文章编号：0427-7104(2020)05-0511-06

基于卷积循环神经网络的中国民族复音音乐的乐器活动检测

李子晋[1]，蒋超亚[2]，陈晓鸥[2]，马英浩[2]，韩宝强[1]
（1. 中国音乐学院 音乐学系，北京 100101；2. 北京大学 王选计算机研究所，北京 100080）

摘 要：针对中国民族复音音乐的乐器活动检测问题，提出了一种基于卷积循环神经网络(CRNN)的复音乐器活动检测方法，该方法属于事件检测类，在秒级的时间分辨率上识别乐器活跃的起止时间及乐器种类.同时，在中国音乐学院的 DCMI 数据库基础上，构建了 3 种不同的面向 10 种中国民族乐器的复音数据集进行训练和评估.通过实验，我们将 CRNN 模型与 CNN 模型进行了比较，验证了模型的特点和优势.

关键词：卷积循环神经网络；中国民族音乐；乐器活动检测
中图分类号：TP391 文献标志码：A
DOI:10.15943/j.cnki.fdxb-jns.2020.05.001

乐器识别是音乐信息检索(Music Information Retrieval，MIR)领域一个活跃的研究问题.乐器识别技术除了可直接应用于基于乐器的音乐浏览与检索，还在基于乐器的音乐转写、推荐、相似性计算、曲风识别、音源分离等领域有重要应用价值.已有的乐器识别方法大致可以分为两类：第 1 类是把乐器识别视为分类问题，其特点是数据样本的时长较短，如秒级，每个样本的内容及标签有单乐器单标签[1]、多乐器(复音)单标签(主乐器标签)[2]和多乐器多标签[3]3 种情况；第 2 类是把乐器识别视为音频事件检测问题，其特点是测试样本数据的时长较长，如几十秒或者更长，每个样本的标签，除了乐器类别以外，还可能涉及乐器出现的起止时间信息.这类乐器识别研究任务也被称为是乐器活动检测(Instrument Activity Detection，IAD)[4-6].本文的工作属于第 2 类.

已有的乐器识别研究大多面向西洋乐器，基于中国民族乐器的乐器识别的研究不多[7-8].本文在中国音乐学院的中国民族乐器数据库(Database of Chinese Musical Instruments，DCMI)的基础上[8]，构造了一个适用于进行乐器活动检测研究的数据集，并给出了一种采用卷积循环神经网络(Convolutional Recurrent Neural Network，CRNN)实现的中国民族音乐的乐器活动检测方法.本文解决乐器活动检测问题采用的策略是先将 IAD 任务简化为短时长分类任务，即将时长较长的数据样本分割成多个短时长的数据样本，然后训练一个针对短时长样本的分类器，通过分类器实现乐器类别的确定，然后通过后处理过程实现乐器出现时间的分段和起止时间的标定[5-6].

1 相关工作

Han 等[3]提出了一个在真实复音音乐中识别主乐器的卷积神经网络 ConvNets 框架，他们针对 11 种乐器，用 6 705 个含单主乐器标签的、时长 3 s 的定长音乐片段作为训练集训练网络，并用 2 874 个时长 5～20 s 含多个主乐器标签的可变长音乐文件作为测试集进行主乐器预测.预测结果为片段级的多乐器标签，且对主乐器的标签数量没有限制，但没有给出乐器出现的起止时间.多标签结果是通过聚合在测试音频上滑动的分析窗口所输出的多个测试结果得到的.他们一共实验了两种标签聚合的方法：一种是对每种乐器标签求均值；另一种是对乐器标签求和，然后进行归一化.

Liu 等[6]提出了一种全卷积神经网络模型，该模型能够在训练阶段仅使用片段级标注的数据集，就能

收稿日期：2020-01-09
基金项目：国家艺术基金(01020120180529564031)
作者简介：李子晋(1982—)，女，副教授，E-mail：zijin.li@mcgill.ca

对要测试的音乐片段进行精确的帧级标签预测,以进一步解决音乐中乐器出现事件的定位问题.针对识别 9 种乐器的训练和评估,他们使用了两个数据集: MagnaTagATune 用于片段级训练;MedleyDB 用于帧级评估.MagnaTagATune 是一个包含片段级标注的音乐数据集[9],它包括 25 863 个时长 29 s 的片段,用 188 个标签进行标注.MedleyDB 是一个多音轨乐器数据集[10],由 122 首歌曲组成,其中 105 首是全长歌曲,大部分歌曲的时长介于 3～5 min 之间.它具有 81 种乐器的帧级标注,且乐器出现的时间戳作为乐器起止时间的标签.

Gururani 等[5]给出了一个 IAD 的广义定义,即在细粒度时间尺度上检测某一音轨中乐器的存在或活动.同时,他们还提出了一种时间分辨率为秒级的复音音乐中多乐器活动检测的方法.他们用 3 种类型的深度神经网络(多层感知器(Multi-Layer Perceptron, MLP)、卷积神经网络(Convolution Neural Network, CNN)、卷积循环神经网络(CRNN))来训练识别模型,以识别 18 种乐器;用两个公开的多轨数据集——MedleyDB 和 Mixing Secrets[11]实现识别模型的训练和测试,使用这两个数据集主要是为了增加音频数据的规模.新版的 MedleyDB 数据集[10]包含 330 个音频文件,Mixing Secrets 数据集包含 258 个音频文件.将这两个数据集混合在一起,包含大约 100 种不同的乐器.混合数据集被分成一个训练集和一个测试集,其中训练集由 361 个音频文件组成,测试集由 100 个音频文件组成.

2 方 法

2.1 数据集

乐器活动检测数据集的构建通常有两种方式: 一是采用合成的方式[12];另一种是采用人工标注的方式,如以标注游戏软件为工具,通过众包方式完成标注,还可以通过多音轨格式的音乐数据提取标注数据[6].本文采用比较传统的合成方式.

本文用于训练和测试的复音数据集是在中国音乐学院的中国民族乐器数据库(DCMI)[13]的基础上线性混合生成的.DCMI 包含了 82 种不同类型民族乐器的单乐器音频数据,所有音高乐器都含有 3 个力度(强音、中音和弱音)的音阶音以及主要演奏技法和经典乐曲片段.本文选择了其中 10 种乐器(G 调新笛、箫、唢呐、二胡、古筝、柳琴、琵琶、三弦、扬琴、中阮)共计 5.46 h 的音频数据进行评估.

为了验证本文提出的模型的有效性,我们分别设计生成了 3 种复音数据集,下面分别简称数据集 1、数据集 2、数据集 3. 数据集 1 是由音阶音和演奏技法生成的不具有乐曲旋律的复音片段集;数据集 2 是根据经典乐曲片段生成的具有乐曲旋律的复音片段集;数据集 3 的测试集中的复音段只包含乐器的音阶音和演奏技法,而测试集中的复音段则是由乐曲片段生成.

通过在上述 3 种数据集上进行实验,分别验证模型在仅有音阶音和演奏技法的情况下乐器的检测效果、在含旋律的经典乐曲组成的复音段中的乐器的识别效果,以及在仅使用音阶音和演奏技法组成的复音段进行训练的情况下,模型对于由乐曲片段生成的复音段中的乐器的检测效果,以验证模型的鲁棒性.

对于数据集 1 和数据集 2,我们分别通过在 DCMI 中含有音阶音和演奏技法的单乐器数据库以及包含各种经典乐曲的单乐器数据库里随机选择乐器样本实例,并分别随机生成 3 000 个乐器复音片断,每个片断的时长为 6 min,共 50 h 的音频数据作为训练、评估的数据集.其中,重叠音最多控制在 4 种乐器,每种乐器持续时间至少 10 s,并且构成训练和评估的原始样本完全不重叠.数据集生成过程中记录的乐器类型及起止时间作为数据集的标签数据.

对于数据集 3,我们按照生成数据集 1 的方法随机生成 5 000 个乐器的复音片段作为训练集,按照生成数据集 2 的方法随机生成 1 000 个乐器的复音片段作为测试集,其中每个片断的时长为 6 min,共 100 h 的音频数据.同样,重叠音最多控制在 4 种乐器,每种乐器持续时间至少 10 s,并且构成训练和评估的原始样本完全不重叠.数据集生成过程中记录的乐器类型及起止时间作为数据集的标签数据.本文所用的原始音频段片段的统计信息如表 1 所示.

表 1 本文所用 DCMI 原始音频短片段的统计信息

Tab.1 The statistical information of DCMI original audio short segment used in this paper

乐器	片段个数/个	单个片段时长/s	总时长/s
二胡	62	1～352	2 616
G 调新笛	21	3～90	378
箫	17	1～49	303
唢呐	44	2～35	929
古筝	38	8～94	934
柳琴	46	10～89	1 402
琵琶	50	1～99	861
三弦	29	8～470	2 787
扬琴	14	5～595	2 987
中阮	38	8～1 051	6 453

2.2　算法

近年来，神经网络被广泛用于进行音频事件检测(Audio Event Detection，AED)。最早的 CRNN 是 1987 年由 Waibel 等[14]提出的时间延迟网络(Time Delay Neural Network，TDNN)。1989 年，Lecun 等[15]提出了用于图像分类的卷积神经网络 LeNet。目前 CNN 被广泛应用于音频事件检测领域，本文提出基于 CRNN 的方法识别复音音乐中的不同乐器，并与传统的 CNN 方法进行比较。

2015 年，Shi 等[16]提出 CRNN 用于光学字符识别(Optical Character Recognition，OCR)，其结构包含卷积层、循环层、转录层，他们提出使用卷积层提取图像的深层特征，同时应用循环层捕捉图像内部的序列信息，最后使用转录层将循环层输出转变为最终输出。

对乐器识别来说，由于音频本身就可视为一种时间序列，因此可使用循环神经网络(Recurrent Neural Network，RNN)捕捉其中的时序信息，由于 CNN 具有较好的信息提取和表达的能力，因此在对乐器识别时，先根据使用滤波器提取的原始频谱特征通过卷积层提取有效信息，然后再输入到循环层中，将循环层最后时刻的输出输入全连接层。本文中提出的 CRNN 将时长为 1 s 的 2 维音频频谱作为输入，通过多层卷积以及池化后输入两层循环层中，最后使用一层输出节点数为 10 的全连接层进行预测，模型结构如图 1 所示。此外，为与 CNN 结构进行对比，我们将 CRNN 中的两层循环层替换为全连接层并使用 Dropout 层避免过拟合，构造出 CNN 模型，模型结构如图 2 所示。

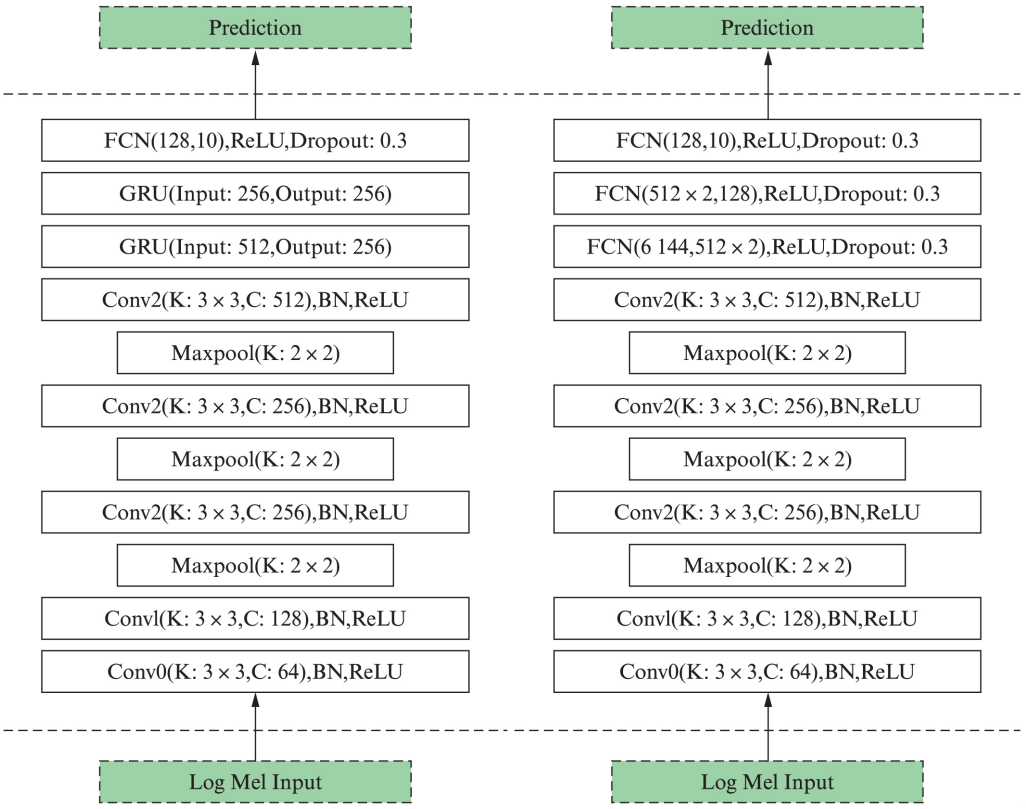

图 1　CRNN 模型的结构

Fig.1　The structure of CRNN model

图 2　CNN 模型的结构

Fig.2　The structure of CNN model

3　实　验

3.1　实验过程

我们将数据集中每个时长为 6 min 的复音段划分为 360 个时长为 1 s 的音频短帧，然后，将窗长设置为 1 024 个采样点，跳步大小设置为 512 个采样点，在每个窗口内提取 128 bin 的 Mel 谱并取对数，由于原始音频数据的采样率为 22 050 Hz，故 1 s 的短帧最终可得到 128×43 的 2 维对数 Mel 谱。

随后使用 Pytorch 分别实现了 CNN 和 CRNN 模型,训练过程中使用二进制交叉熵(Binary cross entropy)作为损失函数并采用 Adam 优化器进行优化,并将初始学习率设置为 0.001.

分别使用两种模型在上文提到的 3 种数据集上进行训练,并在对应的测试集上进行预测,将预测结果汇总后进行后处理,将属于同一复音段的帧级预测结果按时间顺序排列,得到模型的事件级预测.其中帧级预测结果即是模型对时长为 1 s 的短帧的分类预测结果,指出某一帧包含哪几种乐器声音,本实验中的每个短帧大多同时存在 4 种不同乐器的声音.而事件级预测的结果会给出复音段内不同乐器声音出现的起始以及结束的位置,如图 3 所示.

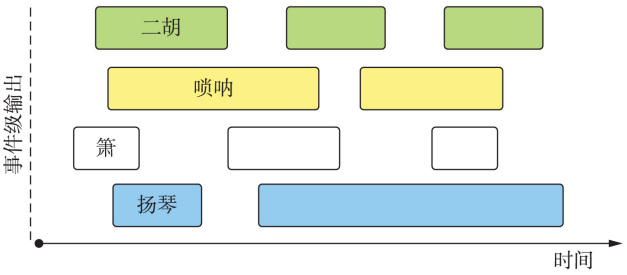

图 3 乐器活动检测中的事件级输出

Fig.3 The event-level output of instrument activity detection

图中每个方块代表一个事件级输出,每个事件输出都具有其起始位置、结束位置以及类别标签.

上述实验均是在 Linux 服务器(处理器为 Intel Xeon E5-2640v3, GPU 为 TITAN X)上进行的.

3.2 实验结果与评估

我们采用 F-measure 值,精确率(Precision)和召回率(Recall)作为评估模型的标准.对两种模型分别计算在不同数据集上的帧级预测结果和事件级预测结果.

精确率 $\lambda_{\text{Precision}}$ 的定义如下:

$$\lambda_{\text{Precision}}=\frac{k_{\text{TP}}}{k_{\text{TP}}+k_{\text{FP}}},$$

其中:TP(True Positive)值 k_{TP} 为将正类预测为正的个数;FP(False Positive)值 k_{FP} 为将负类预测为正的个数.精确率衡量了模型所有预测为正的结果中真正为正类的比例.

召回率 λ_{Recall} 的定义如下:

$$\lambda_{\text{Recall}}=\frac{k_{\text{TP}}}{k_{\text{TP}}+k_{\text{FN}}},$$

其中:FN(False Negative)值 k_{FN} 为将正类预测为负类的个数.召回率衡量了模型将正例预测为正的比例.

F-measure 的定义如下:

$$F1=\frac{2\times\lambda_{\text{Precision}}\times\lambda_{\text{Recall}}}{\lambda_{\text{Precision}}+\lambda_{\text{Recall}}}.$$

F-measure 综合考虑了精确率和召回率,显然 F-measure 越高说明实验方法越有效.

在评估帧级别预测时,对于某一帧来说,若模型在某一类的输出大于阈值 0.5,则判定该类存在.此时,帧级预测中的 k_{TP} 为所有正类样本中模型预测的输出 0.5 的样本的个数,k_{FP} 为所有负类样本中模型预测输出大于 0.5 的样本的个数,k_{FN} 为所有的模型输出小于 0.5 的样本中正样本的个数.而在评估事件级预测时,对于某一预测事件,要同时考虑其起始位置与真实出现的同类型乐器声音事件的起始位置的距离,以及该预测事件的结束位置与真实出现的同类型乐器声音事件的结束位置的距离,这里我们规定了预测事件的起始和终止位置与真实发生的同类别事件的起始和终止位置的误差接受范围为 2 s,即当预测事件的起始位置与对应真实事件的起始位置之差的绝对值不超过 2 s,则认为两者起始位置重叠.因此我们定义事件级预测中 k_{TP} 为在接受范围内预测事件和真实发生同类别事件相重叠的个数,k_{FP} 为在规定的误差范围内没有与之对应的同类型真实事件的预测事件的个数,k_{FN} 为在规定的误差范围内没有与之对应的同类型预测事件的真实事件的个数.

我们对两种模型分别在 3 种数据集上计算上述结果,其中帧级预测的评估结果见表 2,事件级预测的评估结果见表 3.

表 2　CRNN 与 CNN 在 3 种数据集上的帧级预测结果

Tab.2　The frame level prediction results of CRNN and CNN on three datasets

模型	$F1$/%			$\lambda_{Precision}$/%			λ_{Recall}/%		
	数据集 1	数据集 2	数据集 3	数据集 1	数据集 2	数据集 3	数据集 1	数据集 2	数据集 3
CNN	**91.96**	66.36	43.91	**97.19**	**97.10**	48.48	87.27	50.41	50.41
CRNN	91.64	**77.80**	**53.56**	96.42	88.50	**51.63**	**87.32**	**69.40**	**69.40**

注：表中加黑的数据是明确显示预测结果更好的模型.下同.

表 3　CRNN 与 CNN 在 3 种数据集上的事件级预测结果

Tab.3　The event level prediction results of CRNN and CNN on three datasets

模型	$F1$/%			$\lambda_{Precision}$/%			λ_{Recall}/%		
	数据集 1	数据集 2	数据集 3	数据集 1	数据集 2	数据集 3	数据集 1	数据集 2	数据集 3
CNN	80.10	44.12	28.67	74.90	45.56	22.45	**86.09**	42.77	27.25
CRNN	**81.46**	**57.17**	**36.19**	**75.88**	**54.85**	**25.60**	85.62	**59.69**	**36.69**

3.3　实验结果分析

通过实验我们发现：对于只包含乐器音阶音以及弹奏技法的数据集 1，不论是 CNN 还是 CRNN，它们的帧级识别的 F-measure 值均大于 91%，事件级的识别率同样也很高；对于使用不同乐器的经典乐曲片段构造的数据集 2，相较于 CNN 的 66.36%的 F-measure 值，CRNN 取得了 77.80%这一更好的效果；在我们构造的数据集 3 上，使用只包含乐器音阶音和乐器的演奏技法的数据进行训练，对不同乐器的经典乐曲段合成的复音段进行乐器活动检测时，CNN 的帧级预测的 F-measure 值为 43.91%，事件级预测的为 28.67%；CRNN 的帧级预测的 F-measure 值为 53.56%，事件级预测的为 36.19%.由于训练集的来源不同于测试集，因此不论 CNN 还是 CRNN 的效果都不够理想，但是我们依然可以发现 CRNN 具有更好的鲁棒性.

4　结　语

本文基于中国音乐学院的 DCMI 合成了 3 种不同的数据集，并基于此将本文提出的 CRNN 模型与传统 CNN 模型对比实验后，发现相较于 CNN，CRNN 能够更好地识别具有旋律信息的多乐器复音音频中的乐器，因此 CRNN 对于含旋律的音乐片段的乐曲识别检测更加有效且更具有鲁棒性，而从数据集 3 上的实验我们也发现简单旋律的数据集对复杂旋律样本的泛化能力不足，后续我们会继续尝试改进 CRNN 模型，提出更加有效且更具有泛化能力的乐器活动检测方法.

参考文献：

[1] HERRERA P, PEETERS G, DUBNOV S. Automatic classification of musical instrument sounds[J]. *Journal of New Music Research*, 2003, **32**(1): 3-21.

[2] FUHRMANN F, HARO M, HERRERA P. Scalability, generality and temporal aspects in automatic recognition of predominant musical instruments in polyphonic music[C]// Proceedings of the 10th International Society for Music Information Retrieval Conference. Kobe, Japan: ISMIR, 2009: 1-6.

[3] HAN Y, KIM J, LEE K. Deep convolutional neural networks for predominant instrument recognition in polyphonic music[J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017, **25**(1): 208-221.

[4] SELL G, MYSORE J G, CHON S H. Musical instrument detection detecting instrumentation in polyphonic musical signals on a frame-by-frame basis[R/OL]. (2006-12-15)[2019-05-22]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.142.8908&rep=rep1&type=pdf.

[5] GURURANI S, SUMMERS C, LERCH A. Instrument activity detection in polyphonic music using deep neural networks[C]// Proceedings of the 19th International Society for Music Information Retrieval

Conference. Paris, France: ISMIR,2018: 569-576.

[6] LIU J Y, YANG Y H. Event localization in music auto-tagging[C]// Advanced Cosmetic Multi-Media Meeting. Amsterdam, Netherlands: ACM,2016: 1048-1057.

[7] YU J, CHEN X O, YANG D S. Chinese folk musical instruments recognition in polyphonic music[C]// 2008 International Conference on Audio, Language and Image Processing(ICALIP). Shanghai, China: IEEE: 1145-1152.

[8] **沈骏,胡荷芬.中国民族乐器的特征值提取和分类**[J].**计算机与数字工程**,2012,**40**(9): 119-121.

[9] LAW E, WEST K, MANDEL M I, et al. Evaluation of algorithms using games: The case of music tagging[C]// Proceedings of the 10th International Society for Music Information Retrieval Conference. Kobe, Japan: ISMIR,2009: 387-392.

[10] BITTNER R M, SALAMON J, TIERNEY M, et al. Medleydb: A multitrack dataset for annotation-intensive mir research[C]// Proceedings of the 15th International Society for Music Information Retrieval Conference. Taipei, China Taiwan: ISMIR,2014: 155-160.

[11] GURURANI S, LERCH A. Mixing secrets: A multitrack dataset for instrument detection in polyphonic music[C]// Proceedings of the 18th International Society for Music Information Retrieval Conference. Suzhou, China: ISMIR,2017: 1.

[12] HEITTOLA T, KLAPURI A, VIRTANEN T. Musical instrument recognition in polyphonic audio using source-filter model for sound separation[C]// Proceedings of the 10th International Society for Music Information Retrieval Conference. Kobe, Japan: ISMIR, 2009: 1-5.

[13] LI Z J, LIANG X J, LIU J Y, et al. DCMI: A database of Chinese musical instruments[C]// PAGE K. Proceedings of the 5th International Conference on Digital Libraries for Musicology. Paris,France: ACM, 2018: 1-2.

[14] FUKUSHIMA K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. *Biological Cybernetics*, 1980,**36**(4): 193-202.

[15] LECUN Y, BOSER B E, DENKER J. Backpropagation applied to handwritten zip code recognition[J]. *Neural Computation*, 1989,**1**(4): 541-551.

[16] SHI B G, BAI X, YAO C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017,**39**(11): 2298-2304.

Instrument Activity Detection of China National Polyphonic Music Based on Convolutional Recurrent Neural Network

LI Zijin[1], JIANG Chaoya[2], CHEN Xiao'ou[2], MA Yinghao[2], HAN Baoqiang[1]

(1. *Musicology Department, China Conservatory of Music, Beijing* 100101, *China*;
2. *Wangxuan Institute of Computer Technology, Peking University, Beijing* 100080, *China*)

Abstract: Aiming at the instrument activity detection of Chinese national polyphonic music, a polyphonic musical instrument activity detection method based on Convolutional Recurrent Neural Network(CRNN) is proposed. This method is a kind of event detection method which identifies the starting and ending time of musical instruments and musical instrument types in the second temporal resolution. At the same time, based on the Database of China Conservatory of Music(DCMI), three different polyphonic instrument detection datasets which contain 10 kinds of China national instruments were constructed for training and evaluation. Through experiments, we compare the CRNN model with the CNN model, and verify the advantages of the CRNN model.

Keywords: convolutional recurrent neural network; Chinese national music; instrument activity detection