Creative Education Studies 创新教育研究, 2022, 10(9), 2292-2299
Published Online September 2022 in Hans. http://www.hanspub.org/journal/ces
https://doi.org/10.12677/ces.2022.109361

Hans汉斯

钢琴音乐的组合频谱特征表示研究

胡丽敏[1]，桂　浩[2]，汤健雄[2]

[1]武汉音乐学院，湖北　武汉
[2]武汉大学计算机学院，湖北　武汉

收稿日期：2022年8月11日；录用日期：2022年9月19日；发布日期：2022年9月26日

摘　要

钢琴教育作为素质教育的代表性种类，日益普及。人工智能在语音识别领域有了全新的发展，钢琴教育也在将从中受益。使广大的钢琴学习者在人工智能的帮助下进行有指导的钢琴练习，是很有研究意义的问题。利用人工智能解决钢琴教育的智能化陪练问题，实际是将学习者演奏钢琴的音频信号转化成数字信号和真实的数字信号进行对比的过程，包括音级识别、自动音乐转录AMT。本文提出了一种组合频率和周期的多重特征表示方法作为音乐数据的特征表示，采用多特征的表示方法的识别效果往往优于单一频谱特征的表示。

关键词

人工智能，自动音乐转录，钢琴教育，组合频谱特征

Research on the Representation of Combined Spectrum Characteristics of Piano Music

Limin Hu[1], Hao Gui[2], Jianxiong Tang[2]

[1]Wuhan Conservatory of Music, Wuhan Hubei
[2]School of Computer Science, Wuhan University, Wuhan Hubei

Received: Aug. 11th, 2022; accepted: Sep. 19th, 2022; published: Sep. 26th, 2022

Abstract

As a kind of quality education, piano education is becoming more and more popular. Artificial intelligence has made progresses in the field of speech recognition, and piano education will also

文章引用: 胡丽敏, 桂浩, 汤健雄. 钢琴音乐的组合频谱特征表示研究[J]. 创新教育研究, 2022, 10(9): 2292-2299.
DOI: 10.12677/ces.2022.109361

benefit from it. It enables the majority of piano learners to carry out guided piano practice with the help of artificial intelligence. How to use artificial intelligence to solve the problem of intelligent accompaniment in piano education is actually a process of converting the audio signal of the learner playing the piano into a digital signal and comparing it with the real digital signal, including sound level recognition and automatic music transcription (AMT). In this paper, a multi-feature representation method of combined frequency and period is proposed as the feature representation of music data, and the recognition effect of multi-feature representation method is often better than that of single-spectrum feature representation.

Keywords

Artificial Intelligence, Automatic Music Transcription, Piano Education, Combined Spectrum Characteristics

1. 引言

随着素质教育的重视程度被提升，钢琴教育作为素质教育中的代表性的一个部分，在广大中小学生中普及开来。由于钢琴教育是一种特殊的技能教育，这决定了广大琴童仅仅是在有限的课程时间内受到专业的钢琴指导，而大多数的练习时间往往得不到专业的练习指导，这在某种程度上制约了钢琴音乐的普及。在倡导素质教育理念的大背景下，对于钢琴普及教育的受教方，即业余琴童及其家长，需要给他们更多的辅助性指导、陪伴性关怀。具体到日常钢琴练习的过程中，需要更好地解决业余学习者可能存在的一些问题，包括“识别错音”、“识别错节奏”、“规范性陪伴练习”等。目前人工智能在语音识别领域有了全新的发展，音乐教育也在将从中受益。使广大的钢琴学习者在人工智能的帮助下进行有指导的钢琴练习，是一个有很有研究意义的问题。利用人工智能解决钢琴教育的智能化陪练问题，实际是将学习者演奏钢琴的音频信号转化成数字信号和真实的数字信号进行对比的过程，这一类将声学音乐信号转化成普通的音乐符号或者是数字音乐信息的过程被称为自动音乐转录(Automatic Music Transcription, AMT)。

AMT 技术可以在人和音乐之间实现广泛的交互，包括音乐教育(例如，钢琴陪练系统)、音乐即兴创作(例如，口述即兴音乐想法和自动音乐伴奏)、音乐制作(例如，音乐内容可视化和基于内容的智能编辑)、音乐信息检索(例如，根据旋律、低音、节奏或和弦进行对音乐进行索引和推荐)和(音乐学例如，分析爵士乐即兴创作和其他非标记音乐)。本文的研究内容就是围绕着钢琴演奏音频中的钢琴音乐转录和多乐器演奏音频中的钢琴音乐转录展开的。

2. 钢琴乐理和音乐信号特征表示

音乐是由一组触动人类听觉器官的声音现象组成，究其本质是这些现象如何能够和谐地形成优美的组合。音乐的几大关键特质分别是音高、节奏、旋律、音阶等等。

2.1. 音乐特征

音阶(scale)，在音乐乐理中，音符按照基频或者音高的某种序列组成了音阶。如果是按音高的升序排

列则称为升序音阶，反之则成为降序音阶。节奏(rhythm)，节奏的概念与韵律、停顿和节拍相关，节奏是在一个连续的循环中，以相等的时间间隔排列的节拍或音符。旋律(melody)，旋律是音乐中的线性连续的音调，可以被看作单一的特性。从字面的意思来说，旋律是音高和节奏的组合。和弦(chord)，当三个或三个以上的音高相同的音符同时演奏时，就形成了和弦。和弦为音乐增加了音量，同时和弦也被当作音乐创作的基本元素。

2.2. 钢琴乐理及音乐特征分析

十二平均律(12 equal temperament, 12-TET)。平均音律是一种通过将一个八度划分为相同的步长来近似地表示音程的音乐音律或调律系统。这意味着在钢琴琴键上任何相邻的音符通过频率的对数方法表示时，所感知到的步长是相等的。在西方古典钢琴音乐中，最常见的音律系统就是十二平均音律。十二平均律将八度所有音级安装频率比相等原则分为 12 个部分，频率比为 $\sqrt[12]{2}\approx 1.05946$。这意味着通过这 12 个半音中的半音 d 的频率为 f_d，则 d 之后的第 k 个半音的基频可以表示为如下公式[1]：

$$f_n = f_d \times \left(\sqrt[12]{2}\right)^k \tag{2-1}$$

在自动音乐转录系统中，对音频信号的物理特征抽取是非常重要的一步，这会直接影响转录的效果。

钢琴的音高(pitch)是一种重要的感知属性，钢琴按键中的和与其相隔七个键的音符相同但是音高不同。音高的转录准确性直接影响了转录模型的好坏。钢琴的节奏(rhythm)是钢琴演奏好坏的重要评判标准，同时钢琴的节奏特性的转录也是 AMT 技术研究中的热点问题。钢琴的旋律(melody)是指钢琴音乐中的线性连续的音乐音调片段，旋律中包含了音高、节奏、音色等很多音乐元素。

钢琴区中往往包括大量的和弦，当三个或三个以上音符被同时演奏时，就形成了和弦。钢琴中的和弦有大三和弦、增三和弦、加音和弦等等很多种，一般是三和弦或者七和弦，和弦会呈现出一定的音程关系，一般会有保留根音和延留音，先现音两种使用手段，而和弦也会引起转录过程中音符基频之间的谐波干扰问题。

音符识别会受到按下钢琴键后能量逐渐衰退的影响，同时演奏钢琴的外界环境因素影响，这些问题都是钢琴音乐转录过程中的挑战。

2.3. 音乐信号特征表示

主流语音识别相关的研究首先要解决的问题就是语音特征的表示，即对音频信号中的有效信息予以保留，丢弃其他干扰识别的信息[1] [2] [3]。

- 傅利叶变换(FourierTransform, FT)是一种数字信号处理工具，它将音频信号转换成另一种表示形式，借助 FT 将基于时间能量的音频信号转化成基于频率和时间的频域波形图像。
- 梅尔倒谱(mel-frequency cepstrum, MFC)是基于对数频谱在非线性梅尔标度上的线性变换。将声音信号转化成适合人类感知的范围，与普通的倒谱分析不同的就是，梅尔标度上的频带是相等的，可以很好地适应人类的听觉系统反应。
- 恒定 Q 变换(Constant Q Transform, CQT)指将时域信号 $X[n]$转换为时频率表示的一种技术，使频率区域的中心频率以几何间隔排列，并且它们的 Q-factors (指中心频率与带宽之比)都是相等的。
- 其他音乐信号处理方法，还包括；短时傅里叶变换(STFT)、和谐常数 Q 变换(HCQT)等等。

但是，采用单一的频谱特征会出现单个帧级区域内的多个不同的音高的谐波峰值的干扰现象，这会导致识别错音的结果[4]。谐波在钢琴音乐中也称为泛音(源于琴弦振动理论)，其频率通常是基波的整数倍，与基音混合形成复合波，在形成独特钢琴音色的同时，为音级辨识带来噪声和干扰。

3. 钢琴音乐的组合频谱特征表示

在几种常见的数字信号处理方法的基础上，针对当前音乐转录存在的和弦之间的谐波干扰问题，本文提出了一种组合频率和周期的多重特征表示方法作为音乐数据的特征表示。而采用多特征的表示方法的识别效果往往优于单一频谱特征的表示。

首先提出了一种组合频率域特征和时间域特征的多特征数据表示方法，这种方法可以有效缓解传统频谱特征中出现的谐波干扰的情况。多音级估计(Multi-Pitch Estimation, MPE)问题中的音高信息是由音频数据中快速变化的部分决定的，而缓慢变化的部分并不能作为音高特征的代表，所以本文的多特征表示方法中增加了高通滤波器组将缓慢变化的部分进行了过滤移除。

基于组合频率和周期的多特征表示方法的钢琴音乐自动转录模型，模型结构如图 1 所示：

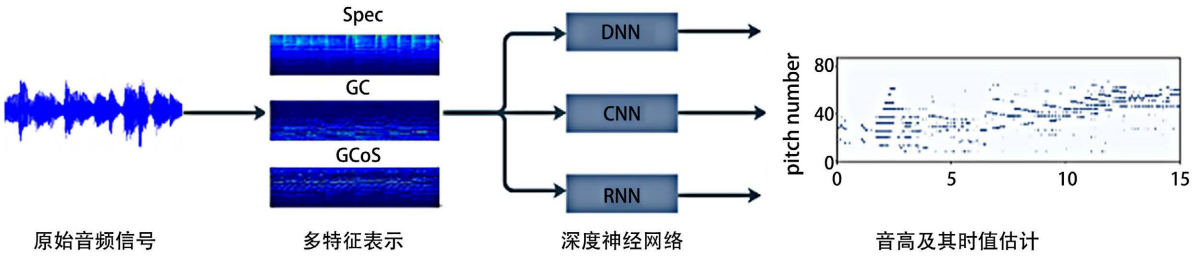

Figure 1. Multi-feature-based piano music transcription model
图 1. 多特征表示的钢琴音乐转录架构模型

模型分为两个阶段，数据特征表示和神经网络训练阶段。首先，模型的输入是钢琴演奏音频，输入信号经过预处理过程，分别提取原始音频信号的频谱特征(spectrogram，信号在时间域中的波形转变为频率域的频率分布曲线)、倒谱特征(generalized cepstrum (GC)，频谱图上周期性信号的单谱线表示)和频谱的广义倒谱特征(generalized cepstrum of spectrum (GCoS)，倒谱特征上的生成倒谱)等几种频谱表征信息[5]。第二阶段，通过卷积神经网络进行训练，输出每一帧的音符估值，所有的帧的结果整合最终得出一个完整的音高和时长信息。

音高作为音频信号中的一个对象，不能仅仅被频谱信号确定下来，还需要其他的非常规信息。音高可以被描述为频率、周期性与和谐性的组合，这意味着时频和周期的表示同样重要。此外，音高信息往往是每个数据表示中快速变化的部分，而缓慢变化的部分与音高无关。

在这种组合特征的表示方法中，可以认为某一时刻 t_0 频率为 f_0 的基音对象的一个音高可以由以下两种数据形式决定：时间频率 $Z(f, t)$和时间倒频率 $Z(q, t)$。因此，本文设计了一种数据驱动的建模方法来表示神经网络中的协调约束，这种方法将 $Z(f, t)$和 $Z(q, t)$作为数据输入。其中时间频率的表示形式有频谱图(spectrum)和频谱的广义倒谱图像(GCoS)，时间－倒频率的表示形式有倒谱图像(GC)。具体的计算方式如下，声学信号表示为 X，窗口函数表示为 h，x 是与时间 n 有关的量，表示为 $X[n]$。X 的短时傅里叶变换的振幅部分如下公式：

$$X\left[k,n\right]=\left|\sum_{m=0}^{N-1}X\left[m+nH\right]h\left[m\right]\mathrm{e}^{-\frac{j2\pi km}{N}}\right| \tag{3-1}$$

给定一个 N 阶的 DFT 矩阵 $\boldsymbol{F}$，高通滤波器 $\boldsymbol{W}_f$ 和 $\boldsymbol{W}_t$，以及非线性激活函数 σ，三种特征表示的公式如下：

$$\boldsymbol{Z}_0\left[k,n\right]=\sigma_0\left(\boldsymbol{W}_f\boldsymbol{X}\right) \tag{3-2}$$

$$\boldsymbol{Z}_1[q,n]=\sigma_1\left(\boldsymbol{W}_t\boldsymbol{F}^{-1}\boldsymbol{Z}_0\right) \tag{3-3}$$

$$\boldsymbol{Z}_2[k,n]=\sigma_2\left(\boldsymbol{W}_f\boldsymbol{F}\boldsymbol{Z}_1\right) \tag{3-4}$$

为了适应音高的检测尺度，将上述的频率和倒频率表示采用对数的标度表示，通过一组滤波器来完成：

$$\boldsymbol{Z}_0=Q_f\left|\boldsymbol{Z}_0\right| \tag{3-5}$$

$$\boldsymbol{Z}_1=Q_t\left|\boldsymbol{Z}_1\right| \tag{3-6}$$

$$\boldsymbol{Z}_2=Q_f\left|\boldsymbol{Z}_2\right| \tag{3-7}$$

Q_f和 Q_q是将功率谱图和倒谱图的横轴的频率和倒频转换成对数表示的滤波器组，产生的对数(倒)频率特征相应的被称为 $\boldsymbol{Z}_0,\boldsymbol{Z}_1$ 和 $\boldsymbol{Z}_2$。如图 2 所示，为 MAPS 数据集中的 ENSTDkAm 目录下的其中一首歌的节选片段的对应的三种特征表示的情况。

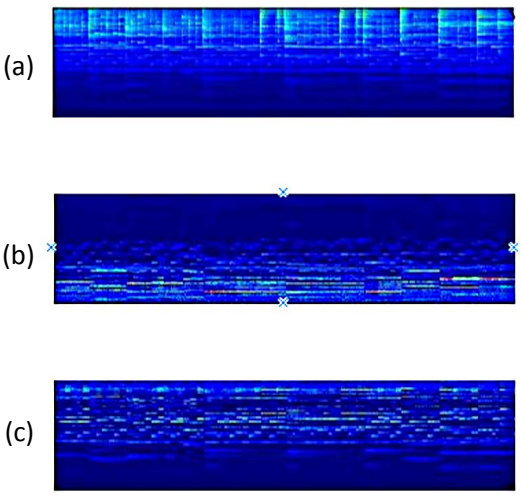

Figure 2. Three features of music data representation (Spec, GC, GCoS)
图 2.音乐数据表示的三种特征(Spec, GC, GCoS)

最后将 $\boldsymbol{Z}_0$、$\boldsymbol{Z}_1$、$\boldsymbol{Z}_2$ 三种特征通过特定的方式组合输入深度神经网络中进行训练，这就是组合频率和周期性的多特征表示方式。音频信号被转化成三种组合频率和周期的图像输入到深度神经网络中进行训练，三种特征图像的高度确定，长度不定。本章后面的内容针对多重特征的表示方法分别设计了基于 CNN 和 DNN 的钢琴音乐转录模型，在 MAPS 上表现效果比普通的频谱特征的效果要好。

4. 多音级估计深度学习模型设计

多特征表示方法实际是从多个维度提取出的特征频谱或倒谱图像，在卷积神经网络[6] [7]中一般是采用的二维的卷积核，由于多特征的输入是多通道的，本文在在二维卷积的基础上设计了多通道的三维卷积，将三种不同的特征组合作为输入。从单个特征的二维视角来看，定义其中的频谱二维图像 $X\in\mathbb{R}^{M\times N}$ 和一个二维的滤波器 $W\in\mathbb{R}^{m\times n}$，单个图像的卷积运算公式为：

$$y_{ij}=\sum_{u=1}^{m}\sum_{v=1}^{n}w_{uv}\cdot x_{i-u+1,j-v+1} \tag{4-1}$$

要利用卷积核捕捉到特征图像中的各种音乐特征，就要知道卷积核的形状对各个特征的捕捉效果，卷积操作对特征图像的应用的时间尺度和空间尺度。特征图像的横向是代表时间或者与时间有关的物理

量，卷积核的宽度决定了对时间特征的捕捉效果。所以，本文将特征图像的尺寸设为 $M*N$，卷积核的尺寸设为 $m*n$。M 和 m 代表音频信号的能量，N 和 n 代表时间帧数。钢琴音频识别采用小型卷积核($m \ll M$ ，$n \ll N$)进行建模，因为钢琴音频的特征在于频率的变化时间不会很长，卷积核可以捕捉到小时间尺度的音高和时值信息(图 3)。

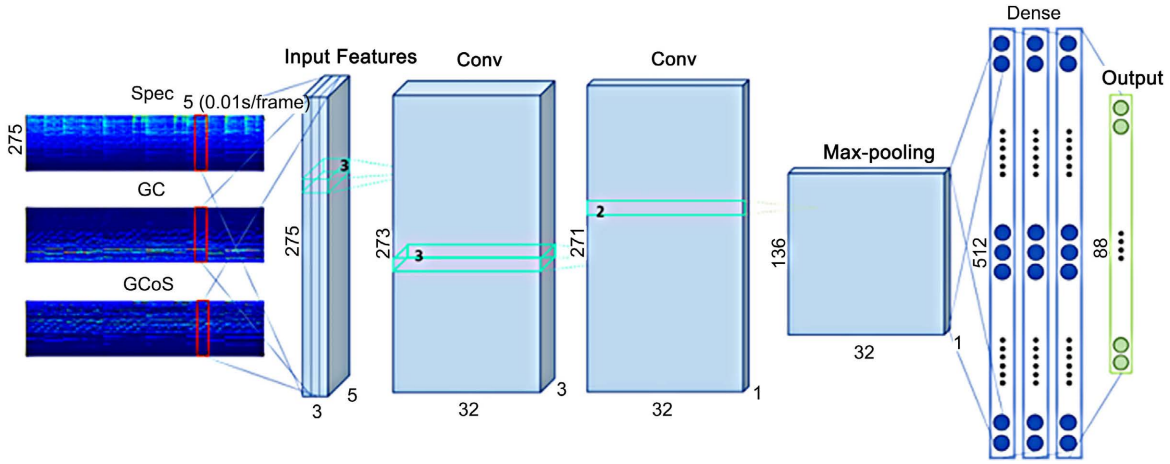

Figure 3. Multi-feature-based deep neural network
图 3. 基于多特征表示的卷积神经网络结构

本文采用指数线性单元[8] (scaled exponential linear units, SELU)作为激活函数，采用 SELU 作为激活函数可以在没有对数据进行批量归一化的时候进行训练。为了避免过度拟合，本文使用了 dropout 和 early-stop。初始网络参数为高斯分布，均值为 0，std 为 0.05。输出为 88 × 1 的二值钢琴卷，其中值位 1 表示这一帧含有此音高。输出层采用 sigmoid 函数建模，通过最小化输出的二进制交叉熵和地面真值进行优化。网络的输出是具有相同维数的向量；每个 bin 的值为激活音高的可能性，取值范围为[0, 1]，从 0.5 处阈值进行二值预测：

$$\overline{y}_t = y_t\big|_{y_t[i]>0.5} \tag{4-2}$$

5. 实验与结果分析

5.1. 实验数据集和环境

实验采用了单钢琴乐数据集 MAPS。本文中采用 MAPS 数据集主要是为了研究组合频谱特征的表示，由于更专注于实时演奏的音乐信息，因此本文采用了 MAPS 中的音乐片段集(MUS)部分内容。MUS 中每一首音乐都有规范的命名 MAPS_MUS_description_instrName.wav，对齐的 MIDI 标注与其对应的音乐文件名相同。

操作系统为 Ubuntu16.04，处理器为 Intel Xeon(R)CPU E5-2683 v3，频率为 2 GHz。实验采用了 GPU 加速，GPU 版本为 NVIDIAGTX1080Ti，显存大小为 11 GB，显存频率 11 MHz，实验过程中使用显存 10,997 MB。代码运行环境为 Python 3.7，Tensorflow1.13，CUDA9.0。

5.2. 实验评价标准

音级转录是一个多分类的问题评价。对于多标签多分类问题，可在时间维度上采用 N 个二分类器对其进行评价。在二分类中，问题的实际结果被定义为正(Positive)、负(Nagative)两类，模型的预测结果也被定义为正(True)、负(False)两类。

5.3. 实验过程和结果分析

本实验环节的首要目的就是验证所提出的组合频谱特征在既定的音乐转录模型中的转录效果是否比单一的频谱特征表现要好，其次就是验证组合频谱中，哪种组合形式的转录效果最优。在单一频谱特征中，实验只采取一种频谱图进行输入，相当于单通道实验，本次实验中涉及单一特征除了前文提到的 Z_0 (Spectrum)，Z_1 (Generalized Cepstrum)和 Z_2 (Generalized Cepstrum Of Spectrum)之外，还使用了主流的 CQT 变换得出的频谱特征。组合频谱特征采用多通道的形式表示，组合频谱包括[Z_0, Z_1]，[Z_0, Z_2]，[Z_1, Z_2]和[Z_0, Z_1, Z_2]。对将以上的特征组合方式输入到基于卷积神经网络的音乐转录模型中，分别计算最终的精确度、召回率、F1 值。得到的结果如表 1 和图 4，所有结果均为百分比值。

Table 1. Comparison of experimental results based on different features of MAPS dataset
表 1. 基于 MAPS 数据集不同特征的实验结果比较

特征表示	准确率	召回率	F1 值
CQT 频谱	70.91	66.36	68.53
Z_0	68.32	66.65	67.62
Z_1	71.23	61.24	70.05
Z_2	70.57	68.42	69.65
[Z_0, Z_1]	75.25	65.75	70.36
[Z_1, Z_2]	79.23	66.23	70.26
[Z_0, Z_2]	78.12	62.57	71.18
[Z_0, Z_1, Z_2]	**80.78**	**69.27**	**73.31**

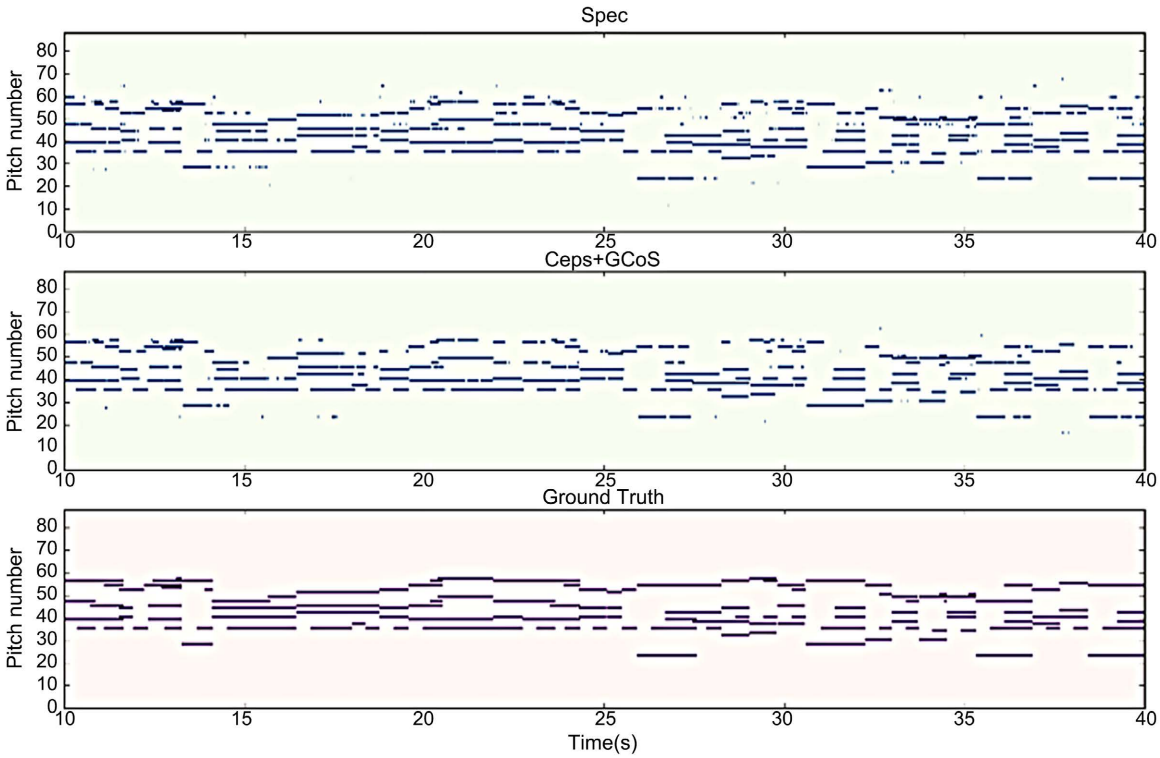

Figure 4. Transcription results of audio segments in MAPS dataset
图 4. MAPS 数据集中的音频截取片段转录结果

如图 4，在截断的数据集上的整体表现也与主流的转录模型的效果相当。实验证明，基于组合频谱的实验方式能够提升模型性能，同时能够降低需要的训练时间；基于组合频谱的特征表示方式，有效证明了组合频谱能够降低和弦音符的谐波干扰，提升识别性能。

6. 总结

本文综合利用钢琴基本乐理、声音信号处理、音乐声学模型、深度神经网络和自动音乐转录等知识，对当下的基于深度学习的钢琴音乐转录方法进行了深入地探讨。针对目前钢琴的多音级估计方法中存在的音符之间的谐波干扰问题，本文提出了一种多重频谱特征表示方法，该方法采用多个通道来共同表示基音的谐波结构和时间频率信息。提出的基于组合频率周期的多特征表示方法在深度卷积网络模型中的整体转录表现优于单一的频谱特征。三种频谱特征组合达到了最优的性能，同样也证明了多特征的表示对于转录的结果是有利的，组合数据的输入在深度学习中具备更高的灵活性。

在自动音乐转录领域中，演奏者与钢琴音乐的交互分为很多方面，本文的研究工作重点只关注了钢琴音乐的音高和时值的转录，除此之外，还包括旋律、节奏等音乐方面的特征，甚至包括演奏者的情绪动态等非音乐性的特征等等，这些方面将是后续转录工作的研究重点内容。

基金项目

湖北省教育厅科学研究计划资助(D20192401)。

参考文献

[1] Sigtia, S., Benetos, E. and Dixon, S. (2015) An End-to-End Neural Network for Polyphonic Piano Music Transcription. *IEEE/ACM Transactions on Audio Speech & Language Processing*, **24**, 927-939. https://doi.org/10.1109/TASLP.2016.2533858

[2] Kelz, R., Dorfer, M., Korzeniowski, F., *et al*. (2016) On the Potential of Simple Framewise Approaches to Piano Transcription. *Proceedings of the* 17*th International Society for Music Information Retrieval Conference*, New York City, 475-481.

[3] Su, L. (2017) Between Homomorphic Signal Processing and Deep Neural Networks: Constructing Deep Algorithms for Polyphonic Music Transcription. 2017 *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (*APSIPA ASC*), Kuala Lumpur, 12-15 December 2017, 884-891. https://doi.org/10.1109/APSIPA.2017.8282170

[4] Su, L. (2018) Vocal Melody Extraction Using Patch-Based CNN. 2018 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Calgary, 15-20 April 2018, 371-375. https://doi.org/10.1109/ICASSP.2018.8462420

[5] Hawthorne, C., Stasyuk, A., Roberts, A., *et al*. (2018) Enabling Factorized Piano Music Modeling and Generation with the Maestro Dataset.

[6] Jansson, A., Humphrey, E., Montecchio, N., *et al*. (2017) Singing Voice Separation with Deep U-Net Convolutional Networks. *Proceedings of the* 18*th ISMIR Conference*, Suzhou, 23-27 October 2017, 745-751.

[7] Chen, L.C., Papandreou, G., Kokkinos, I., *et al*. (2018) DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **40**, 834-848. https://doi.org/10.1109/TPAMI.2017.2699184

[8] He, K., Gkioxari, G., Dollar, P. and Girshick, R. (2017) Mask R-CNN. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 2980-2988. https://doi.org/10.1109/ICCV.2017.322