计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201200081

基于生成对抗网络去影像的多基频估计算法

黎思泉　万永菁　蒋翠玲
华东理工大学信息科学与工程学院　上海 200000
(siquan_li@163.com)

摘　要　多基频估计被广泛应用于音乐结构分析、乐音辅助教育、信息检索等各个领域。为了满足准确识别乐曲中随机和弦的需求,提出了基于生成对抗网络去影像的多基频估计算法。首先将完整音频切分成音符段,提出了一种谐音指纹图提取音符段频谱特征;然后通过卷积神经网络识别谐音指纹图当前的主导基频,将已识别出的主导基频作为干扰下一个基频识别的影像,并通过生成对抗网络去除干扰影像,对已去除干扰影像后的谐音指纹图进行新一轮的多基频估计;最后通过逐级迭代去影像操作实现完整和弦的多基频估计。对随机二音和弦及随机三音和弦组成的钢琴音频数据库进行实验,结果表明,所提算法与经典频谱迭代删除算法和大型词袋和弦识别算法相比,能够适应随机和弦的识别,在不同的音域范围内鲁棒性高,整体正确率有明显提升。

关键词:多基频估计;谐音指纹图;生成对抗网络;卷积神经网络;基频影像
中图法分类号　TP183

Multiple Fundamental Frequency Estimation Algorithm Based on Generative Adversarial Networks for Image Removal

LI Si-quan, WAN Yong-jing and JIANG Cui-ling
Department of Information Science and Engineering, East China University of Science and Technology, Shanghai 200000, China

Abstract　Multiple fundamental frequency estimation is widely used in music structure analysis, music aided education, information retrieval and other fields. In order to meet the requirements of accurate identification of random chords in music, a multiple fundamental frequency estimation algorithm based on generative adversarial networks is proposed. Firstly, the complete audio is divided into note segments, and a homophonic fingerprint is proposed to extract the spectrum characteristics of the note segment. Then, the current dominant fundamental frequency of the homophonic fingerprint is identified by convolution neural network, and the identified dominant fundamental frequency is considered as the image that interferes with the next fundamental frequency recognition. Then, the interference image is removed by generative adversarial networks, and the homophonic fingerprint image affected by interference is processed in a new round. Finally, the multiple fundamental frequency estimation of complete chords is realized by iterative de imaging operation step by step. Experiments on the piano audio database composed of random two tone chord and random three tone chord are carried out. The results show that, compared with the classical spectrum iterative deletion algorithm and the large vocabulary chord recognition algorithm, the algorithm in this paper can adapt to the recognition of random chords, has high robustness in different ranges, and improves the overall accuracy significantly.

Keywords　Multiple fundamental frequency estimation, Homophonic fingerprint, Generative adversarial networks, Convolution neural network, Fundamental frequency image

1　引言

多基频估计的任务是识别同时发声的所有音符,得出每个音符的基频频率值。多基频估计被广泛应用于钢琴音频检索、钢琴辅助教学等场景[1-3],多基频估计的研究能够深入音乐理论的了解,现有的多基频估计算法还不能满足实际的需求,随着人工智能的发展,采用深度学习的多基频估计算法成为了研究热点。

Humphrey 等[4]提出通过卷积神经网络[5-6](Convolutional Neural Network,CNN)进行和弦识别,实现了常用和弦的分类。这种识别方式突破了经典的多基频估计方法,针对基频的结构进行识别。Korzeniowski 等[7]对 Humphrey 等提出的算法作了进一步的改善,运用条件随机场(conditional random field)提取特征,提升了和弦的识别准确率,不足之处在于,其对非常用和弦的识别准确率非常低。为了提高非常用和弦的识别准确率,Zhang 等[8]利用残差网络结合随机森林算法进行音频识别。Deng 等[9]提出和弦大型词汇表(Large-Vocabulary,LV)技术,采用循环神经网络(Recurrent Neural Network,RNN)[10]进行和弦识别,可以分出更多和弦种类,但不能满足当今繁杂钢琴曲目中的更多非常规和弦的分类。在音频检索领域中占据重要地位的 SIMIR 会议中,Meseguer-brocal 等[11]提出引入控制机制的 U-Net 网络,

到稿日期:2020-12-08　返修日期:2021-05-06　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
通信作者:蒋翠玲(cuilingjiang@ecust.edu.cn)

进一步提高了和弦分类的准确率，Lieck 等[12]在琴键层次进行建模，使模型的音频分析有所提升。

Klapuri[13]提出的频谱迭代删除算法（Spectral Smoothness，SS）实现了对多种基频任意组合的识别，但其识别准确率有待提高，在 Klapuri 的基础上，Chen[14]采用基于单原子和多原子音符字典谱分解的方法进行主导基频的计算，避免了涉及不谐和因子的计算，但准确率仍较低。

随着生成对抗网络（Generative Adversarial Networks，GAN）的不断成熟，基于 GAN 网络的信号处理技术在音频处理领域也得到了广泛应用[15-18]。这些算法的特点是将音频频谱特征转换成图片，训练输入图片逼近目标图片，从而实现特定的音频处理任务。

本文提出了一种改进的基于生成对抗网络的多基频估计算法。本文算法训练卷积神经网络作为分类器，提取谐音指纹图中当前的主导基频。然后选择一种 GAN 网络——pix2pix 网络[19]作为修改器，去除谐音指纹图已提取的主导基频影像，并增强未提取基频信息，得到新的谐音指纹图后再次识别新的主导基频，通过分类器识别主导基频，通过修改器更新主导基频，从而识别出所有基频。

2 基于生成对抗网络去影像的多基频估计算法

本文算法基于频谱迭代删除思想，经过逐层去影像不断识别单个主导基频，从而实现多基频和弦的估计。图 1 是三音和弦的多基频估计流程图，将音符段处理为谐音指纹图，将谐音指纹图输入到分类器中，在识别到三音和弦的当前主导基频信息后，根据已识别的基频从单音模板库中调用对应单音模板作为去影像材料，与初始图片以及初始音频平均能量值一起作为一张 RGB 图片的 3 个通道，将已识别的基频信息作为冗余影像，将 RGB 图片输入到修改器得到新的谐音指纹图，再次识别新的主导基频，经过 3 次去影像后，图片能量低于阈值，实现了完整的多基频估计。图片能量是图片中所有像素值的和，去影像操作会大幅度减少图片能量，经过实验验证，图片能量低于原始图片能量的 40 dB 时，认为图片中不存在基频。

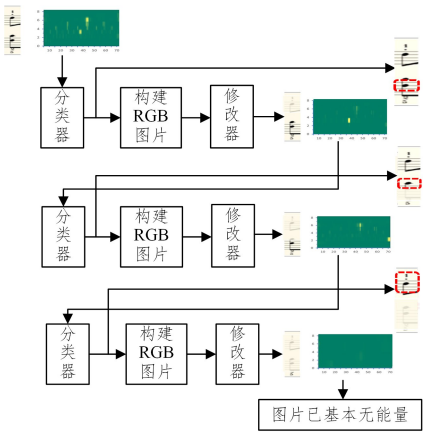

图 1 多基频估计流程图

Fig. 1 Flow chart of multiple fundamental frequency estimation

2.1 谐音指纹图

本文提出了以谐音结构为出发点的频谱提取算法，通过基频与泛音的关系排列位置关系，不考虑时域信息，通过频谱能量构建出谐音指纹图。如图 2 所示，对钢琴音频进行 onset 检测，根据检测结果将音频切分成音符段，然后将音符段转化到频域，构造每一个音符段的谐音指纹图。本文为了避免 onset 检测产生的误差，采用数据库标签以及人工核对的方式来保证 onset 的精准。谐音指纹图基于十二平均律来确定频谱区间的中心频率位置，88 个钢琴琴键分别对应一个基频，基频的位置即频谱区间中心频率位置。考虑到最高频琴键区的谐音频率会超过最高基频频率，因此需要对谐音指纹图的频谱区间向高频处进行拓展。本文基于十二平均律在钢琴的第 88 个琴键后再扩充 20 个虚拟琴键，扩充的 20 个虚拟琴键对应 20 个虚拟中心频率位置，由此实现频谱区间的拓展。拓展后，谐音指纹图共含有 108 个中心频率。为了提高图片的分辨率，基于十二平均律规则对 108 个中心频率进行插值，得到式(1)：

$$f_j = f_0 * 2^{\frac{\frac{j}{M}-49}{12}}, j=1,2,\cdots,108M \tag{1}$$

其中，M 是扩展分辨率的倍数，M 可以有效地在十二平均律的尺度上进行频谱分辨率的拓展。通过 f_j 划分频谱区间，求出每个区间对应的频谱幅值之和，如式(2)所示：

$$E_j = \sum_{f=(f_j-f_{j-1})/2}^{(f_{j+1}-f_j)/2} A(f), j=1,2,\cdots,108M \tag{2}$$

其中，$A(f)$是音频进行短时傅里叶变换得到的频谱值，E_j 是以 f_j 为中心的频谱区间的频谱幅值之和。对 E_j 的值进行最大值归一化，用色度卡表示，对 E_j 进行排列，区间中心频率 f_j 相差一个八度的纵向排列，中心区间频率在一个八度范围内的横向排列得到谐音指纹图。如图 3 所示，图 3 是图 2 中一个包含 a，g 和 d^2 的谐音指纹图。

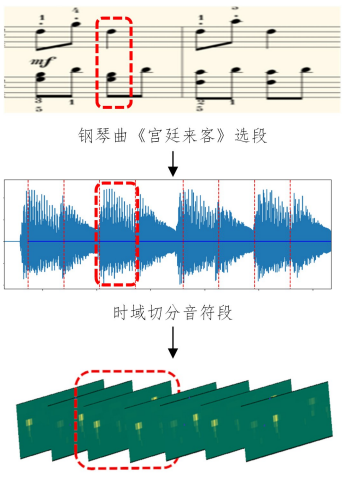

图 2 谐音指纹图的构造流程图

Fig. 2 Flow chart of generating homophonic fingerprint

图 3 谐音指纹图

Fig. 3 Homophonic fingerprint

2.2 谐音指纹图主导基频识别

本文提出的音符识别算法，通过逐步识别谐音指纹图中的主导基频，来实现所有的音符识别。分类器识别主导基频，分类器的结构基于经典卷积神经网络结构来实现，改进的卷积神经网络的结构如图 4 所示，输出层是 88 维的向量矩阵，分别对应钢琴的 88 个琴键基频信息。由于输入音频特征图片的差异性，谐音指纹图的分辨率较小，因此本文算法不直接沿用经典卷积神经网络的架构，而是采取固定池化层数量的方式，防止图片池化过度失去特征信息，只改变卷积层的层数和位置，卷积核的大小采用经典的 3×3。卷积层的层数越多，CNN 拟合的能力就越强，但过多的卷积层层数会导致过拟合、计算时间过长等问题，本文通过实验确定卷积层层数以及输入图片的分辨率大小，实验结果以及分析见 3.1 节，将识别准确率高于 95%的神经网络结构作为分类器。

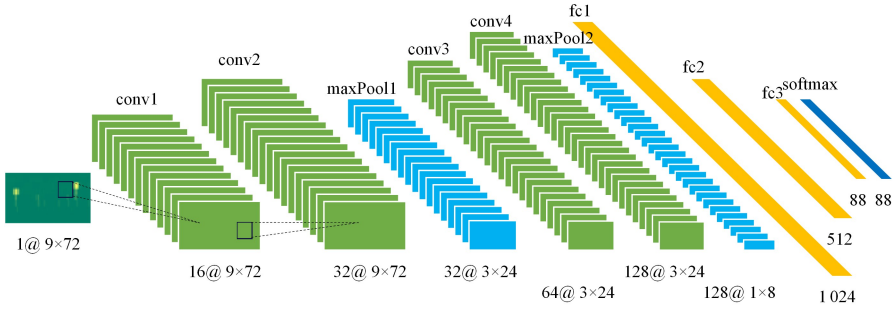

图 4 CNN 分类器的结构

Fig. 4 Structure of CNN classifier

2.3 谐音指纹图去影像

在识别到主导基频之后，主导基频以及对应的谐波信息成为图片中的冗余信息影像。修改器的作用是实现谐音指纹图去影像，并在此基础上增强剩余基频的信息。本文提出了一种基于 pix2pix 的网络修改器。pix2pix 改进条件生成对抗网络，把一维条件向量的输入调整为条件图片输入，将构造的条件图片与数据库中的真实图片进行训练。本文使用每个通道包含的不同信息的三通道 RGB 图片作为条件图片，运用 pix2pix 优秀的图片修改功能，修改输入的图片，实现消除已知基频并增强剩余基频的目的。条件图片的构造如图 5 所示。R 通道是待削弱基频的谐音指纹图，G 通道是从单音模板库中选择的谐音指纹图。单音模板库包含 88 张音频图片，对应 88 个主导基频。每个单音模板图片由 8 个相同单音谐音指纹图求和并取平均得到。单音模板作为 pix2pix 的条件输入参与去除影像的过程，是输入 pix2pix 网络的附加信息，对网络起提示作用，对 pix2pix 网络起决定作用的是训练数据库的组成，训练数据库中包含模板响度、持续时间等差别较大的音频，能够使 pix2pix 网络适应不同的钢琴音频。B 通道体现全局能量特征，取待修改谐音指纹图的能量平均值。

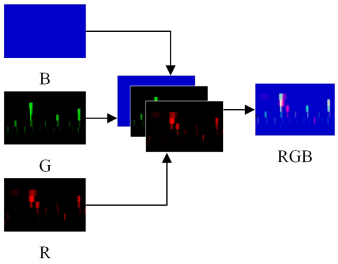

图 5 包含 3 种信息的三通道 RGB 图片

Fig. 5 Three channel RGB picture with three kinds of information

图 5 中亮度代表频谱区间的能量，所显示的亮点就是由 R 通道亮点和 G 通道亮点叠加生成的，直观显示了需要消除的主导基频的信息位置。

训练方式如图 6 所示，生成器和辨别器非同步训练，生成器训练时固定判别器的权值，判别器训练时固定生成器的权值。将三通道 RGB 图片作为生成器的输入，生成新的谐音指纹图，将真实弹奏的剩余基频信息图片作为真实图片，通过两组图片对 pix2pix 网络进行训练，实现在不同条件下对图片的修改。经过训练后，修改器的去影像效果如图 7 所示。原始音频的谐音指纹图经过修改器，不是简单地删减能量，修改器在去除已识别基频的基础上能够还原未识别基频的细节能量信息。

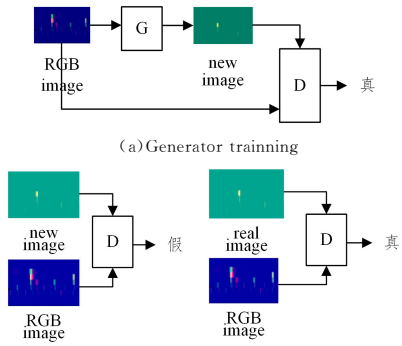

图 6 修改器的训练方式

Fig. 6 Modifier training method

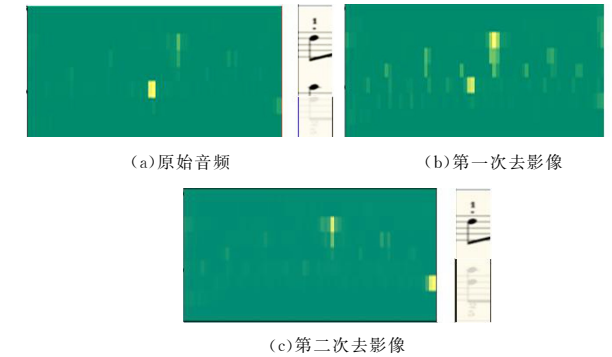

图 7 修改器去影像效果

Fig. 7 Image removal effect of modifier

3 实验结果

实验过程使用的硬件设备是 NVIDIA TITAN Xp 显卡(12GB 显存)以及 Intel Xeon CPU E5-2650 v4 处理器。实验分为 3 个部分:分类器识别准确率测试、修改器去影像能力测试以及整体准确率评估。选取 MAPS 数据集作为数据库,MAPS 由 Valentin Emiya 在 ParisTech 录制而成,包含单音符乐音、常用和弦、随机和弦[20]。根据琴键区分基频的范围如图 8 所示,钢琴将琴键划分为大字二组、大字一组、大字组、小字组、小字一组、小字二组、小字三组、小字四组、小字五组。不同分组的琴键基频结构区别较大:低音区的基频附带大量的谐波能量,基频能量缺失;中音区的基频较清晰;高音区半波分量明显。因此,低音区和高音区的识别难度更大。

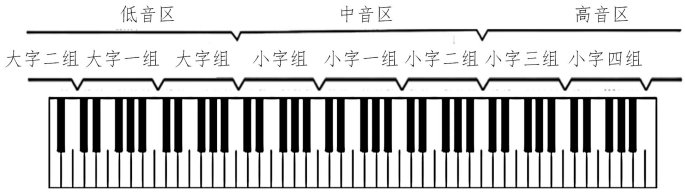

图 8 钢琴琴键分布图

Fig. 8 Piano key distribution

3.1 分类器识别准确率测试

本文在经典卷积神经网络架构的基础上,对神经网络卷积层和池化层的位置和层数进行调整,设计了 5 种不同的卷积神经网络结构,如表 1 所列,K 代表卷积层,括号中的参数是卷积核的张量维度,分别表示输入通道数、输出通道数、卷积核长、卷积核宽;P 代表池化层,括号中的参数代表池化窗口大小,结构中池化窗口的大小都是 3×3。

表 1 5 种不同的卷积神经网络结构

Table 1 Five different convolution neural network structures

struct1	struct2	struct3	struct4	struct5
K:(1,16,3,3)	K:(1,16,3,3)	K:(1,16,3,3)	K:(1,16,3,3)	K:(1,16,3,3)
P:(3,3)	K:(16,32,3,3)	P:(3,3)	K:(16,32,3,3)	K:(16,32,3,3)
K:(16,32,3,3)	P:(3,3)	K:(16,32,3,3)	P:(3,3)	P:(3,3)
P:(3,3)	K:(32,64,3,3)	K:(32,64,3,3)	K:(32,64,3,3)	K:(32,64,3,3)
W:(?,512)	P:(3,3)	P:(3,3)	K:(64,128,3,3)	K:(64,128,3,3)
W:(512,88)	W:(?,512)	W:(?,512)	P:(3,3)	K:(128,256,3,3)
	W:(512,88)	W:(512,88)	W:(?,512)	P:(3,3)
			W:(512,88)	W:(?,512)
				W:(512,88)

将不同分辨率的谐音指纹图输入到表 1 所列的卷积神经网络中进行训练,根据式(1)和式(2)调整谐音指纹图的分辨率。为了使识别和弦的单个基频更加精准,低音位置的基频会产生较多的谐波能量干扰,因此选取最小的琴键号作为主导基频的标签。选取 MAPS 数据集的 80%作为训练集,将剩下的 20%作为测试集,测试结果如图 9 所示。

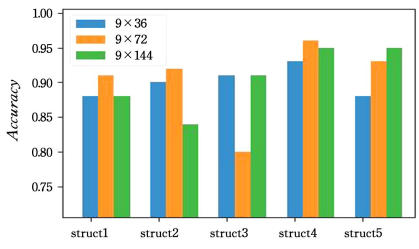

图 9 本文算法测试结果的柱状图

Fig. 9 Column chart of algorithm test results in this paper

结构 5 的网络深度和计算量都大于结构 4,在输入不同分辨率大小的图片时,结构 5 的准确率没有提升甚至有所降低,而结构 4 的运算量已经满足准确分类的需求。当分辨率为 9×72 时,结构 5 在测试集中存在过拟合,导致准确率降低。当分辨率为 9×144 时,准确率有微小的提升,但是计算量成倍增长,运算时间延长。考虑到实际应用中对运算量的要求,选定谐音指纹图的分辨率为 9×72,将结构 4 所示的卷积神经网络结构作为分类器结构。

3.2 修改器去影像能力测试

实验测试集采用 MAPS 数据库中包含 2 个基频或 3 个基频的一共 5 612 个音符段进行测试。钢琴琴键分区如图 8 所示,数据库中大字二组以及小字五组的数据较少,因此测试集的基频范围为大字一组到小字四组。

表 2 列出了去影像能力测试的实验结果,经典频谱迭代算法(Spectral Smoothness,SS)[13]不断识别并删除主导基频,因此选取该算法作为对比算法。包含多个基频的音符段首先消除已识别的一个基频,在图像上凸显出下一个暂未识别出的基频,然后用训练完毕的分类器来识别下一个暂未识别出的基频,以该识别正确率作为修改器去影像能力的评价指标。如表 2 所列,SS 算法的去影像能力在稳定性以及准确率上都远不及基于生成对抗网络去影像的修改器,SS 算法基于谐波结构来删除信息,其对参数优化的要求更高,在大字组区域的识别效果较差。基于生成对抗网络去影像的修改器实现了在消除已识别基频的同时,保留并丰富了原来的基频信息。钢琴音频大字组的谐波分量比较复杂,且基于生成对抗网络去影像的识别算法的自适应性能十分优异,因此在大字组区域也能有效消除基频影像,本文算法中的修改器也有较强的去影像能力。

表 2 经典迭代删除算法与本文算法的去影像能力对比

Table 2 Comparison of image removal ability between classical iterative deletion algorithm and this algorithm

	Algorithm	大字一组	大字组	小字组	小字一组	小字二组	小字三组	小字四组
二音和弦识别的第二个基频准确率	SS	0.47	0.50	0.67	0.77	0.71	0.62	0.58
	本文算法	**0.87**	**0.89**	**0.94**	**0.92**	**0.91**	**0.89**	**0.78**
三音和弦识别的第二个基频准确率	SS	0.44	0.53	0.60	0.72	0.68	0.65	0.58
	本文算法	**0.61**	**0.83**	**0.90**	**0.88**	**0.88**	**0.83**	**0.79**
三音和弦识别的第三个基频准确率	SS	0.40	0.49	0.59	0.68	0.66	0.55	0.51
	本文算法	**0.80**	**0.84**	**0.89**	**0.90**	**0.85**	**0.80**	**0.65**

3.3 识别准确率的比较

测试集采用 MAPS 数据库中的音频数据，包含常用的三音和弦、随机三音和弦、二音和弦和随机双音组合。音频段中最低音位置落在不同的音域范围，具有不同的识别难度，落在低音区的音频段有较多的谐音分量，对后面的基频有较大的干扰，落在高音区的基频谐波丢失，谐波结构不完整，因此将测试集中最低音的位置作为分类标准。

将测试集分为双音和三音两个子测试集进行测试，对比 3 种算法的正确率：文献[10]提出的基于大型词汇表(LV)和 RNN 的和弦识别算法、文献[13]提出的经典迭代删除算法(SS)和本文算法。本文基于经典迭代删除算法进行改进，因此选取经典迭代删除算法作为其中的一种对比算法。近年来，乐音识别研究中基于深度神经网络进行分类的算法一直走在前沿，文献[10]提出的一种大型词汇表技术有效提升了随机和弦的识别率，因此选取文献[10]提出的算法作为一种对比算法。

测试结果如表 3 和表 4 所列，三音的复杂度比双音大，三音测试集的评估参数整体比双音测试集低。LV+RNN 算法在三音识别的小字一组上的正确率较高，常用和弦在该分组中的占比较大，因此该分组能保持较高的正确率，而其他分组的正确率显著下降，整体准确率较低。本文算法在不同的音域都保持着较高的正确率，但识别三音准确率有所下降，相比 SS 算法和 LV+RNN 算法，本文算法的识别稳定性相对更高，面对测试集中较复杂的数据，也能够保持较高的识别正确率。

表 3 3 种算法的识别准确率对比

Table 3 Comparison of recognition accuracy of three algorithms

	Algorithm	大字一组	大字组	小字组	小字一组	小字二组	小字三组	小字四组
二音和弦测试集	SS	0.46	0.53	0.56	0.68	0.54	0.64	0.62
	LV+RNN	0.56	0.62	0.73	0.74	0.78	0.64	0.64
	本文算法	**0.60**	**0.73**	**0.87**	**0.85**	**0.83**	**0.79**	**0.72**
三音和弦测试集	SS	0.40	0.41	0.51	0.55	0.57	0.60	0.63
	LV+RNN	0.44	0.52	0.68	**0.80**	0.72	0.67	0.60
	本文算法	**0.52**	**0.68**	**0.81**	0.75	**0.73**	**0.70**	**0.65**

表 4 3 种算法的评估参数

Table 4 Evaluation parameters of three algorithms

	Algorithm	准确率	查全率	F1 分数
二音和弦测试集	SS	0.63	0.65	0.639
	LV+RNN	0.71	0.69	0.699
	本文算法	**0.84**	**0.87**	**0.854**
三音和弦测试集	SS	0.58	0.65	0.613
	LV+RNN	0.68	0.70	0.689
	本文算法	**0.79**	**0.81**	**0.799**

结束语 本文将生成对抗网络技术应用于钢琴音频的识别中，将包含多个基频信息的钢琴音频段转化成谐音指纹图，然后通过卷积神经网络构建的分类器识别单个音频，再运用基于 pix2pix 技术的生成器对音频的基频信息进行迭代删除，进而实现钢琴音频的识别。相比经典频谱迭代删除算法和大型词汇表和弦识别技术，本文方法能实现琴键号的精确识别，能有效识别随机和弦，具有更好的音符识别鲁棒性。实验测试结果表明，在三音和二音的识别中能有效提取出所有基频信息，鲁棒性高，实用性广。在后续的研究中会进一步优化神经网络结构，实现非成对数据的训练，减少构建数据库的工作量。

参 考 文 献

[1] SUN M. Applied research on music recognition technology[J]. Consumer Electronics, 2020(4): 62-63.

[2] CHEN Y W, LI K, HAN Y, et al. Musical Note Recognition of Musical Instruments Based on MFCC and Constant Q Transform[J]. Computer Science, 2020, 47(3): 149-155.

[2] LIU Y, ZHAO T Z, JIANG Y Q, et al. Improved piano music recognition algorithm based on autocorrelation function[J]. Journal of Wuhan University of Technology, 2018, 40(2): 208-213.

[3] WAN Y, WANG X L, ZHOU R H, et al. Piano multi note estimation algorithm based on spectral envelope nonnegative matrix decomposition[C]// Proceedings of the 5th Academic Exchange Meeting Commemorating the 50th Anniversary of the Institute of Acoustics, Chinese Academy of Sciences. 2014: 283-287.

[4] HUMPHREY E J, BELLO J P. Rethinking Automatic Chord Recognition with Convolutional Neural Networks[C]// International Conference on Machine Learning & Applications. IEEE, 2013.

[5] ALEX K, ILYA S, GEOFFREY E. ImageNet Classification with Deep Convolutional Neural Networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[6] QUAN Z. Convolutional Neural Networks[C]// The 3rd International Conference on Electromechanical Control Technology and Transportation. 2018: 434-439.

[7] KORZENIOWSKI F, WIDMER G. A Fully Convolutional Deep Auditory Model for Musical Chord Recognition[C]// Interna-

tional Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2016.

[8] ZHANG X L, PENG Y. Audio recognition method based on residual network and random forest[J]. Computer Engineering and Science, 2019, 41(4): 727-732.

[9] DENG J Q, KWOK Y K. Large vocabulary automatic chord estimation using bidirectional long short-term memory recurrent neural network with even chance training[J]. Journal of New Music Research, 2018, 47(1): 53-67.

[10] RAZVAN P, CAGLAR G, KYUNGHYUN C, et al. How to Construct Deep Recurrent Neural Networks[J]. arXiv: 1312.6026, 2014.

[11] MESEGUER-BROCAL G, PEETERS G. Conditioned-U-Net: Introducing a control mechanism in the U-Net for multiple source separations[J]. arXiv: 1907.01277, 2019.

[12] LIECK R, ROHRMEIER M. Modelling hierarchical key structure with pitch scapes[C]// Proceedings of the 21st International Society for Music Information Retrieval Conference. Montréal, Canada, 2020.

[13] KLAPURI A P. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness[J]. IEEE Transactions on Speech and Audio Proceessing, 2003, 11(6): 804-816.

[14] CHEN J. Research on multi fundamental frequency estimation of piano music[D]. Chengdu: University of Electronic Science and technology, 2016.

[15] YU L, WU H J, JIANG W K. Multi channel speech enhancement based on beamforming and Gan networks[J]. Noise and Vibration Control, 2018, 38(z1): 591-596.

[16] LIU H, LI Y, YUAN H Q, et al. Speech signal separation based on generated countermeasure network[J]. Computer Engineering, 2020, 46(1): 302-308.

[17] LI Y P, CAO P, SHI Y, et al. Speech conversion based on variational auto encoder and auxiliary classifier in non parallel text [J]. Fudan Journal (Natural Science Edition), 2020, 59(3): 322-329.

[18] CHENG X Y, XIE L, ZHU J X, et al. A review of generative countermeasure network Gan[J]. Computer Science, 2019, 46(3): 74-81.

[19] PHILLIP I, JUNYAN Z, TINGHUI Z, et al. Image-to-Image Translation with Conditional Adversarial Networks[J]. arXiv: 1611.07004, 2018.

[20] EMIYA V, BADEAU R, DAVID B. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(6): 1643-1654.

LI Si-quan, born in 1996, master. His main research interests include computer learning and audio analysis.

JIANG Cui-ling, born in 1976, Ph.D, lecturer. Her main research interests include artificial intelligence and image processing.

(责任编辑:柯颖)