[1] Zijin LI, Chaoya JIANG, Xiaoou CHEN, Yinghao MA, Baoqiang HAN, Detection of Chinese Instrumental Quartet based on Convolutional Recurrent Neural Network, Conference on Sound and Music Technology in China 2019, Dec. 26-29, 2019, Haerbin, China

Abstraction: Instrument recognition is an active research problem in the field of music information retrieval (MIR). Musical instrument recognition technology can be directly applied to music browsing and retrieval based on Musical Instruments, but also has important application value in music transcribing, recommendation, similarity calculation, wind recognition, source separation and other fields based on Musical Instruments. The existing methods of instrument recognition are mainly in two categories: one is to regard instrument recognition as a classification issue, another as an event detection task. Aiming at the latter, this paper proposes a method of polyphonic instrument recognition based on Convolutional Recurrent Neural Network (CRNN), which can identify the active start and end times and the types of the instrument in terms of the temporal resolution of seconds. At the same time, on the basis of DCMI database of China Conservatory of Music, three different polyphonic instrument recognition data sets for 10 Chinese national instruments were constructed for training and evaluation. Through experiments, we compare CRNN model with CNN model, and verify the characteristics and advantages of the model. Finally, we analyze the characteristics and complexity of event detection based on instrument recognition.

基于卷积循环神经网的复音音乐中国民族乐器检测

李子晋 1[1]，蒋超亚 2[2]，陈晓鸥 3[3]，马英浩 4[4]，韩宝强 5[5]

（[1] 李子晋 1 中国音乐学院，北京 邮编 100000）

（[2] 蒋超亚 2 北京大学王选计算机研究所，北京 邮编 100000）

（[3] 陈晓鸥 3北京大学王选计算机研究所，北京 邮编 100000）

（[4] 马英浩 4 北京大学，北京 邮编 100000）

（[5] 韩宝强 5 中国音乐学院，北京 邮编 100000）

摘 要： 乐器识别一直是音乐信息检索（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，一类是把乐器识别视为分类问题，另一类是把乐器识别视为事件检测问题。针对后者，本文提出了一种基于卷积循环网络（CRNN）的复音乐器识别方法，该方法在秒级时间分辨率上，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DCMI数据库基础上，构建了三种不同的面向10种中国民族乐器的复音乐器识别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。最后我们分析了基于乐器识别的事件检测的特点和复杂性。

关键词： 卷积循环网络，乐器识别，音乐信息检索

中图分类号： TBC **文献标志码：** A

1 引言

乐器识别一直是音乐信息检索（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，第一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]三种情况；第二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（IAD）[4，5，6]。本文的工作属于第二类。

已有乐器识别工作大多面向西洋乐器，基于中国民族乐器的乐器识别研究工作不多[7，8]。本文在中国音乐学院的民族乐器数据库的基础上[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实现的中国民族乐器检测方法。本文解决乐器检测问题采用的策略是将先将IAD任务简化为短时长分类任务，先将时长较长的数据样本分割成多个短时长数据样本，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]。

本论文的组织结构如下。第2节为与本文相关工作的介绍，第3节为本文所构造的数据集及所提出的方法，第4节为实验及实验结果评估，第5节对本文进行总结和讨论。

2 相关工作

Yoonchang Han 等人[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6705个含单主乐器标签的、时长3秒的定长音乐片段作为训练集训练网络，并用2874个时长5-20秒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。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

Jen-Yu Liu 等人[6]提出了一种全卷积神经网络结构，该结构能够在训练阶段仅使用片段级标注的数据集，就能对测试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[17]，它包括25863个29秒的片段，用188个标签进行标注。MedleyDB是一个多音轨乐器数据集[18]，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到5分钟之间。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。

Siddharth Gururani 等人[5] 给出了一个乐器活动检测（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。同时提出了一种时间分辨率为1秒级至音轨级的复音音乐中多乐器活动检测的方法。他们用三种类型的深度神经网络（多层感知器MLP、卷积网CNN、卷积循环网CRNN）来训练检测模型，以检测18种乐器。两个公开的多轨数据集，MedleyDB和Mixing Secrets[16]用于检测模型的培训和测试。使用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[15]包含330个音频文件，Mixing Secrets数据集包含258个音频文件。这两个数据集结合在一起，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条音频文件组成，测试集由100条音频文件组成。

3 方法

3.1 数据集

乐器识别数据集的构建基本有两种方式，一是采用合成的方式[11]，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[]。还有一种方式是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。

本文的训练和测试的复音数据集是在中国音乐学院民族乐器数据库DCMI [12]基础上线性混合生成的。DCMI数据库包含了82种不同类型民族乐器的单乐器音频数据，所有音高乐器，都含有三个力度（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小时的数据进行评估。

为了验证本文提出的模型的有效性，我们分别设计生成了三种复音数据集，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。而数据集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。

我们通过在上述三种数据集上进行实验，分别验证我们的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检测效果，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有乐曲片段组成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3000个乐器复音片断，每个片断的时长为6分钟，共50小时的音频数据作为训练、评估的数据集。其中，重叠音最多控制在四种乐器，每种乐器持续时间至少10秒，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5000个乐器复音片段作为训练集，按照生成数据集2的方法随机生成1000个乐器复音片段作为测试集，其中的每个片断的时长为6分钟，共100小时的音频数据。同样，重叠音最多控制在四种乐器，每种乐器持续时间至少10秒，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。

表3.1 本文使用DCMI原始音频短片段的统计信息

Tab3.1 Statistical information of DCMI

乐器	片段个数	单个片段时长(s)	总时长（s）
二胡	62	1-352	2616
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	1402
琵琶	50	1-99	861
三弦	29	8-470	2787
扬琴	14	5-595	2987
中阮	18	8-1051	6453

3.2 算法

近年来，神经网络被广泛用于进行音频事件检测，本文提出使用卷积循环网络CRNN检测复音音乐中的不同乐器并与较为常用的卷积神经网络CNN相比较。最早的卷积神经网络是1987年由Alexander Waibel等提出的时间延迟网络（Time Delay Neural Network, TDNN）[9]。1989年，Yann LeCun提出用于图像分类的卷积神经网络LeNet [10]。目前CNN被广泛应用在音频事件检测领域，本文提出基于CRNN的方法将其应用在乐器检测领域，并将其与传统CNN方法进行比较。

CRNN(Convolutional Recurrent Neural Network)最早被提出时被应用于文本识别(OCR)领域，在2015年由BaoGuang Shi提出的CRNN结构[14]包含卷积层循环层，转录层，该文章中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。

对于乐器识别领域，由于音频本身就可视为一种时间序列，因此可使用循环网捕捉其中的时序信息，而由于卷积网具有较好的信息的取起和表达的能力，因此在对乐器识别时，先根据将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，然后在输入到循环层中，将循环层最后一个时刻的输出输入全连接层。本文中提出的CRNN结构将时长为1s的二维音频频谱作为输入，通过多层卷积以及池化后输入两层中循环层中，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3.2。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3.1。

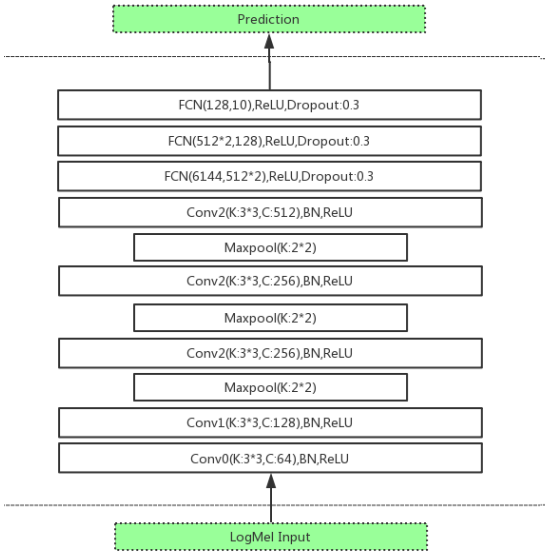

图3.1 CNN模型结构

Fig. 3.1:The structure of CNN

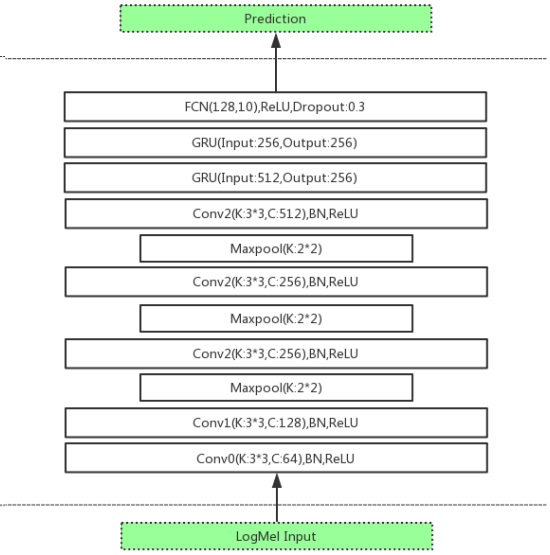

图3.2 CRNN模型结构

Fig. 3.2:The structure of CRNN

4　实验

4.1 实验细节

我们将数据集中的每个长为6分钟的复音段划分为360个时间长度为1秒的音频短帧，然后使用mel滤波器在窗长为1024跳步为512的采样点上提取128bin的mel谱并取对数，由于原始音频数据的采样率为22050，故1秒的短帧最终可得到128×43的二维LogMel谱。

随后使用Pytorch分别实现了CNN和CRNN两种不同模型，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，并将初始学习率设置为0.001。

分别使用两种模型在上文中提到的3种不同数据集上进行训练，并在对应测试集上进行预测，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别预测。其中帧级别的预测结果即是模型对于长为1s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对多同时存在4中不同乐器的声音。而事件级预测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，如下图。

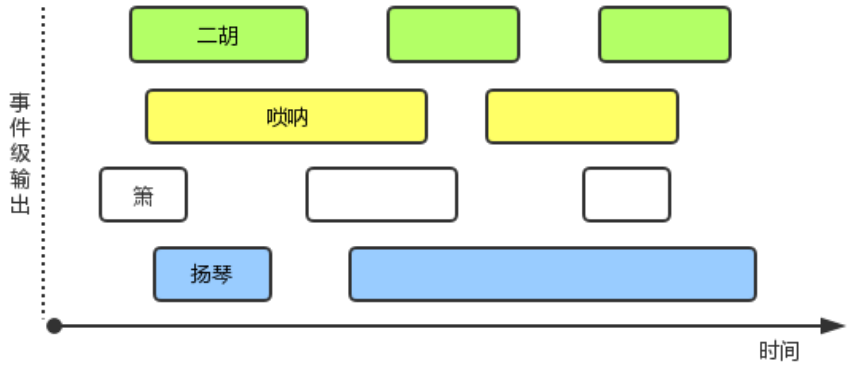

图4.1 乐器检测中的事件级输出，其中每个方块代表一个事件级输出，每个事件输出都具有其起始位置，结束位置，以及类别标签。

Fig. 4.1:The event-level output of the detection of instrument , each square represents an event-level output . Each event-level output has start time, end time, and its category tags

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

4.2 实验结果与评估

我们采用F-measure(F1)值，精确率(Precision)和召回率(Recall)作为评估模型的

标准。对于两种不同模型分别计算在不同数据集上帧级预测结果以及事件级预测结果。

对于精确率的定义如下：

$$\text{Precision} = \frac{TP}{TP + FP}$$

其中TP(True Positive)值为将正类预测为正的个数，FP(False Positive)值为将负类预测为正的个数。则精确率衡量了模型所有预测为正的结果中真正为正例的比例。

对于召回率的定义如下：

$$\text{Recall} = \frac{TP}{TP + FN}$$

其中FN(False Negative)值为将正类预测为负类的个数。则召回率衡量了模型对于正例预测为正的的比例。

对于F-measure的定义如下：

$$\text{F} - \text{measure} = \frac{2*Precision*Recall}{Precision+Recall}$$

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，对于某一帧若模型在某一类输出大于阈值0.5，则判定该类存在。则帧级预测中的TP值为所有正类样本中模型预测的输出0.5的样本的个数，FP值为所有负类样本中模型预测输出大于0.5的样本的个数，FN值为所有的模型输出小于0.5的样本中正样本的个数。而在评估事件级别预测时，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起始和终止位置的误差接受范围为2s，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2s，则认为两者起始位置重叠。因此我们定义事件级预测中TP值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。FP值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。FN值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。

我们对两种不同模型分别在三种数据集上计算上述结果，其中帧级预测的评估结果见表4.1，表4.2，表4.3，事件级预测的评估结果见表4.4，表4.5，表4.6。

表4.1 CRNN与CNN在数据集1上的帧级别预测结果

Tab4.1 The results of frame level on data set 1

	F-measure	Precision	Recall
CNN	**91.96 %**	**97.19 %**	87.27 %
CRNN	91.64 %	96.42 %	**87.32 %**

表4.2 CRNN与CNN在数据集2上的帧级别预测结果

Tab4.2 The results of frame level on data set 2

	F-measure	Precision	Recall
CNN	66.36 %	**97.10 %**	50.41 %
CRNN	**77.80 %**	88.50 %	**69.40 %**

表4.3 CRNN与CNN在数据集3上的帧级别预测结果

Tab4.3 The results of frame level on data set 3

	F-measure	Precision	Recall
CNN	43.91 %	48.48 %	40.13 %
CRNN	**53.56 %**	**51.63 %**	**49.53 %**

表4.4 CRNN与CNN在数据集1上的事件级别预测结果

Tab4.4 The results of event level on data set 1

	F-measure	Precision	Recall
CNN	80.10 %	74.90 %	**86.09 %**
CRNN	**81.46 %**	**75.88 %**	85.62 %

表4.5 CRNN与CNN在数据集2上的事件级别预测结果

Tab4.5 The results of event level on data set 2

	F-measure	Precision	Recall
CNN	44.12 %	45.56 %	42.77 %
CRNN	**57.17 %**	**54.85 %**	**59.69 %**

表4.6 CRNN与CNN在数据集3上的事件级别预测结果

Tab4.6 The results of event level on data set 3

	F-measure	Precision	Recall
CNN	28.67 %	22.45 %	27.25 %
CRNN	**36.19%**	**25.60 %**	**36.69 %**

4.3 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，他们的帧级别识的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。而在我们构造的第三种数据集上，使用只包含乐器音阶音和乐

器的演奏技法的数据训练，对不同乐器的经典乐曲段合成的复音段进行乐器识别时，CNN的帧级别预测的F-measure为43.91%，事件级F-measure为28.67%，CRNN的帧级别预测F-measure为53.56%，事件级预测的F-measure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都没有取得较好的效果，但是我们依然可以发现CRNN具有更好的鲁棒性。

5. 讨论

本文基于中国音乐学院民族乐器数据库DCMI合成三种不同数据集，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能够更好的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检测方法。

参考文献

[1] P. Herrera, G. Peeters, and S. Dubnov, “Automatic classification of musical instrument sounds,” J. New Music Res., vol. 32, no. 1, pp. 3–21, 2003.

[2] Ferdinand Fuhrmann, Mart´ın Haro, Perfecto Herrera, Scalability,Generality and Temporal Aspects in Automatic Recognition of Predominant Musical Instruments in Polyphonic Music, ISMIR 2009, October 26-30, 2009, Kobe, Japan

[3] Yoonchang Han, Jaehun Kim, and Kyogu Lee, Deep convolutional neural networks for predominant instrument recognition in polyphonic music, JOURNAL OF LATEX CLASS FILES, VOL. 14, NO. 8, MAY 2016

[4] Greg Sell, Gautham J. Mysore, Song Hui Chon, Musical Instrument Detection Detecting instrumentation in polyphonic musical signals on a frame-by-frame basis, December 15, 2006

[5] Siddharth Gururani, Cameron Summers, Alexander Lerch, Instrument Activity Detection in Polyphonic Music Using Deep Neural Networks, ISMIR 2018, September 23-27, 2018, Paris, France

[6] Jen-Yu Liu, Yi-Hsuan Yang, Event Localization in Music Auto-tagging, MM 2016, October 15-19, 2016, Amsterdam, Netherlands

[7] Jun Yu, Xiaoou Chen, Deshun Yang, Chinese Folk Musical Instruments Recognition in Polyphonic Music, IEEE ICALIP 2008, July 7-9, 2008, Shanghai, China

[8] 沈骏, 胡荷芬, 中国民族乐器的特征值提取和分类, 《计算机与数字工程》, 2012年第9期, 第119-121页

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

[10] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W. and Jackel, L.D., 1989. Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), pp.541-551.

[11] Toni Heittola, Anssi Klapuri and Tuomas Virtanen, Musical Instrument Recognition in Polyphonic Audio Using Source-Filter Model for Sound Separation, ISMIR 2009, October 26-30, 2009, Kobe, Japan

[12] Zijin Li, Xiaojing Liang, Jingyu Liu, Wei Li, Jiaxing Zhu, Baoqiang Han, DCMI: A Database of Chinese Musical Instruments, DLfM ’18, Sep 2018, Paris, France

[13] Long J , Shelhamer E , Darrell T . Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 39(4):640-651.

[14] B. Shi, X. Bai, and C. Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. TPAMI, 39(11):2298–2304, 2017

[15] Rachel M Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. Medleydb: A multitrack dataset for annotationintensive mir research. In Proc. of the International Society for Music Information Retrieval Conference (ISMIR), pages 155–160, 2014.

[16] Siddharth Gururani and Alexander Lerch. Mixing secrets: A multitrack dataset for instrument detection in polyphonic music. In Late Breaking Demo (Extended Abstract), Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Suzhou, 2017

[17] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie. Evaluation of algorithms using games: The case of music tagging. In Proc. ISMIR, 2009. http:// mirg.city.ac.uk/codeapps/the-magnatagatune-dataset.

[18] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello. Medleydb: A multitrack dataset for annotation-intensive mir research. In Proc. ISMIR, pages 155–160, 2014. http://medleydb.weebly.com