Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2022, 11(4), 348-355
Published Online November 2022 in Hans. http://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2022.114037

Hans汉斯

人工智能语义分割技术在钢琴教育中的应用研究

胡丽敏[1]，桂　浩[2]，陈开一[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22日；录用日期：2022年10月19日；发布日期：2022年10月28日

摘　要

目前，素质教育越来越被重视，作为素质教育代表的音乐教育也越来越被关注，但是音乐教育却极大受限于人工教育资源。人工智能在音乐教育中的辅助，从计算机的角度讲就是信号类型转换的过程。例如对于学者弹琴，需要将钢琴的信号转换特定的数字信号与真实的谱子进行对比纠错，从而识别错音、错节奏的现象并实时校正。这一规范技术过程被称为自动音乐转录AMT (Automatic Music Transcription)。本文采用谐波常数Q变换、CFP等不同的音乐数字特征表示方法，将原始的音乐信号转换为频谱图，作为网络结构的特征输入，改进了语义分割模型DeepLabv3+，融合了U-Net的U型结构对多乐器音乐进行转录，该算法在钢琴音乐MPAS数据集上达到了良好的识别效果。

关键词

人工智能，自动音乐转录，钢琴教育，语义分割

Application of Artificial Intelligence Semantic Segmentation Technology in Piano Education

Limin Hu[1], Hao Gui[2], Kaiyi Chen[2]

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

Received: Aug. 22nd, 2022; accepted: Oct. 19th, 2022; published: Oct. 28th, 2022

Abstract

At present, quality education is more and more valued, and music education as a representative of

文章引用: 胡丽敏, 桂浩, 陈开一. 人工智能语义分割技术在钢琴教育中的应用研究[J]. 人工智能与机器人研究, 2022, 11(4): 348-355. DOI: 10.12677/airr.2022.114037

quality education is also more and more concerned. But music education is greatly limited by artificial educational resources. The help of artificial intelligence in music education is the process of signal type conversion from the perspective of computer. For example, for scholars to play piano, it is necessary to convert the piano signal to a specific digital signal and compare it with the real spectrum to correct errors, so as to identify the phenomenon of wrong sound and wrong rhythm and correct it in real time. This standardized technical process is called Automatic Music Transcription (AMT). The algorithm comprehensively makes use of digital feature representation methods such as harmonic constant Q transformation and CFP. It converts the original music signal into a spectrum chart as a feature input of the network structure. It improves semantic segmentation model DeepLabv3+ and incorporates U-Net's U-shaped structure to transcribe multi-instrument music. The algorithm achieves good performance on piano music MPAS datasets.

Keywords

Artificial Intelligence, Automatic Music Transcription, Piano Education, Semantic Segmentation

1. 引言

近年来，我国素质教育中最具代表性的音乐教育越来越被社会所关注。国内教育资源尤其是专业的指导老师数量供不应求，故此广大的学者只能局限在时间有限的课堂中得到专业指导，而在多数练习时间中却得不到专业的矫正。在素质教育迅速发展的今天，数量庞大的学者需要得到更关切、更正确、更直接的辅助性指导。课上受到指导的时间由一节课的课时所限制，而在课外需要长时间、大量地练习。在练习过程中，他们需要一个强有力的“指导老师”来解决他们存在的问题，譬如错音、错节奏、弹琴指法规范性等等。目前，机器学习在声学领域取得了良好发展。

广大的学者如何能够从人工智能的辅助与监督中得到规范而严格的指导是一项很有研究意义的课题。利用人工智能来进行音乐辅助练习，其实就是将声学信号转换为数字信号后，由 AI 模型进行分析的过程。对于学者弹琴，需要将钢琴的信号转换成特定的数字信号格式与真实的谱子进行对比纠错，从而识别错音、错节奏的现象并实时校正。这一规范技术过程被称为自动音乐转录(Automatic Music Transcription, ATM)或自动演奏识别。作为音乐信息检索中的基本问题，ATM 能够为哼唱识别，乐曲匹配提供有力的技术支撑。ATM 包括 F0 检测[1]、多音级估计、音符起止点检测等。

在单音音乐转录中，机器学习在音高识别方向取得了重大突破。但是对于多音音乐转录中的音高识别和乐器匹配问题，仍在研究探索阶段。本文的研究重点就是乐器中音高识别和匹配问题。

2. 基于语义分割网络的自动音乐转录方法

2.1. 方法概述

本文提出的模型如图 1 所示。

语义分割网络模型首先将输入的原始音频数据经过数字信号处理变换为频谱图，组合频谱图(CFP 与 HCQT)经过数据预处理后提取时频域信息和谐波信息。在输入数据的特征中，长宽分别代表着时域和频域维度，通道数则代表着谐波信息。网络模型基于 DeepLabV3 以及 DeepLabV3+中提出的全卷积网络结

构和编码–解码器结构(Encoder-Decoder Block)。原始的 DeepLabV3 只处理单一图片并进行像素密集分类，一般输出在单通道上的二分类，属于单任务模式。为了能够识别多乐器特征，本文将输出映射到 N + 1 个维度上，以识别 N 种不同类型的乐器并附带有一个乐器无关通道。因此对于前 N 个维度，每个维度都代表着一种乐器的转录结果。所以该模型能够同时预测多维度乐器输入。输出在每个通道上的像素值基于 0 到 1 之间，代表着当前音高在本乐器中是否活跃。

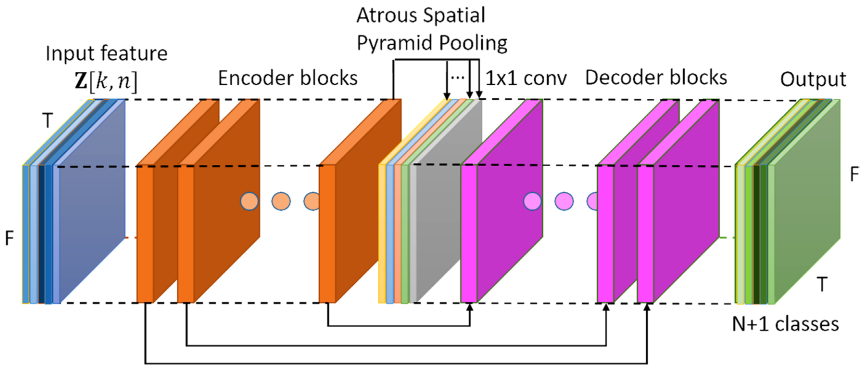

Figure 1. Semantic segmentation network model
图 1. 语义分割网络模型

2.2. 音乐特征提取

音频数据需要经过预处理过程才能作为网络结构的输入。本文的数据特征表示基于文献[2]提出的组合频谱 CFP 和文献[3]提出的 HCQT 变换。文献[2]已经证明了时频域中的组合多特征表示方法能够有效减小谐波干扰，提高 ATM 的最终预测性能。此外，为了更好地捕获谐波信息，本文还借鉴了文献[2] [3]中提出的沿着通道维度方向叠加谐波信息的特征表示办法 HCQT 变换。

假设输入的音频数据经过短时傅里叶变换后得到得特征矩阵为 $X \in R^{F*T}$，其中 F 代表着频谱图中有效频率范围的频域维度，T 代表着时间维度。考虑公式(1)和(2)中数据特征表示办法。

$$Z_f[k,n] = Q_f \left| W_f X \right|^{\gamma_f} \tag{1}$$

$$Z_q[k,n] = Q_q \left| W_q F^{-1} Z_f \right|^{\gamma_q} \tag{2}$$

其中 Z_f，$Z_q \in R^{F*T}$，k，n 分别是数组中时间维度和频率维度的索引，矩阵 F^{-1} 是离散傅里叶变换的逆变换矩阵，W_f 和 W_q 是两个高通滤波器，高通滤波器能够丢弃低变部分，γ 是元素级的幂阶 ReLu 激活函数，即 $|x|^{\gamma} = |\mathrm{ReLu}(x)|^{\gamma} = |\max(0,x)|^{\gamma}$。本文基于文献[4]的设定 $(\gamma_f, \gamma_q) = (0.24, 0.6))$。$Q_f, Q_q \in R^{F*2F}$ 是两个三角滤波器组，它们分别实现时域和频域到有效对数频域的特征映射。两个滤波器组都包含 352 个，每八度 48 个半音的三角滤波器，覆盖范围为 27.5 Hz (A0)到 4487 Hz (C8)。因此，最终的频率表示中，频域范围 F = 352，时域范围 T 根据输入的音频长度确定。简单来讲，Z_f 代表的是原始音频数据的基频和基频高次谐波的幂阶频谱图(Spectrum)，而 Z_q 代表的是基频和基频低次谐波的广义倒谱图，文献中指出，对于多音级识别模型来说组合 Z_f 与 Z_q 能够明显地提供比频谱图更好的性能[5] [6]。

2.3. 语义分割转录模型

模型结构如图 1 所示，模型具体参数见图 2。输入数据特征为 $Z_{HCFP} \in R^{2m*F*T}$ 其中，$2m$ 代表着模型通道数，在 $HCFP$ 中表示即是谐波分量，F 和 T 分别代表频域和时域，本文中采用的时间步长为 256，频域跨度为 352。

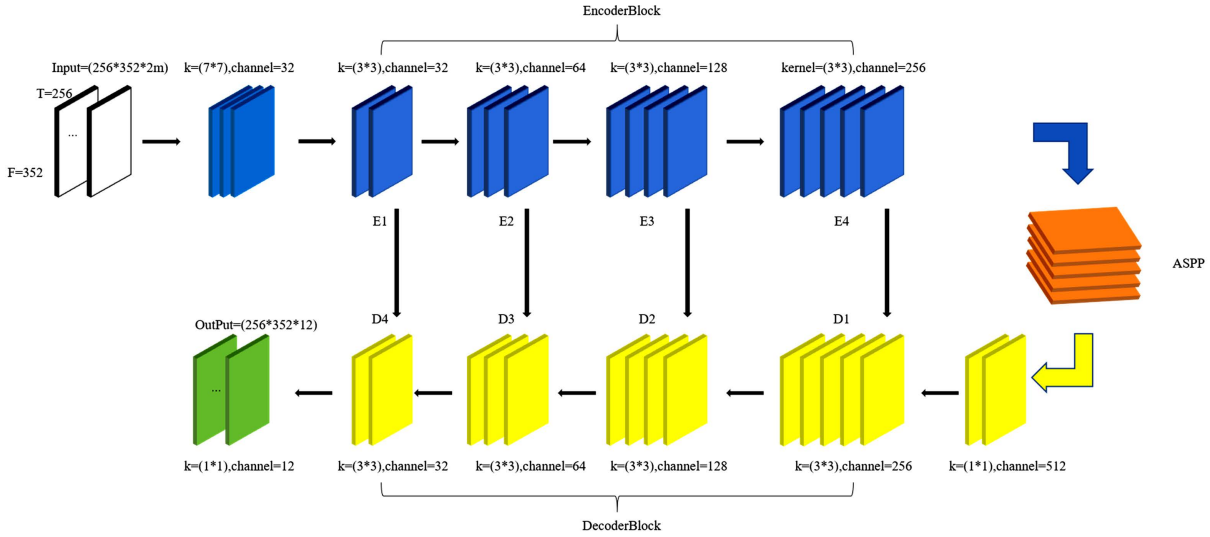

Figure 2. Semantic segmentation network model details
图 2. 语义分割网络模型细节

输入特征首先经过一层尺寸为 7*7 的卷积层，扩展通道数到 32，然后输入到编码器序列。编码器总共有四个模块 E1、E2、E3 和 E4，每个模块都由若干层编码块组成，编码块中包含两个空洞卷积层和一个可选的跳跃全连接层[7]，编码块的结构如图 3 所示。

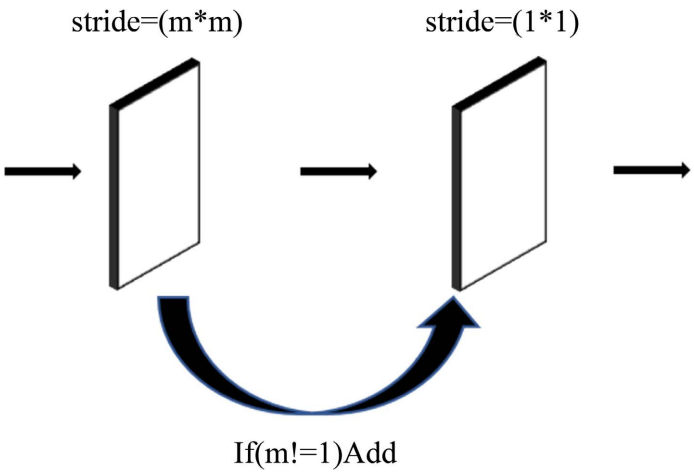

Figure 3. Encoding block structure
图 3. 编码块结构

编码块中的两层空洞卷积层使用的尺寸均为 3*3，考虑音频数据中低维数据在频率和时间维度上都很敏感，即每个音符在图像中可是被视作为小物体，因此采用的膨胀率均为 1。当第一个空洞卷积层的步长不为 1 的时候，跳跃连接层就会被激活。输入音频经过一层卷积层后 featuremap 变为 256*352*32。进而输入到编码器 E1。E1 中通道数均为 32，第一个编码块步长为 2，跳跃连接层被激活，第二个编码块步长为 1，经过 E1 后特征大小、维度都不变。E2 中通道数均为 64，第一个编码块步长为 2，跳跃连接层被激活，后续编码块步长均为 1，音频特征经过第二层编码块后 featuremap 变为 256*352*64。类似地，经过 E3、E4 后，音频特征大小最终被编码为 256*352*256。编码器提取了音乐的高纬度特征，降低了音频数据的时间和频率分辨率。

音乐特征经过编码器之后输入到 ASPP 结构中，用于音乐转录模型的 ASPP 结构如图 4 所示。首先经过 1*1*512 卷积层将音频特征升维到 512，经过四层膨胀率为 1，2，4，8 的膨胀卷积层后并联在一起，再经过一层 1*1*512 的卷积层输出到解码器结构。ASPP 结构消除了全连接层特征大小的限制，简化了网络结构的设计，尤其是对针对类似音乐这种长度不固定的特征。对于音乐数据当模型一步步进行下上采样过程的时候，频率分辨率也会越来越小，因此当频率分辨率过小的时候采用较大膨胀率的空洞卷积仍然能够提取原始音频数据的局部特征信息，从而更好地进行时频域分离，来达到预测不同乐器种类中，音高的识别效果。

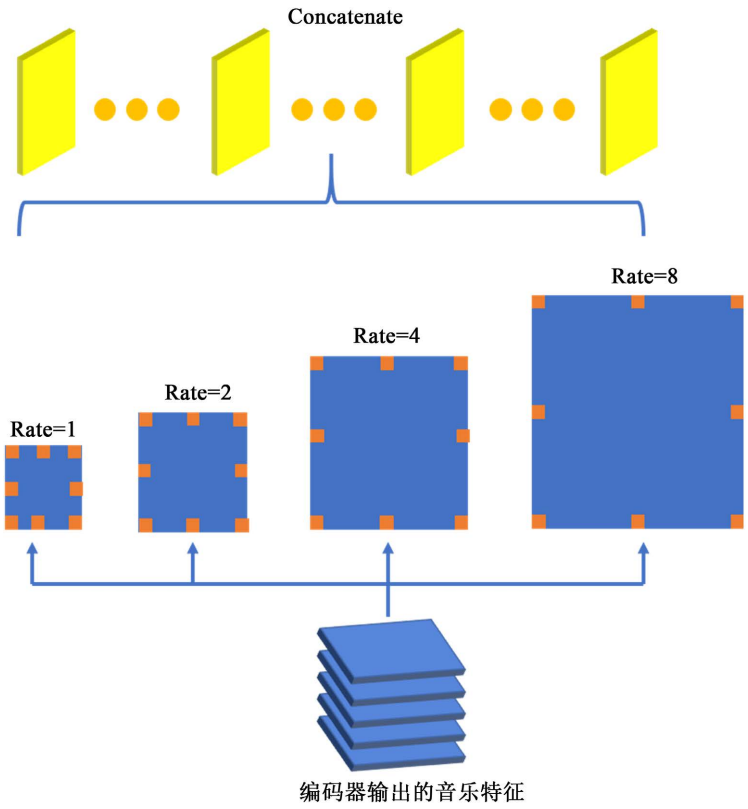

Figure 4. ASPP structure for music transcription model
图 4. 用于音乐转录模型的 ASPP 结构

解码器也包含四个模块 D1、D2、D3 和 D4，每个模块都包含一个解码块。解码块与编码块结构完全一样，只不过第二个卷积被替换成为了转置卷积。经过 ASPP 结构处理后的音乐特征大小为 256*352*512。经过一层 1*1 的卷积核后降维到 256，然后合并编码块中 E4 的特征输入到 D1 中。D1 中解码块的通道数为 256，经过解码块之后并联 E3 的特征，经过一层大小为 1*1，通道数为 256 的卷积层后跳跃连接解码块之前的音频特征。D2、D3 和 D4 的结构与 D1 一模一样，通道数逐渐递减为 128、64 和 32，最后再经过一层通道数为 12 的卷积层输出分类结果。解码器从物理意义上是对音频数据的上采样，逐层恢复音频数据的时间和频率分辨率。

编码器和解码器之间借鉴了 U-Net 结构，但是对像素敏感类型的音乐特征信号，在频域中一个微小的改变都会对最终音高预测的结果产生巨大影响，这是因为在变换过程中，时域到频域是经过对数变换处理了的，当从频域恢复到时域就变成了指数变化。这里的跨步连接能够将低维的高频率分辨率信息特征保留下来，输入到高纬度特征中，这样一来低维的高频率分辨率信息和多维度信息就都被保留下来了。

为了防止过拟合现象，本文模型中在卷积层之后均对数据进行了正则化处理和 Dropout 层，其中失活率 Dropout 为 0.4。模型统一用 ReLu 激活函数。

最后模型采用焦距损失函数(Focal Loss Function) [8]来解决最后预测中的分类不平衡的问题[9]。由于输出的结构是稀疏矩阵，即当前时间激活的音高少而未激活的音高多，因此导致大多数像素值为 0，传统的损失函数，例如二元交叉熵，模型最终将会倾向于将大部分像素值预测为 0。焦距损失函数能够提供一个权重因子，用于平衡激活和未激活音符之间的比例协调性，因此焦距损失函数更有助于声乐旋律的提取[7]。

3. 实验与分析

3.1. 实验数据集和环境

1) 实验数据集

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 Z_{CFP} 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。

2) 实验环境

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 10997 MB。代码运行环境为 Python 3.7，Tensorflow1.13，CUDA9.0。

3.2. 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。

3.3. 实验过程和结果分析

多音级估计实验(Conventional MPE，以下统称 C-MPE)，讨论音高识别的准确性，实验在 MPAS 数据集上进行；

将 CFP 特征表示方法作为中提出的模型的输入，评估 C-MPE 在 MPAS 数据集中的表现性能。实验结果见表 1 和图 5。

F-score 能够反映精度和召回率的调和平均，是体现模型性能的平均指标。本文提出的模型在 MAPS 数据集上的 F-score 达到了 87.92%，高出了文献[10] 6.5%。

为了获取更直观的效果，本文随机选取了 MAPS 数据集中 ENSTDkCl 目录下的一首片段音乐 MAPS_MUS-bk_xmas5_ENSTDkCl.wav，截取了其中 10~30 s 的片段，输入到训练好的模型并将结果可视化，如图 6 所示。其中，蓝线代表的是 TP，绿线代表的是 FP，红线则代表了 FN。

Table 1. Experimental results of C-MPE based on MAPS
表 1. 基于 MAPS 数据集的 C-MPE 实验结果

方法	精度 P	召回率 R	F-score
文献[9]	91.93	75.21	81.44
Z_{CFP}	87.52	87.74	87.92

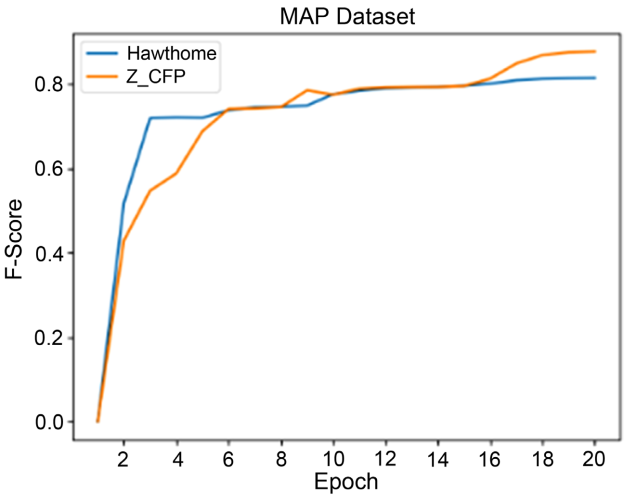

Figure 5. C-MPE experimental training process based on MAPS
图 5. 基于 MAPS 的 C-MPE 实验训练过程

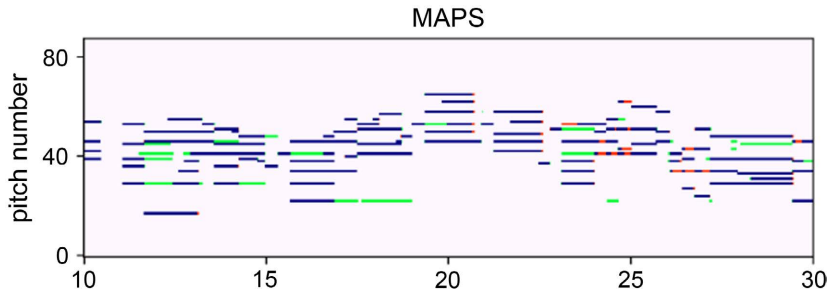

Figure 6. Model prediction results
图 6. 模型预测结果

对于 C-MPE 实验结果，即使钢琴演奏者们使用了延音踏板，该首钢琴曲中大多数音符都能够被正确识别，并且模型能够比较准确地捕捉它们的起始和结束时间。

4. 总结

本文基于谐波常数 Q 变换和组合频谱，改进了 DeepLabV3 语义分割网络，提出一种新的语义分割模型用于转录钢琴复调音乐，取得了良好的效果。实验证明，该方法有效解决了复调音乐产生的谐波重叠与干扰，填补了目前复调音乐转录的空白。

钢琴教育是一种特殊技能教育，家长不能领会其中的教学内容和教学意义，制约了钢琴教学、钢琴练习过程中的规范性，无法实现监督，阻碍了钢琴教育的进一步普及，弹奏练习者需要更多的辅助性指导、陪伴性关怀。

本文深入研究了如何使用人工智能技术，让计算机真正听清楚、听懂弹奏者的钢琴练习，识别和指出“错音”、“错节奏”等常见问题，分析其产生的原因，并为弹奏练习者提供个性化的、陪伴性学习指导。

本文目前涉及到音级估计和乐器匹配，将来还可以考虑更多的音乐特征，比如旋律提取、情感分析等，一方面可以进一步提高转录性能，另一方面为弹奏者的练习提供更丰富的辅助内容和形式。人工智能已经开始打开一个全新的世界，从机器学习、模式识别算法，到基于智能专家系统自动分析推理的反馈式、交互式学习模型，钢琴教育也将从这种新技术所带来的效率提升中受益。

基金项目

湖北省教育厅科学研究计划资助(D20192401)。

参考文献

[1] Klapuri, A. (2006) Introduction to Music Transcription. In: Klapuri, A. and Davy, M., Eds., *Signal Processing Methods for Music Transcription*, Springer, Boston, MA, 3-20. https://doi.org/10.1007/0-387-32845-9_1

[2] Wu, Y., Chen, B., Su, L., *et al*. (2018) Automatic Music Transcription Leveraging Generalized Cepstral Features and Deep Learning. *International Conference on Acoustics, Speech, and Signal Processing*, Calgary, 15-20 April 2018, 401-405. https://doi.org/10.1109/ICASSP.2018.8462079

[3] Sigtia, S., Benetos, E. and Dixon, S. (2020) 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

[4] Peters, G. (2006) Music Pitch Representation by Periodicity Measures Based on Combined Temporal and Spectral Representations. *International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, 14-19 May 2006, 53-56.

[5] Su, L. and Yang, Y. (2015) Combining Spectral and Temporal Representations for Multipitch Estimation of Polyphonic Music. *IEEE Transactions on Audio, Speech, and Language Processing*, **23**, 1600-1612. https://doi.org/10.1109/TASLP.2015.2442411

[6] 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

[7] Lu, W.T. and Su, L. (2018) Vocal Melody Extraction with Semantic Segmentation and Audio-symbolic Domain Transfer Learning. *Proceedings of the* 19*th ISMIR Conference, Paris, France, September* 23-27, 2018, 521-528.

[8] Thickstun, J., Harchaoui, Z., Foster, D., *et al*. (2018) Invariances and Data Augmentation for Supervised Music Transcription. *International Conference on Acoustics, Speech, and Signal Processing*, Calgary, 15-20 April 2018 2241-2245. https://doi.org/10.1109/ICASSP.2018.8461686

[9] Chen, L.-C., Papandreou, G., Schroff, F., *et al*. (2021) Rethinking Atrous Convolution for Semantic Image Segmentation.

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