma Hadji、Konstantinos G. Derpanis 和 Allan D. Jepson，“通过全局时序对齐与循环一致性进行表示学习”，发表于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），线上，2021年，页码 11068–11077。[13] Chien-Yi Chang、De-An Huang、Yanan Sui、Li Fei-Fei 和 Juan Carlos Niebles，“D3TW：用于弱监督动作对齐与分割的判别式可微动态时间规整”，发表于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），美国加利福尼亚州长滩，2019年，页码 3546–3555。

[14] Curtis Hawthorne、Erich Elsen、Jialin Song、AdamRoberts、Ian Simon、Colin Raffel、Jesse H. Engel、SageevOore 和Dou- glas Eck，“起始音与音符时值节拍框架：双目标钢琴转录”，载于国际音乐信息检索学会会议（*ISMIR*）论文集，法国巴黎，2018年，第50–57页。[15] Rainer Kelz、Matthias Dorfer、FilipKorzeniowski、Sebastian B¨ock、AndreasArzt和 Gerhard Widmer，“简单逐帧方法在钢琴转录中的潜力”，载于国际音乐信息检索学会会议（*ISMIR*）论文集，纽约市，纽约州，美国，2016年，第475–481页。[16] Kin Wai Cheuk、Yin-Jyun Luo、EmmanouilBenetos和DorienHerremans，“通过加性注意力重新审视起始音与音符时值节拍框架模型”，载于国际联合神经网络会议（*IJCNN*）论文集，中国深圳，2021年。[17]Ruchit Agrawal、DanielWolff 和 Simon Dixon，“用于结构感知演奏-乐谱同步的卷积注意力神经框架”，*IEEE*信号处理快报，第29卷，第344–348页，2021年。[18] Rachel M. Bittner、Brian McFee、Justin Salamon、PeterLi和 JuanP. Bello，“用于多声部音乐F0跟踪的深度显著性表示”，载于国际音乐信息检索学会会议（*ISMIR*）论文集，中国苏州，2017年，第63–70页。[19] Mehran Maghoumi、Eugene Matthew Taranta和Joseph LaViola，“DeepNAG：深度非对抗性手势生成”，载于智能用户界面国际会议（*IUI*）论文集，美国得克萨斯州College Station，2021年，第213–223页。[20] Graham E.Poliner和 Daniel P.W. Ellis，“用于多声部钢琴转录的判别模型”，*EURASIP*信号处理进展期刊，第2007卷，第1期，2007年。[21] John Thickstun、Za¨ıd Harchaoui 和 Sham M. Kakade，“从头学习音乐特征”，载于学习表征国际会议（*ICLR*）论文集，法国土伦，2017年。[22] ChristofWeiß、Frank Zalkow、VloraArifi-M¨uller、Meinard M¨uller、Hendrik Vincent Koops、Anja Volk和 Harald Gro- hganz，“《冬之旅》数据集：多模态音乐分析场景”，*ACM*计算与文化遗产期刊（*JOCCH*），第14卷，第2期，第25:1–18页，2021年。[23] Curtis Hawthorne、Andriy Stasyuk、Adam Roberts、Ian Si- mon、Cheng-Zhi Anna Huang、Sander Dieleman、Erich Elsen、Jesse H. Engel和Douglas Eck，“利用MAESTRO数据集实现因子化钢琴音乐建模与生成”，载于学习表征国际会议（*ICLR*）论文集，美国路易斯安那州新奥尔良，2019年。[24] Zhiyao Duan、BryanPardo 和 ChangshuiZhang，“通过建模频谱峰值与非峰值区域进行多重基频估计”，*IEEE*音频、语音与语言处理汇刊，第18卷，第8期，第2121–2133页，2010年。[25] JoachimFritsch和MarkD. Plumbley，“使用受约束非负矩阵分解与乐谱合成的乐谱信息辅助音频源分离”，载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*）论文集，加拿大温哥华，2013年5月，第888–891页。[26] Marius Miron、Julio J. Carabias-Orti、Juan J. Bosch、EmiliaG´omez和 Jordi Janer，“多通道管弦乐录音的乐谱信息辅助源分离”，电气与计算机工程期刊，第2016年卷，第8363507:1–8363507:19页，2016年。