引用格式：连志成，程皓楠，张加万.面向巴松演奏音乐的精准音频乐谱比对方法研究[J].中国传媒大学学报（自然科学版），2022，29(02)：54-62.
文章编号：1673-4793(2022)02-0054-09

面向巴松演奏音乐的精准音频乐谱比对方法研究

连志成[1]，程皓楠[2]，张加万[1*]

（1.天津大学智能与计算学部，天津 300350；
2.中国传媒大学媒体融合与传播国家重点实验室，北京 100024）

摘要：音频乐谱比对技术是一种将音频音乐与对应乐谱符号进行对齐的技术，是音乐信息检索（MIR）领域的重要研究方向。由于巴松的器乐、演奏、曲式特点，现有的音频乐谱比对方法无法精准处理巴松音频乐谱对齐任务。本文提出了一种面向巴松演奏的由粗到精、逐层细化的分段式高精度音频乐谱比对方法，针对巴松演奏音乐构建了首个由巴松独奏音频和对应乐谱组成的多曲式分类BSAMS（Bassoon-Solo-Audio-Midi-Score）数据集，并手工标注了音符起始时间和音符对应关系。具体来说，首先基于动态时间规整和音符起始点检测，设计了一种基准点和候选点生成算法，实现了音频乐谱对齐的粗略估计。其次，提出了一种基于支持向量机模型的音频乐谱点对筛选算法。最后，为了对精准匹配的对齐结果进行校验修正，提出了一种基于音乐理论的匹配修正算法，从而进一步提升了比对的准确度。在BSAMS数据集上对不同类型音乐进行实验，结果表明，本文提出的方法相比于传统通用音频乐谱比对方法可以达到在准确度上平均提高32.5%。
关键词：音乐信息检索；音频乐谱比对；巴松演奏音乐；精准对齐；分段式
中图分类号：TP391　　**文献标识码：**A

Research on accurate audio-to-score alignment method for bassoon music

LIAN Zhicheng[1], CHENG Haonan[2], ZHANG Jiawan[1*]

*(1. College of Intelligence and Computing, Tianjin University, Tianjin 300350, China;
2. State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China)*

Abstract: Audio-to-score alignment is a technology that aligns music audio with its corresponding score symbols, which is important in the field of music information retrieval (MIR) and has important practical significance for the development of music analysis and other fields. Due to the characteristics of bassoon′s instrument, performance, and musical forms, the existing audio-to-score alignment methods cannot accurately complete the task of bassoon′s audio-to-score alignment. To solve the above problems, we propose an accurate audio-to-score alignment method for bassoon music. First of all, we take the lead in constructing the BSAMS (Bassoon Solo Audio Midi Score) dataset composed of musical forms classified bassoon solo audio and corresponding scores for bassoon music, and manually annotate the onset of the notes and the correspondence of the audio and score. In order to achieve high-precision

基金项目：国家自然科学基金项目（62172295）；国家重点研发计划项目（2019YFC1521200）
作者简介(*为通讯作者)：连志成（1998-），男，硕士研究生，主要从事计算音乐研究。Email：lzc@tju.edu.cn；张加万（1976-），男，博士，教授，主要从事可视化、计算机图形学研究。Email：jwzhang@tju.edu.cn

audio-to-score alignment, based on the BSAMS dataset, we design a segmented accurate audio-to-score alignment method from coarse to fine. Specifically, based on Dynamic Time Warping and onset detection, a reference point and candidate point generation algorithm is designed to find a rough estimate of alignment. Secondly, an audio-score point pair screening algorithm based on the Support Vector Machine model is proposed to obtain accurate matching. Finally, a music theory based matching correction algorithm is designed to correct the alignment results. Experimental results on the BSAMS dataset demonstrate that the alignment accuracy increases by 32.5% on average compared with the traditional general audio-to-score alignment method.

Keywords: music information retrieval; Audio-to-Score alignment; bassoon music; accurate alignment; segmented

1　引言

音频乐谱比对是一种将音频信号与对应的乐谱符号进行对齐的方法，是音乐信息检索（Music Information Retrieval，MIR）领域的重要研究课题之一。随着数字音乐的发展，数字乐谱和器乐演奏音频的数量不断积累，建立数字乐谱和真实世界音乐演奏音频之间的对应和同步关系逐渐成为数字音乐发展的关键环节之一。

近年来，国内外研究人员在音乐演奏、音乐分析、音乐教育等领域展开了一系列音频乐谱比对方法技术的探索。针对不同乐器类型[1]、音乐形式[2]、性能要求[3]以及结构变化[4]，提出了多种音频乐谱比对方法。根据面向的器乐类型，现有方法可以分为面向通用乐器（或乐器组）演奏的音频乐谱比对方法和面向特定乐器（或乐器组）演奏的音频乐谱比对方法[5]。面向通用乐器的音频乐谱比对方法基于不同器乐演奏场景的音乐共性特征求解音频到乐谱符号对齐的过程[6, 7]。这类方法可以有效应对音乐演奏中固有的真实演奏音乐偏离乐谱的问题，但由于不同乐器的自身特点和演奏方式存在较大差异，面向通用乐器演奏的音频乐谱比对方法在处理特定乐器时往往存在低精度问题。

在面向特定乐器的音频乐谱比对方法中，早期研究人员对具有硬起音、易发音特点的乐器展开探索，在钢琴[8-12]、小提琴[1]等乐器的音频乐谱比对中已经取得较好的对齐结果。但是，针对巴松这类发音较难的软起音管乐器[13, 14]（如图1所示），如何构建精准音频乐谱比对方法，仍是这一领域亟待解决的难题。现有方法难以实现面向巴松的音频乐谱对齐高比对精准度，主要面临以下三方面困难与挑战：

图1　巴松结构示意图

（1）软起音、发音难等器乐特性。巴松发音主要为软起音，这导致音符的起始位置往往难以精准确定[15-18]，为音符级高精度音频乐谱对齐造成障碍。

（2）连音、吐音、颤音等丰富的演奏方式。多样化的演奏方式是导致巴松音频乐谱对齐困难的主要原因[19]，例如吐音导致的非预期静默片段和颤音导致的音符内频率周期性变化导致演奏音频偏离乐谱，从而提升了比对难度。

（3）缺乏曲式完备的巴松音频数据集。巴松演奏音乐的曲式多样，不同曲式的音乐存在速度、演奏方式上的较大差异。然而现有的巴松音频乐谱数据集相对匮乏，缺乏细致的曲式分类，导致相同方法在不同曲式中的比对结果精度存在较大差距。

针对上述问题，本文提出了一种由粗到精、逐层细化的分段式音频乐谱比对方法（如图2所示）。讨论顺序大致如下：第2章构造了首个由巴松独奏音频和对应乐谱组成的包含多曲式分类的BSAMS（Bassoon

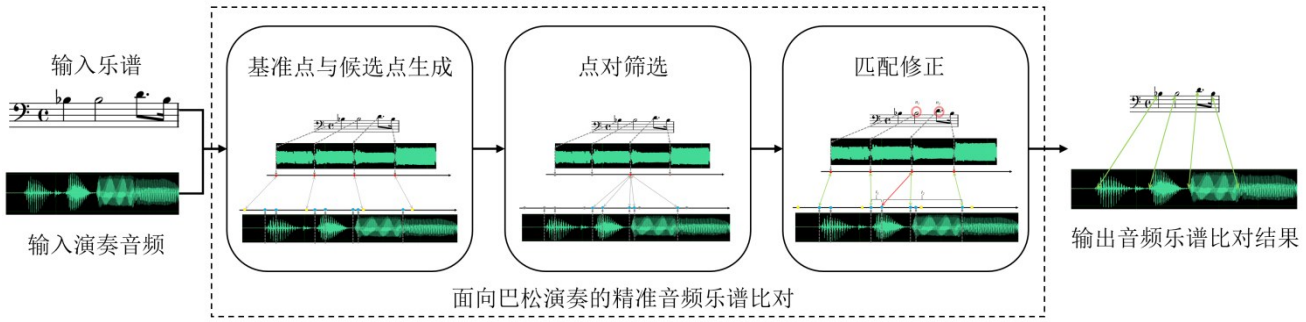


图2 算法整体流程

Solo Audio Midi Score)数据集;第3章提出了一种基于DTW(Dynamic Time Warping)的基准点与候选点生成算法,实现音符位置的粗略估计,设计了一种基于SVM(Support Vector Machine)的点对筛选算法,提高了音频乐谱在音符层次匹配的准确度;第4章 通过BSAMS数据集对不同类型音乐进行实验验证。

2 BSAMS数据集构建

针对巴松演奏数据集相对匮乏且类别不够全面的问题,本节设计并构建了巴松独奏音频和对应MIDI乐谱的BSAMS数据集。构建的巴松演奏数据集应满足两方面需求:(1)体现巴松演奏特点。由于巴松的曲目形式丰富,演奏方式多样,因此需要构建一个可以体现巴松演奏特点的数据集,同时区分曲目的形式和速度。(2)音频与乐谱音符精准对应标注。数据集用于音频乐谱比对方法的研究,要求该数据集的巴松独奏录音音频有精准的音符起始点标注以及和乐谱中的音符的对应关系。

为了满足第一个要求,首先通过对现有巴松演奏曲目的分析和整理,找到巴松演奏的五个重要的音乐片段曲式类别:练习片段、乐曲旋律片段、乐队片段和协奏曲片段。因为巴松在交响乐中使用场景居多,乐队片段占据相对最主要的部分,可以将乐队片段进行细分为:中国交响曲目的乐队片段和外国交响曲目的乐队片段。本文按照这六个类别划分,分别找到每个类别中具有代表性的巴松曲目演奏片段,并按照音乐片段的速度快慢和大致的演奏难度为每个片段进行了相应的标签标注,按照速度分为快速、中速和慢速,其次针对巴松的演奏方式,考虑到研究的重点为巴松的音符级研究,将标签分类为:连音、吐音、连音和吐音三种演奏方式,并将每个带有上述演奏方式的曲目分别进行相应的标签标注。

针对第二个需求,在乐谱方面采用MIDI编曲软件编写得到巴松MIDI乐谱。巴松演奏录音音频方面,录制音频的采样率为22050 Hz,录制音频格式为双声道WAV格式。音频乐谱对齐的标注方式为手工标注,首先标注音频中每个音符的起始位置,然后提取乐谱中的每个音符,将两者的一一相互对应关系记录于文件中。

具体来说,BSAMS数据集包含了18个不同的曲目片段和32对音频乐谱对,共计1118个音符。图3中的饼状图展示了BSAMS数据集中的音频乐谱对的曲目类型分布、速度分布、难度分布。综上,本文所构建数据集包含巴松独奏音阶琶音片段、乐曲旋律片段、

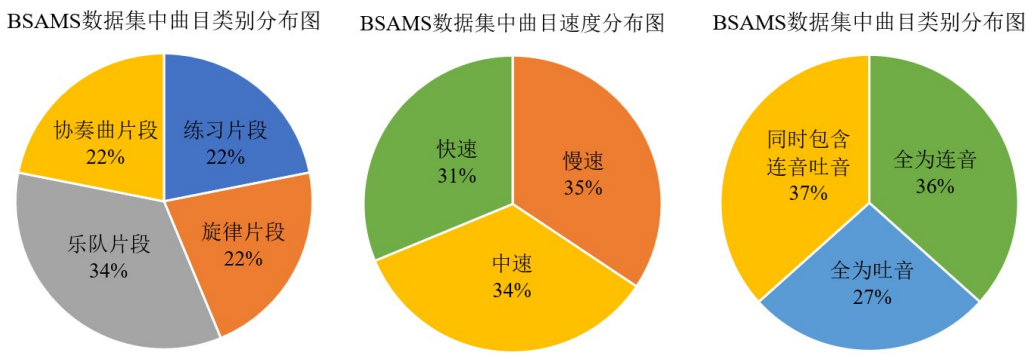


图3 BSAMS数据集曲目片段分布统计图

中国曲目乐队片段、外国曲目乐队片段和协奏曲片段，并具有速度、演奏方式标签以及音频乐谱精准对齐标注标签。本节构建的BSAMS数据集满足了体现巴松特点和用于音符级音频乐谱比对研究的要求，为后续展开的算法设计工作提供了良好的数据支撑。

3　精准音频乐谱比对方法

3.1　基于DTW的基准点生成算法

本节提出了一种基准点生成算法，将巴松演奏的音频和乐谱进行粗略的对齐，对于乐谱中的每一个音符，在对应录音音频中找到其粗略估计的音符起始位置。首先基于音频到音频对齐的思路，基于DTW算法找到MIDI转录音频和演奏音频之间的粗略对齐，以确定演奏音频中的粗略估计的音符起始位置作为基准点。

首先将MIDI转录合成为音频信号，同时基于MIDI协议标记出合成音频信号中每个音符的起始时间，基于DTW算法求解两段音频之间的时间点对匹配路径。具体算法流程如下：

（1）初始化MIDI合成音频的音符起始点时间序列 t_{midi}，每512个采样点取1帧为音符起始帧，得到音符起始点的帧序列集合 n_{midi}。

（2）经DTW算法得到对齐序列 p,q，该序列为MIDI合成音频和演奏音频之间的非递减帧序列，对给定的 $i \in \{1,\ldots,F\}$，$p[i]$ 与 $q[i]$ 形成匹配对，F 为音频帧序列的帧数。

（3）对于每个 $n_{midi}[k]$ 找到最大范围对应的 $[i_{k1},i_{k2}]$，使得 $p[i_{k1}]=n_{midi}[k]$ 且 $p[i_{k2}]=n_{midi}[k]$ 作为给定MIDI合成音频帧对应的匹配下标范围，其中 k 为给定乐谱中的音符数目。

（4）由每个MIDI音符起始点得到的匹配演奏音频中的帧范围 $(q[i_{k1}],q[i_{k2}])$，通过对应演奏音频中的时间点 $(t_{audio}[k1],t_{audio}[k2])$ 计算给定范围对应的时间轴中间点 $t_{base}[k]=(t_{audio}[k1]+t_{audio}[k2])/2$ 作为基准点，得到演奏音频的粗略估计音符起始点，即基准点序列 t_{base}。

为使该过程得到的基准点序列尽量准确，需要对DTW的参数基于BSAMS数据集进行优化，主要参数包括特征向量和向量距离计算函数。特征向量主要考虑梅尔倒谱系数特征、chroma_stft特征、chroma_cqt特征以及chroma_cens特征；向量距离计算主要考虑欧式距离和余弦距离。经实验，最终确定选取的特征为chroma_cqt色度特征。

本节得到的演奏音频中的音符起始基准点序列一方面作为粗略的音频乐谱对齐结果，另一方面为后续候选点的筛选范围提供参考。

3.2　基于起音检测的候选点生成算法

为了解决音频的精准音符起始位置问题，本节结合巴松自身的器乐特点，找到音符起始点的相对准确位置，为进一步精准对齐提供向后迭代所需数据。本节提出了一种基于音符起始点的检测算法，生成尽量靠近音符真实起始点的点位作为候选点。

考虑到巴松的音符演奏特征在一定程度上属于软起音，即音符起始位置的能量上升过程有较长且缓慢的能量上升过程，该过程相对于具有明显硬起音的钢琴而言较长，而相对软起音特点明显的小提琴等乐器较短，且演奏过程中有些音符可能会使用自然颤音的演奏方式，即在演奏某个音符时该音符会出现周期性的频率变化，因而需要采用适当的起音检测算法以提高检测的准确程度，抑制误检的发生。在起音检测算法中，基于能量的起音检测算法对硬起音效果较好，对软起音效果欠佳；结合相位的起音检测算法对软起音有一定的改进效果，但难以解决颤音的问题；基于频谱通量的起音检测可以有效应对软起音的问题，在一定程度上可以抑制颤音的误检。因而本文采取基于频谱通量的起音检测算法。为了适应巴松的音符起始特点，将起音检测得到函数曲线提取的峰值点，以及经过回溯得到峰值点附近的低点共同作为候选点。具体算法如下：

（1）首先基于以下公式计算音频特定频率成分能量变化的幅度即频谱通量（即谱波动）：

$$SF(n)=\sum_{m=1}^{m=M}H(X(n,m)-X(n-\mu,m)) \tag{1}$$

其中，n 为音频帧，m 为频域中两个离散谱线之间的间隔，X 为反映频率成分能量的函数。H 为半波整流函数，由以下公式得到：

$$H(x)=\frac{x+|x|}{2} \tag{2}$$

在实验中，选取参数 $\mu=1$，选取梅尔倒谱系数特征作为频谱成分能量的计算方式。

（2）通过峰值提取算法处理上一步得到的频谱通量曲线。峰值提取公式如下：

$$\begin{cases} SF(n) = \max(SF(n - pre\_max : n + post\_max)) \\ SF(n) \geq \text{mean}(SF(n - pre\_avg : n + post\_avg)) + \delta \\ n - n_{previous\_onset} > combination\_width \end{cases} \tag{3}$$

实验设定 pre_max 和 $post\_max$ 大小为 30 ms，pre_avg 为 100 ms，$post\_avg$ 取值为 70 ms，$combinition\_width$ 取值为 30 ms，其中 δ 为可调参数，实验选取 $\delta = 0.07$ 作为参数值，最终得到点集 $\{n_{peak}\}$。

(3)对点集 $\{n_{peak}\}$ 中的每个点，当 $n > 1$ 循环向前迭代 $n := n - 1$，若 $SF(n) > SF(n-1)$ 重复此循环，直至 $n = 1$ 或 $SF(n) \leq SF(n-1)$，得到点集 $\{n_{back-track}\}$。

(4)将点集 $\{n_{peak}\}$ 和 $\{n_{back\_track}\}$ 合并，得到集合 $\{n_{candidate}\} := \{n_{peak}\} \bigcup \{n_{back\_track}\}$。该集合为候选点帧集合，将候选点的帧转换为音频中对应的时间并将集中的点排序最终得到候选点的时间序列 $t_{candidate}$。

3.3 基于SVM的点对筛选算法

在得到巴松演奏音频中的基准点序列和候选点序列后，需进一步对得到的候选点序列进行筛选，从而得到更精确的音符起始位置。本节提出了一种基于SVM的点对筛选算法，首先计算得到乐谱中某一音符起始点和演奏音频中某一点的匹配置信度。然后，基于匹配置信度分别计算乐谱中每个音符起始点潜在匹配候选点，并依据置信度进行筛选。

首先，设计了基于支持向量机SVM模型的点对匹配相似度度量算法。主要步骤分为特征向量的构建、数据正负集构建和SVM模型的训练三部分。在特征向量的构建方面，总体上采取通过比对演奏音频中给定的点和MIDI合成音频中给定的点，以及附近小范围内的音频之间的相似性，以得到点对的相似性特征。

具体来说，对MIDI合成音频中的音符起始点，截取其附近的音频序列，同时截取演奏音频中候选点附近的音频序列，将两个序列进行预处理和比对，构建多个特征向量。DTW算法可以较好地度量两个给定时间序列之间的相似度，且不要求两条时间序列等长。基于以下平均DTW路径距离公式计算MDD(Mean DTW Distance)：

$$\text{MDD}(X,Y) = \frac{1}{L}\sum_{i=1}^{L}\left\| X[p[i]] - Y[q[i]] \right\|_2 \tag{4}$$

其中 X 和 Y 为音频帧序列，L 为匹配路径长度，p 和 q 为由公式(5)计算得到的匹配路径。

$$p,q = \arg\min_{p,q}\sum_{i=1}^{L}\left\| X[p[i]] - Y[q[i]] \right\|_2^2 + \Phi(i) \tag{5}$$

基于公式(4)，计算6类特征值，构建特征向量，具体计算如下：

$$\begin{cases} \text{MDD}(XM[n_1 - n_t : n_1 + n_t], XA[n_2 - n_t : n_2 + n_t]) \\ \text{MDD}(XA[n_2 - n_s : n_2], XA[n_2 : n_2 + n_s]) \\ \text{MDD}(XM[n_1 - n_{\text{rn}} : n_1], XA[n_2 - n_{al} : n_2]) \\ \text{MDD}(XM[n_1 : n_1 + n_{\text{rn}}], XA[n_2 : n_2 + n_{\text{ar}}]) \\ \text{MDD}(XM[n_1 - n_{\text{ln}} : n_1 + n_{\text{rn}}], XA[n_2 - n_{\text{al}} : n_2 + n_{\text{ar}}]) \\ \text{MDD}(\text{SC}(XM[n_1 - n_{\text{ln}} : n_1 + n_{\text{rn}}]), \text{SC}(XA[n_2 - n_{\text{al}} : n_2 + n_{\text{ar}}])) \end{cases} \tag{6}$$

其中 $XM[\,]$ 为MIDI合成音频的帧序列，$XA[\,]$ 为演奏音频的帧序列，n_1 为MIDI合成音频中待比对的音符起始点在音频中所在帧，$SC()$ 为截取音频中的非静默片段并拼接的函数，n_2 为演奏音频中待比对的点在音频中所在帧的序号，n_s 为固定短窗帧数，选取 $n_s = 5$，n_t 为固定长窗帧数，选取 $n_t = 10$，n_{ln} 为由MIDI协议获取合成音频中音符起始点以左的音符或休止符的时长对应帧数，n_{rn} 为音符起始点以右的音符的时长对应帧数，n_{al} 为由合成音频中 n_{ln} 按合成音频时长和演奏音频的时长的比例得到的近似帧数，即：

$$\begin{cases} n_{\text{al}} = \dfrac{\text{Len}(\text{Trim}(XA, l))}{\text{Len}(XM)} n_{\text{ln}} \\ n_{\text{ar}} = \dfrac{\text{Len}(\text{Trim}(XA, l))}{\text{Len}(XM)} n_{\text{rn}} \end{cases} \tag{7}$$

其中，Len为取音频长度的函数，Trim 为截去输入音频开头和结尾的静音片段的函数，l 为最高不超过的响度分贝值，取 $l = 20$dB。

在特征向量设计的过程中，需重点研究以下几方面内容：

(1)从音频帧的角度，设计特征比对MIDI中音符起始点和演奏音频中给定点之间的附近一个小区间(选取20帧，步长512，对应时间0.46s)的音频相似度，对应特征1。

(2)从音符模型的角度，设计特征比对音频中给定点左右的等长音频区间的相似度，以供参考该点处于音符中还是音符的端点(选取5帧，步长512，对应时间0.12s)，若该点处于音符中或静默片段则左右音频区间相似度较高。对应特征2。

(3)从音符的角度，设计特征比对临近音符的相似度。取MIDI合成音频和演奏音频左右两边相邻的音符长度，同时取对应演奏音频中相应的音频长度，

分别对比左音符（或休止符）、右音符、左右音符的相似度，分别对应特征3、4、5。

（4）结合巴松演奏多吐音的特点，对相邻音符音频片段做先删去静默片段再拼接的操作，以降低吐音造成的静音阶段在MIDI合成音频中无法对应造成的影响。同时将MIDI合成音频做同样操作以处理左音符为休止符的情况。

在得到特征向量后，进一步构造训练SVM的数据集。在时间轴上，将所有手工标注的音频乐谱点对作为正集，将与标注点相邻的两个点，以及标注点与左相邻点的中点、标注点与右相邻点的中点，共四个点与MIDI中音符起始点分别构成四组点对作为数据集的负集。

最后，采用SVM模型对得到的数据进行训练。SVM模型是一种二分类模型，其主要思想是找到数据空间中的一个可以将所有数据样本划开的超平面，并且使得样本集中所有数据到这个超平面的距离最短。具体来说，通过采用在空间中寻找间隔最大化的分离超平面的方式，对样本进行分类，同时通过样本点到超平面的距离可以反映其属于相应类别的概率。由于SVM在小样本训练集上能够得到比其它算法好很多的结果，因此采取SVM模型，模型的输入为提取到的特征向量，采用高斯核函数，并通过网格搜索来优化参数，模型输出为二分类：点对匹配或不匹配，并得到类别对应的概率作为置信度。

具体来说，模型的参数是基于网格搜索得到的，最终采取径向基核函数，设置参数惩罚系数为0.8，参数核函数系数为0.5，类别比重设置正负权重之比为3.8∶1，得到的SVM分类准确率为0.81。最终基于样本点到SVM模型决策超平面的距离得到概率值，用以判别音频和乐谱点对的匹配置信度。

在得到音频和乐谱点对的匹配置信度后，对每个乐谱中的音符找到对应演奏音频中一定范围内的候选点，通过SVM模型度量该音符在MIDI合成音频中音符的起始点和范围内的所有候选点形成的点对之间的匹配置信度，将得到的置信度由高到低排序，采用置信度最高的点对作为筛选结果。其中，演奏音频中的范围由基准点序列确定。具体算法流程如下：

（1）对每个音符的MIDI合成音频起始点$t_{midi}[k]$，通过其对应基准点$t_{base}[k]$，确定候选点的选取边界范围$(t_{base}[k-1], t_{base}[k+1])$。找到所有满足以下边界范围的$t_{base}[k-1] \le t_{candidate} \le t_{base}[k+1]$候选点，得到该音符用于匹配筛选的候选点集合$\{t_{candidate_match}\}$。

（2）根据训练的SVM模型分别计算候选点集合$\{t_{candidate_match}\}$中每个点与音符的MIDI合成音频起始点$t[k]$的匹配置信度。

（3）将置信度从大到小排序得到MIDI合成音频音符起始点$t_{midi}[k]$的最高置信度匹配点$t_{candidate_match}$，若该$t_{candidate_match}$点在与其进行匹配置信度计算的所有MIDI合成音频音符起始点中也有最高的匹配置信度，即双向最高匹配置信度，则该点即为的$t_{midi}[k]$的匹配点$t_{match}[k]$，若范围内无双向最高置信度的候选点或无候选点，则设置$t_{base}[k]$为$t_{midi}[k]$的匹配点$t_{match}[k]$。

（4）将所有筛选出的匹配点按照顺序排列得到匹配点序列t_{match}，与音符的MIDI合成音频起始点t_{midi}一起，构成音频乐谱匹配点对。

综上所述，本节为音频乐谱中点对精心构建了用于表征音频相似度的特征向量，并训练了SVM模型，将输出的匹配概率用于衡量置信度，得到了获取点对匹配置信度的算法。通过筛选出当前最佳的音频乐谱音符起始位置匹配点对，已得到较为精准的巴松音频乐谱对齐结果。但该结果仍旧存在一些问题，需要通过进一步的算法进行修正。

3.4 匹配修正算法

虽然已得到较精准的针对巴松演奏的音频乐谱对齐结果，然而该结果仍旧存在问题：（1）起音检测算法存在一定偏差，可能存在错检或漏检的现象，导致候选点集不够全面和准确，导致求得的匹配点存在偏差。（2）基于SVM模型得到的点对匹配置信度存在一定偏差，可能导致匹配失误的情况，导致求得的匹配点存在偏差。（3）巴松的演奏过程中存在部分音符发音困难的情况，在正确演奏的情况下也有可能出现音频和乐谱的偏离，导致求得的匹配点存在偏差。

针对以上三点问题，本节提出一种基于音乐规律的匹配修正算法。由于在音乐演奏实践中，大部分情况下临近音符之间的速度不会出现较大变化，因而临近的音符之间，音符起始点的时间差值之比可近似看作相邻音符之间时值之比。基于上述音乐特点，设计以下算法：

（1）对$1 < k < K$的点，按照以下公式计算得到每个点的λ值λ_k：

$$\lambda_k = \frac{t_{match}[k+1] - t_{match}[k]}{t_{match}[k] - t_{match}[k-1]} \cdot \frac{t_{midi}[k] - t_{midi}[k-1]}{t_{midi}[k+1] - t_{midi}[k]} \quad (8)$$

(2)理论上若无音乐节奏变化和其他偏差的理想状态,对所有k,λ_k值应等于1。当存在音乐节奏变化和上述偏差问题时,设定阈值$\lambda_{low} = .5$,$\lambda_{high} = 2$,若$\lambda_{low} < \lambda_k < \lambda_{high}$,则判断为正常,可获取每个超出正常范围的最长音频段。

(3)对每个非正常的音频段,截取MIDI合成音频$[t_{midi}[i-1], t_{midi}[j+1]]$段的音频,以及演奏音频$[t_{match}[i-1], t_{match}[j+1]]$段的音频,由以下公式得到其匹配点$t_{match}$:

$$t_{match}[k] = t_{match}[k-1] + (t_{midi}[k] - t_{midi}[k-1]) \cdot \frac{t_{match}[j+1] - t_{match}[i-1]}{t_{midi}[j+1] - t_{midi}[i-1]} \quad (9)$$

其余正常匹配点的匹配结果保持不变,得到最终的音频乐谱匹配点。

4 实验结果与分析

4.1 算法参数验证

为了验证基准点生成算法中DTW算法采用的主要特征,在BSAMS数据集上对色度特征和梅尔倒谱特征进行对比,最终选取基于CQT变换的色度特征chroma_cqt作为DTW算法提取的算法特征。实验结果如表1所示,对比不同特征的对齐准确率和平均每个音符对齐的时间偏差,采用chroma_cqt特征向量时,准确率最高,且平均音符时间偏差最小,因而采用chroma_cqt作为用于巴松音频乐谱粗对齐DTW算法采用的特征向量,以得到较准确的基准点和初步对齐结果。

表1 特征参数选取验证结果(容错0.1s)

特征	准确率	偏差(s)
chroma_stft	0.565	0.127
chroma_cqt	0.620	0.107
chroma_cens	0.531	0.124
mfcc	0.303	0.356

4.2 音频乐谱比对准确度对比

为了验证本方法的对齐准确度,从BSAMS数据集中每个曲目抽取一首音频乐谱对进行对齐测试,得到以下整体测试结果。表2中分别列出了传统DTW算法和本方法在BSAMS数据集上的实验结果。本方法相较传统DTW算法曲目对齐准确率在容错时间为0.1s情况下整体提升32.5%,平均时间偏差整体下降35.6%。

为进一步对比,对本文三个部分算法所得准确率和偏差分别进行实验,其中阶段一为基准点生成算法,阶段二为候选点筛选算法,阶段三为匹配修正算法。按照曲目所属类别(曲目类型、速度和演奏方式)进行归纳分析,如图4所示。首先根据曲目类型进行分析,如图4(a)所示,对每个曲目类型对应的音频乐谱对进行测试,得到各算法阶段平均准确度和音符平均时间偏差变化的折线图。可以发现本算法对各个类型曲目都有明显的对齐准确度和精准度提升,协奏曲片段整体由于难度大,音符类型、速度节奏变化多,效果不够理想,但对比第一阶段采用的通用对齐算法仍有一定提升。练习片段、中外乐队片段、旋律片段大体上可以实现较高的精准度。

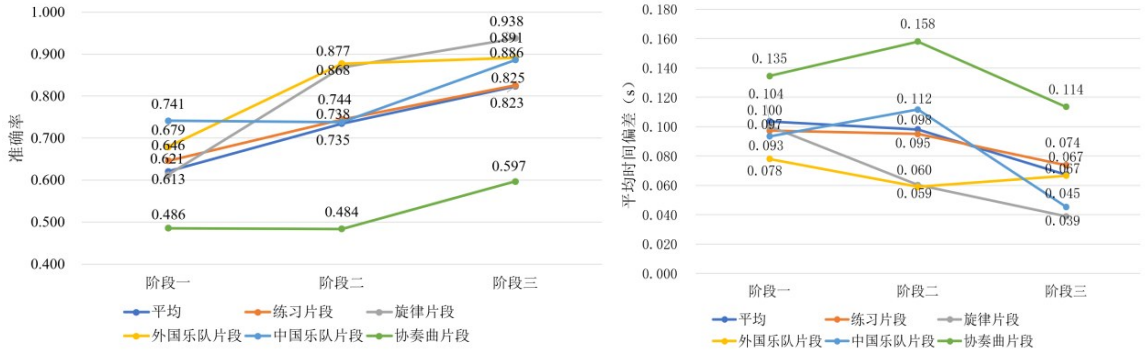
图4(b)展示了根据曲目速度分类的对齐准确率和时间偏差结果。可以观察到,对于慢速和中速的巴松独奏曲目片段,本算法有比较好的表现和效果,可以达到较高准确率。对于速度较快的曲目由于音符较为密集,效果不够理想。

表2 本文提出方法对比传统DTW算法对齐准确率比较

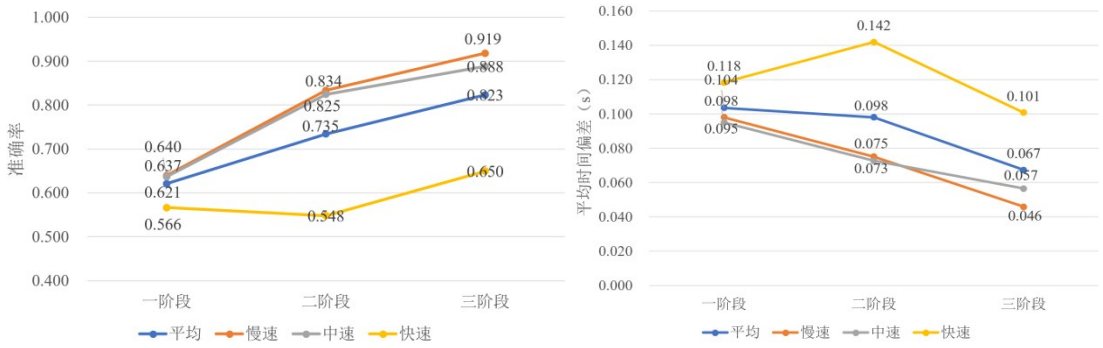
曲目片段	准确率			平均时间偏差(s)		
	DTW	本方法	提升率	DTW	本方法	下降率
曲目平均	0.621	0.823	32.5%	0.104	0.067	35.6%

针对不同演奏方式,如图4(c)所示,对每个曲目难度对应的音频乐谱对进行测试,得到各算法阶段平均准确度和音符平均时间偏差变化的折线图。可以看出本方法对连音演奏和吐音演奏的巴松音乐具有同样显著的提升效果。

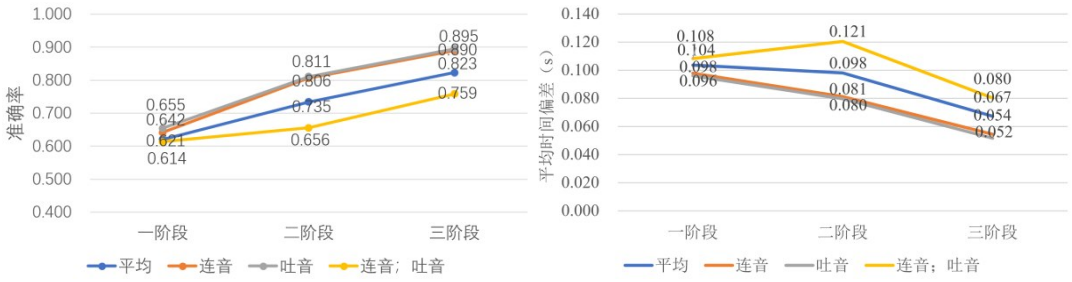
综上所述,本文提出的面向巴松的音频乐谱比对算法,与传统方法相比,对各种曲目类型的准确率和精准度均实现了大幅提升,在BSAMS数据集的中低速度和中低难度的曲目片段中,实现了音频乐谱的精准对齐。



(a)不同曲目类型比对结果准确率和时间偏差



(b)不同演奏速度比对结果准确率和时间偏差



(c)不同演奏方式比对结果准确率和时间偏差

图 4　BSAMS 数据集中不同曲目比对准确度结果

5　总结与展望

音频乐谱对齐是 MIR 领域的重要课题和基础任务，本文针对现有巴松演奏音频乐谱对齐方法精准度较低的问题，提出了一种面向巴松演奏的精准音频乐谱比对方法。构建了首个由巴松独奏音频和对应乐谱组成的包含多曲式的 BSAMS 数据集，手工标注了音符起始时间和音符对应关系。并基于 BSAMS 数据集，设计了一种由粗到精、逐层细化的分段式精准音频乐谱比对方法。首先基于 DTW 和音符起始点检测，设计了一种基准点和候选点生成算法，得到粗略估计的对齐；其次，提出了一种基于 SVM 模型的音频乐谱点对筛选算法，得到精准匹配的音符起始点；最后，设计了一种基于音乐理论的匹配修正算法，进行对齐结果的修正。通过在 BSAMS 数据集上对不同类型音乐进行实验，结果表明，本文提出的方法相比于传统通用音频乐谱比对方法在精准度上有显著提升。

未来的工作考虑以下三点内容：首先，当前算法在面对高难度快速巴松乐曲时，实现精准音频乐谱对齐仍旧较为困难，为提升此类型曲目的对齐准确度，需要设计更加具有针对性的方法。其次，各种乐器都有各自的乐器特性和演奏特色，基于各种乐器的自身属性设计更加合适的方法以提高准确度，是值得探索的研究方向。最后，考虑到音频乐谱比对技术的应用场景，基于面向巴松的音频乐谱比对方法，开发精准音符起始点标注系统，以为 MIR 领域的研究提供更为丰富的数据集。

参考文献(References)

[1] Syue J L, Su L, Lin Y J, et al. Accurate audio-to-score alignment for expressive violin recordings[C]. Proceedings of ISMIR, 2017: 250-256.

[2] Chen C, Jang J S R. An effective method for audio-to-score alignment using onsets and modified constant Q spectra [J]. Multimedia Tools and Applications, 2019, 78(2): 2017-2044.

[3] Arzt A, Widmer G, Dixon S. Automatic page turning for musicians via real-time machine listening[C]. Proceedings of ECAI, 2008: 241-245.

[4] Agrawal R, Wolff D, Dixon S. Structure-aware audio-to-Score alignment using progressively dilated convolutional neural networks[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.

[5] Niedermayer B. Accurate audio-to-score alignment: data acquisition in the context of computational musicology[D]. Linz:Johannes Kepler University, 2012.

[6] Niedermayer B, Widmer G. A multi-pass algorithm for accurate audio-to-score alignment [C]. Proceedings of ISMIR, 2010: 417-422.

[7] Raffel C, Ellis D P W. Optimizing DTW-based audio-to-MIDI alignment and matching[C]. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2016: 81-85.

[8] Arifi V, Clausen M, Kurth F, et al. Automatic synchronization of music data in score-, MIDI-and PCM-format[C]. Proceedings of ISMIR, 2003: 1-2.

[9] Sako S, Yamamoto R, Kitamura T. Ryry: A real-time score-following automatic accompaniment playback system capable of real performances with errors, repeats and jumps[C]. Proceedings of International Conference on Active Media Technology, 2014: 134-145.

[10] Müller M, Kurth F, Röder T. Towards an efficient algorithm for automatic score-to-audio synchronization[C]. Proceedings of ISMIR, 2004:1-8.

[11] Kwon T, Jeong D, Nam J. Audio-to-score alignment of piano music using RNN-based automatic music transcription [C]. Proceedings of Sound & Music Computing Conference, 2017:1-6.

[12] Agrawal R, Dixon S: Learning frame similarity using siamese networks for audio-to-score alignment[C]. Proceedings of EUSIPCO. 2020: 141-145.

[13] 郭轩. 巴松管的音乐表现与音响色彩探索[J]. 科技传播, 2010, 8: 44.

[14] 高佳思. 探究巴松管的演奏及其音乐表现[J]. 北方音乐, 2020(10):50-51.

[15] Böck S, Widmer G. Maximum filter vibrato suppression for onset detection[C]. Proceedings of Digital Audio Effects, 2013:7.

[16] Goto M. An audio-based real-time beat tracking system for music with or without drum-sounds[J]. Journal of New Music Research, 2001, 30(2): 159-171.

[17] Zhou R, Mattavelli M, Zoia G. Music onset detection based on resonator time frequency image[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2008, 16(8): 1685-1695.

[18] Bello J P, Sandler M. Phase-based note onset detection for music signals[C]. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003, 5: V-441.

[19] Lerch A. An introduction to audio content analysis: applications in signal processing and music informatics[M]. NJ, USA:Wiley-IEEE Press, 2012.

编辑:王谦,王雨田